Department of Econometrics
Eramsus School of Economics

Quantitative Finance Master Thesis

Mortgage Prepayments

Charles Ewing, 573203

Academic Supervisor : **Dr. H.J.W.G. Kole**
Firm Supervisor: **Menno van Boven**
Firm Supervisor: **Victor Sorokin**

Abstract

Mortgage loans constitute a substantial amount of most Dutch banks' balance sheets. In a mortgage contract, the borrower has the option to repay more than the contractual amount at any point before the due date. These unexpected cash flows make hedging optimally somewhat tricky, which affects the Asset and Liability Management of a bank; for this reason, it is essential to forecast them accurately until the interest reset date. As the forecasts are needed until the reset date, one must make a multi-step forecast over multiple periods into the future. Machine Learning models are growing in popularity due to their primarily non-parametric approach and ability to capture non-linearity accurately. Despite their plaudits, they have some shortcomings. One of which is the lack of interpretability. Machine Learning models and, in particular, Random Forests generally become less interpretable as they increase inaccuracy. Interpretability of models is critical in the Banking industry, so it can be seen that the models are fair and do not act on historical biases in the data used to train the model. This thesis aims to make a highly accurate and interpretable multi-step mortgage prepayment forecasting model so Knab can optimally hedge prepayment risk in a compliance-friendly way.

Contents

1 Introduction

As of mid-2020, the total outstanding amount of mortgage debt in the Netherlands is over €740 Billion according to Sta. With the large market, it is essential for any financial institution with mortgage loans or mortgage-linked products to model mortgage payments and, in particular, banks. Traditionally, a bank generates profit by providing long-term loans (such as mortgages) funded by short-term liabilities (deposits), earning money from the spread between the rates. When a bank provides a mortgage loan to a borrower, they are exposed to three main components of risk in the loan. The first risk is that the borrower will default on the loan, also known as credit risk. Second is the liquidity risk from the loan because the bank does not know how much capital will incoming to due to prepayments. Lastly, interest rate risk is the potential for investment losses resulting from a change in interest rates.

Most banks are generally risk-averse Nishiyama (2007); therefore, they hedge these risks in order to lock in the profit from the spreads. This thesis will focus on interest rate risk, and to hedge this, the bank will enter into an interest rate swap. Dutch regulators allow mortgagors to prepay their mortgage early. Therefore, the cash flows over the mortgage loan term may differ from the scheduled amortisation plan. Hence, the interest rate swap will no longer be hedged as the bank will now pay back more interest than they will receive. Therefore to avoid this over-hedging, the mortgage lender wants to forecast prepayments so they can be taken into account in the hedge.

The prepayment model must be both accurate and interpretable. Accuracy will allow it to fully capture cash flows so that the bank can hedge appropriately and model the market value of the mortgage. The interpretability of the model is also highly important as the model must be able to explain why it gives a specific prepayment rate for each loan part. It is crucial from both a compliance and moral perspective, as if the model is predicting a prepayment rate that is more difficult to hedge for the bank, this, in turn, would make the mortgage more expensive for the client. Therefore, the model must show that it does not act on racial, sexist or any other bias that could have existed in the historical data that has been used to train the model.

This thesis aims to develop a highly interpretable, time-efficient model for Conditional Prepayment Rate (CPR) for each loan part for each month until the interest reset date. For Knab, the model must be both accurate and time-efficient. Accurate so that interest rate risk can be optimally hedged and time-efficient because Knab updates model parameters monthly for various models; therefore, models must run relatively fast to continue to be feasible.

In practice, most banks use Multinomial Logistic Regression (MNL) to find the probability of different prepayment types before using an expectation approach to obtain CPR on a loan part level. The first issue with this model is that it assumes a linear relationship between the response variable and the explanatory variables; therefore, it cannot capture non-linear relationships and clustering that may occur. The second shortcoming is the independent and identically distributed (i.i.d) assumption for all loan parts in the model. This is not the case because loan parts of the same borrower are correlated as the borrower does not treat his loans independently but acts on which is most beneficial on a portfolio level.

Further models have been investigated to overcome the MNL's inability to capture non-linear relationships and clustering. The first is the Multinomial Logistic Regression Tree (MNL Tree) model, which will make a decision tree-like structure with different MNL models that fit each terminal node of the tree. If we split the data into groups that behave differently, then each MNL model will better explain more data variance as the modelled data will act more similar. Furthermore, whilst doing the splits, more of the non-linearity can be captured because of the non-parametric nature of decision trees, meaning no distribution is assumed. The second model used is a Random Forest (RF); this is a non-parametric model that captures non-linear relationships. However, RF's are often referred to

as "Black Boxes" due to their lack of transparency when producing output. For this reason, its use when modelling CPR in practice is limited. The last classification model is the Multinomial Logistic Regression Forest model, which aims to build upon the advantages of the MNL Tree whilst making the model more stable and hence more accurate. To overcome the second shortcoming of the i.i.d assumption for the MNL model, modelling on borrower level data instead of loan part level has been investigated. Assuming each borrower acts i.i.d is more rational than assuming loan parts are.
The expectation approach commonly used in practice could potentially lead to significant errors as estimated values are multiplied. Therefore, a regression approach has been considered using the regression equivalent of each classification model considered. Hence, Linear Regression, Linear Regression Tree, Linear Regression Forest and a regression Random Forest.
Both deterministic and stochastic multi-step forecasting methods has been evaluated. Stochastic forecasting can account for the path dependency of mortgage prepayments and stochastic independent variables; both cannot be accounted for when things are modelling deterministically. However, modelling stochastic may not be computationally feasible with a large data set. Additionally, implementing a stochastic model within Knab's Asset-Liability Management system is not possible. Therefore, the focus of the thesis will be to optimise the deterministic model accuracy whilst investigating the possible loss in accuracy versus the stochastic model.

1.1 What this research adds to the literature

The contents of this paper make the following contributions to literature.

- Modelling CPR using regression.
- Comparison between modelling CPR using traditional expectation versus regression approach.
- Modelling prepayments on a mortgagor level.
- Comparison between deterministic and stochastic multi-step forecasting prepayments from both accuracy and computational point of view.
- Modelling multi-step forecasting prepayment with Random Forest.

2 General mortgage prepayment background

2.1 Types of mortgages

In the Netherlands, a borrower can have several different types of mortgages. Moreover, a mortgage can be split into various loan parts of different types. The main different mortgage types are as follows:

- Linear:

  Pay a fixed amount of the loan each month; however, the borrower's interest decreases linearly, implying that the total amount paid each month decreases with duration.

- Annuity:

  The total amount paid each month remains constant throughout the mortgage. The interest paid each month decreases with duration, implying that the loan paid each month increases with duration.

- Interest – only:

  The payment only goes towards the interest for a specified amount of time. After the specified time, both the interest and loan is paid off over the rest of the duration of the mortgage.

- Savings and investment:

  The mortgage is linked to a savings account into which the borrower deposits cash each month. The amount of interest for the savings account is equal to the interest paid on the mortgage, so they cancel each other out. At the end of the duration of the mortgage, the loan is then paid off in full (from the savings account).

  As of 1 January 2013, mortgage interest rate tax relief has only been permitted on annuity or linear mortgages.

2.2 Types of prepayment

Prepayment is an accounting term for settling a debt or instalment loan before its official due date. In the Netherlands, mortgagors' loans can prepay the loan at any time within the term. A borrower can make multiple types of mortgage prepayment, which have differing effects on CPR.

- Full Prepayment:

  The borrower repays all remaining notional on the loan prior to its contractual due date. Therefore, the loan ceases to exist after this point. Usually occurs due to house sale.

- Partial Prepayment:

  The borrower repays more than the contractual amount; however, not the full notional. Mainly occurs to lower LTV (lowering mortgage rate), tax reasons or to generally reduce notional owed.

- Non-Cash Prepayment:

  The borrower requests that their mortgage rate be reset and the interest reset date changed. The borrower will pay a penalty that compensates for the loss in market value in the loan for this prepayment; this offset amount saved by a lower interest rate; therefore, a borrower will usually only make non-cash prepayment if they want to extend the interest reset date (possibly to take advantage of low-interest rates as they think rates may increase in future).

2.3 Reasons for prepayment

- Property Sale:

  When a property sale occurs, this would usually lead to a prepayment because the cash from the sale can be used to pay off outstanding mortgage debt. Commonly, the date of the house sale does not line up with the interest reset or maturity date of the mortgage, generating a prepayment. A client can also opt to reattribute the mortgage to the new property. This will mainly be the case when the mortgage rate in the market is higher than the previous rate or due to favourable interest guarantees in the old policy. This would not lead to a prepayment.

- Mortgage Refinancing:

  This occurs when the borrower pays off the existing mortgage or loan part and replaces it with a new one. Borrowers usually do this for economic reasons, and the new loan has a lower mortgage interest rate than the current mortgage rate. A less common reason is to access increased equity of the property. Refinancing can occur internally (the client does not change the mortgage provider) or externally (the client does change the mortgage provider).

- Partial prepayments:

  A client with sufficient personal funds can reduce mortgage debt by repaying part of the outstanding mortgage. Reasons to do so could include:

  1. Economic motives (e.g. the savings interest rate is lower than the mortgage rate, or to bring the savings balance below the wealth tax hurdle),
  2. Risk perception (the mortgage debt is higher than the value of the residence, the result of which means a residual debt would remain after any forced sale)

2.4 Prepayment penalty

When borrowers make a prepayment, they will usually pay a penalty fee. The penalty interest compensates the mortgage provider for missing future interest margin income. However, there are some instances where a prepayment comes penalty-free.

1. The prepayment amount is within the penalty-free threshold. This amount is 10% of the original principal amount per calendar year for Knab mortgages.
2. Full prepayment is made due to a property sale.
3. Any amount of payment made at the interest reset date or maturity date of the mortgage is not classified as prepayment and is therefore penalty-free.
4. If current mortgage interest is higher than the borrower's mortgage rate.
5. In situations where prepayment is in the form of an insurance payout after the loss of the property due to damage or using a term life insurance payout after mortality.
6. If prepayment is on a bridge loan (a short-term loan given to borrowers so they can meet current obligations usually charged at high rates).

3 Literature

3.1 Review of prepayment motives

The four most common drivers known for mortgage prepayment behaviour were first discussed by Brennan and Schwartz (1985) and are as follows:
Firstly, Refinance incentive is the difference between the current mortgage rate and the mortgagor's rate. If the mortgagor's rate is much higher than the current rate, generally, there is a higher probability of prepayment. This effect has been shown to be significant in Richard and Roll (1989), Perry et al. (2001), Charlier and Van Bussel (2003), Meis (2015). Secondly, Burnout explains how after the mortgagor has been exposed to an interest rate environment in which it is optimal to prepay, and they did not, following this in the future, the mortgagor has a lower probability of prepayment. Kang and Zenios (1992) has shown there is a difference in prepayment behaviour between "fast" and "slow" borrowers. Thirdly, Bussel (1998), Alink (2002), Charlier and Van Bussel (2003), Hayre (2003) have shown that there is clear seasonality in prepayment behaviour. Reasons for this seasonality could be that people usually move house in the spring/summer, which often coincides with prepayment. Fourthly, Kang and Zenios (1992) , Clapp et al. (2001), Deng et al. (2000) have shown seasoning (mortgage age) as a significant driver of prepayment. CPR vs mortgage age would generally be a skewed bell curve that stabilises at expiry. When mortgagors have just received the loan, they will be less likely to prepay or move house as utility is typically maximised upon receipt of the loan. As mortgage age increases, the probability they have gained excess money to prepay or move house has increased. CPR then levels off as the probability of the mortgagor moving house decreases.
The rest of the prepayment motives specified in the literature can be assigned to three categories summarised in the table on the next page. Firstly loan-specific drivers are highlighted in yellow, which are pure to do with the dynamics of the loan itself; next are borrower-specific drivers highlighted in blue, which are based on the characteristics of the mortgagor taking the loan. Lastly, Macroeconomic drivers are highlighted in brown, which are related to the current state of the economy.

3.2 Review of current literature

Existing literature models prepayment in two different ways; the first is on a portfolio level. The most simplistic model assumes a flat prepayment rate for all mortgages; this neglects all prepayment determinants explained in 3.1. Richard and Roll (1989) use "The Goldman Sachs model" otherwise known as the four-factor model, which multiplies the four main prepayment drivers, refinancing incentive, age of the mortgage, seasonality, Premium burnout to give CPR. This CPR is applied to all loans in the portfolio. This model is often used in industry, along with the PSA Standard Prepayment Model Hayre (2001), which linearly increases at a rate 0.2% from 0% at the origin of the loan until month 30 then remains at a constant 6% prepayment rate. This is done to capture that mortgagors are less likely to prepay or refinance when they have just begun the mortgage, a.k.a the seasoning effect.
A significant issue with modelling on a portfolio level is that it does not capture the heterogeneous of the different mortgagor's prepayment rates. This can be incorporated by using cohort analysis which segments the mortgages into (more) homogeneous groups. Then the techniques described above can be used on the various cohorts. However one should be careful that each cohort contains a sufficient number of loans to obtain meaningful results or, on the contrary, not too many such that it can no longer capture the heterogeneity.

Prepayment motives			
Prepayment motive	Neg/Pos	Description	Literature
Prepayment penalties poxy	-	This is a normalized proxy for size of penalty borrower would face if they where to prepay. This would disincentives borrowers to prepay.	Charlier and Van Bussel (2003), Subotniaya (2018)
Lock-in effect	-	If locked in to a low rate compared to market rate mortgagors that move houses will keep there mortgage if portable or transfer to new if assumable. If neither and difference between rates in large enough this could discourage moving, lowering CPR.	Saito (2018), Hayre (2003), Clapp et al. (2001)
Loan-to-Value (LTV)	-	Literature finds a a lower LTV implies a higher CRP this could be due to the seasoning effect. However, this relationship seems counter intuitive because the higher the LTV the higher the mortgage rate which would imply lower refinancing incentive.	Clapp et al. (2001), Alink (2002), Deng et al. (2000), Calhoun and Deng (2002)
Down payment at origin	-	Larger the down payment to original ratio the lower the coupon and lower the CPR.	Goodarzi et al. (1998)
Loan through intermediaries	+	More likely to prepay because intermediary earn commission to alert when optimal to prepay.	Alink (2002), Saito (2018)
National Mortgage Guarantee (NHG)	-	NHG Mortgages are granted a lower coupon as they are insured by the government. A lower rate implies lower refinancing incentive therefore lower CPR.	Meis (2015)
Time till reset date	+	There is greater incentive to prepay earlier because cash saved will be for longer time.	Alink (2002)
Market value of loan	+	This is because if market value increase this has mean that rate has rose on mortgage this will imply higher prepayment.	Clapp et al. (2001)
Original loan value	+	If original loan is higher this will imply that borrower needs to borrow more meaning that they are less likely to have free cash to be able to prepay mortgage.	Clapp et al. (2001), Charlier and Bussel (2003)
Property Type	Depends	Family homes have have been shown to have higher prepayment rate than commercial property's this could be due to that family's would move house more often.	Charlier and Bussel (2003), Charlier and Van Bussel (2003)
Number of partial prepayments	+	Higher the number of prepayments the more likely they are to do it again because they are clearly pay close attention to market montage rates.	Wanders et al. (2021)
Number of resets	+	For a similar reason to above the borrower pays close attention to his mortgage.	Wanders et al. (2021)
Indicator if defaulted before	-	If borrower has defaulted before then most likely they has money issues therefore don't have an excess of cash to prepay.	Wanders et al. (2021)
Indicator if prepaid before	+	If borrower prepaid before they will more likely do it again as they pay attention to market rates and will likely be able to get enough cash to prepay again.	Mieras et al. (2021)
Normalized price incentive	+	This is a proxy for length of till rest times by refinance incentive. This accounts for how much mortgagor would be saving and for how long implying a positive relationship.	Expert opinion
Age of borrower	+	It is found to be postive becasue as borrower gets older they will more likely have more excess money to prepay and will be more likely selling there house for a bigger or smaller house (depending on age), therefore they will be more likely to prepay.	Charlier and Van Bussel (2003), Alink (2002), Clapp et al. (2001)
Location of house	Depends	It has been shown that different areas have different CPR's in general. This could be because certain areas are cheaper houses that are sold more often for "upgrades".	Sirignano et al. (2018), Alink (2002)
Income	+	Generally more income would imply more ability to prepay however, this is only provided at loan origin therefore inaccurate.	Clapp et al. (2001),Deng et al. (2000)
Creditworthiness (FICO Score)	+	This has been shown to be positively related to CPR beacuse generally higher FICO means your better at managing money so will prepay when optimal.	Alink (2002), Clapp et al. (2001)
Minority indicator	-	Non-white indicator has been shown to have a negative relationship with CPR.	Clapp et al. (2001)
Media effect	+	Recent "all-time low" rates have a higher CPR due to the media attention.	Hayre (2003)
Unemployment rate	+	Higher unemployment rate more chance the borrower will have to sell or refinance the house therefore implying prepayment.	Sirignano et al. (2018), Meis (2015), Clapp and LaCour-Little (2001)
Divorce rate	+	If divorce occurs then this will likely result in a house sale implying a prepayment.	Meis (2015), Clapp and LaCour-Little (2001)
House price inflation	+	When house price appreciation occurs home sales increase implying increase in CPR.	Meis (2015)

Applying cohort analysis captures many prepayment drivers and more granular links between sub-populations of mortgages and prepayment behaviour than on portfolio level. However, Chinchalkar and Stein (2010) shows modelling at a loan level can better capture the drivers and heterogeneous of prepayment and capture the market value of each loan, which cannot be done at a portfolio or cohort level. The previous literature at the loan level can be thought of in two different categories: optional theoretical models and exogenous models Charlier and Van Bussel (2003).

3.2.1 Option Theoretical Models

Findley and Capozza (2003) were first to apply an optimal option-theoretic model to mortgage prepayment. They viewed it as a call option that is only repaid if, and only if, the option is in the money (contractual rates are higher than current interest rates). This model assumes that all mortgagors are rational and have the capabilities to do so. It has been shown that in reality, mortgagor's do not prepay optimally because prepayment may be influenced by non-financial factors as stated in 3.1 and therefore may not always prepay when optimal to do so. Kau and Keenan (1995) claim this is because housing can be seen as consumption goods instead of financial assets. There have been many extensions of the theoretical option model to try and explain the irrational mortgagor behaviour. Dunn and McConnell (1981) and Brennan and Schwartz (1985) have incorporated a Poisson-driven process term; however, the mortgage prepayments these yields are not necessarily non-optimal. Archer and Ling (1993) show that some prepayments occur when 'out-of-the money' and mortgagors do not always prepay when it may be optimal to do so. They conclude that transaction costs may cause the lack of prepayments for 'in-the-money mortgages. However, Stanton (1995) found that when heterogeneity prepayment costs are modelled, the observed prepayment costs are considerably lower than ones implied by the model.

3.2.2 Exogenous Models

The most common model used in the banking industry is the Multinomial Logit (MNL) model, academically Campbell and Dietrich (1983), Zorn and Lea (1989), Capone and Cunningham (1992) and Calhoun and Deng (2002) have shown significant predictability power for predicting prepayments. However, the MNL model assumes observations are independent, which is likely not the case, especially when looking at the loan-part level. Pravinvongvuth and Chen (2005) conclude that the main drawback of the MNL model is its inability to account for the correlation between mortgage path and perception variance of mortgages of differing length (due to term or prepayments). A popular model in the literature is the Proportional Hazards (PH) model, which Green and Shoven (1986), Schwartz and Torous (1989) have used to predict prepayments. It has been shown that the MNL AND PH model are very similar; in fact, Shumway (2001) has shown that the PH model can be interpreted as a logit model done by firm-year. The PH model's primary disadvantage is its inability to model non-terminating and terminating events such as a partial and full prepayment. A further reason why one may choose MNL over PH is that Clapp et al. (2001), Clapp and An (2006) have shown that the proportionality assumption for the hazard rate may not be true.
More recently Lijmbach et al. (2021) predicted prepayments using the recursive partitioning algorithm designed by Strobl et al. (2008), to extend the MNL model to an MNL-Tree model in which MNL models are fit to the observations in terminal nodes of a decision tree. The reason being is that the tree will help capture clustering in the data, which an MNL model is unable to do on its own. In addition, it helps exploit the non-linearity's which exist when modelling mortgage prepayment.
A less common approach to modelling prepayment would be to use a Markov model, in which the transition probability between states depends on the current state. Although this is strong from

a theoretical point of view Meis (2015) concluded that it does not yield significant results for predicting prepayment. An additional, less common approach would be to model prepayment from a Bayesian point of view, Bhattacharya et al. (2019), Popova et al. (2008) constructed a Bayesian mixture model for which they use a Markov Chain Monte Carlo to estimate parameters. This issue with the Bayesian approach is the model is dependent on the quality of available covariates.
Machine learning methods are increasing in popularity, but work on its application to prepayment is still limited. Sirignano (2016) shows the relationship between prepayment behaviour and risk factors is non-linear. This implies that machine learning techniques should be more appropriate for modelling as most of them do not assume a distribution , unlike the traditional prepayment model. Early adoption of a machine learning model by LaCour-Little et al. (2002),shows how a Non-parametric Kernel Regression can outperform a logit model. Sirignano (2016) has shown that Neural Networks have significant predictive power for prepayments for both pool-level and loan-level data. Saito (2018) has shown that Random forests (RF) outperform Neural Networks. Furthermore, Blumenstock et al. (2020), and Saito (2018) have shown both models outperform hazard and logistic models, respectively, in addition to being, are robust across different periods, including stressed periods.

3.2.3 Improving the interpretability of "Black Boxes"

A drawback to the machine learning models mentioned is their lack of "interpretability". Lijmbach et al. (2021) have attempted to solve this using a Multinomial Logit Regression Tree. The model recursive partitions the data similarly to a small decision tree then fits the MNL model to the data in terminal nodes. Each node MNL model will have its own parameters. The idea comes from Strobl et al. (2008) and is designed in order to capture the non-linear relationships and clustering whist maintaining interpretability. Lijmbach et al. (2021) have been unable to outperform MNL on loan-level; however, it has on a portfolio level. Taking into account, the limited small tree size Lijmbach et al. (2021) have used due to time constraints makes this model is very promising.
Saito (2018) used Partial Dependence Plots to improve the interpretability of the prepayment forecasts using the RF. A plot for each input variable is made by varying the dependant variant value whilst keeping other variables the same. A general relationship between the output variable and the input variable we are considering can be observed from the graph. Although Partial Dependence Plots give a good insight into the general working of the model Goldstein et al. (2014) developed Individual conditional expectation plots which allow the ability to drill down to individual observations, allowing the ability to identify subgroups and corrections between input variables.
Guidotti et al. (2018) conducted a review of all methods for explaining Black Boxes. These methods can be broken into three categories: model explanation, outcome explanation, and model inspection. The model inspection provides a representation for how models work like Saito (2018) and Goldstein et al. (2014) have done. The outcome explanation uses a local point of view to the interpretable model. Examples of these include Saliency Masks and Local Interpretable Model-Agnostic Explanations; these are heavily used in image recognition due to their ability to find the most informative regions Phang et al. (2020). Model explanation gives globally interpretable models that behave similarly to black boxes but are more interpretable for humans. The global nature of these models makes them more appropriate for modelling prepayment behaviour.
There are two main methods of doing this; the first is via single tree approximation; however single trees can still be challenging to interpret depending on the depth and other hyper-parameters. The second way is through Rule Extraction. Bénard et al. (2020) developed SIRUS (Stable and Interpretable Rule Set), which generates a (modified) random forest that creates a rule for each hyperrectangle of each tree. It ranks the importance of the rules based on how often they appear. The

most significant rules are kept and aggregated together, making a shortlist of highly interpretable rules. Akyüz and Birbil (2021) have created RUX (Rule Extraction), which uses Linear programming to extract the rules most critical for classification. Furthermore, they have developed RUG (Rule Generation), which uses Linear programming to create an interpretable model from scratch. Both Bénard et al. (2020) and Akyüz and Birbil (2021) have found accuracy similar to noninterpretable models with a short number of rules. However, when using Akyüz and Birbil (2021) models one could explore whether a model built from scratch to be "transparent" will yield better results for predicting prepayments.

3.2.4 Multi-step prepayment forecasting

To value Mortgage-Backed Securities (MBS), the standard method is to forecast the cash flows of the MBS, discount them, and add them together. In practice, this is usually done on a loan level and using Monte-Carlo simulation, then aggregating the results as can be seen in Schwartz and Torus (1989), Busschers (2011), Kang and Zenios (1992). This can allow for path dependency of mortgages, Wanders et al. (2021) have included state dependency variables in a Monte Carlo simulation to forecast prepayments over three years. The main issue with the Monte-Carlo simulation is the computation time, especially when running on loan-level data. There is no current literature readily available to evaluate that forecasts CPR over the mortgage term deterministically.

4 Data

This section gives an insight to the different methods for treating the data and the motivation for doing so. Then a short description of the data set is given followed by explanation for how and why adjustments are made for unbalanced data. Please note, that in this section and beyond the scales on graphs have been excluded for the banks privacy reasons.

4.1 Data treatment

This section describes the two different methods of data aggregation that will be used and compared for all models in the methodology. The reasons for the aggregated to appropriate level data will be split into training, testing and validation sets. The training set is the data that the models will be trained on, the testing set is the data the model will be evaluated on and performance metrics are generated from. The validation set is for hyperparameter tuning, which is further explained in 5.2.5.

4.1.1 Loan-level

In academic literature and industry, prepayment is usually modelled on loan-level data described in 3.2. There are a few reasons for this; firstly, the most granular level the data goes to, giving the maximum information possible. The Central Limit Theorem states that the larger the size of the sample, the smaller the confidence band around the results , implying that generally, the more data, the more accurate the results are. Furthermore, it can handle the heterogeneous of the different mortgagor's prepayment rates. It is on the loan part level already, so no further data aggregation will be needed.

4.1.2 Mortgage-level

The main issue with using an MNL model is the assumption of the data being independent and identically distributed. This is not the case for our data because one mortgagor usually has multiple loan parts; these loan parts will therefore be highly correlated. Therefore, by aggregating data to mortgagor-level would make the i.i.d assumption more appropriate. As all borrower information is the same, in the process of combining loan parts, little information is lost apart from the method of prepayment. The only information that is being aggregated is the prepayment method (i.e to what loan part(s)).
As described in 3.2.1 prepayment is not always optimal; however, it is rational to think, given the mortgagor prepays, they will optimally make the prepayment, so the information loss may not matter. Therefore to aggregate the prepayment amount back from mortgagor CPR to loan-part CPR, use a model which assumes rational prepayment. However, one could argue with this assumption because tax incentives may mean the optimal way for one mortgagor may not be the optimal way for others to prepay. Given the out of date nature of income data, finding the current tax band is difficult. A new date called "restructuring date" can also be formed when new loan parts are split up or combined. This may give some further insight into prepayment behaviour.

4.2 Data description

The data set has mainly been taken from Knab's internal database, where the data is monthly and ranges from October 2015 to October 2021. Knab is Aegon's Bank and they only buy and sell mortgages to Aegon, therefore we have also used data from Aegon to verify if the loan part that has disappeared from Knab's data has truly fully prepaid or has loan been sold to Aegon. The first 4

years of data will be used to fit the model and the last two years will then be used to evaluate the long-term forecasting accuracy of the model.

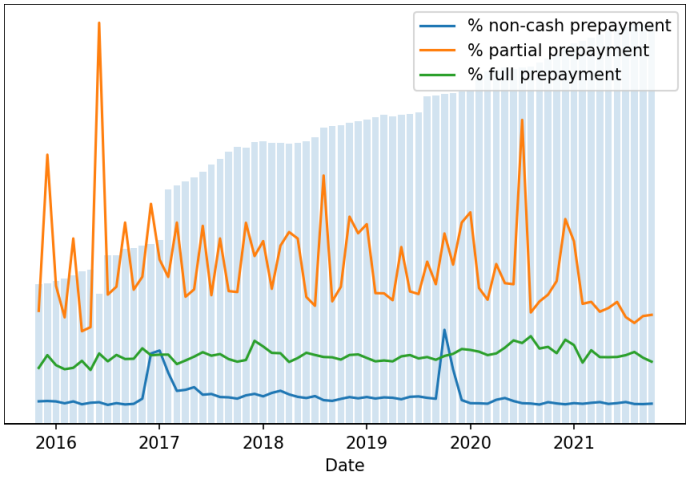


Figure 1: Summary of number of loans and percentage of prepayment types over time

Figure 2 shows seasonality in CPR, with it increasing at the end and middle of the year. The common reason for the increase in CPR in the middle of the year is due to the 8% (of yearly salary) holiday allowance given by employers in May. This increase in disposal would incentivize borrowers to prepay. Furthermore, the expected reason for the increase in CPR at the end of the year is due to December 31st being the end of fiscal year in the Netherlands. Figure 1 shows a similar seasonality in partial prepayments increasing at the end and middle of the year, showing partial prepayment is the main driver for seasonality in CPR. Non-cash prepayments are quite stable, and show no obvious seasonality however have two outliers, first at the end of 2017 and a larger number of non-cash prepayments in October 2019, which coincides with when Knab announced that it would be last month a client could make a non-cash prepayment penalty-free. It is penalty free because the penalty fee is accounted for in the new rate, meaning a client would not be getting as low a mortgage rate as if they chose to pay the penalty. After this month, we see a change in prepayment behaviour, in which fully prepayments increase in frequency and non-cash prepayments decrease.

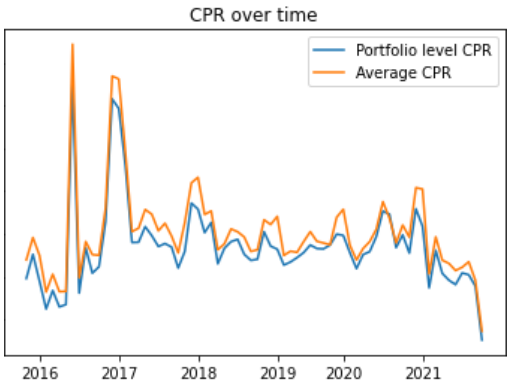


Figure 2

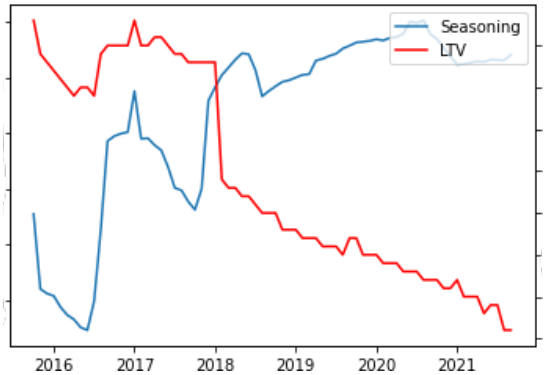


Figure 3

This is most likely from clients moving mortgage provider (and therefore fully prepaying) to one with the option for penalty free non-cash prepayment. The drivers for prepayments will remain the same; however, there will be a fundamental change in the levels of probability of states. To account for this change in state, an indicator variable which is one when after October 2019 and zero otherwise, this will shift the probability's according to the new level shown by the data.
Figure 1 shows the number of loans increases over time signifying that the bank is growing. Furthermore, we can see increased shifts in the number of loans, seen at the start of 2017 where the number of loans increased, this is due to Knab buying mortgage loans from its parent company Aegon. Genrally Knab were buying mortages of similar similar characteristics in terms of seasoning and LTV, shown in figure 3. This was not the case in June 2016 where a sharp decrease in number of loans which was then followed by number of loans dramatically increasing again the month after. Figure 3 show's the bank made a clear business in June 2016 decision selling more younger mortgages with lower LTV's and buying the opposite, which is surprising as these two variables generally negatively correlated. Another shift in strategy from the bank can be seen at start of 2018 where a decease in average LTV coincides with a in average seasoning in portfolio while number of loans remains the same. Meaning the bank is selling short dated mortgages and replacing them with longer dated mortgages.
To ensure these sold mortgages are not seen as prepayments, the loans ids are checked in Aegon's data system, and if they do not exist in Aegon's data, the loan part has been fully prepaid.
The anomalies which occured in 2016 and 2017 seen clearly in figure 2 to our knowledge, have no reasonable explanation. Therefore, it can not be assumed that these acts will not happen again; therefore they can not be removed from the data. Generally, partial is most frequent, followed by full prepayments and non-cash is the least frequent. However, as seen in figure 2, in terms of CPR, full prepayment contributes the most and more or less dominates the CPR due to the size of full prepayment cash flows.

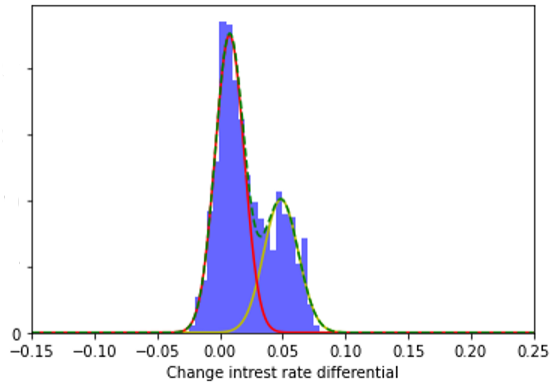


Figure 4: Histogram of change in interest rates between months for each loan part from October 2015 to October 2021

In the Monte Carlo simulation, one must forecast interest rate differential meaning that the market mortgage rate must be forecast. However, in data construction, the interest rate differential variable has been provided by Aegon, and they have a separate model for market rate of which the details are confidential. However, we don't have the past mortgage data Aegon used to construct the interest rate differential. For this reason, interest rate differential will be forecasted by forecasting the change in interest rate differential and adding it to the current interest rate differential; figure 4 shows it's disruption. Disruption is very clearly not normal. However, it appears to be a mixture of normal's in which 2 different world states exist. In one state, rates are relatively stable and other rates are

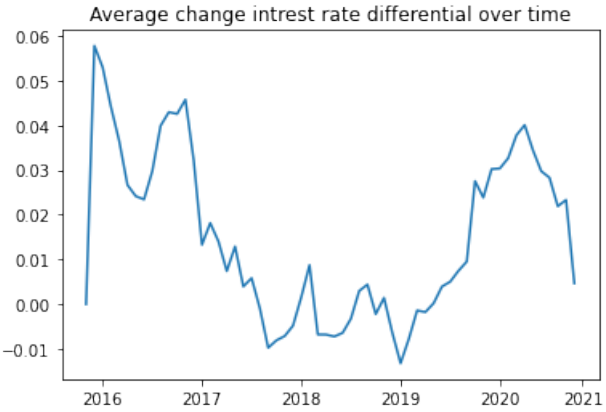


Figure 5: Average change in interest rates for all loan parts in portfolio throughout time

increasing more rapidly. Figure 5 backs up this hypothesis in which the world is roughly in first state from 2017 to middle 2019 and the later state for remaining time.
From figure 7 we can see clear seasonality in the partial prepayment percentage with client tending to partially prepaying more at the end of the year, which is because of suspected. In figure 6 the average compensation percentage for remains extremely low on average about 0.1% across the whole time frame. This is expected as 7 shows that usually, most partial prepayments are below 10%, implying there would be no penalty at all. Even with the partial prepayments that incur a penalty, the likelihood is that wouldn't be much above the threshold and penalty would be would be negligible compared to the prepayment amount.

It can be seen that non-cash compensation seems somewhat random across time. However, there

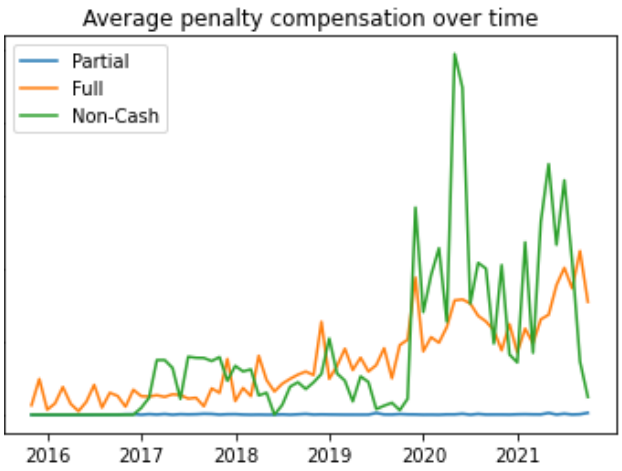


Figure 6

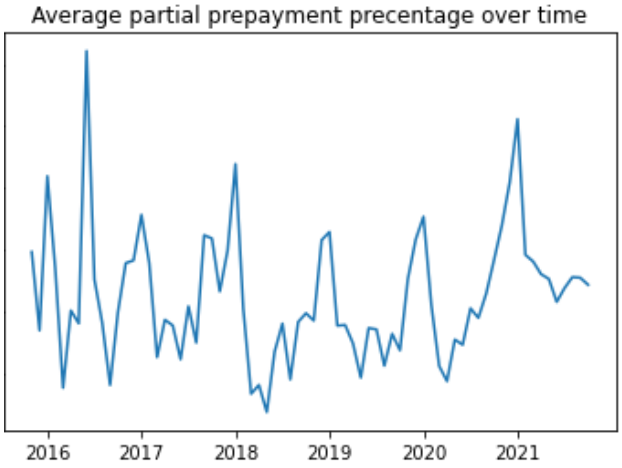


Figure 7

seems to be a change in state at the end of 2019. Compensation increases after October 2019. This is because Knab removed penalty-free non-cash prepayment option explained earlier in the data description. The removal of this option means that a penalty must be paid with a non-cash prepayment, therefore, increasing the penalty compensation for loss in market value. The full prepayment compensation is increasing over time. Furthermore, a sharp increase can be seen in

November 2019, the month after the removal of the penalty free option and can been seen to stay around this new level. This is most likely due to more people fully prepaying with a penalty because they are moving mortgage providers to one that offers this option.
Penalty compensation is very difficult to predict as the reason for not paying the penalty must be predicted, (stated in 2.4), implying one must predict mortality or the event of selling a house. Although there may be some drivers such as age, which becomes a lot more complex, and more personal borrower data would be needed for a model, which would not be allowed from a compliance and ethical point of view. For this reason, it has been decided to take the average of the most recent year's penalty compensation percentage for each prepayment state when forecasting. The reason a year was taken to that only most behavior will be included and not before the change in state occurs.
For the partial prepayment percentage, 3 possible options will be evaluated. First being simple average of most recent years values, similarly to how Saito (2018) has modeled it. Next will be an average with a seasonality factor due to clear seasonality in the data. Lastly, a linear regression model in which the lasso penalty term will be applied to get a model with a limited number of regressors. The reason for a linear model is that partial prepayment percentage will have a limited effect on the cash-flows as full and non-cash should dominate due to the comparative size of cash-flows. Therefore linear regression will be used to approximate a higher order model.
Due to the change in state that occurs when the penalty-free non-cash prepayment option is removed in October 2019 the model must be given enough time to train for this new state. Therefore, the out-of-sample testing period will begin on October 2020 giving a one year testing sample which CPR can be forecasted over.

4.3 Adjustments for Unbalanced data set

In the data set, figure 1 shows that for a large majority of observations no prepayment occurs which makes sense as client will most likely not prepay their mortgage every month or even every couple of months due to time takes for prepayment option to be "in-the-money" once client prepays. Additionally, a large driver of prepayment is house sales and the client is highly unlikely to move house every month. Resulting in large number of no prepayment class observations compared to other classes observations, this is called an unbalanced data set.
Unbalanced data sets are prevalent in real world classification problems as the probability's of rare but important events are often required. For this reason, the area of classification with imbalanced data is heavily researched He and Garcia (2009). Classification Machine Learning techniques do not cope with unbalanced data very well as the algorithms aim to maximize overall classification accuracy. However, by doing this the algorithms have the tenancy to be biased towards classifying the majority correctly as this will have the largest impact on the overall accuracy and the minority class can be considered noise.
Adjustments can be made in two main ways: adjusting the model to better cope with unbalanced data or adjusting the data set to become balanced. To understand how adjustments are made to the model, one must first understand the models, hence the model adjustments will be discussed in 5. The remainder of this chapter will discuss how adjustments can be made to data set.
Modifying the data set to make it balanced can be done in multiple ways, undersampling from the majority classes, oversampling from minority classes or Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a method that creates synthetically manufactured samples within the minority classes by selecting samples that are close to the feature space. Then, a line is drawn between these samples from which a random point on this line is selected as the new synthetic variable. This is done for each of the minority classes resulting in all classes having an equal number of observations.

Oversampling from the minority data set does not add new information to the model. It may cause the model to overfit due to the reliance on these samples within the minority classes. This is why SMOTE is being used over oversampling method. Chawla et al. (2002) has shown a combination of SMOTE and undersampling performs better than undersampling alone; this has been confirmed on prepayment data by Ewing et al. (2021). Due to the size of the data set, it is not feasible to use SMOTE not in conjunction with undersampling. However Saito (2018) has shown that adjustments made on an algorithm level out-preform adjustments of data level for predicting CPR for this reason adjustments have been at algorithmic level.

4.4 Chosen Variables

Section 3.1 describes a large number of prepayment motives have been proved in past literature. The focus of the thesis is to optimize the accuracy of the deterministic model as explained in 1. Consequently, no macroeconomic variables have been included have been used because these variables would need their own model to be foretasted into the future, most of which are stochastic models. Additionally, no borrower specific variables have been taken because either not able to obtain the data or data given is most likely out of date (e.g. Income given when mortgage is taken out has most likely changed).
Now, a large number of Loan specific variables that been generated (which includes large number of variants of variables for example if prepaid before on loan or borrower level). To reduce number of variables Lasso paths have been generated by varying the value of alpha then plotting the resulting coefficients agents corresponding alpha. The variables which shrink to zero for both regression and classification approaches and that experts within Knab agree are not massively important have been removed. The variables mainly removed were all the borrower level variables in addition to the Normalized price incentive variables. The remaining variables are listed in the Appendix C.

5 Methodology

This section will describe the methods used to obtain forecasts for CPR until the interest reset date. Predicting CPR has traditionally been done using an expectation approach by predicting the probability of prepayment types which are each multiplied by the corresponding prepayment amount for each type and then summed together Saito (2018), Schwartz and Torus (1989). For calculating a future expected value when multiple different states of the world can occur in the future, the method is most commonly used to take the expectation approach. Which, in theory, seems like a more intuitive method and should better proxy how real life works.
However, from an interest rate point of view, Knab is only concerned about the part of the prepayment cash flows that the penalty fee has not compensated for. Consequently, it should account for the penalty compensation amount for each type of prepayment, in addition to a partial prepayment percentage to obtain a prepayment amount for partial prepayment. All the additional terms that need to be estimated bring their own uncertainty, exacerbated when combined to give the final CPR. Therefore, regression will also be considered when CPR is obtained straight from the model.
To project CPRs for each loan until interest reset date. In this iterative process, forecasted CPR will affect the following step input variables, consequently affecting the next step's CPR. Therefore one step ahead, CPR forecasts will be used with input variables moved forward by one step after each iteration. For example, one will be added to the seasoning variable after each iteration; this new seasoning value will be used as input to calculate CPR for the next period.

5.1 Structure to obtain one step ahead CPR

This subsection will describe how one step ahead CPR is obtained for both regression and expectation approaches. Section 5.3 will extend then extend this until the interest reset date for each loan.

5.1.1 Expectation Approach

Prepayments can take different forms, which each have a different corresponding cash flow, as explained in 2.1. Therefore each prepayment state has been treated as separate dependent variables yielding the output of the classification model (described below) for each borrower i at time t has been classified as:

$$Y_{i,t} = \begin{cases} 0 \text{ if No prepayment} \\ 1 \text{ if Non-cash prepayment} \\ 2 \text{ if Partial prepayment} \\ 3 \text{ if Cash Full Prepayment} \end{cases}$$

Using the dependant variables the one step ahead prediction is calculated in the following steps.

1. Run one of the classification model described 5.2 the output of this model will give the state probabilities.
2. $CPR_{t+1} = P(\hat{Y}_{3,t+1}) *$ (Full prep pen comp factor at t+1) $+ P(\hat{Y}_{2,t+1}) * PPA_{t+1} *$ (Partial prep pen comp factor at t+1) $+ P(\hat{Y}_{1,t+1}) *$ (Non-cash prep pen comp factor at t+1)

Where P(x) is the probability of classification x, which is the output of step 1, PPA Partial Prepayment amount is the percentage of notional partially prepaid. Saito (2018) is the only time I have seen Partial Prepayment Amount (PPA) modelled in current literate. The impact on overall portfolio-level CPR will be low if PPA is predicted accurately due to its size and low variation in PPA amounts. Due to this, an average of the last 12 months' PPA for assumed as the PPA value. Prep pen comp factor is one minus fraction for which the penalty fee compensates for the loss in the market value of the loan, which when multiplied by the expected prepayment amount.

5.1.2 Regression Approach

The regression approach is much more straightforward than the exception one because penalty compensation and partial prepayment percentage have already been considered in the construction of CPR. This CPR will be used as dependent output in the regression, as stated below:

$$Y_{i,t} = CPR_{i,t} \tag{1}$$

The one major disadvantage with the regression approach is that depending on the chosen model; there is the possibility that predicted CPR is outside of the limit of zero and one (because you cant negatively prepay or prepay more than all your remaining notional).

5.2 Exogenous and endogenous models

5.2.1 Multinomial Logit Model

The first model that will be used is the Multinomial Logit Model (MNL) for reasons stated 3.2.2, but the main reason is that it is the most commonly used model for modelling prepayment in the industry; in fact Aegon, (Knab's pare net company) currently uses it. Therefore, the MNL model will be used as the baseline model to evaluate performance of other models.
The probability of that mortgagor or loan-part (depending on data treatment) i makes a payment of type j at time t denoted as $\pi_{i,t,j}$ and given by:

$$\pi_{i,t,j} = P[Y_{i,t} = j \mid X_{i,t}] = \frac{e^{X'_{i,t}\beta_j}}{\sum_{s\in K} e^{X'_{i,t}\beta_s}} \tag{2}$$

where β_j is the vector of coefficients specific for state j. Since $\sum_{j=1}^{K} \pi_{i,t,j} = 1$ then to ensure parameter identification $\beta_0 = 0$.
When estimating MNL model maximum likelihood is used. Additional to the notation provided before, denote d_i and m_i as the origination date of mortgage or loan-part i and the maturity date respectively, K as the set of payment types and T the total number of time periods. Then the set of observed prepayments for a borrower or loan-part i until time $t-1$ is: $\mathcal{Y}_{i,t-1} = [y_{i,d_j}, y_{i,d_i+1}, ..., y_{i,t-1}]$ and $\mathcal{N}_t = [j, d_i \leq t \leq m_i, 3 \notin \mathcal{Y}_{i,t-1}]$ is the set of the mortgages or loan-part in the portfolio at time t. Since the probability of $Y_{i,t}$ is not independent of its past observations. This gives a likelihood function:

$$L\left(\mathcal{Y}_T\right) = \prod_{t=1}^{T} \prod_{j\in J} \prod_{i\in\mathcal{N}_t} P[Y_{i,t} = j|\mathcal{Y}_{i,t-1}, X_{i,t}]^{I[y_{i,t}=j]}. \tag{3}$$

Implying a log-likelihood:

$$\log L\left(\mathcal{Y}_T\right) = \sum_{t=1}^{T} \sum_{j\in J} \sum_{i\in\mathcal{N}_t} I[y_{i,t} = j] \log\left[\frac{e^{X'_{i,t}\beta_j}}{\sum_{j=0}^{K} e^{X'_{i,t}\beta_j}}\right]. \tag{4}$$

To obtain the optimal estimates $\hat{\boldsymbol{\beta}}$ the likelihood function $L\left(\mathcal{Y}_T\right)$ is maximized with respect to coefficients $\boldsymbol{\beta}$ which is equivalent to maximizing log-likelihood $\log L\left(\mathcal{Y}_T\right)$, as described below:

$$\hat{\boldsymbol{\beta}} = \text{argmax}_\beta L\left(\mathcal{Y}_T\right) \tag{5}$$

Marginal Effects

The interpretation of the estimated coefficients $\hat{\boldsymbol{\beta}}$ derived from equation 4 is not so straight forward and not exactly. This is because the coefficients give the effect in probability relative to one of the prepayment types. However, a change in an independent variable will affect the predicted probability of all states through equation 2. For this reason, we introduce marginal effects, which show each independent variable's effect on the prediction, given that the remaining independent variables remain constant. To find the marginal effect of a variable, one must find the marginal effect for each mortgagor\loan-part and find the mean, for this will give an idea of the effect of the variable on prediction. Calculating the individual marginal effect of the variable involves finding the rate of change in the prediction probability of a state concerning the variable in question by taking the gradient from which one can imply the direction. The marginal effects of a mortgagor or loan-part i with respect to variable x_q for a prepayment type j is calculated as:

$$\frac{\partial \pi_{i,t,j}}{\partial x_q} = \frac{\partial \left(\frac{e^{X'_{i,t}\beta_j}}{\sum_{s \in K} e^{X'_{i,t}\beta_s}} \right)}{\partial x_q} = \pi_{i,t,j} \left(\beta_{j,q} - \sum_{s \in K} \pi_{i,t,s} \beta_{j,s} \right) \tag{6}$$

5.2.2 Linear Regression

Linear Regression (LR) is a simplistic model approach that assumes a linear relationship between CPR and explanatory variables. Resulting in the following model:

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t,1} + ... + \beta_n x_{i,t,n} + \varepsilon_{i,t} = X'_{i,t}\boldsymbol{\beta} + \varepsilon_{i,t} \tag{7}$$

$$\boldsymbol{Y}_t = \boldsymbol{X}'_t \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8}$$

where $\varepsilon \sim N(0, \epsilon)$. To obtain and estimate for the coefficients $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}), \tag{9}$$

where the objective function $L(\boldsymbol{\beta})$ is given by:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left| y_{i,t} - \sum_{j=1}^{p} X_{i,t,j}\beta_j \right|^2 = \left\| \mathbf{Y}_t - \mathrm{X}_t \boldsymbol{\beta} \right\|^2. \tag{10}$$

5.2.3 LASSO Regularization

Least Absolute Shrinkage and Selection Operator (LASSO) is a regularisation technique that was first introduced by Tibshirani (1996). The LASSO penalty term reduces the number of variables the model regresses on, which reduces the chance of over-fitting. LASSO has been picked over other regularisation techniques such as Elastic Net or Ridge because the least essential coefficients in the regression can be set to zero instead of tending to zero. This is important for Kanab as the aim is to get a simple, interpretable (and accurate) model as possible, therefore as few regressors as possible for an accurate model.

A lasso penalty term can be applied to both LR and MNL likelihood functions, in the new likelihood function to be maximised below:

$$L_{LASSO}\left(\mathcal{Y}_T\right) = L\left(\mathcal{Y}_T\right) - \lambda \sum_{j \in K} \|\beta_j\|_1 \tag{11}$$

Where λ is a hyperparameter that decides the rate at which the coefficients "shrink" to zero. The larger λ the more regularization and therefore the fewer regressors that will be included in the model. K-fold Cross-Validation will be used to find optimal hyperparameters, this process is explained in more detail in 5.2.8.

Basis-Variance Trade off

Let $Y_{i,T} = f(X_{i,T}) + \varepsilon$ for some T where $f(X_{i,T})$ is the true function, then the aim is to find $\hat{f}(X_{i,T}|\mathcal{Y}_{i,t}, X_{i,t})$ which minimizes the mean squared error. The mean squared error of a model can be decomposed into three terms the bias, variance and, nosie, given in the equation below (and proved in the appendix):

$$\mathrm{E}_{D,\varepsilon}\left[\left(Y_{i,T} - \hat{f}(X_{i,T}|D)\right)^2\right] = \left(\mathrm{Bias}_D\left[\hat{f}(X_{i,T}|D)\right]\right)^2 + \mathrm{Var}_D\left[\hat{f}(X_{i,T}|D)\right] + \sigma^2 \tag{12}$$

where $D = \{\mathcal{Y}_{i,t}, X_{i,t}\}$, for notation reasons. The bias is the distance between the average prediction of the model and the actual value; the variance is volatility in prediction overtraining sets. The noise σ^2 is an irreducible error which is not dependent on model or training data but can be thought of as the randomness that exists in the world. In Machine Learning, there is a bias-variance tradeoff; if one wants to reduce the model's error, noise can be reduced, but basis and variance can. However, one can only be reduced at the cost of increasing the other.
By applying a Lasso penalty term to the models, the variance is decreased; however, the basis of the model increases.

5.2.4 Decision Tree

DTs have not to be used directly as a classification or regression model. However, they have been used within other models; therefore, it is critical to understand them.
Decision Trees (DTs) are most often used for regression or classification prediction; however, they can extract clustering and non-linear patterns in large databases that are important for discrimination Myles et al. (2004). This way, the data can be split into distinct groups with different prepayment behaviour implying it would make sense to fit separate models to these groups. DTs use a divide and conquer approach. Each node of the decision tree assesses which independent variable and at what value in that variable will result in an optimal split. Data is then split, creating two more nodes; this process continues recursively until specific stopping criteria are met. The nodes that the ends with are called leaf nodes, and the one that starts the tree is called the root node. This process results in a flow chart-like structure where each branch is the list of rules for each path leading to each leaf node.
More formally, a partition set $P_{i,t}$ is generated for each independent variable $X_{i,t}$ by taking the midpoint of each consecutive data point in the independent variable. Then at each node, the optimal partition is found by finding which value in the partition set on which specific independent variable minimises the "error rate". The error rate is calculated differently for classification vs regression problems.
In regression problems, the "error rate" is usually taken as the sum of residuals in each region and is defined below:

$$\text{Error rate} = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y}_{R_w})^2 \tag{13}$$

Where m is the leaf and $\hat{y}_{R_m}$ is the mean value of instances in the m'th leaf. In classification problems, the error rate is impurity and can be measured using metrics. Raileanu and Stoffel (2004)

shown that it rarely matters which measure is used; however, Gini is preferred to Entropy because it is faster. The reasoning is that Entropy uses logarithms, but Gini does not. The Gini impurity is defined below as:

$$\text{Gini} = \sum_{i=0}^{3} \hat{p}_i(1 - \hat{p}_i) \tag{14}$$

Where $\hat{p}_i$ is the probability of an observation training data with label i to be correctly classified. The Gini impurity measures incorrectly classifying an observation if it was randomly assigned to a class according to the class distribution in the set. Therefore minimising Gini gives the optimal split. After the splitting occurs, the number of nodes then doubles; this process repeats until the stopping criteria.

Stopping criteria can be multiple different hyperparameters which include; the minimum number of observations in every leaf, maximum depth of the tree or the threshold decrease in the impurity after a split. A stopping criterion is implemented so that terminal nodes do not overfit. Otherwise, after each node is split, the Gini or MSE. will continue to get smaller as partition spaces get smaller and will tend to zero. Only max depth from one to four will be considered to maintain interpretability.

DTs are non-parametric models, meaning there is no prior distribution assumption, letting the data speak for itself. A further advantage of DTs is their ease of interpretability; one can easily find precisely why the model has given its prediction. However, the main disadvantage is its high variance meaning a slight change in the training data could lead to a significant change in the structure of the DT. DTs have a low basis and tend to overfit the data, making them often inaccurate.

Algorithm 1 Decsion Tree

procedure DT
    Create node, n = createnode()
    while Stopping criteria has not been breached **do**
        Calculate E, error rate or Gini (depending on problem type) using equation 13 or 14 respectively.
        if E $< \mathcal{E}$ **then**:
        The node is leaf node
         return
        Find binary split which results in biggest reduction of E
        Set the child node as the parent node
    Find alpha α which minimizes CCP using equation 15

Decision Tree Purning

Breiman et al. (1984) first introduced Tree Pruning; it is often used on DTs to prevent overfitting. Mingers (1989) has shown that pruning increases the accuracy of the DT. Accuracy of predictions is not the aim of the DT in this instance. However, the Puring will also exaggerate the Patterns seen in the data, in turn making each leaf node more homogeneous.

Minimal Cost-Complexity Pruning (CCP) is a method of pruning; it accounts for both the number of errors in classification and the complexity of the trees. The idea is that a penalty term (which accounts for several terminal nodes) would be added to the error rate. CCP is essentially the upper threshold of this error rate with the penalty. Meaning new nodes are only grown if they bring enough

of a decrease in error rate, such that the CCP is not breached.

$$\text{Cost complexity criterion}(T) = \sum_{w=1}^{|T|} \sum_{i \in R_w} (y_i - \hat{y}_{R_w})^2 + \alpha |T| \tag{15}$$

where $|T|$ is the terminal number of nodes in tree T. Cost complexity criterion for classification, is calculated very similarly expect Gini is used instead of Error rate, therefore giving total misclassification rate of the terminal nodes being considered at T.

5.2.5 Linear Tree Models

The main disadvantage of using Linear models is their inability to exploit data clustering and account for non-linear relationships between dependent and independent variables. Lijmbach et al. (2021) have used the recursive partitioning algorithm designed by Strobl et al. (2008). The algorithm first finds the most unstable variable using a statistical test. Once found the value of split is found by minimising the objective function 5, the procedure repeats until all variables are stable. Implementation of the model must be in Python due to constraints set out by Knab, whereas the package for this algorithm is only available in R. This makes it not feasible to implement this algorithm with given time constraints.

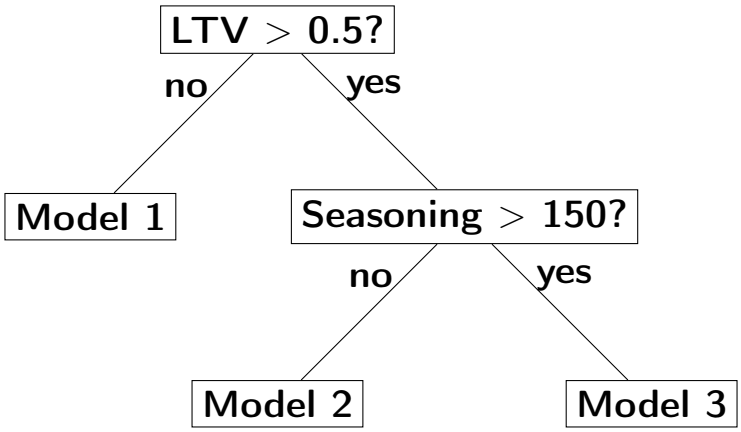


Figure 8: Example of Tree which segments data in step 1 of the Linear Trees

An alternative method implemented for classification is the hybrid CART-logit, first posed by Chan and Loh (2004) and includes a 2-step process. The first step splits the data using a classification DT resulting in piecewise linear segments. Then MNL models can fit the data in each leaf node. Chan and Loh (2004) extended the model to introduce pruning to the decision tree, and more recently Dumitrescu et al. (2020) added a regularisation term to obtain Penalised Logistic Tree Regression (PLTR). Similarly, for the regression case, Karalic (1992) and Chaudhuri et al. (1994) use the same 2-step process; splitting the data using regression DT, then they fit Linear and Polynomial regression models respectively to the terminal nodes.

Steinberg and Cardell (1998) argues that DTs excel in the detection of local data structure. Therefore, the discovery of patterns becomes more localised as the tree grows in-depth, limiting the ability to do further statistical analysis at the nodes. This is not an issue for this research as there is an emphasis on interpretability; a max depth of 3 has been enforced.

Both regression and classification methods have been us ed, for regression approach, a Linear Regression Tree (RL-Tree) and classification a Multinational Logit Tree (MNLL-Tree) are both variations

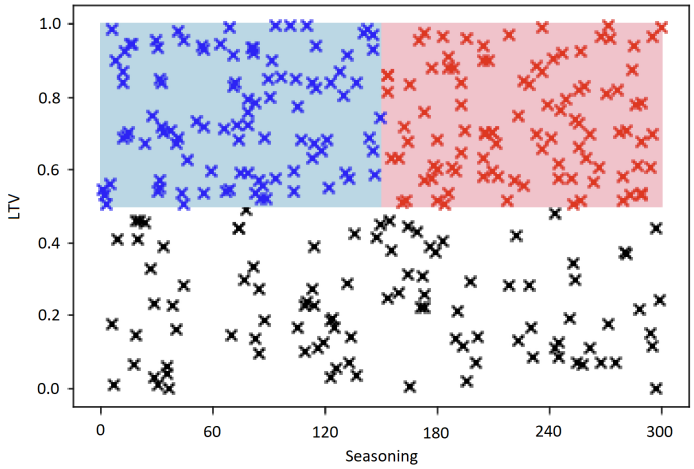


Figure 9: Graphical representation of example model in Figure 8. Please note data has been generated randomly, and no insights can be taken from it.

of a Linear Tree. Current literature is limited to binary classification; the MNL-Tree will be extending the literature to multinomial output.
Like DTs, a drawback to the Linear Trees is their instability meaning a small change in the training data could lead to a large change in the structure of the DT in step 1. In turn, leading to different splits in the data and, hence, impacts the models in each of the nodes.
Figure 9 shows a graphical representation of how the DT splits the data in step one of the example Linear Tree given in figure 8. The data in different regions in step two will then fit into a unique model.

5.2.6 Random Forest

Random Forests (RF) is a supervised Machine Learning technique used for regression and classification problems. Meaning the model is given both dependent and independent variables to train the model on. An RF is a collection of DTs, which are described in 5.2.4, the output from each DT is aggregated together (averaged for regression or majority voting for classification) to give the output of the RF, as visualised figure 10.
RF has a lower variance but a higher bias than a single DT due to the aggregation of trees. However, this exact reason why DTs have a lower bias is why they are likely overfitting the data. Therefore, if the increase in bias is not "overshining" the variance reduction, this will increase prediction accuracy, which is usually the case for an RF. The "overshining" refers to the variance-bias trade-off.
In the case of a RF Louppe (2014) has shown that the variance of the model is inversely proportional to the number of trees in the forest. Implying that increasing the number of trees reduces the overall variance. However, Breiman (2001) has shown there is a performance plateau where computational costs out weight the reduction invariance. The optimal number of trees will be found using K-fold cross-validation as described in 5.2.8.
Each of the trees within the RF is created by choosing by a random subset $m \approx \sqrt{p}$ of all explanatory variables p, which are used to create the split at each node. This process is called bootstrapping, it makes each DT more unique and reduces the correlation between DTs. Furthermore, the random construction of the trees gives the variables with less predictive power a greater ability to be involved

in the fit as sometimes dominant predictors will not be included.
RFs are an Ensemble method as they combine several model predictions, making a stronger and more stable prediction. Hastie et al. (2001) has shown that random forest performs better for classification than for regression and this is because misclassification error is less sensitive to variance than is a mean-squared error. In the research paper both a regression and a classification approach has been considered.

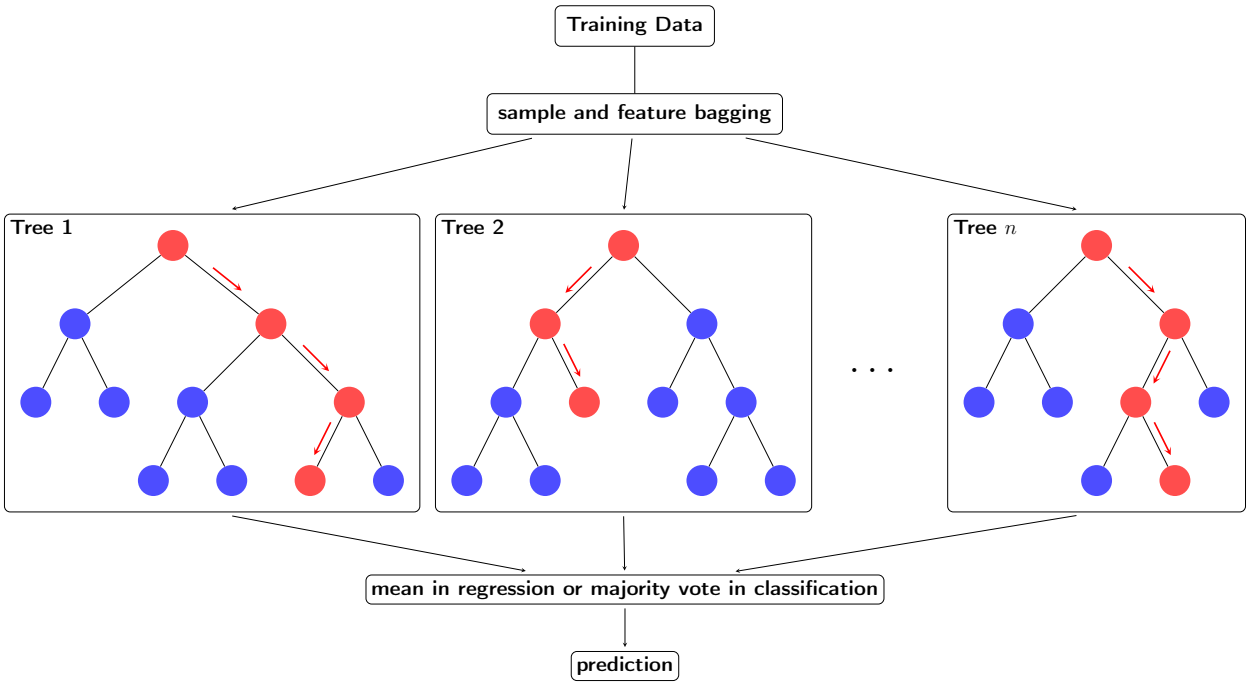


Figure 10: Visualisation of how a Random Forest works

RFs are often referred to as "black-boxes" as it can be difficult to explain why the RF gives certain output due to the ensemble. However, it is possible to rank each input variable based on importance. One method to give some insight into the RF is to obtain the how important each feature is with the RF by measuring which explanatory variable has biggest impact of the decrease in the error. Therefore feature importance is measured using:

$$F(X_{i,t}) = \sum z \in f \delta G(P_{i,t}, z) \tag{16}$$

where z is the node, f is f'th tree in forest, $P_{i,t}$ is the partition set for $X_{i,t}$ and G() is Gini from 14 for classification or Error rate 13. The larger the decrease in impurity and hence $F(X_{i,t})$ the more important the variable in the RF.

An advantage of using a RF vs LR as the regression model is that the model's predictions will inherently be bounded between zero and one. In a regression context, RF predictions are the average predictions of the prediction of the trees in the forest. Each DT prediction for observation X_t can be thought of a weighted average of the response values $Y_{0,t}, .., Y_{N,t}$ in the training data, given below:

$$\hat{f}(X_{i,t}) = \sum_{n=0}^{N} w_n(X_{i,t}) Y_{n,t} \tag{17}$$

where $w_i(X_{i,t})$ are no negative weights with the constraint that they sum to 1, $\sum_{n=0}^{N} w_i(X_{i,t}) = 1$. Which therefore implies that the output is bounded:

$$\underset{0\leq n\leq N}{\text{Min}}\, Y_{n,t} \leq \hat{f}(X_{i,t}) \leq \underset{0\leq n\leq N}{\text{Max}}\, Y_{n,t} \tag{18}$$

which ensures that the predicted CPR is bounded between zero and one.

Model Based Adjustments For Unbalanced Data

It was briefly discussed in 4.3 that the model could be adjusted to handle the unbalanced data better. Classification methods such as DTs and RFs use majority voting to decide output meaning. However, when the data set is unbalanced, the majority class tends to dominate in the voting system, creating a considerable bias toward the majority class and a large false-negative rate. Chen et al. (2006) have used threshold adjustments to combat this problem, which is when you adjust the classification boundary towards minority class. For example, in a binary DT classification problem, one may make it such that a node only needs 25% of observations to be considered that class instead of the normal 50%. However, this method can become more complex with multiclass classification; for example, consider data set with multiple minority classes and their corresponding thresholds are breached.
Another method that is used is Cost-Sensitive Learning Ling and Sheng (2008), which a penalty is placed on misclassifying minority classes, therefore placing a larger emphasis on model correctly classifying them, to mintage the harshness of the classifier. To implement cost-based learning to DTs, weights are incorporated into two parts of the algorithm. First, weights are used to weight the Gini criterion for finding splits in the class. Lastly, weights are applied in the terminal nodes when calculating the probability of each class.
The Gini is used to find the optimal splitting rule in DT classification, which can be thought of as the probability of incorrectly classifying an observation if it was randomly assigned to a class according to the distribution of the dependent variables in the set. Chen and Breiman (2004) describes how using this metric to split data can be biased on unbalanced data; therefore new variables are defined to create metrics correct for bias. Let $t_c = \sum_{i=0}^{3} n_i$ be the number of observations in the potential child node c, where n_i is the number of observations in class i. Then the impurity of the child node c is defined as:

$$i_c = 1 - \sum_{i=0}^{3} \left(\frac{n_i}{t_c}\right)^2 \tag{19}$$

where t_c is the total number of observations in the parent node. Therefore, the total impurity for the split is given by:

$$Gini = \sum_{c\in C} \left(\frac{t_c}{t_p}\right) i_c \tag{20}$$

where C in number of child nodes in the potential split and t_p is total number of observations in parent node. Implying that Gini is the sum of the fraction of number of observations in parent node which which are in the child nodes multiplied by there corresponding impurity of the child node,
Saito (2018) shows that when there exists unbalanceness in the data set the Gini is effecred, in turn effects the optimal split given by the DT. Lets define $t_c^W = \sum_{i=0}^{3} w_i n_i$ as the weighted number of observations in child node child class c, where w_i is the weight associated to class i. Then the impurity of the child node c is defined as:

$$i_c^w = 1 - \sum_{i=0}^{3} \left(\frac{w_i n_i}{t_c}\right)^2 \tag{21}$$

Therefore, the total impurity for the split is given by:

$$Gini = \sum_{c \in C} \left(\frac{t_c^w}{t_p^w} \right) i_c^w \tag{22}$$

where t_p^w is weighted total number of observations in parent node. Saito (2018) has shown that by including weights in the Gini imporoves the performance of the model.

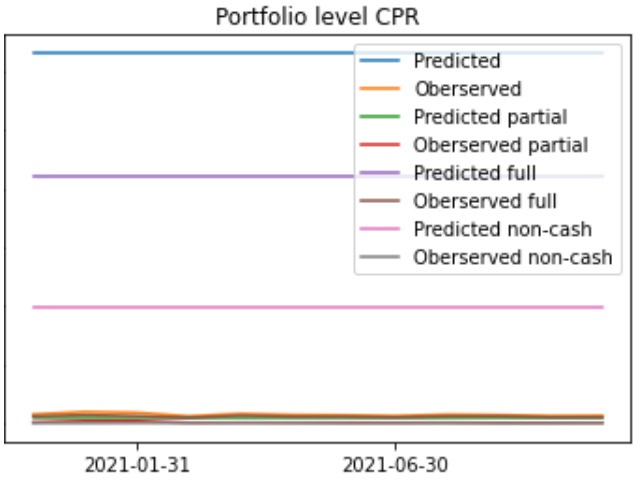


Figure 11: Random forest classification 1 year out of sample forecast without adjusting the probability

Figure 11 shows the massive overestimation in the probability's that occurs without making the weighted adjustments the the output probability's.

5.2.7 Linear Forests

The literature shows RFs to be effective at predicting prepayments Saito (2018). However, a possible problem with RFs is that they may not effectively capture parametric relationships between the response and the predictors, given the non-parametric nature of the model. RFS may fail to accurately make predictions for extrapolated data which are out of the domain of the data set. This thesis aims to produce a long-term forecasting model; therefore, extrapolation is massively essential. Given our limited data size, the likelihood is that the model will have to extrapolate at some points, for example, rising rates.
Furthermore, in 5.2.5 Linear Trees are described; however, as explained in 5.2.4, a considerable disadvantage with DTs is their high variance. This issue persists in Linear Trees because in step one of the Linear tree algorithm, a DT is fit, and if it drastically changes from a new input, this will also drastically change the Linear Tree. As we know, RFs reduced this variance by using an ensemble of trees, by the same logic and inspiration from Zhang et al. (2019) Linear Trees will be extending to Linear Forests using an ensemble to reduce variance and chance of overfitting.

5.2.8 Hyerparameter Tuning

Hyperparameters are parameters that are used to control how the model will learn. For example, in the case of a random forest, the number of trees or maximum depth of the trees. Hyperparameter

tuning is the process of finding optimal hyperparameters for the model. If this is done over the training set, this will lead to overfitting as the model has been finely tuned to the exact data in the training set and may not perform so well out-of-sample. A method to avoid this is to use K-fold cross-validation. In this method, the training and validation set are combined and then split into K sets. The model will be then trained using K-1 sets as the training set, and the remaining set is called the validation set. The trained model is then run with the validation set, and the percentage of correctly classified samples (called accuracy score) is recorded. This process is repeated K times using one of the K sets as the validation set. The K accuracy scores are averaged; this is the metric that will then be optimized over.
This K-fold cross-validation is repeated for various hyperparameters to find optimal. There are two main ways can search through the set of values used as hyperparameters Probst et al. (2019). The first is Grid Search systematically loops through all values in a given parameter space. Whereas Random Search randomly values from a given distribution. Bergstra and Bengio (2012) have shown random search has been shown to be more efficient on learning algorithms, not including RF's. However, I believe the same would be true for RF's as not all hyperparameters are equally important, and Random Search can account for this. For this reason, Random Search will be used.

5.3 Multi-step Forecasting

Forecasting over multiple periods can be done in a recursively or In this section, the different methods for forecasting CPR over the lifespan of each mortgage will be explained. They will be evaluated on both accuracy and run-time.

5.3.1 Deterministic

Different adjustments need to be made to each model to be able to apply multi-step deterministic multi-step.

5.3.1.1 MNL/MNL-Tree

Estimates for $\hat{\boldsymbol{\beta}}$ are obtained in way described above 5.2.1. The equation for $\pi_{i,t,j}$ as described in equation (2). To forecast the MNL and MNL-Tree, a model dependent on time till maturity must be obtained. To do this, the independent variables must be made a function of time till maturity. This can be easily done for deterministic variables such as Seasoning or Seasonality. However, it is more difficult for stochastic variables such as market mortgage rate and usually have to be modelled stochastically. There these variables will be assumed as constant. Therefore, this will yield MNL and MNL-Tree models that will be a function of time until maturity. Forecasts can easily be extrapolated from this.

5.3.1.2 Random Forest

As RFs are "black-boxes" this makes then There are two main methods for forecasting RFs deterministically; the first is done "directly", for which a separate model is developed to forecast each overtime. This is very computationally expensive; for this reason, it will not be used. The other method is to use a Multi-Input Multi-Output (MIMO) RF, which instead of returning a single classification like traditional RFs, it will output a vector of classifications. These multiple classifications will give the forecasts till maturity. However, the issue with MIMO RF is that it will produce a vector of fixed length for all observations; this will not be suitable for modelling mortgage prepay-

ments because each observation will have a different time till maturity; therefore, it needs a different forecast length.

5.3.2 Monte-Carlo

The deterministic forecasting is unable to deal with stochastic variables as explained 5.2.1 when forecasting using Monte-Carlo (MC), these stochastic variables can also be forecast using stochastic models and then incorporated in the MC. Another advantage to forecasting via MC is the ability to capture path dependency. For example, if the model predicts a prepayment int, the probability of prepayment in t+1 will decrease. The deterministic forecasting cannot model this as it never "predicts" a prepayment but instead finds the expected prepayment as a combination of probability and prepayment amounts. The general structure of the Monte-Carlo simulation is:

1. Run steps for the one-step forecast as stated in 5.1. Note the model will be fit only once, and then the new one-step ahead independent variables are input into the model.
2. Move deterministic variables by one period.
3. Run stochastic models to forecast one step ahead stochastic variables.
4. Use these new variables as the input variables and run step one again.
5. Iteratively run steps 1-4 until the time until maturity is zero.
6. Repeat steps 1-5 one hundred and fifty times for each loan and then average the CPR and penalty terms at each forecast time period.

5.4 Procedure for choosing regerssors

Knab aims to get a simple, accurate model with variables that have only shown significant predictive performance in the past and can be interpreted in an economically correct manner. To achieve this two different approaches, the first for the Machine Learning models in which one used 3-fold cross-validation to find the optimal hyperparameters resulting in optimal regressors (in the case of Lasso penalty models). Then for non Machine Learning models (MNL and LR) first model with all variables (excluding ones removed from initial regularization). Then variables are iteratively removed based on which variables have insignificant coefficients (for all states in MNL models). Once completed, the correct combination of variables is found using trail and error, which maximizes Adjusted R-squared (McFadden's for MNL) whilst minimizing Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). Adjusted McFadden's R-squared is used for both is defined as:

$$R^2_{Adj.McFadden} = 1 - \frac{log\left(L_c\right) - K}{log\left(L_{null}\right)} \tag{23}$$

where L_c is maximized likelihood function, L_{null} is the maximized log-likelihood of null function in which only an intercept is included and K is the number of additional parameters relative to the null model Long et al. (1997). It is preferred to optimize Adjusted over standard McFadden's R-squared because it penalizes you for adding regressors that do not fit the model.
In addition to the adjusted R-squared scores, Information Criteria will be used for all non-Machine Learning models, which measure the quality of a statistical model, taking into account how well the

model fits the data and the complexity of the model. Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are defined as:

$$BIC = Klog(n) - log\left(L_c\right) \tag{24}$$

$$AIC = 2log(n) - log\left(L_c\right) \tag{25}$$

Where n is the number of data points Long et al. (1997). BIC can be prone to underfitting where as AIC can be prone to overfitting due to how they penalize the number of parameters. Therefore to minimize the chance of both underfitting and overfitting one must minimize both BIC and AIC, however if AIC and BIC disagree BIC is preferred as penalizes complexity more yielding a simpler model.

5.5 Model Performance Evaluation Methods

Prepayments are notoriously hard to predict due to the external influences causing seemingly non-rational behaviour. Additional banks or other financial institutions which hold large amounts of mortgages are not so interested in forecasting the exact amount each loan part prepays. They are much more interested in a portfolio level and how much prepayment occurs because different loan parts will even each other off. Therefore, the hedge happens on a portfolio level. However, this is not possible to optimize this source whilst training for several reasons. First, because the algorithm would then have to group the predictions with the same dates before a weighted average is used to calculate on a (sub-)portfolio level, however, the date is not an input or accounted for in the model. Furthermore, the same issue occurs for the classification approach because penalty compensation terms and partial prepayment amounts are time-dependent. Therefore obtaining a fair comparison for regression and classification models is not possible in the training data. Therefore once the optimization procedure on training data (described in 5.4) is completed, one can use the model to forecast a portfolio for a number of months which can be evaluated vs the observed prepayments. To combine CPR from loan or borrower level to portfolio level, a weighted averaged approach given below can be used:

$$\mathrm{CPR}_{p,t} = \frac{\sum_{n=0}^{N}\left[\mathrm{CPR}_{n,t}\mathrm{Not}_{n,t}\right]}{\sum_{n=0}^{N}\mathrm{Not}_{n,t}} \tag{26}$$

where $\mathrm{CPR}_{n,t}$ is predicted CPR for nth loan or borrower at time t, N number of loans or borrowers at time t and is $\mathrm{Not}_{n,t}$ is the remaining notional the for nth loan or borrower at time t. It depends on the model chosen for forecasting whether loan or borrower is used.
For evaluating how the predictive performance of the models out of sample on portfolio level three main metrics will be used. The is called normalized delta which aims to measure how accurate the model is at predicting prepayment amounts for whole year. It is calculated as:

$$\Delta_T = \frac{\sum_{t=t_1}^{T}\sum_{n=0}^{N}\left(\text{Pred CPR}_{n,t}\text{Pred Not}_{n,t} - \text{Obs CPR}_{n,t}\text{Obs Not}_{n,t}\right)}{\sum_{t=t_1}^{T}\sum_{n=0}^{N}\left(\text{Obs CPR}_{n,t}\text{Obs Not}_{n,t}\right)} \tag{27}$$

where N is number of loans in portfolio at t_1, $T = t_1 + t_w$ where w is number of periods being forecast. Pred $\mathrm{CPR}_{n,t}$ is the CPR predicted by the model and Obs $\mathrm{CPR}_{n,t}$ is observed CPR in real data for loan n at time t. Pred $\mathrm{Not}_{n,t}$ is the predicted notional for loan n at time t, which is affected by the previous Pred $\mathrm{CPR}_{n,t}$. This issue with the normalized delta metric is that it does not consider how well the prepayments are predicted through out the year, just as a whole at the end.

The other two metrics use and R squared approach which measures how well the model explains the variance in the data thought out the year. R squared is generally calculated using:

$$R^2 = 1 - \frac{\sum_{t=t_1}^{T}(d_t - f_t)^2}{\sum_{t=t_1}^{T}(d_t - \bar{d})^2} \tag{28}$$

where f_t is the forecasted variable at time t for which R squared is being calculated for. d_t is the corresponding observed variable at time t and $\bar{d}$ is the average of the observed variables for data period $T - t_1$.
The first is called Cash flow R squared, which calculated the R squared of the predicted prepayment cash flow's similarly to equation 27. Lastly, a portfolio level CPR R squared is calculated by transforming forecasted CPRs to portfolio level using equation 26 and then calculating the CPR over the forecasting period.

6 Results

6.1 Linear Regression

Linear regression was fit to training data using the iterative process described in 5.4, to find the optimal combination of independent variables. This resulted in the model with coefficients in table 1 and 2. The tables show that similar variables are highly significant in the LR model for loan and

Table 1: Loan Level Coefficients

Variables	Coefficients	P-value
Seasoning	3.222 x10^{-5}	0.000
Seasoning^2	-8.334 x10^{-8}	0.000
NHG	-0.0007	0.000
Interest Rate Differential	0.0027	0.000
Prepaid Before Indicator	-0.0018	0.000
Burnout As Sum	-1.274 x10^{-9}	0.000
After Anomaly Month	0.0007	0.000
Annuity Indicator	-0.0047	0.000
Linear Indicator	-0.0041	0.000
Bullet Indicator	-0.0046	0.000
Savings Indicator	-0.0053	0.000
House Indicator	-0.0039	0.000
January Indicator	0.0119	0.000
February Indicator	0.0107	0.000
March Indicator	0.0103	0.000
April Indicator	0.0100	0.000
May Indicator	0.0104	0.000
June Indicator	0.0126	0.000
July Indicator	0.0108	0.000
August Indicator	0.0109	0.000
September Indicator	0.0106	0.000
October Indicator	0.0099	0.000
November Indicator	0.0111	0.000
December Indicator	0.0126	0.000

Table 2: Borrower Level Coefficients

Variables	Coefficients	P-value
Seasoning	9.341 x10^{-5}	0.000
Seasoning^2	-4.843 x10^{-7}	0.000
Seasoning^3	6.287 x10^{-10}	0.000
Interest Rate Differential	0.0021	0.000
Burnout As Sum	-6.485 x10^{-10}	0.000
Prepaid Before Indicator	-0.0032	0.000
Annuity Indicator	0.0005	0.001
Linear Indicator	0.0018	0.000
Bullet Indicator	0.0006	0.000
House Indicator	-0.0032	0.000
January Indicator	0.0057	0.000
February Indicator	0.0039	0.000
March Indicator	0.0032	0.000
April Indicator	0.0031	0.000
May Indicator	0.0035	0.000
June Indicator	0.0046	0.000
July Indicator	0.0041	0.000
August Indicator	0.0041	0.000
September Indicator	0.0034	0.000
October Indicator	0.0030	0.000
November Indicator	0.0039	0.000
December Indicator	0.0057	0.000

borrower level data. The variables that remain significant in both models agree on the direction of the corresponding coefficients. The month indicator variables show the models capture the seasonality that exists in data shown in 2. The Strong negative coefficient of the Seasoning squared variable captures some of the non-linearity in the data. Furthermore, it captures the idea that the likelihood of a borrower moving house after a certain amount of time starts to decrease because owners are settled in their house and have no reasons to move out (i.e. having kids). Two variables which are highly significant on loan level but not borrower are NHG and After Anomaly indicators which are negative and positive, respectively. The Dutch government insures NHG or National Mortgage Guarantee loans. They have lower interest rates attached to them, lowering the probability that it is optimal to prepay the loan and therefore lowering CPR. The After Anomaly indicator is one after the option of a non-cash prepayment without a penalty was removed; otherwise its 0. As explained in 4.2 There appears to be a behaviour change when an option is removed. A slight increase in the number of full prepayments and a decrease in non-cash prepayments results in an increase in CPR overall, which is confirmed by the direction of the coefficient.
The tables imply for both borrower and loan-level data that there is a highly significant negative cor-

relation between if (loan part or borrower) has prepaid before and CPR. This goes against intuition because, in theory, if someone prepays before, they understand how prepaying can be optimal and will be more likely to prepay again. However, given the small data period for borrowers, there is not enough time to see borrowers prepay again because the borrower needs to build up enough capital to pay the costs of prepaying again. This is backed up by the fact that the number of prepayments greater than prepaying more than once is minimal.
For categorical variables, which have been transformed into binary variables like loan types, to interpret them, you compare them against each other as a pose to looking at the direction of the coefficient. For Loan and borrower level data, linear loans have the highest coefficient implying that they had the highest CPR in the past. In linear loans, the total amount paid each month decreases; therefore, as the loan ages, the mortgagor may have more disposable income giving them the ability to prepay. The savings indicator is the lowest for loan data and is not significantly different from zero on the borrower level. Consequently, both models also agree that loans tied to saving accounts are less likely to prepay. Generally, the interest rate on your savings corresponds to the mortgage interest. Since 2012 interest rates have been falling, implying a decreasing mortgage rate on the loan, making it sub-optimal to prepay. However, with rising interest rates on the near horizon, this may invert, as these loans are more sensitive to interest rates.
Both models agree on a w shape for seasonality, with the coefficient being largest at the end and the start of the year, then low in autumn and spring before going high again in June. This agrees with what we have observed in the data.
Seasoning has a Strong positive correlation with CPR in both models. This agrees with the PSA model Hayre (2001) which increases CPR for the first 30 months. Additionally, this makes economic sense because the longer a borrower lives in a house, the more likely they are to move house or generate enough disposable income to partial prepay, increasing expected CPR. Both models agree that Interest Rate Differential has a strong positive relationship with CPR, implying that borrowers prepay more when the option to prepay is "in-the-money" (i.e. beneficial).
Burnout is the sum of all the past Refinance Incentive for a loan (or borrower); therefore, if burnout is large, it implies that in the past, the borrower has not prepaid when it may have been optimal to do so; otherwise, the interest rate would of been reset and refinance incentive would be zero, giving a lower burnout. The variable is by no means perfect because it could be heavily linked to age. However, this version works the best in the linear regression. The tables 1 and 2 show there exists a Strong negative relationship between Burnout and CPR, implying that someone who had not exercised prepayment when it was optimal to do so in the past, meaning they will be less likely to prepay in future.
The tables both agree that if the mortgage is for a family house instead of an apartment or business property, they will be less likely to prepay as they are more likely to invest in property in the long term. The last thing to note is that in both models, a constant term is not significantly different to zero. This can make sense because, in the case of seasonality, all twelve indicator variables are included, meaning can adjust the baseline level.

Linear Regression with Lasso penalty term

To find optimal level of penalization λ, 3-Fold cross-validation has been performed on the training data optimizing the out-of-sample R-squared. The process has been described in further detail in 5.2.8. This resulted in a λ of 0.0005.

6.2 Multinomial Logit Model

First, to reduce the number of variables in the MNL model, the lasso penalty term was fit with a total of 30 different values for lambda ranging from 0.0001 to 10000. As lambda is increased, the less critical (and less critical of highly correlated) variables for prediction coefficients shrink to zero faster than the more important variables. It was chosen by an expert judgment that at a lambda of 10, remove all variables that had zero for coefficients for all of the states. This results in a lot of the expected highly correlated variables dropping out. The details of which variables had been removed can be found in the appendix.

Then the iterative process of finding optimal regressors described in 5.4 is performed, yielding the MNL model with coefficients described in table 3. The resulting MNL model has a Mcfadden R^2 of 11% , a AIC of 1,129,430 and BIC of 1,483,430

Section 5.2.1 explains why the coefficients in the MNL model can be difficult to interpret; therefore the Average Marginal Effects (AME) are used instead to gain an understanding of the model. For the interpretation, Partial, Full and Non-cash states will mainly be considered because the classification approach uses these states to create CPR.

There exist positive AME family house indicator for Partial but negative for both Non-cash and full; if the property is a family house, the owners are more likely to see themselves there more long term, therefore reducing the probability of full prepaying (to move) and more likely to partially prepay to reduce the interest rates over the long term. Refinance incentive and interest rate differential aim to capture the similar idea of prepayment probability increases when it is optimal to prepay. The direction of both partial and non-cash AMEs agree with this hypothesis. However, for full prepayment, the AME for refinance incentive and interest rate differential differ in the direction. The reason why there might be a weaker relationship in this state is that full prepayments are predominately driven by moving house, and the incentive to prepay is not going to be the primary motivation in moving house.

LTV has a positive AME for partial, implying borrowers will partially prepay more when LTV is high to move to a lower mortgage rate bucket. There exists a decreasing relationship between LTV and probability of non-cash prepayment, which will most likely be due to its correlation with Seasoning, i.e. the older the mortgage the more the borrower will have paid back and therefore a lower LTV.

The After Anomaly Indicator shows the shift in the state, which happens as penalty-free non-cash option is removed. A decrease in the probability of Non-cash and an increase of Full is observed, which makes sense because the removal of this option means fewer people may have the money to be able to pay the penalty in non-cash prepayment. The removal of the option may be driving borrowers to move to a bank that offers it and, therefore, full prepay.

Seasoning and Month Till Maturity are inversely correlated; therefore, it is somewhat surprising for the AMEs to act in the same direction for partial and full states. However, this could be possible because the time till maturity could capture that for loans with a larger time till maturity, it will be more beneficial (if in-the-money) to make non-cash prepayment because the borrower will be saving money on interest payments for longer. The AME of Seasoning shows that the older the mortgage, the more likely they are to move property and, therefore fully prepay.

The loan type indicators confirm the argument made in 6.1 that Linear are most likely to partially prepay and Saving loans the least. The seasonality dummy variables show some clear seasonality in partial prepayment probability, where we see large values for AME at the end/start of the year for tax reasons, then decrease in spring and autumn before increasing again in summer when employees receive holiday pay. There is no economic reason why a borrower would make a non-cash prepayment in one month vs another; this is demonstrated in the data where there doesn't seem to be a pattern in Ames. For full prepayments, seasonality peaks in the summer months between June and September,

Table 3: Multinomial Logit Model summary table

	Coefficients				Average Marginal Effects			
Variables	Partial	Non-cash	Full	None	Partial	Non-cash	Full	None
Seasoning	-0.0001	-0.0046	0.0203	0	$-3.354x10^{-5}$	-7.084x10^{-5}	0.0001	0.0008
	(0.912)	(0.000)	(0.000)					
Seasoning^2	$-3.116x10^{-5}$	$1.223x10^{-5}$	$-7.07X10^{-5}$	0	-1.171x106-7	2.047x10^{-7}	-4.823x10^{-7}	-1.054 x10^{-6}
	(0.000)	(0.000)	(0.0229)					
House Property Type Indicator	0.4406	0.0788	-0.5303	0	0.0004	-0.0005	-0.0026	-0.0001
	(0.000)	(0.000)	(0.000)					
Refinance Incentive	4.231x10^{-6}	7.45X10^{-5}	-7.075X10^{-6}	0	8.16x10^{-9}	1.126x10^{-7}	-7.952x10^{-8}	-4.128x10^{-8}
	(0.000)	(0.000)	(0.0418)					
Month Till Maturity	-0.0009	0.0014	-9.077X10^{-6}	0	-1.222 x 10^{-6}	3.32x10^{-5}	2.943$x10-6$	0.0004
	(0.000)	(0.000)	(0.824)					
Annuity Indicator	-1.2866	-0.9757	0.4203	0	-0.0015	-0.0206	0.0033	0.0187
	(0.000)	(0.000)	(0.0053)					
Linear Indicator	-1.0754	-0.3258	0.3922	0	-0.0001	-0.0114	0.0041	0.0074
	(0.000)	(0.000)	(0.000)					
Bullet Indicator	-1.0703	-1.3823	0.1369	0	-0.0008	-0.0282	0.0022	0.0268
	(0.000)	(0.000)	(0.017)					
Saving Indicator	-4.4698	1.1179	0.1558	0	-0.0036	0.0163	0.0015	-0.0142
	(0.000)	(0.000)	(0.007)					
Interest Rate Differential	1.0133	0.1597	0.2288	0	0.0012	0.0021	0.0018	-0.0051
	(0.000)	(0.000)	(0.000)					
NHG Indicator	-0.0849	-0.2888	0.0906	0	0.0003	-0.0040	-0.0013	0.0024
	(0.037)	(0.000)	(0.000)					
LTV	0.7372	-2.2314	0.1729	0	0.0011	-0.0409	0.0014	0.0383
	(0.000)	(0.000)	(0.000)					
After Anomaly Month	-0.2325	-0.1576	0.4144	0	-0.0003	-0.0028	0.0023	0.0008
	(0.043)	(0.002)	(0.000)					
January Indicator	-8.1023	-2.6730	-6.3870	0	-0.0082	-0.0438	-0.0389	0.0912
	(0.000)	(0.000)	(0.000)					
February Indicator	-7.7111	-3.0656	-6.4876	0	-0.0085	-0.0519	-0.0397	0.1004
	(0.000)	(0.000)	(0.000)					
March Indicator	-7.7584	-2.9383	-6.4897	0	-0.0083	-0.0514	-0.0395	0.0994
	(0.000)	(0.0052)	(0.000)					
April Indicator	-7.7588	-3.0506	-6.5300	0	-0.0087	-0.0520	-0.0407	0.1016
	(0.000)	(0.000)	(0.000)					
May Indicator	-7.7394	-2.9894	-6.4278	0	-0.0085	-0.0472	-0.0410	0.0969
	(0.000)	(0.000)	(0.000)					
June Indicator	-7.9062	-2.6363	-6.3106	0	-0.0073	-0.0442	-0.0388	0.0905
	(0.000)	(0.0049)	(0.000)					
July Indicator	-8.0446	-2.7981	-6.4158	0	-0.0077	-0.0467	-0.0399	0.0945
	(0.000)	(0.000)	(0.000)					
August Indicator	-8.2567	-3.8332	-6.3197	0	-0.00052	-0.0462	-0.0389	0.0788
	(0.000)	(0.000)	(0.000)					
September Indicator	-8.3829	-3.0283	-6.4059	0	-0.00089	-0.0490	-0.0397	0.0979
	(0.000)	(0.000)	(0.000)					
October Indicator	-8.3131	-3.0019	-6.5942	0	-0.0092	-0.0471	-0.0416	0.0981
	(0.000)	(0.000)	(0.000)					
November Indicator	-7.9031	-2.7001	-6.4698	0	-0.0102	-0.0466	-0.0405	0.0975
	(0.00440)	(0.000)	(0.000)					
December Indicator	-7.2064	-2.5682	-6.3302	0	-0.0072	-0.0431	-0.0383	0.0888
	(0.000)	(0.000)	(0.0047)					

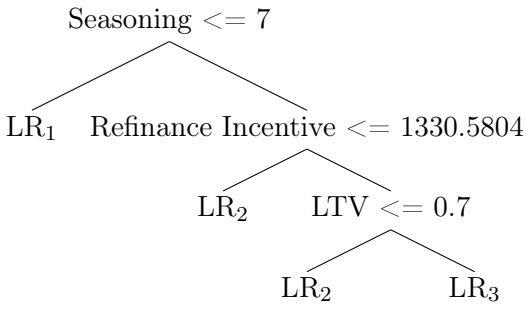
as people generally prefer to move house in the summer.

Multinomial Logit Model with Lasso penalty term

To find optimal level of penalization λ, 3-Fold cross-validation has been performed on the training data optimizing the out-of-sample Gini. The process has been described in further detail in 5.2.8. This resulted in a λ of 0.0001.

6.3 Linear Tree

In the MNL-Tree, we first fit the decision tree from which rules are extracted and used as indicator variables in an MNL model. Max depths ranging from one to three have been investigated to maintain interpretability. Hyperparameter tuning has been applied to each max depth model, in which we use cross-validation to maximize the error giving an optimal Max depth of 2 for the MNL Tree and 3 LR Tree. In step one of the algorithm DTs are formed to which linear models are then fit; the resulting DT'S are obtained below:
LR Tree:



MNL Tree:

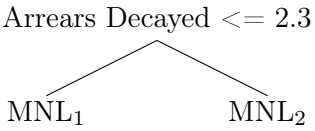


Figure 12 and 13 show the coefficients of the LR Tree and MNL Tree respectively. It can be see large number of non-zero coefficients for all of the models in the nodes. As can be seen even with short trees of a max depth of 3 it can get quite complex to explain the models. Interpreting 12 is more complex than LR beacasue you have multiple LR to stat nd you ahve to consider the spliting rule(s) that occurs which happen before the node which can have and effect on the coefficients.

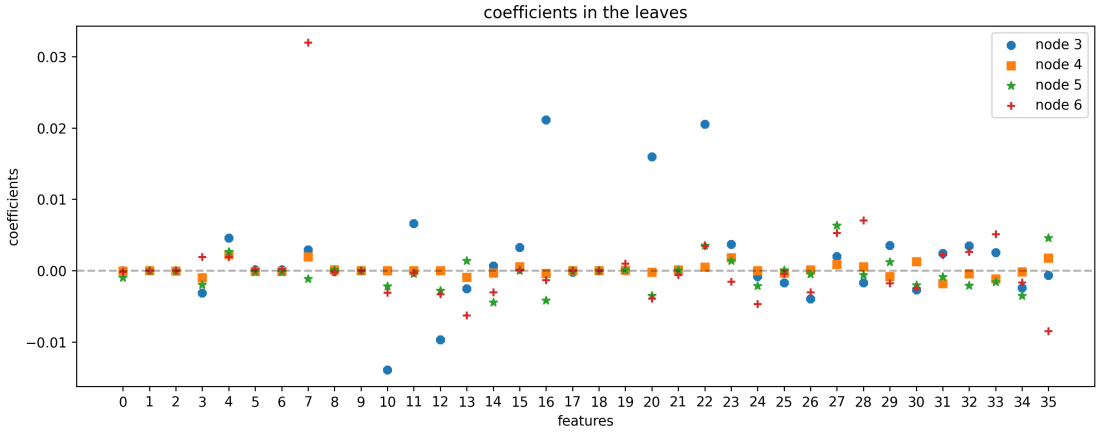


Figure 12: LR Tree Coefficients

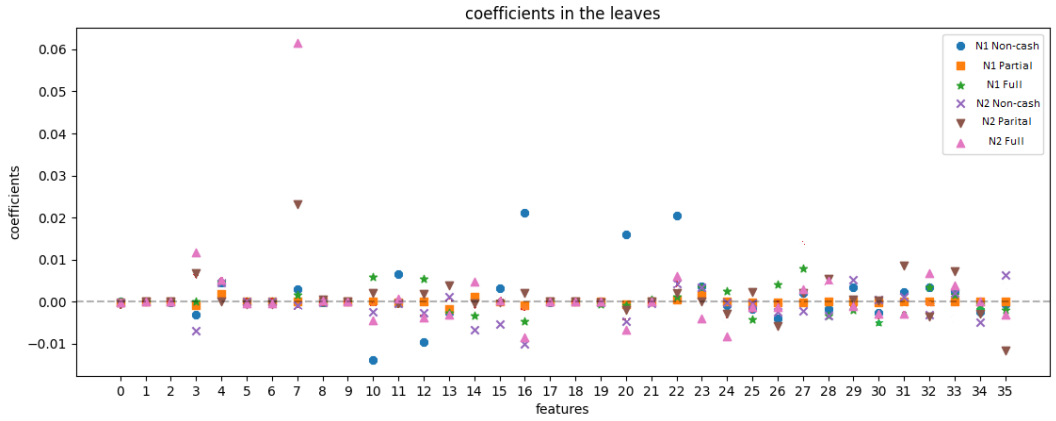


Figure 13: MNL Tree Coefficients

6.4 Random Forest

To reduce the probability of overfitting 3-fold cross-validation was performed; further details on motivation can be found in 5.2.8. The resulting optimal hyperparameters can be found below in table 4.

Table 4: Random Forest Hyperparameter table

		Treatment of Data	
Modeling Approach	Hyperparameter	Loan Level	Borrower Level
Classification	n_estimators	700	800
	min_samples_split	20	100
	min_samples_leaf	20	20
	max_leaf_nodes	48	18
	max_features	auto	auto
	max_depth	30	70
	class_weight	balanced	balanced
	ccp_alpha	0.01	10
Regression	n_estimators	800	1800
	min_samples_split	1000	15
	min_samples_leaf	4	15
	max_leaf_nodes	37	10
	max_features	auto	sqrt
	max_depth	10	90
	ccp_alpha	0	0.01

In order to gain some insight into which variables are important in the RF model for predicting CPR, a feature importance plot has been included.
Please note labels for variables can be found in Appendix C. Figure 14 shows that important variables for RFs can vary based on the type of approach used. For regression, the Number of Past Partial Prepayments and If any prepayment occurred previously are both very important for forecasting CPR, whereas it is not important for forecasting prepayment type through classification approach. In contrast, Arrears Decayed is Highly important for the Classification approach but not at all for regression.

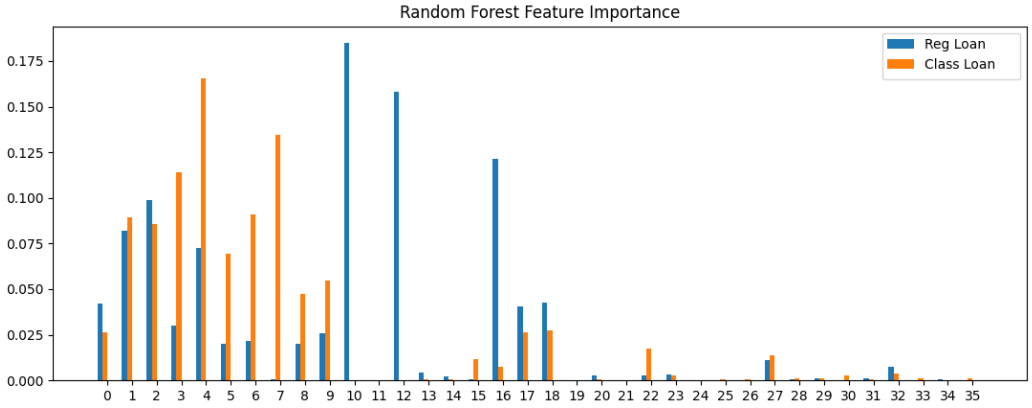


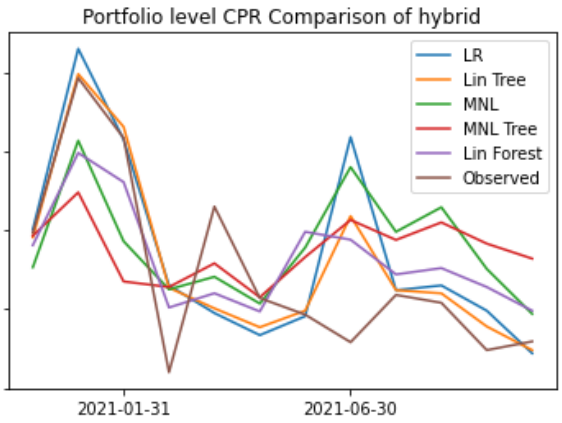
Figure 14: Feature importance for Random Forests on loan level

As explained in 5.2.6 it can be difficult to understand why this is the case; however, an interesting observation is that Arrears Decayed is not significant in the MNL model, implying that there must be a non-linear relationship between this variable and prepayment type which the linear MNL model is not able to capture.
Interestingly enough, the variables that both types of RFs agree are important are linked to commonly accepted and widely used variables in the literature. Richard and Roll (1989) model is heavily used forms the foundation of most prepayment models, in which they make use of four factors; Seasoning, seasonality, Burnout, and Refinance incentive. RFs agree that Seasoning, Month Till Maturity and Month Till Reset are all important and can all be thought of as being linked to one of four factors, Seasoning. Additionally, Interest Rate Differential and Refinance Incentive are important and heavily related to the Refinance Incentive factor. Lastly, the graph shows that all Burnout variables are relatively important. Therefore the only factor not being represented in the RFs is seasonality, with all of the month indicator variables having very low importance scores. With December and June having a slight increase in importance over the other months, this is warranted by the data, which shows increases in those months. However, they are still of quite low importance compared to the other non-seasonality variables.
From the graph, one can see that loan type is quite important in the classification and regression models, which agrees with finding in the LR and MNL models, which finds them significant.

6.5 Forecasting

A one-year multi-step forecast has been obtained for the various models outlined in the Methodology. The forecasting period starts in November 2020 and ends in October 2021. Forecasting will be done on a borrower or loan level depending on the type of model used; this will then be aggregated to the portfolio level, which will then be evaluated using metrics in 5.5.
First, the deterministic models will be discussed. The MNL model is the industry standard therefore, this can be the baseline model which the results will be compared to it. In figure 18 the MNL does a relatively good job, with CPR R Square scores of just over 30% meaning that on a portfolio level 30% t 30% of the CPR in data be explained by the MNL loan level model. Additionally 40 % of the projected cash flows on portfolio level can be be explained by the MNL loan level model. Furthermore, figure 17 the normalized delta is 4% meaning that the loan level MNL model estimates 4% more cash than observed for prepayments for the year forecasting period. From looking at figure

16 the MNL fails to fully capture the great increase increase which occurs in December and January, in addition to overestimation of the CPR which occurs in the summer months.



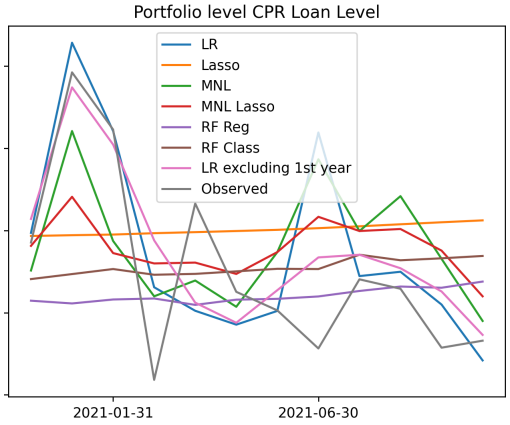


Figure 15 **Figure 16**

When the lasso penalty term is added to the MNL model, it produces less variation in the forecasted CPR, producing predictions in a simpler shape than MNL but with smaller extremes. Figure 17 shows an increase in normalized delta and figure 18 a reduction in both types of R squared values, which all point to an overall worse fit.

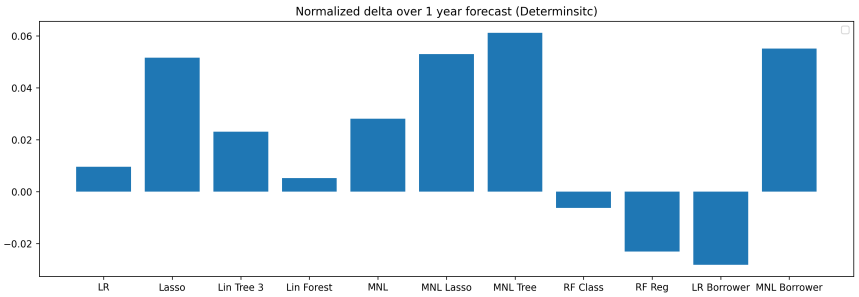


Figure 17: Loan Level Deterministic Forecasting Normalized Deltas

In figure 16 both of the RFs do not capture any of the seasonality in the data, with both of them resulting in a slightly increasing straight line for CPR on portfolio level. Figure 17 shows the RFs have some of the best-normalized deltas meaning that they do a good job of forecasting the exact level of prepayments. However, when looking at figure 18 the R squared of the predicted prepayment cash flows are with neither model obtaining a score above 20%. Furthermore, looking at their CPR R squared scores, they are inferior, with both obtaining negative scores, meaning that naively assuming that the average CPR on portfolio level would have given a better fit. The main issue with the RF models is their poor ability to capture seasonality; this could be due to the model's architecture, which recursively performs a binary split. Therefore the model is very good at finding clustering in the data; however, if data goes up incrementally but not enough to exceed the threshold, it will remain in the same cluster giving the same output.
2).

Now looking at the LR in figure 16 one can see that it fits the forecasts relatively well. It captures

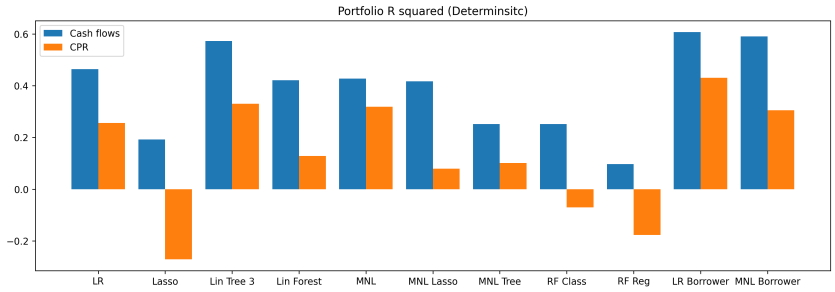


Figure 18: Loan Level Deterministic Forecasting R Squared

the CPR increase in December and January better than any model so far and significantly better than the MNL. It fails to capture the large decrease followed by a sharp increase in February and March. However, it does a reasonable job of averaging them out. The LR has a normalized delta of less than 1%, which is very good compared with the models. However, the R squared are not scores are not as high as expected, with the portfolio CPR R squared score being lower than the MNL model. The main reason for this is because of the overestimation in June, which has been skewed by the large anomaly which happened in June 2016 (seen in figure 2). To check if this was the case, we fit the same model with the same variables, only excluding the first year of data in training. This resulted in the pink line on figure 16. Now the graphs see smother increase over summer months (as expected according to past data seen in figure The main issue with the LR model that has just been demonstrated is the chance of overfitting the training data given the short data period. This promoted LR with a lasso penalty term; however, the resulting model gives an extremely sparse model, which results in a model with an extremely poor fit, which almost just looks like a horizontal line on portfolio level. This result shows that the LR is highly likely to be overfitting; however, it is not easy to test given a short data period. The RL with Lasso model has the worst fit-out of the deterministic forecasted with a portfolio CPR R squared of about -25% shown in figure 18.

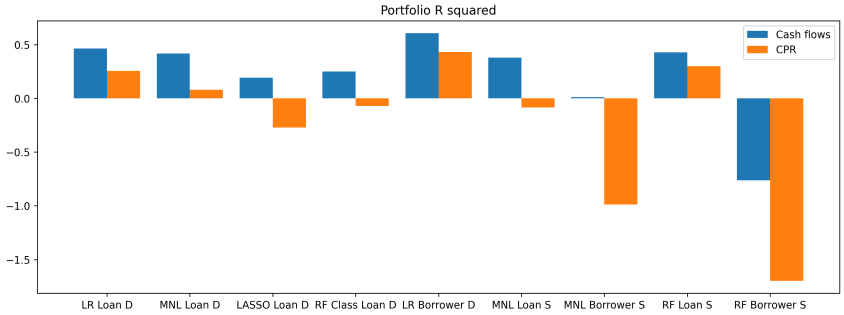


Figure 19: R Squared comparison for stochastic and deterministic forecasts

Now moving on to the Hybrid tree models, which include Linear trees and forests. Figure 15 shows the compassion of the Hybrid tree models, and the LR Tree has a very similar fit to LR; however, it seems to handle the June anomaly better. 18 confirms this because both R squared values have increased against the standard LR and MNL. The MNL Tree is seen to have a worse fit than the MNL with lower R square values and a higher normalized delta; this is mainly because it fails to correctly capture the increase in CPR in December and particularly January. This could be because

when the data is split using the tree part, now overarching seasonality that exists in the full data is not as Strong in the two subgroups; therefore, the models underestimate the CPR. The Linear Forest does much better than the RF, showing that a model that can capture incremental change in data rather than shift to a whole new node is better for our data set. However, Linear Forest seems to perform worse than the Linear Tree. Ensemble models are not universally better than corresponding "single" counterparts. The Ensemble will only give better prediction if the single models suffer from instability; this is unlikely to be true in our data set with many observations and variables. The MNL Tree does not forecast CPR and the normal MNL model. Clearly, there is not enough data for the clustering to occur for classification. Even at the lowest depth tree, data segmentation does not improve the fits of the 2 models.

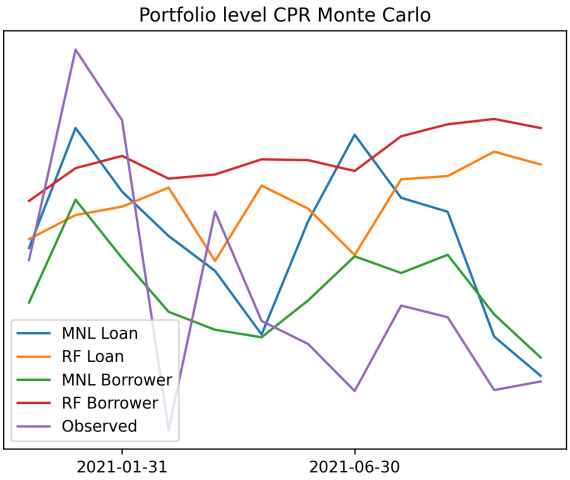


Figure 20

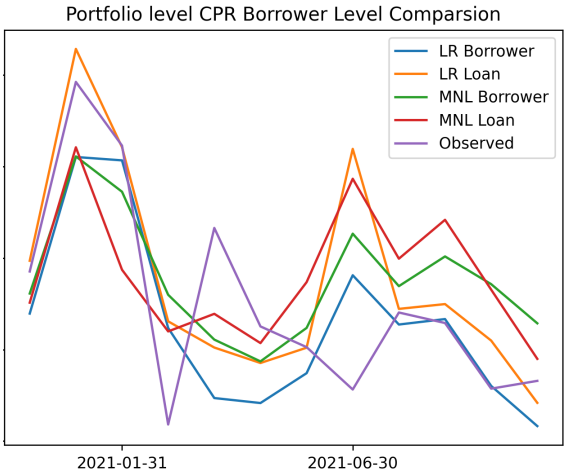


Figure 21

Figure 20 shows how the stochastic models are forecasted over time. It only makes sense to stochastically model classification models because they output the probability of different states, which can then be sampled from. 19 shows how the, generally that Monte Carlo is worse than its deterministic counterpart based on the R Squared. I believe that this is because the error carried forward in the Monte Carlo simulation is worse because if CPR is wrong in the first month, this will impact the notional in the time period, which affects CPR in the following time period. Therefore, this approach of choosing a state that occurs punished model in the month after. Implying that taking a deterministic approach would be less wrong most of the time s, it can be considered an average. Th fits get progressively worse as we move from right to left on Figure 20. Interestingly, the RF performs considerably more than its deterministic version. implying the RFs predictions are more accurate than their probability and the CPR's trending upwards.
Figure 21 shows the one year forecast of the LR and MNL models on both loan and borrower levels. On the Borrower level, the RL is more or less below the loan level LR forecast for forecasted dates. It does not capture the increase in December; however, it better handles anomalies and ends the last half of the forecast. Furthermore, figure 18 has the best portfolio R squared for both by both measures. The Borrower MNL forecasts very similarly to loan-level one, but it has lower CPR R squared, but a higher prepayment cash flow R squared.

7 Conclusion

Throughout the thesis, various methods and models to produce multi-step forecasting of model for prepayment rate have been investigated to find which gives the most accurate CPR on portfolio level.

Chapter 1 gives a background to the problem and why a bank needs to forecast mortgage prepayments, and what the research adds to the current literature. Chapter 2 gives a more detailed background on mortgage information, in which mortgage and prepayment types are described. Furthermore, the various reasons a borrower may prepay their mortgage are explained and how the prepayment penalty works. In the next chapter, a literature review is conducted in which the various prepayment motives have been proved to be helpful when predicting CPR. Additionally, a comprehensive review of all the current literature on predicting prepayment, including portfolio-level models, option theoretic models, exogenous models, multi-step forecasting models and a brief review of making Machine Learning models more interpretable.

In chapter 4, the two different methods of data aggregation are described and the reason for choosing each. A thorough descriptive analysis was carried out, in which anomalies were discovered. After which, it is explained how our data set is unbalanced and why this can be an issue for Machine Learning classification models. Next, it discusses how adjustments for the imbalanced data set can be made to the data or the model and why model-based adjustments have been preferred.

In chapter 5, the research methodology was laid out; first, the two different approaches (regression vs classification) explained that classification is usually used for forecasting CPR. However, because additional terms in the classification approach all bring their own error, a straightforward regression might be more accurate. The multiple models are then explained, including Linear models, Linear Trees, Random Forests, and Linear Forests, each of which has regression and classification variants. In the Random Forest, it discussed the purpose of adjusting for an unbalanced for the Random Forest classification model. Furthermore, it is discussed how Hyperparameter Tuning is used to helping find the optimal models. It is explained how the process of the multi-step CPR foresting takes place for both deterministic and stochastic forecasting. Additionally, the process for finding optimal parameters is described. Before the chapter finishes with the model performance metrics, it describes how the CPR should be evaluated on a portfolio level as a pose to the modelling level.

Chapter 6 shows the results of the analysis. It has been found that common drivers in the literature such as burnout, seasoning, refinance incentive and seasonality are significant drivers of prepayment in our data set. Contrary to current literature, it has been found that using regression to forecast CPR directly is an effective method of modelling prepayment. Furthermore, modelling CPR on borrower level as a pose to loan-level yields more accurate results for Linear Regression. Given our data set, the linear models (Multinomial Logit and Linear Regression) outperform the Random Forest, with the Random Forest failing to detect seasonality. However, there exists the possibility of overfitting due to the short data period and evidenced when removing larger outlier months model fits forecasted CPR improves. It has been shown that Hybrid models such as Linear Trees and Linear Forests do not make enough of a justifiable increase in forested CPRs worth using a more complex model to forecast CPRs. Lastly, forecasting CPRs deterministically as opposed to stochastically yields more accurate CPRs.

8 Future Research

The main limitation of the research is the data length, with only data from the end of 2015 to 2021 available. Fitting models to more extensive data periods where more "states" exist may improve the performance of RFs, as it could better account for the clusters. An example of a different scenario is how would this model perform in a rising interest rate environment? Furthermore, the testing period is only limited to one year, which is not ideal because the test data could potentially have anomaly months in it, making it hard to tell if the model is performing poorly or not. An excellent example of this could be in June, as all models over-predict CPR. This persists even when the first year of training data is excluded (which includes the June anomaly(which skews the model)).
One further model was intended to be used but did not get approved by Knab in time. They were the RUX and RUG models developed by Akyüz and Birbil (2021) which could be used to improve the interpretability of the RF. RF has not been shown to perform well given our dataset; however, this could be different given a larger data period. Additionally, it could be used in the first step of the Linear Trees to increase the stability of the model.
It could be interesting to look into macroeconomic variables. This study did not include them in our model because the main focus was on deterministic forecasting. In deterministic forecasting, stochastic variables must remain constant. Therefore it did not make sense to forecast for a long data period assuming macroeconomic variables are constant. Therefore I propose using macroeconomic models to forecast these variables into the future, which will hopefully increase the accuracy of the stochastic forecasting models. Another way to improve the stochastic model is to use an EM algorithm on the change in interest rate differential to find the transition probability of going from one state to another. At the moment, we assume a normal mixture model for which the probability of the next state does not depend on the current state, which economically does not make sense.
The purpose of forecasting mortgage prepayments is so they can be taken into account when hedging interest rate risk. The bank will have different costs depending on whether the model has false negative or false positive; therefore, this could be taken into account using cost-sensitive learning.

References

StatLine key data sectors; national accounts. `https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84097NED/table?ts=1607329979735`. Accessed: 2021-11-20.

M. Hakan Akyüz and S. Ilker Birbil. Discovering classification rules for interpretable learning with linear programming. *ArXiv*, 2021.

B. J. Alink. Mortgage prepayments in the netherlands. *Ph. D. thesis, University of Twente*, 2002.

W.R. Archer and D.C. Ling. Pricing mortgage-backed securities: Integrating optimal call and empirical models of prepayment. *Estate and Urban Economics Association*, 21(4):373–404, 1993.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, 2012.

Arnab Bhattacharya, Simon P. Wilson, and Refik Soyer. A bayesian approach to modeling mortgage default and prepayment. *European Journal of Operational Research*, 274(3):1112–1124, 2019.

Gabriel Blumenstock, Stefan Lessmann, and Hsin-Vonn Seow. Deep learning for survival and competing risk modelling. *Journal of the Operational Research Society*, pages 1–13, 2020.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification And Regression Trees*. 10 1984.

Michael J. Brennan and Eduardo S. Schwartz. Determinants of gnma mortgage prices. *Real Estate Economics*, 13:209–228, 1985.

R. Busschers. Cash flow modelling for residential mortgage backed securities: a survival analysis approach. Master's thesis, University of Twente, 09 2011.

Arjan Van Bussel. Valuation and interest rate risk of the mortgages in the netherlands. *Ph. D. thesis, University of Maastricht*, 1998.

Clément Bénard, Gérard Biau, Sébastien da Veiga, and Erwan Scornet. Sirus: Stable and interpretable rule set for classification, 2020.

Charles A. Calhoun and Yongheng Deng. *A Dynamic Analysis of Fixed- and Adjustable-Rate Mortgage Terminations*, pages 9–33. Springer US, Boston, MA, 2002.

Tim S Campbell and J Kimball Dietrich. The determinants of default on insured conventional residential mortgage loans. *Journal of Finance*, 38(5):1569–81, 1983.

Charles Capone, Jr and Donald Cunningham. Estimating the marginal contribution of adjustable-rate mortgage selection to termination probabilities in a nested model. *The Journal of Real Estate Finance and Economics*, 5:333–56, 02 1992.

Kin-Yee Chan and Wei-Yin Loh. Lotus: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, 13:826 – 852, 2004.

Erwin Charlier and Arjan Van Bussel. Prepayment Behavior of Dutch Mortgagors: An Empirical Analysis. *Real Estate Economics*, 31(2):165–204, 06 2003.

Erwin Charlier and Arjan Van Bussel. Prepayment behavior of dutch mortgagors: An empirical analysis. *Real Estate Economics*, 31(2):165–204, 2003.

Probal Chaudhuri, M. Huang, Wei-Yin Loh, and R. Yao. Piecewise-polynomial regression trees. *Statistica Sinica*, 4:143–167, 01 1994.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 06 2002.

Chao Chen and Leo Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, 01 2004.

J Chen, Chen-An Tsai, Hojin Moon, H Ahn, J Young, and C.-H Chen. Decision threshold adjustment in class prediction. *SAR and QSAR in environmental research*, 17:337–52, 07 2006.

Shirish Chinchalkar and Roger M. Stein. Comparing loan-level and pool-level mortgage portfolio analysis. *Moody's Research Labs*, 11 2010.

Deng Y. Clapp, J. M. and X. An. Unobserved heterogeneity in models of competing mortgage termination risks. *Real Estate Economics*, 34(2):243–273, 2006.

John M. Clapp, Gerson M. Goldberg, John P. Harding, and Michael LaCour-Little. Movers and shuckers: Interdependent prepayment decisions. *Real Estate Economics*, 29(3):411–450, 2001.

J.P. Harding Clapp, J.M. and M. LaCour-Little. Expected mobility: Part of the prepayment puzzle. *The Journal of Fixed Income*, pages 68–78, 06 2001.

Yongheng Deng, John M. Quigley, and Robert Van Order. Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, 68(2):275–307, 2000.

Elena-Ivona Dumitrescu, Sullivan Hué, Christophe Hurlin, and Sessi Tokpavi. Machine learning or econometrics for credit scoring: Let's get the best of both worlds. *SSRN Electronic Journal*, 01 2020.

Kenneth B. Dunn and John J. McConnell. Valuation of gnma mortgage-backed securities. *The Journal of Finance*, 36(3):599–616, 1981.

Charles Ewing, Amber Ilyas, Felisja Kuci, Daniel Lingsveld, and Marina Tawfik. Modelling prepayment behaviour. 03 2021.

M. C. Findley and D. R. Capozza. The variable rate mortgage: an option theory perspective. *Journal of Money, Credit and Banking*, 9:356–364, 2003.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, 2014.

Afshin Goodarzi, Ron Kohavi, Richard Harmon, and Aydin Senkut. Loan prepayment modeling. pages 62–69, 1998.

Jerry Green and John B Shoven. The effects of interest rates on mortgage prepayments. *Journal of Money, Credit and Banking*, 18(1):41–59, 1986.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models, 2018.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

Lakhbir Hayre. *Salomon Smith Barney Guide to Mortgage-Backed and Asset-Backed Securities*, page 24. Wiley, 2001.

Lakhbir S Hayre. Prepayment modeling and valuation of dutch mortgages. *The Journal of Fixed Income*, 12(4):25–47, 2003.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Pan Kang and Stavros Zenios. Complete prepayment models for mortgage-backed securities. *Management Science*, 38(11):1665–1685, 1992.

Aram Karalic. Linear regression in regression tree leaves. In *In Proceedings of ECAI-92*, pages 440–441. John Wiley Sons, 1992.

James B. Kau and Donald C. Keenan. An overview of the option-theoretic pricing of mortgages. *Journal of Housing Research*, 6(2):217–244, 1995. ISSN 10527001. URL `http://www.jstor.org/stable/24832827`.

Michael LaCour-Little, Michael Marschoun, and Clark L. Maxam. Improving parametric mortgage prepayment models with non-parametric kernel regression. *Journal of Real Estate Research*, 24 (3):299–328, 2002.

Onno Lijmbach, Jeroen Lemsom, Dylan van Gemeren, and Luc van Breukelen. Predicting mortgage prepayment behaviour for knab by promoting sparsity and exploiting nonlinearity. 03 2021.

Charles X Ling and Victor S Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*, 2011:231–235, 2008.

J.S. Long, J.S. Long, and J. Freese. *Regression Models for Categorical and Limited Dependent Variables*. Advanced Quantitative Techniqu. SAGE Publications, 1997.

Gilles Louppe. *Understanding Random Forests: From Theory to Practice*. PhD thesis, 10 2014.

Janneke Meis. Modelling prepayment risk in residential mortgages. 11 2015.

Romy Mieras, Milad Agha, Annabel de Boer, and Daan Gieles. Modelling default and prepayment behavior of a dutch mortgage portfolio. 03 2021.

John Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, 4(2):227–243, 11 1989.

Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6):275–285, 2004.

Yasuo Nishiyama. Are banks risk-averse? *Eastern Economic Journal*, 33(4):471–490, 2007.

Simon Perry, Stuart D. G. Robinson, and J. W. Rowland. A study of mortgage prepayment risk. 2001.

Jason Phang, Jungkyu Park, and Krzysztof J. Geras. Investigating and simplifying masking-based saliency methods for model interpretability, 2020.

Ivilina Popova, Elmira Popova, and Edward George. Bayesian Forecasting of Prepayment Rates for Individual Pools of Mortgages. 01 2008.

Surachet Pravinvongvuth, Ph.D. and Anthony Chen. Adaption of the paired combinatorial logit model to the route choice problem. *Transportmetrica*, 1:223–240, 01 2005.

Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3), Jan 2019. ISSN 1942-4795. doi: 10.1002/widm.1301. URL `http://dx.doi.org/10.1002/widm.1301`.

Laura Raileanu and Kilian Stoffel. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41:77–93, 05 2004.

Scott F. Richard and Richard Roll. Prepayments on fixed-rate mortgage-backed securities. 1989.

Taiyo Saito. Mortgage prepayment rate estimation with machine learning. Master's thesis, TU Delft, 7 2018.

Eduardo S. Schwartz and Walter N. Torous. Prepayment and the valuation of mortgage-backed securities. *The Journal of Finance*, 44(2):375–392, 1989.

Eduardo S. Schwartz and Walter N. Torus. Prepayment and the valuation of mortgage-backed securities. *The Journal of Finance*, 44(2):375–392, 1989.

Tyler Shumway. Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1):101–24, 2001.

Justin Sirignano. Deep learning for limit order books, 2016.

Justin Sirignano, Apaar Sadhwani, and Kay Giesecke. Deep learning for mortgage risk. 11 2018.

R Stanton. Rational prepayment and the valuation of mortgage-backed securities. *The Review of Financial Studies*, 8(3):677–708, 1995.

Dan Steinberg and Nicholas Scott Cardell. The hybrid cart-logit model in classification and data mining. 1998.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9:307, 08 2008.

Yuliya Subotniaya. Prepayment risk modeling of dutch mortgages: A neural networks approach. Master's thesis, University of Amsterdam, 6 2018.

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

Yorick Wanders, Gijs Propper, Thierry Volker, and Bouwe van der Wal. Modelling and forecasting prepayment rates in a mortgage portfolio. 03 2021.

Haozhe Zhang, Dan Nettleton, and Zhengyuan Zhu. Regression-enhanced random forests. 04 2019.

Peter M. Zorn and Michael Lea. Mortgage borrower repayment behavior: A microeconomic analysis with canadian adjustable rate mortgage data. *Real Estate Economics*, 17:118–136, 1989.

Appendices

A Variance-Bias Decomposition

$$
\begin{aligned}
\mathrm{E}\left[(y-\hat{f})^2\right] &= \mathrm{E}\left[(f+\varepsilon-\hat{f})^2\right] \\
&= \mathrm{E}\left[(f+\varepsilon-\hat{f}+\mathrm{E}[\hat{f}]-\mathrm{E}[\hat{f}])^2\right] \\
&= \mathrm{E}\left[(f-\mathrm{E}[\hat{f}])^2\right] + \mathrm{E}[\varepsilon^2] + \mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right] + 2\,\mathrm{E}\left[(f-\mathrm{E}[\hat{f}])\varepsilon\right] + 2\,\mathrm{E}\left[\varepsilon(\mathrm{E}[\hat{f}]-\hat{f})\right] + 2\,\mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})(f-\mathrm{E}[\hat{f}])\right] \\
&= (f-\mathrm{E}[\hat{f}])^2 + \mathrm{E}[\varepsilon^2] + \mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right] + 2(f-\mathrm{E}[\hat{f}])\,\mathrm{E}[\varepsilon] + 2\,\mathrm{E}[\varepsilon]\,\mathrm{E}\left[\mathrm{E}[\hat{f}]-\hat{f}\right] + 2\,\mathrm{E}\left[\mathrm{E}[\hat{f}]-\hat{f}\right](f-\mathrm{E}[\hat{f}]) \\
&= (f-\mathrm{E}[\hat{f}])^2 + \mathrm{E}[\varepsilon^2] + \mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right] \\
&= (f-\mathrm{E}[\hat{f}])^2 + \mathrm{Var}[\varepsilon] + \mathrm{Var}\left[\hat{f}\right] \\
&= \mathrm{Bias}[\hat{f}]^2 + \mathrm{Var}[\varepsilon] + \mathrm{Var}\left[\hat{f}\right] \\
&= \mathrm{Bias}[\hat{f}]^2 + \sigma^2 + \mathrm{Var}\left[\hat{f}\right].
\end{aligned}
$$

(29)

B Penalty Model

$$
\text{Penalty Fee} = \mathbb{1}\left[\begin{Bmatrix}\text{Prepay}\\ \text{amount}\end{Bmatrix} > \frac{\text{Principle}}{10}\right] \mathbb{1}\left[\text{Curr Rate} < \text{Mor Rate}\right] \sum_{i=1} N\left[\begin{Bmatrix}\text{Prepay}\\ \text{amount}\end{Bmatrix} - \frac{\text{Principle}}{10}\right] \left[\text{Curr Rate} - \text{Mor Rate}\right] \frac{1}{12i^{(1+\text{Curr Rate})}} \quad (30)
$$

C Variable notations

- 0 - Seasoning
- 1 - Month Till Maturity
- 2 - Month Till Reset
- 3 - LTV
- 4 - Interest Rate Differential
- 5 - Refinance Incentive
- 6 - Burnout (Max past Refinance Incentive - current Refinance Incentive)
- 7 - Arrears Decayed
- 8 - Burnout (Max past Refinance Incentive)
- 9 - Burnout (Sum of past Refinance Incentive)
- 10 - Number of Past Partial Prepayment
- 11 - Number of Past Non-cash Prepayment
- 12 - If any prepayment occurred previously
- 13 - NHG
- 14 - Annuity Indicator
- 15 - Linear Indicator
- 16 - Saving Indicator
- 17 - Seasoning^2

- 18 - Seasoning[3]
- 19 - Business Property Indicator
- 20 - Family House Indicator
- 21 - Other Property Indicator
- 22 - Apartment Indicator
- 23 - After Anomaly Month Indicator
- 24 - January Indicator
- 25 - October Indicator
- 26 - November Indicator
- 27 - December Indicator
- 28 - February Indicator
- 29 - March Indicator
- 30 - April Indicator
- 31 - May Indicator
- 32 - June Indicator
- 33 - July Indicator
- 34 - August Indicator
- 34 - September Indicator