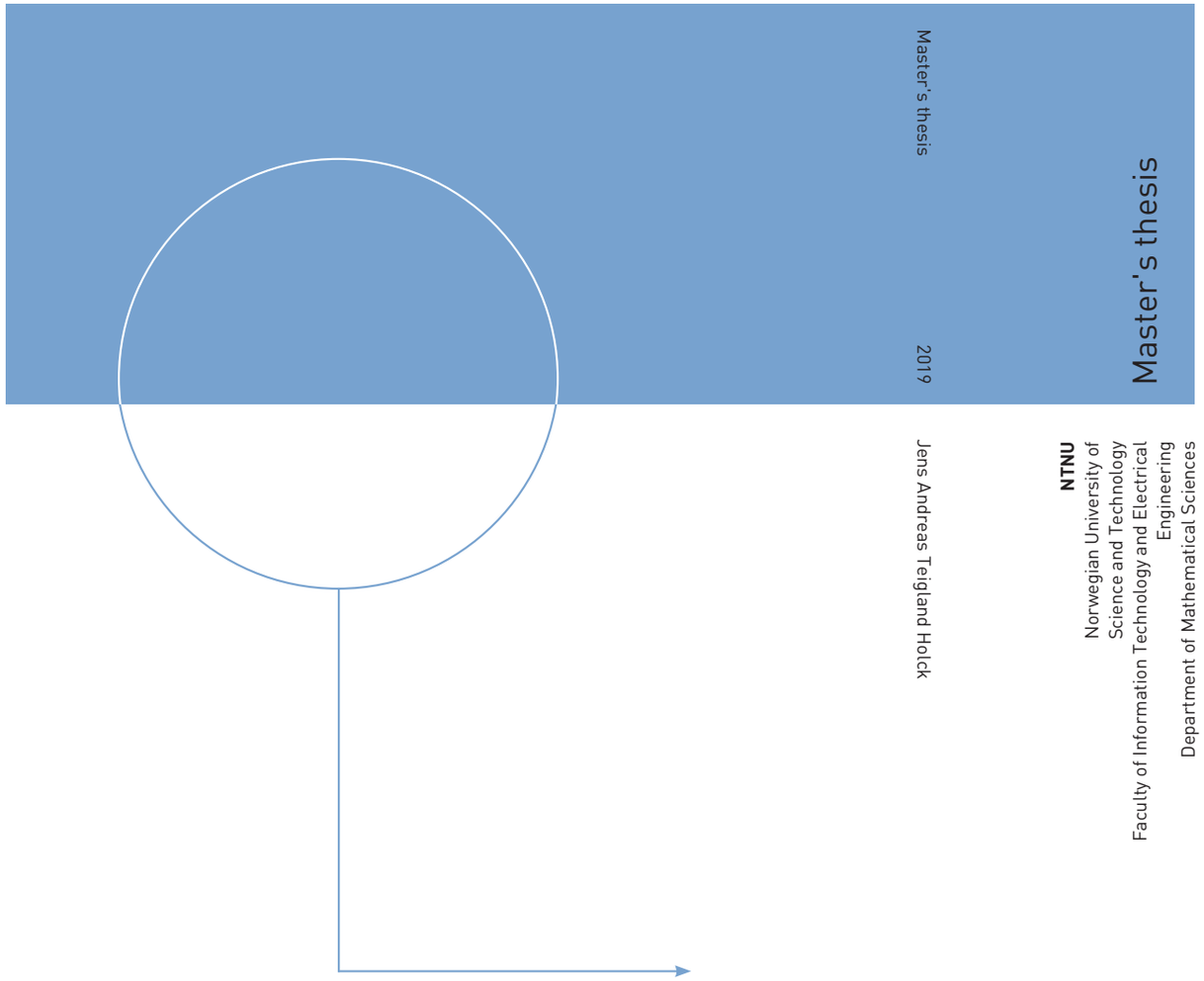


Master's thesis

2019

Jens Andreas Teigland Holck

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

Jens Andreas Teigland Holck

Forecasting the Total Balance Sent to Debt Collection

June 2019

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU

Norwegian University of
Science and Technology

Forecasting the Total Balance Sent to Debt Collection

Jens Andreas Teigland Holck

Master of Science in Applied Physics and Mathematics
Submission date: June 2019
Supervisor: John Sølve Tyssedal

Norwegian University of Science and Technology
Department of Mathematical Sciences

Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) in the field of statistics. The thesis was written at the Department of Mathematical Sciences during the spring semester of 2019 in cooperation with SpareBank 1 Kredittkort AS. It is assumed that the reader is familiar with concepts in statistics, particularly generalized linear mixed models and time series theory.

Thank you to Christian Meland, Jens Morten Nilsen and Christer Dale at SpareBank 1 Kredittkort AS for the opportunity to write this thesis. I would like to express my sincerest gratitude to my supervisor John Sølve Tyssedal for our conversations and his support and guidance throughout the process. Further thank you to my sister, my brother and my girlfriend. Finally, I would like to thank my parents for bringing me into this world.

Trondheim, 10.06.2019,
Jens Andreas Teigland Holck

Abstract

Credit cards have long been a popular form of payment method and the usage is growing worldwide. Credit card companies must constantly assess the risk of money lending and one way to measure risk in conjunction with credit card spending is to track the number of debt collection cases over time. Accurate predictions of the number of customers sent to debt collection in the future is therefore of great interest to credit card companies. The aim of this thesis was to forecast the total balance sent to debt collection each month for the year 2019 based on historical data provided by SpareBank 1 Kredittkort AS in the time period July 2017 to September 2018. Additionally, we aimed to determine factors that make some costumers more prone to delinquency in general.

The data provided was longitudinal with repeated measurements each month for more than 500 000 credit card customers in Norway. A mixed effects logistic regression model was made and used as a classifier to determine whether a customer is sent to debt collection in a given month. The model was then used to classify and count the number of debt collection cases in a month. Multiplying with the average amount a customer owes gave predictions for total balance sent to debt collection each month. The data was highly imbalanced since most customers are not sent to debt collection. We used random undersampling as well as a method to adjust the outputs of the classifier. The model forecasted an increase in the total balance sent to debt collection for the year 2019. A possible improvement of the model would be to collect additional personal information about each customer that could possibly explain why some customers are more prone to delinquency. These may include a customer's monthly income, other types of unsecured debt and marital status.

Sammendrag

Kredittkort har lenge vært en populær betalingsmetode og bruken vokser globalt. Kredittkortselskaper må stadig vurdere risikoen forbundet med pengeutlåning og en måte å måle risiko i forbindelse med kredittkortutgifter er å spore antall inkassosaker over tid. Nøyaktige prediksjoner over antall kunder som blir sendt til inkasso i fremtiden vil derfor være svært interessant for kredittkortselskaper. Formålet med denne avhandlingen var å predikere den totale balansen som blir sendt til inkasso hver måned for året 2019, basert på historiske data fra SpareBank 1 Kredittkort AS i tidsperioden juli 2017 til september 2018. I tillegg ønsket vi å bestemme hvilke faktorer som gjør at noen kunder er mer utsatt for mislighold.

Datasettet var longitudinelt med repeterende observasjoner hver måned for mer enn 500 000 kredittkortkunder i Norge. En mixed effects logistisk regresjonsmodell ble konstruert og brukt som en klassifikator for å bestemme hvorvidt en kunde blir sendt til inkasso i en gitt måned. Modellen ble deretter brukt til å klassifisere og telle antall inkassosaker i en måned. Ved å multiplisere med gjenomsnittsbalansen en kunde skylder ga dette prediksjoner for den totale balansen sendt til inkasso hver måned. Datasettet var svært ubalansert siden de fleste kundene ikke blir sendt til inkasso. Vi brukte tilfeldig undersampling i tillegg til en metode for å justere outputen til en klassifikator. Modellen predikerte en økning i den totale balansen sendt til inkasso for året 2019. En mulig forbedring av modellen vil være å samle tilleggsinformasjon for hver kunde som kan forklare hvorfor noen kunder er mer utsatt for mislighold. Dette kan blant annet være en kundes månedlige inntekt, andre typer usikret gjeld og sivilstatus.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The idea of extending credit on durable goods goes all the way back to the 1800s. Already in the 1920s, some department stores started to offer "courtesy cards" that resembled the credit cards we know today (Evans and Schmalensee, 2005). Today, it is safe to say that the invention of the credit card has been a huge success. For example, in the United States, 75.7% of all consumers possess at least one credit card and the usage as a form of payment is growing worldwide (Greene et al., 2015).

Credit card companies earn their profit in three different ways. First, credit card companies charge stores somewhere between 2% and 3% for each credit card purchase. With billions of transactions daily, this amounts to a huge revenue. Secondly, there are additional credit card fees, such as annual fees that customers must pay to keep their accounts open. Finally, a large portion of consumers do not pay their bills in full each month. The costumers' unpaid credit cards begin to incur interest at rates that are very high. This can be very lucrative for the bank. However, customers that repeatedly fail to pay their bills will eventually be sent to debt collection. It is this group of customers that will be of main interest in this thesis.

Every month, credit card companies will send some customers to debt collection if they have been unable to pay their billings on time. We will refer to these customers as *impaired* customers. The number of impaired customers the credit card company must send to debt collection each month may vary from month to month depending on several factors. These may include seasonal variations, the current unemployment rate and the economy as a whole. The aim of this thesis is to forecast the total balance that SpareBank 1 Kredittkort AS will send to debt collection each month for the year 2019 based on historical data provided by the same company from the time period July 2017 to December 2018. The total balance sent is the sum of all the debt owed by impaired customers. Denote the month December 2018 as T and the balance sent to collection this month as B_T, we thus want to predict

$$B_{T+h} \text{ for } h = 1, 2, \ldots, 12.$$

1.1 The Debt Collection Process

This section describes the process of how a customer is sent to debt collection. After the company issues a credit card to a customer, the he or she receives the credit card transactions billed in what is known as billing cycles. A billing cycle is usually one month. The statement date is the last date of each billing cycle. Finance charges are calculated and added to the customer's balance and the billing statement is prepared. The customer is then given a fixed number of days to pay before the payment due date. Varying from country to country and company to company the period between the statement date and due date is often 14 days or 21 days. Most customers pay their billing in full while others choose to revolve their balance. A revolving customer is one who does not pay the total amount he or she owns at the end of a billing cycle. These types of customers are, for the most part, lucrative for the credit card company as they will have to pay additional fees. The extra amount charged for revolving the balance depends on the total size of the balance and the interest rate of the card. However, if the customer does not pay the minimum amount required before the due date, he or she will receive a dunning on their next statement date. 14 days after this the credit card company will send a due date dunning. If the customer still fails to pay, the customer will receive a statement date collection notice and then a due date collection notice. If the customer still fails to pay, the customer is sent to debt collection. The whole process described is outlined in the timeline shown in figure 1.1.

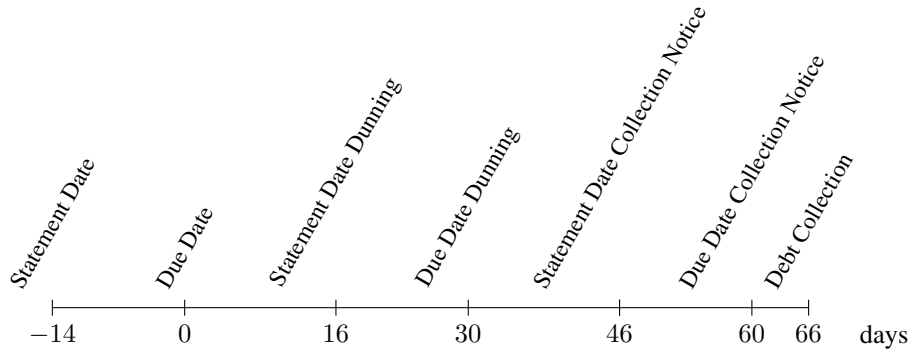


Figure 1.1: A timeline of the debt collection process.

Every month, the credit card company will send some customers to debt collection. Some customers may pay right away, and the debt collection case is removed. This will not be the case for all customers. The credit card company often hires a collection agency to collect their debt for them. Every month, a portion of the total balance sent to collection by the credit card company represent direct losses for the company. Typically, the whole debt collection process usually takes somewhere between 60-90 days for a customer.

1.2 Motivation

For a credit card company, it will be a great advantage to know how the total balance sent to collection will vary each month, especially when creating budgets and projections as they will hopefully be more accurate. It is imperative for the credit card company to always monitor the risk of their portfolio of customers. Credit card companies must walk a thin line. On one hand, it is very beneficial for the company to keep customers who are overdue with their bills as they must pay additional interests and fees. On the other hand, if the risk is too high, customers are not able to pay back what they owe, and this can result in direct losses to the bank. Therefore, credit card companies constantly track risk, which can be measured in several ways. One way is to track the number of customers sent to debt collection. Knowing as early as possible whether or not the number of debt collection cases will increase, or decrease will be a huge advantage for the credit card company. If the credit card company predicts that the number of cases will increase, the company can implement measures at an early stage to counteract this development. For instance, the company may offer re-financing, tighten rules for issuing new cards or reduce the credit limit on the cards issued. Similarly, if the company expects the number of debt collection cases to decrease, they may invest in more advertising or ease the rules for issuing new credit cards.

1.3 The Data Set

The data set is supplied by SpareBank 1 Kredittkort AS, an alliance of 14 different Norwegian banks that provides credit cards to Norwegian consumers. The data set contains information for more than 500 000 customers in Norway in the time period July 2017 to September 2018. That is, for each customer, we have recorded data for a total of 15 months from July 2017 to September 2018. Each row contains recorded information for one customer for one month. Not all customers have 15 registered rows of information, if they for instance became a customer later or have terminated their credit card. In total, the data set contains 8 769 272 rows and 79 columns. Appendix A contains a full list of the column names and a description of each variable. Table 1.1 illustrates the data set and includes a few examples of different customers, as well as a few examples of some variables in the data set. The first column shows each customer's BK_ACCOUNT_ID. This is an ID number

Table 1.1: A detailed explanation of the data set.

BK_ACCOUNT_ID	YearMonth	...	CustomerAge	GENDER_NAME	...	SUM_of_Payment-OverDueFlag	...	DCA0YearMonth	DCA0Ind	BalanceSent
42	201707	...	65	Male	...	0	...	NA	0	0.00
42	201708	...	65	Male	...	0	...	NA	0	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
42	201809	...	66	Male	...	0	...	NA	0	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1521	201707	...	44	Female	...	2	...	NA	0	0.00
1521	201708	...	44	Female	...	2	...	201711	1	47067.67
1521	201709	...	44	Female	...	2	...	NA	1	47067.67
1521	201711	...	44	Female	...	2	...	NA	1	47067.67
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1521	201809	...	45	Female	...	0	...	NA	0	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1542168	201809	...	26	Male	...	0	...	NA	0	0.00

that every customer receives once they acquire a credit card. The BK_ACCOUNT_ID number assigned to a customer is randomly assigned and has no meaning apart from separating

customers. YearMonth shows the month in which all the data was recorded. Note that the data is recorded for the last day of the month. Next, the variables CustomerAge and SUM_of_PaymentOverDueFlag shows the customer's age in the current month and the total number of times the customer has been overdue with his or hers payment in the last 12 months, respectively. The variable DCA0YearMonth shows the month that the customer is sent to debt collection, which is always YearMonth + 3 months. The data set is arranged in this manner such that the response, DCA0Ind, is whether a customer is sent to debt collection 3 months ahead. Table 1.1 also includes some examples of different types of customers. First, the customer with BK_ACCOUNT_ID 42 is a typical example of a customer who either does not use his credit card or pay his billings on time. He is not sent to debt collection for any of the months. The customer with BK_ACCOUNT_ID 1521, however is sent to debt collection in November 2017. Notice that this is registered in the row where YearMonth is August 2017, i.e. three months before the customer was sent to debt collection, YearMonth + 3. The customer with BANK_ACCOUNT_ID 1521 is out of the system for three months, and this is indicated in the response variable DCA0Ind with three 1s in a row. This does not mean, however, that the customer was sent to debt collection in September and October 2017 as well. This issue will be handled in the data preprocessing. Some impaired customers will be permanently dismissed by the credit card company, while others return as customers which is the case for the female with ID 1521. Finally, the customer with BK_ACCOUNT_ID 1542168 is a brand new customer that joined in September 2018 and we therefore only have one observation for this customer. This is to illustrate that the number of observations for each customer is not necessarily the same. Note also that a variable does not need to change for a customer for all observations. Finally, sensitive information such as names and addresses are not included in the data set.

1.4 Visualization of the Data

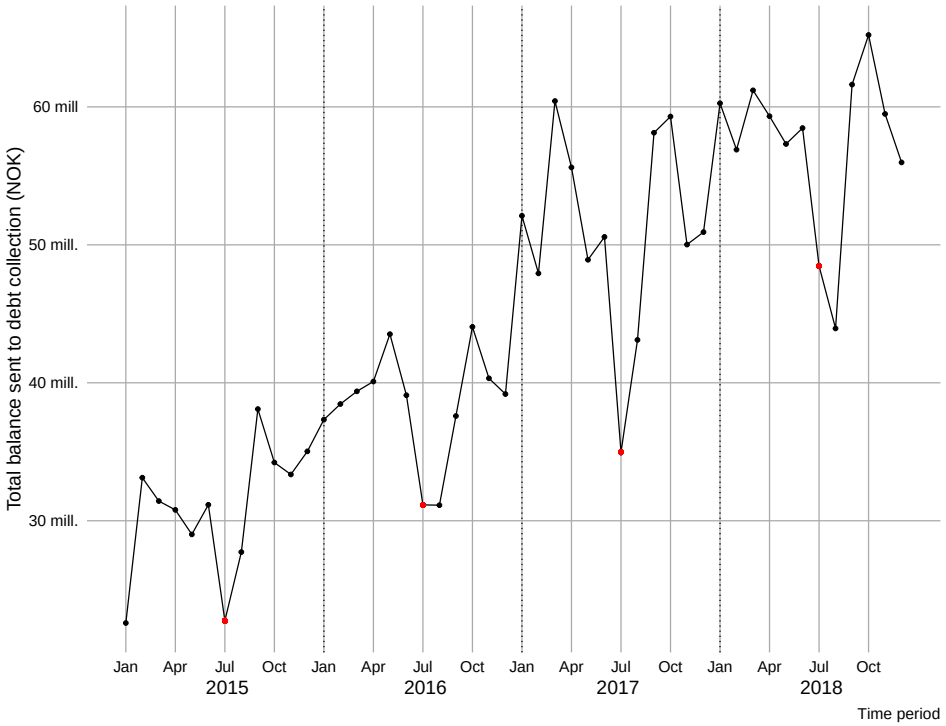
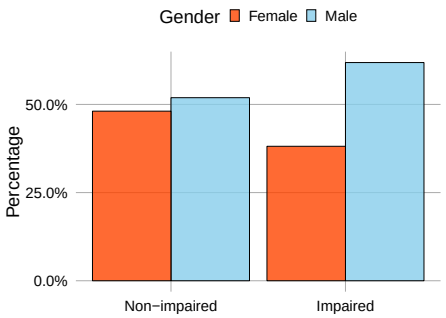


Figure 1.2: The total balance sent to debt collection each month from January 2015 to December 2018. The total balance has generally increased in the time period. The red dots show the drop in total balance sent to collection each July, which we refer to as the *July effect*.

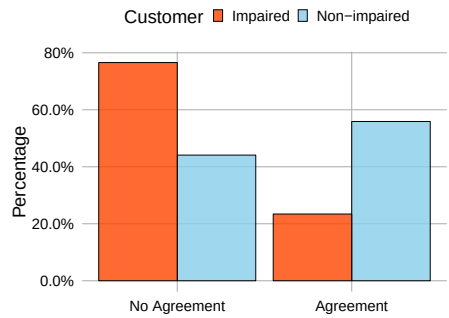
This section is meant to familiarize the reader with the data set. Figure 1.2 shows the total balance sent to collection each month starting in January 2015 and ending in December 2018. The total balance sent to collection ranges from approximately 22 million NOK in January 2015 and up to over 60 million NOK for several months in 2018. There seems to be a clear increasing trend, even though there is great variation in the total balance sent to collection. Furthermore, there seems to be seasonal trends as well. The points marked red in figure 1.2 show that the total balance sent to collection is heavily reduced in July. (This was also the case for July 2018 although the total balance sent to debt collection in August 2018 was even lower than July). A very likely explanation for this is that Norwegian workers receive *feriepenger*, (directly translated as holiday money), usually in June that can be used as down payment on debt. We will refer to this as the *July effect*. Looking at each year individually, there are also high spikes in the spring months, March, April and May, except for the first year. However, the explanation for this is more unclear.

Figure 1.3 aims to visualize impaired customers and how they compare to non-impaired

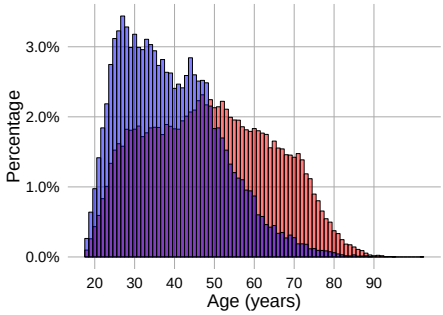
customers. 1.3(**a**) shows that men are more likely to be sent to debt collection than women, with 61.9% of those impaired in the time period July 2017 - December 2018 being male. A possible explanation might be that women are more risk averse than men and that men can be overconfident and are less afraid to take on more debt than women (Croson and Gneezy, 2009). 1.3(**b**) compares customers that have an e-Statement agreement with the credit card company to those who have not. It is generally considered responsible to have this since the customer receives an electronic reminder to pay their bills and most customers find it convenient and easy to use. 1.3(**b**) shows that impaired customers are less likely to have an e-Statement agreement, only 23.4% have it compared to 55.8% of those not sent to debt collection. The blue bars represent impaired customers and the red bars non-impaired customers in (**c**)-(**f**). The age of impaired customers spans from 18 to 89 years old as shown in 1.3(**c**). The bin width is one year. There is a spike of impaired customers at the age of 27 before it starts to decrease. Surprisingly, there is a new peak at 45 years. A possible explanation is that this is a typical age where many people reestablish themselves and builds new relations and with that often comes financial trouble. Another explanation could be that some customers find themselves in a midlife crisis and make irrational financial decisions such as buying a new boat or car. After the age of 45, there is a sharp decrease as most people tend to have a more spacious economy in their 60s and 70s with fewer large expenses as well. 1.3(**d**) shows how long customers possess their credit card before they are impaired. A large number of customers are sent to debt collection fairly quickly. This means that some customers should probably not have been given a credit card in the first place. 1.3(**e**) shows which score the impaired customers have. The score variable is a risk score computed and assigned to each customer with a value between 0 and 7 based on how they use their credit card. A score of 0 indicates a very low risk of delinquency and a score of 7 indicates a very high risk of delinquency. The majority of impaired customer have a high score, although surprisingly few customers have a score of 7. This could possibly be a result of that reaching a score of 7 is very difficult. Over 50% of customers that are non-impaired have a score of 2. Finally, 1.3(**f**) shows how much impaired customers owe. The bin width is 10 000 NOK. The large majority owes somewhere around 30 000 NOK, while some owe much more, above 100 000 NOK.



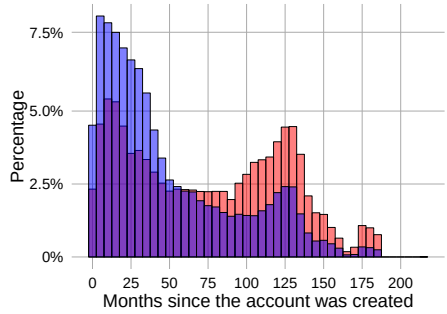
(a) Gender



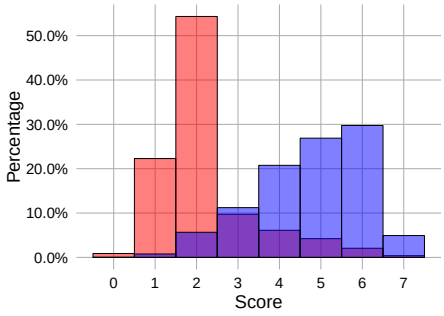
(b) e-Statement agreement indicator



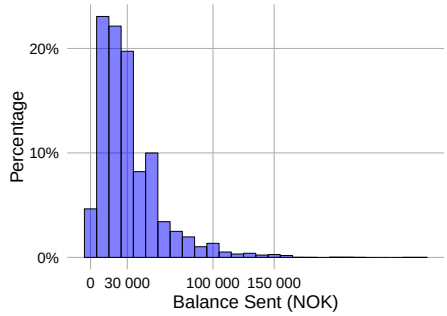
(c) Age of customers



(d) Months since the account was created



(e) Score (value between 0 and 7)



(f) The balance sent to debt collection

Figure 1.3: Visualization of the distribution of impaired customers in terms of gender (**a**), the e-statement agreement indicator (**b**), age (**c**), number of months since the account was created (**d**), score, a value assigned between 0 and 7 (**e**) and the balance owed (**f**).

Additionally, the data is visualized by producing a symmetric correlation plot with all numerical variables in the data set as shown in figure 1.4. The correlation plot is combined

with a significance test that colours all variables that are found to be insignificant white (with a significance value of 0.05). The bar on the right-hand side shows how the colors should be interpreted, with blue indicating positive correlation and red indicating negative correlation. The last row is of most interest as it shows the response variable `DCA0Ind`, which is a binary response equal to 1 if the customer is sent to debt collection and 0 otherwise. Most tiles are white showing no significant correlation with `DCA0Ind`.

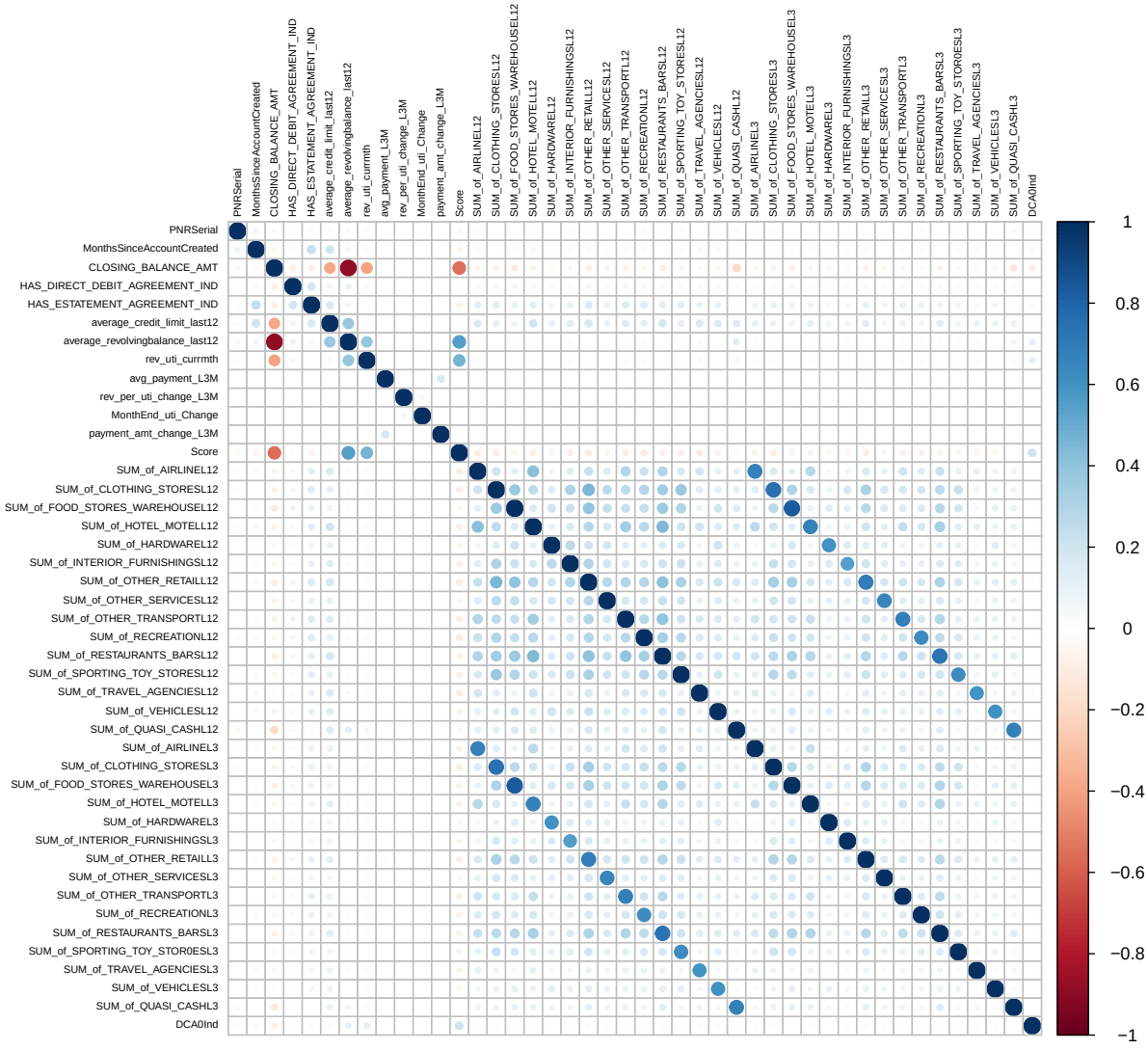


Figure 1.4: Symmetric correlation plot for all the variables in the data set with a significance test coloring all insignificant variables white. The bar on the right-hand side shows how the colors should be interpreted, blue showing positive correlation and red negative correlation.

1.5 Chosen Approach to the Problem

A previous attempt to forecast the total balance sent to debt collection B_T has been done using time series analysis directly on the data (Holck, 2018) shown in figure 1.2, where the predictions for July were modelled as known additive outliers. A different and indirect approach is proposed where we rather model and forecast the total number of customers sent to debt collection for each month, S_T. Notice the similarities between figure 1.2 and 1.5 and how both graphs follow the same pattern. If a model that accurately predicts the

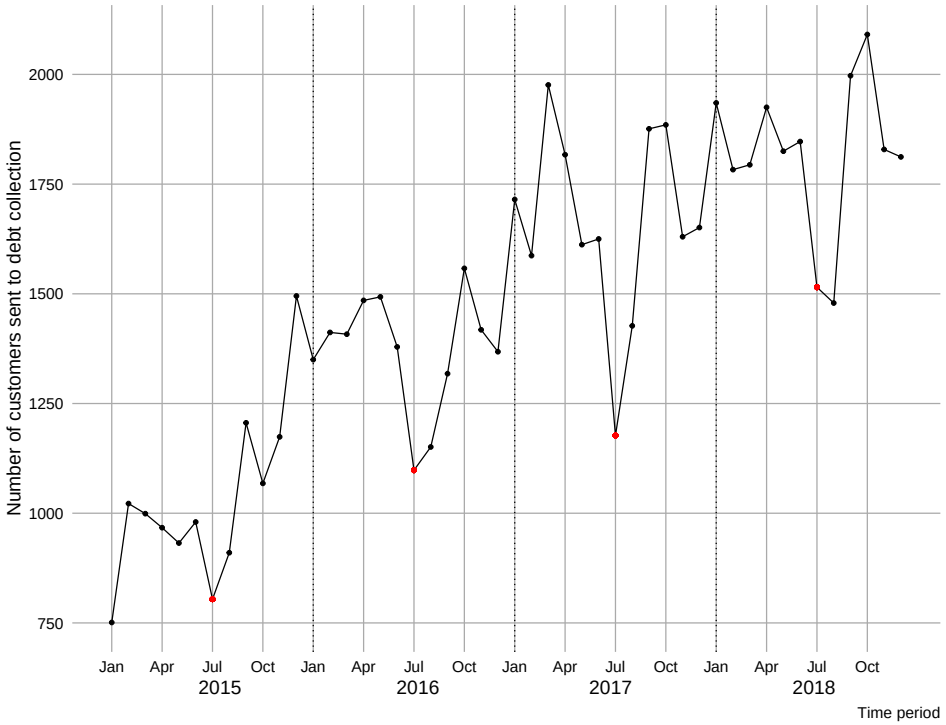


Figure 1.5: The number of customers sent to debt collection each month from January 2015 to December 2018. Notice the red dots illustrating the *July effect* with a reduced number of debt collection cases this month.

total number of impaired customers each month is made, it is a small leap to determine the total balance sent to debt collection for each month. By simply multiplying with the average balance an impaired customer owes, we produce predictions that are very close to the actual balance sent to debt collection, as shown in figure 1.6. (The average balance sent to debt collection chosen is simply the mean of means for each month and is not adjusted for inflation. This is simply to illustrate the high accuracy of the predictions). The solid line shows the actual balance sent to debt collection each month and the dashed line shows the predicted balance sent to debt collection based on the number of customers sent for each month multiplied by the average balance sent to debt collection. The main idea will

be to evaluate a portfolio containing all customers for a given month and based on a trained model on historic data, accurately predict how many in the portfolio will be impaired. We will attempt to create a generalized linear mixed model that can accurately predict how many customers will be sent to debt collection by looking at the current portfolio. A generalized linear mixed model is a reasonable choice since the data is longitudinal as shown in table 1.1. Hopefully, we may capture the random effects for each customer. The response will be a binary one which means we will have a logit connection in our model. The model will therefore be referred to as a mixed effects logistic regression model. Furthermore, the model will be constructed in such a way that given this month's portfolio, the model predicts how many customers will be impaired three months into the future. This is a reasonable justification as the debt collection process is roughly 66-90 days as previously mentioned. A total of 12 mixed effects logistic regression models will be made, one for each month of the year. We will use a variable selection method that extends the well-known LASSO method to be applied to generalized linear mixed models as well, in order to select the explanatory variables that should be included in the mixed effects logistic regression model. For predictions further than three months ahead,

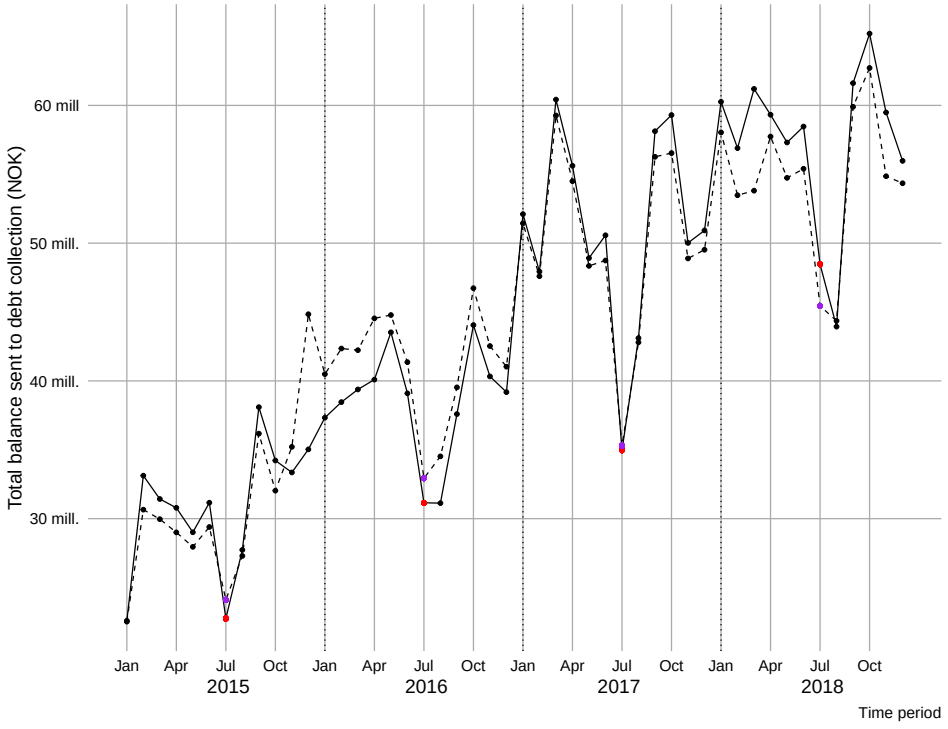


Figure 1.6: The total balance sent to debt collection each month from January 2015 to December 2018 (solid line) with forecasts based on the number of impaired customers in the same time period. The dashed line shows the predicted balance sent to debt collection based on the number of impaired customers that month multiplied by the mean of balances over all months.

we will use various forecasting techniques to predict how the portfolio containing the customers will look like in the future. If we can simulate a likely portfolio for the coming months, we may use our model on these simulated portfolios to accurately predict how many customers who will be impaired. In order to simulate a likely portfolio, we will forecast the chosen explanatory variables for the coming months. This will be done in different ways depending on the explanatory variable in question.

Chapter 2

Theory

This chapter comprises the necessary theory and methods needed to be able to forecast the total balance sent to debt collection each month for the year 2019. The first section considers Generalized Linear Mixed Models, abbreviated as GLMMs. The theory includes how one can estimate parameters for GLMMs and a way to perform variable selection. The GLMM is also specified for a binary response with a logit connection to give what we will refer to as the mixed effects logistic regression model. A section for diagnostic checking of GLMMs is also included as well as how to perform classification with the model. The second section deals with imbalanced data and presents different methods to deal with this issue. The third section contains time series theory and presents most notably an alternative way for modelling seasonality. Finally, the last section combines all the different topics presented and outlines how to use them together in order to forecast the total balance sent to debt collection.

2.1 Generalized Linear Mixed Models

Generalized Linear mixed models can be thought of as an extension of Generalized Linear Models (GLMs) in the sense that it allows mixed effects in the model, or as an extension of Linear Mixed Models (LMMs) as it allows the response to come from different distributions. Consider first a study with N different individuals indexed as $i = 1, \ldots, N$ that we gather data on over a time period. In a balanced study, each individual has the same number of time measurements indexed as $t = 1, \ldots, T$. In some cases though, the number of measurements is not the same for each individual, $t = 1, \ldots, T_i$. The data is therefore longitudinal as repeated measurements are performed for each individual i. We assume that the data is recorded with fixed intervals between each time measurement. The total number of observations in the entire study will thus be

$$\sum_{i=1}^{N} T_i.$$

By definition, longitudinal data is not independent and this dependency within individuals must be accounted for when making statistical models (Hedeker and Gibbons, 2006, p. 2). For each individual i, one records a total of p covariates such that x_{itk} is the kth covariate for individual i at time measurement t. Furthermore, the response for individual i at time t is recorded as y_{it}. An individual i can therefore have different responses at different times. The data layout will then be as shown in table 2.1. We define the covariate vector

Table 2.1: Data layout of N individuals with repeated observations.

Individual	Observation	Covariates	Response
1	1	$x_{111} \dots x_{11p}$	y_{11}
1	2	$x_{121} \dots x_{12p}$	y_{12}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	T_1	$x_{1T_11} \dots x_{1T_1p}$	y_{1T_1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	$x_{N11} \dots x_{N1p}$	y_{N1}
N	2	$x_{N21} \dots x_{N2p}$	y_{N2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	T_N	$x_{NT_N1} \dots x_{NT_Np}$	y_{NT_N}

associated with the fixed effects for individual i at time t as $\boldsymbol{x}_{it}^\top = (1, x_{it1}, \dots, x_{itp})$. We now introduce random effects $\boldsymbol{v}_i^\top = (v_{0i}, v_{1i}, \dots, v_{qi})$ to our model as well. We assume that the correlation between the measurements for individual i comes from sharing unobserved variables. Therefore, there are random effects common to all responses for a given individual that vary from one individual to another (Gad and Kholy, 2012). We denote the kth unobserved variable for individual i at time t as z_{itk} and define the covariate vector associated with the random effects for individual i at time t as $\boldsymbol{z}_{it}^\top = (z_{it1}, \dots, z_{itq})$. Furthermore, we assume the following:

- the responses Y_{it} are conditionally independent given the random effects $\boldsymbol{v}_i$
- the conditional distributions $f(y_{it}|\boldsymbol{v}_i)$ are independent and comes from an exponential family,
$$f(y_{it}|\boldsymbol{v_i}) = \exp\left\{\frac{y_{it}\theta_{it} - \kappa(\theta_{it})}{\phi} + c(y_{it}, \phi)\right\},$$
where θ_{it} is the canonical parameter, ϕ is the dispersion parameter and $\kappa(\cdot)$ and $c(\cdot)$ are known functions.
- the random effects $\boldsymbol{v}_i^\top = (v_{0i}, v_{1i}, \dots, v_{qi})$ are independent and identically distributed, $\boldsymbol{v}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_v})$ where $\boldsymbol{\Sigma_v}$ is a $(q+1) \times (q+1)$ covariance matrix (Groll and Tutz, 2012).

Since the responses y_{it} are conditionally independent given the random effects and are assumed to be generated from a distribution in an exponential family, the conditional means $\mu_{it} = E(y_{it}|\boldsymbol{v}_i)$ are related to the linear predictor $\eta_{it} = \boldsymbol{x}_{it}^\top\boldsymbol{\beta} + \boldsymbol{z}_{it}^\top\boldsymbol{v}_i$ through the link

function g such that $g(\mu_{it}) = \eta_{it}$. Hence, a generalized linear mixed model will then have the form

$$g(\mu_{it}) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \boldsymbol{z}_{it}^\top \boldsymbol{v}_i. \tag{2.1}$$

Equation (2.1) can be written in matrix form for each individual i by collecting all the observations for each individual in a vector, which we will refer to as individual i's cluster. The GLMM for individual i's cluster can then be written as

$$\underbrace{g(\boldsymbol{\mu}_i)}_{T_i \times 1} = \underbrace{\boldsymbol{X}_i}_{T_i \times (p+1)} \underbrace{\boldsymbol{\beta}}_{(p+1)\times 1} + \underbrace{\boldsymbol{Z}_i}_{T_i \times (q+1)} \underbrace{\boldsymbol{v}_i}_{(q+1)\times 1}, \tag{2.2}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT_i})^\top$ is a $T_i \times 1$ vector containing the conditional means, $\boldsymbol{X}_i^\top = (\boldsymbol{1}^\top, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT_i})$ is the $T_i \times (p+1)$ design matrix of fixed effects for individual i, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is a vector containing the fixed effects parameters including the intercept and $\boldsymbol{Z}_i^\top = (\boldsymbol{1}^\top, \boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iT_i})$ is the $T_i \times (q+1)$ design matrix of random effects for individual i. Hence, the model can be thought of as having two parts; the fixed effects for individual i contained in $\boldsymbol{X}_i\boldsymbol{\beta}$ and the random effects for individual i contained in $\boldsymbol{Z}_i\boldsymbol{v}_i$.

2.1.1 Estimation of Parameters for GLMMs

For a generalized linear mixed model, one must estimate both the fixed parameters $\boldsymbol{\beta}$ and the parameters associated with the random effects, $\boldsymbol{\Sigma_v}$. The most popular method is through maximum likelihood estimation. Thus, we need to find the marginal distribution of the Y_{it}s jointly. Recalling our assumptions, the contribution from observations for individual i is

$$\begin{aligned} f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma_v}) &= \int_{\boldsymbol{v_i}} f(\boldsymbol{y}_i|\boldsymbol{v_i}, \boldsymbol{\beta}) f(\boldsymbol{v_i}|\boldsymbol{\Sigma_v}) d\boldsymbol{v}_i \\ &= \int_{\boldsymbol{v_i}} \prod_{t=1}^{T_i} f(y_{it}|\boldsymbol{v}_i, \boldsymbol{\beta}) f(\boldsymbol{v}_i|\boldsymbol{\Sigma_v}) d\boldsymbol{v}_i \end{aligned}$$

where $f(\boldsymbol{v_i}|\boldsymbol{\Sigma_v})$ denotes the density of the random effects assumed to be normally distributed. Since the clusters are independent of each other, the likelihood is then given as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Sigma_v}) &= \prod_{i=1}^{N} f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma_v}) \\ &= \prod_{i=1}^{N} \left[\int_{\boldsymbol{v}_i} \prod_{t=1}^{T_i} f(y_{it}|\boldsymbol{v}_i, \boldsymbol{\beta}) f(\boldsymbol{v}_i|\boldsymbol{\Sigma_v}) d\boldsymbol{v}_i \right]. \end{aligned} \tag{2.3}$$

This likelihood function cannot be solved analytically, except for the special case when we have a Linearized Mixed Model (LMM). Therefore, to estimate the parameters in the GLMM, equation (2.3) must be solved with numerical methods.

There are many ways to estimate the likelihood for GLMMs. Some of the most common methods are; using a pseudo-quasilikelihood approach (see for instance Wolfinger and Oconnell, 1993), penalized quasilikelihood (Breslow and Clayton, 1993) and Markov chain Monte Carlo methods (see for instance Fan et al., 2008). We will use adaptive Gauss-Hermite quadrature (Liu and Pierce, 1994). The adaptive Gauss-Hermite quadrature have been found to be generally more accurate than penalized quasilikelihood methods, but noticeably computationally slower (Bolker et al., 2009). The method is often used for numerical integration in statistics to handle integrals on the very commonly occurring form

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx \approx \sum_{j=1}^{N_{\text{AGQ}}} w_j f(x_j),$$

by approximating the integral as a sum over weighted function points. Here, N_{AGQ} is the number of adaptive quadrature points, x_j are the abscissas determined as the roots of the physicists' version of Hermite polynomials $H_j(\cdot)$, $j = 1, \ldots, N_{AGQ}$ and the x_j's are symmetric around zero. The associated quadrature weights w_j are given by

$$w_j = \frac{2^{N_{\text{AGQ}}-1} n! \sqrt{\pi}}{n^2 [H_{N_{\text{AGQ}}-1}(x_j)]^2}.$$

For further details on the Gauss-Hermite quadrature, see for instance Davis and Rabinowitz, 1975, ch. 2. Additional details on how to efficiently implement this method is presented in Pinheiro and Chao, 2006.

2.1.2 A Random Intercept Model

The GLMM in eq. (2.1) can be simplified if the random effect term only consist of a random intercept term, $\boldsymbol{v}_i = v_{0i}$. The GLMM can then be simplified to

$$g(\mu_{it}) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + v_{0i}, \tag{2.4}$$

where v_{0i} is normally distributed, $v_{0i} \sim N(0, \sigma_v^2)$. The subscript $0i$ is used to indicate that the variable will affect the intercept for individual i. v_{0i} can be thought of as the influence of individual i based on the repeated observations. Notice that the random intercept is constant across time. The random intercept model can be represented in a hierarchical form as well. Assume that there are $p + 1$ fixed effects coefficients where β_0 represents the fixed effect intercept term. Thus, equation (2.4) can be partitioned into the following *within-individual* model

$$g(\mu_{it}) = b_{0i} + b_{1i}x_{it1} + \ldots + b_{pi}x_{itp} \tag{2.5}$$

and *between-individuals* model,

$$\begin{aligned} b_{0i} &= \beta_0 + v_{0i} \\ b_{ki} &= \beta_k \quad \text{for } k = 1, \ldots, p. \end{aligned} \tag{2.6}$$

The *within-individual* model (2.5) suggest that individual i's response at time t is determined by individual i's initial level b_{0i} and slopes b_{ki} for $k = 1, \ldots, p$. The *between-individuals* model (2.6) shows that each individual i has a distinct own initial level which

consist of the population initial level β_0 and an individual contribution v_{0i} (Hedeker and Gibbons, 2006, p. 49). This term will shift the intercept up or down depending on the individual. We will from now on assume that the random effects only consist of the random intercept term v_{0i}.

2.1.3 A Mixed Effects Logistic Regression Model

Consider a random intercept model of the form (2.4) and a binary response $y_{it} \in \{0, 1\}$. We assume that these responses are random variables given the random intercept terms such that $Y_{it}|v_{0i} \sim \text{Bernoulli}(p_{it})$, where $p_{it} = P(Y_{it} = 1)$ is the probability of individual i belonging to class 1 at time t. Thus, we may write

$$Y_{it}|v_{0i} = \begin{cases} 1 \text{ with probability } p_{it} \\ 0 \text{ with probability } 1 - p_{it}. \end{cases}$$

The link function that we will use for a binary response is the logit. Hence, $\mu_{it} = E(Y_{it}|v_{0i}) = p_{it}$ such that

$$g(\mu_{it}) = \log\left(\frac{p_{it}}{1 - p_{it}}\right). \tag{2.7}$$

This connection arises from the cumulative distribution of a logistic(0,1) distribution. Note that individual i may have response 0 at one time and response 1 at another time. Combining (2.4) with (2.7) gives the mixed effects logistic regression model

$$\log\left(\frac{p_{it}}{1 - p_{it}}\right) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + v_{0i}. \tag{2.8}$$

The ratio $\frac{p_{it}}{1-p_{it}}$ is defined as the odds and is useful for interpreting how the fixed effect coefficients affects the model. Re-writing (2.8), we get that

$$\frac{p_{it}}{1 - p_{it}} = \exp(\beta_0 + v_{0i}) \cdot \exp(\beta_1 x_{it1}) \cdot \cdots \cdot \exp(\beta_p x_{itp}).$$

Consider now a unit increase in the 1st covariate x_{it1} to $x_{it1} + 1$,

$$\begin{aligned} \frac{P(Y_{it} = 1|x_{it1} + 1)}{1 - P(Y_{it} = 1|x_{it1} + 1)} &= \exp(\beta_0 + v_{0i}) \cdot \exp(\beta_1(x_{it1} + 1)) \cdot \cdots \cdot \exp(\beta_p x_{itp}) \\ &= \exp(\beta_0 + v_{0i}) \cdot \exp(\beta_1 x_{it1}) \cdot \exp(\beta_1) \cdot \cdots \cdot \exp(\beta_p x_{itp}) \\ &= \frac{P(Y_{it} = 1|x_{it1})}{1 - P(Y_{it} = 1|x_{it1})} \cdot \exp(\beta_1). \end{aligned}$$

Thus, a unit increase in a covariate will change the odds by a factor $\exp(\beta_k)$. If $\beta_k > 0$, the odds will increase and vice versa if $\beta_k < 0$. Solving (2.8) in terms of p_{it} gives that the probability of individual i belonging to class 1 at time t is

$$p_{it} = P(Y_{it} = 1) = \frac{\exp(\boldsymbol{x}_{it}^\top \boldsymbol{\beta} + v_{0i})}{1 + \exp(\boldsymbol{x}_{it}^\top \boldsymbol{\beta} + v_{0i})}. \tag{2.9}$$

Alternatively, (2.9) can be re-written in the two-step formulation presented earlier such that the *within-individuals* model is

$$p_{it} = \frac{\exp(b_{0i} + b_{1i}x_{it1} + \ldots + b_{pi}x_{itp})}{1 + \exp(b_{0i} + b_{1i}x_{it1} + \ldots + b_{pi}x_{itp})},$$

and the *between-individual* model is still

$$\begin{aligned} b_{0i} &= \beta_0 + v_{0i} \\ b_{ki} &= \beta_k \quad \text{for } k = 1, \ldots, p. \end{aligned}$$

2.1.4 Estimation of Parameters for a Mixed Effects Logistic Regression Model

The probability mass function for a Bernoulli distribution can be written as

$$f(y_{it}|v_{0i}) = p_{it}^{y_{it}}(1 - p_{it})^{1-y_{it}} \quad \text{for } y_{it} \in \{0, 1\}.$$

The likelihood function for a generalized linear mixed model was found in (2.3). Using this along with the Bernoulli distribution, the likelihood for a mixed effects logistic regression model becomes

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Sigma_v}) &= \prod_{i=1}^{N} \left[\int_{\boldsymbol{v}_i} \prod_{t=1}^{T_i} \left(p_{it}^{y_{it}}(1 - p_{it})^{1-y_{it}} \right) f(\boldsymbol{v}_i|\boldsymbol{\Sigma_v}) d\boldsymbol{v}_i \right] \\ &= \prod_{i=1}^{N} \left[\int_{\boldsymbol{v}_i} \exp\left(\ln \prod_{t=1}^{T_i} p_{it}^{y_{it}}(1 - p_{it})^{1-y_{it}} \right) f(\boldsymbol{v}_i|\boldsymbol{\Sigma_v}) d\boldsymbol{v}_i \right] \\ &= \prod_{i=1}^{N} \left[\int_{\boldsymbol{v}_i} \exp\left(\sum_{t=1}^{T_i} y_{it} \ln(p_{it}) + (1 - y_{it}) \ln(1 - p_{it}) \right) f(\boldsymbol{v}_i|\boldsymbol{\Sigma_v}) d\boldsymbol{v}_i \right] \end{aligned} \tag{2.10}$$

Notice that the parameters we want to estimate $(\boldsymbol{\beta}, \boldsymbol{\Sigma_v})$ are not "visible" in (2.10) as they lie within the probability $p_{it} = p_{it}(\boldsymbol{\beta}, \boldsymbol{\Sigma_v})$. As previously mentioned, this likelihood cannot be computed analytically, and numerical approximations must be used to determine the parameters.

2.1.5 A Method for Variable Selection

For each individual i, we have recorded p covariates x_{itk}, $k = 1, \ldots, p$ at different time measurements. Ideally, one would like to find a handful of explanatory variables that fit the data well. GLMM models will have computational problems and is therefore usually restricted to a few variables. Too many explanatory variables may also give unstable estimates (Bolker et al., 2009). In order to perform variable selection, we will use the well-known LASSO method, which was first proposed by Tibshirani, 1996 for GLMs and is a penalized regression technique that adds a L_1 penalty term to perform variable selection. This penalty term will shrink fixed effects coefficients towards zero and some will be set to

exactly zero. The extension of the LASSO to GLMMs is presented. The general idea is to maximize the log-likelihood function as usual for maximum likelihood estimation while at the same time constraining the L_1-norm on the fixed effects coefficients $\boldsymbol{\beta}$ to be less than some constant $s \geq 0$, such that

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, l(\boldsymbol{\beta}), \text{ subject to } ||\boldsymbol{\beta}||_1 \leq s \tag{2.11}$$

where $||\cdot||_1$ is the L_1-norm, i.e. for the fixed effects coefficient parameter vector $\boldsymbol{\beta}$

$$||\boldsymbol{\beta}||_1 = \sum_{i=1}^{p} |\boldsymbol{\beta}_i|$$

for p covariates. Alternatively, we can formulate (2.11) as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left(l(\boldsymbol{\beta}) - \lambda ||\boldsymbol{\beta}||_1 \right) \tag{2.12}$$

where $\lambda \geq 0$. The tuning parameters, s and λ must be determined, for example through cross-validation. We need efficient algorithms to solve (2.11) and (2.12). Recall that a generalized linear mixed model can be written as in (2.2). As previously mentioned, one popular method that can be used to maximize the likelihood function for GLMMs is the penalized quasi-likelihood (PQL). We will change the notation slightly from section 2.1.1. Let the covariance matrix depend on some unknown parameter vector $\boldsymbol{\rho}$, $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \boldsymbol{\Sigma}_{\boldsymbol{v}}(\boldsymbol{\rho})$. We define $\boldsymbol{\Phi}^\top = (\phi, \boldsymbol{\rho}^\top)$ where ϕ is the dispersion parameter and a parameter vector $\boldsymbol{\delta}^\top = (\boldsymbol{\beta}^\top, \boldsymbol{v}^\top)$ that contains both the fixed and random effects. The change in notation is done since the likelihood is usually specified by $\boldsymbol{\Phi}^\top$ and $\boldsymbol{\delta}^\top$ for penalized-based concepts (Breslow and Clayton, 1993). The log-likelihood is then found by taking the logarithm of eq. (2.3)

$$\begin{aligned} l(\boldsymbol{\delta}, \boldsymbol{\Phi}) &= \log \left(\prod_{i=1}^{N} \int_{\boldsymbol{v}_i} f(\boldsymbol{y}_i | \boldsymbol{\delta}, \boldsymbol{\Phi}) f(\boldsymbol{v}_i | \boldsymbol{\Sigma}_{\boldsymbol{v}}) d\boldsymbol{v}_i \right) \\ &= \sum_{i=1}^{N} \log \left(\int_{\boldsymbol{v}_i} f(\boldsymbol{y}_i | \boldsymbol{\delta}, \boldsymbol{\Phi}) f(\boldsymbol{v}_i | \boldsymbol{\Sigma}_{\boldsymbol{v}}) d\boldsymbol{v}_i \right) \end{aligned}$$

An approximation to the log-likelihood

$$l_{\text{app}}(\boldsymbol{\delta}, \boldsymbol{\Phi}) = \sum_{i=1}^{N} \log \left(f(\boldsymbol{y}_i | \boldsymbol{\delta}, \boldsymbol{\Phi}) \right) - \frac{1}{2} \boldsymbol{v}^\top \boldsymbol{\Sigma}_v(\boldsymbol{\rho})^{-1} \boldsymbol{v} \tag{2.13}$$

was first derived by Breslow and Clayton, 1993. The term $\boldsymbol{v}^\top \boldsymbol{\Sigma}_v(\boldsymbol{\rho})^{-1} \boldsymbol{v}$ stems from Laplace's method. We now introduce the penalty term to (2.13) to get the penalized log-likelihood

$$\begin{aligned} l_{\text{pen}}(\boldsymbol{\delta}, \boldsymbol{\Phi}) &= l_{\text{app}}(\boldsymbol{\delta}, \boldsymbol{\Phi}) - \lambda ||\boldsymbol{\beta}||_1 \\ &= \sum_{i=1}^{N} \log \left(f(\boldsymbol{y}_i | \boldsymbol{\delta}, \boldsymbol{\Phi}) \right) - \frac{1}{2} \boldsymbol{v}^\top \boldsymbol{\Sigma}_v(\boldsymbol{\rho})^{-1} \boldsymbol{v} - \lambda ||\boldsymbol{\beta}||_1 \end{aligned}$$

Given $\boldsymbol{\Phi}$, this optimization problem becomes

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmax}}\, l_{\text{pen}}(\boldsymbol{\delta}, \boldsymbol{\Phi}). \tag{2.14}$$

A gradient ascent algorithm is proposed and implemented to solve (2.14) by Groll and Tutz, 2012. In order to determine the optimal value of the tuning parameter λ, we will fit models with different values of λs and compute the Bayesian Information Criterion (BIC) (Schwarz, 1978) for each model. The BIC is a common criterion for model selection and is defined as

$$\text{BIC} = k \ln \Big(\sum_{i=1}^{N} T_i \Big) - 2 \ln \hat{L}(\boldsymbol{\delta}, \boldsymbol{\Phi}) \tag{2.15}$$

where $\sum_{i=1}^{N} T_i$ is the sample size, k the number of parameters estimated by the model in total and $\hat{L}$ is the estimated maximum value of the likelihood function. One of the advantages of the BIC is that it penalizes complexity in the model, i.e. the number of parameters in the model. Hence, we choose the value of λ for which the BIC is minimized and investigate possible significant explanatory variables at the chosen λ value.

2.1.6 Diagnostics Checking

Diagnostic checking for generalized linear mixed models is not as straightforward as for generalized linear models (Bolker et al., 2009). We will investigate the residuals and parameters of the model as outlined in the following.

Residual Diagnostics

Residuals from a GLMM is not as easily interpretable as residuals from GLMs, since the expected distribution of the data will change with the fitted values. For Pearson residuals for instance, one reweights the residuals by dividing by the square root of the expected variance, but the residuals will not be readily interpretable because the residuals will not be visually homogeneous, even with a correctly specified model. Instead, we will create interpretable residuals for the mixed effects logistic regression model that are scaled and can be as easily interpreted as residuals from a linear model. This is a simulation-based approach proposed by Hartig, 2019 and the procedure can be outlined as follows:

1. Simulate new data from the fitted mixed effects logistic regression model for each individual i and corresponding observations $t = 1, \ldots, T_i$.

2. For each value of t, compute the empirical cumulative density function $\hat{F}$ for the simulated observations, which describes the possible values (and their probability) at the predictor combination of the observed value, assuming the fitted model is correct.

3. Define the residual as the value of the empirical cumulative density function $\hat{F}$ at the value of the observed data. For instance, if the residual is 0, then all the simulated values are larger than the observed value, while a residual of 0.5 means half of the simulated values are larger than the observed value.

A model that is specified correctly will have observed data that look as if created from the fitted model. Hence, all residuals should appear with equal probability, and the expected distribution of the residuals will be uniform between 0 and 1. For further details, see Dunn and Smyth, 1996; Hartig, 2019.

Parameter Diagnostics

Furthermore, we wish to perform diagnostics on the parameters $(\boldsymbol{\beta}, \sigma_v)$ in the mixed effects logistic regression model. The best method used to test single parameters is through bootstrapping as we avoid the asymptotic assumptions of the likelihood ratio test (Bates et al., 2015). Bootstrapping becomes slightly more complicated when dealing with models that have mixed effects as the response variable must be generated in two steps. The non-parametric bootstrapping is performed in the following manner:

1. Fit the mixed effects logistic regression model (2.8) to the data and find the estimated parameters $(\hat{\boldsymbol{\beta}}, \sigma_v)$ by solving (2.10).
2. Sample the random intercept terms v_{0i}^* from $\hat{v}_{0i}$ with replacement for $i = 1, \ldots, N$.
3. Compute $\boldsymbol{p}_i^* = \boldsymbol{p}_i^*(\hat{\boldsymbol{\beta}}, \sigma_v)$ from equation (2.9) for $i = 1, \ldots, N$ and $t = 1, \ldots, T_i$, where $\boldsymbol{p}_i^* = (p_{i1}^*, \ldots, p_{iT_i}^*)$.
4. Generate bootstrapped responses $Y_{it}^* \sim \text{Bernoulli}(p_{it}^*)$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T_i$.
5. Fit the mixed effects logistic regression model (2.8) to the bootstraped data to obtain the bootstrapped estimates $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}}^*)$.
6. Repeat steps 2 - 5 B times where B is sufficiently large.

2.1.7 Classification with a Mixed Effects Logistic Regression Model

The mixed effects logistic regression model can be used for classification to determine if individual i at time t should belong to class 0 or 1. Consider the problem of classifying the total number of individuals that belong to class 1 in a group consisting of N individuals. Note that this should not be confused with the more common problem of correctly classifying the class an individual i belongs to. From (2.9), we estimate a probability, $\hat{p}_{it} \in (0, 1)$. The classification will be such that

$$\hat{y}_{it} = \begin{cases} 1 \text{ if } \hat{p}_{it} > \hat{\alpha}, \\ 0 \text{ otherwise.} \end{cases} \tag{2.16}$$

where $\hat{\alpha} \in (0, 1)$ is a threshold value. For our purposes, we are only interested that the total number of individuals that are classified as belonging to class 1 agrees with the training set (see table 2.2). Therefore, we choose the optimal $\hat{\alpha}$ that minimizes

$$\hat{\alpha} = \underset{\alpha}{\text{argmin}} \sum_{i=1}^{N} \left[\left| I(p_{it} > \alpha) - I(y_{it} = 1) \right| \right] \tag{2.17}$$

where I is an indicator function. Several algorithms solve this optimization problem.

2.2 Handling Imbalanced Data

This section comprises issues that arise from having *imbalanced* data and methods that can be used to handle these issues. A skewed or imbalanced data set is one where the number of observations from one response class far exceeds the others. Consider still a data set with a binary response, $y_{it} \in \{0, 1\}$. The data set will then be highly imbalanced if for instance 99% of the observations belong to class 0 and only 1% belong to class 1. Class 0 is then referred to as the majority class and class 1 the minority class. Consider a mixed effects logistic regression model trained on such a data set. The result is that the model will almost always classify every instance as belonging to class 0. Moreover, the model will receive a very high accuracy since 99% of instances will be classified correctly. There are several techniques developed to work around this problem, many of which involve making the data less skewed. It should be noted that most imbalanced data problems deal with improving either the sensitivity or specificity. This usually involves using a different

Table 2.2: A contingency table for a binary classifier

	Class positive	Class negative
Assigned positive	True Positives (TP)	False Positives (FP)
Assigned negative	False Negatives (FN)	True Negatives (TN)

performance metric than accuracy for the model. However, for our purposes, we are not interested in False Positives (FP) and False Negatives (FN) and only concern us that the total number of true positives (TP) are correct (see table 2.2).

2.2.1 Random Undersampling

Random undersampling can be used to adjust the class distribution in a data set if the data set is imbalanced. Let 0 be the majority class and let 1 be the minority class. To undersample means reducing the number of observations from the majority class such that the data set becomes less imbalanced. In random undersampling, we randomly remove some of the samples from the majority class. For longitudinal data, this means that we remove all observations for an individual i, i.e. individual i's cluster. A question that arises is how one can determine the right number of clusters from the majority class that should be removed. For instance, we can remove individuals until the training set is evenly balanced with half of the observations being from the majority class and the other half from the minority class. However, an undersample ratio r that is too balanced will reduce the size of the training set as there are not that many instances from the minority class 1. This may lead to other issues. On the other hand, if the undersample ratio is too low, the model will still classify all instances as belonging to the majority class. We will solve this issue by computing the Residual Sum of Squares (RSS) for different undersampling ratios. Let $\Lambda(y_{it}, \hat{y}_{it})$ be the 0-1 loss function at time t such that

$$\Lambda(y_{it}, \hat{y}_{it}) = I\big(\hat{p}_{it} > \hat{\alpha}\big) - I\big(y_{it} = 1\big) = I(y_{it} \neq \hat{y}_{it}),$$

where $\hat{y}_{it}$ is given from (2.16) and $\hat{\alpha}$ from (2.17). Consider now the problem of correctly specifying the number of True Positives (TP) of the minority class in a data set. The RSS

will then be

$$\text{RSS}(r) = \sum_i \Lambda(y_{it}, \hat{y}_{it})^2(r). \tag{2.18}$$

The r in parenthesis is included to indicate that the RSS will depend on the undersampling ratio. Hence, we will choose the undersampling ratio that minimizes

$$r = \underset{r}{\text{argmin}}\, \text{RSS}(r). \tag{2.19}$$

2.2.2 Adjusting the Outputs of a Classifier

Consider again a highly imbalanced data set and the classic binary classification problem where 0 is the majority class and 1 is the minority class. Assume that we have performed random undersampling in order to create a training set that is not as imbalanced as the full data set. Assume further that a mixed effects logistic regression model is trained based on this training set. Applying this model for classification on the full data set may provide sub-optimal results since the model relies on the prior probabilities of belonging to class 0 and 1 in the training set. These prior probabilities may be completely different than the prior probabilities of belonging to class 0 and 1 in the full data set. Since the ratio between the two classes are different in the full data set and the training set, we should adjust the outputs to account for this since the prior class probabilities have changed from the training set to the full data set.

Saerens et al., 2002 proposed a method to improve classification accuracy by adjusting the outputs of a classifier to new prior probabilities. The classifications are based on a collection of observations vectors in a training set at time t, which we denote as $\boldsymbol{X}_t^\top = (\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{Nt})$. Note that $\boldsymbol{X}_i$ should not be confused with $\boldsymbol{X}_t$. $\boldsymbol{X}_i$ contains the covariates recorded for individual i at all time measurements, while $\boldsymbol{X}_t$ contains all recorded covariates at time t for all individuals. Let $p_\tau(1_t)$ denote the prior probability of belonging to class 1, the minority class, in the training set at time t where the subscript τ indicates that the probability is based on the training set. We can estimate this probability as

$$\hat{p}_\tau(1_t) = \frac{N_\tau^{(1_t)}}{N_\tau},$$

where $N_\tau^{(1_t)} = \sum_{i=1}^{N_\tau} I(y_{it} = 1)_\tau$ is the sum of observations where $y_{it} = 1$ in the training set at time t and N_τ is the total number of observations at time t. We estimate $p_\tau(0_t)$ similarly. Depending on the undersampling, we can set this to a fixed probability. We suppose that for the two classes 0 and 1, the total number of training examples are independently recorded according to the within-class probability densities, $p(\boldsymbol{X}_t|1_t)$. Suppose now a classification model is trained and gives an estimated posterior probability of belonging to class 1, $\hat{p}_\tau(1_t|\boldsymbol{X}_t)$. Assume now that we wish to use the mixed effects logistic regression model that is trained on a training set to classify on a real-life data set where the prior probabilities $p(0_t), p(1_t)$ are very different from the trained prior probabilities $p_\tau(0_t), p_\tau(1_t)$. In order to use this model on the real life data set, the posterior probabilities must be adjusted accordingly. Assume first that $p(1_t)$ is known, i.e. a supervised

learning case. We will later discuss how to handle situations where $p(1_t)$ is unknown, an unsupervised learning case.

Prior Probability Known

In the following we assume that the within-class probability densities do not change for the training set and the new data set, $p(\boldsymbol{X}_t|k_t) = p_\tau(\boldsymbol{X}_t|k_t)$, $k_t \in \{0, 1\}$ and that we have estimates of the prior probabilities $\hat{p}(1_t)$, $\hat{p}(1_t)$. Note that only the proportion between positive and negative responses is altered. From Bayes' theorem, we find that the within-class probability density is

$$\hat{p}_\tau(\boldsymbol{X}_t|1_t) = \frac{\hat{p}_\tau(1_t|\boldsymbol{X}_t)\hat{p}_\tau(\boldsymbol{X}_t)}{\hat{p}_\tau(1_t)}. \tag{2.20}$$

Similarly, for the full data set, we get (without the subscript τ)

$$\hat{p}(\boldsymbol{X}_t|1_t) = \frac{\hat{p}(1_t|\boldsymbol{X}_t)\hat{p}(\boldsymbol{X}_t)}{\hat{p}(1_t)}. \tag{2.21}$$

Using our assumption, setting (2.20) equal to (2.21), defining $f(\boldsymbol{X}_t) = \hat{p}_\tau(\boldsymbol{X}_t)/\hat{p}(\boldsymbol{X}_t)$ and solving for $\hat{p}(1_t|\boldsymbol{X}_t)$ gives

$$\hat{p}(1_t|\boldsymbol{X}_t) = f(\boldsymbol{X}_t) \cdot \frac{\hat{p}(1_t)}{\hat{p}_\tau(1_t)} \cdot \hat{p}_\tau(1_t|\boldsymbol{X}_t)$$

Since $\hat{p}(0_t|\boldsymbol{X}_t) + \hat{p}(1_t|\boldsymbol{X}_t) = 1$, we obtain that

$$f(\boldsymbol{X}_t) = \left[\frac{\hat{p}(0_t)}{\hat{p}_\tau(0_t)}\hat{p}_\tau(0_t|\boldsymbol{X}_t) + \frac{\hat{p}(1_t)}{\hat{p}_\tau(1_t)}\hat{p}_\tau(1_t|\boldsymbol{X}_t)\right]^{-1}$$

Hence, the corrected posterior probability of belonging to class 1 is given by

$$\hat{p}(1_t|\boldsymbol{X}_t) = \frac{\frac{\hat{p}(1_t)}{\hat{p}_\tau(1_t)}}{\frac{\hat{p}(0_t)}{\hat{p}_\tau(0_t)}\hat{p}_\tau(0_t|\boldsymbol{X}_t) + \frac{\hat{p}(1_t)}{\hat{p}_\tau(1_t)}\hat{p}_\tau(1_t|\boldsymbol{X}_t)} \cdot \hat{p}_\tau(1_t|\boldsymbol{X}_t) \tag{2.22}$$

Notice that the probability $\hat{p}(1_t|\boldsymbol{X}_t)$ is proportional to $\hat{p}_\tau(1_t|\boldsymbol{X}_t)$. The posterior probability can be thought of as adjusted by a calibration factor to account for the different proportion between classes in the full data set and the training set, although it should be noted that the calibration factor is not a constant since the prior probability of belonging to class 1 based on the training set $\hat{p}_\tau(1_t|\boldsymbol{X}_t)$ is different for each individual i.

Prior Probability Unknown

When the prior probability $\hat{p}(1_t)$ is unknown, we are unable to use equation (2.22) directly to correct the posterior probability. This will be the case when we forecast for $t = T+h$ for $h = 1, 2, \ldots$. Instead, we will forecast $\hat{p}(1_{T+h})$ using time series analysis for $h = 1, 2, \ldots$ as presented in section 2.3.

2.3 A Non-stationary Time Series

A time series is a sequence of observations γ_t taken sequentially in time t. Denote the unknown prior probability of belonging to class 1 as $\gamma_t = \hat{p}(1_t)$. We define a time series $\{\gamma_t\}$ as weakly stationary (Brockwell and Davis, 2002, p. 15) if

- the mean $E(\gamma_t)$ is independent of t
- the covariance function $Cov(\gamma_{t+h}, \gamma_t)$ is independent of t for each lag h.

Suppose now rather that $\{\gamma_t\}$ is an observed, mean-corrected time series that appears to have a time-dependent trend so that the mean $E(\gamma_t)$ depends on t and the covariance function $Cov(\gamma_{t+h}, \gamma_t)$ is time-dependent for each lag h. Such a time series is non-stationary and can be modeled as an Autoregressive Integrated Moving Average process, abbreviated as an ARIMA process. We introduce the difference operator $\nabla = 1 - B$ where B is the backward shift operator, $B^j\gamma_t = \gamma_{t-j}$ for $j = 0, \pm 1, \ldots$ The difference operator can be applied d times such that $\nabla^d = (1 - B)^d$. Applying the difference operator on our time series $\{\gamma_t\}$ d times

$$\nabla^d\gamma_t = (1 - B)^d\gamma_t = w_t \tag{2.23}$$

gives a new stationary time series w_t that we can model as an ARMA process

$$\phi(B)w_t = \theta(B)\varepsilon_t,\ \{\varepsilon_t\} \overset{i.i.d}{\sim} WN(0, \sigma^2), \tag{2.24}$$

where $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ is the autoregressive operator (AR) and $\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$ is the moving average (MA) operator. The integers p and q will therefore determine the order of the AR and MA operator polynomials, respectively. Combining (2.23) and (2.24) gives the ARIMA-process

$$\phi(B)(1 - B)^d\gamma_t = \theta(B)\varepsilon_t. \tag{2.25}$$

Equation (2.25) can be written with all terms as

$$\nabla^d\gamma_t - \phi_1\nabla^d\gamma_{t-1} - \ldots - \phi_p\nabla^d\gamma_{t-p} = \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}.$$

Thus, for an observed time series with T observations, we wish to determine the parameters in an ARIMA model that best fit our observed values. Furthermore, if the time series exhibits seasonal trends, which is common for monthly data, the time series can be modelled as a Seasonal ARIMA model, abbreviated as SARIMA. However, we will propose a different approach for modelling seasonality as described in section 2.3.2.

2.3.1 Identification and Estimation of Parameters

Identification methods are procedures applied to a time series in order to determine a model appropriate for further investigation (Box et al., 1994, p. 183). The first step is to determine suitable values for d and p, q. The value d is determined by differencing the time series until it is sufficiently (weakly) stationary. In many cases, if the time series has a linear trend, $d = 1$. The values p and q are often determined by investigating the sample partial autocorrelation function plot (PACF) and sample autocorrelation function

plot (ACF) of the differenced time series, respectively. However, we will mainly rely on the small-sample-size Akaike Information Criterion, AICC (Hurvich and Tsai, 1989), which measures the goodness of fit for a model, to determine p and q. When the number of observations T is small the more well-known AIC may overfit by choosing too many parameters. Hence, the AICC is preferred. We thus choose the values of p and q that minimize AICC$(\hat{\phi}, \hat{\theta})$

$$\text{AICC}(\phi, \theta) = -2 \ln L(\phi, \theta, S(\phi, \theta)/T) + \frac{2(p+q+1)T}{T-p-q-2}, \tag{2.26}$$

where L and S are defined in (2.27) and (2.28), respectively. For further details regarding the AICC, see for instance Brockwell and Davis, 2002, p. 171-174). Estimation of the parameters of our model $\phi = (\phi_1, \ldots, \phi_p)$, $\theta = (\theta_1, \ldots, \theta_q)$ and σ^2 are based on the likelihood principle. Assume $\{\gamma_t\}$ is a mean-corrected, stationary, Gaussian time series, such that $E(\gamma_t) = 0$. Furthermore, let $\boldsymbol{\gamma}_T = (\gamma_1, \ldots, \gamma_T)^\top$ and $\hat{\boldsymbol{\gamma}}_T = (\hat{\gamma}_1, \ldots, \hat{\gamma}_T)^\top$, where $\hat{\gamma}_1 = 0$ and $\hat{\gamma}_j = \hat{E}(\gamma_j | \gamma_1, \ldots, \gamma_{j-1})$, $j \geq 2$. We use $\hat{E}(\gamma_j | \gamma_1, \ldots, \gamma_{j-1})$ to denote the best linear predictor combination for γ_j in terms of $\gamma_1, \ldots, \gamma_{j-1}$. It can then be shown that the likelihood function can be written as

$$L(\phi, \theta, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^T r_0 \ldots r_{n-1}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{T} \frac{(\gamma_j - \hat{\gamma}_j)^2}{r_{j-1}} \right\}, \tag{2.27}$$

where $r_j = E(\gamma_{j+1} - \hat{\gamma}_{j+1})^2 / \sigma^2$ for $j = 0, \ldots, T-1$. A more comprehensive explanation can be found in for instance Brockwell and Davis, 2002, p. 158-160. Taking the natural logarithm of (2.27), differentiating with respect to σ^2 and setting $\frac{\partial l(\phi,\theta,\sigma^2)}{\partial \sigma^2} = \frac{\partial \ln L(\phi,\theta,\sigma^2)}{\partial \sigma^2} = 0$ gives the maximum likelihood estimation for σ^2,

$$\hat{\sigma}^2 = \frac{1}{T} S(\hat{\phi}, \hat{\theta}),$$

where

$$S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^{T} \frac{(\gamma_j - \hat{\gamma}_j)^2}{r_{j-1}}. \tag{2.28}$$

Furthermore, we determine values $\hat{\phi}, \hat{\theta}$ that minimize $l(\phi, \theta)$

$$l(\phi, \theta) = \ln\left(\frac{1}{T} S(\phi, \theta) \right) + \frac{1}{T} \sum_{j=1}^{T} \ln r_{j-1}. \tag{2.29}$$

2.3.2 An Alternative Approach for Modelling Seasonality

As mentioned, a time series may exhibit seasonal trends as well, in which case the time series can be modelled as a seasonal ARIMA process. This is particularly common for monthly data. However, in situations with few observations T, differencing the time series multiple times can reduce the length of the time series greatly, especially if a seasonal operator is applied as well. An alternative approach is proposed where the seasonal effects

are treated as known reoccurring outliers. We will only focus on additive outliers, i.e. outliers that does not affect subsequent observations. This seems like a justifiable assumption for the time series we will analyze. Suppose $\{\gamma_t\}$ is our observed, mean-corrected time series with T observations. We define $\varphi(B) = \phi(B)(1-B)^d$ such that (2.25) can be formulated as

$$\gamma_t = \frac{\theta(B)}{\varphi(B)}\varepsilon_t \tag{2.30}$$

First, assume the time series has only one known outlier at a known time $t = t^*$ of magnitude ω. The time series can then be modelled as

$$\gamma_t = \omega I_t^{(t^*)} + \frac{\theta(B)}{\varphi(B)}\varepsilon_t, \text{ where } I_t^{(t^*)} = \begin{cases} 1 \text{ if } t = t^*, \\ 0 \text{ if } t \neq t^*. \end{cases} \tag{2.31}$$

The added indicator term will now shift the time series at time $t = t^*$ with a magnitude ω to account for the additive outlier. However, the new parameter ω added to the model must now be estimated as well. We first define $\pi(B) = \theta^{-1}(B)\varphi(B) = 1 - \sum_{i=1}^{\infty}\pi_i B^i$, where we have assumed that the moving average operator is invertible, i.e. the roots of $\theta(z) = 0$ lies outside the unit circle, (Box et al., 1994, p. 69). Further, we multiply (2.31) with $\pi(B)$ to obtain

$$\pi(B)\gamma_t = \omega\pi(B)I_t^{(t^*)} + \varepsilon_t \tag{2.32}$$

Defining $e_t = \pi(B)\gamma_t$ for $t = 1, \ldots, T$ and

$$\pi(B)I_t^{(t^*)} = \xi_{1t} = \begin{cases} 0 & \text{for } t < t^* \\ -\pi_{t-T} & \text{for } t \geq t^* \end{cases}$$

with $\pi_0 = -1$, we end up with a linear regression equation of (2.32)

$$e_t = \omega\xi_{1t} + \varepsilon_t \text{ for } t = 1, \ldots, T, \tag{2.33}$$

where ω now serves as the coefficient in the linear regression equation. Since $\theta(B)\pi(B) = \phi(B)(1-B)^d$, the coefficients $\pi_1, \pi_2, \ldots$ are determined (assuming for simplicity that $d = 1$) recursively by

$$\pi_k = \phi_{k-1} - \phi_k + \theta_k - \sum_{i=1}^{k-1}\pi_i\theta_{k-i} \text{ for } k > 0. \tag{2.34}$$

To determine an estimate for ω, we use the least squares principle on (2.33) by minimizing

$$\min_{\omega}\varepsilon_t^2 = \min_{\omega}(e_t - \omega\xi_{1t})^2.$$

Differentiating with respect to w and setting the derivative equal to zero yields

$$\hat{\omega} = \frac{e_{t^*} - \sum_{t=1}^{T-t^*}\pi_t e_{t^*+t}}{\sum_{t=0}^{T-t^*}\pi_t^2} = \frac{\pi^*(F)e_{t^*}}{\tau^2}, \tag{2.35}$$

where $\pi^*(F) = 1 - \pi_1 F - \pi_2 F^2 - \ldots - \pi_{T_{t^*}} F^{T-t^*}$, F is the forward shift operator, $F^j e_t = e_{t+j}, j = 0, \pm 1, \ldots$ and $\tau^2 = \sum_{t=0}^{T-t^*}\pi_t^2$. We see from (2.35) that the information about

an additive outlier t^* is spread over the proceeding "residuals" $e_{t^*}, e_{t^*+1}, e_{t^*+2}, \ldots$ with generally decreasing weights $1, -\pi_1, -\pi_2, \ldots$, (Box et al., 1994, p. 469-471). Suppose now that the model contains k seasonal additive outliers at known time points $t_1^*, t_2^*, \ldots, t_k^*$ with associated weights $\omega_1, \omega_2, \ldots, \omega_k$ where $t_{j+1}^* - t_j^* = s$ for $j = 1, \ldots, k-1$ and s is the seasonal period between each outlier. For instance, for monthly data the seasonal period will be $s = 12$. The time series can then be modelled as

$$\gamma_t = \sum_{j=1}^{k} \omega_j I_t^{(t_j^*)} + \frac{\theta(B)}{\varphi(B)}\varepsilon_t. \tag{2.36}$$

Similarly, as (2.33) we get

$$e_t = \sum_{j=1}^{k} \omega_j \xi_{jt} + \varepsilon_t \text{ for } t = 1, \ldots, T. \tag{2.37}$$

When there are multiple outliers, however, the estimate of w in (2.35) may become a biased estimate for the outlier at time $t = t^*$ due to the other outliers (Chen and Liu, 1993). We will therefore use an iterative approach to handle the multiple outliers issue based on the work by Chen and Liu, 1993. However, we will do some adjustments as we assume that the additive outliers whereabouts are known and reoccurring.

2.3.3 Estimation of Multiple Known Additive Outlier Weights

Suppose now that the time series has k seasonal additive outliers $t_1^*, t_2^*, \ldots, t_k^*$ and that p and q are determined based on the AICC statistic (2.26). Let $\epsilon > 0$ be a predetermined, constant tolerance chosen by the user as a way to control the accuracy of the parameter estimates. The iterative approach is outlined in the following.

1. Use maximum likelihood estimation (2.27) to estimate $\hat{\phi}_1^{(0)}, \ldots, \hat{\phi}_p^{(0)}, \hat{\theta}_1^{(0)}, \ldots, \hat{\theta}_q^{(0)}$ based on the original time series.

2. Use (2.34) to compute the π_k's so that these can be used to determine $e_t = \pi(B)\gamma_t$ for $t = 1, \ldots, T$.

3. Perform multiple linear regression on (2.37) to estimate $\hat{\omega}_1^{(0)}, \ldots, \hat{\omega}_k^{(0)}$, where $\{e_t\}$ is the output variable and $\{\xi_{jt}\}$ are the input variables.

4. Obtain the adjusted time series

$$\tilde{\gamma}_t = \begin{cases} \gamma_t & \text{if } t \neq t_1^*, \ldots, t_k^*, \\ \gamma_t + \hat{\omega}_j^{(0)} & \text{if } t = t_1^*, \ldots, t_k^*. \end{cases}$$

   by removing the outlier effects.

5. Estimate new parameters $\hat{\phi}_1^{(1)}, \ldots, \hat{\phi}_p^{(1)}, \hat{\theta}_1^{(1)}, \ldots, \hat{\theta}_q^{(1)}$ from (2.27) based on the adjusted series $\tilde{\gamma}_t$. Use the new parameters and (2.34) to find adjusted residuals $\tilde{e}_t$ for

$t = 1, \ldots, T$. If the relative change of the residual standard error

$$\delta e_t = \sqrt{\frac{\sum_{t=1}^{T}(\tilde{e}_t - e_t)^2}{\sum_{t=1}^{n} e_t^2}}$$

is greater than the tolerance $\epsilon > 0$, set $e_t \leftarrow \tilde{e}_t$ and go to step 3 again. If $\delta e_t < \epsilon$, the procedure is stopped and the latest values $\hat{\omega}_1^{(m)}, \ldots, \hat{\omega}_k^{(m)}$ from step 3 are the estimated weights.

2.3.4 Forecasting

Let $\{\gamma_t\}$ still be a non-stationary time series consisting of T observations. Suppose we wish to predict further observations, $\gamma_{T+1}, \gamma_{T+2}, \ldots$. The best linear prediction for h time steps ahead is the conditional expectation $\hat{\gamma}_{T+h} = \hat{E}(\gamma_{T+h}|\gamma_T, \ldots, \gamma_1)$ for $h = 0, 1, 2, \ldots$ with $\gamma_1 = \hat{\gamma}_1$. Furthermore, $\hat{E}(\varepsilon_{T+h}|\gamma_T, \ldots, \gamma_1) = 0$ and $\hat{E}(\varepsilon_t|\gamma_T, \ldots, \gamma_1) = \gamma_t - \hat{\gamma}_t$ for $t = 1, \ldots, T$. Writing out the terms in (2.30) and realizing that $\varphi(B) = \phi(B)(1-B)^d = 1 - \varphi_1 B - \ldots - \varphi_{p+d} B^{p+d}$ is a polynomial of order $p+d$, we find that

$$\gamma_{T+h} = \varphi_1 \gamma_{T+h-1} + \ldots + \varphi_{p+d}\gamma_{T+h-(p+d)} + \varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + \ldots + \theta_q \varepsilon_{T+h-q} \quad (2.38)$$

where the coefficients $\varphi_{,1}, \ldots, \varphi_{p+d}$ are determined recursively. Taking the conditional expectation on both sides of (2.38) gives (Box et al., 1994, p. 131-137)

$$\begin{aligned} \hat{\gamma}_{T+h} &= \sum_{j=1}^{p+d} \varphi_j \hat{E}(\gamma_{T+h-j}|\gamma_T, \ldots, \gamma_1) + \sum_{j=1}^{q} \theta_j \hat{E}(\varepsilon_{T+h-j}|\gamma_T, \ldots, \gamma_1) \\ &= \sum_{j=1}^{p+d} \varphi_j \hat{\gamma}_{T+h-j} + \sum_{j=h}^{q} \theta_j (\gamma_{T+h-j} - \hat{\gamma}_{T+h-j}) \text{ for } h = 0, \pm 1, \pm 2, \ldots. \end{aligned} \quad (2.39)$$

Notice that (2.39) is a recursive formula, where we use $\hat{\gamma}_{T+1}$ to predict $\hat{\gamma}_{T+2}$ and so on. When h is such that at $T + h = t^*$ is a known additive outlier, we add an estimated weight $\hat{\omega}_{k+1}$. Hence,

$$\hat{\gamma}_{T+h} = \begin{cases} \sum_{j=1}^{p+d} \varphi_j \hat{\gamma}_{T+h-j} + \sum_{j=h}^{q} \theta_j (\gamma_{T+h-j} - \hat{\gamma}_{T+h-j}) & \text{for } T+h \neq t^* \\ \sum_{j=1}^{p+d} \varphi_j \hat{\gamma}_{T+h-j} + \sum_{j=h}^{q} \theta_j (\gamma_{T+h-j} - \hat{\gamma}_{T+h-j}) + \hat{\omega}_{k+1} & \text{for } T+h = t^* \end{cases} \quad (2.40)$$

The mean of the previous k estimated weights is used in order to estimate the next weight $\hat{\omega}_{k+1}$,

$$\hat{\omega}_{k+1} = \frac{1}{k} \sum_{j=1}^{k} \hat{\omega}_j. \quad (2.41)$$

2.3.5 Diagnostic Checking

Assume $(\hat{\phi}, \hat{\theta})$ have been determined through maximum likelihood estimation (2.27). We define the residuals as

$$\hat{\varepsilon}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B) w_t$$

where w_t is defined in (2.23). The residuals may be computed recursively as

$$\hat{\varepsilon}_t = y_t - \sum_{j=1}^{p} \hat{\phi}_j y_{t-j} - \sum_{j=1}^{q} \hat{\theta}_j \hat{\varepsilon}_{t-j} \text{ for } t = 1, 2, \ldots, T.$$

In our analysis, we will investigate the sample autocorrelation function of the residuals to determine if the residuals are correlated. Additionally, we will determine if the residuals are identically and independently distributed with mean 0 and variance $\hat{\sigma}^2$ (Box et al., 1994, p. 312).

2.4 Forecasting with the Mixed Effects Logistic Regression Model

This section outlines the procedure for how one can forecast with the mixed effects logistic regression model by combining the theory presented thus far. Assume a mixed effects logistic regression model is created based on a training set where we have performed random undersampling and that estimated parameters of the model have been determined. In order to forecast with the mixed effects logistic regression model, the covariates x_{itk}, determined by variable selection, must be forecasted as well.

Let the covariates x_{itk} be known up to time $t = 1, \ldots, T$ for all individuals i. We know wish to forecast the covariates $x_{i(T+h)k}$ for $h = 1, 2, \ldots$. The explanatory variables will be forecasted in different ways depending on the type of variable. Consider a categorical explanatory variable that can only take a finite number of values. Let this be the kth covariate such that for individual i at time measurement t, we may treat this as a random variable X_{itk}. Furthermore, assume that X_{itk} has the Markov property, i.e. the conditional probability of a future state only depends on the current state, and not those preceding it. For categorical variables, this means that the probability that $X_{i(T+h)k}$ is in a state x given previous states is given by

$$\begin{aligned} P(X_{i(T+h)k} = x | X_{i1k} = x_1, X_{i2k} = x_2, \ldots, X_{i(T+h-1)k} = x_{T+h-1}) = \\ P(X_{i(T+h)k} = x | X_{i(T+h-1)k} = x_{T+h-1}) \text{ for } h = 1, 2, \ldots \end{aligned}$$

Consider now the kth covariate for individual i at time $T + h$. Assume it is only possible to move up one state, down one state or remain in the same state. If for instance individual i is at state j at time $T + h - 1$, he or she can only be in either state $\{j - 1, j, j + 1\}$ at time $T + h$. Assume we want to find the probability that individual i is in state $j + 1$ at time $T + h$ given that individual i was at state j at time $T + h - 1$. We can estimate this probability based on historical data and find the proportion of customers that went from

state j to $j+1$ as

$$\hat{P}(X_{i(T+h)k} = j+1 | X_{i(T+h-1)k} = j) = \frac{\sum_i I\big(X_{i(T+h-1)k} = j, X_{i(T+h)k} = j+1\big)}{\sum_{s=\{j-1,j,j+1\}} \Big(\sum_i I\big(X_{i(T+h-1)k} = j, X_{i(T+h)k} = s\big)\Big)} \tag{2.42}$$

In other words, we count the individuals that transitioned from state j to $j+1$ and divide by the number of individuals that transitioned from state j to one of the states, $\{j-1, j, j+1\}$. Other transition state probabilities are estimated similarly.

Consider a different covariate $x_{itk'}$ with the Markov property that can still be in a finite number of M states labelled $1, \ldots, M$, but assume now that it is possible to move from a state j at time $T+h-1$ to any other state $m \in \{1, \ldots, M\}$ at time $T+h$. The estimated probability of moving to state m at time $T+h$ given state j at time $T+h-1$ is then

$$\hat{P}(X_{i(T+h)k'} = m | X_{i(T+h-1)k'} = j) = \frac{\sum_i I\big(X_{i(T+h-1)k'} = j, X_{i(T+h)k'} = m\big)}{\sum_{l=1}^{M} \Big(\sum_i I\big(X_{i(T+h-1)k'} = j, X_{i(T+h)k'} = l\big)\Big)} \tag{2.43}$$

Furthermore, the unknown prior probabilities $\gamma_{T+h} = \hat{P}(Y_{i(T+h)} = 1)$ are estimated using time series analysis as described in section 2.3. Denote the last recorded time measurement as T and assume we want to predict for further months $T+h$ for $h = 1, 2, \ldots$.

$$p_{i(T+h)} = \frac{\exp(\boldsymbol{x}_{i(T+h)}^\top \boldsymbol{\beta} + v_{0i})}{1 + \exp(\boldsymbol{x}_{i(T+h)}^\top \boldsymbol{\beta} + v_{0i})} \text{ for } i = 1, \ldots, N.$$

where the covariates $\boldsymbol{x}_{i(T+h)}$ have been forecasted properly depending on the type of explanatory variable. For an unobserved individual i, we set $v_{0i} = 0$. Once the probabilities $p_{i(T+h)}$ are calculated they are *adjusted* according to eq. (2.22) to give the adjusted probabilities $\tilde{p}_{i(T+h)}$, which we mark by $\sim$. $\hat{\alpha}$ is estimated from eq. (2.17). Finally, the number of individuals with probabilities larger than $\hat{\alpha}$ is counted. Let S_{T+h} be the total number of individuals greater than the threshold value $\hat{\alpha}$ at time $T+h$ where α is determined from (2.17). Then the estimated number of True Positives (TP) will be

$$S_{T+h} = \sum_{i=1}^{N} I(\tilde{p}_{i(T+h)} > \hat{\alpha}), \tag{2.44}$$

for $h = 1, 2, \ldots$. Consider now the problem introduced in chapter 1 of predicting the total balance sent to debt collection B_{T+h} at time $T+h$. The total number of customers sent to

debt collection at time $T + h$ will then be S_{T+h} given by equation (2.44) such that B_{T+h} is

$$B_{T+h} = \bar{A} \cdot S_{T+h} \text{ for } h = 1, \ldots, 12,$$

where $\bar{A}$ is the average amount an impaired customer owes based on the entire time period.

Chapter 3

Analysis of Results

This chapter presents, explains and analyses the results that were obtained by applying the models and techniques presented in chapter 2 on the data set provided by SpareBank 1 Kredittkort AS. The first section gives a detailed description of how the data was preprocessed before the methods presented were applied. The second section analyzes the mixed effects logistic regression model. This includes variable selection, handling of imbalanced data, parameter estimation and diagnostic checking of the model. The third section comprises the time series analysis necessary to forecast the prior probability of belonging to the minority class 1. Finally, the last section fits the model to the year 2018 and forecasts the total balance sent to debt collection for the year 2019 by combining the results found in the other sections.

All modelling and computation were done using the program **R** (R Core Team, 2018).

3.1 Data Preprocessing

The data was processed in many ways before the mixed effects logistic regression model was created. This included handling the imbalance in the data and creating a training and test set. First, the data set was made less skewed by removing some of the customers. Each customer is automatically placed in one of 9 different segments based on how much they use their credit card and if they pay their credit card bills on time and so on. The different segments are shown in table 3.1. Note that a customer can also change segment if their behavioral pattern changes. The second column shows the percentage of customers that fall into the segment and the third column shows the percentage of customers that are sent to debt collection that is from the given segment. Customers that fall into the segment `Not active in last 6 months` ($\sim 5\%$) are customers that have not yet activated or used their cards. The segment `Active in last 6 months` ($\sim 6\%$) are customers that have been issued a credit card less than 6 months ago and have just started to use their card. `Transactors` ($\sim 18\%$) use their credit cards frequently, but always pay the balance due each month. `Revolvers` ($\sim 15\%$) use their cards frequently and

Table 3.1: The different customer segments according to the variable `Segment9Name`.

Segment9Name	**%customers**	**%customers sent to debt collection**
Not active in last 6 months	$\sim 5\%$	$< 0.1\%$
Active in last 6 months	$\sim 6\%$	$\sim 9.1\%$
Transactor	$\sim 18\%$	$< 1\%$
Revolver	$\sim 15\%$	$\sim 48.1\%$
Occasional revolver	$\sim 25\%$	$\sim 32.5\%$
Revolve only	$\sim 1\%$	$\sim 8.8\%$
Last active 4-6 months ago	$\sim 4\%$	$< 0.1\%$
Last active 7-12 months ago	$\sim 5\%$	$< 0.1\%$
Not active in last 12 months	$\sim 20\%$	$< 0.1\%$

revolve their balance while `Occasional revolvers` ($\sim 25\%$) occasionally revolve. Roughly $\sim 1\%$ of customers `Revolve only`. Customers that fall into the segments `Last active 4-6 months ago` and `Last active 7-12 months ago` have not been active in the last 4-6 months and 7-12 months, respectively. Finally, customers in the segment `Not active in last 12 months` have not used their credit card for over a year. Most impaired customers are either revolvers, occasional revolvers or only revolvers as shown in the third column in table 3.1. In addition, roughly 9% of impaired customers fall into the segment `Active in last 6 months`, which means that their credit card are less than 6 months old. These customers represent those that maybe should not have been issued a credit card in the first place. There are also a few impaired customers that fall into some of the other segments, although this should technically not be possible. This can be thought of as errors in the data set. Customers that fell in the segments `Not active in last 12 mths`, `Last active 7-12 mths ago`, `Last active 4-6 mths ago`, `Not active last 6 mths` and `Transactor` were all excluded from the training set (with the exception of the few customers that were sent to debt collection and fell in one of the said groups). It is no need for the model to train on customers that will never be sent to debt collection. Removing roughly 52% of the customer base made the data set less skewed.

Furthermore, the data contains, for most customers, 15 observations recorded in the time period July 2017 to September 2018. As previously noted, the response for individual i in July 2017 is whether the individual was sent to debt collection in October 2017, i.e. three months ahead. This is the case for each observation with the response being whether the customer was sent to debt collection three months ahead. The last observation for September 2018 will therefore tell whether a customer was impaired in December 2018. This is further illustrated in table 3.2, which shows a preprocessed training set. Note that the variables will be scaled as well before constructing the model. Some examples of customers are included as well. The customer with `BK_ACCOUNT_ID 228` is a `Revolver`, but is not sent to debt collection. The customer with `BK_ACCOUNT_ID 4046` is an `Occasional revolver` and is impaired in December 2018. This is shown in the observation for September 2018. Finally, `BK_ACCOUNT_ID 1524527` is a new customer and falls into the segment `Active in last 6 months`. A mixed effects logistic regression model was made for each month in year 2018, using three observations. For instance, the model

Table 3.2: Illustration of a preprocessed training set for predicting the number of impaired customers in December 2018.

BK_ACCOUNT_ID	YearMonth	...	CustomerAge	GENDER_NAME	...	SUM_of_Payment-OverDueFlag	...	DCA0YearMonth	DCA0Ind	BalanceSent
228	201807	...	39	Male	...	0	...	NA	0	0.00
228	201808	...	39	Male	...	0	...	NA	0	0.00
228	201809	...	39	Male	...	0	...	NA	0	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4046	201807	...	25	Male	...	2	...	NA	0	0.00
4046	201808	...	25	Male	...	2	...	NA	0	0.00
4046	201809	...	25	Male	...	2	...	201812	1	881.03
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1524527	201807	...	43	Male	...	0	...	NA	0	0.00
1524527	201808	...	43	Male	...	0	...	NA	0	0.00
1524527	201809	...	43	Male	...	0	...	NA	0	0.00

predicting the number of impaired customers in January 2018 will use the observations from August 2017, September 2017 and October 2018, since the observations for October 2018 shows if a customer was sent to debt collection in January 2018. Thus, the training set contained three time measurements for each customer.

Furthermore, all explanatory variables were also scaled accordingly as some of the variables were on very different scales. The explanatory variables that had a natural lower and upper bound were scaled between 0 and 1. This was the case for variables such as `CustomerAge` and `MonthsSinceAccountCreated`. Other explanatory variables were scaled to have mean 0 and variance 1, if the variable could take on both negative and positive values. Alternatively, all explanatory variables could have been chosen to be scaled between 0 and 1, but this would possibly diminish the interpretation of the fixed effects coefficients β. Once, the preprocessing was complete, the explanatory variables most suited (section 3.2.1) according to our data could be determined and used to create the mixed effects logistic regression model (section 3.2.3).

3.2 Mixed Effects Logistic Regression Model Analysis

3.2.1 Determination of Explanatory Variables

The package **glmmLasso** was used to perform variable selection (Groll and Tutz, 2012). The explanatory variables were all standardized to have mean 0 and variance 1. Furthermore, the explanatory variables used in the mixed effects logistic regression model was determined by adding a L_1-penalty term as explained in section 2.1.5. The method was run with different proportions of positive responses in the training set to investigate if there were any differences in the explanatory variables. Not surprisingly, when the proportion between the minority and majority class became less skewed, the number of explanatory variables increased. If the ratio between positive and negative response are highly skewed, i.e. 1:99, the method will penalize more explanatory variables and set them to exactly zero. Generally, if the ratio between positive and negative responses become less skewed, the LASSO method would include more possible explanatory variables. The LASSO method was run with 15 different undersampling ratios, ranging from a proportion of 50% of customers sent to debt collection to only 6.25%. A total of 12 different models were created, one for each month of the year 2018. This was to investigate whether or not there were any differences in the explanatory variables for each month. This proved not to be the case, and the same explanatory variables were chosen for each month. The approach for determining the explanatory variables was the following; a range of possible values for the tuning parameter $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ were chosen where the values were descending $\lambda_j > \lambda_{j+1}$ for $j = 1, \ldots, n-1$.

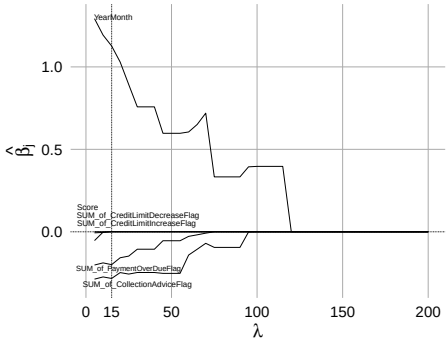


Figure 3.1: Fixed-effects coefficients estimates versus the penalty parameter λ for the month of May. The vertical dotted line shows the optimal value for λ. Note that not all possible explanatory variables are included as there would be far too many lines. The figure is mainly meant to illustrate the variable selection method process.

For each $j = 1, \ldots, n$

1. Fit a mixed effects logistic regression model and compute the estimated fixed-effects parameters $\hat{\boldsymbol{\beta}}$ in the model with λ_j.

2. Compute the Bayesian Information Criterion (BIC$_j$) from eq. (2.15) for the model.

Choose the value of j for which the BIC$_j$ is at a minimum. Use the corresponding value of λ_j and find the significant variables $\neq 0$ at this λ_j value. The model was run with $\lambda_1 = 200$ and $\lambda_{40} = 5$, i.e. the tuning parameter was decreased by 5 units for each run. At $\lambda_1 = 200$, none of the covariates were significant. The appropriate percentage of accounts

sent to debt collection was chosen as described in section 3.2.2. Some of the estimated fixed-effects coefficients are plotted versus the tuning parameter λ for the month of May as shown in figure 3.1. Note that there were more than 70 possible explanatory variables, and most are omitted from figure 3.1 for clarification. The plot shows how more explanatory variables are different from zero as λ decreases. The variable `YearMonth` is the first to be different from zero and others follow as λ decreases. The BIC is minimized at the vertical dotted line, i.e. at $\lambda_{\text{opt}} = 15$. There were some possible explanatory variables not included in the model that were significant for some of the months. However, we decided not to include them since having too many explanatory variables in a GLMM generally makes the model more unstable (Bolker et al., 2009). The explanatory variables chosen by our variable selection method, as well as a description, justification and possible explanation for its presence in the model, are presented in the following.

YearMonth

This variable shows the year and month on the format YYYYMM and is scaled to be between 0 and 1. This is the time-trend in the model and the fixed-effects coefficient is positive.

CustomerAgeSquared

This variable was constructed and is the customer's age squared. The variable selection method suggest that a linear relationship between $\text{logit}(p_{it})$ and the customer's age squared is reasonable. One would expect that customers who are older generally have a more solid economy as they have worked for many years, while younger customers may take more risk and are less stable financially. It is therefore reasonable that the fixed effect coefficient for `CustomerAgeSquared` is negative.

MonthsSinceAccountCreatedSquared

For a credit card company, it may be hard to determine whether a customer should have been given a credit card in the first place. Many of the impaired customers have just recently received their credit card as well, whereas customers who have had their credit card for many years tend to be more stable and pay their bills. The fixed-effects coefficient is negative, which is also reasonable.

Score

As previously mentioned, the `Score` variable is a risk score value computed through a serious of computations to give each customer an assigned value between 0 and 7 based on how they use their credit card. For example, if the revolving balance utilization the last 3 months is greater than some value, a customer's score is increased by +1. A `score` of 0 indicates that the customer has a very low risk of delinquency, whereas a score of 7 indicates a very high risk of delinquency, almost 20 times as high delinquency rate as the average. It is therefore reasonable that the fixed effect coefficient for the `Score` value is positive, as a customer with a higher score value is more likely to be impaired.

`SUM_of_PaymentOverDueFlag`

The variable records the total number of times a customer has been overdue with his or her payment in the last 12 months. It is therefore reasonable that the fixed effect coefficient estimate for `SUM_of_PaymentOverDueFlag` is positive as well.

Hence, the mixed effects logistic regression model used had six explanatory variables, including the fixed-effect intercept, in addition to the random intercept term unique for each customer.

3.2.2 Determination of the Undersampling Ratio

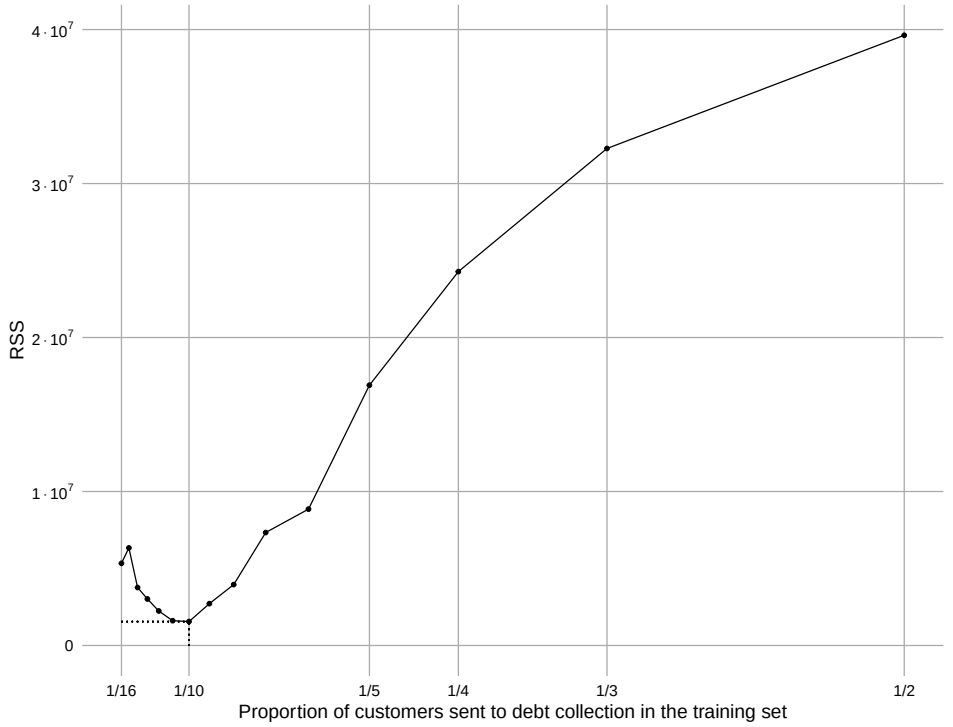


Figure 3.2: The residual sum of squares versus proportion of customers sent to debt collection. The RSS is minimized when the percentage of impaired customers in the training set is 10%.

The data set provided by SpareBank 1 Kredittkort AS was highly skewed with 0 (non-impaired) as the majority class and 1 (impaired) as the minority class. This is not surprising as most customers are not sent to debt collection every month. Roughly $\sim 1\%$ of the responses were positive indicating that a customer had been sent to debt collection. In order to combat this imbalance, random undersampling was performed as described in section 2.2.1. We will refer to the undersampling ratio r as the percentage of customers

sent to debt collection in the training set. This was to decide the optimal percentage of impaired customers that should be included in the training set by solving equation (2.19). If the ratio in the training set between impaired and non-impaired customers is too skewed, e.g. 1:99, the model will classify virtually all customers as non-impaired. If the ratio is perfectly balanced, i.e. 1:1, the training set will be much smaller since there are far fewer instances of impaired customers. The training set may become so small that the model will not have enough instances to train on. It should also be noted that the explanatory variables will be scaled very differently if the undersampling ratio in the training set is very different from the actual ratio in the real data set. A method was proposed to determine the optimal ratio of undersampling, r. For 10 different undersampling ratios, 12 models were created, one for each month of the year 2018. The Residual Sum of Squares (RSS) was computed as in eq. (2.18) and plotted versus different percentages of customers sent to debt collection. Figure 3.2 shows the plot. The undersample ratio ranges from $1/2$ to $1/16 = 0.0625$. The RSS is very high when the ratio between the minority and majority class is too evenly distributed. The RSS decreases when the undersampling ratio becomes smaller and reaches a minimum when the proportion of customers sent to debt collection in the training set is $1/10 = 10\%$. This was therefore chosen as the percentage of impaired customers in the training set that was used when creating the mixed effects logistic regression model. When the undersampling ratio becomes even smaller the RSS starts to increase again as shown in figure 3.2.

3.2.3 Estimation of Parameters

Once the undersample ratio was determined to be 10%, we used the **glmer** function in the package **lme4** (Bates et al., 2015) with 10 quadrature points in the adaptive Gauss-Hermite quadrature to estimate the parameters in the mixed effects logistic regression model. As noted, 12 models were created, one for each month of the year with 3 observations per customer. For instance, the data for customers from January, February and March was used to predict the total number of impaired customers in June. The grouping factor was each customer's BK_ACCOUNT_ID. The explanatory variables chosen in section 3.2.1 means that the model can be written as

$$\begin{aligned}\log\left(\frac{\hat{P}(Y_{it}=1)}{1-\hat{P}(Y_{it}=1)}\right) = (\hat{\beta}_0 + \hat{v}_{0i}) &+ \texttt{YearMonth}_{it}\hat{\beta}_1 + \texttt{CustomerAgeSquared}_{it}\hat{\beta}_2 \\ &+ \texttt{MonthsSinceAccountCreatedSquared}_{it}\hat{\beta}_3 \\ &+ \texttt{Score}\hat{\beta}_4 + \texttt{SUM\_of\_PaymentOverDueFlag}\hat{\beta}_5\end{aligned}$$

Note that the random individual effect v_{0i} represents the deviation for customer i from the group trend. Table 3.3 shows the estimated fixed-effects coefficients for each of the 12 models. The standard errors are written in parenthesis and is computed from the bootstrapping of the parameters (see section 3.2.4). Note that there is some variation in the coefficients depending on the month in question.

Table 3.3: Estimated fixed-effects parameters for every month. Standard errors are written in parenthesis.

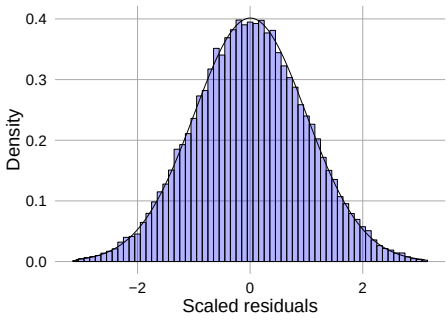
Month	`Intercept`, $\hat{\beta}_0$	`YearMonth`, $\hat{\beta}_1$	`CustomerAgeSquared`, $\hat{\beta}_2$	`MonthsSinceAccount-CreatedSquared`, $\hat{\beta}_3$	`Score`, $\hat{\beta}_4$	`SUM_of_Payment-OverDueFlag`, $\hat{\beta}_5$
January	-9.497 (0.291)	4.178 (0.143)	-2.277 (0.361)	-2.949 (0.253)	1.087 (0.230)	1.015 (0.039)
February	-9.366 (0.310)	4.220 (0.154)	-3.120 (0.383)	-3.467 (0.391)	1.206 (0.243)	0.963 (0.041)
March	-9.125 (0.290)	4.025 (0.147)	-2.847 (0.404)	-3.758 (0.358)	1.515 (0.235)	0.857 (0.035)
April	-9.253 (0.277)	4.031 (0.132)	-2.780 (0.370)	-2.676 (0.263)	0.927 (0.238)	0.999 (0.039)
May	-9.225 (0.286)	4.152 (0.146)	-1.992 (0.351)	-2.657 (0.248)	1.135 (0.230)	0.900 (0.035)
June	-8.933 (0.271)	4.134 (0.138)	-2.799 (0.353)	-3.816 (0.365)	1.052 (0.232)	0.843 (0.035)
July	-8.654 (0.274)	4.016 (0.148)	-3.321 (0.407)	-2.097 (0.276)	0.819 (0.234)	0.876 (0.039)
August	-9.030 (0.306)	3.898 (0.146)	-2.422 (0.409)	-2.961 (0.305)	0.543 (0.268)	1.027 (0.044)
September	-9.701 (0.288)	4.173 (0.141)	-2.419 (0.372)	-3.149 (0.271)	1.079 (0.215)	1.020 (0.040)
October	-9.477 (0.280)	4.058 (0.132)	-2.525 (0.332)	-2.910 (0.257)	1.618 (0.239)	0.960 (0.036)
November	-9.742 (0.316)	4.153 (0.146)	-2.902 (0.382)	-2.937 (0.282)	1.927 (0.239)	0.893 (0.038)
December	-9.221 (0.276)	4.003 (0.138)	-2.863 (0.365)	-2.721 (0.249)	1.420 (0.237)	0.878 (0.036)

3.2.4 Diagnostic Checking

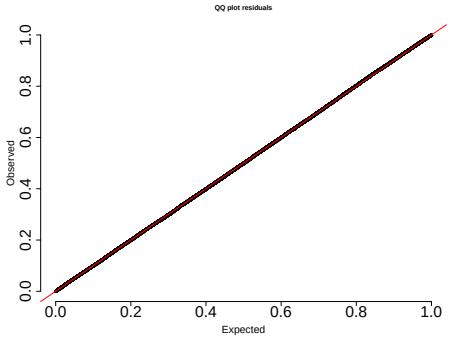
This section comprises first the diagnostic checking of the residuals and then the fixed-effects parameters in the mixed effects logistic regression model. The scaled residuals were simulated a total of 1000 times as outlined in section 2.1.6.

Residual Diagnostics

The uniformly distributed rescaled residuals were transformed to be normally distributed in figure 3.3, which shows the distribution of the scaled residuals for the month of January, presented in a histogram and a Q-Q plot. The residual plots for the other months are similar and is therefore not included. The mean of the scaled residuals is -0.0047 and the variance is 0.98. The shape of the histogram in (**a**) suggest that the residuals appear to be normally distributed and symmetric around zero. The normality is further backed by the Q-Q plot in (**b**) which shows that the scaled residuals follow the straight line marked red very well.



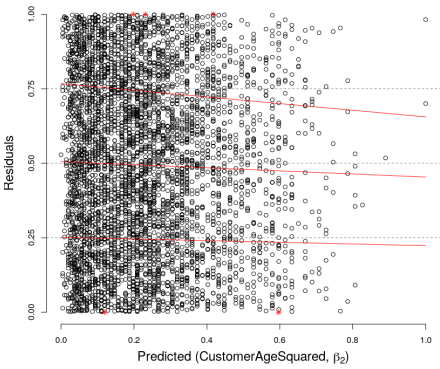
(a) Simulated scaled residuals for the mixed effects logistic regression model for the month of January.



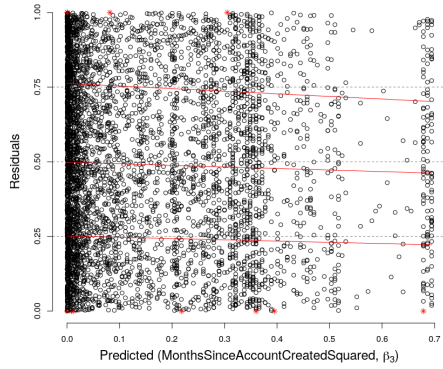
(b) Q-Q plot of the simulated scaled residuals for the mixed effects logistic regression model

Figure 3.3: Simulated scaled residuals from the mixed effects logistic regression model for the month of January shown in a histogram and Q-Q plot. The scaled residuals appear to follow a normal distribution and the residuals follows the straight line in the Q-Q plot.

The scaled uniform residuals were also plotted versus the explanatory variables `CustomerAgeSquared` and `MonthsSinceAccountCreatedSquared` for the month of January. Figure 3.4(**a**) shows the scaled uniform residuals versus the explanatory variable `CustomerAgeSquared` for 2000 randomly sampled observations. The plot provides 0.25, 0.50 and 0.75 quantile lines across the plot that should match the red lines. The red lines in (**a**) generally follows the dotted lines, but are slightly below the quantile lines towards the end. This is also true for the red lines in 3.4(**b**), showing the scaled uniform residuals versus `MonthsSinceAccountCreatedSquared`. Notice that outliers are marked red.



(a) Scaled uniform residuals versus the explanatory variable `CustomerAgeSquared`.



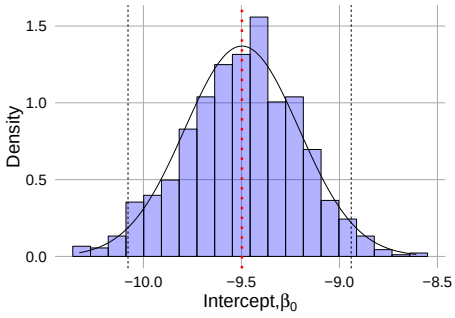
(b) Scaled uniform residuals versus the explanatory variable `MonthsSinceAccountCreatedSquared`.

Figure 3.4: Simulated scaled uniform residuals versus the explanatory variables `CustomerAgeSquared` (**a**) and `MonthsSinceAccountCreatedSquared` (**b**). The red lines are slightly below the quantile lines towards the end in both (**a**) and (**b**).

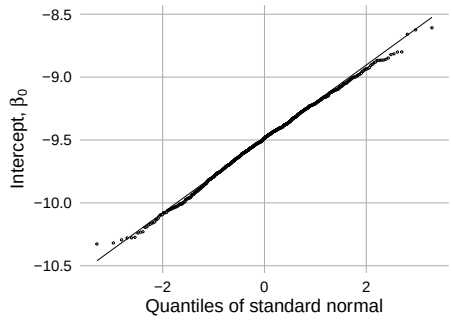
Parameter Diagnostics

Non-parametric bootstrapping of the parameters $(\boldsymbol{\beta}, \sigma_v)$ in the mixed effects logistic regression model was performed. The process is further described in section 2.1.6. The bootstrap procedure was done $B = 1000$ times. Histograms and Q-Q plots for the fixed-effects coefficients are shown in figure 3.5 and 3.6 for the month of January. Similar results are obtained for the other months. The red dotted line in the histograms show the mean bootstrapped fixed-effect coefficient and the black dashed lines show a 95% percentile bootstrap confidence interval based on the replicates, $(\hat{\beta}^*_{k,(0.025)}, \hat{\beta}^*_{k,(0.975)})$, where $\hat{\beta}^*_{k,(0.025)}$ and $\hat{\beta}^*_{k,(0.975)}$ denotes the 0.025 and 0.975 percentile of the bootstrapped coefficients for $k = 0, 1, \ldots, 5$. The histogram and Q-Q plot for 1000 bootstrapped `Intercept` coefficients are shown in figure 3.5(**a**) and 3.5(**b**). The shape of the histogram appears to be fairly normal, although there is a high spike right of the mean. The tails in the Q-Q plot are somewhat off. The 95% percentile confidence interval is rather wide, $(\hat{\beta}^*_{0,(0.025)}, \hat{\beta}^*_{0,(0.975)}) = (-10.079, -8.940)$. The histogram for the vari-

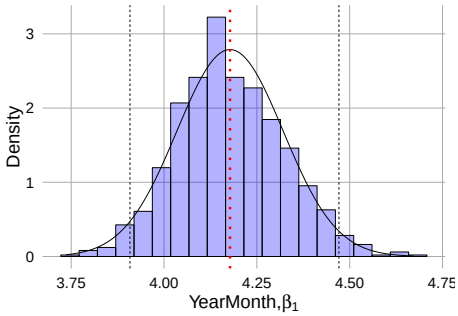
able `YearMonth` has a high spike to the left of the mean but looks otherwise normally distributed, as shown in 3.5(**c**). 3.5(**d**) shows that the tails for `YearMonth` are slightly above the straight line. The 95% percentile confidence interval is $(\hat{\beta}^*_{1,(0.025)}, \hat{\beta}^*_{1,(0.975)}) = (3.908, 4.471)$. 3.5(**e**) and (**f**) analyze the variable `CustomerAgeSquared`. The histogram appears to follow the normal distribution line but is not entirely symmetric around the mean. The Q-Q plot shows that the bootstrapped parameters are below the straight line at the tails. The 95% percentile confidence interval for `CustomerAgeSquared` is rather wide, $(\hat{\beta}^*_{2,(0.025)}, \hat{\beta}^*_{2,(0.975)}) = (-3.013, -1.590)$.



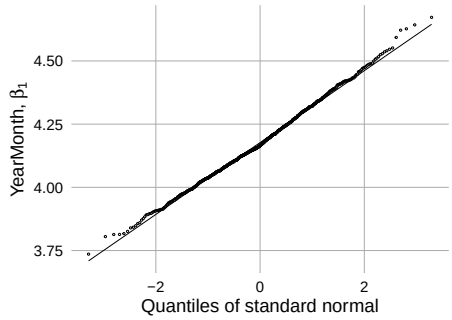
(a) Histogram of `Intercept`, β_0



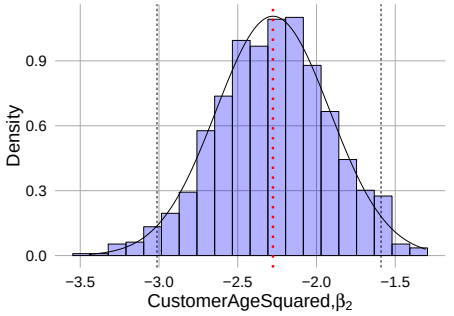
(b) Q-Q plot of `Intercept`, β_0



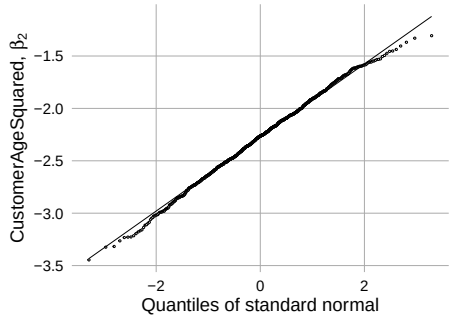
(c) Histogram of `YearMonth`, β_1



(d) Q-Q plot of `YearMonth`, β_1

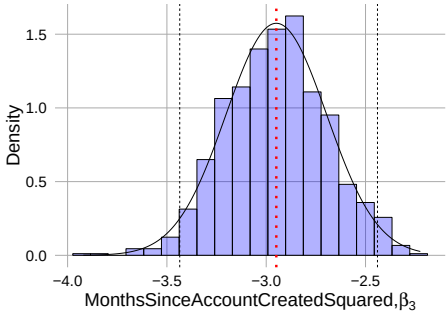


(e) Histogram of `CustomerAgeSquared`, β_2

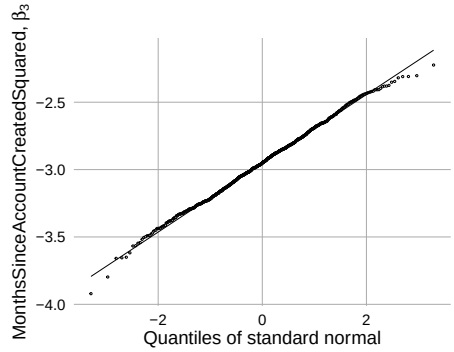


(f) Q-Q plot of `CustomerAgeSquared`, β_2

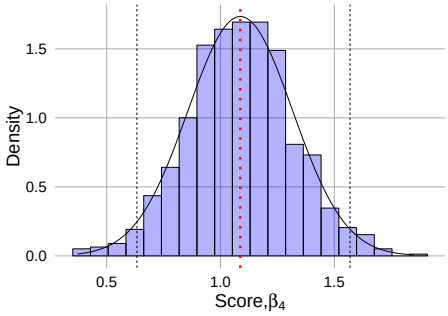
Figure 3.5: Histograms and Q-Q plots for the first three fixed-effects coefficients. The red dotted line in the histograms shows the mean value for the coefficient and the black dashed lines shows a 95% percentile bootstrap confidence interval.



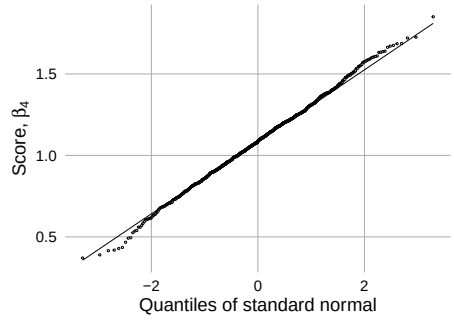
(a) Histogram of `MonthsSinceAccountCreatedSquared`, β_3



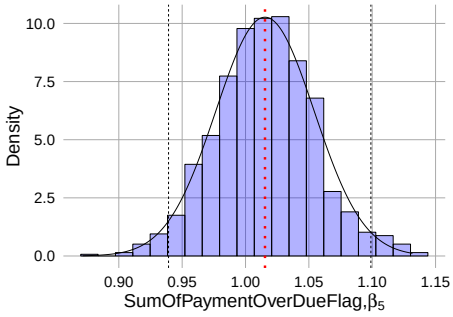
(b) Q-Q plot of `MonthsSinceAccountCreatedSquared`, β_3



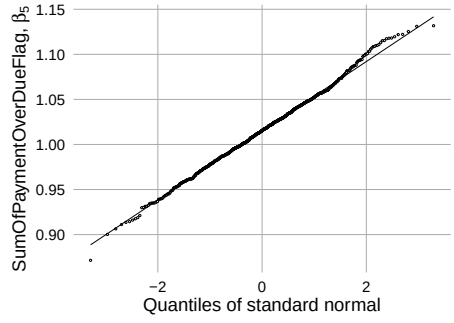
(c) Histogram of `Score`, β_4



(d) Q-Q plot of `Score`, β_4



(e) Histogram of `SUM_of_PaymentOverDueFlag`, β_5

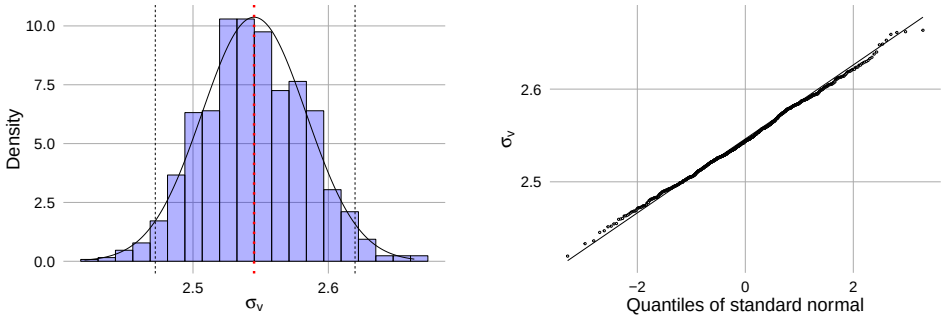


(f) Q-Q plot of `SUM_of_PaymentOverDueFlag`, β_5

Figure 3.6: Histograms and Q-Q plots for the last three fixed-effects coefficients based on 1000 bootstrap replicates. The red dotted line shows the mean value for the coefficient and the black dashed lines shows a 95% percentile bootstrap confidence interval.

Figure 3.6 shows the diagnostic checking of the last three explanatory variables included in the model. 3.6(**a**) shows that the `MonthsSinceAccountCreatedSquared` also has a high spike to the right of the mean, but is otherwise fairly symmetric. The start and end tail is not entirely on the line either as shown in 3.6(**b**), while $(\hat{\beta}^*_{3,(0.025)}, \hat{\beta}^*_{3,(0.975)}) = (-3.435, -2.440)$. Furthermore, the `Score` variable is not entirely symmetric around the mean as illustrated in 3.6(**c**) and both the start and end tail in the Q-Q plot does not follow the straight line (3.6(**d**)). The bootstrap confidence interval for the `Score` coefficient is $(\hat{\beta}^*_{4,(0.025)}, \hat{\beta}^*_{4,(0.975)}) = (0.633, 1.569)$. Finally, 3.6(**e**) and 3.6(**f**) shows that the `SUM_of_PaymentOverDueFlag` is not entirely symmetric around its mean either, but the tails are close to the line in the Q-Q plot. The bootstrap confidence interval is more narrow for this variable, $(\hat{\beta}^*_{5,(0.025)}, \hat{\beta}^*_{5,(0.975)}) = (0.939, 1.099)$.

Furthermore, we performed non-parametric bootstrapping on the random-effects parameter estimate, σ_v, i.e. the conditional standard deviation for the grouping factor `BK_ACCOUNT_ID` (see section 2.1.2). The histogram and Q-Q plot for σ_v are shown in figure 3.7. The bootstrap mean is 2.545 and the percentile confidence interval is $(\hat{\sigma}^*_{v,0.025)}, \hat{\sigma}^*_{v,0.975)}) = (2.472, 2.620)$. The histogram in figure 3.7(**a**) is not symmetric around the mean and the tails are somewhat off the straight line in the Q-Q plot in figure 3.7(**b**).



(a) Histogram of σ_v.

(b) Q-Q plot of σ_v.

Figure 3.7: Histogram and Q-Q plot for the standard deviation σ_v based on 1000 bootstrap replicates. The red dotted line shows the mean value for σ_v and the black dashed lines shows a 95% percentile bootstrap confidence interval.

3.3 Prior Probability Unknown - Time Series Analysis

This section comprises the time series analysis that was performed to forecast the adjusted prior probabilities $\gamma_t = p(1_t)$ of belonging to class 1 (impaired) for 2019. We were given the proportion of customers sent to debt collection γ_t for the time period January 2015 to December 2018. Based on this, an ARIMA model, with known additive outliers to cope with the July effect, was created. The historical data was adjusted by removing certain customers that fell in some segments as described in section 3.1. We therefore refer to γ_t as the adjusted prior probabilities. The time series analyzed is shown in figure 3.8. The time series consist of $T = 48$ data points, 12 for each of the 4 years. The red dots show how the adjusted percentage number of impaired customers drops for the month of July. Furthermore, the adjusted percentage sent to debt collection was generally lower in 2015 compared to the other years.

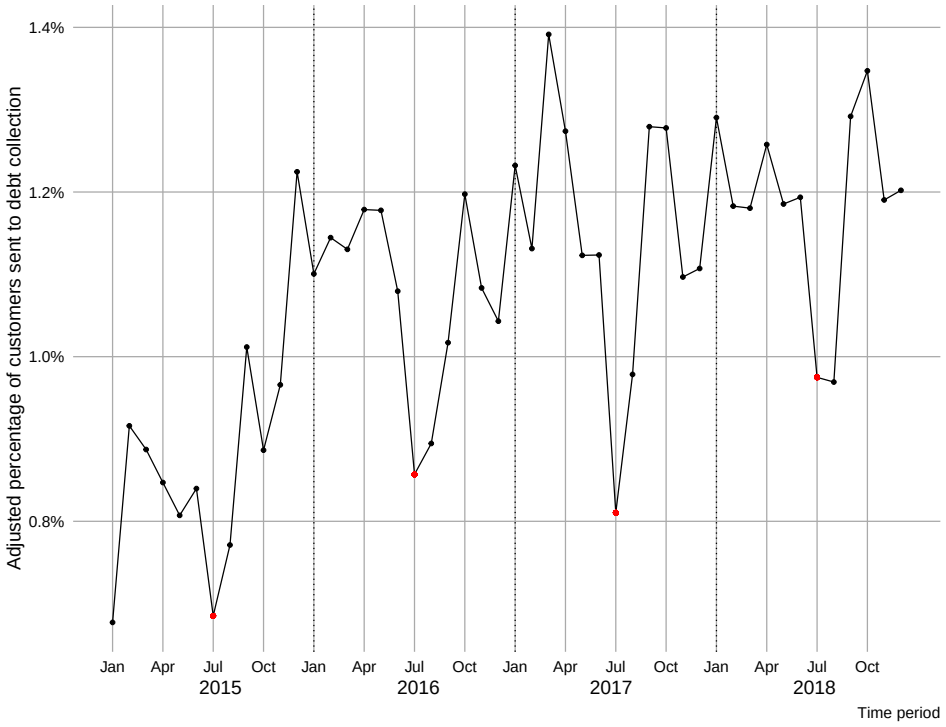


Figure 3.8: Adjusted percentage of customers sent to debt collection each month in the time period January 2015 to December 2018. The adjusted percentage is heavily reduced for each July, marked by red dots.

3.3.1 Identification and Estimation of Parameters

The time series $\{\gamma_t\}$ was modelled as an ARIMA(p, d, q)-process with weights on the known additive outliers for each July month when the number of impaired customers sharply decreased. We first determined p, d, q. Differencing the time series once, $w_t = \gamma_t - \gamma_{t-1}$ gave a sufficiently (weak) stationary time series. This can be shown in figure 3.8, which appears to have a generally increasing trend. Notice that figure 3.8 is very similar in shape to figure 1.2 and 1.5. Hence, $d = 1$. The order of the autoregressive model p and the moving average model q were determined through an investigation of the sample ACF and PACF plots of w_t, as well as the AICC criterion described in section 2.3.1. The ACF, PACF and AICC plots for w_t are shown in figure 3.9.

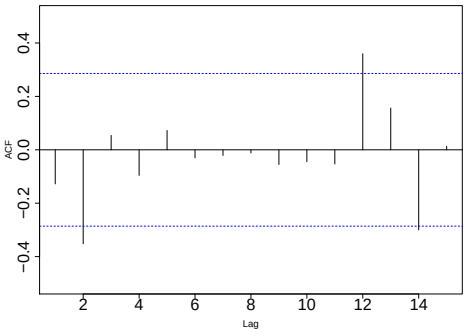
A clear spike at lag 12 is shown in the ACF plot in 3.9(**a**), but this was ignored as we have proposed an alternative way of modelling seasonality as described in section 2.3.2. Furthermore, there were significant spikes at lag 2 for both the ACF and PACF plot. However, the AICC criterion was mainly relied on since the AICC considers that the time series was very short and will thus reduce the risk of overfitting. 3.9(**c**) shows a 3-dimensional plot of the AICC for $p, q \in \{0, 1, \ldots, 5\}$. Notice that the AICC penalizes higher values of p and q. The AICC is minimized for $p = 0, q = 2$ and is slightly higher for $p = 1, q = 1$. These two candidates were investigated. An ARMA(p, q) process is said to be causal if the autoregressive polynomial $\phi(z) \neq 0$ for all $|z| \leq 1$. Thus, for the differenced, causal time series, $w_t = \nabla y_t$, we have that

$$w_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

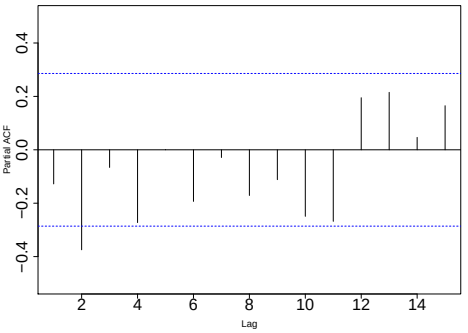
where the coefficients are given by the recursive formula $\psi_0 = 1,\ \psi_j = \theta_j + \sum_{i=1}^{p} \phi_i \psi_{j-i}$ for $j = 1, 2, \ldots$. We found that there are relatively small differences between the two candidates since the third coefficient ψ_3 is small. Therefore, we chose to model the time series with $p = 1, q = 1$.

3.3.2 Fitting the Time Series Model

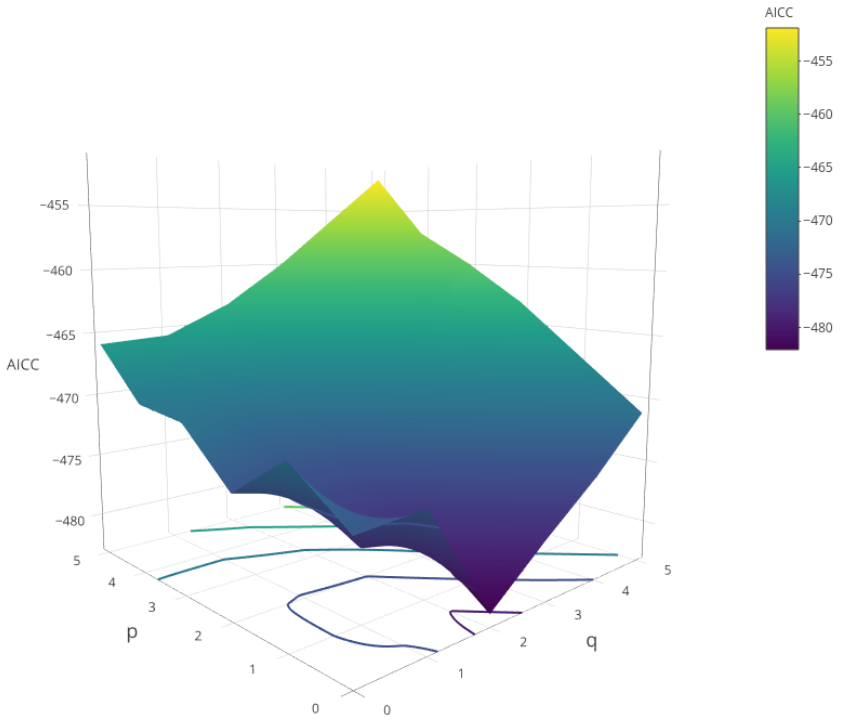
The ARIMA$(1, 1, 1)$ model with additive outlier weights was fitted. The parameters $\hat{\phi}_1, \hat{\theta}_1$ are determined by minimizing equation (2.29). The weights $\hat{\omega}$ were determined through the iterative approach presented in section 2.3.3. There are four weights in total, one for each July in the years 2015 - 2018. The estimated parameters are shown in table 3.4. c denotes the intercept of the model. Notice that $\hat{\theta}_1$ is very close to a unit root. Figure 3.10 shows the observed time series along with the fitted time series. The observed time series is the solid line and the dashed line is the fitted time series. The fitted time series fits the original time series well for the most part but are somewhat off hitting the low and high spikes. The known additive outliers for July are colored purple.



(a) ACF for the differenced time series w_t



(b) PACF for the differenced time series w_t



(c) AICC(p, q) plot for $p, q = \{0, 1, \ldots, 5\}$.

Figure 3.9: Sample ACF plot (**a**) and sample PACF plot (**b**) for the differenced time series w_t. 3-dimensional AICC(p, q) plot (**c**) for $p, q = \{0, 1, \ldots, 5\}$. The AICC is minimized for $p = 0, q = 2$. The clear spike at lag 12 in the ACF plot is ignored since differencing with respect to seasonality would further reduce the length of the time series and a different approach is proposed (see section 2.3.2).

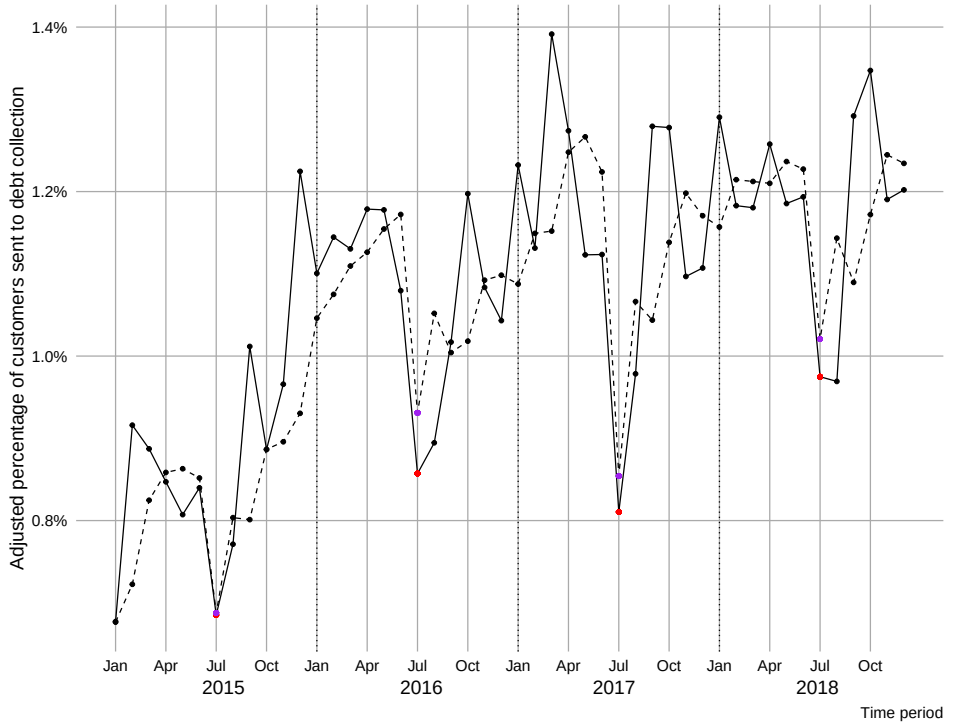


Figure 3.10: Observed time series (solid line) plotted with the fitted time series (dashed line) for the time period 2015 to 2018 for $p = 1, q = 1$. The fitted time series fits the original time series well.

Table 3.4: Estimated parameters in the time series model.

$\hat{\phi}_1$	0.4564
$\hat{\theta}_1$	-0.9995
c	$8.2806 \cdot 10^{-5}$
$\hat{\sigma}^2$	$1.5727 \cdot 10^{-6}$
$\hat{\omega}_1$	-0.002242
$\hat{\omega}_2$	-0.003616
$\hat{\omega}_3$	-0.005087
$\hat{\omega}_4$	-0.003640

3.3.3 Forecasting

Figure 3.11 shows the forecasts of the prior probabilities γ_t for the year 2019 with an ARIMA$(1, 1, 1)$ model. The time series captures the linear trend of the data. The shaded blue area shows an 80% prediction interval for the predictions. The prediction interval becomes wider for forecasts further into 2019. The estimated weight for July 2019 is

taken as the average of the estimated weights according to eq. (2.41). The unknown

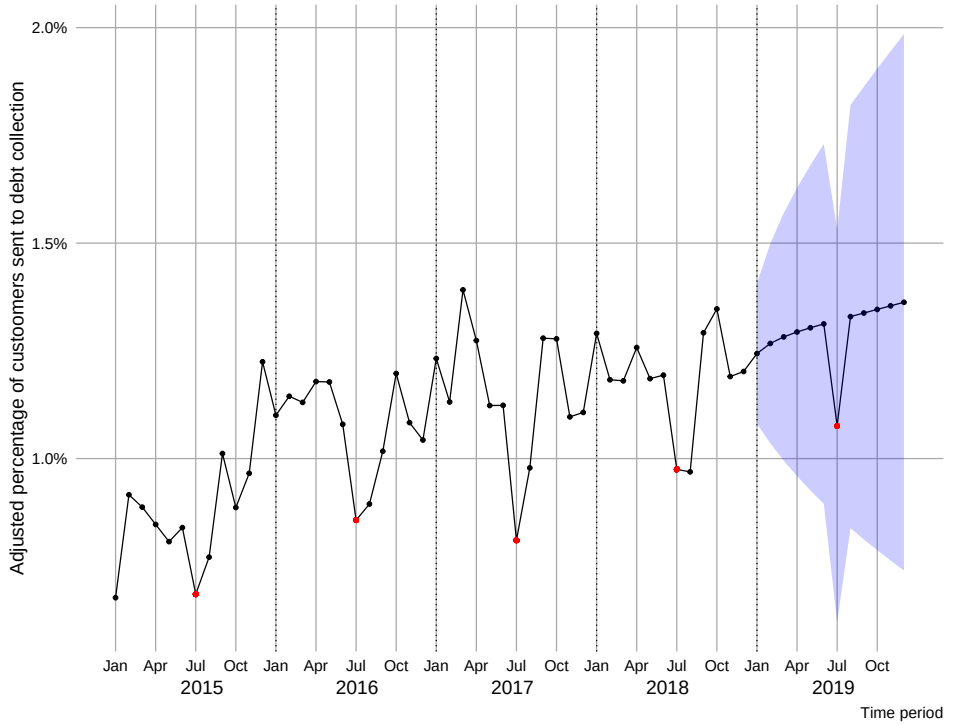


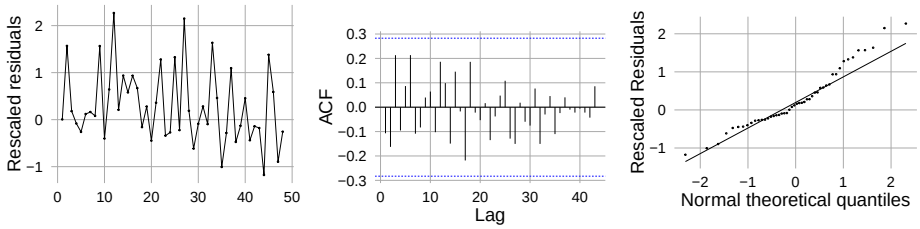
Figure 3.11: Forecasting the prior probabilities of belonging to the minority class for the year 2019 with a 80% prediction interval. The prior probabilities were forecasted to be generally increasing, except for July 2019.

prior probabilities will capture the seasonality in the total balance sent to debt collection. These forecasted values can now be used, in combination with the mixed effects logistic regression model, to forecast the total balance sent to debt collection for the year 2019.

3.3.4 Diagnostic Checking

The goodness of fit of the time series is evaluated and shown in figure 3.12 by investigating the rescaled residuals. It is assumed that the rescaled residuals are independent and identically distributed random variables with mean 0 and variance 1. The rescaled residuals are illustrated through three different plots. First, the rescaled residuals are shown in (**a**). The minimum rescaled residuals is $\hat{\varepsilon}_{\min} = -1.1768$ and the maximum is $\hat{\varepsilon}_{\max} = 2.2681$. The mean of the rescaled residuals is above 0, at 0.28 and the rescaled variance is 0.62. There seems to be a slight decreasing trend, which is further investigated in (**b**). The blue dashed lines shows the bounds $\pm z_{0.025}/\sqrt{T} = \pm 1.96/\sqrt{48} = \pm \approx 0.041$ for a 95% confidence interval. Spikes outside the dashed blue lines shows values significantly different from zero. As shown, none of the rescaled residuals have significant spikes. Finally, (**c**) shows

a Q-Q plot for the rescaled residuals. They do not appear to follow the straight line very well, especially towards the end when some residuals are well above the line. Overall, the model does not fit the time series for all data, with some residuals being very large.



(a) Rescaled residuals

(b) ACF of rescaled residuals

(c) Q-Q plot of rescaled residuals

Figure 3.12: Diagnostic checking of the time series applied to the prior probabilities. (**a**) shows the rescaled residuals, (**b**) is the ACF plot for the rescaled residuals and (**c**) shows a Q-Q plot for the rescaled residuals.

3.4 Forecasting with the Mixed Effects Logistic Regression Model

This section presents the results of forecasting with the mixed effects logistic regression model. The first subsection fit the model to the year 2018. The second subsection presents how the explanatory variables were forecasted. Finally, the last subsection forecasts the total balance sent to debt collection for the year 2019.

3.4.1 Fitted Model for 2018

The average fitted predictions for the total balance sent to debt collection for 2018 are shown in figure 3.13.

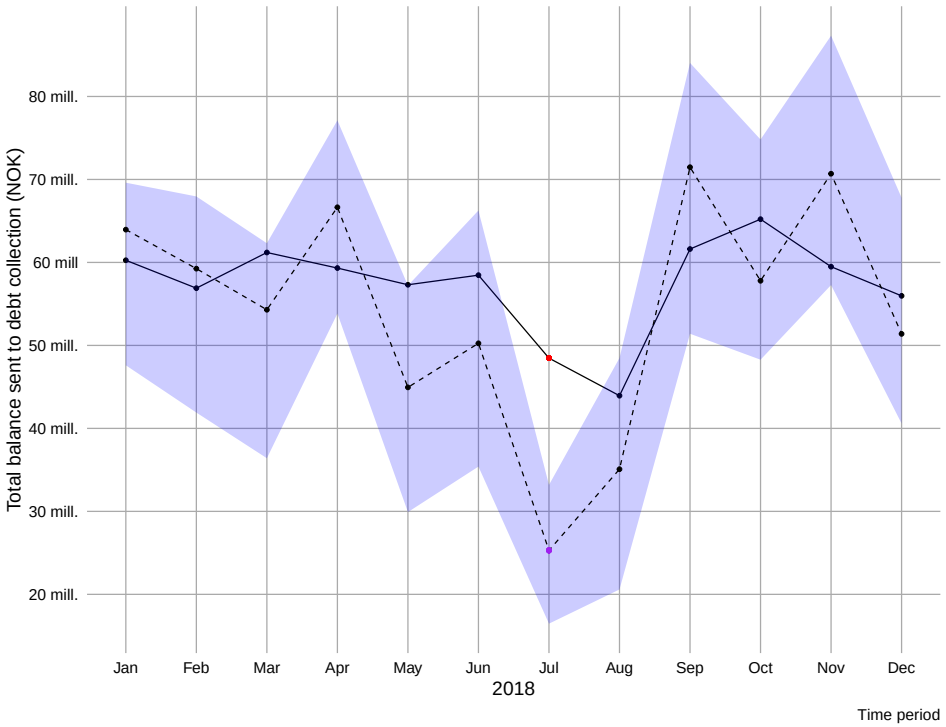


Figure 3.13: Average fitted predictions for the total balance sent to debt collection for 2018. The predicted number of impaired customers each month is multiplied by the average balance sent to debt collection for 2018. The solid line shows the actual balance sent to debt collection and the dashed line are the fitted values. The blue shaded area shows a 95 percentile bootstrap confidence interval.

The blue shaded area shows a 95 percentile bootstrap confidence interval based on bootstrapped parameter $(\hat{\boldsymbol{\beta}}^{*(b)}, \hat{\sigma}_v^{*(b)})$ replicates for $b = 1, \ldots, 1000$. The confidence interval

contains the actual number of impaired customers each month, except for July, although it should be noted that the confidence interval is very wide. The model does not give the same prediction every time the model is run. Due to this instability of the model, we took the average of 50 predictions for each prediction. The instability is further discussed in section 4.1.1. The predicted number of customers sent to debt collection for each month is multiplied by the average balance sent to debt collection for 2018. The average balance sent to debt collection for 2018 was 31529.27 NOK. (It should also be noted that an attempt to draw from a gamma distribution that fitted the balances sent to debt collection did not produce better results). Therefore, we chose to simply use the mean. A prediction that is 100 customers greater or lower than the actual number of impaired customers will therefore give predictions that are roughly 3 million NOK greater or lower than the actual total balance sent to debt collection. The predictions do not fit well, especially for the summer months July and August when the adjusted prior probability of belonging to class 1 is much lower compared to the other months. Although the predictions should be significantly lower for these two months, they are simply too low compared to the actual number of impaired customers. The fitted model fits better for the first four months and the last four months of the year 2018. The predicted number of impaired customers each month is shown in table 3.5.

Table 3.5: Predicted number of impaired customers each month in 2018.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Impaired customers, S_t	1933	1767	1809	1926	1822	1844	1508	1479	1997	2090	1828	1810
Predicted impaired customers, $\hat{S}_t$	2051	1838	1609	2164	1429	1586	788	1181	2317	1853	2174	1664

3.4.2 Development of the Explanatory Variables

The explanatory variables `Score` and `SUM_of_PaymentOverDueFlag` included in the model can only take a finite number of values and was modelled as a Markov chain where each value is a state as described in section 2.4. `SUM_of_PaymentOverDueFlag` can only take the states $\{0, 1, \ldots, 12\}$. (The variable shows how many times a customer has been overdue with their payments in the last 12 months and it is therefore technically not possible to have a value higher than 12). The probabilities of moving from one state to another are estimated from the historical data according to equation (2.42). Most people tend to stay in the same state. If a customer has never received a dunning from the credit card company and is in state 0, the customer is likely not to receive a dunning next month either. The different probabilities of moving from one state to another for the covariate `SUM_of_PaymentOverDueFlag` from January to February is shown in figure 3.14. Note that the Markov chain is only shown up until state 5 as very few customers are in states above this. The last node marked $\cdots$ is to illustrate that the Markov chain continues up until state 12.

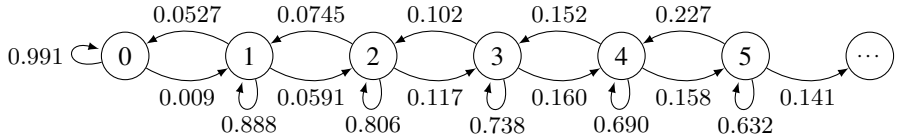


Figure 3.14: Illustration of the different states for `SUM_of_PaymentOverDueFlag` with estimated transition probabilities for customers moving a state from January 2018 to February 2018. The Markov chain is ended at state 5 as there are very few customers in states above this.

More than 99% of customers will stay in state 0 and pay their billing in time, while 0.9% of customers will be overdue and move up to state 1. Customers that are in state 1 have a higher probability of reaching state 2 and customers in state 2 will have a higher probability of reaching state 3. That means that customer that have been overdue with their payments are more likely to be overdue again. Based on these estimate probabilities we simulated the covariates for $\texttt{SUM\_of\_PaymentOverDueFlag}_{i(T+h)}$ for $i = 1, \ldots, N$, $h = 1, 2, \ldots$.

The estimated probabilities for the `Score` covariate were also computed. Each customer is assigned a `Score` between 0 and 7, where 0 indicates a customer with a very low risk of delinquency and 7 indicates a customer with a very high risk of delinquency. For this covariate, it is possible for a customer to move up and down more than one state. Therefore, the estimated probabilities are rather shown in a transition matrix.

$$P_{\texttt{Score}} = \begin{array}{c|cccccccc} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 0 & 0.469 & 0.394 & 0.126 & 0.010 & 0.0008 & 0.0002 & 0 & 0 \\ 1 & 0.015 & 0.662 & 0.304 & 0.017 & 0.002 & 0.0002 & 0.00002 & 0 \\ 2 & 0.002 & 0.114 & 0.839 & 0.034 & 0.009 & 0.002 & 0.00006 & 0 \\ 3 & 0.001 & 0.021 & 0.198 & 0.568 & 0.156 & 0.051 & 0.006 & 0.0001 \\ 4 & 0 & 0.002 & 0.077 & 0.248 & 0.397 & 0.210 & 0.064 & 0.002 \\ 5 & 0 & 0.001 & 0.031 & 0.104 & 0.309 & 0.374 & 0.161 & 0.019 \\ 6 & 0 & 0 & 0.002 & 0.077 & 0.190 & 0.334 & 0.334 & 0.062 \\ 7 & 0 & 0.0004 & 0.0004 & 0.008 & 0.075 & 0.194 & 0.354 & 0.369 \end{array} \tag{3.1}$$

$P_{\texttt{Score}}$ shows the transition probabilities from January to February. In general, customers are most likely to stay in the state they were in and very rarely move more than one state up or down, although there are exceptions. For instance, 12.6% of customers in state 0 move to state 2, and 19.4% of customers in state 7 move to state 5. Based on the estimated probabilities in figure 3.14 and (3.1), a likely portfolio for February 2019 was simulated. The same was done for the other months to produce a complete portfolio.

Additionally, new customers were generated assuming the same percentage increase in the number of customer as the equivalent month in 2018. The covariates of the generated customers were sampled from the portfolio and the covariates
`SUM_of_PaymentOverDueFlag` and `Score` for these generated customers were developed as in the same manner, using figure 3.14 and equation (3.1).

3.4.3 Forecasting for 2019

The explanatory variables were forecasted so that one could use the mixed effects logistic regression model to forecast the total balance sent to debt collection for the year 2019. Figure 3.15 shows the forecasts of the total balance sent to debt collection for each month for the year 2019. The apriori probability of belonging to class 1, i.e. being an impaired customer is reduced for July as shown in figure 3.11. This results in a lower forecast for July 2019. The forecasts are generally higher compared to 2018, predicting that the total balance sent to debt collection will exceed 70 million NOK for several months. The blue shaded area shows 80 percentile bootstrap prediction intervals based on bootstrap replicates of the parameters, $(\hat{\boldsymbol{\beta}}^{*(b)}, \hat{\sigma}_v^{*(b)})$ for $b = 1, \ldots, 1000$. The prediction interval does not consider the uncertainty associated with forecasting the explanatory variables. It should also be noted that the prediction interval does not become wider the further into the future the model forecasts as is the case for time series. This could possibly be the case if we had included all the uncertainty associated with the mixed effects logistic regression model.

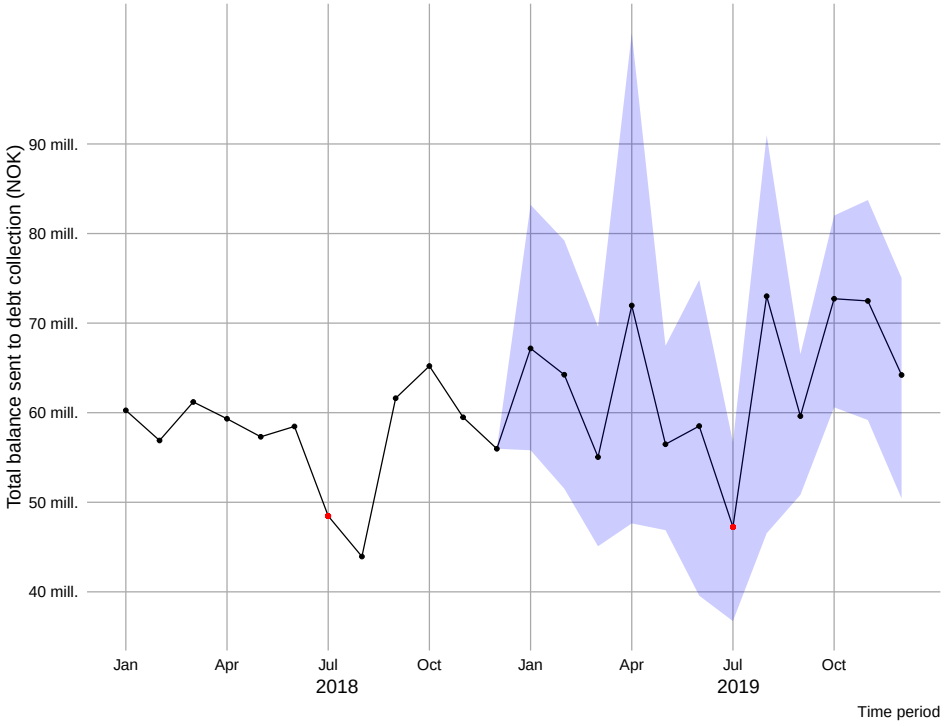


Figure 3.15: Forecasts of the total balance sent to debt collection each month in 2019 based on the mixed effects logistic regression model. The blue shaded area shows a 80 percentile bootstrap prediction interval based on 1000 bootstrap replicates. The forecasts for 2019 are generally higher compared to 2018.

Chapter 4

Summary

This chapter consists of two sections, a discussion and concluding remarks. The first section presents and discusses general issues with the methods and techniques used. Recommendations for possible further work are also presented. Finally, concluding remarks are presented in the last section.

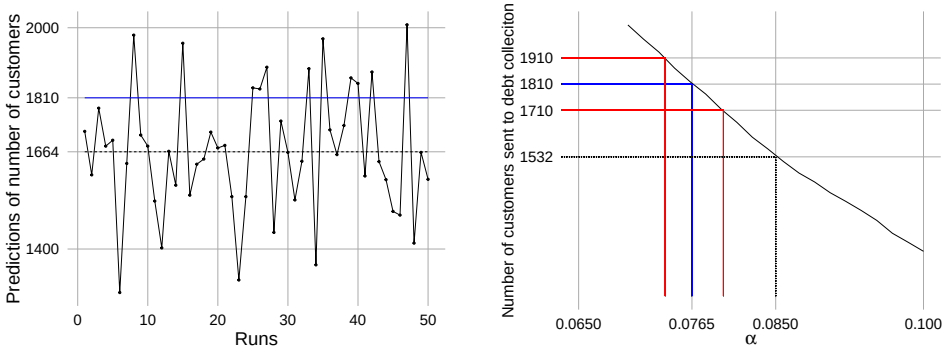
4.1 Discussion

The total balance sent to debt collection each month for the year 2019 was forecasted based on historical data from the time period July 2017 to September 2018 provided by SpareBank 1 Kredittkort AS. A mixed-effects logistic regression model that classified customers to either belonging to class 1 (impaired) or class 0 (non-impaired) was created. Due to a highly imbalanced data set where the number of instances from class 0 was far greater than class 1, the outputs of the classifier were adjusted by using the prior distributions of belonging to class 0 and 1. The prior probability of belonging to class 1 was forecasted using time series analysis. An ARIMA model, where the seasonal trends were modelled as additive outliers, was used to forecast these prior probabilities. Furthermore, the proper covariates chosen by the LASSO method for GLMMs were forecasted as well. Combining these techniques with the mixed-effects logistic regression model produced forecasts for the total balance sent to debt collection for each month in 2019.

4.1.1 Instability of the Mixed-effects Logistic Regression Model

One of the main issues with the mixed-effects logistic regression is instability as the model does not produce the same forecast every time the model predicts. This is probably a consequence of the training and test set that is sampled, which is slightly different every time (depending on which customers are drawn), which again makes the parameters and threshold value α slightly different each time the model is run. In order to combat this, we therefore ran the model 50 times, and used the average prediction based off those runs to account for the instability and investigate the variability in the forecasts. Figure 4.1

shows two plots that analyze the predictions for December 2018; (**a**) shows 50 different predictions for the number of customers sent to debt collection. The dashed line shows the average prediction and the blue line shows the actual number of impaired customers. The predictions vary greatly, showing the instability of the model. The average prediction is not very close to the actual value. This is the case for most months. The predictions for July 2018 is particularly too low, although they should be lower than for the other months. A possible explanation is that the prior probability of belonging to class 1 is much lower for July, possibly too low. The instability of the model is further illustrated in (**b**), which shows



(a) 50 predictions for January 2018.

(b) Number of customers versus α.

Figure 4.1: Analysis of predictions of the number of impaired customers for December 2018. (**a**) shows 50 predictions for the number of customers sent to debt collection, while (**b**) plots the number of impaired customers versus the threshold value α for one of the runs.

the number of customers sent to debt collection versus the threshold value α for December 2018. A total of 1810 customers were impaired this month. A threshold value $\alpha = 0.0765$ would have given an exact prediction. However, we have estimated that $\hat{\alpha} = 0.085$. This gives a prediction of 1532. The red lines show a ± 100 margin for the predictions. In other words, the threshold value has to be in the short interval $\alpha \in [0.0738, 0.797]$ for forecasts to miss by less than 100 customers. Hence, a small change in α will change the predictions significantly. Therefore, it is unlikely to believe that the model can consistently give predictions of the number of impaired customers with a ± 100 margin.

4.1.2 Limitations to the Mixed-effects Logistic Regression Model

The random intercept terms v_{0i} are estimated for all the customers in the training set, but will not be estimated for unobserved individuals in the training set. As a result, the probabilities for unobserved individuals will only be based on the fixed-effects coefficients $\boldsymbol{\beta}$, i.e. the similarities between customers on a population level, not on an individual level. However, one could possibly estimate the random intercept terms for unobserved individuals by comparing with similar customer with an estimated random intercept term. For instance, if customer A is comparable to the customers $(A_1, \ldots, A_n)$ in terms of covariates, it is not unlikely that customer A's random intercept term would be similar. We could therefore estimate customer A's random intercept based on customers $(A_1, \ldots, A_n)$ by for

instance taking the mean of their random intercepts.

Furthermore, the mixed-effects logistic regression model could have been specified differently. For instance, the mixed-effects was reduced to only a random intercept term, v_{0i}. However, we could have possibly included random slopes that would be unique for each customer. An example would be to include a random slope to the time term `YearMonth`, v_{1i}, such that the model

$$\begin{aligned}\log\left(\frac{P(Y_{it}=1)}{1-P(Y_{it}=1)}\right) &= (\beta_0+v_{0i})+\texttt{YearMonth}_{it}(\beta_1+v_{1i})\\ &\quad+\texttt{CustomerAgeSquared}_{it}\beta_2\\ &\quad+\texttt{MonthsSinceAccountCreatedSquared}_{it}\beta_3\\ &\quad+\texttt{Score}\beta_4+\texttt{SUM\_of\_PaymentOverDueFlag}\beta_5,\end{aligned}$$

would have a specified random slope for each customer. We have assumed that 3 months should be enough observations per cluster to determine if a customer was sent to debt collection. Alternatively, one could also have used more observations for each customer, for instance following a customer's behavior in the last 6 months or last year. Another suggestion might be to track customers weekly, thus dividing each month into four observations. This would give more observations per customer (12 observations) while at the same time investigating the most recent time period which is of most interest.

4.1.3 Forecasting for 2019

The explanatory variables included in the model were forecasted such that the mixed-effects logistic regression model could be used to forecast the total balance sent to debt collection in 2019. The explanatory variables `Score` and `SUM_of_PaymentOverDueFlag` were forecasted through Markov chain simulation. The explanatory variables `YearMonth`, `CustomerAge` and `MonthsSinceAccountCreated` were also forecasted. The model assumed a linear relationship between $\text{logit}(p_{it})$ and the square of `CustomerAge`, as well as the square of `MonthsSinceAccountCreated`. An issue that arises is that the portfolio as a whole will become "older" the further into future the portfolio is simulated. We tried to combat this issue by generating new customers as well as removing some customers at random accordingly. At the same time, the prior probabilities of belonging to class 1 were forecasted to increase (except for the month of July). This will generally increase the predictions further into the future. In other words, simulating the covariates $x_{i(T+h)k}, i = 1,\dots,N, k = 1,\dots,p$ will generally decrease the predictions as h increases, while the prior probability $\gamma_T + h$ will increase the predictions as h increases. Of course, it is very hard to determine how the portfolio will look like into the future. For instance, the credit card company may decide to run advertisement that will attract very many new customers one month or competing companies may offer credit cards with more favorable terms such that the credit card company will lose customers. It is therefore very hard to accurately predict how the portfolio will look like in the future.

4.1.4 Problems with Variable Selection for GLMMs

Variable selection for generalized linear mixed models in general is not as straightforward as for GLMs and options are more limited. We computed the BIC criterion for different values of the tuning parameter λ included in the LASSO method and chose the optimal λ on the level which the BIC was minimized. The explanatory variables at this λ value was then evaluated. The advantage of using such an information criterion is that the BIC has its own penalty term that aims to reduce the number of parameters in the model. Nevertheless, other methods for variable selection could have been used.

4.1.5 Adjusting the Outputs of a Classifier

The method proposed by (Saerens et al., 2002) of adjusting the outputs of a classifier was used. However, we did not take into consideration that the longitudinal data was correlated. One could possibly alter the method to account for this correlation.

4.1.6 Additional Noise in the Data Set

It should also be mentioned that there is always some noise the model cannot explain. For instance, the credit card company has had trouble in the past sending out debt collection invoices previously, so that these have been sent at a later time. The result is that the previous month that had very many debt collection cases will register few cases, but it appears that the next month will have very many debt collection cases. In addition, as shown in section 3.1 some customers were sent to debt collection even though they fell into segments such as `Transactor`. It will be hard for the model to identify these as potential impaired customers since this should technically not be possible.

4.1.7 The Debt Register

The Norwegian government have decided to make a *gjeldsregister* (directly translated as debt register) with the goal of gathering all the information about consumers unsecured debt in one place (Finansdepartementet, 2019). Unsecured debt refers to all debt that is not backed by an underlying asset, such as credit card debt. Mortgages, for instance, is secured debt as it is backed by real estate. Banks and other financial institutions will thus have a more complete picture of what a customer owes. The idea is that the debt register will include everything from student loans to credit card debt. The purpose of this register is to prevent debt problems in private households. Ideally, it will be easier for credit card issuers to determine which customers that should be granted a credit card. For instance, a problem that credit card issuers face today is that some customers are dishonest about their own financial situation, and actively does not inform about all their debt in order to receive a credit card. One might think that, in the long run, the number of impaired customers may decrease if customers who should not have been given a credit card in the first place will no longer receive one. If this is the case, the model may no longer be applicable or should be altered. The debt register is scheduled to be operative from the 1st of July 2019.

4.1.8 Additional Usage of the Model

The mixed-effects logistic regression model predicts the probability that customer i is sent to debt collection at time t for $i = 1, \ldots, N$ and $t = 1, 2, \ldots$. Therefore, the model could also be used as a measure to determine whether a customer should be given for instance a credit limit increase on their credit card, or the opportunity to refinance their debt. A customer that has been financially responsible will receive a low probability according to our model and it is therefore a smaller risk to give this customer a credit limit increase on his or her card. On the other hand, a customer that our model has assigned a high probability will likely be a greater risk for the bank to give a credit limit increase on their credit card. This usage of the model can be advantageous for the credit card company.

4.1.9 Recommendations for Further Work

The mixed-effects logistic regression model could be further investigated by implementing some of the ideas suggested in section 4.1.2. Furthermore, the model may be greatly improved by incorporating information such as a customer's monthly income, marital status and overall unsecured debt. This data would then have to be gathered for both new and existing customers. Other models could be considered as well. For instance, a Mixed-Effects Random Forest (MERF) (Hajjem et al., 2014) model could be an interesting option, although random forest methods tend to be biased towards the majority class in an imbalanced data set.

As shown in section 3.1, 9.1% of impaired customers fall into the segment `Active in last 6 months`, which means that they are sent to debt collection fairly quickly after receiving their credit card. This represents customers that possibly should not have been issued a credit card in the first place. The remaining impaired customers are some form of revolver, i.e. a customer who does not pay the total amount he or she owns at the end of a billing cycle. It could also be of interest to make separate models for these two groups and specifically investigate why some customers end up being sent to debt collection after just a few months.

4.2 Concluding Remarks

The aim of this thesis was to forecast the total balance sent to debt collection each month for the year 2019 based on historical data provided by SpareBank 1 Kredittkort AS in the period July 2017 to September 2018. The data was longitudinal with repeated measurements each month for more than 500 000 credit card customers in Norway. Most customers had 15 recorded observations, one for each month in the time period July 2017 to September 2018 with a binary response telling whether the customer was sent to debt collection in three months time. A mixed-effects logistic regression model was made to classify and count the number of customers sent to debt collection. The data set was highly imbalanced as only a few customers are impaired each month. We used random undersampling and adjusted the outputs of the classifier to account for the imbalance in the data set using a method proposed by Saerens et al., 2002. The prior probability of being impaired had to be forecasted in order to forecast with the mixed-effects logistic regression model. This was done by creating an ARIMA$(1, 1, 1)$ time series. The time series showed seasonal trends for the month of July, but due to the short length of the time series, the seasonality was rather modelled as known additive outliers. This captured the seasonality of the time series without reducing the number of observations. Furthermore, the explanatory variables were included based on the LASSO method extended to apply on generalized linear mixed models as well. The explanatory variables were also forecasted. Those that were categorical was modelled as a Markov chain. The time series was fitted to the year 2018 and shows that the model is unstable, as it does not give the same predictions every time. The reason for this and how to combat instability is discussed. The forecasts for 2019 are generally increasing, with a drop in July 2019. Finally, it should also be noted that there are other factors that may make the model obsolete. The debt register planned by the Norwegian government will give credit card companies valuable information about potential customers and may alter who will be issued a credit card. This may have a great impact on the mixed effects logistic regression model.

Bibliography

Bates, D., Mchler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software, Articles 67 (1), 1–48.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., White, J.-S. S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology Evolution 24 (3), 127135.

Box, G. E. B., Jenkins, G. M. J., Reinsel, G. C. R., 1994. Time Series Analysis: Forecasting and Control, 3rd edition. Prentice-Hall, Inc.

Breslow, N. E., Clayton, D. G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88 (421), 9.

Brockwell, P. J., Davis, R. A., 2002. Introduction to time series and forecasting. Springer.

Chen, C., Liu, L.-M., 1993. Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association 88 (421), 284–297.

Croson, R., Gneezy, U., 2009. Gender differences in preferences. Journal of Economic Literature 47 (2), 448474.

Davis, P. J., Rabinowitz, P., 1975. Methods of Numerical Integration Philip J. Davis and Philip Rabinowitz. Academic Press.

Dunn, P. K., Smyth, G. K., 1996. Randomized quantile residuals. Journal of Computational and Graphical Statistics 5 (3), 236.

Evans, D. S., Schmalensee, R., 2005. Paying with Plastic: the Digital Revolution in Buying and Borrowing (2nd Edition). MIT Press.

Fan, Y., Leslie, D., Wand, M., 2008. Generalised linear mixed model analysis via sequential monte carlo sampling. Electronic Journal of Statistics 2, 916938.

Finansdepartementet, Feb 2019. Forskrift om krav til finansforetakenes utlånspraksis for forbrukslån.
URL `http://bit.do/eT6vJ`

Gad, A. M., Kholy, R. B. E., 2012. Generalized linear mixed models for longitudinal data. International Journal of Probability and Statistics 1 (3), 4147.

Greene, C., Schuh, S. D., Stavins, J., 2015. The 2015 survey of consumer payment choice: Summary results. SSRN Electronic Journal.

Groll, A., Tutz, G., 2012. Variable selection for generalized linear mixed models by l 1-penalized estimation. Statistics and Computing 24 (2), 137154.

Hajjem, A., Bellavance, F., Larocque, D., 2014. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation 84 (6), 13131328.

Hartig, F., 2019. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.2.4.
URL `https://CRAN.R-project.org/package=DHARMa`

Hedeker, D. R., Gibbons, R. D., 2006. Longitudinal data analysis. Wiley-Interscience.

Holck, J. A. T., 2018. Forecasting the total balance sent to debt collection. Unpublished, Project thesis at NTNU.

Hurvich, C. M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.

Liu, Q., Pierce, D. A., 1994. A note on gauss-hermite quadrature. Biometrika 81 (3), 624–629.

Pinheiro, J. C., Chao, E. C., 2006. Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics 15 (1), 5881.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Saerens, M., Latinne, P., Decaestecker, C., 2002. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. Neural Computation 14 (1), 2141.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461464.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58 (1), 267–288.

Wolfinger, R., Oconnell, M., 1993. Generalized linear mixed models a pseudo-likelihood approach. Journal of Statistical Computation and Simulation 48 (3-4), 233243.

Appendix A

Variables in the Data Set

Table A.1 provides a description of all the variables in the data set provided by SpareBank 1 Kredittkort AS. The first column is the name of the variable as it occurs in the data set and the second column is a description of the variable.

Table A.1: Description of all the variables in the data set.

Variables	Description
BK_ACCOUNT_ID	Internal account number
PeriodId	Date on format YYYYMMDD
Date	Date on format YYYY-MM-DD
YearMonth	Year and month on format YYYYMM
PNRSerial	Digits 7 and 8 in national identification number
CustomerAge	Customer's age in years
MonthsSinceAccountCreated	Account's age in months
PRODUCT_NAME	Name of product (card type)
STATEMENT_DUE_DAY_OF_MONTH_NUM	Chosen due date (5.,10.,15. or 20.)
ApplicationSalesChannel	Channel of application and / or sale
CAMPAIGN_NAME	Campaign (if any)
CLOSING_BALANCE_AMT	Total amount printed on last statement
DISTRIBUTOR_NAME	Bank Name
GENDER_NAME	Gender
HAS_DIRECT_DEBIT_AGREEMENT_IND	Indicator, direct debit agreement selected (*avtalegiro* in Norwegian)
HAS_ESTATEMENT_AGREEMENT_IND	Indicator, e-statement selected (*e-faktura* in Norwegian)
average_credit_limit_last12	Average credit limit last 12 months
average_revolvingbalance_last12	Average revolving balance last 12 months
avg_rev_bal_L3M	Average revolving balance last 3 months
rev_uti_currmth	Revoling balance divided by credit limit this month
avg_payment_L3M	Average payment last 3 months
rev_per_uti_change_L3M	Change in revolving utilization (revolving balance divided by credit limit) last 3 months
MonthEnd_uti_Change	Change in revolving utilisation by end of month
payment_amt_change_L3M	Change in payment amount last 3 months
RevUti12	Average revolving balance last 12 months divided by average credit limit last 12 months
AvgRevBalL3onL12	(Average revolving balance last 3 months divided by average credit limit last 3 months) divided by (Average revolving balance last 12 months divided by average credit limit last 12 months)
QCashpartL12	Part of sum of transactions in class Quasi Cash last 12 months

Continued on next page

Table A.1 – continued from previous page

Variables	Description
QCashpartL3	Part of sum of transactions in class Quasi Cash last 3 months
QCashL3onL12	(Part of sum of transactions in class Quasi Cash last 3 months) divided by (Part of sum of transactions in class Quasi Cash last 12 months)
TravelpartL12	Sum of transactions in classes, Airline, Hotel_motel and other_transport last 12 months divided by sum of transactions in all classed last 12 months
TravelpartL3	Sum of transactions in classes, Airline, Hotel_motel and other_transport last 3 months divided by sum of transactions in all classed last 3 months
TravelpartL3onL12	(Sum of transactions in classes, Airline, Hotel_motel and other_transport last 3 months divided by sum of transactions in all classed last 3 months) divided by (Sum of transactions in classes, Airline, Hotel_motel and other_transport last 3 months divided by sum of transactions in all classed last 3 months)
Segment9Name	Segment name with 9 segments
Segment23Name	Segment name with 23 segments
Score	Simple risk score between 0 and 7
SUM_of_CreditLimitIncreaseFlag	Number of credit limit increases last 12 months
SUM_of_CreditLimitDecreaseFlag	Number of credit limit decreases las 12 months
SUM_of_PaymentOverDueFlag	Number of months with payment overdue last 12 months
SUM_of_FirstDunningFlag	Number of months with dunning (*purring* in Norwegian) last 12 months
SUM_of_CollectionAdviceFlag	Number of months with collection advice (*inkassovarsel* in Norwegian) last 12 months
SUM_of_CollectionFlag	Number of months with debt collection (*inkasso* in Norwegian) last 12 months
SUM_of_CardFraudFlag	Number of months with card fraud flag (transactions marked as possible fraud) last 12 months
SUM_of_CardLostFlag	Number of months with card lost flag (card marked as lost)(last 12 months
SUM_of_CardStolenFlag	Number of months with card stolen flag (card marked as lost)(last 12 months
SUM_of_AIRLINEL12	Sum of transactions in given class last 12 months
SUM_of_CLOTHING_STORESL12	Sum of transactions in given class last 12 months
SUM_of_FOOD_STORES_WAREHOUSEL12	Sum of transactions in given class last 12 months
SUM_of_HOTEL_MOTELL12	Sum of transactions in given class last 12 months
SUM_of_HARDWAREL12	Sum of transactions in given class last 12 months
SUM_of_INTERIOR_FURNISHINGSL12	Sum of transactions in given class last 12 months
SUM_of_OTHER_RETAILL12	Sum of transactions in given class last 12 months
SUM_of_OTHER_SERVICESL12	Sum of transactions in given class last 12 months
SUM_of_OTHER_TRANSPORTL12	Sum of transactions in given class last 12 months
SUM_of_RECREATIONL12	Sum of transactions in given class last 12 months
SUM_of_RESTAURANTS_BARSL12	Sum of transactions in given class last 12 months
SUM_of_SPORTING_TOY_STORESL12	Sum of transactions in given class last 12 months
SUM_of_TRAVEL_AGENCIESL12	Sum of transactions in given class last 12 months
SUM_of_VEHICLESL12	Sum of transactions in given class last 12 months
SUM_of_QUASI_CASHL12	Sum of transactions in given class last 12 months
SUM_of_AIRLINEL3	Sum of transactions in given class last 12 months
SUM_of_CLOTHING_STORESL3	Sum of transactions in given class last 3 months
SUM_of_FOOD_STORES_WAREHOUSEL3	Sum of transactions in given class last 3 months
SUM_of_HOTEL_MOTELL3	Sum of transactions in given class last 3 months
SUM_of_HARDWAREL3	Sum of transactions in given class last 3 months
SUM_of_INTERIOR_FURNISHINGSL3	Sum of transactions in given class last 3 months
SUM_of_OTHER_RETAILL3	Sum of transactions in given class last 3 months
SUM_of_OTHER_SERVICESL3	Sum of transactions in given class last 3 months
SUM_of_OTHER_TRANSPORTL3	Sum of transactions in given class last 3 months
SUM_of_RECREATIONL3	Sum of transactions in given class last 3 months
SUM_of_RESTAURANTS_BARSL3	Sum of transactions in given class last 3 months
SUM_of_SPORTING_TOY_STORESL3	Sum of transactions in given class last 3 months
SUM_of_TRAVEL_AGENCIESL3	Sum of transactions in given class last 3 months
SUM_of_VEHICLESL3	Sum of transactions in given class last 3 months
SUM_of_QUASI_CASHL3	Sum of transactions in given class last 3 months

Continued on next page

Table A.1 – continued from previous page

Variables	Description
lead1YearMonth	YearMonth+1
lead2YearMonth	YearMonth+2
lead3YearMonth	YearMonth+3
DCA0Ind	Indicator: lead1YearMoth<= DCA0YearMonth <= lead3YearMonth
BalanceSent	Balance sent to DCA (Debt Collection Agency) (in DCA0Month)

Appendix B

Outputs and Results

Outputs of the mixed-effects logistic regression model each month of 2018. The number of quadrature points is `nAGQ`. The `data` is the training set. `Random effects` shows σ_v, the `Number of obs:` in the training set and the number of `groups`. The `Fixed effects` and the `Correlation of Fixed Effects` are also shown.

Mixed-effects Logistic Regression Model for January

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
 10622.2  10682.8  -5304.1  10608.2    42433

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5160 -0.0685 -0.0270 -0.0101  5.4291

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 6.346    2.519
Number of obs: 42440, groups:  BK_ACCOUNT_ID, 14181

Fixed effects:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -9.44355    0.23044 -40.981  < 2e-16 ***
YearMonth                        4.18482    0.13195  31.716  < 2e-16 ***
CustomerAgeSquared              -2.22501    0.43926  -5.065 4.08e-07 ***
MonthsSinceAccountCreatedSquared -2.67862    0.30590  -8.757  < 2e-16 ***
Score                            1.05611    0.26279   4.019 5.85e-05 ***
SUM_of_PaymentOverDueFlag        1.00296    0.03757  26.695  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

```
Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.628
CstmrAgSqrd -0.272 -0.014
MnthsSncACS  0.002 -0.042 -0.267
Score       -0.635  0.142 -0.032 -0.017
SUM_f_PyODF -0.258  0.148  0.068 -0.168 -0.304
```

Mixed-effects Logistic Regression Model for February

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
  9729.0   9788.9  -4857.5   9715.0    38513

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5172 -0.0640 -0.0252 -0.0095  5.3113

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 7.172    2.678
Number of obs: 38520, groups:  BK_ACCOUNT_ID, 12879

Fixed effects:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -9.38233    0.24948 -37.608  < 2e-16 ***
YearMonth                        4.20368    0.14131  29.748  < 2e-16 ***
CustomerAgeSquared              -2.81672    0.48553  -5.801 6.58e-09 ***
MonthsSinceAccountCreatedSquared -2.47446    0.33304  -7.430 1.09e-13 ***
Score                            0.96567    0.28752   3.359 0.000783 ***
SUM_of_PaymentOverDueFlag        0.97455    0.04026  24.208  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.626
CstmrAgSqrd -0.291 -0.012
MnthsSncACS  0.006 -0.029 -0.277
Score       -0.651  0.191 -0.035 -0.038
SUM_f_PyODF -0.213  0.061  0.075 -0.144 -0.305
```

Mixed-effects Logistic Regression Model for March

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
  9767.3   9827.5  -4876.7   9753.3    39923

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.3589 -0.0731 -0.0290 -0.0112  5.3374

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 5.599    2.366
Number of obs: 39930, groups:  BK_ACCOUNT_ID, 13334

Fixed effects:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -9.50657    0.23858 -39.847  < 2e-16 ***
YearMonth                        4.22629    0.14253  29.653  < 2e-16 ***
CustomerAgeSquared              -2.51123    0.43469  -5.777 7.60e-09 ***
MonthsSinceAccountCreatedSquared -3.12670    0.41998  -7.445 9.70e-14 ***
Score                            1.54914    0.26406   5.867 4.45e-09 ***
SUM_of_PaymentOverDueFlag        0.88401    0.03548  24.917  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.642
CstmrAgSqrd -0.256 -0.016
MnthsSncACS  0.009 -0.029 -0.295
Score       -0.659  0.166 -0.035 -0.032
SUM_f_PyODF -0.212  0.056  0.064 -0.153 -0.274
```

Mixed-effects Logistic Regression Model for April

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
 10661.9  10722.5  -5324.0  10647.9    42413

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5067 -0.0671 -0.0264 -0.0101  5.1446

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 6.529    2.555
Number of obs: 42420, groups:  BK_ACCOUNT_ID, 14170

Fixed effects:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -9.26502    0.23199 -39.937  < 2e-16 ***
YearMonth                         4.05177    0.12950  31.287  < 2e-16 ***
CustomerAgeSquared               -2.73496    0.46560  -5.874 4.25e-09 ***
MonthsSinceAccountCreatedSquared -2.82571    0.31395  -9.000  < 2e-16 ***
Score                             0.97072    0.26280   3.694 0.000221 ***
SUM_of_PaymentOverDueFlag         0.99503    0.03774  26.362  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.627
CstmrAgSqrd -0.256 -0.029
MnthsSncACS  0.002 -0.040 -0.258
Score       -0.648  0.191 -0.051 -0.024
SUM_f_PyODF -0.261  0.113  0.051 -0.144 -0.292
```

Mixed-effects Logistic Regression Model for May

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
  9893.9   9954.1  -4940.0   9879.9    39973

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.2821 -0.0737 -0.0297 -0.0115  5.0966

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 5.573    2.361
Number of obs: 39980, groups:  BK_ACCOUNT_ID, 13360

Fixed effects:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -9.28858    0.23755 -39.101  < 2e-16 ***
YearMonth                         4.15367    0.13810  30.077  < 2e-16 ***
CustomerAgeSquared               -1.85315    0.39211  -4.726 2.29e-06 ***
MonthsSinceAccountCreatedSquared -2.50946    0.30438  -8.245  < 2e-16 ***
Score                             1.05123    0.26505   3.966 7.31e-05 ***
SUM_of_PaymentOverDueFlag         0.92987    0.03501  26.558  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.663
CstmrAgSqrd -0.252 -0.019
MnthsSncACS  0.002 -0.031 -0.288
Score       -0.673  0.229 -0.034 -0.018
SUM_f_PyODF -0.210  0.059  0.066 -0.151 -0.289
```

Mixed-effects Logistic Regression Model for June

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
  9960.4  10020.6  -4973.2   9946.4    40223

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4922 -0.0748 -0.0300 -0.0115  4.9467

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 5.644    2.376
Number of obs: 40230, groups:  BK_ACCOUNT_ID, 13451

Fixed effects:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -9.13261    0.23180 -39.399  < 2e-16 ***
YearMonth                         4.22023    0.14270  29.574  < 2e-16 ***
CustomerAgeSquared               -2.32161    0.41496  -5.595 2.21e-08 ***
MonthsSinceAccountCreatedSquared -2.44922    0.30031  -8.156 3.47e-16 ***
Score                             0.90953    0.26017   3.496 0.000472 ***
SUM_of_PaymentOverDueFlag         0.87604    0.03464  25.291  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.649
CstmrAgSqrd -0.270 -0.013
MnthsSncACS  0.006 -0.029 -0.276
Score       -0.649  0.156 -0.001 -0.030
SUM_f_PyODF -0.174  0.048  0.029 -0.153 -0.314
```

Mixed-effects Logistic Regression Model for July

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
  8296.3   8355.0  -4141.1   8282.3    32723

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.2932 -0.0760 -0.0307 -0.0117  5.1420

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 5.742    2.396
Number of obs: 32730, groups:  BK_ACCOUNT_ID, 10951

Fixed effects:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -8.9030     0.2269 -39.242  < 2e-16 ***
YearMonth                         4.1363     0.1499  27.586  < 2e-16 ***
CustomerAgeSquared               -2.5159     0.4752  -5.295 1.19e-07 ***
MonthsSinceAccountCreatedSquared -2.4388     0.3283  -7.429 1.09e-13 ***
Score                             0.8389     0.2456   3.415 0.000637 ***
SUM_of_PaymentOverDueFlag         0.8875     0.0389  22.815  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.681
CstmrAgSqrd -0.293 -0.020
MnthsSncACS -0.001 -0.036 -0.275
Score       -0.521  0.132 -0.029 -0.021
SUM_f_PyODF -0.293  0.128  0.037 -0.156 -0.326
```

Mixed-effects Logistic Regression Model for August

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
  8411.8   8470.5  -4198.9   8397.8    32393

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.3343 -0.0687 -0.0285 -0.0114  4.1737

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 6.591    2.567
Number of obs: 32400, groups:  BK_ACCOUNT_ID, 10829

Fixed effects:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -8.86010    0.25439 -34.828  < 2e-16 ***
YearMonth                         3.83845    0.14001  27.415  < 2e-16 ***
CustomerAgeSquared               -2.50459    0.50226  -4.987 6.14e-07 ***
MonthsSinceAccountCreatedSquared -2.41539    0.35847  -6.738 1.61e-11 ***
Score                             0.41066    0.29249   1.404     0.16
SUM_of_PaymentOverDueFlag         1.00696    0.04261  23.634  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.620
CstmrAgSqrd -0.287 -0.004
MnthsSncACS -0.009 -0.041 -0.287
Score       -0.632  0.167 -0.020  0.002
SUM_f_PyODF -0.264  0.145  0.046 -0.165 -0.316
```

Mixed-effects Logistic Regression Model for September

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
 11074.7  11135.6  -5530.4  11060.7    44013

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5769 -0.0550 -0.0213 -0.0080  4.9294

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 8.081    2.843
Number of obs: 44020, groups:  BK_ACCOUNT_ID, 14703

Fixed effects:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -9.62879    0.21700 -44.372  < 2e-16 ***
YearMonth                         4.05247    0.12738  31.814  < 2e-16 ***
CustomerAgeSquared               -3.12105    0.52359  -5.961 2.51e-09 ***
MonthsSinceAccountCreatedSquared -3.00867    0.35006  -8.595  < 2e-16 ***
Score                             1.13747    0.23867   4.766 1.88e-06 ***
SUM_of_PaymentOverDueFlag         1.02866    0.03921  26.237  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.633
CstmrAgSqrd -0.314 -0.012
MnthsSncACS  0.028 -0.043 -0.288
Score       -0.565  0.233 -0.027 -0.036
SUM_f_PyODF -0.308  0.042  0.031 -0.167 -0.298
```

Mixed-effects Logistic Regression Model for October

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
 11610.0  11671.2  -5798.0  11596.0    46223

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.6769 -0.0716 -0.0288 -0.0112  5.4661

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 5.926    2.434
Number of obs: 46230, groups:  BK_ACCOUNT_ID, 15436

Fixed effects:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -9.24182    0.21955 -42.093  < 2e-16 ***
YearMonth                         3.95081    0.12115  32.610  < 2e-16 ***
CustomerAgeSquared               -3.17146    0.43537  -7.285 3.23e-13 ***
MonthsSinceAccountCreatedSquared -2.99519    0.30288  -9.889  < 2e-16 ***
Score                             1.67835    0.24888   6.744 1.55e-11 ***
SUM_of_PaymentOverDueFlag         0.92427    0.03374  27.394  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.654
CstmrAgSqrd -0.234 -0.025
MnthsSncACS  0.018 -0.050 -0.227
Score       -0.692  0.254 -0.043 -0.034
SUM_f_PyODF -0.236  0.100 -0.004 -0.172 -0.255
```

Mixed-effects Logistic Regression Model for November

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
 10099.2  10159.4  -5042.6  10085.2    40103

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7441 -0.0650 -0.0258 -0.0101  4.9176

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 6.833    2.614
Number of obs: 40110, groups:  BK_ACCOUNT_ID, 13403

Fixed effects:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -9.59971    0.24128 -39.787  < 2e-16 ***
YearMonth                         4.05532    0.13372  30.327  < 2e-16 ***
CustomerAgeSquared               -2.84551    0.48130  -5.912 3.38e-09 ***
MonthsSinceAccountCreatedSquared -2.75558    0.33992  -8.107 5.20e-16 ***
Score                             1.70065    0.27587   6.165 7.06e-10 ***
SUM_of_PaymentOverDueFlag         0.91013    0.03644  24.973  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.626
CstmrAgSqrd -0.253 -0.020
MnthsSncACS -0.003 -0.031 -0.257
Score       -0.686  0.220 -0.040 -0.028
SUM_f_PyODF -0.196  0.041  0.019 -0.156 -0.275
```

Mixed-effects Logistic Regression Model for December

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature,
    nAGQ = 10) ['glmerMod']
 Family: binomial  ( logit )
Formula: DCA0Ind ~ YearMonth + CustomerAgeSquared + MonthsSinceAccountCreatedSquared +
    Score + SUM_of_PaymentOverDueFlag + (1 | BK_ACCOUNT_ID)
   Data: data
Control: glmerControl(optimizer = "optimx", calc.derivs = FALSE,
    optCtrl = list(method = "nlminb",      starttests = FALSE, kkt = FALSE))

     AIC      BIC   logLik deviance df.resid
 10203.6  10263.8  -5094.8  10189.6    39763

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4256 -0.0676 -0.0271 -0.0107  4.4169

Random effects:
 Groups        Name        Variance Std.Dev.
 BK_ACCOUNT_ID (Intercept) 6.819    2.611
Number of obs: 39770, groups:  BK_ACCOUNT_ID, 13288

Fixed effects:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -9.23768    0.23433 -39.421  < 2e-16 ***
YearMonth                         3.92088    0.13137  29.846  < 2e-16 ***
CustomerAgeSquared               -2.98854    0.48975  -6.102 1.05e-09 ***
MonthsSinceAccountCreatedSquared -2.52470    0.33653  -7.502 6.28e-14 ***
Score                             1.27180    0.26993   4.712 2.46e-06 ***
SUM_of_PaymentOverDueFlag         0.90088    0.03759  23.969  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) YrMnth CstmAS MnSACS Score
YearMonth   -0.612
CstmrAgSqrd -0.250 -0.020
MnthsSncACS -0.005 -0.023 -0.299
Score       -0.661  0.175 -0.042 -0.007
SUM_f_PyODF -0.219  0.081  0.008 -0.145 -0.296
```