
# Boosting Performance of Traditional Models in Prepayment Modelling

**Erasmus University Rotterdam**

*Erasmus School of Economics*

Master Thesis Quantitative Finance

*Author*:

D.M. van den Donk (500277)

*Supervisors*:

Prof. Dr. R.L. Lumsdaine (Erasmus University)

M. Mackaij, B. Maarse (Deloitte)

February 4, 2020

**Abstract**

This thesis uses the Freddie Mac Single Family Loan Level Data Set to investigate if a machine learning algorithm called gradient boosting can outperform a multinomial logit model in monthly prepayment predictions for mortgages. If financial institutions can correctly predict prepayments, they can hedge risks and price mortgages better. Additionally, this thesis uses a model interpreter called Shapley Additive exPlanations (SHAP) to interpret the XGBoost model. The XGBoost is better in predicting prepayments than the multinomial logit model. Using SHAP values, this thesis finds that XGBoost is better able to capture the non-linear dependencies of prepayment events on explanatory variables. Although prepayment dynamics are better captured with the XGBoost model, both models are not able to discriminate well between a full prepayment and other prepayment classes on a monthly basis.

# Contents

# 1 Introduction

A mortgage is a contract between a mortgagor and a mortgagee. The mortgagor borrows money from the mortgagee at a specific interest rate to buy property. The property becomes the underlying of the loan, meaning the mortgagee gets ownership of the property in case of a default of the mortgagor. According to the Federal Reserve, total mortgage debt in the US totals 15.5 trillion Dollars (Federal Reserve, 2019), making it a market with a size of approximately 75% of the United States GDP (Bureau of Economic Analysis, 2019).

The prepayment of a mortgage is a (partial) settlement of the mortgage contract before it matures. Hayre (2003) states that indifferent to country, mortgage type or market, four distinct types of prepayments can be defined. These are defaults[1], mobility, refinancing, and voluntary full or partial payoffs, each with its distinct risk drivers.

Mobility is one of the prepayment types, as the sale of a house prompts a full prepayment. Exceptions exist due to, e.g. portability in the Netherlands or assumable loans in the United States. Portability entails taking a mortgage contract from one property to another property and an assumable loan gives the buyer of a house the option to transfer the mortgage that is currently on the house to herself, usually with the same terms. Generally however, these exceptions do not occur.

Additionally, refinancing is a prepayment type. A refinance event happens when a mortgagor keeps the underlying property and prepays her mortgage with another mortgage, usually with better terms. Refinancing is an interesting type of prepayment, since it is solely determined by financial incentive. Generally, refinancing occurs when interest rates fall and mortgagors want to take advantage of a lower interest rate. This is not the only reason, however. Refinancing can occur due to an improved credit score. If the credit score of a mortgagor improves, a lender can offer better rates and hence it is in the interest of the mortgagor to change mortgage. This is a credit-driven refinancing. Also, a so called cash-out refinancing occurs when a mortgagor wants to transform part of the build up equity in the house into cash. According to Hayre (2003), the refinancing rate is typically modelled as a logistic function of the incentive to refinance. The refinancing rate gradually rises when the refinance incentive increases. However, after a certain incentive to refinance, the effect

[1]Defaults make up part of the prepayment space. However, risk drivers of defaults are different than that of other prepayment types and require a different modelling approach. They will only add noise to the model. For this reason, and because the academic literature on defaults is vast and elaborate, e.g. Foster and Van Order (1984); Boyes, Hoffman, and Low (1989); Altman and Saunders (1997); Ghent and Kudlyak (2011); Crosbie and Bohn (2019), defaults are not included in this study.

of a higher incentive is only marginal due to the already increased refinance incentive.

Last, a borrower can opt to voluntarily make a full or partial prepayment. Whereas full prepayments usually invoke a penalty, this is not the case for partial prepayments below a certain threshold, also known as curtailments.

The prepayment option means that banks may expect to lose a portion of the contractual interest rate earnings. This risk can be viewed from two different perspectives, interest rate risk and liquidity risk. Interest rate risk arises from both the missing interest rate payments, as well as the risk of not getting the same or a better interest rate when a new mortgage is given out. Liquidity risk on the other hand, arises from the mismatch in expected cash-flows within the bank. A prepayment model forecasts the prepayment rate or the chance of a certain type of prepayment. These models enable banks to correctly price mortgages and hedge interest rate risk correctly.

Prepayments have been modelled in many different manners. The most common are survival analysis (Jacobs, Koning, & Sterken, 2005), a multinomial logit model (Clapp, Goldberg, Harding, & LaCour-Little, 2001; Vasconcelos, 2010) and an option theoretic framework (Varli & Yildirim, 2015; Goncharov, 2002; Kang & Zenios, 1992). The challenge in predicting prepayments is modelling the behaviour of mortgagors. There are frameworks that assume rational financial behaviour, such as the option theoretic framework. However, prepayments are made, or not made, for a variety of financially irrational reasons (Charlier & Van Bussel, 2001).

The mentioned prepayment models are all linear models that are explainable and have a high degree of interpretability. This is desirable in economics, because then a relationship between variables can be easily explained. For example, with every increase of 1 in the loan age, the probability of a prepayment increases by 0.5 percent points. However, using these models implies that no non-linear relationships are captured correctly. Such relationships are only approximated linearly and thus there is potential for non-linear models to improve upon performance, relative to the aforementioned models.

An example of those non-linear models are machine learning models. Machine learning models are statistical algorithms that minimize a certain cost or loss function by delving into the relationships within the data. Despite recent breakthroughs in and a proven record of machine learning in prepayment forecasting (Sirignano, Sadhwani, & Giesecke, 2015; Riksen, 2017; Guelman, 2012), the financial world is hesitant to implement these models (Brainard, 2018). The opaque nature of machine learning models makes financial institutions prefer easily explainable models that give a clear relationship between, e.g., the prepayment probability and the savings rate. Nonetheless,

aforementioned studies show machine learning approaches work well in prepayment modelling.

One particularly interesting machine learning model is gradient boosting (Friedman, 2001). Gradient boosting can be used in a classification and regression setting. It uses a base model, which is a decision tree. Gradient boosting iteratively fits a new base model on the errors of the previous model. Then the results of the new base model and previous model are combined to create a new model. This process is known as *boosting*. Gradient boosting is highly regarded for its predictive power (Mangal & Kumar, 2016; Ben Taieb & Hyndman, 2014) and might be able to capture the non-linear relationships in prepayment data better than current (linear) models. It is not known to be a good model in a prepayment setting, because as far as the writer of this thesis could find, it is not yet used in a prepayment setting.

Therefore, in order to explore if gradient boosting is indeed a good model for prepayment forecasting the main research question of this thesis is: "Can a gradient boosting algorithm outperform a multinomial logit in monthly prepayment forecasting?". The data set that the models are fitted on is the Single Family Loan Level Data Set from Freddie Mac (Freddie Mac, 2019) enriched with macroeconomic variables. The prepayment classes that are modelled are full prepayment, partial prepayment and no prepayment event. For this research, a specific gradient boosting algorithm called XGBoost (Chen & Guestrin, 2016) is used as a classifier. It is quick compared to other gradient boosting algorithms, has a high performance and is relatively easy to implement.

Although machine learning models are unpopular in financial institutions due to their opaqueness, methods have been invented to increase the explainability of these models. Methods such as Shapley Additive explanations (SHAP) (Lundberg & Lee, 2017) have been introduced into the field of interpretable machine learning. SHAP values of a variable give the attributions of a feature to the output of the model. This provides a researcher with the insight of what contributed to the outcome of a model. This thesis also investigates the inner workings of the XGBoost model with SHAP. By using SHAP values, the relationship of the most important features with the dependent variable is displayed.

The XGBoost algorithm is found to have superior performance over the multinomial logit model in a prepayment setting. However, the XGBoost model overestimates the total number of partial prepayments. Hence, a probability distribution is multiplied with the probabilistic outcome of the XGBoost model to make the model more conservative in predicting partial prepayment. This further increases the forecasting performance of the XGBoost model.

Furthermore, the XGBoost model is analyzed by using SHAP values. The SHAP values provide

explanations on how the model comes to prepayment probabilities. The most important contributor to the model is the partial prepayment flag, which is 1 if a mortgagor has already done a partial prepayment. Most relationships between a feature and the model output are found to be highly non-linear. Monthly income is an important feature for the XGBoost model, but is found to have no influence in the linear model.

This thesis adds to a growing academic literature involving machine learning in prepayment modelling. The XGBoost algorithm performs well and it is possible to explain such a model by using SHAP values. This is a step in the direction of acceptance of machine learning within the financial industry and with financial authorities.

# 2 Literature review

This section discusses the academic literature around prepayment modeling. The literature review is divided into four sections. Section 2.1 elaborates on the risk drivers of prepayments. Section 2.2 discusses models used in prepayment modelling and Section 2.3 introduces machine learning. Finally, Section 2.4 provides academic findings on interpretable machine learning.

## 2.1 Risk drivers of prepayments

Each type of prepayment has its own risk drivers. Clapp et al. (2001) show that modeling refinance and mobility as distinct prepayment types improves predictions significantly. Although the majority of risk drivers has a similar effect on refinancing and mobility, some variables have opposite effects, such as income. In this research, four types of risk drivers are defined: *personal characteristics*, *loan characteristics*, *macroeconomic factors* and *seasonality*.

### 2.1.1 Personal characteristics

Age is found to be a driver of prepayments in several studies. It negatively influences mobility but has no significant effect on refinancing (South & Crowder, 1998; Clapp et al., 2001). This is also true for families when age of the family head is used (Quigley & Weinberg, 1977).

Due to less access to moving or refinancing opportunities, being part of a minority reduces the probability of prepaying (Yinger, 1997). South and Crowder (1998) indeed confirm that in the United States certain races have lower mobility rates, the most notable being African-Americans.

Different income levels have different effects on different types of full prepayments. Lower income

households do not show different refinance behaviour but are more reluctant to move than higher income households, possibly due to an increased percentage that they must finance because of less accumulated wealth. This lowers full prepayment risks for lower income levels (Archer, Ling, & McGill, 2003). Higher income households have a higher opportunity cost for refinancing and hence the probability of refinancing is negatively influenced (Clapp et al., 2001). However, this is compensated by a positive relationship between income and the probability of moving (South & Crowder, 1998; Clapp et al., 2001). The total influence of a higher income on full prepayment probabilities is found to be not significant different from zero (Clapp et al., 2001). Similar to a high income, a poor credit history results in a lower probability of refinancing (Bennett et al., 2001).

Another important factor explaining prepayment behaviour is burnout (Hayre, 2003). It entails that at some point while interest rates are decreasing, the prepayment rate also decreases. This seems odd at first, because the refinance incentive increases. However, most mortgagors have already refinanced and the remaining mortgagors are unable to do so due to not enough equity or creditworthiness. Additionally, Hayre (2003) notes that there is a media effect which entails that in times of prolonged historically low interest rates, the media covers this phenomenon and a bigger part of the population assumes a new mortgage, countering the burnout effect.

### 2.1.2 Loan characteristics

Loan characteristics include all risk drivers that are specific to a certain mortgage. The major risk driver in this category is loan-to-value (LTV). A high initial LTV ratio negatively impacts prepayment rates compared to a low initial LTV ratio (Bennett et al., 2001; Archer et al., 2003). If current LTV is taken, a negative relationship is visible (Deng, Quigley, & Order, 2000). Besides studying the impact of a high LTV ratio, Archer et al. (2003) show that debt-to-income (DTI) ratios have a negative relationship with prepayment probability. Moreover, the contract mortgage rate of the loan is of interest, but this is elaborated on in the section on macroeconomic factors below.

Research shows that mortgagors signal their intended behaviour by choice of mortgage products, e.g. Dunn and Spatt (1988). An example of this choice is the loan term. Clapp et al. (2001) find that mortgages with a maturity of 15 years have a lower prepayment probability than mortgages with a maturity of 30 years. Furthermore, original loan balance has a positive effect on the probability of refinancing because transaction costs are more likely to be covered due to the higher dollar amount benefit of refinancing (Clapp, Deng, An, & Xudong, 2006).

### 2.1.3 Macroeconomic factors

Prepayment decisions can be attributed to personal or loan specific characteristics, but also economic circumstances are an important driver of prepayment risks (Pavlov, 2001). Of these macroeconomic variables driving prepayments, the current market mortgage rate is the most important and also widely used in the academic literature, e.g. Green and Shoven (1986); Deng et al. (2000); Richard and Roll (1989); Clapp et al. (2006). The current interest rate can have a refraining or accelerating influence on full prepayments: the *lock-in effect* and the *refinance incentive*. The former entails that when the market mortgage rate is above the contract rate, the probability of a full prepayment goes down (Green & Shoven, 1986; Clapp et al., 2006). The home owners are reluctant to move due to the higher interest rate on a new mortgage. This is the lock-in effect. The refinance incentive means that when the market mortgage rate is below the contract rate, the probability of a full prepayment rises (Green & Shoven, 1986). In that case, it is rational to obtain a new mortgage with a lower rate and hence with a lower monthly installment; the refinance incentive is high.

Mortgage rate, however, is not the only important macroeconomic risk driver. Deng et al. (2000) show that unemployment rates as well as divorce rates have positive effects on full prepayment rates. The positive effect of unemployment rates on full prepayments is also shown by Pavlov (2001).

Additionally, the appreciation or depreciation of house prices plays an important role in full prepayment behavior. Clapp et al. (2001) find that it can influence the decision of a mortgagor to move and hence to prepay. An appreciated house can offer a substantial surplus, whereas a depreciated house can leave a mortgagor with a debt. Finally, Caplin, Freeman, and Tracy (1997) note that prepayment rates can decline as much as 50% during recessions, making GDP growth an important prepayment driver.

### 2.1.4 Seasonality

Full prepayments occur more in the summer than in the winter, due to e.g. school holidays or a better weather to move (Schwartz & Torous, 1989). Conversely, Charlier and Van Bussel (2001) find that in the Netherlands partial prepayments occur more in the December. Seasonality can be modeled by, e.g. using a dummy for certain months or periods (Charlier & Van Bussel, 2001) or using a sine wave to model yearly seasonality (Spahr & Sunderman, 2001).

## 2.2 Prepayment models

Initially, prepayments were only modeled using option theory, since mortgages entail an option to prepay. Options are often modeled using rationality assumptions. However, empirically it is shown that prepayments do not appear to be rational and hence using option theory might not be optimal (Vandell, 1995). Also, the option theoretic model does not correct for borrower heterogeneity, wheres Deng et al. (2000) show significant heterogeneity among mortgagors exist, especially in prepayments. A shift was made to empirical models that could incorporate exogenous variables, such as a proportional hazard model or a (multinomial) logit model. A proportional hazard model models the survival probability or time to failure of a loan given a set of explanatory variables. Clapp et al. (2001) show that proportional hazard models have certain limitations such as their handling of competing risks and the proportionality assumption. They argue that a multinomial logit might be a more appropriate model to forecast prepayments because, e.g. it handles competing risks better, and show that indeed the multinomial leads to better results.

## 2.3 Machine learning

Two general approaches exist in machine learning modelling. The first approach is modelling one single model on the data, predicting a certain response variable. Another approach is combining several models into a so called *ensemble model*. An ensemble model combines predictions of multiple models into one overall prediction, e.g. by choosing the majority vote in classification problems or averaging predictions in regression problems. One course of action is to build multiple advanced models, but in practice a large number of simpler models is used. Examples of ensemble models are random forests (Breiman, 2001), gradient boosting models (Friedman, 2001) and neural network ensembles (Hansen & Salamon, 1990).

Within ensemble modelling two methods exist: bagging and boosting. Bagging entails independently running multiple models and combining the outputs into one prediction, whereas boosting iteratively trains a new model on the errors of the previous models. Weak learners are commonly used as predictors in boosting models. A weak learner is a model that performs slightly better than random guessing.

Using an ensemble method such as a neural network ensemble, prepayments can be predicted better than by using a logit model or any single neural network Riksen (2017). Sirignano et al. (2015) find in their study that all neural networks, including ensembles, that were investigated outperform a logit model. A neural network is a type of machine learning algorithm that is structured like a

brain to process information in the data. The gain in predictive power in both the neural network ensemble and the single neural networks is significant due to the more complex modelling of relations between features.

Another popular ensemble algorithm is gradient boosting. Gradient boosting is gaining in influence. Studies show that it performs well in several fields, e.g. in insurance loss cost modelling (Guelman, 2012) and in prediction of travelling time (Y. Zhang & Haghani, 2015).

### 2.4 Interpretable machine learning

As mentioned, the "black box" nature of the machine learning models can make financial institutions hesitant to use such models. Linear models, conversely, provide a clear interpretation of the model. However, many studies show that machine learning algorithms outperform simpler linear models, e.g. Sirignano et al. (2015); Riksen (2017); Guelman (2012). Thus, researchers who recognize the potential of machine learning have to consider whether the absence of interpretation is worth the increase in predictive power. In other words, choose between interpretability and accuracy. Fortunately, academic literature on model interpretability of machine learning models is growing and hence interpretability might not immediately require less accuracy.

Two interpretability approaches exist: model specific and model agnostic. Model specific methods are for example the regression weights of a linear model, or specific methods to interpret a neural network. Model agnostic methods are methods that can be used for any model and usually involve analyzing feature input and output pairs.

Within academic literature, two model agnostic methods are commonly used: LIME and SHAP. LIME (Ribeiro, Singh, & Guestrin, 2016) stands for Local Interpretable Model-agnostic Explanations. It uses a local linear model to interpret the influence of each feature to a specific instance of the data. SHAP stands for Shapley Additive exPlanations. Whereas LIME only gives local approximations, SHAP also provides globally consistent explanations. Consistency here means that the final prediction is fragmented into the attributions of each feature, and all the attributions thus sum to the final prediction.

## 3 Data

This section introduces the data used in this thesis. Section 3.1 introduces the Freddie Mac Single Family Loan-Level Data Set, whereas Section 3.2 specifies prepayment definitions. Section 3.3 notes

the data indiscrepancies and how this thesis amends those. Next, Section 3.4 elaborates on the data enrichment and data transformation. Finally, Section 3.5 displays summary statistics.

## 3.1 Freddie Mac Single Family Loan-Level Data Set

For this research, the Freddie Mac Single Family Loan-Level Data Set Sample (Freddie Mac, 2019) is used, which is available for download. Freddie Mac is an American financial services provider that buys mortgages on the secondary mortgage market, pools and subsequently sells these mortgages as mortgage backed securities to investors. The data set contains origination data and monthly performance data on 50,000 US fixed rate single family loan mortgages per year between 2000 and 2017, and 32,793 loans in 1999. This results in data on 932,793 US single family loan mortgages with a total of 44,835,243 monthly observations. The loan terms vary from ten to forty years. For computation purposes this thesis focuses on a sub-sample of 1000 random sampled loans per year, leading to a data set consisting of monthly observations of 19,000 loans with in total 891,492 monthly observations following the loans. This results in approximately 47 monthly observations per loan, whereas the original data set contains approximately 48 monthly observations per loan. The Freddie Mac data set contains information on the loans until one of three things happen: the loan matures, the loan is voluntarily prepaid in full, or it defaults. Information in the data is both loan specific information at loan origination, such as original unpaid principal balance (UPB), location of the property, property type and loan-to-value, as well as monthly information on loans, including loan age, UPB and months to maturity. Moreover, for this thesis, the Freddie Mac data set is enriched with macroeconomic factors to capture macroeconomic dependencies. Data enrichment is elaborated on in Section 3.4.1. Unfortunately the Freddie Mac data set does not contain personal data on mortgagors, such as age or race.

## 3.2 Prepayment types

The goal of this thesis is to model prepayment events and although the data from Freddie Mac is available, it is not yet labeled into prepayment types. In this study, two prepayment types are defined: *full prepayments* and *partial prepayments*. Additionally, a third class is defined as *no event*, if neither of the prepayment events occurs. Information that is available and where we can derive prepayments from is, e.g. the reason for termination of the loan. One of these reasons states that the loan is either prepaid in full or is matured. In this study, full prepayments are defined based on three conditions, 1) the outstanding balance is reduced to zero, 2) the reason for the zero balance

is not default and 3) the installment paid is higher than a threshold times the installment that is legally due. The installment is the contractual repayment of principal balance. If all conditions except the last condition hold, the loan simply matures. The installment is not given, but can be derived from available information in the Freddie Mac data set. It is derived as follows. Since the loans are all fixed rate annuity mortgages and the interest rate, loan term and months to maturity are known, the monthly annuity payment can be derived. When the interest due at month $t$ is subtracted from the expected contractual annuity payment, the resulting amount is the expected contractual installment. This is compared to the decrease in outstanding principal between months $t$ and $t-1$. By using the above three conditions for a full prepayment, the defaulting and maturing loans are filtered out. In this data set, however, there are almost no naturally maturing loans. This is due to the right censoring nature of prepayment data and due to most mortgages in the data being 30-year mortgages whereas the time frame of the Freddie Mac data is only 19 years.

Besides defaults and full prepayments, there are partial prepayments. These are not defined in the data, so this thesis defines partial prepayments as follows. For every loan, when the outstanding principal balance does not go to zero and hence there is no full prepayment,

$$y_t = \begin{cases} \text{Partial prepayment} & \text{if } -\Delta UPB_t > 2.3 * In_t \\ \text{No event} & \text{otherwise} \end{cases} \tag{1}$$

where $\Delta UPB_t$ is the difference in outstanding principal balance between $t$ and $t-1$, and $In_t$ is the expected installment of the loan at $t$. The installment at time $t$ is the to be prepaid amount of the loan at time $t$, or the difference between the annuity and interest due at time $t$. The difference in outstanding principal $\Delta UPB_t$ is expected to be negative, meaning a decrease of outstanding principal, for each month $t$ as each month a mortgagor is expected to pay off part of her debt. A threshold of 2.3 is used because the data contains irregularities and therefore a small decrease in UPB could be caused by these data irregularities instead of a partial prepayment. In some cases there is no payment in one month and a double payment in the consecutive month due to accounting reasons or late payments. Hence, to make sure such errors do not appear as a partial prepayment and to take other irregularities into account, 2.3 times the installment is taken as a threshold for partial prepayments.

Additionally, the data set contains defaults. A variety of academic literature exists on default modeling e.g. Foster and Van Order (1984); Boyes et al. (1989); Altman and Saunders (1997); Ghent and Kudlyak (2011); Crosbie and Bohn (2019). These studies dive deeper in default modelling and

since the scope of this thesis is prepayments and not defaults, defaulting loans are not taken into account for this thesis. The definition of defaulting loans that is used in this thesis is that defaulting loans are loans on property that is in foreclosure. In the Freddie Mac data set, the following reasons for loan termination that are listed are due to foreclosure of the property: third party sale of foreclosed property, charge off, repurchase prior to property disposition, REO disposition or re-performing loan sale. Furthermore, certain loans are modified because the mortgagee was unable to uphold her financial commitments. Modifying loans gives a mortgagor the opportunity to avoid default and hence risk of foregone interest payments is reduced. Also, these modified loans have risk drivers similar to defaulting loans and thus are deleted from the data. The other option is to treat modifying loans as a full prepayment, since new terms are specified and hence it can be seen as a new contract. However, modified loans are similar to defaults because if nothing would change, the mortgagor who qualifies for a modification will most likely default. Therefore, if modified loans are treated as a full prepayment then these loans add noise to the model because they have risk drivers that differ from true full prepayments. These loan modifications are only for loans that are prone to defaulting. Refinancing can also be seen as a modification, but enters the data set as a new mortgage. Hence, modified loans are not refinance loans.

The resulting classes for loan $i$ at time $t$ are

$$y_{it} = \begin{cases} \text{Full prepayment} \\ \text{Partial prepayment} \\ \text{No event} \end{cases} . \tag{2}$$

where the loan terminates if $y_{it}$ is a full prepayment or no event with outstanding principal reducing to zero. The frequency of each class in the Freddie Mac data set is presented in Figure 1. Clearly, there is unbalanced data with no event being the majority class and the next biggest class, partial prepayment, being five percent of the size of the no event class. Full prepayment is less than two percent of the no event class.

The full prepayment rate of the full data set over time is plotted in Figure 2. Full prepayments have a peak in the data set around 2004 and are high from 2002 to 2004. This is possibly due to the combination of two things. One, the fact that the prepayment rate is highest from the first year to the third year (see Appendix Table 10). This includes the fact that the portion of those prepaying loans in the total is higher than later in time, when there are many other loans in the data for a

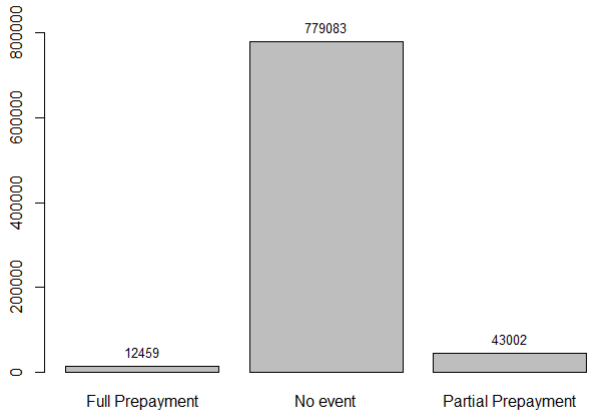



Figure 1: The frequency of prepayment types in the Freddie Mac data set are displayed.

specific time point. Next, it can be attributed to falling interest rates in that period. Also because of relatively low data frequency, the prepayment rate can be more volatile.

Additionally, the partial prepayment rate over time is plotted in Figure 3. Due to data quality reasons that are elaborated on in in Section 3.5, the partial prepayment rate for the first seven months is zero and consequently converges to the partial prepayment rate.

### 3.3 Issues

The Freddie Mac data set contains data discrepancies, of which one has to be amended before analysis. The most important aspect and the aspect that needs to be amended is that the current unpaid balance data is rounded to the nearest thousand in the first six months. This leads to, for these six months, a constant outstanding balance and then a drop of 1000 once the outstanding balance is rounded to the next thousand. Moreover, the outstanding principal is correctly specified starting from the seventh month, leading to a drop or an increase in outstanding balance from the nearest thousand. The drop of unpaid balance will then be labeled as a partial prepayment, but this is not the case. One option is to eliminate these observations from the data and to start from month eight. Information about full and partial prepayments is lost, but the issue is addressed. However, full prepayments at the beginning of the contract have a high foregone interest rate and hence being able to predict these instances does add value to the model. This is solved by labeling all full prepayments that happen as such and the rest as no event. Since there is no method to find out whether a partial prepayment occurred in the first seven months, the instances that are

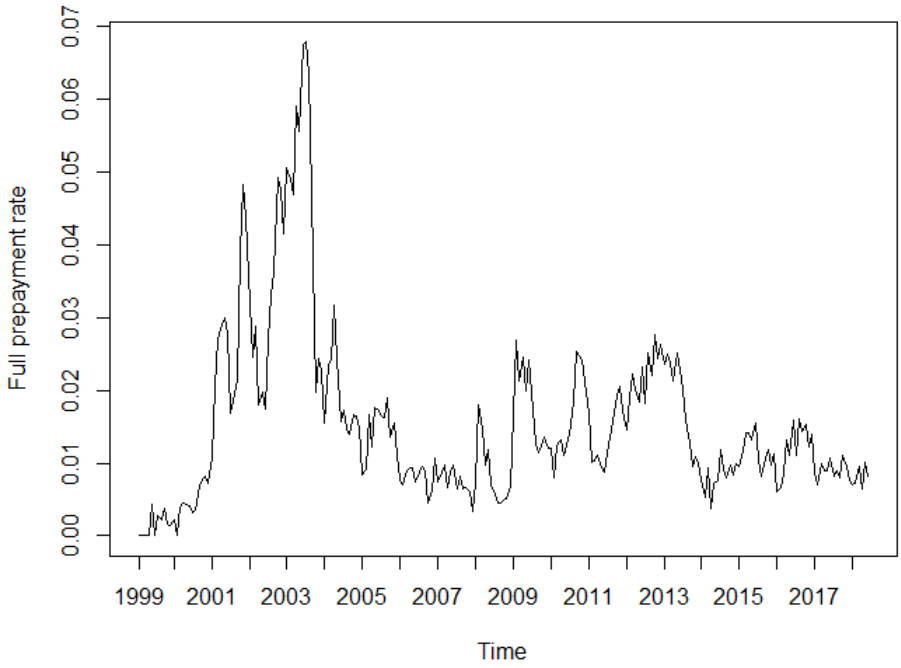



Figure 2: Full prepayment rates over time.

possible partial prepayments are also labeled as no event.

Additionally, the data contains information on long term, e.g. 30 year, mortgages that originated earliest in 1999. Since the period on which the model is trained lasts up until 2017, there are no fully maturing loans in the data. Furthermore, the data is right censored. This is a general issue in prepayment/default modelling since the available information on the loan ends at time of the event. The availability of more data in the lower loan ages leads to better predictions in these lower loan ages than in later loan ages.

In some cases the Freddie Mac data set has no information regarding a specific variable. This is indicated in the Freddie Mac data set by a missing information indicator, usually 9, 99, 999 or "". Data points containing these missing information indicators are excluded from the data set, with two exceptions. For the "first time home buyer flag", more than a quarter of the data points contain no information. This is taken into account as a separate class and expected to have minimal predictive power. The other exception is the reason for loan termination. Although this feature is not used in the model, it is used in the prepayment type labeling process. A minimal number of mortgages in the data have no termination reason in their termination equation. The other features of the termination observation of these mortgages indicate that these loans are prepaid voluntarily

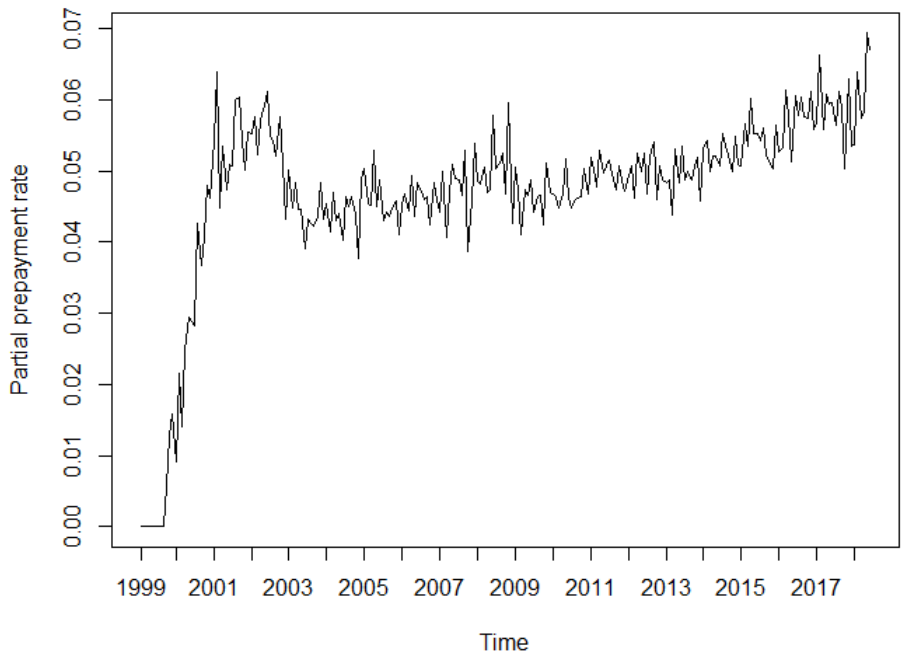



Figure 3: Partial prepayment rates over time.

and hence these instances are labeled as full prepayments.

## 3.4 Data enrichment & transformation

This section discusses how the Freddie Mac Single Family Loan Level data set is enriched and transformed into the final data set. Section 3.4.1 discusses how the data is enriched and Section 3.4.2 elaborates on how the data is transformed to make better use of the information in the data.

### 3.4.1 Enrichment

As discussed in the literature, macro variables drive the risk of prepayments. In this study, Freddie Mac data are enriched with macroeconomic variables. From the St. Louis Federal Reserve Economic Data (FRED) database the following data are obtained:

- The 30-year fixed rate mortgage average in the United States (St. Louis Federal Reserve Economic Data, 2019a). This is weekly data, but for the purpose of this research, it is converted to monthly data by taking the mean over weekly observations of each month.

- Quarterly US real GDP (St. Louis Federal Reserve Economic Data, 2019d). In this research, real GDP growth is calculated and taken as input instead of the level.
- Monthly US cross-country civilian unemployment rate (St. Louis Federal Reserve Economic Data, 2019b)
- Monthly US personal cross-country savings rate (St. Louis Federal Reserve Economic Data, 2019c). The savings rate is used because when the savings rate is high, this might indicate that a (partial) prepayment is less likely due to the preference of saving over spending.

Additionally, from the Quandl page of Freddie Mac the following data is obtained:

- Monthly house price index (HPI) for each US state (Freddie Mac, 2018). As mentioned, Clapp et al. (2001) find that a significant portion of heterogeneity in mortgage pools is due to house price dynamics of different regions. For this reason, the HPI per state is added. For the HPI of extraterritorial areas (Puerto Rico, Guam and US Virgin Islands), the average HPI of the US is used because it is not available in the Freddie Mac data on Quandl. The data is missing for the first half year of 2018, but as the HPI rose four percent in that period, a monthly increase of $\frac{4\%}{6} = 0.66\%$ is taken. To correctly model the incentive of housing sale, the percentage increase in HPI from loan origination is calculated for each month.

### 3.4.2 Transformation

Also, data transformations are performed in order to make best use of the information added. The newly created and transformed variables are:

- The delta mortgage rate, $dM$, is created by taking the difference between the current US mortgage rate and loan specific rate. The idea is to capture the lock-in and refinance effect, mentioned by e.g. Green and Shoven (1986).
- The sign of $dM$ is used as a categorical variable, to indicate an upward or downward movement. Also, no movement is an option.
- The six and twelve year moving averages of $dM$ are calculated. However, the moving averages experience a high degree of correlation. Hence, only the six month moving average is used. Additionally, the difference between the six and twelve month average is used as a new feature. Figure 9 in the Appendix shows that indeed there is no correlation between the difference between the six and twelve month average with the six month average.

- The six and twelve year sums of $dM$ are computed. Following the same rationale as the previous point, the variables used are the six month sum of $dM$ and the difference between the six and twelve month sums of $dM$. If the sum of the difference is close to zero and moving average is low, this indicates that the interest rate is low for a longer period. Also, when the six month difference of $dM$ as well as the difference between six and twelve month sum of $dM$ are negative, this indicates that the interest rate reduces for a longer time period. By using these two variables, this thesis tries to model the burnout effect. When interest rates are persistently lowering, the refinance incentive lowers and thus the burnout effect is expected to have a negative effect on the probability of a prepayment.

- For each month $t$ and loan $i$, UPB percentage is created, which is defined as current UPB divided by original UPB.

- An indicator $I_{i,t}$ is added that indicates whether a partial prepayment has already occurred in the history of loan $i$ at time $t$. The idea is that a mortgagor who has already partially prepaid, has a higher chance of another partial prepayment. This is called the partial prepayment flag.

- Since combined LTV (cLTV) and LTV are highly correlated, cLTV is replaced by the difference between cLTV and LTV to have only the relevant information, but no correlation. Figure 9 in the Appendix shows that indeed there is no correlation between LTV and cLTV.

- Using debt-to-income (DTI) and the monthly annuity, the income of a mortgagor is calculated, where income $= \frac{\text{annuity}}{\text{debt-to-income}}$. Here it is assumed that mortgage debt is the only source of debt. This is not generally true, as the sum of all declared debt payments at origination is used to calculate the DTI. This is done, however, to create a proxy of income. Subsequently, the log of the income is taken. This can be done because debt-to-income and the monthly annuity are always positive and hence debt divided by debt-to-income is also positive.

- Original UPB is divided by 1000 to avoid small $\beta$'s in the multinomial logit model.

The goal of this research is to predict prepayments one month ahead. Hence, all variables relevant variables are lagged one month.

### 3.5 Summary statistics

In Table 1 summary statistics of independent non-categorical variables of the model are displayed. Table 1 contains summary statistics on the continuous independent variables, such as the minimum,

| Feature | Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
|---|---|---|---|---|---|---|
| Delinquency #month | 0.00 | 0.00 | 0.00 | 0.09 | 0.00 | 125.00 |
| Loan age | 0.00 | 12.00 | 28.00 | 37.67 | 54.00 | 232.00 |
| Months remaining | 7.00 | 269.00 | 323.00 | 292.90 | 346.00 | 481.00 |
| Nin-int UPB | 0.00 | 0.00 | 0.00 | 95.05 | 0.00 | 96400.00 |
| FICO score | 300.00 | 698.00 | 745.00 | 735.70 | 780.00 | 832.00 |
| Insurance percentage | 0.00 | 0.00 | 0.00 | 4.67 | 0.00 | 40.00 |
| # Units | 1.00 | 1.00 | 1.00 | 1.03 | 1.00 | 4.00 |
| cLTV | 0.00 | 0.00 | 0.00 | 1.08 | 0.00 | 56.00 |
| Debt to income in % | 1.00 | 25.00 | 33.00 | 33.38 | 42.00 | 65.00 |
| Original UPB | 14.00 | 101.00 | 151.00 | 175.60 | 226.00 | 801.00 |
| LTV in % | 7.00 | 62.00 | 77.00 | 71.35 | 80.00 | 100.00 |
| Interest rate in % | 2.38 | 4.38 | 5.50 | 5.42 | 6.25 | 10.50 |
| Orig loan term | 120.00 | 360.00 | 360.00 | 328.60 | 360.00 | 480.00 |
| # Borrowers | 1.00 | 1.00 | 2.00 | 1.57 | 2.00 | 2.00 |
| $\Delta$ Mortgage rate | -5.82 | -0.32 | 0.02 | -0.03 | 0.30 | 4.36 |
| MR sum of diff 6m | -1.41 | -0.38 | -0.14 | -0.08 | 0.23 | 1.14 |
| MR SD 12-6 | -1.41 | -0.38 | -0.15 | -0.09 | 0.20 | 1.14 |
| MR MA 6m | 3.42 | 3.91 | 4.44 | 4.90 | 5.96 | 8.28 |
| MR MA 12-6 | -0.43 | -0.08 | 0.06 | 0.04 | 0.15 | 0.63 |
| House price index | 39.70 | 98.50 | 104.20 | 107.00 | 113.90 | 271.30 |
| rGDP growth in % | -2.16 | 0.24 | 0.55 | 0.47 | 0.80 | 1.83 |
| Unemployment rate | 3.80 | 4.70 | 5.60 | 6.23 | 7.80 | 10.00 |
| Savings rate | 2.20 | 5.40 | 6.60 | 6.22 | 7.20 | 12.00 |
| UPB percentage | 0.00 | 90.00 | 95.58 | 91.76 | 98.33 | 162.61 |
| Log of monthly income | -6.33 | 7.51 | 7.94 | 7.94 | 8.37 | 12.34 |

Table 1: Summary statistics of all continuous variables of the enriched cleaned Freddie Mac data set used to model prepayments. MR stands for mortgage rate. MR SD 6m and MR MA 6m stands for the six month sum of difference of mortgage rate and the six month moving average, respectively. MR SD 12-6 and MR MA 12-6 represent the difference between the twelve and six month sum difference and moving average, respectively.

maximum, mean and median. There are several variables that are highly skewed, such as number of months delinquent, number of units and insurance percentage. In Table 12 in the Appendix displays an overview of the number of observations in each class of the categorical variables. Table 10 and Table 11 in the Appendix give an explanation of all variables in the Freddie Mac data.

# 4 Methodology

This section discusses the models and methodology used in this thesis. In Section 4.1 the multinomial logit is explained. Moreover, in Section 4.2, gradient boosting is outlined. From Section 4.3 onward, several techniques enhancing model performance are introduced.

## 4.1 Multinomial logit

Following the findings of Clapp et al. (2001), the multinomial logit model is used in this thesis. The multinomial logit model is a generalization of the logistic regression to multiple classes, meaning for each monthly observation $i$ it predicts the prepayment probabilities $\pi_{ik}$ for each prepayment type $k$, where $\sum_{k=1}^{K} \pi_{ik} = 1$. It assumes the model to be time independent. Recall that in this research, three mutually exclusive prepayment classes are defined,

$$y_i = \begin{cases} 1 & \text{if full prepayment} \\ 2 & \text{if partial prepayment} \\ 3 & \text{if no event} \end{cases}, \tag{3}$$

where $y_i$ is the label of a monthly observation. The regression formula of any logistic regression involves regressing the log odds of class $k$ against a baseline class $K$, in this case the no prepayment event class, on the independent variables, so following loosely the notation of Greene (2002),

$$l_i = \log \frac{\mathbb{P}(Y_i = k)}{\mathbb{P}(Y_i = K)} = \beta_{0,k} + \beta_{1,k} x_{1,i} + ... + \beta_{v,k} x_{v,i} = \boldsymbol{\beta}_k' \mathbf{x}_i, \tag{4}$$

for $K - 1$ classes and $v$ explanatory variables, where $x_{v,i}$ is a dependent variable and $\beta_{v,k}$ is the effect of $x_{v,i}$ on class $k$. The vectors $\boldsymbol{\beta}_k$ and $\mathbf{x}_i$ have dimensions $v$ x 1. This can be rewritten for $\mathbb{P}(Y_i = k)$ into

$$\mathbb{P}(Y_i = k) = \mathbb{P}(Y_i = K) e^{\boldsymbol{\beta}_k \mathbf{x}_i}, \tag{5}$$

for $k \in \{1, ..., K - 1\}$. When rewriting for the probability of class $K$, the fact that all probabilities must sum to one is used and the probability of class K can be calculated by

$$\mathbb{P}(Y_i = K) = 1 - \sum_{k=1}^{K-1} \mathbb{P}(Y_i = k) = 1 - \sum_{k=1}^{K-1} \mathbb{P}(Y_i = K) e^{\boldsymbol{\beta}_k \mathbf{x}_i} \Rightarrow \mathbb{P}(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \mathbf{x}_i}}. \tag{6}$$

Equation 6 can be used in combination with Equation 5 to find probabilities of other classes,

$$\mathbb{P}(Y_i = k) = \mathbb{P}(Y_i = K)e^{\boldsymbol{\beta}_k \mathbf{x}_i} = \frac{e^{\boldsymbol{\beta}_k \mathbf{x}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \mathbf{x}_i}}. \tag{7}$$

This thesis uses LASSO regularization (Tibshirani, 1996) to reduce the number of coefficients in the multinomial logit model.

## 4.2 Gradient boosting

This thesis uses a gradient boosting algorithm (Friedman, 2001) to explore if prepayment behaviour can be modelled better than with the multinomial logit. Gradient boosting is successful in many applications, e.g. disease prediction (Bhatt et al., 2013) and face alignment (Xiong & De La Torre, 2013). In Section 4.2.1 the basis of gradient boosting is explained: gradient descent. Next in Section 4.2.2, the methodology called boosting is elaborated on. Then, in Section 4.2.3 the general gradient boosting algorithm by Friedman (2001) is explained. Finally, in Section 4.2.4 the XGBoost algorithm (Chen & Guestrin, 2016) that is used for this thesis is explained.

### 4.2.1 Gradient descent

The basics of gradient descent were introduced by Cauchy (1847). A gradient decent algorithm is an algorithm that minimizes a loss function. We have observation pair $z = (\mathbf{x}, y)$ and a function $F_{\mathbf{w}}(\mathbf{x})$, with parameters $\mathbf{w}$, mapping explanatory variables $\mathbf{x}$ to observation $y$. A differentiable loss function, $Q(z, w)$, is defined to model the performance of $F_{\mathbf{w}}(\mathbf{x})$. This loss function can be modelled as, but is not limited to, a least squares, absolute error or logistic loss function. Section 4.2.4 goes into more detail on the chosen loss function.

In gradient descent, the idea is to use the gradient or derivative of this loss function $\nabla Q_{\mathbf{w}}(z, \mathbf{w})$ with respect to the different function parameters or weights $w$, to find the values of $w$ that minimize this cost function. The starting values of $\mathbf{w}$, $\mathbf{w}_0$, need to be defined for initialisation. In the base case, called Batch Gradient Descent, the value of the gradient is calculated at each available observation pair $z_i$. Then, all gradients are averaged and consequently the new weight is calculated. This is done iteratively, such that,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma \frac{1}{n} \sum_{i=1}^{n} \nabla_{\mathbf{w}} Q(z_i, \mathbf{w}_t), \tag{8}$$

where $\gamma$ is the step function or learning rate and always positive, $\mathbf{w}_{t+1}$ are the optimal parameters found using the derivative of the loss function $\nabla_{\mathbf{w}} Q(z_i, \mathbf{w}_t)$ and $z_i$ is the observation pair $(\mathbf{x}_i, y_i)$

for observation $i$. It is shown that, when step function $\gamma$ is small enough, this iterative process converges to a (local) minimum. (Dennis & Schnabel, 1996).

### 4.2.2 Boosting

Boosting is an ensemble meta-algorithm that uses additive modelling to form a strong learner consisting of many weak learners. A strong learner is a model that is able perform significantly better than random guessing whereas a weak learner only performs slightly better than random better. It started with a question posed by Kearns and Valiant: *"Can a set of weak learners create a single strong learner?"* (Kearns & Valiant, 1989; Kearns, 1988). Schapire (1990) later shows that a weak learner can perform as well as a model with arbitrarily small errors using additive modelling. Rather than bagging, where additive models are formed simultaneously and every model gets an equal vote, boosting sequentially trains models on the errors of the previous model. This way, the emphasis lies on the iterative misclassifications of each model.

### 4.2.3 A gradient boosting machine

Gradient boosting, as introduced by Friedman in 1999, published by *The Annals of Statistics* in 2001 (Friedman, 2001) creates a link between gradient descent and boosting. Consider again observation pair $z_i = (\mathbf{x}_i, y_i)$ and a function $F(\mathbf{x})$. Furthermore, for each observation $i$ we introduce some loss function $L(y_i, F(\mathbf{x}_i))$. The goal is to use multiple weak learners $h(\mathbf{x}; \mathbf{a})$, where $\mathbf{a}$ represents the parameters of function $h(\mathbf{x})$, to form a strong learner $F(\mathbf{x})$. Initially, the best guess for a function $F_0(\mathbf{x})$ is the function $h(\mathbf{x}, \mathbf{a})$ with the parameters $\mathbf{a}_0$ that minimize its loss function,

$$\mathbf{a}_0 = \arg\min_{\mathbf{a}} \sum_{i=1}^{N} L(y_i, h(\mathbf{x}_i, \mathbf{a})). \tag{9}$$

Next, multiple models are added that learn on the errors of their predecessors, which is the boosting process. Gradient boosting differs from other boosting algorithms in its method of finding new weak learners to add to the model. Gradient boosting searches for the steepest-descent step in *function space*, and adds the weak learner that reduces the loss function the most. The most straightforward way of doing this is by taking the gradient of the loss function. However, since the loss function is only defined at data points $\{\mathbf{x}_i\}_1^N$, a gradient that generalizes over the entire feature space $\mathbf{x}$ does not exist. Instead, a weak learner is chosen that best approaches this gradient.

For each new model $m$ that trains iteratively on the errors of model $m-1$, the following is

performed. First, for each observation pair $z_i$, the negative gradient of the loss function with respect to the current model $F_{m-1}(\mathbf{x})$ is calculated. This can be an analytical expression but if that is not feasible, a numerical approximator is used. These gradients are called the *pseudo-residuals*

$$\tilde{y}_i = - \left[ \frac{\delta L(y_i, F(\mathbf{x}_i))}{\delta F(\mathbf{x}_i)} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, \tag{10}$$

which corresponds to the steepest-descent step in function space for that observation. Now, since this gradient can not be generalized, the weak learner $h(\mathbf{x}_i; \mathbf{a}_m)$ that best approaches $\tilde{y}_i$ is chosen as a best alternative. Using least squares as a loss measure between the weak learner and the pseudo-residuals, the solution or weak learner is found by

$$\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(\mathbf{x}_i; \mathbf{a})]^2, \tag{11}$$

where $\beta$ is used to scale the weak learner to the pseudo-residuals. Now that the weak learner that best approaches pseudo-residuals $\{\tilde{y}_i\}_1^N$ is found, it can be added to the current model $F_{m-1}(\mathbf{x})$ using a proper scaling parameter $\rho$. The proper scaling parameter $\rho_m$ to multiply the best weak learner $h(\mathbf{x}_i; \mathbf{a}_m)$ with, is found via a line search and provides

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i; \mathbf{a}_m)). \tag{12}$$

This is the best scaled weak learner to add to the model. Now an updated and hence stronger model can be made by adding that weak learner to the model,

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; \mathbf{a}_m). \tag{13}$$

This is done for a pre-specified number of iterations $M$, or until the results on a specific test set stop improving.

**Algorithm 1:** Gradient boosting

$F_0(\mathbf{x}) = h(\mathbf{x}, \mathbf{a_0})$, where $\mathbf{a}_0 = \arg\min_{\mathbf{a}} \sum_{i=1}^{N} L(y_i, h(\mathbf{x}_i, \mathbf{a}))$

**for** $m = 1$ *to* $M$ **do**

  $\tilde{y}_i = -\left[\frac{\delta L(y_i, F(\mathbf{x}_i))}{\delta F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}$

  $\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(\mathbf{x}_i; \mathbf{a})]^2$

  $\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i; \mathbf{a}_m))$

  $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; \mathbf{a}_m)$

**end**

### 4.2.4 XGBoost

The prepayments class probabilities are modelled using the XGBoost package in R, introduced by Chen and Guestrin (2016). In this section, the workings of the XGBoost algorithm for the 3 class classification problem are explained for the classes full prepayment, partial prepayment and no event. Note that the XGBoost algorithm performs this section automatically. Consider the general case of gradient boosting as explained in the previous section. For XGBoost, the weak learner $f_m(\mathbf{x})$ is a decision tree with $J$ terminal nodes, or leafs. The tree $f_m(\mathbf{x})$ will be denoted as $f_m$ unless the dependence on $x$ is explicit, e.g. a sum over $f_t(\mathbf{x_i})$ for different observations $i$. The loss function used in XGBoost is a combination of the logarithmic (or cross-entropy) loss function and a penalty term $\Omega$. The logarithmic loss function calculates the error of the predicted prepayment probabilities $\hat{y}_i$, e.g. (0.5, 0.2, 0.3), using the actual prepayment class $y_i$, e.g. (1, 0, 0). The penalty term $\Omega$ protects against over-fitting the decision trees on the data by penalizing for complexity. Because of this penalty term for complexity, this loss function is also called the *regularized* loss function,

$$L = \sum_i \ell(y_i, \hat{y}_i) + \sum_m \Omega(f_m) \tag{14}$$

where: $\Omega(f_m) = \gamma J_m + \frac{1}{2}\lambda \|w_m\|^2$

$\ell(y_i, \hat{y}_i) = \sum_i y_i' \hat{y}_i$

$= \sum_i \sum_k y_{ik} \log(\hat{y}_{ik})$

Here, $m \in M$ are the iterations of the XGBoost model, $\hat{y}_{ik}$ is the predicted probability of prepayment class $k$, $y_{ik}$ is 1 if the observation is of class $k$, $J_m$ is the number of terminal nodes (leafs) of the tree of model $m$ and $\mathbf{w}_m$ are the prepayment probabilities that correspond to all leafs $j$ of tree $f_m$.

Additionally, $\lambda$ and $\gamma$ are regularisation parameters and determined by the model. Note that the loss function $\ell$ is differentiable. Additionally, note that the regularized loss function $L$ increases when a new tree is added to the model. Hence, the regularized loss function $L$ only reduces if the added model complexity of adding another tree does not exceed the added value of the new tree to the model and hence prepayments are modelled better. Let $\hat{y}_i^{(t-1)}$ be the predicted prepayment class probabilities of observation $i$ at iteration $t-1$ and let $f_t$ be the tree that is added at iteration $t$. Note that here $t$ is used to distinguish between all iterations $m$ and a specific iteration $t$. The objective is to add the tree $f_t$ that minimizes the loss function, formally

$$\min_{f_t} L^{(t)} = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t). \tag{15}$$

In order to find the best split in their trees, vanilla gradient boosting machines calculate every tree from scratch. XGBoost works well because rather than vanilla gradient boosting machines, it uses the second order approximation to optimize Equation 15 (Chen & Guestrin, 2016),

$$L^{(t)} \simeq \sum_{i=1}^{N} [\ell(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t), \tag{16}$$

where $g_i = \frac{\delta}{\delta \hat{y}_i^{(t-1)}} \ell(y_i, \hat{y}_i^{(t-1)})$ is the first order derivative of loss function $\ell$ and $h_i = \frac{\delta^2}{\delta^2 \hat{y}_i^{(t-1)}}{}^2 \ell(y_i, \hat{y}_i^{(t-1)})$ is the second order derivative of loss function $\ell$ at every monthly prepayment observation $i$, both with respect to the predicted prepayment probabilities of the previous iteration $\hat{y}_i^{(t-1)}$. The objective of each iteration $t$ is to find decision tree $f_t$ that minimizes this loss function. Hence, the terms that do not depend on this new tree $f_t$ can be removed from the objective function. What remains is

$$\tilde{L}^{(t)} = \sum_{i=1}^{N} [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t), \tag{17}$$

which is the simplified loss function that depends on the first and second order derivatives, $g_i$ and $h_i$, of the loss function $\ell$ with respect to the predicted prepayment probabilities $\hat{y}_i^{(t-1)}$ at every observation $i$, the new tree $f_t$ and the complexity of this new tree $\Omega(f_t)$.

Now that the objective function is defined, the second procedure that is of interest is to find the split points that split the tree into leafs and branches. Define $I_j = \{i | f_t(\mathbf{x}_i) = j\}$ as the set of observations $i$ in leaf or terminal node $j$ of the tree $t$. Equation 17 can be rewritten, by expanding

$\Omega$ and summing over the instances of each leaf $j$, $i \in I_j$, where $j$ are the possible leafs, as

$$\begin{aligned}\tilde{L}^{(t)} &= \sum_{i=1}^{N}[g_i f_t(\mathbf{x}_i) + \frac{1}{2}h_i f_t^2(\mathbf{x}_i)] + \gamma J_t + \frac{1}{2}\lambda \sum_{j=1}^{J}{}_t w_j^2 \\ &= \sum_{j=1}^{J}{}_t[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2] + \gamma J_t,\end{aligned} \tag{18}$$

where the second part follows from separating all observations $i$ in observation sets $I_j$ and that the value for $f_t(\mathbf{x}_i)$ of observation $i$ in observation set $I_j$ is $w_j$. Recall that the variable factors here are the number of leafs $J_t$ for the model $f_t$ that is proposed to reduce the loss function, the instances $I_j$ and the corresponding values of the leafs $w_j$. The derivatives $g_i$ and $h_i$ are constant. They depend on the loss function of the previous prediction and hence do not depend on the proposed tree $f_t$. The values of $g_i$ and $h_i$ for every observation $i$ can thus be calculated before proposing the new tree and can be filled in for every proposed tree $f_t$ in Equation 18. Given a tree structure $f_t$ with known observation sets $I_j$ in leaf $j$, the optimal value for weight $w_j^*$ in leaf $j$ which are the optimal prepayment probabilities, can be found by solving the quadratic equation in Equation 18 for $w_j$. This gives

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}, \tag{19}$$

which can then be inserted in Equation 18, resulting in an minimum value for the loss function (an optimal $w^*$ means the smallest loss) for tree structure $f_t$ of

$$\tilde{L}^{(t)}(f_t) = \frac{1}{2}\sum_{j=1}^{J_t}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma J_t. \tag{20}$$

where $J_t$ are the number of leafs of model $f_t$. Equation 20 can be used to assess the loss reduction of a particular split. Say that a split breaks an observation set $I$ into two observation sets, namely $I_L$ and $I_R$ where $I = I_L \cup I_R$. The loss reduction of this split can be easily calculated using Equation 20 by

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] + \gamma, \tag{21}$$

which is simply a summation of constant terms, meaning this is a less computationally heavy method of calculating the loss than vanilla gradient boosting. The constant terms are split based

on feature distributions and consequently the features are placed into (new) observation sets, $I_L$ and $I_R$. Consequently, for both sets, the algorithm tries to find a further loss reduction by splitting $I_R$ and $I_L$ as well.

### 4.3 Overfitting

Machine learning algorithms such as XGBoost capture non-linear relations between variables in the data. However, it is possible that such algorithms capture noise and spurious relations in the data that do not represent any real relationship and are only in the data by chance. This way the error of the model can seem low, whereas if the model is tested on out-of-sample data, it would have a significantly higher error. This is known as *overfitting*. To solve this issue, the unbalanced data are split and twenty percent of the data are randomly taken as a *hold-out* or *test* set for out-of-sample testing, totaling 166,909 data points. The hold-out set is used to test model performance of both the XGBoost and multinomial logit models. The hold-out set is not used in training or tuning the model, hence the model has not seen the hold-out set before evaluating.

### 4.4 Unbalanced data

The very nature of prepayments creates an unbalanced data set. When tackling machine learning problems, unbalanced data are a problem. Unbalanced data, or unbalanced classes, create a challenge in measuring model accuracy. If the model is trained on an unbalanced data set, the model tends to predict the majority class. Why this is the case, is simple. Take as an example prepayment data. Assume only one percent of the data are a prepayment event. Only predicting "no prepayment event" would result in a model accuracy of 99%. However, no instances of the prepayment class would be predicted correctly. This would greatly miss the purpose of the model, which is predicting those prepayment events.

Multiple solutions are available to counter the unbalanced data problem. One can choose to give weights to the observations in the XGBoost algorithm, use oversampling (increase minority class) or undersampling (decrease majority class). Due to ample data, this thesis uses undersampling of the number of no event class instances into the combined number of instances of the full prepayment and partial prepayment classes to solve the unbalancedness issue.[2] After balancing, the training data set contains 87,326 data points. This is not all, however. The next section outlines how this

[2]As mentioned, an alternative approach might be to give weights to all observations, where the minority class instances receive more weight than the majority class instances. For robustness, this approach was also tried but resulted in poorer performance.

thesis solves evaluation issues due to unbalanced data. Note that in this hold-out set the ratio of prepayment events and no-prepayment-event data is the same as for the entire data set, meaning the hold-out set is also an unbalanced set. The hold-out set is not balanced because prepayments events are not balanced and that is what this thesis tries to predict. The training sample is balanced because then the model can better capture the relationships between variables.

## 4.5 Evaluation metrics

Balancing the training data ensures that the model does not favor the majority class of no event over the minority classes full and partial prepayment. However, an issue that still exists is finding a suitable evaluation metric. As mentioned, pure accuracy might tell a misleading story. The probabilistic classification equivalent of accuracy is the log loss, which provides a measure of fit but does not provide an accurate representation of model performance.

### 4.5.1 Precision, recall and F1 score

Other evaluation metrics include precision and recall. Precision of class $C$ is the fraction of correctly classified instances of class $C$ over the total instances classified as class $C$. On the other hand, the recall of class $C$ is the fraction of the correctly classified instances of class $C$ over the total instances of class $C$. Usually to increase precision, recall is reduced. This makes sense because to increase the precision, one only takes the observations of which one is very certain and hence the number of predictions of the class of interest goes down. A researcher has to make a choice as to whether she prefers precision over recall, or if she wants a balancing function of the two. An option is to balance precision and recall is the F1 score. The F1 score is the harmonic average of precision and recall. It is used extensively in literature, e.g. Fujino, Isozaki, and Suzuki (2008); Sepúlveda and Velastin (2015); D. Zhang, Wang, Zhao, and Wang (2016). Since a high recall and a high precision are both signs of a well performing model, a high F1 score also means a well performing model. The F1 score is always between 0 and 1. The goal of this thesis is to correctly predict prepayments and in doing so in this case there is no clear preference for precision or recall, hence the goal is to maximize both. Therefore, this thesis uses the F1 score as one of the evaluation metrics. Additionally, sensitivity and specificity are given for both prepayment models. Sensitivity is the same as recall, it's the true positive rate. If it is high, there are not many false negatives. Specificity is the true negative rate, hence if it is high, there are not many false positives.

### 4.5.2 Confusion matrix

Precision, recall and specificity can be determined from a confusion matrix. A confusion matrix gives a global representation of the actual prepayment classes against the predicted prepayment classes. It is called a confusion matrix because it indicates where the model has confused one class for another. For every predicted class, it shows how many observations are actually from that prepayment class and how many observations are from other prepayment classes.

### 4.5.3 Brier score

Furthermore, the Brier score (Brier, 1950) is used as an alternative to accuracy to measure the ability of the model to capture prepayment behaviour. In the literature, the Brier score is used extensively, e.g. Gerds and Schumacher (2006); Rufibach (2010). The formula to calculate the multi-class Brier score is

$$BS = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}(\hat{y}_{i,k} - y_{i,k})^2, \tag{22}$$

where $N$ is the number of instances that is used to evaluate the model and $C$ is the number of classes in the model, in this case three: full prepayment, partial prepayment and no event. The prepayment probability of instance $i$ for class $k$ is represented by $\hat{y}_{i,k}$. Similarly, the actual prepayment event of instance $i$ for class $k$ is represented by $y_{i,k}$, which is 1 if the actual prepayment event is $k$ and 0 otherwise. The Brier score can be seen as a probabilistic mean squared error.

The reason for choosing the Brier score as an evaluation metric instead of the widely used logloss is twofold. The XGBoost model is optimized by using the logloss or cross entropy error as a loss function. By assessing both models with an evaluation metric that is not used in optimizing, a fair comparison is made. Additionally, the Brier score provides a smoother loss function than logloss, penalizing a higher error less severely. According to Wilks (2010), using the Brier score for assessing models forecasting unbalanced events is sound if the number of observations exceed one thousand. Using this insight, it is suitable to use the Brier score as an evaluation metric for this thesis.

The Brier score itself provides the average error of a model, but in order to compare models it is useful to compute the Brier skill score. The Brier skill score is calculated by

$$BSS = 1 - \frac{BS_m}{BS_{ref}}, \tag{23}$$

where $BS_{ref}$ is the Brier score of a reference model, in this case the multinomial logit, and $BS_m$ is the Brier score of the model that is compared to the reference model, in this case that of the XGBoost. Since the Brier Score can be seen as an error measure, a Brier Skill Score of 1 means a very good model against a bad reference model. A score of 0 indicates a similar performance between models and a negative score states that the reference model, the multinomial logit, performs better.

### 4.6 Hyperparameter tuning

The XGBoost algorithm has a variety of hyperparameters. Performance can increase significantly with the right set of parameters, hence an important aspect of modelling with XGBoost is tuning the hyperparameters. The parameters that are tuned in this model are the *learning rate*, *maximum tree depth*, *gamma*, *sub-sample percentage*, *column sample percentage* and *minimum child weight*.

- The learning rate, or step length, is $\rho_m$ from Equation 12 and Equation 13. The learning rate shrinks the weights predicted by each tree. In XGBoost this is a fixed value in order to avoid overfitting and thus make the model more conservative.

- Maximum tree depth indicates the maximum amount of edges between the root node of a decision tree and its nodes. Increasing the maximum tree depth can lead to more complex models and to overfitting.

- Gamma states the minimum loss reduction needed in order for a tree to make a new split. If the value for gamma is higher, trees become more shallow.

- Sub sample percentage specifies the percentage of training instances randomly sampled every time a new tree is made. If set to 0.5 then for the creation of each tree 50% of the data points are randomly selected to be used to train the tree.

- Column sample percentage does the same as sub sample percentage, but then in the other dimension of the data set. It sets the percentage of variables used to create each new tree. If set to 0.5, then for the creation of each tree 50% of the variables are randomly selected to be used to create the tree.

- Minimum child weight gives the minimum number of instances needed in each leaf. If this value increases, it makes the model more conservative.

For this thesis, hyperparameter grid search is performed. This entails setting up a grid of many possible hyperparameters and creating a model with all pre-specified hyperparameters using cross

validation. For the learning rate, the proposed settings were 0.01, 0.05, 0.1, 0.15 and 0.2. Maximum tree depth was found by using a grid of 5 to 25 with step size 2. For Gamma, the proposed settings were 0 to 10, with step size 1. For both column sample, sub sample and the proposed values are 0.5 to to 1 with step size 0.1. Finally, for minimum child weight the proposed settings are 1, 3, 5 and 7. All combinations are examined by using cross validation and the final hyperparameters are shown in the results section.

### 4.7 K-fold cross validation

K-fold cross validation is applied to tune the hyperparameters of the model in order to avoid tuning the parameters to noise. K-fold cross validation involves splitting the data in $k$ folds. One fold is left out and the other folds are grouped and balanced. It is important that all data transformation and balancing is done after splitting, because the validation set should resemble the real world. Consequently, the model is fitted on the data and tested on the fold that is left out by using a suitable evaluation metric. This is an out-of-sample test, meaning the data on which the model predicts is not used to train the model. It is done $k$ times, until each fold is left out, and thus tested on, once. After the process, the $k$ model evaluation metrics are averaged, resulting in one evaluation metric for the out-of-sample test for that model. This thesis applies 5 fold cross validation to find the best performing hyperparameters.

### 4.8 SHAP

For this thesis, SHapley Additive exPlanations (Lundberg & Lee, 2017), also called SHAP values, are used as a model interpreter. Using SHAP values as model interpreters is a model agnostic method of explaining a model, i.e. SHAP values can be computed for many different models. SHAP values provide an understanding in the relationships between the features and the outcome of the model. The SHAP methodology uses Shapley values (Shapley, 1953) from game theory. Shapley values give the value of a player $i$ in a game by evaluating the marginal increase of the value of a game when adding player $i$ to all possible coalition sets of players. Similarly, SHAP calculates the added value of a feature as the weighted increase or decrease in the value of a model outcome when adding a feature $\{z\}$ over all subsets of features that exclude that feature, namely $S \subseteq F \setminus \{z\}$. The SHAP values for features in a model indicate the attribution of those features to the prediction outcome.

The SHAP value for a feature $z$ and model $f$ is calculated by

$$\phi_z(f) = \sum_{S \subseteq F \setminus \{z\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left( f_{S \cup \{z\}}(x_{S \cup \{z\}}) - f_S(x_S) \right), \tag{24}$$

where $F$ is the total set of features for model $f$, $S \subseteq F \setminus \{z\}$ are all possible subsets of $F$ excluding feature $z$, $f_S(x_S)$ is the function trained on features of subset S and the same holds for function $f_{S \cup \{z\}}(x_{S \cup \{z\}})$ on the set $S \cup \{z\}$. The exclamation mark ! stands for factorial. These SHAP values can be different for each observation and these different SHAP values can be aggregated in a graph, showing the SHAP value for each available value of feature $z$. Figure 4 shows the perspective of SHAP with the data and a model. It is used to interpret the XGBoost model.

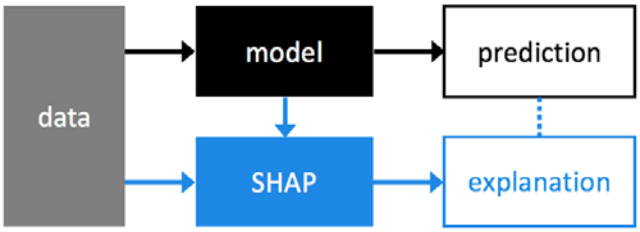



Figure 4: SHapley Additive exPlanations.

For a model interpreter to be useful, it must contain the following three properties:

- Local accuracy: local accuracy states that output of the model interpreter must be the same as the output of the model. The explanation model has to be a truthfull explanation. Hence, the SHAP value of each feature of an observation must sum to the model output of that observation.

- Missingness: missingness states that if a specific feature is not included in the observation, i.e. missing, its SHAP value is zero. The feature then does not contribute to the model. This is different from a feature being zero, which does provide information to the model.

- Consistency: consistency states that if the model changes and the contribution of a specific feature to the model increases or stays the same, its SHAP value should not decrease.

According to Lundberg and Lee (2017), SHAP is the only model interpreter that has the desirable properties local accuracy, missingness and consistency.

Due to the three class prepayment problem, the SHAP methodology provides three global graphs per feature, one for each class with the relationship of the class with that feature. It is useful to have three graphs because a regulator can also be interested in why you did not make a prediction,

compared to why you did make a prediction. However, this provides a challenge in comparing the SHAP output with the log odds from the multinomial logit. Hence, a Poisson regression is performed for all prepayment classes to find the relationship between a specific feature with the log of the probability of each class. For all three classes $k$,

$$\log \mathbb{P}(Y_i = k) = \beta_{0,k} + \beta_{1,k}x_{1,i} + ... + \beta_{v,k}x_{v,i} = \boldsymbol{\beta}_k' \mathbf{x}_i, \tag{25}$$

where $\mathbb{P}(Y_i = k)$ is the probability of class $k$, $\beta_{v,k}$ the coefficient for variable $v$ and class $k$ and $x_{v,i}$ is variable $v$ for observation $i$. The Poisson regressions form the basis of the multinomial logit, as $\log \frac{\mathbb{P}(Y_i=k)}{\mathbb{P}(Y_i=K)} = \log \mathbb{P}(Y_i = k) - \log \mathbb{P}(Y_i = K)$.

A limitation of SHAP for XGBoost in a multinomial setting is that it does not provide contributions to the probabilistic outcomes of the model, but rather to the log odds of that class. The log odds can, naturally, be transformed into probabilities once the log odds for other classes are also known. The log odds of other classes, however, do not follow directly from the SHAP values of one class. Hence, when commenting on the SHAP contributions of the features on the model, this thesis focuses on the direction of the SHAP values, rather than the exact contributions the of the log odds. The transformation of log odds to probability is monotone, hence a higher log odds of a full prepayment indicates a higher probability of a prepayment and this is of interest. After all, the direction of the contributions is what counts on a global level to see whether, e.g., a higher income gives a higher probability of a full prepayment.

# 5 Results

This section outlines the empirical results of this study. In Section 5.1, the results from the XGBoost model are discussed. In Section 5.2, a comparison is made between the cross sectional performance of the XGBoost algorithm and the multinomial logit model. Additionally, in Section 5.3 the multinomial logit model and the XGBoost model are compared when determining the actual prepayment rate over time. Finally, in Section 5.4 the SHAP values are shown, discussed and compared to the coefficients of the Poisson regression.

## 5.1 The XGBoost model

### 5.1.1 Hyperparameter tuning

Table 2 displays the hyperparameters that contribute to the highest F1 score. These hyperparameters are found by using a grid search. The performance gain of using these parameters is significant, with the F1 score of this set of parameters being 60% higher than that of the worst performing parameter set.

| Hyperparameter | Value |
|---|---|
| Learning rate | 0.15 |
| Max tree depth | 13 |
| Gamma | 5 |
| Sub sample | 0.8 |
| Column sample | 0.8 |
| Minimum child weight | 1 |

Table 2: Hyperparameters used to fit the final XGBoost model.

### 5.1.2 Initial results

Table 3 shows the results for XGBoost on the prepayment data set. When looking at precision and recall of the XGBoost model, no event has the highest values compared to the other classes, although the recall of partial prepayment and no event are similar. For this model, the F1 score of full prepayment is 0.097 and the F1 score of partial prepayment is 0.514. Recall the definition of a confusion matrix from Section 4.5.2. Table 4 displays the confusion matrix of the XGBoost model. The XGBoost model overestimates the number of full prepayments in the test set: a total of 3295 predicted full prepayments compared to the 2437 original full prepayments. Of those 3295 full prepayment predictions, 278 are actually a full prepayment observation, giving a precision of 0.084 or an 8.4% precision rate. Additionally, of the 2437 original full prepayment observations, 278 are correct. This gives a recall rate of 11.4%.

### 5.1.3 Distribution specification

From Section 5.1.2, the XGBoost model overpredicts both partial and full prepayments. This gives rise to the idea that shrinking the forecast probabilities of those classes might prove beneficial to the outcome of the model. Hence, using cross validation, the best prior probability distribution to

| Class | No Event | Full Prepayment | Part Prepayment |
|---|---|---|---|
| Sensitivity | 0.897 | 0.114 | 0.876 |
| Specificity | 0.745 | 0.981 | 0.915 |
| Precision | 0.980 | 0.084 | 0.363 |
| Recall | 0.897 | 0.114 | 0.876 |
| F1 | 0.936 | 0.097 | 0.514 |
| Balanced acc. | 0.821 | 0.548 | 0.896 |
| Brier score | | 0.200 | |

Table 3: Evaluation metrics of the XGBoost model.

| | | **Actual class** | | |
|---|---|---|---|---|
| | | No Event | Full Prepayment | Part Prepayment |
| **Predicted class** | No Event | 135447 | 1767 | 1014 |
| | | (89.7%) | (72.5%) | (12.0%) |
| | Full Prepayment | 2978 | 278 | 39 |
| | | (02.0%) | (11.4%) | (00.4%) |
| | Partial Prepayment | 12625 | 392 | 7426 |
| | | (08.3%) | (16.1%) | (87.6%) |

Table 4: Confusion matrix of the XGBoost model on the hold out set. For each class the predicted and actual values are shown. In brackets, the percentages of the number of observations of the actual class is displayed. The column percentages sum to one.

multiply with the probabilistic outcomes of the model is found. The goal of using this probability distribution is similar to using a threshold in binary classification; to have a higher probabilistic threshold for selecting a certain prepayment class. In order to use evaluation criteria such as the Brier Skill Score, the prepayment class probabilities resulting from combining the model and the distribution are normalized, meaning they are multiplied by a constant so that they sum to 1. This is done because the Brier score is evaluated based on predicted probabilities of each class.

Table 5 shows the probability distributions and the corresponding average F1 scores of all classes. The average F1 score is used because, using 5-fold cross validation, there are 5 F1 scores per class per distribution. More variations than shown in Table 5 are also tested but these variations give lower average F1 scores. The standard deviation of the F1 scores of the 5 cross validation folds is displayed in brackets next to the average score. In the left column the distribution for respectively no event, full prepayment and partial prepayment is given, with on the right for all three classes their average F1 scores and standard deviations. For this research, the distribution $(0.45, 0.48, 0.07)$ is chosen as probability distribution to multiply with the probabilistic outcomes of the XGBoost model, due to

the highest overall F1 scores. The choice for $(0.45, 0.48, 0.07)$ rather than $(0.48, 0.45, 0.07)$, which has the same full prepayment and partial prepayment F1 score, is made because it has the highest F1 score for the no event class. Although the F1 score of full prepayment and no event is higher for $(0.45, 0.51, 0.04)$, $(0.46, 0.49, 0.04)$ and $(0.43, 0.48, 0.09)$ than for $(0.45, 0.48, 0.07)$, with a lower standard deviation for the F1 score of full prepayment for $(0.45, 0.51, 0.04)$, the F1 score for partial prepayment for $(0.45, 0.48, 0.07)$ is significantly higher and hence $(0.45, 0.48, 0.07)$ is selected as a distribution. The biggest advantage of having a distribution such as the one that is chosen, is the increase of the F1 score of partial prepayments. The results of adding the distribution to the results of the XGBoost model are displayed in Table 6. When comparing the results to the original XGBoost model, the F1 scores for all classes are improved. The F1 score of partial prepayment has increased significantly, whereas the F1 scores of Full Prepayment and No Event only increased marginally.

| Distribution | NE | NE sd | FP | FP sd | PP | PP sd |
|---|---|---|---|---|---|---|
| 0.49, 0.46, 0.04 | 0.967 | (0.001) | 0.098 | (0.002) | 0.654 | (0.006) |
| 0.48, 0.48, 0.04 | 0.966 | (0.001) | 0.099 | (0.003) | 0.653 | (0.006) |
| 0.46, 0.49, 0.04 | 0.965 | (0.001) | 0.101 | (0.004) | 0.651 | (0.006) |
| 0.45, 0.51, 0.04 | 0.964 | (0.001) | 0.101 | (0.003) | 0.649 | (0.006) |
| | | | | | | |
| 0.48, 0.45, 0.07 | 0.966 | (0.001) | 0.098 | (0.003) | 0.667 | (0.003) |
| 0.46, 0.46, 0.07 | 0.965 | (0.001) | 0.099 | (0.004) | 0.667 | (0.003) |
| **0.45, 0.48, 0.07** | **0.963** | **(0.001)** | **0.100** | **(0.004)** | **0.667** | **(0.003)** |
| 0.44, 0.49, 0.07 | 0.962 | (0.001) | 0.100 | (0.004) | 0.667 | (0.003) |
| | | | | | | |
| 0.46, 0.44, 0.1 | 0.963 | (0.001) | 0.096 | (0.004) | 0.652 | (0.003) |
| 0.45, 0.45, 0.1 | 0.962 | (0.001) | 0.098 | (0.004) | 0.652 | (0.003) |
| 0.44, 0.47, 0.09 | 0.961 | (0.001) | 0.100 | (0.005) | 0.652 | (0.003) |
| 0.43, 0.48, 0.09 | 0.960 | (0.001) | 0.101 | (0.004) | 0.652 | (0.003) |
| | | | | | | |
| 0.45, 0.42, 0.12 | 0.961 | (0.001) | 0.095 | (0.004) | 0.635 | (0.003) |
| 0.44, 0.44, 0.12 | 0.960 | (0.001) | 0.097 | (0.005) | 0.635 | (0.003) |
| 0.43, 0.45, 0.12 | 0.958 | (0.001) | 0.098 | (0.005) | 0.635 | (0.003) |
| 0.42, 0.47, 0.11 | 0.957 | (0.001) | 0.100 | (0.004) | 0.635 | (0.003) |
| | | | | | | |
| 0.33, 0.33, 0.33 | 0.936 | (0.001) | 0.094 | (0.006) | 0.508 | (0.004) |

Table 5: This Table shows the different distributions that are multiplied with the probabilistic output of the XGBoost model with the evaluation scores of the corresponding posterior distribution. The distributions for respectively no event (NE), full prepayment (FP) and partial prepayment (PP) are displayed on the left with their corresponding average F1 scores and standard deviations (sd, in brackets) on the right. The distribution at the bottom represents the uniform distribution and the distribution used for this research is displayed in bold.

## 5.2 Model comparison

In Table 6 the evaluation metrics of both the XGBoost prepayment model and the multinomial logit prepayment model can be found. The regularization parameter, $\lambda$, that is used for the multinomial logit model is $1.403 \times 10^{-3}$. All lambdas and corresponding percentage of explained deviance can be found in Figure 11 and Table 19 in the Appendix. Figure 11 shows that the explained deviance of the multinomial logit model reduces exponentially when the lambda is higher than the chosen lambda. The coefficients of both the Poisson regression to form the multinomial logit and the coefficients of the multinomial logit can be found in respectively Tables 13, 14 and 15, and Tables 16, 17 and 18 in the Appendix.

| | XGBoost | | | Multinomial Logit | | |
|---|---|---|---|---|---|---|
| Class | No Event | Full Ppmt | Part Ppmt | No Event | Full Ppmt | Part Ppmt |
| Sensitivity | 0.958 | 0.146 | 0.674 | 0.869 | 0.007 | 0.879 |
| Specificity | 0.591 | 0.975 | 0.982 | 0.734 | 0.998 | 0.87 |
| Precision | 0.970 | 0.082 | 0.674 | 0.978 | 0.040 | 0.271 |
| Recall | 0.958 | 0.146 | 0.674 | 0.869 | 0.007 | 0.879 |
| F1 | 0.964 | 0.105 | 0.674 | 0.920 | 0.011 | 0.415 |
| Bal. Acc. | 0.774 | 0.560 | 0.828 | 0.802 | 0.502 | 0.874 |
| Brier score | | 0.142 | | | 0.234 | |

Table 6: Evaluation metrics of both the XGBoost model multiplied with the probability distribution and the multinomial logit model. Additionally the Brier score is given. Ppmt stands for prepayment.

Similar to the XGBoost model, in the multinomial logit model no event also has the highest values between classes for precision, but the second highest score for recall. No event does, however, achieve a significantly higher F1 score than other classes. The F1 score of Full Prepayment is 0.011 and the F1-score of Partial Prepayment is 0.415. Table 8 shows the confusion matrix of the multinomial logit model and Table 7 shows the confusion matrix of the XGBoost model with distribution. Contrary to the XGBoost model, the multinomial logit the model underestimates the amount of full prepayments, a total of 427 predicted compared to the 2492 original full prepayments.

When comparing both models, the XGBoost model is outperforming the multinomial logit on almost all evaluation metrics that are of interest. The F1 scores of Full Prepayment en Partial Prepayment are higher for the XGBoost model than that of the multinomial logit model. The Brier score of the XGBoost model is lower than that of the multinomial logit model, meaning its error is lower. This is reflected in the Brier Skill Score, which is 0.397.

The goal of this research is to investigate whether a gradient boosting machine such as XGBoost can outperform a multinomial logit model in a prepayment setting and hence if it is better able to capture prepayment behaviour. Combining the above observations provides evidence that the XGBoost model of this thesis indeed leads to better predictions of prepayment behaviour than the multinomial logit model. The F1 scores for the XGBoost model are higher than the F1 scores of the multinomial logit model. However, the F1 score for the full prepayment class of the XGBoost model is very low. This is due to a low precision and low recall, of respectively 8.2% and 14.6%. Thus, even a well tuned XGBoost model does not have much discriminatory power between monthly observations for a full prepayment and the other classes. This result can be expected, as features do not differ much between consecutive months and a prepayment remains a behavioural decision.

| | | **Actual class** | | |
|---|---|---|---|---|
| | | No Event | Full Prepayment | Part Prepayment |
| **Predicted class** | No Event | 144543 | 1927 | 2495 |
| | | (95.69%) | (79.07%) | (29.43%) |
| | Full Prepayment | 3933 | 354 | 247 |
| | | (02.6%) | (14.53%) | (02.91%) |
| | Partial Prepayment | 2574 | 156 | 5737 |
| | | (01.70%) | (06.40%) | (67.66%) |

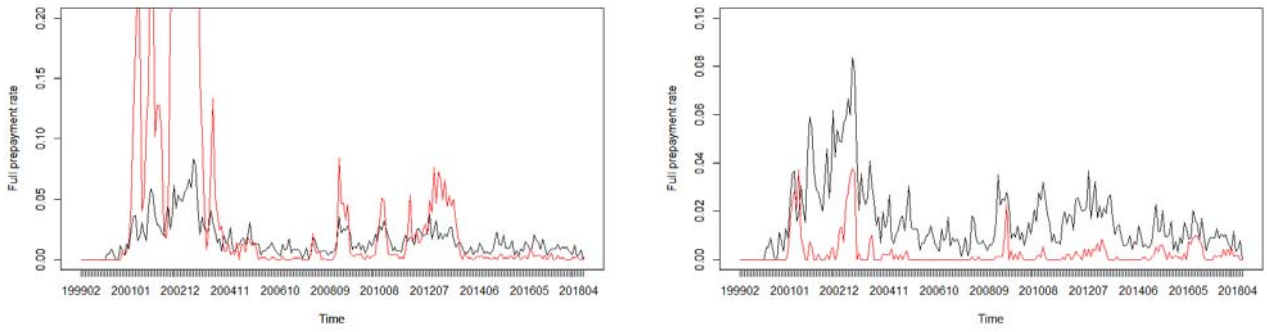
Table 7: Confusion matrix of the XGBoost model with prior distribution on the hold out set. For each class the predicted and actual values are shown. In brackets, the percentages of the number of observations of the actual class is displayed. The column percentages sum to one.

| | | **Actual class** | | |
|---|---|---|---|---|
| | | No Event | Full Prepayment | Part Prepayment |
| **Predicted class** | No Event | 131216 | 1887 | 1013 |
| | | (86.87%) | (77.43%) | (11.95%) |
| | Full Prepayment | 365 | 16 | 17 |
| | | (00.24%) | (00.66%) | (00.20%) |
| | Partial Prepayment | 19469 | 534 | 7449 |
| | | (12.89%) | (21.91%) | (87.85%) |

Table 8: Confusion matrix of the multinomial logit model on the hold out set. For each class the predicted and actual values are shown. In brackets, the percentages of the number of observations of the actual class is displayed. The column percentages sum to one.

### 5.3 Model results over time

Because banks are eventually interested in determining the prepayment rate over time, the prepayment rate over time is calculated. The full predicted and actual prepayment rates over time are plotted in Figure 5. From Section 5.2.1 it is already clear that the XGBoost model overestimates the amount of full prepayments. Interestingly, as shown in Figure 5a, these full prepayment forecasts happen in months where the prepayment rate is already high. The most extreme case is between 2000 and 2004, where the model hugely overestimates the prepayment rate. This is potentially because the model trains on the high prepayment rate data points and this increases the probability of prepayment for data with features corresponding to that time period. The mortgage rate is decreasing drastically in this period. The model likely overreacts to this because this is rare in the data and only happens again during the financial crisis of 2008, where the model also predicts a peak prepayment rate. Contrarily, in months with an average or low prepayment rate, the XGBoost




(a) XGBoost full prepayment rate over time.

(b) MNL full prepayment rate over time.

Figure 5: The actual prepayment rate over time of the hold out sample is plotted in black, whereas the predicted prepayment rate is plotted in red. The left plot is of the XGBoost model and the right plot is of the multinomial logit model.

model underestimates the prepayment rate.

From the confusion matrix in Section 5.2.1 it is already clear that the multinomial logit model underestimates the number of full prepayments. Judging from Figure 5b, the prepayment rate is indeed underestimated. The multinomial logit model underestimates the prepayment rate in all months, except for April and March 2001 and January 2018.

The mean absolute error of the forecasted prepayment rate against the realized prepayment rate for the XGBoost model is 0.033 and for the multinomial logit model it is 0.014. It is thus straightforward to conclude that the MNL model performs better when the prediction results are viewed in a temporal context. However, the mean actual prepayment rate is 0.016, which is low. This, in combination with the fact that in Figure 5b it is shown that the MNL model hugely underpredicts the prepayment rate, confirms that the MNL model not a better model. Indeed, the error is lower, but also it rarely predicts prepayments. Additionally, the error of the XGBoost is extremely high because of the period between 2000 and 2005, where the predicted prepayment rate is extremely high.

## 5.4 Model explanation/SHAP

The final part of this research entails how SHAP adds to the interpretability of the XGBoost model. With SHAP, relations between the independent and dependent variables can be found. In Figures 6, 7 and 8, the SHAP plots of respectively the full prepayment, partial prepayment and no event classes can be found. For each class the six most important variables are displayed. Those six

variables have the highest average absolute SHAP value. In Figures 12, 13 and 14 in the Appendix, the influence of the twelve most important variables can be found. Recall that the SHAP values for a specific class are the marginal attributions of a specific variable to the probability of that class for each instance. The SHAP attributions are plotted for all existing observations of that specific variable. This results in the global influence of the variables on the outcome of the XGBoost model. Recall that the SHAP values are attributions to the log odds. The blue dots in the figures are the individual SHAP attributions and the red line is a smoothed average over all observations.

Globally, the partial prepayment flag, which is yes if another partial prepayment has occurred, and no if a partial prepayment has not occurred (yet), is the most important variable. It has the highest average absolute SHAP value. Figure 6a displays that the probability of a full prepayment decreases when another partial prepayment has already occurred (flag=1). Figure 7a shows that this is possibly due to the the fact that the partial prepayment flag is such a strong indicator for another partial prepayment. Therefore, the probability shifts from full prepayment and no event to partial prepayment.

Recall that for the first seven months in each contract, no prepayments occur. This is visible in Figures 6d, 7d and 8d. The effect on partial prepayment probabilities is negative and so large that it is not visible in the figure. The absence of partial prepayments in the first seven months of the contract inflates the probabilities of full prepayment and no event relative to eight months and higher.

Additionally, all figures show some degree of dispersion of SHAP attributions. This indicates interaction with other variables. A middle class income, e.g., only indicates a high probability of a full prepayment when the original unpaid principal balance is bigger than $200,000. The dispersion shows that the XGBoost model finds many interactions in the data.

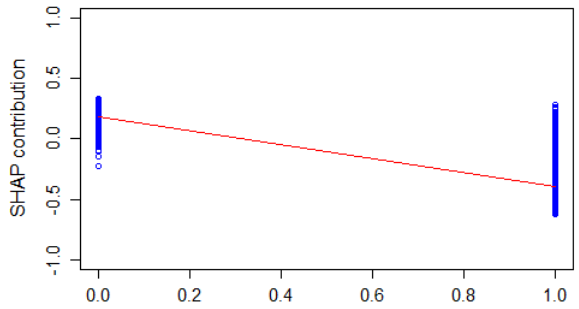
### 5.4.1 Full prepayments

Figure 6 shows the SHAP values for the log odds of full prepayment for the most important variables. The percentage unpaid principal balance is a good example of a non-linear relationship with full prepayment probability. When the percentage unpaid principal balance is 100, meaning nothing is paid off, this positively influences the probability of a full prepayment. When more UPB is paid off, the probability of a full prepayment declines and after approximately sixty percent of the original UPB is paid off, the probability of a full prepayment rises again. Figure 6c shows that as fewer months remain in the contract, the probability of a prepayment increases. Moreover, at 354 to 360

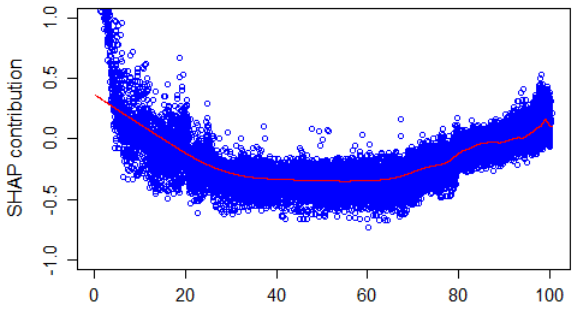
months remaining, a small positive shock is observed. This is most likely due to the absence of partial prepayment in the first 7 months in combination with the fact that most mortgages have a duration of 30 years (360 months) or shorter. For loan age, the probability of a full prepayment increases until approximately one hundred months, from where it is constant. Contract interest rate is also an important feature. The relationship with full prepayment probability is linear. There is, however, a high degree of dispersion. This indicates many interactions with other variables. Interestingly, the full prepayment probability declines after eight and a half percent. This decline is only minor. It indicates that after a contract interest rate of eight percent, although the SHAP values are dispersed, the biggest collection of SHAP values is around 0.3. Finally, the log of income shows a parabola relationship with prepayment probabilities, where for both high and low income families the full prepayment probability is higher.

### 5.4.2 Partial prepayments

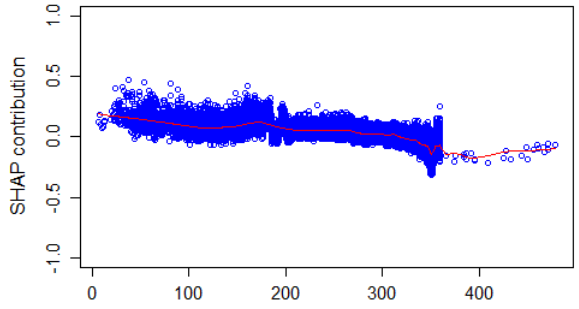
Figure 7 shows the SHAP values for the log odds of partial prepayment for the most important variables. The SHAP values for the partial prepayment class show a high degree of dispersion. The increase in partial prepayment probability when the partial prepayment flag is 1 is already mentioned and clearly visible here. For the percentage UPB paid there is a lot of dispersion. Also, the range of the figure is adjusted to account for the range of the relationship between the variable and the outcome. As more principal is paid off, the probability of a partial prepayment increases. At some point, the probability decreases again, possibly due to the fact that at some point a partial prepayment becomes a full prepayment as the principal goes to zero. Again, for both months remaining and loan age it is visible that in the first seven months no partial prepayments are observed. The SHAP values of loan age for the first seven months are very low, indicating the absence of partial prepayments in the first seven months. From eight months the prepayment probability is relatively high and declines with loan age. For both loan age and months remaining, the degree of dispersion is high, which indicates many interactions. The effect of the contract interest rate on the partial prepayment probability is relatively constant, but increases abruptly around nine percent. Finally, the partial prepayment probability initially increases with the log of income, but then decreases and eventually stays constant.



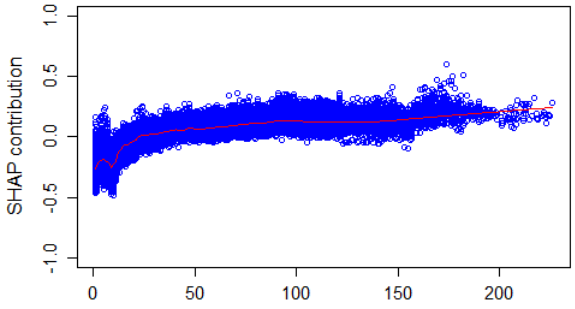

(a) Partial prepayment flag



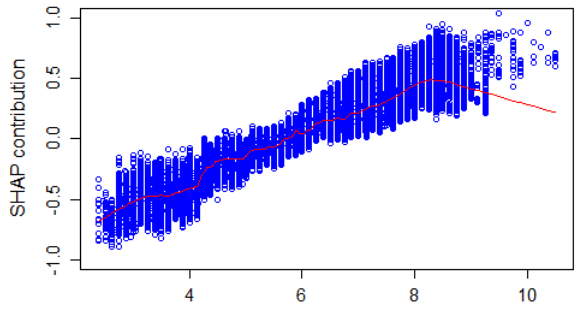

(b) Percentage UPB paid



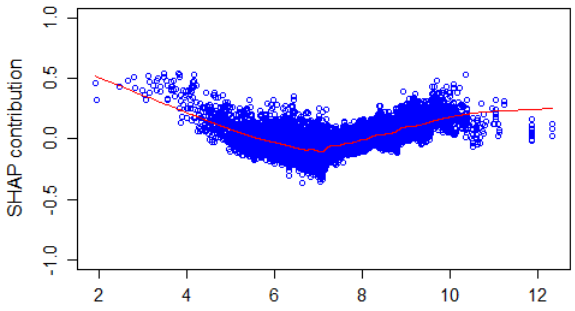

(c) Months remaining




(d) Loan age

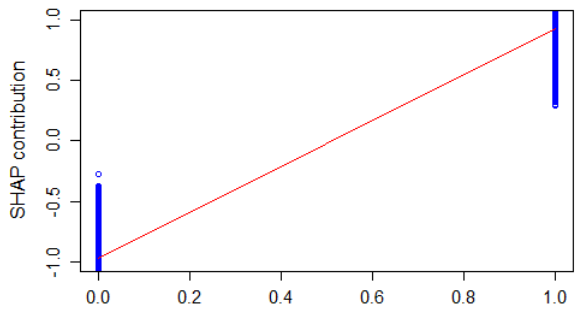



(e) Contract interest rate

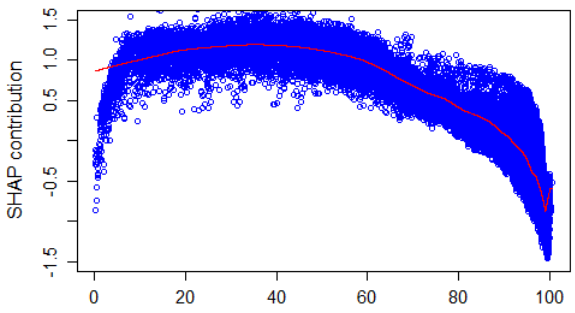



(f) Log of monthly income

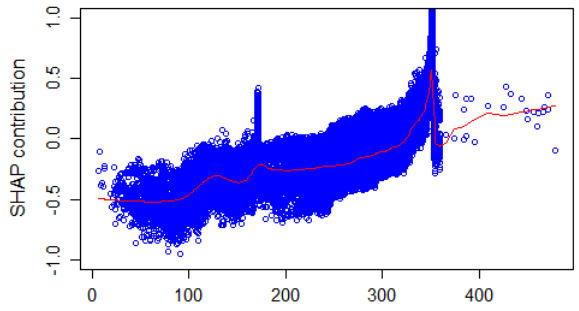
Figure 6: Global SHAP attributions for the "full prepayment" class for the six most contributing variables.



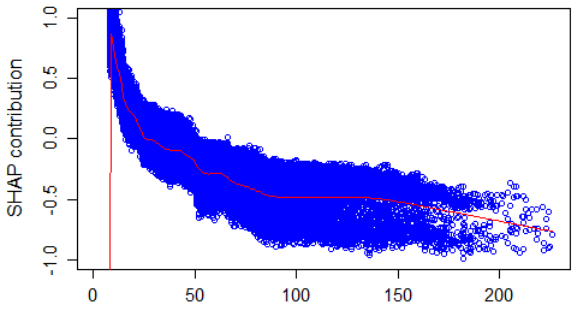

(a) Partial prepayment flag



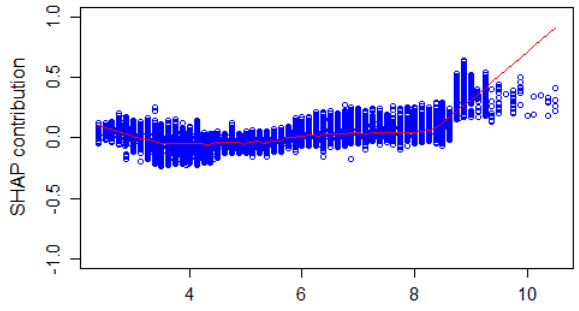

(b) Percentage UPB paid



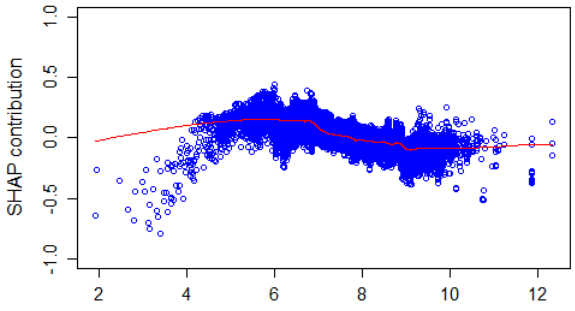

(c) Months remaining




(d) Loan age




(e) Contract interest rate




(f) Log of monthly income

Figure 7: Global SHAP attributions for the "partial prepayment" class for the six most contributing variables. Percentage UPB paid has a different scale than the other figures.

### 5.4.3 No prepayment event

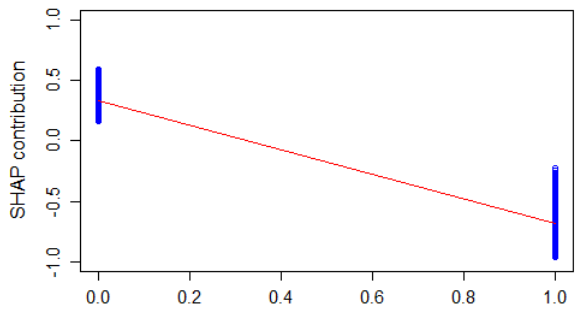
Figure 8 shows the SHAP values for the log odds of no prepayment event for the most important variables. Most results from these figures follow directly from the other prepayment class SHAP figures. Interestingly, the big increase for partial prepayment probability as more principal is paid, comes from the no prepayment event probability. The range in Figure 8b is changed to adjust for the range of the entire relationship. As more principal is paid off, the probability of no event declines heavily, although the dispersion is high. As there are fewer months remaining for the contract, the probability of no prepayment event increases. This is possibly due to the fact that the partial prepayment probability decreases as there are fewer months remaining and hence either a full prepayment or no event happens. Moreover, mortgagors who consistently partially prepay, have already fully prepaid their mortgage when the months remaining becomes small and hence have no SHAP value for these feature values. Adding that mortgagors who prepay are inclined to prepay again, it seems reasonable that partial prepayment probabilities decline when months remaining decline and hence the no event probability increases. Additionally, the difference between the first seven months of the contract and the later contract months is greater for the probability of no prepayment event than that for a full prepayment. Hence, the probability of no prepayment event increases from the fact that in the first seven contract months there is no partial prepayment.

## 5.5 Comparison coefficients with SHAP

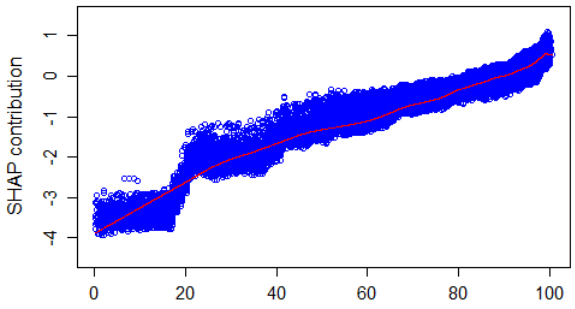
Recall that instead of comparing the log odds coefficients from the multinomial logit, the coefficients from the Poisson regressions are used to relate SHAP values of each class to regressions of each class. These Poisson regressions form the basis of the multinomial logit and using these regressions gives the additional benefit of being able to compare a coefficient for each class with the SHAP values of each class. Table 9 shows the Poisson coefficients of the six most important variables according to the SHAP model.

Similar to the SHAP values, the coefficients for no prepayment event and partial prepayment have a higher absolute value than those of full prepayment. The only exception is the contract interest rate, which coefficient for full prepayment has a higher absolute value. This is expected, as this relationship in the SHAP plots is a linear relation.

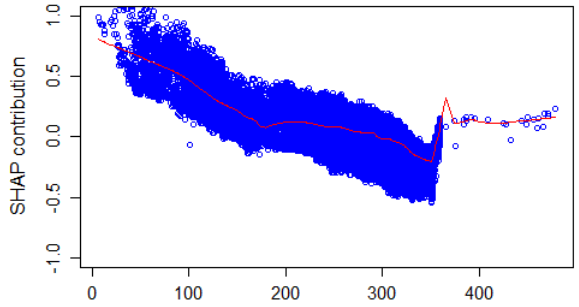
The directions of the SHAP relationships and coefficients are similar for the partial prepayment flag. Also the extreme influence for partial prepayment is captured by the MNL model. The MNL model has difficulties with the non-linear relationship between model output and percentage UPB,



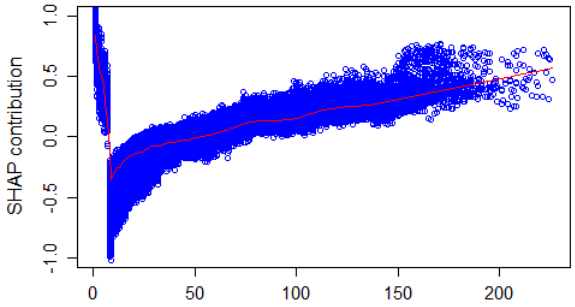

(a) Partial prepayment flag



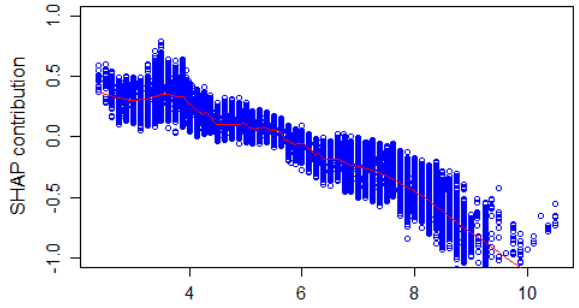

(b) Percentage UPB paid



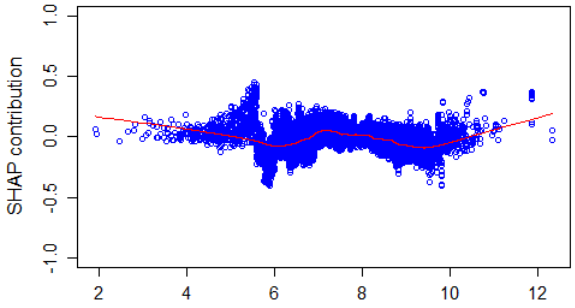

(c) Months remaining




(d) Loan age




(e) Contract interest rate




(f) Log of monthly income

Figure 8: Global SHAP attributions for the "no event" class for the six most contributing variables. Percentage UPB paid has a different scale than the other figures.

| | No Event | Full Prepayment | Partial Prepayment |
|---|---|---|---|
| Partial prepayment flag | -11.123 | -0.903 | 20.157 |
| % Principal paid off | 0.043 | -0.005 | -0.038 |
| Months remaining | -0.003 | -0.001 | 0.004 |
| Loan age | 0.010 | -0.001 | -0.008 |
| Interest rate | -0.260 | 0.380 | -0.120 |
| Log monthly income | . | . | . |

Table 9: The coefficients of the regularized Poisson regressions for all 3 prepayment classes. Points indicate that the variable is excluded from the regularized regression.

although the mainly upward direction of the no prepayment event class and mainly negative direction of the partial prepayment class agree. For loan age and months remaining, the MNL model has very small beta's. The XGBoost model captures many interaction effects that the MNL has no capacity of capturing without explicit modelling. The relationship of contract interest rate with prepayment probabilities has similar directions for both models, except for partial prepayment. Where the SHAP values indicate a constant or slightly increasing relationship, the Poisson regression shows a negative relation. Interestingly, the regularized Poisson regression excludes the log of monthly income as a variable, meaning that in the Poisson regression the log of monthly income is found to have no significant predictive power. This possibly is due to the fact that people with middle-class incomes have SHAP values of 0 and hence do not add to the probabilities. These middle incomes form the majority of the data. The higher and lower incomes do have an effect, as is seen in the SHAP plots.

# 6 Conclusion & discussion

This thesis has shown that a well tuned XGBoost model has superior performance over the multinomial logit in a prepayment setting and hence is better able to predict prepayment events. The F1 scores for all prepayment classes are higher for the XGBoost model than for the multinomial logit model. The F1 scores of the XGBoost model increase when a distribution is added that limits the number of partial prepayments. However, even the XGBoost model has a poor performance for predicting full prepayments. The discriminatory power of the model between full prepayments and the other prepayment classes for monthly observations is low. When plotting the predicted prepayment rate over time one can see that the XGBoost model either overpredicts or underpredicts the full prepayment rate. The XGBoost model is highly sensitive to periods where the mortgage

rate declines. The multinomial logit severely underpredicts full prepayments. In general, it rarely predicts full prepayment and this is confirmed by the prepayment rate over time.

For the second part of the research, this thesis uses SHAP to investigate why the XGBoost model makes a certain prediction. It is shown that many relationships between variables and the prepayment probabilities are non-linear and that this is captured by the XGBoost model. All SHAP figures display a degree of dispersion. This shows interaction between variables in the XGBoost model. Whether a partial prepayment has already occurred is found to be the most important contributor to the model in terms of average absolute SHAP values. The thesis confirms other research that states that full prepayments are more likely as the loan ages.

Additionally, this thesis finds that extreme incomes have a distinct effect on prepayment probabilities. Comparing the Poisson regression coefficients with the SHAP values, the two match roughly. The SHAP values do display a high degree of non-linearity which is not in the Poisson regression coefficients. Monthly income does not match, however. The regularized Poission regression shrinks the coefficients of monthly income to zero whereas it is found to be an important feature by the SHAP methodology.

A limitation of this research is that the Freddie Mac data does not include personal characteristics such as age or race. The XGBoost model possibly predicts prepayment better when these factors are in the model. Further research can focus on data sets including these personal characteristics, e.g. non-public data sets from financial service providers, to see if the model improves. Additionally, in the Freddie Mac data set no partial prepayments are observed for the first seven months. This makes the XGBoost model not forecast partial prepayments in these months. Further research can focus on the first months of the mortgage contract, or other data can be used where these first months are not limited to full prepayments and no prepayment events.

## References


Altman, E. I., & Saunders, A. (1997, 12). Credit Risk Measurement: Developments Over the Last 20 Years. *Journal of Banking & Finance*, *21*(11-12), 1721–1742. doi: 10.1016/S0378-4266(97)00036-8

Archer, W. R., Ling, D. C., & McGill, G. A. (2003). Household Income, Termination Risk and Mortgage Pricing. *The Journal of Real Estate Finance and Economics*, *27*(1), 111–138. doi: 10.1023/A:1023663530674

Bennett, P., Peach, R. W., Peristiani, S., Bennett, P., Peach, R., & Peristiani, S. (2001). Structural Change in the Mortgage Market and the Propensity to Refinance. *Journal of Money, Credit and Banking*, *33*(4), 955–75.

Ben Taieb, S., & Hyndman, R. J. (2014, 4). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting*, *30*(2), 382–394. doi: 10.1016/j.ijforecast.2013.07.005

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., ... Hay, S. I. (2013, 4). The global distribution and burden of dengue. *Nature*, *496*(7446), 504–507. doi: 10.1038/nature12060

Boyes, W. J., Hoffman, D. L., & Low, S. A. (1989, 1). An Econometric Analysis of the Bank Credit Scoring Problem. *Journal of Econometrics*, *40*(1), 3–14. doi: 10.1016/0304-4076(89)90026-2

Brainard, L. (2018). *What Are We Learning about Artificial Intelligence in Financial Services?* Philadelphia. Retrieved from `https://www.federalreserve.gov/newsevents/speech/brainard20181113a.html (2nd of May, 2019)`

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Brier, G. W. (1950, 1). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, *78*(1), 1–3. doi: 10.1175/1520-0493(1950)078¡0001:vofeit¿2.0.co;2

Bureau of Economic Analysis. (2019). *Gross Domestic Product, Second Quarter 2019 (Second Estimate); Corporate Profits, Second Quarter 2019 (Preliminary Estimate) — U.S. Bureau of Economic Analysis (BEA).* Retrieved from `https://www.bea.gov/news/2019/gross-domestic-product-2nd-quarter-2019-second-estimate-corporate-profits-2nd-quarter`

Caplin, A., Freeman, C., & Tracy, J. (1997, 11). Collateral Damage: Refinancing Constraints and Regional Recessions. *Journal of Money, Credit and Banking*, *29*(4), 496. doi: 10.2307/2953710

Cauchy, A. (1847). Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes rendus de l'Académie des sciences*, 536–538.

Charlier, E., & Van Bussel, A. . (2001). *Prepayment Behavior of Dutch Mortgagors: An Empirical Analysis* (Tech. Rep.).

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *CoRR*, *abs/1603.02754*. Retrieved from `http://arxiv.org/abs/1603.02754`

Clapp, J. M., Deng, Y. ., An, & Xudong. (2006). Unobserved Heterogeneity in Models of Competing Mortgage Termination Risks. *Real Estate Economics*, *34*(2), 243–273.

Clapp, J. M., Goldberg, G. M., Harding, J. P., & LaCour-Little, M. (2001). Movers and Shuckers: Interdependent Prepayment Decisions. *Real Estate Economics*, *29*(3), 411–450.

Crosbie, P., & Bohn, J. (2019). Modeling Default Risk. *Moody's MVK "White Paper" series*, 471–506. Retrieved from `https://www.worldscientific.com/doi/abs/10.1142/9789814759595_0020` doi: 10.1142/9789814759595_0020

Deng, Y., Quigley, J. M., & Order, R. (2000, 3). Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options. *Econometrica*, *68*(2), 275–307. doi: 10.1111/1468-0262.00110

Dennis, J. E., & Schnabel, R. B. (1996). Convergence results for properly chosen steps. In *Numerical methods for unconstrained optimization and nonlinear equations.* (pp. 120–125).

Dunn, K., & Spatt, C. (1988). Private information and incentives: Implications for mortgage contract terms and pricing. *The Journal of Real Estate Finance and Economics*, *1*(1), 47–60. doi: 10.1007/BF00207903

Federal Reserve. (2019). *Mortgage Debt Outstanding.* Retrieved from `https://www.federalreserve.gov/data/mortoutstand/current.htm`

Foster, C., & Van Order, R. (1984). An Option-Based Model of Mortgage Default. *Housing Finance Review*, *3*.

Freddie Mac. (2018). *House Price Index - All States and US National.* Retrieved from `https://www.quandl.com/data/FMAC/HPI-House-Price-Index-All-States-and-US-National`

Freddie Mac. (2019). *Freddie Mac Single Family Loan-Level Data Set Sample.* Retrieved from

http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Fujino, A., Isozaki, H., & Suzuki, J. (2008). Multi-label Text Categorization with Model Combination based on F1-score Maximization. In *Proceedings of the third international joint conference on natural language processing: Volume-ii.*

Gerds, T. A., & Schumacher, M. (2006, 12). Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*, *48*(6), 1029–1040. doi: 10.1002/bimj.200610301

Ghent, A. C., & Kudlyak, M. (2011, 9). Recourse and Residential Mortgage Default: Evidence from US States. *Review of Financial Studies*, *24*(9), 3139–3186. doi: 10.1093/rfs/hhr055

Goncharov, Y. (2002). *An Intensity-Based Approach for Valuation of Mortgage Contracts Subject to Prepayment Risk* (Tech. Rep.).

Green, J., & Shoven, J. B. (1986). The Effects of Interest Rates on Mortgage Prepayments. *Journal of Money, Credit and Banking*, *18*(1), 41–59.

Greene, W. H. (2002). The Multinomial Logit Model. In *Econometric analysis* (Fifth Edition ed., pp. 720–723). New Jersey: Pearson Education.

Guelman, L. (2012, 2). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*(3), 3659–3667. doi: 10.1016/j.eswa.2011.09.058

Hansen, L., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. doi: 10.1109/34.58871

Hayre, L. (2003). Prepayment modeling and valuation of Dutch mortgages. *The Journal of Fixed Income*, *12*(4).

Jacobs, J., Koning, R. H., & Sterken, E. (2005). *Modelling Prepayment Risk* (Tech. Rep.). University of Groningen.

Kang, P., & Zenios, S. A. (1992). Complete Prepayment Models for Mortgage-Backed Securities. *Management Science*, *38*(11), 1665–1685.

Kearns, M. (1988). *Thoughts on Hypothesis Boosting* (Tech. Rep.). University of Pennsylvania. Retrieved from https://www.cis.upenn.edu/ mkearns/papers/boostnote.pdf

Kearns, M., & Valiant, L. G. (1989). Crytographic limitations on learning Boolean formulae and finite automata. In *Proceedings of the twenty-first annual acm symposium on theory of computing - stoc '89* (pp. 433–444). New York, New York, USA: ACM Press. doi:

10.1145/73007.73049

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (Tech. Rep.). Retrieved from `https://github.com/slundberg/shap`

Mangal, A., & Kumar, N. (2016). Using big data to enhance the bosch production line performance: A Kaggle challenge. In *Proceedings - 2016 ieee international conference on big data, big data 2016* (pp. 2029–2035). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/BigData.2016.7840826

Pavlov, A. D. (2001). Competing Risks of Mortgage Termination: Who Refinances, Who Moves, and Who Defaults? *The Journal of Real Estate Finance and Economics*, *23*(2), 185–211. doi: 10.1023/A:1011158400165

Quigley, J. M., & Weinberg, D. H. (1977, 10). Intra- Urban Residential Mobility: A Review and Synthesis. *International Regional Science Review*, *2*(1), 41–66. doi: 10.1177/016001767700200104

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why Should I Trust You? Explaining the Predictions of Any Classifier* (Tech. Rep.). San Francisco, CA, USA: University of Washington. doi: 10.1145/2939672.2939778

Richard, S. F., & Roll, R. (1989, 4). Prepayments on fixed-rate mortgage-backed securities. *The Journal of Portfolio Management*, *15*(3), 73–82. doi: 10.3905/jpm.1989.409207

Riksen, R. (2017). *Using Artificial Neural Networks in the Calculation of Mortgage Prepayment Risk* (Unpublished doctoral dissertation).

Rufibach, K. (2010, 8). Use of Brier score to assess binary predictions. *Journal of Clinical Epidemiology*, *63*(8), 938–939. doi: 10.1016/j.jclinepi.2009.11.009

Schapire, R. E. (1990). The Strength of Weak Learnability. *Machine Learning*, *5*, 197–227.

Schwartz, E. S., & Torous, W. N. (1989, 6). Prepayment and the Valuation of Mortgage-Backed Securities. *The Journal of Finance*, *44*(2), 375. doi: 10.2307/2328595

Sepúlveda, J., & Velastin, S. (2015). F1 Score Assesment of Gaussian Mixture Background Subtraction Algorithms Using the MuHAVi Dataset. In *6th international conference on imaging for crime prevention and detection (icdp-15).* Institution of Engineering and Technology. doi: 10.1049/ic.2015.0106

Shapley, L. (1953). A Value for n-Person game. In Kuhn H. W. & A. W. Tucker (Eds.), *Contributions to the theory of games (am-28), volume ii.* (pp. 307–318). Princeton University Press.

Sirignano, J. A., Sadhwani, A., & Giesecke, K. (2015). *Deep Learning for Mortgage Risk* (Tech.

Rep.).

South, S. J., & Crowder, K. D. (1998, 2). Leaving the 'Hood: Residential Mobility between Black, White, and Integrated Neighborhoods. *American Sociological Review*, *63*(1), 17. doi: 10.2307/2657474

Spahr, R. W., & Sunderman, M. A. (2001). *The Effect of Prepayment Modeling in Pricing Mortgage-Backed Securities* (Vol. 3; Tech. Rep. No. 2).

St. Louis Federal Reserve Economic Data. (2019a). *30-Year Fixed Rate Mortgage Average in the United States.* Retrieved from `https://fred.stlouisfed.org/series/MORTGAGE30US`

St. Louis Federal Reserve Economic Data. (2019b). *Civilian Unemployment Rate.* Retrieved from `https://fred.stlouisfed.org/series/UNRATE`

St. Louis Federal Reserve Economic Data. (2019c). *Personal Savings Rate.* Retrieved from `https://fred.stlouisfed.org/series/PSAVERT`

St. Louis Federal Reserve Economic Data. (2019d). *Real Gross Domestic Product.* Retrieved from `https://fred.stlouisfed.org/series/GDPC1`

Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso* (Vol. 58; Tech. Rep. No. 1).

Vandell, K. D. (1995). How Ruthless Is Mortgage Default? A Review and Synthesis of the Evidence. *Journal of Housing Research*, *6*, 245–264. doi: 10.2307/24832828

Varli, Y., & Yildirim, Y. (2015). Default and prepayment modelling in participating mortgages. *Journal of Banking & Finance*, *61*, 81–88. doi: 10.1016/J.JBANKFIN.2015.09.003

Vasconcelos, P. (2010). *Modelling Prepayment Risk: Multinomial Logit Model Approach For Assessing Conditional Prepayment Rate* (Unpublished doctoral dissertation). University of Twente.

Wilks, D. (2010, 10). Sampling distributions of the Brier score and Brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, *136*(653), 2109–2118. doi: 10.1002/qj.709

Xiong, X., & De La Torre, F. (2013). Supervised Descent Method and its Applications to Face Alignment.
doi: 10.1109/CVPR.2013.75

Yinger, J. M. (1997). *Ethnicity: Source of Strength? Source of Conflict?* Rawat Publications.

Zhang, D., Wang, J., Zhao, X., & Wang, X. (2016, 1). A Bayesian hierarchical model for comparing average F1 scores. In *Proceedings - ieee international conference on data mining, icdm* (pp. 589–598). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ICDM.2015.44

Zhang, Y., & Haghani, A. (2015, 9). A gradient boosting method to improve travel time

prediction. *Transportation Research Part C: Emerging Technologies*, *58*, 308–324. doi: 10.1016/j.trc.2015.02.019

# A Appendix

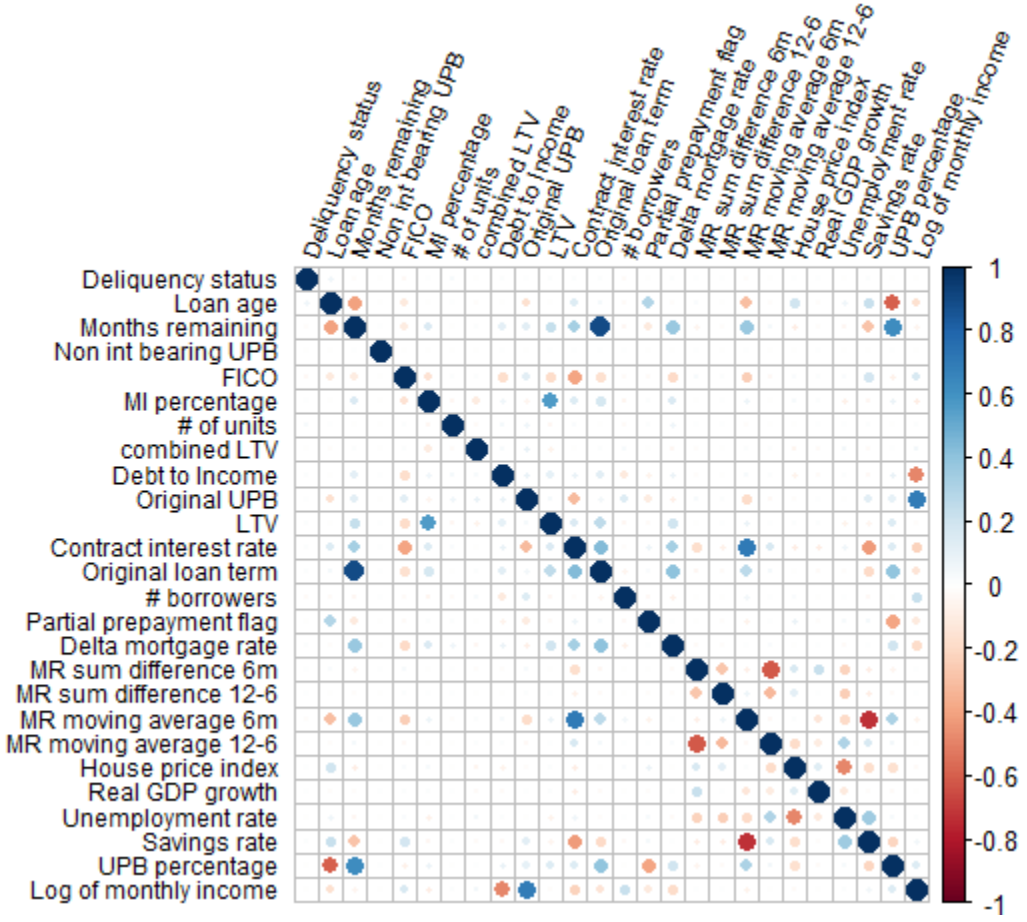



Figure 9: Correlation matrix of continuous variables of the Freddie Mac Single Family Loan Level Data Set enriched with macroeconomic variables including transformed variables.

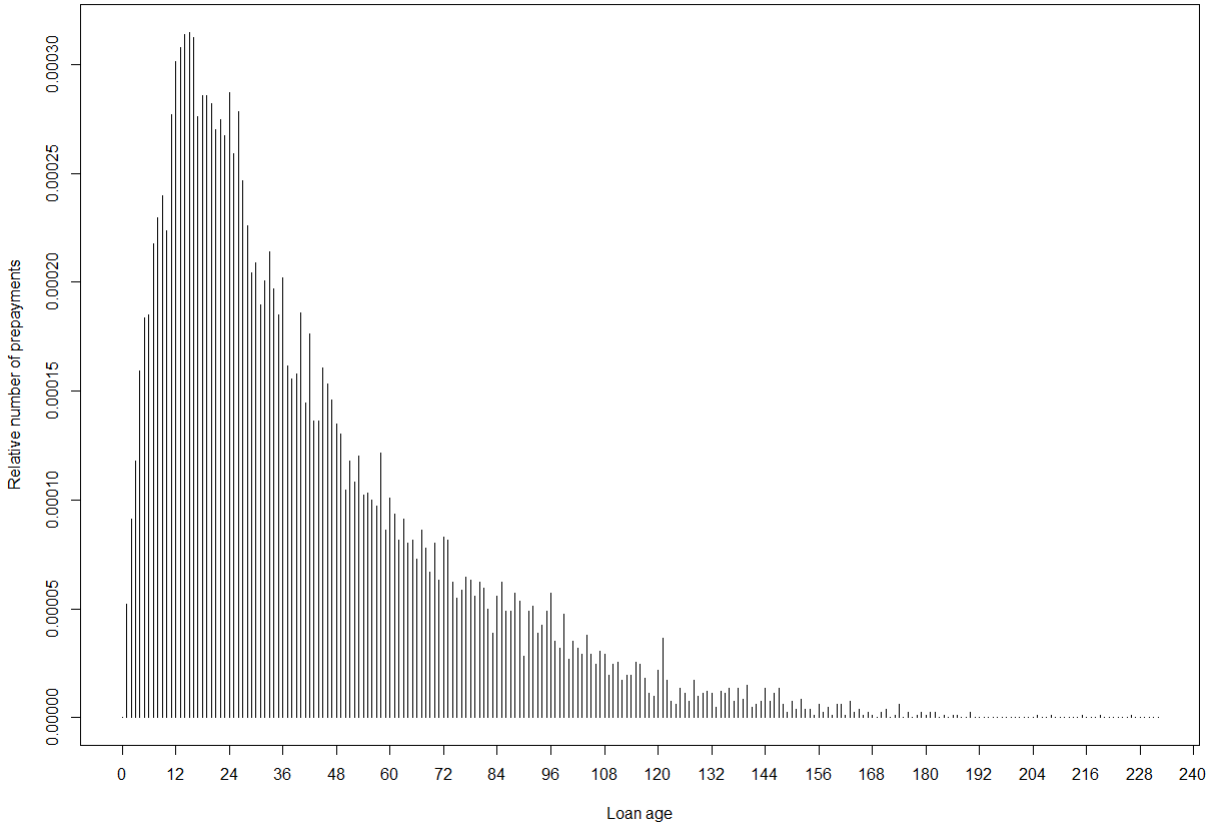



Figure 10: The number of full prepayments per contract months.

| Name | Value | Description |
|---|---|---|
| Credit Score | 301-850 or 9999 | Representation of a borrowers creditworthiness. This is the score at time of origination. |
| First Payment Date | YYYYMM | Date of first scheduled mortgage payment |
| First Time Homebuyer | Y, N, 9 | Indicates whether borrower is first time homeowner |
| Number of Units | 1, 2, 3, 4, 99 | Number of units in the property, 99 = NA |
| Occupancy status | P, S, I, 9 | Denotes if the mortgage type is primary resident, Second home or Investment property. 9 = NA |
| Original Combined LTV | 0-200%, 999 | Original mortgage + secondary mortgages |
| Original Debt to Income | 1-65% | Monthly debt payments divided by total monthly income |
| Original Unpaid Principal Balance | Rounded to 1.000 | The unpaid principal balance of the mortgage on the note date. |
| Original Loan to Value | 6-105%, 999 | Dividing loan amount by lesser of the property's value or purchase price, 999 = NA |
| Original Interest rate | % | Interest rate of the loan. |
| Channel | R, B, C, T, 9 | Retail, Broker, Correspondent, Third Party, Origination Not Specified, 9 |
| Property State | XX | Abbreviation indicating the state of the location of the mortgaged property. |
| Property Type | CO, PU, MH, SF, CP, 99 | Denotes whether the property type secured by the mortgage is a condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or Single Family home. |
| Loan Sequence number | XXXXXX | Unique identifier assigned to each loan, F1 is FRM |
| Loan Purpose | P, C, N, 9 | Indicates if the mortgage is cash-out refinance mortgage, no cash-out refinance mortgage or a purchase mortgage |
| Original Loan Term | 123 | A calculation of the number of scheduled monthly payments based on the first payment date and the maturity date. |
| Number of Borrowers | 01, 02, 99 | The number of borrowers who are obligated to repay the mortgage. 01 = 1 borrower, 02 >1 borrower, 99 = NA. |

Table 10: Explanation of the Freddie Mac Single Family Loan Level Data Set variables that are available at loan origination.

| Name | Value | Description |
|---|---|---|
| Loan Sequence Number | XXXXXX | Unique identifier assigned to each loan. |
| Monthly Reporting Period | YYYYMM | As-of month for loan information contained in loan record. Delinquency happened on month before. |
| Current Actual UPB | 12345 | Current actual UPB reflects the mortgage ending balance. |
| Current Loan Delinquency Status | XX, 0, 1, 2, 3, . . . , R | A value indicating the number of days the borrower is delinquent, based on the due date of last paid installment. N is in months and R means REO disposition. |
| Loan Age | 123 | Number of months since origination of the mortgage. |
| Remaining Months to Contractual Maturity | 123 | Number of months until contractual maturity date of the mortgage |
| Zero Balance Code | 1, 2, 3, 6, 9, 15 | Reason the balance has gone to zero. 1 = prepaid or matured (voluntary payoff), 2 = third party sale, 3 = short sale or charge off, 6 = repurchase prior to property disposition, 09 = REO Disposition, 15 = Note sale/re-performing sale. |
| Current Interest Rate | 12345 | Reflects the current interest rate on the mortgage note, taking into account any loan modifications. |
| Current deferred UPB | 12345 | The current non-interest bearing UPB of the modified mortgage. |

Table 11: Explanation of the monthly Freddie Mac Single Family Loan Level Data Set variables that follow the loan over time.

| Feature | Class | Class | Class | Class | Class | Class |
|---|---|---|---|---|---|---|
| FT buyer | Yes | No | NA | | | |
| | 91023 | 445429 | 266968 | | | |
| Occupancy | Inv. Prop | Prim. Prop | Sec. home | | | |
| | 48029 | 718165 | 37226 | | | |
| Channel | Broker | Cor. | Retail | Unspecified | | |
| | 42453 | 130753 | 399106 | 231108 | | |
| State | Alaska | Alabama | Arkansas | Arizona | California | Colorado |
| | 1871 | 9848 | 5886 | 19866 | 85697 | 18446 |
| | Connecticut | DC | Delaware | Florida | Georgia | Hawaii |
| | 8988 | 1710 | 3177 | 45368 | 27297 | 3162 |
| | Iowa | Idaho | Illinois | Indiana | Kansas | Kentucky |
| | 8020 | 4611 | 34934 | 21994 | 9246 | 12684 |
| | Louisiana | Mass. | Maryland | Maine | Michigan | Minnesota |
| | 8435 | 18499 | 18172 | 3587 | 30378 | 20402 |
| | Missouri | Mississippi | Montana | N Carolina | N Dakota | Nebraska |
| | 19491 | 3085 | 3102 | 27177 | 2094 | 4418 |
| | New Hamp. | New Jersey | New Mexico | Nevada | New York | Ohio |
| | 4647 | 21917 | 4979 | 5708 | 34373 | 32177 |
| | Oklahoma | Oregon | Penn. | Rhode Is. | S Carolina | S Dakota |
| | 8154 | 13936 | 30155 | 2297 | 12839 | 1585 |
| | Tennessee | Texas | Utah | Virginia | Vermont | Washington |
| | 12212 | 51553 | 8652 | 24653 | 3046 | 23826 |
| | Wisconsin | W Virg. | Wyoming | Extrater. | | |
| | 18311 | 2623 | 1461 | 2671 | | |
| Property type | Condo | Co-op | MH | PUD | SF home | |
| | 57761 | 1514 | 6094 | 130773 | 607278 | |
| Loan purpose | CO-ref | No CO-ref | Purchase | | | |
| | 222341 | 256009 | 325070 | | | |
| # Borrowers | One | Two | | | | |
| | 341161 | 462259 | | | | |
| Ppmt flag | Yes | No | | | | |
| | 147064 | 656356 | | | | |
| MR direction | Up | Neutral | Down | | | |
| | 346007 | 4784 | 452629 | | | |
| Month | January | February | March | April | May | June |
| | 67461 | 67787 | 68016 | 68571 | 68931 | 69209 |
| | July | August | September | October | November | December |
| | 64283 | 64859 | 65428 | 65865 | 66300 | 66710 |

Table 12: Summary statistics of categorical variables of the enriched cleaned Freddie Mac data set used to model prepayments. For each categorical variable the number of observations that belong to each class or category of that variable are displayed.

| | No Event | Full Prepayment | Partial Prepayment |
|---|---|---|---|
| Delinquency status | 0.1271 | 0.1648 | -0.2919 |
| Loan age | 0.0095 | -0.0014 | -0.0081 |
| Months remaining | -0.0027 | -0.0011 | 0.0038 |
| Non-interest rate UPB | . | . | . |
| FICO | -0.0013 | 0.0006 | 0.0007 |
| Non-first time homebuyer | 0.0285 | 0.0232 | -0.0517 |
| First Time homebuyer | 0.0243 | -0.0755 | 0.0512 |
| MI percentage | . | . | . |
| Number of units | 0.0195 | -0.1130 | 0.0935 |
| Occupancy: Primary resident | -0.0927 | 0.2176 | -0.1249 |
| Occupancy: Second home | -0.0759 | 0.0282 | 0.0477 |
| Combined LTV | 0.0002 | 0.0010 | -0.0012 |
| DTI | 0.0052 | 0.0040 | -0.0092 |
| Original UPB | -0.0007 | 0.0018 | -0.0011 |
| LTV | 0.0026 | 0.0003 | -0.0029 |
| Interest rate | -0.2600 | 0.3797 | -0.1198 |
| Channel - Correspondent | . | . | . |
| Channel - Retail | . | . | . |
| Channel - Third party | -0.0223 | 0.0174 | 0.0049 |
| St. Alabama | . | . | . |
| St. Arkansas | 0.0169 | -0.0024 | -0.0145 |
| St. Arizona | 0.0114 | -0.0007 | -0.0107 |
| St. California | 0.0093 | 0.0512 | -0.0605 |
| St. Colorado | -0.0720 | 0.1180 | -0.0460 |
| St. Connecticut | . | . | . |
| Washington DC | . | . | . |
| St. Delaware | 0.1460 | 0.0319 | -0.1779 |
| St. Florida | 0.0373 | -0.1154 | 0.0780 |
| St. Georgia | 0.0949 | -0.0187 | -0.0762 |
| St. Hawaii | 0.0257 | -0.1224 | 0.0967 |
| St. Iowa | -0.0126 | 0.1999 | -0.1873 |
| St. Idaho | 0.1384 | 0.0364 | -0.1748 |
| St. Illinois | -0.1275 | 0.0777 | 0.0498 |
| St. Indiana | . | . | . |
| St. Kansas | -0.0202 | -0.0255 | 0.0457 |
| St. Kentucky | -0.0008 | 0.0129 | -0.0121 |
| St. Louisiana | 0.0350 | -0.2531 | 0.2181 |
| St. Massachusetts | -0.0173 | 0.0415 | -0.0241 |
| St. Maryland | . | . | . |
| St. Maine | -0.0251 | 0.0297 | -0.0046 |
| St. Michigan | -0.0104 | 0.1075 | -0.0971 |
| St. Minnesota | -0.0330 | 0.0602 | -0.0272 |
| St. Missouri | -0.0204 | 0.0102 | 0.0102 |

Table 13: Coefficients of the grouped regularized Poisson regressions. 1/3

| | No Event | Full Prepayment | Partial Prepayment |
|---|---|---|---|
| St. Mississippi | . | . | . |
| St. Montana | . | . | . |
| St. North Carolina | 0.0337 | 0.0310 | -0.0647 |
| St. North Dakota | . | . | . |
| St. Nebraska | 0.0276 | 0.0676 | -0.0952 |
| St. New Hampshire | -0.0086 | 0.0103 | -0.0017 |
| St. New Jersey | -0.0314 | -0.0837 | 0.1151 |
| St. New Mexico | 0.2092 | -0.0205 | -0.1887 |
| St. Nevada | -0.0095 | 0.0057 | 0.0038 |
| St. New York | 0.1613 | -0.1253 | -0.0360 |
| St. Ohio | -0.0692 | -0.0282 | 0.0974 |
| St. Oklahoma | 0.0463 | -0.0600 | 0.0137 |
| St. Oregon | 0.0220 | -0.0165 | -0.0055 |
| St. Pennsylvania | -0.0270 | -0.0803 | 0.1074 |
| St. Rhode Island | 0.0257 | 0.0452 | -0.0709 |
| St. South Carolina | 0.0035 | -0.0427 | 0.0392 |
| St. South Dakota | -0.0515 | 0.0083 | 0.0432 |
| St. Tennessee | . | . | . |
| St. Texas | -0.0046 | -0.1634 | 0.1679 |
| St. Utah | -0.0716 | 0.0987 | -0.0272 |
| St. Virginia | . | . | . |
| St. Vermont | -0.0737 | -0.0395 | 0.1132 |
| St. Washington | 0.0046 | 0.0017 | -0.0063 |
| St. Wisconsin | -0.2211 | 0.1214 | 0.0997 |
| St. West Virginia | . | . | . |
| St. Wyoming | -0.1307 | -0.0724 | 0.2030 |
| Extraterritorial areas | 0.6149 | -0.2177 | -0.3972 |
| Property Type - Co-op | . | . | . |
| Property Type - MH | 0.2548 | -0.2222 | -0.0326 |
| Property Type - PUD | -0.0193 | 0.0171 | 0.0022 |
| Property Type - SF | -0.0220 | -0.0469 | 0.0689 |
| Purpose: NCO refinance | -0.0358 | 0.0009 | 0.0350 |
| Purpose: Purchase | -0.1173 | -0.0056 | 0.1229 |
| Original loan term | . | . | . |
| Number of borrowers | -0.0110 | 0.0785 | -0.0676 |
| Partial prepayment flag | -11.123 | -0.9034 | 20.157 |
| $\Delta$ mortgage rate | 0.0004 | -0.0003 | -0.0001 |
| Mortgage rate direction: down | 0.0071 | -0.0048 | -0.0023 |
| Mortgage rate direction: up | -0.0223 | 0.0179 | 0.0043 |
| Mortgage rate sum of diff 6m | 0.2154 | -0.3452 | 0.1298 |
| Mortgage rate SD 12-6 | 0.1198 | -0.1894 | 0.0696 |
| Mortgage rate moving av. 6m | 0.1149 | -0.2508 | 0.1359 |
| Mortgage rate MA 12-6 | -0.0035 | 0.0035 | -0.0001 |
| House price index | -0.0036 | 0.0047 | -0.0011 |

Table 14: Coefficients of the grouped regularized Poisson regressions. 2/3

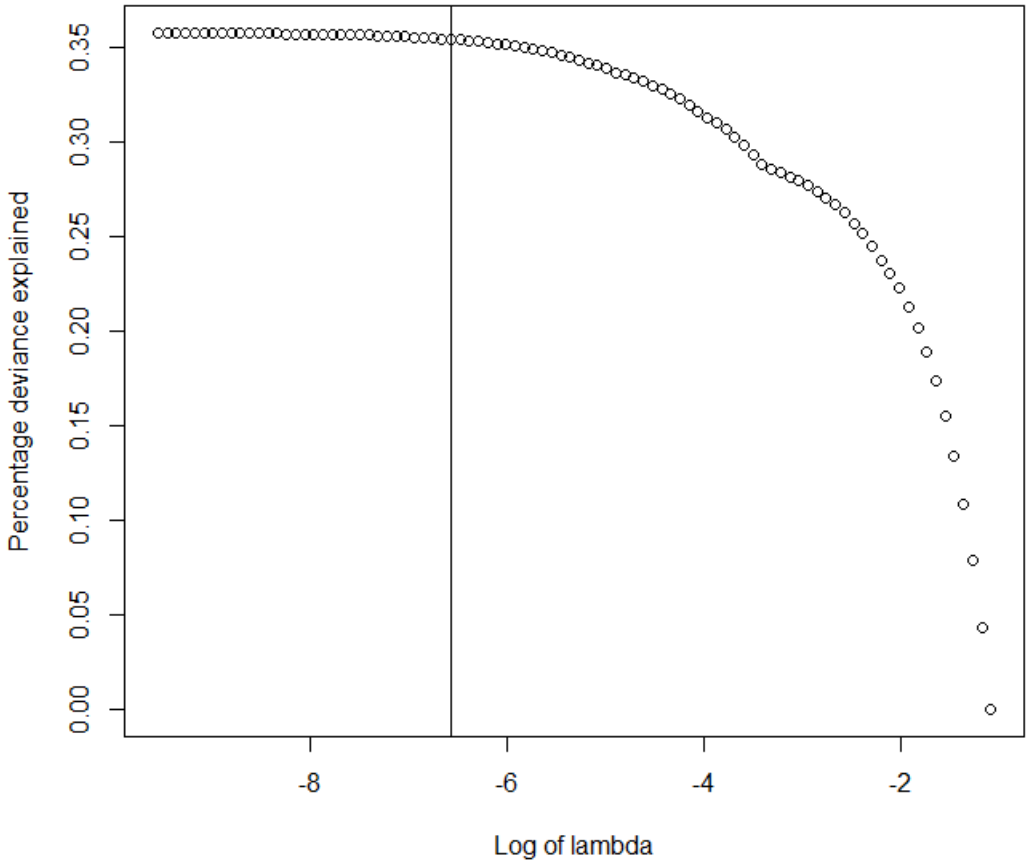



Figure 11: Percentage deviance explained for every lambda in the regularized multinomial logit. The log of lambda is taken because of its skewness towards low values. The vertical line corresponds with lambda = 0.001.

| | No Event | Full Prepayment | Partial Prepayment |
|---|---|---|---|
| Real GDP growth | -0.0232 | 0.0422 | -0.0190 |
| Unemployment rate | -0.0189 | 0.0188 | 0.0001 |
| Savings rate | -0.0020 | 0.0009 | 0.0011 |
| % principal paid off | 0.0432 | -0.0049 | -0.0383 |
| February | 0.0058 | 0.0105 | -0.0163 |
| March | . | . | . |
| April | . | . | . |
| May | -0.0115 | 0.0080 | 0.0035 |
| June | 0.0710 | -0.0351 | -0.0359 |
| July | -0.0055 | 0.0020 | 0.0035 |
| August | . | . | . |
| September | 0.0062 | 0.0183 | -0.0245 |
| October | . | . | . |
| November | 0.0088 | -0.0018 | -0.0070 |
| December | 0.0587 | -0.0809 | 0.0222 |
| Log monthly income | . | . | . |

Table 15: Coefficients of the grouped regularized Poisson regressions. 3/3

| | Full Prepayment | Partial Prepayment |
|---|---|---|
| Delinquency status | 0.0377 | -0.4190 |
| Loan age | -0.0109 | -0.0176 |
| Months remaining | 0.0016 | 0.0065 |
| Non-interest rate UPB | . | . |
| FICO | 0.0019 | 0.0020 |
| Non-first time homebuyer | -0.0053 | -0.0802 |
| First Time homebuyer | -0.0998 | 0.0269 |
| MI percentage | . | . |
| Number of units | -0.1325 | 0.0740 |
| Occupancy: Primary resident | 0.3103 | -0.0322 |
| Occupancy: Second home | 0.1041 | 0.1236 |
| Combined LTV | 0.0008 | -0.0014 |
| DTI | -0.0012 | -0.0144 |
| Original UPB | 0.0025 | -0.0004 |
| LTV | -0.0023 | -0.0055 |
| Interest rate | 0.6397 | 0.1402 |
| Channel - Correspondent | . | . |
| Channel - Retail | . | . |
| Channel - Third party | 0.0397 | 0.0272 |
| St. Alabama | . | . |
| St. Arkansas | -0.0193 | -0.0314 |
| St. Arizona | -0.0121 | -0.0221 |
| St. California | 0.0419 | -0.0698 |
| St. Colorado | 0.1900 | 0.0260 |
| St. Connecticut | . | . |
| Washington DC | . | . |
| St. Delaware | -0.1141 | -0.3239 |
| St. Florida | -0.1527 | 0.0407 |
| St. Georgia | -0.1136 | -0.1711 |
| St. Hawaii | -0.1481 | 0.0710 |
| St. Iowa | 0.2125 | -0.1747 |
| St. Idaho | -0.1020 | -0.3132 |
| St. Illinois | 0.2052 | 0.1773 |
| St. Indiana | . | . |
| St. Kansas | -0.0053 | 0.0659 |
| St. Kentucky | 0.0137 | -0.0113 |
| St. Louisiana | -0.2881 | 0.1831 |
| St. Massachusetts | 0.0588 | -0.0068 |
| St. Maryland | . | . |
| St. Maine | 0.0548 | 0.0205 |
| St. Michigan | 0.1179 | -0.0867 |
| St. Minnesota | 0.0932 | 0.0058 |
| St. Missouri | 0.0306 | 0.0306 |

Table 16: Coefficients of the multinomial logit model. The coefficients represent the log odds for each class against reference class "no event". 1/3

| | Full Prepayment | Partial Prepayment |
|---|---|---|
| St. Mississippi | . | . |
| St. Montana | . | . |
| St. North Carolina | -0.0027 | -0.0984 |
| St. North Dakota | . | . |
| St. Nebraska | 0.0400 | -0.1228 |
| St. New Hampshire | 0.0189 | 0.0069 |
| St. New Jersey | -0.0523 | 0.1465 |
| St. New Mexico | -0.2297 | -0.3979 |
| St. Nevada | 0.0152 | 0.0133 |
| St. New York | -0.2866 | -0.1973 |
| St. Ohio | 0.0410 | 0.1666 |
| St. Oklahoma | -0.1063 | -0.0326 |
| St. Oregon | -0.0385 | -0.0275 |
| St. Pennsylvania | -0.0533 | 0.1344 |
| St. Rhode Island | 0.0195 | -0.0966 |
| St. South Carolina | -0.0462 | 0.0357 |
| St. South Dakota | 0.0598 | 0.0947 |
| St. Tennessee | . | . |
| St. Texas | -0.1588 | 0.1725 |
| St. Utah | 0.1703 | 0.0444 |
| St. Virginia | . | . |
| St. Vermont | 0.0342 | 0.1869 |
| St. Washington | -0.0029 | -0.0109 |
| St. Wisconsin | 0.3425 | 0.3208 |
| St. West Virginia | . | . |
| St. Wyoming | 0.0583 | 0.3337 |
| Extraterritorial areas | -0.8326 | -1.0121 |
| Property Type - Co-op | . | . |
| Property Type - MH | -0.4770 | -0.2874 |
| Property Type - PUD | 0.0364 | 0.0215 |
| Property Type - SF | -0.0249 | 0.0909 |
| Purpose: NCO refinance | 0.0367 | 0.0708 |
| Purpose: Purchase | 0.1117 | 0.2402 |
| Original loan term | . | . |
| Number of borrowers | 0.0895 | -0.0566 |
| Partial prepayment flag | 0.2089 | 3.1280 |
| $\Delta$ mortgage rate | -0.0007 | -0.0005 |
| Mortgage rate direction: down | -0.0119 | -0.0094 |
| Mortgage rate direction: up | 0.0402 | 0.0266 |
| Mortgage rate sum of diff 6m | -0.5606 | -0.0856 |
| Mortgage rate SD 12-6 | -0.3092 | -0.0502 |
| Mortgage rate moving av. 6m | -0.3657 | 0.0210 |
| Mortgage rate MA 12m6 | 0.0070 | 0.0034 |
| House price index | 0.0083 | 0.0025 |

Table 17: Coefficients of the multinomial logit model. The coefficients represent the log odds for each class against reference class "no event". 2/3

| | Full Prepayment | Partial Prepayment |
|---|---|---|
| Real GDP growth | 0.0654 | 0.0042 |
| Unemployment rate | 0.0377 | 0.0190 |
| Savings rate | 0.0029 | 0.0031 |
| % principal paid off | -0.0481 | -0.0815 |
| February | 0.0047 | -0.0221 |
| March | . | . |
| April | . | . |
| May | 0.0195 | 0.0150 |
| June | -0.1061 | -0.1069 |
| July | 0.0075 | 0.0090 |
| August | . | . |
| September | 0.0121 | -0.0307 |
| October | . | . |
| November | -0.0106 | -0.0158 |
| December | -0.1396 | -0.0365 |
| Log monthly income | . | . |

Table 18: Coefficients of the multinomial logit model. The coefficients represent the log odds for each class against reference class "no event". 3/3

| Lambda | Relative deviance explained | Lambda | Relative deviance explained |
|---|---|---|---|
| 0.070 | 0.267 | 0.016 | 0.320 |
| 0.064 | 0.270 | 0.014 | 0.323 |
| 0.058 | 0.274 | 0.013 | 0.325 |
| 0.053 | 0.277 | 0.012 | 0.328 |
| 0.048 | 0.279 | 0.011 | 0.330 |
| 0.044 | 0.281 | 0.010 | 0.332 |
| 0.040 | 0.283 | 0.009 | 0.333 |
| 0.036 | 0.285 | 0.008 | 0.335 |
| 0.033 | 0.288 | 0.007 | 0.338 |
| 0.030 | 0.293 | 0.006 | 0.34 |
| 0.028 | 0.298 | 0.005 | 0.343 |
| 0.025 | 0.302 | 0.004 | 0.347 |
| 0.023 | 0.306 | 0.003 | 0.350 |
| 0.021 | 0.31 | 0.002 | 0.352 |
| 0.019 | 0.313 | **0.001** | **0.354** |
| 0.017 | 0.316 | 0.000 | 0.357 |

Table 19: All lambda's and the corresponding percentage deviance explained. The lambda used for the regularized multinomial logit is in bold. This table corresponds to Figure 11 in the Appendix.

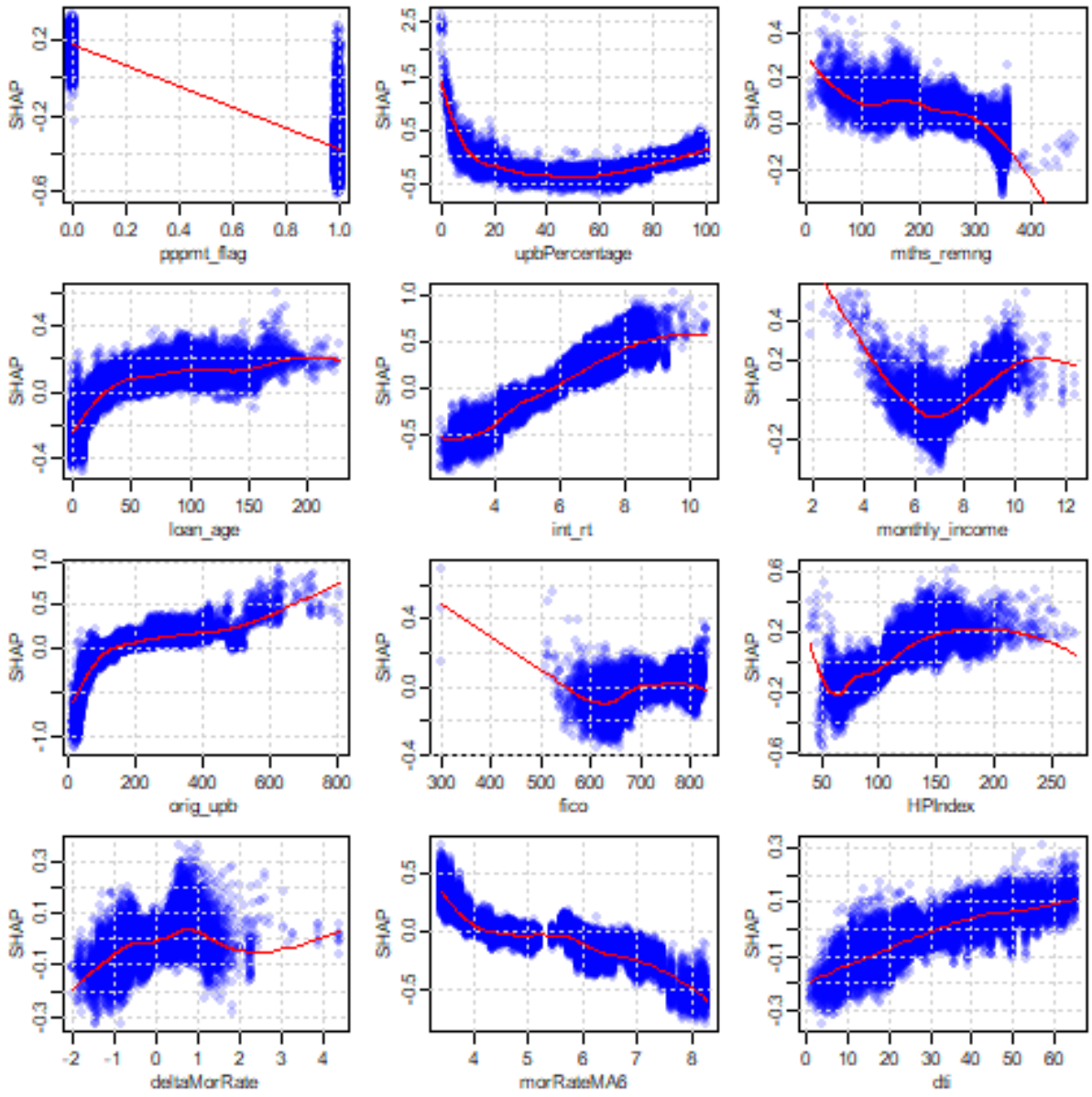



Figure 12: SHAP values of the 12 most important variables for class full prepayment in the XGBoost model.

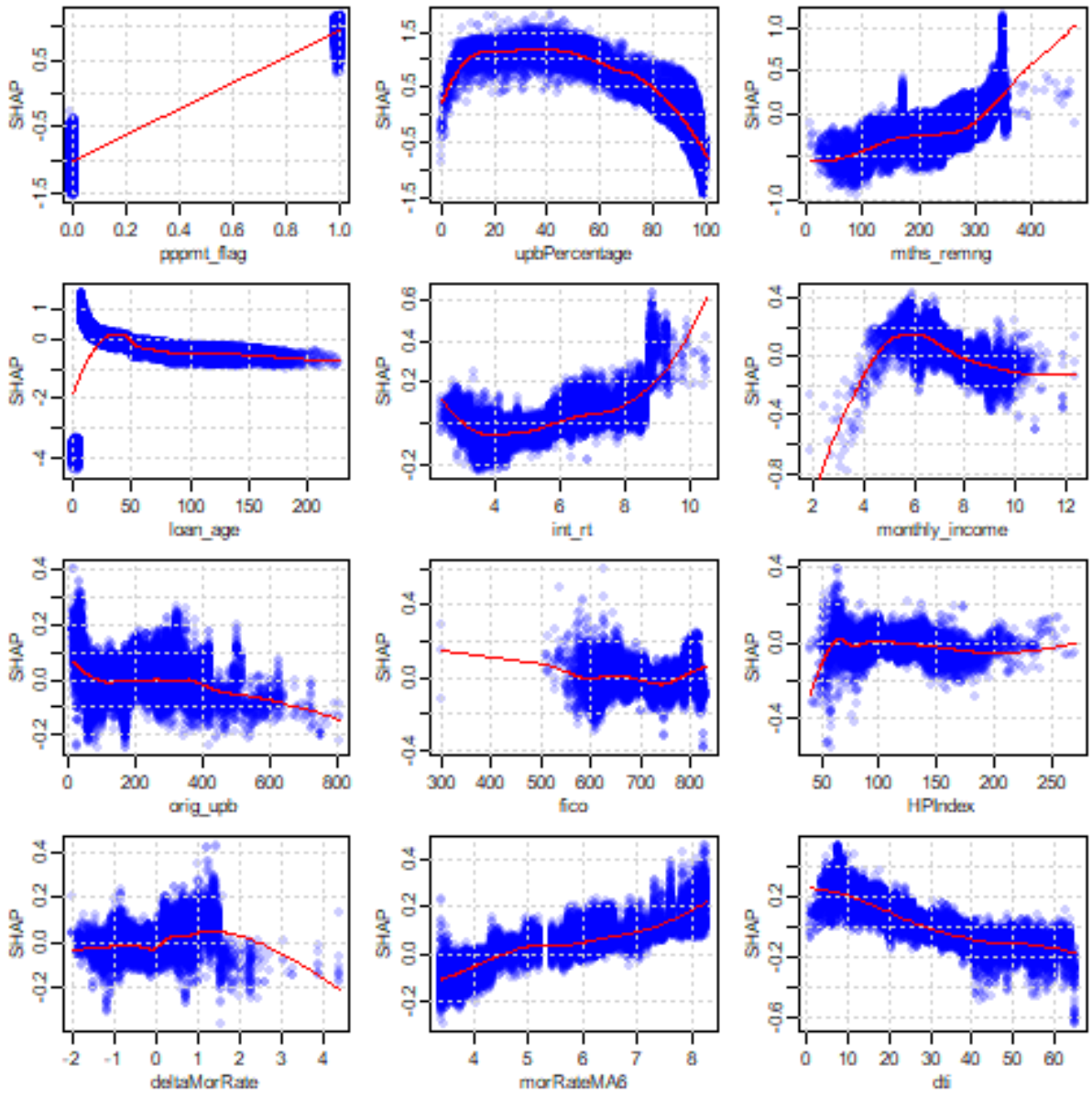



Figure 13: SHAP values of the 12 most important variables for class partial prepayment in the XGBoost model.

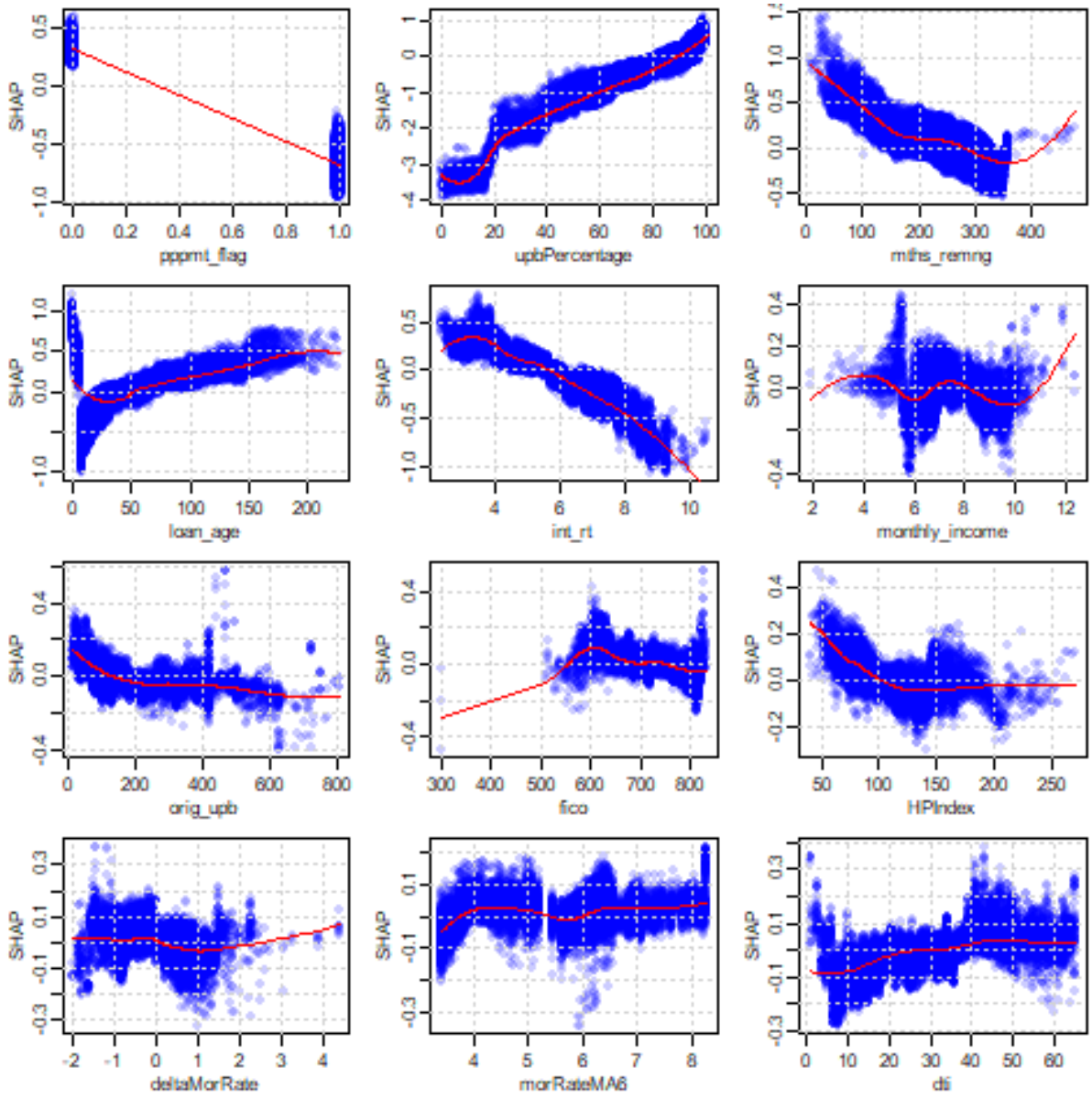



Figure 14: SHAP values of the 12 most important variables for class no event in the XGBoost model.