Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

Jonas Hyllseth Ryen

Predicting Customer Churn with Data Analytics

How to use data analytics to indicate customer churn for a small online-based company.

Master's thesis in Engineering and ICT
Supervisor: Bjørn Haugen
June 2020

NTNU
Norwegian University of Science and Technology

LEARNLINK

Jonas Hyllseth Ryen

Predicting Customer Churn with Data Analytics

How to use data analytics to indicate customer churn for a small online-based company.

Master's thesis in Engineering and ICT
Supervisor: Bjørn Haugen
June 2020

Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

NTNU
Norwegian University of
Science and Technology

Summary

Effective use of data is increasingly becoming a competitive advantage for businesses. Although a wide range of studies has been conducted to understand the benefits of applying data-driven knowledge, few of them provides practical approaches. While large enterprises have the resources to develop a practical approach themselves, many small businesses struggle to find the time and resources to solve problems from data analytics. This thesis presents a concrete approach to solve problems through data analytics for small online-based companies. The online tutoring agency Learnlink was chosen as the subject. Customer, student and tutor data was analyzed in order to solve one of Learnlink's hardest problems: how to indicate premature cancellations for individual customers, also referred to as customer churn.

The methods that were applied to achieve the main objective can be summarized as *collect, analyze and evaluate*. Relevant data was *collected* from the company's database, payment solution and customer relationship system, assembled and cleaned, and made available for analysis in Google Data Studio visualizations. With a near real-time analytics tool at hand, an *exploratory analysis* was conducted in order to investigate patterns and correlations in the data. Seeming relations were tested for significance before integrated into four different prediction models. By classifying customers based on indications for premature cancellations and comparing the predicted classes with actual churn, the results from the analysis could be *evaluated*.

Throughout the analysis, it became clear that both demographic, behavioural and tutor data could provide indications of whether a customer was probable to prematurely cancel their subscriptions. The analytics tool that was developed throughout this thesis should be sufficient to work as a foundation for solving further business problems with data as well as help mitigating customer churn for the subject company. Furthermore, other small companies should be able to implement the same system in order to solve problems with data. For future work, the results can be improved by introducing more data sources and to experiment with a wider variety of prediction models.

Sammendrag

Flere og flere bedrifter oppnår konkurransefortrinn ved å ta i bruk data på en effektiv måte. Til tross for at det er gjennomført mange studier for å forstå fordelene ved å ta i bruk datadrevet kunnskap, er det få av disse som beskriver praktiske fremgangsmåter for å gjøre dette. Store selskaper har ofte nok ressurser til å utvikle slike praktiske fremgangsmåter på egenhånd, men mange små bedrifter sliter med å sette av nok tid og ressurser til å legge til rette for problemløsing ved hjelp av dataanalyse. Denne oppgaven presenter en konkret fremgangsmåte som kan benyttes av små bedrifter som leverer nettbaserte tjenester til å løse problemer ved hjelp av dataanalyse. Det nettbaserte leksehjelpsselskapet Learnlink ble valgt som utgangspunkt for oppgaven. Data fra kundeforhold, elever og lærere ble analysert for å løse en av de aller vanskeligste utfordringene Learnlink står ovenfor: hvordan de kan få indikasjoner på hvilke kunder som kommer til å avslutte kundeforholdet for tidlig, vanligvis kalt *churn*.

Metodene som ble brukt i denne oppgaven, kan oppsummeres som *samle*, *analysere* og *evaluere*. Relevant data ble *samlet* fra selskapets database, betalingsløsning og kundeoppfølgingssystem, slått sammen og renset, og gjort tilgjengelig for analyse gjennom visualiseringer i Google Data Studio. Med et analyseverktøy som viste data nærmest i sanntid, kunne en *utforskende analyse* gjennomføres for å finne mønstre og korrelasjoner i dataen. Tilsynelatende relasjoner ble testet for signifikans, før de ble integrert i fire ulike prediksjonsmodeller. Ved å klassifisere kundene basert på indikasjonene for tidlig avslutning og sammenligne disse klassene med faktisk churn, kunne resultatene fra den utforskende analysen *evalueres*.

Analysen viste at både demografisk data, oppførsel og data knyttet til læreren kunne indikere om en kunde hadde høy sannsynlighet for å avslutte kundeforholdet for tidlig. Analyseverktøyet som ble utviklet gjennom arbeidet med oppgaven bør være godt nok til å fungere som et grunnlag for datadrevet problemløsing for Learnlink, samt bidra til arbeidet med å redusere *churn*. Andre små selskaper vil også kunne bygge det samme systemet for å løse problemer ved hjelp av data. For videre arbeid kan resultatene forbedres ved å legge til flere datakilder og eksperimentere med et bredere spekter av prediksjonsmodeller.

Preface

This thesis was developed at the Department of Mechanical and Industrial Engineering at the Norwegian University of Science and Technology (NTNU) during the spring of 2020. Some preliminary research was completed during the fall of 2019.
I thank my supervisor, Bjørn Haugen, for providing helpful advice and guidance throughout the writing process, and for his positive inclination to engage in a subject that is unusual for a master thesis in this department. Also, I am grateful to NTNU for providing the resources to complete the thesis. I would also like to thank the employees of Learnlink AS for contributing with background knowledge, help and feedback throughout the development process, and especially Chief Technology Officer Johannes Berggren for facilitating the technical assistance that was necessary to complete the thesis.

Jonas Hyllseth Ryen, author

Table of Contents

List of Figures

Nomenclature

Churn	=	When a customer cancels the customer relationship.
Churn rate	=	Percentage of total customer base lost during a certain time period.
Fast churn	=	A premature cancellation of the customer relationship.
Data analytics	=	Visualisation and interpretation of data.
Prediction	=	Forecasting values based on previous events.
Classification	=	Labeling items based on predicted values.

Chapter 1

Introduction

1.1 Problem description

Effective use of data is increasingly becoming a competitive advantage for businesses [2] [3]. Data provides knowledge and let companies make quick decisions based on facts. Access to frequently updated and detailed data enables companies to make more informed decisions faster than their competitors. Many companies utilize this to build a company-wide culture that is data-driven, where the goal is that all employees are empowered by data in their day-to-day work[4].

Companies providing services through online channels have the opportunity to gather vast amounts of data from their users. This data range from visitor data (non-registered users), data delivered by third-parties like Google or Facebook, usage data from registered users and data gathered through offline communication like phone calls. In the age of big data[5], even small companies generate data that far exceeds the limit that their computing power, available storage space and human resources are ready to handle. In order to make this data useful, companies need to establish tactics for retrieving, collecting, cleaning, analyzing and storing data.

Data is, however, most useful when being applied to solve problems or to cover knowledge gaps. Recent developments have made the power of data analytics available to not only large enterprises with departments dedicated to data processing, but also to small companies with tight budgets. [6] Cloud-based databases updated in real-time, advanced plug-and-play analytics software offered online and pre-trained machine learning algorithms are all available today, starting at a few dollars per month. Today's challenge for a small company is rather limitations of data, shortage of time and integration with business strategy. [7] In order to justify spending time on data analytics, one has to decide: what problems are most important to solve, and how do we proceed to solve them?

One of the most common, yet hard-to-solve problems of consumer-product businesses

is customer churn, or *the premature cancellation of a customer relationship*. Churn is when a customer stops buying a product from the company, like when a Netflix subscription is cancelled or you make a switch from your local grocery store in favour of the new supermarket. The word *retention* is often also used when addressing this problem, and refers to retaining customers - in other words, retention is the opposite of churn. Businesses often rely on recurring payments from the same customers, and keeping customers from leaving or switching to a competitor can be the key to a more sustainable business model. Customer churn has a direct impact on profits, is an indicator of customer satisfaction, and it is often more lucrative to retain customers than acquiring new ones. Consequently, finding ways to understand, predict and mitigate customer churn are important activities for most companies.

During the phase of early growth, companies can sustain high churn rates and still grow fast. However, as the number of customers becomes higher, the newly acquired customers become a decreasingly smaller part of the total. Sales and marketing activities become less important for sustaining the overall revenue than stopping existing customers from leaving. Hence, retaining customers becomes a priority for mature companies; it is profitable to keep existing customers than to lose them and reach new ones through marketing. As lowering churn rates become more important as the size of the company customer base grows, large gains can be achieved by lowering the churn rate when the company is still small and not when the problem escalates.

Churned customers share one common trait across any business; they have all been customers at some point. Which means, they have been interacting as customers and left information about their behaviour, their purchases, their interactions with customer support and often personal information. As customers who churn by definition are the ones a company has collected the most data about, finding out why customers churn is a sensible problem to solve with data analytics.

The purpose of this thesis is to contribute to solving one of the most difficult problems for a small online-based company through data analytics. Through the writing of this thesis, the goal was to not only help with the churn problem, but to also establish a foundation for data-driven problem solving for the subject company. Learnlink, an online tutoring company founded in 2016, was chosen as the subject. We investigated customer and tutor data, and used this to establish a framework for classifying the risk for individual customers to churn. The goal was that by the completion of the thesis, the Learnlink team could use this knowledge to impose measures for high-risk customers before churn, and thus improve customer loyalty, reduce churn rates and build a more sustainable business.

1.2 Objectives

1.2.1 Main objective

Build a data analytics tool for a small company to detect whether a customer relationship is on track or is probable to be prematurely cancelled by the customer.

1.2.2 Secondary objectives

In order to accomplish our main objective, the technical infrastructure must be in place. We will need to be able to navigate data from the database and other data sources easily and to make visualisations to observe patterns and trends. A prerequisite for drawing the right conclusions is that irregularities and errors are removed - data sets need to be *clean*. After cleaning and structuring data from all data sources, we need to determine exactly what to extract from our data: A clear, unambiguous definition for our key term *churn* must be established. After completing this preliminary work, we will be ready to embark on the actual analysis; we will investigate the data through visualizations, looking for patterns and correlations. Relations will be checked for significance. In order to assess the practical value of the knowledge we derive from the analysis, we will build a prediction model based on these assumptions and test the model against actual events. After completing this step, we will know whether our main objective is achieved.

1. Choose data sources and establish connections between the necessary software for collecting and cleaning the relevant data.
2. Establish a robust definition of *churn* and a measure that can indicate whether the churn is premature.
3. Perform an exploratory analysis of the data set with focus on differences between fast-churn and slow-churn customers.
4. Choose variables that best describe the indication for a premature cancellation.
5. Set up prediction models based on the findings in the exploratory analysis.
6. Evaluate predictions by comparing to actual events..

1.3 Limitations

The thesis should be regarded as preliminary work that establishes a method, routines and a framework for gathering data and analyzing customer churn. Results will be limited by the amount of data that is available. The end result will be a set of indicators, not predictors, and the system that will be built will require maintenance and continuous improvement in order to reach the full potential.

This thesis is reliant on access to customer data. The actual gathering of data is performed by the technical team in the company and not a subject of this thesis. Thus, the thesis is constrained by the tools and architectural choices made by the Learnlink team. The most significant constraint is assumed to be the choice of database. The author of this thesis is not entitled to request significant changes to the technical stack for the Learnlink platform. Similarly, the technical team in Learnlink has made some choices regarding the structure of the data flow. Segment.io has been chosen as the platform for routing data to the desired destination, as it is cost-efficient and easy-to-use. Apart from this, most services regarding data are chosen from Google's BI platform in order to make integration smoother and costs low.

There are many constraints on the data that is collected for this thesis. As described in more detail in the Background section, the nature of the business has changed dramatically over the past years. The company has completely changed the business model two times during the analysis period, and has experienced a rapid growth that results in a customer base that has an unusual high ratio of recently acquired customers to the total customer base. The GDPR privacy regulations also impose limitations to our analysis, as some data has to be excluded from the set that might have been interesting, including geographical data. Performing this analysis is both challenging and very interesting due to these constraints, but there is little doubt that modifications to the resulting prediction models will be necessary after the thesis is completed. This thesis and the corresponding data models should be seen as foundational work that can continued by the subject company after completion. Our thesis is also limited by the observation period when it comes to comparing actual events with predictions. The observation period is set to approximately 1,5 months, from mid-April to the end of May 2020.

Given the wide variety of prediction models and algorithms, applying them to predicting churn is worthy of a thesis in itself; prediction modelling from data is a broad and developed field.. Even though the last two objectives of this thesis are concerned with establishing predictions, they should be viewed merely as a form of validation of the findings from the analysis. Data analysis is the main focus in this thesis. The priority is not to build advanced prediction models, neither to evaluate, compare or provide an overview of the prediction models available for solving similar problems. The purpose of prediction in this thesis is not to make advancement in the field but to utilize it for validation. Consequently, our investigation of prediction modelling will be limited.

1.4 Privacy

All data used in this project was anonymised before the analysis. Learnlink AS is General Data Protection Regulation (GDPR) compliant and is anonymising data on requests from customers and after inactivity. Historical data used in this thesis has been stripped of all data points that could be used to directly or indirectly identify individuals. No personal information for individual customers is included in this thesis, and it will not be possible to track any information back to individual students or tutors based on the information collected here.

Chapter 2

Background

2.1 Empirical focus

2.1.1 Gap in literature

A wide range of literature regarding data analytics, and especially big data, has become available over the past years, but there is little research describing how to implement the available solutions in businesses [8] Most research and literature is focused on the underlying concepts and not the implementation and the challenges with using data analytics in practice. Large enterprises struggle to integrate data analytics into the overall strategy [9]. Startups struggle to integrate data analytics into their development, even though they are aware of the opportunities and regard them as valuable. It seems that many companies regard analytics as something that should be postponed and dealt with in the future [7]. It is difficult to find literature that suggests that data is not useful and that it should not be important for small companies. The gap in recommendations and step-by-step methods to use data analytics in practice is worthy of attention.

2.1.2 Subject company and problem

We will now elaborate on the choice of Learnlink AS as the subject. Young companies face a significant challenge when aiming to solve problems through data analytics. They are often in a constant process of change as the company structure is developing, and the management tools have to be developed accordingly. They have tight budgets and limited time. A large and established company is able to use external consultants to set up data infrastructure, while early-stage companies have neither the financial resources nor human resources to make large investments in data infrastructure. They are in need of finding a cost-efficient, flexible and easy way to visualize their data.

As a company that delivers service through online channels, Learnlink is a good fit for a data analytics project. The company has already been gathering and storing data for

years, has basic data infrastructure in place and has come a long way in creating a data-driven culture among employees. The technical team knows the challenges at hand, but like many other early-stage companies, they have strict development deadlines and limited resources. Thus they can assist in the project with experience, advice and knowledge, but are prevented from carrying out the project themselves. Nonetheless, the importance of forecasting their income stream and getting the problem under control ensures that the project will receive the necessary attention and assistance.

The topic of churn rates was partly chosen based on inquires from Learnlink management and partly due to the application a solution might have for other similar companies. As mentioned in the problem statement, the churn rate is a core part of the business model. Other problems that were considered was demand prediction, evaluating tutoring lesson quality and analyzing variations in the effectiveness of sales calls. Below follows comments from Learnlink management on the problem chosen for this thesis.

> *Understanding and predicting churn rates is crucial to our financial planning. A few percentage points increase in the churn rate can be the difference between profitability and bankruptcy in the long run. It determines how much we are able to spend for marketing and the number of hires we can make. Nonetheless, churn rates have shown to be hard to predict, especially with the rapid changes we are making to our service and to the business model. Industry standards do not suffice as we have an offering that is significantly different from our competitors.*
>
> *Decreasing the churn rate is one of the main challenges for the Learnlink team at the moment. Most of the measures we have taken in the past have not had any effect. Decisions on how to mitigate churn are too often based on gut feelings rather than hard facts, and we struggle to separate the effective measures from the insignificant ones.*
>
> *We hope that this thesis will uncover patterns that can be used to detect whether a customer has a high risk for churning and obtain better prerdictability for our company.*
>
> - Product manager, Learnlink AS[10]
>
> *Utilizing the power of data analytics has always been an ambition, but never the highest priority for our team. The development team is constantly entangled in fixing issues and improving business-critical features, and with only two full-time developers we are not able to allocate the time that is required to dig as deep as we want into building data infrastructure. Nevertheless, we have acknowledged that leveraging data will be crucial for us in the future and are currently collecting all the data we can. A few projects in the past have been successful with preliminary work to establish data flows and*

automatically updated dashboards. Our most common key performance indicators are updated continuously in a Google Data Studio dashboard, and we use these for guiding our priorities and making decisions. Another obstacle has been using statistics to separate irrelevant data from real patterns and to single out the most important questions to query. Data science is not really an engineering problem but a business problem, although the technical team has to be involved. I hope that this project will establish a foundation for harvesting knowledge from our data in the future and that we can continue the work on both the churn analysis and other components after completion.

- Chief technology officer, Learnlink AS [10]

2.1.3 Learnlink AS

Learnlink was founded in 2016 by three university students who experienced the tutoring market in Norway as inefficient and underdeveloped. Tutoring was only available close to universities, and companies operating in the market had not utilized the developments in online education tools. By 2019, Learnlink had grown to be the fourth largest tutoring provider in the country and the top-grossing provider of online tutoring services for students in elementary through high school. According to the company's management, Learnlink had 5 MNOK revenue from 11 000 tutoring lessons in 2019, employing 150 university students as tutors.

As of May 2020, Learnlink had six full-time and two part-time employees. Their business is organized through Learnlink.no, a two-sided web platform connecting students with qualified tutors. Students are often represented by parents, who are paying the bills and often administering how often lessons take place. This way, tutoring through Learnlink involves four different stakeholders: the students, who aspire to advance in school, parents, who want their children to advance, tutors who help students and get paid, and the Learnlink team which works as a quality-assuring intermediary.

Tutors

Learnlink tutors are university students with strong academic results and a desire to maintain a steady income from a flexible and relevant part-time job. As the tutoring is done online, there are no geographical limitations, and even though most tutors live in Oslo or Trondheim, there are Spanish tutors living in Barcelona and French tutors in Paris. Aspiring tutors register at Learnlink's website and upload diploma from high school, police records and other relevant documentation before their applications are reviewed. Strong applicants are called into online interviews and ultimately qualified as tutors. All tutors comply with confidential agreements about their student's activities and personal information. Most tutors teach 3-4 students, although all lessons are one-on-one and kept separate. Tutors are matched with students based on their academic and personal profiles, as well as preferences from parents or students. Tutors are free to structure the tutoring as they see fit, but have access to a wide range of resources through the Learnlink web platform.

Students and parents

The students range from children in the lowest levels of elementary school to middle school and up to high school. Some students have previously graduated or failed to graduate from high school and are retaking exams, but most are still in school. The vast majority of students are lagging behind the rest of the class and are taking tutoring lessons to catch up. The goal of tutoring lessons is often to increase motivation and feeling of accomplishment through more personalized learning than the student receives in a classroom with thirty other students. As the tutoring lessons are online, students from all over Norway use Learnlink. In most cases, parents are the ones to initiate tutoring, to manage lesson schedules and communicate with tutors as well as paying the bills. Presumably, they would also be active in the decision about ending the tutoring and churning. Parents receive reports

from the tutor after every lesson and are this way able to follow the progress over time.

Tutoring

Most students have 2-4 lessons per week, divided into 1-2 sessions. The student and the tutor meet at the agreed time in a virtual classroom, where they can write, share screen and see and hear each other. Contents of the sessions vary, often split between going through assignments the students have completed since last time and learning new subjects. After the session, the tutor sends the tutoring report to the student's parents.

2.1.4 Interesting attributes regarding a tutoring company

Uncovering patterns for customers buying educational services can be interesting for public educational institutions as well. Even though the mechanisms for "losing" students are different in the public sector, factors like student motivation, their willingness to engage in learning activities and progress will be interesting to investigate.

Tutoring companies are subject to unusually powerful seasonal variations. Two months of summer holidays with low activity periods in the beginning and the end completely halt the market, while exams creates short periods of very high demand. These effects are likely to emerge in the churn data. Moreover, as Learnlink only offers tutoring for students in elementary through high school, all customers naturally grow out of the service. While other industries can retain customers for decades, all tutoring customers will indisputably stop using the service when they finish school.

2.2 Methods

Methods used for writing this thesis are closely connected, and the last steps rely on results from the previous one. The activities were carried out in the order they are represented here.

2.2.1 Interviews

In order to pick the right problem to solve, attain a better understanding of the chosen problem as well as gaining insight into the underlying structure of the platform, interviews with employees were carried out. The interview with Learnlink's Chief Technology Officer contributed to the description of the technological stack used by the company and the architecture of the Learnlink database, as well as the choices for software to the data pipeline. Other employees participated by describing their challenges with the current access to data and information about attempts to understand and solve problems regarding churn. All interviews were done one-on-one and supplemented by a continuous dialogue during the implementation period.

2.2.2 Practical implementation of data analytics dashboards

The most important method used in this thesis and a prerequisite for carrying out the analysis is the implementation of data analytics dashboards. The implementation involves connecting data sources to online software in order to access and visualize data near real-time. The approach is practical because it is carried out similarly to how the company and other companies would have done on their own. An alternative to this approach would be to export historical data to perform analysis locally. This would be sufficient for covering the analysis and prediction need for this thesis, but would not be useful for Learnlink in the future. Yet another approach would have been to evaluate ways for connecting data sources and perform analysis without actually doing this in practice. Even though this would have provided a more thorough evaluation of different analytics systems, it might fail to uncover some problems that could arise during the actual installation.

2.2.3 Exploratory analysis

The exploratory analysis involved visualizing different parts in order to investigate patterns and trends. After the implementation of the data dashboards, data from the different sources like the database, the payment system and the customer relationship management system could be collected and displayed together. The basis for the exploratory analysis was a list of roughly 40 parameters derived from these data sources. Every parameter was explored in Google Data Studio through pie charts, column charts and bubble charts for interesting patterns or correlations. When patterns were found in the visualizations, they were tested for statistical significance, and significant correlations were included in the prediction models.

Significance

Significance was calculated using a student t-test in the following manner.

Let Z be the part of the dataset that our hypothesis concerns. The average or median for the whole dataset - the assumed actual value over time - is denoted by μ. The number of observations in Z, usually individual customers, is given by n, and $\bar{x}$ is the average or median for the subset that is tested for the significant change. Standard deviation is denoted by σ. The probability that the measured difference in Z is due to change is then given by

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{2.1}$$

The significance level is then given by taking the TDIST function of $1 - t$. The TDIST function is the Excel version of the student t-distribution function. For all calculations in this thesis, we have used the t-distribution with n - 1 degrees of freedom and one tail. Hypothesises with significance levels above 95% will be regarded as significant enough to be used in the prediction model as a standalone parameter. Results with significance levels above 90% will be used in combination with other parameters. When displaying the significance level in the Results section, it will be displayed as an error, which means that a value closer to zero means higher significance.

Causation and correlation

We can not necessarily assume causation when we find significant correlations in the data. When the results from the exploratory analysis are ready, the logic of the hypothesises will be discussed and evaluated. There will always be patterns in a data set, and checking for logical sense will help distinguish mere random patterns from actual relations.

2.2.4 Evaluating the analysis through prediction

Testing churn predictions that are derived from the exploratory analysis against actual events that are yet to happen will show how useful the analysis can be in practice. Even though we find patterns in the data, the patterns can be arbitrary, or for some reasons only relevant to the historical data and not applicable to future events. The usefulness of the data analysis will be drastically lower if it cannot be used to provide indications for churn in the future. In order to be able to test predictions, we have to develop a simple prediction model. Criteria for the model were that it should be based on the patterns discovered through the exploratory analysis and that it should be easy to both understand why some customers are chosen as high churn probability while others don't. This will make it easier to improve the model in the future.

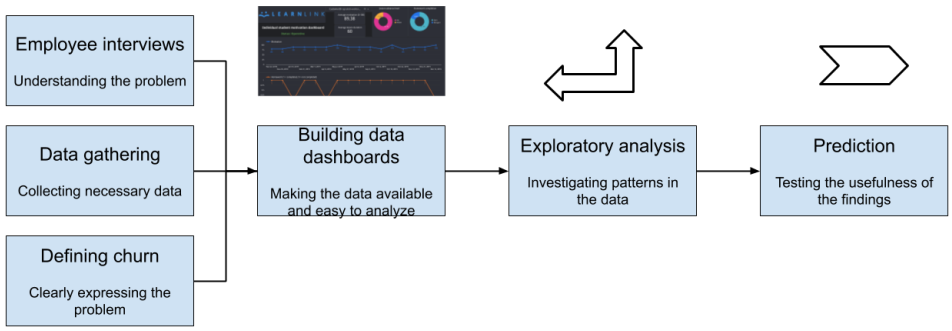


Figure 2.1: An overview of the methods used for writing this thesis.

2.3 The Learnlink platform architecture

This section will provide the necessary background information about Learnlink's technology stack and architecture.

2.3.1 Firestore and document-oriented databases

Learnlink is using Firestore as the database for their web app, and this is where most of the data will be retrieved from.

About document-oriented databases

Firestore is a NoSQL ("Not only SQL") database that is a part of the Firebase development platform, first developed by Firebase Inc. and later acquired by Google. The database is real-time and offered as back-end-as-a-service, as it is entirely hosted in the cloud. Firestore is included in a subcategory of NoSQL databases known as document-oriented databases [11]. Contrary to traditional SQL databases, document-oriented databases like Firestore store data as key-value-pairs rather than strictly defined tables. In traditional relational databases, objects can be divided across different tables. In document-oriented databases, all information about an object is stored at the same place, which removes the need for object-relational mapping when loading the database[12].

The CTO of Learnlink mentions that the fact that the key-value-pairs in Firestore are so similar to Javascript objects and thus can be directly exported was one of the main reasons for choosing this database. As the Learnlink front-end is written in Javascript, this allowed the team to develop the entire platform, including front-end and back-end, in one programming language. It is common to use languages like PHP combined with Javascript for frontend and SQL in the database, which would require the developers to master several programming languages. This also allows for closer integration between the front-end web apps and the database. Firestore is convenient in that it updates the web apps continuously and real-time without the need for refreshing the web page. In summary, the Firestore database gives the Learnlink team access to a series of advanced features without the need for building it themselves[13].

Document-oriented databases and data analytics

SQL
Relational databases are considered to support a wider variety of queries than document-oriented databases. The most common language for data analytics is SQL, which is not supported in a document-oriented database. Consequently, most plug-and-play data analytics or business intelligence software is made for relational databases and do not provide the same support for NoSQL databases[14]. However, NoSQL databases can be preferable in big data projects because they do not have the same strict schema as relational databases and do not have to satisfy the ACID-properties (Atomicity, Consistency, Isolation and Durability)[15] and can be used to store unstructured data [15]. The unstructured data can be preprocessed and structured in the application that will use the data rather than

in the database itself.

Storing in multiple locations
Document-oriented databases support storing the same data in different places. In relational databases, this was considered bad practice, but when it comes to storing and accessing big data efficiently, many paths to the same data source can reduce query time [14].

Flexibility
Relational databases have strict rules for changing the structure of the data in the same table (schema). This can give rise to trouble when working with large amounts of unexplored data, as different rows in a data set may have different structure and attributes. Document-oriented databases have more flexible schemas and are thus handier when working with large, unstructured data sets [14].

Speed
Document-oriented databases provide faster indexing and thus faster query response. Instances have shown more than 100x the query speed compared to relational databases [14].

Scalability
Sharding means partitioning large data sets into smaller and more manageable data parts, referred to as shards. A database that supports sharding can be scaled almost without limits, which is useful when working with large data sets or for small businesses that plan and build for fast growth. Document-oriented databases support sharding, while relational databases do not [14].

2.3.2 Firestore SQL export

At the time of initiation of this thesis, there was no pre-made integration that easily could be set up to export data from Firestore to BigQuery, the tool for structuring the data before queries. To solve this problem, Learnlink's CTO Johannes had made an export function to complete the export, made publicly available on GitHub. As Google did not have a permanent solution that could automatically refresh the data with frequent intervals, the function has been used by other companies around the world, which probably face the same problem as Learnlink.[11].

2.3.3 Learnlink platform architecture

The Learnlink platform architecture is shown in figure 2.2. There are several front-end applications for different uses. App.learnlink.no is used by customers and tutors for administration of their customer relationship; access to payment information, lesson schedule, learning resources and reports from completed tutoring lessons. The admin-platform is used by the Learnlink team when following up customers and tutors. Online.learnlink.no is a virtual classroom where the actual tutoring happens. The virtual classroom is a video chat software with support for drawing and writing as well as recording footage from the tutoring lessons to be viewed later.

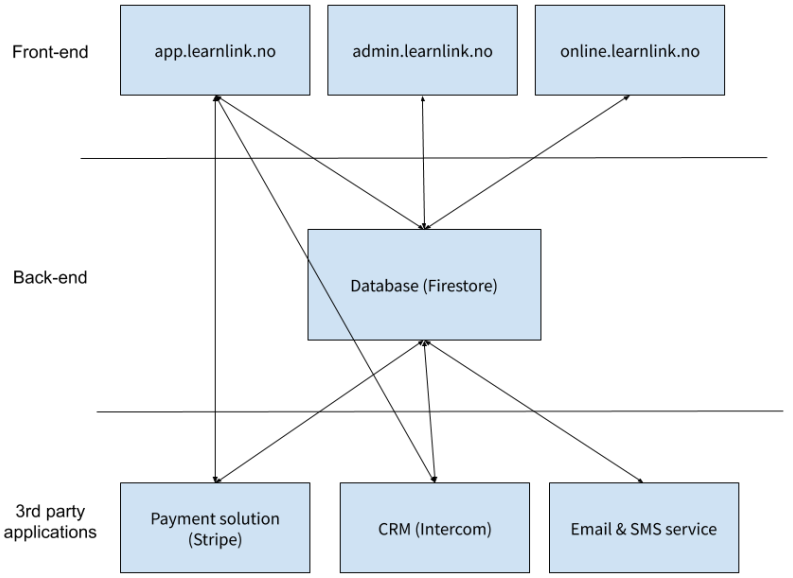


Figure 2.2: Overview showing how parts of the Learnlink software platform are sending data to the other parts.

All of these front-end applications are connected to the database, in addition to some third party applications as shown [13].

Most out-of-the-box analytics tools are event-based: instead of reading data from the database, they receive information about user events. These events are activated in the front-end applications when users push buttons, send messages and so forth. A widely used event-based analytics service is Google Analytics. The challenge with event-based analytics arises when you are dependent on tracking changes that arise from something else than user activity in the applications. As Learnlink is a two-sided platform, every user's data is connected to and updated because of *other users'* actions. So when a tutor registers a lesson for one of their students, the student's profile is updated accordingly, and an email with a summary from the lessons should is sent out. Another challenge surfaces when tutoring is completed on another video conversation tool than Learnlink's own; there is no "evidence" on the platform that the actual tutoring found place, however the completion of the lessons should be reflected on both the tutor and the student's accounts. Our final example is payment, that is handled through third-party payment provider Stripe and is automatically collected at the time of completion for a lesson. Due to the more complicated circumstances around a two-sided platform, only event-based analytics was never an option for giving a complete picture for the Learnlink team. A visualization tool needed to be able to read both event-based analytics and database data and show them in the same charts.

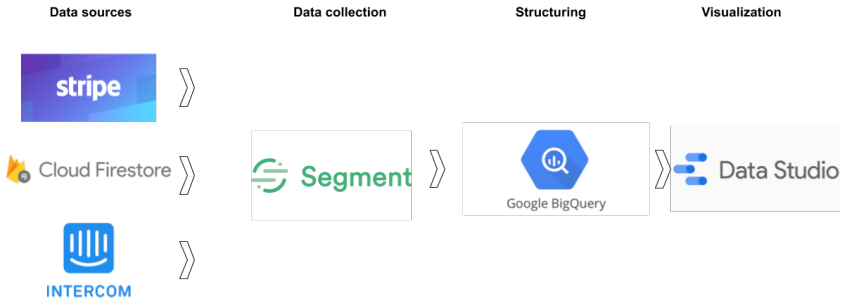


Figure 2.3: The data flow from Learnlink's data sources to data dashboards.

2.3.4 Data flow

Figure 2.3 describes the data flow in the Learnlink application. Customer data is saved in the Firestore database. Information about payments and subscriptions comes from the payment provider Stripe. Information about customer correspondence comes from the Customer Relationship Management (CRM) system Intercom. This data is sent to Segment, which is a data pipeline software - it gathers data from different sources and sends it to the desired destinations. If the Learnlink team would want to attach more data sources to their analytics tool in the future, this can be connected to Segment and sent together with the rest of the data. Segment also handles event-tracking from Learnlink.no: When users perform certain actions, messages that the events are triggered are sent to Segment. Segment data is forwarded to Google BigQuery, where the data is rewritten to SQL. Data in BigQuery is structured in tables. Tables can then be used as data sources in analytics tools like Google Data Studio. Dashboards made in Google Data Studio can then be shared with others [13] and is the tool used to display dashboards and perform the exploratory analysis.

2.4 Data

This section will describe the data in the form it is retrieved in BigQuery, and how it is preprocessed before being sent to Google data studio for the exploratory analysis.

2.4.1 Data collection

Most of the data that will be used in this thesis is collected from data points that are derived from customer, student and tutor activity on the Learnlink web platform. Information about the student is collected at the beginning of the customer relationship in order to find the right tutor. Subject, school level and reason for requesting tutoring are typical data points collected here. Tutor profiles are displayed to customers before their first lesson and is designed to give an impression of the tutor's strengths and preferred learning strategies. Consequently, tutor profiles contain information about the tutor and can be used in the analysis. During the lifespan of the customer relationship, data is gathered through lesson reports. Motivation, homework assignment and completion and difficulty levels are gathered from reports. Payment and subscription information is collected from the payment provider Stripe. Information about customer support activity, as complaints, satisfaction survey responses and number of conversations, is collected through the CRM system Intercom.

2.4.2 Tables

Table	Rows	Variables	Description
Balances	5705	6	Lesson balances
Categories	89	9	List of tutoring subjects (math etc.)
Lessons	16726	24	Tutoring lessons
Projects	3445	45	Customer/tutor relationships
Reports	6015	42	Lesson progress reports
Users	5774	46	Users (tutors and students)
Intercom conversations	64945	18	Customer support conversation activity
Stripe	3364	16	Payment information from Stripe

Table 2.1: Tables from the database, Stripe and Intercom in BigQuery.

Balances

Every customer has a lesson balance that records how many lessons the customer has available for tutoring. Customers who buy more lessons than they complete have positive balances, and customers who complete more lessons than they pay for have negative balances until they pay their debt. This table contains balances for every user and metadata such as the last transaction update and how negative balances are handled. Balances can be of relevance for churn because they indicate whether a customer use more or fewer lessons than they initially planned for. Balances were introduced with the subscription business model in August 2019 and thus there is no balance data prior to this.

Categories

Every student receives tutoring in one or more subjects. Subjects are referred to as categories in the database. They are divided on grade level - for example, ”Math 8th to 10th grade” and ”Math 1st to 7th grade”. At higher levels, there can be several courses within one subject, like different math classes for students attending the first year of high school or various foreign languages. Consequently, there are more *categories* per year in the higher levels than in the lower levels. All subjects also have a level which indicates whether they are elementary school, middle school or high school subjects.

Lessons

The lessons table contains information about all the tutoring lessons. Examples are time, date, duration and whether the lesson has been paid or cancelled.

Projects

Every relationship between a tutor and a student is called a *project*. A tutor usually has several projects, one with each student, and a student might have more projects if they have different tutors in different subjects. The projects table contains information about the number of lessons, tutor wage, relevant milestones like the initiation of the project and pricing information.

Reports

After every tutoring lesson, the tutors fill out a lesson evaluation report. Reports are sent to parents who can review progress, follow student motivation and check whether the student is completing the assigned homework. The most recent reports also include all official subject goals (”læreplanmål”) and how the student’s progress on each of the goals is over time. This feature is however just recently launched and there is not enough data to include subject goals this analysis. Information about motivation and overall goal achievement has been present in evaluation reports since 2017 and is used in the exploratory analysis.

Users

All customers and tutors have user profiles with person-specific information. Data points include personal information, contact info, activity and information that is specific to the user type (whether it is a tutor or a customer). The table is included as demographic information might be relevant when looking a churn rates.

Stripe

The Stripe table contains information about payments and subscriptions. Examples of relevant data points are whether invoices are paid on time, the number of payment reminders and total revenue per customer.

Intercom conversations

The most common way for customers and tutors to get in touch with Learnlink employees is through the Intercom chat or email. All conversations are stored in this table, together with user-specific information such as customer satisfaction and the number of email notifications received.

2.4.3 Excluded data

The Learnlink platform gathers more data than what is included in the tables listed above. The reason for excluding some of the data we have available is to increase the chances of deriving useful conclusion from our dashboards. Examples of data that is excluded is visitor event-data from website landing pages and data from the online tutoring virtual classroom. The website has thousands of visitors per month, clicking buttons and scrolling pages, but we assume this to be of little importance compared to other events during the customer lifetime. Virtual classroom video material would be interesting to analyze, but presents a great challenge due to the immense size of the data. Terabytes of data every day and thousands of lessons every month requires us to apply more sophisticated big data techniques for analysis, which is beyond the scope of this thesis.

2.5 Defining churn

To be able to compare data and to perform queries, we will need to establish a quantitative measure of churn.

Churn was not unambiguously defined when the work with this thesis was initiated. Due to seasonal variations, distinguishing between an inactive customer and a lost one is not straightforward. Moreover, Learnlink has changed its business model twice over the past years. In order to compare data collected during different time periods, we would need to have a definition that is insensitive to changes in the pricing structure. Before proceeding to possible definitions, we will go through the different business models.

Business Models (BMs)

BM1: 2017 - June 2018: Pay-as-you-go
Customers pay a flat fee per lesson and there is no volume discount. There was no commitment to buy a certain number of lessons and the customer could quit at any time without any delay.

BM2: July 2018 - July 2019: Packages
Customers choose a *package* with a certain number of lessons, ranging from 10 to 80 lessons, with volume discounts. Lessons are paid for right after completion and not in advance, and a cancellation fee applies if the customer quits before all lessons in the package are used.

BM3: August 2019 - today: Subscriptions
The current business model is based on subscriptions with recurring payments. Customers choose between 3 different subscriptions with different number of lessons included per month and the subscriptions with more lessons are cheaper. Payments are in advance and recurring. In order to stop the recurring payments, a customer must proactively cancel their subscription by contacting a customer support representative. Subscriptions can be paused for one or two months. Customers can downgrade their subscriptions by changing to one with fewer lessons per month. Widespread downgrading can lead to significant revenue loss for the company.

Irregularities

Some customers take breaks from tutoring and then reengage a few months later. Others make payments but do not complete lessons, yet others pause their subscriptions and complete lessons saved up from previous months, so are having lessons while not making payments. Most customers take breaks during the summer holidays, but the duration of those breaks vary from 1 to 4 months. Some customers take breaks for 1-2 months in December and January.

Criteria for the definition

The data set is already sparse, so our aim should be to find a definition that is compatible with data from all three business models. Summer holidays can not count as churn, as this would make the churn rate every summer 100 per cent and corrupt the data. Optimally, no pauses should count. Downgrades should be partial churns.

The definition we choose should be the one that will be most effective for completing the objectives. As the definition is derived from the underlying data and not the other way around, objective one is unaffected. In the exploratory analysis, we want to compare customers based on how fast they churn, so the definition should be time-sensitive ("*when* did the customer churn" and not "*did* the customer churn"). Asking how probable it is that a customer eventually will churn does not make any sense, as all customers are determined to churn at some point. Instead, finding out *how likely* customers are to churn within a given time-frame or finding the expected time to churn will be more useful.

In other words, we need to translate the following questions to a quantitative language that can be used to perform queries:

"Has this customer churned (or is it active)?" How fast did this customer churn?

Possible churn definitions

Payment stop
Payment one month, then no payment the month after.

Payment stop covers all business models, counts pauses as churn and one-time-orders or prepayment of lessons as immediate churn. The definition can be somewhat inaccurate as customers can keep having lessons even though they are not making payments. The frequency and volume of lessons is not taken into account, so downgrades are not included in this definition.

Subscription cancelled
A customer cancels their subscription.

Subscription cancelled leaves out all customers without a subscription. As BM1 and BM2 did not have explicit cancellations, this definition rules out large parts for the data set. The definition does not count downgrading subscription as churn. Due to the incompatibility with older business models, this definition can be useful in the prediction model, but not in the analysis.

Lesson stop
A customer has not attended any lessons for the previous n months, but did attend lessons during month n-1 and the months prior to this.

Lesson stop is compatible with all business models, as lessons have been the core product across the whole time period. Pauses and holidays do not count as churn as long as lessons are resumed after a break. Downgrading a subscription is not counted as churn.

One might argue that a customer is active as long as lessons are completed, regardless of payments, which supports this definition.

After comparing the three above definitions, we can see that definition *lesson stop* covers most use-cases and is compatible with all business models. Note that downgrades are not counted as churn, so reductions in activity without a complete stop are not covered in our data.

Measuring the time until churn

A measure of how fast the customer has churned will be of importance when determining premature cancellations.

Lifetime
The lifetime of a customer is the time between their first lesson and their last lesson.

Even though lifetime is not an explicit churn definition, it covers the most important aspect; the time period of which the customer is generating revenue for the company. Whether a customer has churned or not is binary, but lifetime is continuous and is more useful for comparisons.

Lifetime is compatible with all the three business models. Pauses are insignificant, and one-time-orders are can be filtered out by only looking at the lifetime span over 1 month. Summer holidays do not count as churn. However, this definition does not take downgrading into consideration.

Lifetime has a neat relationship with churn rate, which makes it possible to use them interchangeably. Using this formula, it is also possible to calculate the average lifetime based on the percentage churn rate of the whole customer body, even though all customers have not churned yet.

Let the average lifetime be L_a, and the churn rate c.
The average lifetime in months will then be

$$L_a = 1 + (1 - c) + (1 - c)^2 + ... = \sum_{n=0}^{\infty} (1 - c)^n \tag{2.2}$$

This is a geometric series with $r = 1 - c$, so using the formula for summing infinite series, we obtain

$$L_a = \frac{a_0}{1 - r} = \frac{1}{1 - (1 - c)} = \frac{1}{c} \tag{2.3}$$

Using the above equation, lifetime can be used as a metric for measuring the speed of the churn rate.

We have now established a precise definition of churn and a measure that can differentiate fast and slow churns.

2.5.1 Adjusting our definition to obtain prediction results

A challenge with the chosen definition of churn, is that it takes one whole month to know whether a customer has churned or not. Additionally, as the definition carries a special condition during summer holidays, it takes even longer time to determine whether a customer has churned in May and June. In order to obtain results faster for evaluating the prediction model, we can choose another definition for the prediction model. A definition that is not applicable in the past because of the adjusted business model is *subscription cancellation*. At the time of writing, however, all customers who wish to stop having tutoring lessons must cancel their subscription. Subscription cancellation is recorded immediately. In order to be able to obtain the best comparison possible for prediction results versus actual events, we will use the *subscription cancelled* definition for evaluating the prediction model.

2.6 Prediction model

Objective five and six deal with validating the results from the exploratory analysis by forming simple predictions that are tested against actual events. This thesis is written during the spring of 2020, and the analysis is carried out in March and April. The aim is to form predictions that can provide indications of which customers are more likely to churn during May 2020.

The purpose of the prediction model is to provide indications of whether a customer relationship is "on-track" or whether there is a high risk of a premature cancellation. A foundation for the prediction model will be established from the correlations in the exploratory analysis. In order for the model to be practical for the Learnlink team, the model should identify *individual* churn risk as opposed to an aggregated risk. A model that will predict a correct aggregated churn percentage every month would be useful for financial planning purposes, but will not enable the team to solve the problem by imposing measures affecting individual customers.

Many employ artificial intelligence and machine learning to improve predictions over time. As we are facing a large number of parameters that can be relevant to the lifetime of a customer, our model must be multivariate. Collins (2014)[16] conducted a review of reports from multivariate prediction models and concluded that the majority of the models lacked external validation, meaning that the models were not tested on other data than the data that was the foundation of the model. We will work around this challenge by testing on data from events that are yet to occur when the model is built.

One representation of the major classes of prediction algorithms group them as follows[17]:

1. Decision trees: A tree-structured algorithm that groups data based on a series of functions.
2. Neural networks: Acyclic networks inspired by the human brain. This is a form of unsupervised prediction, which means that labelling data before the prediction is unnecessary.
3. Instance-based learning: The whole training set is used to build a function that classifies data.

It should be noted that this is no exhaustive list of models or algorithms. We will further only discuss the prediction algorithms that will be used in this thesis.

2.6.1 Prediction model criteria

Accuracy

The accuracy of our model is the number of correctly predicted values compared to the total number of values predicted. Abbott (2014) [18] proposes Percent Correct Classification (PCC) as the main metric to assess accuracy. With PCC, all errors are handled equally and the score is determined based on *whether errors exist* rather than *how they occur*.

$$PCC = \frac{Correct\,predictions}{Number\,of\,predicted\,values} \tag{2.4}$$

Precision and recall

When predicting churn, errors where the model fails to identify potential high churn risk customers (false negative) has potentially more severe consequences than an error where low-risk customers are identified as high risk (false positive). A false negative might result in losing a customer, while a false positive may result in unnecessary measures imposed to stall a churn. We assume that anti-churn measures are somewhat effective and less expensive than losing a customer, and will choose false positives rather than false negatives. We measure precision as the ratio of correct positive predictions to the total positive predictions [19]. In other words,

$$P = \frac{True\ positives}{True\ positives + false\ positives} \tag{2.5}$$

Recall, also called sensitivity, is a measure on how many of the true positives we have predicted.

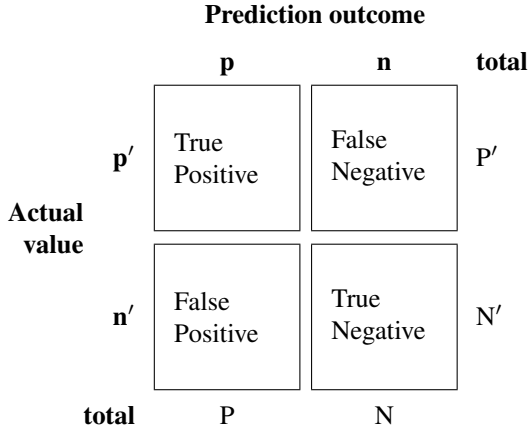
$$R = \frac{True\ positives}{True\ positives + false\ negatives} \tag{2.6}$$

Simplicity

As stated in the Introduction section, we aim to establish a model that can be matured by the Learnlink team in the future and used when more data becomes available. A model that is easy to implement and can be frequently adjusted will be easier to maintain and improve. Third-party algorithms that have already been trained and shown to work are preferable to writing code from scratch for the algorithm.

Overall performance

A supplementary measure to the performance metrics measured above is the confusion matrix, providing an overview of *what kind* of errors the model makes [19].



2.6.2 Classification

Classification is the process of predicting values and assigning labels or classes based on these values. Regression is often used when predicting numerical values, while classification is used when predicting categorical data. We often separate between monoclass classification models using solely yes/no-labels, and multiclass models which have more than two classes.

When predicting churn, monoclass can be used as the answer to the question ”Will this customer churn?”. Multiclass algorithms can differentiate between ”low risk”, ”medium risk” and ”high risk”.

Contrary to another technique that is often used, clustering, classification groups data based on predefined labels. Classification is a type of supervised learning, which means that a training data set with both input and output values are used, while clustering only uses input values. For the purpose of this thesis, we will use classification, as we have predefined labels we want to use.

Decision tree algorithms

Decision trees are a form of supervised machine learning algorithms which are both used for regression and classification. For classification, the algorithms use simple if/then-rules to classify samples. The name is derived from a way of visualizing the algorithm as a tree, where the first decision is made in the root, and the algorithm traverses down the branches to reach a decision in one of the leaf nodes.

Decision trees can be constructed manually based on prior knowledge, without machine learning. Advantages to this approach is that you can leverage prior knowledge and gain insights into how changes in parameters of the model affect outcomes and thus understand more of the underlying effects. Machine learning algorithms can lack transparency and it can be hard to understand how the algorithm reaches certain conclusions. The disadvantage is that the model will not learn by itself and has to be updated when new patterns are found.

Logistic regression algorithm

Logistic regression is a supervised classification algorithm. The algorithm is usually based on Machine Learning and uses the sigmoid function instead of a linear function, which is useful for dealing with outliers in the data set: a linear regression model will give too much weight on extreme values. Logistic regression can be used for binary classification or for multilinear functions. [20] Many regression algorithms handle non-numeric input values.

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{2.7}$$

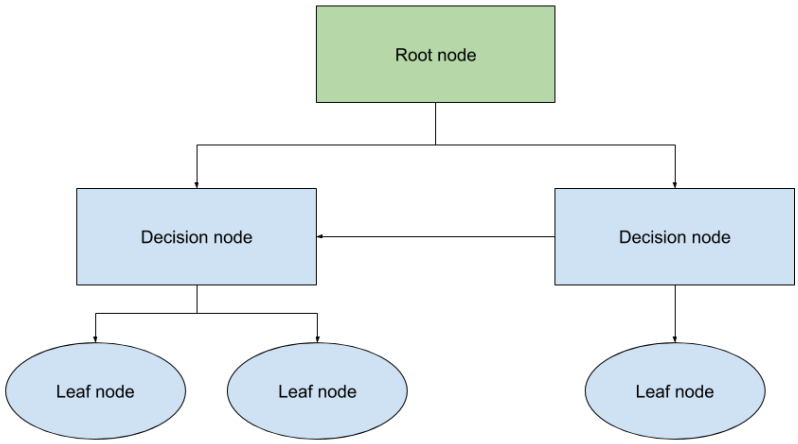


Figure 2.4: A standard decision tree.

2.6.3 Regression models

Based on the above consideration of prediction techniques, we proceed with two different prediction models and use both for evaluating the classification results. First, we will utilize correlations from the exploratory analysis to construct decision trees manually. Secondly, we will feed the variables that have shown to have significant correlation with the churn rate to a pre-built logistic regression algorithm in BigQuery. Both of the resulting algorithms are fairly simple and will be easy for the Learnlink team to develop further.

Classification model I - Manual decision tree

After retrieving a set of variables that correlate with customer churn / lifetime, we will write an algorithm based on if/then-statements in BigQuery. The model will apply labels based on the answers to these statements and apply the same labels that are used in model I. As the model is built based on the exploratory analysis, more details and figures are presented in the Results chapter.

Classification model II - Logistic regression in BigQuery

The first classification model is made using logistic regression in BigQuery. BigQuery's premade machine learning algorithms are developed and trained by Google and are open to use in the BigQuery ML library. The ML library let the user build models with standard

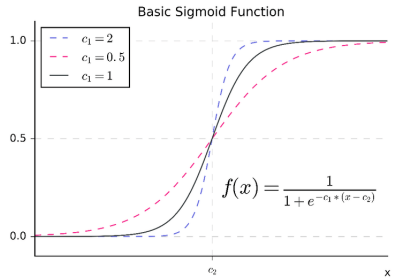


Figure 2.5: The characteristic S-curve of the sigmoid function, used in logistic regression.[1]

SQL and perform queries on data that is already queried in BigQuery, which is useful for us as we already have exported all our data to BigQuery for the exploratory analysis in Google Data Studio.

The model was made after the exploratory analysis, and the parameters included were the ones that were found to have significant correlation with customer lifetime. The queries used for the classification model are included in the Appendix. Customers were grouped into three categories: Customers with less than 60 days lifetime were labelled ”fast churn”, customers with less than 180 days lifetime were labelled ”Medium value” and the rest (more than 180 days lifetime) were labelled ”High value”. The data set, constituted of all customers registered after December 31st 2017, was then divided into a training and evaluation set and a prediction set. All customers who already have churned, meaning they do not have an active subscription, were put in the training and evaluation set, while all active customers were put in the prediction set.

After the model is trained, the performance of the model is evaluated using the evaluation set. When the evaluation is complete, performance metrics are calculated. After evaluation, we can proceed to run the prediction command which predicts actual events for May 2020.

Chapter 3

Results

3.1 Data analytics dashboards for customer churn and lifetime

The development of the dashboards to achieve objective one was structured in two phases: *research and planning* and *implementation*.

3.1.1 Phase 1 - Research and planning

Employee interviews were conducted in order to understand the feature requirements from the dashboards. The following factors were important to the Learnlink employees:

- **Reliability**:As the dashboards are intended to be used for planning and decision making, it is important that the numbers presented are trustworthy, and available when needed.
- **Data freshness**: In order to be able to use the dashboards for decision making, the employee needs to know that the data presented reflects the current situation.
- **Accuracy**: Errors and deviations in the dashboards because of irregularities or dirty data may occur. Some deviation is tolerable, but a large offset provides a false foundation for decisions and could be harmful to the company.

After requirements for the dashboards were established, the technical team was consulted to provide insight into the technical platform and to provide access to the data sources.

3.1.2 Phase 2 - Implementation

An early version of every dashboard was implemented in Google data studio and made accessible for the employees. By comparing the data displayed to actual experienced events and other systems, they could locate errors and inconsistencies. The first version

of the dashboards had numerous errors, but enabled employees to give feedback that was important for further development.

Building BigQuery tables

The first part of the implementation job consisted of exporting data from the relevant data sources to BigQuery. As both BigQuery and Firestore are cloud-based Google products, exporting data from one to the next was of minor complexity. The data from the database was structured in tables similar to the tables in the database, but data from several places in the database was put together in the same table in BigQuery. In order to increase loading speed and to minimize complexity, only the data that was needed for one specific dashboard was extracted. As opposed to gathering all the data in one large table, this approach led to several data tables with less data and a more specific purpose for each of them.

The following tables were constructed from the original data sources:

- **Churn**: Data derived from the projects, user, lesson, balances, account, categories and report tables, aggregated on a per user basis. The table was used to look at customer-specific variables like demographic attributes.
- **Tutors**: Data derived from tutor profiles, like university background and experience for tutors.
- **Lessons and reports**: Data from lessons and reports that can show changes over time as the customer relationship develops.

Each of the tables were assigned to one data dashboard, each giving an overview a specific area of the customer relationship.

Data pre-processing

In order to minimize errors and anomalies, the tables had to be cleaned before the analysis. The following procedures were carried out.

Removing null values

To look at how the lifetime varies with respect to one parameter, samples where this specific parameter is not collected should be excluded from the data set. As some of the demographic data originates from a collection method that was put into practice during the summer of 2019, customers that were acquired prior to this point lack some of the parameters. This includes decision-maker, gender, parent occupation, acquisition channel and price sensitivity. Consequently, the lifetime of these customers can not be longer than from the time the change was implemented (July 2019) to the day of the analysis, which gives an upper bound of around 6 months with the time of analysis in April. In order to not draw the erroneous conclusion that the median lifetime of a customer is higher if these values are null (since more customers from an earlier time period have these values as “null”), every customer with the parameter in question is removed from the correlation analysis. Hence, the median/average that is used for comparison varies because the underlying data set is different. Furthermore, some customers are not willing to provide certain

information, or the information was not relevant and hence not collected. These customers were also excluded from the data set when perform analysis for the attribute in question.

Excluding recent customers

New customers will necessarily have shorter lifetimes than customers who have had customer relationships for several years. However, it was established even before the initiation of this thesis that lifetimes in the early days of the company, on average, were shorter - churn rates have been improving steadily as the company has matured. This could have originated from the difference in the business model, where customers had to make fewer commitments in order to start tutoring, and a less streamlined customer experience. The team pointed to customer interviews and direct feedback as proof for a higher customer satisfaction at the point of this analysis than during the first years. Choosing to rely on these assumptions, data from customers who signed up before July 2018 and after December 31st 2019 was excluded from the analysis. This led to an upper bound of 17 months on the customer lifetime.

Taking growth into consideration

Learnlink experienced significant growth in the number of customers and number of lessons during the time period that was analyzed. High growth introduces a challenge for the analysis because the data set is tail-heavy; with more samples from customers that signed up for the last few months prior to the analysis than the same number of previous months. The increase in revenue has been around 2,5 x per year, so we can assume that the activity increase was similar. An argument against making changes based on this behavior is that the most recent data is of most relevance today. Furthermore, the tail-heavy data set will adjust the median and average lifetime for the whole customer base, but will not affect whether a variable tend to correlate with lower or higher lifetime compared to the median. We have not made any adjustments to the model due to this, but it should be noted as a potential source of error.

After completing the implementation, our end result was three interactive data dashboards in Google Data Studio, displaying data from a wide range of attributes. Variables unaffected by time, like demographic attributes, were displayed using pie charts or column charts. Time-dependent variables were displayed as time series or column charts. Correlations between attributes were discovered through bubble charts. All charts displayed the actual value for every measurement, the median, the standard deviation and the sample size.

Churn dashboard | Lessons and reports dashboard | Tutor dashboard

Figure 3.1: A sample page from each of the three dashboards used in the exploratory analysis.

3.2 Exploratory analysis

In order to efficiently display hypotheses and corresponding results in subsequent tables, we will be using the following terminology. If a factor is expected to have a negative effect on customer lifetime, it is denoted *N* for *negative*. So if the hypothesis from a chart is that factor X indicates a shorter customer lifetime, the hypothesis is described with *X - N (negative)*. Correspondingly, a factor Y that is assumed to have a positive indication on customer lifetime is denoted with *Y - P (positive)*.

In the tables showing significance, only the most relevant results have been included: results showing either significant negative or positive difference from the median. This will enable us to focus on the patterns that have been uncovered and not get lost in the high number of possible combinations of parameters. Necessarily, more analysis and calculations than the ones displayed here have been carried out, but are for practical purposes not listed.

3.2.1 Natural churn

Some customers will churn naturally as they have completed the purpose of their tutoring. This is assumed to be true for all students retaking exams and students who are attending the last year in high school. Showing the number of students in these groups provides an impression of the amount of churn that the Learnlink team will be unable to mitigate.

Natural churn indicators	
Students born before 2002	42
Students retaking exams	17
Students in high school subjects	41

Table 3.1: Table of students who will churn ”naturally”.

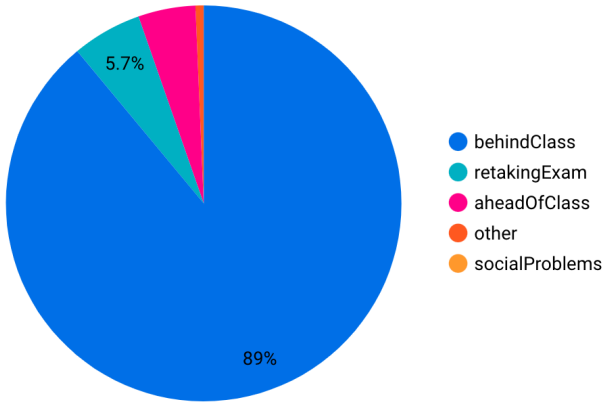


Figure 3.2: Students grouped by problem. In this analysis, only *retakingExam* is relevant, showing the ratio of students who have already finished school and is retaking exams to improve grades.

3.2.2 Trends by season

Since the tutoring market has high seasonal variations, it is relevant to take a closer look at how lifetime varies dependent on the time of the year for the first and the last lesson. By looking at what time the last lesson occurred, we can be able to uncover what months *most* customers cancel and when the most valuable customers disappear.

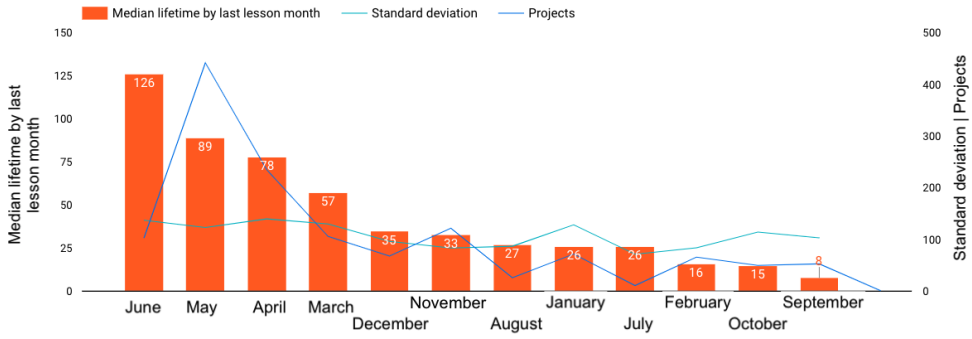


Figure 3.3: Median lifetime, in days, shown by which month the last lesson occurred.

Hypothesis	μ	n	$\bar{x}$	σ	Significance
June - positive	63	92	126	143	0,0000
April - positive	63	373	78	135	0,0163
May - positive	63	237	89	105	0,0001
March - positive	63	138	57	135	0,3012
September - negative	63	53	8	103	0,0001
October - negative	63	50	15	114	0,0023
February - negative	63	66	16	84	0,0000
July - negative	63	11	26	72	0,0596

Table 3.2: Registration month

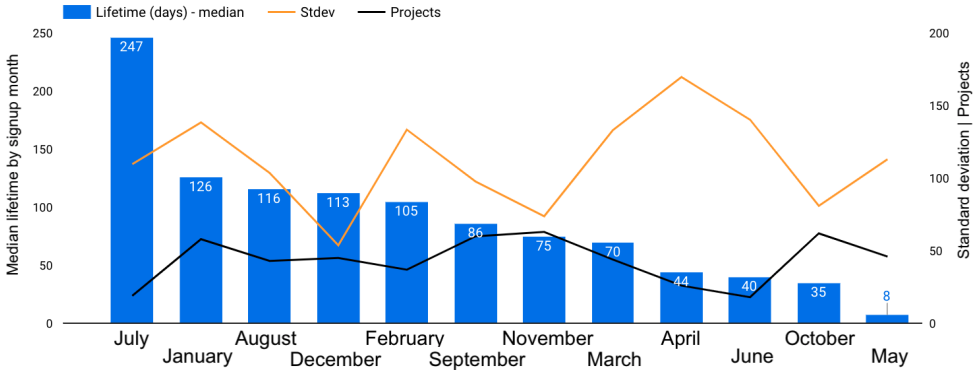


Figure 3.4: Median lifetime, in lessons, shown by which month the first lesson occurred.

Hypothesis	μ	n	$\bar{x}$	σ	Significance
July - positive	87	19	247	110	0,0000
January - positive	87	58	126	139	0,0185
August - positive	87	43	116	104	0,0371
February - positive	87	37	105	54	0,0250
December - positive	87	45	113	54	0,0012
May - negative	87	46	8	110	0,0000
October - negative	87	62	35	73	0,0000
June - negative	87	18	40	130	0,0724
April - negative	87	26	44	162	0,0946

Table 3.3: Last lesson month effects on lifetime.

3.2.3 Trends by demographics and subjects

The following figures and tables show how demographics indicate customer lifetime. For the school subject analysis, subjects are compared across each level because of the variations in lifetime between different levels.

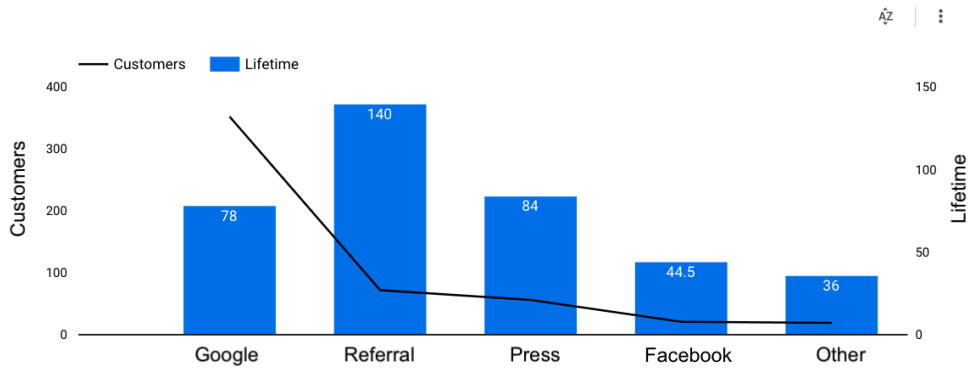


Figure 3.5: Acquisition channel as indicator of lifetime.

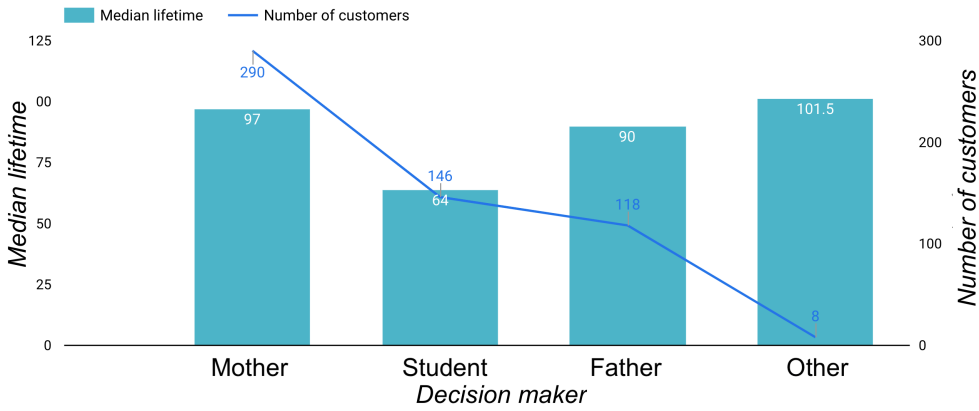


Figure 3.6: Decision maker as indicator of lifetime.

Hypothesis	μ	n	$\bar{x}$	σ	Significance
Referral - positive	84	72	140	192	0,0079
Facebook - negative	84	21	45	52	0,0013
Other - negative	84	19	36	25	0,0000
Student - negative	84	146	64	76	0,0009

Table 3.4: Acquisition channel and decision maker lifetime indicator.

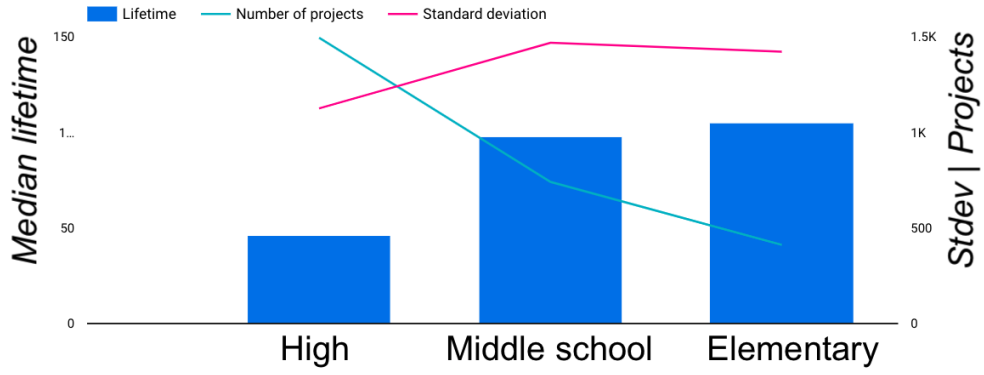


Figure 3.7: Lifetime shown by level.

Hypothesis	μ	n	$\bar{x}$	σ	Significance
High school - negative	72	1495	46	113	0,0000
Middle school - positive	72	741	93	143	0,0001
Elementary school - positive	72	412	98	139	0,0000

Table 3.5: School level and lifetime.

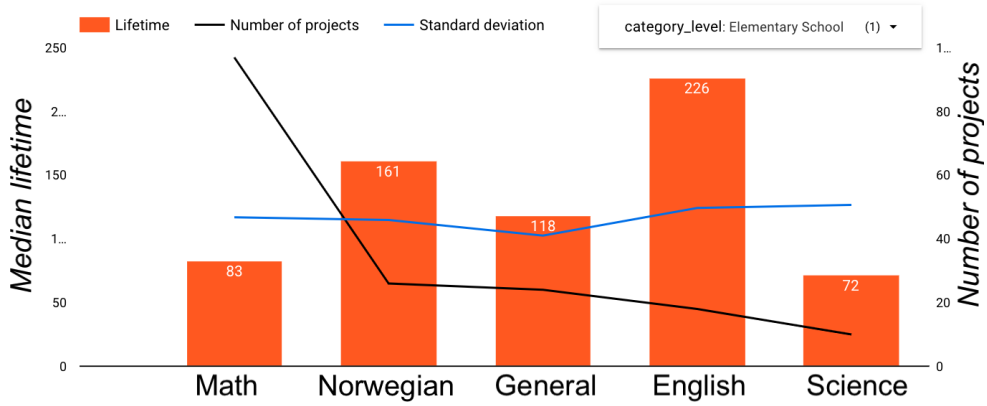


Figure 3.8: Lifetime by subject, for elementary school students.

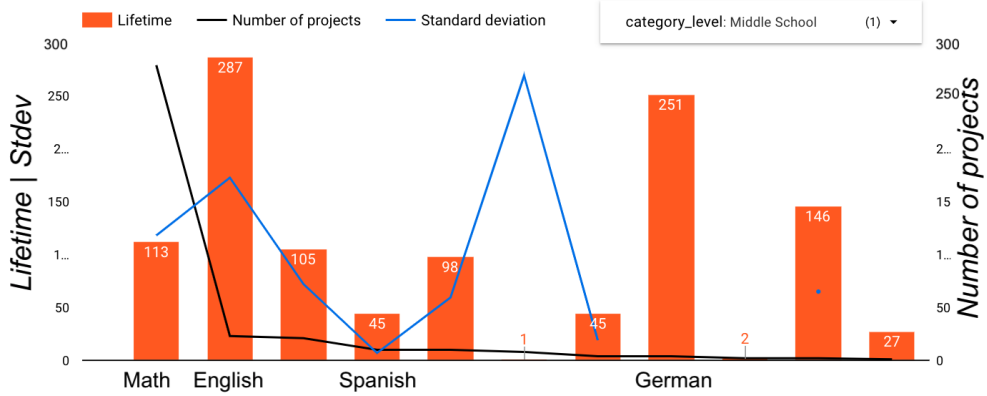


Figure 3.9: Lifetime by subject, for middle school students. Subjects with little relevance are not labeled.

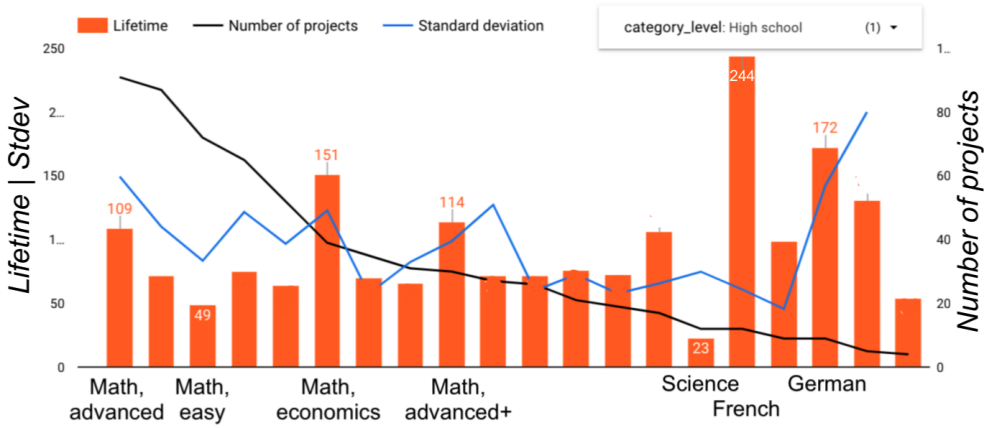


Figure 3.10: Lifetime by subject, for high school students. Label and value is only displayed for subjects with significant results.

Hypothesis	μ	n	$\bar{x}$	σ	Significance
HS: Science - N	76	12	23	75	0,0162
HS: Math, easy - N	76	72	49	84	0,0040
HS: Math, advanced - P	76	91	109	142	0,0146
HS: Math, economics - P	76	39	151	1124	0,0003
HS: Math, advanced+ - P	76	30	114	101	0,0242
HS: German - P	76	9	179	145	0,0328
HS: French - P	76	12	244	61	0,0000
MS: English - P	105	23	287	164	0,0000
ES: English - P	102	18	226	123	0,0003

Table 3.6: Subject as indicator on lifetime. HS = high school, MS = middle school, ES = elementary school.

3.2.4 Trends throughout the customer lifespan

It is likely that events that occur throughout the customer relationship can affect how fast the customers churn. While some customers have clear, time-specific goals for their tutoring, others use tutoring to retain a certain level of motivation or to prevent falling behind the rest of the class. For students who lack motivation, a change in the subjectively perceived value of the tutoring service can possibly end in a churn.

Motivation

Motivation is reported by tutors after every lesson, and is rated on a scale from 0 to 100 where 100 is the highest motivation possible.

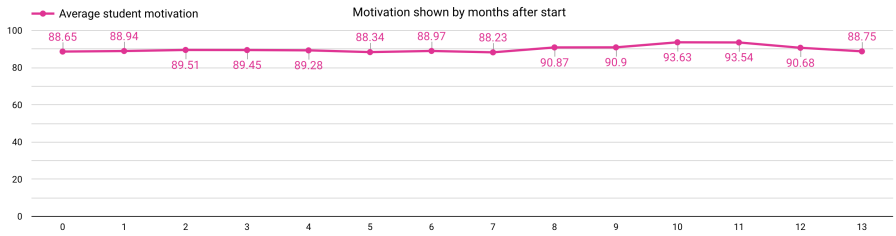


Figure 3.11: Motivation shown by number of months after first lesson.

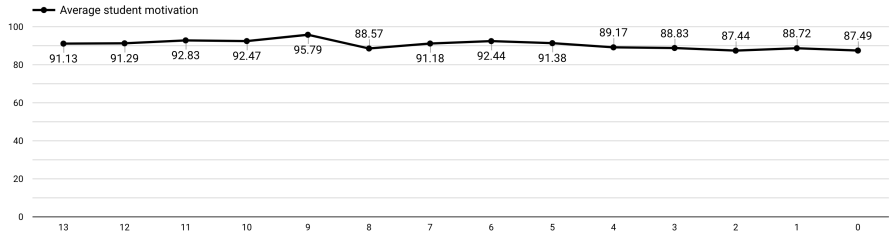


Figure 3.12: Motivation shown by number of months before last lesson (churn).

Hypothesis	μ	n	$\bar{x}$	σ	Significance
Decline 1 mth before - N	89,67	611	88,79	15,77	0,0000
Decline 2 mths before - N	89,67	299	88,06	16,96	0,0509
Decline churn month - N	89,67	531	87,63	16,31	0,0028

Table 3.7: Motivation as indication of churn in relation to the churn month.

We are further looking at whether students with overall lower motivation have shorter or longer lifetimes than students with higher motivation. In the following table, we are

grouping students based on their average motivation over time. Lifetime in this table is denoted by months, not days. The groups are constructed by rounding the median lifetime to the closest ten. In other words, 80 shows all students with median motivation between 75 and 84.

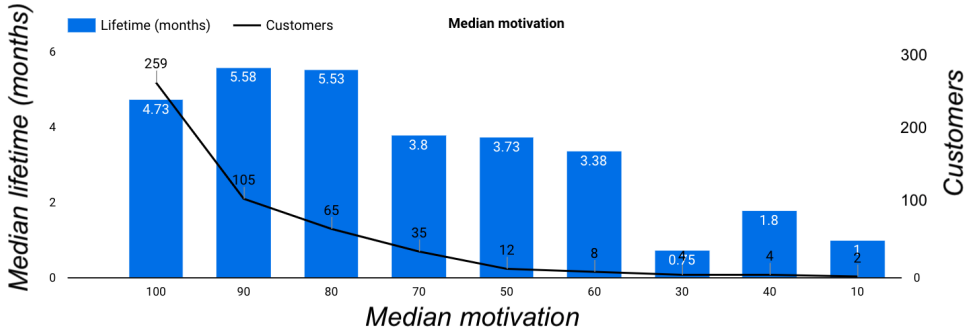


Figure 3.13: Lifetime by motivation, grouped.

Hypothesis	μ	n	$\bar{x}$	σ	Significance
Average motivation $<$ 75 - N	4,97	65	3,47	3,47	0,0004
Average motivation 75-84 - P	4,97	76	5,75	5,15	0,0954
Average motivation $>$ 85 - P	4,97	101	5,49	5,55	0,1743

Table 3.8: Motivation as an indicator of lifetime.

Continuity

Continuity, or the frequency of lessons, was not found to have any significant impact on lifetime for the months prior to churn. It is a surprising finding and is hence included here.

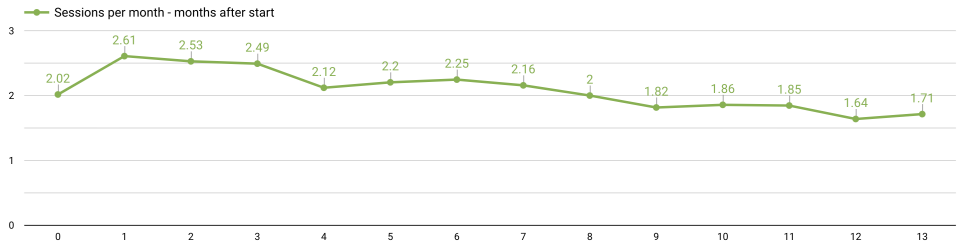


Figure 3.14: Average number of sessions shown by number of months after first lesson.

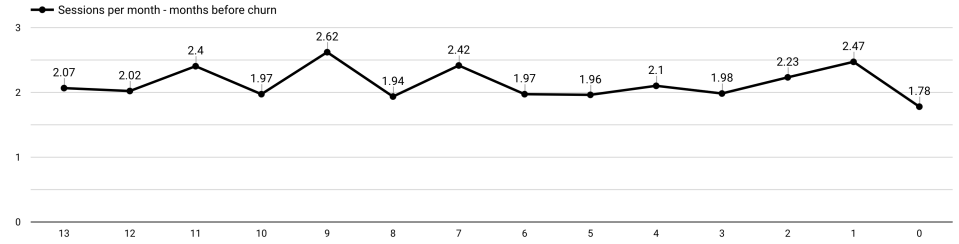


Figure 3.15: Average number of sessions shown by number of months before last lesson (churn).

Homework completion

No significant change in homework completion was found during the months prior to churn. One might assume that not completing the assigned homework would be negative, but we were not able to find any significant negative impact from this. However, a surprising finding was that it seems to be negative for a student to not be *assigned* any homework.

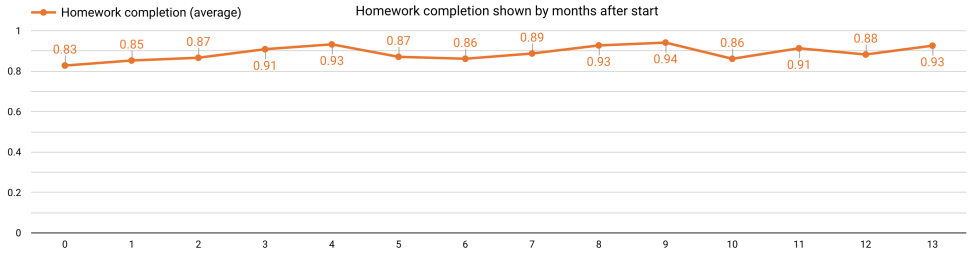


Figure 3.16: Average homework completion shown by number of months after first lesson, where 1 indicates that all homework has been completed while 0 indicates that the homework was unfinished.

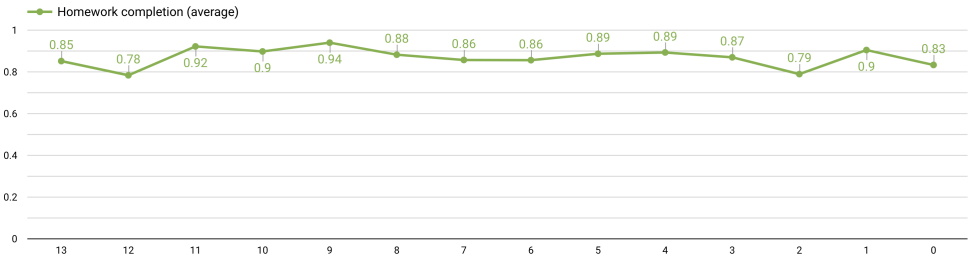


Figure 3.17: Average homework completion shown by number of months before last lesson (churn), where 1 indicates that all homework has been completed while 0 indicates that the homework was unfinished.

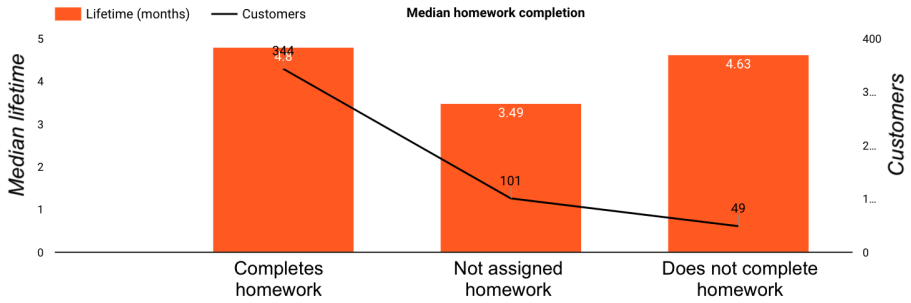


Figure 3.18: Lifetime in relation to homework assignment and completion.

Hypothesis	μ	n	$\bar{x}$	σ	Significance
On average not completed - N	4,81	44	4,38	4,55	0,2670
Student is not assigned homework - N	4,81	98	2,97	3,21	0,0000

Table 3.9: Median homework completion/assignment as an indicator of lifetime.

Lesson difficulty

Students at a lower difficulty level seem to have lower probability of a premature cancellation. Simultaneously, we found negative impact from a *reduction* in difficulty level.

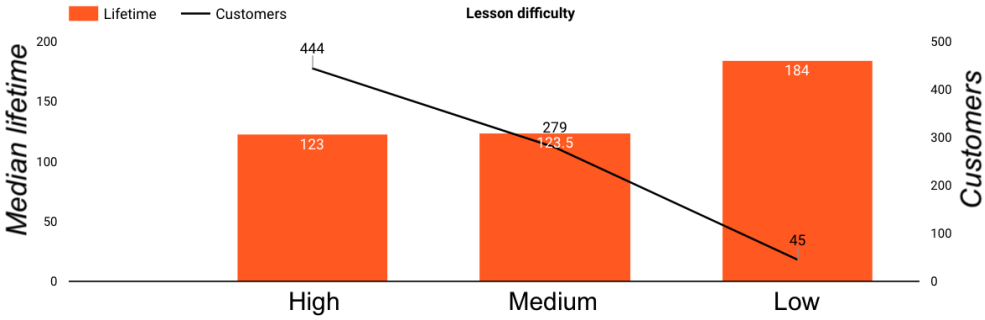


Figure 3.19: Lifetime based on median difficulty level in lesson.

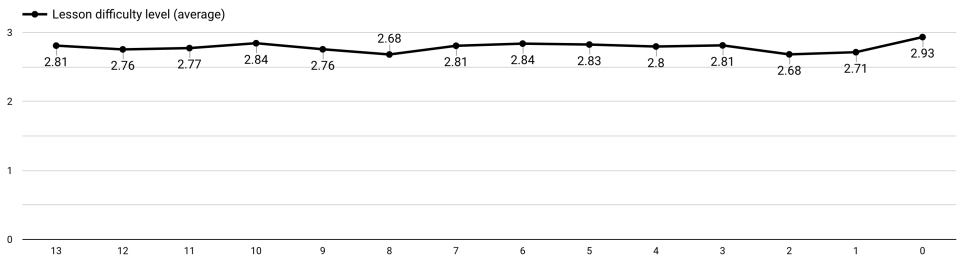


Figure 3.20: Average difficulty level shown by number of months after first lesson.

Hypothesis	μ	n	$\bar{x}$	σ	Significance
Low lesson difficulty - P	141	45	184	154	0,0399
Difficulty drop 2 mths before - N	2,79	68	2,66	0,61	0,0417

Table 3.10: Lesson difficulty level and lifetime.

3.2.5 Tutor attributes effect on lifetime

It is reasonable to assume that different tutors have unequal performance for tutoring, and that a tutor's performance affects student and parent satisfaction and thus the duration of the customer relationship. We found that tutors need experience from surprisingly many students before it can have a positive effect on lifetime, and that tutors who seem to write shorter texts on their tutor profile tend to have students with shorter lifetimes.

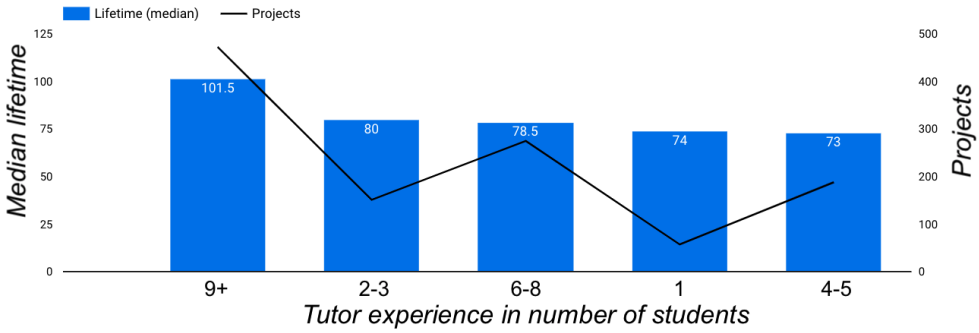


Figure 3.21: Tutor experience (in number of students) and indication on lifetime.

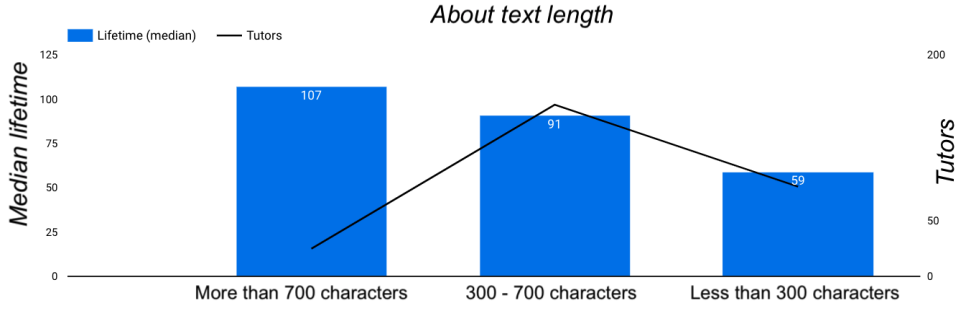


Figure 3.22: Tutor experience (in number of students) and indication on lifetime.

Hypothesis	μ	n	$\bar{x}$	σ	Significance
9 + students - P	85	473	102	135	0,0032
About text $<$ 300 characters - N	85	81	59	119	0,0264

Table 3.11: Tutor attributes and indication on lifetime.

3.3 Prediction results

During the prediction period April 19th to May 31st 2020, 66 Learnlink customers churned. Out of these 66, 21 could be regarded as natural churn while the other 43 were regarded as ”ordinary” churn. These numbers will be compared to the predicted values in order to determine the performance of the predictions.

3.3.1 Decision tree classification

Three decision tree algorithms were developed based on the exploratory analysis: Natural churn, high risk and low risk. If a customer qualified for neither of these labels, the customer was labeled ”medium risk”. Factors deciding natural and unconditional churn was put in the first model, factors found to be indicating a lower customer lifetime was used in the high risk model and factors found to indicate higher customer lifetime was put in the ”low risk” model. The natural churn model was run first, then the ”high risk” model and then the ”low risk” model.

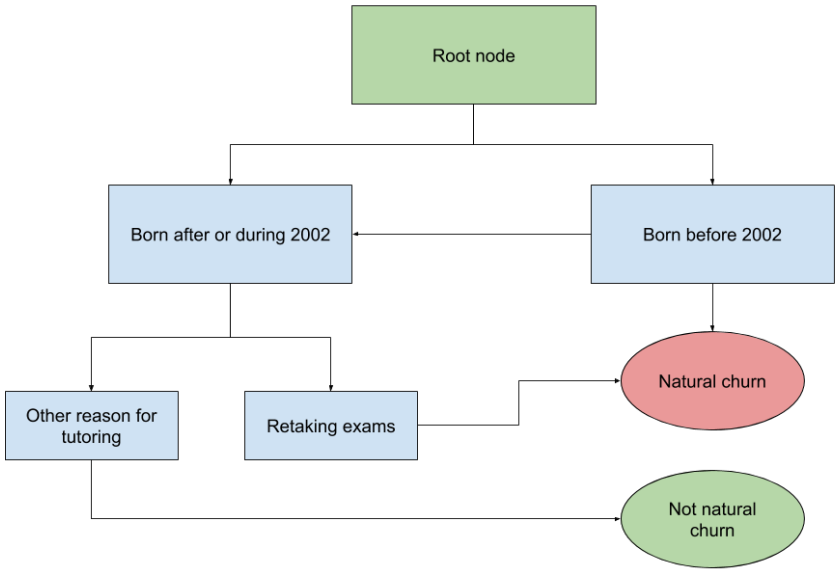


Figure 3.23: Natural churn decision tree.

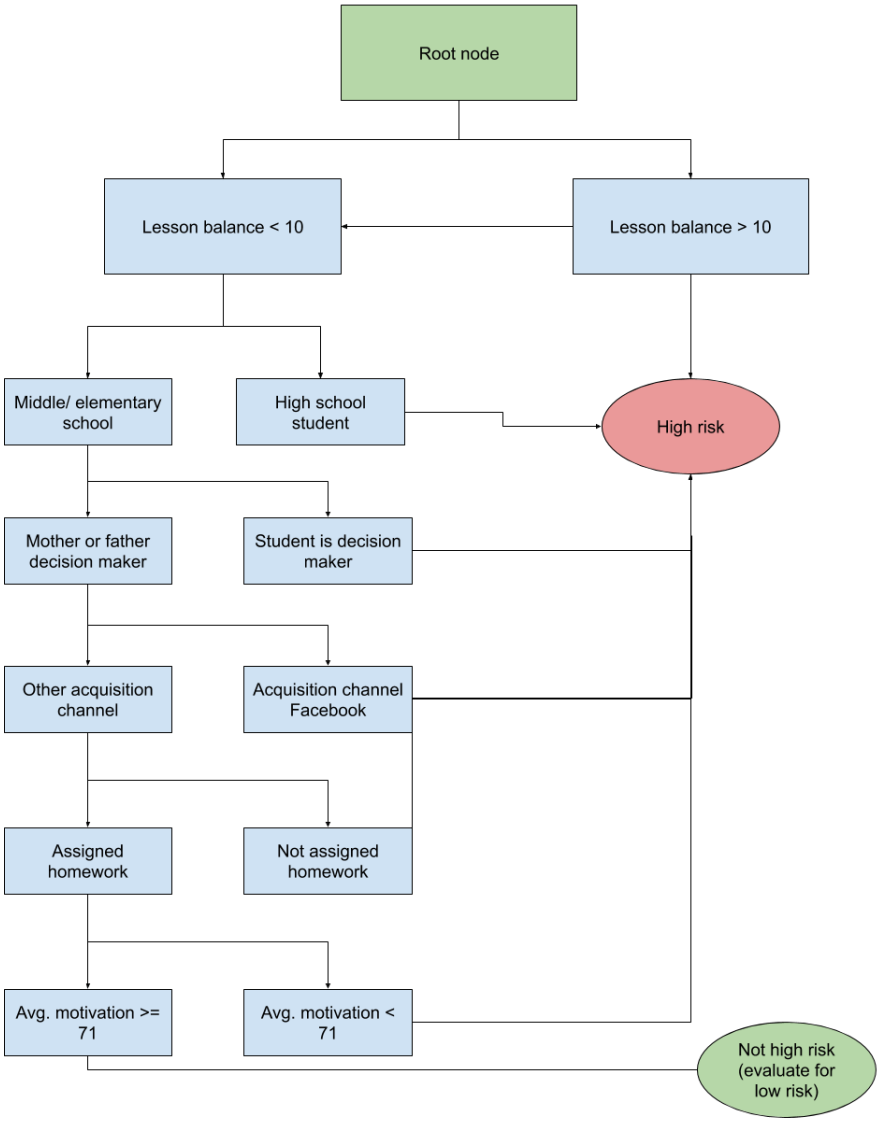


Figure 3.24: High risk decision tree.

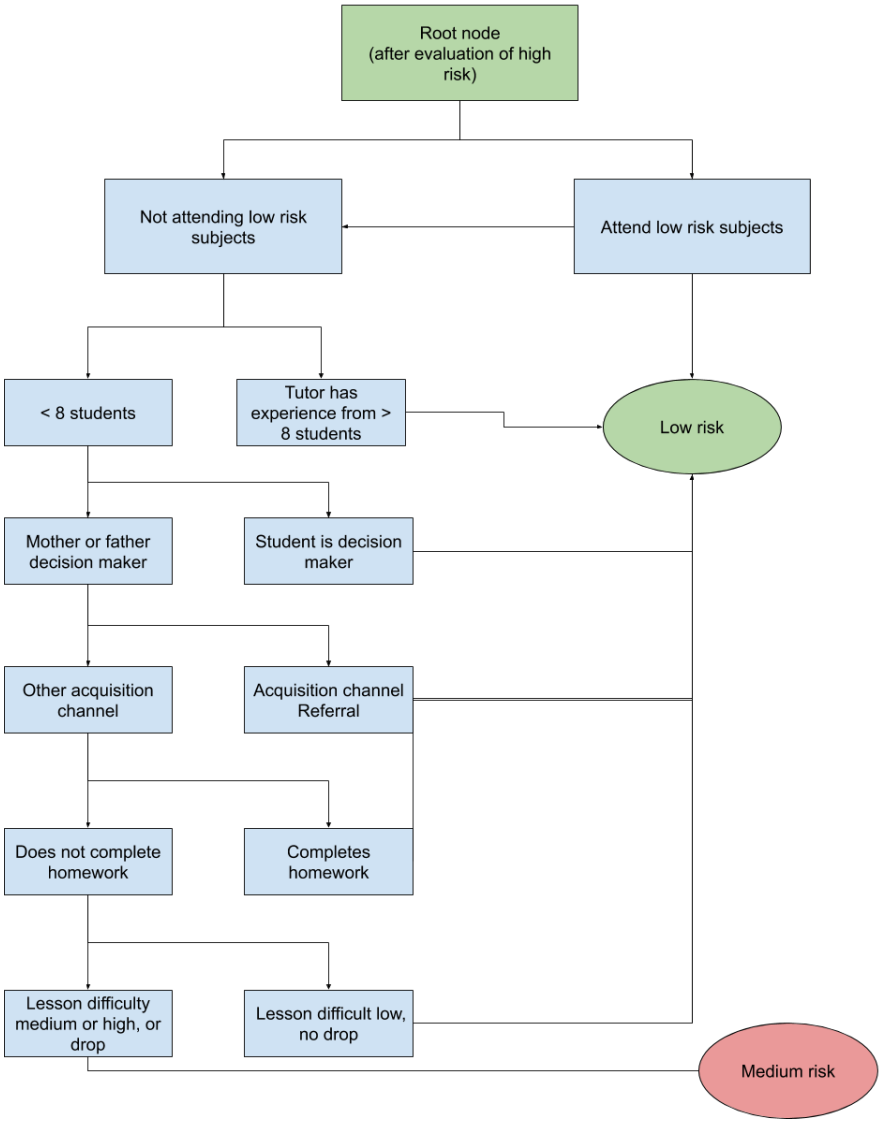


Figure 3.25: Low risk decision tree.

3.3.2 Decision tree algorithm performance

Natural churn

The algorithm classified 43 customer as ”natural churn”. Out of these 43 customers, 21 customers churned during the observation time period.

Measure	Description	Value
Total	Numbers of customers in set	374
Positives	Predicted to churn	43
True positives	Predicted to churn, and churned	21
False negatives	Not predicted to churn, but churned	22
True negatives	Not predicted to churn, did not	0
True negatives	Not predicted to churn, did not	331

Table 3.12: Results from the natural churn decision tree algorithm.

Performance measure	Value
PCC	94%
Precision	49%
Recall	100%

Table 3.13: Performance measures for the natural churn decision tree algorithm.

		Prediction outcome		
		p	**n**	**total**
Actual value	**p′**	49% 21	0% 0	21
	n′	51% 22	100% 331	353
	total	43	331	

High churn risk

The algorithm classified 65 customer as ”fast churn”. Out of these 65 customers, 26 customers churned during the observation time period.

Measure	Description	Value
Total	Numbers of customers in set	329
Positives	Predicted to churn	65
True positives	Predicted to churn, and churned	26
False positives	Predicted to churn, did not churn	39
False negatives	Not predicted to churn, but churned	17
True negatives	Not predicted to churn, did not	247

Table 3.14: Results from the churn decision tree algorithm.

Performance measure	Value
PCC	83%
Precision	40%
Recall	60%

Table 3.15: Performance measures for the churn decision tree algorithm.

		Prediction outcome		
		p	**n**	**total**
Actual value	**p′**	40% 26	6% 17	43
	n′	60% 39	94% 247	286
	total	65	331	

Low churn risk

The algorithm classified 110 customers as ”low risk”, of whom 12 (11%) churned. As this model was built for choosing the customers who were probable to *not* churn, a *positive* here refers to *not churning*, and a *negative* refers to churn. This is opposite to how it was referenced to above, where the models chose customers who *were* likely to churn.

Measure	Description	Value
Total	Numbers of customers in set	329
Positives	Predicted to not churn	110
True positives	Predicted to not churn, did not	98
False positives	Predicted to not churn, did churn	12
False negatives	Not classified as ”low risk”, did not churn	145
True negatives	Not classified as ”low risk”, churned	31

Table 3.16: Results from the churn decision tree algorithm for ”low risk” labelling.

Performance measure	Value
PCC	39%
Precision	89%
Recall	40%

Table 3.17: Performance measures for the churn decision tree algorithm for ”low risk” labelling.

		Prediction outcome		
		p	**n**	**total**
Actual value	**p′**	89% 98	82% 145	243
	n′	11% 12	18% 31	43
	total	110	176	

3.3.3 Classification model with logistic regression in BigQuery

The classification model was built utilizing only the data points found to have a correlative effect in the exploratory analysis. Below is performance measures for the training of the algorithm, as well as a comparison with actual events.

Results from training and evaluation

Actual labels \ Predicted labels	Fast churn	High value	Medium value	% samples
Fast churn	68.12%	5.8%	26.09%	55.2%
High value	22.22%	38.89%	38.89%	14.4%
Medium value	15.79%	18.42%	65.79%	30.4%

Figure 3.26: Confusion matrix from the logistic regression evaluation set.

Aggregate metrics

Threshold	0.0000
Precision	0.5712
Recall	0.5760
Accuracy	0.6320
F1 score	0.5677
Log loss	0.8972
ROC AUC	0.7907

Figure 3.27: Performance metrics from the logistic regression evaluation set.

Logistic regression performance

The logistic regression model was trained with data from 1088 different customers. This group was selected with the same restrictions used in the exploratory analysis. Customers that were active at the time of implementation were excluded from training, as they were

to be used in the prediction. Customers were labelled ”high churn”, ”medium value” or ”high value” (equivalent to ”low risk”) based on their lifetime. Before the model was constructed, the group was divided into training and evaluation sets. The training set was used to let the algorithm learn, while the evaluation set was use to measure the performance of the model.

The algorithm classified 158 of the customers in the prediction set as ”fast churn”. Out of these 158 customers, 27 customers churned during the observation period.

Measure	Description	Value
Customers in set		329
Total	Numbers of customers in set	158
True positives	Predicted to churn, and churned	27
False positives	Predicted to churn, did not churn	131
False negatives	Not predicted to churn, but churned	16
True negatives	Not predicted to churn, did not	131

Table 3.18: Results from the logistic regression algorithm in BigQuery.

Performance measure	Value
PCC	55%
Precision	17%
Recall	63%

Table 3.19: Performance measures for the logistic regression algorithm.

		Prediction outcome		
		p	**n**	**total**
Actual value	**p′**	17% 27	9% 16	33
	n′	83% 131	91% 155	286
	total	158	331	

Chapter 4

Discussion

We will first discuss the results and the achievement of the secondary objectives, before turning to the overall objective and discussing the general achievement.

4.1 Building data analytics dashboards for an early-stage company

4.1.1 Prerequisites for building data infrastructure

An understanding of the underlying technology of the Learnlink platform was crucial for building the data dashboards. The Learnlink Chief Technology Officer illuminated on limitations and opportunities through interviews early on, which made developing the dashboards much easier. Specifically, the challenges that arose from using a document-oriented database had to be understood in order to connect the relevant software. One of the pitfalls avoided was trying to use out-of-the-box analytics software that was not compatible with data derived from document databases. This knowledge also made it easier to detect error sources; some errors were rooted in Google data studio, yet others were due to inconsistencies in the Learnlink database. After a few early iterations on the dashboards, the process for understanding the origin of an error was the following:

1. Check the dashboard for errors. A typical error is to try to combine grouped and non-grouped data points, another is combining data with non-compatible date fields.
2. Check the data source in Google data studio for invalid calculated fields.
3. Investigate the BigQuery table and re-run the query to detect anomalies.
4. Re-run the Firestore-to-SQL-export function.

If neither of these steps uncovers the error source, it is likely that it originates from the database and not from the data analytics funnel. In these cases, a developer with prior knowledge to the back-end architecture should assist in finding the error. A recurring error in the database originated from the fact that Javascript is a weakly typed language, which means there is no type declaration on some of the variables. When some variables have no type or the wrong type, an error can develop throughout the system and in turn affect the export function.

Understanding the business dynamics in Learnlink has also been important for building dashboards. As became evident when establishing the definition of churn, there are countless details that can have a great effect on the knowledge derived from the dashboards. As an example, if our dashboards would not take different business models into account, we would draw the wrong conclusions. The interviews prior to the implementation and a continuous dialogue with employees were very helpful during both the dashboard construction period and the analysis period.

4.1.2 Choosing tools

As the Learnlink team was already using BigQuery and a clear wish was to keep using the current technology, our choice of tools was limited and thus not the main focus for this thesis. Other small companies could, however, benefit from doing more research on combinations of different analytics software. A company with a less complicated business structure and/or a relational database could use a more plug-and-play software. As mentioned previously, companies using relational databases have a wider range of solutions to pick from and an easier set-up process.

The choice of Google Data Studio as a visualization tool worked out well; with an intuitive interface, the dashboard developer can create reports through trial and error. Quick one-time reports called ”*explorers*” can be used for testing functionality and investigating the structure of the data in a data source. The relationship between explorers, data sources and reports are easy to understand, and setting up the first report is quick and hassle-free. Google data studio is free to use but requires BigQuery, which has usage costs depending on the amount of data processed. The costs are however small with the amount of data we have used in the analysis, and other small companies will probably find the costs manageable. Some tasks were, however, unnecessarily difficult to carry out in Google Data Studio. Timestamp management was complicated and you have to be very careful about how you format data and time in order to not create inconsistencies. Blending two or more data sources can be hard and poses challenges when you are creating filters and date pickers. A better solution instead of blending data sources is to create custom tables through tailored BigQuery queries, as was done with the *Churn* dashboard in this project. Showing numbers as percentages of other data fields should be trivial, but there is no straightforward way to do this as you are unable to save a data field as a variable and re-use it in other locations. For other companies who are looking for a low-cost and low -entry data visualization tool, Google data studio is a fair choice. However, for larger institutions who plan to invest in data visualization and have several employees working with developing dashboards, heavier tools should be considered. Other data visualizations tools that should

be assessed are Tableau, Klipfolio and Microsoft Power BI.

4.1.3 Structuring data in BigQuery

When faced with the choice between a *more-is-more*-approach, grouping all relevant parameters into one massive data source, and composing a higher number of smaller tables, going with the latter is definitely advised. Handling smaller data sources made the implementation in Data Studio more rapid, errors were easier to discover and loading times increased. As complex data blending seem to lead to the data dashboards crashing for no good reasons in some cases, blending sources should be kept to a minimum. Hence, tables should be updated continuously when working with reports and tailored to one specific dashboard. Keeping the number of data sources per report to a minimum is also advisable as it makes fault detection easier.

Data in the dashboards was updated every second hour. Google data studio enabled refreshing from every 12th hour to every hour, but the data export from the database to BigQuery was only run every second hour. Hence, data freshness could be improved by rewriting this script. The employees did not have any clear demands for data freshness time, apart from a wish that the data was as up-to-date as possible. If cost reduction is desired while maintaining an acceptable data freshness level, refreshing can be set to 1 hour for the most frequently used dashboards and to 12 hours for the less frequently used ones.

4.1.4 Errors and changes

Trustworthiness is important for ensuring that the visualizations are to be considered reliable. Interviews conducted prior to the implementation indicated that as soon as a few errors commence, employees start to distrust the data and look for other sources of information. Consequently, errors should be detected, reported and fixed continuously to avoid discrediting the knowledge derived from the dashboards.

An important source of obvious errors at the beginning of the work with the thesis was changes done to the Learnlink platform. As the company was still making major changes to the platform during the writing of this thesis, several tables in the database were changed during the period. This illustrates the importance of developing data dashboards in-house. As the platform being analyzed is being developed, so must the data dashboards in order to ensure that the data will still be displayed. Consequently, software developers should be aware of how the data dashboards gather data and understand the potential effects of making changes in the database.

4.1.5 Generalization of data dashboard development

Based on the results, it should be possible for other small companies to use the same building blocks to make data dashboards available for employees. Dashboards with charts are much easier to use for decision-making than extracting database data to spreadsheets or manually calculating patterns by investigating the database. In practice, the only limits to what questions you can ask and display answers to in the dashboards, are that answers have to be measurable and you need the correct data. Queries and calculations are flexible and available ad-hoc.

All software solutions used in this project are low-cost and easy to use if you have a background in software engineering. By using the Firestore-to-BigQuery export, Segment.io, BigQuery and Google Data Studio, building near-real-time visualizations is neither time-consuming nor expensive. While the visualizations do require maintenance as long as the data sources are being developed, the major part of the job is to connect the correct data sources and clean the data for anomalies and duplicates. Consequently, building a data visualization foundation can be achieved during an intensive project period of 1-2 months, with a few hours per week of maintenance succeeding this intensive period. Based on the error rates and the usefulness of continuously receiving employee feedback, the implementation should preferably be done in-house or by someone else closely connected to the company.

4.2 Exploratory analysis

4.2.1 Natural churn

By performing the natural churn analysis, we identified 42 students who were more than 18 years old and 17 students retaking exams. Some customers were in both groups, which resulted in a total of 43 students who were in *at least* one of the groups. This shows that using only the age of the student would have been sufficient; data for whether the student is retaking exams or not is mostly unnecessary. As the highest level where Learnlink offers help is high school, it was assessed as highly probable that these students would end their customer relationship after their exams. This information was used in making the *natural churn* decision tree prediction model in order to split non-preventable and preventable customer churn.

4.2.2 Trends by season

Tutoring businesses see high seasonal fluctuations in customer activity and revenue due to the school year timeline. Learnlink employees reported before the analysis that the most profitable months tend to be during exam season at the end of the school year - April and May - and the least profitable months during summer holidays.

Data from the exploratory analysis showed that most customers churn in May and April for the spring semester and in November for the fall semester. As these months are at the end of the semester and the months where mid-terms and exams are completed, it is reasonable to assume that many customers choose to quit after the important tests. Completing tests or exams might be the end-goal of their tutoring and it might not be the student's intention to continue after the goal is achieved. We saw that customers that churn in April, May and June tend to have had a longer customer relationship than customers churning during other months. Specifically, customers that churn in September, October and February tend to have significantly shorter lifetimes than the ones churning during the rest of the year. These months are just after the beginning of each semester, and customers churning at this point have likely signed up during the beginning of the semester and thus cancelled prematurely.

We saw how the lifetime varied based on the month the first lesson was completed. Customers with a lesson in July, August, December, January and February have significantly higher lifetimes than customers starting up at other times of the year. This is not surprising, as these months are at the beginning of the semester. We can see that the median lifetime for a customer with the first lesson in January is 126 days - about 4 months - which means a cancelled customer relationship in May. Starting with lessons in January to prepare for exams in May is a probable chain of events. Significantly lower lifetimes for customers starting up in May and October (months right before exams or mid-terms) support this logic.

Seasonal trends - impact on strategy

We have uncovered big differences in a median lifetime for customers in the months with the highest lifetime versus the shortest, and these trends make logical sense. In order to seek to maximize the lifetime of customers, the intensity of marketing and customer acquisition activities should be varied according to the expected lifetime for customers acquired during the given month. Even though demand in the market might be the same or higher in May compared to January, a customer relationship lasts 16 times as long when initiated in January. If Learnlink is to earn a profit from May customers, these need to be acquired at low costs, while acquiring customers in January is worth spending resources on. These observations should be taken into account when setting the marketing budget and revising financial plans.

Natural churn, seasonal churn and unwanted churn

Even though this thesis is limited to estimating future churn, a short discussion of the opportunities to mitigate churn is in place.

Our results show that some customers end their relationship with Learnlink because of reasons out of the company's control. Naturally churning customer have completed their exams and are finishing high school. In order to retain these customers, a larger strategic initiative is necessary - the company needs to expand its services to offer learning services for students not attending school as well.

Expected customer lifetime is, as we have discovered, highly seasonal. Even though measures can be taken to stop customers from quitting, market trends will be hard to change. Customers cancelling their subscriptions in May or June are likely to be at the end of their planned customer life-cycle regardless of the measures imposed.

We have previously seen that customers cancelling right after the beginning of the semester have the shortest lifetime. There is no obvious logical reason to stop a tutoring subscription in February when exams are in May. The customer has already made a commitment by paying for the first month of tutoring. Exams are several months away and the goal has not been achieved yet. Hence, we can assume that the reason for cancelling is either dissatisfaction with the service or expectation mismatch. These factors should be possible for Learnlink to control and not wanted for either the customer or the company. If we believe the data, it should be possible to extend the customer lifetime by several months by overcoming the obstacles during the first months of the customer relationship. The greatest potential can be assumed to lie in stopping unwanted churn at the beginning and in the middle of the semester.

4.2.3 Trends by demographics

Discovering demographic groups with high lifetimes can be useful for marketing, as campaigns and communication can be targeted at the most profitable audience. Similarly, short lifetime customer groups can be excluded from targeting through online advertisements.

Levels and subjects

Lifetime trends by level found in the analysis make logical sense and should be of little surprise: high school students, who are older and have fewer years to go before they are too old to be in the target audience, have a shorter lifetime than students in middle school and elementary school.

More interesting results are found after investigating the differences in lifetime for different subjects. For high school subjects, students who receive science ("Naturfag") tutoring show significantly shorter lifetimes than other students. Harder math subjects seem to correlate with higher lifetimes. A logical reason for more advanced mathematics subjects to have higher lifetime is that students in these subjects face harder challenges and need more help to overcome these. Another explanation can be that the students who voluntarily choose harder subjects may have higher motivation or and determination to complete their courses.
Two of the foreign language subjects, German and French, show significantly higher median lifetimes. Correspondingly, English students in elementary and middle school tend to have long customer relationships. We have limited data for foreign language subjects in Middle school, so we are unable to conclude whether this effect appears here as well. The long customer lifetimes in language subjects may arise from how the subjects are mastered. It takes time to build up a vocabulary, learn grammar and earn vocal skills in a foreign language, and students who need help with these subjects might understand that there is no quick-fix and have a longer perspective than students in other subjects.

Decision maker and acquisition channel

Our data indicates that there could be a relationship between who the family decision-maker and contact person is and the duration of the customer relationship. When a parent is involved, churn is lower than if the student is responsible for communicating with the tutor and the Learnlink team. A reason for this might be that students themselves are more impulsive and lack the long term perspective, so that parent involvement would increase lifetime. Another explanation can be that students pay for tutoring themselves and are more price-sensitive, or that the value of the tutoring is communicated better to the parents when they are more closely involved. The analysis of acquisition channel differences shows that customers who have heard about Learnlink through friends and family (referral) have a significantly higher lifetime than others. Customers who first heard of Learnlink through Facebook tend to churn faster. It is logical that recommendation from friends and family has a higher impact than advertisements through social media, lead to a higher determination for referral customers to sustain their tutoring plans.

4.2.4 Trends relating to events during the customer lifespan

Motivation and customer lifetime correlates. Students with a median motivation below 75 have lower lifetimes than students with higher motivation. However, students with very high motivation (close to 100) and high (around 90) seem to behave similarly. Furthermore, motivation seems to decrease during the months prior to churn. More specifically,

we can find a motivation drop the month prior to churn and the month with the last lesson. This seems logical: the reasons for many customers to use tutoring is to obtain higher motivation. If the student motivation drops, the service is no longer as valuable. Another observation is that motivation rises during the first months and then drops. Students might experience very fast progress during the beginning, but after a while, more work has to be put in to keep gaining the same amount of progress. As the initial enthusiasm cools down, motivation drops again. We should point out that the motivation data point is subjective by nature. The tutor who writes the report makes the judgment of the student's motivation and might have different baselines than other tutors. Motivation is in general hard to measure objectively, but an improvement could be achieved by communicating how to link the motivation rating to common signals like ”giving up on an assignment” or ”asking for a break”.

One could assume that students who do not complete their homework might be more probable to churn. However, our data does not indicate that this is the case. We can not find any significant reduction in homework completion during the months prior to churn. Neither do students who in general do not complete their homework churn faster than others. Surprisingly, another pattern emerges; students who are not assigned homework churn faster than others. As tutoring lessons usually are conducted once or twice per week, not being assigned extra homework is probable to reduce the effect and over time the student will make slower progress than with extra homework.

The fact that students with more frequent lessons did not have different lifetimes than others is surprising. On the one hand, one could assume that students who completed more lessons would see faster progress and experience more value from the service. Hence, they might stick to attending tutoring lessons for a longer period of time. On the other hand, more frequent lessons mean higher costs per month. Customers with high bills may have higher requirements for the effectiveness of the service and might quit if their expectations are not met. By looking at our data, we can either assume that none of the previously mentioned effects are important, or that they cancel out.

Indications from lesson difficulty level can seem ambiguous: on the one hand, difficulty level seems to have a slight drop two months prior to churn before a rise during the churn month. On the other hand, students with overall low difficulty level in their lessons seem to have higher lifetime than students with medium or high difficulty level. One can make arguments for both cases. Students who have high difficulty level on their lesson, in other words learning the advanced parts of the curriculum, will probably have better results on tests and feel less that they are lagging behind. This can be a positive experience, but can also give the student and the parents the feeling that “the job is done” and that the student is doing well by herself. Hence, it gives a reason for quitting tutoring classes. However, the low difficulty level can mean that the student is still struggling with the basic parts of the curriculum. This can have a negative effect on motivation (which we have seen drops prior to churn) and reduce the determination for tutoring. Moreover, parents can feel that they are not receiving the value from the service that they expect. In other cases, a low difficulty level can mean that the student is making progress and indeed have a very high

need for tutoring classes. But regardless of the average difficulty level, a drop in difficulty means that progress has halted and that the tutoring is ineffective. It is logical for lack of progress to give rise to a churn.

4.2.5 Tutor attributes and behaviour as an indication of lifetime

One could assume that new tutors quickly understood the mechanisms for delivering on student and parent needs, so that the most inexperienced tutors would have very short lifetimes for their customers. Surprisingly, the effect of tutor experience first arises when a tutor has experience from 9 or more different students. With an average of 2-3 students at a time, it would take a tutor at least 1,5 years to reach experience from nine different students. Regardless, it makes sense that experienced tutors have more satisfied and more loyal students. A strategic impact for the company from these findings would be to give incentives that reward more experienced tutors, as losing experienced tutors can have negative effects on average customer lifetimes and thus revenues. Similarly, providing incentives for new tutors to gain experience faster can make more tutors reach high experience faster.

We found a significant correlation between the number of characters in the tutor's "about me"- text on their profile and lifetime for their customers. The tutor profile is filled out during the application process, and one can assume that the amount of effort put into the application indicates the degree of determination to be employed.

4.2.6 Overall observations from the exploratory analysis

As mentioned in the Background chapter, our data set is tail-heavy due to the high growth in the number of customers. The further away from the present we look, the less data is available. When checking for differences in the data set between customers acquired during the past year and the year before, no significant differences in demographics or subjects were discovered. We can assume that the data we have analyzed provide an acceptable representation of the customer segment today as well as during previous years.

Drivers versus indicators When discussing the implications of our findings, we should distinguish between *drivers* - parameters that can be affected in order to reduce unwanted churn and increase customer lifetimes - and pure *indicators*, which can only help us discover potential fast-churn customers. Although our main goal for this thesis is to establish a method for indication, improving churn rates will be the main goal for the company in the long term. The most prominent drivers seem to be tutor experience and student homework assignment. If Learnlink succeeds in retaining tutors and enabling them to gain more experience faster, it is likely that customer lifetimes will increase. Similarly, requesting all tutors to assign some kind of homework to their students might give positive effects. Motivation might be increased after a drop by changing the tutor, introducing new learning methods or give more positive feedback. Demographic factors can not be altered for customers that are already acquired, but the data can be used to target the customer groups who are the most profitable. The same holds for subjects and school levels - they serve as

indicators for when existing customer churns, but can and should be taken into consideration when laying the company strategy.

Causation and correlation It is important that we point out that none of the connection we have found can prove causation. We have only established correlations and can derive indications to shorter lifetimes, not reasons. In order to determine whether a causation is possible, one should use logic. An example is the connection that was found between the length of the text in the tutor profiles. It makes sense that tutors who put little effort into their tutor profiles also put little effort into tutoring, which results in less satisfied customers. It is, however, unlikely that this factor is a driver for longer lifetimes - asking tutors to fix their profiles would not increase lifetime for the customer. However, a stricter evaluation of tutor profiles can be included in the recruitment process to avoid tutors who have a casual attitude and thus obtain a more determined tutor staff.

4.3 Prediction

4.3.1 Natural churn decision tree

The decision tree algorithm for natural churn was constructed to detect customers that would inevitably quit for reasons out of the company's control, like finishing school or completing exams. Throughout our limited observation period, nearly half of the customers labelled "natural churn" cancelled their subscription. Our immediate observation is that this churn rate is much higher than for the rest of the customer group. As the school year lasts longer than our observation period, we will not know whether all of the labelled customers will churn. Reasons for customers to keep having lessons might be that they are wrongly labelled when signing up (their birth date is wrong) or that they are one year older than the rest of their class. When it comes to customers who have final exams, they might choose to keep the subscription until they receive their results or they might have more exams in the fall semester.

Accuracy for this algorithm was high - just 6% of the samples were labelled wrong. There were only false positives and no false negatives. The only way false negatives could arise in this situation would be if data points were missing or misunderstood. This is hard to detect, as you would have to investigate the data points by talking to the customer directly after the customers cancel their subscription. We might, therefore, assume that the recall is somewhat lower than 100% and that we probably have some false negatives even though it is not visible in our performance tables. 49% precision is fairly good for such a short observation period. It would be interesting to increase the observation period to see if the precision kept increasing further into June.

When it comes to simplicity, implementing and understanding this algorithm is straightforward. The underlying assumptions were derived from the exploratory analysis and aligned with common sense. Gathering and displaying level and age data was no difficult task when the data dashboards were constructed, and writing the algorithm was easily done in BigQuery. It should be feasible for the Learnlink team to continue using this algorithm and develop it further if they find more indications for natural churn.

4.3.2 Churn decision tree

After running the data for active customers through the natural churn decision tree, our churn decision tree algorithm was used in order to predict preventable churn. Our assumption was that all the resulting 329 customers could continue their tutoring lessons given they were motivated enough, but that some of them were more likely than other to decide to put their tutoring to a halt. The algorithm classified 65 customers as "fast-churn" customers, meaning that they were more probable than others to cancel prematurely. 110 customers were classified as "low risk", indicating that they were less probable to churn.

Out of the 65 customers, 26 or 40% chose to cancel during the observation period. The churn rate for the total body of customers (all labels) when disregarding natural churn was

13%, which means that our high-risk group was three times as likely to cancel compared to a randomly chosen customer. Out of the customers labelled as ”low risk”, 11% churned, slightly lower than for the body as a whole.

We will first discuss the ”high risk” model. Given the short observation period and the comparably high ratio of “fast churn”-labelled customers, it is not surprising that there were few false negatives for our prediction and many true negatives. As discussed in the Background section, our aim is to minimize false negatives and limit false positives to a controllable level. This was achieved. Accuracy, precision and recall levels seem to be high enough to be of practical value to the Learnlink team. If the model was used to impose measures on customers, we would have 17 customers who were given some sort of incentive to not quit their subscription. If the incentive had a cost - for example, if customers who were labelled fast-churn received a discount - the expected effectiveness of the measure should be taken into account when calculating the expected profit from reducing churn. In other words, the cost of the discount needs to be evaluated if it is to be used.

The low churn risk model had a high precision value, but low accuracy and recall. The high precision level is not surprising given that most customers will *not* churn in any given time-period of 1,5 months. As the churn rate in this model differs only slightly from the overall churn rate, more development is necessary before relying on the results.

This algorithms for low and high had more parameters and were more intricate than the natural churn algorithm. The actual implementation is manageable and consists of straightforward conditional statements. However, the background for the assumptions originate from a much more thorough exploratory analysis and it is thus more time consuming to build. When it is already established, however, finding more parameters that seem to indicate churn and adding them to the model should be straight forward and can easily be carried out by Learnlink employees in the future.

An advantage of the decision tree algorithms is that it is easy to grasp the logic behind the algorithm's choices. By visualizing the algorithms as trees, one can easily simulate the algorithm manually for individual customers in order to understand the logic. As a result, it will also be easier to evaluate individual parameters used in the algorithm again when more data is available, thus making it possible to revise the complete model and not only to add features. If you were to keep adding branches when discovering correlations for new parameters, you might end up with a model that predicted too many positives and not enough negatives.

It is a fine balance between predicting enough positives and not predicting too many. With too many positives, potential measures can be evaluated as too expensive as they are imposed on many customers unnecessarily. If the number of positives is too low, however, you will end up with false negatives and not impose measures on customers that might have been “saved” from cancelling. It is unlikely that the balance obtained through this first try is optimal, and different balances should be experimented with based on potential measures and budgets. The balance can be adjusted through removing or adding parameters that correlate with faster churn.

A challenge when developing the algorithm further would be to evaluate the effectiveness of the individual parameters and not the model as a whole. A method for carrying out an analysis like that would be to compare the attributes of the group of customers that actually cancelled their subscription (the actual positives) and comparing to the attributes in the model. Also, by splitting the algorithm into more levels where each level is calculated separately before passing the data over to the next level, one can identify how many customers are labelled at each level. This is not very different from what we have done with the natural churn algorithm and this algorithm. Splitting up to test the different levels is advisable to do if developing the algorithm further, as some of the parameters might not be as important as others.

Our model has only used binary decisions. In some cases, sending customers to three or four different destinations may make sense, like if you would want to increase the solution of the results. As an example, one could use motivation to split customers into "fast churn", "medium churn", "slow churn" and "very slow churn".

4.3.3 Logistic regression

The logistic regression model was carried out with a very different approach than the decision trees. Our aim was to utilize the strengths of pre-trained machine learning algorithms to predict churns. This approach relies on the exploratory analysis but is only dependent on a choice of parameters to use for the prediction and not limit values like the decision trees. Hence, building the model was easier, but we have less control over the decisions the model makes.

The regression model was programmed to label customers with three labels: "fast churn", "medium value" and "high value" (which is equivalent to "slow churn"). In the tables for showing performance metrics for the prediction, we have grouped "medium value" and "high value". 158 customer, nearly half of the data set, was predicted to fall into the "fast churn" category. We can see from the table with training data performance that most of the customers in the evaluation set was labelled "fast churn", so it makes sense that the prediction was this high. It is however not practical with such a high number of predicted churn unless a very large avalanche of cancellations is due, because as mentioned when discussing the decision tree models, it makes it harder to impose anti-churn measures. In order to reduce the number of samples that would be put in the "fast churn" label, one could change the limits in the original data set through a trial and error process.

For the training data, the model performed fairly well with 57% precision, at 58% recall and 63% accuracy. The number of true positives was 68% and false negatives around 30%, which is somewhat high. The model did not, however, perform as well in practice. All values were lower when compared to the actual data, and especially precision suffered from the high amount of positives predicted. There was also a surprisingly high number of false negatives when taking into consideration the high number of positives predicted.

The model can probably be improved by introducing more data and experimenting with the label limits. Regardless, the lack of transparency that results from using a pre-

written, pre-trained machine learning algorithm will make it harder to keep developing this model. It is also not given that providing more variables will improve the performance of this model.

4.3.4 Overall prediction results

Our three prediction models based on two algorithms performed overall better than pure random guesses would, which support the underlying knowledge derived from the exploratory analysis. The decision tree algorithms, although the simplest to implement, clearly outperformed the logistic regression model in this case. These algorithms were more closely based on the results from the analysis, which also indicates that we should trust the exploratory analysis.

If evaluating prediction algorithms was our goal, we should have included more algorithms to compare. Our prediction results provide a glimpse of the opportunities for applying prediction to business problems, and it is likely that we would have found both better versions of our current algorithms and new algorithms with better performance. As validations of our findings in the exploratory analysis, two algorithms and three models was sufficient to show that the knowledge we have acquired is not purely random and that the effectiveness of different algorithms varies. Applying the most advanced technology to solve the problem is not always the best approach, which is shown by the superior performance of the simple decision tree algorithm.

All three algorithms can probably be improved by including more data and finding more parameters through deeper analysis. As more parameters will increase the number of positives with our current structure, increasing the amount of data is most advised, a pathway that is already paved as Learnlink keeps growing and acquiring new customers.

The amount of data used to make predictions was limited. All models could probably have been improved by increasing the data set, and for other applications, the advice is to aim for more data to choose from. Our limited data restricted the use of some parameters as the differences were not significant.

Throughout the analysis and the prediction, it has become clear that for a seasonal market like the tutoring market, mid-semester and end-semester churn are two completely different cases. We handled this by establishing a separate measure and a prediction model for natural churn. However, taking this one step further and separating the analysis into two parts would be interesting. One part should be dedicated to mid-semester churn and could look more closely on tutor performance, while the other should investigate the mechanisms that fall into play when exams are over.

Prediction results and the Corona crisis

Just before the observation period started, the Corona outbreak reached its peak in Norway and the government imposed heavy restrictions on many parts of society. Some of the effects applied to Learnlink's customer and tutors; schools were closed and home-schooling was put into effect, Universities were closed and all exams were cancelled. It is likely that these changes affected our data. Some parents chose to take care of all home-schooling themselves during this period, reducing the need for some of the tutors. Other had increased demand for help and wanted more assistance from tutors during this period. These two effects can have cancelled out. The cancellation of exams, however, had two direct effects on Learnlink's business: Fewer short-term exam-focused customers that would usually have churned naturally during May were acquired, and some customers experienced lower demand and cancelled their subscriptions early.

Chapter 5

Conclusion

We have built a data-driven tool that provides an indication of whether customers are likely to prematurely stop their customer relationship in Learnlink, and have thus achieved the main objective. The tool that has been developed throughout this thesis should be sufficient to work as a foundation for solving further business problems with data as well as help mitigate customer churn. It should, however, be noted that the tool could be improved through deeper analysis, more data and investigating other analysis and prediction methods.

Defining customer churn was made more complicated by the frequent change of business model in the subject company. Defining a cancellation as having no more lessons came out as the best approach, one which was compatible with all business models. The exploratory analysis was the most comprehensive step of our process, as there were few limits to factors that could be investigated. Throughout the analysis, several interesting patterns emerged and were checked for significance. It became clear that both demographic, behavioural and tutor data could provide indications of whether a customer was probable to prematurely cancel their Learnlink subscriptions. These findings were validated with four prediction models based on two different algorithms. From the prediction results, we could conclude that the findings in the exploratory analysis were likely to be of importance and not entirely random and that the decision tree algorithm had the best performance.

The prediction models can clearly have value in themselves, and not only as a validation technique. Using these models to indicate customer churn can help the company calculate expected churn rates, and they can use the indicators to impose targeted measures to stop customers from cancelling subscriptions. Combined with the knowledge that already has been extracted from the data dashboards and the potential knowledge that comes with more data and more analysis, the Learnlink team should be better equipped for handling the churn problem in the future.

It is clear from our findings that today, not only large enterprises can utilize the power

of data analytics. The tools are available at low costs, and by investing time in connecting the right software and establishing an infrastructure, it can be maintained with little effort. The knowledge that can be derived is limited by the data that is collected and the time that is spent for analysis. Other than that, there are few limiting factors.

5.0.1 Errors, biases and improvements

As mentioned, the amount of data available has been a limiting factor for this project. Optimally, one would want data that is collected over a long time period with little variation in other factors that the ones that should be investigated. This is, however, not realistic, especially not for a small company in the growth phase. The changes of business model and other fields where the company has matured might have affected the churn rates in certain periods. One might expect that during periods of big change, like when switching business model, churn rates could have been higher than during more stable times. We did not have enough data at hand to know if this is correct.

The definition of churn that was chosen, *lesson stop*, enabled us to compare data from periods with different business models. However, the definition did not count activity reduction as churn, which could be misleading. When a customer reduces the number of lessons and thus their monthly spend considerably, the comparable serious consequences as a churn applies. Optimally, our definition should cover these events as well as complete halts in payments. By extending the *subscription cancellation* definition to include downgrades, other mechanisms that lead to revenue loss can be detected. For instance, a downgrade that results in a revenue drop of more than 50 per cent for the customer could be regarded as a churn. It should be noted that this definition is more complicated, so using it for analyzing historical data from before August 2019 might still be impractical.

The thesis could have been improved by having a longer observation period. In order to validate the model further, churn data for the rest of 2020 should be compared to our predictions.

Utilizing some of the data that was on purpose excluded from this analysis might also give better results. For instance using data from the text-message provider would enrich the view of the communication volume for customers, which might give indications of the number of support tickets the customer is engaged in. Data from customer satisfaction surveys could also be included.

5.0.2 Applications

Other small companies that can gather customer data through web channels should be able to achieve similar results using the same approach. All tools used to build the infrastructure are available, either for small fees or for free. Implementing the tools does not require extraordinary technical skill and should be feasible for software developers in

most companies. When the tools are implemented, other employees can be trained to complete analysis without understanding the complete setup. After reviewing the results from investigating this one specific problem of customer churn, it is clear that many companies can benefit from using data analytics for decision making and business strategy. The approach can be used to solve churn problems for similar companies with different customer groups or to understand other mechanisms. Other possible problems to address in Learnlink could be demand forecasting, tutor recruitment and selection, lead qualification or irregularity detection in tutoring lessons. Use cases for other companies vary based on what kind of data they have access to. A prerequisite for using this approach would be that you have enough data to reach statistical significance for some of the results. There is no clear limit here as significance depends on the variations, but a data set with at least 1000 customers is recommended. It is also necessary that you have access to enough parameters to distinguish the samples from each other. Most variables are hard to gather for events that have happened previously, so those who plan to use data analytics in the future should map out the necessary data points and starting gathering the data as soon as possible.

For a small company that aims to use data analytics to gain insight and solve problems, the following procedure could be used.

1. Map out all data that might be relevant for analysis, and make sure this data is collected.
2. Prioritize areas for analysis and business problems to solve based on the data at hand and the seriousness of the problems.
3. Translate the problems to quantitative questions and establish clear, unambiguous definitions.
4. Choose analytics software based on the current technology stack and available resources, and connect all data sources to one data destination that can be queried. Clean the data.
5. Explore the available data by looking for patterns, and establish hypothesises that are checked for significance.
6. If you are investigating recurring events, use the knowledge to construct a decision tree of condition statements and test the outcomes against actual events.
7. Evaluate the results, make adjustments and develop the model further. Keep searching for new patterns as the data set grows.

Knowledge derived from this analysis could also be of use to other learning institutions, public as well as private. Most educational institutions aim to minimize student drop-out. We have discovered that drops in motivation influence the willingness to quit more than the overall motivation level. An institution can monitor student motivation either by self-reporting or through activity assessments, and impose measures for students who seem to have a sudden change in motivation. The understanding of tutor attributes can also be utilized when employing student assistants or teacher for student follow-ups.

Our data indicate that it takes time to build up enough experience to create better results for students. Nine students for a tutor is equivalent to at least 1,5 years of experience with a typical tutoring frequency, which can point to that hiring student assistants for one or two semesters might not be a good idea. We have also seen that tutors who show limited commitment when filling out their tutor profile, equivalent to an application, have students that more often resign from learning activities prematurely. Based on these finding, a recommended strategy for a university that wants to minimize student resignation would be to pick dedicated students from the lower levels to train as assistants and to keep those assistants engaged in different subjects until they finish their degree. Our data also shows that handing out assignments to students is effective in itself even though the students might not always complete the assignments in time.

5.1 Further work

5.1.1 Further develop the prediction models

There might be a reason to have different prediction models for end-of-the-semester churn and for mid-semester churn, as the metrics that define churn in these periods are different. End-of-the-semester churn might be more defined by demographics, while mid-semester churn is likely to be more defined by bad customer experiences and lack of satisfaction. This should be investigated further.

We have seen that the prediction models can be used to indicate higher churn probability. If the probability is not high enough to justify imposing anti-churn measures for the selected customer group, iterations should be made in order to increase the precision of the model. As imposing measures on the customer base would alter the data, there has to be made a choice of whether the goal is to make iterations on the model or to stop the predicted customers from cancelling. Alternatively, one can divide the customer group and impose measures on some of them. This will help validate whether the measures work or not.

5.1.2 Analysis of virtual classroom data

The virtual classroom online.learnlink.no is used for thousands of tutoring lessons every month. This produces massive amounts of data that could be analyzed in order to understand more about the educational aspects of the tutor and student relationship. As the recording of one single tutoring lesson will appropriate several gigabytes of storage space, storing all this video data will be expensive and not recommended. Another approach is to use a big data framework to analyze the stream of video data that is produced by the virtual classrooms. This way, metadata can be derived and stored instead of the raw data, and it would be feasible to extract interesting knowledge from the tutoring data.

Bibliography

[1] Leibovich-Raveh T. A basic sigmoid function; 201. Available from: https://www.researchgate.net/figure/A-Basic-sigmoid-function-with-two-parameters-c1-and-c2-as-commonly-used-for-subitizing_fig2_325868989.

[2] McKinsey. How companies are using big data and analytics. McKinsey Quarterly April 2016. 2016;2. Available from: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/how-companies-are-using-big-data-and-analytics.

[3] Hagiu A, Wright J. When Data Creates Competitive Advantage. Harvard Business Review, YEAR = 2020, Volume = 1, URL = https://hbrorg/2020/01/when-data-creates-competitive-advantage;.

[4] Newman R. How we scaled data science to all sides of AirBnB. Venturebeat. 2015;June. Available from: https://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/.

[5] McKinsey. The age of analytics: Competing in a data-driven world. McKinsey Quarterly December 2016. 2016;4. Available from: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.

[6] Chaudhuri S, Dayal U, Narasayya V. An overview of business intelligence technology. Communications of the ACM. 2011;54. Available from: https://dl.acm.org/doi/fullHtml/10.1145/1978542.1978562.

[7] Berg V, Birkeland J, Pappas I, Jaccheri L. The Role of Data Analytics in Startup Companies: Exploring Challenges and Barriers. Conference on e-Business, e-Services and e-Society I3E. 2018;p. 205–216. Available from: https://www.researchgate.net/publication/328245398_The_Role_of_Data_Analytics_in_Startup_Companies_Exploring_Challenges_and_Barriers.

[8] Mikalef P, Pappas IO, Krogstie J, Giannakos M. Big data analytics capabilities: a systematic literature review and research agenda. Inf Syst E-Bus Manage. 2018;16:547–578. Available from: https://rdcu.be/b4v6I.

[9] O'Toole T. The Best Approach to Data Analytics - Harvard Business Review webinar; 2020. Available from: https://hbr.org/2020/03/whats-the-best-approach-to-data-analytics.

[10] Tveit T, Berggren J. Personal communication. Pilestredet 75C, 0354 OSLO, NORWAY; 2019-2020.

[11] Ganesh A. NoSQL for the serverless age: Announcing Cloud Firestore general availability and updates. Google Cloud Product News. 2019;January 31st. Available from: https://cloud.google.com/blog/products/databases/announcing-cloud-firestore-general-availability-and-updates.

[12] Drake M. A Comparison of NoSQL Database Management Systems and Models. Digital Ocean. 2019;August 9th. Available from: https://web.archive.org/web/20190813163612/https://www.digitalocean.com/community/tutorials/a-comparison-of-nosql-database-management-systems-and-models.

[13] Berggren J. Personal interview. Pilestredet 75C, 0354 OSLO, NORWAY; 2019.

[14] Raisinghani J. Should you use MongoDB or SQL databases for analytics? Holistics Blog. 2018;August 28th. Available from: https://www.holistics.io/blog/should-you-use-mongodb-or-sql-databases-for-analytics/.

[15] Elmasri R, Navathe SB. Fundamentals of Database Systems; 2015.

[16] Collins G. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology. 2014;14. Available from: https://rdcu.be/b3wiA.

[17] Leskovec J, Rajaraman A, Ullman JD. Mining of Massive Datasets; 2014.

[18] Abbott D. Applied Predictive Analytics; 2014.

[19] Tharwat A. Classification assessment methods. BMC Medical Research Methodology. 2018;Available from: https://www.sciencedirect.com/science/article/pii/S2210832718301546.

[20] Kleinbaum DG, Klein M. Logistic Regression - A self-learning text; 2010.

Appendix

:

A BigQuery table queries

A.1 Churn

```
SELECT
    project.ID AS project_ID,
   TIMESTAMP(project.approved) AS project_approved,
   TIMESTAMP(project.cancelled) AS project_cancelled,
    project.categories__0 AS project_categories,
    categories.title AS project_category_title,
    project.channel AS project_channel,
   TIMESTAMP(project.closed) AS project_closed,
    project.completed AS project_completed,
   TIMESTAMP(project.created) AS project_created,
    project.customerAID AS project_customerAID,
    project.customerUID AS project_customerUID,
    project.description AS project_description,
    project.fee AS project_fee,
    project.sellerUID AS project_sellerUID,

    customerAccount.birthdate AS customer_birthdate,
    customerAccount.certificate AS customer_certificate,
    customerAccount.email AS customer_email,
    customerAccount.phone AS customer_phone,

    customerUser.customerType AS customer_customerType,
    customerUser.decisionMaker AS customer_decisionMaker,
    customerUser.decisionTime AS customer_decisionTime,
    customerUser.diagnosis AS customer_diagnosis,
    customerUser.contacted AS customer_contacted,
    customerUser.priority AS customer_priority,
    customerUser.gender AS customer_gender,
    customerUser.genderPreference AS
        customer_genderPreference,
    customerUser.parentOccupation AS
        customer_parentOccupation,
```

```
  customerUser.priceSensitivity AS
      customer_priceSensitivity,
  customerUser.problemOrigin AS customer_problemOrigin,
  customerUser.referrer AS customer_referrer,
  customerUser.preferredStart AS customer_preferredStart,
  customerUser.studentBirthYear AS
      customer_studentBirthYear,
  customerUser.studentGender AS customer_studentGender,
  customerUser.studentInterests AS
      customer_studentInterests,
  customerUser.seller AS customer_seller,

  sellerAccount.city AS seller_city,
  sellerAccount.email AS seller_email,
  sellerAccount.phone AS seller_phone,

  stripe.subscriptions__total_count AS
      subscriptions_total_count,
  stripe.discount AS discount,
  stripe.subscriptions__plan__nickname AS subscription,
  stripe.subscriptions__plan__active AS
      active_subscription,
  stripe.subscriptions__canceled_at AS canceled_at,
  stripe.subscriptions__amount AS lesson_price,
  stripe.subscriptions__latest_invoice AS latest_invoice,

  balance.balance AS balance,
  balance.updated AS balance_last_updated,

  lesson.first_lesson_date AS first_lesson_date,
  lesson.last_lesson_date AS last_lesson_date,
  lesson.num_paid_lessons AS num_paid_lessons,
  lesson.num_lessons AS num_lessons,

  report.avg_motivation AS average_motivation


FROM
  `learnlink-prod.firestore.projects` project
LEFT JOIN
  `learnlink-prod.firestore.accounts` customerAccount
  ON project.customerUID = customerAccount.uid
LEFT JOIN
  `learnlink-prod.firestore.categories` categories
  ON project.categories__0 = categories.doc_ID
```

```
LEFT JOIN
  `learnlink-prod.firestore.users` customerUser
  ON project.customerUID = customerUser.uid
LEFT JOIN
  `learnlink-prod.firestore.accounts` sellerAccount
  ON project.sellerUID = sellerAccount.uid
LEFT JOIN
  `learnlink-prod.firestore.stripeCustomerAccounts` stripe
  ON project.customerUID = stripe.metadata__uid
LEFT JOIN
  `learnlink-prod.firestore.balances` balance
  ON project.customerUID = balance.uid
LEFT JOIN
  `learnlink-prod.firestore.users` sellerUser
  ON project.sellerUID = sellerUser.uid
LEFT JOIN (
  SELECT
    report.projectID,
    AVG(report.motivation) AS avg_motivation
  FROM `learnlink-prod.firestore.reports` AS report
  GROUP BY report.projectID) as report
  ON project.ID = report.projectID
LEFT JOIN (
  SELECT
    lesson.projectID,
    COUNT(lesson) AS num_lessons,
    MIN(TIMESTAMP(lesson.startTime)) AS first_lesson_date,
    MAX(TIMESTAMP(lesson.startTime)) AS last_lesson_date,
    COUNT(lesson.paymentID) AS num_paid_lessons
  FROM `learnlink-prod.firestore.lessons` AS lesson
  WHERE lesson.cancelled = 0
  GROUP BY lesson.projectID) as lesson
  ON project.ID = lesson.projectID
ORDER BY lesson.num_lessons DESC;
```

A.2 Lessons and reports

```
SELECT
subject.subject_title AS subject_title,
subject.sent_at AS sent_at,

chapter.timestamp AS chapter_timestamp,


  lesson.customerUID AS lesson_customerUID,
  lesson.duration AS lesson_duration,
  TIMESTAMP(lesson.endTime) AS lesson_endTime,
  lesson.location AS lesson_location,
  lesson.paymentID AS lesson_paymentID,
  lesson.projectID AS lesson_projectID,
  lesson.reportID AS lesson_reportID,
  TIMESTAMP(lesson.reversed) AS lesson_reversed,
  lesson.sellerUID AS lesson_sellerUID,
  TIMESTAMP_SECONDS(lesson.startTime) AS lesson_startTime,
  lesson.status AS lesson_status,
  lesson.subtracted AS subtracted,
  lesson.transferred AS transferred,
  lesson.amount/100 AS amount,
  lesson.ID AS lesson_ID,
  TIMESTAMP_SECONDS(lesson.cancelled) AS lesson_cancelled,
  lesson.rating AS lesson_rating,

  report.doc_ID AS reportID,
  report.motivation AS motivation,
  report.personalComment AS personal_comment,
  report.nextLessonPropDate AS next_lesson_proposed_date,
  report.improvementPlan AS improvementplan,
  report.week1 AS week1_plan,
  report.week2 AS week2_plan,
  report.week3 AS week3_plan,
  report.week4 AS week4_plan,
  report.week5 AS week5_plan,
  report.week6 AS week6_plan,
  report.lessonLengthRec AS lesson_duration_actual,
  report.recommendedPackage AS recommended_subscription,
  report.longTermGoal1 AS longTermGoal1,
  report.longTermGoal1 AS longTermGoal2,
  report.longTermGoal1 AS longTermGoal3,
  report.notes AS lesson_notes,
  report.positiveTopics AS positiveTopics,
  report.infoToLearnlink AS infoToLearnlink,
```

```
  report.studiedTopics AS studiedTopics,
  report.membershipRec AS membershipRec,
  report.goalGrade AS goalGrade,
  report.negativeTopics AS negativeTopics,
  report.homework AS homework,
  report.nextLessonPropStartTime AS
      nextLessonPropStartTime,
  report.goalLater AS goalLater,
  report.initialAbsence AS initialAbsence,
  report.lastLesson AS is_last_lesson,
  report.firstReport AS is_first_report,
  report.initialGrade AS initialGrade,
  report.lessonNumberRec AS lessonNumberRecorded,
  report.goal AS goal,
  report.goalToday AS goalToday,
  report.goalAbsence AS goalAbsence,
  report.nextTime AS nextTime,
  report.character AS reportCharacter,
  TIMESTAMP_SECONDS(report.created) AS sent_time,

  lessonagg.first_lesson_date AS first_lesson_date,
  lessonagg.last_lesson_date AS last_lesson_date,
  lessonagg.num_paid_lessons AS num_paid_lessons,
  lessonagg.num_lessons AS num_lessons



FROM `learnlink-prod.firestore.lessons` lesson
LEFT JOIN `learnlink-prod.javascript_112.open_subject`
    subject ON
subject.user_id = lesson.customerUID
LEFT JOIN `learnlink-prod.javascript_112.open_chapter`
    chapter
ON subject.user_id = chapter.user_id
LEFT JOIN `learnlink-prod.firestore.reports` report ON
lesson.ID = report.lessonID
LEFT JOIN (
  SELECT
     lesson.projectID,
     COUNT(lesson) AS num_lessons,
     MIN(TIMESTAMP_SECONDS(lesson.startTime)) AS
         first_lesson_date,
     MAX(TIMESTAMP_SECONDS(lesson.startTime)) AS
         last_lesson_date,
     COUNT(lesson.paymentID) AS num_paid_lessons
```

```
FROM `learnlink-prod.firestore.lessons` AS lesson
WHERE lesson.cancelled = 0
GROUP BY lesson.projectID) as lessonagg
ON lesson.projectID = lessonagg.projectID
```

A.3 Tutors

```
SELECT
postalCode,
notes,
tutor.uid AS tutor_UID,
registerDate,
employee,
identified,

num_lessons,
first_lesson_date,
last_lesson_date,

num_projects,

profile.tutorText AS tutorText,
profile.aboutText AS aboutText,
profile.interestsText AS interests,
profile.minutesSold AS minutesSold,
profile.numberOfCustomers AS numberOfCustomers,
profile.completedMotivatorCourse AS motivatorCourse,
profile.tagLine AS studies


FROM `learnlink-prod.firestore.users` tutor
LEFT JOIN (
SELECT
    lesson.sellerUID,
    COUNT(lesson) AS num_lessons,
    MIN(TIMESTAMP(lesson.startTime)) AS first_lesson_date,
    MAX(TIMESTAMP(lesson.startTime)) AS last_lesson_date,
FROM `learnlink-prod.firestore.lessons` AS lesson
  WHERE lesson.cancelled = 0
  GROUP BY lesson.sellerUID) as lesson
  ON tutor.uid = lesson.sellerUID
LEFT JOIN (
SELECT
   COUNT(project) AS num_projects,
   project.sellerUID
FROM `learnlink-prod.firestore.projects` AS project
GROUP BY project.sellerUID) as project
ON tutor.uid = project.sellerUID
LEFT JOIN `learnlink-prod.firestore.profiles` profile
ON profile.uid = tutor.uid
```

```
WHERE tutor.seller IS TRUE
```

B Prediction

B.1 Decision trees

Natural churn decision tree

```
CASE WHEN CONTAINS_TEXT(customer_customerType, "Exam")
    THEN 1
WHEN birth_numeric < 2002 THEN 1
ELSE 0
END
```

Fast churn decision tree

```
CASE WHEN
balance_hours > 10 THEN 1
WHEN CONTAINS_TEXT(category_level,"High") THEN 1
WHEN CONTAINS_TEXT(customer_decisionMaker,"student") THEN 1
WHEN CONTAINS_TEXT(customer_referrer,"facebook") THEN 1
WHEN homework_completion IS null then 1
WHEN average_motivation < 71 then 1
ELSE 0
END
```

Low churn risk decision tree

```
CASE WHEN CONTAINS_TEXT(customer_referrer,"referral") THEN
    1
WHEN NOT CONTAINS_TEXT(category_level,"High") AND
    CONTAINS_TEXT(project_category_title,"Engelsk") THEN 1
WHEN CONTAINS_TEXT(customer_decisionMaker,"mother") THEN 1
WHEN CONTAINS_TEXT(customer_decisionMaker, "student") THEN
    0
WHEN NOT CONTAINS_TEXT(category_level,"High") AND
    average_motivation < 90 AND average_motivation > 79
    THEN 1
WHEN lesson_difficulty_drop = 1 THEN 1
WHEN tutor_students > 9 then 1
ELSE 0
END
```

B.2 Logistic regression in BigQuery

Structuring data for prediction in BigQuery

```
SELECT
   project.ID AS project_ID,
   categories.title AS subject,
   categories.level AS level,
  EXTRACT(month FROM TIMESTAMP(project.created)) AS
      signup_month,


  CASE
     WHEN ROUND(lesson.diff/(86400),0) < 60  THEN 'Fast
         churn'
     WHEN ROUND(lesson.diff/(86400),0) < 180  THEN 'Medium
         value'
     ELSE 'High value'
  END AS lifetime_months,

ROUND(30*(num_lessons-1)/(1+lesson.diff/86400),1) AS
    lessons_per_month,


   report.avg_motivation AS average_motivation,

  LENGTH(tutor.aboutText) AS aboutText,
   tutor.minutesSold AS minutesSold,
   tutor.numberOfCustomers AS numberOfCustomers,


  CASE
     WHEN stripe.subscriptions__plan__active THEN
         'prediction'
     ELSE 'training'
  END AS dataframe
FROM
   `learnlink-prod.firestore.projects` project
LEFT JOIN
   `learnlink-prod.firestore.categories` categories
  ON project.categories__0 = categories.doc_ID
LEFT JOIN
   `learnlink-prod.firestore.users` customerUser
  ON project.customerUID = customerUser.uid
LEFT JOIN
```

```
  `learnlink-prod.firestore.stripeCustomerAccounts` stripe
  ON project.customerUID = stripe.metadata__uid

LEFT JOIN `learnlink-prod.firestore.profiles` tutor
ON tutor.uid = project.SellerUID

LEFT JOIN (
  SELECT
    report.projectID,
    AVG(report.motivation) AS avg_motivation
  FROM `learnlink-prod.firestore.reports` AS report
  GROUP BY report.projectID) as report
  ON project.ID = report.projectID
LEFT JOIN (
  SELECT
    lesson.projectID,
    COUNT(lesson) AS num_lessons,
    MAX(lesson.startTime) - MIN(lesson.endTime) AS diff,
    COUNT(lesson.paymentID) AS num_paid_lessons
  FROM `learnlink-prod.firestore.lessons` AS lesson
  WHERE lesson.cancelled = 0
  GROUP BY lesson.projectID) as lesson
  ON project.ID = lesson.projectID
WHERE num_lessons > 0 AND EXTRACT(year FROM
    TIMESTAMP_SECONDS(project.created)) > 2017 AND
    lesson.diff IS NOT NULL
ORDER BY project.created ASC;
```

Build model

```
CREATE OR REPLACE MODEL
  `learnlink-prod.datastudiojonas.prediction_model`
OPTIONS
  ( model_type='LOGISTIC_REG',
    input_label_cols=['lifetime_months'],
     auto_class_weights=TRUE
  ) AS
SELECT
  *
FROM
  `learnlink-prod.datastudiojonas.prediction_input_view`
WHERE
  dataframe = 'training'
```

Predict with model

```
SELECT
```

```
  *
FROM
  ML.PREDICT (MODEL
      `learnlink-prod.datastudiojonas.prediction_model`,
    (
    SELECT
      *
    FROM
      `learnlink-prod.datastudiojonas.prediction_input_view`
    WHERE
      dataframe = 'prediction'
    )
  )
```

Jonas Hyllseth Ryen

Predicting Customer Churn with Data Analytics

NTNU
Norwegian University of
Science and Technology

LEARNLINK