UNIVERSITY OF TWENTE

MASTER THESIS

Forecasting Mortgage Prepayment

Author:
Tim van der Star

Supervisor:
Dr. Berend ROORDA
Dr. Ir. Wouter van HEESWIJK

External supervisor:
Tom van der HEIJDEN

Master Industrial Engineering and Management
Specialization Financial Engineering and Management

UNIVERSITY OF TWENTE

Management Summary

Specialization Financial Engineering and Management
Master Industrial Engineering and Management

Master of Science

Forecasting Mortgage Prepayment

by Tim van der Star

Embedded into the Dutch mortgage contract is the option for mortgagors to prepay part of their residential mortgage outside their scheduled contractual payments. There are three types of prepayments that mortgage providers have to consider, namely partial, full and arbitrage prepayments. Mortgage prepayments alter the expected cash flows from the mortgage and due to their stochastic nature it is hard for mortgagees to make a correct or accurate valuation for their mortgage portfolio on an aggregate level. We study how different models, determinants and undersampling techniques help predict the observed prepayment rate for the Dutch portfolio book of Allianz. The current prepayment model at Allianz forecasts a mortgage prepayment cash flow over the period of 2014-2021 with a total error of -19.3% compared to the actual cash flows seen in the Allianz portfolio. Furthermore, the average yearly forecasted cash flow error by the Allianz model is 22.1%, which Allianz would like to reduce. The aim of this thesis is therefore to compute a model which is able to do so.

We perform a literature research to identify prepayment determinants and use these determinants to perform a preliminary data analysis on the Allianz mortgage data. We investigate the relevancy of these determinants on the prepayment rate for all models. After this we construct three models, namely a logistic regression model, random forest model and a neural network model, and investigate their ability to forecast each type of prepayment separately. Furthermore, we explore the ability of each model in forecasting the average and total prepayment cash flow error.

We undersample the training set in order to decrease the data imbalance towards the non-prepayment class. Through undersampling the training data set we alter the data set size multiple times and use these various data sets as training sets for the model. We create data sets where the prepayment observations (the minority class) are present for 10%, 20%, 30%, 40% and 50% of the training data set and explore the effect of undersampling and their ability to help improve the predictive power of each model. By evaluating each model on multiple portfolio error and loan level metrics we can deduce which model is able to best replicate the observed conditional prepayment rate (CPR) for each of the three separate prepayment types, namely partial, full and arbitrage, present in the Allianz portfolio.

Concerning the models that give insight into relevant prepayment variables we find that the logistic regression model is the only model to give interpretable and clear relationships between the modelled variables and each of the three prepayment rates. Furthermore, we find that all models are very imprecise at predicting prepayments on an individual loan-level and thus we reduce the relevance of these results.

We find that the random forest model trained on the undersampled training set where the minority class represents 30% of all observations on the basis of weighted root mean square error (WRMSE) produces the lowest partial prepayment CPR compared to the observed partial prepayment CPR. This model had a WRMSE of 0.205%. Concerning full prepayments we find that the random forest model trained on the data set where the minority class (in this case full prepayments) represents 50% of the training data has the lowest WRMSE error compared to the observed full prepayment CPR, with it being 0.581%. Regarding the arbitrage prepayment CPR (which takes place when refinancing) we find that the random forest model without undersampling performs best with a WRMSE of 0.392%. Reviewing the cash flow estimation we find that the neural network model trained on the original imbalanced training set, without undersampling, is able to achieve a yearly cash flow error of 17.5% which is the lowest error of all models and 4.5% lower than the benchmark Allianz model of 22.1%. We therefore achieve the goal of this thesis in computing a model which has a lower error than the benchmark Allianz model.

Contents

Chapter 1

Introduction

Starting 2021 the total outstanding household mortgage debt of the Netherlands equalled €748 billion (*StatLine - Kerngegevens sectoren; nationale rekeningen*). When comparing European countries by mortgage debt, €748 billion lands the Netherlands a 4th place spot, allowing only countries such as the UK, Germany and France ahead (*Total residential mortgage lending by country Europe 2020 | Statista*). The Netherlands has a much higher mortgage debt than most other countries in Europe, despite its relatively small population. This is the result of a political initiative to increase the owner-occupancy rate in the Netherlands. The aim of the initiative was to make the mortgage interest rate tax-deductible. With the Netherlands achieving the second highest share of the population of owner occupancy with a mortgage in Europe the initiative has shown its value (*Europe: owner occupiers with and without mortgage | Statista*).

The purchase of a house is an expense that not many individuals can afford without extra financing. In order to finance the purchase of a house, an individual must find an entity that is willing to lend the appropriate amount of money. Such a financing deal is referred to as a mortgage. The two main parties involved in the financing deal are the mortgagor, a person who borrows funds in order to purchase a house, and the mortgagee, the entity that lends the funds. There are many different ways in which the mortgagor can pay back the mortgagee. This is often done on a monthly basis, in which the mortgagor pays a part of the principal plus added interest. Although the details of a mortgage are contract specific, there are ways in which the mortgagor can deviate from the contract, bringing with it different types of risk for the mortgagee. The biggest risks the mortgagee faces, concerning the behaviour of the mortgagor, are default risk and prepayment risk. Default risk implies that the mortgagor is not able to meet its contractual monthly obligations stated in the contract, meaning part of the mortgage is not repaid. Prepayment risk is the risk of the mortgagor prepaying part of the mortgage (or the whole mortgage) earlier than anticipated, hence, reducing the life of the mortgage. The implication of this is that the mortgagee must reinvest the principal in the market, more often for a lower interest rate. This creates a mismatch for the mortgagee between the expected future cash flows of its mortgages and actual cash flows. Additionally, prepayment risk brings with it liquidity risk and interest rate risk. Liquidity risk arises due to the mismatch in cash flows between the assets and liabilities of the firm whilst interest rate risk is a consequence due to the mortgagors' possibility of refinancing. Hence, the mortgagee misses out on future interest payments and usually receives lower interest payments. This again creates a mismatch between the interest payments received and payments projected. It is important for the mortgage lender to model default rates and prepayment rates accurately in order to predict the future cash flows correctly. In turn this will reduce the asset and liability mismatch as the entity can alter its liability strategy according to the previously predicted future cash flows. In turn

this improves the risk management side of the entity. It must be noted that prepayment risk arises when interest rates fall, due to more prepayments being made and when interest rates increase as the number of prepayments decrease.
One such mortgage lender that funds mortgages is Allianz. Allianz entered the Dutch mortgage market in 2011 and slowly increased their mortgage portfolio in he Netherlands, acquiring their first €1 billion in mortgages in 2017. From 2017 onwards, however, Allianz has greatly increased its mortgage portfolio and as of July 2021 Allianz has an outstanding mortgage portfolio of €6 billion.
Allianz, like many other financial entities, uses a logistic regression model to calculate and predict the prepayment rate. However, the model only uses a small number of variables and is not able to predict the prepayment rate accurately enough. The main reason for this is that prepayments rarely occur, making input data for the logistic regression model highly imbalanced and therefore challenging to model. A solution must be found in order to reduce or correct for the data imbalance. Furthermore, with the increase in data collection over the past decade, machine learning algorithms are nowadays seen as viable replacements for the regression models. We aim to find out if there are machine learning models that, combined with new and addition data variables, can improve on the benchmark logistic regression model currently being used by Allianz to predict the mortgage prepayment rate.

1.1 Problem Statement

The current model at Allianz is not able to give accurate results concerning the forecasting of the conditional prepayment rate (CPR) and the accompanying prepayment cash flows. The Allianz model currently has an average yearly error of 22% compared to the actual cash flows of the Allianz portfolio, which they would like to decrease. It is in the interest of Allianz to improve this model in order to reduce the error between the observed and model forecasted prepayment cash flows so that the prepayment liquidity risk can be reduced. We do research into the alternative models that can replace the benchmark model at Allianz. Furthermore, we investigate which variables effect each type of prepayment and lastly we look at the impact of correcting for the data imbalance on the modelled outcomes.

1.2 Research Question

In order to achieve the goal of this thesis we formulate the following main and subquestions were. The research question of this thesis states: *"Are there any machine learning models that, when applied to Allianz mortgage data of the Netherlands, can forecast the prepayment rate more accurately than the benchmark model at Allianz over the time horizon of 2014-2021?"*
We set up several sub-questions that will help us to answer the main research question. By answering these questions we get a better understanding of the broader context. We propose the following sub-questions:

1. What are the main prepayment drivers stated in literature?
2. What are the main drivers of prepayment when analysing the Allianz mortgage portfolio?
3. Which machine learning models give insight to (new) variables that are relevant and important for prepayment modelling?

4. What is the effect of correcting for the data imbalance problem on the various model results?
5. Which machine learning models can be used to estimate the future prepayment rates?

1.3 Outline

In Chapter 2 we perform a literature study on the main drivers of mortgage prepayment. Furthermore, in this chapter we research what the main prepayment models are and how prepayment modelling has changed over time, where we focus on the use of machine learning models within prepayment literature. We follow this chapter by explaining the appropriate technical details of how the various classifying models to be used in this thesis function. In Chapter 4 we provide information of the available data set at Allianz, give relevant statistics of this data set and show how we will calculate the prepayment rate. In Chapter 5 we analyse what the main drivers are within the Allianz data set. Furthermore, in this chapter we discuss the effect of altering the data set size and how these changes can be corrected. In this chapter we also show how we evaluate each constructed model. Chapter 6 shows the results from the various models. We look at the performance of each model, the effect of the data set size on these models and lastly discuss if and how each model is able to provide insight into the important main drivers for prepayment. In the second to last chapter we give concluding remarks on the results found in Chapter 6, while in the last chapter we review the limitations of this thesis and give recommendations for future research.

Chapter 2

General Background

This chapter firstly gives general information on different facets of mortgage prepayment. This is done by giving a short overview of the Dutch mortgage market, hereafter we elaborate on determinants of mortgage prepayment and furthermore give information on previous historical research into prepayment modelling. It is followed by a section in which we give background information on different machine learning models and clarify why these models were chosen for this research.

2.1 The Dutch mortgage market

The Dutch mortgage market has changed significantly since the 2008 housing crisis. The average house price in 2020 was valued at €334,488 which is more than 30% higher than the average in 2008 (€254,918) (*StatLine - Bestaande koopwoningen; gemiddelde verkoopprijzen, regio*). However, during the period between 2010 and 2020 the total mortgage debt of the Netherlands only increased by 15% (*Wonen - Nederlandse Vereniging van Banken (NVB)*).
In 2010 roughly 80% of all outstanding mortgages had been provided by a bank whereas the mortgage market share of banks decreased to 57% in 2016 (*Staat van de Woningmarkt - Jaarrapportage 2020 | Rapport | Rijksoverheid.nl*) (*Grootbanken winnen terrein op hypotheekmarkt*). To add fuel to fire, a historic low for bank market share was reached midway 2020 when only 45% of new mortgages were being provided by the top 3 biggest banks (Rabobank, ABN Amro and ING) (*Marktaandeel grootbanken hypotheken bereikt historisch dieptepunt*). The decline of the market share of Dutch banks in the mortgage industry has allowed the market share of insurers, pension funds and other third party mortgage lenders to increase significantly. 'De Nederlandsche Bank' (DNB) shed light on this matter in 2016 (Bank, 2016) mentioning that pension funds and insurance companies have been incentivized to find long-term investments with low risk and a favourable yield due to the low interest rate climate. The confidence of investors in the Dutch mortgage market and the low number of write-offs has made the Dutch mortgage market attractive to pension funds and insurance companies.
The Dutch mortgage climate has historically been characterised by a high loan-to-value (LTV) ratio due to the Dutch tax system allowing for the deductibility of mortgage interest payments from taxable income (called 'hypotheekrenteaftrek' in the Netherlands). It was considered normal to have an LTV of 110% however, as of 2013 the maximum allowed LTV ratio has gradually been reduced and since 2018 the maximum allowed LTV for new mortgages is 100%. The reduction in the maximum LTV ratio has had a positive impact on the average LTV in the Netherlands as the average LTV of new mortgage loans has decreased from 94% in 2010 to 84% in 2019 (*Wonen - Nederlandse Vereniging van Banken (NVB)*). Despite this high LTV ratio the

Dutch households have a good reputation in paying off their debt, ranking amongst the lowest mortgage default rates in Europe.
Throughout the years the mortgage tax system in the Netherlands has allowed its inhabitants to deduct a large percentage of mortgage interest payments from their income (Box 1). The OECD (2011) reported the Netherlands as having the most generous tax relief system, in regards to house debt financing, of all OECD countries (Andrews and Sánchez, 2011). However, since 2013 the Dutch government has gradually been reducing the maximum mortgage interest payments deductible from taxable income. In 2019 the maximum percentage of mortgage interest payments deductible from a mortgagors income was 49%, every subsequent year since then the maximum deductible rate has been reduced with 3%. This ends in 2023 with the maximum deductible rate being 37.05%. Other major changes in the tax system, that have taken place over the course of the past decade, are with regards to the different mortgage types eligible for tax deduction. Before 2013 all mortgage types were eligible for tax deductibility, however, since then only the mortgages of type annuity and linear are tax deductible. This has had a drastic effect on the types of mortgages taken out by Dutch citizens. In the years up to 2013, amortizing mortgages (annuity and linear type mortgages) were accountable for 5% of newly issued mortgage debt (Uhe, 2019), with the mortgage of type bullet (interest-only) being accountable for more than half of new mortgages. However, since 2013 the linear and annuity type mortgages have accounted for more than half of new mortgage debt (*Bijlage Hypotheekschuldmonitor en update aanpak aflossingsvrije hypotheken | Kamerstuk | Rijksoverheid.nl*). Data shows that nowadays the most popular mortgage type under people of the Netherlands is the annuity mortgage (*Wonen - Nederlandse Vereniging van Banken (NVB)*).
The three most financed mortgage types since 2013 are: the annuity mortgage, linear mortgage, bullet (interest-only) mortgage.

- An annuity mortgage entails that the mortgagor pays a fixed monthly amount for the total duration of the loan. This fixed monthly payment is made up of an interest component and a principal component. Close to the start of the mortgage a large part of each monthly payment is made up of interest as the outstanding principal is still large. Throughout the life of the loan the principal decreases, hence reducing the interest part of the payment. This means that a larger portion of the monthly payments will start consisting of payments towards the principal, and less for the interest. The large interest payments at the start of the mortgage allow the mortgagor to deduct this from their income, thus reducing their income tax. However, as the interest component decreases towards the end of the mortgage so does the deductible portion eligible for the income tax .

- For a linear mortgage the principal component of the monthly payments stays fixed throughout the life of the mortgage loan and the interest component linearly decreases each month. At the start of this type of mortgage the total monthly payments are quite high and decrease over the mortgages' lifetime. The benefits of such a mortgage are that the total interest expenses over he whole lifetime are lower than other mortgage repayment types. Furthermore, the fixed principal component means that the mortgagor is able to start paying off the total principal relatively quickly, reducing mortgage debt. However, throughout the life of the mortgage the interest deductible from the income tax reduces.

- The bullet mortgage allows the mortgagor to only pay the interest component each month. However, at the end of the mortgage the mortgagor must repay the entire principal at once. Due to the high tax incentive this type of mortgage was very popular before 2013. This type of mortgage is often linked to savings account in order to help mortgagors repay the principal at the end of the mortgage.

2.1.1 NHG

The inhabitants of the Netherlands have the opportunity to take out a mortgage with an additional insurance called 'Nationale Hypotheek Garantie' (NHG). This insurance protects the mortgagor from any residual debt in the circumstance that a forced sale of their property were to happen due to special circumstances, such as the death of a partner or a job loss. In the case of a forced sale, amounting to a lower selling value than the outstanding principal on the mortgage the mortgagee has to pay 10% of this loss while the rest is covered by NHG. In order to be protected by this insurance, the mortgagor must pay an upfront fee of 0.9% (90 basispoints) of the loan value (Bonsema, 2019). Although this fee may discourage mortgagors from taking the insurance, the reduction in credit risk for the mortgagee results in a lower interest rate, hence benefiting the mortgagor.

2.1.2 Allianz Mortgages

In 2011 Allianz entered the Dutch mortgage market and has since been able to increase its offering of mortgage products. Due to the low interest rate climate mortgagors have been choosing mortgages with a long fixed interest period, such as the 20-year and 30-year fixed interest rate mortgages. Insurance companies have a strong presence in these segments as these long term, low risk mortgages fit in well into their strategy and risk appetite. This has allowed insurers like Allianz to increase its market share of the Dutch mortgage industry. As of July 2021 Allianz Benelux has roughly 33,000 mortgage clients and a Dutch mortgage portfolio of roughly €6 billion euros.

2.1.3 Mortgage funding

There are multiple ways of funding mortgage loans. Banks most often fund their mortgage loans through two different ways, they may finance the mortgage loans by using the credit balances of the savers' accounts at the bank or it do so through the capital market. In the capital market banks go into a 'swap' agreement with a counterparty. These swap agreements are based on the market swap rate (also known as the funding rate). The bank will most often agree to pay the counterparty a fixed interest rate while the counterparty will supply the bank with the current market rate (EURIBOR swap rate). If the market rate changes, the parties must pay the difference depending on whether the market rate increases or decreases.
Insurers, like Allianz, can also fund mortgage loans through internal financing from available assets which are the result of collected insurance premiums.

2.2 Mortgage Prepayment

2.2.1 Prepayment option risk

Dutch mortgagors have the option to prepay part of their mortgage. A prepayment takes place if the mortgagor pays the mortgagee an extra part of the principal above the contractual payment for that time period, thus decreasing the outstanding mortgage debt. The amount of the principal prepaid or the frequency of prepayment is up to the mortgagor. In the Netherlands, however, mortgagors have the option of prepaying 10% of the total principal yearly, penalty-free. Although this option is not stated in any law, most mortgage providers allow for this 10% yearly prepayment option. Above this amount the mortgagor will receive a penalty. In the situation that the mortgagor relocates and sells the property, they do not receive a prepayment penalty.

The prepayment option gives the mortgagor the right to prepay the principal loan before the obligatory termination date. Exercising the prepayment option reduces the time to termination of the mortgage, in turn decreasing the expected cash flows for the mortgagee. As mortgagors prepay their mortgage at random moments in time it is challenging for mortgagees to make a correct or accurate valuation of each mortgage (Kolbe, 2008).

The prepayment option embedded in the mortgage contract poses different types of risk for the mortgagee. When a mortgagor prepays part of the mortgage loan, the outstanding debt is reduced or redeemed altogether if the prepayment is large enough. This will result in a lower periodic payment for the mortgagor as the mortgagor will not have to pay any interest over the prepaid principal. Furthermore, a prepayment will decrease the life of the mortgage. This will result in different cash flow payments for the mortgagee than previously expected and calculated. In the case of a loan redemption the mortgagor has the option to refinance the mortgage at the current market rate. In the event that the current market rate is lower than the mortgage rate, the resulting cash flow for the mortgagee will be lower.

Banks hedge this by engaging in interest rate swaps, whereby they exchange a fixed interest rate in return for the variable market rate. However, in the case of a prepayment it may occur that the bank receives lower interest payments from its mortgagors than that it has to pay for the swap.

2.2.2 Determinants of mortgage prepayment

There are many determinants that influence mortgagors' decision to fully prepay their mortgage. These variables can be grouped into three categories, namely variables with loan characteristics, borrower characteristics and macro-economic characteristics. Although many determinants for mortgage prepayment have been researched over time, we only give a few of the most common determinants below (Clapp et al., 2001) (*MMI Fund Analysis FY 2005 Appendix A: Econometric Analysis of Mortgages A-1 Appendix A: Econometric Analysis of Mortgages*).

- Loan characteristics: loan age, loan amount, mortgage rate, penalty, geographical location, property type, loan-to-value (LtV)
- Borrower characteristics: age, income, credit worthiness, employment status, marital status
- Macro-economic: house prices, mortgage rates, interest rates

Research by (Alink, 2002), (Charlier and Van Bussel, 2003) and (Elsing, 2019) mention the mortgagor's age and property type as an important explanatory variable for forecasting the prepayment rate. (Charlier and Van Bussel, 2003) find a relation between the mortgagor's age and property type on the prepayment rate, naming it the 'upgrading effect'. They find that young people in an apartment have the increasing probability of upgrading their property as time passes and hence prepaying their mortgage. Although this effect is largely present in young people, the effect is greatest for mortgagors in their mid-thirties and forties.
(Alink, 2002) mentions the LTV as an important explanatory variable for the prepayment rate. This is backed by (Elsing, 2019) and (Sirignano, Sadhwani, and Giesecke, 2015), who empirically show that the mortgagors with a low LTV have a higher probability of prepaying. The reasoning behind this is, according to Elsing, that mortgagors with a low LTV are more likely to have additional money available to make a prepayment.
In regards to the Dutch mortgage market multiple empirical studies present different findings. (Bussel, 1998), (Charlier and Van Bussel, 2003) and (Hayre, 2003) mention seasonality as being an important determinant for mortgage prepayment. This is also mentioned by (LaCour-Little, Marschoun, and Maxam, 2002), but is not specific to the Netherlands. Seasonality is the effect in which prepayments take place closer to the end of the year. Possible reasons for more prepayments being made in the months November and December are due to the inflow of extra salary, which is often gifted as bonuses to employees (called '13e maand' in the Netherlands). Furthermore, it can also be fiscally beneficial for households to receive the prepayment penalty (achieved if you prepay more than 10% of the principal) so that this can be deducted from their yearly income. For these reasons the prepayment rate is lower in the months January and February.
All papers examined by (Jacobs, Koning, and Sterken, 2005) name refinancing incentive as a key determinant for mortgage prepayment. (Charlier and Van Bussel, 2003) even name it the most important determinant for prepayment. As also mentioned in other prepayment literature (not specific to the Netherlands) the decrease in mortgage market rates creates an incentive for the mortgagor to refinance (LaCour-Little, Marschoun, and Maxam, 2002). This determinant looks at the difference between the mortgage rate of the mortgagor and that of the market rate. If the current mortgage rate is far enough below the contract rate the mortgagor will most likely refinance their mortgage, which means 'trading' in their current mortgage for a new mortgage with a new principal and differing interest rate. Although there is no correct interest rate difference at which to refinance, the rule of thumb lenders suggest is refinancing if you can reduce the interest rate by at least 1%. In the case of the Netherlands, mortgagors will have to pay a refinancing penalty. However, not all mortgagors will optimally refinance in the event that refinance incentives are available. This is known as the burnout effect.
The burnout effect takes into account that not all mortgagors will behave rationally when presented with a refinancing incentive. The most aware mortgagors will refinance immediately when the mortgage rate drops to a level far below their own mortgage rate (Charlier and Van Bussel, 2003). The other mortgagors may not be aware that a refinancing opportunity has been presented, or they may be slow to react, thus acting irrationally[1] in the eyes of the researchers. The burnout effect takes

[1]This behaviour is called irrational in literature when in fact the behaviour depicted is only suboptimal. Researchers do not take external factors such as job change, divorce and illness into account that may be of importance. In this thesis we will use the term 'irrational' in order to stay consistent with literature.

into account that mortgagors, who have been given refinancing incentives in the past, will be slower to prepay when presented with these incentives in the future (Gan and Gan, 2009) .
Lastly, all the papers analysed in (Jacobs, Koning, and Sterken, 2005) mention seasoning as a key determinant for mortgage prepayment. Seasoning explains the behaviour between the prepayment rate and the age of the loan. It is stated in literature that this relationship often has a s-shaped curve (Hayre, 2003). The number of prepayments after the inception of a mortgage usually remains low due to factors such as the interest rate, family composition and employment staying unchanged. Over time, however, the number of prepayments will gradually increase until they reach a stable state and remain constant (Charlier and Van Bussel, 2003). The prepayment rate is said to have reached a "seasoned" level.

2.2.3 Conclusion sub-question 1

1. What are the main prepayment drivers stated in literature?

In section 2.2.2 we state all the relevant determinants found in literature specific to prepayment modelling. Although not all determinants were found in literature specific to the Netherlands most, if not all of these determinants can be applied to the Dutch mortgage market. Below we summarise all the relevant determinants and classify them into loan, borrower and macro-economic characteristics:

- Loan characteristics: loan age, loan amount, mortgage rate, penalty, geographical location, property type, loan-to-value (LtV), seasonality, refinancing incentive, burnout effect
- Borrower characteristics: age, income, credit worthiness, employment status, marital status
- Macro-economic: house prices, mortgage rates, interest rates

2.2.4 Prepayment models: optimal vs. exogenous

(Kau and Keenan, 1995) write in their overview article on mortgage option pricing that prepayment can be considered as an American-style call option, as the borrower has the right to gain the whole house at any time by paying off the entire mortgage. As with prepayment, defaulting can also be regarded as an option, namely that of a European put option. According to the article, prepayment can be separated into two distinct classes, namely as an optimal (endogenous) prepayment and an exogenous prepayment. An optimal prepayment takes place as a result of the borrower minimizing the market value of the loan, independent of the borrowers individual characteristics. (Bussel, 1998) investigates this model and finds that optimal prepayment behaviour is based on external drivers such as the market interest rate and mortgage rate available to the mortgagor. The prepayment behaviour assumption van Bussel makes is that a borrower prepays a loan when "the present value of the loan exceeds the outstanding debt plus any transaction costs" (Bussel, 1998). (Kuijpers and Schotman, 2007) applies the optimal prepayment theory to the Dutch mortgage market, developing a model for the valuation of limited callable mortgages, taking into account the yearly penalty fee.
However, empirical results found by (Bussel, 1998) show that the mortgagors do not always follow optimal prepayment behaviour, which is also mentioned before in subsection 2.2.2. This was already acknowledged by (Dunn and Mcconnell, 1981)

who find that mortgagors do not optimally call the prepayment option, sometimes exercising their option when the market interest rate is above the mortgage rate stated in the contract. Dunn and McConnell state the reason for mortgagors exercising sub-optimal calls is often due to job relocation or other behavioural variables[2].
The stochastic nature of prepayments, which is due to existence of borrower-specific variables such as job relocation, has led to the development of exogenous modelling for prepayments. This type of modelling looks at the relationship between the prepayment rate and different explanatory variables. The availability of borrower-specific variables which are loan specific has allowed researchers to investigate the prepayment on a loan level. Furthermore, by pooling the loan level data researchers can find the prepayment rate on a portfolio level.

2.2.5 Prepayment modelling over time

Literature on prepayment modelling started with research towards the optimal prepayment model. (Dunn and Mcconnell, 1981) make a model in which prepayments are only interest rate driven and base it on the Cox-Ingersoll-Ross (CIR) term structure model. Dunn and McConnell further incorporate some non-financial termination features into the model through a Poisson-driven process. (Brennan and Schwartz, 1985) build on this model by incorporating sub-optimal prepayment behaviour into it. Furthermore, they extend on the model by using a two factor model to value both the mortgage and the prepayment option. (Bussel, 1998) builds on this framework and applies it to the Dutch mortgage market using a nonlinear and nonparametric model. Although these models are all based on the optimal prepayment model, every researcher finds that mortgagors, to some extent, exhibit irrational behaviour.
With many optimal model researchers mentioning the role of irrational behaviour in mortgagors, research shifted towards exogenous prepayment modelling that incorporate a behavioural framework. The field of exogenous prepayment modelling can be further subcategorized into two alternative models, the logit model and a range of survival models. The most commonly used survival model is the proportional hazard model, which was first developed by (Cox, 1972). The first to implement this model for prepayment modelling were (Green and Shoven, 1986), in which they measure the time until termination of a mortgage, based on one predictor variable, namely the refinancing incentive (interest rate). Although no specific loan or borrower characteristics are incorporated in the model, the (Green and Shoven, 1986) mention that these characteristics are the primary determinants for prepayment, but due to lack of data the researchers are unable to research them. (Schwartz and Torous, 1989) extend on the model by (Green and Shoven, 1986) and add more explanatory variables, such as seasoning, seasonality and other demographic variables to name a few. (Charlier and Van Bussel, 2003) build on the model by (Green and Shoven, 1986) by incorporating more explanatory variables and by applying the model to the Dutch mortgage market.
The availability of detailed loan-level data has allowed researchers to incorporate more explanatory variables and research whether certain loan and borrower characteristics are important determinants for mortgage prepayment, which was previously often not possible due to insufficient data. Towards the end of the 1990's the first publications of prepayment modelling using a logit model are published. Publications such as (Archer and Ling, 1993), (Archer, Ling, and McGill, 1996)(Archer,

[2]In the eyes of researchers such variables are deemed as irrational even though such variables are perfectly rational and normal.

Ling, and McGill, 1997) and (Green and Lacour-Little, 1999) are among the first to model the probability of mortgage prepayment using a binary outcome. They model the prepayment rate, incorporating a multitude of loan- and borrower-specific characteristics. The first extensive research modelled with Dutch mortgage data is done by (Alink, 2002), who uses SNS bank data in the period from 1993-2001 in his logit model (Jacobs, Koning, and Sterken, 2005).

2.3 Machine Learning models

The logit type models have become the main baseline model within risk management departments of mortgage providers due to the easy interpretability, implementation and efficiency ("Forecasting Loan Default in Europe with Machine Learning*"). However, the popularity of machine learning models has increased greatly over the past few decades due to the increase in abundance and affordability of computational power. The increase in data collection and improvement in computational power has allowed machine learning algorithms to become viable options for all types of financial forecasting, including modelling prepayment behaviour. Some examples in which alternative algorithms for prepayment modelling are used are (Sirignano, Sadhwani, and Giesecke, 2015), (Saito, 2018) and (Habib, 2020). A logistic regression model can be regarded as a simple machine learning algorithm according to the definition given by (Samuel, 1959), yet there are more complex models that are viable replacements for the logit type models.

2.3.1 Random Forest

Logistic regression is often an easy technique to implement and to interpret. However, one disadvantage is the assumption of linearity between the dependent and independent variables. A deep learning study performed by (Sirignano, Sadhwani, and Giesecke, 2015) uncovered a nonlinear relationship between the dependent variable, prepayment behaviour, and the independent variables. Empirically they find that in their data, variables such as the refinance incentive, unemployment rate and LTV ratios are non-linear with the prepayment rate (to name a few). Furthermore, they find that the non-linearity in some cases is caused by variable interactions. A complex relationship between the dependent and independent variables is hard to obtain using a logistic regression and for such cases one is better off using a different algorithm that is able to take this relationship into account. For example, a study on Brazilian loan default data using various machine learning techniques finds that the random forest classification technique outperforms the baseline logistic regression model (Aniceto, Barboza, and Kimura, 2020). Not only does the study show that the random forest classifier performs better that the logistic regression model, it performs better that other tested machine learning models. (Aniceto, Barboza, and Kimura, 2020) suggest that banks can improve default prediction models by investigating different machine learning algorithms.
A study performed by (Luo, 2019) compares the random forest classifier with five different models including, the Artificial Neural Network, Support Vector Machines and logistic regression and considers the random forest as the best classifier due to it having the lowest error rate on the data. The advantages of a random forest classifier are that it has excellent predictive power and that no user prespecification is needed between the dependent and independent variables. This characteristic is regarded to be beneficial as previous literature stated the irrational behaviour seen in mortgagors (Dunn and Mcconnell, 1981) (Charlier and Van Bussel, 2003) (Brennan

and Schwartz, 1985). For these reasons this thesis will use a random forest model to predict the mortgage prepayment rate.

2.3.2 Neural Network (NN)

The last classifier that will be considered in this research is the neural network (NN). Research done by (Riksen, 2017) on mortgage prepayment prediction finds that the neural network classifier was able to capture the relation between the explanatory variables and the prepayment rate. Although neural networks have gained popularity in the past few years, research published in 1998 already examined the predictive capability of neural networks. This study performed on the prepayment of residential mortgages finds that the neural network model was able to predict prepayments with over 70% accuracy while the logit model only achieved a 50% accuracy rate (Waller and Aiken, 1998). When applied on differing credit risk data sets, the neural network classifier gives high accuracy predictions when compared to other statistical classifiers (Damrongsakmethee and Neagoe, 2019) (Assef et al., 2019). Furthermore, (Luo, 2019) states that the neural network is only beaten by the random forest classifier when considering the error rate as evaluation metric. Artificial neural networks perform well with complex and nonlinear data and do not require specified relationships among the dependent and independent variables. Hence, it is a viable option for modelling mortgage prepayment behaviour. Thus, besides using the logistic regression and random forest model, we will also use model the NN and compare the results between the three models.

2.4 Summary

This chapter started out by giving an introduction on the mortgage market of the Netherlands. We share some statistics that characterise the Dutch mortgage market and give a short explanation on the three main types of mortgages. We give some insights on Allianz and explain the different ways mortgagees can fund mortgages. Next we elaborate on the prepayment option risk that is embedded into mortgage contracts before we perform a literature research on the determinants of prepayment, answering the first sub-question of this research. We follow this by literature over prepayment modelling over time and lastly we state the past research done on the usage of the machine learning models random forest and neural network in combination with mortgage prepayment.

Chapter 3

Technical Background

This chapter describes the inner workings of all three the models to be used in this research, namely that of the logistic regression, random forest and the neural network. We give the relevant mathematical formula's regarding each model and further elaborate on different properties of each model. This is done so that the reader is familiar with each model.

3.1 Logistic Regression

A logistic regression is a useful predictive algorithm that can be used to classify the dependent variable into one, or more, categories. The output is bounded between zero and one, whereas the linear regressor predicts continuous values for the dependent variable. For this thesis we consider the logistic regression as the baseline model for which the other models will be compared to. A logistic regression fits into our situation well as we want to predict if the mortgagor makes a prepayment (classified as one) or does not make a prepayment (classified as zero). The probability of a prepayment taking place is denoted as follows, $P(Y_{i,t} = 1|X_{i,t})$, which is the conditional probability of a prepayment $Y_{i,t}$ (independent variable) taking place given the set of explanatory independent variables $X_{i,t}$. Furthermore, we denote each mortgage i at time t. The following equation is called the logistic function:

$$log\left(\frac{P(Y_{i,t} = 1|X_{i,t})}{1 - P(Y_{i,t} = 1|X_{i,t})}\right) = \beta X_{i,t} \tag{3.1}$$

The left side of the equation is considered the logit function or log-odds and β are the independent variable coefficients. If we solve for $P(Y_{i,t} = 1|X_{i,t})$ we get:

$$P(Y_{i,t} = 1|X_{i,t}) = \frac{e^{\beta X_{i,t}}}{1 + e^{\beta X_{i,t}}} = \frac{1}{1 + e^{-\beta X_{i,t}}} \tag{3.2}$$

which is a sigmoid function in which the output resembles that of a binomial probability distribution with an output between 0 and 1. As a logistic regression predicts probabilities, we can use the maximum likelihood estimator (MLE) to estimate the parameters of the regression. For a more detailed explanation please refer to Appendix A.1.

3.2 Random Forest

The subsequent model that will be used for forecasting the prepayment rate is the random forest classifier. The random forest model operates by constructing a multitude of decision trees and combining them to one final classification.

3.2.1 Decision Trees

A decision tree is a supervised learning algorithm that uses a tree like structure to predict for the dependent variable Y using multiple splits on the input variables $X_1, X_2 ... X_n$. A decision tree resembles that of a flowchart in which we make decisions to eventually reach a class label of the dependent variable. The splits in the dataset of a decision tree take place at the tree nodes. At every node we ask ourselves a question in order to split up samples with similar characteristics. One such split regarding mortgage data that we could ask at a node would be *is the mortgage principal more that €300,000?*. The 'yes' or 'no' flows from this node we call the branches. Eventually after asking many questions, related to different attributes, we end up with a subset of the original dataset which represent the path we took from the starting node until the end. This we call a leaf node. The goal of the decision tree is to separate data samples that have similar characteristics with the same classification label (dependent variable). The higher percentage of the same classification label we have at the leaf node, the better the predictive power of the decision tree. An example of a simple decision tree for playing tennis is given in figure 3.1.

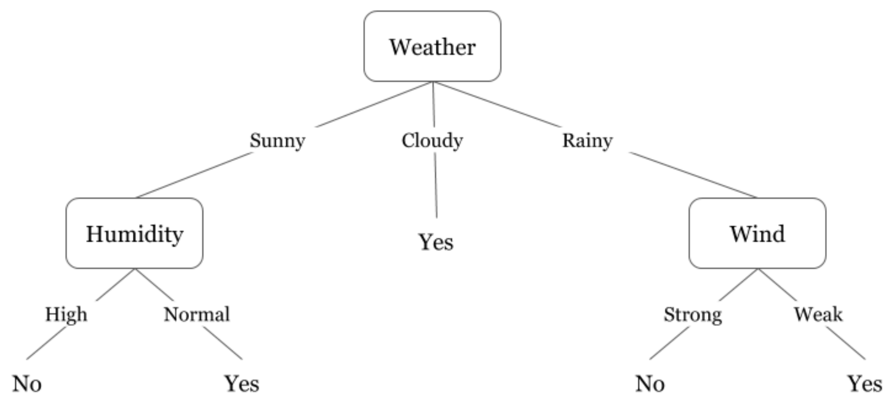

FIGURE 3.1: A schematic representation of a decision tree for playing tennis, based on the outside weather (Waiganjo Wagacha, 2003).

3.2.2 Random Forests

The random forest classifier works by constructing multiple decision trees and pooling the results. This technique of creating multiple models and combining them to improve results is called an ensemble learning method. The random forest classifier makes use of the ensemble method called bagging, in which each decision tree is modelled on a 'new' training data set. This data set is chosen by taking a random sample with replacement of size N from the available training data set. Thus, each decision tree is modelled on 'different' data. Bagging decorrelates each tree from one another. In addition to using the bagging ensemble method, the random forest uses

a different method to split the tree. Whereas a decision tree considers every predictor (independent variable) p at a split, the random forest considers a random sample of m predictors from the available p predictors as possible split candidates. A new sample of m predictors is taken at each split. The rule of thumb for how many predictors m are chosen from the total number of p predictors is $m = \sqrt{p}$. This rule of thumb is backed by (Breiman, 2001). Thus, at each split the algorithm is not allowed to consider a majority of the available predictors (James et al., 2013).
Choosing a random sample of in total $m = \sqrt{p}$ predictors is what distinguishes the random forest classifier from a group of decision trees that have been enhanced with the bagging ensemble method. Suppose we have one very strong predictor. By only using bagging, most (if not all) decision trees would use this strong predictor in the top split. Consequently, all trees would be similar to one another and hence all the predictions would be highly correlated. Bagging on its own, will in general not lead to a substantial reduction in variance over a single decision tree in this setting (James et al., 2013).
(Breiman, 2001) introduces a notation Θ_k for all the random choices which are made when fitting the kth tree. This notation takes into account both the random selection of data and the random selection of predictors, giving the following notation for each classifier:

$$h(\mathbf{x}, \Theta_k) \tag{3.3}$$

where $k = 1, 2, ..., K$, $\mathbf{x}$ is an input vector, and $\{\Theta_k\}$ are independent identically distributed random vectors (Breiman, 2001). For each classifier $h(\mathbf{x}, \Theta_1), h(\mathbf{x}, \Theta_2), ...,$ $h(\mathbf{x}, \Theta_K)$ a prediction is made for the dependent variable Y using input data drawn from the explanatory variables $\mathbf{X}$. From this prediction we can define the margin function as follows

$$mr(\mathbf{X}, Y) = P_\Theta(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X}, \Theta) = j). \tag{3.4}$$

The margin function is the probability of correct predictions minus the maximum probability of incorrect predictions. Thus, the larger the margin, the more certain the prediction. From the margin function we can define the generalization error as follows

$$PE* = P_{\mathbf{X},Y}(mr(\mathbf{X}, Y) < 0) \tag{3.5}$$

where the subscripts $\mathbf{X}, Y$ indicate that the probability space is over $\mathbf{X}, Y$ (Breiman, 2001). The probability given in equation 3.5 gives the probability of switching from one class to another for different random samples of the data set. Thus, a high probability means that the assigned class of the data is not stable for different samples. After this, (Breiman, 2001) proves that the generalization error converges to the population generalization error as the number of trees are increased. This is given in the following formula:

$$P_{\mathbf{X},Y}(P_\Theta(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X}, \Theta) = j) < 0) \tag{3.6}$$

which gives the probability of switching from one class to another. It is implied by (Breiman, 2001) that as the number of trees in a random forest increases, the random forest is less likely to overfit the data.
(Amit and Geman, 1997) state that the accuracy of a random forest depends on the strength of each tree classifier and the correlation between each tree. As stated by (Breiman, 2001), the strength of a set of classifiers is given by

$$s = E_{\mathbf{X},Y} mr(\mathbf{X}, Y) \tag{3.7}$$

which is the expected value of the margin function, whereby a high value is preferred. Furthermore, (Breiman, 2001) shows that there is an upper bound for the generalization error

$$PE* \leq \frac{\bar{\rho}(1 - s^2)}{s^2} \tag{3.8}$$

where $\bar{\rho}$ is the average correlation between the trees of the random forest. Equation 3.8 shows how the upper bound depends on the strength and correlation, as previously mentioned by (Amit and Geman, 1997). Decreasing the upper bound is desirable and can be achieved when the strength, s, is close to 1. Furthermore, a lower correlation between the trees also decreases the upper bound.
In summary, by showing the theory developed by (Amit and Geman, 1997) and (Breiman, 2001) we show that the generalization error does not increase when the random forest is comprised of a larger number of trees. Furthermore, the increase in trees of the random forest does not overfit the data. These reasons show that by increasing the number of trees in a random forest, there are no negative side effects.

3.2.3 Splitting Criteria

When growing a classification decision tree we are interested in predicting the class proportions of the training data at every particular terminal node region. In growing the tree we split the tree using binary recursive splitting, which means we start at the top of the tree and go down the tree level by level according to the splitting logic. Literature shows that there are two popular measures that can be used as splitting criteria, namely the Gini index and the cross-entropy (Nembrini, König, and Wright, 2018) (Hamza and Larocque, 2005). Both evaluation criteria are very similar to one another. This thesis will use the Gini index as splitting criterion as it is computationally faster than the cross-entropy criterion. This is because the formula for the cross-entropy is more complex than the Gini index. The formula for the Gini-index is defined by

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3.9}$$

where $\hat{p}_{mk}$ is the proportion of training observations in the mth region that are from the kth class (James et al., 2013). The formula is a measure of the total variance across the K classes. The Gini index will take on a small value if a tree contains nodes with a high purity, meaning that the labels (prepayment versus non-prepayment) are homogeneous.

3.3 Artificial Neural Network

An artificial neural network is a machine learning algorithm that resembles the functions of a human brain in the sense that it contains a network of neurons (also known as nodes) in which an incoming signal is transformed into an output signal. A neural network often consists of an input layer, one or more hidden layers and an output layer. A schematic representation of such a neural network can be seen in figure 3.2. The signals from all the nodes in the previous layer are sent to each node in the subsequent layer.

The artificial neural network consists of two distinct phases concerning the flow of information through the network, namely the forward propagation phase and back propagation phase.

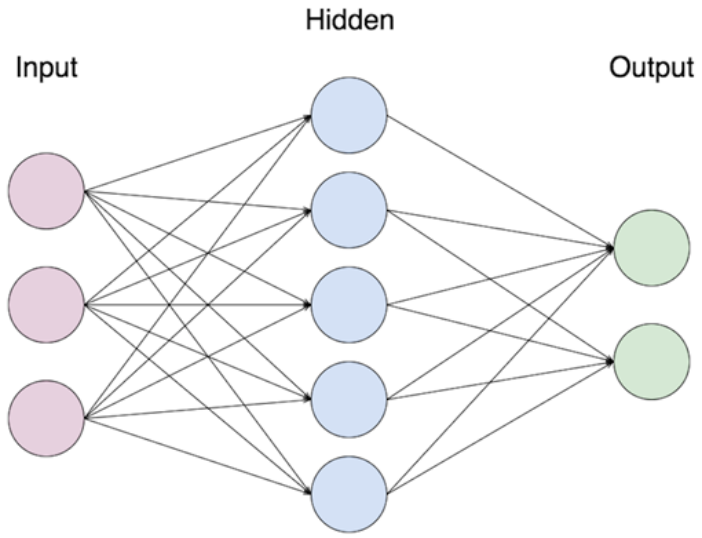

FIGURE 3.2: A schematic representation of an artificial neural network.

In the forward propagation phase the input nodes of the neural network receive numerical data from an outside data set and pass the information through to the hidden layer of neural network, where all the computations take place. Each neuron in the first hidden layer receives an input signal, multiplied by a weight, from the neurons of the previous layer. The summation of these signals is then passed through the activation function and then leaves the node as an output signal, which in turn is used as an input signal for the nodes in the next layer, where the whole process is repeated. Figure 3.3 shows the computational process that happens at every node of the hidden layer. The net input for every node can be given by the following equation

$$a_j = \sum_i x_i w_{ij} + \beta_j \tag{3.10}$$

where x_i is each input, w_{ij} is the weight given to the input between node i to j and β is the bias, the term given for unknown parameters. The weighted net input is then passed through the activation function before leaving the node as output. The activation function $h(\cdot)$ is shown in the equation below

$$z = h(a_j). \tag{3.11}$$

The weights for each input are found by the neural network during training, however, the activation function must be chosen by the user. Without an activation function the neural network would essentially turn into a linear regression model. The activation function is needed for non-linear transformations and allows the neural network to perform complex tasks.

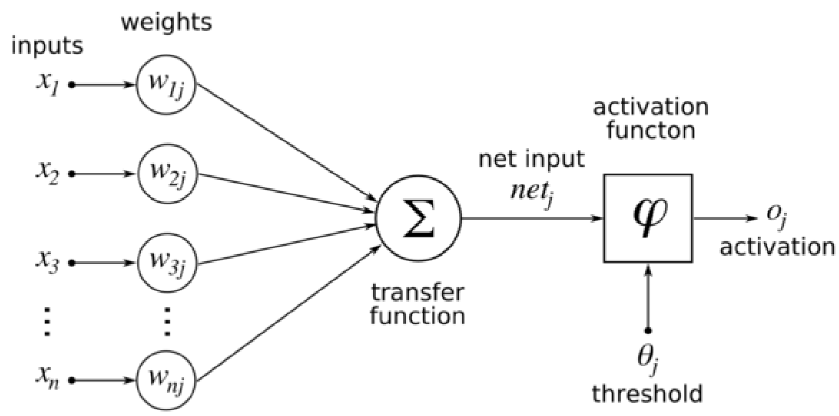

FIGURE 3.3: A schematic representation of the computational processes in each node of a neural network.

3.3.1 Activation function

There are a number of activation functions which can be chosen by the user. These activation functions can be chosen for both the hidden layers in the neural network and the output layer. It should be noted that the chosen activation function should be differentiable, as this property will be needed for the back propagation phase of the training process.

- **Sigmoid function**

The sigmoid activation function returns an S-shaped curve in the range $[0,1]$ and is essentially the same output function present in a logistic regression. The formula is given below

$$h(a_j) = \frac{1}{1+e^{-a_j}}. \tag{3.12}$$

The sigmoid function is a popular activation function due to its easy to use nature and interpretability. Furthermore, it is a non-decreasing function with a smooth output. Small changes in the node weights will not affect the output significantly, which is a desirable property as later on we change the node weights in order to improve the model during training runs.
However, the model has problems during the backpropagation phase (this phase is explained in subsection 3.3.4) in which the sigmoid activation function will tend to saturate at both tail ends, where the values reach 0 and 1. Secondly, the fact that the sigmoid function is not zero-centered means that the neural network training time increases.

- **Hyperbolic tangent function**

The function takes in a real-valued input and squashes the range to $[-1,1]$ and is given by the formula

$$tanh(a_j) = \frac{sinh(a_j)}{cosh(a_j)} = \frac{e^{a_j} - e^{-a_j}}{e^{a_j} + e^{-a_j}}. \tag{3.13}$$

The hyperbolic tangent function is zero-centered and hence is an improvement to the sigmoid function. However, it should be noted that the function will still saturate at both tail ends 0 and 1, thus this problem will still persist when using this activation function. This function is considered useful when performing a classification prediction between two distinct classes. The comparison between the tangent activation function and the sigmoid function can be found in Figure. 3.4.

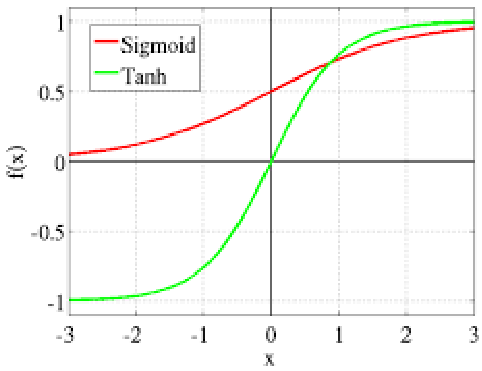

FIGURE 3.4: A graph depicting the difference between the hyperbolic and sigmoid activation functions

- **Rectified Linear Unit**

The rectified linear unit (ReLu) function rectifies the saturation problem persistent in the sigmoid and tangent functions. The formula is given as

$$h(a_j) = max(0, a_j). \tag{3.14}$$

This function simply has a threshold at zero and is linear with slope 1 when a_j is above zero. This function is found to accelerate the gradient descent compared to the sigmoid and tangent function. Furthermore, the function involves simple operations making it a highly efficient function, which is not the case for the other functions. The disadvantage of the ReLu is that many nodes can have zero activation due to the negative input values not passing through the threshold. If this is the case, the node is considered 'dead'. Once a node has a zero output it cannot recover due to the zero gradient threshold of the ReLu function.

- **Leaky ReLu**

The Leaky ReLu tries to fix the 'dying' problem found in the ReLu function. Instead of having zero for $x < 0$ the leaky ReLu will have a small negative slope. This is given by the following formula

$$h(a_j) = \begin{cases} a_j & a_j > 0 \\ \alpha a_j & a_j \leq 0 \end{cases}. \tag{3.15}$$

Where α is usually a small number such as 0.01. Although people have reported successes with this activation function, the results are not always consistent.

- **Softplus activation function**

The softplus function is largely similar to the ReLu and is much smoother around zero than the ReLu function. This function has gained popularity in the past years and is often used for K-class classification problems. The function is characterized by the following formula

$$h(a_j) = ln(1 + e^{a_j}). \tag{3.16}$$

Although it has gained popularity, the softplus is computationally inferior to the ReLu function due to the more difficult exponential calculation term.

In this thesis we opt for the sigmoid function as we regard interpretability and easy use higher than the additional training time that is needed. Furthermore, the smooth output given by this activation function is a quality that is needed during model training.

3.3.2 Output function

The output layer is the last layer of the neural network and is responsible for the final output. It is the rightmost layer in Figure. 3.2. The output layer also has an activation function, which is dependent on the purpose for which the neural network was built for. Most often it is dependent on the type of cost function. However, for a classification problem the softmax activation function is usually preferred.

3.3.3 Loss function

In order to assess each neural network model we make use of something known as a loss function. This function allows us to measure the performance during each training run. With a loss function we look at the difference between the predicted values and the correct values. The lower the loss, the better the predictions made by the model. The default loss function used for classification problems is the binary cross-entropy loss function (Brownlee, 2016). This loss function works well with a sigmoid activation function. We give the formula for binary cross-entropy loss function below.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i log(p(y_i)) + (1 - y_i) log(1 - p(y_i)) \tag{3.17}$$

In which y_i represents the actual class, which in this case would be a prepayment, and $p(y_i)$ is the probability of that class.

3.3.4 Backward propagation

Backpropagation is an algorithm that is used to minimize the error from the loss function. During the backpropagation phase we in essence travel back from the output layer to the hidden layer(s) to change the weights w at each node in order to decrease the prediction error of the output layer given by the loss function. The delta change of the weights leads a different loss function error. This algorithm works by calculating the gradient of the loss function, known as gradient descent. As we have multiple weights the gradient is a vector of partial derivatives with respect to the weights (*Reducing Loss: Gradient Descent*). The backpropagation process is repeated until we reach our desired output.
As mentioned we calculate the gradient of the loss function during backpropagation. However, activation functions such as the sigmoid and the hyperbolic tangent will

tend to saturate at tail ends 0 and 1. At these tail end the local gradient for the gradient descent is (close to) zero meaning that nearly no signal will flow through.

3.4 Summary

In this chapter we shed light on the technical details of the three models to be used in this thesis, namely the logistic regression, the random forest and the neural network. For each of the three models we firstly explain how they work and then elaborate on the mathematical formula's each model uses in order to predict the dependent variable.

Chapter 4

Data

This chapter will elaborate various aspects of the data set which will be used to train and test the model on. In depth information will be given about the data on a portfolio-level and loan-level. Furthermore, details will be given on the relationship between the explanatory variables and the dependent variable. The final section in this chapter will focus on the various processing steps taken in order to balance the data set.

4.1 Portfolio Data Properties

The data set to be used for this research is provided by Allianz Benelux. As mentioned in section 2.1.2 Allianz only started issuing mortgages in 2011, making it a relatively young mortgage provider. The implication of this is that a full mortgage cycle has not yet taken place. This does not have to be a negative thing, however, most prepayment literature is based on a full mortgage cycle, which can take up to thirty years. Therefore, there might be differences between the relationships of some explanatory variables with the prepayment rate for the Allianz data compared to literature.

The data that will be used for this thesis consists on various monthly data files which all contain different contract specific variables. The data time period to be used is from June 2012 up to July 2021. The data prior to June 2012 was discarded due to incomplete data files. Since 2011, Allianz has seen a substantial growth in its mortgage portfolio. In June 2012 the outstanding portfolio balance amounted to €33 million whereas in July 2021 this amounted to roughly €6 billion. In this same time period the number of loan parts increased from 380 to roughly 60,000[1]. This growth is depicted in figure 4.1. It must be noted that the outstanding balance at the beginning of 2018 was roughly €1.2 billion, meaning in the time period between 2018 and July 2021 the outstanding balance has quadrupled in value. Although this growth is a great achievement, the down side is that the large production of mortgages decreases the average loan age even further for an already young mortgage portfolio. We consider this a downside because the mortgage portfolio will be less mature, also called less 'seasoned' (this will be further elaborated in subsection 5.1 under the header 'Seasoning'). A less seasoned portfolio will most often lead to less prepayments. As mentioned in literature and noted in subsection 2.2.2 the number of prepayments at mortgage inception is usually low due to factors such as the interest rate, family composition and employment status remaining unchanged. In that respect the Allianz data will differ to that of other big Dutch mortgage lenders such as the ING and ABN AMRO. These competitors have fully matured portfolios in

[1]The Allianz portfolio currently consists of just under 60,000 loan parts, this number is not to be mistaken with the total number of loan parts shown in table 4.2, which shows the total number of unique loan parts in the whole data set ranging over multiple years.

which mortgage cycles are constantly ending and restarting. These portfolio's will most likely have higher prepayment rates that are more representative for the Dutch population. The models in this research that will be modelled using the Allianz data will most likely be biased towards lower prepayment rates when compared to that of fully matured mortgage portfolios.

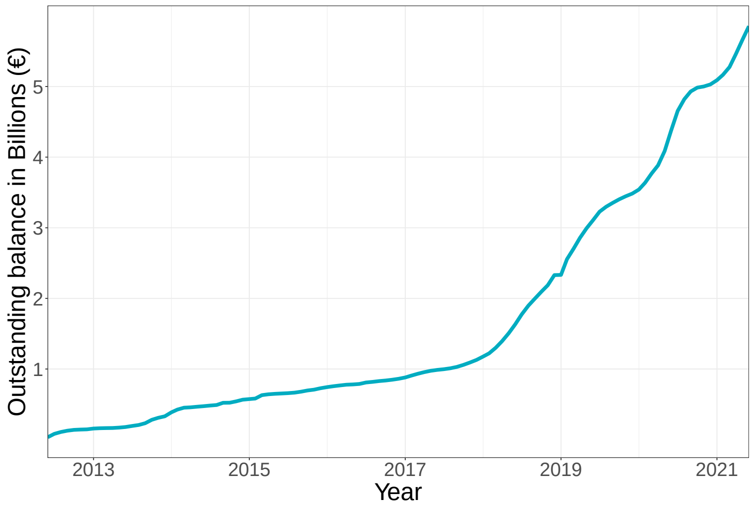

FIGURE 4.1: The outstanding balance of the Allianz Benelux mortgage portfolio over time.

4.1.1 Loan Level Properties

The available data at Allianz consists of various monthly files which all contain different contract specific variables. The various monthly files are monitored and new variables are added regularly. Since inception in 2011, the number of reported monthly variables has increased. Currently the monthly files consist of 631 variables. We can discard a large portion of these variables due to irrelevance (examples of such variables would be house number, house number addition and ending date of mortgage) or highly incomplete data (as such variables, for example, have only been introduced in the reporting files in recent months). Collecting all the monthly data observations and discarding all irrelevant variables we end with a complete data set consisting of $1,660,000$ observations and 57 variables. This data set consists of $34,984$ unique contracts with $63,830$ unique loan parts. To this data set we add additional macro-economic data such as the house price index (HPI) per month (as reported by the Dutch bureau of of statistics CBS), the unemployment rate and the monthly number of houses sold as these have been mentioned by literature to be important determinants of mortgage prepayment (as previously mentioned in subsection 2.2.2. In table 4.1 we give the average and maximum values for multiple important mortgage variables for the whole data set.
To further elaborate on monthly mortgage data, mortgagors have the opportunity to construct mortgage loans consisting of multiple loan parts. Each loan part may have varying characteristics. For example, each loan part may have a different redemption type, interest-rate fixed period (IFP) and capital amount. The combination of loan characteristics and the application date of the loan all influence the loan part mortgage rate. Table 4.2 shows the differences between the redemption type and IFP on the loan part rate.

	Average	Maximum
Mortgagor age	41	97
Income	€53,150	€1,180,900
Total principal	€193,950	€795,000
Remaining capital	€108,800	€750,000
Loan age	23	110
Mortgage rate	2.64	6.93
Burnout	6.67	84
Initial LtMV	0.89	1.04

TABLE 4.1: The average and maximum values of multiple mortgage variables.

IFP class	5	10	15	20	25	30	**Total**
Annuity (#)	886	4,070	1,695	23,719	1,575	9,816	41,761
Linear (#)	83	409	199	2,695	135	1,146	4,667
Bullet (#)	209	2,288	593	9,049	355	4,908	17,402
Total (#)	1,178	6,767	2,487	35,463	2,065	15,870	63,830
							Average
Average loan part rate	1.42%	2.66%	2.11%	1.94%	1.78%	2.15%	2.0%
Average loan part capital	€27,436	€96,552	€108,280	€108,946	€102,606	€112,584	€92,734

TABLE 4.2: Various mortgage characteristics for each interest-fixed-period (IFP) class.

Table 4.2 shows the number of loan parts per interest rate fixed period for mortgages with different redemption types. Furthermore the average mortgage rate and average original capital is shown for each inter fixed period. Looking at the total number of loan parts per IFP we see that the 10-, 20- and 30- years are the most popular. The reason for this is because Allianz prefers longer maturity mortgages as these have a better match with their long-term liabilities. Between the redemption types we see that the annuity type mortgages are chosen most often. This type of redemption has been the most popular type in the Netherlands for multiple years now as the monthly costs at the beginning of the mortgage are cheaper than those of a linear mortgage.

Apart from being contractually obliged to pay monthly costs, comprising of an interest part and a principal part, mortgagors can prepay part of their loan. The three types of prepayment, partial prepayment, full prepayment and arbitrage prepayment make up the total prepayment. Prepayment is calculated as an annual rate, called the conditional prepayment rate (CPR). Equation 4.1 shows the formula for the CPR for month t.

$$CPR_t = 1 - (1 - SMM_t)^{12} \tag{4.1}$$

Where SMM is the single monthly mortality rate (SMM), which is a monthly rate for the total prepayments. The SMM for a specific month is computed as follows:

$$SMM = \frac{\text{€ } Total\ amount\ prepaid}{\text{€ } Total\ remaining\ outstanding\ debt} \tag{4.2}$$

The total amount prepaid per month is a summation of the amounts prepaid through a partial, full and arbitrage prepayment. A partial prepayment takes place when the amount repaid is more than is contractually obliged for that month, but less that the outstanding balance of the mortgage. A mortgagor may make multiple partial prepayments over the lifetime of the loan. However, as mentioned in section 2.2.1, a mortgagor may only partially prepay more than 10% of the principal per year, penalty free. This type of prepayment is also called a curtailment. As the name suggests, a full prepayment takes place when the whole outstanding principal is redeemed in advance. A full prepayment usually only occurs when a mortgagor divorces, relocates or passes away. The last type of prepayment is not an actual prepayment. An arbitrage prepayment takes place when a mortgagor refinances the mortgage, taking advantage of lower market rates. An arbitrage prepayment has a similar effect as a full prepayment, due to the fact that the old loan is terminated. However, with an arbitrage prepayment the old loan is replaced by a new loan with a lower mortgage rate. This type of prepayment is not very popular in the Netherlands as it brings with it additional costs. Table 4.3 shows the total number of observations per prepayment type. Furthermore, the total amount, average amount and average loan age are shown.

	Number of observations	Total amount	Average Amount	Average Loan Age (months)
Full prepayment	5,801	€572,034,633	€98,610	40
Partial prepayment	12,199	€85,986,101	€7,049	26
Arbitrage prepayment	911	€95,032,913	€104,317	57

TABLE 4.3: Prepayment statistics of the Allianz data set.

In total we have $18,911$ observations in which some form of prepayment takes place. From table 4.3 we see that the average amounts of full and arbitrage prepayments are much larger than the partial prepayments. This is obvious as a full and arbitrage prepayments payoff the whole remaining principal.

It should be noted that during the time frame in which Allianz has been selling mortgage products (2011-present) the Dutch mortgage rates have been in extremely low territory. This is mostly due to the low interest rate climate which, in turn, has affected the swap rate. An important part of mortgage funding, as explained in subsection 2.1.3, is based on the swap rate. A lower swap rate will lead to a lower mortgage rate climate. In figure 4.2 we show the historical mortgage rates in the Netherlands based on new mortgage contracts with a maturity longer than ten years (the rates have been published by the DNB). On the same graph we see the 6 months EURIBOR swap rate denominated in euros. As can be seen from the graph the swap rate and mortgage rates have been decreasing steadily since 2011. Furthermore, we see that the mortgage rates have never been as low as August 2021.

4.2 Summary

In this chapter we give information about the data properties of the Allianz data set, on both the portfolio-level and loan-level. Furthermore, we explain what the

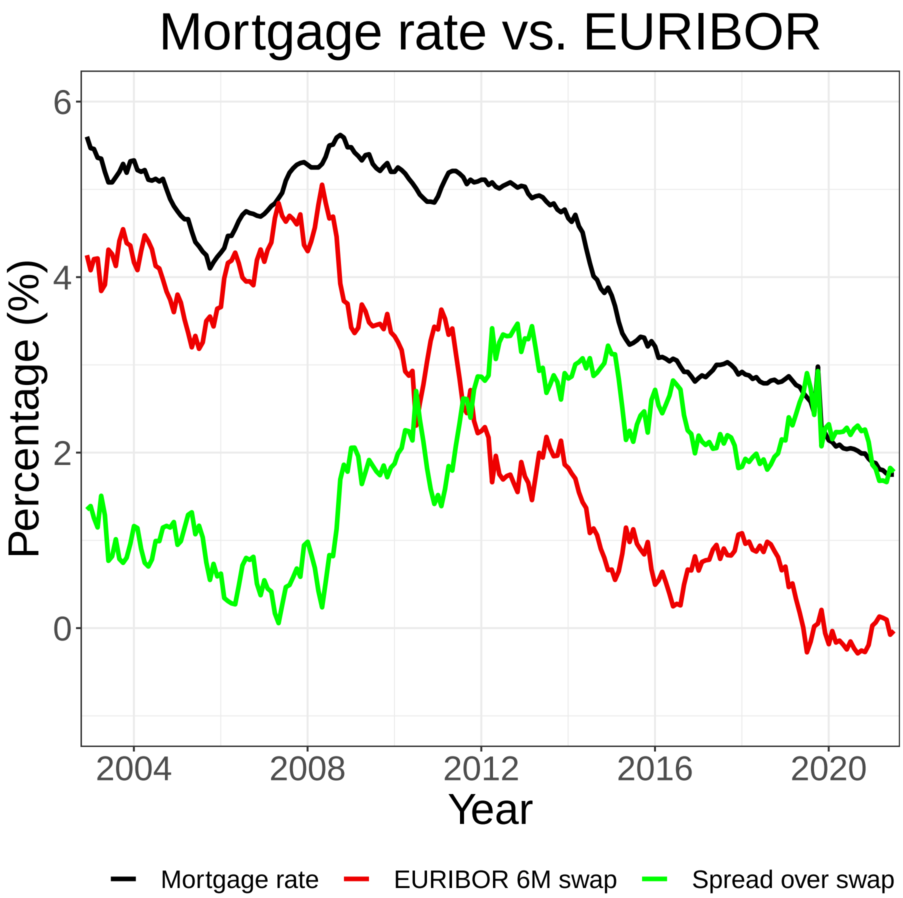

FIGURE 4.2: A graph showing the mortgage rate for all new contracts with a maturity longer than ten years, versus the EURIBOR swap rate over time. Added is the spread over swap which is taken by subtracting the swap rate from the mortgage rate.

conditional prepayment rate (CPR) is and how it is calculated. Lastly, information is shown on the number of prepayments within the Allianz portfolio.

Chapter 5

Preliminary Analysis

In this section we perform a preliminary analysis on the Allianz data that will be used for modelling. We look at the relationship between relevant determinants found in literature and the conditional prepayment rate. We further elaborate on data transformations performed, how we account for data imbalance and lastly show how we will be evaluating the results.

5.1 Explanatory variables

The Allianz data set contains many independent variables which can be subdivided into variables relating to loan characteristics, borrower characteristics or macroeconomic factors. With a clear overview of the portfolio statistics and an explanation of the conditional prepayment rate (CPR) this section will analyse the relationship between different determinants mentioned in literature and the CPR at Allianz, subdivided into the various forms of prepayment. This will be done through a visual aid in the form of line- and bar-plots whereby different types of prepayment will be analysed. Determinants not shown in this section will be placed in Appendix B.

Property Type

We compute a categorical variable for the various property types used as collateral by mortgagors. We distinguish between two types of classes, namely between apartment properties and non-apartment properties. Both apartments with and without garage are classified as apartment properties whereas non-apartment properties are all other possible collateral types. Collateral types that fall under non-apartment properties vary greatly and range from house boats to farms. However, nearly all of the non-apartment properties consist of family homes.
As mentioned in literature in section 2.2.2 by (Alink, 2002) (Charlier and Van Bussel, 2003) (Elsing, 2019) the property type is found to be an important determinant in forecasting the prepayment rate. This effect can also be found in the loan level data from Allianz, shown in Figure 5.1. Although the arbitrage and partial prepayments do not differ greatly between categorical variables, the CPR percentage of full prepayments for mortgagors living in an apartment is 2.5% higher. This suggests that mortgagors living apartments tend to move out and upgrade their homes.

FIGURE 5.1: CPR(%) for apartment properties versus non-apartment properties

Mortgagor Age

Many papers in literature also mention the mortgagors age as a determinant for prepayment. The mortgagor's age for contracts with two borrowers is calculated as the average age of the two. Figure 5.2 shows the relationship between the mortgagors age and the CPR. Although the total CPR increases for mortgagors age from 20 to 35 the CPR stagnates and does not keep increasing, except for the large spike at a mortgagor age of roughly 85 years old. The number of mortgagors in the Allianz portfolio that are older than 80 is relatively small. Furthermore, the remaining capital on the mortgage loans for these mortgagors is not very high as older mortgagors do not tend to start new mortgage loans. The reason for more full prepayments being made by older mortgagors could be related to mortgagors passing away.

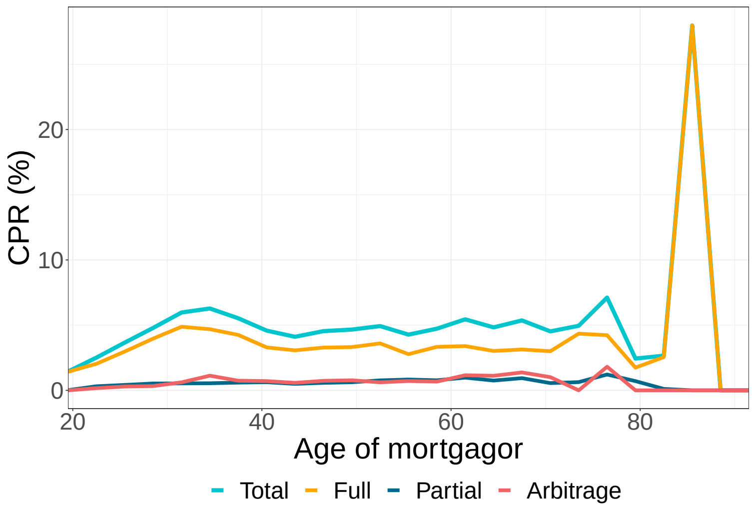

FIGURE 5.2: The CPR (%) for per age of the mortgagor within Allianz.

Upgrading Effect

The upgrading effect mentioned in 2.2.2 by (Charlier and Van Bussel, 2003) shows that young people in an apartment have an increasing probability of upgrading their property and hence prepaying their mortgage. Figure 5.3 shows that this effect is also somewhat present in the Allianz portfolio. We in fact see that the percentage of mortgage observations making a form of prepayment that are aged between 25 - 40 is higher for mortgagors living in an apartment than mortgagors living in 'other' property types. (Charlier and Van Bussel, 2003) mention that the upgrading effect is greatest for mortgagors in their mid-thirties and forties, which is somewhat the case for the Allianz mortgagors. From the figure it is clear that between the ages of 30 - 35 the percentage of prepayment observations has the biggest difference.

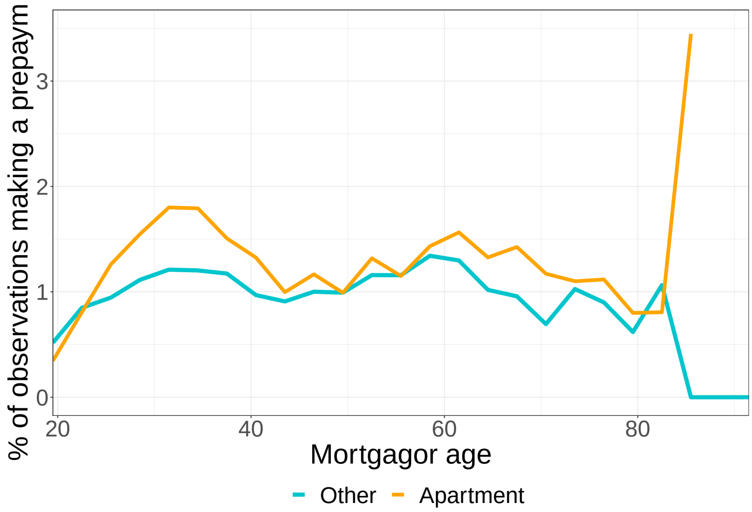

FIGURE 5.3: A graph depicting the difference in prepayment observations between mortgagors living in an apartment and in other property types per mortgagor age.

Initial Loan-To-Market-Value

We define the initial LtMV as the total mortgage principal divided by the market value of the property at loan origination. This is given by the following formula:

$$Initial\ LtMV = \frac{Value\ Total\ Principal}{Market\ Value\ Property} \tag{5.1}$$

The relationship between the intitial LtMV and the CPR is shown in figure 5.4. As can be seen, the CPR for mortgagors with an initial LtMV of 0.25 is nearly double that of mortgagors with a higher original LtMV. There are two possible reasons mentioned in literature (Elsing, 2019) (Sirignano, Sadhwani, and Giesecke, 2015) as to why mortgages with a lower initial LtMV have a higher CPR. One reason suggests that mortgages with a high initial LtMV have less opportunities to refinance due to the large loan outstanding, which would incur a higher refinancing penalty. The other reason suggests that mortgagors with a lower initial LtMV are more likely to have additional funds to their disposal, thus can bare the cost or prepaying. Regarding the peak CPR at a LtMV of 0.225, we see that it is caused by a higher CPR for partial prepayment. Looking at the data we find that in the number of mortgages with such a low initial LtMV are quite low relative to mortgages with a higher initial LtMV. Furthermore, the amount of partially prepaid balance is higher. However, as the numbers are so low, one extra partial prepayment or a larger partial prepayment will have a big influence on the partial CPR. As is this case for the LtMV of 0.225. The outstanding balance for observations with a LtMV of 0.275 is double that of the observations with a LtMV of 0.225, meaning it is less prone to changes in the prepayment amounts.

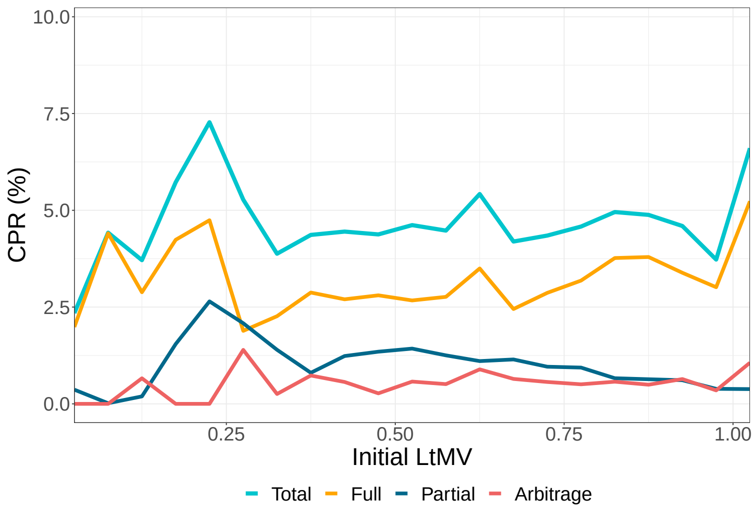

FIGURE 5.4: The CPR per initial LtMV at mortgage inception.

Seasonality

Literature defines seasonality to be mortgage prepayments made in the months November and December. This effect is captured using a dummy variable. The percentage of observations with prepayment per month are shown in figure 5.5, where it is obvious that the number of observations for partial prepayments is significantly higher in the months November and December. Although October is not regarded as an 'end of year' month the percentage of partial prepayments made in this month is also far higher than the months earlier in the year. Possible reasons for the seasonality effect are, as stated in section 2.2.2, the inflow of extra salary due to the '13th month' and the fiscally beneficial prepayment penalty being deducted from mortgagor's yearly income. Due to the high partial prepayment rate in October we include this month for the seasonality dummy.

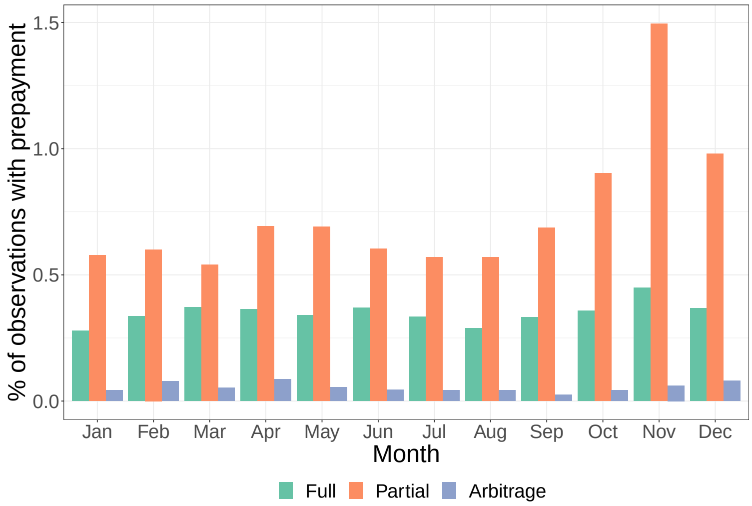

FIGURE 5.5: Percentage of data observations making a prepayment for each month of the year.

Refinancing Incentive

A very prominent prepayment determinant mentioned in literature is the refinancing incentive. Some even mentioning the refinance incentive as the most important prepayment determinant. In order to calculate the refinance incentive we must establish the market interest rate for each type of mortgage. The first step in this process is grouping the mortgages into buckets on the remaining fixed interest rate time to maturity (IRFP ttm) of the loan. These buckets are shown in the table 5.1.

IRFP ttm buckets	IRFP ttm (months)
60	$M \leq 90$
120	$90 < M \leq 150$
180	$150 < M \leq 210$
240	$210 < M \leq 300$
360	$M > 300$

TABLE 5.1

We define the market rate for a certain type of mortgage as the average rate of all the mortgages with the same redemption type and remaining fixed interest rate time to maturity. The refinancing incentive for each mortgage is then determined by subtracting the market rate of the previous month from the mortgage rate.
Figure 5.6 shows the percentage of observations with prepayment versus the refinance incentive. There is a clear positive relationship between refinance incentive and the number of prepayments. Although one would only expect prepayments to take place for a positive refinance incentive, prepayments are also made for a negative refinance incentive. We see that there is an arbitrage spike for observations with a refinance incentive of -0.7. Analysing the data shows that the are very few observations with a refinance incentive below -0.3. Digger deeper we see that there are just north of 100 observations with a refinance incentive of -0.7, where one loan made an arbitrage prepayment. This one arbitrage prepayment is the sole cause for

the red spike on the left side of figure 5.6. Although speculative, there could be external factors at play as to why this loan made an arbitrage prepayment, even though the interest rate would become higher. Such reasons could be due to illnesses, death or change in marital status. We see that as the refinance incentive increases so does the percentage of arbitrage prepayments. This is logical as mortgagors have a chance to refinance and take advantage of lower mortgage rates.
Furthermore, from figure 5.6 we see that that the percentage of partial prepayments is more or less independent from the refinance incentive. For mortgages with a high positive refinance incentive as for a negative refinance incentive we do not see very big changes. This is logical as the refinance incentive has no influence on the number of partial prepayments.
The percentage of full prepayments increases steadily with increasing refinance incentive. In the period in which Allianz has been issuing mortgages, they have only been in a decreasing interest rate climate. Reviewing the data we see that the loans with a higher refinance incentive have a higher loan age, suggesting that these loans were issued close in the beginning years of of the company. We therefore regard the loan age as the main cause for the percentage increase of full prepayments. As mentioned in subsection 2.2.2, literature states that the reason an increased loan age (mentioned as seasoning) leads to a higher number of full prepayments is because external factors such as the family composition, employment status and marital status, among other things, changes.
The last noticeable point in figure 5.6 is that there are mortgages with a refinance incentive of bigger than 4%. However, the number of observations with such a high refinance incentive is very small, being just lower than 100. As the refinance incentive increases further the number of observations decreases even further. We see that zero percent of these mortgagors made any sort of prepayment. If a mortgagor were to make a prepayment then this would immediately have a big impact for the curve.

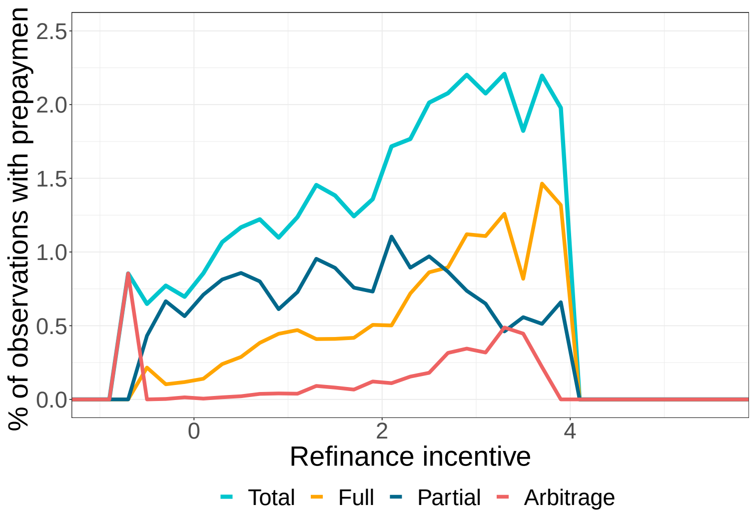

FIGURE 5.6: Percentage of data observations making a prepayment based on the refinance incentive (also called the delta rate) of the mortgage.

Burnout Counter

Mortgagors will for a certain period of time not be aware of the market rate is below their mortgage rate. Mortgagors that are slow to react when a refinancing incentive presents itself will be slow to prepay. The variable burnout counter measures the number of months the mortgage has been presented with a refinance incentive but has not taken advantage of the opportunity to prepay. Financial institutions mention that in the past the golden rule on when to refinance a mortgage was if the market rate was 2% below the mortgagors' rate. However, such institutions currently advise people to refinance if the market rate is 1% below the mortgagors' rate. Thus, we start the burnout counter if the market rate is 1% below the mortgage rate. This effect is also present in the Allianz portfolio. Figure 5.7 shows that once the burnout counter starts, many mortgagors are slow to react to the refinance incentive. The burnout counter has a upwards slope with the CPR, which is inline with literature.

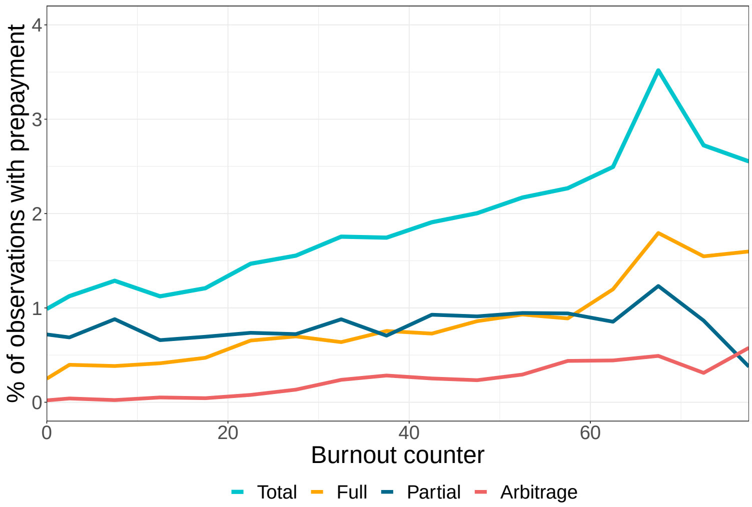

FIGURE 5.7: Percentage of data observations making a prepayment based on the burnout counter.

Seasoning

The last determinant that is discussed in section 2.2.2 is seasoning. This determinant looks at the CPR per loan age of each mortgage. Figure 5.8 shows the CPR of the portfolio based on the loan age. It is clear that the positive effect between the loan age and CPR. It must be noted that the outstanding principal does somewhat decrease at a larger loan age, thus every prepayment made will have a larger effect on the CPR. However, literature states that this relationship should produce an s-shaped curve. Which is not the case for the portfolio of Allianz. A possible reason for this is that the Allianz portfolio is still very young and has not matured. If we were to reproduce this plot in 20 years we should expect the relationship to resemble that of an s-shape.

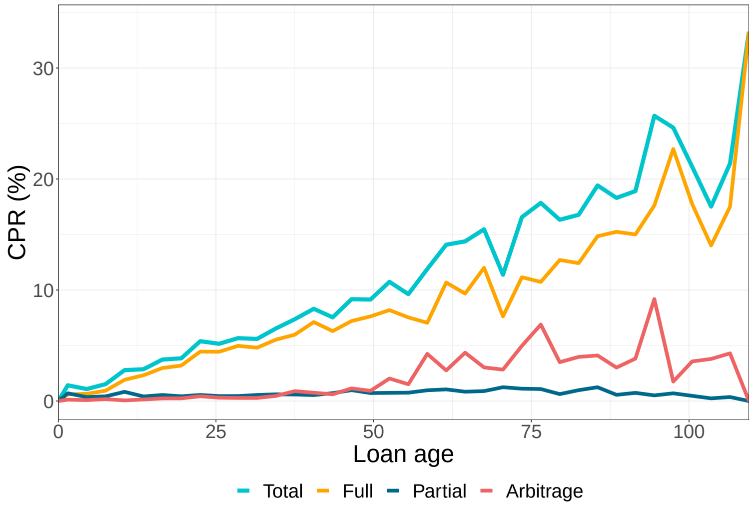

FIGURE 5.8: A graph showing the CPR per loan age, in months, of the mortgage.

5.1.1 Conclusion sub-question 2

2. What are the main drivers of prepayment when analysing the Allianz mortgage portfolio?

We analyse the different drivers of prepayment for the Allianz portfolio in section 5.1. Here we find that there multiple determinants for the three different types of prepayment are visible from the data.
We find that the main drivers for partial prepayments are the initial LtMV and seasonality. We see that for low initial LtMV levels that the partial CPR is at its highest point and gradually declines as the initial LtMV rises. Furthermore, we see that during the last months of the year the percentage of partial prepayments increases greatly, showing that seasonality has a big effect on the number of partial prepayments.
There are a number of determinants that, through analysis of the Allianz data, can be considered the main drives for full prepayments. These are the property type, mortgagor age, refinancing incentive, burnout and seasoning. We see that the full prepayment CPR for mortgagors living in an apartment property type is greater than that of mortgagors living in other property types. Secondly, we find that the mortgagor's age also has an effect on the full prepayment CPR, increasing from the ages 20-35, before decreasing and then staying roughly at the same level. For increasing refinancing incentive and burnout we see that percentage of full prepayment observations also increases. Lastly, the increasing loan age (seasoning) also drastically increases the full prepayment CPR.
As for arbitrage prepayment we find that the mortgagor age, refinancing incentive, burnout and seasoning are important drivers found in the Allianz data. We find that as with full prepayments the arbitrage prepayment CPR increases from the ages 20-35, before decreasing and then staying roughly at the same level. We see that both increasing refinancing incentive and burnout increase the percentage of arbitrage prepayments. Lastly, the increasing loan age (seasoning) also increases the arbitrage prepayment CPR.

5.2 Data Transformation

We apply a number of data transformations to variables that will be used by all three classifiers. This is done for two reasons. Firstly, to achieve symmetry in skewed variables and secondly, in order to achieve linearity between the independent variables and the logit of the dependent variable. From data analysis we know that the following variables are right skewed: total income, age, loan age, refinance incentive, total principal, remaining loan part principal. Whilst the variable 'number of houses sold' is left skewed. We transform the right skewed variables by taking the log and the left-skewed variable by taking the square root. Lastly we transform the 'refinance-incentive' variable. Instead of calculating the refinance-incentive by subtracting the market rate from the mortgage rate we take the logarithm of the mortgage rate divided by the market rate (LaCour-Little, Marschoun, and Maxam, 2002).
We take the logarithm of these variables as this is an easy and common technique for transforming right-skewed data. Through the logarithm we 'pull in' more right tailed data relative to the median and transform the data to a more normal distribution. One can imagine this is particularly useful for variables such as 'total income' where most mortgagors have similar total incomes, yet there are enough mortgagors that have a total income of multiple times the amount of the median total income. We take the square root of left-skewed variables as this has somewhat the opposite effect of the logarithm, where smaller numbers are inflated and larger numbers become more stabilised.

5.3 Data Splitting

In order to develop and evaluate a good classifying model we split the total data set into multiple smaller ones, whereby each data set serves a different purpose. The first data split that was performed was to create an 'in-sample' data set and an 'out-of-sample' data set. The in-sample data set contains all observations except for latest two months which are June 2021 and July 2021. Thus the in-sample data set contains observations from June 2012 up to and including May 2021, while the out-of-sample data set only contains the observations from June and July 2021.
We perform a second split on the in-sample data, splitting this data set into a training data set and test data set. The training set contains 80% of the in-sample data while the test set contains the other 20%. We perform this split on each reporting month, meaning that 80% of the observations each month are subset into the training set while the other 20% is subset into the test set. This split was chosen due to two main reasons, firstly because it is a common split found in literature and secondly because we require a large training set. We require a large training set as this training set has multiple purposes. Other than being used for training the each model, the training set will also function as 'validation' set for k-folds cross validation. Furthermore, the training set will also be used for undersampling, which is explained in subsection 5.4. We show the number of observations for each data set in table 5.2. Furthermore in the table we show the number of partial, full and arbitrage prepayments made in each data set.
Instead of creating an extra separate validation set (a set used for tuning model-specific (hyper)parameters) we opt to use k-folds cross validation instead. In doing so the 'validation' set in k-folds cross validation is also used for training the model, thus more training data can be used to train the model which in turn is considered advantageous. In k-folds cross validation we split the training set into k groups, also called folds, of roughly the same size. We leave the first fold out, train the

model parameters on the other *k - 1* folds and then validate the model with tuned hyperparameters on the fold left out. The error rate is then computed for the left out fold. We repeat this process *k* times and average out the *k* error rates. The error rate for the left out fold is computed as the number of misclassified observations and is given in the formula below:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i \tag{5.2}$$

where $Err_i = I(y_i \neq \hat{y}_i)$. If we were to withhold observations and partition the data further into a validation set, then the observations from the validation set (common validation set sizes in literature comprise of 15% of all observations) would never be used in training the model, which could be considered 'wasteful'. Although we have millions of data observations, when undersampling the data we reduce the number of observations greatly. For example, the undersampled arbitrage prepayment training set where 50% of the observations are arbitrage prepayments (which can also be called the 50% arbitrage prepayment undersampled data set) only contains a total of 1400 observations. With such a low number of observations every observation counts. For such situations k-fold cross validation is more efficient at using and re-using the data than constructing a validation set.

Data set name	Partial (#)	Full (#)	Arbitrage (#)	Total (#)
Train	9,187	4,243	701	1,239,681
Test	2,245	1,016	154	309,753
Out-of-sample	767	542	56	114,551

TABLE 5.2: A table showing the number of observations present in each data set for each prepayment type.

5.4 Imbalanced Data

As shown in table 4.3 there have close to 19,000 observations in which a prepayment occurs. Although this may seem like a lot, we have over 1.6 million data observations in the whole data set. Thus, a prepayment occurs in just over 1% of observations. The relative class imbalance between observations with and without a prepayment leads to a skewed class distribution. This is problematic as classifiers tend to ignore the minority class and pick up the patterns of the majority class. This causes these classifying models to become biased towards the majority class. The imbalanced data problem is not at all uncommon and has elaborately been discussed and analysed in literature (He and Garcia, 2009). In literature we find an abundance of papers not only analysing the imbalance problem but also reviewing and proposing solutions to the problem (Kotsiantis, Kanellopoulos, and Pintelas, 2005) (King et al., 2001) (Sun, Wong, and Kamel, 2009) just to name a few. There are two popular types of solutions to the imbalanced data problem, namely methods implemented during the pre-processing phase on the data level and methods implemented on an algorithmic level. Most papers agree that both approaches are equivalent in addressing the imbalance problem (López et al., 2012).
The two most popular methods in minimizing the data imbalance on a data level involve a type of undersampling or oversampling technique. Whereas on an algorithmic level some type of cost-sensitive learning is applied. Each method has its advantages and disadvantages and analysing each available method would be a

thesis on its own. We propose two methods to combat the data imbalance, one on the data level and one on the algorithmic level. These methods have been chosen due to their interpretability, easy implementation and popularity.
The data level method to reduce the class imbalance will be done using an undersampling technique. The main idea behind undersampling is to balance the class distribution by randomly eliminating the number of majority class observations. This will in turn balance the data and reduce the relative imbalance between the minority and majority class, if-not eliminate it altogether. There are two main downsides to undersampling. Firstly, by eliminating majority class observations we discard potentially useful data and secondly, by undersampling we alter the probability distribution and shift it towards the minority class. During undersampling we alter the prior distribution of the training set resulting in a biased posterior probability of the classifying algorithm. This problem is called a prior probability shift and can be corrected, as has been shown in literature (Dal Pozzolo et al., 2015). This correction will be implemented in this thesis and will be further elaborated below.
The following paragraphs show how to acquire the unbiased probability, p, in order to correctly classify the unbiased probability after undersampling. Assume we have a random binary variable s that takes the value of 1 if the observation is in the new balanced (or undersampled) training sample and 0 otherwise. Furthermore we assume that s is independent from input x given the class y: $p(s|y,x) = p(s|y)$, which in turn implies $p(x|y,s) = p(x|y)$ meaning that removing observations at random will not affect the within-class distributions. For each observation in this thesis we denote y being a prepayment as $y = 1$ and no prepayment as $y = 0$. Thus $y = 1$ represents the minority class and $y = 0$ represents the majority class. Through Bayes Rule and using $p(s|y,x) = p(s|y)$ we can write the following equation:

$$P(y=1|x,s=1) = \frac{P(s=1|y=1)P(y=1|x)}{P(s=1|y=1)P(y=1|x) + P(s=1|y=0)P(y=0|x)} \tag{5.3}$$

With the knowledge that $p(s=1|y=1) = 1$ we can rewrite equation 5.3 as follows:

$$p(y=1|x,s=1) = \frac{p(y=1|x)}{p(y=1|x) + p(s=1|y=0)p(y=0|x)} \tag{5.4}$$

As shown in (Dal Pozzolo et al., 2015) we denote $\beta = p(s=1|y=0)$ as the probability of choosing a negative observation with undersampling. We further denote $p = p(y=1|x)$ as the posterior probability of the positive class based on the original data set and $p_s = p(y=1|x,s=1)$ as the posterior probability of the positive class after sampling. We can now rewrite equation 5.4 as follows:

$$p_s = \frac{p}{p + \beta(1-p)} \tag{5.5}$$

We can rewrite equation 5.5 to get an expression for p:

$$p = \frac{\beta p_s}{\beta p_s - p_s + 1} \tag{5.6}$$

Which is the posterior probability of the positive class on the original data set as a function of the posterior probability of the positive class after sampling, p_s. β has a lower bound of $\frac{N+}{N-}$, where $N+$ and $N-$ indicate the number of positive and negative observations in the data set. The upper bound for β is 1. β indicates the degree of undersampling being done, where $\beta = 1$ means all negative observations are used

in the training sample and $\beta < 1$ indicates that a subset of the negative observations are used in the training sample.
In order to retain the classification accuracy of the model we also have to adjust the probability threshold, otherwise we would have different misclassfication costs between the imbalanced and balance data models. From (Elkan, 2001) we let τ be the threshold for the unbiased probability p.

$$\tau = \frac{\beta \tau_s}{(\beta - 1)\tau_s + 1} \tag{5.7}$$

Where we denote τ_s as the threshold used to classify an observation after undersampling. To summarise, due to undersampling a higher percentage of observations are predicted as positive. However, undersampling has made the posterior probabilities biased due to a change in the prior distribution. By obtaining p, found in equation 5.6, as the unbiased probability after undersampling and by using the threshold τ as the classification threshold we can correctly classify the unbiased probabilties after undersampling.
We undersample the training data set multiple times to observe the effect of undersampling on the loan-level and portfolio evaluation metrics. Not only do we undersample the data set separately for each prepayment type, we undersample it so that the minority class (the prepayment class) accounts for 10%, 20%, 30%, 40% and 50% of the total number of training observations. We then use these undersampled data sets as training data for each model. Thus, for each prepayment type we have five undersampled data sets. Using the different undersampled data sets as training sets we can evaluate the effect of the data set size for the different models. Table 5.3 shows the number of observations per training set. The first row header indicates that 10% of the observations account for the minority class (the prepayment class). For example, the 30% full prepayment training set has a total of $14,175$ rows with $4,253$ of those rows being of the class full prepayment. This represents 30% of the total observations of that training set.

Minority class presence (%)	Partial prepayment data set	Full prepayment data set	Arbitrage prepayment data set
10%	91,955	42,546	7,043
20%	46,021	21,271	3,504
30%	30,667	14,175	2,348
40%	23,008	10,657	1,769
50%	18,265	8,428	1,387

TABLE 5.3: A table showing the total number of observations for each training set for each different degree of undersampling. The row header indicates which percentage of the training set the minority class represents.

5.5 Model Evaluation

We perform the model evaluation using R and the packages available within this environment. R was chosen because it can easily be used for both manipulating data sets and programming. We distinguish between two models for each machine learning algorithm; the baseline model and the improved model. The baseline model for each classifier will use the training data set as input data and will be tested using

the in-sample test data and the out-of-sample data. As mentioned in subsection 5.4 this training set is highly imbalanced. The improved model for each classifier will use the various undersampled data sets as input and identical to the baseline model will also be tested on the in-sample test data and the out-of-sample data. Thus we will forecast different types of prepayment on in-sample test data and out-of-sample data.
We can evaluate each model in two separate ways, through loan-level metrics and secondly through portfolio metrics. On the loan-level we evaluate each classifier on its forecasting power whilst on the portfolio level we evaluate the overall difference between the actual CPR and the predicted CPR. We prioritize the portfolio level results due to their importance for Allianz. The following section will go into greater detail on these metrics.

Portfolio Level

Predicting the CPR per month for the portfolio is important. We calculate the CPR by multiplying the forecasted probabilities per prepayment type from each classifier by each specific prepayment factor. From this we can easily calculate the SMM and thus also the CPR. The CPR for each model can be evaluated by comparing it to the actual CPR. There are several metrics that can be used to compare the predicted CPR from each classifier and the actual CPR.
The root mean squared error (RMSE) is a widely adopted metric that calculates the square root of the average squared difference between the predicted values and the actual values. Although useful, it is prone to outliers. The formula gives more weight to outliers further from the mean. The formula is given below.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2} \tag{5.8}$$

The mean absolute error (MAE), on the other hand, is robust to outliers. It measures the average magnitude of the errors regardless of its direction and gives all individual differences an equal weight.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \tilde{y}_i| \tag{5.9}$$

We want to further extend on these metrics by taking into account the large increase in mortgages from the past few years. We alter the RMSE and MAE metrics into weighted RMSE (WRMSE) and weighted MAE (WMAE), these are shown below.

$$WRMSE = \sqrt{\sum_{i=1}^{N} w_i(y_i - \tilde{y}_i)^2} \tag{5.10}$$

$$WMAE = \sum_{i=1}^{N} w_i|y_i - \tilde{y}_i| \tag{5.11}$$

Where w_i is the weight given to every month with $\sum_{i=1}^{N} w_i = 1$. The monthly weight is calculated by the number of observations for that month divided by the total number of observations. With this weighted scheme the errors of the monthly CPR from the past three years will be given a larger weight than those in 2012-2014. This is done because the mortgage portfolio from 2012-2014 was very young and thus very prone to outliers.

Loan level

Although we evaluate the three classifiers on portfolio level we also look at the performance on a loan level, by using different metrics. Each metric serves a different purpose and can help us find out how well our model is performing. On the loan level we have a classification problem and would like to forecast whether or not an observation belongs to one of the prepayment classes or not. We classify an observation as being a prepayment if the forecasted probability is higher than a certain threshold. By making use of a confusion matrix we will be able to count the correct and incorrect predictions per class. A confusion matrix is a visual table that shows the prediction results of our prepayment classification problem). From the confusion matrix we obtain four values, namely the number of True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN). The TP count value is the number of observations that were predicted by the classifier to be part of the positive class (1) and were actually part of the positive class (1). FP is the value for the number of observations that were predicted to be part of the positive class but are actually part of the negative class. The same metrics are counted for the negative class, the TN is the number of observations predicted as part of the negative class and are actually part of the negative class. The FN is the number of observations predicted to be of the negative class but were actually part of the positive class. The main outline of the confusion matrix is shown in figure 5.9.

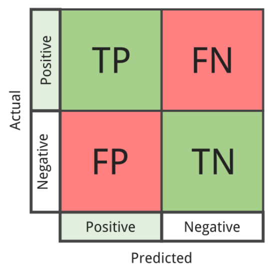

FIGURE 5.9: The confusion matrix gives the number of true positives, false positives, true negatives and false negatives.

The confusion matrix lays the basis on which many other evaluation metrics can be derived. One of these metrics is the accuracy and is calculates as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.12}$$

Although this metric gives a good overview of the number of correctly classified observations out of the total number of observations, it it has its limitations when using imbalanced data. If we were to keep the imbalanced data, the cases of prepayment to no prepayment would be roughly 1:99. If the model were to accurately classify all the observations as non-prepayment the accuracy of the model would be 99%,

nearly perfect. However, the aim is to predict if mortgagors will prepay and thus this classifier would be deemed useless.
The precision metric looks at how many true positives the classifier predicted out of all predicted positives. However, this metric doesn't give us any information on the false negatives. The formula for precision is shown below.

$$Precision = \frac{TP}{TP + FP} \tag{5.13}$$

To account for the lack of information on the false negatives we use the recall metric. This metric tells us the number of positive observations correctly identified out of the total number of actual positive cases. Recall and precision have an inverse relationship, if recall increases the precision will decrease. The formula for recall is shown below.

$$Recall = \frac{TP}{TP + FN} \tag{5.14}$$

Specificity, also known as the true negative rate, gives us more insight into the negatives cases. It is calculated by the number of true negatives out of all actual negative cases. The formula is shown below.

$$Specificity = \frac{TN}{TN + FP} \tag{5.15}$$

From the precision and recall metrics we can calculate the F1 score. The F1 score, is the trade-off between precision and recall and acts as a harmonic mean between the two metrics. For this research we would like a high F1 score.

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.16}$$

Based on the true positive rate (which is synonymous for the recall metric) and the false positive rate (FPR = 1 - Specificity) we can construct a Receiver Operating Characteristic curve (ROC curve). Which is a useful curve for calculating the area under the curve (AUC) metric. In essence a ROC curve is constructed by measuring the true positive rate and the false positive rate for many thresholds. However, doing so would be time consuming work. Luckily constructing a ROC curve can easily be done with most programming packages.
We mention that the AUC can b calculated form the ROC curve. This is done by, as the name suggests, calculating the area under the ROC curve. The AUC metric can be interpreted as the probability that the model ranks a randomly chosen positive sample (in which a prepayment takes place) higher than a random negative sample in which no prepayment takes place. The AUC can range between 0 and 1 where a higher score is better. For example, a prepayment classifier with an AUC of 0.80 will be able to distinguish between observations with and without prepayment 80% of the time.
In addition to the AUC value the Brier score can be used as a supplement. The Brier score looks as how close the predicted probability is to the actual case. The lower the Brier score the better the prediction from the classifier. The formula for the Brier score is given below, where N is the number of observations, f_t is the predicted probability per observation and o_t is the actual outcome of the observation where 1 would mean a prepayment and 0 no prepayment.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{5.17}$$

5.6 Prepayment Cash Flow Estimation

Apart from testing the models on loan-level metrics and portfolio level metrics we also estimate the prepayment cash flow from the various constructed models. By doing so we can compare the cash flow estimation of the Allianz model and our models to the actual prepayment cash flow of the portfolio.
The actual cash flows for the mortgage portfolio is the addition of the cash flows from all partial prepayments, full prepayments and arbitrage prepayments. In each case we have a probability of making a prepayment, which in the case of the Allianz mortgage portfolio is either 0 or 1 for each type of prepayment. Furthermore, only one type of prepayment can take place per data observation. The amount prepaid for a partial prepayment is a certain percentage of the outstanding remaining capital. We previously mentioned that up to 10% of the remaining capital can be partially prepaid per year penalty free, hence the percentage of partially prepaid capital is usually lower than this amount. This differs from full and arbitrage prepayments where the total outstanding capital is fully prepaid. The cash flow prepaid per month for each mortgage j in month t is shown in the formula below:

$$\begin{aligned} CF_{j,t} = {} & [P(partial) * P(CP_{j,t}|P(partial) = 1) * OC_{j,t}] \\ & + [P(full) * OC_{j,t}] + [P(arbitrage) * OC_{j,t}] \end{aligned} \tag{5.18}$$

Where $OC_{j,t}$ is the outstanding capital for mortgage j at time t. Summing for all mortgages per month gives the total actual prepaid cash flow. This equation can also be used to estimate the cash flows from each constructed model. When doing so we alter two components of the equation 5.18, namely the probability of each type of prepayment and the amount prepaid in case of a partial prepayment. Each constructed model estimates a probability of prepaying for each type of prepayment, which will be used in the cash flow estimation. This probability is between 0 and 1 instead of having only possible values of 0 or 1. Furthermore, we compute the percentage of average partial prepayment in relation to the outstanding capital. We sum all yearly prepayment cash flows of the best performing constructed models (chosen on the basis of the portfolio metrics) and compare them to that of the Allianz model and the actual observed prepayment cash flows. We compute the error of each model relative to the observed cash flows and give the results per year. We use the same time range as was mentioned in subsection 5.3. However, we exclude the cash flows from the years 2012 and 2013 as the actual prepaid amounts from these years are very small which lead to unreliable estimations from the models and therefore also unreliable results. Lastly, we do not distinguish between the in-sample and out-of-sample time range in the results.

5.7 Summary

In this chapter we analyse the main drivers of prepayment present in the Allianz mortgage portfolio. We find multiple drivers present in the Allianz data that have either a positive of negative relationship with each of the three prepayment types.

This allows us to conclude the second sub-question of this thesis. Furthermore, we elaborate on how the different training or testing samples will be split from the over data available. Information is then given on our answer to the data imbalance problem, namely undersampling. We give theory into how undersampling effects the sample probability distributions and show which corrections will be implemented in this thesis to account for this. Moving on, we elaborate on the different portfolio level and loan level metrics that will be used in order to determine the performance of each model. Lastly, we give the formula for the prepayment cash flow estimation which will help deduce which model is closest to replicating the actual observed prepayment cash flow.

Chapter 6

Results

This section will elaborate on the results for each of the three different types of classification models, namely the logistic regression, random forest and the neural network. For each of these different classifiers we analyse a multitude of metrics in order to assess the performance of each model on each of the prepayment types.
In the first part of the results we model and forecast for each of the three prepayment types separately using the binomial logistic regression, random forest and neural network. All the results will be written following the same order for each different machine learning model, whereby we start with the partial prepayment forecasting, followed by full prepayment forecasting and end with arbitrage prepayment forecasting. For each of the prepayment models we firstly use the imbalanced training data, which as explained in section 5.4 is highly imbalanced towards the non-prepayment classes. The models trained on the imbalanced training data will be called the baseline models. The baseline models will then be tested on two different data sets, which as elaborated in subsection 5.3 are the in-sample test data and the out-of-sample test data. Both of these sets have the same degree of imbalance as the training set.
In order to examine the effect of correcting for the data imbalance problem, we retrain all these models for each prepayment type with the undersampled training data. These models that are trained on the data sets with the various degree of undersampling will be called the 'improved models'. As explained in section 5.3, the prepayment type in question for each model will account for 10%, 20%, 30%, 40% and 50% of the total training observations. The improved models using a different degree of undersampling will be called the 10%-, 20%-, 30%-, 40%- and 50%-undersampled models. However, the same in-sample and out-of-sample testing data sets will be used for forecasting as with the baseline model. Although the improved models will be trained on a more balanced data set, they will again be tested on the highly imbalanced in-sample and out-of-sample test data set.
Furthermore, the variables for both the baseline and improved logistic regression models are chosen using best subset selection and are chosen on the basis of the lowest Akaike Information Criterion (AIC) score. By using best subset selection we only use the variables that are statistically significant to train and test the model.
We use the same 19 variables as input for every classifier. The used variables were chosen based on the literature review performed in section 2.2.2, on the condition that enough data for each variable was available. In addition to the variables found in the literature review we add extra variables which we believe may help in predicting the prepayment rate. Table 6.1 shows all 19 variables used by all models. For a detailed description of each variable we refer to Appendix C.

ID	Determinant	Data type
1	Property Type	Dummy
2	Mortgagor Age	Numeric
3	Initial LTV	Continuous
4	Loan Age (Seasoning)	Numeric
5	Refinance Incentive	Continuous
6	Seasonality	Dummy
7	Geographical Location	Dummy
8	Burnout	Numeric
9	Starting Loan part Principal	Numeric
10	House Price Index Ratio (HPI-ratio)	Continuous
11	Marital Status	Dummy
12	Risk Class	Dummy
13	Remaining Loan Part Principal	Continuous
14	Income	Continuous
15	Dutch	Dummy
16	LTI	Continuous
17	Number of Houses Sold	Numeric
18	Unemployment Rate	Continuous
19	Redemption Type	Dummy

TABLE 6.1: Table showing the determinants that will be used by all models

During the second half of the results we show the results of the prepayment cash flow estimation for all constructed models and that of the benchmark Allianz model. We compare these results with the observed prepayment cash flows from the years 2014 until present. We look at the error between the actual and forecasted prepayment cash flows on a yearly basis and furthermore look at the total error of the total prepayment cash flows over the whole time range of 2014 until present.

6.1 Logit Model

6.1.1 Baseline Logit Model

For each of the three types of prepayment, namely partial, full and arbitrage prepayment, a separate binomial logistic regression was conducted after best subset selection had taken place. All models are tested on both the in-sample and out-of-sample data.
Below, we show the coefficients for all used predictor variables for the baseline partial, full and arbitrage prepayment models. The coefficients show the log odds change in mean partial, full and arbitrage prepayment for each increased unit of the dependent variable. We note that all variables have been standardized before regression, which is useful for interpreting numeric variables. Furthermore we acknowledge that binary variables can only increase with one unit, no more.
From table 6.2 we see that the variables 'original capital', 'loan age', 'HPI ratio', 'total income' and 'seasonality' all have a positive log-odds of increasing the mean partial prepayment, with the HPI ratio having the highest odds. When reviewing the relationship between the HPI-ratio and the partial prepayment CPR we find that for a higher HPI-ratio the partial CPR increases greatly, nearly doubling for larger HPI

ratios. A positive log-odds for the variable 'original capital' makes sense as mortgagors with a higher mortgage loan are mostly likely well off and therefore have extra funds to make a partial prepayment. The positive relationship between seasonality' and the partial prepayment rate is not unexpected considering the increased number of partial prepayments in the November and December, as shown in figure 5.5. Furthermore, we see that the variables 'initial LTMV', 'delta rate' (also known as refinance incentive), 'mortgagor age', and 'LTI' all have a negative log-odds with partial CPR. From figures 5.4 and 5.6 we see that for higher values for these variables the partial CPR decreases. Thus, the coefficients are in accordance with Allianz portfolio data analysis. The negative log-odds for the initial LTMV is also in line with literature, stated by (Alink, 2002) (Sirignano, Sadhwani, and Giesecke, 2015).

Reviewing the coefficients for the full prepayment baseline model, found in table 6.3, we find that the variables with the highest positive log-odds are 'delta rate' and 'loan age'. From figures 5.8 and 5.6 it is evident that the full prepayment CPR increases drastically as the refinance incentive or the loan age increases. This is not illogical as the longer the mortgage existed, the more probable external factors such as the family composition, job location and marital status could change, which could lead to a full prepayment. A possible reason for the increased CPR for full prepayment for mortgagors with a higher refinance incentive could be because this refinance incentive gives mortgagors an extra reason to move houses. It may be extra appealing for mortgagors to move houses if mortgage rates are much lower than the one they are currently paying.

We see that the variable 'mortgagor age' has the highest negative log-odds, which is mostly what was seen during data analysis in figure 5.2. We see that as the age of mortgagors increases that the CPR for full prepayment decreases, with the exception of mortgagors that are older than 80 years old. After the the age of 35-40 years old mortgagors have the family composition and the house they want and are able to afford and often don't want to move houses. This leads to a decrease in CPR for full prepayments. This story changes after the age of 80 where there is a high chance that the massive peak has to do with mortgagors either moving to nursing houses or because of death. Both of these actions lead to a full prepayment.

Lastly, we analyse the log-odds for arbitrage prepayment, as found in table 6.4. We see that the refinance incentive has a very high log-odds compare to the other variables. This is not strange as literature mentions the refinance incentive to be one of the most important variables (if not the most important variable) for arbitrage prepayments. This relationship was also found during our data analysis, where larger refinance incentives lead to a larger percentage of mortgagors making a arbitrage prepayment.

	Estimate	z value	Pr(>|z|)
original capital	0.82	21.12	0.00
initial LtMV	-1.38	-20.37	0.00
delta rate	-1.36	-12.07	0.00
loan age	0.66	13.99	0.00
total income	0.50	5.60	0.00
age	-0.96	-9.62	0.00
married	-0.26	-10.43	0.00
LTI	-0.32	-19.00	0.00
HPI ratio	1.08	8.20	0.00
seasonality	0.59	27.04	0.00

TABLE 6.2: Coefficient results for the baseline partial prepayment logit model.

	Estimate	z value	Pr(>|z|)
original capital	-0.95	-11.39	0.00
remaining capital	0.25	3.56	0.00
delta rate	2.22	9.77	0.00
loan age	1.20	13.27	0.00
m count burnout	-0.01	-3.78	0.00
property	0.39	10.10	0.00
total income	0.37	2.74	0.01
age	-1.87	-12.45	0.00
married	-0.27	-6.97	0.00
LTI	-0.12	-4.98	0.00
HPI ratio	0.75	3.95	0.00
unemployment rate	-0.08	-4.22	0.00
seasonality	0.22	6.30	0.00

TABLE 6.3: Coefficient results for the baseline full prepayment logit model.

	Estimate	z value	Pr(>|z|)
original capital	-0.25	-2.00	0.05
delta rate	5.28	12.52	0.00
loan age	1.10	4.73	0.00
total income	2.12	7.26	0.00
age	1.32	3.63	0.00
LTI	0.17	6.88	0.00
unemployment rate	-0.48	-7.95	0.00
seasonality	0.21	2.47	0.01

TABLE 6.4: Coefficient results for the baseline arbitrage prepayment logit model.

Figure 6.1 shows the predicted and actual CPR throughout the years for all three prepayment types of the baseline logistic regression model. For the actual observed partial CPR and predicted CPR we see that both lines resemble the same shape however, the predicted CPR of the logit model underfits the data. Although the model is able to capture the yearly seasonality component present at the end of the year, the model is not able to capture any other differences in CPR throughout the year. The model is able to somewhat find the general pattern for the full prepayment CPR over time, it again underfits the data significantly. It is understandable that this happens as the actual full CPR itself looks to be quite noisy. However, looking at the forecasting for the arbitrage prepayment CPR the model has a hard time forecasting the observed arbitrage prepayment CPR. A reason for this could be that of all the three arbitrage forms, this form takes place the least amount of times and therefore the training data is the most imbalanced for this prepayment type. The portfolio level metrics accompanying the figure 6.1 are shown in table 6.5.

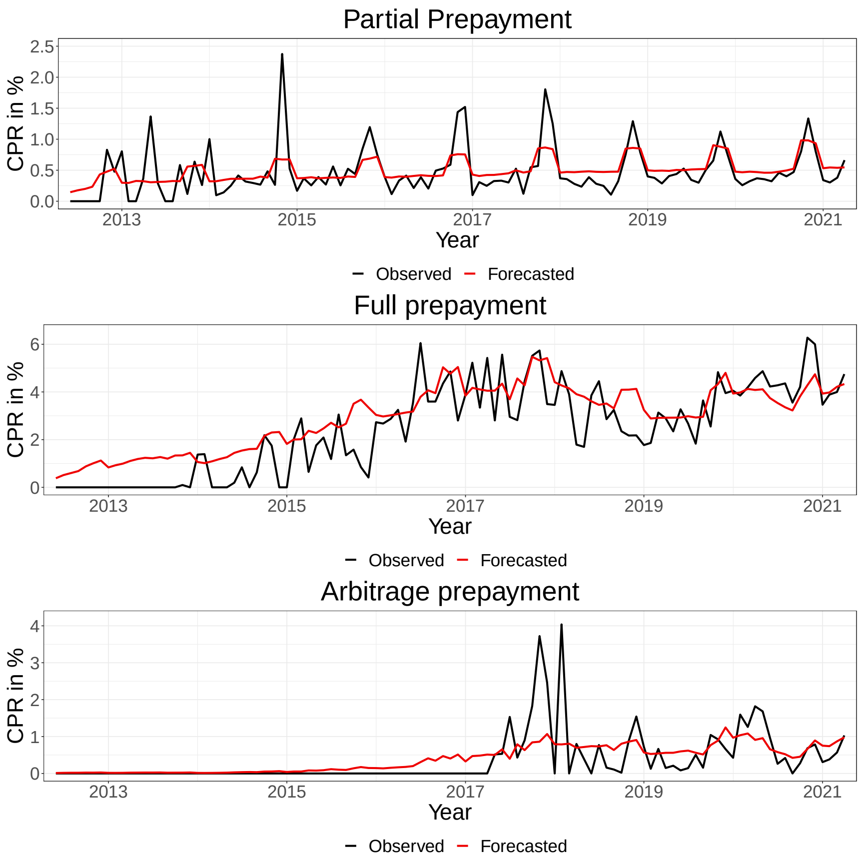

FIGURE 6.1: Plot showing the partial, full and arbitrage CPR over time for the baseline logistic regression model.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
Partial	0.317	0.214	0.231	0.172
Full	1.126	0.914	0.988	0.767
Arbitrage	0.558	0.301	0.538	0.361

TABLE 6.5: Table showing the portfolio metrics for the in-sample test data of the baseline logit model for all types of prepayment.

We do not show the loan-level metrics for the baseline model as no prepayments of any kind were predicted, leading to a very high accuracy (as only non-prepayments were predicted) and the precision, recall and F1 metrics being zero.

Out-of-sample results

We further test the performance of our model on 'out-of-sample' data which contains two subsequent months that are not present in the training set. As mentioned in section 5.3 these months are June and July 2021. The portfolio level results are shown in table 6.6. Examining the metrics we see a lower deviation in the results when compared to the in-sample results. However, it is the two most recent months of the entire data set which means it is also the most matured. In turn this would make these two months easier to forecast, which would explain the lower portfolio errors. Looking at the portfolio metrics for the partial prepayment, we see that they are extremely low. This is because the two out-of-sample months are June and July which, when looking at figure 6.1 dont exhibit much deviance as there is no seasonality component. The 'out-of-sample' portfolio performance will mostly be used to compare the different classifier models with one another.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
Partial	0.026	0.025	0.026	0.025
Full	0.654	0.647	0.656	0.649
Arbitrage	0.505	0.505	0.505	0.505

TABLE 6.6: A table showing portfolio level metrics of the 'out-of-sample' results of the baseline logistic regression.

6.1.2 Improved Logit Model

For the improved logit models we undersample the original training set specific to each of the three prepayment types, so that the minority (prepayment) class takes up a larger percentage of the training set. As mentioned in section 5.4 we get five different undersampled training sets for each prepayment type which will be used for their respective partial, full, and arbitrage prepayment logistic regression models. By using the undersampled data sets we observe the effect this has on the model performance. As mentioned in section 5.4 we use a different number of more balanced training sets, using undersampling, best subset selection, and correcting for the bias made by this process. We look at the effect of this per prepayment type and and show the results. Additionally we also look at the loan-level metrics for these improved models.

Partial Prepayment

As we train the model on various different undersampled data sets we do not see a substantial improvement in the portfolio metrics for partial prepayment on the in-sample test set. Looking at results in table 6.7 we find that only the model trained on the 50% undersampled set performs better than the baseline model, but only fractionally. As for the out-of-sample portfolio scores we see the opposite pattern. Nearly all models perform better than the baseline model, with the exception of the model trained on the 50% undersampled data. The models trained on data sets with a higher degree of undersampling somewhat overfit the data more than the models trained on lower degrees of undersampling, leading to higher errors in out-of-sample data. Here the model trained on the 10% undersampled data has the lowest errors. We show the results from this model in table 6.8.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
10 percent	0.318	0.218	0.228	0.164
20 percent	0.322	0.220	0.232	0.164
30 percent	0.323	0.218	0.234	0.165
40 percent	0.322	0.215	0.233	0.164
50 percent	0.316	0.214	0.225	0.159

TABLE 6.7: A table showing the portfolio metric results for the undersampled data sets for partial prepayment.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
10 percent	0.019	0.019	0.019	0.019

TABLE 6.8: A table showing the portfolio metric results for the out-of-sample undersampled data sets for partial prepayment.

Reviewing the loan-level metrics of the partial prepayment model trained on the 50% undersampled data set we see that these metrics are not are overall not spectacular. Table 6.9 shows the loan-level metrics just for the 50% undersampled model while the loan-level metrics for all other partial prepayment models trained on the various undersampled data sets can be found in Appendix E.2. From table 6.9 we see that we have normal scores for accuracy and specificity and very low scores for the precision metric and a somewhat low score for the recall. Regarding the low precision and recall scores, we see that the model predicts a larger false positive score than a true positive score and an even larger false negative score than the true positive score. As the F1-score is based on a combination of the precision and recall scores this metric is also very low. We reflect on this in the section 6.4.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
50 percent	0.717	0.011	0.590	0.726	0.011	0.648	0.0001

TABLE 6.9: A table showing the loan metric results for model trained on the 50% undersampled data set for partial prepayment.

On the basis of the lowest portfolio metrics we choose the model trained on the 50% undersampled training set to be the best model for the partial prepayment model. We show the coefficients from this model in table E.3 in the Appendix as they do not differ greatly to those found for the baseline model in table 6.2.

6.1.3 Full prepayment

Looking at the undersampled results for full prepayment we see that the model trained on 10% undersampled data performs best and narrowly scores smaller errors on all portfolio metrics than the baseline model. The results for this model can be found in table 6.10. Looking at the full prepayment CPR over time shown in figure 6.2 we see that the undersampled models underfit the data less than the baseline model shown in figure 6.1. The out-of-sample results also perform better than the out-of-sample baseline portfolio metrics. We show the result of the model trained on the 10% undersampled data set in table 6.11. Although the loan-level results have improved compared to the baseline model, the precision and recall scores from the 10% undersampled are extremely low. They are shown in the Appendix in table

E.6. As for the variable coefficients, they are not very much different to those of the baseline model and thus are also given in the Appendix, in table E.7.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
10 percent	1.080	0.877	0.959	0.754

TABLE 6.10: A table showing the portfolio metric results for the model trained on the 10% undersampled data set for full prepayment.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
10 percent	0.587	0.585	0.588	0.586

TABLE 6.11: A table showing the portfolio level results for the logistic regression models trained on the 10% undersampled data set for full prepayment, tested on the out-of-sample data.

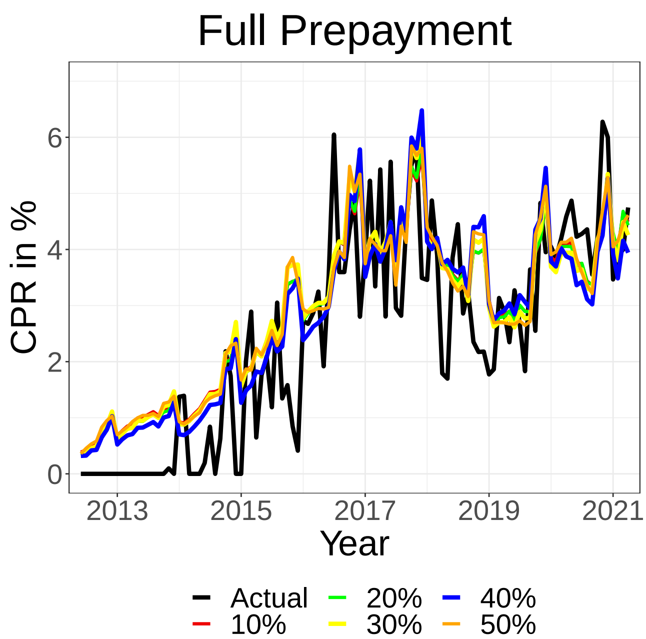

FIGURE 6.2: Plot showing the full CPR over time for the observed and undersampled logistic regression models

The model trained on the 10% undersampled model has the lowest portfolio metrics for both the in-sample and out-of-sample test data. Hence, we choose this to be the best model.

6.1.4 Arbitrage prepayment

Last but not least we review the results for the improved arbitrage prepayment model. Likewise as with the full prepayments we see that the model trained on the 10% undersampled data set performs best, having fractionally smaller portfolio errors than the baseline model. In order not to become repetitive we only show the best undersampled model in table 6.12, which is the model trained on the 10% undersampled data set. The full results are found in table E.8 of the Appendix and the arbitrage prepayment CPR over time can be found in figure 6.3. However, the same improvement cannot be seen for the out-of-sample portfolio metrics, which perform substantially worse than the baseline model.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
10 percent	0.554	0.295	0.536	0.358

TABLE 6.12: A table showing the portfolio metric results for the undersampled data sets for arbitrage prepayment trained on the 10% undersampled data set.

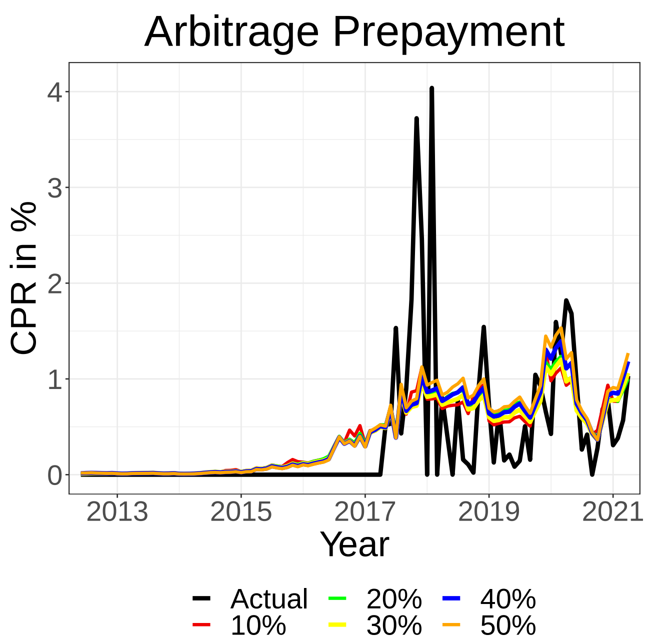

FIGURE 6.3: Plot showing the arbitrage CPR over time for the observed and undersampled logistic regression models

Again we show the accompanying loan-level metrics of the best performing improved arbitrage model (shown in table 6.13). We see that due to the high degree of imbalance in the test data that we have a high accuracy, but very low precision, recall and F1-score. This shows that the logistic classifier has a hard time predicting the true positives out of all predicted positives. Furthermore, the model has a hard time identifying the the number of true positives out of all actual positive cases.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
10 percent	0.981	0.004	0.162	0.981	0.004	0.859	0.000

TABLE 6.13: A table showing the loan metric results for the model trained on the 10% undersampled data set for arbitrage prepayment.

On the basis of the in sample test data we find the 10% undersampled training set to be the best model for the arbitrage prepayment. Likewise as with the improved partial and full prepayment models we show the coefficients for this model in table E.11. The coefficients for this model differ a bit to those of the baseline full prepayment model. We see that being dutch gives a larger odds for making an arbitrage prepayment than non-dutch people. The most logical reason for this could be that dutch citizens are more informed on the dutch mortgage market and the available options for refinancing than non-dutch citizens. This is not surprising because non-dutch citizens most often don't speak the language and are not familiar with all of the dutch financial rules, thus they are less likely to refinance and make an arbitrage prepayment. Lastly, we see that the odds for the 'initial LtMV' have a large negative impact on arbitrage prepayments. Thus, a higher LtMV leads to a lower odds of making an arbitrage prepayment. A reason for this is that mortgagors with a high initial LtMV most often do not have a lot of extra money to spare for prepayments, otherwise these mortgagors would have borrowed a smaller loan. To make an arbitrage prepayment one must also pay a fine, which might be too high for mortgagors with a high initial LtMV.

	Estimate	z value	Pr(>|z|)
redemption type Linear	0.36 5	2.41	0.02
original capital	-0.91	-2.69	0.01
initial LtMV	-1.24	-4.14	0.00
delta rate	5.82	11.52	0.00
loan age	1.26	4.85	0.00
total income	2.59	5.95	0.00
Dutch	1.12	2.31	0.02
LTI	0.22	4.84	0.00
unemployment rate	-0.48	-6.95	0.00
seasonality2	0.24	2.29	0.02

TABLE 6.14: A table showing the variables coefficients for the arbitrage prepayment model which was trained using the 10% undersampled data set.

6.2 Random Forest

6.2.1 Random Forest Model

As with the logistic regression baseline model we run a random forest model for each of the three different types of prepayment. Likewise we use the same variables for the baseline random forest as used in subsection 6.1.1, these are shown in table 6.1. Unlike a logistic regression, the determinants used in a random forest model do not have coefficients. The variables all receive a 'variable importance' score which is based the mean decrease in Gini impurity. This is calculated by the sum over the number of splits that include that certain feature to the number observations it splits, more detail is given in subsection subsection 3.2.3.
We show the variable importance scores for all three types of prepayment baseline models. These are given in figures 6.4, E.1 and E.2 of the Appendix. What is evident from figure 6.4 is that the most important variable for the partial random forest model is the remaining capital. Unfortunately, the model does not indicate the relationship between the variable and the partial prepayment rate. There is no logical explanation for this variable to be so important. The same can be said for the 'LTI', no obvious relationship can be found between the 'LTI' and the partial CPR. Furthermore we see that the seasonality determinant is one of the least important variables, even being close to zero. This is odd as figure 5.5 shows that in the months October, November and December the percentage of partial prepayments is significantly

higher than in other months. The logistic regression for the partial prepayment base model in section 6.2 did give the seasonality variable a high positive coefficient. Further investigating the variable importance results we see that the random forest is biased towards continuous variables and deems categorical variables to be unimportant. We review this problem in Chapter 8.
We see similar results for the variable importance plots for the full and arbitrage prepayment models. We place these figures and those of the improved models in the Appendix as the variable plots show unreliable variable importance results that are not interpretable.

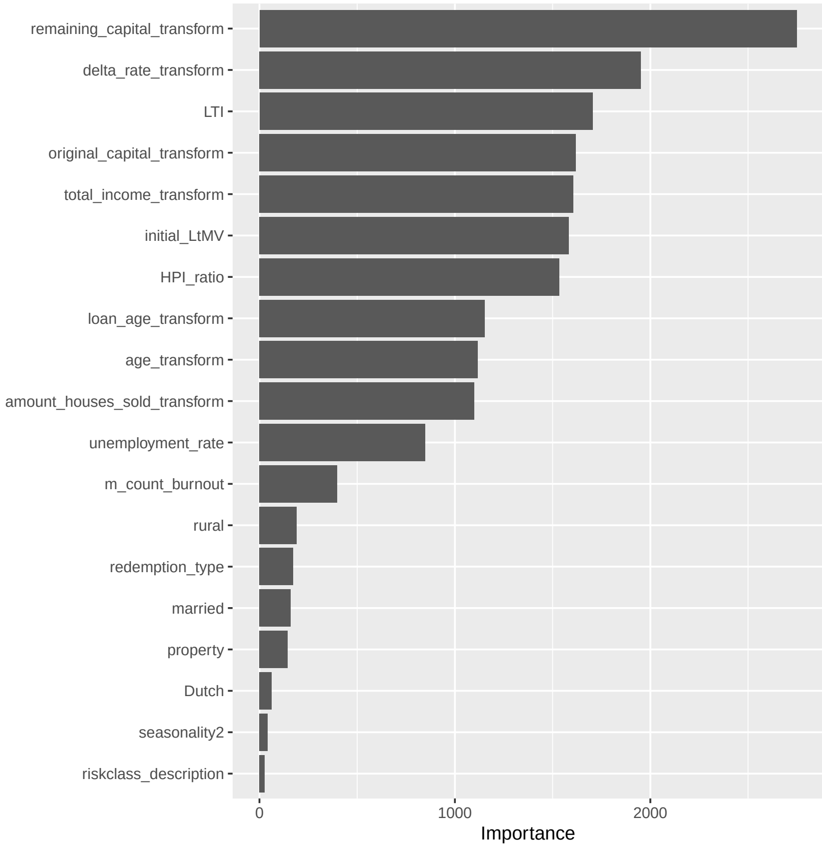

FIGURE 6.4: A figure showing the variable importance for all variables used in the baseline partial prepayment model

When comparing the random forest portfolio results, in table 6.15, with those of the logistic regression baseline model in table 6.5 we see that the all metrics of the baseline random forest model perform equally well for partial repayment and better for the full and arbitrage prepayments. This is understandable when reviewing the CPR over time for both models in figures 6.1 and 6.5, where it is evident that the random forest model is able to capture the general pattern of all three types of prepayment better (as opposed to the logistic regression model). We recall that the baseline logistic regression model severely underfitted the data for both full prepayment CPR and arbitrage prepayment CPR.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
Partial	0.297	0.235	0.230	0.184
Full	1.183	0.891	1.074	0.834
Arbitrage	0.376	0.211	0.392	0.286

TABLE 6.15: Random Forest baseline model portfolio metrics for all three types of prepayment

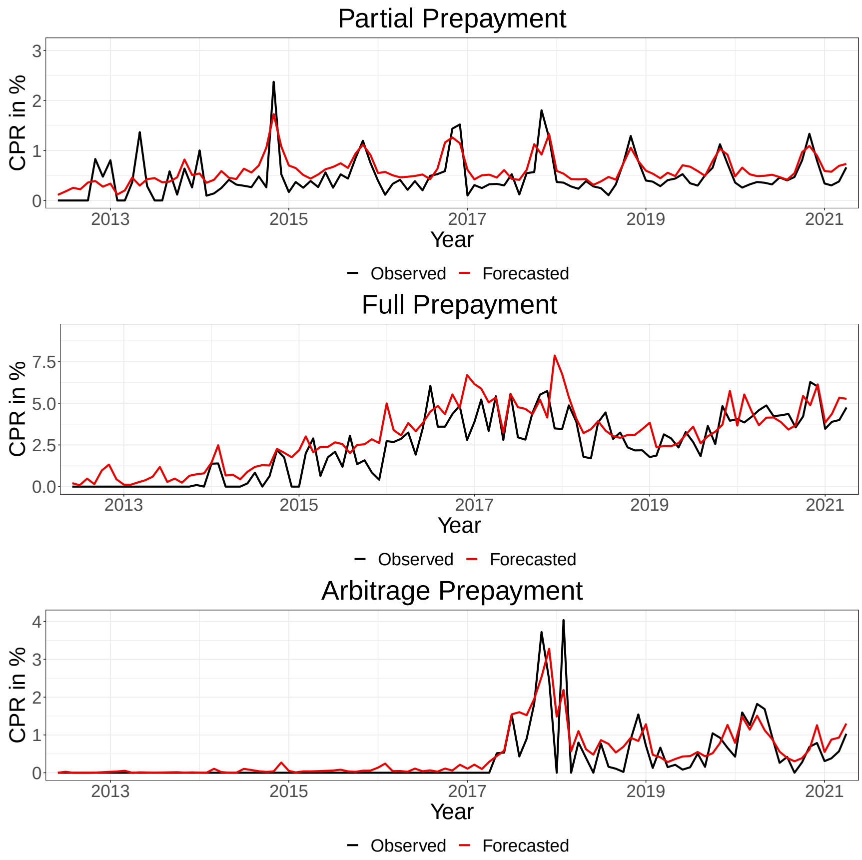

FIGURE 6.5: Plot showing the partial, full and arbitrage CPR over time for the baseline random forest model

Out-of-sample results

As with the logistic regression we review the performance of the random forest model on months the model has never seen before. We see that the model performs considerably worse on the out-of-sample data, especially for full and arbitrage prepayments. This suggests that the random forest overfits the data, leading to much worse results for out-of-sample observations.

6.2.2 Improved Random forest Model

Likewise as with the logistic regression model we improve the baseline model by undersampling the data set for each prepayment type so that the minority class takes up a larger percentage of the training set. Through this we observe the effect of undersampling on the model and hopefully improve the performance. As mentioned in subsection 5.4 we use a different number of more balanced training sets, using undersampling, and correcting for the bias made by this process. We look at the effect of this per prepayment type and and show the results. Additionally we also look at the loan-level metrics for these improved models.

Partial Prepayment

Reviewing the models trained on the various undersampled data sets we see similar results as with the improved models for logistic regression partial prepayment, where a small improvement is seen. The model trained on the 30% undersampled data set performs best of all improved models on the basis of the RMSE and weighted RMSE, and performs better than the baseline model. From figure 6.6 we see that the model trained on the 10% undersampled data set (red line) has higher seasonality peaks and is closer to the observed partial prepayment CPR for the months November and December. However, the 30% undersampled model (yellow line) has a lower base partial prepayment CPR throughout the year, which is closer to the actual observed partial prepayment CPR. The out-of-sample portfolio metrics perform even better and outperform the out-of-sample results for the improved partial logistic regression models.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
10 percent	0.291	0.223	0.212	0.163
20 percent	0.291	0.208	0.209	0.151
30 percent	0.287	0.199	0.205	0.145
40 percent	0.295	0.196	0.223	0.143
50 percent	0.300	0.190	0.232	0.136

TABLE 6.16: Random Forest improved model loan-level metrics for partial prepayment

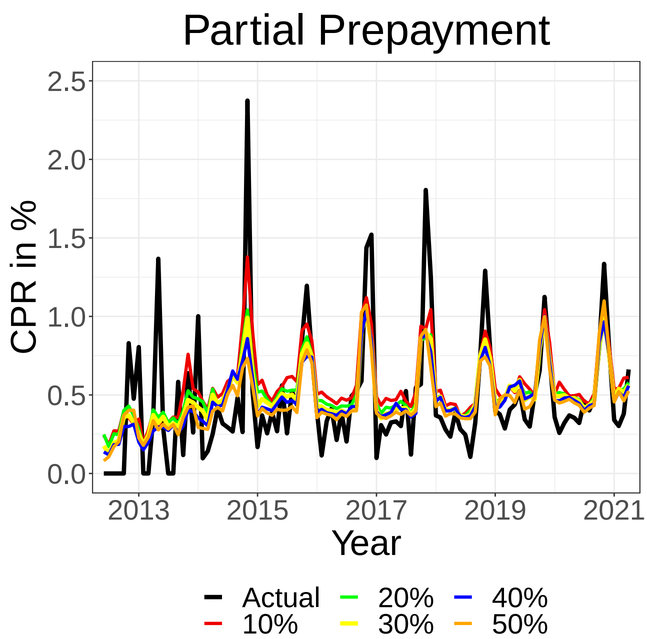

FIGURE 6.6: Plot showing the partial CPR over time for the undersampled random forest models

Showing the loan level results for best improved model we see similar metrics scores to that of the improved partial prepayment logistic regression model, however all metrics seem to be slightly better. What is peculiar is that the brier score is far lower than any of the previous brier scores we have seen. This is probably due to the fact that prior probabilities for observations not making a partial prepayment are closer to 0. With the shifting of the prior probability to the posterior probability this has decreased even further, leading to a large improvement in the brier score. As previously pointed out the precision metric is again very small, leading to a very low F1-score.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
30 percent	0.928	0.050	0.492	0.931	0.090	0.819	0.007

TABLE 6.17: Random Forest improved model loan-level metrics for partial prepayment

Full Prepayment

We see a big improvement in the portfolio level metrics for the full prepayment improved models. The results shown in table 6.18 show a substantial improvement as the degree of undersampling in the data sets increases. The model trained on the 50% undersampled data set performs best, having he lowest errors in all four metrics. Visually we see the difference between the actual and predicted full CPR in figure 6.7. Although less clear, we see that as the degree of undersampling increases that the monthly full prepayment CPR decreases and approaches that of the observed full prepayment CPR, which would explain the decrease in errors seen in the portfolio metrics.
The same improvement in out-of-sample portfolio metrics cannot be seen. The table in Appendix E.16 shows that there is great volatility between the models trained on various undersampled data sets. Overall the models trained on the 20% and 30% undersampled models show better metrics than the in-sample portfolio metrics but the other models perform substantially worse.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
10 percent	1.137	0.908	0.958	0.748
20 percent	1.048	0.848	0.860	0.668
30 percent	0.993	0.814	0.808	0.645
40 percent	0.914	0.732	0.668	0.637
50 percent	0.876	0.704	0.581	0.582

TABLE 6.18: Random Forest improved model loan-level metrics for full prepayment

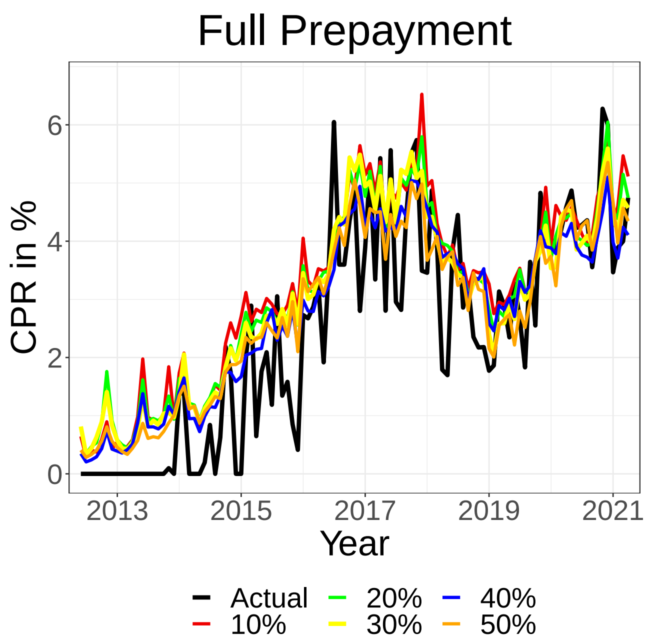

FIGURE 6.7: Plot showing the full CPR over time for the undersampled random forest models

On the exception of the accuracy metric we see that the improved random forest model for full prepayment performs better than that of the improved logistic regression model. Table 6.19 shows these results, where we again see a very low Brier score compared to that of the improved logistic regression model.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
50 percent	0.669	0.008	0.799	0.669	0.016	0.814	0.003

TABLE 6.19: Random Forest improved model loan-level metrics for full prepayment

Arbitrage Prepayment

Reviewing the portfolio metrics of the improved arbitrage prepayment model we see that no improvements are made for all four portfolio metrics compared to the baseline model shown in figure 6.15. We see from figure 6.8 that the 50% undersampled model is closer to the observed arbitrage prepayment CPR around the end of 2002 and beginning of 2021 than the baseline model shown in figure 6.5. As for the out-of-sample metrics only the 30% and 40% undersampled data sets perform better than both the baseline random forest model and the improved arbitrage prepayment model based on the portfolio metrics.

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
10 percent	0.436	0.263	0.445	0.329
20 percent	0.455	0.258	0.448	0.309
30 percent	0.463	0.256	0.452	0.301
40 percent	0.452	0.243	0.439	0.286
50 percent	0.470	0.273	0.497	0.354

TABLE 6.20: Random Forest improved model portfolio metrics for arbitrage prepayment

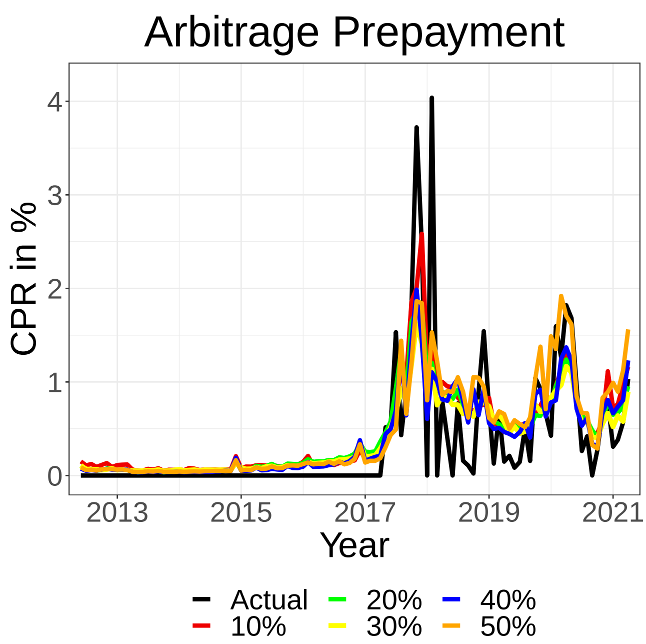

FIGURE 6.8: Plot showing the arbitrage CPR over time for the under-sampled random forest models

The accompanying loan metric results for the random forest arbitrage prepayment model trained on the 50% undersampled data set can be found in table 6.21. In its totality this model performs best on the loan metrics compared to those of the improved models for partial and full prepayment. However, we see a very high recall and low precision. This shows that the model is still predicting to many false positive cases but on the other hand a much lower number of false negative cases. Thus, it is able to correctly identify a large portion of positive observations out of the total number of actual positive observations.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
50 percent	0.825	0.00245	0.863	0.825	0.00488	0.907	0.00050

TABLE 6.21: Random Forest improved model loan-level metrics for arbitrage prepayment

6.3 Neural Network

Regarding the neural network classifier we follow the same structure as was done with the logistic regression and the random forest classifier.

6.3.1 Neural Network Model

The baseline neural network model performs best on all types of prepayment, when compared to the random forest and logistic regression classifiers, except for arbitrage prepayment. For arbitrage prepayment we see that the random forest model has lower errors for all portfolio metrics. The results of the baseline neural network model are shown in table 6.22. The baseline neural network graph showing the predicted CPR over time compared to the observed CPR over time is similar to that of the random forest model.
As for the out-of-sample results, which are shown in table E.21 in the Appendix, they perform better on the partial and full prepayment portfolio metrics than the logistic regression out-of-sample portfolio metrics and perform much better on all types of prepayment than the random forest baseline model.

	RMSE (%)	MAE (%)	Weighted RMSE (%)	Weighted MAE (%)
Partial	0.296	0.204	0.214	0.153
Full	0.961	0.768	0.863	0.683
Arbitrage	0.496	0.256	0.497	0.328

TABLE 6.22: Neural network baseline model portfolio metrics for all three types of prepayment

As with the logistic regression and random forest baseline models we run the neural network baseline model for each of the three different types of prepayment. Again, we use the same variables (which are shown in table 6.1). The variable importance for the neural network model is determined by the absolute value of the weighted connections between nodes of the model (Gevrey, Dimopoulos, and Lek, 2003). The variable importance table resembles that of the random forest model in the sense that most continuous variables have been given a higher variable importance than categorical variables. We see that the 'loan age' variable has the highest variable importance. However, from figure 5.8 we see that with increasing loan age the partial prepayment CPR stays constant. It is therefore odd that this variable is given the highest variable importance. Making the variable importance plot of the neural network model even less credible is again the fact that the 'seasonality' variable has a very low variable importance score. We see similar results for the variable importance tables for the full and arbitrage prepayment models. Due to this we refer to the Appendix for these plots.

However, from figure 5.8 we see that with increasing loan age the partial prepayment CPR stays constant. It is therefore odd that this variable is given the highest variable importance. Making the variable importance plot of the neural network model even less credible is again the fact that the 'seasonality' variable has a very low variable importance score. We see similar results for the variable importance tables for the full and arbitrage prepayment models. Due to this we refer to the Appendix for these plots.

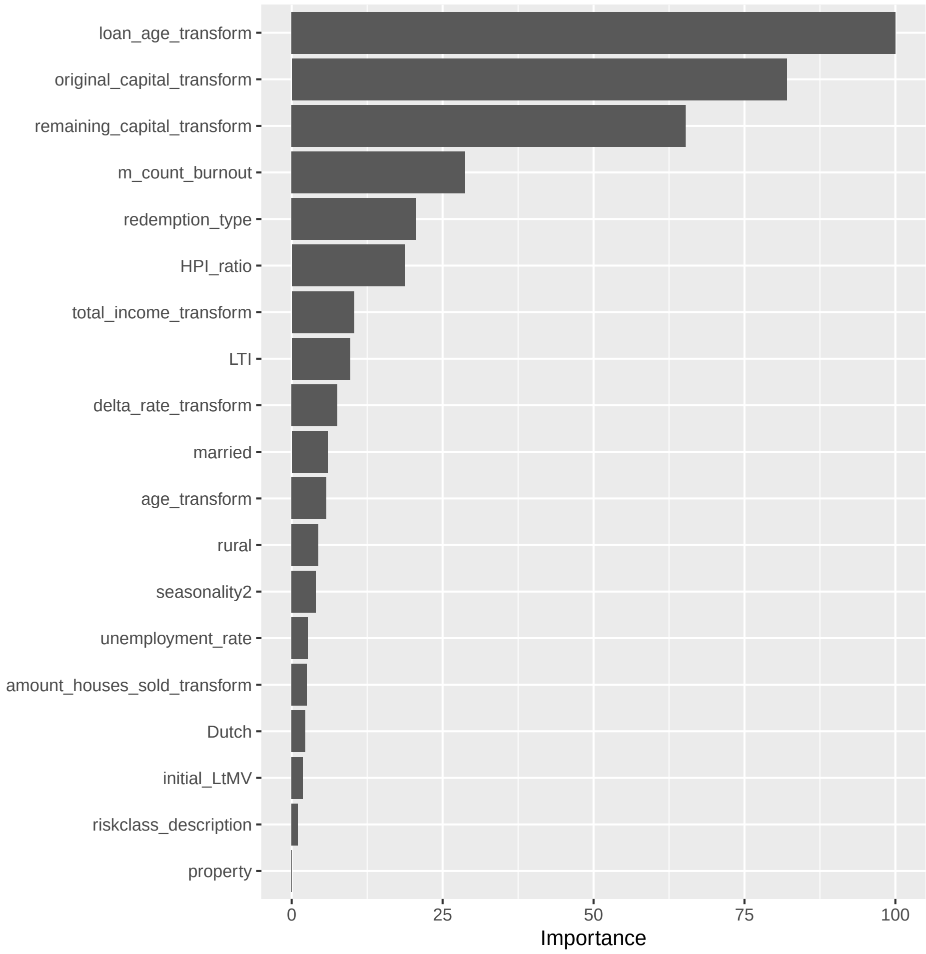

FIGURE 6.9: A figure showing the variable importance for all variables used in the baseline partial prepayment model.

6.3.2 Improved Neural Network Model

We try to improve the neural network by using a different number of less imbalanced data sets through undersampling, these are the same undersampled training sets that were also used for the logistic regression model and the random forest model. We retrain the neural network models and tune the hyperparameters so that we have the optimal configuration for each of the undersampled sets. Furthermore, we do not alter any other variables. The portfolio metrics and the accompanying loan-level metrics will be shown in the same fashion as was done for the improved logistic regression model and improved random forest model.

Partial Prepayment

The improved models for the partial prepayment neural network model do not show any improvement in portfolio metrics relative to the baseline model and thus we place the portfolio metrics in the Appendix. The results can be found in table E.22 and the accompanying partial prepayment CPR over time can be found in figure 6.10. The same observations can be seen for the out-of-sample data, where the baseline model outperforms the improved model. These results are shown in table E.23.

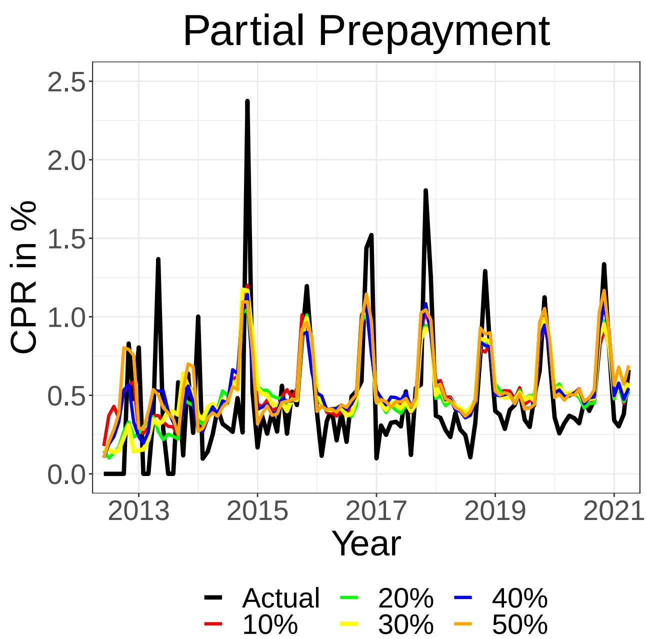

FIGURE 6.10: Plot showing the partial CPR over time for the under-sampled neural network models

Full Prepayment

Reviewing the improved full prepayment models we find that only the model trained on the 10% undersampled data set (where the minority class represents 10% of all observations) performs better than the baseline neural network model. We show this result below while the other results can be found in table E.25 in the Appendix. The most likely reason that this model performs best is that the predicted full prepayment CPR is closer to the observed full prepayment CPR at the beginning years of the portfolio (around 2012-2014), which can be seen from figure 6.11. However, this model performs much worse on the out-of-sample data, seen in table E.26, leading to believe that the model somewhat overfits the in-sample data.

	RMSE (%)	MAE (%)	Weighted RMSE (%)	Weighted MAE (%)
10 percent	0.927	0.746	0.718	0.671

TABLE 6.23: Neural network improved model portfolio metrics for full prepayment

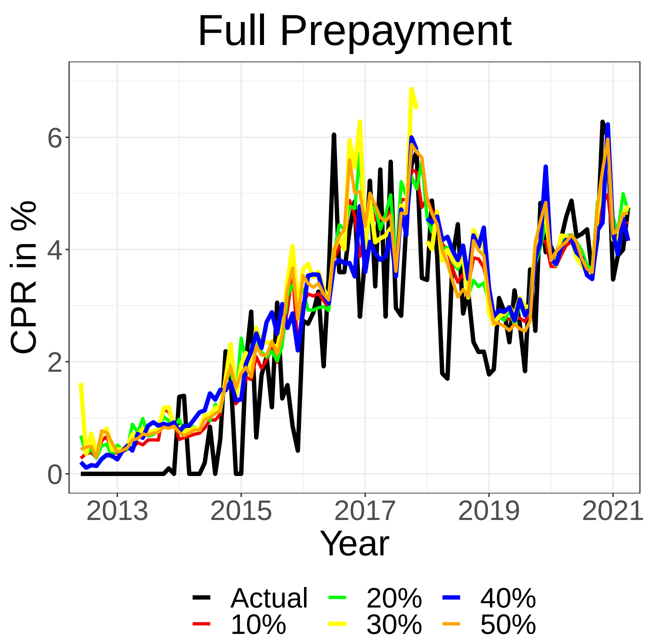

FIGURE 6.11: Plot showing the full CPR over time for the undersampled neural network models

The accompanying loan-level metrics of the 10% undersampled model are shown below. Where we see similar in that both the precision and recall metrics are very low whilst the accuracy of the model is very high. The low precision and recall scores again lead to a low F1-score, which was also the case for the previous loan-level results. Furthermore, we see that the brier score is very low meaning that the predicted probabilities are not very far from the observed probabilities.

	Accuracy	Precision	Recall	Specificity	F1 score	AUC	Brier
10 percent	0.996	0.044	0.012	0.999	0.009	0.768	0.003

TABLE 6.24: Neural network improved model loan-level metrics for full prepayment for the 10% undersampled model

Arbitrage Prepayment

The arbitrage prepayment models trained on undersampled data sets are unable to improve the portfolio level metrics. This is the case for both the in-sample and out-of-sample portfolio metrics, so we only show these results in Appendix E.3. The main reason that these results perform worse than the baseline model is because they predict a higher arbitrage prepayment CPR over time than that of the baseline model. Although the undersampled models are able to capture some of the arbitrage prepayment CPR peaks (such as the one at the end of 2017) the average CPR is higher and seems further away from the observed arbitrage prepayment CPR.

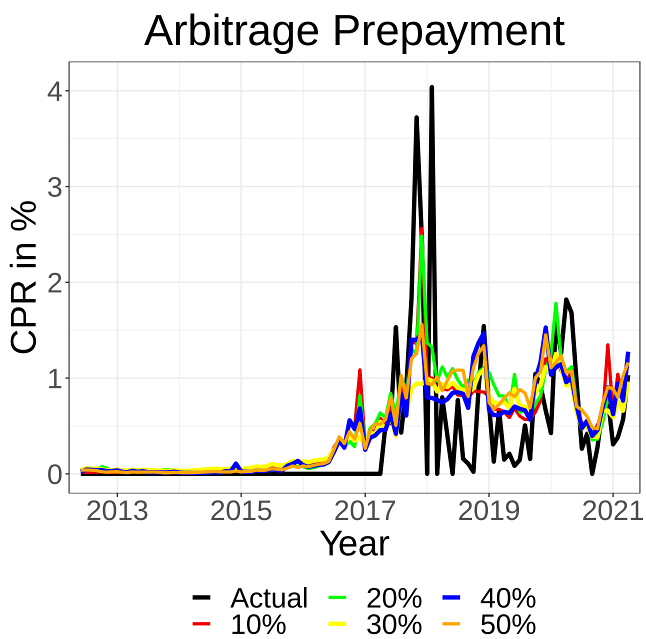

FIGURE 6.12: Plot showing the arbitrage CPR over time for the undersampled neural network models

6.4 Concluding remarks

Regarding the loan-level results we see a reoccurring pattern in which all models provide low loan-level scores for the metrics precision and in most instances also a low recall score. As the F1-score is based on the precision and recall metrics, it is also low. This shows that our models predict and classify too many observations as a prepayment while those observations are not actually prepayments. Low precision values have been seen in previous prepayment modelling research (Saito, 2018). As the models are tested on the highly imbalanced data sets it seems unpreventable that many non-prepayment observations are classified as prepayment observations, which is a common problem in highly imbalanced datasets. A possible reason for the low precision values could be that prepayment and non-prepayment classes overlap one another, making it hard for models to distinguish between the two classes. In order to decrease the separability problem new variables must be found that are able to distinguish between the two classes in a better manner. The results of this thesis show that modelling for loan-specific prepayments still give undesirable results and that the models used for this thesis are not precise enough to classify individual prepayments.

6.5 Concluding sub-question 3

3. Which machine learning models give insight to (new) variables that are relevant and important for prepayment modelling?

We give the results to the this question in sections 6.1, 6.2 and 6.3. We find that the logistic regression is able to give insights into relevant and important variables for prepayment. This is done through the use of log-odds variable coefficients that give interpretable results that clearly show the effects of each used variable in the

model for each of the three prepayment types. Furthermore, due to the use of subset selection the models only incorporate the variables that are statistically significant for that prepayment type. Lastly, as already discussed in section 6.1.1 we find the used variables logical and mostly in line with previous literature and analysis.
The same cannot be said for the random forest model. This model gives a variable importance score to all variables used. However, the variable importance score does not state the relationship with the prepayment type, thus the user has to perform additional analyses in order to find out what kind of effect (positive or negative) the variables have on the prepayment rate. As already mentioned in section 6.2.1 we find that the random forest has a bias towards continuous variables. The effect of this is that binary variables have a much smaller variable importance that continuous variables. Due to these reasons we find that the random forest does not give any relevant insight into variables hat are important for prepayment modelling.
Lastly, we find the same results for the neural network models. Again, the variable importance does not state the relationship each variable has with the prepayment type and as with the random forest, the model seems to also have a bias towards continuous variables. Due to the same reasons we also do not find that the neural network model gives any relevant insight into the variables that are important for prepayment modelling.

6.6 Concluding sub-question 4

4. What is the effect of correcting for the data imbalance problem on the various model results?

We implement the technique of undersampling so that the number of non-prepayment observations decrease, leading to a larger percentage of prepayment observations in the undersampled training sets. As already mentioned we evaluate all models trained on the undersampled training data in which the the prepayment class in question accounts for 10%, 20%, 30%, 40% and 50% of the data set.
The effect of undersampling on the logistic regression models for each of the types of prepayment was mixed. We see that all partial prepayment models trained on the undersampled data sets perform worse than the baseline partial prepayment model with the exception of the model trained on the 50% undersampled data set. However, this model only fractionally performs better than the model without undersampling. Looking at the weighted RMSE we get 0.231% for the imbalanced baseline model and 0.228% for the 50% undersampled model. In the case of full prepayment we see that the model trained on the 10% undersampled data set shows smaller errors than the baseline model. However, here again the improvement is small, with a weighted RMSE 0.988% as opposed to 0.959%. The other full prepayment models trained on a higher degree of undersampling all perform worse. As with full prepayment, the exact same is seen for arbitrage prepayment where only the model trained on the 10% undersampled data set performs better than the baseline model.
Looking at the effect of undersampling on the random forest models we see that the results are mixed. All partial prepayment models trained on the undersampled data sets perform better than the baseline model, we see that the weighted RMSE for the 30% undersampled model is 0.205% whereas that of the baseline model is 0.230%. The effect of undersampling is even bigger for full prepayments where the baseline model has a weighted RMSE of 1.074% whereas the 50% undersampled model has a

weighted RMSE of 0.581%. However, reviewing the results for arbitrage prepayment we see that undersampling does not improve the portfolio level results.
Lastly, for the neural network models we see that undersampling again shows mixed results. We find that undersampling does not improve the error for both the partial prepayments and arbitrage prepayments. Whereas for full prepayment the model we see that only the model trained on the 10% undersampled data set is fractionally better than the baseline full prepayment model.
To conclude we find that the effect of undersampling is mixed. It has a positive, albeit small, effect for the logistic regression model. For the random forest the positive effect is substantial for both partial and full prepayments, however does not have a positive effect for arbitrage prepayments. Lastly, we find that the effect of undersampling is the smallest on the neural network in which only the full prepayment model saw a small improvement in the error metrics.

6.7 Concluding sub-question 5

5. Which machine learning models can be used to estimate the future prepayment rates?

On the basis of the best portfolio metrics for the in-sample results we see that the random forest partial prepayment model trained on the 30% undersampled data set performs best at forecasting the partial prepayment CPR on the basis of the RMSE and weighted RMSE. Concerning full prepayment we find that again the random forest classifier performs best, with the model being trained on the 50% undersampled data set showing the lowest error (again on the basis of the RMSE and weighted RMSE) and being able to replicate the observed full prepayment CPR the best. The story repeats itself for arbitrage prepayment, where the random forest classifier performs best. However, here the baseline random forest models shows the lowest errors. The best models from each classifier for each of the three types of prepayment are given in tables 6.25, 6.26 and 6.27. We want to add that the neural network model comes close for both the partial and arbitrage prepayment CPR.

Partial Prepayment

	RMSE (%)	MAE (%)	Weighted RMSE (%)	Weighted MAE (%)
LR 50 percent	0.316	0.214	0.225	0.159
RF 30 percent	0.287	0.199	0.205	0.145
NN baseline	0.296	0.204	0.214	0.153

TABLE 6.25: A table showing the results of the best model for each of the three classifiers for partial prepayment. The first row is logistic regression model where a prepayment represents 50% of the observations. The second row is the random forest model where 30% of the observations are prepayments and the last row is the neural network baseline model with no undersampling.

Full Prepayment

	RMSE (%)	MAE (%)	Weighted RMSE (%)	Weighted MAE (%)
LR 10 percent	1.080	0.877	0.959	0.754
RF 50 percent	0.876	0.704	0.581	0.582
NN 10 percent	0.927	0.746	0.718	0.671

TABLE 6.26: A table showing the results of the best model for each of the three classifiers for full prepayment.

Arbitrage Prepayment

	RMSE (%)	MAE (%)	Weighted RMSE (%)	Weighted MAE (%)
LR 10 percent	0.554	0.295	0.536	0.358
RF baseline	0.376	0.211	0.392	0.286
NN baseline	0.496	0.256	0.497	0.328

TABLE 6.27: A table showing the results of the best model for each of the three classifiers for arbitrage prepayment.

6.8 Cash Flow Estimation results

We examine the results for the portfolio cash flow estimation of various models and compare it to the portfolio cash flow estimation of the Allianz model and the observed cash flow over time. We estimate the cash flows of every baseline model and the best undersampled model, meaning we calculate the cash flows for a total of six models. Reviewing the portfolio metrics for the total prepayment rate for the models trained on undersampled data sets in tables E.12, E.20, E.31 we see that the models with the lowest overall errors are the random forest model trained on the 50% undersampled data set, the baseline neural network and the neural network trained on the 10% undersampled data set.

Year	Allianz model error (%)	Base LR error (%)	LR 50% error (%)	Base RF error (%)	RF 50% error (%)	Base NN error (%)	NN 10% error (%)
2014	20.7	65.4	57.1	74.3	52.9	29.2	35.0
2015	56.0	61.6	56.4	62.8	49.1	37.7	43.1
2016	-0.9	11.5	9.8	28.1	10.5	11.7	11.5
2017	-19.7	-3.7	-3.5	21.3	-3.0	8.7	5.8
2018	-19.2	26.5	28.6	22.8	17.6	23.6	28.4
2019	-12.5	18.2	18.7	15.3	5.7	7.4	13.4
2020	-31.9	-11.5	-7.2	-0.0	-0.7	-10.2	-7.2
2021	-16.2	9.0	16.5	25.0	19.6	11.3	14.4
abs(mean) without 2012-2013	22.1	25.9	24.7	31.2	19.9	17.5	19.8

TABLE 6.28: A table showing the yearly cash flow estimation error for multiple models compared to the observed prepayment cash flow.

Reviewing the table we see that for the year 2014 and 2015 all models have a very large error compared to the observed prepayment cash flows. This is most likely due to the young nature of the portfolio. Although prepayments occur, they are very small in size. This can also be seen by observing a figure such as figure E.4, where only partial prepayments take place up to midway 2014. Due to the young nature of the portfolio and the small size of the portfolio the models have a hard time estimating the prepayment cash flows. As the portfolio seasons over the years and the portfolio balance grows we see lowers errors in estimated prepayment cash flows of the model. We calculate the absolute mean of the yearly model errors and see that there are three models that perform better than the Allianz model. These models are the random forest model trained on the 50% undersampled data set, the base neural network and the neural network trained on the 10% undersampled data set. These models achieve a mean error rating of 19.9%, 17.5% and 19.8% compared to the 22.1% of the Allianz model. The other three models perform worse on the yearly cash flow estimation. By taking a weighted mean, based on the number of loans in each year, we find that all models have a smaller error than the Allianz model. These results can be found in table E.32 in the Appendix E.4. This suggests that as the portfolio size increases that the newly constructed models perform better than the original Allianz model.
We also measure the total prepayment cash flows of each model and compare them to the observed prepayment cash flows over the range of 2014 until present. These results can be found in table 6.29, were we see that the Allianz model overall estimates 19.3% less prepayment, over the period between 2014 until present, than was actually observed. When reviewing the other models we find that their total prepayment cash flow error is much smaller, with the neural network base model having just a 4% error to the observed prepayment cash flows. The predicted prepayment cash flows versus the actual prepayment cash flows of both the best random forest model and the Allianz model are shown in figure 6.13. The left hand plot shows the observed prepayment cash flows (turquoise) of the Allianz portfolio versus the predicted cash flows (orange) of the Allianz model while the right hand plot shows the observed prepayment cash flows (turquoise) of the Allianz portfolio versus the predicted cash flows (orange) of the neural network baseline model trained on the fully imbalanced training data set. Observing the left hand plot we see that the Allianz model has been under-forecasting the prepayment cash flows since mid-2018. Since then the difference between the observed and forecasted cash flows has only increased, which has lead to the 19.3% error given in table 6.29. The right hand plot shows that although the baseline neural network model was over-forecasting (although only slightly) the prepayment cash flows from roughly 2015-2020, the error between the observed and forecasted has decreased, if not gone completely.

Model name	Error (%)
Allianz model	-19.3
LR base	5.4
LR 50% undersampled	8.2
RF base	14.2
RF 50% undersampled	7.3
NN base	4.0
NN 10% undersampled	7.2

TABLE 6.29: A table showing the estimated prepayment cash flow error over the range 2014 until present.

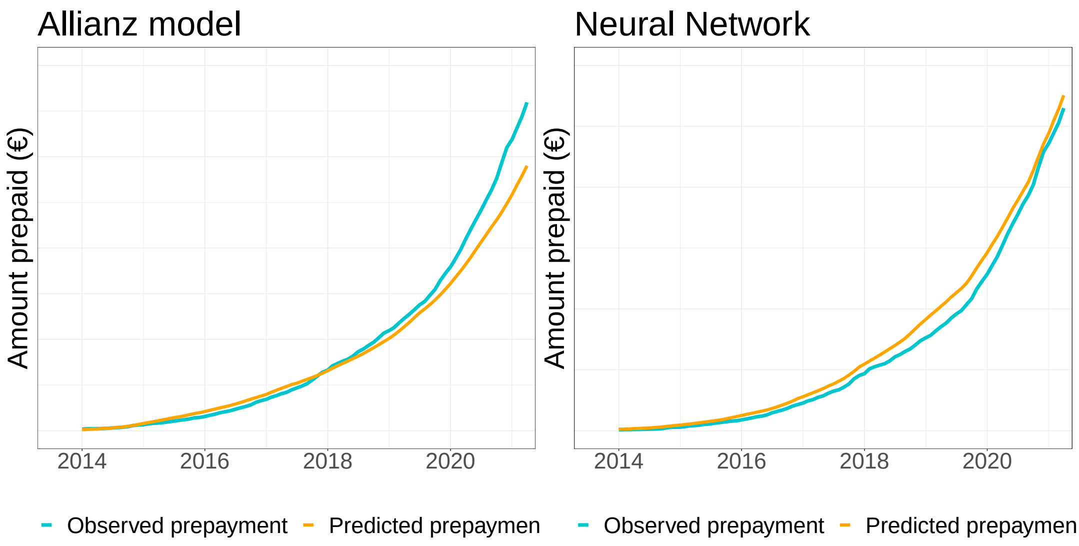

FIGURE 6.13: LHS: Plot showing the observed prepayment cash flow (turquoise) of the Allianz portfolio versus the predicted cash flow (orange) of the Allianz model. RHS: Plot showing the observed prepayment cash flow (turquoise) of the Allianz portfolio versus the predicted cash flow (orange) of the neural network baseline model.

6.9 Summary

In this chapter we give the results of the portfolio- and loan-level results for each of the three prepayment types for the three different machine learning models, namely the logistic regression, the random forest and the neural network. We start off by giving the portfolio-level results along with the variable coefficients for the baseline logistic regression model. We find that the logistic regression shows interpretable results for the prepayment variable coefficients. We follow with results for the logistic regression model which has been trained on multiple undersampled data sets and find that this reduces the portfolio-level errors of the model. In the same order we analyse the results for the random forest model, first showing the results for the baseline model and then analysing the results for the random forest model trained on multiple undersampled data sets. We find that undersampling also reduces the portfolio-level errors for this model. However, the random forest model does not show clear and interpretable prepayment variable coefficients. The same reporting structure is used for the neural network model. We find that undersampling only gives a small improvement for the portfolio-level metrics of full prepayment. Likewise, as with the random forest model, the neural network does not give interpretable prepayment variable results. We find that all loan-level results mostly improve with undersampling but all show extremely low precision and F1-scores. This shows that the models have a hard time precisely predicting individual prepayments. Last of all we compare the yearly average and total prepayment cash flow error of the benchmark Allianz model to the best logistic regresion, random forest and neural network models. We find that the baseline neural network model performs best in predicting the yearly and total prepayment cash flows.

Chapter 7

Conclusion

This thesis set out to investigate if there were any machine learning models that, when applied to mortgage data were able to forecast the prepayment rate more accurately than is currently being done by the benchmark logistic regression model at Allianz. In order to aid in answering this research question we set up a series of sub-questions. By performing a literature review we were able to find relevant prepayment drivers. Furthermore, through a literature search we were able to find suitable machine learning models that could show promising results for prepayment modelling. These are the logistic regression, random forest and the neural network. With the literature in mind we analysed the data to find the main prepayment drivers, specific to the Allianz mortgage portfolio. After elaborating on the model performance metrics, cash flow estimation and choosing undersampling as a suitable method to adjust for the imbalance in the data set we give the results to all the models. We find that the logistic regression model is the only model to give insight into variables that are relevant and important to prepayment modelling. Furthermore, we find that undersampling has mixed effects on each of the three machine learning models. Undersampling decreases the portfolio-level errors for both the logistic regression and random forest model, however, only reduces the errors for the neural network for full prepayment.

Continuing, we find the model with lowest conditional prepayment rate (CPR) error on the basis of the weighted RMSE for each of the three prepayment types. The random forest trained on the 30% undersampled training set has the lowest weighted RMSE for partial prepayment, with it being 0.205%. Regarding full prepayment we see that the random forest trained on the 50% undersampled training set has the lowest weighted RMSE at 0.581% when compared to the actual full prepayment CPR. As for arbitrage prepayment it is the baseline random forest model with the lowest weighted RMSE error, being 0.392%.

Lastly we answer the main research question:

> *"Are there any machine learning models that, when applied to Allianz mortgage data of the Netherlands, can forecast the prepayment rate more accurately than the benchmark model at Allianz over the time horizon of 2014-2021?"*

This is done by modelling the prepayment cash flow estimation error of each constructed model with that of the Allianz logistic regression prepayment model and to the actually observed prepayment cash flows over time. We look at the absolute yearly mean prepayment error of each model and to the total cash flow prepayment error over the time period of 2014-2021. In regards to the absolute yearly mean prepayment error we find that only the baseline neural network model, the neural network model trained on the 10% undersampled data and the random forest model trained on the data set with equal minority and majority class observations (50% undersampled) have a lower yearly error than the Allianz model. The Allianz model has an absolute yearly error of 22.1% whereas the baseline neural network model, the neural network model trained on the 10% undersampled data and the random forest model trained on the 50% undersampled data have an absolute yearly error of 17.5%, 19.8% and 19.9% respectively. Reviewing the total prepayment cash flow error over the lifetime period of 2014-2021 we find that the Allianz model has an error of -19.3% in regards to the observed prepayment cash flow. All other models have a smaller error, with the baseline neural network model having the lowest with a total prepayment cash flow error of 4.0%. This is far below that of the Allianz benchmark model, achieving the goal set out at the start of this thesis in finding a machine learning model that is able to forecast the prepayment rate more accurately than the benchmark Allianz model.

In conclusion we find that the baseline neural network model shows the most promising results overall. Not only did this model have low errors when predicting for each individual prepayment type, it also had the lowest yearly mean prepayment cash flow error when compared to the actual prepayment cash flow and also had the lowest lifetime total error when compared to the actual prepayment cash flow. However, the lack of insight into relevant drivers of prepayment for the neural network model make the model less interpretable. This aspect of the model must be improved in order for the model to become practical and useful.

Chapter 8

Limitations and future research

In this section we elaborate on the various shortcomings in this research. Furthermore, we comment in which direction future research into prepayment modelling should point.

8.1 Limitations

This study into prepayment modelling has some limitations, the largest limitation of this research is the young portfolio book of Allianz. The biggest consequence of this would be that the Allianz data is not completely representative of the Dutch population. This topic has been a recurring theme throughout this thesis and has been mentioned multiple times. Although Allianz has been funding mortgagors since 2011, the main increase in outstanding mortgages came after 2017. This means that a large part of the book is just shy of four years in age. As we previously mentioned one of the main drivers for prepayment, according to literature, is the seasoning variable. The older the age of the mortgage the larger the chance that external factors such as the marital status, family composition or job location change for a household. Such factors contribute largely to mortgagors making a full prepayment. Due to the young maturity of the Allianz portfolio this change is less visible. If we were to redo this research in 15-20 years, when the portfolio is fully matured and longer term mortgages are ending, we would most likely find higher prepayment rates that would be more representative for the Dutch population. A possible solution in the meantime would be to find mortgage data that is representative of the Dutch population that has the needed determinants incorporated into the data set and has a matured data. Analysing this data could give results as to how the prepayment rate could look like in the future for the Allianz data.
Additionally, having such a young portfolio has its limitations for the effect of certain determinants for the prepayment rate. Since 2011, when Allianz started started funding mortgages, mortgage rates have only been decreasing. Thus, we have only been able to research the effect of decreasing mortgages rates on important determinants as mentioned in literature, such as the refinancing incentive and the burnout counter. It is important to analyse what happens to prepayment rates once mortgages rates start (gradually) increasing. The addition of periods in time where mortgage rates have been increasing would be beneficial to study these determinants. Not only would it be beneficial for analysing the impact on certain determinants, it would also help with modelling the effect of long-term scenario's for Allianz. Without data containing rising interest rates it is harder to model outlook scenario's in which funding rates increase. Thus, with this data one should expect that the models calculating long-term scenario's would be more complete.
The big data set proved to be problematic at times during data manipulation and analysis. However, it should not be perceived as a limitation for this research. On

the other hand it can be conceivable that the increasing data set size could become a limitation for future research, mainly because the programming language R does not have the capabilities to work with very large data sets. Other programs such as SQL could be more suitable for the data manipulation phase of this research in the future.

Another problem that hindered the progress of this research was the lack of allocated computational power. Although we made use of special shared servers at Allianz, the computational heavy models such as the random forest and neural network often took very long to complete and hindered others from using the server in question. This meant that running the random forest and neural network models had to done during the evening, night or during weekends so that no other employees were hindered in the their daily tasks. With the increased number of outstanding loans the number of data points will only increase at a greater rate in the coming period. The data set will most likely double in size as of mid 2022. This will negatively impact the running time of models as the computational limits of the server will be tested. A possible and feasible solution for improving the computational power would be to allocate a special server for this project. Another solution could be to subset the total data and use samples that are smaller in size but still representative for the whole data set. A possible sampling technique that could be helpful is stratified sampling. Although we mention this in the limitations chapter we want to clarify that this problem only delayed the modelling progress but was not an actual constraint for the modelling phase. If non of the above mentioned solutions is implemented then this problem will become a constraint for modelling in the future.

One other limitation that was found during this thesis was the variable importance bias of random forest models. Analysing this problem we see that a 2007 study (Strobl et al., 2007) finds that when a random forest model is trained using various different types of variables that the random forest classifier has a bias towards continuous variables with a larger scale of measurement. In turn this makes the variable importance unreliable. This was also observed and mentioned during this research in the random forest variable importance graphs, where categorical variables were often given a very low importance score. This was the case with the seasonality determinant, which we know from the preliminary analysis is a very prominent variable for partial prepayment. Something similar was also found in the neural network models, where continuous variables seemed to be given a higher preference above categorical variables. Without improving this aspect of the random forest and neural network models, they will be less interpretable and transparent. In turn this will hinder the implementation of these models in the financial sector.

In this research we looked at the performance of three different types of models on prepayment modelling. Furthermore, the effects of undersampling and the data set size were evaluated for each model. In totality this led to a cornucopia of results and parameters for each model to train or investigate, meaning that numerous aspects of each model were not fully optimized due to lack of time, for example. Instead of reviewing three models and not optimizing or examining every aspect for the model in great detail, one could opt to examine just two models but also investigate the effect of different variable importance metrics for the random forest and neural networks. This might achieve better results for the models in question regarding not only the loan-level and portfolio metrics but also the results for the main drivers of prepayment. In the long run this might achieve a better understanding for the importance of some variables, which would in turn lead to prepayment models that are more accurate.

8.2 Future directions

The acceptability of machine learning models in the financial world play a crucial role in the implementation of these models. For now many financial institutions still consider machine learning models to be black box models that are not interpretable and lack explanatory outputs. In order to make them implementable, research must be conducted at Allianz stating which interpretable requirements are needed for machine learning models to become implementable. From this a standardized framework must be made constructed stating these requirements so that machine learning models can be accepted and used for modelling and risk management. This was evident for both the random forest and neural network models, which lacked explanatory output to show the relationship between possible prepayment drivers and the prepayment rate itself. In order for such models to become implementable in the future research must be done in order to establish a way to make the relationship between the variables and the prepayment form interpretable and clear.

Although we performed out-of-sample tests for the portfolio level model results this was not done with the cash flow estimation. With the addition of a full new year worth of prepayment data it is necessary to evaluate the performance of each model and compare it with the benchmark Allianz model for the out-of-sample data. For the out-of-sample testing we suggest to look at both the yearly estimation error as the influence on the total error over the whole time range.

In the future we suggest to incorporate the chosen (newly constructed) prepayment models to help the valuation of mortgage buckets in the future. The idea would be to use the best prepayment model, including all the chosen determinants, and combine it with a interest rate Monte Carlo simulation over a long time period. This would help to examine the effect of the interest rate paths on the forward-looking prepayment rate and also on the future mortgage cash flows. In turn the net present value (NPV) for certain mortgage buckets can be calculated and compared to the NPV of different interest rate scenario paths. Incorporating such Monte Carlo simulations and calculating the NPV of different scenario's will help Allianz to 'stress test' their mortgage portfolio in regards to the prepayment rate. This will help them gain more knowledge on expected mortgage cash flows in the future under different interest rate scenario's, which will help in calculating the mortgage assets. In turn this should help in reducing the long-term asset and liability mismatch, improving the risk management side of Allianz.

Bibliography

Alink, Bjorn Johan (2002). “Mortgage prepayments in the Netherlands. Explanatory research on prepayment variables and the effects on asset & liability management and securitisation”. In:

Amit, Yali and Donald Geman (Oct. 1997). “Shape Quantization and Recognition with Randomized Trees”. In: *Neural Computation* 9.7, pp. 1545–1588. ISSN: 08997667. DOI: 10.1162/neco.1997.9.7.1545. URL: https://dl.acm.org/doi/abs/10.1162/neco.1997.9.7.1545.

Andrews, Dan and Aida Caldera Sánchez (2011). “The Evolution of Homeownership Rates in Selected OECD Countries: Demographic and Public Policy Influences”. In: *OECD Journal: Economic Studies* 2011. DOI: 10.1787/19952856. URL: http://dx.doi.org/10.1787/19952856http://dx.doi.org/10.1787/eco_studies-2011-5kg0vswqpmg2.

Aniceto, Maisa Cardoso, Flavio Barboza, and Herbert Kimura (Dec. 2020). “Machine learning predictivity applied to consumer creditworthiness”. In: *Future Business Journal* 6.1, pp. 1–14. ISSN: 2314-7210. DOI: 10.1186/s43093-020-00041-w. URL: https://doi.org/10.1186/s43093-020-00041-w.

Archer, Wayne R and David Ling (1993). “Pricing Mortgage-Backed Securities: Integrating Optimal Call and Empirical Models of Prepayment”. In: *Real Estate Economics* 21.4, pp. 373–404. URL: https://econpapers.repec.org/RePEc:bla:reesec:v:21:y:1993:i:4:p:373-404.

Archer, Wayne R, David Ling, and Gary A McGill (1996). “The effect of income and collateral constraints on residential mortgage terminations”. In: *Regional Science and Urban Economics* 26.3-4, pp. 235–261. URL: https://econpapers.repec.org/RePEc:eee:regeco:v:26:y:1996:i:3-4:p:235-261.

Archer, Wayne R., David C. Ling, and Gary A. McGill (June 1997). “Demographic versus Option-Driven Mortgage Terminations”. In: *Journal of Housing Economics* 6.2, pp. 137–163. ISSN: 10511377. DOI: 10.1006/jhec.1997.0206.

Assef, Fernanda et al. (Oct. 2019). “Classification Algorithms in Financial Application: Credit Risk Analysis on Legal Entities”. In: *IEEE Latin America Transactions* 17.10, pp. 1733–1740. ISSN: 15480992. DOI: 10.1109/TLA.2019.8986452.

Bank, De Nederlandsche (2016). *Kredietmarkten in beweging*. Tech. rep., p. 44.

Barbaglia, Luca, Sebastiano Manzan, and Elisa Tosetti. “Forecasting Loan Default in Europe with Machine Learning*”. In: (). DOI: 10.1093/jjfinec/nbab010. URL: https://academic.oup.com/jfec/advance-article/doi/10.1093/jjfinec/nbab010/6313398.

Bijlage Hypotheekschuldmonitor en update aanpak aflossingsvrije hypotheken | Kamerstuk | Rijksoverheid.nl. URL: https://www.rijksoverheid.nl/documenten/kamerstukken/2020/06/11/bijlage-hypotheekschuldmonitor-en-update-aanpak-aflossingsvrije-hypotheken.

Bonsema (2019). *Dutch RMBS: a primer ABS In-Depth*. Tech. rep.

Breiman, Leo (Oct. 2001). “Random forests”. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324. URL: https://link.springer.com/article/10.1023/A:1010933404324.

Brennan, Michael J. and Eduardo S. Schwartz (Sept. 1985). "Determinants of GNMA Mortgage Prices". In: *Real Estate Economics* 13.3, pp. 209–228. ISSN: 15406229. DOI: 10.1111/1540-6229.00351. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/1540-6229.00351https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.00351https://onlinelibrary.wiley.com/doi/10.1111/1540-6229.00351.

Brownlee, J (2016). *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End*. Machine Learning Mastery. URL: https://books.google.nl/books?id=BgmqDwAAQBAJ.

Bussel, Petrus Johannes Michaël van (1998). "Valuation and interest rate risk of mortgages in the Netherlands". In:

Charlier, Erwin and Arjan Van Bussel (2003). "Prepayment behavior of Dutch mortgagors: an empirical analysis". In: *Real estate economics* 31.2, pp. 165–204. ISSN: 1080-8620.

Clapp, John M et al. (2001). *Movers and Shuckers: Interdependent Prepayment Decisions*. Tech. rep., pp. 411–450.

Cox, D R (May 1972). "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2, pp. 187–220. ISSN: 00359246. URL: http://www.jstor.org.ezproxy2.utwente.nl/stable/2985181.

Dal Pozzolo, Andrea et al. (2015). "Calibrating Probability with Undersampling for Unbalanced Classification". In: DOI: 10.1109/SSCI.2015.33.

Damrongsakmethee, Thitimanan and Victor Emil Neagoe (2019). "C4.5 Decision Tree Enhanced with AdaBoost Versus Multilayer Perceptron for Credit Scoring Modeling". In: *Advances in Intelligent Systems and Computing*. Vol. 1047. Springer, pp. 216–226. ISBN: 9783030313616. DOI: 10.1007/978-3-030-31362-3{_}22.

Dunn, Kenneth B and John J Mcconnell (1981). *A Comparison of Alternative Models for Pricing GNMA Mortgage-Backed Securities*. Tech. rep. 2.

Elkan, Charles (2001). "The Foundations of Cost-Sensitive Learning". In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 973–978. ISBN: 1558608125.

Elsing, Hugo (2019). "Modeling and pricing the prepayment option of Dutch residential mortgage loans". PhD thesis. Erasmus University Rotterdam, p. 55.

Europe: owner occupiers with and without mortgage | Statista. URL: https://www.statista.com/statistics/957803/homeowners-with-and-without-an-outstanding-mortgage-in-eu-28-per-country/.

Gan, Jumwu and Junwu Gan (2009). *Burnout from pools to loans: Modeling refinancing prepayments as a self-selection process*. Tech. rep.

Gevrey, Muriel, Ioannis Dimopoulos, and Sovan Lek (Feb. 2003). "Review and comparison of methods to study the contribution of variables in artificial neural network models". In: *Ecological Modelling* 160.3, pp. 249–264. ISSN: 0304-3800. DOI: 10.1016/S0304-3800(02)00257-0.

Green, Jerry and John B Shoven (1986). *The Effects of Interest Rates on Mortgage Prepayments*. Tech. rep. 1, pp. 41–59. URL: http://www.jstor.org/journals/ohio.press.html..

Green, Richard K. and Michael Lacour-Little (Sept. 1999). "Some Truths about Ostriches: Who Doesn't Prepay Their Mortgages and Why They Don't". In: *Journal of Housing Economics* 8.3, pp. 233–248. ISSN: 10511377. DOI: 10.1006/jhec.1999.0247.

Grootbanken winnen terrein op hypotheekmarkt. URL: https://www.banken.nl/nieuws/8048/grootbanken-winnen-terrein-op-hypotheekmarkt.

Habib, Yousef (2020). "Modelling the Prepayment Risk of Dutch Mortgages". PhD thesis. Erasmus Univeristy Rotterdam, p. 52.

Hamza, Mounir and Denis Larocque (Aug. 2005). "An empirical comparison of ensemble methods based on classification trees". In: *Journal of Statistical Computation and Simulation* 75.8, pp. 629–643. ISSN: 00949655. DOI: 10.1080/00949650410001729472.

Hayre, Lakhbir S (Mar. 2003). "Prepayment Modeling and Valuation of Dutch Mortgages". In: *The Journal of Fixed Income* 12.4, pp. 25–47. ISSN: 1059-8596. DOI: 10.3905/jfi.2003.319337. URL: https://jfi.pm-research.com/content/12/4/25https://jfi.pm-research.com/content/12/4/25.abstract.

He, Haibo and Edwardo A. Garcia (Sept. 2009). "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9, pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

Ife, Tac /. *MMI Fund Analysis FY 2005 Appendix A: Econometric Analysis of Mortgages A-1 Appendix A: Econometric Analysis of Mortgages*. Tech. rep.

Jacobs, J P A M, R H Koning, and E Sterken (2005). *MODELLING PREPAYMENT RISK*. Tech. rep.

James, Gareth et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Kau, James B. and Donald C. Keenan (1995). "An overview of the option-theoretic pricing of mortgages". In: *Journal of Housing Research* 6.2, pp. 217–244.

King, Gary et al. (2001). "Logistic Regression in Rare Events Data". In: URL: http://GKing.Harvard.Edu..

Kolbe, Andreas (2008). *Valuation of mortgage products with stochastic prepayment-intensity models*.

Kotsiantis, Sotiris, D Kanellopoulos, and P Pintelas (Nov. 2005). "Handling imbalanced datasets: A review". In: *GESTS International Transactions on Computer Science and Engineering* 30, pp. 25–36.

Kuijpers, Bart H M and Peter C Schotman (2007). *Optimal prepayment of Dutch mortgages**. Tech. rep. 1, pp. 137–155.

LaCour-Little, Michael, Michael Marschoun, and Clark Maxam (Jan. 2002). "Improving Parametric Mortgage Prepayment Models with Non-parametric Kernel Regression". In: *Journal of Real Estate Research* 24.3, pp. 299–328. ISSN: 0896-5803. DOI: 10.1080/10835547.2002.12091098. URL: https://www.tandfonline.com/doi/abs/10.1080/10835547.2002.12091098.

López, Victoria et al. (June 2012). "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics". In: *Expert Systems with Applications* 39.7, pp. 6585–6608. ISSN: 0957-4174. DOI: 10.1016/J.ESWA.2011.12.043.

Luo, Cuicui (Oct. 2019). "A comprehensive decision support approach for credit scoring". In: *Industrial Management and Data Systems* 120.2, pp. 280–290. ISSN: 02635577. DOI: 10.1108/IMDS-03-2019-0182.

Marktaandeel grootbanken hypotheken bereikt historisch dieptepunt. URL: https://www.vastgoedactueel.nl/marktaandeel-hypotheken-grootbanken-dieptepunt/.

Nembrini, Stefano, Inke R. König, and Marvin N. Wright (Nov. 2018). "The revival of the Gini importance?" In: *Bioinformatics* 34.21, pp. 3711–3718. ISSN: 14602059. DOI: 10.1093/BIOINFORMATICS/BTY373.

Reducing Loss: Gradient Descent. URL: https://developers.google.com/machine-learning/crash-course/reducing-loss/gradient-descent.

Riksen, R.Q. (2017). "Using Artificial Neural Networks in the Calculation of Mortgage Prepayment Risk". In: URL: https://scripties.uba.uva.nl/scriptie/634160.

Saito, Taiyo (2018). "Mortgage Prepayment Rate Estimation with Machine Learning". PhD thesis. TU Delft, p. 86.

Samuel, A L (1959). "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3.3, pp. 210–229. DOI: 10.1147/rd.33.0210.

Schwartz, Eduardo S and Walter N Torous (1989). *American Finance Association Prepayment and the Valuation of Mortgage-Backed Securities*. Tech. rep. 2, pp. 375–392.

Sirignano, Justin A, Apaar Sadhwani, and Kay Giesecke (2015). *Deep Learning for Mortgage Risk*. Tech. rep.

Staat van de Woningmarkt - Jaarrapportage 2020 | Rapport | Rijksoverheid.nl. URL: https://www.rijksoverheid.nl/documenten/rapporten/2020/06/15/staat-van-de-woningmarkt-jaarrapportage-2020.

StatLine - Bestaande koopwoningen; gemiddelde verkoopprijzen, regio. URL: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83625NED/table?ts=1618324596348.

StatLine - Kerngegevens sectoren; nationale rekeningen. URL: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84097NED/table?ts=1607329979735.

Strobl, Carolin et al. (2007). "Bias in random forest variable importance measures: illustrations, sources and a solution". In: *BMC bioinformatics* 8. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-25. URL: https://pubmed.ncbi.nlm.nih.gov/17254353/.

Sun, Yanmin, Andrew K.C. Wong, and Mohamed S. Kamel (June 2009). "Classification of imbalanced data: A review". In: *International Journal of Pattern Recognition and Artificial Intelligence* 23.4, pp. 687–719. DOI: 10.1142/S0218001409007326.

Total residential mortgage lending by country Europe 2020 | Statista. URL: https://www.statista.com/statistics/614792/outstanding-residential-mortgage-lending-europe-by-country/.

Uhe, Jasper (2019). *Financial Framework and Crisis Effect on the Mortgage Investments of Insurance Companies Developments, Risks, and Financial Conglomerates*. Tech. rep.

Waiganjo Wagacha, Peter (2003). *Decision Trees Learning*. URL: https://www.researchgate.net/publication/283569105_Decision_Trees_Learning.

Waller, Bennie and Milam Aiken (Dec. 1998). "Predicting Prepayment of Residential Mortgages: A Neural Network Approach". In: *International Journal of Information and Management Sciences* 9.4, pp. 37–44. ISSN: 10171819.

Wonen - Nederlandse Vereniging van Banken (NVB). URL: https://www.bankinbeeld.nl/thema/wonen/.

Wonen - Nederlandse Vereniging van Banken (NVB). URL: https://www.bankinbeeld.nl/thema/wonen/.

Appendix A

Technical Formula's

A.1 Maximum Likelihood Estimator

We show the formula for the maximum likelihood estimator (MLE) below in formula A.1:

$$\mathcal{L}(\beta, Y, X) = \prod_{i=1}^{N} P(Y_{i,t} = 1|X_{i,t})^{Y_{i,t}}(1 - P(Y_{i,t} = 1|X_{i,t}))^{1-Y_{i,t}} \tag{A.1}$$

By taking the log-likelihood we can transform this equation into the following:

$$l(\beta, Y, X) = \sum_{i=1}^{n} Y_{i,t}(log(P(Y_{i,t} = 1|X_{i,t}) + (1 - Y_{i,t})log(1 - (P(Y_{i,t} = 1|X_{i,t})) \tag{A.2}$$

As we know the formula for $P(Y_{i,t} = 1|X_{i,t})$ is:

$$P(Y_{i,t} = 1|X_{i,t}) = \frac{e^{\beta X_{i,t}}}{1 + e^{\beta X_{i,t}}} = \frac{1}{1 + e^{-\beta X_{i,t}}} \tag{A.3}$$

If we fill in equation A.3 into A.2 we get the following:

$$l(\beta, Y, X) = \sum_{i=1}^{n} -log(1 + e^{\beta X_{i,t}}) + \sum_{i=1}^{n} Y_{i,t}(\beta X_{i,t}) \tag{A.4}$$

To find the maximum likelihood estimate we differentiate the log-likelihood with respect to the parameters B, set the derivative to zero and solve the equation.

Appendix B

Determinant plots

HPI

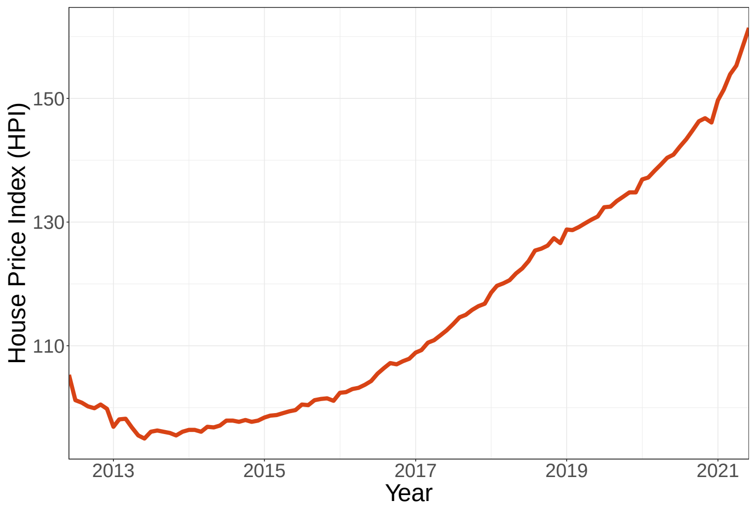

FIGURE B.1: Plot showing the House Price Index over time.

Indexed actual LtMV

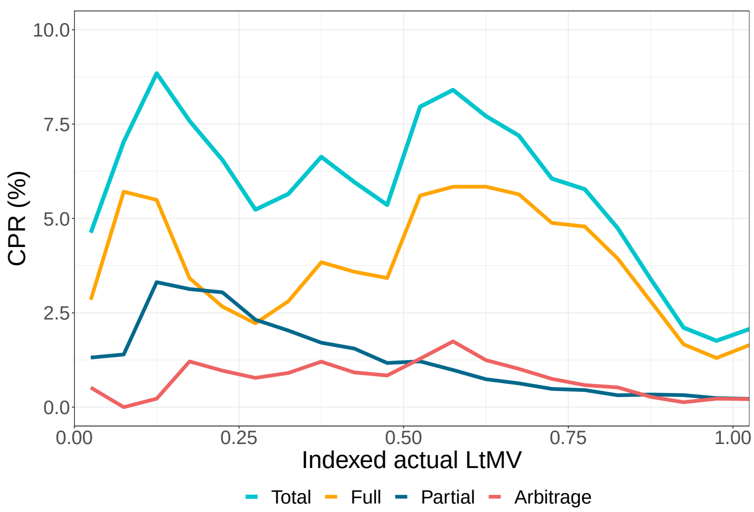

FIGURE B.2: Plot showing the CPR for each indexed actual LtMV.

HPI ratio

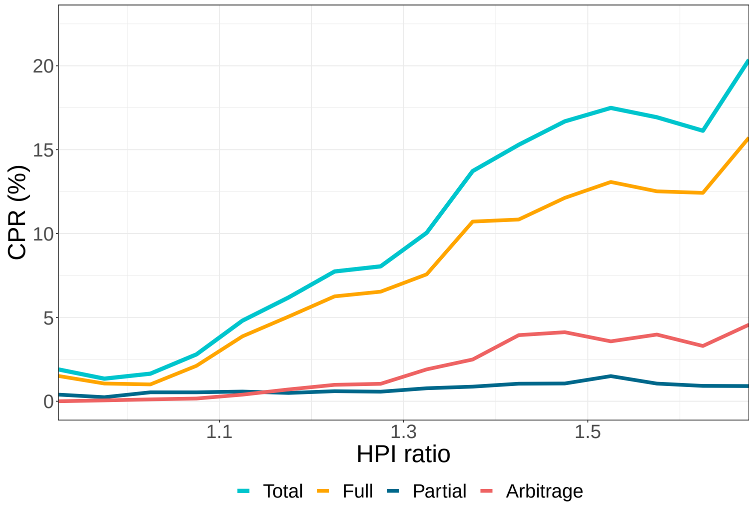

FIGURE B.3: Plot showing the CPR for each HPI ratio.

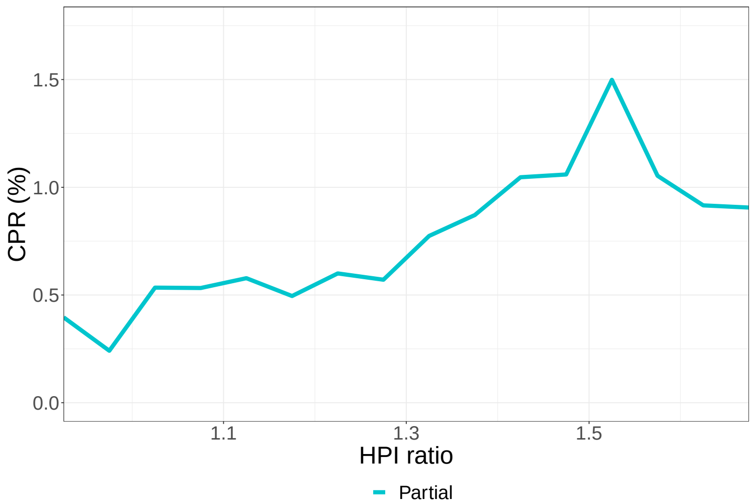

FIGURE B.4: Plot showing the partial prepayment CPR for each HPI-ratio.

Loan to Income

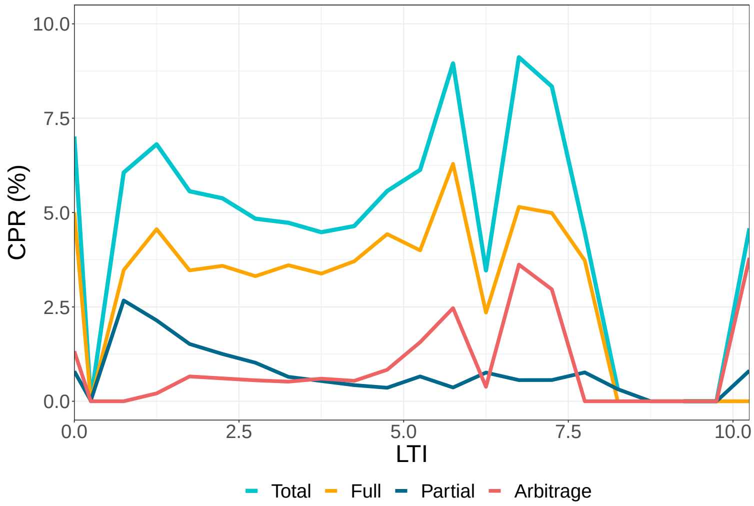

FIGURE B.5: Plot showing the CPR for each LTI ratio

Appendix C

Variable Description

In this Appendix we give a detailed description of every variable used for modelling in the logistic regression, random forest and neural network models.

Property Type

A categorical variable that looks at the building type of each observation. This variable only distinguishes between mortgagors with a flat type property (with or without a garage) and the other property types.

Mortgagor Age

This is a continuous variable which looks at the log age of each mortgagor of each observation. If a mortgage contract has more than one mortgagor linked to the contract the mean age of both mortgagors is calculated.

Initial LtMV

This is a continuous variable which looks at the loan-to-market-value of the observation at the time of mortgage inception.

Loan age (Seasoning)

This continuous variable counts the log age of the mortgage, in months, since its inception.

Refinancing Incentive (Delta rate)

The refinancing incentive, also called the delta rate, is a continuous variable that calculates the logarithm of the mortgage rate divided by the market interest rate for a mortgage with similar characteristics as the mortgage in question. The refinancing incentive is positive is positive if the market interest rate is lower than the mortgage rate in question and negative if the market interest rate is higher than the mortgage in question.

Seasonality

This categorical variable, which is also called 'seasonality2', looks at whether the reporting date is situated in the months October, November or December. Literature usually only incorporates the months November and December into the seasonality variable, however, from figure 5.5 in the preliminary analysis it was visible that the month October also contained a larger percentage of mortgagors that made a prepayment.

Geographical Location

This categorical variable, which is called 'rural' in the data set, looks at whether the mortgage is located in a rural province in the Netherlands or in a urban province. The urban provinces are considered to be 'Zuid-Holland', 'Noord-Holland' and 'Utrecht'.

Burnout

This variable is closely linked to the 'refinancing incentive' variable which was discussed earlier. It is a continuous variable which counts the number of months in which the mortgage (and mortgagor) has been given the opportunity to refinance but has not taken advantage of this opportunity. We say that a refinancing incentive is possible if the market interest rate is lower than 1% of that of the interest rate of the mortgage in question.

House Price Index Ratio (HPI-ratio)

A continuous variable that looks at the HPI-ratio between the reporting date at mortgage inception and that of the observation of the mortgage in question. For example, if the HPI is at 102 at mortgage inception and the observation of that mortgage in question has a HPI of 104 then we calculate the HPI-ratio as the following $(104/102) * 100 = 1.019$.

Marital Status

Also called 'married', is a categorical value which looks at the marital status of the mortgagor. We distinguish between 'single' mortgagors and 'non-single' mortgagors.

Risk class description

A categorical variable which distinguishes between the multiple risk classes of each mortgage. We distinguish between NHG mortgages and mortgages with different LTV buckets.

Remaining Outstanding Loan Part Principal

This continuous variable looks at the log remaining principal outstanding of the loan part in question. This variable is also called 'remaining capital' in the data set.

Loan part principal

This variable, also called 'original capital', looks at the log principal of each specific loan part and not at the total contract capital of the mortgage. This is again a continuous variable.

Income

This variable calculates the log total yearly income of the mortgagors linked to the mortgage. If only one mortgagor is linked to the mortgage contract then only this persons' income is taken into account. For two mortgagors we add both incomes. Again this is a continuous variable.

Dutch

This categorical value looks at the nationality of the mortgagor and distinguishes be Dutch and non-Dutch mortgagors linked to the contract. All mortgages in question are all situated in the Netherlands.

LTI

This continuous variable looks at the loan-to-income of the mortgage contract. As the name already mentions, this continuous variable looks at the ratio of the total mortgage principal to the total income of the mortgagors connected to the mortgage contract.

Number of houses sold

This variable looks at the number of houses sold per month in the Netherlands and take the square root of this number. These numbers are given by the Dutch Bureau of statistics, also called Centraal Bureau van Statistiek ('CBS').

Unemployment Rate

This variable looks at the monthly unemployment rate of the the workforce in the Netherlands. These numbers have been taken from the Centraal Bureau van Statistiek ('CBS') of the Netherlands. This is a continuous variable.

Redemption Type

This categorical variable distinguishes between the type of redemption form of each mortgage. We distinguish between three different types of redemption forms for each mortgage, namely the linear redemption type, the annuity redemption type and the bullet redemption type. Each loan part within a mortgage contract can have a different redemption type.

Appendix D

Statistics

	riskclass description	Mortgagor age	Partial pp (#)	Risk class	Fraction (%)
1	>95% MW	38.70	9.00	1595	0.56
2	60% t/m 80% MW	44.64	99.00	13688	0.72
3	80% t/m 90% MW	40.51	12.00	2535	0.47
4	90% t/m 95% MW	40.92	3.00	644	0.47
5	NHG	40.21	11830.00	1612468	0.73
6	t/m 60% MW	55.89	237.00	32815	0.72
7	t/m 80% MW	38.85	9.00	240	3.75

TABLE D.1: Table showing the number of partial prepayments per risk class. This is also represented as a fraction of all observations for that risk class.

	riskclass description	Mortgagor age	Full pp (#)	Observations (#)	Fraction (%)
1	>95% MW	38.70	1.00	1595	0.06
2	60% t/m 80% MW	44.64	28.00	13688	0.20
3	80% t/m 90% MW	40.51	10.00	2535	0.39
4	90% t/m 95% MW	40.92	5.00	644	0.78
5	NHG	40.21	5671.00	1612468	0.35
6	t/m 60% MW	55.89	83.00	32815	0.25
7	t/m 80% MW	38.85	3.00	240	1.25

TABLE D.2: Table showing the number of full prepayments per risk class. This is also represented as a fraction of all observations for that risk class.

Appendix E

Results

E.1 Logistic regression

E.1.1 Partial Prepayment

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	0.019	0.019	0.019	0.019
20 percent	0.023	0.017	0.023	0.017
30 percent	0.021	0.020	0.021	0.020
40 percent	0.022	0.022	0.022	0.021
50 percent	0.031	0.025	0.031	0.025

TABLE E.1: Portfolio metrics for the partial prepayment model trained on various undersampled data sets. The model was tested on the out-of-sample data set.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
10 percent	0.993	0.031	0.014	0.998	0.019	0.651	0.0001
20 percent	0.990	0.047	0.021	0.997	0.014	0.651	0.0001
30 percent	0.963	0.030	0.131	0.969	0.024	0.651	0.0001
40 percent	0.854	0.017	0.344	0.858	0.017	0.652	0.0001
50 percent	0.717	0.011	0.590	0.726	0.011	0.648	0.0001

TABLE E.2: A table showing the loan metric results for the undersampled data sets for partial prepayment.

	Estimate	Std. Error	z value	Pr(>|z|)
redemption typeBullet	-0.12	0.04	-3.06	0.00
redemption typeLinear	0.61	0.05	12.49	0.00
original capital	0.54	0.05	11.06	0.00
initial LtMV	-1.33	0.11	-12.51	0.00
delta rate	-1.44	0.17	-8.35	0.00
loan age	0.78	0.07	11.26	0.00
rural	-0.10	0.03	-3.17	0.00
total income	0.53	0.11	4.76	0.00
age	-0.32	0.15	-2.18	0.03
married	-0.12	0.04	-3.31	0.00
Dutch	-0.28	0.10	-2.87	0.00
LTI	-0.23	0.02	-9.85	0.00
HPI ratio	0.49	0.21	2.34	0.02
amount houses sold	-0.57	0.20	-2.84	0.00
seasonality2	0.60	0.03	17.39	0.00

TABLE E.3: A table showing the variables coefficients for the partial prepayment model which was trained using the 50% undersampled data set.

E.1.2 Full Prepayment results

	RMSE (%)	MAE (%)	weighted RMSE (%)	weighted MAE (%)
10 percent	1.080	0.877	0.959	0.754
20 percent	1.081	0.871	0.964	0.761
30 percent	1.134	0.906	0.995	0.785
40 percent	1.126	0.892	1.070	0.843
50 percent	1.108	0.883	0.961	0.734

TABLE E.4: A table showing the portfolio metric results for the undersampled data sets for full prepayment.

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	0.587	0.585	0.588	0.586
20 percent	0.686	0.685	0.687	0.686
30 percent	0.887	0.885	0.888	0.886
40 percent	0.990	0.989	0.991	0.990
50 percent	0.531	0.529	0.532	0.530

TABLE E.5: A table showing the portfolio level results for the logistic regression models trained on the undersampled data sets for full prepayment, tested on the out-of-sample data.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
10 percent	0.996	0.024	0.004	0.999	0.003	0.749	0.000
20 percent	0.974	0.016	0.111	0.977	0.014	0.749	0.000
30 percent	0.902	0.011	0.329	0.904	0.011	0.748	0.000
40 percent	0.804	0.009	0.530	0.805	0.009	0.747	0.000
50 percent	0.659	0.007	0.712	0.659	0.007	0.748	0.000

TABLE E.6: A table showing the loan metric results for the undersampled data sets for full prepayment.

	Estimate	Std. Error	z value	Pr(>|z|)
redemption typeLinear	-0.22	0.06	-3.41	0.00
original capital	-1.01	0.09	-11.32	0.00
remaining capital	0.30	0.08	3.98	0.00
delta rate	2.24	0.25	8.84	0.00
loan age	1.18	0.10	11.73	0.00
m count burnout	-0.01	0.00	-3.66	0.00
property	0.40	0.04	9.41	0.00
total income	0.33	0.13	2.62	0.01
age	-1.93	0.17	-11.36	0.00
married	-0.27	0.04	-6.38	0.00
LTI	-0.14	0.03	-5.44	0.00
HPI ratio	1.06	0.23	4.56	0.00
amount houses sold	0.50	0.26	1.93	0.05
unemployment rate	-0.06	0.02	-2.85	0.00
seasonality2	0.19	0.04	4.76	0.00

TABLE E.7: A table showing the variables coefficients for the full prepayment model which was trained using the 10% undersampled data set.

E.1.3 Arbitrage Prepayment results

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	0.554	0.295	0.536	0.358
20 percent	0.568	0.304	0.551	0.370
30 percent	0.567	0.297	0.545	0.363
40 percent	0.568	0.302	0.553	0.377
50 percent	0.571	0.308	0.567	0.393

TABLE E.8: A table showing the portfolio metric results for the undersampled data sets for arbitrage prepayment.

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	0.559	0.559	0.560	0.560
20 percent	0.530	0.530	0.530	0.530
30 percent	0.615	0.614	0.615	0.615
40 percent	0.725	0.724	0.726	0.725
50 percent	0.784	0.782	0.784	0.783

TABLE E.9: A table showing the portfolio level results for the logistic regression models trained on the undersampled data sets for arbitrage prepayment, tested on the out-of-sample data.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
10 percent	0.981	0.004	0.162	0.981	0.004	0.859	0.000
20 percent	0.936	0.004	0.461	0.937	0.004	0.860	0.000
30 percent	0.906	0.003	0.591	0.906	0.003	0.862	0.000
40 percent	0.865	0.002	0.669	0.865	0.002	0.855	0.000
50 percent	0.792	0.002	0.773	0.792	0.002	0.856	0.000

TABLE E.10: A table showing the loan metric results for the undersampled data sets for arbitrage prepayment.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-14.72	2.22	-6.63	0.00
redemption typeBullet	-0.20	0.11	-1.85	0.06
redemption typeLinear	0.36	0.15	2.41	0.02
original capital	-0.91	0.34	-2.69	0.01
remaining capital	0.51	0.31	1.68	0.09
initial LtMV	-1.24	0.30	-4.14	0.00
delta rate	5.82	0.50	11.52	0.00
loan age	1.26	0.26	4.85	0.00
total income	2.59	0.43	5.95	0.00
married	0.18	0.11	1.60	0.11
Dutch	1.12	0.49	2.31	0.02
LTI	0.22	0.05	4.84	0.00
unemployment rate	-0.48	0.07	-6.95	0.00
seasonality2	0.24	0.10	2.29	0.02

TABLE E.11: A table showing the variables coefficients for the arbitrage prepayment model which was trained using the 10% undersampled data set.

E.1.4 Total Prepayment results

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	1.304	1.070	1.191	0.959
20 percent	1.306	1.067	1.210	0.972
30 percent	1.315	1.075	1.197	0.969
40 percent	1.301	1.051	1.258	1.002
50 percent	1.311	1.076	1.189	0.947

TABLE E.12: A table showing the portfolio level results for the logistic regression models trained on the undersampled data sets for total prepayment, tested on the in-sample data.

E.2 Random Forest Results

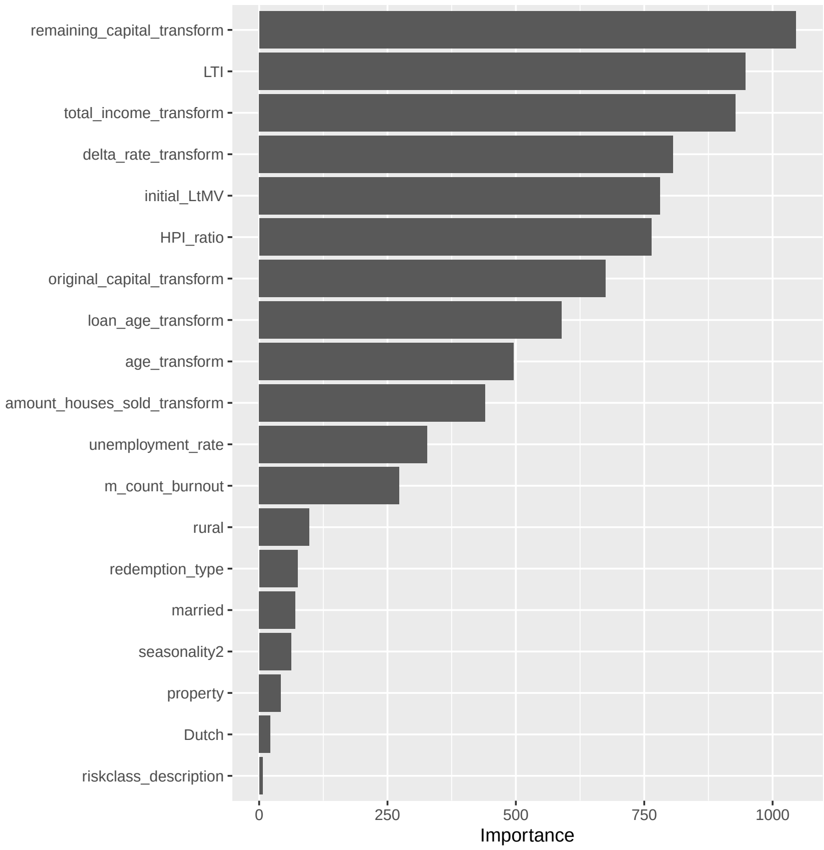

FIGURE E.1: A figure showing the variable importance for all variables used in the baseline full prepayment model

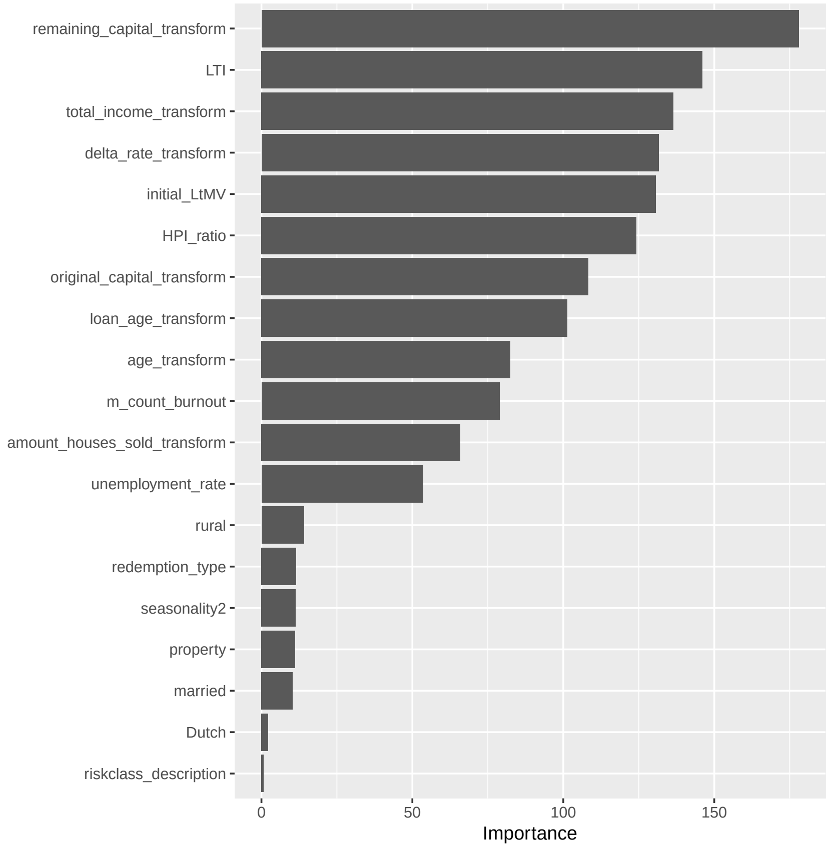

FIGURE E.2: A figure showing the variable importance for all variables used in the baseline arbitrage prepayment model

	RMSE	MAE	weighted RMSE	weighted MAE
Partial	0.791	0.786	0.792	0.788
Full	6.471	6.360	6.497	6.386
Arbitrage	3.447	3.415	3.456	3.424
Total	9.035	8.904	9.069	8.938

TABLE E.13: A table showing the portfolio level results for the baseline random forest model for all types of prepayment, tested on the out-of-sample data.

E.2.1 Partial Prepayment Results

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	0.242	0.239	0.242	0.240
20 percent	0.084	0.082	0.085	0.082
30 percent	0.014	0.013	0.014	0.013
40 percent	0.068	0.067	0.068	0.067
50 percent	0.120	0.120	0.119	0.120

TABLE E.14: A table showing the portfolio level results for the random forest models trained on the undersampled data sets for partial prepayment, tested on the out-of-sample data.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
10 percent	0.985	0.140	0.199	0.991	0.165	0.828	0.007
20 percent	0.964	0.076	0.360	0.968	0.126	0.825	0.007
30 percent	0.928	0.050	0.492	0.931	0.090	0.819	0.007
40 percent	0.870	0.033	0.596	0.872	0.062	0.816	0.007
50 percent	0.758	0.021	0.718	0.759	0.041	0.810	0.007

TABLE E.15: Random Forest improved model loan-level metrics for partial prepayment

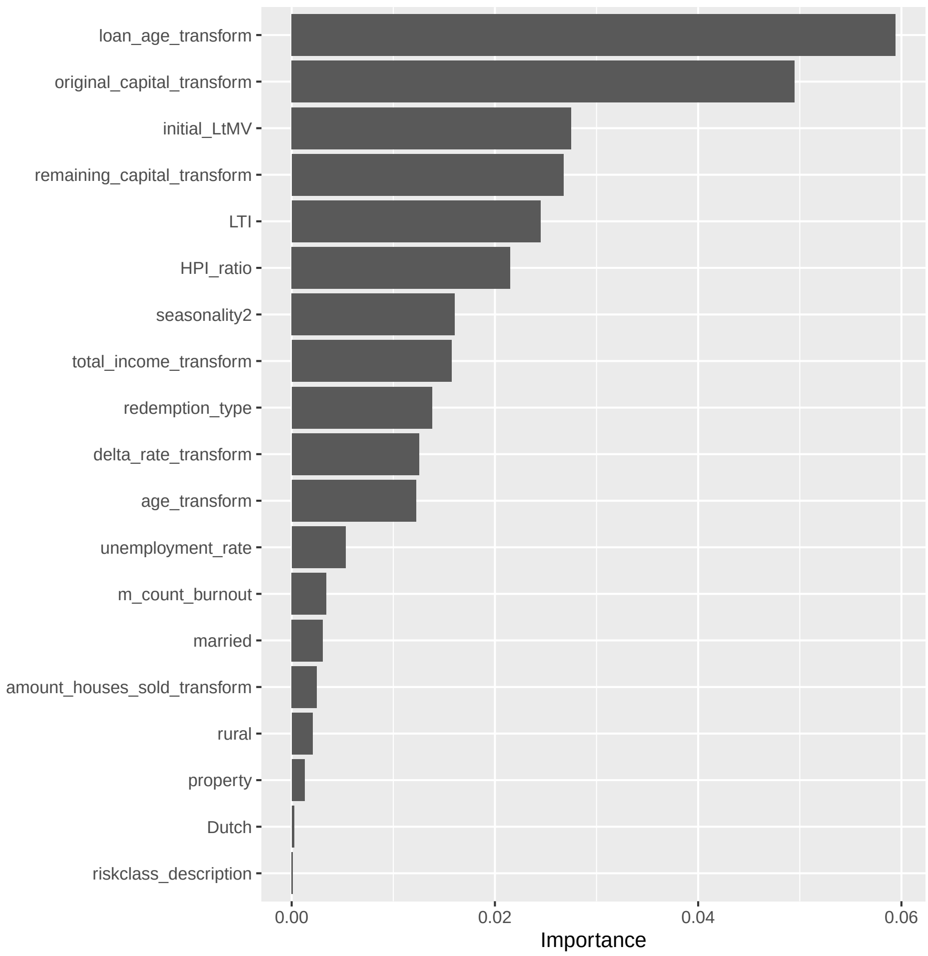

FIGURE E.3: Plot showing the variable importance of the 50% undersampled partial prepayment model

E.2.2 Full Prepayment Results

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	1.882	1.873	1.885	1.877
20 percent	0.432	0.417	0.434	0.419
30 percent	0.422	0.422	0.422	0.422
40 percent	1.021	1.021	1.021	1.021
50 percent	1.482	1.480	1.483	1.481

TABLE E.16: A table showing the portfolio level results for the random forest models trained on the undersampled data sets for full prepayment, tested on the out-of-sample data.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
10 percent	0.994	0.020	0.006	0.999	0.009	0.652	0.005
20 percent	0.977	0.011	0.046	0.981	0.018	0.665	0.005
30 percent	0.931	0.011	0.148	0.935	0.020	0.654	0.005
40 percent	0.854	0.009	0.266	0.857	0.017	0.655	0.005
50 percent	0.655	0.008	0.572	0.655	0.015	0.673	0.005

TABLE E.17: A table showing the loan-level results for the random forest models trained on the undersampled data sets for full prepayment, tested on the out-of-sample data.

E.2.3 Arbitrage Prepayment Results

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	0.645	0.645	0.645	0.645
20 percent	0.309	0.308	0.309	0.308
30 percent	0.194	0.191	0.194	0.191
40 percent	0.225	0.224	0.224	0.224
50 percent	0.358	0.347	0.356	0.345

TABLE E.18: A table showing the portfolio level results for the random forest models trained on the undersampled data sets for arbitrage prepayment, tested on the out-of-sample data.

	Accuracy	Precision	Recall	Specificity	F1	AUC	Brier
10 percent	0.980	0.011	0.448	0.980	0.022	0.923	0.0004
20 percent	0.937	0.005	0.649	0.937	0.010	0.919	0.0004
30 percent	0.901	0.004	0.740	0.901	0.007	0.918	0.0004
40 percent	0.862	0.003	0.812	0.862	0.006	0.915	0.0004
50 percent	0.823	0.002	0.864	0.823	0.005	0.909	0.005

TABLE E.19: Random Forest improved model loan-level metrics for arbitrage prepayment

E.2.4 Total Prepayment results

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	1.396	1.156	1.232	1.007
20 percent	1.277	1.044	1.095	0.856
30 percent	1.196	0.966	1.011	0.792
40 percent	1.114	0.888	1.017	0.786
50 percent	1.100	0.871	0.989	0.759

TABLE E.20: A table showing the portfolio level results for the random forest models trained on the undersampled data sets for total prepayment, tested on the in-sample data.

E.3 Neural Network Results

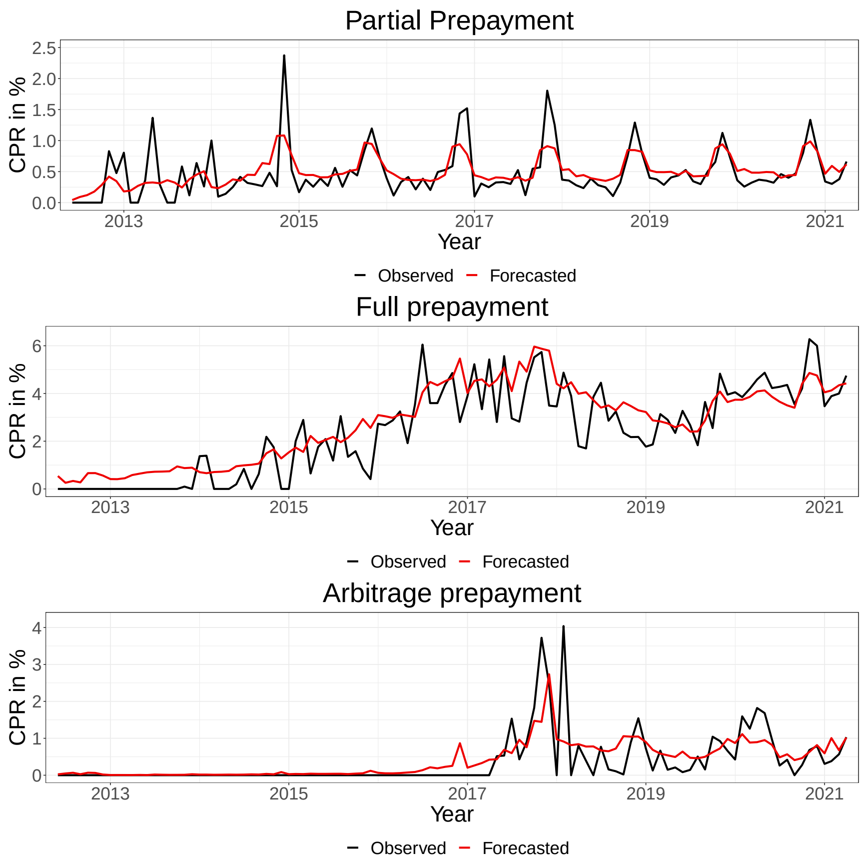

FIGURE E.4: Plot showing the CPR for all three types of prepayment for the baseline neural network model

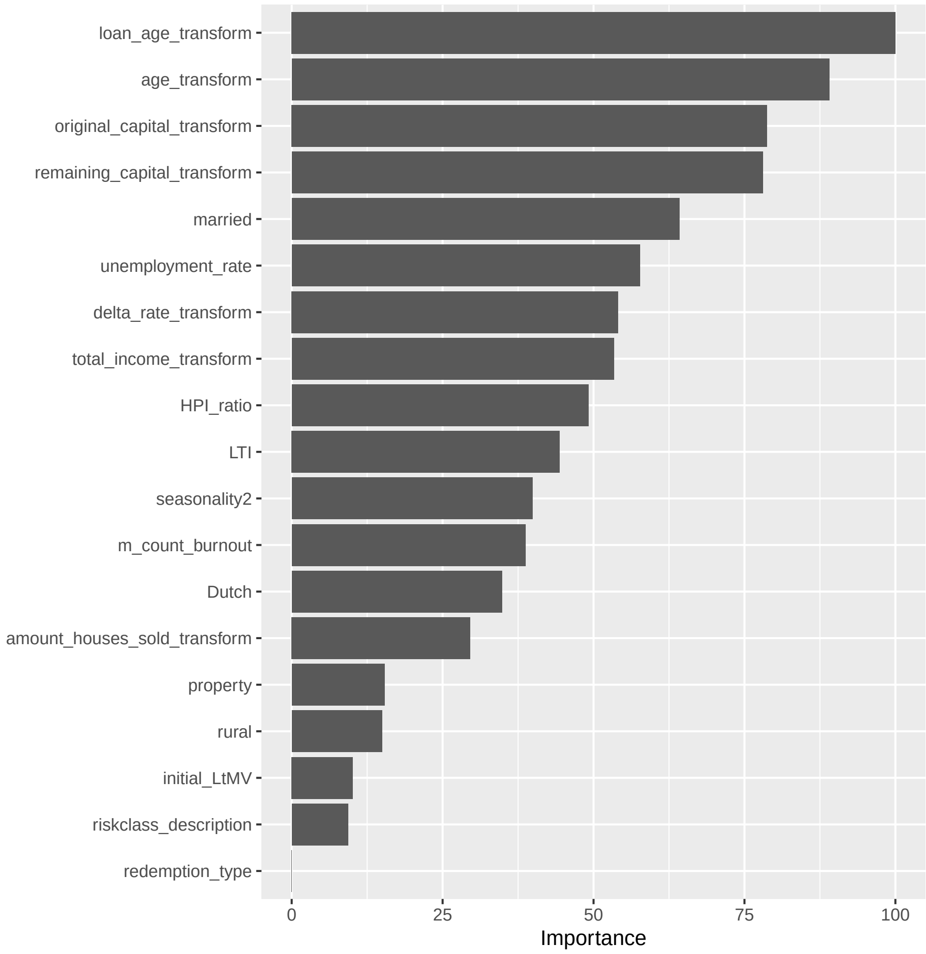

FIGURE E.5: A figure showing the variable importance for all variables used in the baseline full prepayment model

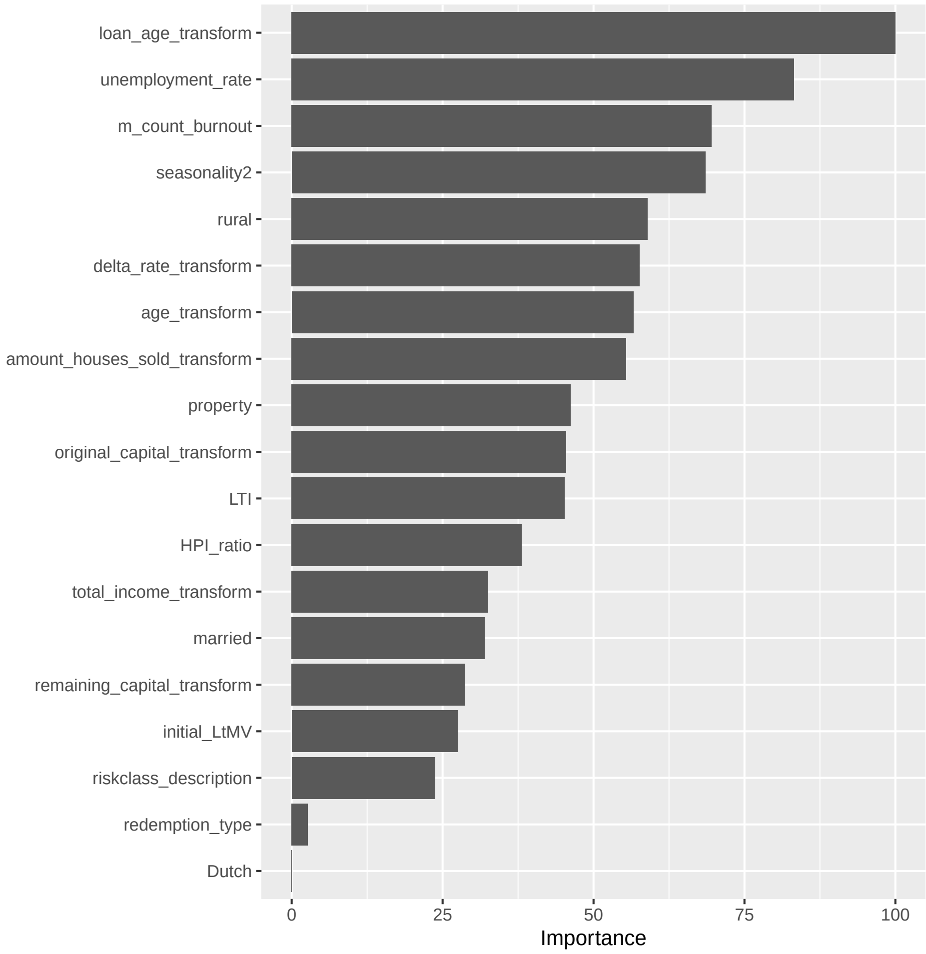

FIGURE E.6: A figure showing the variable importance for all variables used in the baseline arbitrage prepayment model

	RMSE	MAE	Weighted RMSE	Weighted MAE
Partial	0.126	0.125	0.125	0.125
Full	0.864	0.843	0.867	0.847
Arbitrage	0.854	0.847	0.851	0.844
Total	0.315	0.301	0.314	0.299

TABLE E.21: Table showing the portfolio metrics for the out-of-sample test data of the baseline neural network model for all types of prepayment using.

E.3.1 Partial Prepayment results

	RMSE	MAE	Weighted RMSE	Weighted MAE
10 percent	0.296	0.216	0.216	0.160
20 percent	0.301	0.215	0.220	0.163
30 percent	0.303	0.218	0.220	0.164
40 percent	0.291	0.214	0.211	0.160
50 percent	0.299	0.217	0.218	0.166

TABLE E.22: Neural network improved model portfolio metrics for partial prepayment

	RMSE	MAE	Weighted RMSE	Weighted.MAE
10 percent	0.331	0.331	0.110	0.331
20 percent	0.246	0.247	0.246	0.246
30 percent	0.458	0.458	0.458	0.458
40 percent	0.690	0.684	0.688	0.682
50 percent	2.172	2.169	2.170	2.167

TABLE E.23: Neural network improved model portfolio metrics for partial prepayment tested on out-of-sample data.

	Accuracy	Precision	Recall	Specificity	F1-score	AUC	Brier
10 percent	0.980	0.084	0.182	0.986	0.058	0.800	0.007
20 percent	0.966	0.061	0.262	0.971	0.050	0.776	0.007
30 percent	0.922	0.041	0.428	0.926	0.037	0.794	0.007
40 percent	0.862	0.031	0.589	0.864	0.029	0.795	0.007
50 percent	0.763	0.021	0.678	0.764	0.020	0.784	0.007

TABLE E.24: Neural network improved model loan-level metrics for partial prepayment

E.3.2 Full Prepayment results

	RMSE	MAE	Weighted RMSE	Weighted MAE
10 percent	0.927	0.746	0.718	0.671
20 percent	1.022	0.805	0.890	0.692
30 percent	1.134	0.893	0.954	0.711
40 percent	1.097	0.839	1.011	0.754
50 percent	1.001	0.799	0.873	0.668

TABLE E.25: Neural network improved model portfolio metrics for full prepayment

	RMSE	MAE	Weighted RMSE	Weighted MAE
10 percent	3.000	2.997	9.013	2.999
20 percent	0.575	0.567	0.577	0.569
30 percent	3.016	3.014	3.018	3.017
40 percent	1.097	0.839	1.011	0.754
50 percent	12.135	12.134	12.133	12.132

TABLE E.26: Neural network improved model portfolio metrics for full prepayment tested on out of sample data

	Accuracy	Precision	Recall	Specificity	F1 score	AUC	Brier score
10 percent	0.996	0.044	0.012	0.999	0.009	0.768	0.003
20 percent	0.962	0.015	0.166	0.965	0.014	0.760	0.003
30 percent	0.885	0.011	0.382	0.887	0.011	0.760	0.003
40 percent	0.784	0.009	0.569	0.785	0.009	0.763	0.003
50 percent	0.639	0.007	0.736	0.639	0.007	0.761	0.003

TABLE E.27: Neural network improved model loan-level metrics for full prepayment

E.3.3 Arbitrage Prepayment results

	RMSE	MAE	Weighted RMSE	Weighted MAE
10 percent	0.522	0.288	0.517	0.375
20 percent	0.526	0.298	0.526	0.364
30 percent	0.559	0.309	0.543	0.361
40 percent	0.531	0.287	0.533	0.364
50 percent	0.552	0.311	0.565	0.405

TABLE E.28: Neural network improved model portfolio metrics for arbitrage prepayment

	RMSE	MAE	Weighted RMSE	Weighted MAE
10 percent	0.167	0.163	0.165	0.162
20 percent	0.732	0.728	0.731	0.727
30 percent	1.402	1.402	1.402	1.402
40 percent	3.430	3.424	3.434	3.428
50 percent	4.584	4.575	4.589	4.580

TABLE E.29: A table showing the loan-level results for the neural network models trained on the undersampled data sets for arbitrage prepayment, tested on the out-of-sample data.

	Accuracy	Precision	Recall	Specificity	F1 score	AUC	Brier
10 percent	0.976	0.007	0.318	0.976	0.006	0.871	0.000005
20 percent	0.931	0.004	0.500	0.931	0.004	0.858	0.00001
30 percent	0.896	0.003	0.630	0.896	0.003	0.872	0.000004
40 percent	0.846	0.002	0.721	0.846	0.002	0.868	0.000004
50 percent	0.795	0.002	0.805	0.795	0.002	0.869	0.000004

TABLE E.30: Neural network improved model loan-level metrics for arbitrage prepayment

E.3.4 Total Prepayment results

	RMSE	MAE	weighted RMSE	weighted MAE
10 percent	1.190	0.960	1.214	0.881
20 percent	1.309	1.034	1.205	0.958
30 percent	1.334	1.093	1.183	0.948
40 percent	1.290	1.048	1.242	0.993
50 percent	1.280	1.042	1.182	0.930

TABLE E.31: A table showing the portfolio level results for the neural network models trained on the undersampled data sets for total prepayment, tested on the in-sample data.

E.4 Cash Flow Estimation Results

Year	Allianz model error (%)	Base LR error (%)	LR 50% error (%)	Base RF error (%)	RF 50% error (%)	Base NN error (%)	NN 10% error (%)
2014	0.7	2.5	2.7	2.4	1.7	0.9	1.3
2015	2.4	2.8	2.7	2.8	1.8	1.4	1.8
2016	-0.0	0.4	0.2	1.5	0.2	0.8	0.4
2017	-1.3	-0.5	-0.7	1.4	-0.3	0.1	0.0
2018	-2.2	2.7	2.2	2.4	1.7	2.7	2.6
2019	-2.7	4.2	3.9	3.5	1.4	3.4	3.1
2020	-10.3	-4.2	-3.9	-0.6	-4.3	-3.8	-2.2
2021	-2.1	1.5	2.0	3.6	0.8	1.4	2.4
mean	-1.9	1.2	1.1	2.1	0.4	0.8	1.2
Absolute mean	2.7	2.4	2.3	2.3	1.5	1.8	1.7

TABLE E.32: A table showing the yearly cash flow estimation weighted error for multiple models compared to the observed prepayment cash flow.