Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
NTNU Business School

Jesper Ødegård
Marcus André Glover Meek

Explainable Machine Learning for Customer Churn Prediction

A study of customer characteristics and the effect of external and macroeconomic factors

Master's thesis in Economics and Business Administration
Supervisor: Ranik Raaen Wahlstrøm
May 2023

NTNU
Norwegian University of
Science and Technology

Jesper Ødegård
Marcus André Glover Meek

Explainable Machine Learning for Customer Churn Prediction

A study of customer characteristics and the effect of external and macroeconomic factors

Master's thesis in Economics and Business Administration
Supervisor: Ranik Raaen Wahlstrøm
May 2023

Norwegian University of Science and Technology
Faculty of Economics and Management
NTNU Business School

NTNU
Norwegian University of
Science and Technology

Preface

This thesis is created as a part of our 2-year master's at NTNU Handelshøyskolen. There have been many challenges throughout, and we look forward to even more challenges ahead.

We thank our collaborative bank for access to the data and the opportunity to work on and create this exciting thesis. We would also like to extend a great thanks to our supervisor Ranik Raaen Wahlstrøm at the Faculty of Economics and Management at NTNU Business School, for his efficiency, valuable insights and great suggestions along the way.

The content of this assignment is the responsibility of the authors.

Abstract

This thesis is delimited to the financial data of customers in a bank with the help of a collaborative bank and external and macroeconomic data. The thesis investigates what characteristics in loan customers can influence the likelihood of their churn, with the help of prediction through the machine learning (ML) methods; Logistic Regression and XGBoost. The variable analysis of SHAP and LASSO are used to better understand the ML models and the importance of variables.

The findings of this thesis indicate that the age of the loan customer and credit risk, as indicated by PD, are key factors in explaining customer churn. Additionally, loan-to-value (LTV), repayment plan, customer duration, and the loan balance of "boligkreditt", significantly influenced the likelihood of customer churn alongside other moderately impactful variables. However, variables that might be expected to have an impact, such as area of residence, size of the households and others showed no significant influence. The characteristics influencing churn in valuable customers differed from those in other customers. For valuable customers, the age of the loan customer had a lesser impact, while income, DTI, and repayment loan balance played a more significant role.

The variable analysis confirms that external and macroeconomic factors influence customer churn. Factors such as the housing price index and policy rate held significance, especially for higher-valued customers. Regarding the ability to predict customer churn using ML methods, both XGBoost and LR showed weak predictive results. However, there were indications that churn predictions within banks could be made with an increased amount of data and reduced imbalances.

In conclusion, this thesis contributes to understanding the factors influencing loan customers' churn and the predictive abilities of ML methods on financial banking data. Improvements in data quality, model performance, and incorporation of qualitative data are suggested for future studies to achieve more robust and actionable results.

Sammendrag

Denne avhandlingen er avgrenset til finansielle data om kunder fra en samarbeidende bank, og eksterne og makroøkonomiske data. Avhandlingen undersøker hvilke egenskaper hos lånekunder som kan påvirke sannsynligheten for kundefrafall ved hjelp av følgende maskinlæringsmetoder (ML) for prediksjon; Logistisk Regresjon og XGBoost. Variabel-analysene av SHAP og LASSO brukes for å forstå resultatene fra ML-modellene og variablenes betydning.

Resultatene fra denne avhandlingen indikerer at alderen til lånekunder og kredittrisiko, som indikert av PD, er sentrale faktorer for å forklare kundefrafall. I tillegg påvirker følgende variabler sannsynligheten for kundefrafall betydelig; lån-til-verdi (LTV), tilbakebetalingsplan, kundens varighet og boligkreditt, sammen med andre moderat påvirkende variabler. Imidlertid viste variabler som er forventet å ha en innvirkning, slik som bostedsområde, størrelse på husholdninger og andre, ingen signifikant innflytelse. Egenskapene som påvirker frafall hos verdifulle kunder, var forskjellige fra de hos andre kunder. For verdifulle kunder hadde alderen til lånekunden mindre innvirkning, mens inntekt, DTI (gjeldsgrad) og lånebalanse hadde en mer betydelig rolle.

Variabelanalysen bekrefter at eksterne og makroøkonomiske faktorer påvirker kundefrafall. Faktorer som boligprisindeks og styringsrente hadde betydning, spesielt for kunder med høyere verdi. Når det gjelder evnen til å forutsi kundefrafall ved hjelp av ML-metoder, viste både XGBoost og logistisk regresjon svake prediktive resultater. Det var derimot indikasjoner på at prediksjon av kundefrafall innenfor banker kunne gjøres med en økt mengde data og reduserte ubalanser.

Konklusjonen er at denne avhandlingen bidrar til å forstå faktorene som påvirker kundefrafall blant lånekunder, og de prediktive evnene til ML-metoder på finansielle bankdata. Forbedringer i datakvalitet, modellprestasjon og innlemming av kvalitative data foreslås for fremtidige studier for å oppnå mer robuste og handlingsrettede resultater.

Table of Contents

List of Figures **vii**

List of Tables **viii**

1 Introduction **1**

1.1 Research questions 3

1.2 Structure of the thesis 4

2 Literature **5**

2.1 Customer churn and factors 5

2.2 External and macroeconomic factors 7

3 Data **9**

3.1 Data description 9

3.2 Variables 10

3.2.1 Target variable 10

3.3 Data pre-processing 11

3.3.1 Data cleaning 11

3.3.2 Transformation of variables 14

3.4 Final dataset 15

4 Method **16**

4.1 Data sampling and Walk-forward testing 16

4.1.1 Imbalanced data 17

4.2 XGBoost . . . 18
4.2.1 Hyperparameter tuning - XGBoost . . . 19
4.3 Logistic regression . . . 20
4.4 Variable analysis . . . 21
4.4.1 SHAP . . . 21
4.4.2 LASSO . . . 22
4.5 Evaluation metrics . . . 23
4.5.1 AUC score . . . 23
4.5.2 Brier score . . . 23

5 Results and discussion **24**

5.1 Model performance . . . 24
5.1.1 XGBoost . . . 24
5.1.2 Logistic regression . . . 26
5.2 Variable analysis . . . 27
5.2.1 SHAP . . . 27
5.2.2 LASSO . . . 32
5.3 Valuable customers . . . 35

6 Conclusion **41**

6.1 Discoveries and implications . . . 41
6.2 Strengths and weaknesses . . . 43
6.3 Recommendations for future studies . . . 44

References **46**

Appendix **52**

A Appendix **52**

A Housing price index 52

B Definition of variables 53

C Descriptive statistics of variables 54

D Visual and descriptive representation of the folds 55

E Hyperparameter XGBoost 56

F Visualisation of churn rate by age 57

B Appendix **58**

A SHAP Beeswarm plots of folds 1 - 8 58

List of Figures

1 External and macroeconomic factors 10

2 Beeswarm plot of top 10% customers 35

3 Waterfall plot of a churned 28-year-old customer 38

4 Waterfall plot of churned 58-year-old customer 39

5 External factors - Trend 52

6 Churn rate by age pre-dataprocessing 57

7 Fold 1: Beeswarm Plot 58

8 Fold 2: Beeswarm Plot 59

9 Fold 3: Beeswarm Plot 59

10 Fold 4: Beeswarm Plot 60

11 Fold 5: Beeswarm Plot 60

12 Fold 6: Beeswarm Plot 61

13 Fold 7: Beeswarm Plot 61

14 Fold 8: Beeswarm Plot 62

List of Tables

1 Description of data reduction 12

2 Hyperparameter table description 20

3 Performance of XGBoost models 24

4 Performance of LR models 26

5 Normalised SHAP values 28

6 Logistic regression coefficients 32

7 Definition of variables 53

8 Descriptive statistics 54

9 Visual representation of the folds 55

10 Descriptive statistics of the folds 55

11 Tuned hyperparameters 56

1 Introduction

The Norwegian banking industry has undergone a significant change over the past 20 years, causing some of the leading financial institutions to undergo major digital transformation. So much so that the consumer council in Norway had to impose a minimum offer for non-digital consumers (Hålien, 2022).

Piccinini et al. (2015) found that consumers now expect better quality information, faster responses and more interactive ways of communicating. This is due to the increased use of digital technologies, which has resulted in the producer-consumer relationship becoming more consumer-centric. The rapid growth of digital transformation has brought with it new business strategies and methods to enhance customer relationships, such as Customer Relationship Management systems (CRM), chatbots and more. CRM can often be defined as the process of planning, monitoring and executing the company's relationship with customers and distributors through the use of technologies, data and strategies.

The topic of churn rates has been around for a while. Both old and new ways of market strategies have had the benefit of focusing on churn rates. Digital transformation has distanced the physical interaction with customers. Sharma and Panigrahi (2011) refer to churning as a customer leaving one company for another. "Churn" is also referred to as a customer who terminates their relationship with a service provider, such as a retail bank. When a customer churns, the relationship with the company is ended. There can be multiple reasons for churn to take place, such as better alternatives or lack of satisfaction with the current service provider.

Acquiring new customers and retaining old ones is important in retail banking. Studies show that there is a higher cost in acquiring new customers than retaining old customers, which results in a substantial impact on a bank's profitability (Hundre et al., 2013; Reichheld & Kenny, 1990; Van den Poel & Larivière, 2004). Reichheld and Kenny (1990) show that a customer retention rate of 5% yielded higher margins and faster growth, as well as greater control within their own branches. This "minor" improvement in customer retention could lead to profit swings between 25% and 80% profit. Hence, churn prediction models are highly relevant in retail banking.

One of the innovative insights this thesis offers is the use of external and macroeconomic variables for churn prediction. We investigate the following external and macroeconomic variables, suggested by previous literature to influence society and consumer behaviour in multiple ways; Average Interest Rate across banks, Housing Price Index, Unemployment rate, Consumer price index, and Policy rate. The effect of these variables can be either directly or indirectly affecting people, such as loss of job for oneself, or by inflation reducing purchasing power (Aursand, 2022; Burda & Wyplosz, 2013, p.5-7). We have yet to see any of these variables being implemented into a churn prediction and characteristic analysis within banks. We therefore see it as highly relevant to include such financial data that can inflict the banks themselves, as well as consumers directly and indirectly.

There has been little research done around the prediction and emphasis on the valuable customers themselves and rather on predicting churn in general (Lemmens & Gupta, 2020). Hence, we will further enhance this thesis with the new key insight of focusing on valuable customers and their characteristics. Valuable customers are scored based on the probability of defaulting (PD) on payments or loans, as well as Loan to Value (LTV), which is the loan sum divided by market asset value. This creates an understanding of the risk customers bring to the bank and their value.

As of Spring 2023, an added benefit of researching this now and emphasising on external and macroeconomic factors, is the ongoing financial turbulence in the world economics (Almås et al., 2023). It is, therefore, essential for businesses today to assess their customers, behavioural patterns and economics, as well as big data, in combination to maintain their competitiveness.

In this thesis, we study the financial industry of banking and predict the customer churn through the use of a collaborative Norwegian banks' data, as well as external and macroeconomic data. Further, compared to other studies, we contribute to the literature by not only investigating the characteristics of customers who churn in general but also focusing on the characteristics of customers who are most valuable for banks. We seek to find indicators that can be of good value to the collaborative bank for them to enhance their business strategies and counter-act possible churn of their customers.

Research has been done within the field of churn rates, CRM data, customer surveying and product and marketing. It is therefore important to clarify that this thesis will be delimited to the financial data of customers in a bank as well as external and macroeconomic data.

1.1 Research questions

With opportunities to predict certain outcomes, and gain a broader understanding of the customers through the data businesses possess, we see it as relevant to investigate the following case:

- What characteristics in loan customers can influence the likelihood of their churn?

To answer our problem, we use Extreme Gradient Boosting (XGBoost), SHapley Additive exPlanations (SHAP), Logistic Regression (LR) and the Least Absolute Shrinkage and Selection Operator (LASSO). XGBoost and LR are both machine learning (ML) methods used to help us predict future churn. LASSO performs variable selection and regularisation to the prediction model. By using SHAP, we can look deeper into the different variables and their effect on the model itself.

In addition to the main problem, we also want to answer the following research questions:

- Can we predict the churn of customers based on machine learning models?
- Do external and macroeconomic factors affect the churn?
- Are the characteristics that influence the likelihood of churn different for the most valuable customers compared to other customers?

We have chosen these research questions to help get a deeper understanding of the problem itself. By understanding the characteristics of the customers who leave and whether macroeconomic factors affect the churn rates, we can gain a deeper

understanding of the predictions made. This can give the banks the opportunity to be aware of potential customer churn and whether there are highs or lows in the economies.

1.2 Structure of the thesis

The rest of the thesis is organised as follows. Chapter 2 presents different previous literature that is central to our field of study. Chapter 3 explains the data and the pre-processing of it. It will also explain the different variables we have created and why. Furthermore, Chapter 4 introduces the methodology.

We present and discuss the results in Chapter 5 before we finally conclude our research and findings in Chapter 6, as well as provide suggestions for further research and analysis.

2 Literature

2.1 Customer churn and factors

Customer churn prediction and the understanding of the characteristics of customers who churn is the first and one of the most important steps in retaining customers. Customer churn can be observed in two groups, where one group is defined as voluntary churners, and the other is non-voluntary churners. Non-voluntary churners are forced to leave a company as a result of problems with payments or violation of contractual terms. Voluntary churners are more challenging to recognise due to their conscious decision to leave a company (Hadden et al., 2007).

Reichheld and Kenny (1990) described how customer retention is seen as a leading business strategy for long-term profits and balance growth. Retaining customers over a longer period of time offers a reduction in costs as selling to a new customer can cost between five to ten times more than to an existing customer. A customer who has been with the bank for five years is more valuable than a new customer. The same can be said of a customer with a 10-year relationship being more valuable than a five-year relationship customer. Dawkins and Reichheld (1990) uncovered through the consultancy business, Bain & Co. that companies lose on average 15% to 20% of their customers each year. They state that customer satisfaction surveys and the banks' increased commitment to service are not enough but that retention should instead be analysed and tracked. The customer retention rate gives performance feedback whilst customer defection analysis gives an explanation, making for a combination of analytical tools to be used in order to understand and improve the performance of customers' defections (Dawkins & Reichheld, 1990).

Lemmens and Gupta (2020) described multiple issues with churn prediction models which try to minimise misclassifications of all customers who are likely to churn, rather than minimise churn of profitable customers. A business with a wide customer base have both profitable and costly customers. To enhance profits using the churn prediction model, it becomes crucial to focus on targeting the profitable customer segment. Gorgoglione and Panniello (2011) described the problem with

focusing on the accuracy of the churn prediction models rather than creating personalised actions to deal with customer defection, and further emphasise that a good prediction model should give managers the opportunity to take appropriate actions to prevent customer churn. Both Gorgoglione and Panniello (2011) and Lemmens and Gupta (2020) described important issues with customer churn models. For a business, it is important to focus the prediction on its key customers, the ones who are most profitable, and additionally have the opportunity to take effective actions. Without implementing these two, the prediction models would be ineffective.

Lukas and Nöth (2019) studied the German mortgage loan market and found that borrowers are more likely to search for additional offers when interest rates rise and are less likely to search for additional offers when the interest rates fall. The difference in interest rates for borrowings is likely to be minimal, but customers may be concerned that their financial situation could be impacted by differences in interest rates (Levesque & McDougall, 1996). Ongena et al. (2021) found that the average price differs between new customers and present customers, where the new customers on average get offered 26 basis points lower price. They also found that the difference between interest rates offered at initiation has changed over time, with a decrease in difference of interest rates offered at initiation. The interest rates increase in difference after initiation in comparable loan products, which could be due to a more transparent lending market. Additionally, they emphasise that as customers' age increases, and how long they have been a customer with the bank lengthens, their probability of staying with the bank also increases. This indicates that older customers appear to be loyal regardless of the bank's increase in interest rates. Bilal Zorić (2016) highlights that banks should adopt a customer-facing strategy, focusing on the long term and acknowledging that young customers may initially be unprofitable in their early stages but become more profitable later.

Van den Poel and Larivière (2004) studied customer attrition in a European financial services company that offers banking and insurance services. They found that the probability that customers change their financial service provider change over time depending on their relationship duration with their current service provider. Individuals had a higher probability of switching service providers earlier in their

relationship with the current one, but after seven years, the probability of switching stabilises for about 15 years. After 20 years, the probability of switching increases rapidly before it stabilises again around 40 years. According to Van den Poel and Larivière (2004), there are two critical periods of a customer's lifecycle, one in the early years of becoming a customer, and one after a period of 20 years.

In the study conducted by Ongena et al. (2021) they found that higher credit risk, measured by PD, was a contributor to a higher price difference with the customers. Additionally, they found that higher values of PD were a contributor to a higher probability of switching banks. According to Lessmann et al. (2015), PD has been widely researched and used for credit scoring of customers and applied to predictive models that estimate the likelihood of a customer exhibiting undesirable actions in the future. In the context of customers' choice of a bank, the relevance of location can be considered essential. Levesque and McDougall (1996) mentioned that Thwaites and Vere (1995), Laroche and Taylor (1988) and Anderson et al. (1976) found location to be an important determinant in a customer's choice of bank, and that convenience and accessibility are important elements that contribute to an overall customer experience.

2.2 External and macroeconomic factors

Van der Drift et al. (2023) discussed that decreasing interest rates has been an important factor in the increase of house prices over the past two decades. With higher house prices, current homeowners are enabled to allocate more financial resources towards the acquisition of their next home. Houses are commonly acquired through a mortgage, and it could be argued that mortgage payments play a crucial role as a determining factor in the decision-making process, as opposed to the price of the house itself. A decrease in the interest rate would automatically reduce the expenses associated with the mortgage, and make it possible to increase the mortgage while keeping their payments unchanged. Basten and Koch (2015) studied the effect of house prices on mortgage demand and supply in Switzerland. They found that higher housing prices resulted in a growth in mortgage demand. For refinancing

loans, they found that the impact of house prices is arguably less significant for refinancing requests compared to new mortgages.

Policy rates influence both the interbank interest rates and the interest rates offered to customers for loans and deposits (Norgesbank, n.d.-a). Koeniger et al. (2022) studied how the transmission of monetary policy can affect the housing market in Germany, Italy and Switzerland. They discovered that in the occurrence of favourable monetary policy shocks, the housing market tends to increase the number of transitions from renting a home to buying a home, both in the short term and a long-lasting period. They emphasise that the attributes of the mortgage market and the availability of public housing influence the transmission of monetary policy to the housing market.

The unemployment rate can affect society as high unemployment rates can cause rifts, insecurity and compassion for others within the society (Burda & Wyplosz, 2013, p.7). Foote et al., 2010 studied the borrower's decision to default and found that increased unemployment rates contribute to increased default in mortgages, as a larger number of individuals are at risk of losing their jobs and becoming financially restricted during economic downturns. They also found that declining housing prices contributed to an increase in defaults. Negative changes in income can lead to negative equity. Consequently, they become unable to sell their home for a sufficient amount to repay the mortgage.

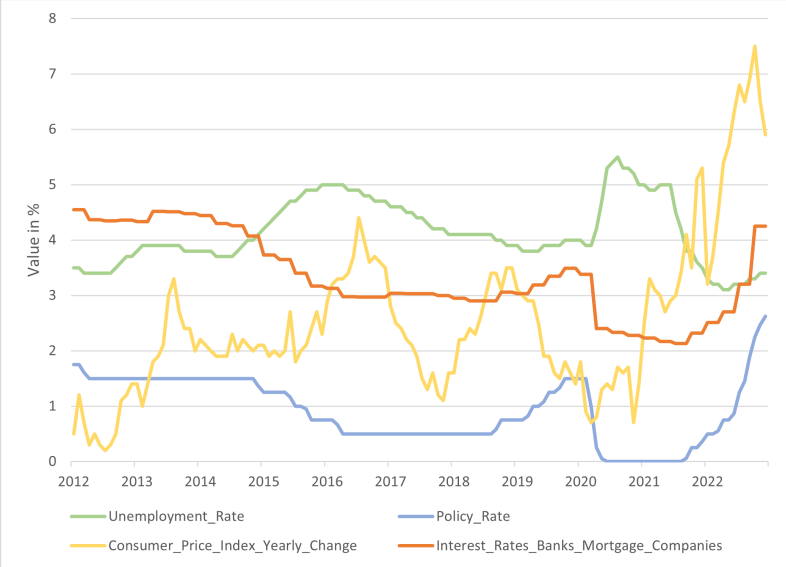
The consumer price index is a unit of measurement on inflation. Inflation is when prices of goods and services increase over time. High inflation is often seen as a negative thing, especially when the wages do not increase in line with the inflation. High inflation can cause a weaker purchasing power per individual and less predictability, making them more strict on their income and outcomes (Aursand, 2022). Armantier et al. (2015) studied expected inflation and behaviour and discussed how it is necessary for households to consider expected inflation when making decisions regarding mortgage financing or refinancing, and debt managing. Their study did not indicate that households made their decision based on expected inflation and argued that studying this specific subject is challenging due to the influence of other factors on the decision-making.

3 Data

3.1 Data description

Our data contains monthly information about mortgage customers given by our collaborative bank from January 2010 to February 2023, including information on loans, interest rates, and demographics. In total, there are 1 680 159 monthly observations of 29 088 customers. In our analyses, we treat each month and each customer as a separate observation. That is, we use customer-month observations. The number of customer-month observations for each customer varies based on the duration of their relationship with the bank. As a result, there are more customer-month observations for customers who have been a customer with the bank for a longer period. The dataset is expanded to contain external and macroeconomic data that is considered relevant for our thesis, this includes unemployment rates, housing price indexes, consumer price index, policy rate (monthly average) and interest rates for a selection of banks and credit companies. The external and macroeconomic data was collected from the following organisations in Norway; Statistics Norway, the central bank of Norway, and Real Estate Norway (Eiendomnorge, n.d.; Norgesbank, n.d.-b; SSB, 2023a, 2023b, 2023c). We merged the dataset with the external and macroeconomic data and aligned each observation with the corresponding month in the original dataset. The final dataset is a combination of both internal bank data and external and macroeconomic data, which makes it unique.

Figure 1: External and macroeconomic factors



The above figure presents the macroeconomic trends in Norway for the chosen variables, from 2012 to 2022.

3.2 Variables

The dataset used in our models contains 42 variables, where 37 are associated with the customer, that is economic, demographic and geographic variables, and 5 are external and macroeconomic variables. Table 7 in Appendix A describes each variable in the dataset. Further descriptive statistics and visual representations can be found in Appendix A.

3.2.1 Target variable

In our dataset, customer churn is defined as the fulfilment of a customer's loan, specifically when the loan balance reaches zero. Even if the loan balance reaches zero, customers who have other unrelated products are still considered churned. The target variable in our dataset is the dummy variable *"Churned_6_Lag"*, which is 1 if the customer has churned six months ago and 0 otherwise. There are two reasons why we introduce the six months lag. First, the last monthly observations of churned customers have many missing and inaccurate values, the reason for this is that data has been recorded both during and after a customer has churned. Second,

we believe that the customers' decision to leave a bank is not made in the same month as the actual departure, but rather some months earlier. For instance, (Levesque & McDougall, 1996) found that the decision to switch banks was often driven by problems with satisfaction that occurred in the past six months.

3.3 Data pre-processing

To ensure that we have a suitable dataset for analysing customer churn in retail banking, we had to pre-process the data. This involved transforming data that was not in a proper format, and also cleaning and removing data points that were either missing or problematic for the model.

3.3.1 Data cleaning

The dataset contained observations that were seen as unrealistic and problematic for the analysis. We had to remove these observations to get a realistic dataset. Since the dataset did not contain too many variables and all of the variables were interpretable as raw data points, we concluded that a reasonable approach was to set limits for each separate variable. This meant carefully observing each variable and choosing a reliable maximum and minimum value. Table 1 describes the removal of observations.

Table 1: Description of data reduction

	Observations
Original dataset	1 680 159
Less: Year earlier than 2012 and later than 2022	1 439 532
Less: Weighted_Average_Interest_Rate less than 0 and larger than 0.1	1 418 295
Less: Repayment_Loan_Sum less than 0	1 417 618
Less: Boligkreditt_Balance less than 0	1 408 107
Less: LTV less than 0 and larger than 1	1 401 866
Less: Age younger than 18 and older than 78	1 297 971
Less: Seniorlaan_Balance less than 0	1 297 967
Less: Other_External_Debt less than 0 and larger than 50 000 000	1 297 937
Less: Other_Internal_Debt less than 0	1 297 924
Less: Observations after churn	1 235 900
Data sample (pre-missing value removal)	1 235 900

The above table presents the gradual process of removing observations in the dataset. The numbers represented on the right is the amount of observations remaining after cleaning the given variable on the left.

Observations before 2012 were incomplete and had multiple missing values, we removed observations that were before 2012 and after 2022. For the variable *"Weighted_Average_Interest_Rate"* we excluded data points outside the range of 0 and 0.1 since they were considered outliers. Variables that had information about loan values that were either very high or below 0 were removed. Values outside the range of 0 to 1 for *"LTV"* were considered outliers and were removed. Furthermore, we removed observations where the value of *"Age"* is 17 years or younger, which we assume is incorrectly registered data. This is because the bank is legally prohibited from granting loans to individuals under 18 years old. Based on our preliminary analysis of churn rate distributed by age found in Figure 6 Appendix A, we excluded observations of individuals older than 78. We concluded that the rise in churn rate among individuals after reaching the age of 78 can be attributed to natural causes, such as death. Finally, to avoid bias in the model, we removed the remaining observations after the registered value 1 in *"Churned_6_Lag"*.

In addition to the reduction of observations, we had to handle missing values in the dataset. Some missing values were replaced, while others were removed. *"Weighted_Average_Interest_Rate"*, *"Repayment_Plan"*, *"Repayment_Loan_Balance"*, *"Boligkreditt_Balance"*, *"Boligkreditt_Credit_Limit"* and *"Seniorlaan_Balance"* had an explanation for the missing values, due to the fact that repayment loan, boligkreditt and seniorlån are three different loan products offered by the bank, where some customers may have one or all of these products. This can lead to missing values for some of the variables in each observation if the customer does not have all three products. Therefore, we had to give each missing data point a value of 0 if either of the other products were active. We clarified with the bank that *"Weighted_Average_Interest_Rate"* was only related to boligkreditt and repayment loan, and not with seniorlaan. We replaced the missing values of *"Weighted_Average_Interest_Rate"* with 0 if the customer only had seniorlaan as a loan product. The missing values of the variables *"Other_Internal_Debt"* and *"Other_External_Debt"* were concluded to be missing due to the fact that the customers did not have any other debt, and therefore the missing value was set to 0. For variables associated with county, the process of assigning each customer a variable within county resulted in some missing values. This was due to misspellings or missing values in the original dataset. The missing values for county were given the value *"Other_County"*.

We had four variables, *"PD"*, *"LTV"*, *"Size_Household"*, and *"Income"*, where we had a large number of missing values that had to be handled in a proper way. Ringdal (2013, p.262) argues that a valid approach to deal with missing values is to replace missing values with a mean value attached to other observations, where another variable has a strong coherence with the observation that is missing. In this context, missing values of *"Income"* could have a strong coherence with education, but since our dataset doesn't include education and there are no other obvious variables with a strong coherence, we decided to not replace values with a mean based on other observations. If we were to replace values with a mean, the effect of the models could make it non-representative in terms of analysing customer churn. To fix the problem with missing values we isolated each customer based on a customer ID, and then filled in missing values for *"Size_Household"* and *"Income"* based on the mean for

the specific customer. For *"PD"* and *"LTV"* we did the same approach, but instead of using mean, we used a forward and a backward fill technique. For a forward fill, this will imply that a missing value would be filled in with the last non-missing value in the series of the data, and for a backward fill, this will imply that a missing value would be filled in with the first non-missing value in a series of data. We applied the forward fill before the backward fill. With the approach of handling missing values based on the isolated customer, we avoid filling in missing values that are based on other customers' observations, which in fact could be unrealistic and not valid. It is important to notice that this was done in the modelling of the data, where we separated the training set and the testing set. So the filling of missing values for the isolated customer was done separately for the training set and the testing set. This was to avoid data leakage between training and testing.

3.3.2 Transformation of variables

The dataset contains multiple variables that have information which is categorised as True/False, Yes/No and other two-category information such as Female/Male. We transformed these variables to become dummy variables of 1 and 0. We created new columns for every *"County"* and added the value 1 or 0 for the customer if their residence was located in the respective *"County"*. We also created the dummy variables *"Credit_Card_Dummy"*, *"Sex_Dummy"*, and *"Deposit_Dummy"*. The three dummy variables *"Age_group1"*, *"Age_group2"* and *"Age_group3"* were also created and assigned the value 1 if the customers' age (given by variable "Age") was 18-39, 40-64, and above 65, respectively, years.

We also created two new variables based on other variables that we had in our dataset. The variable Dept-to-income (*"DTI"*), was created by summing variables that included loan balance and dividing it by *"Income"*. The variable *"PD_LTV"* was created by multiplying *"PD"* and *"LTV*. To minimise the effect of extreme values, we log-transformed the variables *"Age"*, *"Customer_Relationship_Duration"*, *"Income"*, *"Repayment_Loan_Sum"*, *"Boligkreditt_Balance"*, *"Boligkreditt_Credit_Limit"*, *"Seniorlaan_Balance"*, *"Other_Internal_Debt"* and *"Other_External_Debt"*.

3.4 Final dataset

After data processing, our final dataset contains a total of 1 038 181 observations and 42 variables. The total of churned customers is 10 941. A description of each variable can be found in Appendix A, specifically Table 7.

4 Method

Deciding what method to use for churn prediction and why, is crucial for the value and validity of the study. Different methods can give different results for a given dataset. To further enhance our results, we have performed a Walk-forward testing approach and tuned parameters to counteract under-and overfitting. In this chapter, we present methods used in our study and how they work, as well as how we performed the Walk-forward testing and tuning of parameters.

4.1 Data sampling and Walk-forward testing

Complex prediction models can suffer from overfitting as they tend to memorise the training set rather than learning. This results in a high accuracy of the training set, but a noticeably lower accuracy on the test set. Applying this to the real world can lead to weaker predictive performance. Validation can help get better performance on the estimated model with the use of observations that are not included in training (Kirkos, 2015). James et al. (2021, p.181) describes the use of k-Fold Cross-Validation as a validation technique, where the data is separated into k groups, the model is fitted on the k-1 groups, and the first group is used for validation. This continues for k times, where a new validation set is used each time.

To address the issue of overfitting in our model, we needed to consider the time aspect within our dataset. We use an out-of-sample approach where future observations will be held out-of-sample. The use of ordinary k-Fold Cross-Validation technique would not be suitable for our dataset, because it would be considered cheating as it uses future observations to predict the past. To avoid the problem with overfitting in our model, we have applied a Walk-forward testing approach. This works by moving a fixed set of training and test sets periodically by each fold, with a chosen number of folds applied (Börjesson & Singull, 2020). We have split the Walk-forward testing into an 8-Fold with a 3-year training and 1-year test starting from 2012. A visual representation of the folds and observations can be found in Table 9 and Table 10.

We have also had to use a different approach to answer the following research question "Are the characteristics that influence the likelihood of churn different for the most valuable customers compared to other customers?" In order to achieve this, we created a proxy for identifying valuable customers within the bank, based on their credit risk. *"PD"* and *"LTV"* were multiplied to obtain the variable *"PD_LTV"*, which is an indicator of how risky a customer might be, in terms of the probability of default, weighted on the loan-to-value. We filtered out the top 10% of customers according to *"PD_LTV"*, where a lower value of *"PD_LTV"* would be seen as a "valuable" customer. For this approach, we conducted training and testing on the identical sample, with the purpose of finding what characteristics that contribute to the likelihood of churn among the top 10% of customers.

4.1.1 Imbalanced data

Classification of a binary target variable such as churn may involve imbalanced data. Customer-month observations classified as churned are in the minority in relation to customers classified as not churned. López et al. (2013) refers to this situation as the "class imbalance problem" and points out that it is usually the minority class that is the most important to be learned from. Neslin et al. (2006) describes that when managing customer churn, it is important to predict the most likely scenario for each customer in terms of churning. By identifying which customers who are more likely to churn, rather than the ones who are not, it is possible to offer targeted incentives to retain customers and save money. This assumes that the prediction can be done with an acceptable accuracy. In our dataset, each customer has multiple observations depending on how long they have been a customer with the bank. This means that churning is a highly rare scenario for each observation in the dataset. Resampling techniques such as oversampling can be used to deal with imbalanced data in churn prediction, simply by taking the observations of the minority class used for training the model and adding it to the testing of the model to increase the share of the minority class used for testing (Verbeke et al., 2012).

We chose not to make use of a resampling approach presented by Verbeke et al. (2012). The use of oversampling techniques to create a more balanced dataset could

possibly harm and create bias in the model. It would therefore, not be suitable for a future prediction such as our thesis. If a prediction model is trained on a balanced dataset, it is trained to believe that reality is different from the unbalanced dataset (test set) it is supposed to predict. The test set cannot be balanced because the bank does not know in advance who will "churn" or not.

4.2 XGBoost

XGBoost is a scalable end-to-end tree boosting system introduced by Chen and Guestrin (2016) and is based on the ML technique introduced by Friedman (2001) called Gradient boosting. The main difference between XGBoost and Gradient Boosting is that XGBoost uses a more regularised formalisation to prevent and control over-fitting, which makes it more flexible. XGBoost not only captures non-linear relationships but also utilises an ensemble approach to iteratively enhance its performance through training on the data (Wahlstrøm, 2023). XGBoost is described by Wahlstrøm (2023) as the following:

$$\arg\min_{f_j(\mathbf{X})} \sum_{n=1}^{\psi N} \left[l(y^n, p(f(\mathbf{x}^n)^{[j-1]})) + g_n f_j(\mathbf{x}^n) + \frac{1}{2} h_n f_j^2(\mathbf{x}^n) \right] + (\gamma T_j + \frac{1}{2}\lambda \|w_j\|^2) \tag{1}$$

In XGBoost, we consider a training set denoted as $\{\mathbf{y}, \mathbf{X}\}$, where $\mathbf{X} = \{x_i^n\}_{n=1,...,N, i=1,...,I}$ is the matrix with the values of I variables for N customer-months. $\mathbf{x}^n$ is the vector of variable values for customer-month n, and $\mathbf{y} = \{y^n\}_{n=1,...,N} \in \{0,1\}^N$ represent the actual classifications of non-churn (0) or churn (1) for all the N customer-months. This approach employs an ensemble method that undergoes iterative training to gradually enhance its performance with respect to the data. In each subsequent iteration j, a weak learner defined by $f_j(\mathbf{X})$ is obtained by minimising the residuals of the XGBoost model $f(\mathbf{X})^{[j-1]}$ from the preceding iteration j-1 over ψN customer-months chosen randomly for each iteration. Here $\psi \in (0, 1]$ represents the ratio of the subsample.

The logistic loss function is given by

$$l(y^n, p(f(\mathbf{x}^n)^{[j-1]})) \tag{2}$$

The gradient statistics for both the first and second order are expressed as

$$g_n = \delta_{f(\mathbf{x}^n)^{[j-1]}} l(y^n, p(f(\mathbf{x}^n)^{[j-1]})) \tag{3}$$

$$h_n = \delta^2_{f(\mathbf{x}^n)^{[j-1]}} l(y^n, p(f(\mathbf{x}^n)^{[j-1]})) \tag{4}$$

To help prevent overfitting, the model contrains the minimisation to favour simpler trees by using a penalised decision tree as a weak learner, given by the penalty term:

$$\Omega(f_j(\mathbf{X})) = \gamma T_j + \frac{1}{2}\lambda \|w_j\|^2 \tag{5}$$

where γ and λ are penalty parameters and the number of tree leaves are defined by T_j whereas the weight of the leaves are determined by the magnitude $\|w_j\|^2$.

After the weak learner $f(\mathbf{X})^{[j]}$ is obtained, the contribution of $f(\mathbf{X})^{[j]}$ is integrated into the model through

$$f(\mathbf{X})^{[j]} = f(\mathbf{X})^{[j-1]} + v f_j(\mathbf{X}) \tag{6}$$

where $\nu \in (0, 1]$ is the convergence speed, ensuring that the model introduces only minor modifications during each iteration. In summary, the XGBoost model iteratively decreases the residuals over time.

The scalability, speed and real-life use case of XGBoost makes the model a suitable choice for our thesis. However, it is important to be aware that Gradient Boosting and decision tree models, which XGBoost is based on, are considered to be black-box. These models go over multiple different trees making the predictions less explainable and harder to comprehend for the end user (Sagi & Rokach, 2021).

4.2.1 Hyperparameter tuning - XGBoost

XGBoost allows for several hyperparameters which are determined in advance to achieve a better result and performance. By performing tuning on the training data

and not on the test data, we avoid data leakage and gain a better out-of-sample prediction performance. In order to optimise the model output and reduce overfitting, we chose a Bayesian optimisation that helps avoid manually tuning and testing each parameter. Bayesian hyperparameter optimisation has proven to outperform other optimisation techniques such as manual search, random search, and grid search (Snoek et al., 2012; Xia et al., 2017). The parameters tuned through Bayesian optimisation are given in Table 2, and their respective values can be found in Table 11.

Table 2: Hyperparameter table description

Hyperparameter	**Description**	**Type**	**Search space**
$\boldsymbol{\nu}$	Shrink feature weights to prevent overfitting	decimal	[0.05, 0.3]
$\boldsymbol{\psi}$	The fraction of training instances used for each iteration	decimal	[0.5, 1]
$\boldsymbol{\gamma}$	Minimum loss reduction required for a split in tree leaves	decimal	[0, 1]
$\boldsymbol{J}$	Number of iterations	integer	[30, 100]
$\boldsymbol{w}$	Minimum sum of instance weight needed in a child	integer	[1, 10]
$\boldsymbol{\lambda}$	Controls the weight of the L2 regularization to prevent overfitting	integer	[1, 5]
max_depth	Maximum depth of a tree	integer	[2, 6]

The above table gives a short description of the different parameters used to tune our XGBoost model. The last column presents values and the interval used for the parameters.

4.3 Logistic regression

LR builds on the fundamental ideas of linear regression analysis. Compared to other more complex ML models that are seen as black boxes, LR is considered a more interpretable model. LR operates on the probability of a target binary variable of 0 or 1 as the response. In our case we have the binary variable of churn being $Y = 1$ and not churned being $Y = 0$. The predicted probabilities $\hat{y}$ are given by the following formula from (Paraschiv et al., 2023):

$$\hat{y} = \iota \oslash (\iota + \exp(-\mathbf{X}\mathbf{w} - \iota w_0)) \tag{7}$$

where $\mathbf{X}$ is a matrix of values for each variable $i = 1, ..., I$ obtained from the dataset

of N customer-months. $\mathbf{w}$ is a vector that contains the coefficients, w_0 is the intercept coefficient in the model, $\oslash$ represents the Hadamard (element-wise) division, and ι consists of a $N \times 1$ vector of ones. We estimate the coefficients $\mathbf{w}$ and w_0 by minimising the negative of the log-likelihood function given by (Paraschiv et al., 2023):

$$\ell(\mathbf{w}, w_0) = \sum_{n=1}^{N} [y \odot (\mathbf{X}\mathbf{w} + \iota w_0) - \log(\iota + \exp(\mathbf{X}\mathbf{w} + \iota w_0))] \tag{8}$$

The binary classifications of the model are explained by y, which is a vector of actual classifications. The customer-month observations are given by $n = 1, ..., N$, and $\odot$ represents Hadamard (element-wise) product. In addition, we calculate the significance of each estimated LR coefficient by using Wald statistic (Hosmer et al., 2013).

4.4 Variable analysis

4.4.1 SHAP

In order to accommodate the black-box issue and get a more comprehensive understanding of the different variables and their effect on the XGBoost model, we use SHAP (Lundberg et al., 2019). SHAP is derived from the Shapley (1953) values, and game theory. It is used to explain the prediction of the model through the computation of each variable (Molnar, 2022, chap.9.6). Game theory is a mathematical solution to behaviour and decision-making. The Shapley values calculate all contributions from each player through all existing coalitions of players. The Shapley values are obtained through the following formula presented by Aas et al. (2021):

$$\phi_i(v) = \phi_i = \sum_{S \subseteq \mathbf{x}^n_{/i}} \frac{|S|!(I - |S| - 1)!}{I!} \left(f(S \cup \{x_i^n\}) - f(S)\right), \quad i = 1, \ldots, I \tag{9}$$

This is explained as the weighted amount that player i gets by their contribution differences for all subsets not including player i. $S \subseteq \mathbf{x}^n = \{1, ..., I\}$ is a subset containing $|S|$ players. f is a gain function and $f(S)$ is the given contribution

function explaining the aggregate expected value attainable by the players in S through cooperation. $\big(f(S \cup \{x_i^n\}) - f(S)\big)$ explains each player's demand for a fair compensation equal to their contribution within the coalition.

Aas et al. (2021) also explains how the Shapley game theory can be implemented with an ML approach, where the individual variable can be explained as the player, and the prediction can be explained as the amount the player gets. The prediction is explained by the following function:

$$f(\mathbf{x}^n) = \phi_0 + \sum_{i=1}^{I} \phi_i^n \tag{10}$$

where $i = 1, ..., I$ are the predictors. This formula explains the difference between the Shapley values prediction $y^n = f(\mathbf{x}^n)$, and the global average prediction. $\phi_0 = f(\emptyset) = \mathrm{E}[f(\mathbf{X})]$ and is the fixed value which is independent of the actions taken by any of the predictors, and is often zero. ϕ_i^n is the SHAP value of the predictor i for observation n (Aas et al., 2021; Wahlstrøm, 2023).

To visualise the SHAP values we use a beeswarm and waterfall plot obtained with the use of the Python package "SHAP", created by Lundberg (2018). Wahlstrøm (2023) explains how the values provided in the different plots provide the ability to determine the weight and direction of the predictors' influence on the predictions.

4.4.2 LASSO

LASSO was introduced by Tibshirani (1996) and is used to minimise the residual sum of squares which allows for greater variable selection. LASSO proceeds to remove the explanatory variables which do not have an effect on the model, before estimating the LR model (Paraschiv et al., 2023). LASSO is presented by Paraschiv et al. (2023) as following:

$$-\ell(\mathbf{w}, w_0) + \lambda \|\mathbf{w}\|_1 \tag{11}$$

where $\ell(\mathbf{w}, w_0)$ is the log-likelihood which is previously presented in equation (8).

Higher λ values indicate a more important and stronger predictive ability per variable that is added to the model as it is a positive tuning parameter. $\|\mathbf{w}\|_1$ is the l_1-norm of $\mathbf{w}$ which combined with λ makes for the penalty term $\lambda\|\mathbf{w}\|_1$ (Paraschiv et al., 2023). Our LASSO regularisation uses a λ-value of 20.

4.5 Evaluation metrics

4.5.1 AUC score

To explain the predictive ability of our model, we use what is called Area Under the Curve (AUC) which is a summary of Receiver Operating Characteristic (ROC). ROC is a commonly used metric that represents the relationship between the true positive rate and the false positive rate across various discrimination thresholds. AUC serves as a comprehensive metric for evaluating binary classification and is a popular visualisation of discriminating threshold between the binary variable. $AUC = 1$ gives a perfect classification, whilst an $AUC = 0.5$ will be an average classification, making the prediction random. This means that a higher AUC value indicates a stronger predictive ability for the model (Kvamme et al., 2018).

Hosmer et al. (2013, p.177) mentions a rule of thumb when looking at the AUC score. AUC between 0.5 and 0.7 is considered poor and close to random. AUC between 0.7 and 0.8 is acceptable, AUC between 0.8 and 0.9 is excellent and AUC above 0.9 is considered outstanding.

4.5.2 Brier score

We also evaluate our model based on the Brier (1950) score. Paraschiv et al. (2023) mentions how we can think of the Brier score as a cost function that quantifies the average squared difference between the predicted probabilities assigned to potential outcomes by a set of predictors and the actual outcome, making it a loss function. Consequently, a lower Brier score indicates a model with better predictive power. Brier score has a value between 0 and 1.

5 Results and discussion

5.1 Model performance

In this chapter, we will be presenting the predictive power of our ML models. Furthermore, we present the XGBoost and LR prediction results and discuss the results against previous literature.

5.1.1 XGBoost

Our XGboost model is based on an 8-fold Walk-forward testing approach. This results in eight different prediction scores, as shown in Table 3, where Fold 8 is the closest relative prediction to the current date. The table contains train and test scores for both AUC and Brier. It also presents a difference in percentage between the train and test scores of AUC.

Table 3: Performance of XGBoost models

Fold	**Train_AUC**	**Test_AUC**	**Diff %**	**Train_Brier**	**Test_Brier**
1	0.6367	0.6334	-0.52 %	0.0101	0.0095
2	0.6318	0.6067	-3.97 %	0.0149	0.0151
3	0.6994	0.5676	-18.84 %	0.0101	0.0111
4	0.7291	0.5997	-17.75 %	0.0103	0.0124
5	0.6122	0.6172	0.82 %	0.0118	0.0121
6	0.6340	0.6268	-1.14 %	0.0117	0.0110
7	0.6618	0.6116	-7.58 %	0.0115	0.0099
8	0.6295	0.5602	-11.02 %	0.0110	0.0065
Mean	**0.6543**	**0.6029**	**-7.86 %**	**0.0114**	**0.0110**

The above table presents the XGBoost models AUC and Brier scores for each fold based on the train and test set. It also shows a difference in percentage between train AUC and test AUC.

Our performance results of the Brier score show an average of 0.0114 in train and a average score of 0.0110 in test. The predictive results given by AUC show a difference in the spread for all folds between -18.84% to 0.82%. This results in a total average difference between the predictive test and train score of -7.86%.

We can see a significant gap between train AUC and test AUC in certain folds shown in Table 3, more precisely fold 3, 4 and 8. With the help of Diff%, we can get an indication of just how big of a difference there is, as well as looking at the AUC scores of both train and test. These folds indicate a problem of overfitting as the train AUC is significantly larger than the test AUC.

Even as fold 3, 4 and 8 indicate overfitting, we note that the other folds have minor differences, and the average AUC scores and Diff%, make for little overfitting. This means that our use of the Walk-forward testing approach and Bayesian optimisation has helped mitigate overfitting. Bayesian optimisation adapts the chosen parameters given in Table 2 to each fold in our Walk-forward testing approach. The result of our testing and tuning is given in the average, which has a smaller difference in total, thus indicating little overfitting as a whole.

The models' average predictive performance is 0.654 for train and 0.603 for test, which could be considered poor and close to random, according to the rule of thumb described by Hosmer et al. (2013, p.177). This indicates that our XGBoost model has a low predictive ability. We can, therefore, not, predict with certainty the customer churn for our collaborative bank. Neslin et al. (2006) mentions how prediction can be possible with an acceptable accuracy, and taking Hosmer et al. (2013, p.177) rule of thumb into consideration, we can not say this is an acceptable accuracy. However, the AUC score is above 0.5 in train and test, which can indicate a minor potential for prediction.

5.1.2 Logistic regression

Table 4 presents the performance of our LR model. It introduces the AUC scores, the difference between AUC train and test, and the Brier score, split over the eight folds from the Walk-forward testing approach.

Table 4: Performance of LR models

Fold	Train_AUC	Test_AUC	Diff %	Train_Brier	Test_Brier
1	0.6103	0.5857	-4.20 %	0.0099	0.0092
2	0.6022	0.5845	-3.03 %	0.0102	0.0105
3	0.6120	0.5467	-11.95 %	0.0101	0.0111
4	0.6026	0.6000	-0.44 %	0.0104	0.0123
5	0.6035	0.6066	0.52 %	0.0113	0.0117
6	0.6079	0.6005	-1.22 %	0.0117	0.0110
7	0.6152	0.5956	-3.29 %	0.0115	0.0099
8	0.6141	0.6191	0.80 %	0.0108	0.0062
Mean	**0.6085**	**0.5924**	**-2.72 %**	**0.0107**	**0.0102**

The above table presents the LR models AUC and Brier scores for each fold based on the train and test set. It also shows a difference in percentage between train AUC and test AUC.

The average Brier score in our model is 0.0107 for train and a slightly smaller value of 0.0102 for test. The train AUC for our LR models has a low spread, with the highest value being 0.615 in fold 7 and the lowest being 0.602 in fold 2. Test AUC has a higher spread, where fold 3 has the lowest score of 0.546 and fold 8 has the highest score of 0.619. We see, based on the spread and Diff% of each fold, that the model does not seem to suffer from overfitting in most of the folds, except for fold 3. Fold 3 has a difference of -11.95%, which can indicate overfitting. This is similar to our XGBoost model performance which also struggles with overfitting in fold 3.

The difference between the average train AUC and the average test AUC is -2.72%. The total average train AUC has a score of 0.608, and test AUC has a score of 0.592. The LR model has minor challenges with overfitting compared to the XGBoost model and has even scores across the eight folds. Similarly to the XGBoost model, our LR models' predictive ability can be defined as poor and close to random, according to Hosmer et al. (2013, p.177).

Gorgoglione and Panniello (2011) emphasise that a good prediction model should give managers the opportunity to take appropriate actions to prevent customer churn. Based solely on the AUC and Brier scores given by the performance of our ML models in Table 3 and Table 4, we can not predict well enough to give the stakeholders, or management, of our collaborative bank, satisfactory predictions in order to take appropriate actions. We can, however, give indications of some predictions.

5.2 Variable analysis

To gain a better understanding of the black-boxes in our ML models, we present and analyse the SHAP and LASSO values with respective visualisations. The visualisations will also give us a more in-depth understanding of the different characteristics of each customer and the weight they have on our different models.

5.2.1 SHAP

Table 5 presents normalised SHAP values for each fold of the XGBoost model. The table shows the variable importance obtained from the SHAP values. In the table, each fold assigns a value between 0 and 100 to the variables, with the most important variable receiving a value of 100, while the remaining variables are assigned a percentage value relative to the most important variable. The model is colour graded to enhance the interpretability, where a darker colour indicates a higher score and thus higher variable importance. A mean column is added, which calculates the average score of each variable based on the scoring of each fold. The mean column

helps with getting an overall understanding of the importance of each variable in the model.

Table 5: Normalised SHAP values

	Variable	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Mean
Demographic	Size_Household	0.00	0.00	17.88	1.56	0.00	0.00	0.49	0.00	2.49
	Income_Log	0.00	0.00	25.24	8.58	1.33	0.32	2.46	0.00	4.74
	Age_Group_1	14.25	4.60	11.11	0.00	0.00	0.00	0.00	0.00	3.74
	Age_Group_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Age_Group_3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Age_Log	100.00	100.00	46.09	22.35	34.38	100.00	100.00	54.96	69.72
	Sex_Dummy	0.00	0.00	8.63	0.88	0.00	0.16	0.36	0.00	1.25
Loan-related	Repayment_Plan	0.00	0.00	77.30	9.59	0.00	0.01	1.51	0.00	11.05
	Weighted_Average_Interest_Rate	15.62	5.39	49.24	1.93	0.00	8.33	19.84	10.41	13.85
	LTV	4.03	1.12	62.54	10.08	0.01	4.65	76.64	24.79	22.98
	PD	20.50	11.58	100.00	100.00	100.00	20.87	47.75	100.00	62.59
	Credit_Card_Dummy	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01
	Deposit_Dummy	0.00	0.00	0.00	0.26	0.00	0.00	0.03	0.00	0.04
	Customer_Relationship_Duration	10.56	2.48	40.17	1.87	0.00	0.00	17.26	0.00	9.04
	Customer_Relationship_Duration_Log	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Repayment_Loan_Balance_Log	3.29	0.00	13.59	12.57	0.15	1.01	1.20	0.00	3.97
	Boligkreditt_Balance_Log	56.59	0.00	22.94	3.27	0.00	7.62	3.01	0.00	11.68
	Boligkreditt_Credit_Limit_Log	32.53	0.00	0.00	0.96	0.42	0.00	3.36	0.00	4.66
	Seniorlaan_Balance_Log	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Other_Internal_Debt_Log	0.00	0.00	0.00	0.18	0.00	0.00	0.08	0.00	0.03
	Other_External_Debt_Log	10.10	0.34	5.30	2.28	5.21	0.26	2.82	0.00	3.29
	DTI	9.59	0.59	14.75	1.36	3.21	6.75	30.45	7.38	9.26
Geographic	Agder	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Innlandet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Møre og Romsdal	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Nordland	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Oslo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Other_County	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Rogaland	0.00	0.00	1.12	0.79	0.00	0.00	2.58	0.00	0.56
	Svalbard	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Troms og Finnmark	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01
	Trøndelag	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.00	0.03
	Vestfold og Telemark	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.02
	Vestland	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01
	Viken	0.00	0.00	0.00	2.44	0.00	0.00	0.12	0.00	0.32
External and macroeconomic	Unemployment_Rate	0.00	0.00	37.75	23.47	0.15	0.00	0.41	0.00	7.72
	Policy_Rate	0.00	0.00	0.00	7.76	0.00	11.15	0.10	0.00	2.38
	Housing_Price_Index	13.03	0.00	10.64	14.08	0.00	1.44	0.47	0.00	4.96
	Consumer_Price_Index_Yearly_Change	0.00	0.00	21.71	5.81	0.00	0.37	0.36	0.00	3.53
	Interest_Rates_Banks_Mortgage_Companies	0.00	0.00	0.00	5.60	0.00	0.00	5.20	0.53	1.42

The above table presents a heatmap of the normalised SHAP values, where a darker colour indicates a higher variable importance. Each fold assigns a value between 0 and 100 to the variables, with the most important variable receiving a value of 100, while the remaining variables are assigned a percentage value relative to the most important variable.

In order to provide a more comprehensive explanation of the results presented in Table 5, we enhance the analysis by including SHAP beeswarm plots, which can be found in Appendix B. The beeswarm plots introduce the SHAP values and their impact on the model output horizontally, with negative SHAP values indicating a reduced likelihood of churn and the opposite for positive SHAP values. In addition to the horizontal axes, the beeswarm plot introduces a vertical axis which represents the size of each variable's value.

We can see that *"PD"* and *"Age_Log"* dominate the average impact of the folds. *"PD"* has a mean of 62.59, whereas *"Age_Log"* has a mean of 69.72, making *"Age_Log"* slightly above *"PD"* in importance. The variable *"PD"* seems to be important for multiple folds, where it is the most important variable in 4 of the total 8 folds. If we study the beeswarm plots in Appendix B attached to each fold, mainly for folds 3, 4 and 5, we can see that higher values of *"PD"* indicate churn and lower values indicate non-churn. This result is similar to what Ongena et al. (2021) found in terms of customers with higher credit risk being more likely to switch banks. *"Age_Log"* was the most important variable in folds 1, 2, 6 and 7. From the corresponding SHAP beeswarm plots found in Appendix B, we can see that lower values for *"Age_Log"* indicate churn, and higher values indicate non-churn. This is in line with the results from Ongena et al. (2021), where increased age and length of customer relationship increases the likelihood to remain with the bank. Van den Poel and Larivière (2004) also found that the probability of switching was higher earlier in the relationship with their service provider. From the variable *"Customer_Relationship_Duration* we can see that lower values indicate churn and higher values indicate non-churn. This corresponds to both Van den Poel and Larivière (2004) and Ongena et al. (2021) results, but it is important to notice that an early churn of a customer could be due to other factors, naturally making for lower valued *"Customer_Relationship_Duration"* to indicate churn.

Variables associated with loans, such as *"Seniorlaan_Balance_Log"*, *"Other_Internal_Debt_Log"*, and *"Other_External_Debt_Log"* appears to have little or no effect on the model output, whereas *"Boligkreditt_Balance_Log"* appears to have a larger effect. From Figure 9 we see that larger values of *"Boligkreditt_Balance_Log"*

indicates non-churn, and some lower values also have an effect on the probability of churn. *"Repayment_Plan"* appears to have a large average impact on the folds, but it is mainly associated with high values in folds 3 and 4. From the SHAP plot Figure 10 in Appendix B, higher values of *"Repayment_Plan"* indicates non-churn and lower values indicate churn. This could be due to the fact that a loan customer with a repayment plan naturally will churn with less time left on the repayment plan, and the opposite with a longer time left, as long as the customer is not affected by other factors. In Table 5, we see that *"LTV"* has an overall large impact on the model output. We can see in Figure 13 that both high and low values of *"LTV"* indicates a higher likelihood of churn, and mid-range values indicate a lower likelihood of churn. The lower values could have the same explanation as *"Repayment_Plan"*, where less time left in the repayment plan would give lower *"LTV"* values. Higher values of *"LTV"* increases the likelihood of churn and can be explained by the fact that high values may involve increased credit risk. This is in relation to what we found concerning *"PD"*.

"Weighted_Average_Interest_Rate" has an overall large impact on the model output. Figure 13 shows that lower values indicate less likelihood of churn, and higher values indicate a higher likelihood of churn. This corresponds to what Lukas and Nöth (2019) found in relation to how borrowers are more likely to search for additional offers when interest rates rise and less likely when interest rates fall. If we take into consideration that a decreased difference of interest rates offered at initiation, and increased difference after initiation, found by Ongena et al. (2021), the explanation could be due to a transparent lending market. We can assume that customers with lower interest rates would not have a wide range of competitive alternatives, and might not tend to switch banks. Customers with a higher interest rate might be able to find better alternatives.

From Table 5 we see that counties, where the customer have their residence, have little to no impact on the model. This does not correspond to what Levesque and McDougall (1996) mentions, where location is important when choosing a bank. This could be due to the digital transformation of banks, where more bank services are available online, and the need for physical branch visits has diminished. It is

important to notice that in this analysis, we can only interpret the location based on the switching of banks. The location of the bank may play a significant role in the decision-making process when selecting a bank.

For external and macroeconomic variables, we see that all 5 variables have an impact on the model output, where *"Unemployment_Rate"* and *"Housing_Price_Index"* have the largest average impact. From Figure 10 we can see that lower values of *"Housing_Price_Index"* can indicate a lower likelihood of churn, and higher values can indicate a higher likelihood of churn. In relation to what Basten and Koch (2015) found regarding that higher housing prices resulted in a growth in mortgage demand. It can be argued that increasing housing prices could influence the decision to take out new mortgages due to a stronger housing market, which could potentially provide better alternatives and influence the possibility of switching banks.

By observing Table 5 we can see that multiple variables seem to have a higher impact on the model output in folds 3 and 4, but a lower impact on other folds. For external and macroeconomic variables, we see a larger impact in folds 3 and 4. In Table 3, we see that the AUC in fold 3 and fold 4 indicate overfitting, which could possibly have an impact on the SHAP-values in fold 3 and fold 4. Another possible explanation for the varied overall impact of fold 3 and fold 4 could be attributed to the specific time period they include, spanning from 2014 to 2017. The significant decline in oil prices during 2014 contributed to pushing the Norwegian economy into a recession, resulting in a decline in economic growth and a rise in unemployment (Brander, 2019). The impact of external and macroeconomic variables in fold 3 and fold 4 could be affected by this. If we were to examine this further, it would be natural to look at the variable *"Rogaland"*, which is a county that was significantly affected during this period. We can see that *"Rogaland"* has a slight impact on the model in folds 3 and 4.

5.2.2 LASSO

Table 6 contains LR coefficients and z-scores in parentheses for each variable selected by LASSO regularisation. LASSO variable selection was applied for each fold.

Table 6: Logistic regression coefficients

	Variable	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8
Demographic	Size_Household			0.01(0.44)	-0.0(-0.08)	0.01(0.52)	-0.04(-1.75)		-0.01(-0.79)
	Income_Log	0.02(0.16)		0.05(0.54)	0.08(0.99)		0.18(2.96)	0.08(1.65)	
	Age_Group_1	0.13(0.48)	0.09(0.75)	0.09(0.78)				-0.09(-0.68)	0.03(0.2)
	Age_Group_2	-0.08(-0.42)					-0.08(-1.45)	-0.16(-1.94)	-0.1(-1.33)
	Age_Group_3								
	Age_Log	-0.62(-2.06)	-0.87(-3.69)	-0.96(-4.28)	-0.96(-7.18)	-0.92(-7.64)	-0.82(-7.51)	-0.84(-5.34)	-0.65(-4.49)
	Sex_Dummy	-0.19(-2.14)	-0.17(-2.19)	-0.23(-3.05)	-0.11(-1.64)	-0.12(-2.06)	-0.06(-1.18)	-0.03(-0.81)	-0.04(-1.0)
Loan-related	Repayment_Plan	-0.02(-2.43)	-0.01(-1.4)	-0.01(-1.41)	-0.01(-1.66)	-0.01(-2.3)	-0.01(-2.62)	-0.02(-4.21)	-0.02(-5.1)
	Weighted_Average_Interest_Rate								
	LTV							0.47(3.55)	0.56(4.78)
	PD				2.88(12.03)	2.42(10.77)	2.09(10.79)	1.94(11.01)	1.83(10.1)
	Credit_Card_Dummy							-0.08(-1.36)	-0.08(-1.78)
	Deposit_Dummy								
	Customer_Relationship_Duration	-0.0(-1.94)	-0.0(-3.1)	-0.0(-2.74)	-0.0(-1.32)	-0.0(-1.67)	-0.0(-3.25)	-0.0(-3.72)	-0.0(-4.43)
	Customer_Relationship_Duration_Log	0.07(1.46)	0.09(2.06)	0.09(2.18)	0.03(0.8)	0.06(1.73)	0.05(1.72)	0.06(2.15)	0.06(2.33)
	Repayment_Loan_Balance_Log	-0.03(-1.41)	-0.01(-0.79)	-0.01(-0.7)	-0.01(-0.5)	-0.02(-0.93)	-0.02(-1.07)	-0.03(-1.97)	-0.04(-2.75)
	Boligkreditt_Balance_Log	-0.06(-1.42)	-0.04(-3.06)	-0.05(-3.46)	-0.11(-4.42)	-0.12(-4.79)	-0.15(-7.09)	-0.12(-5.59)	-0.13(-6.15)
	Boligkreditt_Credit_Limit_Log	-0.02(-0.57)			0.07(2.7)	0.07(2.8)	0.08(3.99)	0.04(1.69)	0.03(1.23)
	Seniorlaan_Balance_Log								-0.08(-1.18)
	Other_Internal_Debt_Log				0.03(1.37)	0.04(3.01)	0.02(1.87)	0.01(0.6)	-0.02(-1.23)
	Other_External_Debt_Log	0.02(2.12)	0.01(1.63)	0.01(2.01)	0.01(0.98)	0.01(1.51)	0.0(0.08)	-0.0(-0.07)	-0.0(-0.41)
	DTI	-0.0(-0.03)	-0.0(-0.03)	-0.0(-0.03)	-0.0(-0.03)	0.04(4.1)	0.03(2.1)	0.02(2.16)	0.01(2.01)
Geographic	Agder								
	Innlandet								
	Møre og Romsdal								
	Nordland								
	Oslo								-0.09(-1.85)
	Other_County								
	Rogaland						0.18(1.89)	0.19(2.25)	
	Svalbard								
	Troms og Finnmark								
	Trøndelag								
	Vestfold og Telemark								-0.28(-2.6)
	Vestland								
	Viken				0.13(2.07)		-0.04(-0.79)	-0.13(-3.01)	-0.16(-3.51)
External and macroeconomic	Unemployment_Rate			-0.12(-1.65)	0.02(0.13)	-0.03(-0.29)			-0.07(-1.18)
	Policy_Rate						0.32(4.4)	0.11(2.39)	0.11(1.62)
	Housing_Price_Index	0.02(2.8)	-0.01(-2.39)		0.01(2.35)	0.01(2.63)	-0.0(-0.86)	-0.0(-1.1)	0.0(1.66)
	Consumer_Price_Index_Yearly_Change	-0.04(-0.51)	-0.16(-1.82)			0.05(1.47)		-0.09(-2.7)	-0.1(-3.37)
	Interest_Rates_Banks_Mortgage_Companies				0.16(0.68)				

The above table presents coefficients which have been chosen by LASSO from the LR model. The table contains coefficients and z-scores in parentheses. Positive and negative coefficients help to explain how the values in each variable are associated with churn.

Each fold has a different amount of selected variables with a spread from 12 to 25 variables, where fold 2 contains 12 selected variables and fold 8 contains 25. Table 6 shows that the first 3 folds have a smaller number of selected variables, specifically 15, 12 and 13 variables. In contrast, the last three folds have a higher number with 19, 22, and 25.

Out of a total of 40 variables used in this model, only 9 variables have been included in every fold. Each of the 9 variables are associated with the customer, where 6 of them are related to loan and the other 3, *"Customer_Relationship_Duration"*, *"Sex_Dummy"* and *"Age_Log"* are related to the demographics of the customer. The variable *"Housing_Price_Index"* was selected in 7 folds, and *"Bolig_Credit_Limit_Log"* was selected in 6 folds. In contrast, 12 variables were not selected in any of the folds, where 9 of these variables are associated with county, and only 4 variables of county were selected by LASSO.

We define variables as significant for a significance level of lower than 5%, that is, when the z-score is larger than 1.96 or smaller than -1.96. The variable *"Age_Log"* was the only variable that was selected in every fold and is also defined as significant for every fold. *"PD"* is significant for all the 5 folds that it was included in. *"Size_Household"* is selected for 5 folds, but is not significant in any of the folds. This is also the case for *"Unemployment_Rate"*, where the variable is not significant in any of the selected 4 folds.

We can see that a positive coefficient of *"PD"* indicates that higher values increases likelihood of churn. LR coefficients in the variable *"Age_Log"* are negative for all folds and indicate that a lower age has an increased likelihood of churn, and higher age has a decreased likelihood of churn. The results in *"PD"* and *"Age_Log"* corresponds to what we found in Table 5 and the supplemented beeswarm plots. The variable *"Weighted_Average_Interest_Rate"* is not selected by LASSO in any of the folds. This is in contrast to the SHAP values in Table 5 from the XGBoost model where *"Weighted_Average_Interest_Rate"* has a relatively large impact on the model output.

"LTV" was only selected in folds 7 and 8, which indicates that higher values increase the likelihood of churn. In the SHAP values for fold 7, we found that both higher and lower values increased the likelihood of churn, whereas mid-range values decreased the likelihood of churn. The LR-coefficients have a similar result in terms of emphasising that higher values increase the likelihood of churn. The results seem to give the same interpretation, but the beeswarm plot Figure 13 gives a deeper understanding of how both low and high values can increase likelihood of churn.

The variable *"Customer_Relationship_Duration_Log"* is selected in all folds, but only significant on a 5% level in folds 2, 3, 7 and 8. In these folds, we can see that the LR coefficients have an impact, and that higher values of the variable *"Customer_Relationship_Duration_Log"* increases the likelihood of churn. This does not corresponds with what we found in the SHAP values associated with the XGBoost model. It is important to notice that we have both *"Customer_Relationship_Duration"* and *"Customer_Relationship_Duration_Log"* included in the dataset, and it appears that the XGBoost model considers the relationship between the variable *"Customer_Relationship_Duration"* and the target variable as more important than the logarithmic transformed variable. In contrast, it appears that the LASSO variable selection consider the logarithmic transformed variable as more important. *"Boligkreditt_Balance_Log"* is selected for all folds and is significant for all folds except the first. The LR coefficients indicates that lower values increases the likelihood of churn. This corresponds with what we found in the SHAP values in Table 5 and Figure 9.

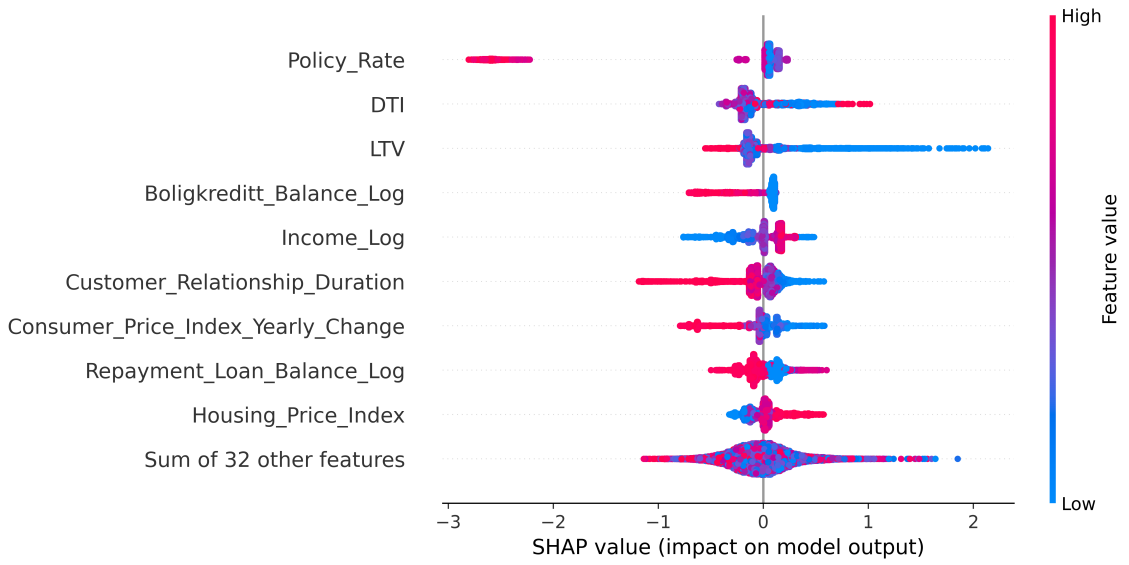
For external and macroeconomic variables we see that the variable *"Policy_Rate"* is selected in folds 6, 7 and 8, and is significant for all three. Higher values of *"Policy_Rate"* indicates an increased likelihood of churn. This corresponds to what we found in Figure 12, where higher values of *"Policy_Rate"* increases the likelihood of churn. For *"Consumer_Price_Index_Yearly_Change"* we see that the variable is selected in 5 folds, but is only significant in folds 7 and 8. Lower values indicate an increased likelihood of churn.

We notice how external and macroeconomic factors seem to have some effect on the models. The most noticeable external and macroeconomic factors are *"Housing_Price_Index"*, *"Consumer_Price_Index_Yearly_Change"* and *"Policy_Rate"*. The big spike in external and macroeconomic variable importance within fold 3 and 4 can be related to the significant decline in the oil price during 2014.

5.3 Valuable customers

Figure 2 presents a SHAP beeswarm plot of the top 10% most valuable customers. We proxy the bank's valuation of customers based on a measure *"PD_LTV"*, which is the product of *"PD"* and *"LTV"*. For the analysis in this chapter, we analyse the top 10% of the customer-months with the lowest *"PD_LTV"*.

Figure 2: Beeswarm plot of top 10% customers



The above figure presents a beeswarm plot for the top 10% valuable customers based on the value of *"PD_LTV"*, which is the product of PD and LTV. The beeswarm plot introduces the SHAP value and their impact on the model output horizontally, with negative SHAP values indicating a reduced likelihood of churn and the opposite for positive SHAP values. In addition to the horizontal axes, the beeswarm plot introduces a vertical axis representing the size of each variable's value.

"LTV" is one of the variables with a higher average score in Table 5, which can also be seen in this plot ranking as number 3 in impact, with low values indicating a high probability for churn. Analysing *"LTV"* in itself gives, however, little information

as *"LTV"* reduces for each down-payment on the customers' loan, until they have fulfilled their loan and are, per our definition, no longer a loan customer. Higher values of *"LTV"* however, indicate less likelihood of churn, but we have seen in certain SHAP beeswarm folds from the XGBoost model that high and low *"LTV"* can indicate a greater chance for churn.

"DTI" is a highly rated variable with an interesting combination of both high and low values indicating a higher chance for churn whilst the mid-range values indicate a lower chance for churn. As *"DTI"* refers to Dept-to-Income, it is also interesting to note that *"Income_log"* makes an appearance in this plot with mainly higher income and some with lower income indicating churn, whilst most of the low-income customers indicate less probability for churn. Looking at these two together, we can make an assumption that those who have a higher-than-average income might be interested to look into other opportunities elsewhere. We can also assume that those customers who might struggle more with their monthly debt payment because of their lower income can be tempted to find better offers suited for them.

Income can also be seen in connection with age. We have previously pointed out how age has a significant variable importance to our models. We see from the beeswarm plot how most of the lower-income customers of the top 10% have a lower chance to churn, but an increase in income results in an increased chance for churn. We can make an assumption that younger customers tend to get less favourable loans as their income is lower as well as having fewer options to choose from. As their income increases, we believe they can be more likely to look for better offers elsewhere. This assumption of age is also based on the descriptive statistics in Figure 6 and the overall weight of *"Age_Log"* as shown in Table 5.

"Customer_Relationship_Duration" and *"Boligkreditt_Balance_Log"* are interesting variables that show up in Figure 2. Both of the variables indicate a form of customer loyalty based on the duration spent being a customer of the bank, and the use of other financial products, in this case, *"Boligkreditt"*. A higher value *"Customer_Relationship_Duration"* and *"Boligkreditt_Balance_Log"* indicates a decreased likelihood of churn.

The variable importance of *"Boligkreditt_Balance_Log"* and *"Customer_Relationship_Duration"* corresponds with Bilal Zorić (2016) and Reichheld and Kenny (1990) emphasis on valuable customers, as a long-term customer have a higher value than younger customers and the longer your relationship with the bank, the more valuable you are.

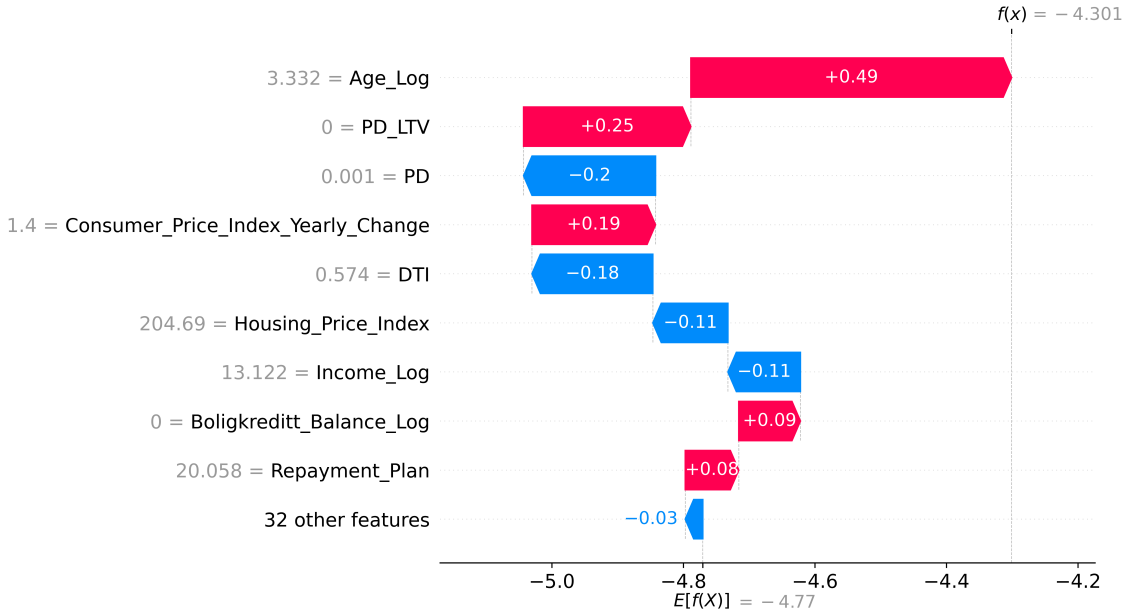
The *"Policy_Rate* has a fairly high effect on the churn and we can assume this corresponds to the thought of older customers having more money. We can assume that the older customers can receive more, and potentially better offers, making them more willing to look elsewhere. Especially when the policy rate is low. This assumption around policy rate also helps us understand the big spike and importance of policy rate in Figure 2.

"Policy_Rate" has the highest impact on this model with a high value of *"Policy_Rate"* indicating a lower likelihood of churn. Considering that the analysis of the top 10% customers is based on low-risk customers, we can assume that lower risk can give these customers better price levels on mortgage loans even when policy rates are increasing, and therefore the effect on policy rate might not create incentives to switch banks. However, low-risk customers might have other and better alternatives, which could possibly lead to churn. The fact that a higher policy rate leads to a lower likelihood of churn can indicate that the customers are remaining with the bank due to other factors, such as loyalty. In addition to *"Policy_Rate"*, we also notice the following external and macroeconomic variables, *"Consumer_Price_Index_Yearly_Change"* and *"Housing_Price_Index"* has a impact in Figure 2.
"Consumer_Price_Index_Yearly_Change" tells us that a higher consumer price index indicates less likelihood of churn, and the *"Housing_Price_Index"* tells us that higher values indicates an increased likelihood of churn.

To gain a deeper understanding of the valuable customers, we present in Figure 3 and Figure 4 two waterfall plots based on two top 10% customers who have churned and the most important variables that made them churn. Waterfall plots give a local interpretation, meaning we look at one single customer per plot. Figure 3 presents a top 10% customer at the age of 28 that has churned. Figure 4 presents the same but with a customer aged 58. The grey values given on the left of the variables are the

value of this customer's presented variables. Red bars in the waterfall plots indicate a higher impact on churn, whilst the blue bars indicate less impact on churn.

Figure 3: Waterfall plot of a churned 28-year-old customer

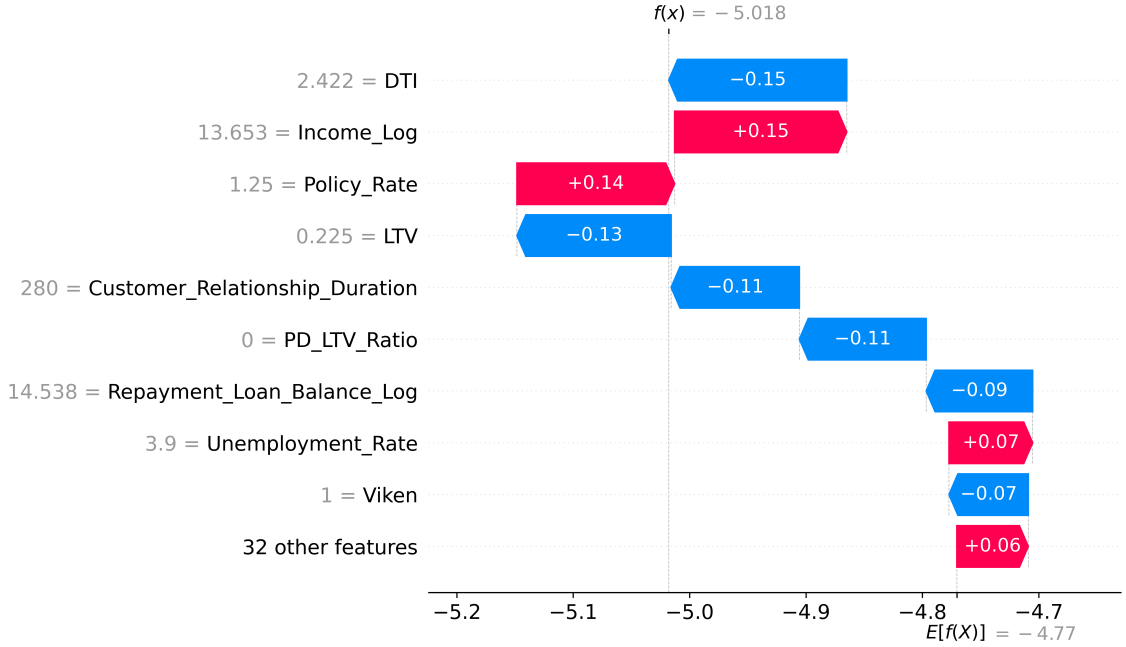


The above figure presents a SHAP waterfall plot. The waterfall plot extracts a unique customer from our "Top 10% customers" at the age of 28. The values on the left of the variables indicate this individuals' specific variable values. For example Age = 3.332 = 28, with the consumer price index being at 1.4 on the extracted moment. Red bars indicate a higher impact on churn, whilst the blue bars indicate less impact on churn. The size of the bar represents the significance of the variables' impact on the individual.

In the case of the 28-year-old customer, Figure 3 shows what we have seen in previous figures such as Figure 6 and Table 5 that *"Age_Log"* has a significant impact on churn. *"PD_LTV"* indicate a high impact on the individual's churn, but in contrast *"PD"* is seen to not indicate the impact of churn. The impact of *"PD_LTV"* could be a result of lower values of *"LTV"*, which would increase the likelihood of churn, as described under Table 5.

We note the external and macroeconomic variables *"Housing_Price_Index"* and *"Consumer_Price_Index_Yearly_Change"* having an impact on the churn. *"Consumer_Price_Index_Yearly_Change"* indicates an increased impact of churn, whereas *"Housing_Price_Index"* indicates a decreased impact of churn. From Table 8 we can see that the values of these two variables is below the median of the dataset.

Figure 4: Waterfall plot of churned 58-year-old customer



The above figure presents a SHAP waterfall plot. The waterfall plot extracts a unique customer from our "Top 10% customers" at the age of 58. For a better understanding of the plot, refer to the caption in Figure 3.

In the case of the 58-year-old customer, Figure 4 shows how *"DTI"* is the highest rated variable but without impact on the churn. *"Income_Log"* and *"Policy_Rate* are seen as having a significant impact on the churn. At the time of churn for this customer, the policy rate was 1.25 which is higher than the median Table 8. We note here that a higher value of *"Policy_Rate"* increases the impact of churn, which is corresponding with that we found in Figure 2. For the 58-year-old customer we can see that a value of 280 months in *"Customer_Relationship_Duration"* reduces the impact of churn, this is higher than the median, and is corresponding to what we found in Figure 2.

We can see through the waterfall plots that the younger customer have variables that revolve around PD, DTI and age, as well as income and other loans. We can assume this is because of their young age resulting in averagely weaker finance, thus making their loans and the ability for down-payments worse. On the contrary, the 58-year-olds' variables emphasise the financial strength the customer has, their down-payments and the *"Customer_Relationship_Duration"*, which can imply that

the customer has more money to spend and thus a lower threshold to assess other offers.

We have previously discussed the relevancy and importance of focusing on the characteristics and the value of each customer through a literature review. Gorgoglione and Panniello (2011) and Lemmens and Gupta (2020) mention how it is important to focus the prediction on the key customers, also seen as the ones who are most profitable. We have seen, through Figure 2, Figure 3 and Figure 4, that there is a notable difference between a valuable customer and a general customer. We see different variables in the beeswarm plots, where the variable importance and their values are also different. Some of the different variables that are rarely presented in the general 8 beeswarm plots compared to Figure 2 are; *"Consumer_Price_Index_Yearly_Change"*, *"Repayment_Loan_Balance_Log"*, *"Income_Log"* and *"Policy_Rate"*. Based on these findings we make the assumption that the most valuable customers pay more attention to the external and macroeconomic factors and their impacts on society, thus potentially being more reactive or proactive. The results of our findings implicate the importance of emphasising and focusing on valuable customers and their characteristics as mentioned by, Gorgoglione and Panniello (2011) and Lemmens and Gupta (2020), since they present different variable importance than the general customer analysis does.

6 Conclusion

The aim of this thesis has been to study what characteristics in loan customers can affect the churn by the use of ML methods. We have used two different variable selection tools to gain a deeper understanding of the black-boxes and the variable importance extracted from the ML methods. Here we present our conclusions based on the research questions stated in the introduction. We also discuss the strengths and weaknesses of this thesis, and, finally, ideas for future studies.

6.1 Discoveries and implications

Our main problem is introduced as: *"What characteristics in loan customers can influence the likelihood of their churn?"*

Our thesis revealed that the age of the loan customer and the credit risk indicated by PD were the key factors in explaining customer churn. These factors had the greatest impact on our XGBoost model. In addition, we found LTV, repayment plan, the length of customer duration, and the balance of the loan product boligkreditt, to have a large impact on the likelihood for customer churn. We also found that the weighted average interest rate and DTI had an impact on the likelihood for customer churn. Our thesis revealed that several characteristics, such as the customers' county of residence, household size, gender, possession of credit card and deposit account, did not have a substantial impact on explaining customer churn. In addition, we found that external and macroeconomic factors played a role in influencing the likelihood of customer churn. The policy rate, housing price index and consumer price index had a low impact, whereas the unemployment rate had a greater impact on customer churn.

In addition to the main problem, we also have three research questions with one being as follows: *"Can we predict the churn of customers based on machine learning models?"*

Our ML methods have proven to give weak predictive results, and thereby preventing us from confidently predict customer churn. Both XGBoost and LR struggle to give

acceptable results in our analysis based on the given performance measures. The results given in the two mentioned models are adequately similar, helping us reduce bias in the choice of different, and specific, ML methods.

Although the predictive ability is considered weak, we can see slight indications that churn predictions within banks can be done. The results present slight tendencies for predictive ability, which, with an increased amount of data and reduced imbalances, can make for better results.

We also answer the following research question: *"Are the characteristics that influence the likelihood of churn different for the most valuable customers compared to other customers?"*

Our thesis demonstrates that the characteristics influencing the likelihood of churn in valuable customers differ from those in other customers. In contrast to other customers, the age of the loan customer did not have a large impact on the likelihood of churn. We found that the income of the customer, DTI, and repayment loan balance had a larger impact on the likelihood of churn on valuable customers than with other customers. The impact of external and macroeconomic factors on the likelihood of churn was more significant for valuable customers compared to other customers. The valuable customer was influenced by the consumer price index, the housing price index and the policy rate. Among all variables associated with valuable customers, the policy rate had the most significant impact on the likelihood of churn. This is in contrast to other customers, where the policy rate had a low impact. In contrast to the impact seen on general customers, the unemployment rate had little effect on the likelihood of churn among valuable customers.

Finally, we answer the research question: *"Do external and macroeconomic factors affect the churn?"*

Throughout the analysis our models, we have found that external and macroeconomic factors do, in fact, affect the customer churn. The housing price index is a notable factor which has been presented multiple times within the fold plots. This makes sense due to the nature of our dataset being housing loan customers and the previously mentioned effect on loan customers.

The policy rate has also proven to be an important macroeconomic factor, more specifically for higher-valued customers. Other external and macroeconomic factors are presented within the analysis of valuable customers, indicating that the factors influence valuable customers more than the general customer base.

6.2 Strengths and weaknesses

In addition to the new insight we give to the research community, we also want to emphasise what benefits businesses can get from this. Successfully retaining customers have previously been proven to increase the profits of businesses significantly.

We have managed to prove that external and macroeconomic factors can affect the customer churn, meaning businesses can take preliminary actions when a country's financial situation changes. Such actions can be the likes of having a below-average and competitive interest rate. The thesis also presents the general characteristics of their customers and the likelihood of churn. This can give our respective collaborative bank the opportunity to get to know their customers even more and counteract possible churn through changes based on their variable values. We also presented characteristics for the most valuable customers, as well as characteristics of two individual valuable churned customers. The two individual churned customers can help create a better understanding of their characteristics, and how to retain them. This approach can be applied to multiple customers as a useful tool for further analysis and improved retention.

Although our thesis can benefit the business aspect, we can not exclude the fact that there are multiple weaknesses. We have previously mentioned how the dataset has suffered from imbalances. There are multiple missing values in different variables, meaning we have had to delete more than we naturally desired. This also resulted in the dataset being smaller, which, in return, made for complications as we possess a scarce number of observations of churned customers.

We define churned customers as those who have a loan balance of zero. This also goes for those who have fully repaid their loan, but they might not have actually "churned", resulting in a less ideal situation. Our definition of churn has also proven

to further complicate our results as some variables naturally decrease with time. This results in variables such as LTV and repayment plan having a higher importance than what they maybe should have. This can disrupt the overall interpretation of our models.

The results given of the ML models' predictive abilities are also a problem for the thesis as it makes the validity of our results and assumptions weaker. Another problem we encountered is that customers often make decisions to churn based on emotions and behaviour, as well as their personal relationship with the bank. This makes for factors similar to customer service, userfriendly webpage and traffic crucial information in understanding customer churn.

6.3 Recommendations for future studies

We have delimited our thesis to financial, external and macroeconomic data, thus excluding CRM data, surveys and other qualitative data. We would highly recommend similar research be carried out on other banks but with a combination of quantitative and qualitative data, to gain a better understanding of the customers and their churn. This can potentially result in prediction models with increased performance. We see it as relevant and important for the study to gain both depth and validity, in order for businesses to make better decisions. Achieving successful retention of customers should be achieved through the use of both quantitative and qualitative data to gain a complete picture of their customers.

Furthermore, according to Song et al. (2004), customers have the potential to transition from being non-churners to becoming churners over time, or vice versa. We, therefore, believe that performing an analysis on individual customers, rather than customer-month observations, could provide more insights. This approach allows for incorporating changes in customers' behaviour over time into the models.

It could also be interesting to expand the study by implementing data from multiple different banks. This will help generalise the results more, and potentially increase the predictability of chosen ML methods. Combining the amount of data with a more in-depth analysis of external and macroeconomic factors could potentially

make for new discoveries which could change the way banks and other financial service providers adapt their business strategies.

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, *298*. https://doi.org/10.1016/j.artint.2021.103502

Almås, I., Bache, I. W., Børsum, Ø., Fjære-Lindkjenn, J., & Longva, P. (2023, May 10). *Vurdering av finansiell stabilitet 2023 - 1. halvår* [Financial stability assessment 2023 H1]. Retrieved 22nd May 2023, from https://www.norges-bank.no/tema/finansiell-stabilitet/vurdering-av-finansiell-stabilitet/2023-1/?tab=129824

Anderson, W. T., Cox, E. P., & Fulcher, D. G. (1976). Bank selection decisions and market segmentation. *Journal of Marketing*, *40*(1), 40–45. https://doi.org/10.2307/1250674

Armantier, O., Bruine de Bruin, W., Topa, G., van der Klaauw, W., & Zafar, B. (2015). Inflation expectations and behavior: Do survey respondents act on their beliefs? *International Economic Review*, *56*(2), 505–536. https://doi.org/10.1111/iere.12113

Aursand, P. (2022, September 15). *Hva er inflasjon?* [What is inflation?]. Retrieved 29th April 2023, from https://www.ssb.no/priser-og-prisindekser/konsumpriser/artikler/hva-er-inflasjon

Basten, C., & Koch, C. (2015). The causal effect of house prices on mortgage demand and mortgage supply: Evidence from switzerland. *Journal of Housing Economics*, *30*, 1–22. https://doi.org/10.1016/j.jhe.2015.07.001

Bilal Zorić, A. (2016). Predicting customer churn in banking industry using neural networks. *Interdisciplinary Description of Complex*, *14*(2), 116–124. https://doi.org/10.7906/indecs.14.2.1

Börjesson, L., & Singull, M. (2020). Forecasting financial time series through causal and dilated convolutional neural networks. *Entropy*, *22*(10), 1094. https://doi.org/10.3390/e22101094

Brander, A. S. (2019, May 2). *Nytt oljeprisfall vil bremse veksten i norsk økonomi* [A new drop in oilprice will slow down the growth of norwegian economy].

Retrieved 19th May 2023, from https://www.norges-bank.no/bankplassen/arkiv/2019/nytt-oljeprisfall-vil-bremse-veksten-i-norsk-okonomi/

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Burda, M. C., & Wyplosz, C. (2013). *Macroeconomics: A european text* (Sixth edition). Oxford University Press.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Dawkins, P., & Reichheld, F. F. (1990). Customer retention as a competitive weapon. *Directors & boards*, *14*(4), 42–.

Eiendomnorge. (n.d.). *Månedsrapporter - Eiendom Norge* ["Monthly Reports - Real Estate Norway"]. Retrieved 24th May 2023, from https://eiendomnorge.no/boligprisstatistikk/statistikkbank/rapporter/manedsrapporter/

Foote, C., Gerardi, K., Goette, L., & Willen, P. (2010). Reducing foreclosures: No easy answers. *NBER Macroeconomics Annual*, *24*, 89–138. https://doi.org/10.1086/648289

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. Retrieved 21st May 2023, from https://www.jstor.org/stable/2699986

Gorgoglione, M., & Panniello, U. (2011). Beyond customer churn: Generating personalized actions to retain customers in a retail bank by a recommender system approach. *Journal of Intelligent Learning Systems and Applications*, *03*(2), 90. https://doi.org/10.4236/jilsa.2011.32011

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, *34*(10), 2902–2917. https://doi.org/10.1016/j.cor.2005.11.007

Hålien, J. (2022, September 12). *Forbrukerrådet: Banktilbudet for ikke-digitale kunder er ikke godt nok.* [Consumer Council: The banking offer for non-digital customers is not good enough]. Retrieved 21st April 2023, from https://www.aftenposten.no/norge/i/2B1w8r/forbrukerraadet-banktilbudet-for-ikke-digitale-kunder-er-ikke-godt-nok

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (1st ed.). Wiley. https://doi.org/10.1002/9781118548387

Hundre, S., Kumar, P. R., & Kumar, G. D. (2013). Customer retention–key to success for organization: A case study of banking industry. *Research Journal of Agricultural Science*, *4*(1), 702–705. https://www.academia.edu/38765691/Customer_Retention_Key_to_Success_for_Organization_A_Case_Study_of_Banking_Industry

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in r*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Kirkos, E. (2015). Assessing methodologies for intelligent bankruptcy prediction. *Artificial Intelligence Review*, *43*(1), 83–123. https://doi.org/10.1007/s10462-012-9367-6

Koeniger, W., Lennartz, B., & Ramelet, M.-A. (2022). On the transmission of monetary policy to the housing market. *European Economic Review*, *145*. https://doi.org/10.1016/j.euroecorev.2022.104107

Kvamme, H., Sellereite, N., Aas, K., & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, *102*, 207–217. https://doi.org/10.1016/j.eswa.2018.02.029

Laroche, M., & Taylor, T. (1988). An empirical study of major segmentation issues in retail banking. *International Journal of Bank Marketing*, *6*(1), 31–48. https://doi.org/10.1108/eb010824

Lemmens, A., & Gupta, S. (2020). Managing churn to maximize profits. *Marketing Science*, *39*(5), 956–973. https://doi.org/10.1287/mksc.2020.1229

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Levesque, T., & McDougall, G. H. (1996). Determinants of customer satisfaction in retail banking. *International Journal of Bank Marketing*, *14*(7), 12–20. https://doi.org/10.1108/02652329610151340

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141. https://doi.org/10.1016/j.ins.2013.07.007

Lukas, M., & Nöth, M. (2019). Interest rate changes and borrower search behavior. *Journal of Economic Behavior & Organization*, *163*, 172–189. https://doi.org/10.1016/j.jebo.2019.03.020

Lundberg, S. (2018). *Welcome to the SHAP documentation*. Retrieved 24th May 2023, from https://shap.readthedocs.io/en/latest/index.html

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). Explainable AI for trees: From local explanations to global understanding. (arXiv:1905.04610). https://doi.org/10.48550/arXiv.1905.04610

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book

Neslin, S., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research American Marketing Association ISSN*, *43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Norgesbank. (n.d.-a). *Styringsrenten* [The policy rate]. Retrieved 29th April 2023, from https://www.norges-bank.no/tema/pengepolitikk/Styringsrenten/

Norgesbank. (n.d.-b). *Styringsrenten månedsgjennomsnitt* [Policy rate monthly average]. Retrieved 20th April 2023, from https://www.norges-bank.no/tema/Statistikk/Styringsrente-daglig/Styringsgrente-manedlig/

Ongena, S., Paraschiv, F., & Reite, E. J. (2021). Counteroffers and price discrimination in mortgage lending. https://doi.org/10.2139/ssrn.3935746

Paraschiv, F., Schmid, M., & Wahlstrøm, R. R. (2023). Bankruptcy prediction of privately held SMEs using feature selection methods. https://doi.org/10.2139/ssrn.3911490

Piccinini, E., Gregory, R., & Kolbe, L. (2015). Changes in the producer-consumer relationship - towards digital transformation. *Wirtschaftsinformatik Proceedings 2015*. https://aisel.aisnet.org/wi2015/109

Reichheld, F. F., & Kenny, D. W. (1990). The hidden advantages of customer retention. *Journal of Retail Banking*, *12*(4), 19–24. Retrieved 21st May 2023, from https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=01952064&v=2.1&it=r&id=GALE%5C%7CA9295626&sid=googleScholar&linkaccess=abs

Ringdal, K. (2013). *Enhet og mangfold: Samfunnsvitenskapelig forskning og kvantitativ metode* (3. udg) [Unit and diversity: Social science research and quantitative method.]. Fagbokforlaget.

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522–542. https://doi.org/10.1016/j.ins.2021.05.055

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the theory of games* (pp. 307–317). Princeton University Press.

Sharma, A., & Panigrahi, D. P. K. (2011). A neural network based approach for predicting customer churn in cellular network services. *International Journal of Computer Applications*, *27*(11), 26–31. https://doi.org/10.5120/3344-4605

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, *25*. Retrieved 21st May 2023, from https://proceedings.neurips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

Song, H. S., Kim, J. K., Cho, Y. B., & Kim, S. H. (2004). A personalized defection detection and prevention procedure based on the self-organizing map and association rule mining: Applied to online game site. *Artificial Intelligence Review*, *21*(2), 161–184. https://doi.org/10.1023/B:AIRE.0000021067.66616.b0

SSB. (2023a, May 4). *Renter i banker og kredittforetak* [Interest rates in banks and mortgage companies]. Retrieved 20th April 2023, from https://www.ssb.no/bank-og-finansmarked/finansinstitusjoner-og-andre-finansielle-foretak/statistikk/renter-i-banker-og-kredittforetak

SSB. (2023b, May 9). *Arbeidskraftundersøkelsen* [Labour force survey]. Retrieved 20th April 2023, from https://www.ssb.no/arbeid-og-lonn/sysselsetting/statistikk/arbeidskraftundersokelsen

SSB. (2023c, May 10). *Konsumprisindeksen* [Consumer price index]. Retrieved 20th April 2023, from https://www.ssb.no/priser-og-prisindekser/konsumpriser/statistikk/konsumprisindeksen

Thwaites, D., & Vere, L. (1995). Bank selection criteria — a student perspective. *Journal of Marketing Management*, *11*(1), 133–149. https://doi.org/10.1080/0267257X.1995.9964334

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, *157*(1), 196–217. https://doi.org/10.1016/S0377-2217(03)00069-9

Van der Drift, R., de Haan, J., & Boelhouwer, P. (2023). Mortgage credit and house prices: The housing market equilibrium revisited. *Economic Modelling*, *120*, 106136. https://doi.org/10.1016/j.econmod.2022.106136

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

Wahlstrøm, R. R. (2023). Explainable artificial intelligence (xAI) for interpreting machine learning methods and their individual predictions. https://doi.org/10.2139/ssrn.4321303

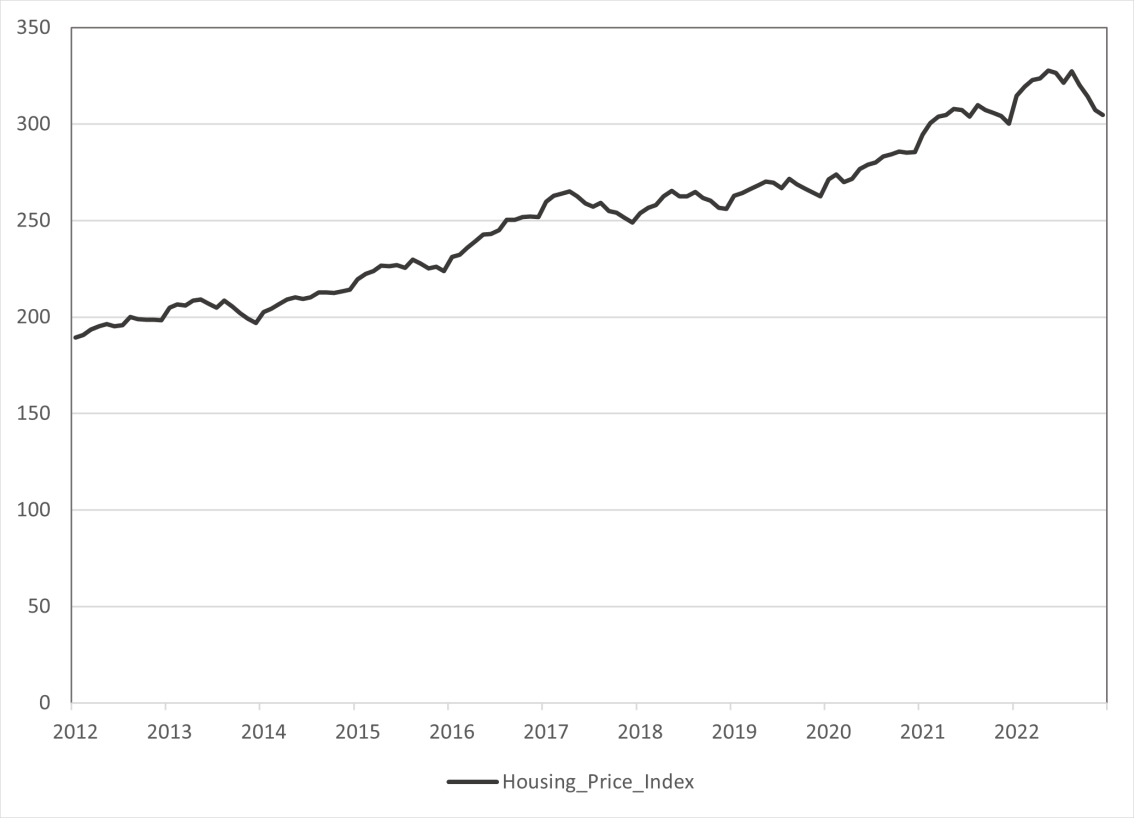
Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Appendix

A Appendix

A Housing price index

Figure 5: External factors - Trend



The above figure presents the external variable "Housing_Price_index" trend in Norway from 2012 to 2022.

B Definition of variables

Table 7: Definition of variables

	Variable name	Definition
Target variable	Churned_6_Lag	Dummy variable: 1 if the customer has churned
Demographic	Size_Household	The customer's houshold size
	Income_Log	The customer's income
	Age_Group_1	Dummy variable: 1 if the age of the customer is between 18 and 39
	Age_Group_2	Dummy variable: 1 if the age of the customer is between 40 and 64
	Age_Group_3	Dummy variable: 1 if the age of the customer is 65+
	Age_Log	Age of customer log-transformed
	Sex_Dummy	Dummy variable: 0 if male and 1 if female
Loan-related	Repayment_Plan	Repayment plan for the customers mortgage loan
	Weighted_Average_Interest_Rate	Weighted average interest rate for "Nedbetalingslån" and "Boligkreditt"
	LTV	The loan to value on the customers current loan
	PD	The probability of default
	Credit_Card_Dummy	Dummy variable: 1 if the customer has a credit card
	Deposit_Dummy	Dummy variable: 1 if the customer has a deposit account
	Customer_Relationship_Duration	Length of the customer relationship with the bank in months
	Customer_Relationship_Duration_Log	Length of the customer relationship with the bank in months log-transformed
	Repayment_Loan_Balance_Log	The customer's repayment loan balance log-transformed
	Boligkreditt_Balance_Log	The customer's "boligkreditt" balance log-transformed
	Boligkreditt_Credit_Limit_Log	The customer's "boligkreditt" credit limit log-transformed
	Seniorlaan_Balance_Log	The customer's "seniorlån" balance log-transformed
	Other_Internal_Debt_Log	The customer's other internal debt log-transformed
	Other_External_Debt_Log	The customer's other external debt log-transformed
	DTI	The customer's debt to income
	PD_LTV	The probability of default multiplied with loan to value
Geographic	Agder	Dummy variable: 1 if the customer has its residence in Agder
	Innlandet	Dummy variable: 1 if the customer has its residence in Innlandet
	Møre og Romsdal	Dummy variable: 1 if the customer has its residence in Møre og Romsdal
	Nordland	Dummy variable: 1 if the customer has its residence in Nordland
	Oslo	Dummy variable: 1 if the customer has its residence in Oslo
	Other_County	Dummy variable: 1 if the customer has its residence in "Other_County"
	Rogaland	Dummy variable: 1 if the customer has its residence in Rogaland
	Svalbard	Dummy variable: 1 if the customer has its residence in Svalbard
	Troms og Finnmark	Dummy variable: 1 if the customer has its residence in Troms og Finnark
	Trøndelag	Dummy variable: 1 if the customer has its residence in Trøndelag
	Vestfold og Telemark	Dummy variable: 1 if the customer has its residence in Vestfold og Telemark
	Vestland	Dummy variable: 1 if the customer has its residence in Vestland
	Viken	Dummy variable: 1 if the customer has its residence in Viken
External and macroeconomic	Unemployment_Rate	Unemployment rate
	Policy_Rate	Policy rate
	Housing_Price_Index	Housing price index
	Consumer_Price_Index_Yearly_Change	Consumer price index yearly change
	Interest_Rates_Banks_Mortgage_Companies	Interest rate for banks and mortgage companies

The above table presents the different variables within our dataset including a short definition of each variable.

C Descriptive statistics of variables

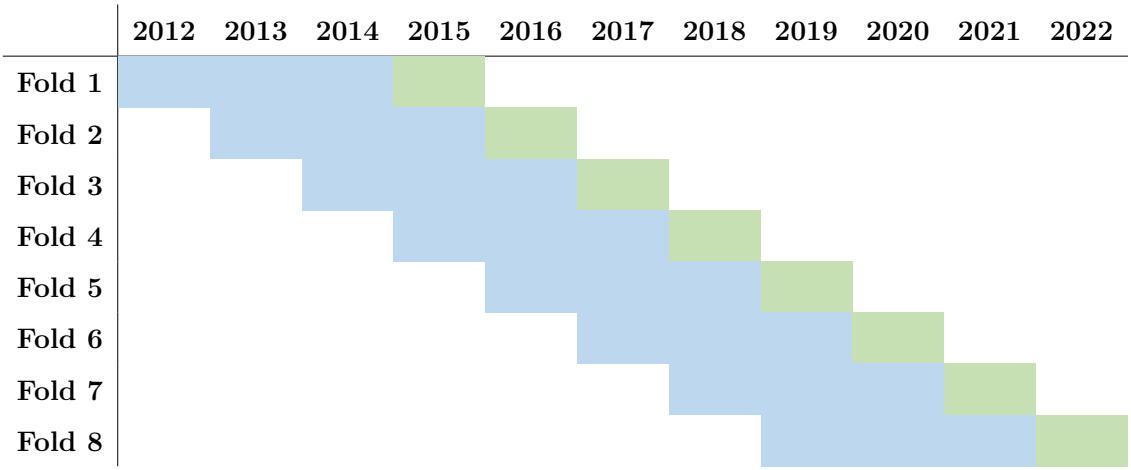
Table 8: Descriptive statistics

	Variable	Mean	Std	Min	Median	Max
Demographic	Size_Household	2.27	1.22	1.00	2.00	13.00
	Income	946471.85	591512.82	0.00	850000.00	40824762.00
	Age_Group_1	0.14	0.35	0.00	0.00	1.00
	Age_Group_2	0.61	0.49	0.00	1.00	1.00
	Age_Group_3	0.25	0.43	0.00	0.00	1.00
	Age_Log	3.96	0.26	2.89	3.99	4.36
	Sex_Dummy	0.34	0.47	0.00	0.00	1.00
Loan-related	Repayment_Plan	17.38	12.67	0.00	22.55	49.07
	Weighted_Average_Interest_Rate	0.03	0.01	0.00	0.03	0.10
	LTV	0.48	0.21	0.00	0.50	1.00
	PD	0.01	0.05	0.00	0.00	1.00
	Credit_Card_Dummy	0.13	0.34	0.00	0.00	1.00
	Deposit_Dummy	0.92	0.27	0.00	1.00	1.00
	Customer_Relationship_Duration	105.78	89.79	0.00	82.00	502.00
	Customer_Relationship_Duration_Log	4.11	1.28	0.00	4.41	6.22
	Repayment_Loan_Balance_Log	9.99	6.59	0.00	13.91	17.84
	Boligkreditt_Balance_Log	3.30	5.87	-4.61	0.00	17.37
	Boligkreditt_Credit_Limit_Log	3.55	6.22	0.00	0.00	17.37
	Seniorlaan_Balance_Log	1.10	3.76	-1.71	0.00	16.64
	Other_Internal_Debt_Log	0.04	0.66	0.00	0.00	13.88
	Other_External_Debt_Log	5.48	6.27	0.00	0.00	17.62
	DTI	29.18	7270.54	0.00	2.94	2623562.50
	PD_LTV	0.01	0.03	0.00	0.00	0.99
Geographic	Agder	0.01	0.10	0.00	0.00	1.00
	Innlandet	0.01	0.11	0.00	0.00	1.00
	Møre og Romsdal	0.01	0.12	0.00	0.00	1.00
	Nordland	0.01	0.08	0.00	0.00	1.00
	Oslo	0.32	0.47	0.00	0.00	1.00
	Other_County	0.00	0.07	0.00	0.00	1.00
	Rogaland	0.05	0.21	0.00	0.00	1.00
	Svalbard	0.00	0.01	0.00	0.00	1.00
	Troms og Finnmark	0.01	0.12	0.00	0.00	1.00
	Trøndelag	0.10	0.31	0.00	0.00	1.00
	Vestfold og Telemark	0.04	0.19	0.00	0.00	1.00
	Vestland	0.06	0.23	0.00	0.00	1.00
	Viken	0.37	0.48	0.00	0.00	1.00
External and macroeconomic	Unemployment_Rate	4.14	0.59	3.10	4.00	5.50
	Policy_Rate	0.90	0.60	0.00	0.75	2.62
	Housing_Price_Index	254.40	38.62	189.42	258.02	327.72
	Consumer_Price_Index_Yearly_Change	2.60	1.48	0.20	2.20	7.50
	Interest_Rates_Banks_Mortgage_Companies	3.30	0.75	2.13	3.06	4.55

The above table presents the different variables: mean, standard deviation, minimum, median and maximum. This gives a better understanding of the size of values we have within each variable.

D Visual and descriptive representation of the folds

Table 9: Visual representation of the folds



The above table gives a visual representation of our Walk Forward testing approach. Blue = train set (3 years), and Green = test set (1 year), creating one full fold within a time period of 4 years, which is replicated 8 times, increasing by one year for each fold in both the train and test set.

Table 10: Descriptive statistics of the folds

	Train			**Test**			**Total**	
	Observations	**Churned**	**Ratio**	**Observations**	**Churned**	**Ratio**	**Observations**	**Churned**
Fold 1	63 877	641	1.00 %	30 512	285	0.93 %	94 389	926
Fold 2	79 420	818	1.03 %	37 611	398	1.06 %	117 031	1 216
Fold 3	92 991	951	1.02 %	41 601	468	1.12 %	134 592	1 419
Fold 4	109 724	1 151	1.05 %	41 768	523	1.25 %	151 492	1 674
Fold 5	120 980	1 389	1.15 %	88 765	1 049	1.18 %	209 745	2 438
Fold 6	172 134	2 040	1.19 %	95 466	1 061	1.11 %	267 600	3 101
Fold 7	225 999	2 633	1.17 %	103 700	1 026	0.99 %	329 699	3 659
Fold 8	287 238	3 136	1.09 %	109 179	677	0.62 %	396 417	3 813

The above table further describes the Walk Forward testing approach visual representation in Table 9. This table presents the number of observations within each fold and the number of churned customers within. A Ratio column is given to better understand the ratio between the number of churned and the number of observations.

E Hyperparameter XGBoost

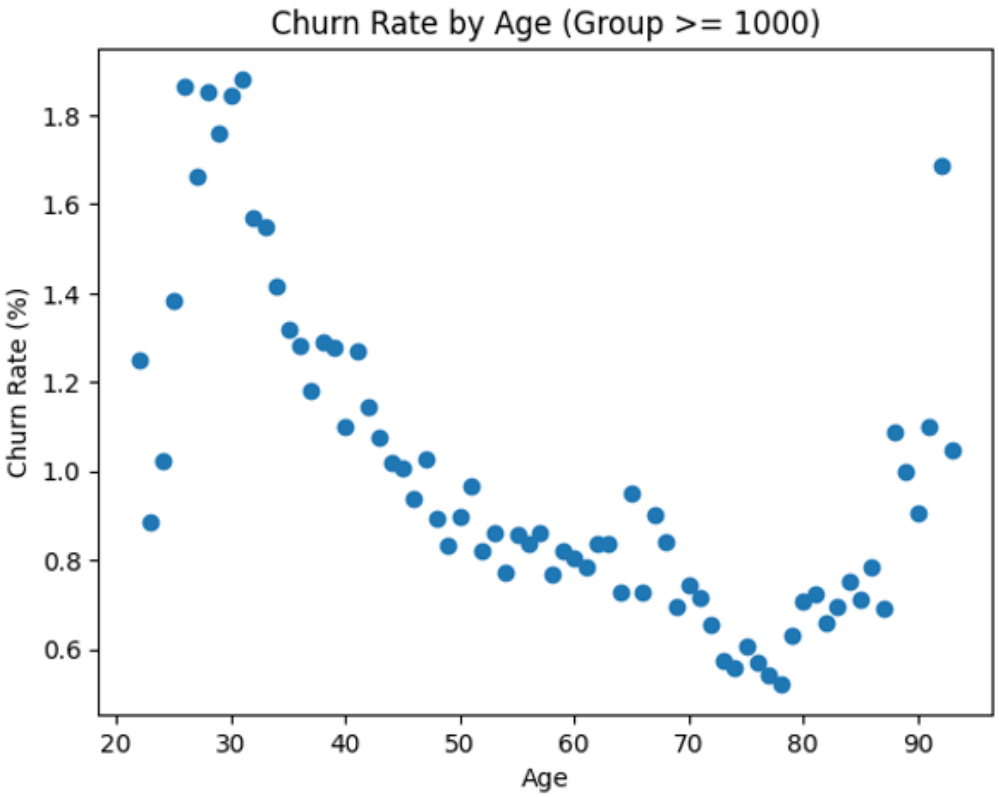
Table 11: Tuned hyperparameters

Fold	$\boldsymbol{\nu}$	$\boldsymbol{\psi}$	$\boldsymbol{\gamma}$	$\boldsymbol{J}$	$\boldsymbol{w}$	**max_depth**
1	0.024	0.800	0.154	100	9	2
2	0.032	0.599	0.058	50	9	2
3	0.103	0.500	0.000	96	10	2
4	0.222	0.657	0.898	100	1	4
5	0.025	0.698	0.000	50	1	4
6	0.083	0.587	0.441	56	1	2
7	0.088	0.500	0.000	50	8	4
8	0.063	0.796	0.191	78	3	2

The above table presents the tuned hyperparameters for each fold in the XGBoost models.

F Visualisation of churn rate by age

Figure 6: Churn rate by age pre-dataprocessing

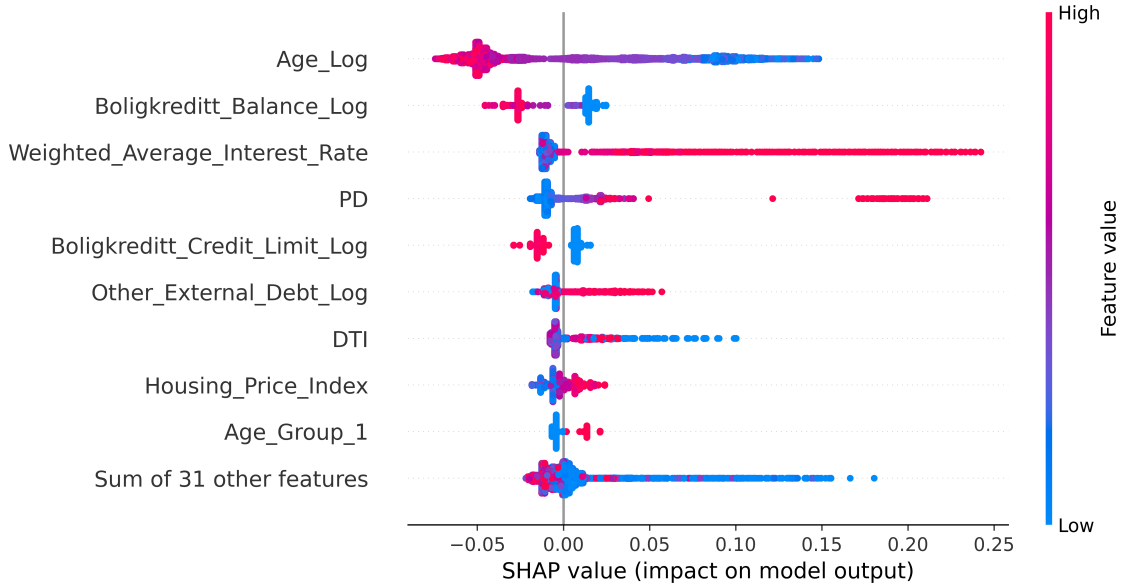


The above figure presents the churn rate by age given through our raw dataset. The churn is given as a percentage of churn within their respective age.

B Appendix

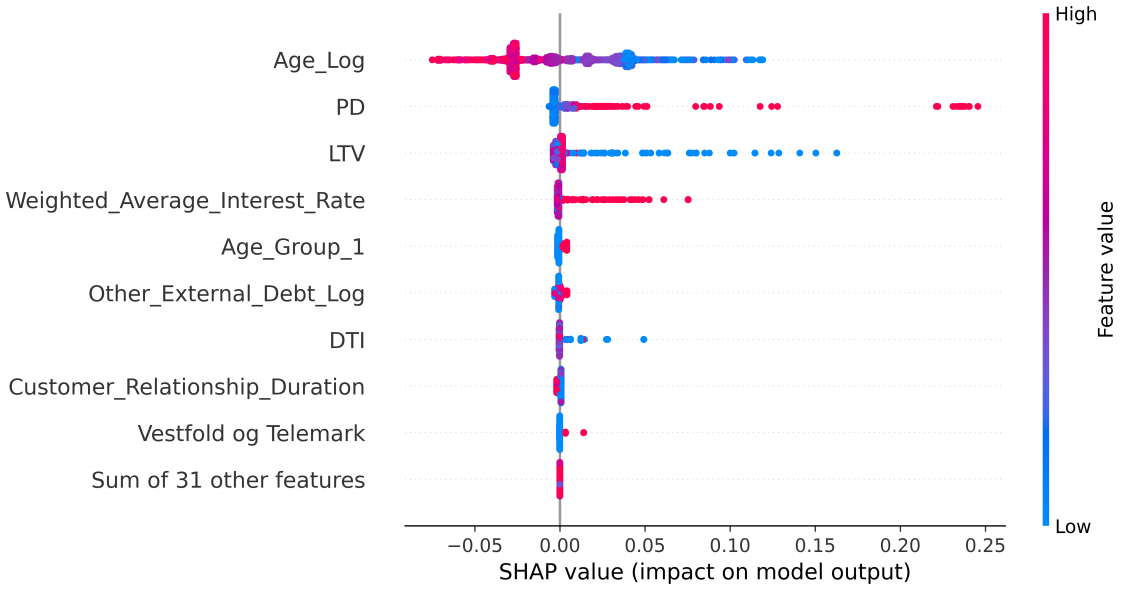
A SHAP Beeswarm plots of folds 1 - 8

Figure 7: Fold 1: Beeswarm Plot



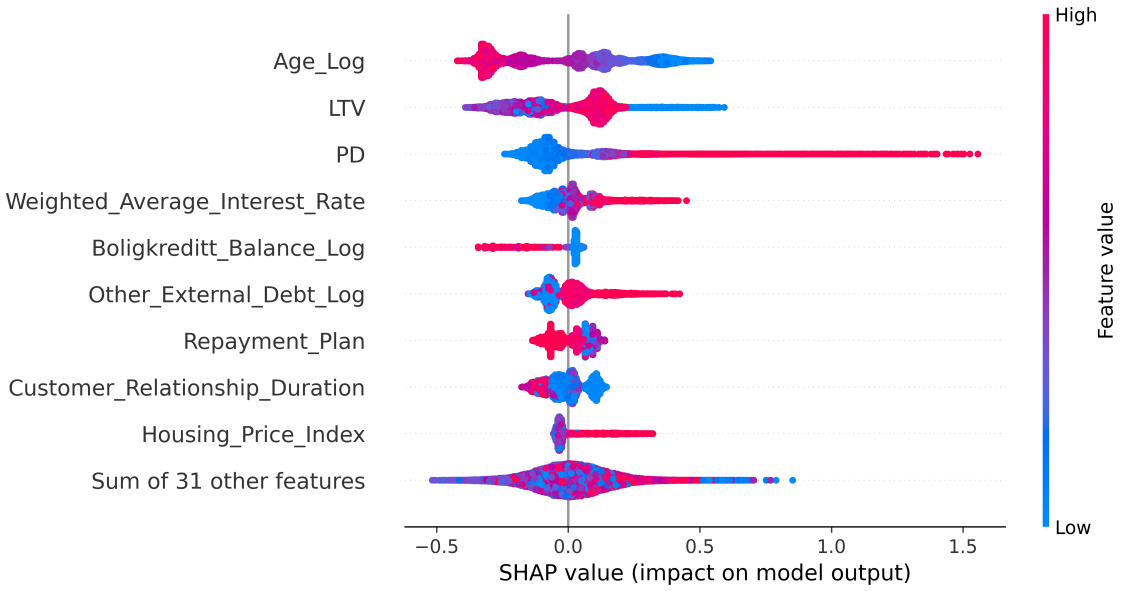
The above figure is a SHAP beeswarm visualisation of fold 1. The beeswarm plots introduce the SHAP value and their impact on the model output horizontally, with negative SHAP values indicating a reduced likelihood for churn and the opposite for positive SHAP values. In addition to the horizontal axes, the beeswarm plot introduces a vertical axis representing the size of each variable's value (This applies to the other following beeswarm plots).

Figure 8: Fold 2: Beeswarm Plot



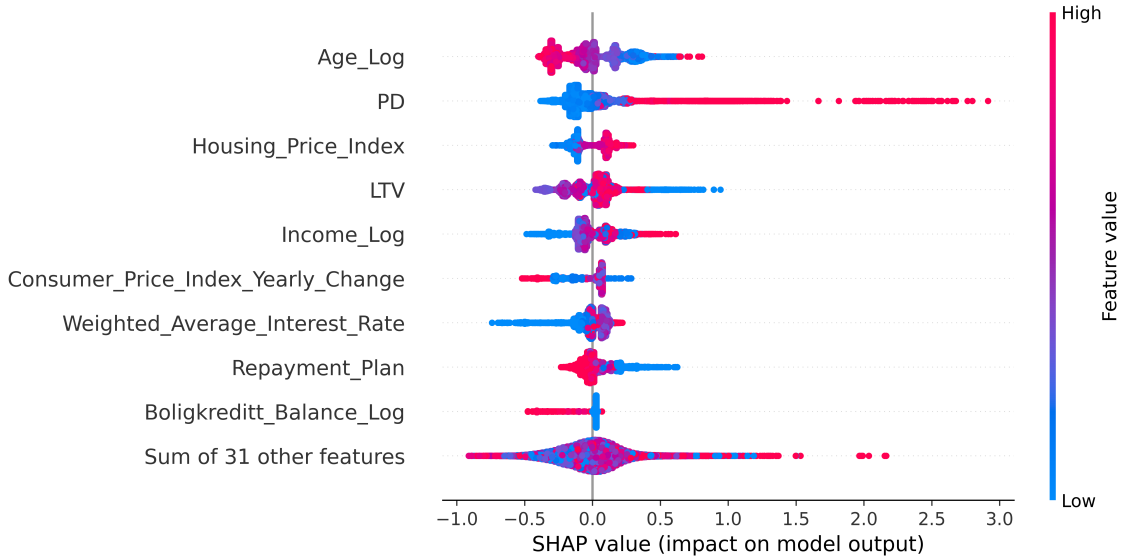
The above figure is a SHAP beeswarm visualisation of fold 2. To better understand the plot, look at the caption to Figure 7.

Figure 9: Fold 3: Beeswarm Plot



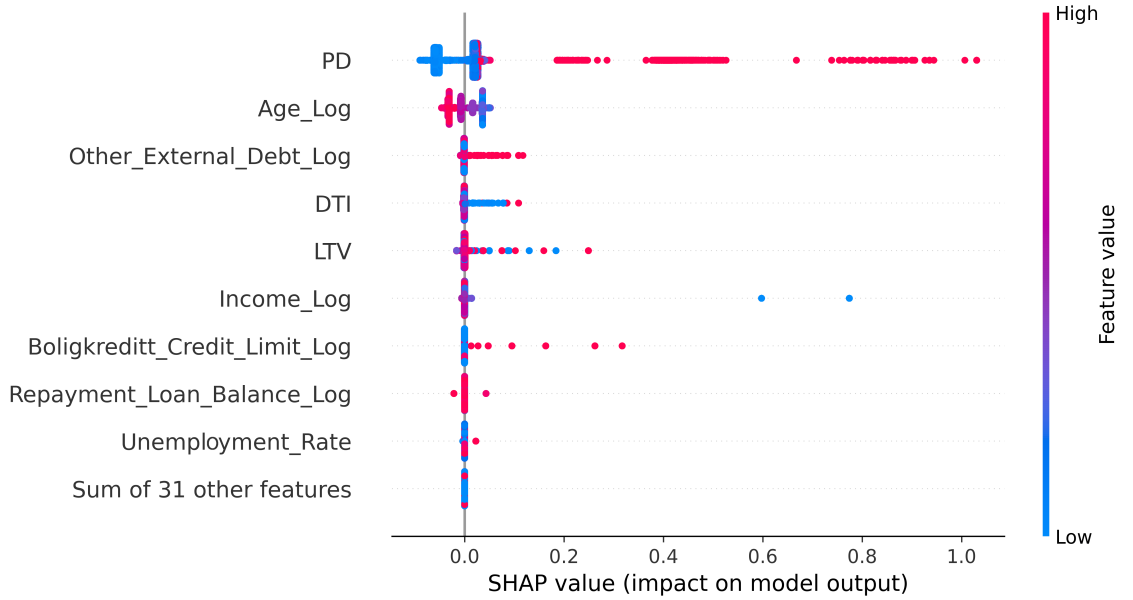
The above figure is a SHAP beeswarm visualisation of fold 3. To better understand the plot, look at the caption to Figure 7.

Figure 10: Fold 4: Beeswarm Plot



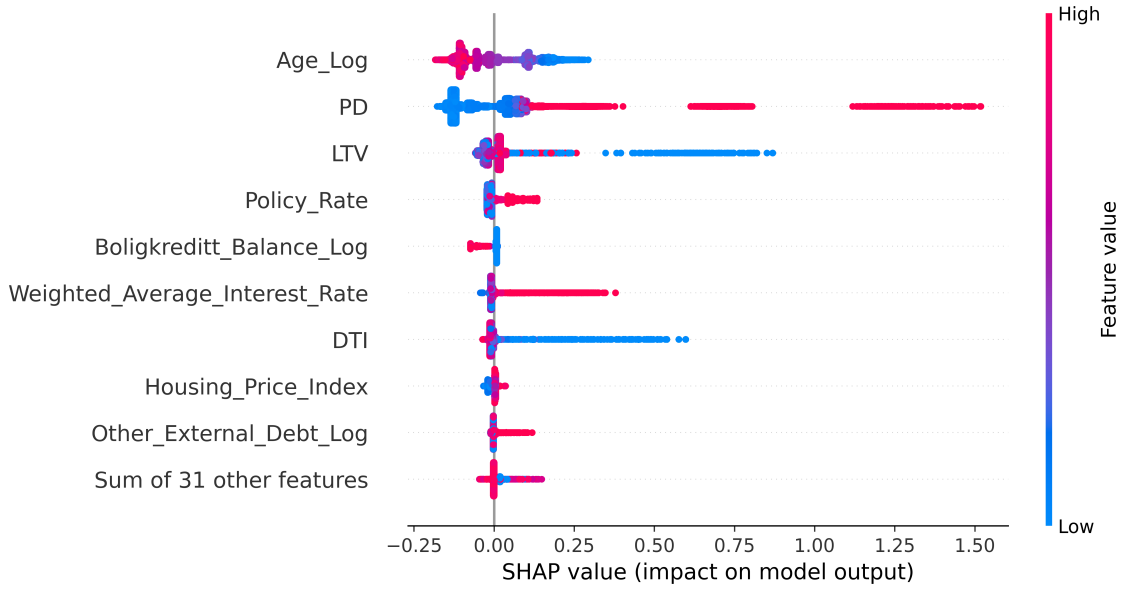
The above figure is a SHAP beeswarm visualisation of fold 4. To better understand the plot, look at the caption to Figure 7.

Figure 11: Fold 5: Beeswarm Plot



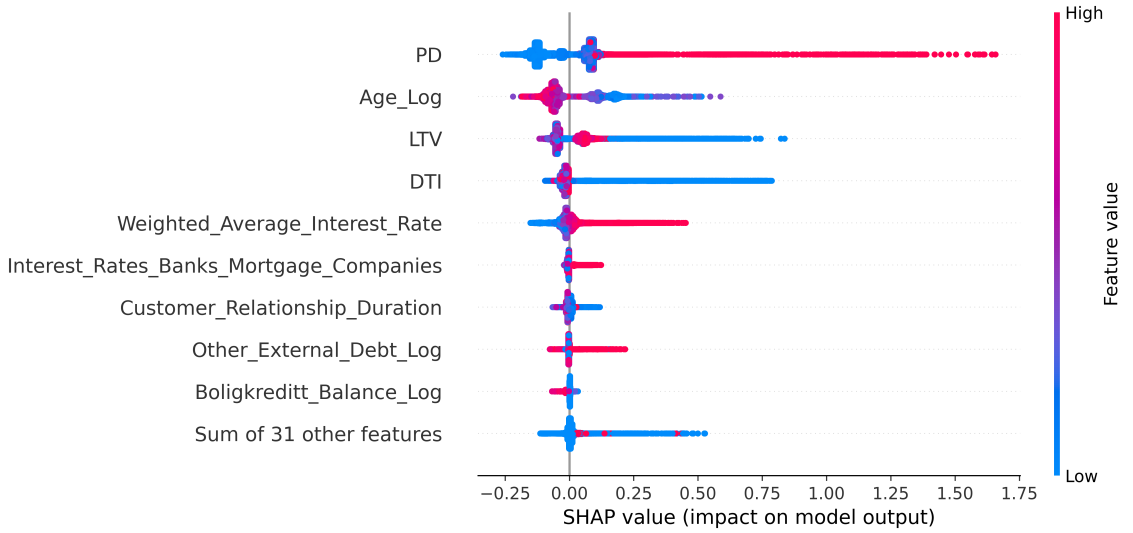
The above figure is a SHAP beeswarm visualisation of fold 5. To better understand the plot, look at the caption to Figure 7.

Figure 12: Fold 6: Beeswarm Plot



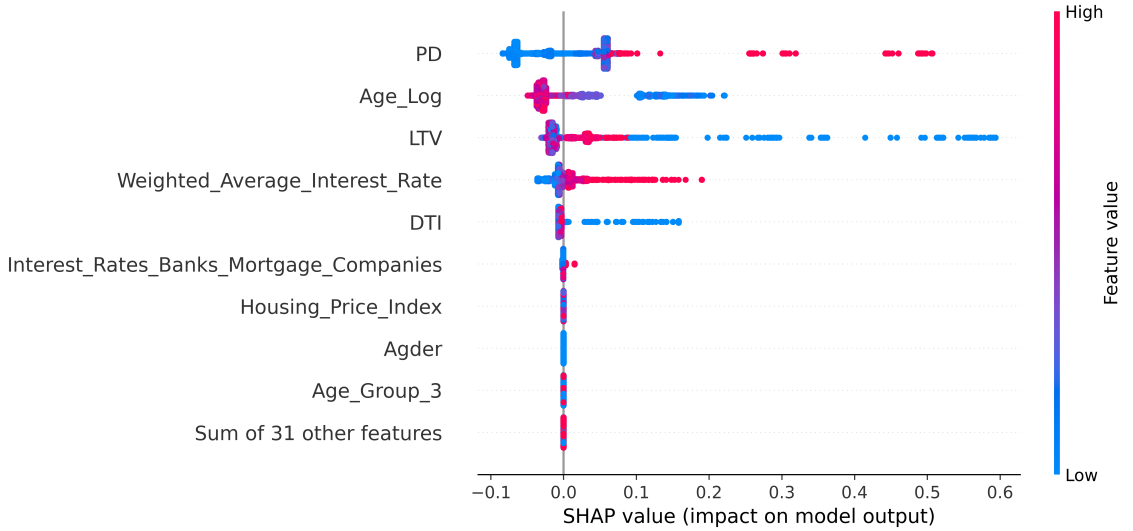
The above figure is a SHAP beeswarm visualisation of fold 6. To better understand the plot, look at the caption to Figure 7.

Figure 13: Fold 7: Beeswarm Plot

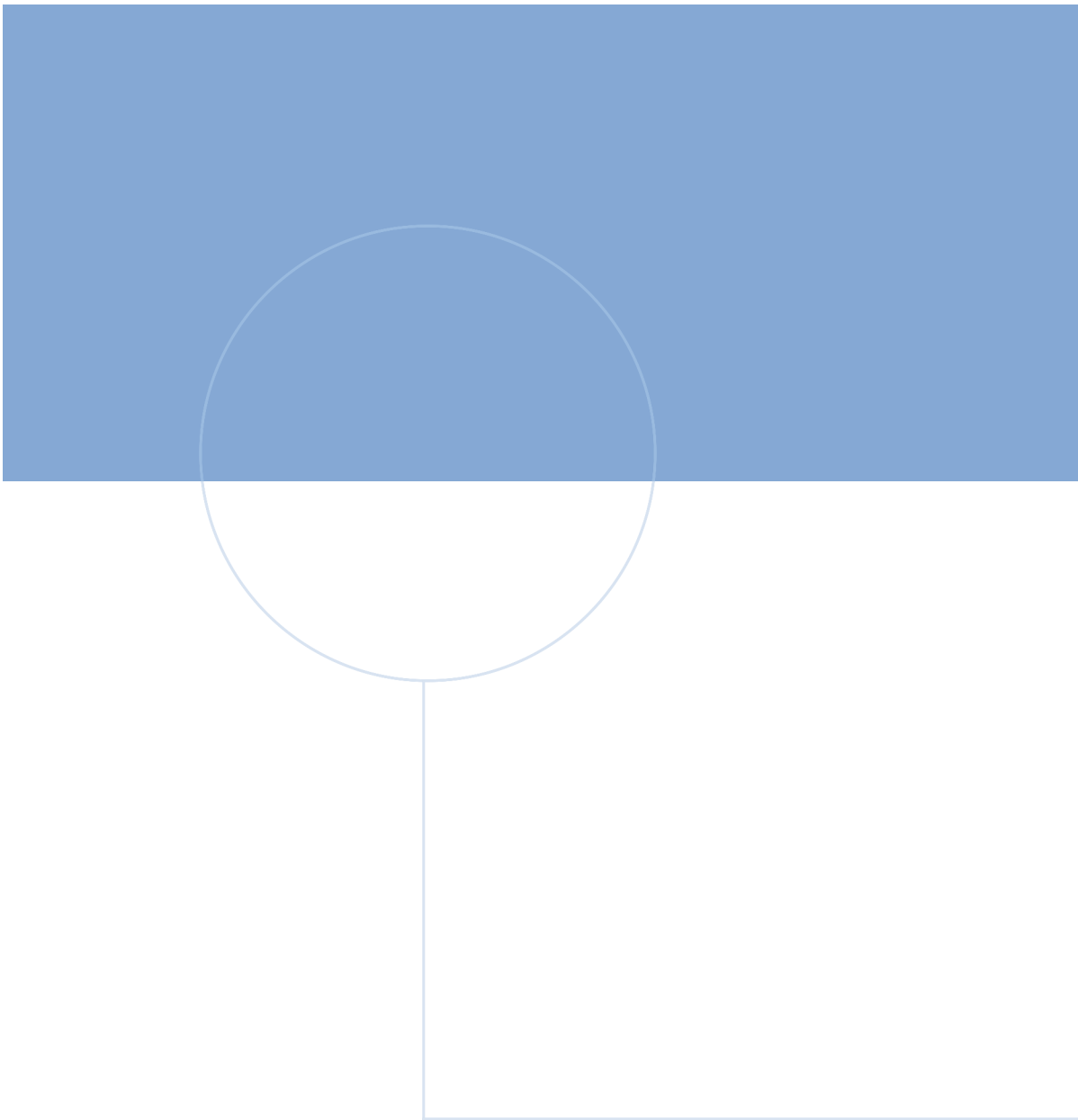


The above figure is a SHAP beeswarm visualisation of fold 7. To better understand the plot, look at the caption to Figure 7.

Figure 14: Fold 8: Beeswarm Plot



The above figure is a SHAP beeswarm visualisation of fold 8. To better understand the plot, look at the caption to Figure 7.



NTNU

Norwegian University of
Science and Technology