Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Erlend Øien

Enhancing Online Lecture Recommendations Through Exploration of Student Behaviour Across Topics

Master's thesis in Computer Science
Supervisor: Özlem Özgöbek
June 2023

NTNU
Norwegian University of
Science and Technology

Erlend Øien

Enhancing Online Lecture Recommendations Through Exploration of Student Behaviour Across Topics

Master's thesis in Computer Science
Supervisor: Özlem Özgöbek
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

The availability of free, quality online education has increased globally; however, learner retention and dropout rates remain concerning in Massive Open Online Courses (MOOCs). Successful learners often exhibit self-structured and controlled learning, demonstrated by their more frequent review of course material as part of Self-Regulated Learning (SRL) strategies. Despite its importance, learners do not generally revisit course material frequently. As video lectures have been an important learning resource for Technology-Enhanced Learning (TEL), how users interact with the videos has been of interest. The viewing behaviours have been shown to correlate with student performance, dropout rates and perceived difficulty. However, research has mainly considered only a few educational topics, often related to Computer Science (CS), leaving questions regarding the general applicability in-video viewing behaviour unanswered. The narrow topic focus has largely been the case for learning resource Recommendation Systems (RSs) as well, where considering contextualised, sequential student behaviour is left unexplored.

Therefore a Sequence Aware Recommendation System (SARS) which considers both in-video viewing behaviour and learning resource topics is proposed for improving online lecture recommendation quality across educational domains. Furthermore, an analysis of lecture-*re-consumption* across diverse topics is conducted. In particular, the predictive power of in-video viewing behaviour for predicting lecture revisits is explored. Lastly, to which degree Recommendation Systems are aligned with users' re-consumption behaviour is examined to quantify to which extent recommendations may influence reviewing behaviour. These experiments establish a statically significant improvement in lecture recommendation accuracy using in-video viewing behaviour and related lecture topics. Furthermore, the analysis of revisiting behaviour indicates the importance of considering lectures' intrinsic properties for viewing behaviour, where the results highlight differences in in-video viewing and revisiting behaviour across educational topics. However, re-consumption prediction using in-video viewing features shows promise. Lastly, the novel re-consumption alignment analysis illustrates sequence-aware models' ability to distinguish individual re-consumption behaviour, providing a more personalised and calibrated learning experience.

Sammendrag

Tilgangen til gratis, nettbasert utdanning av høy kvalitet har økt globalt, men i MOOCs er det fortsatt bekymringsfullt at mange elever faller fra undervies. Studenter som lykkes, lærer ofte på en selvstrukturert og kontrollert måte, noe som kommer til uttrykk ved at de oftere går gjennom læringsmaterialet som en del av sine Selv-Regulerte Læringsstrategier. Til tross for at repitisjon er viktig, er det ikke vanlig at studentene repeterer læringsressursene så ofte. Ettersom video forelesninger har vært en viktig læringsressurs for teknologi-basert læring (Technology-Enhanced Learning (TEL)), har det vært av interesse hvordan brukerne interagerer med videoene. Deres video-interaksjonsmønstre har vist seg å korrelere med studentenes prestasjoner, frafall og opplevd vanskelighetsgrad. Forskningen har imidlertid i hovedsak bare tatt for seg noen få emner, ofte relatert til datateknologi, noe som åpner for spørsmål angående video-interaksjonsmønstres generelle anvendelighet. Det smale emnefokuset har i stor grad også vært tilfelle for anbefalingssystmer (RS) for læringsressurser, der kontekstualisert, sekvensiell studentatferd ikke er blitt utforsket.

Derfor foreslår denne oppgaven et anbebalingssystem modellert på sekvensiell data (Sequence Aware Recommendation System (SARS)) og som tar hensyn til både video-interaksjonsmønstre og temaene diskutert i video forelesninger, for å forbedre kvaliteten på nettbaserte forelesningsanbefalingessystemer på tvers av forskjellige emner. I tillegg ble repitisjon av video forelesning analysert på tversav på tvers av ulike emner. Konkret undersøkes video-interaksonsmønstre sin evne til å predikere hvorvidt en videoforelesning vil bli sett på nytt eller ikke. Til slutt undersøkes til hvilken grad anbefalingssystemer stemmer overens med brukernes repitisjonsmønstre, for å kvantifisere i hvilken grad anbefalinger kan påvirke repitisjonsatferd. Eksperimentene viser en statistisk signifikant forbedring i nøyaktigheten av forelesningsanbefalingene ved hjelp av video-interaksjonsdata og relaterte forelesningstemaer. Videre viser analysen av repitsjonsatfer at det er viktig å ta hensyn til forelesningers iboende egenskaper når det gjelder video-interaksjonsatferd, og resultatene fremhever forskjeller i seer- og gjenbesøksatferd på tvers av temaer. Prediksjon av repitisjon basert på vidoe-interaksjonsmønstre er imidlertid lovende. Til slutt illustrerer den innovative analysen av anbefalingssystemers kalibreringsevne for repitisjonspreferanser at sekvens-baserte modeller kan skille mellom studentenes individuelle repitisjonsmønstre, noe som gir en mer personlig tilpasset og kalibrert læringsopplevelse.

Preface

This research project and the resulting report is submitted to *Norwegian University of Science and Technology (NTNU)* as a master's thesis for the course *TDT4900 - Computer Science, Master's Thesis*, amounting to 30 credits. The project's outline, direction and progress were supervised by associate professor Özlem Özgöbek at the Department of Computer Science (IDI) at NTNU. The research and corresponding findings are partially related to my preliminary work [1].

I would like to thank Özlem for encouraging me to find an area of research which I think is exciting and for many hours of brainstorming and advice provided over the course of the project. To my fellow, study hall neighbours - You've been an encouragement and a delightful distraction throughout the year. Last, but not, least, to my better half and housekeeper - Thank you for keeping me grounded, tolerating my absent-mindedness and reminding me of what's important in life. Yes, we did in fact make it!

Contents

Figures

Tables

Acronyms

AP Average Precision

BERT Bidirectional Encoder Representations from Transformers

BPR Bayesian Personalised Ranking

BST Behaviour Sequence Transformer

CF Collaborative Filtering

CLM Causal Language Modelling

CS Computer Science

CTR Click-Through Rate

Exp.1 Experiment 1 - Next-Lecture Prediction

Exp.2 Experiment 2 - Re-consumption Behaviour

Exp.3 Experiment 3 - Alignment

FFN Feed-Forward Network

FN False Negative

FP False Positive

FWER Family Wise Error Rate

GCN Graph Convolutional Network

GHSS Government, Health & Social Science

GRU Gated Recurrent Unit

HHRDE Humanities, History, Religion, Design & Education

iALS implicit Alternating Least Squares

KL Kullback-Leibler

KNN K-Nearest Neighbours

KT Knowledge Tracing

KU Knowledge Unit

LA Learning Analytics

LMF Logistic Matrix Factorisation

LOO Leave-One-Out

LSTM Long Short-Term Memory

MAP Mean Average Precision

MF Matrix Factorisation

MLM Masked Language Modelling

MOOC Massive Open Online Course

NDCG Normalised Discounted Cumulative Gain

PCA Principal Component Analysis

PLA Predictive Learning Analytics

RNN Recurrent Neural Network

RS Recommendation System

SARS Sequence Aware Recommendation System

SOHE Soft One-Hot Encoding

SPM Sequence Pattern Mining

SPOC Small Private Online Courses

SRL Self-Regulated Learning

STEM Science, Technology, Engineering & Math

SVD Singular Value Decomposition

SVM Support Vector Machine

TEL Technology-Enhanced Learning

TN True Negative

TP True Positive

TPE Tree-Structured Parzen Estimator

ZPD Zone of Proximal Development

Chapter 1

Introduction

This chapter firstly presents the general background for the research, as well as the motivation in Section 1.1, before presenting the resulting main research goal and the related research questions in Section 1.2. Following it, the thesis' research approach is described in Section 1.3, before the main contributions are highlighted in Section 1.4. Concluding the chapter, an outline of the thesis is presented in Section 1.5.

1.1 Background and Motivation

As the internet became increasingly available, so did information and the possibilities for digital education. The idea of democratised and openly available education was quickly formed in the early 2000s, and later formalised as MOOCs in 2008 [2]. Despite the intention of enabling quality education to more people, some barriers to adoption of MOOCs have been language, accessibility and cost. Consequently, most MOOC users are existing students of higher education in developed countries [2–4]. Moreover, MOOCs have had low completion rates [5], also compared to other e-learning options [6]. The exact retention rates differ but are reported to be below 10% [6, 7]. Some of the reported reasons for course dropout have been related to the lack of available instructions on how to use MOOCs [4] in addition that the learning environment is not personalised to the individual learner [6]. Furthermore, users have reportedly widely different intentions when enrolling in a MOOC, from exploring the field to taking a certification [5, 6]. These diverse user intentions require a more personalised learning environment, as they will have different platform usage patterns. A potential solution for overcoming the lack of instructions and offering a more personalised learning experience is to use a Recommendation System (RS) for learning resource recommendations. As educational videos and lectures are important sources of information in e-learning environments [8–10], recommending relevant lectures to the users can improve navigational efficiency and overall learning experience.

To better understand user preferences, lecture interactions and general learning behaviour, in-lecture viewing interactions have been of interest for over two decades [11]. Moreover, different types of viewing behaviour such as the number of rewinds or the total time spent on the video, have been useful for measuring student engagement [12, 13], predicting student performance [14–16], measuring perceived difficulty [17, 18] and cognitive load [19], as well as both course and in-lecture dropout [20, 21]. Although they have been used to some de-

gree in RSs for recommending *learning resources* such as exercises or lectures [22], the wide range of viewing behaviours and their relations in that regard. Furthermore, modelling users' sequential nature of interactions with learning resources has been promising for RS [23], but granular user behavioural features have not been taken into account. Moreover a general issue in previous learning resource Recommendation Systems and lecture viewing behaviour; they have only been evaluated on a narrow set of homogeneous domains, most often computer science related [24] or a small set of users and learning resources. Additionally, some studies have illustrated users' behavioural viewing preferences [25, 26] and relation to physical lecture properties [17, 27], but their relation to the intrinsic lecture topics and domain properties have largely been left unexplored. Moreover, backward seeks in lectures have intuitively been associated with in-video replay and revision, but it has also been shown to relate to more concrete *frame seeking* behaviour [18], which illustrates a different user intention. Therefore a more nuanced viewing behaviour measurement of in-video re-consumption is needed to improve the understanding of user intentions.

Within MOOCs, Self-Regulated Learning (SRL) strategies have been found to be crucial for student performance, retention and personal goal attainment [28, 29]. Furthermore, poor SRLs skills have been identified through surveys as a hindrance for MOOC adoption [4, 29]. *Revisting* learning resources is deemed important on an educational psychology level for knowledge retention and long-term learning [30]. Additionally, revision is a common manifestation of SRL strategies [28], where in particular, lecture re-consumption is positively correlated with most identified SRL strategies [9, 28] as well as course completion [31]. Even though re-consumption is arguably important in learning, it has been shown to be infrequent in some Technology-Enhanced Learning (TEL) scenarios [26]. So the relation between SRL and revisiting behaviour emphasises the importance of understanding the underlying factors in Technology-Enhanced Learning (TEL) for improving the learning experience and increasing MOOC retention. Despite that revisitation has been studied in multiple domains [32], it has only been considered for RSs in more recent years. One key insight of these *re-consumption*-aware RSs studies, is that re-consumption behaviour and its importance for recommendation accuracy is highly domain dependent [32–34]. Moreover, despite that lecture viewing behaviours have been shown to be informative for indicating re-consumption [18], they have not been studied for re-consumption prediction to the author's knowledge. Moreover, viewing behaviour's relation to re-consumption is neither evaluated across a diverse set of educational topics nor corrected for the user and lecture behavioural biases when studied [18].

Lastly, for Recommendation Systems to be useful for users in TEL environments, it should also be able to adapt to their individual *learning style*. One aspect of learning styles is re-consumption behaviour, but little has been researched regarding RSs ability to align to individual users' re-consumption preferences. Some proposed learning resources RSs have modelled explicit re-consumption behaviour in their systems [22, 35], but little work has been done to measure the alignment of them, or of other re-consumption-aware RSs. This is of interest for recommending more relevant items regarding users' preferences, in addition to providing potential corrective recommendations for improving SRL behaviours.

1.2 Research Goals and Questions

Following the discovered gaps in knowledge, the main research goal of this thesis is first of fall to improve recommendation accuracy in topic-diverse, large-scale educational contexts

to provide a better learning experience for users. Secondly, the research aims to better understand how and why users re-consume educational videos or not. Thirdly, to accommodate for individual users' re-consumption behaviours, the work aims to understand the degree to which various RSs are affected by such behaviour. Moreover, a goal is to accurately measure how well RSs align with users' re-consumption behaviour to understand their potential impact on improving SRL behaviour.

To achieve the mentioned, overarching research goals, three sets of research questions are proposed below.

RQ1 How is general recommendation accuracy affected by different recommendation techniques and the inclusion of side information?

- *RQ1a* How do both conventional RS and SARS recommendation accuracy differ on large-scale educational datasets?
- *RQ1b* How is the recommendation accuracy of a SARS affected by the inclusion of in-video viewing patterns on educational datasets?
- *RQ1c* How do user bias adjustment of behavioural features improve SARS' recommendation accuracy on educational datasets?

RQ2 What factors influence re-consumption behaviour in Technology-Enhanced Learning environments?

- *RQ2a* Which aspects of users' interaction history, lecture topic diversity and in-lecture interaction behaviours are related to re-consumption behaviour?
- *RQ2b* To what extent are viewing patterns predictive for re-consumption behaviour?

RQ3 How do recommendation systems handle re-consumption behaviour and adapt to user preferences?

- *RQ3a* To what degree do different RSs recommend already interacted with items, and does side-information affect it?
- *RQ3b* How do RSs adapt to individual users' re-consumption behaviour without explicit re-consumption preference modelling?

1.3 Research Approach

To properly address the research questions, the proposed approach is to answer them through three sets of experiments, each for the respective set of research questions. The following sections describe the general research and experiment approaches on a high level.

Experiment 1 - Next-Lecture Prediction

Of the identified relevant in-video viewing features and corresponding datasets, preprocessing steps are necessary for feature extraction and data cleaning. Furthermore, as the experiment

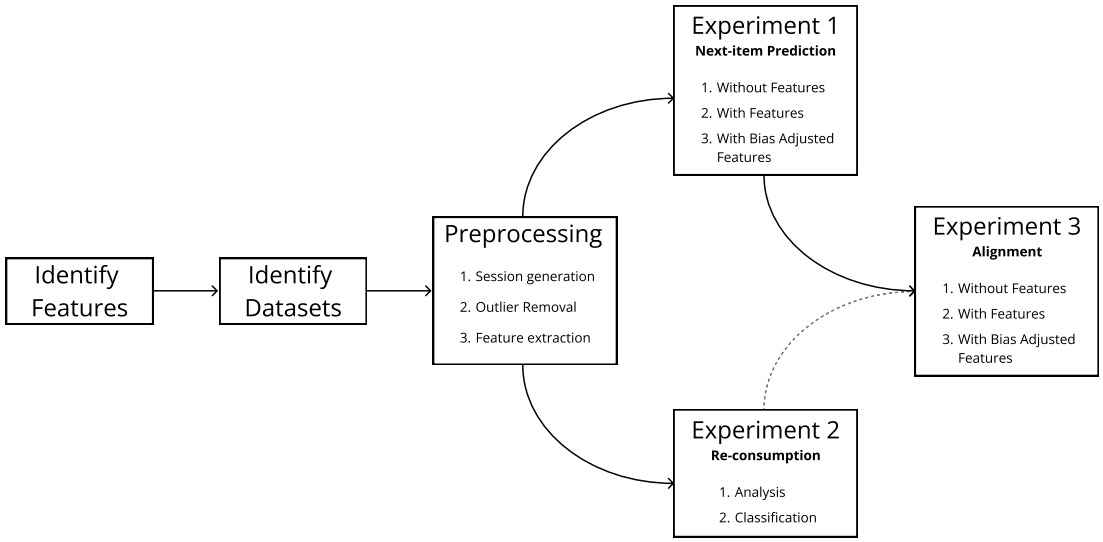


Figure 1.1: Proposed research approach for the thesis.

evaluates recommendation accuracy, the relevant evaluation methodologies and metrics must be decided.

The intention of the first partial experiment, Experiment 1.1, is to evaluate various conventional and Sequence Aware Recommendation System on the identified datasets excluding any side information, to properly address *RQ1a*. The motivation is to first explore how different modelling assumptions impact recommendation accuracy. The recommendation accuracy of the models is in addition compared using statistical significance testing to quantify any meaningful results [36, p.268-272].

Experiment 1.2 addresses *RQ1b*, by evaluating the sequence-aware models from Experiment 1.1 including side information. The individual model variants are then compared to the results from Experiment 1.1 and quantified using statistical significance testing to establish if side information provides an improvement or degradation in recommendation accuracy [36, p.268-272]. The best-performing model is used to search for an optimal feature subset, including both categorical and continuous features. This model's recommendation results are then compared to the equivalent models with all features and without any respectively.

The last partial experiment, Experiment 1.3, addresses *RQ1c*. Instead of using the *raw* viewing patterns in Experiment 1.2 as side-information, Experiment 1.3 includes in-video viewing behaviour features adjusted by users' learning styles, in addition to the categorical, lecture-related features, using the same sequential models as in Experiment 1.2. The recommendation results are then compared to the equivalent models in Experiment 1.2 using statistical significance testing, to address *RQ1c* [36, p.268-272]. Additionally, the results are compared to the equivalent models in Experiment 1.1 to further contextualise potential improvements or declines.

Experiment 2 - Re-consumption Behaviour

Experiment 2 - Re-consumption Behaviour (Exp.2) intends to explore in-video viewing behaviours' relations to re-consumption behaviour and to predict revisits of lectures using viewing

behaviour.

Three sub-questions are proposed to collectively address *RQ2a*. The first question (Q1) is: Is re-consumption a substantial fraction of a user's interaction history across topics? The second question (Q2) is: Are there re-consumption frequency differences between topics? The last question (Q3) is: Are In-lecture interaction behaviours statistically different between first-time views and re-consumption views? The sub-questions are to be addressed using various significance testing methods and visualisations to provide quantifiable arguments for addressing *RQ2a* [36, p.268-272].

The other aspect of the relationship between in-video viewing and re-consumption behaviour is if the viewing features can be used to predict re-consumption whether or not a user will revisit the lecture. To tackle the related research question *RQ2b*, one must first provide a balanced dataset of user-lecture interactions, labelled according to the definition of a re-consumption. Then various classical and state-of-the-art classification models can be fitted and evaluated on the dataset.

Experiment 3 - Alignment

The last set of experiments is structured to answer various Recommendation Systems' recommendation accuracy and alignment related to individual users' re-consumption behaviour.

Based on the various models' recommendations made in Experiment 1 - Next-Lecture Prediction (Exp.1), the partial Experiment 3.1 rather evaluates how inclined the individual models are to recommend lectures the user has already viewed. The motivation is to provide a baseline of models to which degree they are inclined to recommend already viewed lectures versus *novel*, unseen lectures. Therefore, the individual users' viewing history is considered as the relevant item set, to address **RQ3a**. The recommendation accuracy of users' interaction history is evaluated using traditional ranking metrics and compared using statistical significance testing to establish significant differences [36, p.268-272].

Building upon the results in the previous experiment, to address **RQ3b**, users with existing re-consumption behaviour are more closely examined. The motivation is that if a user indicates a preference towards re-consuming a lecture, the algorithms should be able to adapt to that user's re-consumption preference. This alignment problem relates to the issue of *calibrated recommendations*, either through re-ranking efforts [37] or explicit re-consumption parameters or objectives [22, 35]. By evaluating the recommendations made to users with existing re-consumption behaviours, one can measure the *native* calibration of the proposed methods without re-ranking. Further on, the effect of side information on calibration is evaluated. Therefore this experiment utilises a subset of the recommendations made in Exp.1 and applies calibration-related evaluation methods to them, with statistical significance testing of the results [36, p.268-272].

1.4 Contributions

This research project makes contributions to multiple aspects of TEL, including Learning Analytics (LA), learning resource recommendation and re-consumption behaviour in education. The contributions are mainly derived from the proposed experiment groups. Briefly summarising the main contributions, a new sequence-aware model which incorporates both user in-

video behavioural features and item features are proposed in the first experimental group, Exp.1. Moreover, a time-aware behaviour bias correction approach is proposed for correcting users' in-video viewing behaviour. Further on, Exp.2 contributes to the understanding of re-consumption behaviour across different educational topics, and the utility of in-video viewing behaviour for predicting lecture re-consumption.

Additionally, Experiment 3 - Alignment (Exp.3) contributes to the understanding of how RSs' recommendations are affected by re-consumption behaviour in educational datasets and their degree of alignment to individual re-consumption behaviours. In that regard, a novel approach for mapping re-consumption alignment to the problem of calibration is proposed, as well as an alternative, explainable re-consumption alignment metric. Moreover, another measurement for in-video replaying behaviour is proposed to better reflect the users' in-video behavioural intentions.

1.5 Thesis outline

Chapter 1 - Introduction The chapter provides the initial motivation and background, the resulting research goals and questions, as well as the research approach and main contributions.

Chapter 2 - Background Theory The chapter introduces topics regarding digital education, educational psychology, Recommendation Systems, calibration and evaluation methods.

Chapter 3 - Related Works An introduction to the related works of side-information inclusion in SARS, as well as studies regarding LA and in-video viewing behaviour, in addition to research on learning resource RSs and re-consumption prediction, recommendation and calibration.

Chapter 4 - Datasets Publicly available learning resource datasets, as well as the reasoning for the chosen datasets, are provided in this chapter. The globally applied preprocessing steps are then described, including data cleaning, feature extraction and visualisation.

Chapter 5 - Method and Experiments This chapter provides the specific methodology for each of the proposed sets of experiments, as well as their results.

Chapter 6.1 - Discussion The research questions are in this chapter addressed through a discussion of the experimental results and their limitations, including limitations of the general methodology and chosen datasets.

Chapter 6 - Conclusion Finally, the conclusion summarises the main results and contributions of the thesis, in addition to proposing areas of future work.

Chapter 2

Background

This chapter gives an introduction to the relevant topics touched by this project. Section 2.1 describes the Technology-Enhanced Learning (TEL) as well Massive Open Online Course (MOOC) and Self-Regulated Learning (SRL). In the following Section 2.2, relevant machine learning models related to classification and sequence modelling are presented, before describing Recommendation System (RS) in more detail in Section 2.3. Lastly, various evaluation methods for traditional classification and ranking problems, as well some relevant statistical tests in Section 2.4.

2.1 Technology-Enhanced Learning

A commonly used term regarding technology-related learning is *e-learning*, but it can be defined in many different ways, on multiple different axes [38]. A more specific term under the larger e-learning umbrella is TEL which describes technologies used to improve the utility of learning and teaching [39]. In a formal, physical educational setting like traditional classrooms, one example of TEL is the use of *lecture capture* tools to give students the opportunity to review educational material [40], which generally is a type of *blended learning* where parts of the course are provided both online and offline, e.g. the lectures are held for both present and online students [41]. Online *flipped classrooms* is a specific case of blended learning [41], where the lecture is substituted with in-class discussions. To prepare for these discussions, the students are expected to view online lectures or perhaps answer a short, related quiz [42]. Considering TEL through online-only courses, there are examples such as Khan Academy [1], but also more formal learning environments like Massive Open Online Course (MOOC)s which are created and offered in co-operation with physical universities [43]. As an alternative to MOOCs and to improve the retention rates, Small Private Online Coursess (SPOCs) are sometimes offered by universities as an addition or a part of a physical education [44].

2.1.1 Massive Open Online Course

Although the idea of publicly available, online, democratised education was present in the early 2000s, the shape and form of today's MOOC platforms was first introduced by Stanford as *xMOOC* in 2011. Among other factors, xMOOC influenced the creation of the still active

[1] `https://www.khanacademy.org`

MOOC platform (MIT) EdX[2] in 2012 and platforms like Udacity[3] and Coursera[4] further down the road. Some of the core ideas of MOOCs is that a larger set of people can have access to free, quality education. [2] Generally, a MOOC is a free, online course offered by a university with all of the learning resources needed to pass an exam to receive a certification. As there is generally no limitation for the number of enrolments, a close tutor-student relationship is infeasible. Therefore scaleable, peer-based learning through the MOOC-platform's forums are encouraged [45]. Despite the adaption and offerings of MOOCs have grown, where over 1000 universities offer more than 14,000 MOOCs, they have low retention rates [5, 6, 44]. Moreover, statistics show that the users of MOOCs often are more highly educated [2] and coming from more developed countries [3], where factors such as cultural, accessibility, lack of instructions and usefulness are identified as barriers for adaption for some demographics [29].

2.1.2 Self-Regulated Learning

Within educational psychology, Self-Regulated Learning (SRL) theory can be defined as strategies and processes exhibited by students to regulate and control their learning. In physical learning environments, such skills are not as critical as one follows a fixed time schedule with lecture time slots and deadlines for assignments. For online learning and MOOCs in particular, no such structure is enforced on the student, in addition to higher expectations of independent and autonomous learning. Therefore the ability to self-regulate becomes increasingly important to obtain learning goals in online environments. Some of the self-reported manifestations of defined SRL strategies which are positively correlated with the presence of SRL behaviour are various reviewing strategies, like revisiting assignments after completing a lecture or after passing an assignment [28].

2.2 Supervised Machine Learning Methods

Within traditional machine learning, the general idea is to *train* a proposed model on historical data, i.e. *fitting* it, with the assumption that the model can extract generalised parameters which can be applied to *predict* unseen data. One can coarsely partition the types of machine learning problems into four categories: *unsupervised*, *semi-supervised*, *supervised* and *reinforcement*, where the problem differs in terms of to which degree *the ground truth* is available. In particular for supervised learning, the ground truth, i.e. a numeric value for *regression* problems or one or multiple *labels* for classification problems. A more formal formulation of the problem; Given observed pairs $(\mathbf{x}, y)$, where an underlying function $f(\mathbf{x}) = \mathbf{y}$ describes their relationship perfectly. The objective is therefore to learn a function $\hat{f}$ such that $\hat{f}(\mathbf{x}; \theta) = \mathbf{y} \approx \mathbf{y}$ based on an observed paired dataset $(\mathbf{X}, \mathbf{Y})$, where θ is the *model parameters* learned from the historical data.

2.2.1 Classification models

A simple supervised model is a linear regression model predicting $\hat{y}$, i.e. a scalar value for an *unseen* sample $\mathbf{x}$ such that $y \approx \hat{y} = \alpha\mathbf{x} + \beta$, where y is the ground truth and α and β are model

[2] https://www.edx.org
[3] https://www.udacity.com
[4] https://www.coursera.org

parameters which are *fitted* to the historical data X. To measure the error of the model, one can calculate the sum of the average squared distance for every prediction $\hat{y}_i$ given $\mathbf{x}_i$.

As in the name of the model, it can only fit linear relationships between the input $\mathbf{x}$ and the target output y. A generalisation of such *linear models* is an *logistic regression model*, where it in a classification problem can be defined as using the *maximum likelihood estimation* on the probability of given $\mathbf{x}$, what is the likelihood that it is labelled $\mathbf{y}$? More formally, it learns the model parameters, i.e. fits the historical *training* by optimising $Pr(Y = y|\mathbf{x}; \theta)$ which essentially creates *linear decision boundaries* between the target labels. [46, p. 119] Though this can be an effective model, it might be *biased* in terms of its inability to describe all types of relations between inputs and outputs. Moreover, some classification problems may have an infinite number of decision boundaries which perfectly separates the target labels from each other. This is the motivation for finding the *optimal* decision boundary, or *hyperplane* in multi-dimensional problems in which *optimally* separates the labels. [46, p. 129] Support Vector Machines (SVMs) are models with the objective of maximising the distance between the decision boundary and the identified *support vectors* - the observations closest to the hyperplane. When the target labels, or *classes* overlap, a *linear* SVM cannot find a hyperplane which perfectly separates the data. This motivates two extensions, the first is the idea of *soft-margin* SVM [47], where the model allows some observations to be wrongly classified *if* it provides a better generalisation. [48, p. 417] The other approach is to use the *kernels*, which essentially maps the input from a linearly non-separable input space, to a non-linear separable space [49]. Different kernels exist, depending on the input and output relations. [48, p. 423]

2.2.2 Recurrence Modelling

Another modelling approach is to consider multiple weaker supervised models which individually are not great predictors, *weak learners*, but as a collective provide a weighted, improved target prediction, either in terms of classification or regression. There are multiple approaches to create such a collective, or *committee*, where *gradient boosting* method uses numerical approximation techniques like Stochastic Gradient Descent (SGD), to iteratively add a model with the parameters providing the minimum calculated loss across the training set until a *sufficient* until convergence or some other stopping-criteria [50].

For some problems, a more *sequentially* dependent model of data is needed, such as in Natural Language Processing (NLP) where the placement, meaning and choice of words are related to the words coming before and after it. Such sequential problems have motivated the idea of using *recurrent* cells to predict sequentially dependent inputs. A single Recurrent Neural Network (RNN)-cell consists of a hidden state h_t which is calculated based on the previous cell's hidden state h_{t-1}, the readout r_t and the input token x_t of an input sequence of length n. By sequentially calculating the hidden states h_τ, for $1 \leq \tau <= n+1$, any token $x_{\tau-1}$ can be *encoded*. The readout r_τ of a cell is tradionally $\tanh h_{t+1}$. [51] As for longer sequences, the *vanishing gradients* and *exploding gradients* problems occur, complicating modelling important dependencies across longer sequences. To tackle these issues, variations of the recurrent cells such as Long Short-Term Memorys (LSTMs) [51] and Gated Recurrent Units (GRUs)-cells have been proposed. In particular GRU-cells use a *gating-mechanism* to selectively choose what *memory* should be updated for each individual cell.

2.2.3 Attention and Transformers

Although the *attention* mechanism was used with RNNs, to provide nuances in the importance of individual *tokens* in an input, several issues related to recurrent modelling remain. For instance, they are not easily trained in a parallel matter and there are vanishing gradients related to longer sentences [51]. This motivated to only use *self-attention* to model the sequential dependencies between the tokens of the input, reducing model complexity, handling longer sequences better and the training is more parallel compared to RNNs [52]. These types of attention-based models are often referred to as *transformers*.

For learning the semantically meaningful embedding of inputs, *pre-training*, the traditional, *self-supervised learning* approach only considers the previous inputs in the sequence for predicting the next, so-called Causal Language Modelling (CLM). To take advantage of both the left and right context of a token in a bi-directional manner, another problem definition was proposed, as the next-input prediction problem becomes trivial with access to *the future*. The alternative approach is then to mask parts of the input sequence and use this *masks* as the prediction targets of the model instead. This approach is called Masked Language Modelling (MLM) and was popularised by the introduction of BERT [53], which showed how it could be used to *pre-train* language models to learn the embedding representations of words, such that down-stream tasks like classification, can be done by simply *fine-tuning* a single *hidden layer* to the specific task and related training samples, instead of re-learning all of the semantic word representations. [53] BERT's results on a large set of NLP tasks have inspired countless transformers like, RoBERTa [54] and XLNet [55]

2.3 Recommendation Systems

The main purpose of Recommendation Systems is to use historical interactions between *users* and *items* to infer their *preferences* and provide *relevant* recommendations given those preferences or explicitly created user requirements. The quality of an item interaction can be provided as *explicit feedback*, e.g. a rating from one to five, or implicit feedback such as whether or not the video was viewed or how long it was viewed for. In practical scenarios, explicit user feedback is hard to obtain. Moreover, ratings, either explicit or implicit, in general often follow long-tailed distributions, where a few items are given plenty of feedback, whereas most items do not contain much feedback. [56]. Generally, the recommendation problem can be viewed as predicting the rating of a given user-item interaction or as providing a list of "top-k' relevant items for the user. Regardless, it results in an item prediction problem, as the rating predictions would be used to rank the provided k items. [56, 57].

To provide these recommendations, several different types of RSs exist and corresponding classifications. Some commonly used groups of recommendation methods are *content-based*, Collaborative Filtering (CF), and *knowledge-based* methods, as well as *hybrid* methods, combining any of the methods [58]. Content-based methods utilise the users' preferences towards specific features of items, e.g. movie genres or actors. Collaborative Filtering on the other hand solely uses the ratings of users to recommend items. [58]. More specifically, *item-based* Collaborative Filtering (CF) predicts the rating of an item v for a given user u, using a set of similarly rated items by that given user. An *user-based* CF on the other hand, uses the ratings provided by similar users to predict the rating of item i for the user u.

Collaborative Filtering methods can either be implemented directly as *neighbourhood*-based

methods as proposed above, also called *memory-based*, which are generalisations of the K-Nearest Neighbours (KNN) algorithm. [56] For large, sparse user-item rating matrices, the computation of the neighbourhood-based ratings is inefficient. Creating a *model* of the user preferences can therefore be an alternative approach, where many classification models can be adapted to the CF case [56, p. 86]. An example of a *model-based* CF method is to apply dimensionality reduction techniques to the rating matrix, representing them as *latent factors*, which reduces the sparsity and therefore improves the efficiency. Moreover, by representing both user rows and item columns by low-dimensional latent factors, one can approximate the original rating matrix due to the correlations between rows and columns in the rating matrix, but with lower memory usage and higher efficiency without using any neighbourhood methods. [56, p. 47, 91] More formally, the rating matrix approximation by *Latent Factor Models* can be summarised as

$$R \approx UV^{\intercal}, \tag{2.1}$$

where R is of size $m \times n$. U and V are the low-rank latent representations of the users and items of size $m \times k$ and $n \times k$ respectively. [56, p. 93] Moreover, RSs can be seen as a generalisation of classification or regression problems, e.g. ratings as labels or continuous values. [56, p. 93]

Although there are various methods for obtaining these low-rank, latent representations through dimensionality reduction, most can be defined as Matrix Factorisation (MF) methods, such as Non-negative MF and Singular Value Decomposition (SVD), where they mostly differ in the chosen objective functions and the imposed constraints on U and V[56, p. 96]. When considering implicit feedback, some additional constraints are implied. With explicit feedback, both positive and negative ratings are present, but with implicit feedback, only positive feedback is available. For instance, if a user does not view a video does not necessarily indicate that the user dislikes the video. Furthermore, the quantity or numerical value of implicit feedback, e.g. the number of times a video is viewed, indicates *confidence* rather than their *preference* which is exhibited through explicit feedback. Moreover, as the user's true preferences or intentions are never explicitly given, implicit feedback is inherently noisy as it can only provide indicators for them. [59]

On another note, the *knowledge-based* methods use explicit user requirements or infer user preferences through domain knowledge, instead of learning their preferences as with content-based and CF [56, 60]. While the other two mentioned methods perform poorly with sparse data, the *cold-start problem*, knowledge-based methods avoid it by using the users' stated preferences. [60]

2.3.1 Sequence-Aware Recommendation Systems

As data collection has become easier, *contextual* methods have been proposed to improve the quality of the recommendations when user ratings are sparse [56]. Examples of types of context can be the user's location, the time of day of or the device used for the interaction. [57] A subset of context-aware Recommendation Systems, is the models which only include temporal context, *time-aware* RS [57]. In comparison, Sequence Aware Recommendation System (SARS) mainly do not focus on the specific timestamps of interactions, but the relative ordering of them [61]. Most often, the objective SARSs is formalised as the *top-k* recommendation problem, providing a prioritised list of relevant items. More specific for sequence-aware models is that the recommendations can either be alternatives, e.g. dinner suggestions, or recommendations to be consumed in the given order, like a series of educational videos. Moreover, the input

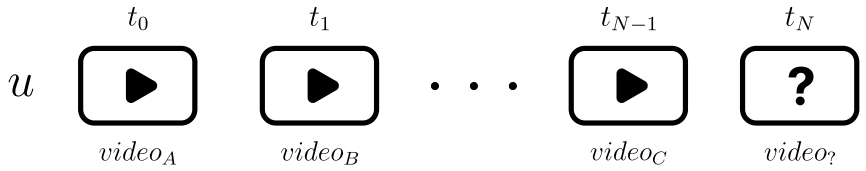


Figure 2.1: An example sequence of viewed videos of user u for a next-item prediction task, where each video is viewed at time step t_τ, where $0 \leq \tau \leq N-1$ and t_N is the next-item predicted for a given user.

to the recommendation system is no longer a user-item rating matrix, but rather ordered sets of interactions, *sequences*, for each user which may contain additional properties such as the type of the interaction or timestamp. [61]

Sequence-aware Evaluation

As Sequence Aware Recommendation Systems impose order constraints on the interactions, traditional data splitting and evaluation strategies might not be applicable, depending on the strength of the constraints. For instance, in the time-unaware evaluation, one can randomly sample the dataset into a train and test set, e.g. by user or item or both, whereas the training set can be evaluated using randomised cross-validation strategies. For SARS on the other hand, the interactions, i.e. implicit ratings are ordered by definition and a time-aware splitting strategy must be employed, e.g. by a *hold-out* strategy. For instance, one can retain either a fixed or relative set of each user's most recent interactions as the unseen, test set as the defined targets for SARS is the future interactions of the user. This is often referred to as the *next-item prediction* task, given the N user actions before it. Figure 2.1 illustrates a single video recommendation example of the next-item prediction task. As a consequence of having only one train-test split, the evaluation may lead to biased results. As an alternative to random cross-validation techniques, one can alleviate the bias by splitting the data into equally sized, potentially overlapping "sliding windows" generate several test sets, or repeatedly sample a subset of users and consider their most recent actions as the test set for larger datasets. [61]

2.3.2 Recommendation systems as Technology-Enhanced Learning (TEL)

As the availability and usage of TEL have grown, in particular large-scale solutions like MOOCs, the amount available student data has also increased, and the cost of collecting it has decreased This data is helpful to understand the students learning behaviour and processes, but it must be processed through well defined LA techniques. [62] In addition, it can provide insights into which methods to provide, and how to apply them for creating a personalised learning experience. Consequently, RSs have been proposed to help personalise different aspects of TEL [63], such as MOOC platform recommendation, course recommendation, learning resource recommendation and *learning path* recommendation[24]. Some examples of learning resources are relevant exercises, articles or videos, where a learning path consists of a sequence of proposed learning materials and the order to consume them [64]. Therefore the learning resource RSs is a specific case of learning path recommendation, where the length of the recommended learning path is one. The main argument for personalising the learning environment is to increase the efficiency and effectiveness of learning processes, as well as the quality. [62]

2.3.3 Calibration

Besides evaluating recommendation accuracy of recommendation systems, other research has focused on aspects regarding *diversity*, *coverage*, *adaptivity*, *novelty* and *fairness* to name a few [36, 65]. Another aspect which was formalised in [65] is the *calibration* of a Recommendation System. The intention is that the users' historic interactions across a set of classes C should be *proportionally* reflected in the provided recommendations of each individual user. So for example in the space of movie recommendations, if one movie genre, e.g. *comedy*, dominates the user's past interactions, it should proportionally dominate the recommendations for that user. In difference to diversity, it does not attempt to. To measure to which the degree the RS is calibrated, [65] proposed to use Kullback-Leibler (KL) divergence [66] between the users' historic class distribution and the recommendations. More formally, the definitions of [66] is adapted to the user-instance scenario, where the user u's distribution over the classes c, $p(c|u)$ is defined as

$$p(c|u) = \frac{\sum_{i \in \mathcal{H}} w_{u,i} \cdot p(c|i)}{\sum_{i \in \mathcal{H}} w_{u,i}}, \tag{2.2}$$

where $\mathcal{H}$ is the set of items which the user has interacted with and $w_{u,i}$ is the weight of that item. $p(c|i)$ is the assumed known distribution of each class c [65]. The distribution of the made recommendations $q(c|u)$ is defined as

$$q(c|u) = \frac{\sum_{i \in \mathcal{R}} w_{r(i)} \cdot p(c|i)}{\sum_{i \in \mathcal{R}} w_{r(i)}}, \tag{2.3}$$

where $\mathcal{R}$ is the set of recommended items and $w_{r(i)}$ is the weight of that item, e.g. based on its rank [65]. The KL divergence between the two is then defined as

$$C_{KL}(p,q) = KL(p||\tilde{q}) = \sum_{c} p(c|u) \log_2 \frac{p(c|u)}{\tilde{q}(c|u)}, \tag{2.4}$$

where $\tilde{q}(c|u)$ is the smoothed distribution $\tilde{q}(c|u) = (1-\alpha) \cdot q(c|u) + \alpha \cdot p(c|u)$ with $\alpha = 0.01$ to avoid divergence for $q(c|u) = 0$ and $p(c|u) > 0$. C_{KL} is small if the recommendation distribution $p(c|u)$ is similar to the target distribution $q(c|u)$. Some of the relevant properties of the metric are that it is zero for identically evaluated distributions, sensitive to small differences when $p(c|u)$ is small and it prefers more uniform distributions. [65]. To evaluate an entire dataset, C_{KL} is defined as

$$C_{KL} = \frac{\sum_{u \in \mathcal{U}} C_{KL}(p,q)}{|\mathcal{U}|}, \tag{2.5}$$

where $\mathcal{U}$ is the set of all users.

2.4 Evaluation

In the following sections, several traditional classification evaluation metrics are defined, as well as ranking metrics. Further on, some non-parametric statistical significance tests are described.

2.4.1 Classification Metrics

In the realm of binary classification, there are four outcomes of the predictions. If the true label was positive, it can be either categorised as positive, i.e. True Positive (TP), or negative, False

Negative (FN). If the label instead was negative, it could then also be labelled as positive and result in a False Positive (FP), or correctly labelled as negative, i.e. True Negative (TN). [67]. Their relations are often presented in a *confusion matrix*, which is illustrated in Figure 2.2.

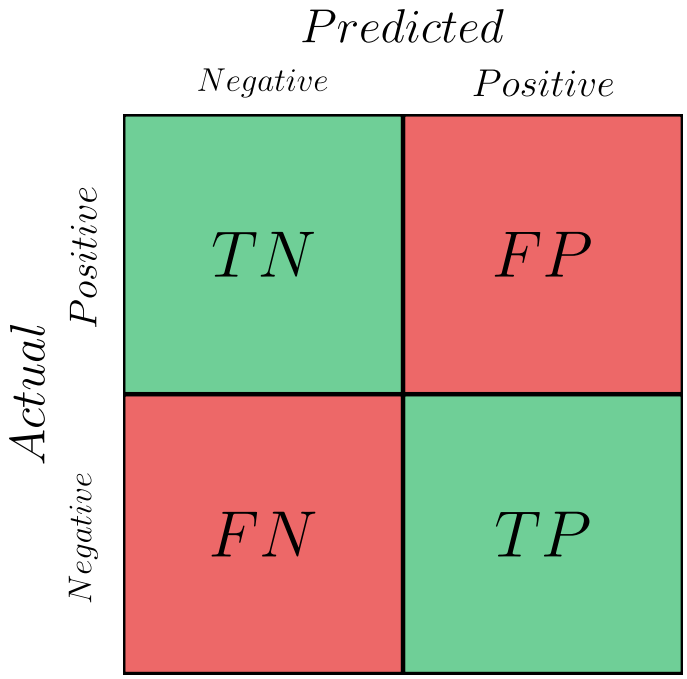


Figure 2.2: Example confusion matrix for binary classification

Based on these four possible classification outcomes for each considered instance, one can define the accuracy, precision, recall for binary classification as

$$Accuracy = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN} \tag{2.6}$$

$$Precision = \frac{\#TP}{\#TP + \#FP} \tag{2.7}$$

$$Recall = \frac{\#TP}{\#TP + \#FN} \tag{2.8}$$

$$\tag{2.9}$$

where # indicates the number of observations which fall into the respective square of the confusion matrix [67].

Moreover, as a classification algorithm would prefer both high accuracy and high recall, F_score measures a trade-off between the two, calculating the weighted, harmonic mean. Formally, it is defined as

$$F_\beta = (1+\beta^2)\frac{Precision}{Recall + \beta^2 \cdot Precision} = \frac{(1+\beta^2)\#TP}{(1+\beta^2)\#TP + \#FN + \#FP}, \tag{2.10}$$

where β is a hyperparameter for weighing the importance of either precision or recall [67].

2.4.2 Ranking metrics

For Recommendation Systems the *ranking* of the recommended items are of interest as the end-user will often be recommended a list of items, prioritised by the RS. For implicit feedback data, the exact predicted scores of items might not be meaningful to the user, but the ranking of them can be [36]. Using the definitions of [68], one can assume a recommendation algorithm A which provides the predicted ranks $R(A,u) \subseteq \{1,\dots,n\}$ of the relevant items for a user u, from a total of n items. A ranking metric M then maps these rankings to a scalar which is then averaged across all users considered and presented, i.e.

$$\frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} M(R(A(u)). \tag{2.11}$$

where $\mathcal{U}$ is the set of all users in this specific case [68].

By these definitions, the following metrics recall and Average Precision (AP) and NDCG with *cut-off* at position, i.e. rank, k is defined as

$$Recall(R)_k = \frac{|r \in R : r \leq k|}{|R|} \tag{2.12}$$

$$AP(R)_k = \frac{1}{\min(|R|,k)}\sum_{i=1}^{k} \delta(i \in R) Prec(R)_i \tag{2.13}$$

$$NDCG(R)_k = \frac{1}{\sum_{i=1}^{\min(|R|,k)} \frac{1}{\log_2(i+1)}}\sum_{i=1}^{k} \delta(i \in R)\frac{1}{\log_2(i+1)} \tag{2.14}$$

respectively, where precision at rank k is defined as $Prec(R)_k = \frac{|r\in R: r\leq k|}{k}$ and the respective algorithm A and u is omitted for sake of simplicity [68]. Although each of the mentioned ranking metrics is reported as an average across all users, Mean Average Precision (MAP) is an alternative name to AP when measured in this way.

2.4.3 Statistical significance

Statistical tests, i.e. hypothesis testing, are desired to establish any statistical significance of the evaluation metrics when comparing different RSs. In general terms, a *null hypothesis* h_0 and a*lternative hypothesis* h_a must first be defined. Then a relevant test statistic T where its distribution is then derived, e.g. a student T-distribution for the Student T test. Given a probability threshold α, h_0 is rejected if the probability of the observed, calculated statistic t_{obs} of the set of observations, i.e. *sample*, is at least as extreme as the given threshold. The choice of α will decide the likelihood of a false positive, i.e. rejecting h_0 when it is valid. The calculated likelihood of observed t_{obs} is often referred to as the p-value. [69] Moreover, hypothesis tests can be categorised as parametric or non-parametric, where the non-parametric tests do not make assumptions regarding the underlying distributions of the samples, while parametric tests do. Moreover, when there is some relation between the observations to apply hypothesis testing on, e.g. they are from the same test subject, *paired* hypothesis tests should be used. [36, p. 268] Paired Student's t-test is an often used paired test, but it is a parametric test where one of the assumptions is that the observations are normally distributed for smaller sample sizes[36, 70, p. 268]. When such assumptions are invalid or there are only a few observations,

the non-parametric *Wilcoxon signed rank test*[71] is a good alternative which uses the ranks of the pairwise differences in its test statistic calculation [72]. The test statistic is defined as

$$z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{2.15}$$

where $T = \min(R^+, R^-)$. Moreover, R^+ is the sum of the ranks of the differences between paired observations and R^- is the sum of the negative paired differences, defined as

$$R^+ = \sum_{d_i>0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i) \tag{2.16}$$

$$R^- = \sum_{d_i<0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i), \tag{2.17}$$

where $\text{rank}(d_i)$ is the given rank of the paired difference of the i-th out n paired observations. [72],

Another set of statistical tests has also been proposed to verify the normality of a distribution. An adoption of the one-sample Kolmogorov normality test is the Kolmogorov-Smirnov test [73], which have a two-sample definition to examine whether or not two, continuous distributions are identical. More formally, its test statistic is defined as

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)|, \tag{2.18}$$

where F_m is the empirical distribution, estimating the underlying cumulative distribution, of the first sample of size m and G_n of the second sample of size m. Furthermore, sup denotes the *supremum* function.

In the case of multiple hypotheses testing. i.e. multiple null and corresponding alternative hypotheses, the likelihood of incorrectly rejecting a true h_0 increases. Moreover, the likelihood of having at least one false positive is referred to as the Family Wise Error Rate (FWER) [69]. A commonly used correction method to maintain the FWER is the *Bonferroni*-correction where one compares the p-value to $\frac{\alpha}{m}$, where m is the number of null-hypotheses to test [74]. A correction method which is uniformly better is the *Holm-Bonferroni* correction as the resulting FWER is always smaller or equal compared to the Bonferroni-correction method [74].

Chapter 3

Related Works

As the thesis covers multiple areas of interest, this chapter first introduces related works of side information inclusion in SARS in Section 3.1,. Following it is a study of related learning resource RSs in Section 3.2, considering hybrid methods, learning behaviour related and SARS for learning resources. Furthermore, related work regarding in-video viewing behaviour in blended learning and MOOCs are presented in Section 3.3. Lastly, some work related to re-consumption and calibration for RSs is described in section 3.4. As these topics are not disjoint, the deemed most relevant or informative part of each specific approach is used to determine their respective subsection.

3.1 Side information fusion in Sequence-Aware Recommendation Systems

As the field of RS usually handles data scenarios with high sparsity [75], enriching user-item interactions with side information like user features, item features and contextual features like time and device can improve recommendations [75]. The inclusion of such rich information has consequently been a topic in SARS to further improve them. One work studied how to include multi-modal features such as images and text in a parallel RNN for session-based recommendations. They evaluated multiple different views of incorporating the item features but using separate GRUs for each item feature, with element-wise multiplication of the resulting hidden states which are individually projected to a shared output. [76]. The proposed model Behaviour Sequence Transformer (BST) [77] utilised both contextual features as well as user and item features in the output context of their transformer network for Click-Through Rate (CTR) prediction. Moreover, some heuristically deemed important item features, like item category, were embedded and concatenated with the item embedding as the input into the transformer layers. Side information like context, user, other item features and various cross-features were concatenated together with the output context vector of the transformer layer and fed into a Feed-Forward Network (FFN). An offline evaluation of the model showed improvements over a Wide-&-Deep model with incorporated sequential features.

Another context enriched SARS proposed in [78] for video recommendation for YouTube studied the effect of various fusion methods of contextual features. The first insight is that first-order FFNs, e.g. single hidden layer FFNs are inefficient in finding low-rank relations of concat-

enated features. Using the recency of the interaction and client device information as context for a LSTM, they showed that element-wise multiplication, i.e. crossed-features, is more effective than concatenation either before or after the LSTMs layers. The results of [79], though less extensive, contradict the above findings, indicating that concatenation of features is superior to element-wise multiplication using GRU-cells. The latter conclusion is further supported by the findings in [80]. Using an XLNet [55] model as the core of a RS, they showed that feature concatenation before the transformer blocks gave a better recommendation accuracy than element-wise multiplication. Moreover, the generalised element-wise multiplication requires the same embedding size, adding potentially more complexity to the model. Their results indicated that both Soft One-Hot Encoding (SOHE) [81] and feature-wise layer normalisation was necessary for the side information enriched models to outperform the non-enriched versions [80]. Other self-attention-based approaches, like [82] simply concatenated the embeddings of the side-information with the item embedding, and contextual features were concatenated and put as the first token in a given item sequence. As their item space was heterogeneous, they padded missing or irrelevant features related to a given item type with zeros. In a cross-attention, context and attribute-aware approach [83], the authors showed a statistically significant outperformance with the proposed model compared to multiple SARS, both with and without user and item attributes, and context. Their context and attributed adopted version of SASRec [84] was also drastically improved compared to the base version.

3.2 Learning Resource Recommendation Systems

Various surveys of recommendation systems for learning have highlighted several commonly applied methods for learning resource recommendations. [63] highlights that *ontology* representation of the domain with attributes and respective relationships is the most common approach. Furthermore, CF and *rule-based* approaches are the most common, but graph and knowledge-based approaches were becoming increasingly popular to address sparsity and unstructured data. A more recent survey of RSs for MOOCs specifically, shows that learning resource recommendation is the second most researched topic after course recommendations. Moreover, the authors point out an increase in the use of neural methods and Sequence Pattern Mining (SPM) techniques of the more recently studied works, where earlier work mainly relied on content-based and hybrid methods. The following sections detail some of the approaches including hybrid, user behaviour and graph-based approaches.

3.2.1 Knowledge-Based Hybrid Methods

Due to the domain's hierarchical and relation-rich nature, modelling resource recommendations as graph problems or utilising ontology and knowledge graphs in RSs has been common [24, 63, 85, 86]. With a dataset of explicit ratings of learning resources in different computer science topics, [87] proposed a hybrid RS using an ontology for learner behaviour and learning resources. The learner ontology included the learners' knowledge level as well as their learning style based on the Felder-Silverman Learning Style model [88], retrieved through a knowledge orientation test and an online survey respectively. The learning resource ontology contained the type of resource and the format, e.g. text or video for 240 unique resources. Utilising item-based CF incorporating both ratings and ontological similarities between learning resources. The provided top N recommendations are then re-ranked using SPM (GSP algorithm) based historical sequential resource patterns, outperforming each of the model components alone

(regular item-based CF and ontology adjusted item-based CF). Moreover, a larger portion of the 50 test subjects displayed a satisfaction of RS.

Another ontology-based approach [89], proposed the use of an ontology of the course content to recommend *learning paths* of Knowledge Unit (KU) which takes topic or knowledge mastery into account. By considering educational videos and their difficulty as well as self-reported knowledge mastery, the RS will adapt its model of the users' topic mastery and change its recommendations accordingly. The implementation of the RS also included learning path and knowledge mastery visualisations. Based on online testing with 34 test subjects, more than 80% found the RS to be useful.

Rather than confining the recommendation space to only resources, one can instead consider the recommendation of knowledge concepts when they are related to learning resources as in [90]. This allows for a wider definition of relevance and a potentially more accurate model for offline evaluation. By looking at heterogeneous relations, including teacher-courses relations, video-concept relations and user-video relations, they defined four *meta-paths* to consider from the modelled Heterogeneous Information Network (HIN) of the dataset. A Graph Convolutional Network (GCN) is applied to the assumed relevant meta-paths and the attention mechanism is applied to the learned meta-path representations. Using knowledge concepts' number of clicks, they utilise extended MF including concept representations, to predict knowledge concept ratings. The ablation study of the importance of each individual meta-path showed that two users viewing the same video was more significant than relations through the same course, same knowledge concept or the same teacher. Comparing the model to both conventional and Sequence Aware Recommendation Systems, which illustrated an outperformance on all metrics.

Another approach which accounts for educational video recommendation across courses and topics was proposed in [91]. With a goal of reducing the cognitive overload of users, the authors propose a RS utilising both the relations between educational videos across courses through related KUs, and extracted learning preferences and video viewing patterns. Based on user-based CF of learning preferences and viewing patterns, a candidate video set is generated. Each video in the video set is then extended using the cross-curriculum video-knowledge map to a video subgraph with related videos, ordered by relevance. The results indicate a higher recommendation accuracy and more relevant videos based on knowledge relevance than individual viewing pattern-based CF methods, where duration is sliced or manually normalised. Although a relevant method and exploring a relatively large dataset, the three examined courses are highly related and do not illustrate how they would perform in a more diverse course domain setting.

A similar cross-curriculum method *MOOCex* was proposed in [92] with the goals of improving decision-making, contextualising learning paths and knowledge concepts, and increasing diversity and learner flexibility. By collecting more than 4100 videos from over 40 MOOCs, mainly Computer Science (CS), they could employ SPM techniques on the extracted topics from the video transcripts. The proposed model uses a hybrid approach, first providing a candidate video set using a content-based approach with a *tf-idf* [93] weighted transcript similarity calculation. The retrieved videos are then re-ranked according to topic similarity and the mined topic transition patterns from the various syllabi. To better inform the user of its choices, the recommended videos and the similarity distances are visualised in a user interface where the current path is highlighted. In addition, each video is colour coded according to the

course it is a part of, as well as the related topic for explanation purposes, as well the next-in-syllabus videos if relevant enough. The evaluation shows that the proposed model compared to a content-based approach is better at recommending non-trivial videos, i.e. videos in the same syllabus section and therefore enriching the user experience. Additionally, the proposed model has a higher course diversity in its recommendation, which may provide a more flexible and diverse learner scenario. Both metrics show significant improvement over the baseline content-based method.

Though the mentioned approaches do illustrate improved recommendation accuracy over regular CF or content-based methods, they also require deep insight into individual course structures and corresponding KU across courses. The upfront cost costs of knowledge graphs can be a hindrance to the adaption of the proposed methods. Moreover, there lacks a discussion of which in-video viewing patterns [91] are chosen and how they should be utilised.

3.2.2 Learning Behaviour-Based Recommendations

Other work has focused more on using domain-specific theory to model users' learning behaviour and preferences. One paper studied higher-level learner behaviour and utilised it for the recommendation of a diverse set of learning resources [94]. The RS is a hybrid approach where the users are first clustered according to their learning style, modelled by eight dimensions. Based on these clusters, SPMs techniques are applied to calculate similarities between learning resources for each learner cluster. Using the calculated similarities, the relevant learning resources can be recommended to each user for each cluster. The list can then be re-ranked / filtered according to how likely the user is to click on a given learning resource. Considering over 20 courses, roughly 20,000 users and 20 different learning resources, The proposed model outperforms the individual components of the hybrid model (traditional item-based CF and clustering + item-based CF) but underperforms on recall due to the tendency filtering.

[95] also looked into users' learning ability for learning resource recommendation, but only based on high-level learning behaviour. The proposed approach firstly uses a user-user CF approach by generating a similarity matrix based on gender and profession, where the user characteristics are selected through correlation analysis. This similarity matrix is then adjusted by using a learner ability calculation based on previous test performance. The ability measure is calculated such that high-performing students are less similar to poorly performing students, and low-performing students are more similar to high-performing students. The intuition is that the higher-performance students' behaviour can have a positive effect on low-performing students. The evaluation was done in an online fashion for a single CS course with 126 students, and it showed improved recommendation accuracy compared to a traditional user-based CF. Moreover, students spent more time with the videos recommended by the proposed model compared to the baseline, potentially indicating a higher relevance of videos.

For a more granular, sequence-aware learning behaviour approach, [35] proposed a RS for exercise recommendation using reinforcement learning. They designed three learner behaviour-based goals related to learners' long-term learning performance. More specifically, the first goal was related to users' *exploit-explore*-preference. The second is regarding the smoothness of exercise difficulty, meaning that the transition from one exercise to another exercise is not a step-change in difficulty. The last goal was associated with the engagement of the user, i.e. that the user is recommended tasks which are within the user's expected range of difficulty, neither too easy nor too difficult. The three rewards are then linearly combined with adjustable

parameters according to real-world needs. By embedding both the content and the concepts of each exercise, they propose a new exercise recommendation framework, one variation using the *Markov property*, and the other using recurrence through GRU cells. Visualisations of two exercise datasets illustrate that users' concept coverage increases with the session lengths, measured by the number of exercises. Moreover, shorter study sessions have more dramatic changes in exercise difficulty than longer ones. The last insight is that longer study sessions have a mix of both harder and easier exercises. In offline evaluation, the proposed method drastically outperforms the proposed methods, both conventional and reinforcement learning based RSs. Additionally, the recurrent variant is generally better than the one using the Markov property. Simulating student performance, the online evaluation illustrates that the exploit-explore reward is effective in terms of concept coverage and that the smoothness of exercise difficulty and engagement rewards are generally effective as well for providing a more gradual learning curve and individually adapted difficulty levels.

In [22], the authors also explored learning behaviour for recommending, in this case, lectures. They based their user-based model on more formal educational psychology theory, namely Zone of Proximal Development (ZPD). In short, it describes the optimal learning zone where students are given resources which are not too easy, nor too difficult, but it is slightly beyond their current ability. [22] Firstly they explored differences in learning ability per student based on the completion proportion of viewed lectures and the related quiz performance and clustered the users into three groups based on ability (active, potential and inactive students). Secondly, they created a RS based on LinUCB, a *context-bandit* algorithm shown to balance users' exploit-explore preferences well [22]. The difficulty degree of a video using in-video viewing patterns such as the average completion rate and number of rewinds, and the number of users who answered the video quiz correctly. Along with the number of videos studied and the student's ability, a personalised exploration coefficient was calculated and utilised in the improved LinUCB model. In addition to evaluating the model's recommendation accuracy, they also evaluated the *adaptivity* of the model to learners' ability, i.e. the difference between the recommended videos' difficulty and the ones which were viewed. Additionally, they evaluated the diversity of the recommended videos with respect to the average difference between the pairwise compared recommendation lists of users. Considering only the active students-cluster, the results showed that the proposed model outperforms both item-based and user-based CF. Moreover, comparing performance across the different user clusters, the model had the highest precision for the active students, who had both high learning abilities and had viewed the most lectures. The inactive students, with the lowest ability and who had viewed the least amount of lectures, had the lowest precision. Regarding the non-utility metrics, they evaluate the exploit-explore trade-off by manually adjusting the personalised exploration coefficient, they show that the personalised coefficient improves diversity more than globally set exploration coefficients.

As the authors of [22] utilised a subset of the dataset in [14, 15, 25], there is not much topic diversity to make the approach generally applicable to all educational domains. Although they had access to multiple granular in-video viewing features, they only included the completion rate and rewind frequency. Some of the reasoning was based on educational psychology theory, but some were less argumentative and moreover lacked quantifiable arguments. Furthermore, the used accuracy metrics are not standard as the precision metric only considered positive items as both viewed and correct quiz answers.

3.2.3 Sequence-Aware Recommendation Systems for Learning Resources

Some work has also studied the use of other SARS besides SPM [96], and other conventional, adapted RSs for resource recommendation with the inclusion of some context or side-information as contextual information is deemed valuable for learning resource recommendation and RSs for TEL in general [63]. Moreover, some graph methods can be considered as sequence-aware methods [61], but they are not considered in this study.

Most notably, [97] explored the usage of an LSTM with various methods of including the time spent on each resource to recommend the next learning resource by its URL. The time spent is calculated by the difference in timestamps between consecutive navigation events. By modelling time as categorical or continuous, as well as fusing the time feature with the embedding through concatenation, or as *bucketed output*, similar to contextual post-filtering. Compared to a next-in-syllabus model, *n*-gram model and a *MostPop* baseline, the non-time augmented LSTM was superior with respect to their measured accuracy. Moreover the four different time augmented LSTMs showed improvement over the base version, where the discretized, categorical time feature fusion without the bucketed output had the highest accuracy.

Improving their previous work, they expanded the research to 13 different courses in [23] in several different domains, to evaluate the general applicability of the approach. By characterising each course by the number of navigational logs, students, resources (course pages) and measured navigational entropy, they trained a regular LSTM, as well as a time augmented LSTM as in their previous work [97] on each of the courses. In ten out of 13 courses, the time augmented LSTM model performed the best according to accuracy. Moreover, a linear regression of the course features with a target of the relative improvement of the time augmented LSTM over the next-in-syllabus method illustrated that the raw number of navigational events as well as a high navigational entropy had a positive contribution to the relative outperformance.

3.3 In-video viewing behaviour

In a lecture capture setting, [40] studied both quantitative and qualitative efforts related to the effects of providing lecture captures. Another work [19] studied the effect of intrinsic video characteristics and the relation to cognitive load. Further on, it explored how different in-video viewing patterns are related to the video characteristics, showing that some viewing patterns could have a mitigating effect on high, non-positive cognitive load, though self-reported.

At a lower level, studying in-video viewing behaviour and patterns has been of interest for a long time. A study from 1999 [11], illustrated early on the highly skewed distribution of some viewing patterns, e.g. pausing, and seeking by using Kolmogorv-Smirnov tests [98, p. 595]. Moreover, the data showed that the average time spent viewing a video was about 6 minutes, though with a standard deviation of close to 9 minutes. On average, a user viewed 2 videos per study session, where a session lasted approximately 20 minutes on average. By using Markov chains, the authors attempted to predict the mean duration of some in-video viewing behaviours with some success. The various in-video viewing patterns explored in the related works are highlighted in Table 3.1. The following sections describe related works in in-video viewing behaviour analysis in blended learning and MOOCs respectively, as lectures blended learning may be only supplementary sources, while for MOOCs they are primary sources of knowledge.

Blended Learning in-Video Viewing Behaviour

A preliminary study in [99] in another flipped computer science course explored whether learning behaviour could be predictive of course performance. They clustered users by their online lecture completion proportion into five clusters, analysing the individual users' test scores of the related video quizzes. Using an Analysis of Variance (ANOVA) [98, p. 865] parametric test, they found statistically significant differences in the means quiz performances of each cluster, though they did not discuss the validity of the assumptions made required by an ANOVA [70].

In the authors follow-up study [42], they studied two semesters of the same programming-course offering and clustered users in a given study cycle based on their viewing punctuality, average lecture completion proportion and the normalised number of lectures viewed in the cycle, resulting in two clusters with a relatively balanced split. The further analysis described the clusters as "low video engagement", consisting of users with initial low engagement and decreasing as the semester progressed. The other cluster illustrated users with generally high lecture engagement, with a high initial viewing rate and maintaining it throughout the semester. One insight they provide is that users with prior experience with programming were negatively correlated with the lecture engagement level. Furthermore, there was a statistically significant difference in course performance between the two groups by the *Chi-square* test [98, p. 257], though small, favouring high video engagement. Some of the limitations noted by the authors are not accounting for the learning activities practised outside of the analysed platform.

Another flipped classroom study also explored punctuality, in addition to revisiting behaviour [26]. Through analysing three semesters of the same course offering, the initial insight is that there can be stark differences in engagement and behaviour between each course offering. On the other hand, a *Kruskal-Wallis* test [98, p. 981] did not indicate any difference in course grade distributions across the three semesters. Furthermore, the number of visits per video is related to the temporal distance to relevant assignments, i.e. older videos are more likely to be reviewed. Regarding revisiting behaviour, the data showed it was relatively infrequent and clearly biased by individual users. In addition to punctuality, they studied lecture competition proportion, global video coverage and the average number of visits per video, and the *Spearman rank* correlation [100] to the course grade. The video completion rate was shown to be statistically significant and positively correlated with course performance for one-course offering, where the number of visits per video was both positively and negatively correlated depending on the course offering. Moreover, the study suggests a need to control or adjust differences in viewing behaviour. Lastly, they did not quantify the effect of differences in course offerings on learning behaviour. Proposed as future work, they suggest analysing more fine-grained features and the nature of revisiting behaviour.

Another flipped classroom paper [101], studied the use of a single lecture, where the different in-video viewing patterns were examined, and how they related to different topics within the lecture. Some of the insights were that there are frame seeking behaviour when there is a related assignment to a segment of a video. Moreover, pausing and seeking behaviour was related to the information provided in the lecture, as well as the perceived importance of it. Additionally, the revisiting behaviour was deemed different from the initial viewing behaviour because of the intrinsic nature of the lecture. Furthermore, the data showed a non-normal distribution of pause duration.

Another study of video analytics [10], included for each lecture a corresponding quiz to relate navigational efforts to student performance. During a time-limited experiment, users displayed

a more interactive behaviour to seek out the relevant information in the lectures. The results indicated that rewind behaviour was most correlated to the relevant segments in the lectures. In a larger experiment, they studied eleven students and their points of interest in the lectures. The findings are similar to [101], as segments which are most viewed are also deemed relevant for any assignments and required higher-order cognitive skills, i.e. information rich.

Some of the main limitations of the above studies in blended learning is that they consider only small datasets with low topic diversity and only high-level revisiting behaviour.

Table 3.1: Summarised in-video features explored in the LA-related studies mentioned. The abbrevation definitions are: **RW**: Rewind, **FW**: Forward, **F**: Frequency, **D**: Duration, **SK**: Skip, **RP**: Replay, **T**: Total, **P**: Played, **C**: Completion, σ: Change, μ: Average

Paper	Seek	Pausing	Time spent	Playback rate	Objective
[17]	RW-F, FW-F	F, D	RP, SK,	A, σ	LA
[18]	RW-F, FW-F	F, D	RP, SK	F, σ	LA
[14]	RW-F	F, D	-	σ	KT
[25]	RW-F, FW-F	F, D	T, P, C	A, σ	PLA
[15]	RW-F, FW-F	F, D	RP, SK	σ	KT
[16]	RW-F, FW-F	F	T, C	-	KT
[22]	RW-F	-	C	-	RS
[102]	RW-F, FW-F	F	T*, C	-	RS
[42, 99],	-	-	C	-	LA
[20]	RW-F, FW-F	F	RP, SK	σ	KT, PLA & dropout
[101]	RW-F, FW-F	F	-	σ	LA
[11]	D	F, D	Played	-	PLA
[103]	F	F	-	σ	LA
[91]	F	F	T	-	RS

*Counts the number of times the video is played, not the actual duration of it

MOOC in-Video Viewing Behaviour

In the more specific case of in-video learning analytics for MOOCs, multiple studies have been shown to be relevant. A platform-oriented study [103] focused on efforts to visualise various relevant aspects of a MOOC offering, like demographic data and temporal popularity of the course lectures. For in-video viewing behaviour, their findings based on two courses illustrated that one, the viewing patterns depends on the type of the video, e.g. lectures, assignment or lab/experiment videos. Moreover, most peaks in interaction behaviour were related to transitions, e.g. slide transitions, but also video intrinsic properties like in-video questions can cause abnormalities. The second insight is that seeking behaviour may be different despite it being triggered by the same actions, like an in-video question. The third finding regards initial views and revisits, as initial views generally have more pausing behaviour, while revisits exhibit more frequent seeking behaviour. Another insight, the popularity of the video is also related to the temporal distance to the exam, i.e. inversely related to recency. Lastly, the results show clear demographic dependencies for in-video viewing behaviour. More specifically, users from the US had a statistically different distribution of navigational events than users from China.

Another paper [16], studied the effects of video playback rate on student performance, as well as other viewing patterns, using a time-budged-based model. Exploring a dataset of over 350 lectures from six different courses with varying video lengths and domains, before and after an online study. Regarding the playback rate, roughly one in five students altered the speed at least once, whereas only a few students decreased the playback rate. The experiment studied two groups of students, where one group's initial playback rate was set to 1.0x, while the other was initially set to 1.25x for each lecture they viewed, but they could still actively alter it. The general insights were that playback rate manipulation can improve student importance, controlled for both course and student heterogeneity. As pointed out in the work, some of the grade improvement is because the users will attempt or explore more of the course content. On the other hand, users were less likely to exhibit self-regulatory study behaviours like pausing when viewing at an increased playback rate but slightly more inclined to rewind the lecture.

In courses which are highly dependent on lectures as learning resources, the users perceived difficulty of videos can provide valuable information for instructors as well as Knowledge Tracing (KT). A preliminary study on the correlation between various viewing patterns and perceived video difficulty was done in [18]. By aggregating viewing patterns on a per-lecture level and grouping them by interactive and non-interactive behaviours, they provided some valuable insights considering two courses. Regarding playback rates, videos with initially higher than default playback rates on average had a lower perceived video difficulty. Secondly, more frequent decreases in playback rate correlate with a higher perceived difficulty. Moreover, the amount of playback rate decrease is negatively correlated with the perceived difficulty, while neither frequency nor amount of increase of playback rate is significantly correlated with perceived difficulty.

Regarding pausing behaviour, the results indicate a statistically significant, non-linear correlation between perceived difficulty and pausing frequency and duration. A higher pausing frequency and duration are positively correlated with perceived difficulty, where the pausing frequency is more significant, while the perceived video difficulty does not increase on average for a median pause duration longer than one minute. For skipping behaviour, the frequency of forward skipping has a linearly, negative correlation on difficulty, relating highly frequent forward seeking to "skimming"-behaviour. The length of the skipped segments is non-linearly positively correlated, not in line with expectations as the assumption that the user would find the lecture easy or not relevant. [18].

The rewinding behaviour tells a similar story, where the main finding suggests that a higher average replayed length per rewind event indicates a higher perceived video difficulty. Moreover, viewing sessions with many rewind events most often occurred within a short time span, indicating a "frame-seeking" behaviour instead of a reviewing/re-watching behaviour [18]. Lastly, the results also showed that pausing and seeking behaviour both were highly skewed as in [11].

Since the study only focused on the occurrence or non-occurrence of a given viewing pattern, a follow-up study looked into more nuanced in-video viewing behaviours [17]. By clustering the video viewing sessions by eight different viewing patterns into nine explainable clusters, where sessions with few interaction patterns account for 65% of the sessions. Out of these clusters, both the *Replay*-cluster and *FrequentPause*-cluster are associated with higher perceived difficulty, while the *SpeedUp*-cluster has significantly lower video difficulty, but not the *HighSpeed*-cluster. The difference is the users who actively increase the playback rate versus those

who have an initial higher playback rate. Moreover, viewing sessions with significantly longer pauses do not indicate a significant increase in perceived video difficulty, compared to sessions with few to non-viewing patterns.

Furthermore, they studied which of the viewing patterns associated with the clusters can be associated with revisiting behaviour. More specifically, 60-70% of sessions with in-video dropout, i.e. not completing the video, were later revisited, whereas only 20-24% of completed videos were. Furthermore, viewing sessions in the *Replay* and *FrequentPause* clusters were significantly more likely to revisit the lecture, while *SpeedUp* and *Passive*-clusters were significant. Additionally, a decrease in playback rate, large skips and long pauses were not significantly related to revisiting behaviour. They did also identify course differences, where a viewing session apart of the *Inactive*-cluster was significantly more likely to revisit the lecture for one course but significantly *less* likely to revisit the video in the other course. [17]

Lastly, relating in-video viewing patterns to student performance, by dividing the users into *strong* and *weak* students by their performance and assignment completion rate. A *Chi-square* test [98, p. 257] showed that higher-performing students exhibited fewer viewing patterns, while weaker students tended to show different types of skipping behaviour as well as more frequent and longer pauses. Moreover, despite that replay patterns were shown to be related to perceived difficulty, there were not any significant differences between the two users partitions. Delving into the pausing patterns of the weaker students, contextualising the pausing events with the content of the lecture, showed that information-rich parts in the video, i.e. presenting example code, were the most common reason for the pauses, but there was not a significant difference between the examined types of lecture segments. [17]

Though both of these studies provide some valuable insight across a large set of users, the results are only related to computer science topics. In addition, the perceived difficulty is measured by self-reporting through a post-video survey, and not an objective measurement like a post-video quiz. Though they analysed a specific content-based scenario for one viewing pattern, they did not generally account for the differences intrinsic nature of the videos, nor the temporal tendencies of lecture viewing as a course progresses.

Another set of studies also explored various aspects of in-video viewing patterns, with many overlapping measures as the two previous studies. Utilising a dataset considering two courses, the authors' initial study researched the correlation of the different patterns, as well as utilising them to predict student performance [14]. They considered nine different viewing patterns for each video watched with a corresponding post-video question. The statistical analysis showed that a user who in total, including pauses, spent more time on a video performed significantly better on the post-video quiz. On the other hand, the completion rate was not shown to be significantly correlated to quiz performance in general. Excluding pause duration, the time spent playing the video was also statistically significant, indicating an increase in time spent playing the lecture improves score performance, regardless of pauses.

For pausing behaviour, the study showed that frequent pausing indicated a significantly higher test score, where the ratio of total pause duration to video length was not statistically indicative of either an increase or decrease in test score. The playback rate on the other hand was roughly identical for users answering incorrectly and those answering correctly, but still significantly different, with a slightly higher playback rate for correct answers. Moreover, the results showed that most users do not change the playback rate, but those who answered correctly had a statistically significant tendency to change it more often. [14]

Lastly, the difference in frequency of rewinds was statistically significant between the two groups, where users answered correctly rewinding more frequently. The frequency of skips where not significantly different. In general seeking and pausing frequency distributions were both highly positively skewed, as reported in [11, 18]. Their KT models were shown superior to MF and user-based CF baselines in predicting video quiz performances, across different types of users partitioned by their course completion rate.

In another work considering the same datasets, the authors mined the raw clickstreams of in-video viewing patterns with SPM motivated techniques to predict the score on the post-video quiz [15]. By aggregating each click event on a video into a sequence, which was then statistically analysed before frequent patterns were identified using a probabilistic mixture model. Some general insights were that there are highly significant course differences in the duration of different viewing patterns and correlated to the intrinsic nature of the courses. Furthermore, users have a statistically significant tendency to skip more content than they review. More generally, the durations where the user plays the video are statistically longer than the paused durations.

After *motif*, i.e. sequence extraction, grouping the most significant ones into four categories and relating the characteristics of each category to the average related video quiz performance. In the *Reviewing* category where users saw a part of the video before reviewing parts of it, half of the motifs were significantly positively correlated with test performance. In addition, half of the motifs in the *Skimming* category are significantly negatively correlated with test performance, in contrast to the initial findings in [14] which only considered the frequency of forward skips. The *Speeding* category is more nuanced, where some motifs which have a higher playback rate are significantly associated with higher test performance for one course. Other motifs have an "increase-then-slow-down"-playback rate pattern, which can be both positively and negatively associated with score performance dependent on the motif and course. [15]

A final study of this feature-rich dataset explored the field of Predictive Learning Analytics (PLA) studied behavioural biases related to users and lectures. When visualising some of the same nine behavioural patterns as in [14], there are significant lecture-specific variations. Considering a subset of the more active users with respect to video coverage, the visualisation also displays large deviance from the mean behaviour, illustrating the user biases. Applying multivariate linear regression to some of the users shows some interesting relations between the viewing patterns. Specifically, the regression indicates that seeking, playback rate, both changes and magnitude and pausing behaviour come at the cost of a reduced video completion rate. On the other hand, the time spent playing the video is positively affected by the number of pauses and rewinds, potentially indicating the users were reviewing some content. [25]

Some of the main limitations of the in-video viewing behaviour analysis in the mentioned works is that they mainly consider CS courses. Moreover, even though some findings of interesting viewing behaviours are correlated across studies, others are contradictory, complicating generalisations of causes and effects for a given viewing behaviour across domains. In addition, the relations between various in-video viewing behaviours may be linear or non-linear, whereas some distributions are highly skewed.

3.4 Re-consumption and Calibration in Recommendation Systems

Historically most work on RSs have studied to recommend relevant *novel* items to users, but long-term behaviour is generally a mix of re-consumption and seeking novelty [32]. More generally, the trade-off between these two behaviours is often referred to as *exploitation* (repetition) and *exploration* (novelty). However, in some domains such as education, re-consumption of items can be deemed more of an interest and importance for KT and modelling user behaviour. As knowledge gained will be forgotten as time goes by, revision can improve both the retention and understanding of the given topics [30]. Although forgetting behaviour has been an important topic in KT [104, 105], it has not been explored as much for re-consumption of learning resources in Recommendation Systems. As several terms for re-consumption are used in previous research, re-consumption, revisit, review repetition and *exploit* is used interchangeably in this work.

For predicting repetition behaviour, there are mainly two aspects of repetition-related recommendation and prediction: At a given time step, predict whether or not the user will repeat a previous interaction or given that the user will re-consume an item, which previously interacted items will it re-consume? For the former problem, [33] analysed the binary problem of predicting whether or not the next interaction is a repetition or not, proposing two models utilising domain-independent features. Their focus is on short-term behaviour and therefore only a sliding window of the most recent interactions is kept. The paper utilises two item features: their popularity by the number of interactions, and their re-consumption ratio. For a sliding window, the average of both features was found to be positively correlated with re-consumption probability. In addition, a static, user re-consumption ratio feature was used based on their interaction sequences. A last feature measured the proportion of re-consumptions in a given time window, where the re-consumption probability was found to be approximately linearly correlated with the window re-consumption proportion. Based on their proposed linear and quadratic methods, a SVM, discriminant analysis and two proposed methods, linear and quadratic, respectively. Moreover, experiments with varying window lengths illustrate domain-specific re-consumption behaviours, where two out of four datasets had negatively correlated prediction accuracy with window size. Furthermore, feature importance analysis showed that their importance is not conclusive and varies between both models and datasets.

Their follow-up study [34] explores which item in the given time window will be re-consumed, given that a re-consumption is likely. The proposed personalised pairwise method utilised a dynamic item recency and familiarity feature, in addition to the previously used item re-consumption rate and item popularity. In addition, the sliding window is an adjustable parameter for not considering the most Ω recent interactions. The proposed model consistently outperforms the baselines, including state-of-the-art models. Based on the results, time-sensitive features are found to be useful for predicting the item to be re-consumed. Moreover, the feature importance analysis indicates that the item re-consumption rate is the most significant feature. An issue is not clear if it is calculated in a time-aware method or not, which potentially biases the inference. The same issues apply to the proposed features in their initial study. A combination of the proposed linear model in the initial study for re-consumption prediction and the current paper's model shows promising results in predicting both if and what to re-consume.

RepeatNet [106] is a session-based method, combining recommendations for *explore* and *exploit* behaviours, by having separate prediction heads for each and linearly combining the probabilities. The approach uses a GRU to encode the sessions, which is then fed to a self-attention layer

to predict whether or not the next-item interaction is a re-consumption. These probabilities are multiplied with the output of two separate attention-based decoders fed the output of the GRU encoded session. The exploit decoder considers only items which are already interacted with, while the exploitation encoder considers only *novel* items. In addition to outperforming the other SARSs using only the exploitation mode, the results showed further improvement in recommendation accuracy with the fusion of both modes.

In learning resource recommendation, no previous work has been done on explicit re-consumption repetition to the author's knowledge but explicit modelling of exploit-explore preferences has been explored. [22] calculated a personalised exploit-explore coefficient with regard to lecture recommendation based on the learners' ability. [35] on the other hand used an adjustable exploit-explore objective as their multi-objective reinforcement model for exercise recommendation, also based on ability related to exercise performance for a given topic.

Research regarding calibration of RS is still relatively new, where the problem of the calibration was more formally defined in [65]. The study illustrates the similarities between the problem of calibration to other "beyond-accuracy"-metrics such as fairness, diversity and serendipity [36], but it also highlights the main differences. Moreover, it shows how KL-divergence is a more appropriate metric for measuring calibration compared to other metrics in related works. Lastly, it introduces a greedy algorithm with an adaptable calibration coefficient, illustrating the method's effectiveness in re-ranking the recommendations to align with the users' class proportionality. A more recent study proposed to model the calibration problem as a minimum-cost flow problem [37]. The proposed re-ranking method illustrates an improved calibration-accuracy trade-off regardless of the degree of calibration over the baselines, including the previously proposed greedy method [65]. Moreover, it outperforms the greedy approach across the evaluated ranking metrics regardless of recommendation list length.

Chapter 4

Datasets

This chapter first describes publicly available learning resource datasets in Section 4.1, including the reasoning and further description of the two chosen datasets. Thereafter, the preliminary, global preprocessing steps are described in detail in Section 4.2, including session generation and outlier removal. The feature extraction methods and considered features, both viewing features and lecture features, are described in Section 4.2.2, as well as the resulting dataset statistics. As similar methods applied to the same datasets, were used in the preliminary work [1], some of the sections may have some resemblance.

4.1 Learning Resource Datasets

Previous work on RS in the education domain has primarily relied on closed, unavailable datasets and more publicly available, large-scale datasets have been sought after. [24] Of the publicly available learning resource datasets, there are multiple exercise datasets such as the *ASSISTments* datasets [107, 108], *JUNIY* [109] and *KDDCup* datasets [110] [111] which have been predominately used for KT research [104], but they do not contain other types of learning resources as well. A dataset that does is *OULAD* [112] which consists of many types of learning resources related to different courses, enrolments and demographics, though they only contain the number of resource visit interactions aggregated to a daily level. The *VLEngagement* [113] dataset, on the other hand, does contain more granular interaction data of a large set of scientific videos, including explicit and implicit ratings and many video-specific features like those discussed concepts and speaker rate. Due to privacy and technical limitations, the only in-video behavioural feature they provide is the watch time. [113].

Another large-scale, learning resource dataset is *EdNet* [114] where different levels of user interactions granularity are available in different versions of the dataset. The levels of granularity range from the navigation level between learning resources to in-resource interactions such as play and pause events when watching a lecture or when selecting or erasing an option for a multiple-choice exercise. [114]. Based on data from the MOOC platform XuetangX[1], the general purpose dataset MOOCCube was published [115], containing a large set of courses, lectures and related concepts, as well as in-video viewing behaviours. Some of the limitations of MOOCCube is the coarse granularity and the lack of data types, motivating the authors

[1] `https://www.xuetangx.com`

to re-collect the data and apply new preprocessing techniques for generating a richer, larger and more diverse dataset MOOCCubeX [116]. Lastly a more recent educational, large-scale video dataset is *PEEK* [117], which contains interactions logs at a video fragment level, with related concepts, allowing for more granular knowledge tracing, video recommendation and per-video engagement analysis. Despite the per-fragment interaction level logs, only the normalised watch time per fragment is logged, where it is further discretised as a binary label for student engagement.

To answer the proposed research questions, the main criteria for an eligible educational dataset is that it must contain in-video interaction logs, such as pause, skip and play events. Moreover, it should be large-scale, containing logs from numerous students and videos, as well as across multiple topics, preferably from different domains. Considering the mentioned, commonly used learning resource datasets, only a few of them meet these criteria. More concretely only EdNet [114], MOOCCube [115] and MOOCCubeX [116] contain granular, in-video interaction logs. MOOCCube has become less relevant as MOOCCubeX is an improved, larger and contextualised version of the same data source [116]. Therefore EdNet and MOOCCubeX were used for the experiments outlined in Section 1.3, although MOOCCubeX is the only platform with videos from multiple educational domains.

4.1.1 EdNet

A large-scale, dataset *EdNet*[114] was published in 2020, with a large variety of user interactions from the South-Korean English learning platform Santa *Santa*[2]. The data was collected over a three-year period, containing data of over 784,000 users with an average of over 441 interactions per user, totalling more than 130 million interactions in total. Regarding the learning resources, it has over 13,000 exercises and more than 1000 lectures where each resource has an encoded *tag* or *skill* related to one of the 293 topics on the platform. In addition to learning resource interactions, it also has other user interaction logs related to purchases, reading explanations, as well as resource meta-data video lengths and first time made available and interaction context such as client device used. The dataset is structured hierarchically, where each hierarchy level is increasingly granular and created with different tasks in mind, e.g. KT or dropout prediction. The most granular level *KT4* is used in this project as it is the only level which contains interaction logs within a lecture viewing, i.e. when the user enters and quits the video. Moreover, the logs contain play and pause events, with corresponding in-video timestamps. [114]

4.1.2 MOOCCubeX

As mentioned *MOOCCubeX* [116] is an improved version of the originally large-scale dataset MOOCCube [115] collected from the Chinese MOOC-platform *XuetangX*[3]. In difference to EdNet, MOOCCubeX contains user interactions across over 70 expert-identified educational fields with over 4200 courses distributed across the fields. The courses consist of lectures and exercises partitioned into sections of each syllabus, totalling to 230,000 lectures and more than 350,000 exercises. Most of these resources have potentially multiple related *concepts* and each concept is related to one of the fields, totalling more than 630,000 unique concepts. Moreover, the dataset contains more than 3.3 million enrolled users with corresponding user

[2] `https://www.aitutorsanta.com/`
[3] `https://www.xuetangx.com/`

demographic-related profiles, which have in total accumulated more than 296 million interaction records. For a user viewing a lecture, the authors have aggregated the click events to *viewing segments*, indicating which parts of the user viewed at what time and at what playback rate. In addition, MOOCCubeX consists of other entities and meta-data as well such as teacher relations to courses, university relations, learning resource discussion logs and prerequisite knowledge. To date, it is one of the largest, public educational datasets created with no specific analysis task in mind to accommodate the stated needs of standardised datasets in TEL research. [116]

4.2 Preprocessing

Although preliminary preprocessing steps have been taken by the dataset authors of EdNet and MOOCCubeX, several challenges arise related to how the user-lecture interactions are logged. Three main preprocessing steps were applied to both datasets: Generation of *user interaction sessions*, removal of outlier users and extracting the in-video viewing patterns. In particular, (in-video) viewing *features* and *behaviour* are used interchangeably in this work, referring to how users interact with videos in terms of usage, pausing, skipping etc. Moreover, this work does not distinguish between online lectures and videos, where every video is considered a lecture and the opposite as the datasets do not distinguish these terms.

4.2.1 Session generation

Before extracting the in-video viewing patterns, the start and end of a given user-lecture interaction must be defined, as neither of the datasets has it stored explicitly and correctly. All of the in-video viewing interactions are grouped together for a given user lecture interaction to a *user lecture session*, further simply referred to as a *session*. Moreover, the raw logs of the datasets must be mapped to watching segments to be able to extract the viewing patterns.

Ednet

Since Ednet's KT4 partition contains all types of resources and interactions, only the actions regarding lectures are kept, where exact duplicate records are discarded. In addition to the interaction logs, the dataset contains some metadata on the lectures, such as deployment date, lecture length and a single numerically encoded tag. Each record contains the type of action, the local timestamp of the action, which lecture the action is applied to and at which millisecond in the video the action happened. The duration, speed or. There are four types of actions, namely *enter*, *quit*, *play* and *pause*.

Because of irregularities in the logs, one cannot correctly use the enter and quit actions to specify the start and end of a user lecture session. This is because a large number of the user-lecture interactions do not contain both, or any, of the defining actions, and other inconsistencies where one or more related video action records are logged after a "quit" action of the given video.

Therefore, the definition of the start and end of a user lecture interaction was determined by the duration of the pause until the next chronological action. To best infer the duration of the pause until the next action was taken, assuming that the user did in fact take a break. Due to some outliers of the play actions, a new user lecture session was also generated if a play

event lasted longer than roughly twice the lecture length. So the preprocessing steps were the following: sort all action records by their local timestamp. For each user, group together consecutive action records regarding the same lecture. If the pause between two consecutive action records is larger than *some threshold* }, create a new user lecture session. The implication is that the pause duration is so large that it is rather a "break" than a pause.

A perfect duration threshold is difficult to define, as it could depend on multiple factors such as the users' preferences and physical context when they are using the platform. For instance, [17] excludes pauses larger than 10 minutes as they interpret them as "breaks" instead of pauses for in-video viewing behaviour measures. Other work defined the length of a *study session* using outlier detection of the time between interactions, *inter-activity* period and the maximum allowed *passive* time, i.e. watching a video. Based on the analysis, where the third quartile was 16 minutes, they chose 30 minutes as the maximum pause duration. The datasets considered in this project have more granular event tracking, but no consistent known lecture lengths. Furthermore, the proposed *session* definition only considers a single lecture interaction, not multiple. Therefore, 20 minutes was chosen as the session threshold to allow for some flexibility and maintain a relative trade-off between the number of in-video interactions per user lecture session and the number of consecutive user lecture sessions,

This resulted in some user lecture sessions containing only one action recorded and could therefore not be used for any in-video pattern extraction as the start or end of the given action was undefined. Therefore, in cases where a consecutive user lecture interaction, if it contains more than 1 user lecture session, remove the sessions containing only a single action record, excluding 3917 sessions. The argument to keep the interaction is that one assumes the interaction did in fact happen, and it is meaningful for the user interaction sequence to keep it intact to better replicate the user's actual interaction sequence and meaningful lecture relations. Then within each user lecture session, the action records are transformed into *viewing segments*, with start and end-points in the given lecture. As the logs do not contain data on the playback rate used, the assumption is that is constant at 1 based on the infrequent use of it in [16], although the learning platform does support variation in playback rate.

MOOCCubeX

The creators of MOOCCubeX have already done the latter preprocessing step of creating viewing segments, so a given record contains the server timestamp, the playback rate and the start and end point of the given watching segment the user has viewed. Using this information, the assumed end server timestamp is inferred. The user lecture session generation is, therefore, more accurate compared to for EdNet, as one can use the pause duration between the end timestamp and the following start timestamp. A pause duration threshold of 20 minutes is used here as well. The effect on sequence length distributions is minimal as shown in the empirical cumulative distribution functions in Figure 4.1.

Removing outliers

During data exploration, it was apparent that there are some user behaviour outliers. More specifically, for MOOCCubeX there were multiple users who had viewed a single or multiple lecture(s) thousands of times. As this seemed like unlikely user behaviour, user which had seen

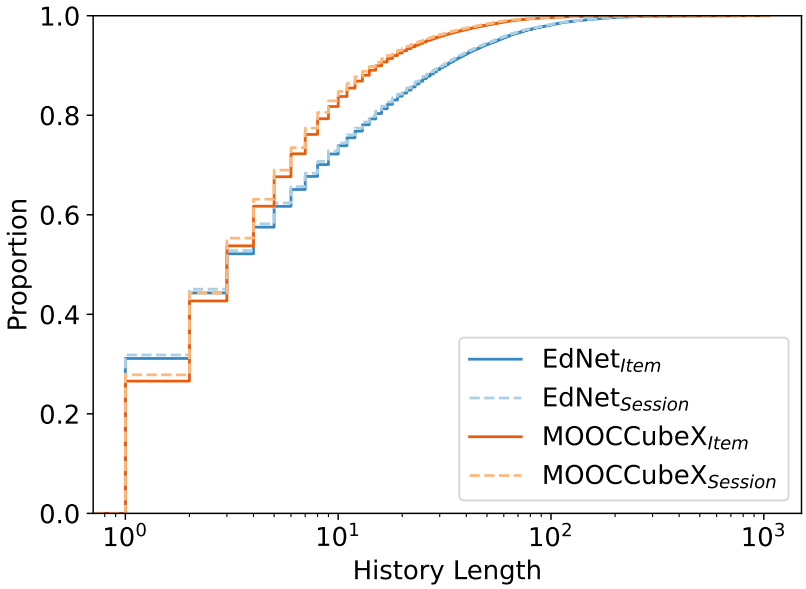


Figure 4.1: The effect of session generation on user history length distributions, on a logarithmic scale. The dashed lines indicate the sequence lengths when considering sessions and the solid line considers only the consecutively viewed lectures

at least one video more than 50 times were excluded. For MOOCCubeX, there are two resource identifiers, one unique within a given course offering ($video_id$) and one which is globally unique ($ccid$). The repetition count was based on the $ccid$, though empirically showed little difference. Although the threshold is relatively arbitrary, it was set as a threshold between likely user behaviour and the total dataset retention. In total, this meant that 5100 users were excluded from MOOCCubeX. For EdNet, this behaviour was only identified for two users.

Table 4.1: Effects of preprocessing steps for EdNet and MOOCCubeX specifically.

	EdNet			**MOOCCubeX**		
	# Interactions*	$\|\mathcal{U}\|$	$\|\mathcal{I}\|$	# Interactions*	$\|\mathcal{U}\|$	$\|\mathcal{I}\|$
Raw (unique) records	5,009,098	42,828	971	25,748,664	310,360	193,624
Session Generation*	539,331	42,828	971	21,304,672	310,360	193,624
Outlier removal	538,649	42,826	971	2,221,362	305,260	186,865
Feature extraction	538,649	42,826	971	2,213,674	304,807	186,670

*The definition of *interactions* changes from the number of recorded in-video events, to the number of sessions after the session generation.

4.2.2 Feature Extraction

As the research in related works illustrated that various viewing behaviour is related to different aspects of learning behaviour and success, the most commonly studied viewing behaviours as studied in [14–16, 18, 22, 25, 101, 102] are explored further. Due to dataset differences in

data availability, some of the definitions and resulting features are adapted. In the following sections, the specific feature definitions and extraction methods are explained in further detail, including the usage and preprocessing of topics related to the individual lectures.

In-video viewing behaviour

Some of the previous studies of in-video viewing behaviour overlap in the behaviours measured, feature definitions and datasets explored [14, 15, 25], whereas some use the raw duration features [17, 18] and others may normalise them due to large differences in lecture duration [14, 25]. Based on related analysis work on in-video viewing patterns summarised in Table 3.1, the following features are selected to be explored further:

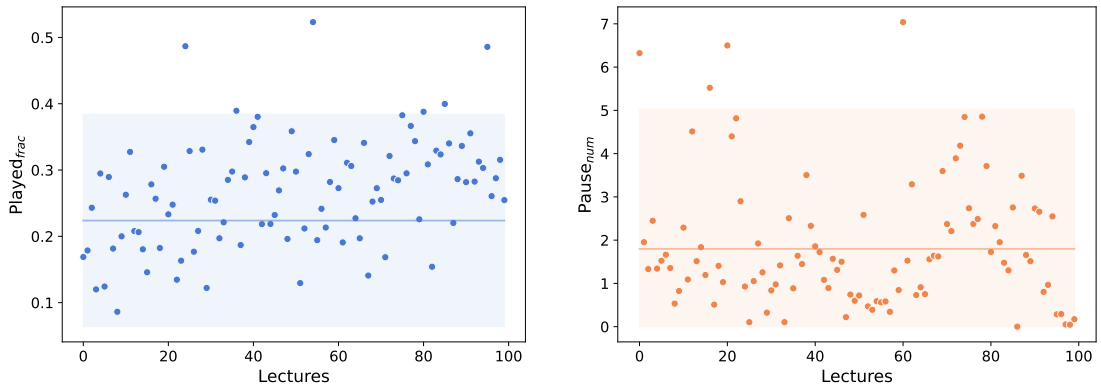
- **Skip**$_{frac}$: The total duration of (forward) skipped lecture segments, i.e. intervals, as a fraction of the lecture length. Multiple skips of the same video segment is only counted once, so it is bounded by 1. *When we compute the proportion of skipped video content, we only consider the proportion that is skipped by forward seeks.*
- **Replay**$_{frac}$: The total duration of overlapping viewed lecture segments as a fraction of the lecture length. The same overlapping segment can be included multiple times, so it is not bounded by 1.
- **Play**$_{frac}$: The total duration of played lecture segments as a fraction of the lecture length, including repeated segments. It is therefore not bounded by one.
- **Spent**$_{frac}$: The total time spent on the lecture, including pauses, as a fraction of the lecture length and therefore not bounded by 1.
- **Completed**$_{frac}$: The proportion of the lecture content the user has viewed, which makes it bounded by 1.
- **Pause**$_{num}$: Number of pauses, where 2 s $<$ pause duration $<$ 20 minutes .
- **Pause**$_{median}$: The median pause duration, where 2 s $<$ pause duration $<$ 20 minutes.
- **Rewind**$_{num}$: The number of backward seeks. A backward seek is counted if the following (play) action or interval has an earlier lecture timestamp than the action or interval in question.
- **Forward**$_{num}$: The number of forward seeks. A forward seek is counted if the following (play) action or interval has a higher lecture timestamp than the assumed end of the action or interval in question.
- **SegReplay**$_{num,\,t}$: The number of overlapping, viewed lecture segments, where the overlap is of at least t seconds.
- **PBR**$_{\mu}$: The duration-weighted average playback rate. Only applicable to MOOCCubeX.
- **PBR**$_{\sigma}$: The standard deviation σ of the playback rate. Only applicable to MOOCCubeX.
- **PBR**$_{eff}$: The "effective" playback rate, defined as $\text{PBR}_{\mu} - \text{PBR}_{init}$, where PBR_{init} is the initial playback rate when the user starts the lecture which by default is 1.

They are all features shown to be interesting for describing user interaction behaviour in terms of user engagement [13], perceived video difficulty [18] [17] and KT [105]. As more complex, non-linear relations between multiple of features indicate that they collectively are more informative [105], they are all explored further. Specifically for the definition of pauses, following [17, 18], pauses shorter than 2 seconds are not considered for pause-related features in MOOCCUbeX, but they are considered in Pause_{median} for EdNet. Due to the large variance in recorded video lengths in both EdNet and MOOCCubeX, the duration-based features, features with subscript $_{frac}$, are normalised by the highest recorded in-video timestamp per video. As

the normalisation is done using data from a training set, to avoid bias in the inference, the normalised features are not guaranteed to be upper bounded by 1.

In addition to the previously researched features, this work introduces $\text{SegReplay}_{num,\,t}$ with the intent to better differentiate between actual replaying behaviour and frame seeking. An example of a segment replay is if a user views a lecture segment $[0\,\text{s}, 34\,\text{s})$, then rewinds and watches $[29\,\text{s}, 52\,\text{s})$. This overlapping view segment would count as a segment replay for $\tau <= 5\,\text{s}$. With an increasing τ, one could assume that it is more likely that the user intended to replay some part of the video. The characteristic is a mix of the number of seeks and the replay fraction of the video, where all combinations of viewed video segments are compared for $\text{SegReplay}_{num,\,t}$. As $\text{SegReplay}_{num,\,t}$ correlates with Rewind_{num} for $t \to 0$ and naturally becomes more sparse as $t \to \infty$. Therefore $\text{SegReplay}_{num,\,60}$ is used as the given overlap threshold provides a trade-off between sparsity and actual in-video repetition

To present some of the feature characteristics related to specific lectures as the intrinsic properties of lectures have shown to affect how users interact with the lecture [10, 12, 19, 25], a subset of both datasets are used to visualise the variance in behaviours. To evaluate the variance in viewing behaviour related to individual lectures, 100 lectures are randomly selected, and each of them has been viewed by at least 100 different unique users. The average of the representative continuous and discrete features Played_{frac} and Pause_{num} respectively for each lecture is presented in Figure 4.2. As visible in both illustrations, some lectures are viewed for longer amounts of time than others, whereas others cause differences in pause frequency relative to the global average. The visualisations show similar trends in lecture variation as in [25].



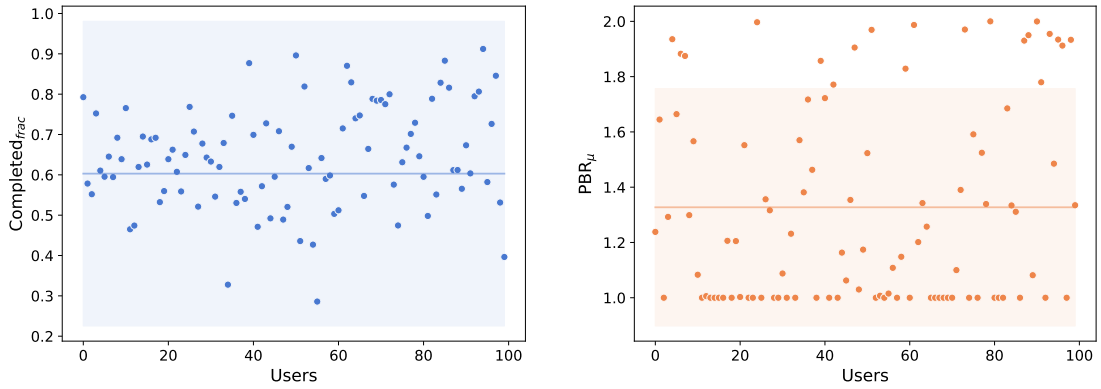
(a) The average time spent playing a video as a fraction of the video length Played_{frac} for each selected lecture in EdNet.

(b) The average number of pauses Pause_{num} per each selected lecture in MOOCCubeX.

Figure 4.2: The intrinsic lecture bias related to Played_{frac} and Pause_{num} in EdNet and MOOCCubeX respectively. The average of all sessions is represented by the horizontal line, and the global uncertainty given by one standard deviation, is visualised as the highlighted background, where it is limited to highlight positive values.

To visualise individual users viewing behaviour preference bias as shown in [25], users with a minimum level of diverse, viewing activity in terms of different lectures viewed, are selected to alleviate any outlier behaviour caused by a single session. In particular, 100 users who have seen at least 50 different lectures are randomly sampled and averaged per user. Another

pair of features Completed$_{frac}$ and PBR$_{\mu}$ averaged for each user is presented in Figure 4.3b. Examining the user averages of Completed$_{frac}$ in Figure 4.3a, clearly some users tend to watch only a small fraction of lectures, while others view on average mostly the entire lectures. Figure 4.3b on the hand illustrates that most users do not alter the video speed, as reported in [16], but some students do on average watch lectures at very high speeds.



(a) The average completion rate Completed$_{frac}$ for each selected user in EdNet.

(b) The mean average playback rate PBR$_{\mu}$ of individual, selected users in MOOCubeX.

Figure 4.3: The individual user viewing behaviour bias related to FEAT1 and FEAT2 in EdNet and MOOCCubeX respectively. The average of all sessions is represented by the horizontal line, and the global uncertainty given by one standard deviation, is visualised as the highlighted background, where it is limited to highlight positive values.

Bias adjusted features

In [25], they highlighted the longitudinal nature of a course's impact on in-video viewing behaviour. This can affect the inference of similarity between users' learning styles through viewing behaviour. As an example, given a user u_a which normally has a high SegReplay$_{num,\,t}$ for its viewing sessions, for a session $l_t^{(u_a)}$, u_a's measured SegReplay$_{num,\,t} = 0$. However, another user u_b has on average low SegReplay$_{num,\,t}$ and does not replay any segments of the same lecture in its session $l_t^{(u_b)}$. Given it is the same lecture, the discrepancy for u_a may indicate a need for a different type of learning behaviour for the given lecture, but not for u_b. Based on their previously exhibited preferences, one could argue that they should not be related positively, i.e. have high similarity for any behaviour-similarity based Recommendation Systems. The example generalises to any abnormally measured behaviour by any of the viewing features, compared to previously exhibited viewing behaviour. This motivates an additional preprocessing step to generate another dataset where the in-video viewing patterns are adjusted for users' individual, expanding bias for each feature. Formally, for a given viewing feature γ, user u, lecture l, at time step t of a user lecture session $l_t^{(u)}$, the bias-adjusted version of the feature $\hat{\gamma}$ is defined as

$$\hat{\gamma}_{l_t^{(u)}} = \gamma_{l_t^{(u)}} - \mu_{\gamma,u,\,t-1}, \tag{4.1}$$

where $\mu_{\gamma,u,\,t-1}$ is the user's average of viewing pattern γ up til the given user lecture session at time step t.

Categorical feature extraction

To relate in-video viewing behaviour to intrinsic lecture properties on a large scale, the related *topics* discussed in the lectures can be used. More concretely for EdNet, lecture topics are available as pre-encoded, expert-annotated "tags", one for each lecture. MOOCCubeX's lecture topics on the other hand consist of a *concept* and a related *field* by a many-to-one relation, where each lecture can have multiple related *topics*. The concepts are more fine-grained topics extracted from the lecture transcripts and each MOOC was labelled a given *field* by three experts [116].

After the preprocessing steps in Section 4.2, EdNet contains **259** unique tags. To analyse the concepts and fields individually, each unique concept-field mapping representation is split into separate *concept* and *field* features. Consequently, MOOCCubeX contains 171,993 unique concepts with a many-to-one mapping to 74 different fields. Notably, the concepts are in several languages, including English and Chinese, whereas the fields are only in Chinese. Moreover, several of the concepts are not semantically meaningful, i.e. random numbers or symbols. However, not every lecture have a tag or concept-field relations, where the overall dataset characteristics regarding the frequency of topics are presented in Table 4.2.

Table 4.2: The distribution of lecture topics in EdNet and MOOCCubeX, *tags* and *fields* respectively. Annotated lecture proportion describes the fraction of lectures which has at least one related topic. Topic sparsity is the proportion of user lecture sessions which does not contain any lecture topics

	Annotated Lecture Proportion	Topic Sparsity	Median #Topics
EdNet	0.5953	0.2802	1
MOOCCubeX	0.3617	0.5279	10

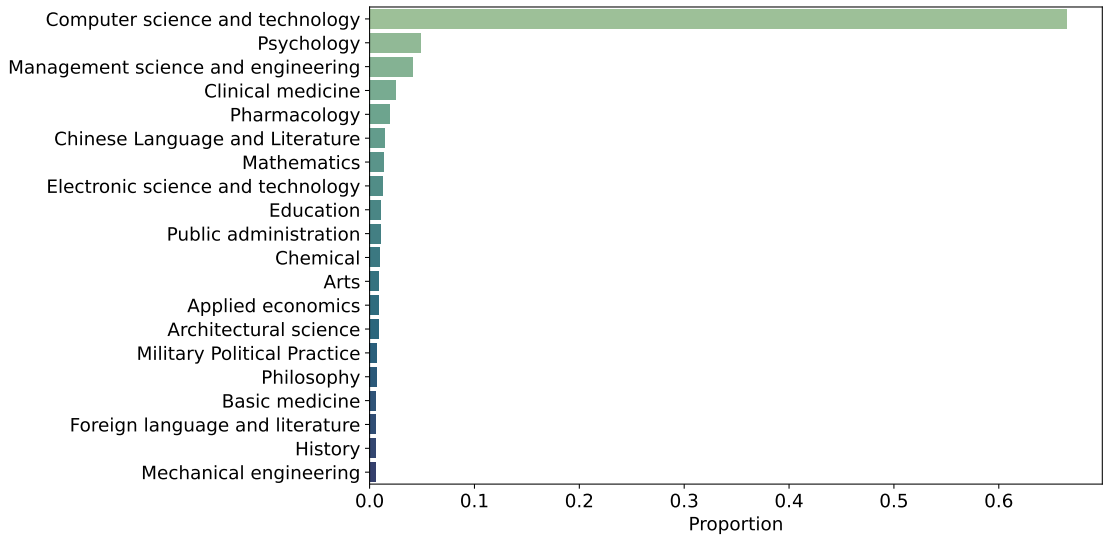


Figure 4.4: The average number of sessions of the 20 most viewed fields of MOOCCubeX

The number of sessions per field in MOOCCubeX is positively skewed as visible in 4.4, where the most frequent field per lecture is used, complicating a fair comparison across all fields.

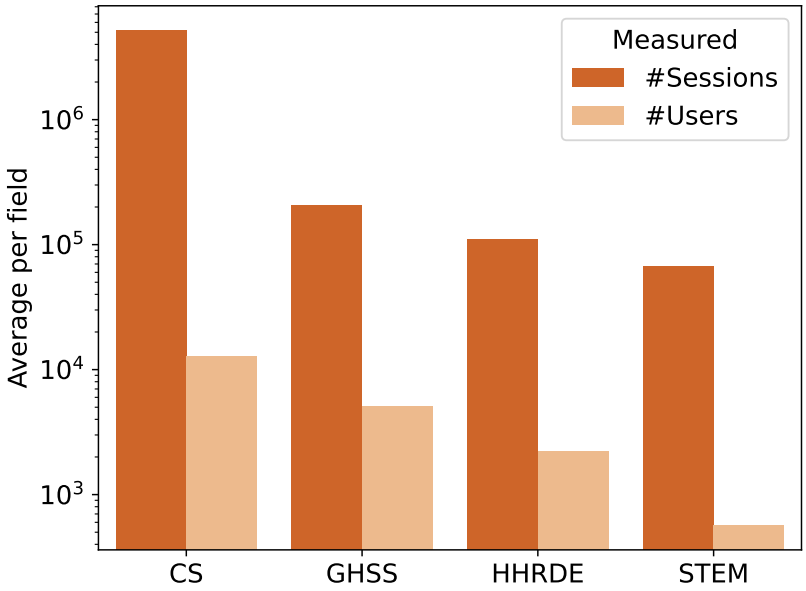


Figure 4.5: The average number of sessions and unique users per field, for each category

Therefore, each field is manually categorised into the larger *domain categories*. As there is no common, global standard of domain categorisation, the classification is done according to the categories defined in the first EdX summary study [43]: CS, Science, Technology, Engineering & Math (STEM), Government, Health & Social Science (GHSS) and Humanities, History, Religion, Design & Education (HHRDE). The classification mapping between each field and the corresponding domain category is accessible in the code repository[4]. As the number of fields per category is not equally distributed, the average number of sessions and users per field in each of the categories is displayed in Figure 4.5 on a logarithmic scale.

To take advantage of the semantic meaning of both concepts and fields in MOOC, they are separately embedded using the pre-trained embeddings for Chinese offered by the library *fastText* [118], which do support out-of-vocabulary embeddings in the case of the English and symbols-only concepts. As some of the concepts and fields are sentences, the *sentence embedding*-feature is used, where the original embeddings of dimension 300 are reduced to 64 using the library's embedding-reduction tool which is based on Principal Component Analysis (PCA) [118].

[4] `https://github.com/erlendoeien/enhancing-lecture-rs`

Chapter 5

Experiments

In this chapter, the details of each experiment proposed in 1.3 are presented. For each of the experiments, the problem statement is first presented, followed by additional, specific preprocessing steps to those applied in 4.2. Then the experiment-specific setup such as chosen algorithms, hyperparameter search configurations, data splitting strategies and evaluation metrics are described. Given the setup, the results of the experiment are presented, following a discussion of the results and how they address the corresponding research questions, as well as any limitations of the experiment.

All of the experiments are run on NTNU's IDUN compute cluster [119], which results in a varying degree of performance, dependent on the GPU of the experiment. Therefore, the specifically used GPU for each experiment evaluation is presented in Appendix A. The hyperparameter search spaces and corresponding parameters for each of the experiments are available in the code repository[1].

[1] https://github.com/erlendoeien/enhancing-lecture-rs

5.1 Experiment 1 - Next-Lecture Prediction

In this section, the necessary prerequisites to execute the set of experiments in Experiment 1 are outlined, including experiment-specific preprocessing, RSs used and the evaluation methodology.

5.1.1 Transformers4Rec architecture

Transformers4Rec [80] is an open-source framework which enables a complete pipeline from data modelling to inference, specifically created for modelling and including side information in SARS. Using the framework's modular approach, the proposed architecture includes both optionally pre-trained item embeddings and additional embedding fusion computation. The general architecture is illustrated in Figure 5.1, As Figure 5.1 shows, the overall model flow illustrates the interaction sequence of a single user. The input features, including the lecture identifier, topic features and video interaction features are illustrated and going through separate embedding processes.

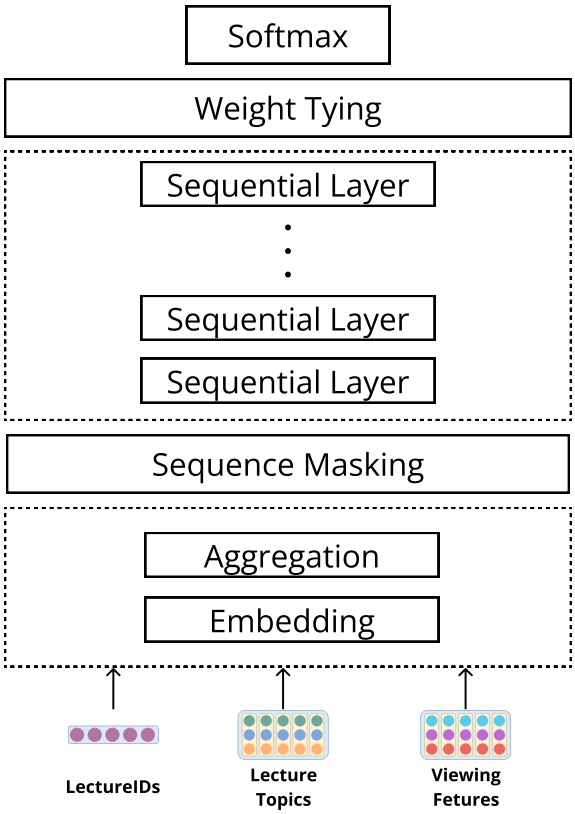


Figure 5.1: Proposed main model architecture for SARS, inspired by [80]

The framework distinguishes categorical, e.g. user, item or contextual features, and continuous features, e.g. the extracted viewing patterns. As MOOCCubeX has semantically meaningful topics, using pre-trained embeddings is an option. The categorical features are individually embedded, whereas in the case of multi-label features, like concepts or fields, the average of the concept and field embeddings are used for each session. The continuous features on the other hand can either be projected using a $\text{FFN}_{continuous}$, SOHE or both. The continuous

and categorical features can then be merged by either concatenation or element-wise, which requires identical embedding sizes. The resulting tensor can then optionally be projected using another FFN_{agg}. [80]

The motivation to add additional computation to the continuous features and the merged tensor is to learn and extract non-linear interactions between the various features. As in-video viewing patterns have been shown to have both linear [25] and non-linear relations [18], a $\text{FFN}_{continuous}$ with non-linear activation function may improve utilisation of informative patterns and reduce the importance of less informative viewing patterns. The motivation for applying a FFN_{agg} to the merged tensor is to emphasise potential non-linear relations between the various lecture topics and the associated viewing patterns. In detail, the continuous feature $\text{FFN}_{continuous}$, has a single hidden layer which is four times the size of the input feature vector, i.e. either the number of continuous features or the combined embedding size of the SOHE. The merged tensor projection has also a single hidden layer FFN_{agg} with a width four times the sequential model embedding size or in this case input size. If a FFN_{agg} is not applied, and the merged tensor dimensions are not aligned with the sequential model size, it is either way projected to the expected model size using a linear transformation.

5.1.2 Problem definition

Adapting the problem statement of a sequence-aware next-item prediction problem stated in [120], by considering a set of users $\{u_1, u_2, \ldots, u_{|\mathcal{U}|}\} = \mathcal{U}$ and a set of lectures $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$. Further, denote a *list* of user lecture sessions for a given user u as $\mathcal{S}_u = [l_1^{(u)}, \ldots, l_t^{(u)}, \ldots, l_{n_u}^{(u)}]$ which represents the temporally ordered *sequence* where $l_t^{(u)}$ is the lectured viewed by user u at the relatively indexed time step t. n_u denotes the total user lecture session sequence length for user u. In this context, the next-lecture prediction task can be formalised as predicting the lecture viewed at time step $n_u + 1$ for user u given the user lecture session history $\mathcal{S}_u$ by modelling the probability of viewing each possible lecture l stated as:

$$Pr(l_{n_u+1}^{(u)} = l | \mathcal{S}_u), \tag{5.1}$$

where Pr denotes the probability. [120]

5.1.3 Experiment setup

The specific experiment-specific preprocessing, algorithms, model configurations and hyperparameter tuning, as well as the evaluation metrics, are described in the following sections.

Preprocessing

As the main focus of the evaluation is to use interaction histories for next-item prediction, *cold users* are excluded. Following general practice, the users with less than five interactions are excluded [121]. The effect is an 8.15% reduction in the number of user lecture sessions for EdNet and a 17.1% reduction for MOOCCubeX. Further on, the sequential models require a fixed length sequence input [80]. As the sequence length is a trade-off between information and model complexity, the 30 first interactions are kept for each user, as a middle ground of the differences in sequence length distributions between EdNet and MOOCCubeX. As platform behaviour can depend on the users' experience with the platform [29], the 30 first, rather than the

most recent interactions are chosen to avoid platform experience bias. In addition, it reduces model complexity and limits the effect of crawling, bot-like behaviours which were identified [23]. Consequently, 61.6% of EdNet's interactions are retained, while 79.6% of MOOCCubeX's interactions. The resulting dataset characteristics are presented in Table 5.1. Despite relatively short sequence lengths, they are similar to commonly used benchmarks for sequence-aware evaluation [121]. In addition, sequences with less than 30 interactions are padded accordingly [80].

Table 5.1: Dataset statistics after preprocessing. The number of items is excluding the padding token

	#Users	#Items	#Interactions	sparsity	mean_length	median_length
EdNet	18,194	951	304,754	0.9823	16.7502	14
MOOCCubeX	116,661	158,358	1,461,684	0.9999	12.5293	9

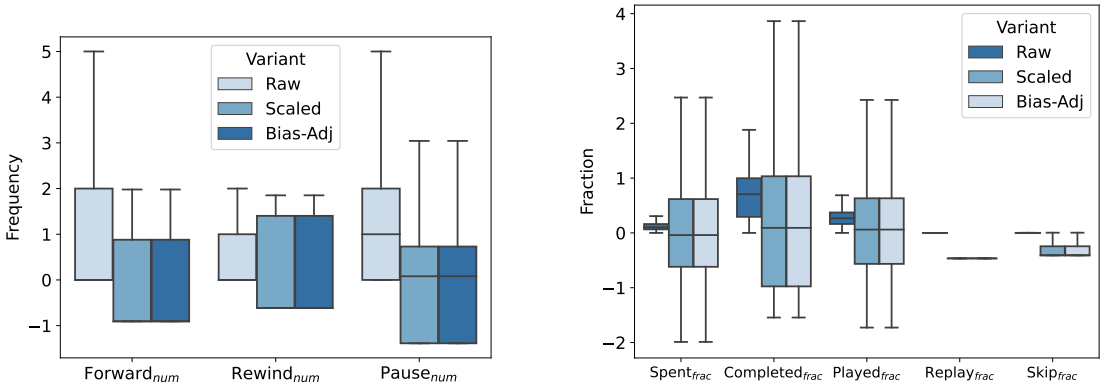
Furthermore, the in-video viewing features are scaled to approximate a Gaussian distribution. Due to most features' highly non-normal distribution, they are approximately standard scaled using the non-linear Yeo-Johnson method [122] from Scikit-Learn[2] [123], which attempts to maximise the goodness of fit of the individual features by raising them to the power of a λ for each feature. Due to most features' long tails, an interquartile-range-based scaling method was attempted, but the preliminary results did not indicate a drastically more Gaussian-like distribution and were therefore not used. For the bias-adjusted evaluations, the in-video features were individually adjusted using Equation 4.1, before too scaling them using Yeo-Johnson.[3] Missing features, both categorical and continuous, were replaced with zeros. As the SARS required a fixed length side-information input as well, only the 10 first related concepts and fields were kept for MOOCCubeX as it is the median length as shown in Table 4.2. Interactions with lectures which have less than 10 related concept-field mappings are padded with zeros as well [80]. Lastly, for the sequential models, each user's interaction history with the corresponding side information is aggregated into a list for each feature, ordered by the global timestamp. For the conventional models, the implicit user-lecture matrix is created for each of the data splits, where the entries are binary indicating whether or not the user has interacted with the given lecture or not.

The Figures 5.2a and 5.2billustrate the effect of the scaling and bias adjustments on representative in-video viewing features using EdNet as an example. Due to the long tails of the distributions, which are highlighted in Figure 5.3, outliers are not included to make the visualisations informative in Figure 5.2.

As illustrated in these figures, the scaling in some cases improves the variance, as the non-scaled feature visualisations falsely have **dense** interquartile ranges due to the long tails, as visible in Figure 5.3. However, the tails are still long despite the scaling effort and the features are less explainable and intuitive for further analysis. Lastly, there is no visual difference between bias-adjusted and scaled features versus the not-bias-adjusted, scaled ones.

[2] `https://scikit-learn.org/`

[3] To avoid any model bias, the scaling estimators for the raw and bias-adjusted datasets respectively were fitted on the training data.



(a) Distributional effect of the preprocessing steps on discrete in-video viewing features

(b) Distributional effect of the preprocessing steps on discrete in-video viewing features

Figure 5.2: Box plots illustrating the distributional effects on discrete and continuous in-video viewing features respectively.

Algorithms

The following algorithms which were used in the evaluation, are all implemented in Python [124]. The CF methods are implemented using the *implicit* library[4], while the SARSs are implemented using the Transformers4Rec framework[5] [80].

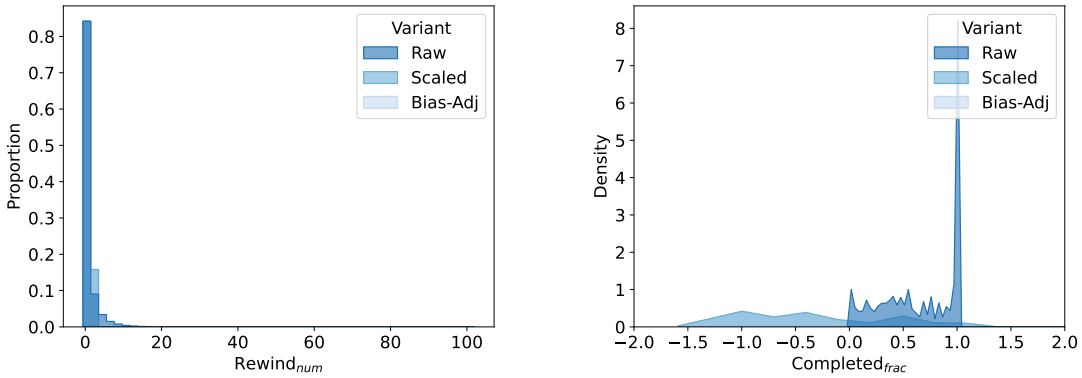
Random: Naïve baseline, emphasises the differences in prediction space size and establishes the minimum performance to expect. Implemented using PyTorch [125]

MostPop[126]: Also a naive, non-personalised baseline recommending lectures ranked by their total number of interactions. It illustrates a simple, easily implemented method, in addition to how skewed the given dataset is towards popularity. Implemented using PyTorch[6] [125]

Syllabus [23]: Based on the last user lecture session, recommend the next consecutive lectures in the current course's syllabus. The method is only applicable to MOOCCubeX as EdNet does not contain multiple courses. Further on, the syllabus information in MOOCCubeX is very sparse. Based on the preprocessed dataset, only 10.4% of the lectures have a known syllabus mapping. Those lectures make up 6.66% of the total user lecture sessions. Therefore, missing syllabus entries or incomplete syllabi are replaced with the *MostPop* recommendations. Though the syllabus entries have "chapters", the naming of them is not consistent. Therefore the order of the recommended syllabus resources in the dataset is not altered and is utilised as is. Implemented using PyTorch [125].

implicit Alternating Least Squares (iALS)[59]: It is a conventional MF method proposed to address the challenges of implicit feedback datasets. The learned latent user and lecture embeddings are updated by an iterative approach using the implicit Alternating Least Squares lossIn addition to being a common model-based CF baseline, recent work has shown that it is still a relevant recommendation method outperforming several recent and more complex

[4] `https://benfred.github.io/implicit`
[5] `https://nvidia-merlin.github.io/Transformers4Rec/main`
[6] `https://pytorch.org`



(a) A histogram of the effect of scaling, and bias adjustments and scaling on Rewind$_{frac}$

(b) An Estimated kernel density visualisation of the effect of scaling, and bias adjustments and scaling on Complete$_{frac}$

Figure 5.3: Histogram and Kernel density estimations are visualised for a representative discrete and continuous viewing feature respectively

models if configured correctly [127].

Bayesian Personalised Ranking (BPR)[128]: In difference to iALS, BPR is a MF method optimised for ranking explicitly, rather than implicitly through the rating prediction problem. This is achieved by using a *pair-wise* loss, where the training samples are triples $(u, l^{(u)}_{pos}, l^{(u)}_{neg})$, where , $l^{(u)}_{pos}$ is a positive item, i.e. lecture viewed by user u, and $l^{(u)}_{neg}$ is a *negative* item, i.e. not viewed by user u. Using these triples, the objective is to provide the user with a personalised total ranking by creating a *total order* $> u$ of all of the lectures by reconstructing it through the pairwise preference comparisons between the lectures. BPR is one of the most commonly used baselines and has illustrated reproducible outperformance of several other MF methods, both conventional and neural [121].

Logistic Matrix Factorisation (LMF)[129]: A probabilistic MF method for implicit feedback, calculating the probability of interacting with an item given the logistic relations. It was first evaluated in the space of artist recommendation, where listening behaviour is highly skewed. The method has been shown to be effective and is included due to its applicability to domains where re-consumption behaviour is relatively common. The intention is therefore that it will give another perspective than the other MF methods for learning resource recommendation where re-consumptions are interesting.

K-Nearest Neighbours: A conventional memory/neighbourhood item-based CF model, as described in Section 2.3, utilising the k most similar lectures by each user's ratings to predict recommend relevant lectures. In this implementation, the implicit feedback is binary, i.e. if the user has interacted with the item or not. To calculate similarity, three similarity metrics were considered: cosine, TF-IDF [93] and BM25 [130] due to their shown effectiveness from information retrieval to balance informative and less informative interactions, as well as to handle varying sequence lengths, i.e. document length in information retrieval.

The following models are SARS and are implemented as a part of the model architecture described in Section 5.1.1 and thereby support the inclusion of side information. Therefore

they are used for experiments Experiment 1.2 and 1.3. The model names indicate what type of sequential layer is used in the RS architecture.

GRU$_{T4R}$[80]: The model is inspired by *GRU4Rec* and its extension *GRU4Rec+* [131], which have been state-of-the art in session-based recommendation. [80] illustrated that GRU$_{T4r}$ performs comparatively to both GRU4Rec and transformer-based RSs, dependent on the dataset and metric. Though contextualised RNNs have been explored in other domains [76], the effect of self-supervised training on contextualised GRUs-based models were not displayed in [80]. Due to the restrictions of GRU's architecture, CLM was used.

BERT$_{T4R}$[80]: The BERT$_{T4R}$ used in [80] is similar to the original [120]. It is included as it is currently assumed to be state-of-the-art for sequence-aware recommendation modelling. The main difference to the self-supervised training approach in [120], is that no last-item-only test samples are included in the training of the model to better align with the evaluation task, as it is trained with MLM.

XLNet$_{T4R}$[80]: As proposed by [80] to use XLNet [55] as the sequence model, it was shown to generally be the best transformer module to use, both with and without side-information. As in [80], the model is trained using a MLM approach, in contrast to the original, *autoregressive* formulation of the pretraining objective used in natural language processing tasks [55]. Regarding the sequence model XLNet, some of the differences to BERT [53], besides the original pretraining approach, is that it is an encoder-decoder model which is not limited to fixed length sequences which BERT is.

Data splitting and model configuration

Following recent sequence-aware evaluation methods [84, 120, 132, 133], a temporal Leave-One-Out (LOO) strategy was used for data splitting. More concretely, the sessions until time step $n_u - 2$ for each user u were used as the *training set*, where the $l^{(u)}_{n_u-1}$ session,i.e. the second most recent user lecture session for each user, was used for *validation* and hyperparameter tuning. The most recent session $l^{(u)}_{n_u}$ was held out for each user, creating the *test set* to evaluate the model's recommendation accuracy.

To find the optimal hyperparameters for each non-naive model, several hyperparameters are jointly explored and tuned on both datasets individually. For the conventional models, the number of training epochs is included as a hyperparameter, while the SARSs are trained for 10 epochs as in [80], with an early stopping criteria with *patience* of 5 and a stopping threshold of 1×10^{-3}. Individual searches are done for each model variant. Following [80, 132, 133], NDCG is used as the optimisation metric, with a cut off at 10. Due to the larger search spaces, a Tree-Structured Parzen Estimator (TPE) sampler is used for a directed hyperparameter search [134], where each model's hyperparameter search includes 150 trials. In addition to the model-specific hyperparameter studies, a search for the optimal subset of viewing features is done using the optimal hyperparameters for the XLNet-model not-biased adjusted, enriched variant XLNet$_{feat}$. The hyperparameter search is implemented using the Optuna[7] library [135]. The training and validation sets are used for the hyperparameter tuning.

As part of the embedding layer in Figure 5.1, each of the individual viewing features was *layer normalised* as suggested by [80], as well as SOHE as it was deemed crucial to improve

[7] `https://optuna.org/`

the recommendation accuracy. Furthermore, the output was projected to the input item space using *weight tying* as in [80] as a regularisation and parameter reduction technique. For each of the models, a label smoothed cross-entropy loss was used, where zero smoothing is equal to regular cross-entropy [80]. Other model configurations such as which optimiser to use, besides the hyperparameters searched, were left to their default values. In addition, any further reference to the sequence-aware models is used without the subscript $_{T4R}$.

Evaluation

To account for the randomness provided in the weight initialisation for both MF and neural models, each of the model-based RSs is evaluated on the same ten different, randomly selected seeds. Using the found hyperparameters, the conventional models are trained for the optimised number of epochs. The SARSs' optimal number of epochs is decided through training the model variant with the found hyperparameters, on the training set with the same early stopping and model selection criteria as in the general hyperparameter tuning. The sequence-aware model is then fitted on the validation set for the given number of epochs and is evaluated on the test set, for each of the different seeds. It has been common in recent and recognised works to use "sampled metrics" [83, 120, 136] for evaluation due to the cost of evaluating. In the case of LOO means that for each relevant lecture, m negative, not-interacted with, lectures are sampled and ranked together, intended to give a performance boost as large item spaces can be expensive to evaluate. Recent studies have shown that the sampled versions of metric not only mislead the magnitude of models' recommendation accuracy, but the comparative performance between models might also be invalid [68, 133, 137]. Therefore each relevant lecture is ranked against the entire lecture set for both datasets, using the *full*-version of the metric. To evaluate the recommendation accuracy, the ranking metrics NDCG, MAP and Recall are used with cut-offs at 5 and 10 since this work is mostly interested in top-ranked lectures. While *recall* will measure the models' ability to generally retrieve the relevant lecture, NDCG measure the models' ability to rank it correctly. Lastly, MAP more strictly penalises incorrect ranking than NDCG [68], so it is included to further emphasise the models' ranking ability.

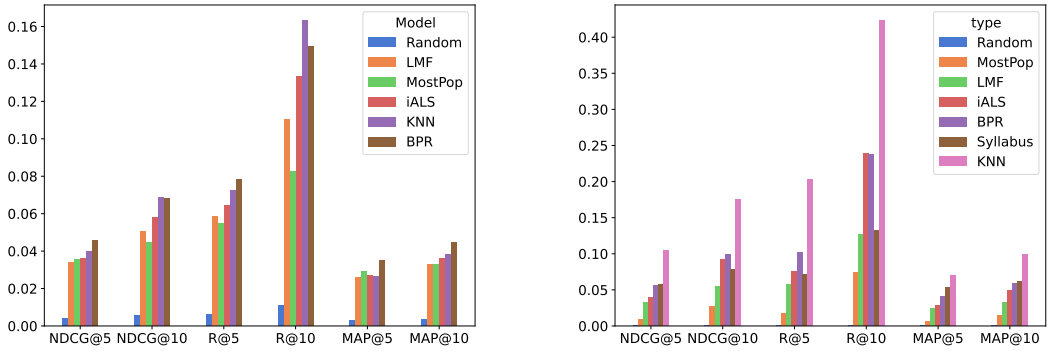
To quantify any statistical improvements or declines in recommendation accuracy, the Wilcoxon signed rank test [71] is used for each metric, with a level of significance of $\alpha = 0.05$. The paired testing is due to the relatedness of the results by evaluating them with identical seeds. As only ten evaluations are available for each model, an assumption of normality is less likely, as well as the independence of the results as they are evaluated on the same datasets, which violates the assumptions of a two-sample, paired Student's T-test [138]. As multiple metrics are tested and thus multiple hypotheses, the FWER is controlled by applying the Holm-Bonferroni correction [74]. Within the baseline experiments, experiment 1.1, the best-performing model and the second-best model per metric are tested. Furthermore, for each of the SARS, the fully enriched models are compared to their respective base variants, whereas the bias-adjusted variants are compared to both the enriched and the base variants to establish any potential benefit of the two enriching methods. Lastly, XLNet$_{feat}$ is not included in the significance testing. The used tests are implemented in SciPy [139] and the Holm-Bonferroni-correction in *statsmodels*[8] [140].

[8] `https://www.statsmodels.org/`

5.1.4 Results

EdNet

The recommendation accuracy results of EdNet are presented in Table 5.2, where it is clear that the conventional CF methods, either neighbourhood (KNN) or MF, perform similarly to a non-personalised *MostPop* baseline. One exception is for R@10 where the CF methods perform drastically better. The baseline results excluding the SARS are presented in Figure 5.4a. For the SARS models without side information, the table shows clearly that they outperform the non-sequence-aware methods by a large margin. More specifically, the XLNet consistently outperforms the others, though not by a statistically significant margin. Another note is that the BERT-based model performs drastically worse than both the XLNet and GRU models.



(a) EdNet baselines' comparison

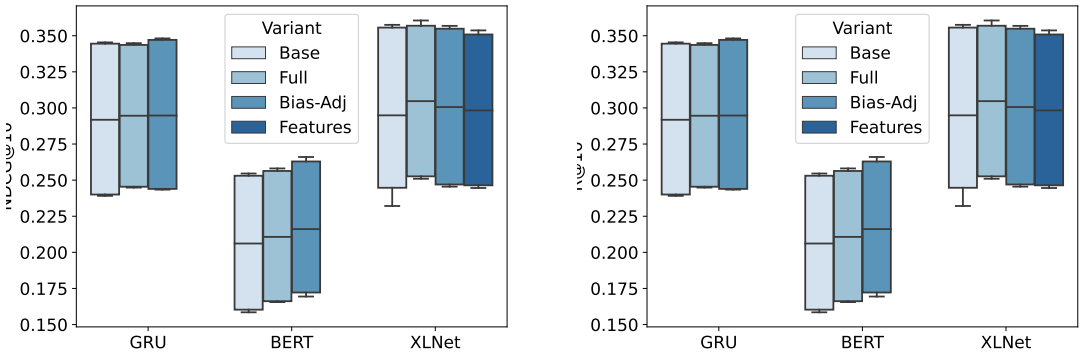
(b) MOOCCubeX baselines' comparison

Figure 5.4: Baselines comparison on various ranking metrics

For the models which incorporate side information, all three perform statistically significantly better than their non-enriched counterparts except for GRU on R@10. Moreover, the enriched XLNet-variant consistently outperforms the enriched GRU on all metrics, whereas BERT strongly underperforms the other two. The average relative change across the metrics for BERT compared to the base variant is 1.4% - 5.5%, while it is 1.2% - 5.5% and -0.2% - 3.9% for XLNet and GRU respectively. Furthermore the optimally found feature subset by the feature selection method described in Section 5.1.3 for XLNet, denoted as XLNet$_{feat}$ does not perform better than using all of the features.

Lastly, the results of the model variants incorporating user behavioural bias-adjusted features are more nuanced. Generally, all three models perform statistically significantly better than their base variants, except for XLNet on R@10. Compared to the enriched, not adjusted variants, the bias-adjusted models perform similarly, or worse. BERT is the only model that has statistically significant, consistent improvements over its biased counterpart. Both XLNet and GRU show a statistically significant decline across all metrics compared to the bias variants, except for GRU on R@5. Specifically, GRU's average relative difference over the full version is -1.6% - 0.8%, compared to XLNet and BERT which has improvements of -2.8% - -0.8% and 2.4% - 3.6% respectively.

The box plots in Figure 5.5 better illustrates the ranges of NDCG@10 and R@10 for the SARS variants. From the diagram, it is clear that BERT benefits from the side information. For XLNet



(a) NDCG@10 variance by model variant

(b) Recall@10 variance by model variant

Figure 5.5: Box plot of EdNet's recall and NDCG variance by SARS

and GRU, the benefit is less obvious. Moreover, the standard deviation of bias-adjusted variants is not consistently lower than their comparative base variants, or their biased counterparts.

MOOCCubeX

For the non-enriched models, the syllabus-based method performs drastically better than MostPop, and even better than the MF methods on MAP and NDCG@5 as visible in both Table 5.3 and 5.4b. In contrast to EdNet, the CF methods perform drastically better relative to MostPop, while KNN results' are 2-4 times higher than the results of the MF methods on some metrics. Out of the MF methods, LMF performs much worse than iALS and BPR on recall and NDCG, particularly with a cut-off at 10. KNN's out-performance is illustrated in Figure 5.4b compared to the other sequence unaware models. Regarding the SARS without side-information, GRU out-performs XLNet by a statistically significant, large margin, except for on R@10. As with EdNet, BERT consistently underperforms the other sequence-aware models.

Examining the enriched model variants, the GRU based model performs drastically better than enriched BERT and XLNet in contrast to with EdNet. Moreover, the feature subset optimised XLNet_{feat} performs relatively better than the *full* XLNet variant, though only slightly. As with the base variants and in EdNet, BERT performs drastically worse than the other two SARS. Comparing the enriched models to their base variants, they all perform statistically significantly better than their baseline counterparts, except for XLNet on R@10. Moreover, the average relative improvement of the enriched BERT compared to the base variant is 2.7% - 5.1%, while it is 0.1% - 1.8% and 0.9% - 1.4% for XLNet and GRU respectively.

The relative performance between the bias-adjusted models is more nuanced, where GRU performs relatively better than BERT and XLNet, except for on R@10. Again, BERT drastically underperforms. More interestingly, GRU has statistically significant *improvements* over the base variants, but statistically significant *decline* compared to the not-biased adjusted version. BERT on the other hand has statistically significant improvements to the respective base and non-bias adjusted counterparts. XLNet's results are more mixed, where the only statistically significant difference is for R@10. Specifically, it performs slightly worse than its fully, enriched variant, and slightly better than its base variant. Put in numbers, the bias-adjusted XLNet's average relative change over the not-bias-adjusted variant is -1.0% - 1.1%, while it is 3.4% - 5.8% for

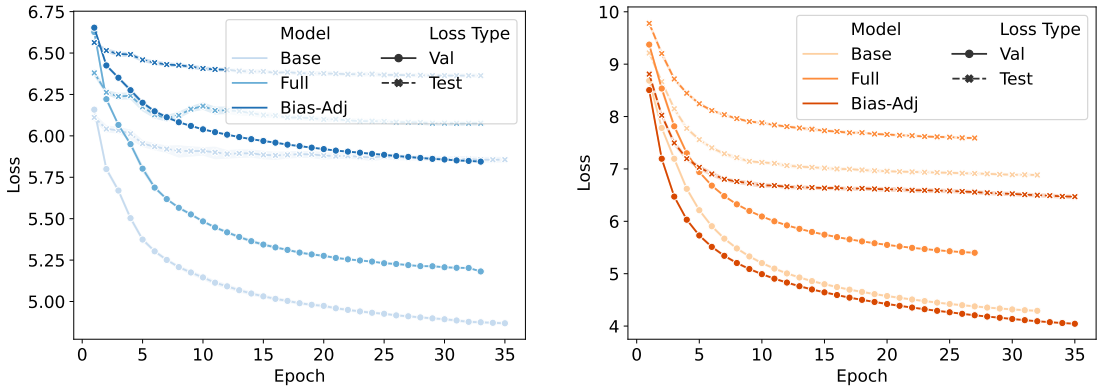
Table 5.2: Results of RQ1, RQ2a and RQ2b for EdNet, where the *Baselines* section contains the models without side-information, the *Full* for methods *with* side-information and *Bias-Adj* section for the models with user bias adjusted side-information. * indicates *not* statistically significant difference second best-performing baseline in the given metric. † indicates *not* statistically significant compared to the base version of the model.‡ indicates *not* statistically significant compared to the full version of the model, evaluated with a significance level of $\alpha = 0.05$ using a Holm-Bonferroni corrected Wilcoxon signed rank test. The best result for each metric per section is highlighted in **bold**, while the second best result is underlined.

		NDCG@5	NDCG@10	R@5	R@10	MAP@5	MAP@10
Baselines	Random	0.0038	0.0054	0.0062	0.0112	0.0031	0.0037
	MostPop	0.0354	0.0444	0.0550	0.0827	0.0290	0.0328
	iALS	0.0362	0.0582	0.0645	0.1333	0.0270	0.0360
	LMF	0.0340	0.0506	0.0587	0.1104	0.0260	0.0327
	BPR	0.0457	0.0684	0.0785	0.1491	0.0352	0.0444
	KNN	0.0396	0.0686	0.0727	0.1634	0.0266	0.0383
	GRU	0.2163	0.2399	0.2714	0.3445	0.1982	0.2079
	BERT	0.1378	0.1603	0.1827	0.2528	0.1231	0.1323
	XLNet	**0.2173***	**0.2427***	**0.2746***	**0.3534***	**0.1985***	**0.2089***
Full	GRU	0.2232	0.2455	0.2747	0.3439†	0.2062	0.2153
	BERT	0.1447	0.1663	0.1889	0.2564	0.1302	0.1390
	XLNet	**0.2278**	**0.2522**	**0.2819**	**0.3578**	**0.2100**	**0.2200**
	XLNet$_{feat}$	0.2221	0.2461	0.2765	0.3512	0.2041	0.2139
Bias-Adj	GRU	0.2206	0.2439	0.2744‡	0.3468	0.2029	0.2124
	BERT	0.1495	0.1717	0.1935	0.2627	0.1351	0.1441
	XLNet	**0.2226**	**0.2471**	**0.2791**	**0.3550†**	**0.2040**	**0.2140**

BERT and -2.8% - -0.8% for GRU.

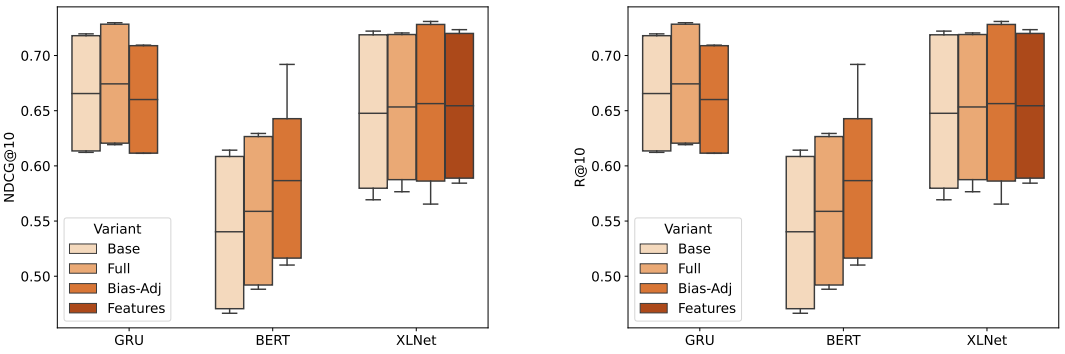
The ranges of performance for the SARSs is illustrated in Figure 5.7, where it is clear that XLNet's performance is less stable, while GRU's performance is. Moreover, it highlights the improvements of the bias-adjusted variants over the base and not-adjusted variants for BERT and XLNet, but less for GRU, even a negative effect in terms of recall. As with EdNet, the box plots demonstrates that the bias-adjusted models do not generally reduce the variance in the measured metrics compared to their respective variant counterparts.

Lastly, to evaluate the hyperparameter-tuned number of epochs and the different seeds' effect on model training, the evaluation and test loss are visualised for the variants of BERT in Figure 5.6 for both datasets. From the plots, the models generally overfit from early on, without a general degradation of the test loss, but slightly improving. Moreover, the models use less than half of the given training budget. Another note is that the enriched variants converge more quickly than the others, which is particularly evident on MOOCCubeX. Lastly, the uncertainty provided by the 95% interval is little and only strictly visible for the test loss of bias-adjusted BERT on EdNet. The other models' loss plots are available in Appendix B.1, where the same trends are present, except that the enriched GRU converges more slowly than the other variants. To summarise the overall results, the number of improvements or declines comparing the various variants for both datasets are presented in Table 5.4.



(a) BERT validation and test loss on MOOCCubeX **(b)** BERT validation and test loss on MOOCCubeX

Figure 5.6: The mean and 95% confidence interval of validation and test losses across each seed evaluation for the variants of BERT.



(a) NDCG@10 variance by model variant **(b)** Recall@10 variance by model variant

Figure 5.7: Box plot of MOOCCubeX's recall and NDCG variance by SARS

5.1.5 Discussion

The results related to the individual research questions RQ1a-c and how they are addressed are discussed below, as well as the limitations of the experiment.

Research Question 1a: Baseline recommendation accuracy

Firstly considering *RQ1a*, the dominance of SARS in other fields [80, 120, 133] is also present in large-scale, learning resource datasets. This further indicates that there are relevant, sequential patterns when interacting with educational videos which are not captured by time-unaware models like conventional CF methods. More interestingly, despite EdNet having much lower sparsity, fewer unique lectures and in general longer user interaction sequences, the general recommendation accuracy is much lower, especially regarding the sequence-aware models as for instance XLNet has an average NDCG@10 of 0.5786 on MOOCCubeX, it has only 0.2173 on EdNet. One difference which might be a factor is that MOOCCubeX has additional playback rate features, which are not available for EdNet, further enriching in-video viewing behaviour.

Table 5.3: MOOCCubeX results for RQ1, RQ2a and RQ2b. * indicates *not* statistically significant over the second best-performing baseline in the given metric. † indicates *not* statistically significant compared to the base version of the model.‡ indicates *not* statistically significant compared to the full version of the model. The test was evaluated using a significance level of $\alpha = 0.05$ with Holm-Bonferroni corrected Wilcoxon signed rank test. The best result for each metric per section is highlighted in **bold**, while the second best result is underlined.

		NDCG@5	NDCG@10	R@5	R@10	MAP@5	MAP@10
Baselines	Random	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
	MostPop	0.0089	0.0274	0.0177	0.0743	0.0061	0.0139
	Syllabus	0.0580	0.0777	0.0718	0.1322	0.0535	0.0618
	iALS	0.0394	0.0920	0.0757	0.2394	0.0278	0.0494
	LMF	0.0320	0.0542	0.0569	0.1264	0.0239	0.0330
	BPR	0.0559	0.0996	0.1013	0.2371	0.0413	0.0593
	KNN	0.1046	0.1754	0.2035	0.4235	0.0698	0.0986
	GRU	**0.5989**	**0.6134**	**0.6738**	0.7182	**0.5739**	**0.5799**
	BERT	0.4486	0.4704	0.5425	0.6095	0.4174	0.4265
	XLNet	0.5588	0.5786	0.6577	**0.7186***	0.5258	0.5341
Full	GRU	**0.6047**	**0.6204**	**0.6801**	**0.7285**	**0.5795**	**0.5861**
	BERT	0.4705	0.4915	0.5623	0.6266	0.4400	0.4487
	XLNet	0.5672	0.5855	0.6627	0.7190†	0.5353	0.5429
	XLNet$_{feat}$	0.5705	0.5887	0.6644	0.7202	0.5391	0.5467
Bias-Adj	GRU	**0.5981**	**0.6115**	**0.6678**	0.7090	**0.5748**	**0.5804**
	BERT	0.4965	0.5174	0.5840	0.6484	0.4673	0.4760
	XLNet	0.5629†‡	0.5842†‡	0.6617†‡	**0.7268**	0.5300†‡	0.5389†‡

Another possible factor is a difference in the path variety between the two datasets because of their differences in lecture accessibility. While each user has access to every video in the video corpus for EdNet, users in MOOCCubeX only have access to the lectures of courses they are enrolled in. Because of the higher availability, combined with higher general lecture coverage, EdNet's users might have a larger variety in the different interaction paths than for MOOCCubeX and therefore the sequence-aware models have more difficulty extracting relevant sequence patterns. Though this must be verified, through for instance measuring the path entropy [23]. Furthermore, it does not explain the discrepancy for the CF methods applied to the datasets.

Moreover GRU, a less complex model than XLNet, achieves similar performance and drastically better performance on MOOCCubeX. Due to relatively short sequences, the benefit of using transformers with regard to maintaining long-term memory and avoiding vanishing gradients [131] is reduced, which could be part of the explanation. Furthermore, the GRU-based model is consistently worse than XLNet for EdNet, which is in line with this hypothesis as the sequences are generally longer.

An interesting note on the *Syllabus*-based method for MOOCCubeX is that despite the sparsely populated syllabus information, it vastly outperforms the *MostPop* baseline, with 2-10 times improved ranking accuracy. As it is likely that a user will return at some point to the lectures in the syllabus, higher cut-off metrics like 10 is less indicative of a linear syllabus-based learning

Table 5.4: The total number of improvements or declines comparing the side-information enriched to the base SARS (*Full2Base*), the bias-adjusted to the enriched (*Bias2Full*) and the bias-adjusted compared to the base versions (*Bias2Base*) across GRU, BERT and XLNet and all metrics. The number in parentheses denotes the number of statistically significant results for a significance level of $\alpha = 0.05$, measured by a Holm-Bonferroni corrected Wilcoxon signed rank test.

Variant Comparison	**EdNet**		**MOOCCubeX**	
	# Improvements	#Declines	# Improvements	#Declines
Full2Base	17 (17)	1 (0)	18 (17)	0
Bias2Full	7 (7)	11(10)	7 (7)	11 (6)
Bias2Base	18 (17)	0	14 (9)	4 (4)

approach. On the other hand, it partially contradicts the findings of [116] which indicated that students do not follow the syllabus. The fact that Recall@10 is not ≈1 for the *Syllabus*-method, indicates that users to some degree do interact with different courses intermittently, emphasising the importance of considering user behaviour on a platform basis, rather than a per-course basis. Notably, to which degree the context-switching between lectures occur is not accurately quantified, though one factor is that the recommendation accuracy of *Syllabus* is naturally low when recommending unavailable *MostPop* lectures and only 7.15% of the target lectures in the evaluation are included in a syllabus.

Lastly, the general underperformance of BERT relative to the other on both datasets is interesting as previous work did not experience significant differences [80] [133]. Though it has been illustrated that BERT have often been under-fitted when evaluated and used as a baseline [133], this is not the case as was illustrated in 5.6 where the over-fitting is clear. Notably, the loss functions indicate that it is the variance of the data which is not properly explained by the model which could be improved by stronger regularisation. This is supported by that the found hyperparameters generally had no or very low dropout rates, of both inputs and within the sequential layers. This further highlights issues with joint, sampled hyperparameter search as dropout was not deemed relevant by the sampling strategy. Other factors for the unexplained variance, might be a less heterogeneous lecture selection for new platform users and that most sequences are relatively short, as well as the misaligned between the training objective of MLM and next-lecture prediction [132]. Moreover, the effect of only using a single evaluation for determining the optimal number of epochs per model variant is minuscule as the .95% confidence interval is barely visible in the visualisations. On another note, as BERT was given the same training budget as XLNet, and GRU, its worse overall performance indicates that BERT is generally less efficient to train as explored in [132, 133], and motivates further exploration of other attention-based models and transformers in particular for recommendation as they evolve.

Research Question 1b: Side Information Recommendation Accuracy

Regarding *RQ1b*, the results show that there is a general, statistically significant improvement of the recommendation accuracy of side-information enriched Sequence Aware Recommendation System. Notable, the improvements are relatively small with a maximum average improvement of 5.5%. The improvement comes at the cost of increased model complexity as more features are included. Moreover, the data collection and processing pipelines for a de-

ployed RS are increased, as well as concerns regarding the privacy of users making it less applicable to real-life scenarios. Furthermore, using additional computing for finding interactions between the side-information, both continuous and aggregated nor using pre-trained embedding was not found to be consistently better in the hyperparameter search. This could partially be explained by [78], illustrating how inefficient first order FFNs are for extracting low-rank interactions. Of the evaluated sequence-aware models, BERT illustrated the largest relative improvements over the respective base variants. The reasons for this are unknown, but as the evaluation and test loss in Figure 5.6 illustrates, it converges at approximately the same rate as the base variant in both cases, but the relative learning ability and generalisation ability differ between the datasets and the respective variants.

Some of the limitations of the side-information inclusion experiments are that the impact of pre-trained embeddings and additional computation is not further quantified. Neither is the general impact of including knowledge-related topics combined with in-video behaviour of lectures for inferring user's navigational behaviours, i.e. how it relates to the lectures they choose to watch. Moreover, even though feature selection exploration was done, a more thorough ablation study could be explored to better quantify the effect of individual features and their potential relation to lecture topics.

Research Question 1c: Bias-Adjusted Recommendation Accuracy

The improvements in adjusting users' behaviour according to their time-aware calculated bias are less clear. While bias adjustments of the in-video features for EdNet are consistently better than the base variant, that is not the case for MOOCCubeX, where the bias-adjusted GRU overall has a statistically significant decline in performance. Moreover, the BERT and XLNet's improvements are less statistically significant, in terms of frequency, than the not-bias-adjusted versions compared to their base variants. When comparing the accuracy of the bias-adjusted variants to the enriched ones, the only model which has overall improved recommendation accuracy is BERT for both datasets, where the results of the other models are mostly statistically significant declines, though less than a relative decline of -2.8%. This indicates that there is no additional gain by adjusting for the user bias regarding to improving the quality of recommendations. On the contrary, the improvement of BERT compared to the not-bias-adjusted versions can indicate improvements in the convergence rate for BERT, although the scaling effects illustrated in Figure 5.2 and 5.2b illustrate long tails and not necessarily a Gaussian-like distribution of every feature, making the training more. Another explanation for the relative declines of GRU and XLNet, is that in the case of users with consistent, active, interaction behaviour, the bias-adjusted features will become more sparse than the not-bias-adjusted features. As a consequence, the information provided by these features might therefore be lost and the behaviour becomes indistinguishable from passive, user behaviour.

Therefore a limitation is that which features to adjust for users' bias and the impact on recommendation accuracy of bias adjusting individual features have not been thoroughly explored. As presented in Section 4.2.2, some features exhibit a larger degree of user bias than others and consequently adjusting them could be more informative when comparing abnormal behaviour by these features. On the other hand, not adjusting for user bias altogether excludes an additional preprocessing step when the results indicate that *not* adjusting for bias is better for sequence-aware models. Furthermore, an analysis of the recommendation quality partitioned by users, lectures or topics was not studied. For instance, clustering the users by in-video inter-

action behaviour, could provide further insights and quantify to which degree different types of behaviours affect the recommendation accuracy.

Limitations of Exp.1

In addition to the limitations mentioned in the previous sections, the experiment has some more general limitations. Firstly, the relevancy of a lecture is only decided by a single target lecture which is the most recently viewed one in the users' history when evaluating the recommendation accuracy. For instance, it does not consider how long the user viewed the target lecture for measuring its relevancy. If a user briefly viewed a lecture before moving on to the next, it could indicate a less relevant lecture than a lecture viewed all the way til the end. Furthermore, even though the user only viewed the targeted lecture, it does not necessarily mean that the other lectures are *not* relevant to some extent. A more continuous and nuanced relevancy measure could be needed, perhaps based on lecture similarity or by lectures viewed in the near future, which could have a degree of relevancy at the time of recommendation as they viewed it later.

For the preprocessing and evaluation methodology, a limitation is that the hyperparameter tuning is done jointly, in a very large search space where some parts of it have major model impacts, such as the number of sequential layers, e.g. BERT-layers, or to include additional continuous feature projections. Though time-consuming, an iterative hyperparameter tuning, with smaller search spaces could be beneficial to provide more confidence that the optimal hyperparameters were found [127]. Moreover, as some work on fixed history lengths SARS, uses the N most recent actions [61, 80], which might indicate a lower lecture viewing diversity as users often have the same starting points and then migrate to other topics or courses [43]. Moreover, limiting the sequences to only the 30 first actions do exclude almost 25% of the user lecture sessions of EdNet, which could be avoided by including longer sequences. The implications of using longer sequences on recommendation accuracy for these datasets are uncertain, as larger parts of the sequences would be padded. Moreover, neither model complexity nor runtime is considered in the evaluation in detail, which in real-life applications would have often favour the conventional models CF. [61]. Moreover, although the models are evaluated across ten different seeds, there is only one test set which is evaluated which can bias the results. Although this has been a common leave-on-out evaluation technique for SARSs, a user-based re-sampling method to generate multiple test sets as proposed in [61] could provide more confidence in the results, likely without too large of a loss in training data as the datasets are relatively large.

5.2 Experiment 2 - Re-consumption Behaviour

In this section, the necessary prerequisites to execute the set of experiments in The specific details of Experiment 2 are described in this section, related to the problem definition, experiment-specific preprocessing and evaluation methods. The section is concluded with the results of the experiments, with a corresponding discussion of how they address *RQ2* and the limitations of the experiment.

5.2.1 Problem definition

For the problem of predicting whether or not a lecture will be re-consumed by a user, the formal definition of a lecture re-consumption is stated. Given a user u with a historic user lecture session sequence $\mathcal{S}_u$, a lecture l is re-consumed if it occurs at least twice in $\mathcal{S}_u$. For each lecture l and user u, the first visit of a lecture is kept and labelled as re-consumed if it occurs later in the sequence or labelled as *not* re-consumed if its the only view of the given lecture. Considering a target label y and an observed in-video viewing behaviour represented as $\mathbf{x} = [x_0, x_1, \ldots, x_f]$, consisting of f different viewing features, the prediction problem can be formalised as estimating the conditional probability

$$Pr(Y = y|\mathbf{X} = \mathbf{x}; \theta), \tag{5.2}$$

where θ is the model specific parameters.

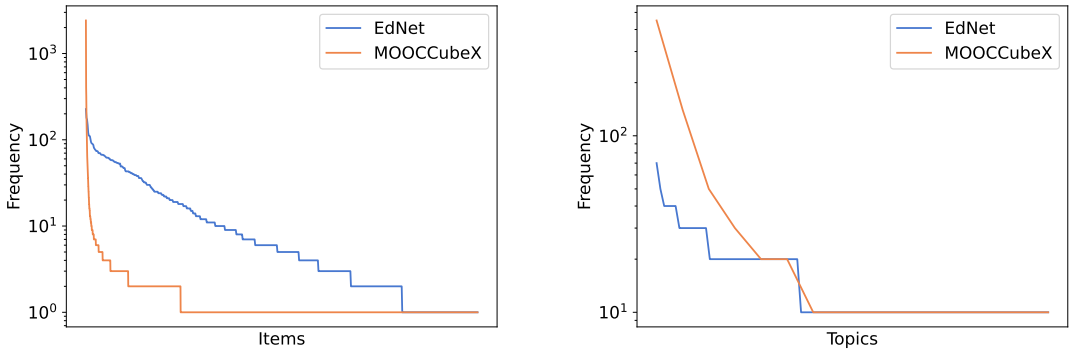
5.2.2 Experiment setup

Firstly the additional, experiment preprocessing steps are specified for each of the specific subquestions proposed for *RQ2a* and for creating a re-consumption prediction dataset. Then the selected algorithms to use for the prediction problem are described, as well as the data splitting methods and model configurations and lastly the evaluation methods for both the individual subquestions and the prediction problem.

Preprocessing

To answer the proposed research questions RQ2a and RQ2b regarding lecture re-consumption analysis and prediction, further preprocessing steps were taken to accurately compare behaviours. More concretely, only the 30 first sessions of users with at least five sessions are examined as in Exp.1. Specifically, EdNet contains 274,396 *novel* sessions, i.e. where the lecture is not re-consumed by a user and 30,358 re-consumption sessions. MOOCCubeX contains 1,284,845 novel sessions and 176,839 re-consumption sessions. The resulting set of datasets is further referred to as $\mathcal{D}_{rep}$

For the last question related to RQ2a regarding viewing behaviours' relation to re-consumption, a subset of the sessions are considered due to individual users revisiting preference [26]. Firstly, to compare viewing behaviour differences, only the first visit and the following re-consumption is compared, in contrast to in [17], as the behaviour might change with the number of re-consumptions. For instance, a user will likely skip larger parts of a lecture the fourth time it views it versus the second as the viewing intentions are likely different, e.g. general revision versus seeking out specific segments. Moreover, due to previously reported user and lecture-specific viewing behaviour [25] and as shown in Section 4.2.2, these should



(a) Item frequencies in EdNet and MOOCCubeX before item correction, on a logarithmic scale.

(b) Topic distributions in corrected downsampled dataset for EdNet and MOOCCubeX, tags and most frequent field respectively

Figure 5.8: Frequencies of items and topics in EdNet and MOOCCubeX, before and after item correction

be corrected. Therefore, for each user, a single, random first and second-time lecture viewing pair is sampled to account for any one user's bias[9]. Considering these viewing pairs, Figure 5.8a illustrates the high lecture popularity bias of the sampled viewing pairs. Therefore, for each of the present lectures in the viewing pairs, only ten per lecture is considered, disregarding lectures less frequently occurring and downsampling lectures which are more frequently occurring. Though ten is relatively arbitrarily chosen, it provides an estimation of behaviour related to intrinsic lecture properties. The effects of the respective user and item correction is presented in Table 5.5, where $\mathcal{U}^{\sigma}_{revisit}$, $\mathcal{I}^{\sigma}_{revisit}$ and $\text{Topic}^{\sigma}_{revisit}$ denotes the standard deviation of the number of revisits per user, lecture and topic respectively, i.e. per *tag* in EdNet and per *field* in MOOCCubeX. As presented in the table as well as in 5.8b, the number of topics is not adjusted for. The set of datasets are further reffered to as $\mathcal{D}_{pred,adj}$

Table 5.5: Results of downsampling for both datasets for re-consumption comparison of viewing features.

		First_{num}	Revisit_{num}	$\mathcal{U}^{\sigma}_{revisit}$	$\mathcal{I}^{\sigma}_{revisit}$	$\text{Topic}^{\sigma}_{revisit}$
EdNet	imbalanced	274,396	30,358	2.62	64.15	133.38
	User Adj	10,379	10,379	0	24.14	48.69
	Item Adj	2440	2440	0	0	10.19
MOOC	imbalanced	1,284,845	176,839	4.34	61.80	4112.59
	User Adj	55,084	55,084	0	27.79	1444.25
	Item Adj	3070	3070	0	0	111.44

To create the re-consumption prediction dataset, the first-time view of a lecture, i.e. a *novel* lecture, for each user, is labelled as either re-consumed or not, dependent on the existence of the given lecture later in the user's historical sequence. For EdNet, there are 248,977 novel lecture sessions which are not re-consumed and 25,419 sessions which are, i.e. 9.26% of the novel lec-

[9]Multiple pairs could have been considered, at the cost of a less diverse set of user behaviours

ture sessions. MOOCCubeX's consists of 9.30% re-consumptions, in particular 119,472 novel lecture sessions which are re-consumed and 1,165,373 which are not. To provide a balanced dataset, the labelled *novel* sessions are downsampled such that for each user, the number of lectures viewed once or re-consumed is equal for each user, where the resulting set of datasets are referred to as $\mathcal{D}_{pred}$. Consequently, EdNet contains 49,714 samples equally split between not-re-consumed and re-consumed novel lecture sessions across 10,355 users and 221,944 sessions across 54,068 for MOOCCubex. All of the available viewing features for each dataset were scaled using the non-linear Yeo-Johnson transformation method [122], fitting the estimator on the *training* data and used in the evaluation.

Algorithms

To evaluate to which degree viewing behaviour is predictive of re-consumption, a four different classification models are chosen. No neural models are utilised, as it has been shown that gradient boosted tree methods are generally superior for tabular data of datasets of around 10,000 - 50,000 samples [141]. Therefore a logistic classification model, a SVM, a XGBoost classification model as well as a *Random* baseline was used in the experiments. Despite its simplicity, it is a common, linear classification model as it does not have to make assumptions of the underlying data which similarly complex models like *Linear Discrimant Analysis* do [46, p. 127] Another traditional classifier is the SVM which attempts to find an optimal *hyperplane* which best separates the target labels. Using regularisation techniques and allowing for *slack*, one achieves a soft-margin SVM, allowing some misclassifications with the benefit of better generalisation when classes overlap. [48] In this experiment, only a linear kernel is considered for the SVM. The SVM and logistic classifier are implemented as a stochastic-gradient optimised model for efficiency, where the Scikit-Learn [123] implementation was used.

Lastly, XGBoost is included as a one of the state-of-the-art classification models for tabular data [141]. It is *gradient boosting* method, i.e an ensemble of *decision trees* as weak learners. The output label of the ensemble is the weighted prediction across all of the individual trees' leaf weights, which together provides a more accurate prediction. For creating the ensemble, each iteration greedily adds a decision tree which improves the prediction the most, according to a regularised objection function. [142]

Data splitting and model configuration

For re-consumption prediction, an additional experiment setup was needed. To split the binary re-consumption dataset $\mathcal{D}_{pred}$ into respective *training* and *test* splits, a 90/10 stratified strategy by user was used to make sure both that the splits are balanced, as well as that individual user's behaviour should not bias any one split. The various hyperparameters of the models were tuned using 10-fold cross-validation on the training split, with user-based stratified splitting to keep the splits balanced. Each model was given 250 trials for the respective hyperparameter search. For each trial, the evaluated hyperparameters were sampled using a Tree-Structured Parzen Estimator sampler [134], implemented in Optuna [135]. The test split was used to evaluate the models in a balanced label scenario.

As the original class distribution is almost 1:10, another, unseen test set was created using the sessions which was not excluded in the per-user downsampling process described in Section 5.2.2 as most users did not have an equal number of novel lecture sessions viewed once or re-consumed. Excluding the sessions included in the training split, the resulting imbalanced

test sets were further randomly downsampled to match the original, imbalanced class-ratio of the prediction dataset $\mathcal{D}_{pred}$. The resulting downsampled datasets contained 30,423 and 3106 for not-revisited and revisited novel lecture sessions respectively for EdNet, and 191,985 and 19,682 respectively for MOOCCubeX.

Evaluation

To address the sub-questions posed as a part of experiment 2.1, the general re-consumption behaviour and for specific topics are evaluated using statistical measures and visualisations. For the last sub-question comparing viewing features differences between the first and second view of a lecture, statistical hypothesis testing was applied. In particular, inspired by [11], the two-sample Kolmogorov–Smirnov test [73] was used to test if the general distributions of the individual features were different, measuring it across different users as in. Additionally, the Wilcoxon signed rank test [71] was used, comparing the difference between each viewing pair of first and second-time sessions for each viewing feature. Both tests are non-parametric due to most features' highly skewed distributions and possible dependence. The feature differences were symmetrical, so no symmetric adjustments were needed for the Wilcoxon signed-rank test. Furthermore, the p-values were corrected using the Holm-Bonferroni method [74]. Lastly, both tests used were from the SciPy library [139]. To have an interpretative analysis and comparison of the viewing features, they were not scaled like in related works [18, 25].

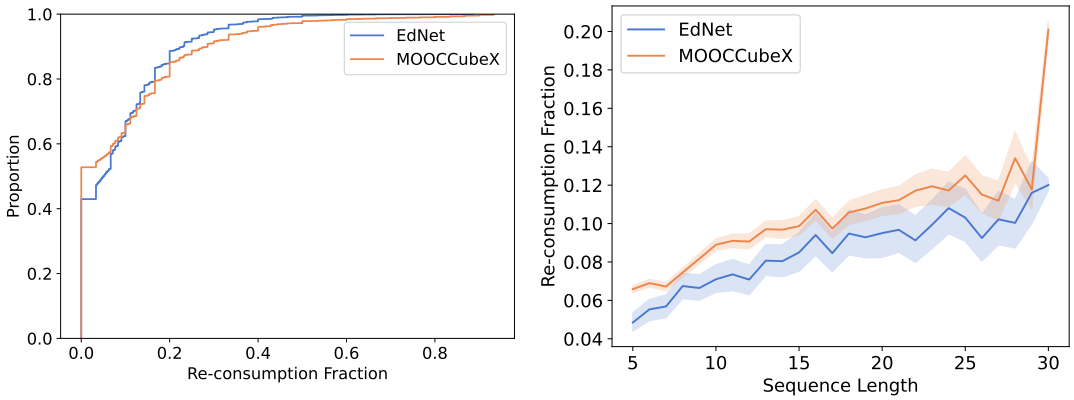
For evaluating the re-consumption classification, the classification accuracy on the balanced dataset is reported by their Accuracy, F1 and Precision, while the imbalanced test set was evaluated on F1, Precision and Recall, all with macro weighting. Moreover, the resulting feature importance calculated by the XGBoost model is evaluated to compare with the individual viewing feature analysis.

5.2.3 Results

Q1: Are re-consumption a substantial fraction of a user's interaction history across topics?

Figure 5.9a presents the empirical cumulative distribution of users' re-consumption fraction, i.e. how many of their interactions are revisits of previous lectures. As illustrated, MOOCCubeX has relatively fewer users who do not re-consume lectures, approximately 47%, whereas 57% of EdNet's users do to some degree. Based on the preprocessing steps in Section 5.2.2, EdNet consists of 9.96% revisits in total, while MOOCCubeX consists of 12.10%. Moreover, 83.49% of EdNet's lectures have been re-consumed, while only 21.03% of MOOCCubeX. Despite this, the raw re-consumption frequency of each user is relatively similar for the two datasets, though MOOCCubeX users in general re-consume more relative to the number of novel lectures they see than EdNet users. Moreover, examining the correlation between sequence lengths in 5.9b, there is a slight positive correlation for both datasets, but relatively large uncertainty.

Furthermore, the expected probability of a lecture to be re-consumed based on the re-consumption frequencies is 8.81% for EdNet and 6.00% for MOOCCubeX. In comparison, the two courses studied in [18] [17] consist of 44% and 50.1% revisits respectively. Moreover, the expected probability of a video to be revisited was 20.1% and 23.5% for the respective courses.



(a) ECDF of what fraction of users' interaction history consists of repetitions, in raw terms

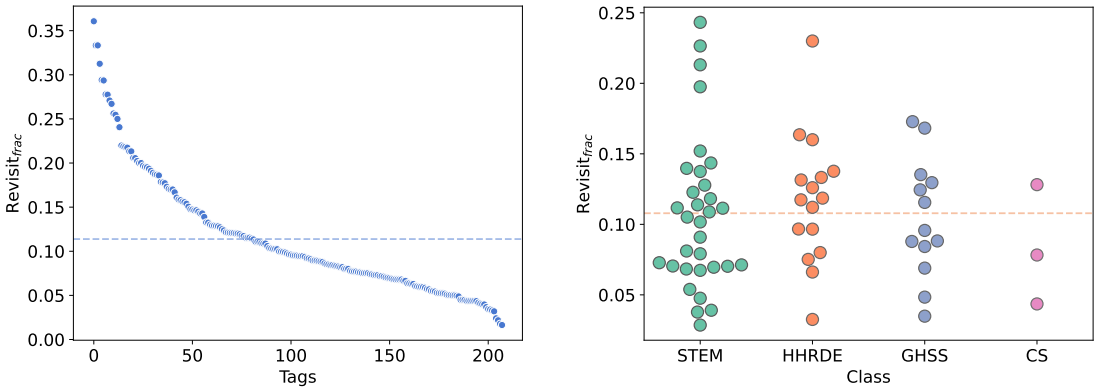
(b) The correlation between the length of users' historic sequences $\mathcal{S}_u$ and the corresponding revisit fraction of the sequence. The highlighted area visualises the 95% confidence interval at each given sequence length

Figure 5.9: The frequency of re-consumption in users' historic interactions sequences and its association to the total number of viewed lectures

Q2: Are there re-consumption frequency differences between topics?

As both datasets have relatively large sets of topics related to each lecture, the differences in re-consumption frequencies for each topic are examined. To alleviate individual users' re-consumption biases the re-consumption proportions of each *known* topic based on the down-sampled version of each dataset, without item bias correction, is presented in Figure 5.10. The first insight is that there are large variances in re-consumption frequencies for tags and fields respectively. In particular for MOOCCubeX, despite the skew of fields per category, there are several topics which are far from the average re-consumption fraction. This is further illustrated in 5.11, where HHRDE has a higher median re-consumption frequency and STEM have the largest variance.

Looking into the 10% fields with the highest and lowest re-consumption proportions, Table 5.6 shows their machine translations and related domain category.



(a) Fraction of all interactions which are revisits for EdNet, for each known topic.

(b) Fraction of all interactions which are revisits for MOOCCubeX, for each known field, where the categories are highlighted.

Figure 5.10: Frequencies of re-consumption per field. The dashed line denotes the global average across topics.

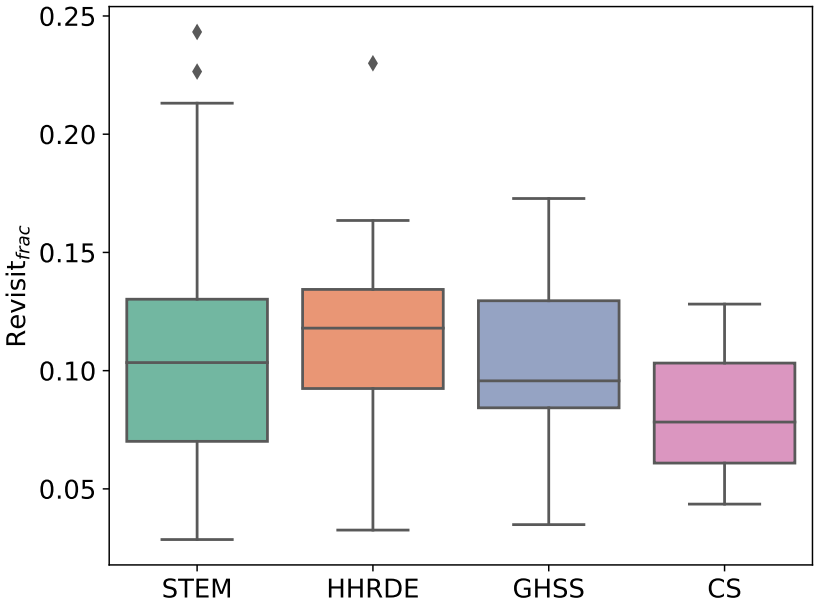


Figure 5.11: The distribution of fields average re-consumption fraction

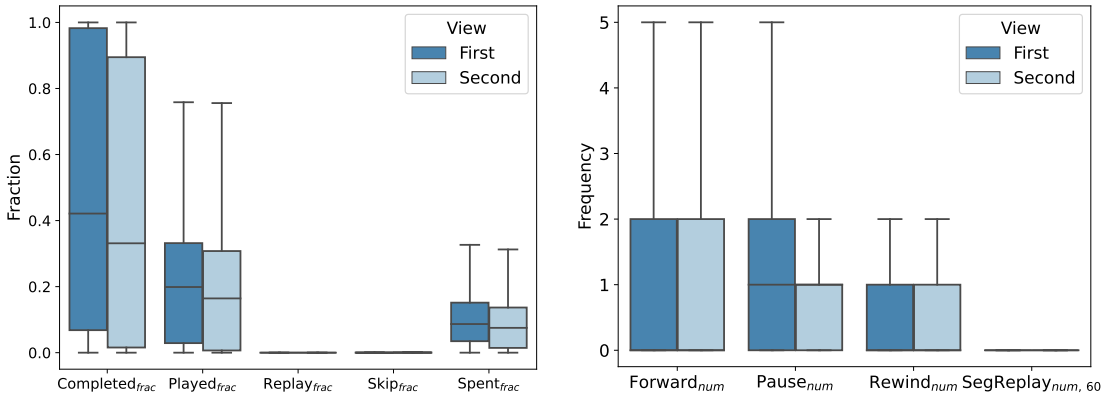
Table 5.6: 10% Most and least re-consumed fields of MOOCCubeX

Most Re-consumed			**Least Re-consumed**		
Class	Field	Revisit	Class	Field	Revisit
STEM	Light engineering and engineering	0.2432	STEM	Instrument science and technology	0.0286
HHRDE	Law	0.2300	HHRDE	History of science and technology	0.0326
STEM	Oil and gas engineering	0.2265	GHSS	Applied economics	0.0349
STEM	Geological resources and engineering	0.2131	STEM	Chemical	0.0378
STEM	Architectural science	0.1975	STEM	Crop science	0.0391
GHSS	Clinical medicine	0.1728	CS	Information and archives management	0.0436

Q3: Are In-lecture interaction behaviours statistically different between first-time views and re-consumption views?

As it has been shown that some in-lecture interaction behaviours are statistically significant between first-time visits and revisits within a single educational domain [17]. To build further on this work, the hypothesis is that some in-lecture interaction behaviours are inherent to a first-time visit or re-consumption, regardless of the educational topic. As the analysis in the previous sections have shown that users and lecture exhibit biases in in-lecture interaction behaviour, as well as lecture re-consumption, the downsampled, item bias corrected dataset was utilised.

To visualise the main differences, the discrete and continuous interaction features are presented in Figure 5.12 and 5.13 for EdNet and MOOCCubeX respectively. As the features are generally positively skewed and long-tailed, the outliers are excluded in the visualisation, with the same argumentation for Pause_{median}. Examining EdNet's continuous features, the completion fraction (Completed_{frac}) and time spent on the video, both including (Spent_{frac}) and excluding pause duration (Played_{frac}), are generally lower for the second time a user views a video compared to the first. For the discrete features, there are fewer visible differences, except for the number of pauses which is generally lower when re-consuming the video.



(a) Discrete features comparison

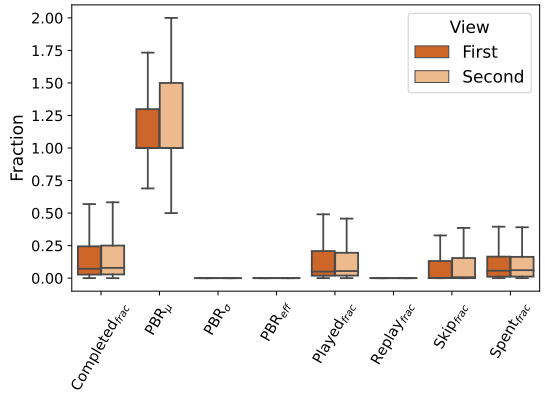
(b) Continuous features comparison

Figure 5.12: Comparing feature ranges for first-time and second-time viewing, i.e. the first repetition of EdNet, excluding Pause_median.

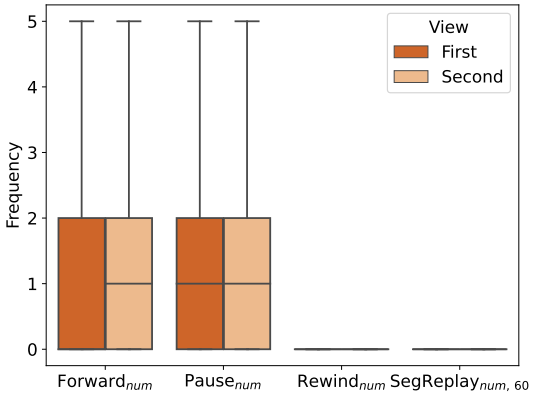
For MOOCCubeX's in-lecture feature distributions in Figure 5.13, there are no visible differences in the distribution between the first time and the second lecture visit in Figure 5.13a. The continuous features are slightly more nuanced. The largest visible difference is in average playback rate (PBR_{μ}), which has more extreme values for re-consumption sessions, but identical median to the first visit sessions. Regarding completion rates, it is slightly higher for re-consumption sessions. The time spent playing is slightly lower for re-consumption sessions, while the total time spent is identical. Moreover, the skipped fraction of the lecture (Skip_{frac} is also slightly higher for re-consumption sessions.

Simply studying some of the distributions according to field categories in Figure 5.14, larger visible differences in terms of per category and with respect to re-consumption behaviour.

To statistically quantify the in-lecture interaction behaviour differences between first-time vis-

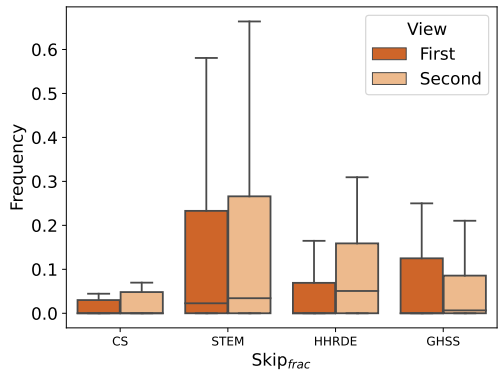


(a) Discrete features comparison

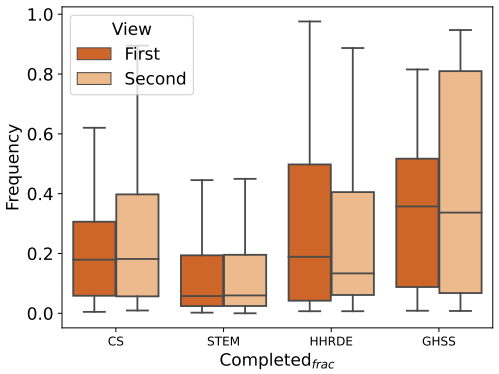


(b) Continuous features comparison

Figure 5.13: Comparing feature ranges for first-time and second-time viewing, i.e. the first repetition of MOOCCubeX, excluding Pause$_{median}$.



(a) Skip$_{frac}$ of MOOCCubeX, comparing the different field categories



(b) Completed$_{frac}$ of MOOCCubeX, comparing the different field categories

Figure 5.14: Comparing features across field categories for first and second view respectively

its and re-consumption, the downsampled, item bias corrected dataset means and standard deviations for each dataset are presented in Table 5.7.

To start with, most features of EdNet have statistically significant differences of distributions and means, of the first and second lecture interactions, as measured by the Kolmogorov–Smirnov test [73] and Wilcoxon signed-rank test [71] respectively. Only the number of forward seeks (Forward$_{num}$), pauses (Pause$_{num}$), and replay proportion (Replay$_{frac}$) are not. Specifically for the number of rewinds (Rewind$_{num}$), the difference of distributions is statistically significant, but not the means measured by the first-time and re-consumption pair for each user. A drastic feature difference is the median pause duration (Pause$_{median}$), where the average for first-time visits are 43 s and 14 s for second-time visits. The same difference is not found in MOOCCubeX, where the average is 43 s and 45 s for first and second-time visits respectively, showing a slight increase for re-consumption interactions.

Compared to EdNet, only one of the considered in-lecture interaction features has a statistically

Table 5.7: Feature means and uncertainty given by 1 standard deviation, ** Indicates K-S distribution test with $\alpha = 0.005$, * $\alpha = 0.05$, ‡ indicates Wilcoxon significance test with $\alpha = 0.005$, † for $\alpha = 0.05$

Feature	**EdNet**		**MOOCCubeX**	
	First-time	Second-time	First-time	Second-time
Spent_{frac}	0.117 ± 0.13	0.1021 ± 0.13**‡	0.1164 ± 0.15	0.1182 ± 0.16
Forward_{num}	1.8439 ± 3.9	1.9045 ± 3.7	1.9622 ± 3.85	2.0515 ± 4.13
Rewind_{num}	0.8066 ± 2.26	0.784 ± 1.94**	0.0844 ± 0.3	0.1098 ± 0.59
Pause_{num}	1.4324 ± 2.61	1.3844 ± 2.71	1.9564 ± 3.75	2.0977 ± 4.16
Pause_{median}	4320.10 ± 9795	1357.50 ± 4120**‡	43.0497 ± 107.23	45.0157 ± 114.36
Completed_{frac}	0.4839 ± 0.4	0.4314 ± 0.4**‡	0.1888 ± 0.25	0.1958 ± 0.26
Played_{frac}	0.2039 ± 0.17	0.1816 ± 0.17**‡	0.143 ± 0.19	0.1495 ± 0.2
Replay_{frac}	0.0162 ± 0.07	0.0158 ± 0.08	0.0125 ± 0.08	0.0246 ± 0.61
Skip_{frac}	0.0106 ± 0.06	0.0358 ± 0.14**‡	0.1239 ± 0.23	0.1276 ± 0.22
$\text{SegReplay}_{num,60}$	0.0393 ± 0.37	0.0311 ± 0.28	0.0098 ± 0.14	0.0088 ± 0.12
PBR_{σ}	-	-	0.0146 ± 0.07	0.0142 ± 0.08
PBR_{μ}	-	-	1.2166 ± 0.36	1.244 ± 0.37†
PBR_{eff}	-	-	0.0096 ± 0.09	0.0105 ± 0.1

significant difference between the first and second-time views. Though the mean and standard deviation of average playback rate (PBR_{μ}) is similar between the two, the sample means are pairwise statistically significantly different, measured by the (paired) Wilcoxon signed-rank test [71]. The difference is better highlighted in Figure 5.13b.

Despite no other statistically significant differences, some of the features' means and standard deviations are quite different from EdNet. While EdNet's average completion rate (Completed_{frac}) is 43% - 48 %, MOOCCubeX's are below 20% for both session types. Notably, the standard deviations are high too; 0.4 for EdNet and ≈0.25 for MOOCCubeX regardless of the session type. The same trend for the average proportion measured features are found in the time spent playing (Played_{frac}), but not for the total time spent (Spent_{frac}) where both the averages and standard deviations are similar for both datasets, regardless of the session type. Moreover, the average number of rewinds (Rewind_{num}) is 0.8 for EdNet's first-time sessions, while only 0.08 for MOOCCubeX's first-time sessions. Furthermore, both the average and standard deviation is higher for second-time visits, while they are both lower for EdNet.

Regarding the number of pauses (Pause_{num}), the results show that MOOCCubeX's average is 1.96, but EdNet's is 1.43, where the standard deviation is also lower. Moreover, MOOCCubeX's re-consumption sessions have slightly more pauses on average, while EdNet has slightly fewer, compared to the first-time sessions. Generally, the replay-related features are low for both datasets, indicating relatively linear watching behaviour and little in-video replaying. Specifically, the Replay_{frac} is just 1-2%, regardless of dataset and session type, but MOOCCubeX's re-consumption standard deviation is much higher, at 0.61, compared to the first-time view standard deviation of 0.08, and EdNet's standard deviation of 0.07-0.08.

Table 5.8: EdNet re-consumption prediction results, where the best result for each metric is highlighted in **bold**, while the second best result is underlined.

Model	Balanced			Imbalanced		
	Accuracy	F1	Precision	F_1	Precision	Recall
Random	0.5018	0.5018	0.5018	0.4012	0.4988	0.4964
Logistic	0.6022	0.6010	0.6035	0.4851	0.5320	0.5896
SVM	0.6040	0.5975	0.6111	0.5162	0.5377	0.5940
XGBoost	**0.6443**	**0.6432**	**0.6460**	**0.5310**	**0.5537**	**0.6399**

Prediction classification

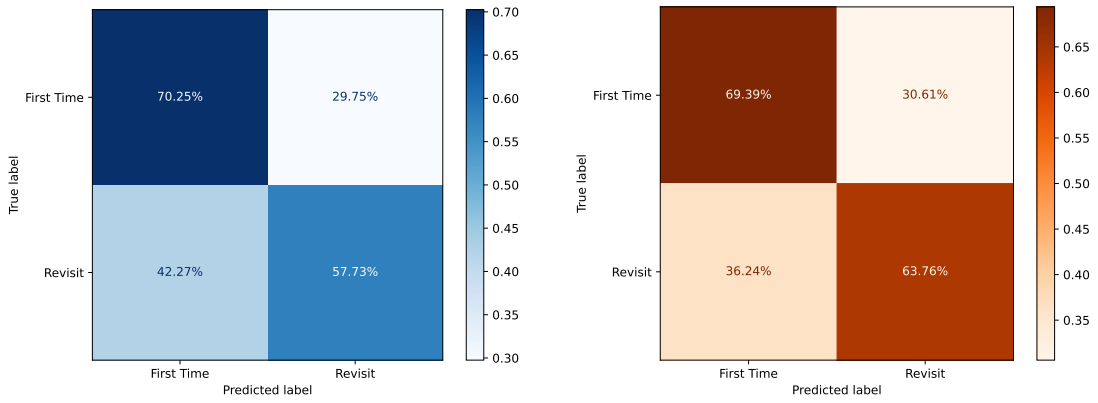
The prediction results of first and second-time views of lectures are presented in Table 5.8 and 5.9 for EdNet and MOOCCubeX respectively. Firstly considering the balanced, downsampled test results, the classification results are relatively comparable for both datasets, where XGBoost illustrates the best classification performance across all metrics and datasets. In particular, XGBoost has 6.67% higher accuracy than SVM on EdNet and 4.03% higher accuracy than the logistic classifier on MOOCCubeX. Moreover, the SVM and logistic classifier perform more similarly, where the SVM has slightly higher accuracy and precision for EdNet, while the logistic classifier performs slightly better for all three metrics on MOOCCubeX.

Table 5.9: MOOCCubeX re-consumption prediction results, where the best result for each metric is highlighted in **bold**, while the second best result is underlined.

Model	Balanced			Imbalanced		
	Accuracy	F1	Precision	F_1	Precision	Recall
Random	0.4987	0.4987	0.4987	0.3987	0.4988	0.4965
Logistic	0.6180	0.6180	0.6181	0.5287	0.5561	0.6497
SVM	0.6167	0.6165	0.6169	0.5189	0.5550	0.6517
XGBoost	**0.6433**	**0.6432**	**0.6433**	**0.5388**	**0.5626**	**0.6658**

Considering the prediction results on the larger, imbalanced datasets, there is a decline as expected in overall classification accuracy, though XGBoost still outperforms the other baselines with a F_1 decline of 17.6% on EdNet and 16.2% on MOOCCubeX from the balanced test case. Comparing the baselines to *Random*'s macro F_1, better-than-random classification accuracy is clear for all three non-naïve classifiers. Moreover, the relative difference between XGBoost and the second-best-performing model for the respective metric is slightly smaller than for the balanced dataset, but it is generally larger for EdNet than for MOOCCubeX across the metrics. In contrast to the balanced dataset, SVM is consistently better than the logistic model for EdNet, while it is only better than the logistic model for recall on MOOCCUbeX. Due to the large imbalance of the first-time and re-consumption labels, examining the confusion matrices in Figure 5.15 of XGBoost is of interest. In particular, it classifies a larger proportion of the revisits correctly on MOOCCube than on EdNet, which is the major contributor to the differences in recall for the two datasets. The recall for the first-time visits is similar for both datasets, where the difference is less than 100 basis points.

Lastly, considering the importance of the different features as defined by the fitted XGBoost



(a) XGBoost classification outcomes on EdNet

(b) XGBoost classification outcomes on MOOCCubeX

Figure 5.15: Confusion matrices for the classification results of XGBoost on the imbalanced datasets for the respective datasets. The matrices are normalised row-wise, i.e. by their true label.

models, there is no significant correlation between the two datasets, though the most important feature is Played_{frac} and Spent_{frac} for EdNet and MOOCCubeX respectively. Moreover Pause_{median} is the second most important feature for EdNet, while one of the least important features for MOOCCubeX. For MOOCCubeX in particular, the playback rate-related features are given less importance, as well as $\text{SegReplay}_{num\,60}$ and Replay_{frac}.

5.2.4 Discussion

The results and how they may address RQ2a and RQ2b are described below, including the limitations of the experiment.

RQ2a: in-video behavioural features and relation to lecture re-consumption

Re-consumption occurrence in the examined datasets is overall infrequent as reported in some papers [26], but much lower compared to other studies [17, 18]. Despite the relatively similar proportion of user lecture sessions being re-consumptions of the two datasets, 10 percentage points more of EdNet's users have re-consumed at least one lecture. This indicates that a minority of the users in MOOCCubeX re-consume drastically more than the remainder user base. Furthermore, the large difference in the proportion of which lectures have been re-consumed between the datasets can indicate that users of MOOCCubeX tend to re-consume the same lectures rather than to re-consume a larger range of different lectures. The larger lecture space and lower lecture accessibility for MOOCCubeX are likely factors for the discrepancy as well. Lastly, the individual re-consumption fraction illustrated in Figure 5.9a illustrates the above points, showing that re-consumption frequency differs on a platform-to-platform basis as well. Interestingly, there is a slight correlation between individual users' history length and re-consumption fraction as illustrated in Figure 5.9b, despite a relatively large uncertainty for both datasets. This may indicate that as the users do use the platform more, a need for some revision occurs, though the correlation with any possible assignments is not provided in the datasets which has been shown to be a motivating factor for revisiting [26]. In addition, it

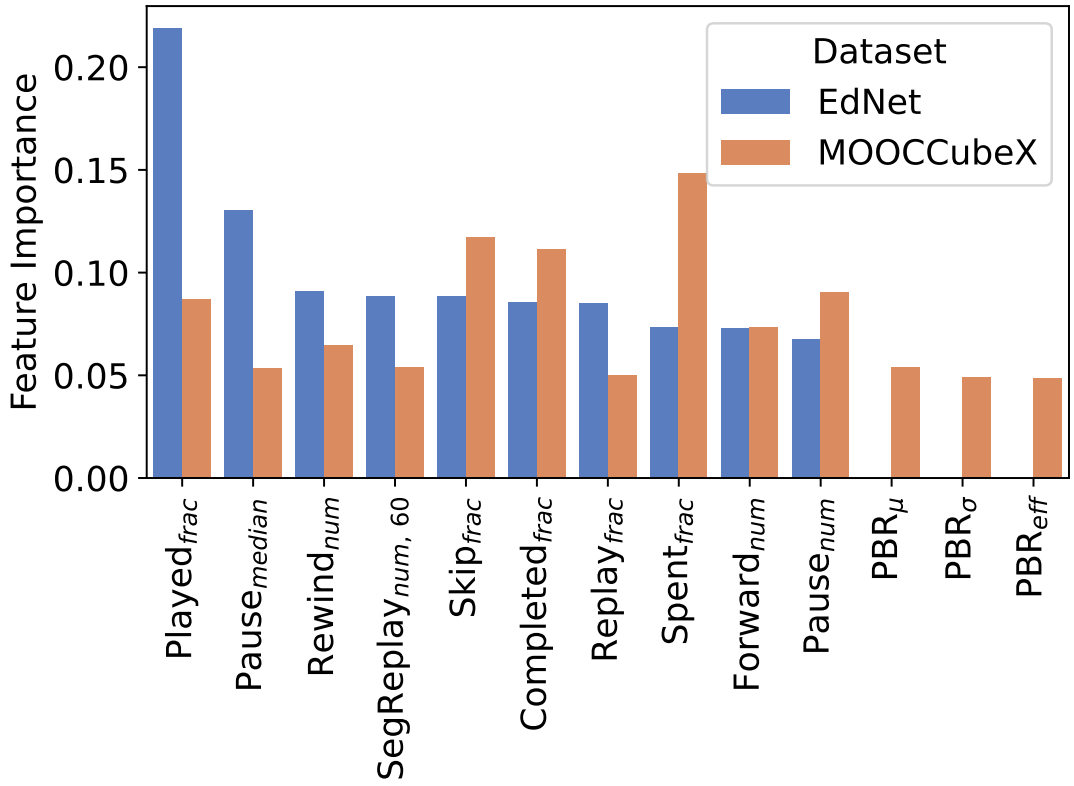


Figure 5.16: Feature importance determined by XGBoost

could be interesting with a more iterative evaluation, considering various users and whether or not they would re-consume at a given relative time step t, to explore the temporal distances between re-consumptions.

Regarding re-consumption frequencies of topics, there are definitive variances in both datasets. As EdNet consists of lectures from a single domain, more specifically for learning languages, it is also interesting that there are such large variances within the domain. This can indicate that more granular topics within a domain have intrinsic properties which demand more reviewing. Looking into MOOCCubeX, there is similarly a large variance between the average re-consumption fraction of fields. Moreover, though slight, there are differences between the field categories as well, where STEM has the largest variance. A contributing factor is the potential mislabelling of fields as it is also the largest category by the number of fields in the dataset. Moreover, as seen in Figure 5.10b, the number of fields per domain category is not balanced, providing a potentially biased representation of the domain category's re-consumption characteristics. Moreover, as STEM fields are the most re-consumed fields among the top 10% of fields, indicates that lectures in these fields might have more difficulty comprehending topics. On the other hand, some STEM fields are also among the least re-consumed fields, illustrating the variance within the broader categories. However, a limitation of the results is that they are not based on equal sets of sessions per topic, though they are corrected for the user's individual bias. Therefore, lectures' topics and their intrinsic properties may skew the comparison of re-consumption behaviour, as well as other lecture properties not considered, like production

style [12].

Finally, considering the in-video behavioural features, adjusted for both user and lecture bias, they are not statistically significant across both datasets. Regardless of statistical significance, most differences between the first and second visit are not drastic, with the exception of Completed_{frac} and Pause_{median} for EdNet. The difference in Pause_{median} might be contributed to noise or external contextual factors, while Completed_{frac} correspond with the intuition that users likely view a smaller portion of the lecture the second time, as they might be reviewing specific parts of the lectures. However, this argument does not hold when considering MOOCCubeX's behavioural features. For EdNet, the Kolmogorov-Smirnov and Wilcoxon signed rank test mostly agree, indicating that the user and lecture bias corrections were effective, as the Wilcoxon test compares first and second-time views per user, per lecture, while Kolmogorov-Smirnov considers the distributions as a whole. The exception is Rewind_{num} which is not statistically significant by the Wilcoxon test, which can indicate an influence of other factors besides the user and lecture biases. Furthermore, the large deviations between the datasets on some of the features like Completed_{frac} and Rewind_{num} further shows that there are inconsistencies across learning platforms, making it more difficult to provide general, domain, demographic and context-independent conclusions of in-video behaviour indicative of re-consumption.

Some of the limitations of the study are that though it considers behaviours across users, lectures and topics, the user and lecture bias corrections heavily reduced the dataset, so the comparison is not across a large set of users and lectures. This is due to the low overall re-consumption rate as discussed previously, providing a less diverse and smaller comparative dataset compared to [17, 18]. Moreover, the frequency of different topics in the datasets is not adjusted for as seen in Figure 5.8b, potentially leading to biased results due to topic-related behaviours. This argument is supported by the visual comparison of the domain categories of MOOCCubeX, which showed larger visual differences per domain category than when examining the features across domains. However, these differences are not statistically tested due to relatively few and unbalanced first and second-time paired views per domain category. Lastly, as most in-video viewing features are sparse and positively skewed, reporting the mean and standard deviation might be less appropriate compared to e.g. the median and interquartile ranges. Moreover, using a statistical test like Anderson-darling [143], as a two-sample test [144] which is more sensitive to the tails of distributions could more accurately determine the differences.

So in summary, some users do re-consume consistently more than others and across a more diverse set of lectures. Additionally, there are some in-video viewing features which significantly differ between the first visits and the following revisit, but the differences are not consistent across datasets, nor are the differences large on average. Moreover, there are indications of domain-influenced in-video viewing behaviour related to re-consumption, but the extent of it is not quantified.

RQ2b: Re-consumption prediction

Firstly, evaluating the performance of XGBoost as one of the state-of-the-art methods for tabular data, the superiority over less complex models like SVM and logistic regression is not drastic. Moreover, in the real-life setting evaluation on the unbalanced dataset, the outperformance is relatively smaller compared to the balanced evaluation. This might be due to a less complex feature selection and engineering process, e.g. no cross-features are used. Con-

sidering the different unbalanced classification outcomes, the difference in accurately recalling re-consumptions, i.e. TPs of the datasets can partially be attributed to the slightly less imbalanced MOOCCubeX. Overall, roughly four out of ten user lecture sessions are wrongly assumed to not be re-consumed, which decreases the ability of the model to intervene in an appropriate way. Considering the consistent FPs across the datasets despite the class imbalance differences, the real-life applied consequences might be an annoyance for the end user as lectures which are not re-consumed are anticipated to be re-consumed.

On the other hand, this could be a TEL tool for promoting reviewing behaviour, and more generally enforcing SRL strategies. By this reasoning, it enables choosing a model with a relatively high FPs for increased recall of to-be re-consumed lectures. Considering the calculated feature importance, despite relatively noisy features like Pause_{median}, it is, perhaps falsely, an important feature as assessed by XGBoost for EdNet. Though some of the traits of the more important features are their low sparsity, it is not consistent for all of the features, nor across the datasets. In contrast to the viewing and re-consumption analysis discussed in the previous section, the balanced dataset used to train the classifiers is not down-sampled to correct for user or lecture bias, and implicitly topic bias. The imbalanced test set is also randomly downsampled to match the original class distribution, which consequently may add further user and lecture bias as the users and lectures which are part of more sessions will have a higher likelihood to be kept, reducing the diversity of the test evaluation. However, further analysis is needed to quantify the potential impact of the dataset composition.

Limitations of Exp.2

In addition to the specific limitations discussed for the in-video feature analysis and re-consumption prediction, some more general limitations are that as the user historical views are limited to their 30 first sessions, students who view a lot of lectures in a relatively short time won't necessarily have the need to review previous lectures until later, e.g. after the cut-off. Though to which degree this occurs is not quantified. Moreover, re-consumption distances are not considered in this analysis, so continuing the lecture a user started earlier in the evening, is treated equally as an introductory lecture reviewed before a final assignment, though the user intentions of the considered revisit are different. This leads to the problem of defining a re-consumption of a lecture. Although [28] defines a lecture revisited only if it had been watched in its entirety in the past, a such definition would deem few lectures as re-consumed by the reported completion rates of the datasets in 5.2.3. However, [17] found that in-video dropout is highly correlated with revisiting the lecture and therefore a partial viewing of a lecture is more reasonable to consider. Notably, [17] did exclude in-video dropout sessions in their re-consumption analysis, though they did not consider the temporal distances of revisits. On a conceptual level, one can consider a continuation of a lecture in a separate session as exhibiting "explore"-behaviour by watching new, though related, content. Moreover, $\text{SegReplay}_{num\,t}$ should then be an indication of in-video repetition in the video revisit, but generally, there was no large difference between the first visit and re-consumption in Table 5.7. On the other hand, considering re-consumption at a topic level, a continuation of the lecture on a different occasion can be considered as an exploit behaviour due to the likely similarity of topics within a lecture. These aspects further highlight the difficulty of defining a re-consumption purely based on viewing logs.

5.3 Experiment 3 - Alignment

In the last experiment, the inclination towards recommending already viewed lectures, i.e. suggesting to re-consume, and the effect of re-consumption behaviour in users' history on RSs' recommendations are explored. Two different evaluation approaches were applied, where the first approach assembles a relative baseline to which extent various models are inclined to recommend lectures from users' interaction history. The second evaluation method studies how well-aligned out-of-the-box models are with respect to individual users' re-consumption behaviour.

5.3.1 Problem definition

The two evaluation scenarios have separate problem definitions. The first is stated as a next-item prediction problem as in Section 5.1.2 for Exp.1, but the set of relevant lectures, i.e. target labels for a sequence $\mathcal{S}_u$, is redefined. In Exp.1, it consisted of a single item, lecture at time step $n_u + 1$ for a user u. However, for the scenario of recommending previously viewed lectures, the set of unique lectures in each user's history $\mathcal{H}_u = \{l^{(u)} : l_t^{(u)} \in \mathcal{S}_u\}$ is used as the relevant item set to indicate to what extent the models recommend to re-consume viewed lectures.

The second scenario of re-consumption alignment is an adaption of the problem of calibration as defined in Section 2.3.3, where the classes C to measure the calibration of, is defined by the exploit-explore behaviour and corresponding recommendations in the following sections.

5.3.2 Experiment Setup

The specific preprocessing steps and evaluation methods related to Exp.3 are described below.

Preprocessing

The evaluation considers the recommendations made by RSs evaluated and Exp.1, so to evaluate their inclination towards recommending re-consumption of already viewed lectures, every user is considered and no further preprocessing is needed. In the second case, only users' who have shown re-consumption behaviour in their interaction history are evaluated, as the main focus is to study if the models are able to adapt to individual users re-consumption preferences and to which degree. Moreover, users who have only shown repeating behaviour are excluded in the evaluation as well to avoid outliers, which is 0.37% of repeating users in EdNet and 2.85% of repeating users in MOOCCubeX.[10] In sum, 9336 repeating users for EdNet and 46,484 repeating users for MOOCCubeX are evaluated in the second scenario. Lastly, the model configurations and data splitting strategies are the same as in Exp.1.

Evaluation

For the former scenario, the task is to evaluate the recommendation accuracy and is consequently evaluated using the same ranking metrics as in Exp.1: NDCG, MAP and Recall, by the same argumentation as in Section 5.1.3. Because the individual interaction history lengths varies to a large degree per dataset and between the datasets, the cut-off is based on each

[10]These users *are* included in the fitting the model, as the recommendations are based on the fitted models in Exp.1.

dataset's median interaction history length of the *training splits* to make the results more comparable. The reasoning is that a large deviance between the cut-off and the interaction history lengths, can introduce a bias in the evaluation metrics because they are directly impacted of each users' history. The median length of the training splits for EdNet is 12 and 7 for MOOC-CubeX.

For evaluating the alignment of the recommendations of the proposed models with regard to users exploit-explore preferences, the problem is mapped to a calibration problem, evaluated using KL-Divergence following [37, 65], adapting the definition in Equation 2.4, with some minor changes. As in [37], the un-smoothed $q(c|u)$ is used rather than $\tilde{q}(c|u)$. In addition, each item is weighted equally in both distributions $p(c|u)$ and $q(c|u)$, i.e. $w_{i,u} = 1$ and $w_{r(i)} = 1$ for all lectures and users [37, 65],. Moreover, in contrast to [65] and [37], the considered classes are user-item interaction dependent, rather than static, item-dependent classes like movie genres. This leads to some challenges as for the definition of the class set C for labelling the dynamic exploit-explore preferences. Firstly, although a user's history can mainly consist of a few lectures viewed multiple times, the provided recommendations can only recommend any lecture once, i.e. a recommendation list consists of unique lectures. Consequently, a comparison of the raw distribution of historical re-consumptions and the recommended lecture distributions would be biased. Therefore, only the re-consumption status of unique lectures are considered. More formally, the user's history $\mathcal{H}$ can be divided into two, disjoint sets $\mathcal{H}_n$ and $\mathcal{H}_rep$, where they consist of lectures viewed once and lectures viewed at least twice respectively.

Secondly, in this proposed binary setting, $\mathcal{H}_n$ can be interpreted as representing the users' *explore*-preference, and $\mathcal{H}_rep$ as their *exploit*-preference. Following this definition, the provided recommendations then have three classes; recommended lectures already seen once, lectures already viewed multiple times and lectures which the user has not viewed before. In the case of utilising three classes for measuring the exploit-explore preferences, the evaluated KL-divergence of any set of recommendations will diverge as the "unseen lecture"-class will always be zero. For formally $q(c = unseen|u) = 0 \; \forall u \in \mathcal{U}$. On the other hand, an interpretation of recommending already seen lectures, regardless of how many times it has been viewed, is a suggestion to re-consume and therefore an indication of an exploit preference. In addition, recommending unseen items can also be interpreted as an indication of a explore preference. By these two interpretations, KL_{novel} is proposed which considers two classes, which have asymmetric definitions for $p(c|u)$ and $q(c|u)$.

KL_{novel} has mainly two limitations as it has asymmetrical definitions for the classes and it does not differentiate between recommended lectures seen once or multiple times. To overcome these drawbacks, KL_{rep} is proposed. In contrast to KL_{novel}, the recommended lectures already viewed *once* are labelled *explore*-class and recommended lectures already viewed *multiple* times are labelled as *exploit*. This provides a better nuance of Recommendation Systems ability to distinguish outlier re-consumption behaviour and general, consistent re-consumption behaviour among users. Recommended, novel, i.e. not-viewed, lectures are not considered.

A comparison of the differences between the two definitions for the recommendations classes is illustrated in Table 5.10 for different sequence lengths and sequence behaviours. The example recommendations are from a naive RS which recommends each of the lectures in the user history. The table illustrates that Recommendation Systems which are strongly inclined to recommend viewed lectures, will be highly calibrated according to KL_{rep}. However, a more

likely scenario where re-consumptions are less frequent as in the third row such naive RSs won't recommend any novel lectures to the user which did not align with their exhibited preference of exploring, despite a perfect calibration measured by KL_{rep}. Moreover, the equivalent scenario in row six for a user with a shorter sequence history than the used cut-off, is no longer an issue, highlighting the impact of history lengths.

Table 5.10: An adaption for re-consumption of the calibration scenarios discussed in [65], where the number of recommended lectures is $N = 10$. p_{novel} and p_{rep} denotes the frequency of each of the exploit-explore classifications of the recommendations list

$\mathcal{S}_u$	$\lvert\mathcal{H}_{rep}\rvert : \lvert\mathcal{H}_n\rvert$	$\mathcal{R}_{10}$	p_{novel}	p_{rep}
A A A A A A A A A B	1:1	A, B, ...	2:8	1:1
A A A A B B B B C D	2:2	A, B, C, D ...	4:6	4:1
A A B C D E F G H I	1:6	A, B, C, D, E, F, G, H, I, J	10:0	1:6
A A A A B	1:1	A, B, ...	2:8	1:1
A A B B C	2:1	A, B, C, ...	3:7	2:1
A A B C D	1:3	A, B, C, D, ...	4:6	1:3

Explainable re-consumption calibration

Although KL-divergence has a lower bound, the metric is not directly **interpretative**. For instance, it only provides a scalar value of the difference between the distributions. In the exploit-explore alignment setting, and more generally for binary-class problems, it does not indicate the direction of the miscalibration. More specifically it does not answer whether or not the RS or re-ranking algorithm overemphasise or undervalue exhibited re-consumption behaviour. Moreover, the magnitude of differences of C_{KL} when comparing models does not have a clear interpretation. Therefore, a more interpretative and directional evaluation "metric" *Recommendation-to-repetition*, i.e. $\mathcal{R}2\mathcal{R}$ is proposed, comparing the unique lecture re-consumption ratio of the user's history and the recommended re-consumption ratio at cut-off k. More formally, the re-consumption ratio of a provided set of recommended lectures $\mathcal{R}_{u,R}$ for user u is defined as

$$\mathcal{R}_{u,R} = \frac{|\mathcal{R}_u \cap \mathcal{H}_{u,rep}|}{|\mathcal{R}_u|}, \tag{5.3}$$

where $\mathcal{H}_{u,rep}$ is the unique, historically re-consumed lectures of user u and R_u is the set of uniquely recommended lectures for user u [11]. When $\mathcal{R}_{u,R} = 1$, the entire recommendation list consists of previously re-consumed lectures, where as when $\mathcal{R}_{u,R} = 0$, the recommendation list does not contain any re-consumed lectures, however it may contain lectures viewed once, i.e lectures within $\mathcal{H}_n$. Moreover, the historic re-consumption ratio $\mathcal{R}_{u,\mathcal{H}}$ of an user u is defined as

$$\mathcal{R}_{u,\mathcal{H}} = \frac{|\mathcal{H}_{u,rep}|}{|\mathcal{H}_u|}, \tag{5.4}$$

[11] Note that this differs from the definition used for ranking metrics in Section 2.4.2 as $\mathcal{R}_u$ is the set of actual items recommended, while R consists of the *rankings* of the recommended relevant items.

where $\mathcal{H}_u$ is the *unique* set of all, previously viewed lectures of user u. . By combining Equation 5.3 and 5.4 the definition of $\mathcal{R}2\mathcal{R}_u$ for an user u is

$$\mathcal{R}2\mathcal{R}_u = \frac{\mathcal{R}_{u,R}}{\mathcal{R}_{u,\mathcal{H}}} = \frac{\frac{|R_u \cap \mathcal{H}_{u,rep}|}{|R_u|}}{\frac{|\mathcal{H}_{u,rep}|}{|\mathcal{H}_u|}}, \tag{5.5}$$

where $\mathcal{H}_{u,rep}$ is the *unique* set of re-consumed lectures for user u. For an entire set of users $u \in \mathcal{U}_rep$ who have exhibited re-consumption behaviour, $\mathcal{R}2\mathcal{R}$ is defined as

$$\mathcal{R}2\mathcal{R} = \frac{1}{\mathcal{U}_{rep}} \sum_{u \in \mathcal{U}_{rep}} \mathcal{R} \in \mathcal{R}_u. \tag{5.6}$$

In general, $\mathcal{R}_{u,\mathcal{H}} \in [0,1]$ making $\mathcal{R}2\mathcal{R}_u$ undefined for $\mathcal{R}_{u,\mathcal{H}} = 0$, but when constraining it to $|\mathcal{H}_{u,rep}| \geq 1$ as in this scenario, it is well-defined. In particular, for a 'perfectly' calibrated RS $\mathcal{R}2\mathcal{R} = 1$, while $\mathcal{R}2\mathcal{R} > 1$ indicates that the recommendation list of a RS on average contain more re-consumptions than what the user has previously exhibited. When $\mathcal{R}2\mathcal{R} < 1$, the RS on average recommends fewer re-consumed lectures than what the user has previously viewed. More generally, it focuses on the magnitude of the *exploit-class* considered for KL_{rep}, where the magnitude refers to the number of uniquely re-consumed lectures, not the frequency of each.

By the same argument as for the re-consumption ranking evaluation, the cut-off for KL_{novel}, KL_{rep} and $\mathcal{R}2\mathcal{R}$ are chosen by the individual training splits of the datasets' medians, 12 and 7 for EdNet and MOOCCubeX respectively. Furthermore, To quantify any significant differences in re-consumption ranking and calibration, between baseline models, and between the different variants the same statistical testing procedure as in Exp.1 is used, where XLNet$_{feat}$ is not statistically tested.

5.3.3 Results

Repetition Ranking metrics

The raw repetition ranking evaluation metrics of the different models are presented in Table 5.11. Of the naive baselines, *MostPop* is similar to most CF baselines. The CF methods have generally higher metrics than SARS, indicating a higher inclination towards. Moreover KNN has a statistically significant higher re-consumption recommendation, indicating it is more inclined to recommend lectures from the users' history, as well as ranking them higher than novel items. Moreover, it is the most affected model of all the variants explored. The reported results of the SARSs are between half and a tenth of the re-consumption recommendation accuracy of KNN for the respective metrics. BERT has higher metrics than the other two, while GRU has the lowest.

Regarding the side-information enriched models, the performance is similar to the base variants, where BERT is the most inclined to recommend already viewed lectures. Moreover, most of the results are not statistically significantly different from the base variants, though all three models mostly have slightly lower recommendation accuracy. The bias-adjusted variants are

also similar to the enriched versions, though BERT shows a statistically significant decline in all metrics. GRU's results, however, are statistically significantly higher than the not-corrected variant, while XLNet has consecutively lower than their enriched counterpart but mostly not statistically significant.

Table 5.11: Repetition ranking results. * indicates *not* statistically significant over the second best-performing baseline in the given metric with $\alpha = 0.05$. † indicates *not* statistically significant result compared to the *base* variant of the model with $\alpha = 0.05$. ‡ indicates *not* statistically significant compared to the *full* variant of the model, with $\alpha = 0.05$, with Holm-Boneferroni corrected Wilcoxon signed rank test. The best result for each metric in each variant section is highlighted in **bold**, while as the second best result is underlined.

	Model	EdNet			MOOCCubeX		
		NDCG@12	R@12	MAP@12	NDCG@7	R@7	MAP@7
Baselines	Random	0.0663	0.1436	0.0389	0.0002	0.0004	0.0001
	MostPop	0.3979	0.5146	0.2454	0.1608	0.1363	0.1459
	Syllabus	-	-	-	0.1660	0.1425	0.1501
	iALS	0.3748	0.5752	0.2048	0.4438	0.4353	0.3049
	LMF	0.3648	0.5273	0.2219	0.3469	0.3432	0.2107
	BPR	0.4029	0.5994	0.2399	0.5552	0.5550	0.3785
	KNN	**0.5018**	**0.6513**	**0.3238**	**0.6856**	**0.6400**	**0.5020**
	GRU	0.1708	0.1057	0.0393	0.0236	0.0146	0.0159
	BERT	0.2227	0.1379	0.0562	0.0322	0.0217	0.0228
	XLNet	0.1979	0.1224	0.0462	0.0247	0.0157	0.0162
Full	GRU	0.1716†	0.1053†	0.0384	0.0237†	0.0148†	0.0162†
	BERT	**0.2185**	**0.1382**†	**0.0538**	**0.0346**	**0.0240**	**0.0244**
	XLNet	0.1966†	0.1245	0.0457†	0.0256†	0.0164†	0.0164†
	XLNet$_{feat}$	0.2014	0.1272	0.0469	0.0253	0.0160	0.0163
Bias-Adj	GRU	0.1729	0.1068†	0.0389†	**0.0285**	**0.0210**	**0.0184**
	BERT	**0.2065**	**0.1347**	**0.0511**	0.0272	0.0172	0.0177
	XLNet	0.1960†‡	0.1230†‡	0.0452‡	0.0249†‡	0.0161†‡	0.0164†‡

Examining the repetition ranking results of MOOCCubeX, *MostPop* show comparative results to the syllabus-based method. Moreover, the *Random*-baseline reports much lower results than for EdNet. Excluding the naïve-baseline differences, the other baseline results are similar to EdNet, where KNN performs statistically significantly better than all other models and variants as well. Generally, LMF reports lower results for both datasets, but the relative difference to the other CF methods is higher in MOOCCubeX. Interestingly, the SARS are not largely inclined to recommend re-consumptions as their ranking results are a magnitude lower compared to the conventional CF methods. Moreover, the large discrepancy between NDCG and precision is not present in contrast to with EdNet, though precision is slightly lower compared to NDCG for all three models.

As with EdNet, the enriched variants are performing similarly to the base versions. Moreover BERT is the only model with statistically significant results, and they all show an increase in re-consumption recommendation inclination in contrast to with EdNet. The results of the biased corrected version on the other hand are more significant, though similar to the base

and enriched variants. Firstly, GRU reports the highest repetition recommendation accuracy, while XLNet reports the lowest. Moreover GRU's results are statistically significantly higher than both of its counterparts, while BERT's results are statistically significantly lower. None of XLNet's reported declines are statistically significant.

Repetition model alignment

With the general effect of re-consumption behaviour on the various models' recommendations, a comparison of the models' degree of calibration to the individual users' re-consumption preferences is presented in Table 5.12. One thing to note is that most results were not statistically significant. Therefore, only the ones which *were* statistically significant are highlighted.

Examining the calibration of the models according to KL_{novel}, the relative miscalibration between models differs. Firstly, the differences between the baseline models in EdNet are relatively smaller than when measuring KL_{rep}. Furthermore, KNN reports the highest miscalibration of all models and variants, across both datasets. On the other hand, the sequence-aware models are in both datasets among the most aligned models, in addition to *Random* whose calibration level is statistically significant compared to XLNet in MOOCCubeX. Examining the side-information enriched models for EdNet, all of them are statistically significantly more aligned, although the increase is small. In contrast, none of the enriched models in MOOCCubeX shows statistically significant results, but the enriched BERT and XLNet are slightly more aligned than their base variants. For the bias-adjusted variants, GRU and BERT both have statistically significant increases in alignment compared to the base variants for EdNet, but only BERT is statistically significantly more aligned than its enriched counterpart. For XLNet, there is no difference on average to the enriched variant. Examining the bias-adjusted variants on MOOCCubeX, both GRU and BERT have statistically significant increases in alignment to both of their counterparts, whereas XLNet is slightly more misaligned, though not significantly.

Regarding KL_{rep}, *Random* reports the overall highest miscalibration measured of all models and variants in both datasets, corresponding to its low ranking results in Table 5.11. In contrast, KNN is the most calibrated by KL_{rep} by a statistically significant margin in both datasets as well, in line with its repetition ranking performance. Furthermore LMF is less calibrated than the other CF methods in either dataset, but the miscalibration difference to the other methods is twice as high in MOOCCubeX than in EdNet. The SARS without side-information are less aligned according to KL_{rep} than the conventional methods, only out-performed by *Random* in both datasets.

For the side-information enriched models, none of the KL_{rep} results is statistically significant for EdNet, while only GRU's slightly improved calibration is statistically significant. Notably BERT is the most calibrated in both datasets. The measured KL_{rep} for the bias-adjusted variants are more varied. Considering EdNet, both BERT and GRU report statistically significant increases in miscalibration; GRU compared to its base variant and BERT compared to both of its other variants. For MOOCCubeX, GRU is statistically significantly more calibrated than both of its counterparts, while BERT is statistically significantly less calibrated as measured by KL_{rep}.

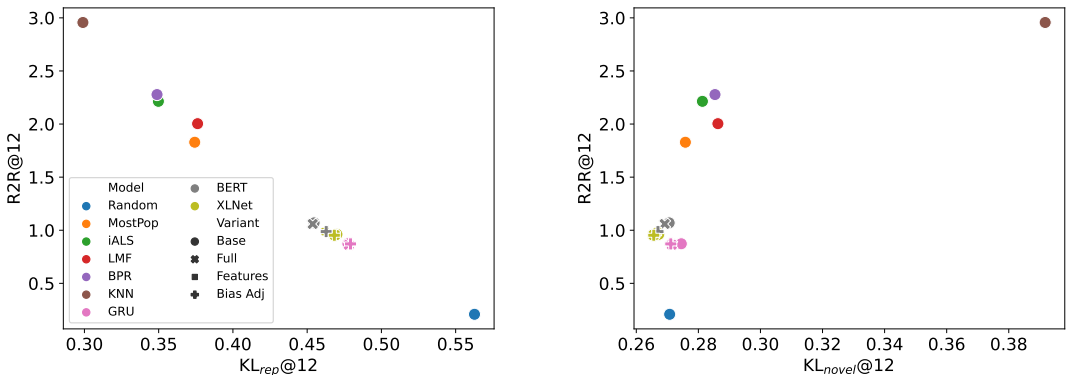
Continuing to the proposed re-consumption calibration metric $\mathcal{R}2\mathcal{R}$, the results provide a different angle to the alignment challenge. Firstly, all of the conventional CF methods overemphasise previously re-consumed lectures, recommending them drastically more frequently in both datasets. In contrast, the SARS are highly aligned in EdNet, but as much in MOOCCUbeX,

Table 5.12: Repetition alignment results. * indicates statistically significant over the second best-performing baseline in the given metric with $\alpha = 0.05$. † indicates statistically significant result compared to the *base* variant of the model with $\alpha = 0.05$. ‡ indicates statistically significant compared to the *full* variant of the model, with $\alpha = 0.05$ using Holm-Bonferroni corrected Wilcoxon signed rank test. The best result for each metric in each variant section is highlighted in **bold**, while the second best result is underlined.

	Model	EdNet			MOOCCubeX		
		KL_{novel}@12	KL_{rep}@12	R2R@12	KL_{novel}@7	KL_{rep}@7	R2R@7
Baselines	Random	0.2708	0.5627	0.2085	**0.4321***	0.4468	0.0006
	MostPop	0.2758	0.3743	1.8287	0.6889	0.3710	**1.4317**
	Syllabus	-	-	-	0.6853	0.3726	1.4440
	iALS	0.2813	0.3498	2.2139	0.6984	0.2703	3.2920
	LMF	0.2863	0.3762	2.0036	0.5901	0.3313	2.5457
	BPR	0.2854	0.3489	2.2777	0.6757	0.2772	3.5502
	KNN	0.3917	**0.2990***	2.9564	0.8407	**0.2342***	4.1415
	GRU	0.2745	0.4774	0.8727	0.4760	0.4350	0.2999
	BERT	0.2706	0.4547	1.0695	0.4874	0.4320	0.3730
	XLNet	**0.2672***	0.4700	**0.9616***	0.4738	0.4354	0.3033
Full	GRU	$0.2714^{\dagger}$	0.4780	$0.8635^{\dagger}$	0.4763	$0.4343^{\dagger}$	0.2994
	BERT	$0.2691^{\dagger}$	**0.4537**	1.0592	0.4865	**0.4320**	$\mathbf{0.3819}^{\dagger}$
	XLNet	$\mathbf{0.2656}^{\dagger}$	0.4686	0.9544	0.4729	0.4354	0.3078
	$XLNet_{feat}$	0.2668	0.4622	**0.9874**	**0.4729**	0.4361	0.3057
Bias-Adj	GRU	$0.2711^{\dagger}$	0.4792	$0.8713^{\ddagger}$	$\mathbf{0.4737}^{\dagger\ddagger}$	$\mathbf{0.4338}^{\dagger\ddagger}$	$\mathbf{0.3289}^{\dagger\ddagger}$
	BERT	$0.2670^{\dagger\ddagger}$	$\mathbf{0.4628}^{\dagger\ddagger}$	$\mathbf{0.9872}^{\dagger\ddagger}$	$0.4753^{\dagger\ddagger}$	$0.4354^{\dagger\ddagger}$	$0.3236^{\dagger\ddagger}$
	XLNet	**0.2656**	0.4683	0.9531	0.4742	0.4356	0.3054

generally recommending fewer of the re-consumed lectures, which correlates with their ranking results in Table 5.11. For EdNet, XLNet is the most aligned model by a statistically significant margin, while out of the sequence-aware models, BERT is the most aligned in MOOCCubeX. For the side-information enriched model variants in EdNet, they show consecutively slight declines in alignment, but only GRU's is statistically significant. Interestingly, the feature selected XLnet is the most calibrated model tested on EdNet as measured by $\mathcal{R}2\mathcal{R}$. BERT has a slight, statistically significant increase in alignment in MOOCCubeX, while GRU and XLnet's slight decreases are not. Lastly, for the bias-adjusted variants, BERT is significantly less calibrated in both datasets, with a sharp relative decline from the enriched variants. On the other hand, GRU is statistically significantly more aligned than its base variant in EdNet and to both of its counterparts in MOOCCubeX. XLNet's decreases in calibration are not statistically significant in either dataset.

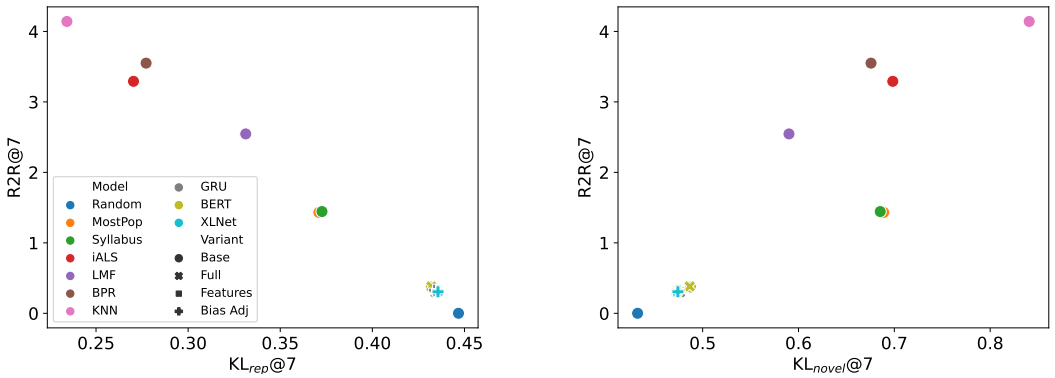
As presented in Table 5.12, the results of KL_{novel} and $\mathcal{R}2\mathcal{R}$ are highly positively correlated, while KL_{rep} and $\mathcal{R}2\mathcal{R}$ are negatively correlated. A further comparison of the metrics for EdNet and MOOCCubeX are presented in Figure 5.17 and 5.18 respectively. The relatively small differences between the variants are highlighted in their clustering in all the diagrams. Moreover, both datasets exhibit close to perfect, negative linear correlation for KL_{rep}, while there are some outliers for KL_{novel} like KNN.



(a) The average $\mathcal{R}2\mathcal{R}$@12 to KL$_{rep}$@12 for each model and variant

(b) The average $\mathcal{R}2\mathcal{R}$@12 to KL$_{novel}$@12 for each model and variant

Figure 5.17: EdNet alignment measures' correlation



(a) The average$\mathcal{R}2\mathcal{R}$@7 to KL$_{rep}$@7 for each model and variant

(b) The average $\mathcal{R}2\mathcal{R}$@7 to KL$_{novel}$@7

Figure 5.18: MOOCCubeX alignment measures' correlation

5.3.4 Discussion

RQ3a: Ranking Accuracy of Already Viewed Lectures

Examining the ranking results of already viewed lectures for the different models provides some interesting insights. The overall trend is that the SARS methods generally recommend less frequently and rank previously seen lectures less than the conventional CF methods for both datasets. While the CF methods perform similarly on both datasets, the relative difference for SARS on MOOCCubeX is much larger than for EdNet, across all metrics. A potential explanation is due to the longer interaction histories of EdNet, making it more likely to recommend a previously seen lecture by pure chance. On the other hand, this should likely have been the case for the CF methods as well, but it is not consistently higher across the models and metrics compared to MOOCCubex. Moreover, the large discrepancy between MAP and NDCG for EdNet, as well as the lower recall for each of the SARS, indicates that they generally rank previously seen lectures higher than novel ones, but they are infrequently recommended.

Why this discrepancy is not present for MOOCCubeX is uncertain, though the large difference of re-consumption lecture diversity presented in 5.2.4 may be a factor. Moreover, of the SARS, BERT is the most inclined to recommend previously seen items.

Considering the side-information enriched and the bias-adjusted variants, the results are mostly not statistically significantly different from the respective base and enriched compared variants. In particular for MOOCCUbeX, only BERT has statistically significant results out of the enriched variants. Moreover, none of the results is either large increases or declines, indicating that the effect of including side-information to recommend either more or fewer viewed lectures is minimal regarding ranking, a fraction of the recommendation list, precision, and regarding out of. For the bias-adjusted versions, all of the GRU and BERT's results are statically significantly different compared to the enriched versions for both datasets. The difference is that GRU's results are increased, while BERT's are decreased. Moreover, none of the results of bias-adjusted XLNet is statistically significant, making the conclusion regarding the effect of side-information inconclusive for SARS' inclination towards recommending previously viewed lectures. In general SARS are shown to prefer recommending novel items over already seen items to a larger degree than conventional CF methods. Whether or not this is a positive for lecture recommendation using sequence-aware methods might depend on the use case. For instance, in a corrective, review-encouraging TEL environment, this might be less optimal.

RQ3b: Re-consumption Alignment

Continuing, with the results of the measured re-consumption alignment of the various models, the three metrics provide different aspects of the re-consumption behaviour. Firstly, considering $\mathcal{R}2\mathcal{R}$, the CF methods are highly inclined towards recommending re-consumed lectures based on $\mathcal{R}2\mathcal{R}$, despite that the utility matrix which they are trained does not consider re-consumption, only binary labels. Furthermore, the SARS are the most aligned for EdNet and MOOCCubeX (except for *MostPop* and *Syllabus*), despite lower inclination towards recommending, general previously seen lectures discussed in the previous section. An interpretation of these findings is that the SARS more accurately distinguish the re-consumed items and recommends, making them generally a better model for re-consumption recommendation. This is further highlighted by the low KL_{novel} for both datasets, which illustrates that the recommendations are not dominated by lectures previously only viewed once. KL_{rep} results are as shown negatively correlated with $\mathcal{R}2\mathcal{R}$, where the CF methods are the deemed more calibrated than the SARS. This may be due to that re-consumption is in-frequent, and as these methods are more inclined to recommend already viewed lectures as discussed, the likelihood of proportional user history distributions and recommendation distributions are generally higher. As SARS generally recommend fewer already viewed lectures, they will have less proportional distributions.

Regarding the inclusion of side information, the relative differences to the base versions are small as in the previous evaluations and more specifically, there are few statistically significant differences for MOOCCubeX. Though small, the enriched variants for EdNet are slightly more calibrated, where all of the KL_{novel} improvements are statistically significant, despite a slightly less inclination to recommend already-seen lectures as discussed in, indicating an improved ability to distinguish the re-consumed lectures 5.3.4. The same correlation is not present for MOOCCubeX, weakening the argument. Moreover, the SARS have generally much lower alignment measured by $\mathcal{R}2\mathcal{R}$, compared to EdNet. This might be partially due to the

sparsity and generally less frequent re-consumption behaviour than in EdNet. Moreover for the bias-adjusted features, interestingly, XLNet is not affected by either not-bias-adjusted or bias-adjusted features, though the recommendations made were generally more accurate according to Exp.1. A user-level analysis of the recommendations might provide insight on why that is. For the other models, despite statistically significant results, the general direction is inconclusive. Therefore general conclusions regarding the effect of accounting for individual user bias' does not necessarily improve the alignment of the models.

Limitations of Exp.3

Some of the more general limitations of Exp.3 are that As the results are based on the preprocessing, model selection and training of Exp.1 as well as the generated recommendations, the same limitations regarding those aspects are applicable to some extent for the results of Exp.1. In addition, the limitations of the definition of a re-consumption discussed in Section 5.2.4 are applicable to the result of Exp.3 as well. Furthermore, the CF methods do not account for re-consumed lectures as the labels are binary. Therefore they are at a disadvantage compared to the SARSs for detecting and varying the weighing of re-consumed lectures. For the evaluation, extreme cases are not considered, i.e. users with no re-consumption behaviour and those who've only re-consumed lectures. The former would likely have more of an impact as around half of the users have not shown re-consumption behaviour, while only 0.37% and 2.85% have only re-consumed lectures for EdNet and MOOCCubeX respectively. For the calibration metrics, only the variety of the re-consumed lectures is accounted for, i.e. the number of uniquely re-consumed lectures, not their magnitude. As in the examples provided in Table 5.10, largely different re-consumption behaviours may not be differentiated as by these evaluation metrics. Some issues of the metrics are reflected in $\mathcal{R}2\mathcal{R}$, where MOOCCubeX's results indicate that *MostPop* and *Syllabus* are the most aligned models. Furthermore, the temporal aspects are not considered either for the metrics, though the timing of re-consumption recommendation is important [61]. Other factors of re-consumption recommendation not considered are which re-consumed lectures were recommended and the relation to the re-consumption frequency of that lecture. Some questions in that regard are for instance at what re-consumption frequency should a lecture be deemed less or more relevant to recommend? Additionally, what is the concrete relation between the recency of viewing a lecture and re-consuming it?

Chapter 6

Conclusion

Considering the overall results of the three sets of experiments and discussion related to various aspects of educational lecture recommendation and viewing behaviour, an overarching discussion of the limitations of the study is presented in section 6.1. In the following Section 6.2, the main contributions of the research are described in further detail. Finally, several aspects which were not investigated in the experiments, as well as possible further research within the field of resource recommendation, in-video viewing behaviours and re-consumption are mentioned in Section 6.3.

6.1 General Discussion

In addition to the limitations discussed for the individual experiments, the choice of datasets, preprocessing and evaluation techniques impose other types of limitations. Regarding the datasets, as EdNet is only a single domain dataset, the *tags* of lectures are likely more related than *fields* of MOOCCubeX which represents different domains. As the tags were encoded initially, the exact semantic similarity cannot be determined. Furthermore, the categorisation of *fields* into the domain categories was done only based on the name of the machine-translated fields, by the author which imposes some uncertainty regarding the validity of some of the fields category labels. Lastly, both datasets are of platforms for Asiatic countries, namely South Korea and China. Moreover, as XuetangX consists of 98% Chinese users, the applicability of the findings to other demographics is reduced as both in-video and general learning behaviour is related to demographics and cultural context [28, 31, 103].

Some of the limitations of the preprocessing steps regarding the use of scaling techniques for transforming the data into normally distributed. Firstly, the different features have various degrees of sparsity and skewness, which could indicate that scaling techniques should be used on a per-feature basis. Moreover, a non-linear feature transformation was used, Yeo-Johnson [122], potentially destroying some of the linear relations between the viewing features. On the other hand, for Exp.1 (and consequently Exp.2), the individual features are later layer normalised as one of the pre-embedding steps of the SARS architecture which was shown to be crucial in previous work [80] indicating that it positively affected the feature distributions. Notably, additional standardisation steps were not applied for the re-consumption prediction in Experiment 2 - Re-consumption Behaviour. Moreover, due to the differences in how in-video interactions are logged, the correlation between the features will differ, as well as the accur-

acy of the measurement of each feature. For instance, as EdNet does not contain playback rate information, the inference of the time spent on a lecture is not necessarily accurate, as the fraction of users who do change the playback rate is unknown. However, the data of MOOCCubeX and previous work [16], that most do not change it. Furthermore, experiments for MOOCCubeX contain playback rate features, providing additional side information regarding user behaviour which are not available for the EdNet experiments.

For the evaluation techniques, a limitation of the statistical significance testing is that the number of test evaluations per model is few for Exp.1 and Exp.3, only 10, and related as only the seed is changed. Moreover, in [36, p .271], the authors mention how the Bonferroni, and by relation Holm-Bonferroni correction can be too restrictive when comparing multiple, correlated metrics of two algorithms. Consequently, the comparisons of the different variants of SARS might have provided more statistically significant results if one chose a less restrictive correction method. On the other hand, the actual improvement or decline of the respective metrics would still remain the same, which was slight in most comparisons. Moreover, different factors such as users' prerequisite knowledge is not considered, though the users will have different starting points determining to which degree lectures are relevant, at both a per segment and per video level. Other contextual or demographic factors of users' or lecture features are not considered, mainly due lack of availability for both datasets, which would skew the comparison of the datasets further. Moreover, in-video dropout and its relation to in-video viewing behaviour or re-consumption behaviour is not taken into account, though it is shown to be frequent [27] and correlated with re-consumption [17].

6.2 Contributions

The main contributions made by this paper are with regard to the posed research questions and goals threefold. Firstly, a solid baseline has been created for the recommendation accuracy on large-scale learning resource datasets through evaluating the performance of various conventional and Sequence Aware Recommendation System. In particular, sequence-aware modelling provided drastically more relevant recommendations than the conventional ones, establishing their dominance within learning resource recommendation. However, less complex RNN-based models can provide similar or better recommendations than transformer-based RS, illustrating that they are still relevant for some recommendation applications. Furthermore, highly granular in-video viewing behaviour fused with lecture intrinsic properties, do provide statistically significantly better recommendations than SARS only considering the lecture embeddings. However, the relative improvement is generally small, with an increased complexity cost for data collection, processing and the model itself. In addition, despite users being biased in their in-video viewing behaviour, adjusting for it does not generally improve recommendation quality.

Secondly, a novel analysis of in-video viewing behaviours across domains and topics has been executed to strictly analyse differences in viewing behaviour when watching a lecture for the first time and when viewing it for the second time. The findings illustrate that statistically significant differences between re-consumption and first-time viewing behaviour across diverse fields of study is not currently present, emphasising the importance of accounting for the domains when analysing or implementing Technology-Enhanced Learning tools. Moreover, the differences were in general slight, even for statistically significant differences within the same educational domain. On the other hand, when accounting for the domain categories, prelim-

inary results indicate that there are differences firstly between domains, but also highlighting the first time and revisiting differences within them. However, as the analysis was done on a smaller and restricted subset due to users' and lecture intrinsic behaviours, an analysis across a more diverse set of users and lectures would provide more confidence in the results. On another note, a novel approach for re-consumption classification utilising in-video viewing behaviour was executed, indicating that in-video viewing behaviour can be used for learning intervention or proactive recommendations for re-consumptions. Despite the infrequency of re-consumptions, a side-effect of the higher misclassifications of non-re-consumed lectures can provide corrective measures for improving reviewing behaviour as it is been deemed important for learning success as a manifestation of Self-Regulated Learning strategies.

Thirdly, to further evaluate how the re-consumption of lectures affects RSs and their inclination towards recommending repetition of lectures, the recommendation accuracy of previously viewed lectures was evaluated. The results indicate that conventional CF methods are most inclined to recommend users to revisit lectures, compared to SARS. This can be applicable for TEL where either exploring or exploiting behaviour, in general, is recommended. Contextualising the SARS with in-video viewing behaviour and lecture topics did not drastically change the results overall nor in one specific direction. Taking the RSs' measured inclination into account, a novel evaluation of re-consumption calibration of RSs was carried out, adapting previously used calibration metrics to re-consumption as well as proposing an additional, more explainable task-specific metric. The evaluation highlights SARS ability to align with users' re-consumption preferences and distinguish the re-consumed lectures from the ones which are not. Furthermore, there are slight increases in alignment by including in-video behaviour, but it is not consistent for learning resources across diverse educational domains. Lastly, the re-consumption calibration evaluation and adaption of calibration metrics is also applicable to other domains where re-consumption is an important factor of user behaviour.

6.3 Further work

As this project covers multiple areas of research, the following paragraphs describe potential future works with regard to general preprocessing steps, lecture recommendation, re-consumption analysis, prediction and alignment, as well as evaluation techniques for SARS and next-item prediction tasks.

Firstly, a more in-depth study of the drawbacks of the SARS lecture recommendation accuracy would be interesting, i.e. clustering the users by in-video viewing behaviour or more generally by activity level or by lecture diversity. Moreover, a thorough ablation study of the effect of additional feature projection and in-video watching features' association with lecture topics could provide insights regarding how such relations may be extracted. Furthermore, including other learning resources in the sequence-aware models could better represent users' complete study workflow, and improve recommendation accuracy as experienced in e-commerce [82]. This is also interesting for re-consumption analysis as for instance completing assignments is associated with reviewing [28]. There are also multiple interesting analysis aspects of a syllabus approach and usage to further understand how the proposed course order is in fact utilised, raising questions such as to what degree is the syllabi used sequentially. Moreover for large-scale diverse MOOC platforms, to what extent do users context switch between topics across fields on large-scale platforms? Regarding feature inclusion in RSs, including user demographics with in-video viewing behaviour, as well temporal features such as recency could be

interesting in general as well as improved feature reduction and engineering methods as a part of the model instead of a simple, inefficient FFNs [78]. Moreover, as reinforcement learning has been shown to be effective for exercise recommendation [35], applying it to lecture recommendation with in-video viewing behaviours is also of interest.

Within the scope of re-consumption analysis, a more advanced feature selection and engineering methodology could be interesting to explore. Though Exp.1 did not see drastic improvements considering random feature subsets, the classifiers in Exp.2 could have a positive effect on more appropriately engineered behavioural features, as well as the inclusion of categorical lecture features. Moreover, quantifying the effect of in-video behavioural bias related to users and lectures could enable a less restricted dataset for comparison, as well providing more confidence in the analysis. Moreover, the temporal aspects of re-consumption are still mainly unexplored on a large-scale, diverse dataset, e.g. the temporal distances between the first session and a re-consumption of a lecture. This is also of interest for re-consumption prediction. Furthermore, a more nuanced definition of a lecture re-consumption and its consequences for re-consumption analysis is another area of interest. A possibility is to look at re-consumption at a topic level, rather than at per video level, or study it at a more granular per video segment level which is enabled by the release of the PEEK dataset [117].

For further study of re-consumption prediction, a natural extension would be to make the CF models re-consumption aware, applying information retrieval weighting techniques to adjust the confidence values of user-lecture interaction. Moreover, using sequence-aware models as in [106] to measure any re-consumption classification improvement when considering users' historic behaviour, as well as its relation to forgetting behaviour and Knowledge Tracing is enticing. Moreover, online evaluation of the consequences of re-consumption aware RSs or re-consumption-based interventions and their impact on student performance and SRL behaviour could further illustrate the importance of revision for learning. This could also require modelling of the timing aspect of re-consumption recommendations, such as *spaced repetition* [145], as it is deemed an important part of re-consumption recommendations. [61] Lastly, due to the large imbalance of first visits and re-consumptions, more specific anomaly detection techniques would be of interest.

Finally, regarding re-consumption alignment, an alignment metric which considers both the magnitude of revisits of individual lectures, as well as the diversity of the re-consumption would be of high priority. Moreover, evaluating the calibration using weighting schemes of both user history and recommendations would enable the inclusion of temporal aspects, such as recency. In general, exploring more nuanced evaluation techniques and their consideration of relevancy instead of a binary next-item target, could in turn provide more nuanced recommendation accuracy results and perhaps better simulate online evaluation.

Bibliography

[1] E. Øien, 'Exploring Sequential Behaviour for Educational Videos,' Unpublished, Trondheim, Norway, Dec. 2022.

[2] M. Rohs and M. Ganz, 'MOOCs and the claim of education for all: A disillusion by empirical data,' *The International Review of Research in Open and Distributed Learning*, vol. 16, no. 6, Dec. 2015, ISSN: 1492-3831. DOI: 10.19173/irrodl.v16i6.2033. (visited on 09/05/2023).

[3] J. Reich and J. A. Ruipérez-Valiente, 'The MOOC pivot,' *Science*, vol. 363, no. 6423, pp. 130–131, Jan. 2019. DOI: 10.1126/science.aav7958. (visited on 09/05/2023).

[4] L. Ma and C. S. Lee, 'Investigating the adoption of MOOCs: A technology–user–environment perspective,' *Journal of Computer Assisted Learning*, vol. 35, no. 1, pp. 89–98, 2019, ISSN: 1365-2729. DOI: 10.1111/jcal.12314. (visited on 09/05/2023).

[5] W. Feng, J. Tang and T. X. Liu, 'Understanding Dropouts in MOOCs,' *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 517–524, Jul. 2019, ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.3301517. (visited on 26/05/2023).

[6] C. Gütl, R. H. Rizzardini, V. Chang and M. Morales, 'Attrition in MOOC: Lessons Learned from Drop-Out Students,' in *Learning Technology for Education in Cloud. MOOC and Big Data*, L. Uden, J. Sinclair, Y.-H. Tao and D. Liberona, Eds., ser. Communications in Computer and Information Science, Cham: Springer International Publishing, 2014, pp. 37–48, ISBN: 978-3-319-10671-7. DOI: 10.1007/978-3-319-10671-7_4.

[7] H. Huang, L. Jew and D. Qi, 'Take a MOOC and then drop: A systematic review of MOOC engagement pattern and dropout factor,' *Heliyon*, vol. 9, no. 4, e15220, Apr. 2023, ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2023.e15220. (visited on 26/05/2023).

[8] S. Rizvi, B. Rienties, J. Rogaten and R. F. Kizilcec, 'Beyond one-size-fits-all in MOOCs: Variation in learning design and persistence of learners in different cultural and socioeconomic contexts,' *Computers in Human Behavior*, vol. 126, p. 106 973, Jan. 2022, ISSN: 0747-5632. DOI: 10.1016/j.chb.2021.106973. (visited on 26/05/2023).

[9] P. G. de Barba, D. Malekian, E. A. Oliveira, J. Bailey, T. Ryan and G. Kennedy, 'The importance and meaning of session behaviour in a MOOC,' *Computers & Education*, vol. 146, p. 103 772, Mar. 2020, ISSN: 0360-1315. DOI: 10.1016/j.compedu.2019.103772. (visited on 26/05/2023).

[10] M. N. Giannakos, K. Chorianopoulos and N. Chrisochoides, 'Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course,' *The International Review of Research in Open and Distributed Learning*, vol. 16, no. 1, Jan. 2015, ISSN: 1492-3831. DOI: 10.19173/irrodl.v16i1.1976. (visited on 24/01/2023).

[11] P. Branch, G. Egan and B. Tonkin, 'Modeling interactive behaviour of a video based multimedia system,' in *1999 IEEE International Conference on Communications (Cat. No. 99CH36311)*, vol. 2, Jun. 1999, 978–982 vol.2. DOI: 10.1109/ICC.1999.765419.

[12] P. J. Guo, J. Kim and R. Rubin, 'How video production affects student engagement: An empirical study of MOOC videos,' in *Proceedings of the First ACM Conference on Learning @ Scale Conference*, ser. L@S '14, New York, NY, USA: Association for Computing Machinery, Mar. 2014, pp. 41–50, ISBN: 978-1-4503-2669-8. DOI: 10.1145/2556325.2566239. (visited on 24/01/2023).

[13] A. S. Lan, C. G. Brinton, T.-Y. Yang and M. Chiang, 'Behavior-based latent variable model for learner engagement,' in *International Conference on Educational Data Mining (EDM)*, Wuhan, China: International Educational Data Mining Society, Jun. 2017.

[14] C. G. Brinton and M. Chiang, 'MOOC performance prediction via clickstream data and social learning networks,' in *2015 IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2015, pp. 2299–2307. DOI: 10.1109/INFOCOM.2015.7218617.

[15] C. G. Brinton, S. Buccapatnam, M. Chiang and H. V. Poor, 'Mining MOOC Clickstreams: Video-Watching Behavior vs. In-Video Quiz Performance,' *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3677–3692, Jul. 2016, ISSN: 1941-0476. DOI: 10.1109/TSP.2016.2546228.

[16] D. Lang, G. Chen, K. Mirzaei and A. Paepcke, 'Is faster better? a study of video playback speed,' in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, ser. LAK '20, New York, NY, USA: Association for Computing Machinery, Mar. 2020, pp. 260–269, ISBN: 978-1-4503-7712-6. DOI: 10.1145/3375462.3375466. (visited on 24/01/2023).

[17] N. Li, Ł. Kidziński, P. Jermann and P. Dillenbourg, 'MOOC Video Interaction Patterns: What Do They Tell Us?' In *Design for Teaching and Learning in a Networked World*, G. Conole, T. Klobučar, C. Rensing, J. Konert and E. Lavoué, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 197–210, ISBN: 978-3-319-24258-3. DOI: 10.1007/978-3-319-24258-3_15.

[18] 'How Do In-video Interactions Reflect Perceived Video Difficulty?' *Proceedings of the European MOOCs Stakeholder Summit 2015*, N. Li, L. Kidzinski, P. Jermann and P. Dillenbourg, Eds., 2015.

[19] J. Costley, M. Fanguy, C. Lange and M. Baldwin, 'The effects of video lecture viewing strategies on cognitive load,' *Journal of Computing in Higher Education*, vol. 33, no. 1, pp. 19–38, Apr. 2021, ISSN: 1867-1233. DOI: 10.1007/s12528-020-09254-y. (visited on 24/01/2023).

[20] T. Sinha, P. Jermann, N. Li and P. Dillenbourg, 'Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions,' in *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 3–14. DOI: 10.3115/v1/W14-4102. (visited on 13/05/2023).

[21] R. A. Adara, 'MOOC as an Alternative for Teaching and Learning During Covid-19 Pandemic: Students' Motivation and Demotivation,' *Celt: A Journal of Culture, English Language Teaching & Literature*, vol. 21, no. 2, pp. 203–223, Dec. 2021, ISSN: 2502-4914. DOI: 10.24167/celt.v21i2.3255. (visited on 26/05/2023).

[22] X. Wei, S. Sun, D. Wu and L. Zhou, 'Personalized Online Learning Resource Recommendation Based on Artificial Intelligence and Educational Psychology,' *Frontiers in Psychology*, vol. 12, 2021, ISSN: 1664-1078. (visited on 23/01/2023).

[23] S. Tang and Z. A. Pardos, 'Personalized Behavior Recommendation: A Case Study of Applicability to 13 Courses on edX,' in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '17, New York, NY, USA: Association for Computing Machinery, Jul. 2017, pp. 165–170, ISBN: 978-1-4503-5067-9. DOI: 10.1145/3099023.3099038. (visited on 26/01/2023).

[24] I. Uddin, A. S. Imran, K. Muhammad, N. Fayyaz and M. Sajjad, 'A Systematic Mapping Review on MOOC Recommender Systems,' *IEEE Access*, vol. 9, pp. 118 379–118 405, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3101039.

[25] M. Shridharan, A. Willingham, J. Spencer, T.-Y. Yang and C. Brinton, 'Predictive learning analytics for video-watching behavior in MOOCs,' in *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2018, pp. 1–6. DOI: 10.1109/CISS.2018.8362323.

[26] S. L. Dazo, N. R. Stepanek, R. Fulkerson and B. Dorn, 'An Empirical Analysis of Video Viewing Behaviors in Flipped CS1 Courses,' in *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '16, New York, NY, USA: Association for Computing Machinery, Jul. 2016, pp. 106–111, ISBN: 978-1-4503-4231-5. DOI: 10.1145/2899415.2899468. (visited on 19/05/2023).

[27] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos and R. C. Miller, 'Understanding in-video dropouts and interaction peaks inonline lecture videos,' in *Proceedings of the First ACM Conference on Learning @ Scale Conference*, ser. L@S '14, New York, NY, USA: Association for Computing Machinery, Mar. 2014, pp. 31–40, ISBN: 978-1-4503-2669-8. DOI: 10.1145/2556325.2566237. (visited on 26/01/2023).

[28] R. F. Kizilcec, M. Pérez-Sanagustín and J. J. Maldonado, 'Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses,' *Computers & Education*, vol. 104, pp. 18–33, Jan. 2017, ISSN: 0360-1315. DOI: 10.1016/j.compedu.2016.10.001. (visited on 26/05/2023).

[29] L. Ma and C. S. Lee, 'Drivers and barriers to MOOC adoption: Perspectives from adopters and non-adopters,' *Online Information Review*, vol. 44, no. 3, pp. 671–684, Jan. 2020, ISSN: 1468-4527. DOI: 10.1108/OIR-06-2019-0203. (visited on 09/05/2023).

[30] S. H. K. Kang, 'Spaced Repetition Promotes Efficient and Effective Learning: Policy Implications for Instruction,' *Policy Insights from the Behavioral and Brain Sciences*, vol. 3, no. 1, pp. 12–19, Mar. 2016, ISSN: 2372-7322. DOI: 10.1177/2372732215624708. (visited on 11/06/2023).

[31] P. J. Guo and K. Reinecke, 'Demographic differences in how students navigate through MOOCs,' in *Proceedings of the First ACM Conference on Learning @ Scale Conference*, ser. L@S '14, New York, NY, USA: Association for Computing Machinery, Mar. 2014, pp. 21–30, ISBN: 978-1-4503-2669-8. DOI: 10.1145/2556325.2566247. (visited on 24/01/2023).

[32] A. Anderson, R. Kumar, A. Tomkins and S. Vassilvitskii, 'The dynamics of repeat consumption,' in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14, New York, NY, USA: Association for Computing Machinery, Apr. 2014, pp. 419–430, ISBN: 978-1-4503-2744-2. DOI: 10.1145/2566486.2568018. (visited on 27/05/2023).

[33] J. Chen, C. Wang and J. Wang, 'Will You "Reconsume" the Near Past? Fast Prediction on Short-Term Reconsumption Behaviors,' *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, Feb. 2015, ISSN: 2374-3468. DOI: 10.1609/aaai.v29i1.9172. (visited on 24/05/2023).

[34] J. Chen, C. Wang, J. Wang and P. S. Yu, 'Recommendation for Repeat Consumption from User Implicit Feedback,' *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 3083–3097, Nov. 2016, ISSN: 1558-2191. DOI: 10.1109/TKDE.2016.2593720.

[35] Z. Huang, Q. Liu, C. Zhai, Y. Yin, E. Chen, W. Gao and G. Hu, 'Exploring Multi-Objective Exercise Recommendations in Online Education Systems,' in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19, New York, NY, USA: Association for Computing Machinery, Nov. 2019, pp. 1261–1270, ISBN: 978-1-4503-6976-3. DOI: 10.1145/3357384.3357995. (visited on 22/05/2023).

[36] G. Shani and A. Gunawardana, 'Evaluating Recommendation Systems,' in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira and P. B. Kantor, Eds., Boston, MA: Springer US, 2011, pp. 257–297, ISBN: 978-0-387-85820-3. DOI: 10.1007/978-0-387-85820-3_8. (visited on 02/02/2023).

[37] H. Abdollahpouri, Z. Nazari, A. Gain, C. Gibson, M. Dimakopoulou, J. Anderton, B. Carterette, M. Lalmas and T. Jebara, 'Calibrated Recommendations as a Minimum-Cost Flow Problem,' in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '23, New York, NY, USA: Association for Computing Machinery, Feb. 2023, pp. 571–579, ISBN: 978-1-4503-9407-9. DOI: 10.1145/3539597.3570402. (visited on 06/05/2023).

[38] A. Sangrà, D. Vlachopoulos and N. Cabrera, 'Building an inclusive definition of e-learning: An approach to the conceptual framework,' *The International Review of Research in Open and Distributed Learning*, vol. 13, no. 2, p. 145, Apr. 2012, ISSN: 1492-3831. DOI: 10.19173/irrodl.v13i2.1161. (visited on 05/06/2023).

[39] C.-w. Shen and J.-t. Ho, 'Technology-enhanced learning in higher education: A bibliometric analysis with latent semantic approach,' *Computers in Human Behavior*, vol. 104, p. 106 177, Mar. 2020, ISSN: 0747-5632. DOI: 10.1016/j.chb.2019.106177. (visited on 05/06/2023).

[40] C. Brooks, C. D. Epp, G. Logan and J. Greer, 'The who, what, when, and why of lecture capture,' in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, ser. LAK '11, New York, NY, USA: Association for Computing Machinery, Feb. 2011, pp. 86–92, ISBN: 978-1-4503-0944-8. DOI: 10.1145/2090116.2090128. (visited on 20/05/2023).

[41] A. Le, S. Joordens, S. Chrysostomou and R. Grinnell, 'Online lecture accessibility and its influence on performance in skills-based courses,' *Computers & Education*, vol. 55, no. 1, pp. 313–319, Aug. 2010, ISSN: 0360-1315. DOI: 10.1016/j.compedu.2010.01.017. (visited on 24/01/2023).

[42] C. Moore, L. Battestilli and I. X. Domínguez, 'Finding Video-watching Behavior Patterns in a Flipped CS1 Course,' in *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '21, New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 768–774, ISBN: 978-1-4503-8062-1. DOI: 10.1145/3408877.3432359. (visited on 24/01/2023).

[43] I. Chuang and A. Ho, *HarvardX and MITx: Four Years of Open Online Courses – Fall 2012-Summer 2016*, SSRN Scholarly Paper, Rochester, NY, Dec. 2016. DOI: 10.2139/ssrn.2889436. (visited on 24/01/2023).

[44] J. Ruiz-Palmero, J.-M. Fernández-Lacorte, E. Sánchez-Rivas and E. Colomo-Magaña, 'The implementation of Small Private Online Courses (SPOC) as a new approach to education,' *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, p. 27, Aug. 2020, ISSN: 2365-9440. DOI: 10.1186/s41239-020-00206-1. (visited on 05/06/2023).

[45] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu and F. M. F. Wong, 'Learning about Social Learning in MOOCs: From Statistical Analysis to Generative Model,' *IEEE Transactions on Learning Technologies*, vol. 7, no. 4, pp. 346–359, Oct. 2014, ISSN: 1939-1382. DOI: 10.1109/TLT.2014.2337900.

[46] T. Hastie, R. Tibshirani and J. Friedman, 'Linear Methods for Classification,' in *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009, pp. 101–137, ISBN: 978-0-387-84857-0 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_4. (visited on 10/06/2023).

[47] C. Cortes and V. Vapnik, 'Support-vector networks,' *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, ISSN: 1573-0565. DOI: 10.1007/BF00994018. (visited on 12/06/2023).

[48] T. Hastie, R. Tibshirani and J. Friedman, 'Support Vector Machines and Flexible Discriminants,' in *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009, pp. 417–458, ISBN: 978-0-387-84857-0 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_12. (visited on 10/06/2023).

[49] B. E. Boser, I. M. Guyon and V. N. Vapnik, 'A training algorithm for optimal margin classifiers,' in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ser. COLT '92, New York, NY, USA: Association for Computing Machinery, Jul. 1992, pp. 144–152, ISBN: 978-0-89791-497-0. DOI: 10.1145/130385.130401. (visited on 12/06/2023).

[50] T. Hastie, R. Tibshirani and J. Friedman, 'Boosting and Additive Trees,' in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer Series in Statistics, T. Hastie, R. Tibshirani and J. Friedman, Eds., New York, NY: Springer, 2009, pp. 337–387, ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_10. (visited on 12/06/2023).

[51] A. Sherstinsky, 'Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network,' *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, Mar. 2020, ISSN: 0167-2789. DOI: 10.1016/j.physd.2019.132306. (visited on 12/06/2023).

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, 'Attention is All you Need,' in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. (visited on 12/06/2023).

[53] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,' in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. (visited on 09/06/2023).

[54] L. Zhuang, L. Wayne, S. Ya and Z. Jun, 'A Robustly Optimized BERT Pre-training Approach with Post-training,' in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, Huhhot, China: Chinese Information Processing Society of China, Aug. 2021, pp. 1218–1227. (visited on 13/06/2023).

[55] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov and Q. V. Le, 'XLNet: Generalized Autoregressive Pretraining for Language Understanding,' in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019. (visited on 09/06/2023).

[56] C. C. Aggarwal, *Recommender Systems*. Cham: Springer International Publishing, 2016, ISBN: 978-3-319-29657-9 978-3-319-29659-3. DOI: 10.1007/978-3-319-29659-3. (visited on 05/06/2023).

[57] P. G. Campos, F. Díez and I. Cantador, 'Time-aware recommender systems: A comprehensive survey and analysis of existing evaluation protocols,' *User Modeling and User-Adapted Interaction*, vol. 24, no. 1, pp. 67–119, Feb. 2014, ISSN: 1573-1391. DOI: 10.1007/s11257-012-9136-x. (visited on 05/06/2023).

[58] R. Burke, 'Hybrid Web Recommender Systems,' in *The Adaptive Web: Methods and Strategies of Web Personalization*, ser. Lecture Notes in Computer Science, P. Brusilovsky, A. Kobsa and W. Nejdl, Eds., Berlin, Heidelberg: Springer, 2007, pp. 377–408, ISBN: 978-3-540-72079-9. DOI: 10.1007/978-3-540-72079-9_12. (visited on 05/06/2023).

[59] Y. Hu, Y. Koren and C. Volinsky, 'Collaborative Filtering for Implicit Feedback Datasets,' in *2008 Eighth IEEE International Conference on Data Mining*, Dec. 2008, pp. 263–272. DOI: 10.1109/ICDM.2008.22.

[60] R. Burke, 'Hybrid Recommender Systems: Survey and Experiments,' *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, Nov. 2002, ISSN: 1573-1391. DOI: 10.1023/A:1021240730564. (visited on 05/06/2023).

[61] M. Quadrana, P. Cremonesi and D. Jannach, 'Sequence-Aware Recommender Systems,' *ACM Computing Surveys*, vol. 51, no. 4, 66:1–66:36, Jul. 2018, ISSN: 0360-0300. DOI: 10.1145/3190616. (visited on 26/01/2023).

[62] W. Greller and H. Drachsler, 'Translating learning into numbers: A generic framework for learning analytics,' *Journal of Educational Technology & Society*, vol. 15, pp. 42–57, 2012, ISSN: 1436-4522.

[63] H. Drachsler, K. Verbert, O. C. Santos and N. Manouselis, 'Panorama of Recommender Systems to Support Learning,' in *Recommender Systems Handbook*, F. Ricci, L. Rokach and B. Shapira, Eds., Boston, MA: Springer US, 2015, pp. 421–451, ISBN: 978-1-4899-7636-9 978-1-4899-7637-6. DOI: 10.1007/978-1-4899-7637-6_12. (visited on 19/04/2023).

[64] Y. Zhou, C. Huang, Q. Hu, J. Zhu and Y. Tang, 'Personalized learning full-path recommendation model based on LSTM neural networks,' *Information Sciences*, vol. 444, pp. 135–152, May 2018, ISSN: 0020-0255. DOI: 10.1016/j.ins.2018.02.053. (visited on 03/02/2023).

[65] H. Steck, 'Calibrated recommendations,' in *Proceedings of the 12th ACM Conference on Recommender Systems*, ser. RecSys '18, New York, NY, USA: Association for Computing Machinery, Sep. 2018, pp. 154–162, ISBN: 978-1-4503-5901-6. DOI: 10.1145/3240323.3240372. (visited on 20/05/2023).

[66] S. Kullback and R. A. Leibler, 'On Information and Sufficiency,' *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951, ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177729694. (visited on 04/06/2023).

[67] C. Goutte and E. Gaussier, 'A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation,' in *Advances in Information Retrieval*, D. E. Losada and J. M. Fernández-Luna, Eds., ser. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2005, pp. 345–359, ISBN: 978-3-540-31865-1. DOI: 10.1007/978-3-540-31865-1_25.

[68] W. Krichene and S. Rendle, 'On Sampled Metrics for Item Recommendation,' in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20, New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 1748–1757, ISBN: 978-1-4503-7998-4. DOI: 10.1145/3394486.3403226. (visited on 04/06/2023).

[69] A. Banerjee, U. B. Chitnis, S. L. Jadhav, J. S. Bhawalkar and S. Chaudhury, 'Hypothesis testing, type I and type II errors,' *Industrial Psychiatry Journal*, vol. 18, no. 2, pp. 127–131, 2009, ISSN: 0972-6748. DOI: 10.4103/0972-6748.62274. (visited on 04/06/2023).

[70] J. Demšar, 'Statistical Comparisons of Classifiers over Multiple Data Sets,' *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 1–30, 2006, ISSN: 1533-7928. (visited on 28/03/2023).

[71] F. Wilcoxon, 'Individual Comparisons by Ranking Methods,' *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945, ISSN: 0099-4987. DOI: 10.2307/3001968. JSTOR: 3001968. (visited on 04/06/2023).

[72] B. Trawiński, M. Smętek, Z. Telec and T. Lasota, 'Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms,' *International Journal of Applied Mathematics and Computer Science*, vol. 22, no. 4, pp. 867–881, Dec. 2012. DOI: 10.2478/v10006-012-0064-z. (visited on 28/03/2023).

[73] J. L. Hodges, 'The significance probability of the smirnov two-sample test,' *Arkiv för Matematik*, vol. 3, no. 5, pp. 469–486, Jan. 1958, ISSN: 1871-2487. DOI: 10.1007/BF02589501. (visited on 04/06/2023).

[74] S. Holm, 'A Simple Sequentially Rejective Multiple Test Procedure,' *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979, ISSN: 0303-6898. JSTOR: 4615733. (visited on 04/06/2023).

[75] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic and E. Duval, 'Context-Aware Recommender Systems for Learning: A Survey and Future Challenges,' *IEEE Transactions on Learning Technologies*, vol. 5, no. 4, pp. 318–335, Oct. 2012, ISSN: 1939-1382. DOI: 10.1109/TLT.2012.11.

[76] B. Hidasi, M. Quadrana, A. Karatzoglou and D. Tikk, 'Parallel Recurrent Neural Network Architectures for Feature-rich Session-based Recommendations,' in *Proceedings of the 10th ACM Conference on Recommender Systems*, ser. RecSys '16, New York, NY, USA: Association for Computing Machinery, Sep. 2016, pp. 241–248, ISBN: 978-1-4503-4035-9. DOI: 10.1145/2959100.2959167. (visited on 23/05/2023).

[77] Q. Chen, H. Zhao, W. Li, P. Huang and W. Ou, 'Behavior sequence transformer for e-commerce recommendation in Alibaba,' in *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*, ser. DLP-KDD '19, New York, NY, USA: Association for Computing Machinery, Aug. 2019, pp. 1–4, ISBN: 978-1-4503-6783-7. DOI: 10.1145/3326937.3341261. (visited on 20/05/2023).

[78] A. Beutel, P. Covington, S. Jain, C. Xu, J. Li, V. Gatto and E. H. Chi, 'Latent Cross: Making Use of Context in Recurrent Recommender Systems,' in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, ser. WSDM '18, New York, NY, USA: Association for Computing Machinery, Feb. 2018, pp. 46–54, ISBN: 978-1-4503-5581-0. DOI: 10.1145/3159652.3159727. (visited on 21/01/2023).

[79] S. Mizrachi and P. Levin, 'Combining context features in sequence-aware recommender systems,' 2019.

[80] G. De Souza Pereira Moreira, S. Rabhi, J. M. Lee, R. Ak and E. Oldridge, 'Transformers4Rec: Bridging the Gap between NLP and Sequential / Session-Based Recommendation,' in *Fifteenth ACM Conference on Recommender Systems*, Amsterdam Netherlands: ACM, Sep. 2021, pp. 143–153, ISBN: 978-1-4503-8458-2. DOI: 10.1145/3460231.3474255. (visited on 15/05/2023).

[81] Y. Li and S. Bengio, 'TIME-DEPENDENT REPRESENTATION FOR NEURAL EVENT SEQUENCE PREDICTION,' 2018.

[82] M. Celikik, A. Peleteiro Ramallo and J. Wasilewski, 'Reusable Self-Attention Recommender Systems in Fashion Industry Applications,' in *Proceedings of the 16th ACM Conference on Recommender Systems*, ser. RecSys '22, New York, NY, USA: Association for Computing Machinery, Sep. 2022, pp. 448–451, ISBN: 978-1-4503-9278-5. DOI: 10.1145/3523227.3547377. (visited on 12/06/2023).

[83] A. Rashed, S. Elsayed and L. Schmidt-Thieme, 'Context and Attribute-Aware Sequential Recommendation via Cross-Attention,' in *Sixteenth ACM Conference on Recommender Systems*, Seattle WA USA: ACM, Sep. 2022, pp. 71–80, ISBN: 978-1-4503-9278-5. DOI: 10.1145/3523227.3546777. (visited on 23/05/2023).

[84] W.-C. Kang and J. McAuley, 'Self-Attentive Sequential Recommendation,' in *2018 IEEE International Conference on Data Mining (ICDM)*, Nov. 2018, pp. 197–206. DOI: 10.1109/ICDM.2018.00035.

[85] A. Khalid, K. Lundqvist and A. Yates, 'Recommender Systems for MOOCs: A Systematic Literature Survey (January 1, 2012 – July 12, 2019),' *International Review of Research in Open and Distributed Learning*, vol. 21, no. 4, pp. 255–291, 2020, ISSN: 1492-3831. DOI: 10.19173/irrodl.v21i4.4643. (visited on 11/06/2023).

[86] S. M. Nafea, F. Siewe and Y. He, 'On Recommendation of Learning Objects Using Felder-Silverman Learning Style Model,' *IEEE Access*, vol. 7, pp. 163 034–163 048, 2019, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2935417.

[87] J. K. Tarus, Z. Niu and A. Yousif, 'A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining,' *Future Generation Computer Systems*, vol. 72, pp. 37–48, Jul. 2017, ISSN: 0167-739X. DOI: 10.1016/j.future.2017.02.049. (visited on 23/01/2023).

[88] R. M. Felder, 'LEARNING AND TEACHING STYLES IN ENGINEERING EDUCATION,'

[89] B. Cheng, Y. Zhang and D. Shi, 'Ontology-Based Personalized Learning Path Recommendation for Course Learning,' in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, Oct. 2018, pp. 531–535. DOI: 10.1109/ITME.2018.00123.

[90] J. Gong, S. Wang, J. Wang, W. Feng, H. Peng, J. Tang and P. S. Yu, 'Attentional Graph Convolutional Networks for Knowledge Concept Recommendation in MOOCs in a Heterogeneous View,' in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20, New York, NY, USA: Association for Computing Machinery, Jul. 2020, pp. 79–88, ISBN: 978-1-4503-8016-4. DOI: 10.1145/3397271.3401057. (visited on 02/02/2023).

[91] H. Zhu, Y. Liu, F. Tian, Y. Ni, K. Wu, Y. Chen and Q. Zheng, 'A Cross-Curriculum Video Recommendation Algorithm Based on a Video-Associated Knowledge Map,' *IEEE Access*, vol. 6, pp. 57 562–57 571, 2018, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2873106.

[92] J. Zhao, C. Bhatt, M. Cooper and D. A. Shamma, 'Flexible Learning with Semantic Visual Exploration and Sequence-Based Recommendation of MOOC Videos,' in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18, New York, NY, USA: Association for Computing Machinery, Apr. 2018, pp. 1–13, ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173903. (visited on 20/05/2023).

[93] 'TF–IDF,' in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds., Boston, MA: Springer US, 2010, pp. 986–987, ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_832. (visited on 09/06/2023).

[94] H. Chen, C. Yin, R. Li, W. Rong, Z. Xiong and B. David, 'Enhanced learning resource recommendation based on online learning style model,' *Tsinghua Science and Technology*, vol. 25, no. 3, pp. 348–356, Jun. 2020, ISSN: 1007-0214. DOI: 10.26599/TST.2019.9010014.

[95] H. He, Z. Zhu, Q. Guo and X. Huang, 'A Personalized E-Learning Services Recommendation Algorithm Based on User Learning Ability,' in *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, vol. 2161-377X, Jul. 2019, pp. 318–320. DOI: 10.1109/ICALT.2019.00099.

[96] C. Bhatt, M. Cooper and J. Zhao, 'SeqSense: Video Recommendation Using Topic Sequence Mining,' in *MultiMedia Modeling*, K. Schoeffmann, T. H. Chalidabhongse, C. W. Ngo, S. Aramvith, N. E. O'Connor, Y.-S. Ho, M. Gabbouj and A. Elgammal, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 252–263, ISBN: 978-3-319-73600-6. DOI: 10.1007/978-3-319-73600-6_22.

[97] Z. A. Pardos, S. Tang, D. Davis and C. V. Le, 'Enabling Real-Time Adaptivity in MOOCs with a Personalized Next-Step Recommendation Framework,' in *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, ser. L@S '17, New York, NY, USA: Association for Computing Machinery, Apr. 2017, pp. 23–32, ISBN: 978-1-4503-4450-0. DOI: 10.1145/3051457.3051471. (visited on 20/05/2023).

[98] D. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition*. CRC Press, 2020, ISBN: 978-1-4398-5804-2.

[99] L. Battestilli, I. X. Domínguez and M. Thyagarajan, 'Toward Finding Online Activity Patterns in a Flipped Programming Course,' in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '20, New York, NY, USA: Association for Computing Machinery, Feb. 2020, p. 1345, ISBN: 978-1-4503-6793-6. DOI: 10.1145/3328778.3372626. (visited on 19/05/2023).

[100] C. Spearman, 'The Proof and Measurement of Association between Two Things,' *The American Journal of Psychology*, vol. 100, no. 3/4, pp. 441–471, 1987, ISSN: 0002-9556. DOI: 10.2307/1422689. JSTOR: 1422689. (visited on 04/06/2023).

[101] J. M. Aiken, S.-Y. Lin, S. S. Douglas, E. F. Greco, B. D. Thoms, M. D. Caballero and M. F. Schatz, 'Student Use of a Single Lecture Video in a Flipped Introductory Mechanics Course,' in *2014 Physics Education Research Conference Proceedings*, Apr. 2015, pp. 19–22. DOI: 10.1119/perc.2014.pr.001. arXiv: 1407.2620 [physics]. (visited on 18/04/2023).

[102] G. Akcapinar and A. Bayazit, 'Investigating Video Viewing Behaviors of Students with Different Learning Approaches Using Video Analytics,' *Turkish Online Journal of Distance Education*, vol. 19, no. 4, pp. 116–125, Oct. 2018, ISSN: 1302-6488. DOI: 10.17718/tojde.471907. (visited on 25/01/2023).

[103] C. Shi, S. Fu, Q. Chen and H. Qu, 'VisMOOC: Visualizing video clickstream data from Massive Open Online Courses,' in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, Apr. 2015, pp. 159–166. DOI: 10.1109/PACIFICVIS.2015.7156373.

[104] G. Abdelrahman, Q. Wang and B. Nunes, 'Knowledge Tracing: A Survey,' *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–37, Nov. 2023, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3569576. (visited on 06/06/2023).

[105] T.-Y. Yang, C. G. Brinton, C. Joe-Wong and M. Chiang, 'Behavior-Based Grade Prediction for MOOCs Via Time Series Neural Networks,' *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 716–728, Aug. 2017, ISSN: 1941-0484. DOI: 10.1109/JSTSP.2017.2700227.

[106] P. Ren, Z. Chen, J. Li, Z. Ren, J. Ma and M. de Rijke, 'RepeatNet: A Repeat Aware Neural Recommendation Machine for Session-Based Recommendation,' *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 4806–4813, Jul. 2019, ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.33014806. (visited on 20/05/2023).

[107] M. Feng, N. Heffernan and K. Koedinger, 'Addressing the assessment challenge with an online system that tutors as it assesses,' *User Modeling and User-Adapted Interaction*, vol. 19, no. 3, pp. 243–266, Aug. 2009, ISSN: 1573-1391. DOI: 10.1007/s11257-009-9063-7. (visited on 11/05/2023).

[108] Z. A. Pardos, R. S. Baker, M. San Pedro, S. M. Gowda and S. M. Gowda, 'Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes,' *Journal of Learning Analytics*, vol. 1, no. 1, pp. 107–128, May 2014, ISSN: 1929-7750. DOI: 10.18608/jla.2014.11.6. (visited on 11/05/2023).

[109] H.-S. Chang, H.-J. Hsu and K.-T. Chen, 'Modeling Exercise Relationships in E-Learning: A Unified Approach,'

[110] J. Stamper, A. Niculescu-Mizil, S. Ritter, G. J. Gordon and K. R. Koedinger, *Algebra I 2006-2007. Challenge Data Set from KDD Cup 2010 Educational Data Mining Challenge*, 2010.

[111] L. Cao and C. Zhang, 'KDD Cup 2015—Predicting Dropouts in MOOC,' in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[112] J. Kuzilek, M. Hlosta and Z. Zdrahal, 'Open University Learning Analytics dataset,' *Scientific Data*, vol. 4, no. 1, p. 170 171, Nov. 2017, ISSN: 2052-4463. DOI: 10.1038/sdata.2017.171. (visited on 11/05/2023).

[113] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz and J. Shawe-Taylor, *VLEngagement: A Dataset of Scientific Video Lectures for Evaluating Population-based Engagement*, Nov. 2020. DOI: 10.48550/arXiv.2011.02273. arXiv: 2011.02273 [cs, stat]. (visited on 02/02/2023).

[114] Y. Choi, Y. Lee, D. Shin, J. Cho, S. Park, S. Lee, J. Baek, C. Bae, B. Kim and J. Heo, 'EdNet: A Large-Scale Hierarchical Dataset in Education,' in *Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin and E. Millán, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 69–73, ISBN: 978-3-030-52240-7. DOI: 10.1007/978-3-030-52240-7_13.

[115] J. Yu, G. Luo, T. Xiao, Q. Zhong, Y. Wang, W. Feng, J. Luo, C. Wang, L. Hou, J. Li, Z. Liu and J. Tang, 'MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs,' in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 3135–3142. DOI: 10.18653/v1/2020.acl-main.285. (visited on 11/05/2023).

[116] J. Yu, Y. Wang, Q. Zhong, G. Luo, Y. Mao, K. Sun, W. Feng, W. Xu, S. Cao, K. Zeng, Z. Yao, L. Hou, Y. Lin, P. Li, J. Zhou, B. Xu, J. Li, J. Tang and M. Sun, 'MOOCCubeX: A Large Knowledge-centered Repository for Adaptive Learning in MOOCs,' in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Virtual Event Queensland Australia: ACM, Oct. 2021, pp. 4643–4652, ISBN: 978-1-4503-8446-9. DOI: 10.1145/3459637.3482010. (visited on 11/05/2023).

[117] S. Bulathwela, M. Perez-Ortiz, E. Novak, E. Yilmaz and J. Shawe-Taylor, *PEEK: A Large Dataset of Learner Engagement with Educational Videos*, Sep. 2021. DOI: 10.48550/arXiv.2109.03154. arXiv: 2109.03154 [cs]. (visited on 31/01/2023).

[118] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch and A. Joulin, 'Advances in pre-training distributed word representations,' in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[119] M. Själander, M. Jahre, G. Tufte and N. Reissmann, *EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure*, Feb. 2022. DOI: 10.48550/arXiv.1912.05848. arXiv: 1912.05848 [cs]. (visited on 30/11/2022).

[120] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou and P. Jiang, 'BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer,' in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19, New York, NY, USA: Association for Computing Machinery, Nov. 2019, pp. 1441–1450, ISBN: 978-1-4503-6976-3. DOI: 10.1145/3357384.3357895. (visited on 23/05/2023).

[121] Z. Sun, D. Yu, H. Fang, J. Yang, X. Qu, J. Zhang and C. Geng, 'Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison,' in *Proceedings of the 14th ACM Conference on Recommender Systems*, ser. RecSys '20, New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 23–32, ISBN: 978-1-4503-7583-2. DOI: 10.1145/3383313.3412489. (visited on 07/03/2023).

[122] I.-K. Yeo and R. A. Johnson, 'A New Family of Power Transformations to Improve Normality or Symmetry,' *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000, ISSN: 0006-3444. JSTOR: 2673623. (visited on 08/06/2023).

[123] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, 'Scikit-learn: Machine learning in Python,' *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[124] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1-4414-1269-7.

[125] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, 'PyTorch: An imperative style, high-performance deep learning library,' in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 721, Red Hook, NY, USA: Curran Associates Inc., Dec. 2019, pp. 8026–8037. (visited on 09/06/2023).

[126] Y. Ji, A. Sun, J. Zhang and C. Li, 'A Re-visit of the Popularity Baseline in Recommender Systems,' in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2020, pp. 1749–1752. DOI: 10.1145/3397271.3401233. arXiv: 2005.13829 [cs]. (visited on 19/04/2023).

[127] S. Rendle, W. Krichene, L. Zhang and Y. Koren, 'Revisiting the Performance of iALS on Item Recommendation Benchmarks,' in *Sixteenth ACM Conference on Recommender Systems*, Seattle WA USA: ACM, Sep. 2022, pp. 427–435, ISBN: 978-1-4503-9278-5. DOI: 10.1145/3523227.3548486. (visited on 08/06/2023).

[128] S. Rendle, C. Freudenthaler, Z. Gantner and L. Schmidt-Thieme, 'BPR: Bayesian personalized ranking from implicit feedback,' in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '09, Arlington, Virginia, USA: AUAI Press, Jun. 2009, pp. 452–461, ISBN: 978-0-9749039-5-8. (visited on 09/06/2023).

[129] C. C. Johnson, 'Logistic matrix factorization for implicit feedback data,' *Advances in Neural Information Processing Systems*, vol. 27, no. 78, pp. 1–9, 2014.

[130] G. Amati, 'BM25,' in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 257–260, ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_921. (visited on 09/06/2023).

[131] B. Hidasi and A. Karatzoglou, 'Recurrent Neural Networks with Top-k Gains for Session-based Recommendations,' in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ser. CIKM '18, New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 843–852, ISBN: 978-1-4503-6014-2. DOI: 10.1145/3269206.3271761. (visited on 23/05/2023).

[132] A. Petrov and C. Macdonald, 'Effective and Efficient Training for Sequential Recommendation using Recency Sampling,' in *Proceedings of the 16th ACM Conference on Recommender Systems*, ser. RecSys '22, New York, NY, USA: Association for Computing Machinery, Sep. 2022, pp. 81–91, ISBN: 978-1-4503-9278-5. DOI: 10.1145/3523227.3546785. (visited on 23/05/2023).

[133] A. Petrov and C. Macdonald, 'A Systematic Review and Replicability Study of BERT4Rec for Sequential Recommendation,' in *Proceedings of the 16th ACM Conference on Recommender Systems*, ser. RecSys '22, New York, NY, USA: Association for Computing Machinery, Sep. 2022, pp. 436–447, ISBN: 978-1-4503-9278-5. DOI: 10.1145/3523227.3548487. (visited on 08/06/2023).

[134] J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, 'Algorithms for Hyper-Parameter Optimization,' in *Advances in Neural Information Processing Systems*, vol. 24, Curran Associates, Inc., 2011. (visited on 20/05/2023).

[135] T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, 'Optuna: A Next-generation Hyperparameter Optimization Framework,' in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19, New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 2623–2631, ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330701. (visited on 09/06/2023).

[136] H. Chen, Y. Lin, M. Pan, L. Wang, C.-C. M. Yeh, X. Li, Y. Zheng, F. Wang and H. Yang, 'Denoising Self-Attentive Sequential Recommendation,' in *Proceedings of the 16th ACM Conference on Recommender Systems*, ser. RecSys '22, New York, NY, USA: Association for Computing Machinery, Sep. 2022, pp. 92–101, ISBN: 978-1-4503-9278-5. DOI: 10.1145/3523227.3546788. (visited on 21/01/2023).

[137] S. Latifi, D. Jannach and A. Ferraro, 'Sequential recommendation: A study on transformers, nearest neighbors and sampled metrics,' *Information Sciences*, vol. 609, pp. 660–678, Sep. 2022, ISSN: 0020-0255. DOI: 10.1016/j.ins.2022.07.079. (visited on 02/02/2023).

[138] D. Kalpić, N. Hlupić and M. Lovrić, 'Student's t-Tests,' in *International Encyclopedia of Statistical Science*, M. Lovric, Ed., Berlin, Heidelberg: Springer, 2011, pp. 1559–1563, ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_641. (visited on 10/06/2023).

[139] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, 'SciPy 1.0: Fundamental algorithms for scientific computing in python,' *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[140] S. Seabold and J. Perktold, 'Statsmodels: Econometric and statistical modeling with python,' in *9th Python in Science Conference*, 2010.

[141] L. Grinsztajn, E. Oyallon and G. Varoquaux, 'Why do tree-based models still outperform deep learning on typical tabular data?' In *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Sep. 2022. (visited on 10/06/2023).

[142] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System,' in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. (visited on 10/06/2023).

[143] M. A. Stephens, 'EDF Statistics for Goodness of Fit and Some Comparisons,' *Journal of the American Statistical Association*, vol. 69, no. 347, pp. 730–737, 1974, ISSN: 0162-1459. DOI: 10.2307/2286009. JSTOR: 2286009. (visited on 08/06/2023).

[144] A. N. Pettitt, 'A Two-Sample Anderson–Darling Rank Statistic,' *Biometrika*, vol. 63, no. 1, pp. 161–168, 1976, ISSN: 0006-3444. DOI: 10.2307/2335097. JSTOR: 2335097. (visited on 08/06/2023).

[145] S. H. K. Kang, 'Spaced Repetition Promotes Efficient and Effective Learning: Policy Implications for Instruction,' *Policy Insights from the Behavioral and Brain Sciences*, vol. 3, no. 1, pp. 12–19, Mar. 2016, ISSN: 2372-7322. DOI: 10.1177/2372732215624708. (visited on 30/05/2023).

Appendix A

GPUs used during Experiment 1

GPUs used in Exp.1 for evaluating the RSs

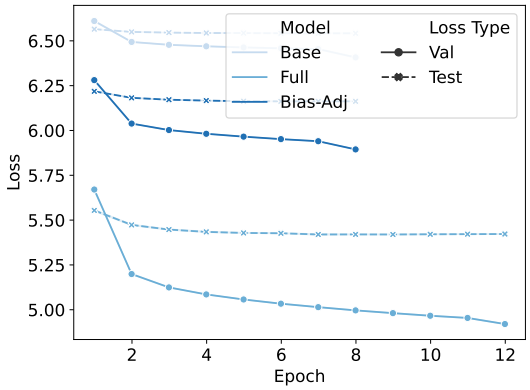
	Model	**EdNet**	**MOOCCubeX**
Baselines	iALS	P100 16GB	P100 16GB
	BPR	P100 16GB	P100 16GB
	GRU	2x P100 16GB	A100 80GB
	BERT	2x P100 16GB	A100 40GB
	XLNet	2x P100 16GB	A100 40GB
Full	GRU	2x P100 16GB	A100 80GB
	BERT	2x P100 16GB	A100 40GB & V100 32GB
	XLNet	2x P100 16GB	A100 40GB
	XLNet$_{feat}$	2x P100 16GB	A100 40GB
Bias adj	GRU	2x P100 16GB	A100 80GB
	BERT	2x P100 16GB	A100 40GB
	XLNet	2x V100 32GB	A100 40GB

Appendix B

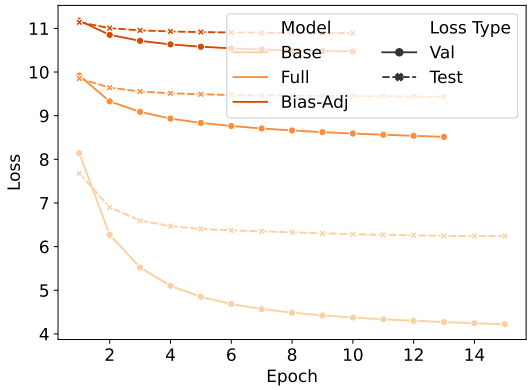
Visualisations

Some visualisations are added in addition to the representative visualisations presented in the main matter. In particular, the additional validation and test losses for GRU and XLNet are presented in Section B.1 as supplementary to Figure 5.6.

B.1 Validation and test loss

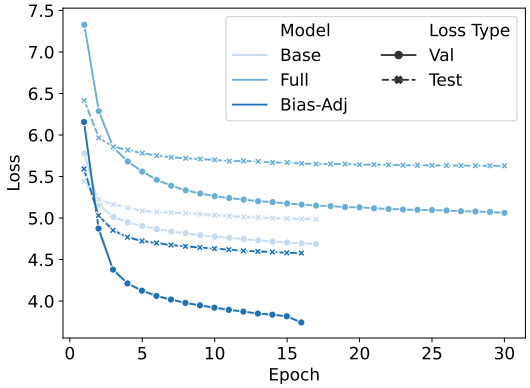


GRU validation and test loss on EdNet

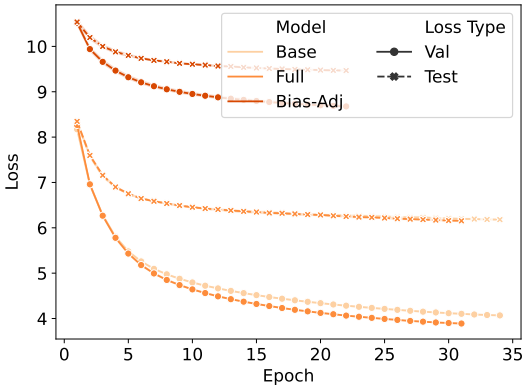


GRU validation and test loss on MOOCCubeX

The mean and 95% confidence interval of the validation and test loss for the variants of GRU explored in Exp.1 in Section 1.3

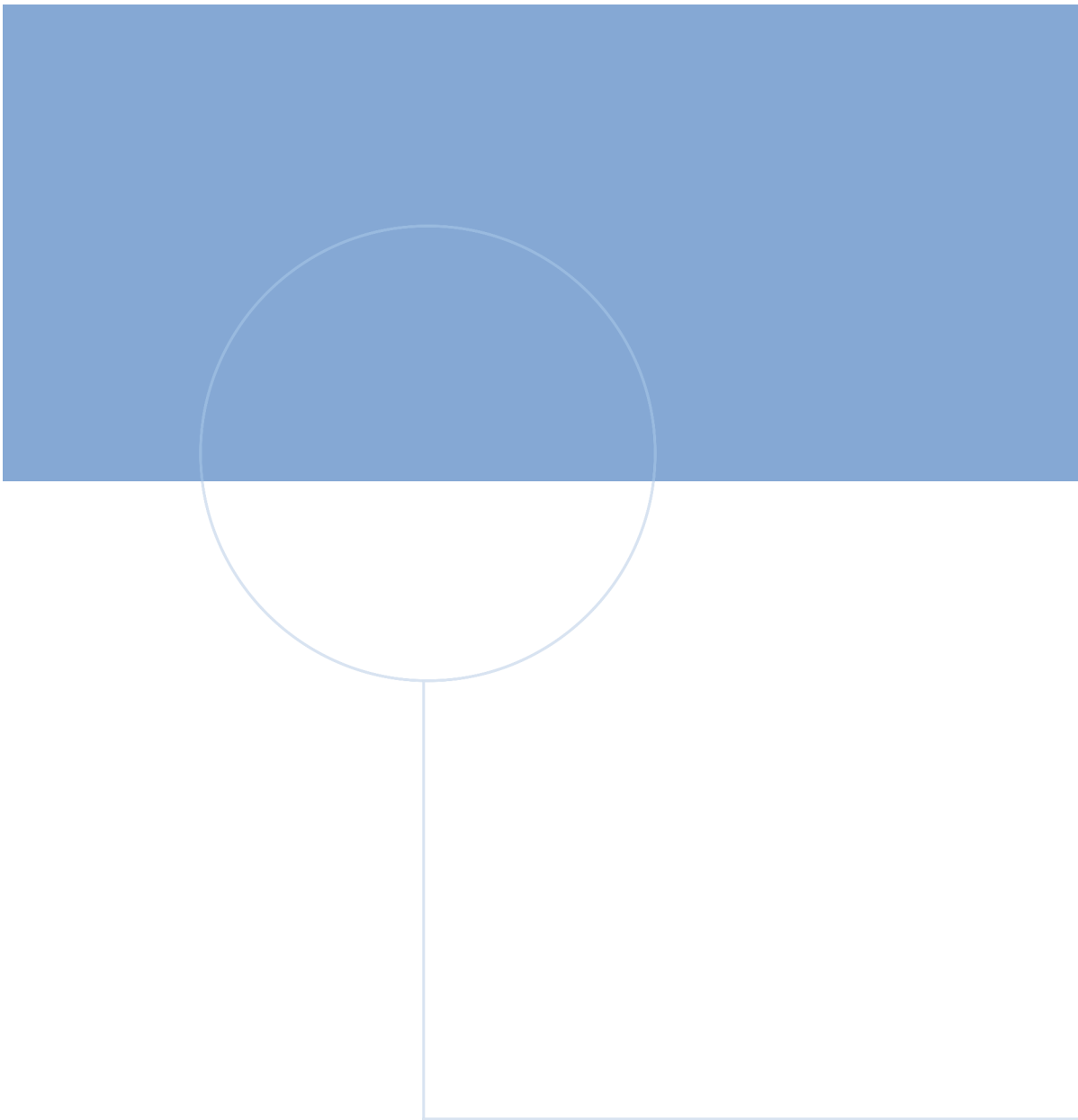


XLNet Validation and test loss on EdNet



XLNet Validation and test loss on MOOCCubeX

The mean and 95% confidence interval of the validation and test loss for the variants of XLNet explored in Exp.1 in Section 1.3



NTNU

Norwegian University of
Science and Technology