Masterthesis

Predicting the Probability of Default using the Random Forest Algorithm

Meikee PAGSINOHIN

Matrikelnummer: 01327477

Summary

Contents

Chapter 1

Introduction

1.1 Motivation

1.2 Structure

1.3 Previous Work

Chapter 2

Credit Risk

2.1 Credit Risk management

Part of the daily business of a financial institution is the credit risk assessment of existing and new customers. The result is used to decide if they want to decline or grant a credit application and, among other things, to set the required regulatory capital. Credit risk assessment is performed during the whole lifetime of an exposure. It starts with the approval of a transaction and is continuously monitored afterward. Corporate clients usually need to submit financial reports regularly, which are then analyzed by their bank advisor and credit analyst, while it is done automatically for retail customers via behavior scoring. The information used during the application scoring is limited because the applicant mainly provides it at the start of a new contract. It generally covers variables about their financial health, e.g., income and outstanding debt. For the behavior scoring model, internal historical data is used, for example, the borrower's payment history and credit utilization. The behavior model generally shows a better predictive performance than the application model. If a decline in financial health or behavior rating is detected, the bank may try to decrease the overdraft limit to regulate the credit risk. In the case of delayed payments, the early collection process starts, where affected customers are contacted and an alternative payment plan will be negotiated. If all interventions fail, defaulted exposures may be sold or outsourced to collection companies for further processing, like the sale of collateral. [1, p. 7]

2.2 Default Rate and Probability of Default

A critical risk measure is the probability of default (PD), which is an estimate of the likelihood of a borrower failing to pay back their financial obligations in a given time period. Depending on the analyzed portfolio, the expected number of defaults can vary. In the corporate segment, individual defaults might already be seen as an indicator of a bank's failing credit assessment process and decision. In contrast, a higher number of defaults can be expected in the retail sector. On the contrary, profit is generated if the income gained from non-defaulted customers covers the loss from the defaulted portion of the portfolio. [1, p. 2]

In the Capital Requirements Regulation (Capital Requirements Regulation, Article 178(1)), the definition of default is stated as:

> A default shall be considered to have occurred with regard to a particular obligor when either or both of the following have taken place:
>
> (a) the institution considers that the obligor is unlikely to pay its credit obligations to the institution, the parent undertaking or any of its subsidiaries in full, without recourse by the institution to actions such as realising security;

(b) the obligor is more than 90 days past due on any material credit obligation to the institution, the parent undertaking or any of its subsidiaries. Competent authorities may replace the 90 days with 180 days for exposures secured by residential property or SME commercial immovable property in the retail exposure class, as well as exposures to public sector entities. The 180 days shall not apply for the purposes of point (m) Article 36(1) or Article 127.

In the case of retail exposures, institutions may apply the definition of default laid down in points (a) and (b) of the first subparagraph at the level of an individual credit facility rather than in relation to the total obligations of a borrower.

A time period has to be defined in which a default event is observed. A common observation window is one year; an illustration is visible in Figure 2.1. The default rate per category (e.g., month, rating grade) is then calculated as the number of defaults divided by the total number of customers (Eq. 2.1). [1, pp. 20-21]

$$DR_i = \frac{d_i}{n_i} \tag{2.1}$$

where:

d_i = number of defaults in class i
n_i = number of observations in class i

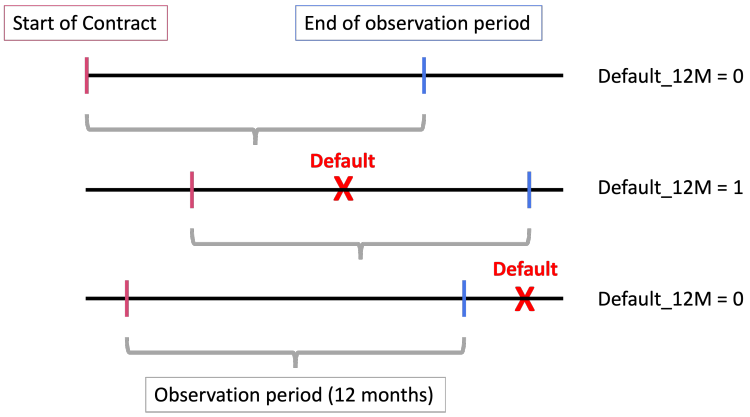

Fig. 2.1: Observation period after reference point

2.3 Regulatory Framework

2.3.1 History of Regulatory Framework

The Basel Committee on Banking Supervision sets the regulatory framework, which consists of all regulators of the most developed countries. The goal is to define a high standard for risk management and internal controls and establish a risk-sensitive calculation process of the regulatory capital for banks worldwide. The First Capital Accord was published in 1988 and has been adapted and reformed numerous times. The New Capital Accord, also known as Basel II, was first issued in 2004 and underwent multiple amendments, especially after the financial crisis until July 2009. The European Union motivated the integration of these regulations by the Implementation Directive CAD 2006. At the end of 2010, a new reform called Basel III was approved. [1, p. 13]

2.3.2 Credit Risk Regulatory Capital

The credit risk capital requirement calculation was significantly improved compared to the First Capital Accord. The total loss of a bank is split into expected and unexpected loss (Figure 2.2). The former should be covered by revenue; for the latter, a bank must allocate an appropriate level of capital. In the original approach, each exposure was assigned to one of four risk categories and then a multiplier ranging from 0-100% was applied. Regulations now allow the Standardized (SA), Foundation or Advanced Internal Rating Based (IRB-F, IRB-A) Approach. The Standard Approach defines five risk buckets for calculating regulatory capital, and it also allows the use of external ratings. For the IRB approach, internal models estimate input parameters of the regulatory formulas, which then result in risk weights for each exposure used in the calculation of the regulatory capital. The formulas are given in Eq. 2.2-2.5. The IRB-F approach only permits the estimation of the PD. In contrast, for the IRB-A approach, the risk parameters Loss Given Default, Exposure at Default, Conversion Factor and Effective Maturity are additionally derived from internal models. While the corporate segment allows for both IRB-F and IRB-A approaches, the retail portfolio is limited to the IRB-A approach. [1, 15-17] [2, p. 59]

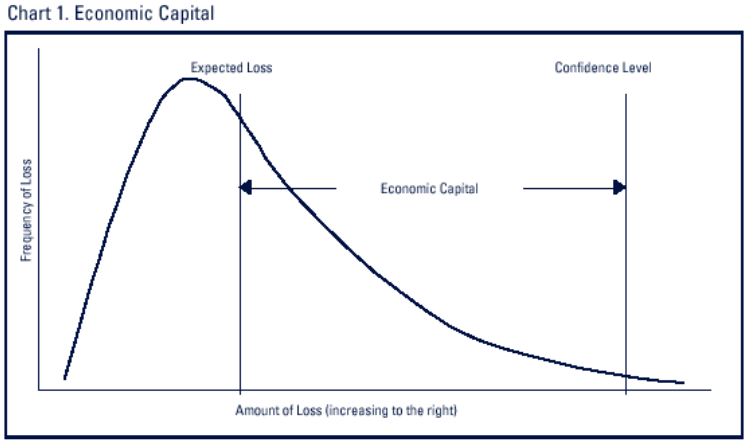

Fig. 2.2: Economical Capital, Expected and Unexpected Losses, Source: [3]

$$\rho = \rho_{min} \times \frac{1 - e^{-k \times PD}}{1 - e^{-k}} + \rho_{max} \times \left((1 - \frac{1 - e^{-k \times PD}}{1 - e^{-k}} \right) \tag{2.2}$$

$$\mathrm{b} = (0.11852 - 0.05478 \times ln(PD))^2 \tag{2.3}$$

$$\mathrm{MA} = \frac{1 + (Maturity - 2.5) \times b}{1 - 1.5 \times b} \tag{2.4}$$

$$\mathrm{K} = \left(\Phi \left[\frac{\Phi^{-1}(PD) + \sqrt{\rho} \times \Phi^{-1}(0.999)}{\sqrt{1 - \rho}} \right] - PD \right) \times LGD \times MA \tag{2.5}$$

where:

ρ = Correlation
ρ_{min}, ρ_{max} = minimum and maximum correlation per class
k = rise coefficient per class
b = Maturity adjustment
MA = Residual Maturity Adjustment Factor
K = Capital requirement

2.3.3 Challenges and Limitations

Good data is of utmost importance for the credit risk assessment. While it will be used for modeling purposes, it is also essential that already-known negative information about customers is available and considered. Examples include internal information, such as a client with a history of fraudulent activity during a credit application or credit bureau data, where negative credit information is made available for all participants. Institutions in Austria providing these kinds of information are Kreditschutzverband (KSV) and CRIF.

Other possible challenges are the constant change in economic conditions and regulatory frameworks, which influence the PD estimates. In addition, during the PD estimation process, the default events of individual borrowers are assumed to be independent, which does not adequately capture behavioral risks (e.g., strategic default) or systemic risks (e.g., market-wide shocks) that can affect multiple borrowers simultaneously. Therefore, PD models must be continuously refined and adapted to accurately reflect the current economic situation. [1, p. 8]

Chapter 3

Traditional models

To estimate a PD model, different types of models varying in complexity are available:

1. Statistical Models: This type utilizes historical data for the estimation process. Techniques such as logistic regression, survival analysis and machine learning algorithms are used to predict default events and analyze contributing risk factors.

2. External Rating Models: Rating agencies develop models that assign credit ratings to borrowers. These models consider various factors, e.g., financial statements and macroeconomic conditions, to evaluate creditworthiness. These types of PD models are only available for a limited portion of borrowers.

3. Expert Judgment: In cases where historical data is limited or only a low number of default events is available, expert judgment will become the most relevant. Experienced credit analysts rely on their expertise and industry knowledge to estimate the PD based on qualitative factors, market conditions and client information.

In practice, a substantial portion of the banking sector employs a combination of multiple types of models in their credit risk assessment.

3.1 Logit and Probit regression

Logistic regression is one of the banking industry's most commonly used statistical models. It is practical when the dependent variable is binary. The model estimates the probability of default by fitting a link function to the explanatory variables. Therefore, it transforms the resulting score, which can take any negative or positive value, to the corresponding PD value between 0 and 1. A high model score means a lower probability of default and vice versa. The logistic function or standard normal cumulative distribution function can be used for the link function, resulting in the logit or probit model. An advantage of the logit model is the heavier tails in the logistic distribution, which would put higher weights on extreme events. [1, pp. 40-42]

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

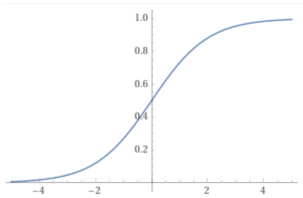

Fig. 3.1: Logistic Function

3.2 Other Models

3.2.1 Linear regression

During the linear regression, the algorithm estimates a linear relationship between the default variable, which assumes either the value 0 (non-default) or 1 (default), and explanatory variables, which can be continuous and categorical independent variables, such as income, employment duration and profession. Unfortunately, due to the binary dependent variable, the residuals are heteroscedastic; therefore, the coefficients' estimation is inefficient. Additionally, the model may output non-logical results, like negative values or a PD over 100%. [1, pp. 39-40]

3.2.2 External Ratings

Scorings of corporate clients are usually performed mainly by a credit analyst and only partly automated due to the low number of default events and the type of information available. External ratings may be used if the financial institution needs more resources to develop and maintain internal models. The most known rating agencies are Standard & Poor, Moody's and Fitch. They provide ratings for a wide range of corporations since most companies request a rating before a sale or registration of a debt issue. An analyst will use their financial statements of the last few years and additional information to derive a rating, which is then discussed in a rating committee. Afterward, the corporation is informed about its rating and the corresponding factors and given the opportunity to respond; finally, the ratings will be published. A disadvantage of external ratings observed in the past is the conflict of interest since the company mainly pays the ratings. It is suspected that good ratings were related to high fees, visible during the financial crisis, where many structured bonds with high scorings deteriorated unexpectedly. [1, pp. 34-36]

3.2.3 Shadow Rating

The Shadow Rating approach aims to estimate a model that produces similar PDs as ratings determined by external rating agencies. For this process, possible input factors, for example, macroeconomic factors and financial statements, need to be defined. The model's output serves as a valuable tool for credit analysts in making the final decisions. [1, pp. 67]

Chapter 4

Advanced Models

Stimulated by the surge in data availability and advancements in computational power, alongside the growing popularity of advanced machine learning models, financial institutions are increasingly motivated to leverage advanced models for predicting the probability of default. The objectives include enhancing model performance and uncovering previously unseen interactions among risk factors. However, challenges such as limited transparency in explaining individual predictions and constraints imposed by regulatory requirements and data protection laws, as discussed in Chapter 7.1.1, force financial institutions to be hesitant to implement these algorithms in the calculation of regulatory capital. [4, p. 4]

These advanced models can be broadly classified into three categories [5, p. 43-45]:

1. Supervised Learning: In this category, a target variable is provided and utilized during the training process. Boosting methods are employed to improve the classification of misidentified observations. Decision Trees and Random Forests are prominent examples.

2. Unsupervised Learning: Here, unlabeled data is used for training, relying solely on features to uncover underlying patterns or relationships within the data without explicit guidance. Examples include the kNN algorithm and Clustering.

3. Unsupervised Learning: In this paradigm, an "agent" learns to make decisions by interacting with an environment. It receives feedback in the form of rewards or punishments based on the actions it takes, allowing it to learn the optimal strategy to maximize cumulative rewards over time.

4.1 Decision Trees

Classification trees are used to separate the categories (default, non-default) as best as possible using explanatory factors. The split is determined by maximizing the homogeneity of the resulting subgroups, so-called branches. The algorithm calculates the measure for each possible threshold for numeric variables, while for categorical variables, it determines it for each unique value. This step is repeated until a stopping condition is met and the final subgroup is called a leaf. To avoid overfitting, the Decision Tree may be "pruned", where some branches are removed. The preliminary decision tree is applied on a separate data set, i.e., the validation sample, and to improve its performance, redundant splits are cut off. An example of a decision tree and the pruning process is visible in Figure 4.1. A Separate validation should then be performed on a third data set since the validation sample became part of the modeling process.

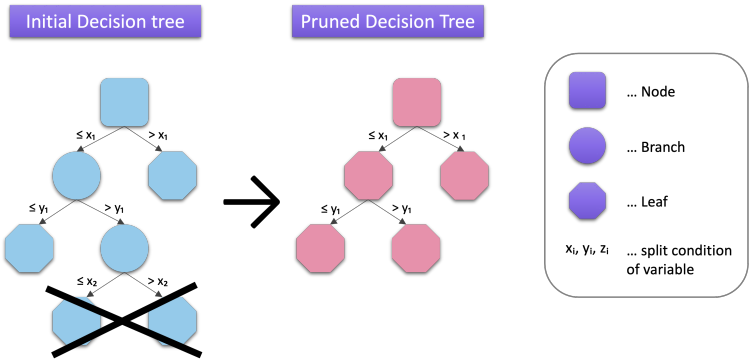

Fig. 4.1: Decision Tree incl. Pruning

The estimated PD is the average default rate per leaf and each terminal node can be classified into default or non-default class using a defined threshold. Popular statistics to measure homogeneity are the Gini index and Entropy index. The Gini index assumes a value between 0 and 1, where 0 means complete purity, 0.5 represents an equal distribution of all classes and 1 shows a random distribution across all classes. The formula is given in 4.1. Decision trees usually perform worse than logistic regression models and are rather used to assess the best variables or segmentation possibilities. [1, pp. 75-76]

$$GINI = \sum_{i=1}^{n} p_i \times (1 - p_i) \tag{4.1}$$

where:

n = number of unique classes in variable
p_i = proportion of observations in class n

4.1.1 Boosted Decision Trees and Random Forests

Boosted Decision Trees (BDT) or Gradient Boosted Decision Trees combine decision trees with boosting techniques to achieve higher predictive performance. This algorithm iteratively builds decision trees, placing more weight on misclassified observations in each iteration, resulting in a better model. BDT can capture complex interactions and non-linear relationships in PD modeling.

Random forests are an ensemble learning method that combines multiple decision trees to make predictions. However, unlike boosted decision trees, random forests build each tree independently, without sequential corrections (Figure 4.2). This approach reduces the risk of overfitting and the variance of predictions. Random forests are known for their robustness, scalability and ability to handle high-dimensional data. [1, p. 88]

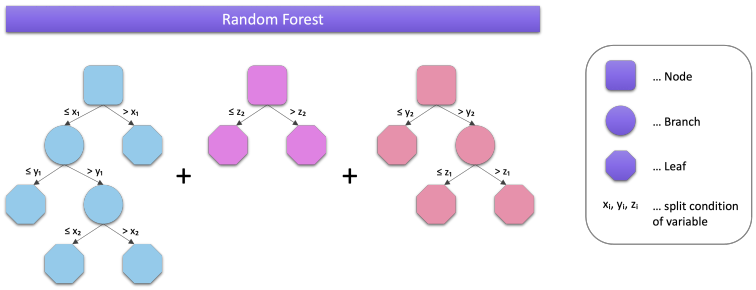

Fig. 4.2: Random Forest

4.2 Other Models

4.2.1 Neural Networks

Neural networks, inspired by the structure and function of the human brain, can learn intricate patterns and nonlinear relationships in data. They consist of multiple layers of interconnected nodes, also called neurons, where each neuron is assigned a simple computation and uses activation functions to pass along a value. The result of the model is a numerical or classification value. Commonly used activation functions are logistic, threshold and tangent hyperbolic functions.

The initial layer is referred to as the input layer, the final layer as the output layer and those in-between are known as hidden layers. An illustration is visible in Figure 4.4. Due to the virtually endless possibility of configurations, there is a possibility of over-parametrization, especially with an increasing number of hidden layers and nodes, also called deep neural networks. [1, pp. 79-80]

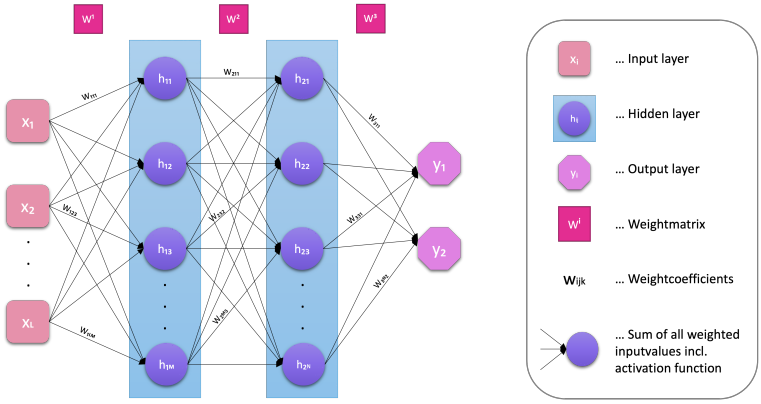

Fig. 4.3: Neural Network

4.2.2 k-Nearest Neighbour

Based on a data set with explanatory factors and observed default events, the unknown PD of a new data entry can be estimated by taking the nearest data points determined via their risk factors and calculating the average default rate of the new data point. The Euclidean metric or Manhattan Distance can be used: The Euclidean distance measures the straight-line distance between two points. In comparison, the Manhattan distance is the sum of the absolute differences between two points in a space, where it can only move along coordinate axes. The number of nearest neighbors "k" is a hyper-parameter. The advantages of this model are the simple

approach and the possibility to update new and outdated data entries dynamically and if k is set as a low number, a credit analyst can view individual scorings manually. [1, p. 83]

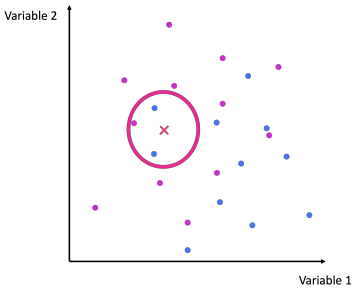

Fig. 4.4: k-nearest Neighbour with k = 3

Chapter 5

Modeling process

5.1 Data Preparation

Explanatory factors can be split into numerical, categorical and indicator variables. Within the corporate segment, these risk factors mostly take the form of numerical variables, such as financial ratios and macroeconomic conditions. The retail sector typically involves categorical factors, including profession, marital status and residential status. Even numerical values like age or employment duration often undergo a binning process, transforming them into categories (e.g., "20-25 years," "25-30 years"). If a categorical variable assumes too many distinct values or one category comprises a minimal proportion of observations (adhering to a rule of thumb of at least 5% per bucket), merging categories might be beneficial.

One effective approach is to consolidate categories with similar default rates or utilizing measures like Weight of Evidence (WoE) and Information Value (IV), as outlined in Equations 5.1 and 5.2. WoE measures the discriminatory power of each value of a risk factor - a positive WoE means a relative low risk and a negative Woe indicates a relative high risk. Meanwhile, IV assesses the whole variable's capability to distinguish between default and non-default events; a higher IV corresponds to better discriminatory power and vice versa.

As a final step, categorical variables must be transformed into dummy variables for use in the modeling process. In this context, if a variable encompasses n distinct values, n-1 dummy variables are generated, as illustrated in Figure 5.1. Omitting one dummy variable is crucial to prevent the introduction of a linear combination during the modeling process. [1, pp. 47-51]

$$\text{WoE} = \ln\left(\frac{\text{Distribution of Non-Default}}{\text{Distribution of Default}}\right) \tag{5.1}$$

$$\text{IV} = \sum_{i=1}^{n}(\text{Distribution of Non-Default} - \text{Distribution of Default}) \times WoE \tag{5.2}$$

where:

n = number of categories or buckets

Categorical Variable	Dummy_Value1	Dummy_Value2	Dummy_Value3	Dummy_Value4
Value 1	1	0	0	0
Value 2	0	1	0	0
Value 3	0	0	1	0
Value 4	0	0	0	1

Fig. 5.1: Dummy encoding

5.1.1 Missing Data Handling

A common problem in real datasets is missing information, which need to be appropriately handled during the modeling process. Popular approaches involve replacing missing values through statistical methods such as mean and median or algorithm-based imputation, such as the k-nearest neighbor Imputer. In the latter case, the missing value gets imputed with the average value of its k-nearest neighbors, a process detailed in Chapter 4.2.2. While another option is to eliminate data entries with missing information, this approach comes with potential drawbacks like information loss and bias. In the case of categorical variables, an alternative method is to treat missing information as a distinct category labeled "Missing", which removes the need for additional adaptations. [6, p. 207]

5.1.2 Erroneous Data Handling

Erroneous data originating from data entry mistakes or inconsistencies, have the potential to introduce noise and bias into the PD modeling process. Expert knowledge is crucial in identifying and rectifying such erroneous data. The best way to reduce incorrect data are control procedures implemented in the data entry systems and a data quality framework, where data validation rules are applied to identify inconsistent or illogical data. These data entries can be treated as missing information for the data preparation process.

5.1.3 Outlier Detection and Treatment

Extreme values, also called Outliers, can significantly impact the estimated PD model, making expert knowledge crucial in this domain as well. Variables resulting from ratio calculations are especially sensible to outliers, particularly when dealing with division by small numbers. A simple technique for outlier detection is the visual inspection, but this would become impractical with the increasing number of variables. To address this, quantitative approaches become invaluable, utilizing statistical measures such as Interquartile Range (IQR), box plots or Z-scores to identify outliers. A boxplot, also known as a box-and-whisker plot, is a visual representation of the distribution and the spread of the variable, with the box representing the IQR and the whiskers denoting the upper and lower limits. An example is displayed in Figure 5.2. The Z-Score, on the other hand, shows how many standard deviations a data point deviates from the mean of a dataset. After the detection, a popular method to treat outliers is winsorization, where all values above or below a certain threshold are capped to the upper and lower limit, thereby minimizing the impact of outliers on the PD model. [6, p. 250]

$$IQR = Q_3 - Q_1 \tag{5.3}$$
$$UpperLimit = Q_3 + 1.5 \times IQR \tag{5.4}$$
$$LowerLimit = Q_1 - 1.5 \times IQR \tag{5.5}$$

where:

$Q_3 = 3.\text{Quartile}$
$Q_1 = 1.\text{Quartile}$

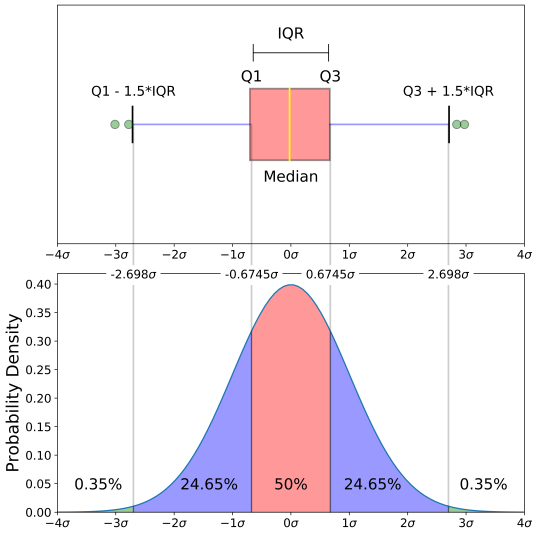

Fig. 5.2: Interquartile range and boxplot, Source: [7]

5.2 Modeling process: Logistic Regression

5.2.1 Variable selection

During the modeling process, the aim is to estimate a model that shows the best performance not only within an in-sample and out-sample data set as well. An approach that incorporates all available information into the scoring function may indeed yield a high discriminatory power within the trained sample. However, this method usually result in multiple variables exhibiting insignificant coefficients: their p-values fall below the designated confidence level, making it impossible to reject the null hypothesis asserting that the coefficient is, in fact, zero. This outcome not only widens the confidence level but also raises concerns about the accuracy of the coefficient's sign. Therefore, the model's performance on a different data set, especially on more recent data, will most likely deteriorate, leading to unstable predictions. To address this challenge, an extensive analysis of variable performance, considerate selection processes and expert knowledge becomes crucial. [1, p. 44]

5.2.1.1 Univariate Analysis

As a first step, all available variables should be considered for the modeling process. Then, an assessment of the missing rate, detection of outliers and the plausibility of values is conducted. If the variable complies with all data quality requirements, its discriminatory power is then evaluated. This evaluation can be executed using either the univariate Gini coefficient or Information value. Detailed explanations for both measures are provided in Chapters 6.2.3 and 5.1, respectively. The remaining risk factors exhibiting satisfactory discriminatory power form the basis of the long list. [1, p. 45]

Example thresholds for each measure are:

- **Missing rate**: $< 20\%$
- **Number of outliers**: $< 5\%$
- **Gini coefficient**: $> 10\%$

- **Information value**: $> 4\%$
- **Correlation coefficient**: $< 25\%$
- **Variance Inflation Factor**: 5

5.2.1.2 Multivariate Analysis

Maintaining a low level of correlation among model variables is essential to avoid issues related to multicollinearity, which can lead to unstable coefficient estimates in the modeling process. The correlation between two variables can be measured using the Pearson or Spearman correlation coefficient as detailed in Equations 5.6 and 5.7. The Pearson correlation coefficient is more appropriate, if a linear relationship is present, while the Spearman correlation coefficient excels in detecting non-linear, monotonic interactions and is more robust against outliers. The coefficient ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. After computing the coefficients for each variable pair, all values can be arranged into a correlation matrix. An example is visible in Figure 5.3. If two variables are highly correlated with a correlation coefficient above a pre-defined threshold, the variable with the lower discriminatory power should be removed from the model list. [1, p. 45]

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \hat{y}_X)(y_i - \hat{y}_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x}_X)^2 \sum_{i=1}^{n}(y_i - \hat{y}_Y)}} \tag{5.6}$$

where:

x_i, y_i = individual observations
$\hat{y}_X, \hat{y}_Y$ = sample mean of X and Y
n = number of paired observations

$$\rho_s = 1 - \frac{6 \times \sum d_i^2}{n \times (n^2 - 1)} \tag{5.7}$$

where:

d_i = difference between the ranks of corresponding observations
n = number of paired observations

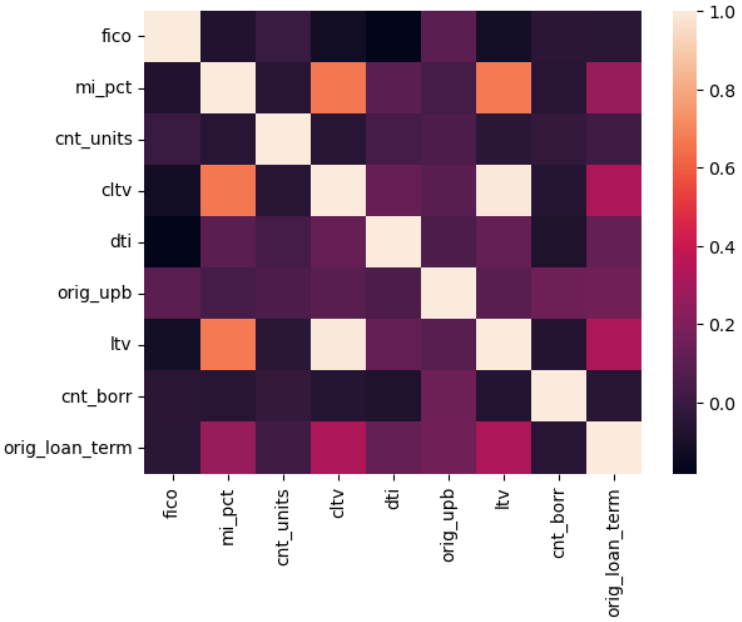

Fig. 5.3: Correlation matrix example

In the case of categorical variables, the Variance Inflation Factor (VIF, Equation 5.8) can be utilised, which measures the collinearity in the regression analysis. As a final refinement, adjustments based on expert judgment should be applied, allowing disqualified variables into the list or removing variables, even if they meet all criteria. Particular attention should be given to analyzing the relationship between explanatory factors and default rates. If the observed behavior contradicts economic rationale, the variable should be excluded from the model list, also known as the shortlist. [1, pp. 45, 53, 54]

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \tag{5.8}$$

where:

R_i^2 = coefficient of determination obtained by regressing the i-th regressor on all the other regressors

Example thresholds for the mentioned measures are:

- **Correlation coefficient**: $< 25\%$
- **Variance Inflation Factor**: 5

5.2.2 Modeling steps

Upon identifying the most promising key factors, there are different approaches to determine the final model variables, specifically, the forward, backward, forward stepwise and backward stepwise selection procedure. The forward approach starts with an empty model and step by step, the variable with the highest discriminatory power is incorporated contingent upon the significance of its coefficient (p-value below a specified threshold). This process continues until no variable satisfies the condition.

Conversely, in the backward selection procedure, the model starts with all variables and they are iteratively removed one by one based on the p-values of their coefficients, until only significant

coefficients remain. The forward stepwise procedure combines elements of both, with an initially empty model that progressively incorporates variables. However, after each addition, any variable losing significance will be removed. The backward stepwise procedure is the opposite process, starting with all variables and then eliminating those with non-significant coefficients. If, upon exclusion, any variables fulfill the significance condition once more, they are reintroduced into the model. [1, p. 45]

5.3 Modeling Process: Random Forest

5.3.1 Variable selection and Modeling steps

During the variable selection process of a normal decision tree, all available features are considered for a split. The evaluation involves assessing the information gain or Gini impurity improvement for each value of a categorical or indicator variable, as well as for every possible split of a numerical risk factor. A split is executed for the maximum improvement and this process iterates for each resulting subsegment until a stopping condition is fulfilled. Variables that have already been used for a splitting condition may be considered for a split again. Stopping conditions include achieving homogeneous subgroups, detecting no significant improvement, reaching the minimum leaf size, completing the maximum allowed splits or attaining the maximum depth of the tree. [8, pp. 2,4]

As described in Chapter 4.1, a Random Forest model is built out of individually trained smaller decision trees and the prediction is a combination of the results of all decision trees, either by calculating the average or using the majority vote. During the training process, only a fraction of all possible features may be selected for the split and optionally, a fraction of the training sample can be used to prevent overfitting. This method is known as Bootstrap Aggregating or Bagging. These steps contribute to a more robust model by avoiding to rely on only a few specific features, which leads to a better generalization. The algorithm demonstrates reduced sensitivity to outliers, as individual decision trees may be affected, but the ensemble tends to mitigate their impact on the overall model. Missing values can still contribute to training the tree by considering other available features or may be randomly removed during the bagging process. A Random Forest model is also relatively robust compared to the logistic regression, because, at each split, it selects a random subset of features and this helps in reducing the correlation between a set of features. [9]

5.3.2 Hyperparameter Tuning

Hyperparameters are pre-set configurations of a machine learning algorithm, which are not trained but actually set prior to the training process. Examples are the number of neighbors for the kNN-algorithm, the count of layers in a neural network model or the number of trees in a random forest. An optimal configuration of the parameters is crucial, because it directly impacts the models performance to distinguish between the classes as well as influences overfitting. Adjusting these parameters is an individualized task, dependent on the unique characteristics of each problem, encompassing factors like the number and type of features, data size and data quality. With increasing number of hyperparameters, the combination possibilities become extensive, making the search for the optimal configuration time-consuming. Therefore, the objective is to quickly identify a good configuration and subsequently search for potentially better settings.

Initially, a wide range of hyperparameter values is defined and the model's performance is assessed through randomly chosen combinations for a predefined number of iterations. The best combination is then utilized for the subsequent grid search. In this phase, a more constrained

range for each hyperparameter, centered around the optimized configuration, is defined. All possible combinations within this scope are evaluated and the superior outcome from both approaches forms the final model configuration. [6, p. 465] [9]

5.3.3 k-Fold Cross Validation

To further avoid overfitting, a cross-validation can be incorporated during the hyperparameter tuning process. The data set is randomly split into equally sized k subsamples, also called folds. The model with the tested hyperparameters is then trained on k-1 subsets and their performance is evaluated on the unused data set. This step is repeated k-times, with a different fold used for evaluation each time. The overall performance is calculated by averaging the metrics across all subsamples. This comprehensive process will be performed for each configuration setting during the random or grid search, contributing to a more robust model. An illustrative sketch is provided in Figure 8.1. [6, p. 470]

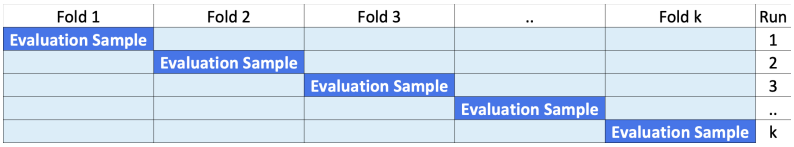

Fig. 5.4: k-Fold Cross Validation

Chapter 6

Validation

6.1 Out-of-Sample and Out-of-Time Validation

Out-of-sample validation involves evaluating the model's performance on a data set that was not used during its development. Taking it a step further, out-of-time validation introduces new data, covering a more recent time period. The objective is to assess the model's ability to make accurate predictions on unseen data. Because the model is estimated on the training sample, it is not uncommon for it to exhibit better performance on in-sample datasets compared to other samples. However, if the performance metrics differs significantly, it could signal overfitting. The full dataset is usually split into a 70% training and 30% testing sample, also referred to as validation sample. Different ratios, such as 60/40 or 80/20, are also popular choices and mainly depends on the dataset size and the number of default events. During the splitting process it is important to keep the number of default events even in both samples to prevent scenarios, where one class is disproportionally larger than the other. This method called stratification reduces the probability of a biased model training and evaluation process. [1, p. 27]

6.2 Model Performance Evaluation

6.2.1 Confusion matrix

The confusion matrix, illustrated in Figure 6.1, is a table comprising four elements, which shows the count of observations correctly (True Positive, True Negative) and incorrectly (False Positive, False Negative) identified cases. A False Positive, denoting a customer predicted to default but survived, is also called Type I Error. A False Negative, where a borrower is expected to survive but defaulted, is also known as Type II error. In practice, a Type II error holds greater severity, as the repercussions of a defaulted exposure outweigh the missed opportunity income from rejecting a non-defaulted application. [1, p. 29.30]

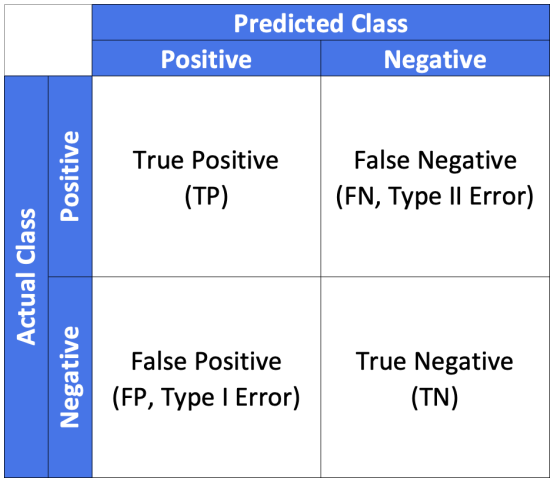

Fig. 6.1: Confusionmatrix

Using the elements of the confusion matrix, the measures Accuracy, Precision, Recall, F1-Score and others can be calculated, as illustrated in Equations 6.2 to 6.7. However, measures like Accuracy and Precision are not recommended for unbalanced data, because they can provide misleading insights about the model's performance. In instances of unbalanced data, a model might attain a high accuracy by simply predicting only the majority class. This high accuracy overshadows the model's ability to correctly identify observations of the minority class. In such scenarios, Recall and F1-score provide a more accurate evaluation of the model's performance. For the transformation from PD to a predicted default flag a threshold has to be selected. To determine an ideal cut-off, the F1-Score can be utilized, where the value is set where the F1-score attains its maximum. [10]

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negative}} \tag{6.1}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positives}} \tag{6.2}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{6.3}$$

$$\text{Negative Predictive Value} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Negative}} \tag{6.4}$$

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population}} \tag{6.5}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6.6}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.7}$$

6.2.2 Receiver Operating Characteristic Curve

The Receiver Operating Characteristic Curve (ROC Curve, see Figure 6.2) is the resulting curve after plotting the proportion of False Positive along the x-axis and proportion of True Positive along the y-axis. A diagonal line between (0,0) and (0,1) represents the random model, while the perfect model's curve would be a step function that starts at (0,0) straight up and moves horizontally to (1,1). The Area Under the ROC-Curve (AUC-ROC Curve) is, as the name suggests, the area below the ROC-curve and the formula for calculating the AUC is Equation 6.8. [10]

$$AUC = A + \frac{1}{2} \tag{6.8}$$

6.2.3 GINI coefficient

The Gini coefficient and AUC are connected via the given Equation 6.9 and a visual presentation is visible in Figure 6.2, the areas are designated as A and B. Therefore, they relate the same information but are differently scaled and the Gini coefficient shows an improved interpretability. While the AUC has a range between 0.5 (random model) to 1 (perfect model), the Gini coefficient takes on values between 0 (no discriminatory power) and 1 (perfect discriminatory power). In general, the AUC can also take on a value below 0.5, but that would indicate, that the model's predictions are less accurate than the random model, indicating an issue in the model's ability to differentiate between the classes. [10]

$$GINI = \frac{A}{A+B} = \frac{A}{\frac{1}{2}} = (2 * A + 1) - 1 = 2 * \text{AUC} - 1 \tag{6.9}$$

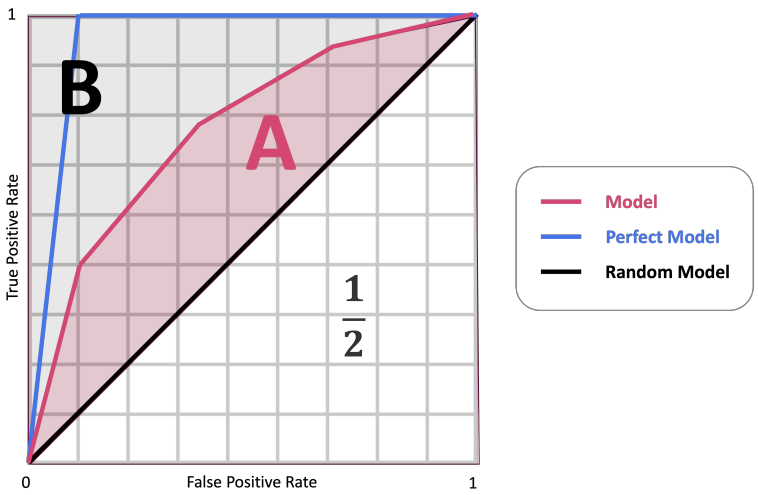

Fig. 6.2: AUC-ROC curve

6.3 Stability Test

Stability testing is performed to assess the robustness and consistency of a PD model over time. It examines whether the model's performance remains stable and reliable when applied to data collected at different time periods. Stability testing helps identifying potential model deterioration or drift over time, which may be caused by changes in the underlying credit conditions or data

characteristics. If significant discrepancies are detected, model recalibration or updates may be necessary to maintain its accuracy and relevance.

Chapter 7

Interpretability

7.1 Importance of Interpretability

Interpretability refers to the capability to explain and understand how a model arrives at its predictions or decisions. Regression models and decision trees are simple to understand and thus very popular in the banking industry. In contrary, more advanced machine learning models show a black box nature; their model logic and output are difficult to explain. Machine learning models' complex structure have advantages and disadvantages. While they can detect non-linear relationships and correlations, and may show improved accuracy or efficiency, they are also prone to overfitting and lack explainability. Their black box nature stems from the model's numerous transformation of input variables, as well as their optimization process. [5, p. 56]

7.1.1 Regulatory and Legal Requirements

Interpretability enables compliance with regulations and consumer protection laws such as the Capital Requirements Regulation (CRR) and General Data Protection Regulation (GDPR). Data protection principles such as purpose limitation, data minimisation and limitation on automated decisions are evident obstacles for complex AI models. In the CRR (Capital Requirements Regulation, Article 144(1)(a)), a requirement of the PD model development is stated as:

> (a) the institution's rating systems provide for a meaningful assessment of obligor and transaction characteristics, a meaningful differentiation of risk and accurate and consistent quantitative estimates of risk;

Regulations mandate that both model developers and users provide explanations for credit-related decisions to their customers. Modelers, along with internal and external auditors, are obligated to validate not only the model's structure but also its results, ensuring whether the model aligns with domain knowledge and expectations. Interpretability helps identifying potential biases, data issues or model limitations. Additionally, a unexplainable model used in production increases operational risk, as it becomes challenging to assess potential consequences, such as bias or fairness, and verify the accuracy of results or detect system errors. To circumvent the constraints imposed by regulatory requirements and consumer protection laws, machine learning models may find application in areas where the model's structure and output are not of utmost priority, such as in the collection process or fraud detection. [5, pp. 57, 58] [1, p. 89]

7.1.2 Data Management

Before the development or deployment of machine learning models, a sound data management process has to be established. The training data must be unbiased and accurately reflect the population the model will be deployed on, meaning that individual groups should not be over- or underrepresented. Failure to correct and validate the data utilized during the training phase

or in production can yield unexpected outcomes or result in a biased model. Machine learning algorithms have the potential to amplify errors, as popular saying goes, "Garbage In - Garbage Out." [5, p. 61]

7.2 Methods for Interpretability Analysis

Techniques to asses the interpretability of advanced models are also called model-agnostic explainability methods. They are algorithm independent, usually applied after model development and assess on global or local level, which means on dataset or data observation level. Depending on the objective, the techniques can be allocated into five categories: feature importance, input variable impact, specific prediction analysis, output analysis and robustness check. [5, p. 62]

7.2.1 Feature Importance

Feature importance measures the contribution of each variable in a predictive model to the overall model performance. If the performance drops significantly when changing the value of a variable while keeping other risk factors constant, implies the importance of that particular feature. Relative feature importance compares the importance of features relative to each other, which helps in prioritizing features based on their influence on the model's predictions. To facilitate a meaningful comparison, the ranges of each variable need to be normalized to the same scale, enabling a direct assessment of their impact. [5, p. 63]

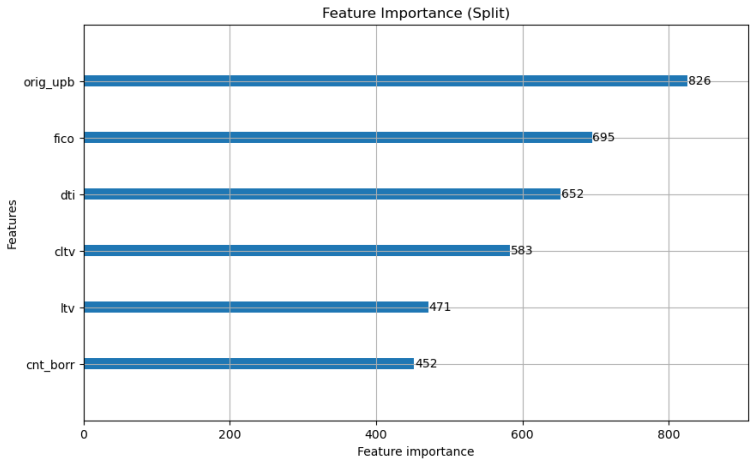

Fig. 7.1: Feature Importance

7.2.2 Input Variable Impact

Exploring the impact of individual variables is carried out through techniques like Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE), illustrated in Figure 7.2. PDP visualises the relationship between a specific feature and the model's predictions while holding other variables constant. They provide insights into the direction and magnitude of the feature's effect on default probability. ICE is an extension of PDP, where it illustrates how predictions change for an individual data point as a specific feature varies. [5, p. 63]

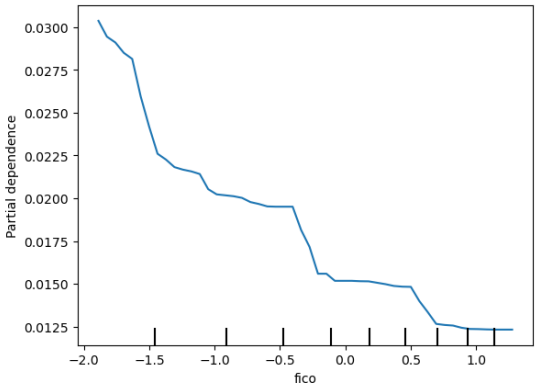

Fig. 7.2: Partial Dependence Plots

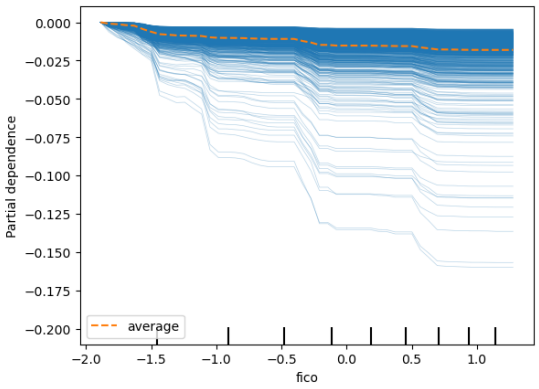

Fig. 7.3: Individual Conditional Expectation

7.2.3 Individual Prediction Analysis

For the interpretation of specific predictions, tools such as Local Interpretable Model-Agnostic Explanations (LIME) and Local rule-based explanations can be utilized. In the LIME process, a local interpretable surrogate model is estimated. A small sample with similar variable values is selected and used to estimate a sparse linear regression model while using the predictions of the machine learning models as target. Similarly, the Local rule-based explanations method builds a set of decision rules to act as a surrogate model in the interpretation process. [5, p. 65-67]

7.2.4 Output Analysis and Robustness Check

During Counterfactual analysis, the feature values are slowly changed to assess, which total changes are necessary to receive a specific prediction. Adversial testing is performed to analyze, how the machine learning model reacts to adversial attacks, which are input data deliberately designed with the aim of causing misclassification or incorrect output. Internal layers of Deep Neural Networks can be computed to detect adversial data to respond accordingly. Alternatively, adversial data can be incorporated in the development sample to include them during the training phase. In a sensitivity test, data with value ranges not captured by the training sample are used to analyze the model predictions and their performance beyond its training scope. [5, p. 65-67]

Chapter 8

Used Data and Results

8.1 Freddie Mac's Single Family Loan-Level Dataset

Freddie Mac, officially known as the Federal Home Loan Mortgage Corporation, is a government-sponsored enterprise (GSE) in the United states and plays a crucial role in the secondary mortgage market. In the primary market, individual customers secure mortgage loans from retail banks. Within the secondary mortgage market, the lender have the option to sell these mortgages to entities like government-sponsored enterprises, e.g. Freddie Mac. This practice provides liquidity to the primary market, subsequently making additional loans available to other customers. GSEs aggregate multiple mortgages into mortgage-backed securities (MBS), which are then sold to various investors, such as insurance companies and hedge funds. [11] [12]

The Single Family Loan-Level Dataset encompasses a broad range of variables and some of the essential features of this dataset include:

1. Customer Information: This category includes borrower details such as Credit Score provided by FICO, first-time homebuyer status, occupancy status and number of borrowers.

2. Financial Attributes: The dataset provides important financial indicators such as the (combined) loan-to-value ratio (CLTV, LTV) and debt-to-income ratio (DTI).

3. Loan and Property Details: It also includes information about the properties associated with the loans, such as mortgage insurance percentage, number of units, original unpaid balance, original Loan Term, property type and loan purpose.

4. Loan Performance: This section captures essential data related to historical loan performance. It includes details on delinquency status, zero balance reason codes, modification information and the current deferred unpaid balance.

5. Expenses and Loss Information: Included in the performance data are financial details such as mortgage insurance recoveries, net sale proceeds, legal costs, and actual loss, among others.

8.1.1 Data Quality, Limitations and Usage

Over the years, Freddie Mac has accumulated vast amounts of mortgage loan data. At the direction of the Federal Housing Finance Agency (FHFA), this data has been made publicly available to enhance transparency and assist investors in analyzing credit performance. Freddie Mac's Single Family Loan-Level Dataset is publicly available on their website, where users are required to register and agree to the dataset's terms of use before accessing the data.[13]

While the Single Family Loan-Level Dataset offers a wealth of information, it is essential to consider its quality and limitations. Freddie Mac explicitly states that they do not guarantee a complete or error-free data set and that it may contain potential biases.The dataset represents

mortgages acquired by Freddie Mac, which may not be fully representative of the entire mortgage market. Furthermore, the dataset is subject to data protection regulations leading to the anonymization or removal of personally identifiable information. The data set is updated on a quarterly basis and thus, contain corrections or updates that may impact the analysis.[14]

8.2 Dataset

The Freddie Mac Dataset consists of the origination and monthly performance data file. The former contains information about the borrower and the mortgage loan collected at the start of the contract. The latter includes monthly snapshots of the mortgage loan's payment, status and loss history. The data preparation and modeling process focuses on the origination data, as the objective is creating an application scoring model. The monthly performance data will be used to approximate a default flag, described in Chapter 8.2.1. The provided data files cover all months since January 1999 and undergo continuous updates every quarter. On average, each month contains about 157.000 mortgage loans, with the highest number of mortgage loans opened in September 2003 with 577.000 accounts. Table 8.1 shows the full list of relevant variables, their description and abbreviation:

Variable Name	**Description**	**Abbr.**
Credit Score	A score from an external source (FICO), indicating the borrower's creditworthiness. The higher the score, the lower the probability of default. Value ranges between 300 and 850 or a value of 9999 will be set.	Credit Score
First Time Homebuyer Flag	Variable is set to 'Y' if borrower purchased the mortgaged property to use as a primary residence and had no ownership in a different property in preceeding three years before purchase.	Homebuyer Flag
Mortgage Insurance Percentage (MI %)	Percentage of loss coverage on the loan, that a mortgage insurer covers after a default. Value ranges between 1% and 55% or a value of 999 will be set.	MI Perc
Number of Units	Number of units in property. Value ranges between 1 and 4 or a value of 99 will be set.	No Units
Occupancy Status	Contains values "Primary Residence", "Investment Property", "Second Home", "Not Available"	Occupancy
Original Combined Loan-to-Value (CLTV)	Ratio: (Original mortgage loan amount + Secondary mortgage loan amount if available) divided by the mortgaged property's appraised value. Value ranges between 1% and 998% or a value of 999 will be set. If the CLTV is lower than CTV, then the value was set to 999.	CLTV
Original Debt-to-Income (DTI) Ratio	Ratio: (Monthly debt payments + housing expenses) divided by (monthly income). Value ranges between 0% and 65% or a value of 999 will be set.	DTI

Original UPB	Unpaid principal balance rounded to the nearest 1.000	UPB
Original Loan-to-Value (LTV)	Ratio: Original mortgage loan amount divided by lesser of the mortgaged property's appraised value. Value ranges between 1% and 998% or a value of 999 will be set.	LTV
Channel	Contains values "Retail", "Broker", "Correspondent", "TPO Not Specified", "Not Available"	Channel
Prepayment Penalty Mortgage (PPM) Flag	Variable is set to 'Y' if borrower is or was obligated to pay a penalty in the event of certain repayments of principal.	PPM Flag
Amortization Type (Formerly Product Type)	Contains values "Fixed Rate Mortgage", "Adjustable Rate Mortgage"	Amort Type
Property State	Two letter statecode of property	State
Property Type	Contains values "Condo", "PUD", "Manufactured Housing", "Single-Family", "Co-op", "Not Available"	Prop Type
Loan Purpose	Contains values "Purchase", "Refinance - Cash Out", "Refinance - No Cash Out", "Not Available"	Loan Purpose
Original Loan Term	Number of scheduled monthly payments.	Loan Term
Number of Borrowers	Number of borrowers obligated to repay the mortgage. Value ranges between 1 and 10 or a value of 99 will be set.	No Borrowers
Super Conforming Flag	Variable is set to 'Y' if mortgage loan exceed conforming loan limits.	Sup Conf Flag
Program Indicator	Contains values "Home Possible", "HFA Advantage", "Refi Possible", "Not Available", "Not Applicable"	Prog Flag
HARP Indicator	Variable is set to 'Y' if loan is part of Freddie Mac's Relief Refinance Program	HARP Flag
Property Valuation Method	Contains values "Relief Refinance Loan", "Non-Relief Refinance loan"	Prop Val Method
Interest Only (I/O) Indicator	Variable is set to 'Y' if loan only requires interest payments at the beginning of contract.	Int Only Flag
Current Loan Delinquency Status	Number of days the borrower is delinquent and calculated under the Mortgage Bankers Association (MBA) method	Delinquency Status
Zero Balance Code	Reason, why the loan's balance was reduced to zero; Contains values "Prepaid or Matured (Voluntary Payoff)", "Third Party Sale", "Short Sale or Charge Off", "Repurchase prior to Property Disposition", "REO Disposition", "Whole Loan sales", "Reperforming sales securitizaitons"	Zero Balance Code

Tab. 8.1: Description of variables

8.2.1 Approximation of default flag

Since the dataset does not directly contain default information, an approximation for the indicator needs to be created. This information is derived from the performance data of the mortgage loan. As a first step, the number of months between the date of the first payment and the date of being in delinquency continuously for 30/60/90/120/180 days was calculated. To imitate the definition of default described in the CRR 178(1) a (Chapter 2.2) as closely as possible, the 90 days delinquency information was selected for further analysis. Due to data irregularities, where the 120 or 180 DPD field is filled in, but the 90 DPD is missing, the minimum of all three variables was used for the next steps. Additionally, to fulfill the definition of default stated in the CRR 178(1) b, the variable *Zero Balance Code* was incorporated. It contains the reason why the loan balance was reduced to zero, displayed in Table 8.2. Therefore, Zero Balance Codes 02, 03, 09 and 15 indicate a negative financial health and were considered in the default approximation.

Zero Balance Code	**Description**
01	Prepaid or Matured (Voluntary Payoff)
02	Third Party Sale
03	Short Sale or Charge Off
96	Repurchase prior to Property Disposition
09	REO Disposition
15	Whole Loan sales
16	Reperforming sales securitizaitons

Tab. 8.2: Description of Zero Balance Code

In the modeling process, a 12-month time span was chosen. Following the identification of default events, the default flag was set based on the following conditions:

- Customer was in delinquency for at least 90 days continuously during the first 12 months in the books.
- Loan balance showed a negative behavior in the *Zero Balance Code* during the first 12 months in the books.

8.3 Sample Creation

8.3.1 Data Exclusions

The first two months of the whole data set showed an unusually low number of observations and default events, leading to their exclusion. Given the 12-month observation period for the default flag, the final 12 months of the dataset were also omitted. Additionally, accounts that were prepaid before the 12-month observation period ended were removed. Lastly, mortgage loans without monthly performance data were also deleted because it was then impossible to approximate a default flag. The breakdown of exclusions based on each reason is detailed in Table 8.3.

Reason	Number of data entries
Remove first 2 months due to unusual low number	5047
Less than 12 months (Last Year)	2127828
Less than 12 months and prepaid	4452984
Missing Monthly Performance data	1484

Tab. 8.3: Number of exclusions

8.3.2 Training, Test and Validation Sample

Figure 8.1 shows the number of observations as well as the default rate per month for the whole sample before and after data exclusions and Figure 8.2 is a separate display of the development and out-of-time sample. If the number of defaults had not been sufficient, an increase in the observation window might have been a solution. Both figures show a satisfying number of monthly defaults with an average default rate of 1,52% in the development sample. A plausible development of the default rate is visible; it follows the expected increase of defaults during the Dot-Com crisis in the late 1990s, the financial crisis in 2007/2008 and the COVID-19 crisis in 2020/2021.

The development sample encompasses data from January 2018 to December 2020, offering a timeframe that covers different economic conditions, including the period before and during the COVID-19 crisis. This selection not only captures diverse economic scenarios but also limits the number of observations due to the limitation of computational power. The out-of-time sample, referred to as the validation sample, is constructed using data points from January to December 2021.

The data preparation and univariate analysis were performed on the whole development data set, then split into 70% training and 30% test samples. The split was stratified on the default flag and the year to ensure a balanced data set for the multivariate analysis and the modeling process of both modeling approaches. Table 8.4 shows the sample sizes and default rates. Complete lists of the number of accounts and defaults per month for the whole sample before and after exclusion are given in Appendix B.

Sample	# Accounts	# Def	% Def Rate
Whole Dataset	39.404.062	208.008	0,53%
Training	3.789.579	57.510	1,52%
Test	1.624.101	24.645	1,52%
Out Of Time	4.121.718	16.848	0,41%

Tab. 8.4: Development and Validation sample

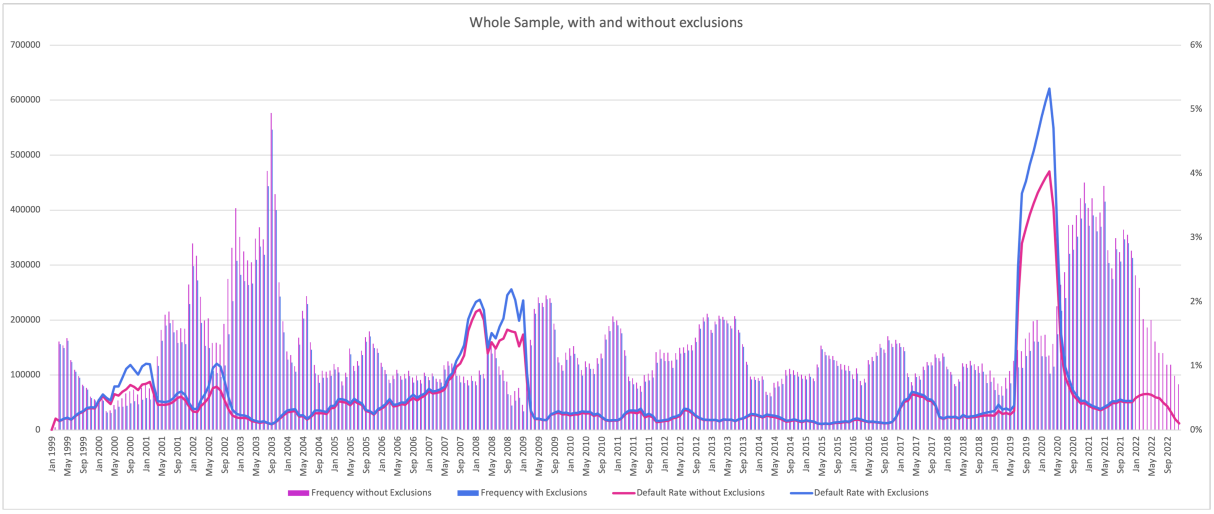

Fig. 8.1: Distribution and default rate of whole sample

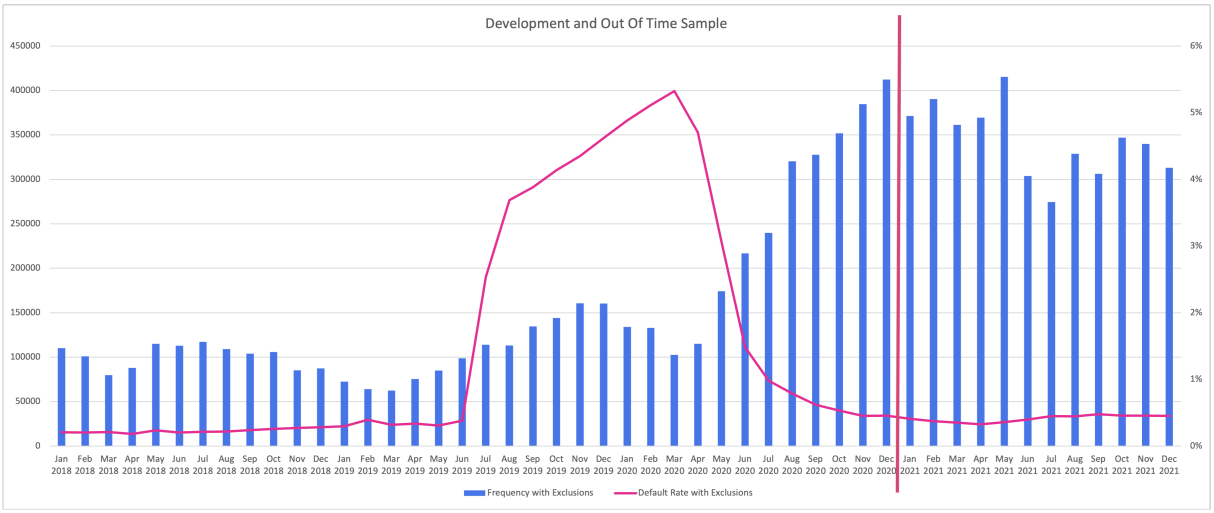

Fig. 8.2: Distribution and default rate of development and validation sample

8.4 Data preparation

8.4.1 Missing and Erroneous Data Treatment

The variables were first split into three types: categorical, indicator/binary and numerical variables. The missing rate of all variables were determined and are given in Table 8.8. *Program Indicator*, *HARP Indicator* and *Super Conforming Flag* exhibit a high proportion of missing values, exceeding 90%, making them unsuitable for inclusion in the model. On the other hand, all other risk factors either have no missing values or possess an acceptable rate of less than 20%. Given the low number of missing values for numerical variables such as *DTI* and *Credit Score* and others with a missing rate below the third decimal, imputation was performed using the median, stated in Table 8.5. Missing data points for the variable *Property Valuation Method* and other categorical risk factors were treated as a distinct category, while missing values for indicator variables were assigned a value of 'Y'.

No treatment of erroneous data was undertaken, as data entries outside pre-defined ranges were already set as 'Not Available' or '999' by Freddie Mac. Consequently, further analysis for potential errors was deemed unnecessary. Figure 8.3 to 8.6 show the distribution plots of relevant risk factors, all plots are given in Appendix C.1.

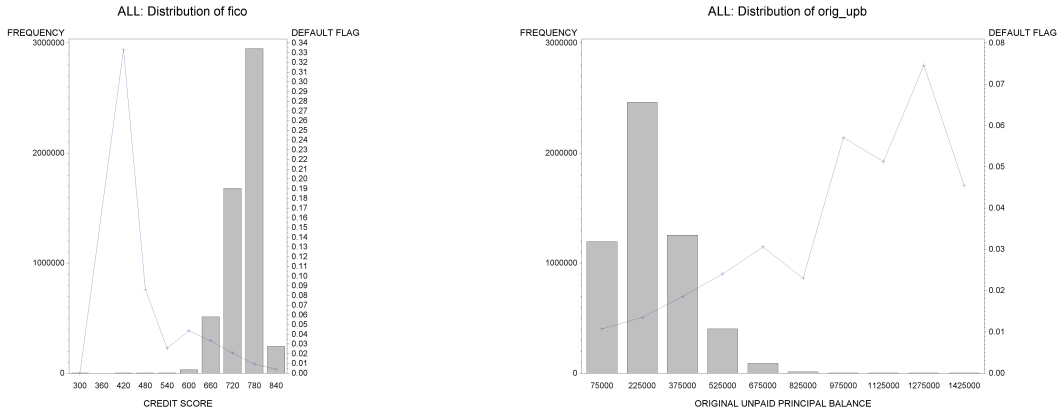

Fig. 8.3: Distribution of Credit Score (fico) and Original UPB

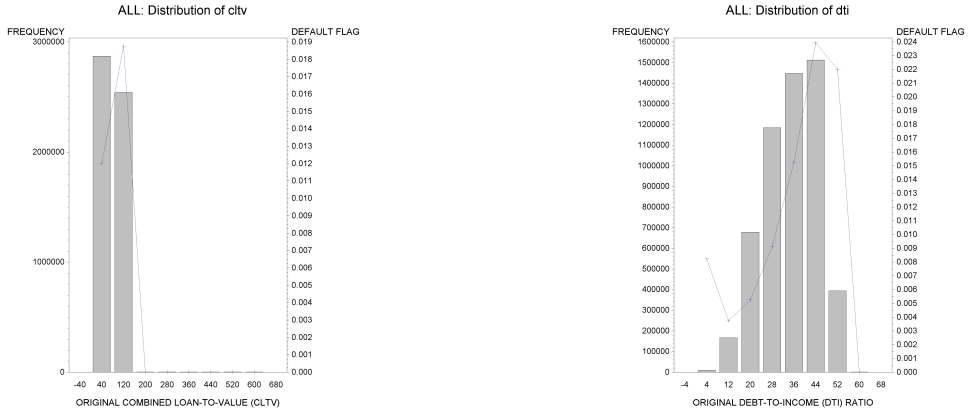

Fig. 8.4: Distribution of CLTV and DTI

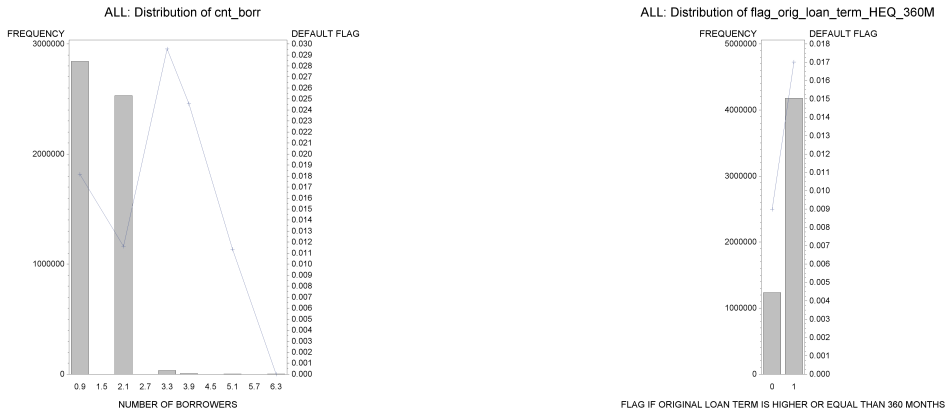

Fig. 8.5: Distribution of Number of borrowers and Flag Original Loan Term $\geq$ 360 months

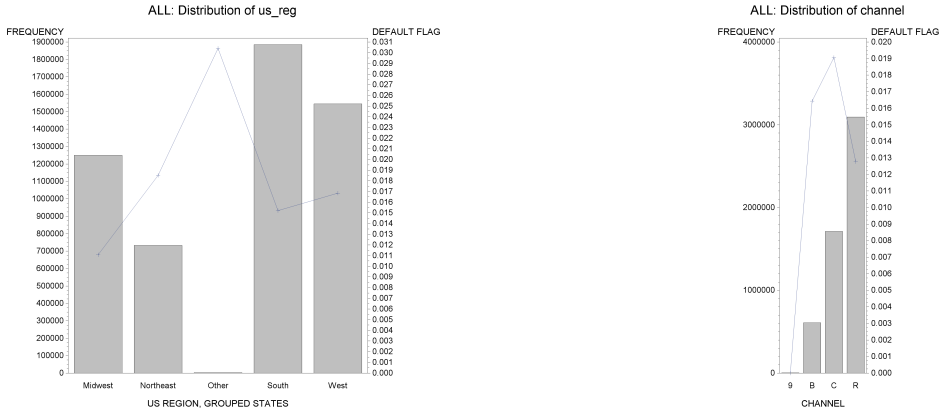

Fig. 8.6: Distribution of US region and Channel

8.4.2 Outlier Treatment

The interquartile approach, detailed in Chapter 5.1.3, was employed to identify outliers and their presentation was visualized through boxplots (Figure 8.7 - 8.9, all plots in Appendix C.2). Upper and lower boundaries were determined using Equations 5.4 and 5.4. Quartiles for all numerical risk factors are outlined in Table 8.5, with resulting limits available in Table 8.6. The proportion of upper and lower outliers were calculated to analyze, if a significant amount of data entries are affected and could potentially impact the modeling process. While outliers were present in all numerical variables, only the risk factor *Original Loan Term* shows a concerning proportion. New risk factors were derived to circumvent this issue. Multiple versions of indicator variables were created using the definition listed in Table 8.7. Additionally, after analyzing the distribution and box plots of all risk factors, a winsorization was executed to evaluate, if outliers affected the discriminatory power negatively. Therefore, data points with values above the upper or below the lower limit, were capped at their respective thresholds. The performance of the adjusted variables did not change significantly and thus, the original versions were retained.

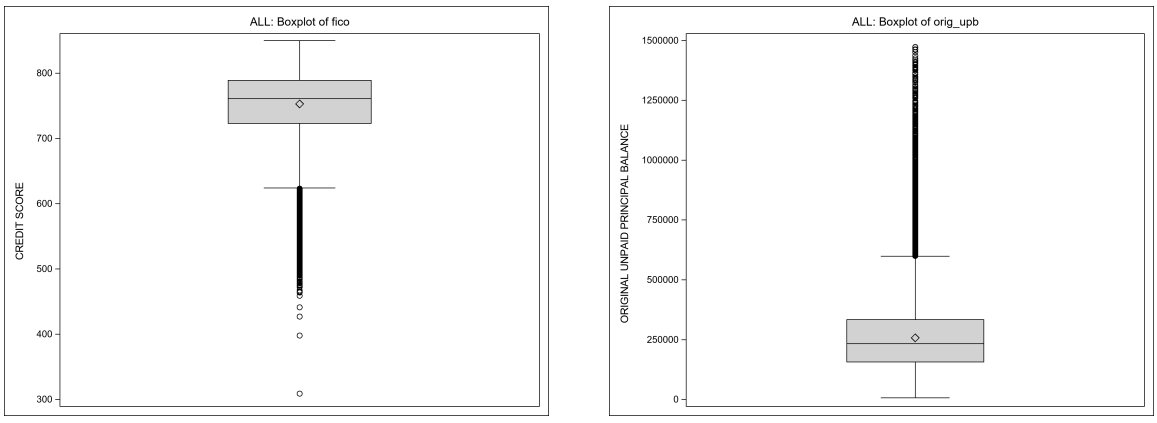

Fig. 8.7: Boxplot of Credit Score (fico) and Original UPB

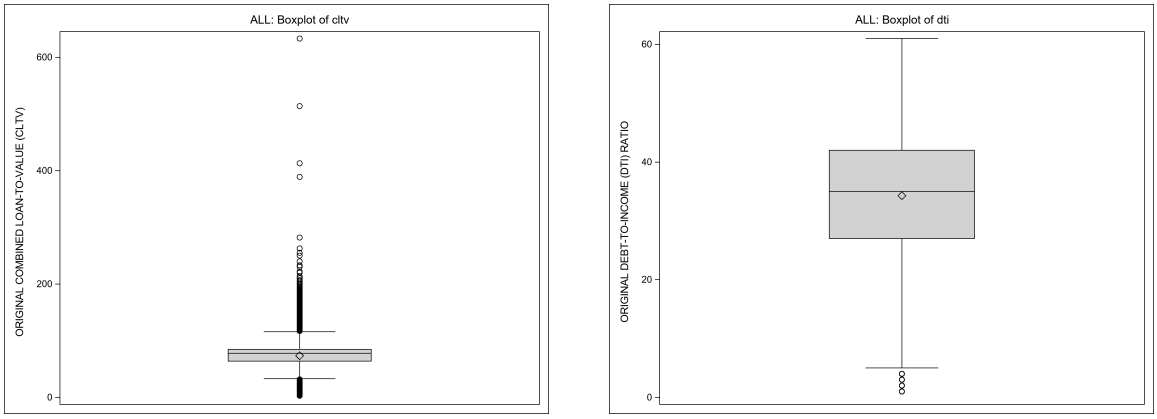

Fig. 8.8: Boxplot of CLTV and DTI

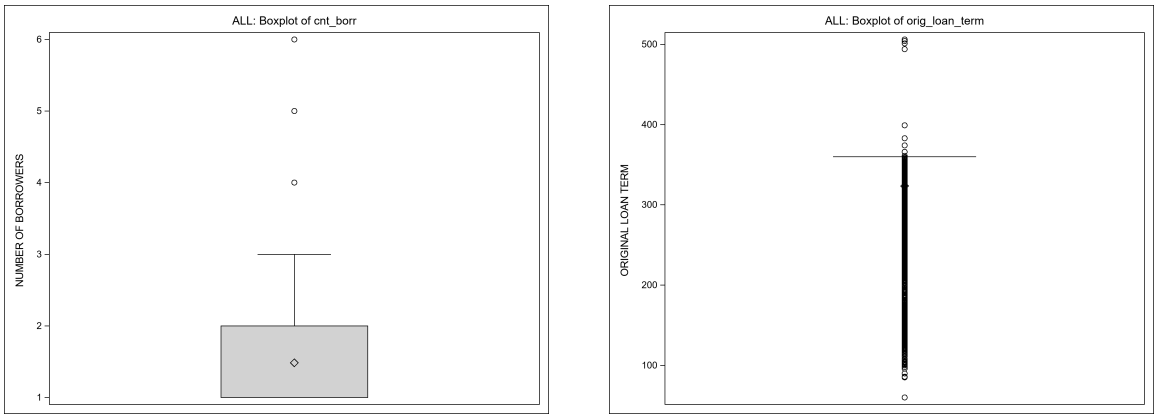

Fig. 8.9: Boxplot of Numner of Borrowers and Original Loan Term

Variable	Sum	Mean	Mode	StdDev	Min	P1	P5	P25	Median	P75	P95	P99	Max
CLTV	398.545.357	73,62	80	17,27	3	25	39	64	78	85	95	97	633
No Borrowers	8.030.443	1,48	1	0,52	1	1	1	1	1	2	2	2	6
No Units	5.566.001	1,03	1	0,22	1	1	1	1	1	1	1	2	4
DTI	184.916.404	34,28	45	9,70	1	12	17	27	35	42	48	50	61
Credit Score	4.074.892.728	752,92	801	43,93	309	637	670	723	761	789	809	817	850
LTV	397.336.584	73,40	80	17,29	3	24	39	64	77	85	95	97	514
MI Perc	37.410.176	6,91	0	11,56	0	0	0	0	0	12	30	30	52
Loan Term	1.751.503.811	323,53	360	69,96	60	180	180	360	360	360	360	360	506
UPB	1.390.864.618.000	256916,67	200.000	131963,97	7.000	54.000	85.000	157.000	233.000	334.000	502.000	663.000	1.473.000

Tab. 8.5: Descriptive statistics

Variable	Lower Boarder	Upper Boarder	# below Lower Boarder	# above Upper Boarder	% below Lower Boarder	% above Upper Boarder	# Outliers	% Outliers
CLTV	32,5	116,5	142518	596	2,63%	0,01%	143.114	2,64%
No Borrowers	-0,5	3,5	0	5461	0,00%	0,10%	5.461	0,10%
No Units	1	1	0	108940	0,00%	2,01%	108.940	2,01%
DTI	4,5	64,5	21294	0	0,39%	0,00%	21.294	0,39%
Credit Score	624	888	16444	0	0,30%	0,00%	16.444	0,30%
LTV	32,5	116,5	145323	394	2,68%	0,01%	145.717	2,69%
MI Perc	-18	30	19	25853	0,00%	0,48%	25.872	0,48%
Loan Term	360	360	1236041	18	22,83%	0,00%	1.236.059	22,83%
UPB	-108500	599500	0	102141	0,00%	1,89%	102.141	1,89%

Tab. 8.6: Interquartile range

8.5 Variable Selection

8.5.1 Univariate Analysis

8.5.1.1 New variables

During the outlier treatment, a few new variables were created. While analyzing the different distinct values of categorical variables, the risk factor *Property states* was grouped into five *US regions* according to their geographical position: Northeast, Southeast, Southwest, Midwest, West and other Regions, e.g., outside of the North American mainland. Table 8.7 lists a summary of the new variables.

Variable Name	**Description**	**Abbr.**
MI Percentage Indicator	Indicator, that Mortgage Insurance Percentage >0%	MI Flag
Loan Term >360 Months	Indicator, that Original Loan Term >360 Months	Loan Term >360m
Loan Term Group	Grouped Variable, Original Loan Term is "<360m", "= 360m", ">360m"	Loan Term Group
Loan Term ≥ 360 Months Indicator	Indicator, that Original Loan Term ≥ 360 Months	Loan Term ≥ 360m
Indicator, that Original Loan Term = 360 Months	Indicator, that Original Loan Term = 360 Months	Loan Term = 360m
Original Combined Loan-to-Value (CLTV) after Outlier Treatment	Original Combined Loan-to-Value (CLTV) after Outlier Treatment	CLTV adj
Original Loan-to-Value (LTV) after Outlier Treatment	Original Loan-to-Value (LTV) after Outlier Treatment	LTV adj
US Region	Grouped variable of "Property State"	US Region

Tab. 8.7: Description of new variables

8.5.1.2 Discriminatory Power

To assess the discriminatory power, distribution plots with default rates, ROC curves and calculations of AUC, GINI coefficients, WoE and IV were conducted. A list of all metrics is presented in Table 8.8 and relevant variable plots are depicted in Figures 8.10 to 8.13. All plots are available in Appendix C.1 and C.3. As expected, the external *Credit Score* provided by FICO shows the highest discriminatory power, followed by financial ratios such as *DTI*, *LTV* and *CLTV*. Other numerical risk factors, except *No Units*, also indicate good predictive power, making them optimal candidates for the modeling process. In contrast, the performance of categorical and indicator risk factors proved underwhelming, with few GINI coefficients surpassing 5% or IV exceeding 3%. Consequently, a GINI threshold of 5% and IV of 3% was established for the selection process of the long list.

To summarize, all variables with a missing proportion below 20%, outlier proportion not succeeding 20%, GINI coefficient above 5% and IV higher than 3% were selected for the long list given in Tab. 8.11.

Variable	Value	% Missing	AUC	GINI	WoE	IV
Amort Type	Fixed Rate Mortgage	0,00%	0,5000	0,0000	0,0000	0,0000
Prop Val Method	ACE Loans	0,07%	0,5402	0,0805	0,4326	0,0486
	Full Appraisal	0,07%	0,5445	0,0890	-0,1126	0,0486
	Other Appraisals (Desktop, driveby, external, AVM)	0,07%	0,5042	0,0084	0,4429	0,0486
	Not Available	0,07%	0,5001	0,0001	0,2385	0,0486
Channel	Not Available	0,00%	0,5000	0,0000	0,0000	0,0358
	Broker	0,00%	0,5047	0,0095	-0,0811	0,0358
	Correspondent	0,00%	0,5411	0,0822	-0,2318	0,0358
	Retail	0,00%	0,5458	0,0917	0,1744	0,0358
Prog Flag	Not Available or Not Applicable	92,48%	0,5178	0,0355	0,0391	0,0176
	HFA Advantage	92,48%	0,5039	0,0078	-0,8763	0,0176
	Home Possible	92,48%	0,5138	0,0277	-0,3369	0,0176
Loan Purpose	Refinance - Cash Out	0,00%	0,5144	0,0289	-0,1380	0,0080
	Refinance - No Cash Out	0,00%	0,5179	0,0358	0,1096	0,0080
	Purchase	0,00%	0,5035	0,0069	-0,0149	0,0080
Occupancy	Investment Property	0,00%	0,5027	0,0053	-0,0838	0,0055
	Primary Residence	0,00%	0,5037	0,0074	-0,0082	0,0055
	Second Home	0,00%	0,5063	0,0127	0,3945	0,0055
Loan Term Group	= 360 Months	0,00%	0,5473	0,0946	-0,1158	0,0615
	> 360 Months	0,00%	0,5000	0,0001	-3,4796	0,0615
	< 360 Months	0,00%	0,5473	0,0947	0,5310	0,0615
Prop Type	Condo	0,00%	0,5024	0,0048	-0,0582	0,0005
	Co-op	0,00%	0,5002	0,0003	0,2712	0,0005
	Manufactured Housing	0,00%	0,5003	0,0005	0,1516	0,0005
	PUD	0,00%	0,5010	0,0020	-0,0073	0,0005
	Single-Family	0,00%	0,5029	0,0059	0,0093	0,0005
US Region	Midwest	0,00%	0,5317	0,0634	0,3194	0,0300
	Northeast	0,00%	0,5151	0,0301	-0,2016	0,0300
	Other	0,00%	0,5002	0,0004	-0,7108	0,0300
	South	0,00%	0,5004	0,0009	-0,0025	0,0300
	West	0,00%	0,5160	0,0320	-0,1064	0,0300
Homebuyer Flag		0,00%	0,5167	0,0333	0,0000	0,0068
Int Only Flag		0,00%	0,5000	0,0000	0,0000	0,0000
MI Flag		0,00%	0,5531	0,1062	0,0000	0,0510
Loan Term = 360m		0,00%	0,5473	0,0946	0,0000	0,0611
Loan Term $\geq$ 360m		0,00%	0,5473	0,0947	0,0000	0,0612
Loan Term >360m		0,00%	0,5000	0,0001	0,0000	0,0002
Sup Conf Flag		96,51%	0,5000	0,0000	0,0000	0,0158
HARP Flag		99,66%	0,5000	0,0000	0,0000	0,0013
PPM Flag		0,00%	0,5000	0,0000	0,0000	0,0000
CLTV		0,00%	0,5872	0,1744	0,0000	0,1087
No Borrowers		0,00%	0,5504	0,1007	0,0000	0,0537
No Units		0,00%	0,5078	0,0156	0,0000	0,0092
DTI		0,35%	0,6354	0,2709	0,0000	0,2709
Credit Score		0,03%	0,6647	0,3294	0,0000	0,3467
LTV		0,00%	0,5854	0,1708	0,0000	0,1051
MI Perc		0,00%	0,5546	0,1091	0,0000	0,0543
Loan Term		0,00%	0,5487	0,0974	0,0000	0,0661
UPB		0,00%	0,5792	0,1583	0,0000	0,0778

Tab. 8.8: Discriminatory power

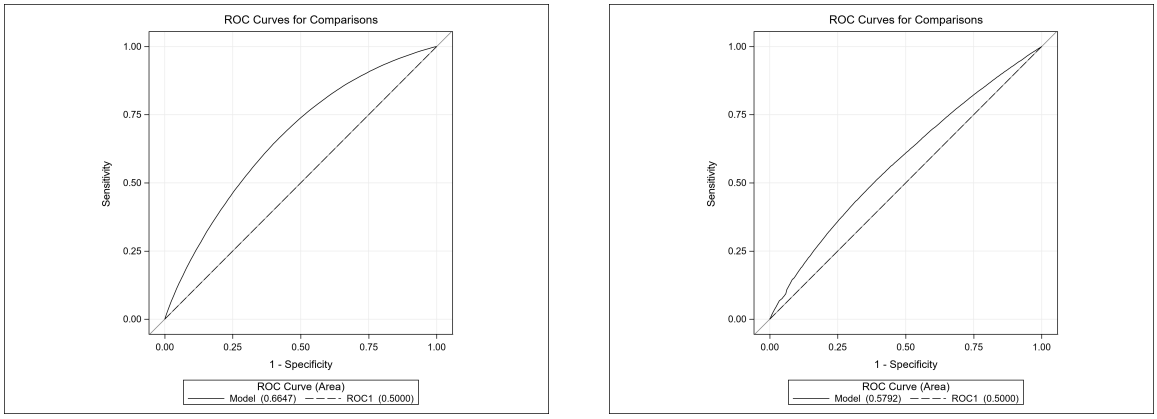

Fig. 8.10: ROC-curve of Credit Score (fico) and Original UPB

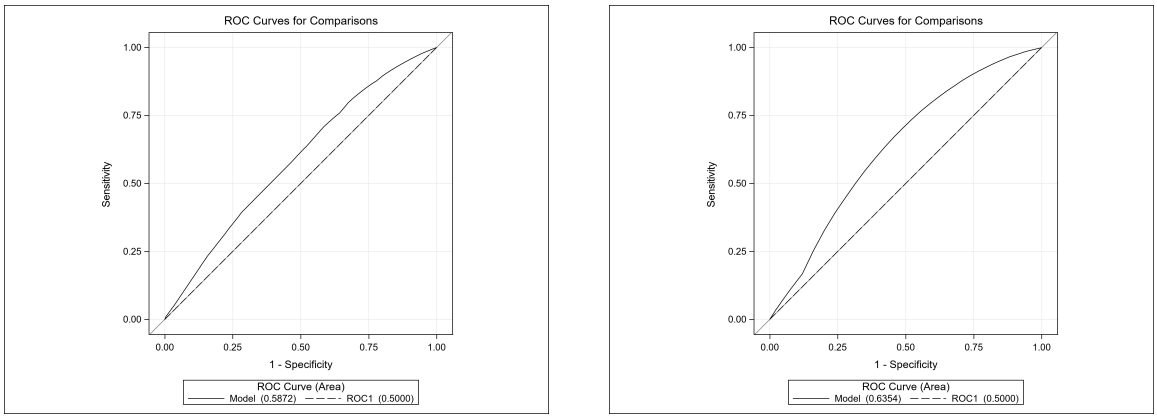

Fig. 8.11: ROC-curve of CLTV and DTI

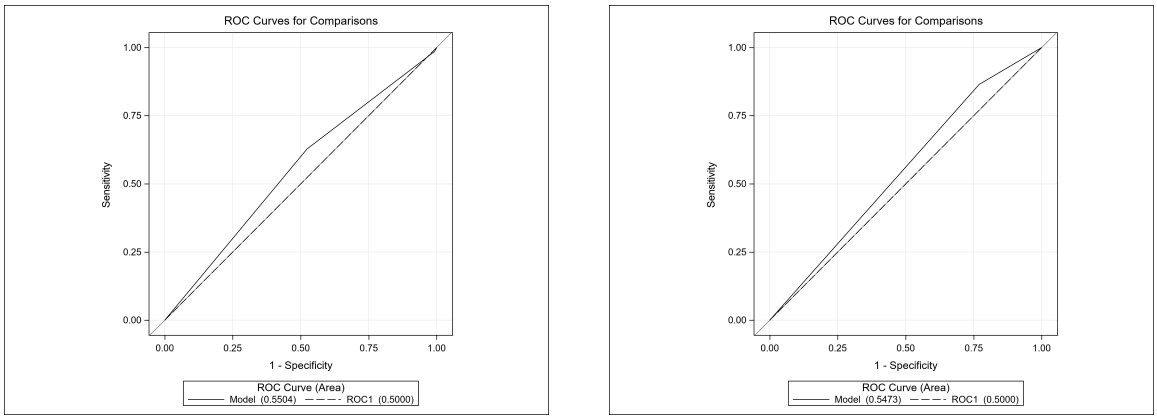

Fig. 8.12: ROC-curve of Number of Borrowers and Original Loan Term $\geq$ 360 months

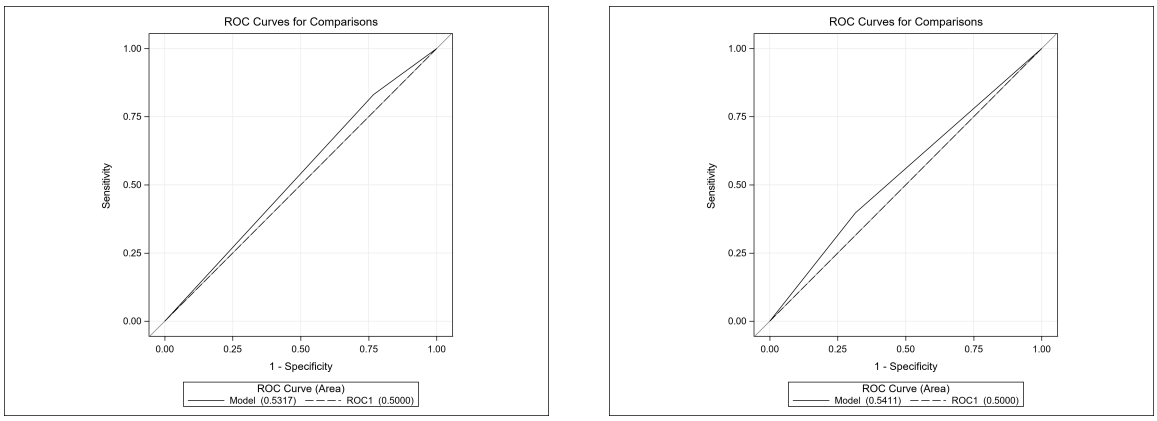

Fig. 8.13: ROC-curve of US region = Midwest and Channel = Correspondent

8.5.2 Multivariate Analysis

The shortlist was created using the following process: First, the correlation matrix with all numerical variables was created. Then, the variable with the highest GINI coefficient was identified and all risk factors with a correlation coefficient above +0.25 or below -0.25 were removed. This step should be repeated with the remaining variables until all features were analyzed. Table 8.9 shows the correlation coefficient for all numeric variables in the long list. To account for potential interaction effects between categorical and indicator variables, the Variance Inflation Factor (VIF) was considered. A high correlation coefficient between *LTV* and *CLTV* is expected and a higher correlation is also visible between *MI Perc*, *Loan Term*, *LTV* and *CLTV*. Executing the described process, *CLTV* was added to the shortlist, while the remaining correlated risk factors were excluded. The VIF values, as shown in Table 8.10, do not indicate severe multicollinearity. In the final step, instead of using the raw variable, the indicator version *Loan Term* $\geq$ *360m* was chosen. The short list is given in Table 8.11.

Variable	Credit Score	MI Perc	No Units	CLTV	DTI	UPB	LTV	Loan Term	No Borrowers
Credit Score	1,00	-0,08	0,00	-0,12	-0,18	0,10	-0,12	-0,05	-0,04
MI Perc	-0,08	1,00	-0,06	0,67	0,10	0,03	0,67	0,26	-0,05
No Units	0,00	-0,06	1,00	-0,05	0,04	0,06	-0,05	0,02	-0,02
CLTV	-0,12	0,67	-0,05	1,00	0,13	0,09	0,99	0,33	-0,06
DTI	-0,18	0,10	0,04	0,13	1,00	0,06	0,13	0,13	-0,08
UPB	0,10	0,03	0,06	0,09	0,06	1,00	0,09	0,16	0,15
LTV	-0,12	0,67	-0,05	0,99	0,13	0,09	1,00	0,33	-0,07
Loan Term	-0,05	0,26	0,02	0,33	0,13	0,16	0,33	1,00	-0,06
No Borrowers	-0,04	-0,05	-0,02	-0,06	-0,08	0,15	-0,07	-0,06	1,00

Tab. 8.9: Correlation matrix

Variable	VIF	p-value
Credit Score	1,09	<.0001
DTI	1,09	<.0001
CLTV	2,08	<.0001
No Borrowers	1,02	<.0001
No Units	1,09	<.0001
Homebuyer Flag	1,57	<.0001
MI Flag	2,02	<.0001
Loan Term $\geq$ 360m	1,22	<.0001
Channel = Not Available	1,00	0.7367
Channel = Broker	1,10	<.0001
Channel = Correspondent	1,09	<.0001
Loan Purpose = Refinance - Cash Out	1,71	<.0001
Loan Purpose = Refinance - No Cash Out	2,08	<.0001
Prop Val Method = ACE Loans	1,49	<.0001
Prop Val Method = Other Appraisals (Desktop, driveby, external, AVM)	1,03	<.0001
Prop Val Method = Not Available	1,00	0.6436
US Region = Midwest	1,30	<.0001
US Region = Northeast	1,23	<.0001
US Region = Other	1,00	<.0001
US Region = West	1,40	<.0001
Occupancy = Second Home	1,06	0.0103
Occupancy = Investment Property	1,16	<.0001

Tab. 8.10: Variance Inflation Factor

Variable	Long List	Short List
Credit Score	Missing Treatment applied	
Homebuyer Flag	Low GINI	-
MI Perc	Missing Treatment applied	Correlated with CLTV
No Units	Low GINI	-
Occupancy	Low GINI	-
CLTV	Missing Treatment applied	
DTI	Missing Treatment applied	
UPB		
LTV	Missing Treatment applied	Correlated with CLTV
Channel		
PPM Flag	Low GINI	-
Amort Type	Low GINI	-
State	Adapted Variable derived (US-region)	-
Prop Type	Low GINI	-
Loan Purpose	Low GINI	-
Loan Term		Correlated with CLTV
No Borrowers		
Sup Conf Flag	High Missing Rate	-
Prog Flag	High Missing Rate	-

HARP Flag	High Missing Rate	-
Prop Val Method		
Int Only Flag	Low GINI	-
US Region		
MI Flag		
Loan Term >360m	Low GINI	-
Loan Term Group	Similar Variable used	-
Loan Term $\geq$ 360m		
Loan Term = 360m	Similar Variable used	-

Tab. 8.11: Long list and Short list

8.5.3 Scaling

Due to the different ranges of risk factors, especially because of *Original Unpaid Balance*, a standardization of all numeric variables was carried out to prevent unusually small or large coefficients. The following formula is used:

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma} \tag{8.1}$$

where:

X = original variable
μ = mean of the variable
σ = standard deviation of the variable

8.6 Modeling

8.6.1 Logistic Regression

Employing the stepwise selection algorithm, all short list variables were inserted into the modeling process, resulting in the model presented in Table 8.12. The order of the risk factors indicates the significance in the model, with all coefficients exhibiting statistical significance indicated by low p-values. The AIC and BIC were calculated at each modeling step to limit the number of explanatory variables. Table 8.13 outlines the relative change per step. A noticeable drop in improvement occurs between step 5 and 6 and thus, prompting the removal of all succeeding variables, followed by a re-estimation of the model. The final model, detailed in Table 8.14, only includes numeric variables. All coefficients align with the expected economic sign; for example, a higher DTI corresponds to a higher probability of default, reflected in a positive coefficient. In contrast, an increased number of borrowers, which means more people are obligated to repay the loan, correlates with a lower probability of default, resulting in a negative coefficient. Performance metrics for the model are provided in Chapter 8.7 as part of the comparison with the Random Forest model.

Parameter	**Coefficient**	**p-value**
Intercept	-4,892	<.0001
Credit Score	-0,542	<.0001
DTI	0,392	<.0001
CLTV	0,223	<.0001
UPB	0,316	<.0001
No Borrowers	-0,245	<.0001
Loan Term $\geq$ 360m	0,144	<.0001
Channel = Broker	0,104	<.0001
Channel = Correspondent	0,279	<.0001
Prop Val Method = Other Appraisals (Desktop, driveby, external, AVM)	-0,206	<.0001
Prop Val Method = Full Appraisal	0,180	<.0001
US Region = Midwest	-0,152	<.0001
US Region = Northeast	0,138	<.0001
US Region = Other	0,869	<.0001

Tab. 8.12: Preliminary Logit Model

Step	**AIC**	**rel. AIC ch.**	**BIC**	**rel. BIC ch.**	**Variable added**
1	577821		577847		Credit Score
2	568369	-1,64%	568408	-1,63%	DTI adjusted
3	562335	-1,06%	562388	-1,06%	CLTV
4	558885	-0,61%	558951	-0,61%	UPB
5	555949	-0,53%	556028	-0,52%	No Borrowers
6	554913	-0,19%	555005	-0,18%	Loan Term $\geq$ 360m
7	554574	-0,06%	554679	-0,06%	Channel = Broker
8	554308	-0,05%	554427	-0,05%	Channel = Correspondent
9	554193	-0,02%	554324	-0,02%	Prop Val Method = Other Appraisals (Desktop, driveby, external, AVM)
10	554066	-0,02%	554210	-0,02%	Prop Val Method = Full Appraisal
11	554008	-0,01%	554166	-0,01%	US Region = Midwest
12	553984	0,00%	554155	0,00%	US Region = Northeast
13	553953	-0,01%	554137	0,00%	US Region = Other

Tab. 8.13: AIC and BIC per step

Parameter	Coefficient	p-value
Intercept	-4,546	<.0001
Credit Score	-0,542	<.0001
DTI adjusted	0,408	<.0001
CLTV	0,261	<.0001
No Borrowers	-0,248	<.0001
UPB	0,341	<.0001

Tab. 8.14: Final Logit Model

8.6.2 Random Forest

The same training and test sample after the data preparation process was used for the development of the Random Forest model. Due to sample size constraints and computational limitations, a 90:10 split ratio was applied to the training sample and a 2-fold cross-validation was executed during hyperparameter tuning. All variables from Table 8.11 were considered for modeling.

As a first step, a baseline Random Forest model was created using the default settings of the LGBMRegressor from the LightGBM library. A wide range for various parameters was established for the Random Search (see Figure 8.14, left), exploring 75 configurations on two folds each, totaling 150 fits on the HP Tuning Sample. The best-performing configuration was then tested on the HP Evaluation sample for comparison with the Baseline model. The best Random Search configuration is detailed in the "Random Search" column of Table 8.16.

Afterward, a more narrow parameter range was defined for the Grid Search (see Figure 8.14, right). Therefore, twelve configurations, each with a 2-fold cross-validation, resulting in 24 fits, were tested. The best hyperparameter settings are presented in Table 8.16. AUC and Gini values for all three models are illustrated in Figure 8.15. Based on these results, the Grid Search configuration was chosen for developing the Random Forest model.

Sample	# Accounts	# Def	% Def Rate
HP Tuning Sample	378958	5751	1,52%
HP Evaluation Sample	3410621	51759	1,52%

Tab. 8.15: Hyperparameter Tuning sample

```
'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000],
'num_leaves': [6, 8, 12, 16, 20, 25, 30, 40],
'max_depth': [2, 4, 6, 8, 12, 16],
'reg_alpha' : [1,1.2, 1.4, 1.6],
'reg_lambda' : [1,1.2,1.4, 1.6],
'feature_fraction': [0.3, 0.5, 0.7, 0.9],
'bagging_fraction': [0.3, 0.5, 0.7, 0.9],
'bagging_freq': [5, 7, 10],
```

```
'n_estimators': [200],
'num_leaves': [25, 30, 35],
'max_depth': [16, 20],
'reg_alpha' : [1.2],
'reg_lambda' : [1.6, 2.5],
'feature_fraction': [0.5],
'bagging_fraction': [0.3],
'bagging_freq': [7],
```

Fig. 8.14: Parameter ranges for Random and Grid Search

Hyperparameter	Random Search	Grid Search
Number of trees in Random Forest	200	200
Number of leaves	25	30
Maximum depth of tree	16	16
L1 regularization term on weights	1	1,2
L2 regularization term on weights	1,6	1,8
Fraction of features to be used in each boosting round	0,5	0,5
Fraction of data to be used in each boosting round	0,3	0,3
Frequency for bagging	7	7

Tab. 8.16: Best Parameter Settings of Random and Grid Search

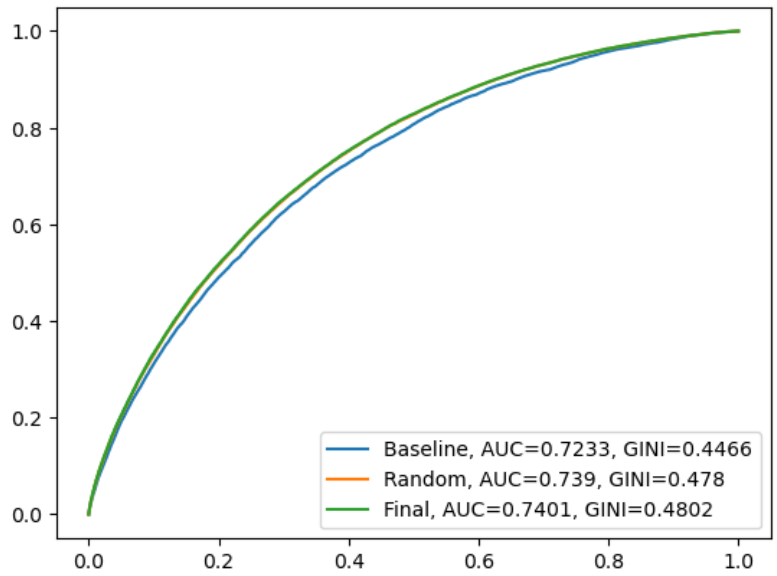

Fig. 8.15: Performance comparison for Hyperparameter Tuning

8.7 Validation and Comparison

8.7.1 Discriminatory power

The performance of both models is evaluated using both the test and an out-of-time dataset. To detect potential overfitting, performance metrics for the training sample are also provided. The Gini values are presented in Table 8.17 and visually represented in Figure 8.16 - 8.17. The Random Forest model exhibits a slightly higher performance in the training and test samples, while the logistic regression model shows a slight advantage in the validation sample. Notably, both models demonstrate higher Gini values on the validation sample, indicating no signs of overfitting.

Sample	Logistic Model	Random Forest
Training	0,4626	0,4708
Test	0,4638	0,4718
Validation	0,4810	0,4734

Tab. 8.17: Comparison of Gini values in training, test and validation sample

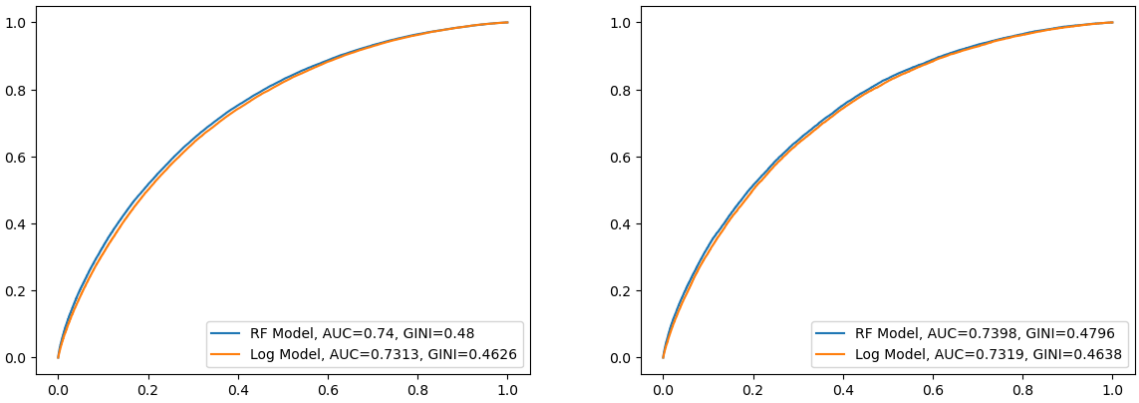

Fig. 8.16: AUC comparison in training and test sample

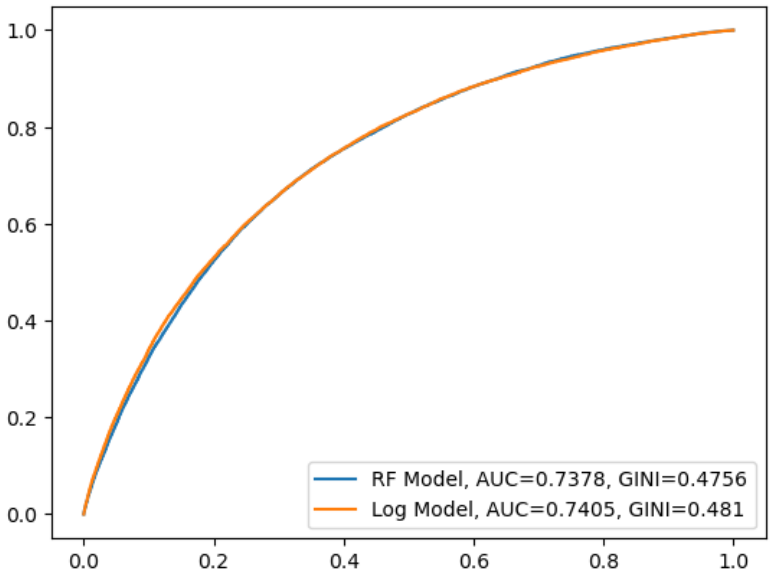

Fig. 8.17: AUC comparison in Validation sample

8.7.2 Classification

To determine an optimal threshold to set a predicted default flag derived from the predicted PD, the F1-score is computed for different values. These values are detailed in Table 8.18, leading to the decision to set the threshold at PD $> 4\%$ for both models. The resulting confusion matrices for the training, test and validation sample are visualized in Figure 8.18 to 8.20. On the x-axis,

the predicted values are presented, while the y-axis represents true values. Focusing on the most critical cases, where default events are misclassified as non-default (lower left corner), the Random Forest model performs slightly worse in this regard compared to the Logistic Regression model on all samples.

	F1-Score	
Threshold	**Logistic Regression**	**Random Forest**
0,005	0,0362	0,0299
0,010	0,0464	0,0360
0,015	0,0561	0,0521
0,020	0,0648	0,0664
0,025	0,0721	0,0788
0,030	0,0777	0,0872
0,035	0,0819	0,0920
0,040	0,0836	0,0929
0,045	0,0830	0,0901
0,050	0,0803	0,0846
0,055	0,0751	0,0761
0,060	0,0687	0,0667
0,065	0,0608	0,0588
0,070	0,0522	0,0517
0,075	0,0438	0,0446
0,080	0,0362	0,0387
0,085	0,0290	0,0329
0,090	0,0234	0,0284
0,095	0,0193	0,0243

Tab. 8.18: F1 score for different threshold values

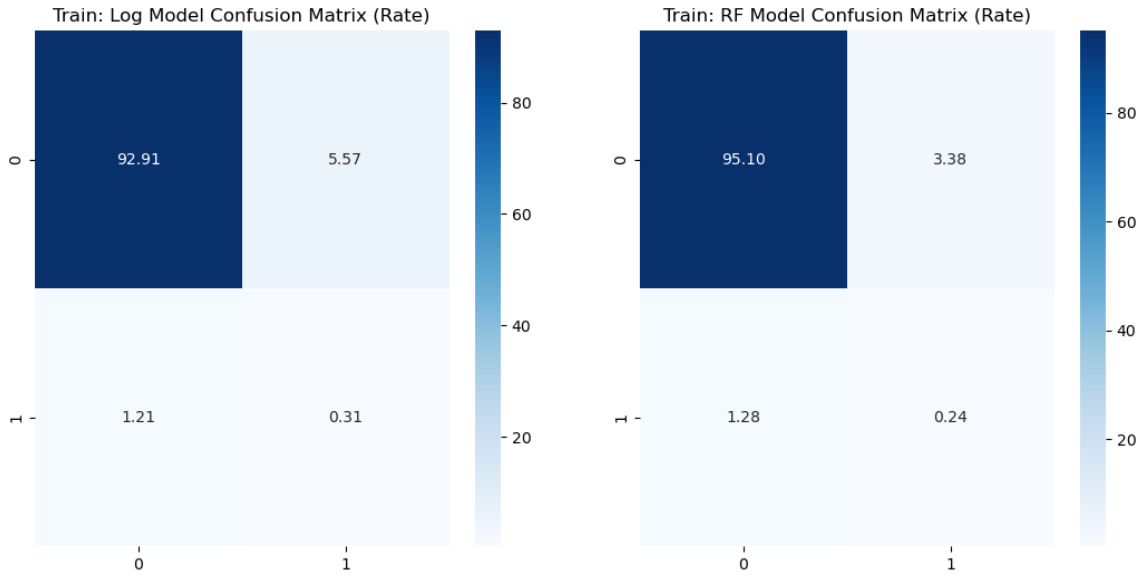

Fig. 8.18: Confusionmatrix in training sample

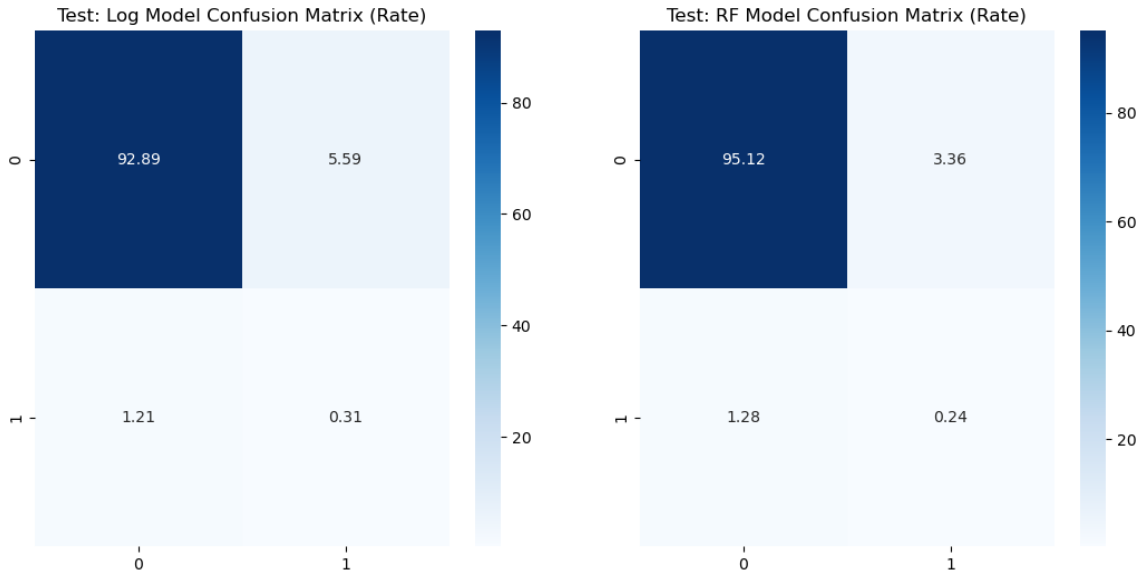

Fig. 8.19: Confusionmatrix in test sample

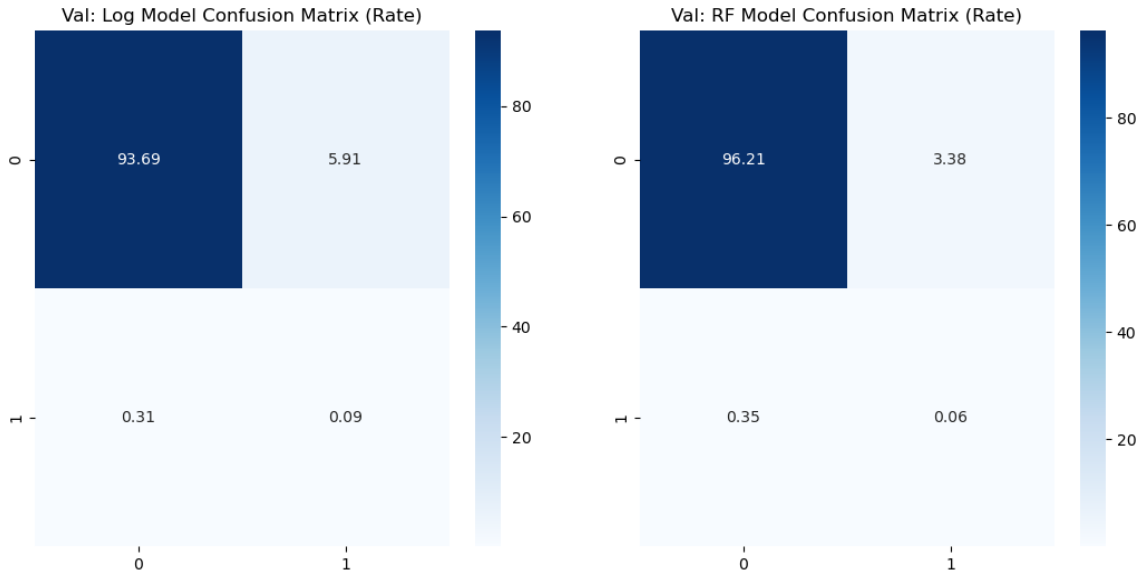

Fig. 8.20: Confusionmatrix in validation sample

8.7.3 Stability Test

A stability test for both models was conducted by assessing the Gini coefficient on the test and validation samples for each year, as detailed in Table 8.19 and illustrated through ROC curves in Figure 8.21. The left side of the Figure is the logistic model, while the right side showcases the Random Forest model. Both models exhibited their lowest performance in 2019. The logistic model demonstrated its peak discriminatory power in 2018, whereas the Random Forest model excelled in the year 2020. Across the years, both models exhibit a Gini difference of up to 8.4%, suggesting a moderate level of stability.

Year	LOG Model	RF Model
2018	0,5202	0,5044
2019	0,4358	0,4440
2020	0,4936	0,5162
2021	0,4810	0,4734

Tab. 8.19: Stability test over time

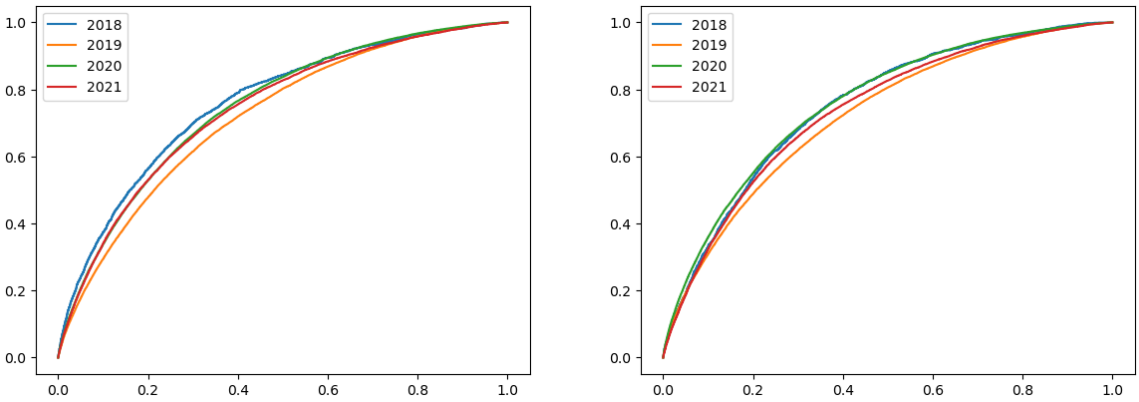

Fig. 8.21: ROC curve of stability test over time

8.7.4 Binning Process

A binning process was executed to provide an overview of the distribution of PDs and default rates. In real-world applications, it is a standard practice to categorize PD values into distinct rating bands, distinguishing between investment and non-investment grades. A straightforward method involves employing fixed ranges for each grade. The specific ranges and the resulting distribution are detailed in Table 8.20 and visually presented in Figure 8.22. These arbitrary defined ranges already provide a satisfactory distribution, accompanied by a desired upward trend in defaults. The logistic model tends to allocate a greater number of customers with PD values below 0.5% compared to the Random Forest model. The Random Forest model exhibits a moderate concentration in the range of 1% to 1.5%.

	# Observation		% Observation		% Default Rate	
Range	**Log**	**RF**	**Log**	**RF**	**Log**	**RF**
(-inf, 0.001)	48.096		1%		0,06%	
[0.001, 0.005)	1.665.964		21%		0,19%	
[0.005, 0.01)	1.983.432	1.734.463	25%	22%	0,43%	0,17%
[0.01, 0.015)	1.326.997	2.829.949	17%	36%	0,75%	0,48%
[0.015, 0.02)	891.664	1.454.260	11%	18%	1,09%	1,02%
[0.02, 0.03)	1.031.273	1.225.736	13%	15%	1,58%	1,74%
[0.03, 0.04)	493.678	388.029	6%	5%	2,24%	2,65%
[0.04, inf)	470.193	278.860	6%	4%	3,32%	4,06%

Tab. 8.20: Distribution Table of PDs and Default Rates for Logistic Regression and Random Forest Model

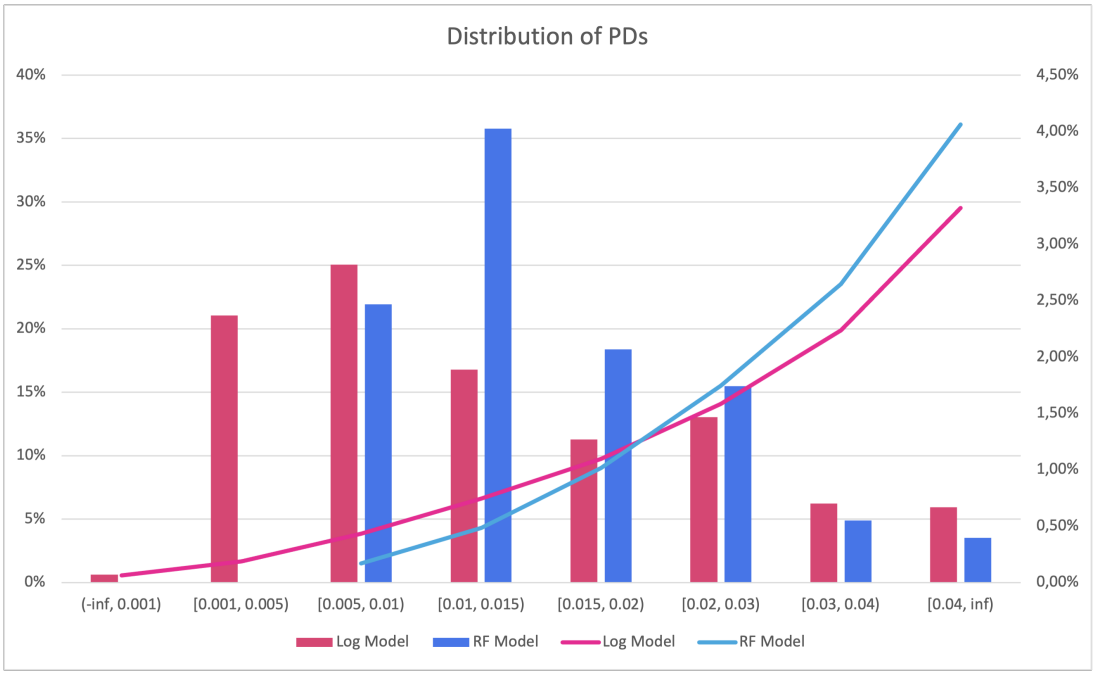

Fig. 8.22: Distribution Plot of PDs and Default Rates for Logistic Regression and Random Forest Model

8.8 Interpretability

8.8.1 Global Interpretation

To interpret the Random Forest model, several techniques described in Chapter 7 were utilized. Figure 8.23 shows the feature importance by split, signifying the frequency with which a variable is employed for a split. Meanwhile, Figure 8.24 depicts feature importance by gain, denoting the information gain derived from the risk factor during a split. Reiterating the findings of the logistic regression model, the most important risk factors in the Random Forest model are *Credit Score* (fico), *UPB*, *CLTV*, *DTI* and *No Borrowers* (cnt_borr) as well. The *LTV* was removed from the logistic regression model due to its high correlation with the risk factor *CLTV*.

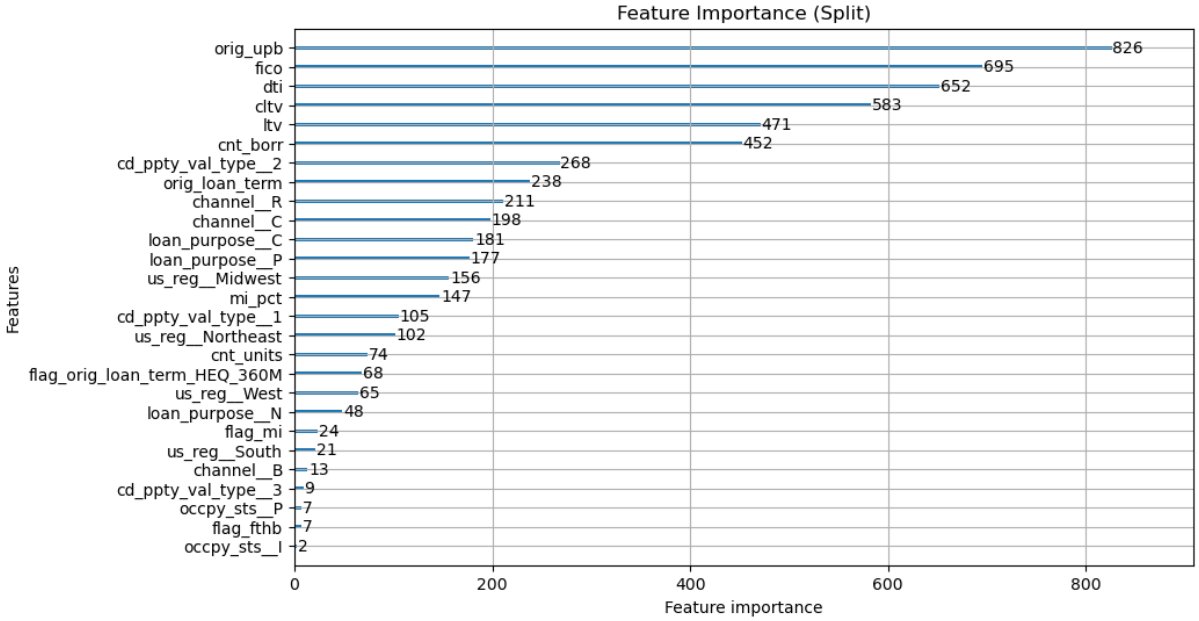

Fig. 8.23: Feature Importance by Split

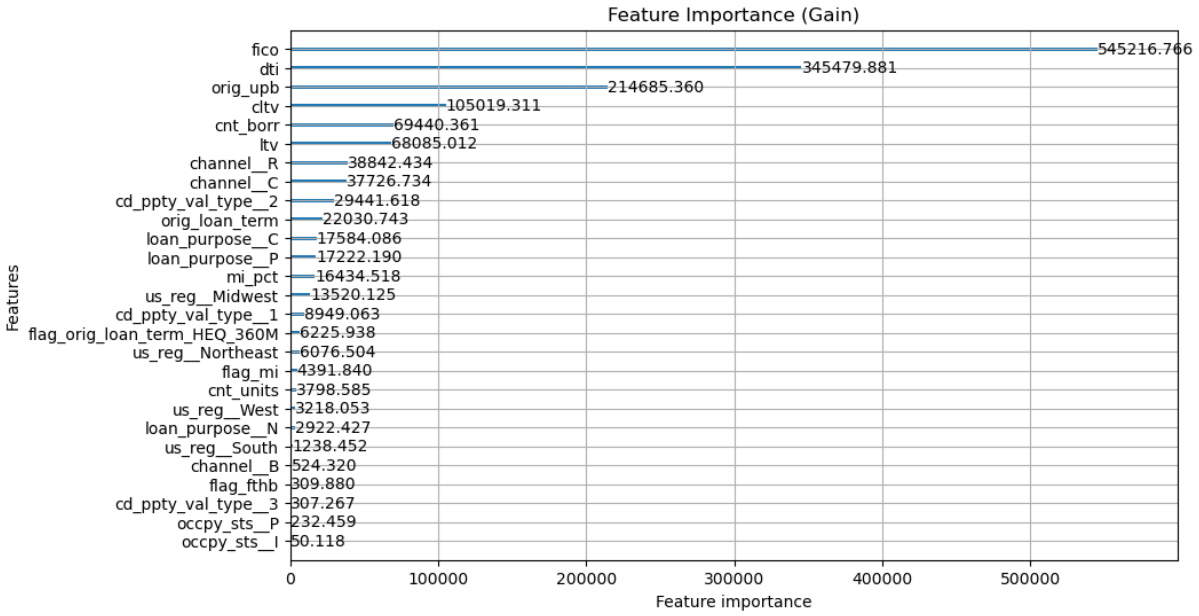

Fig. 8.24: Feature Importance by Gain

8.8.2 Local Interpretation

Two particular instances were selected for the Local Interpretable Model-agnostic Explanation (LIME) technique. The first observation received nearly identical PDs of 1.2547% in both models, while the second observation received vastly different PDs of 59.0619% in the Logit Model and 2.2576% in the RF Model. The score composition for the former is displayed in Table 8.21 and the LIME results are visible in Figures 8.25 and 8.26. The green bars indicate the features

contributing to predicting the class "non-default" class and the red bars signify the contribution to the "default" class.

While the high Credit Score mainly influences the low PD in observation 1 for both models, the lower Credit Score, as well as the higher DTI, resulted in a higher PD in the Logit model. Surprisingly, in the RF model, both features contribute to the classification of "non-default", contrasting the result seen in the first observation.

Variable	Observation 1			Observation 2		
	Original Value	Scaled Value	Score	Original Value	Scaled Value	Score
Intercept			-4,5457			-4,5457
Credit Score	691	- 1,4100	0,7642	309	- 10,1100	5,4796
DTI	29,99	- 0,4421	-0,1802	40,01	0,5910	0,2409
CLTV	73,00	- 0,0359	-0,0094	30,00	- 2,5257	-0,6600
No Borrowers	2	0,9969	-0,2476	1	- 0,9326	0,2317
UPB	200.000	- 0,4313	-0,1469	110.000	- 1,1133	-0,3793
Model Score		-4,3656			0,3672	
PD		1,25%			59,08%	

Tab. 8.21: PD Result of Logistic Regression model

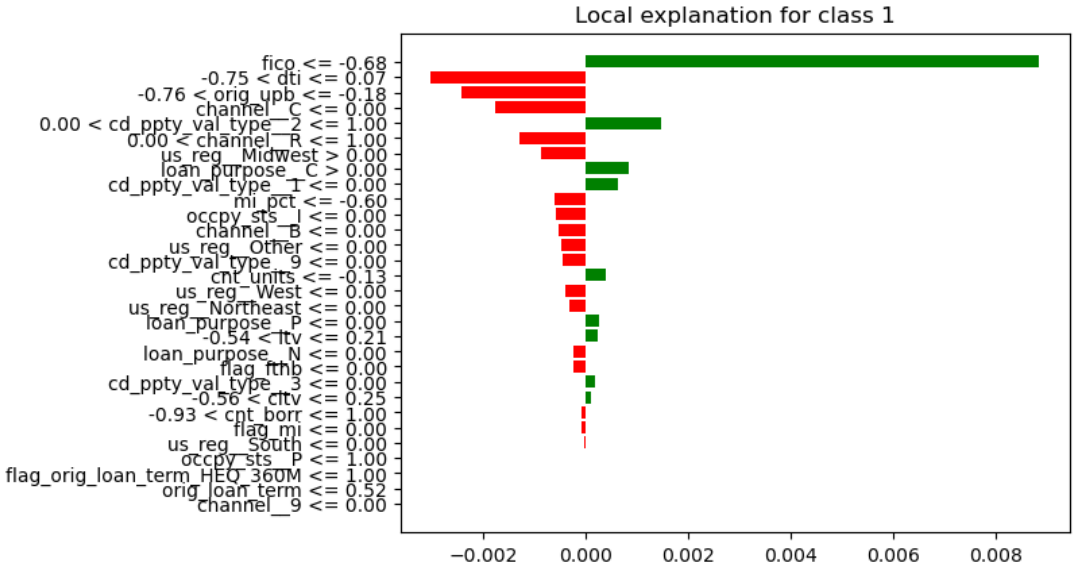

Fig. 8.25: LIME Result of Random Forest, minimum difference in PD

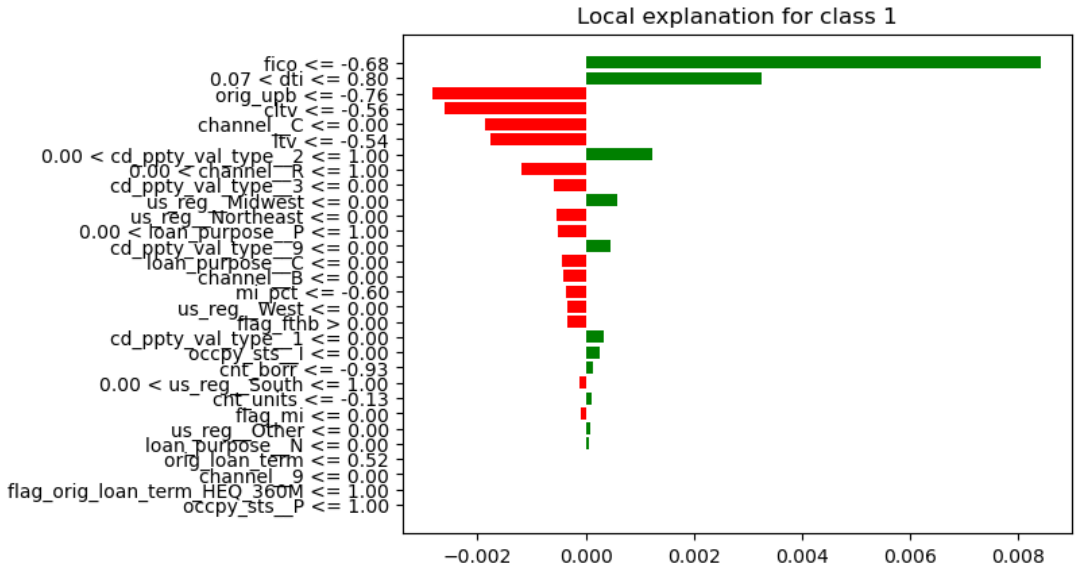

Fig. 8.26: LIME Result of Random Forest, maximum difference in PD

8.8.3 Individual Decision Trees, PDP and ICE plots

Two individual decision trees from the Random Forest model are presented in Figure 8.27 and 8.28. Analyzing the split conditions, they reaffirm the importance of *Credit Score*, *UPB*, *DTI* and *LTV* as primary variables. To assess their influence, Partial Dependence plots of these four features, along with *Prop Val Method = Full Appraisal* and *Loan Purpose = Refinance - Cash Out* are showcased in Figure 8.29 to 8.31. All plots are in Appendix C.4. The observed trends align with the economic rationale: increasing Credit Score corresponds to decreasing PD, while rising UPB, DTI and CLTV are associated with higher PD. Both categorical variables indicate a higher PD if the condition is fulfilled. Individual Conditional Expectation (ICE) plots were generated to further analyze the impact of the four most crucial features. Figure 8.32 to 8.33 provide a granular analysis of each feature's influence on individual predictions. The findings from these ICE plots also align with the anticipated relationship between the respective risk factor and the PD estimation.

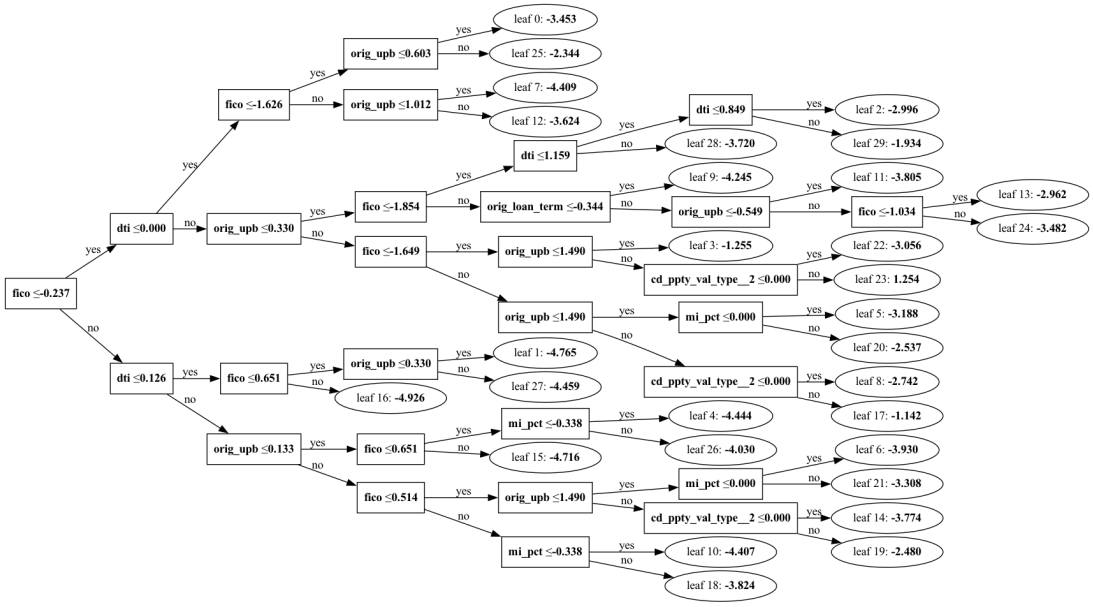

Fig. 8.27: Individual Decision Tree 1 of Random Forest

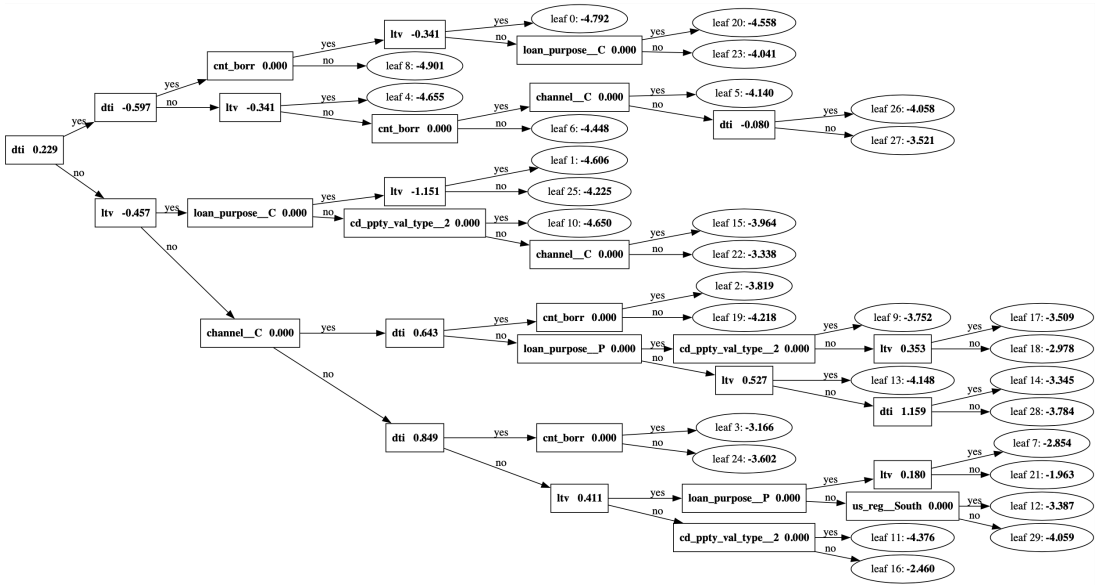

Fig. 8.28: Individual Decision Tree 2 of Random Forest

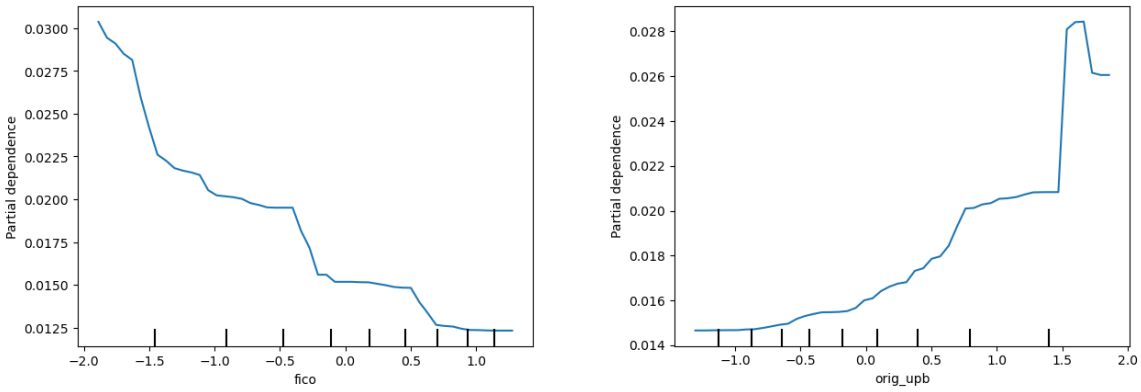

Fig. 8.29: Partial Dependence Plots for Credit score and UPB

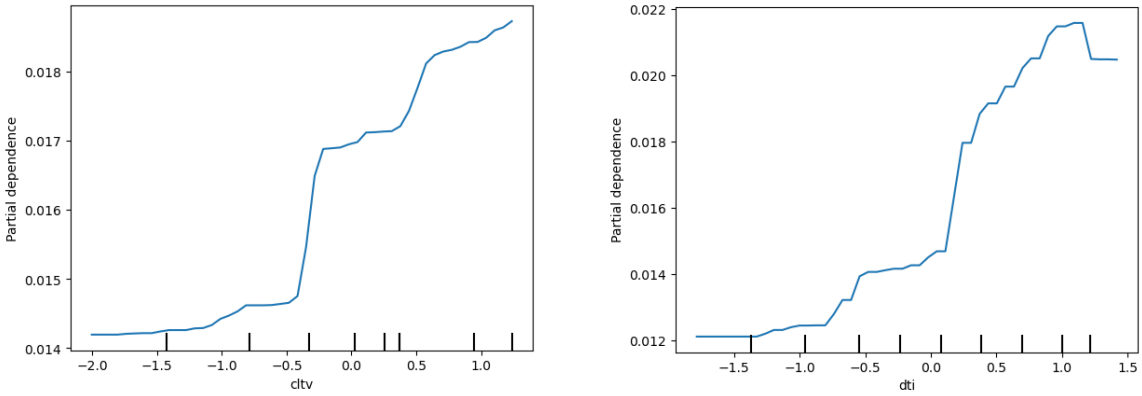

Fig. 8.30: Partial Dependence Plots for CLTV and DTI

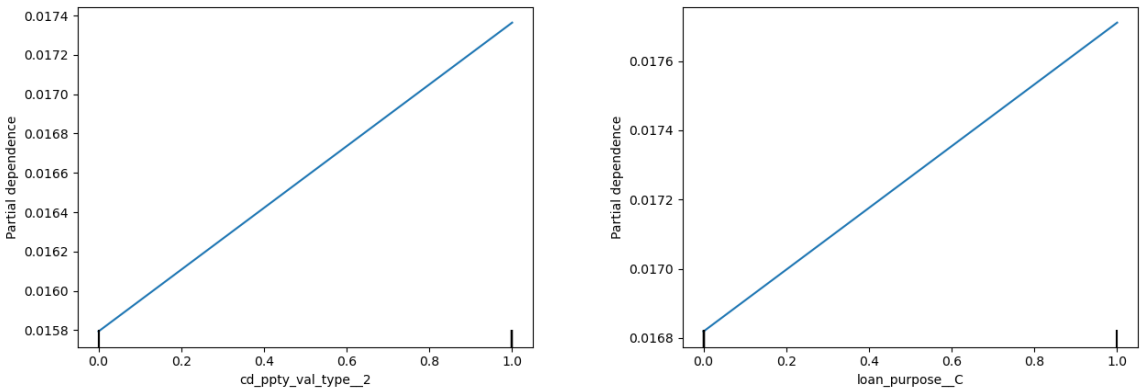

Fig. 8.31: Partial Dependence Plots for Prop Val Method = Full Appraisal and Loan Purpose = Refinance - Cash Out

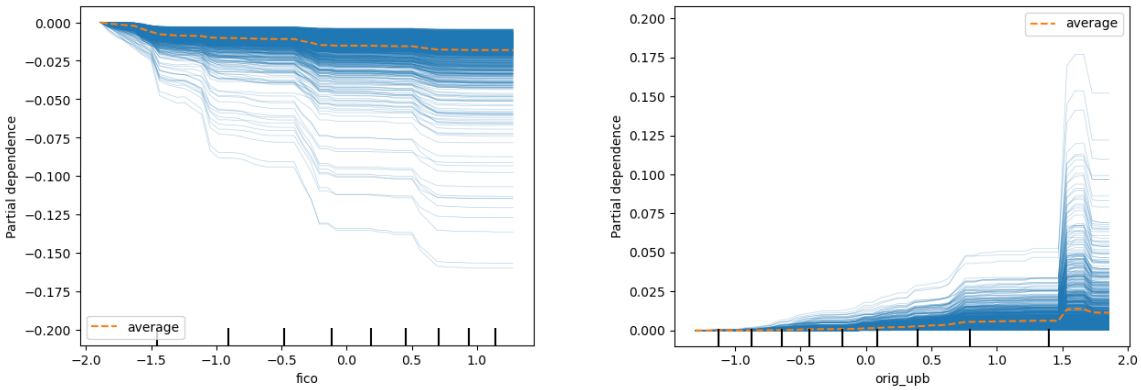

Fig. 8.32: Individual Conditional Expectation Plots for Credit Score and UPB

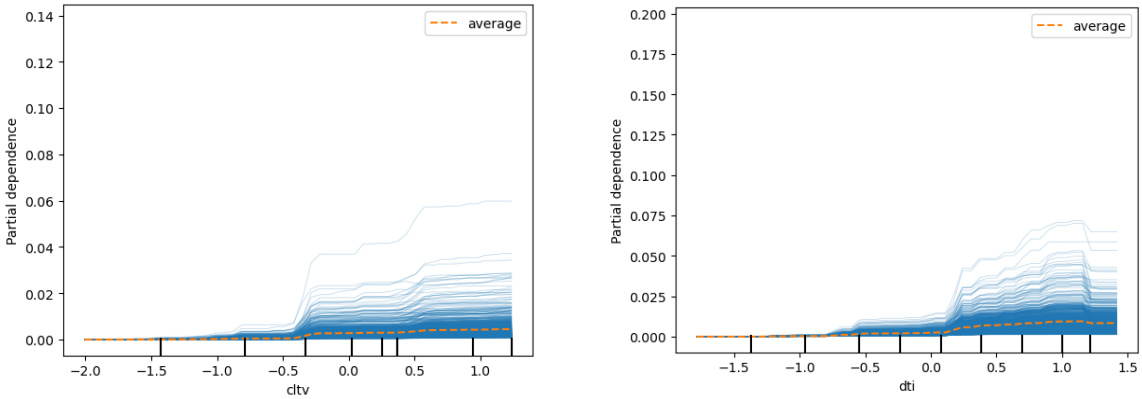

Fig. 8.33: Individual Conditional Expectation Plots for CLTV and DTI

Chapter 9

Conclusio

Appendix A

Variable names in data set

Variable Name	Description
amrtzn_type___FRM	Amort Type = Fixed Rate Mortgage
cd_ppty_val_type___1	Prop Val Method = ACE Loans
cd_ppty_val_type___2	Prop Val Method = Full Appraisal
cd_ppty_val_type___3	Prop Val Method = Other Appraisals (Desktop, driveby, external, AVM)
cd_ppty_val_type___9	Prop Val Method = Not Available
channel___9	Channel = Not Available
channel___B	Channel = Broker
channel___C	Channel = Correspondent
channel___R	Channel = Retail
channel___T	Channel = TPO Not Specified
ind_afdl___9	Prog Flag = Not Available or Not Applicable
ind_afdl___F	Prog Flag = HFA Advantage
ind_afdl___H	Prog Flag = Home Possible
ind_afdl___R	Prog Flag = Refi Possible
loan_purpose___C	Loan Purpose = Refinance - Cash Out
loan_purpose___N	Loan Purpose = Refinance - No Cash Out
loan_purpose___P	Loan Purpose = Purchase
occpy_sts___I	Occupancy = Investment Property
occpy_sts___P	Occupancy = Primary Residence
occpy_sts___S	Occupancy = Second Home
orig_loan_term_3grp___EQ_360M	Loan Term Group = 360 Months
orig_loan_term_3grp___HI_360M	Loan Term Group >360 Months
orig_loan_term_3grp___LE_360M	Loan Term Group <360 Months
prop_type___CO	Prop Type = Condo
prop_type___CP	Prop Type = Co-op
prop_type___MH	Prop Type = Manufactured Housing
prop_type___PU	Prop Type = PUD
prop_type___SF	Prop Type = Single-Family
us_reg___Midwest	US Region = Midwest
us_reg___Northeast	US Region = Northeast
us_reg___Other	US Region = Other
us_reg___South	US Region = South
us_reg___West	US Region = West
flag_fthb	Homebuyer Flag
flag_int_only	Int Only Flag

flag_mi	MI Flag
flag_cnt_units	Number of units >1
flag_orig_loan_term_EQ_360M	Loan Term = 360m
flag_orig_loan_term_HEQ_360M	Loan Term ≥ 360m
flag_orig_loan_term_HI_360M	Loan Term >360m
flag_sc	Sup Conf Flag
ind_harp	HARP Flag
ppmt_pnlty	PPM Flag
cltv	CLTV
cnt_borr	No Borrowers
cnt_units	No Units
dti	DTI
fico	Credit Score
ltv	LTV
mi_pct	MI Perc
orig_loan_term	Loan Term
orig_upb	UPB

Tab. A.1: Variable Name in data set with description

Appendix B

Default Rates of whole data set

	Without Exclusions			With Exclusions		
Date	# Accounts	# Def	% Def Rate	# Accounts	# Def	% Def Rate
Feb 1999	5.011	9	0,18%			
Mar 1999	161.042	223	0,14%	156.078	223	0,14%
Apr 1999	153.734	260	0,17%	149.010	260	0,17%
May 1999	167.212	312	0,19%	162.388	312	0,19%
Jun 1999	127.273	203	0,16%	123.515	203	0,16%
Jul 1999	109.484	238	0,22%	106.094	238	0,22%
Aug 1999	98.295	254	0,26%	95.139	254	0,27%
Sep 1999	82.515	234	0,28%	79.657	234	0,29%
Oct 1999	76.428	253	0,33%	73.493	253	0,34%
Nov 1999	60.623	206	0,34%	58.056	206	0,35%
Dec 1999	56.123	189	0,34%	53.735	189	0,35%
Jan 2000	56.204	249	0,44%	53.766	249	0,46%
Feb 2000	56.034	296	0,53%	53.268	296	0,56%
Mar 2000	34.391	159	0,46%	31.862	159	0,50%
Apr 2000	35.530	143	0,40%	31.021	143	0,46%
May 2000	45.550	255	0,56%	37.472	255	0,68%
Jun 2000	53.557	288	0,54%	42.326	288	0,68%
Jul 2000	57.754	344	0,60%	42.203	344	0,82%
Aug 2000	66.307	421	0,63%	44.062	421	0,96%
Sep 2000	70.303	494	0,70%	48.669	494	1,02%
Oct 2000	73.513	494	0,67%	52.575	494	0,94%
Nov 2000	64.587	403	0,62%	46.573	403	0,87%
Dec 2000	76.632	545	0,71%	54.873	545	0,99%
Jan 2001	83.556	602	0,72%	58.194	602	1,03%
Feb 2001	75.637	569	0,75%	55.461	569	1,03%
Mar 2001	93.130	528	0,57%	77.036	528	0,69%
Apr 2001	134.241	526	0,39%	116.681	526	0,45%
May 2001	182.006	715	0,39%	162.261	715	0,44%
Jun 2001	209.952	826	0,39%	189.900	826	0,43%
Jul 2001	215.387	877	0,41%	194.197	877	0,45%
Aug 2001	199.998	886	0,44%	177.766	886	0,50%
Sep 2001	181.646	907	0,50%	158.541	907	0,57%
Oct 2001	185.047	961	0,52%	159.762	961	0,60%
Nov 2001	184.186	851	0,46%	156.267	851	0,54%
Dec 2001	264.752	964	0,36%	228.886	964	0,42%
Jan 2002	339.452	978	0,29%	298.275	978	0,33%

Feb 2002	316.911	893	0,28%	272.236	893	0,33%
Mar 2002	242.242	910	0,38%	194.813	910	0,47%
Apr 2002	199.192	856	0,43%	153.106	856	0,56%
May 2002	203.055	1.024	0,50%	149.008	1.024	0,69%
Jun 2002	158.258	1.035	0,65%	107.281	1.035	0,96%
Jul 2002	158.900	1.074	0,68%	104.011	1.074	1,03%
Aug 2002	155.575	958	0,62%	96.789	958	0,99%
Sep 2002	193.156	893	0,46%	117.277	893	0,76%
Oct 2002	274.830	886	0,32%	174.301	886	0,51%
Nov 2002	331.594	748	0,23%	234.445	748	0,32%
Dec 2002	403.672	791	0,20%	307.544	791	0,26%
Jan 2003	351.153	653	0,19%	282.039	653	0,23%
Feb 2003	324.902	619	0,19%	270.779	619	0,23%
Mar 2003	308.817	553	0,18%	263.871	553	0,21%
Apr 2003	305.264	423	0,14%	266.294	423	0,16%
May 2003	348.083	436	0,13%	309.593	436	0,14%
Jun 2003	369.023	413	0,11%	333.634	413	0,12%
Jul 2003	346.861	436	0,13%	318.726	436	0,14%
Aug 2003	471.073	494	0,10%	443.731	494	0,11%
Sep 2003	576.871	502	0,09%	546.467	502	0,09%
Oct 2003	428.832	506	0,12%	400.001	506	0,13%
Nov 2003	268.799	473	0,18%	242.628	473	0,19%
Dec 2003	197.711	451	0,23%	177.886	451	0,25%
Jan 2004	143.015	394	0,28%	128.043	394	0,31%
Feb 2004	136.235	390	0,29%	122.803	390	0,32%
Mar 2004	116.305	343	0,29%	105.926	343	0,32%
Apr 2004	168.044	351	0,21%	155.115	351	0,23%
May 2004	216.630	471	0,22%	202.563	471	0,23%
Jun 2004	243.744	396	0,16%	229.097	396	0,17%
Jul 2004	159.733	328	0,21%	146.150	328	0,22%
Aug 2004	118.494	311	0,26%	102.802	311	0,30%
Sep 2004	100.060	262	0,26%	85.824	262	0,31%
Oct 2004	107.624	281	0,26%	94.962	281	0,30%
Nov 2004	105.842	265	0,25%	95.095	265	0,28%
Dec 2004	108.327	364	0,34%	98.716	364	0,37%
Jan 2005	120.026	418	0,35%	109.967	418	0,38%
Feb 2005	114.419	508	0,44%	104.941	508	0,48%
Mar 2005	88.389	388	0,44%	81.208	388	0,48%
Apr 2005	104.399	443	0,42%	96.473	443	0,46%
May 2005	148.232	570	0,38%	137.871	570	0,41%
Jun 2005	116.487	523	0,45%	107.539	523	0,49%
Jul 2005	125.284	520	0,42%	116.951	520	0,44%
Aug 2005	144.449	563	0,39%	136.354	563	0,41%
Sep 2005	168.860	500	0,30%	160.388	500	0,31%
Oct 2005	179.349	500	0,28%	170.283	500	0,29%
Nov 2005	157.105	374	0,24%	149.061	374	0,25%
Dec 2005	148.118	453	0,31%	140.142	453	0,32%
Jan 2006	122.301	409	0,33%	114.776	409	0,36%

Feb 2006	109.021	412	0,38%	101.739	412	0,40%
Mar 2006	91.332	414	0,45%	85.445	414	0,48%
Apr 2006	98.995	363	0,37%	92.806	363	0,39%
May 2006	109.663	413	0,38%	102.906	413	0,40%
Jun 2006	98.158	395	0,40%	91.421	395	0,43%
Jul 2006	101.997	402	0,39%	93.896	402	0,43%
Aug 2006	107.615	489	0,45%	98.261	489	0,50%
Sep 2006	95.510	476	0,50%	86.200	476	0,55%
Oct 2006	98.841	454	0,46%	90.074	454	0,50%
Nov 2006	90.923	458	0,50%	84.692	458	0,54%
Dec 2006	104.364	558	0,53%	98.046	558	0,57%
Jan 2007	96.382	581	0,60%	90.732	581	0,64%
Feb 2007	103.078	588	0,57%	97.824	588	0,60%
Mar 2007	93.133	538	0,58%	87.752	538	0,61%
Apr 2007	92.518	550	0,59%	86.174	550	0,64%
May 2007	117.897	741	0,63%	109.957	741	0,67%
Jun 2007	124.774	963	0,77%	117.090	963	0,82%
Jul 2007	121.108	1.004	0,83%	113.394	1.004	0,89%
Aug 2007	108.066	1.065	0,99%	98.632	1.065	1,08%
Sep 2007	98.932	1.025	1,04%	86.704	1.025	1,18%
Oct 2007	97.364	1.132	1,16%	85.541	1.132	1,32%
Nov 2007	90.554	1.401	1,55%	81.156	1.401	1,73%
Dec 2007	98.369	1.682	1,71%	89.404	1.682	1,88%
Jan 2008	88.259	1.626	1,84%	81.368	1.626	2,00%
Feb 2008	108.319	2.032	1,88%	100.121	2.032	2,03%
Mar 2008	102.239	1.755	1,72%	93.920	1.755	1,87%
Apr 2008	168.379	2.013	1,20%	156.407	2.013	1,29%
May 2008	153.609	2.103	1,37%	139.160	2.103	1,51%
Jun 2008	146.924	1.864	1,27%	130.618	1.864	1,43%
Jul 2008	115.915	1.617	1,39%	100.547	1.617	1,61%
Aug 2008	108.849	1.550	1,42%	89.159	1.550	1,74%
Sep 2008	87.955	1.375	1,56%	65.359	1.375	2,10%
Oct 2008	63.338	974	1,54%	44.376	974	2,19%
Nov 2008	71.031	1.083	1,52%	53.403	1.083	2,03%
Dec 2008	76.467	998	1,31%	58.577	998	1,70%
Jan 2009	46.042	685	1,49%	33.903	685	2,02%
Feb 2009	95.146	707	0,74%	80.191	707	0,88%
Mar 2009	163.902	461	0,28%	153.766	461	0,30%
Apr 2009	220.154	380	0,17%	211.363	380	0,18%
May 2009	241.508	409	0,17%	230.926	409	0,18%
Jun 2009	231.450	359	0,16%	223.715	359	0,16%
Jul 2009	244.868	375	0,15%	238.009	375	0,16%
Aug 2009	239.830	562	0,23%	231.348	562	0,24%
Sep 2009	193.683	493	0,25%	182.402	493	0,27%
Oct 2009	132.521	361	0,27%	122.858	361	0,29%
Nov 2009	117.734	286	0,24%	107.667	286	0,27%
Dec 2009	140.695	348	0,25%	127.610	348	0,27%
Jan 2010	148.725	337	0,23%	134.895	337	0,25%

Feb 2010	152.532	368	0,24%	138.597	368	0,27%
Mar 2010	131.242	319	0,24%	119.348	319	0,27%
Apr 2010	109.134	290	0,27%	99.008	290	0,29%
May 2010	124.837	323	0,26%	114.545	323	0,28%
Jun 2010	121.090	314	0,26%	111.153	314	0,28%
Jul 2010	111.497	244	0,22%	101.766	244	0,24%
Aug 2010	130.747	310	0,24%	120.418	310	0,26%
Sep 2010	138.861	272	0,20%	130.152	272	0,21%
Oct 2010	173.719	264	0,15%	164.283	264	0,16%
Nov 2010	189.301	265	0,14%	179.834	265	0,15%
Dec 2010	206.731	308	0,15%	196.662	308	0,16%
Jan 2011	199.346	290	0,15%	190.598	290	0,15%
Feb 2011	184.790	326	0,18%	176.013	326	0,19%
Mar 2011	145.050	358	0,25%	135.157	358	0,26%
Apr 2011	97.376	268	0,28%	88.728	268	0,30%
May 2011	94.342	256	0,27%	83.067	256	0,31%
Jun 2011	86.287	226	0,26%	75.371	226	0,30%
Jul 2011	79.971	231	0,29%	69.789	231	0,33%
Aug 2011	100.035	199	0,20%	88.038	199	0,23%
Sep 2011	102.568	228	0,22%	90.068	228	0,25%
Oct 2011	108.561	213	0,20%	95.405	213	0,22%
Nov 2011	141.784	174	0,12%	122.457	174	0,14%
Dec 2011	146.548	189	0,13%	129.561	189	0,15%
Jan 2012	140.197	196	0,14%	125.596	196	0,16%
Feb 2012	140.802	203	0,14%	125.227	203	0,16%
Mar 2012	126.206	228	0,18%	113.109	228	0,20%
Apr 2012	139.887	273	0,20%	127.860	273	0,21%
May 2012	149.736	336	0,22%	138.836	336	0,24%
Jun 2012	150.589	467	0,31%	140.210	467	0,33%
Jul 2012	155.802	472	0,30%	144.853	472	0,33%
Aug 2012	154.506	405	0,26%	144.870	405	0,28%
Sep 2012	168.161	355	0,21%	160.047	355	0,22%
Oct 2012	192.072	342	0,18%	184.281	342	0,19%
Nov 2012	204.918	320	0,16%	197.398	320	0,16%
Dec 2012	211.605	330	0,16%	204.715	330	0,16%
Jan 2013	182.099	277	0,15%	176.671	277	0,16%
Feb 2013	197.623	306	0,15%	192.214	306	0,16%
Mar 2013	207.804	275	0,13%	202.476	275	0,14%
Apr 2013	205.659	336	0,16%	200.626	336	0,17%
May 2013	199.240	311	0,16%	194.112	311	0,16%
Jun 2013	189.424	300	0,16%	184.702	300	0,16%
Jul 2013	207.243	285	0,14%	202.238	285	0,14%
Aug 2013	181.888	301	0,17%	177.300	301	0,17%
Sep 2013	156.356	278	0,18%	151.341	278	0,18%
Oct 2013	124.203	255	0,21%	118.690	255	0,21%
Nov 2013	97.241	229	0,24%	91.973	229	0,25%
Dec 2013	96.716	224	0,23%	91.122	224	0,25%
Jan 2014	94.450	203	0,21%	89.338	203	0,23%

Feb 2014	97.778	188	0,19%	91.593	188	0,21%
Mar 2014	69.672	156	0,22%	64.043	156	0,24%
Apr 2014	67.625	141	0,21%	61.151	141	0,23%
May 2014	85.736	172	0,20%	77.204	172	0,22%
Jun 2014	88.681	165	0,19%	79.378	165	0,21%
Jul 2014	93.305	148	0,16%	83.776	148	0,18%
Aug 2014	109.142	143	0,13%	98.847	143	0,14%
Sep 2014	111.339	153	0,14%	101.707	153	0,15%
Oct 2014	108.722	168	0,15%	99.806	168	0,17%
Nov 2014	105.755	144	0,14%	97.360	144	0,15%
Dec 2014	100.666	124	0,12%	93.150	124	0,13%
Jan 2015	98.420	141	0,14%	91.891	141	0,15%
Feb 2015	102.397	135	0,13%	96.637	135	0,14%
Mar 2015	93.638	119	0,13%	88.996	119	0,13%
Apr 2015	119.077	118	0,10%	114.066	118	0,10%
May 2015	153.544	138	0,09%	146.992	138	0,09%
Jun 2015	142.258	139	0,10%	135.831	139	0,10%
Jul 2015	135.591	124	0,09%	129.346	124	0,10%
Aug 2015	134.534	146	0,11%	127.138	146	0,11%
Sep 2015	126.000	141	0,11%	116.403	141	0,12%
Oct 2015	120.117	137	0,11%	110.051	137	0,12%
Nov 2015	118.202	149	0,13%	108.297	149	0,14%
Dec 2015	116.689	140	0,12%	107.229	140	0,13%
Jan 2016	101.329	155	0,15%	92.955	155	0,17%
Feb 2016	112.360	185	0,16%	102.394	185	0,18%
Mar 2016	89.228	136	0,15%	81.486	136	0,17%
Apr 2016	93.296	126	0,14%	86.306	126	0,15%
May 2016	127.305	153	0,12%	119.382	153	0,13%
Jun 2016	132.679	169	0,13%	125.399	169	0,13%
Jul 2016	141.715	165	0,12%	135.158	165	0,12%
Aug 2016	156.698	176	0,11%	149.761	176	0,12%
Sep 2016	146.561	152	0,10%	140.269	152	0,11%
Oct 2016	170.536	183	0,11%	163.355	183	0,11%
Nov 2016	157.177	197	0,13%	150.666	197	0,13%
Dec 2016	163.960	329	0,20%	157.085	329	0,21%
Jan 2017	157.805	523	0,33%	151.056	523	0,35%
Feb 2017	151.029	663	0,44%	143.289	663	0,46%
Mar 2017	103.232	473	0,46%	96.564	473	0,49%
Apr 2017	87.101	483	0,55%	81.078	483	0,60%
May 2017	102.954	561	0,54%	96.230	561	0,58%
Jun 2017	98.018	512	0,52%	91.749	512	0,56%
Jul 2017	115.187	592	0,51%	108.991	592	0,54%
Aug 2017	123.457	569	0,46%	117.377	569	0,48%
Sep 2017	123.275	555	0,45%	117.489	555	0,47%
Oct 2017	137.663	517	0,38%	131.595	517	0,39%
Nov 2017	130.775	259	0,20%	124.861	259	0,21%
Dec 2017	139.655	239	0,17%	133.368	239	0,18%
Jan 2018	115.392	230	0,20%	110.252	230	0,21%

Feb 2018	105.792	204	0,19%	100.986	204	0,20%
Mar 2018	83.637	167	0,20%	79.815	167	0,21%
Apr 2018	92.316	160	0,17%	88.047	160	0,18%
May 2018	121.427	272	0,22%	115.150	272	0,24%
Jun 2018	119.981	230	0,19%	112.922	230	0,20%
Jul 2018	125.680	249	0,20%	117.255	249	0,21%
Aug 2018	119.609	239	0,20%	108.982	239	0,22%
Sep 2018	116.193	247	0,21%	103.900	247	0,24%
Oct 2018	121.184	273	0,23%	105.843	273	0,26%
Nov 2018	100.887	231	0,23%	85.174	231	0,27%
Dec 2018	108.216	247	0,23%	87.424	247	0,28%
Jan 2019	95.556	215	0,22%	72.510	215	0,30%
Feb 2019	84.352	253	0,30%	64.014	253	0,40%
Mar 2019	79.578	199	0,25%	62.536	199	0,32%
Apr 2019	94.982	253	0,27%	75.363	253	0,34%
May 2019	107.504	262	0,24%	85.087	262	0,31%
Jun 2019	125.868	376	0,30%	98.917	376	0,38%
Jul 2019	145.718	2.892	1,98%	113.994	2.892	2,54%
Aug 2019	143.304	4.170	2,91%	113.062	4.170	3,69%
Sep 2019	166.096	5.229	3,15%	134.717	5.229	3,88%
Oct 2019	177.342	5.960	3,36%	143.972	5.960	4,14%
Nov 2019	197.912	6.995	3,53%	160.764	6.995	4,35%
Dec 2019	200.265	7.407	3,70%	160.318	7.407	4,62%
Jan 2020	171.415	6.546	3,82%	134.079	6.546	4,88%
Feb 2020	173.131	6.809	3,93%	133.087	6.809	5,12%
Mar 2020	135.394	5.456	4,03%	102.489	5.456	5,32%
Apr 2020	156.463	5.418	3,46%	115.169	5.418	4,70%
May 2020	225.138	5.340	2,37%	174.067	5.340	3,07%
Jun 2020	264.134	3.218	1,22%	216.754	3.218	1,48%
Jul 2020	287.055	2.357	0,82%	239.976	2.357	0,98%
Aug 2020	372.939	2.526	0,68%	320.333	2.526	0,79%
Sep 2020	373.180	2.025	0,54%	327.813	2.025	0,62%
Oct 2020	390.674	1.875	0,48%	351.855	1.875	0,53%
Nov 2020	421.571	1.740	0,41%	384.723	1.740	0,45%
Dec 2020	449.991	1.885	0,42%	412.331	1.885	0,46%
Jan 2021	403.900	1.521	0,38%	371.299	1.521	0,41%
Feb 2021	421.368	1.456	0,35%	390.387	1.456	0,37%
Mar 2021	387.648	1.268	0,33%	361.276	1.268	0,35%
Apr 2021	395.604	1.206	0,30%	369.583	1.206	0,33%
May 2021	443.977	1.486	0,33%	415.422	1.486	0,36%
Jun 2021	326.988	1.211	0,37%	303.863	1.211	0,40%
Jul 2021	293.959	1.230	0,42%	274.588	1.230	0,45%
Aug 2021	348.859	1.469	0,42%	328.920	1.469	0,45%
Sep 2021	323.413	1.460	0,45%	306.388	1.460	0,48%
Oct 2021	364.471	1.578	0,43%	346.949	1.578	0,45%
Nov 2021	355.292	1.546	0,44%	339.922	1.546	0,45%
Dec 2021	326.148	1.417	0,43%	313.121	1.417	0,45%
Jan 2022	281.587	1.414	0,50%			

Feb 2022	258.392	1.399	0,54%
Mar 2022	201.823	1.127	0,56%
Apr 2022	186.572	1.050	0,56%
May 2022	200.165	1.098	0,55%
Jun 2022	161.083	823	0,51%
Jul 2022	140.327	692	0,49%
Aug 2022	139.773	592	0,42%
Sep 2022	118.730	439	0,37%
Oct 2022	118.928	329	0,28%
Nov 2022	98.497	165	0,17%
Dec 2022	83.166	87	0,10%

Appendix C

Plots of all variables

C.1 Distribution of all variables

C.1.1 Numerical variables

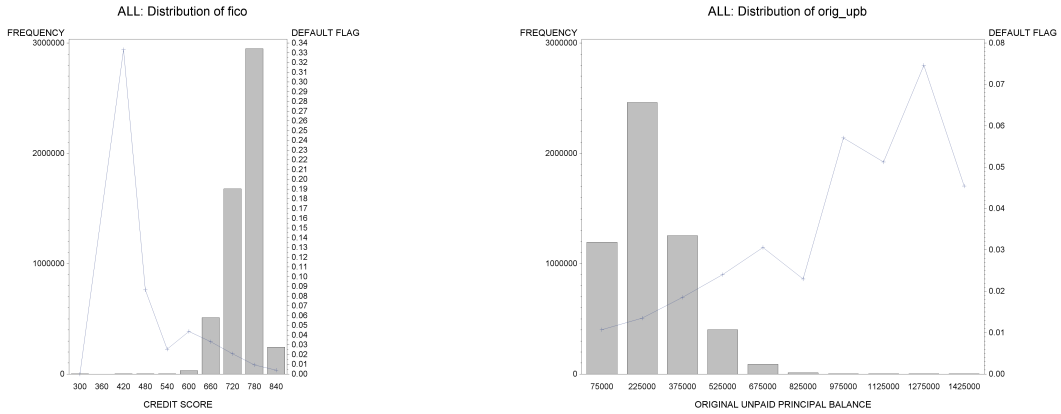

Fig. C.1: Distribution and default rate of Credit Score and UPB

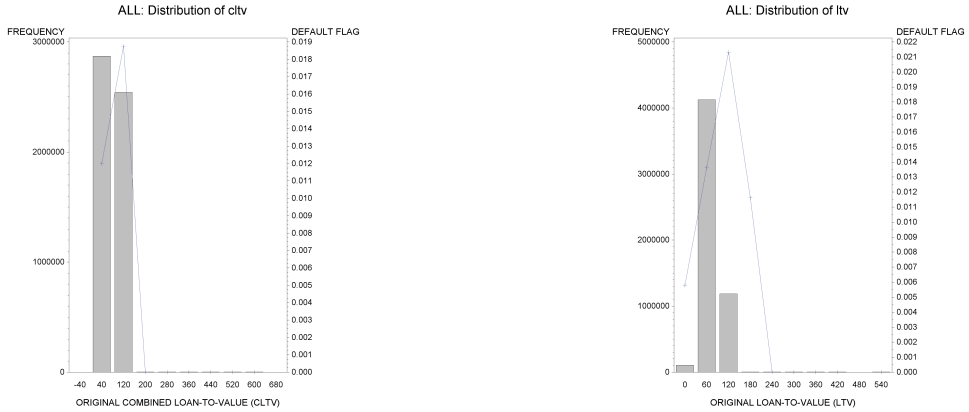

Fig. C.2: Distribution and default rate of CLTV and LTV

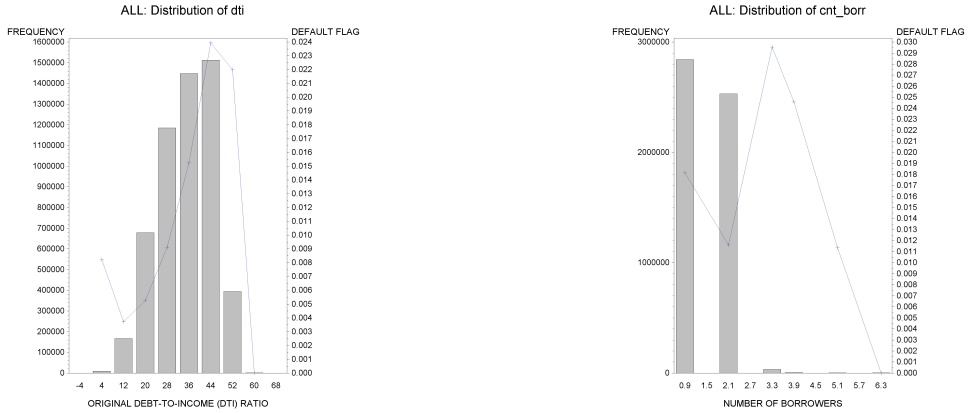

Fig. C.3: Distribution and default rate of DTI and No Borrowers

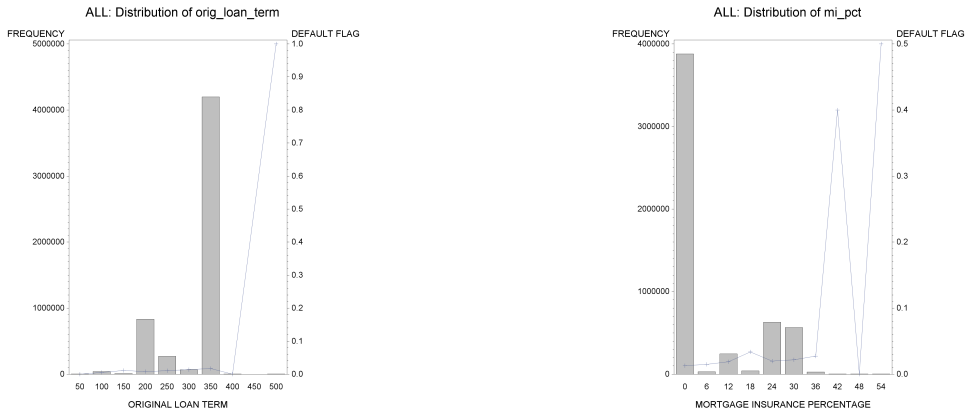

Fig. C.4: Distribution and default rate of Loan Term and MI Perc

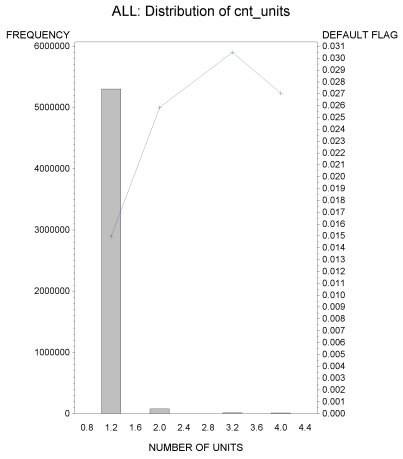

Fig. C.5: Distribution and default rate of No Units

C.1.2 Categorical variables

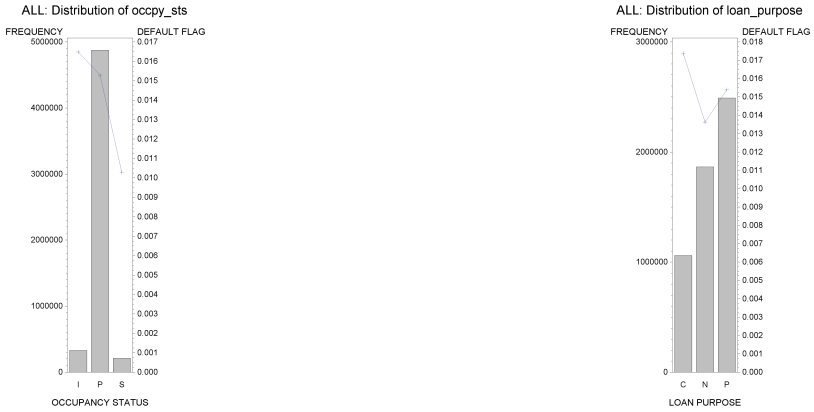

Fig. C.6: Distribution and default rate of Occupancy and Loan Purpose

Fig. C.7: Distribution and default rate of Prop Type and Valuation Method

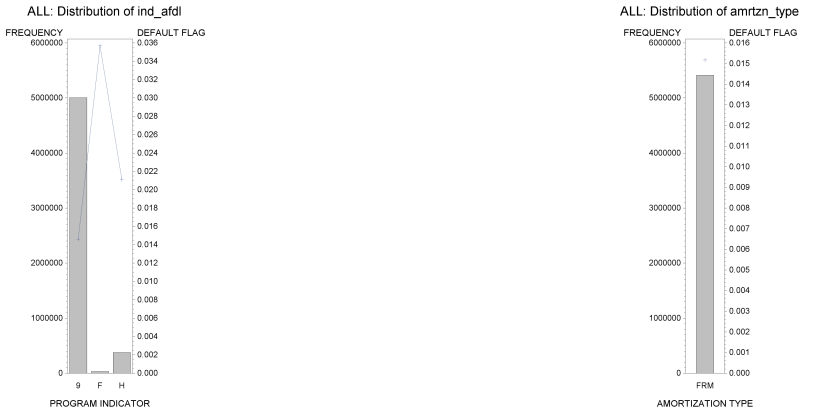

Fig. C.8: Distribution and default rate of Prog Flag and Amort Type

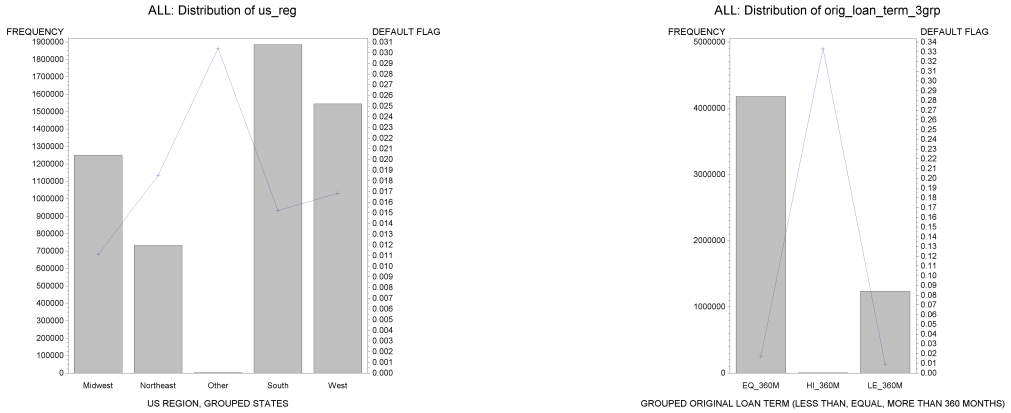

Fig. C.9: Distribution and default rate of US region and Loan Term (grouped)

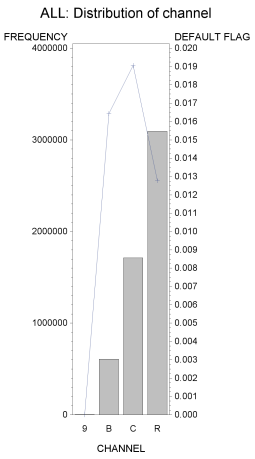

Fig. C.10: Distribution and default rate of Channel

C.1.3 Indicator variables

Fig. C.11: Distribution and default rate of Loan Term $\geq$ 360m and MI Flag

Fig. C.12: Distribution and default rate of Loan Term > 360m and Loan Term = 360m

Fig. C.13: Distribution and default rate of Sup Conf Flag and HARP Flag

Fig. C.14: Distribution and default rate of PPM Flag and Int Only Flag

Fig. C.15: Distribution and default rate of US region and Loan Term (grouped)

C.2 Boxplots of all numerical variables

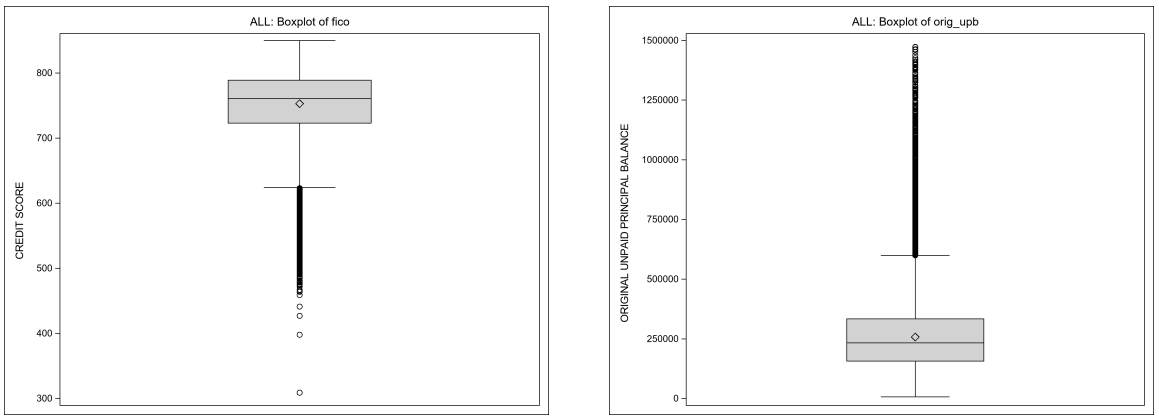

Fig. C.16: Boxplot of Credit Score and UPB

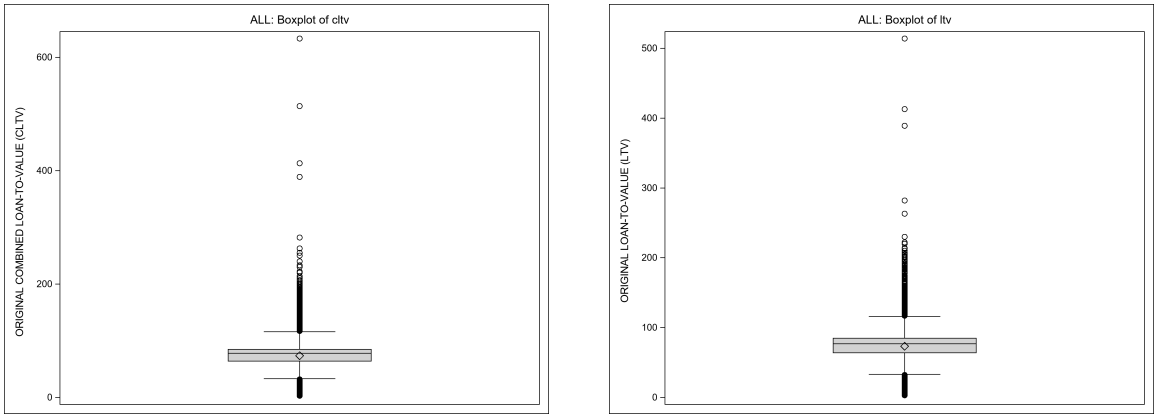

Fig. C.17: Boxplot of CLTV and LTV

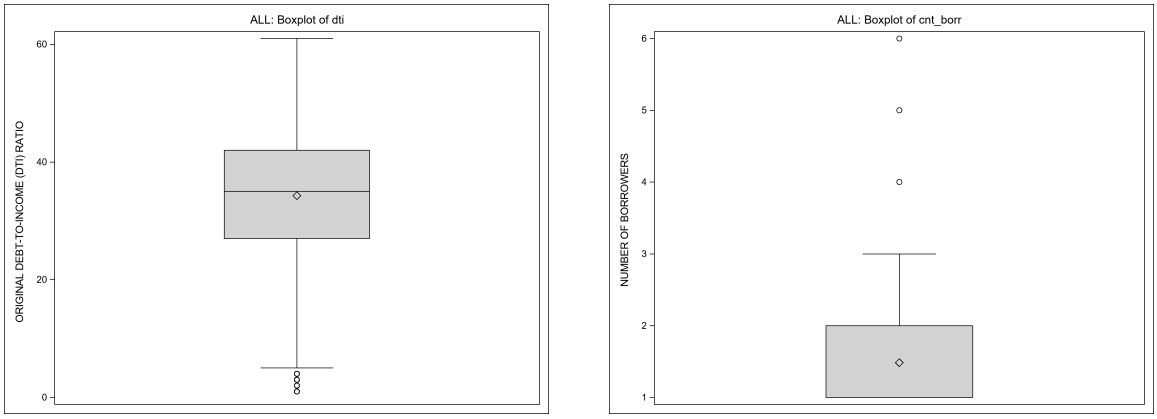

Fig. C.18: Boxplot of DTI and No Borrowers

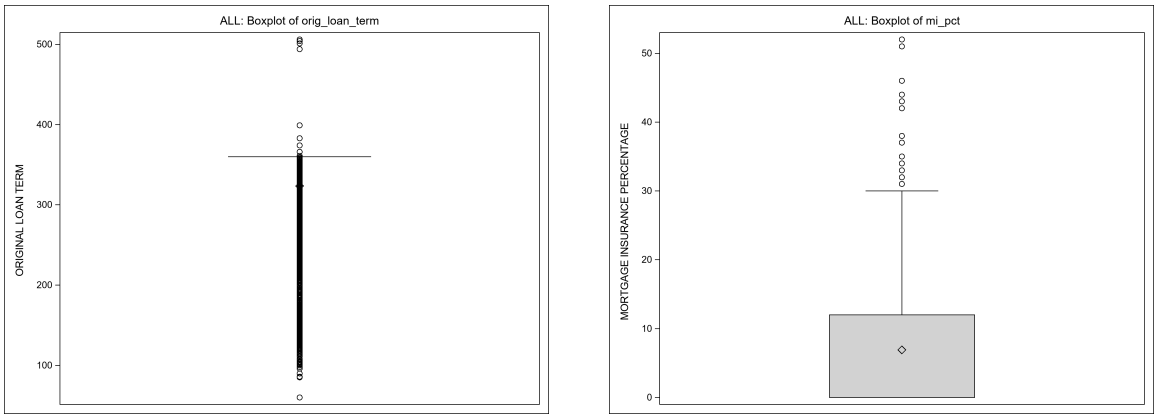

Fig. C.19: Boxplot of Loan Term and MI Perc

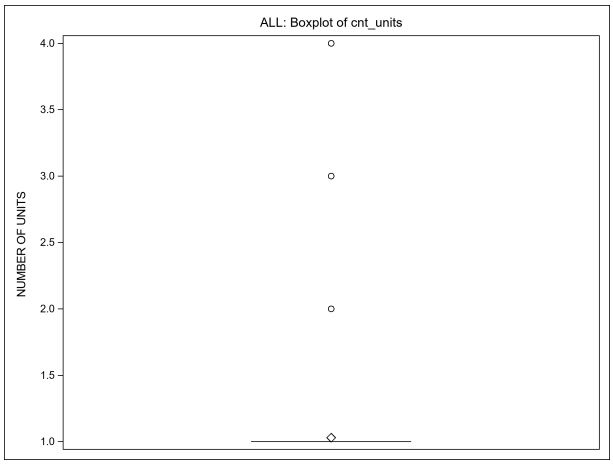

Fig. C.20: Boxplot of No Units

C.3 ROC-curves of all variables

C.3.1 Numerical variables

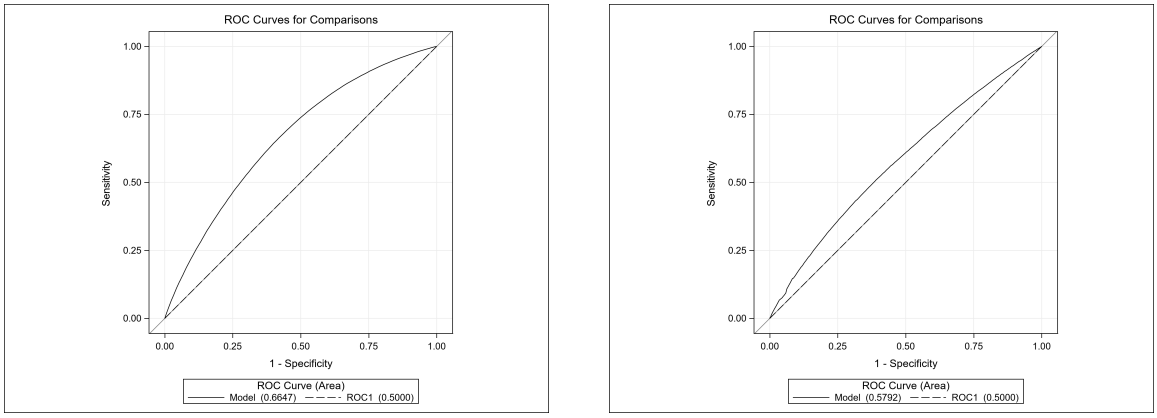

Fig. C.21: ROC-curve of Credit Score and UPB

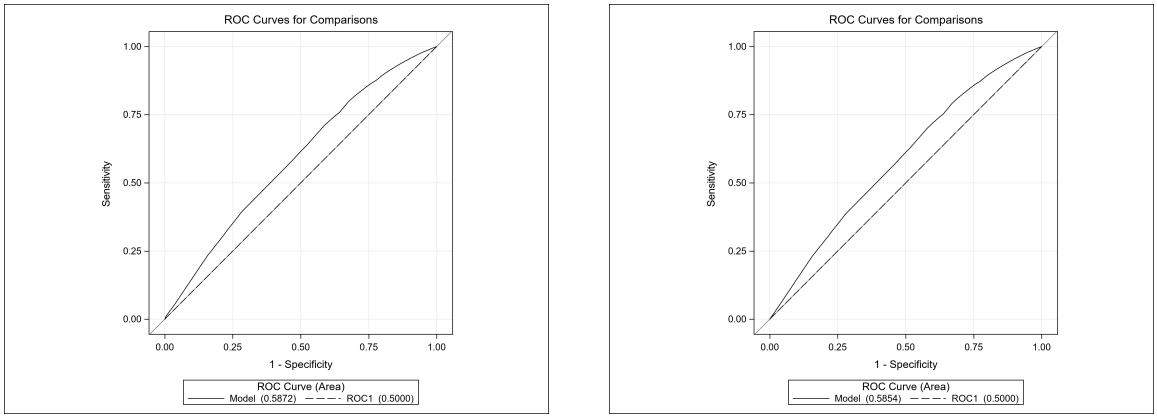

Fig. C.22: ROC-curve of CLTV and LTV

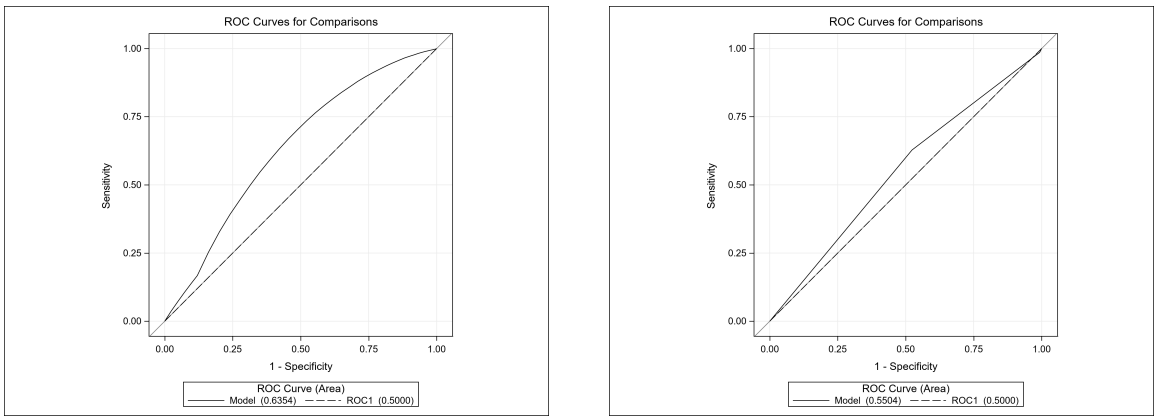

Fig. C.23: ROC-curve of DTI and No Borrowers

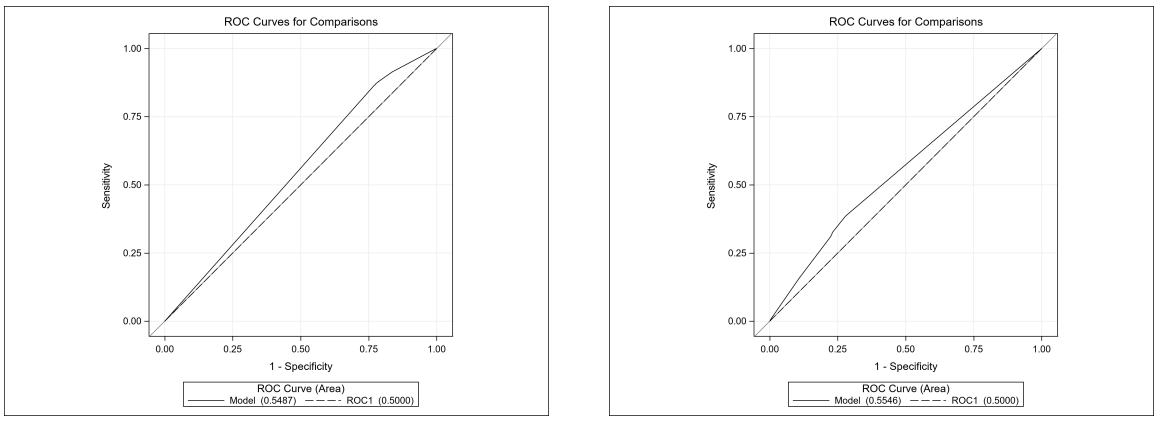

Fig. C.24: ROC-curve of Loan Term and MI Perc

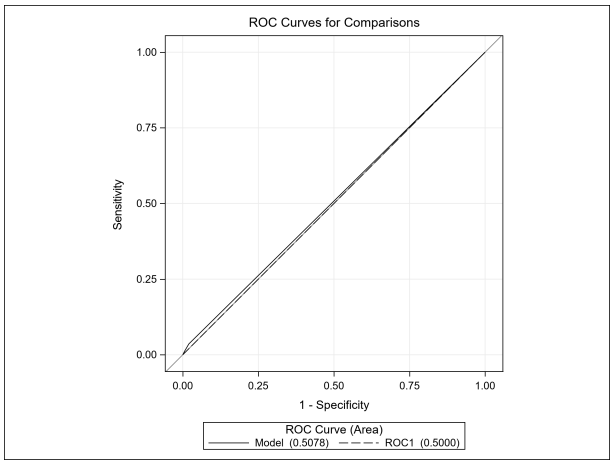

Fig. C.25: ROC-curve of No Units

C.3.2 Categorical variables

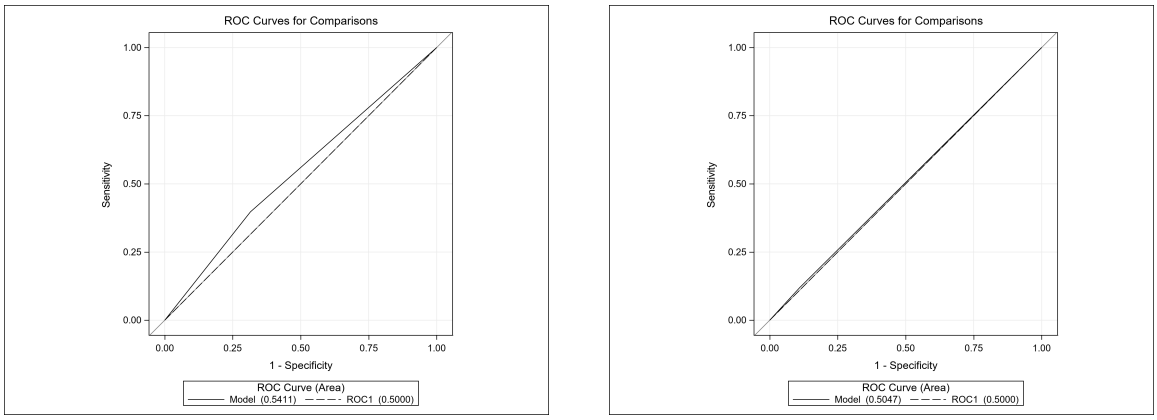

Fig. C.26: ROC-curve of Channel = C, B

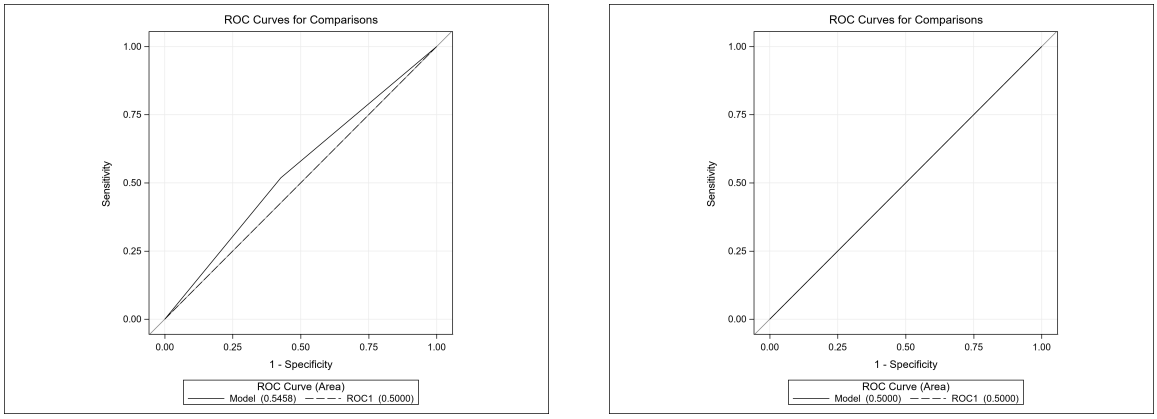

Fig. C.27: ROC-curve of Channel = R, Missing

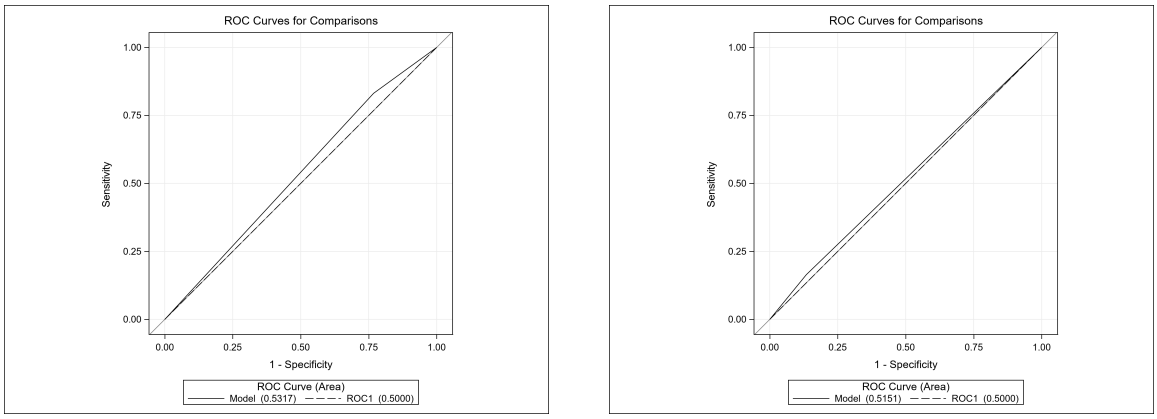

Fig. C.28: ROC-curve of US Region = Midwest, Northeast

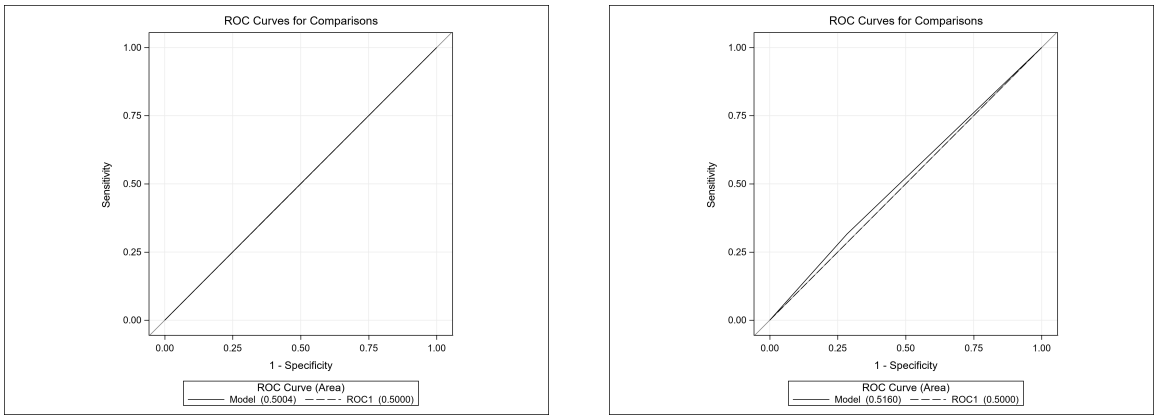

Fig. C.29: ROC-curve of US Region = South, West

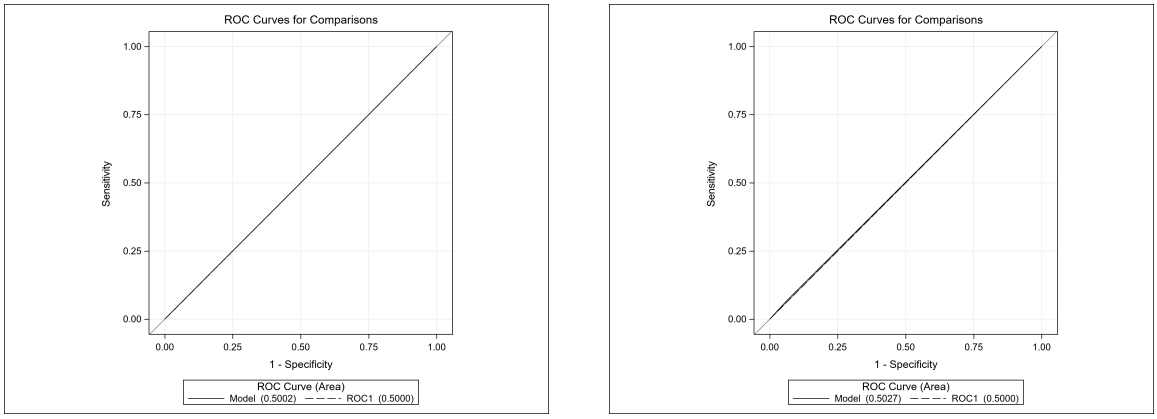

Fig. C.30: ROC-curve of US Region = Other, Occupancy = I

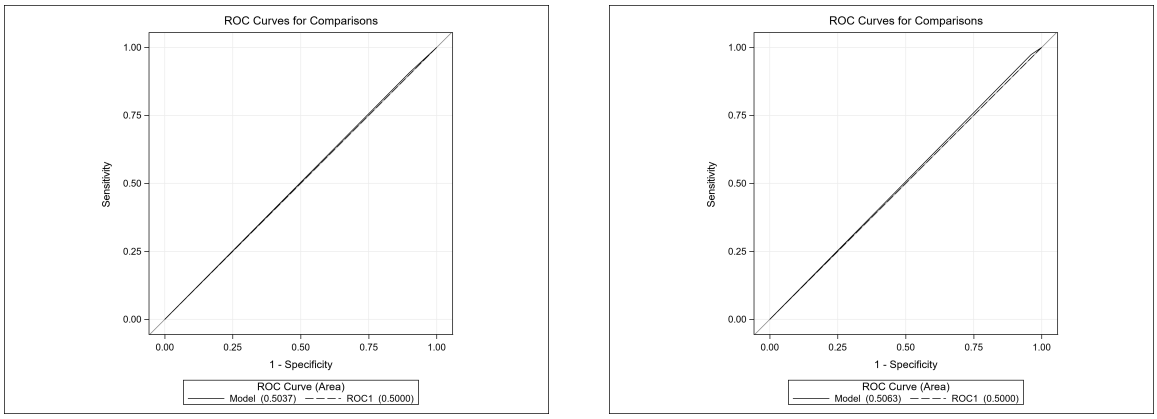

Fig. C.31: ROC-curve of Occupancy = P, S

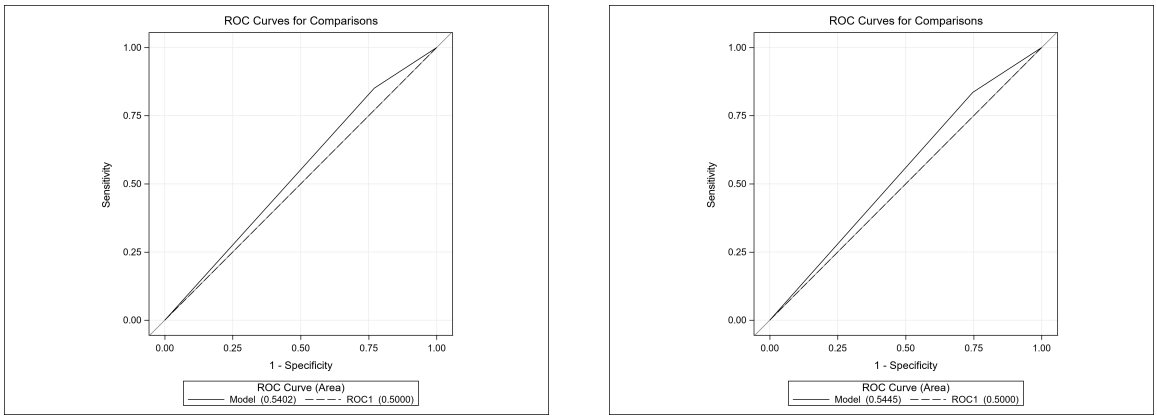

Fig. C.32: ROC-curve of Prop Val Method = 1, 2

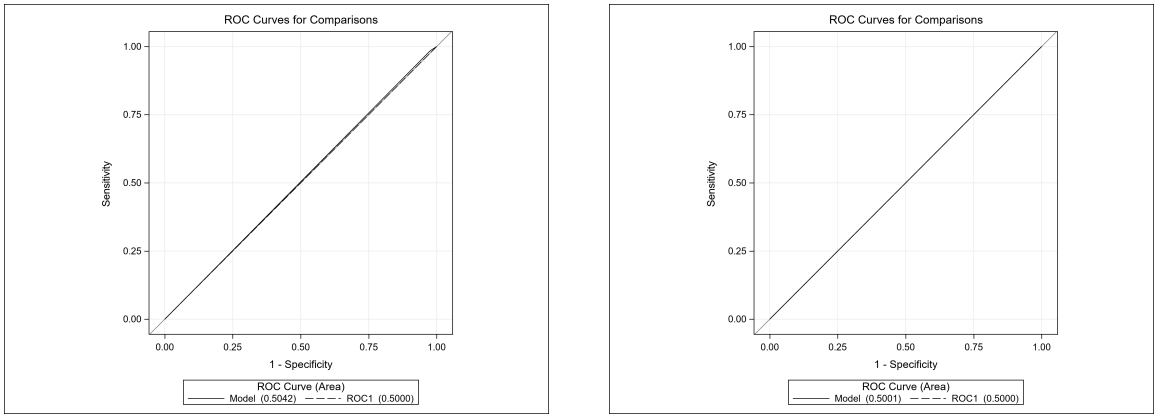

Fig. C.33: ROC-curve of Prop Val Method = 3, Missing

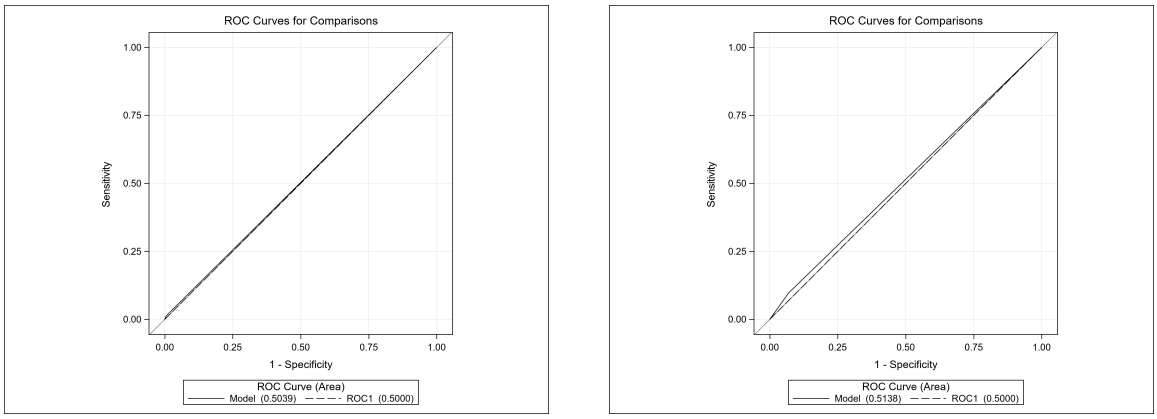

Fig. C.34: ROC-curve of Prog Flag = F, H

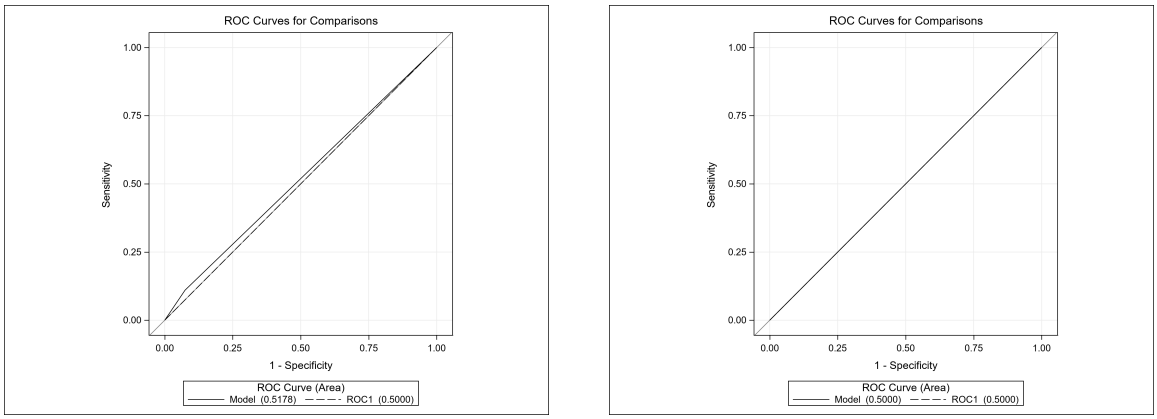

Fig. C.35: ROC-curve of Prog Flag = Missing and Amort Type = FRM

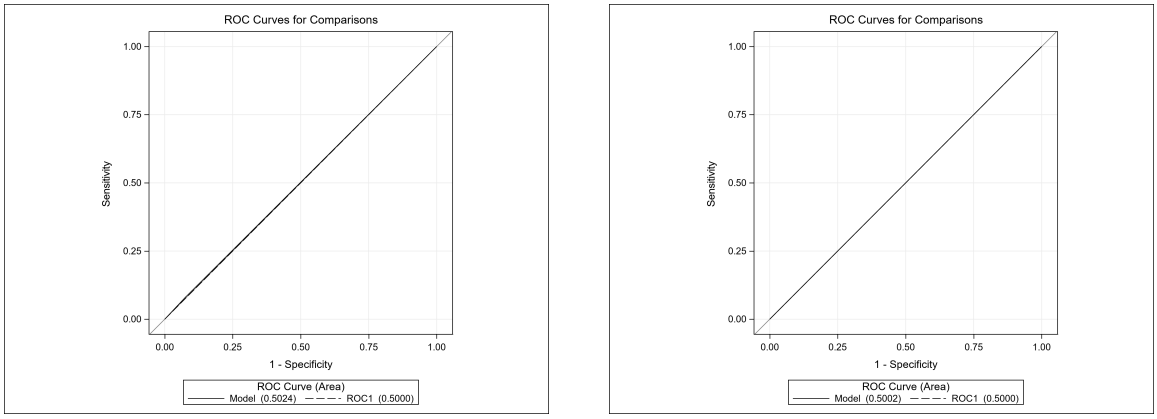

Fig. C.36: ROC-curve of Prop Type = CO, CP

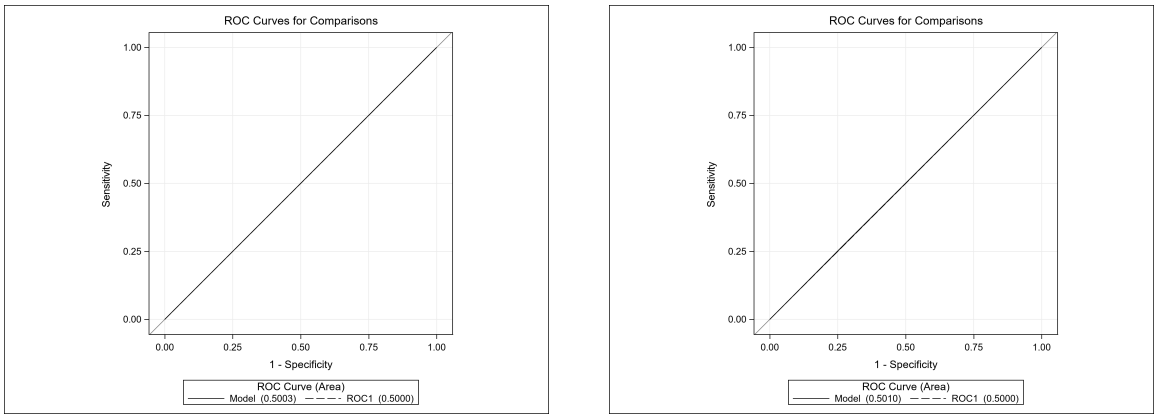

Fig. C.37: ROC-curve of Prop Type = MH, PU

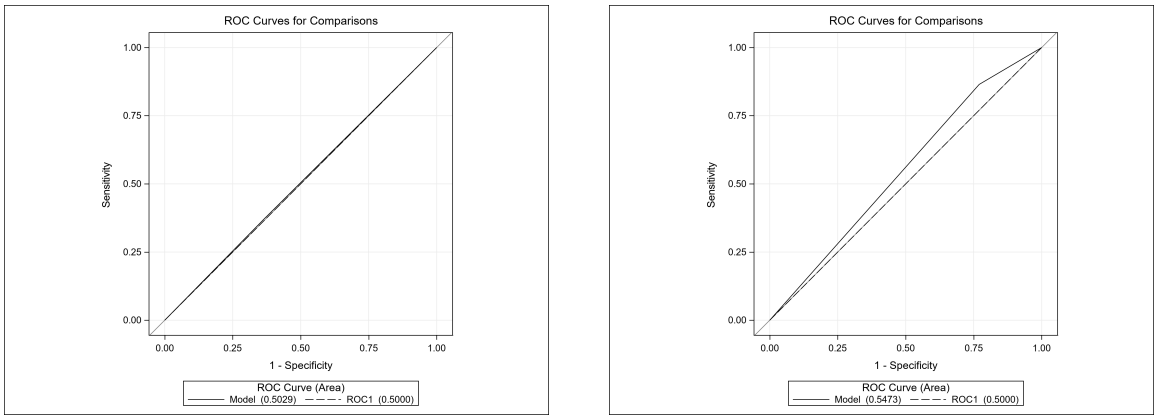

Fig. C.38: ROC-curve of Prop Type = SF and Loan Term = 360m

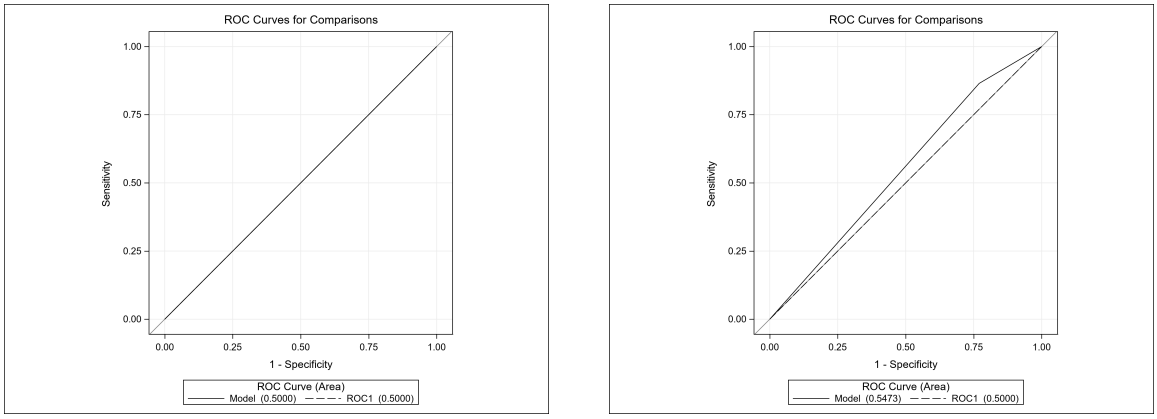

Fig. C.39: ROC-curve of Loan Term $>$ 360m, $<$ 360m

C.3.3 Indicator variables

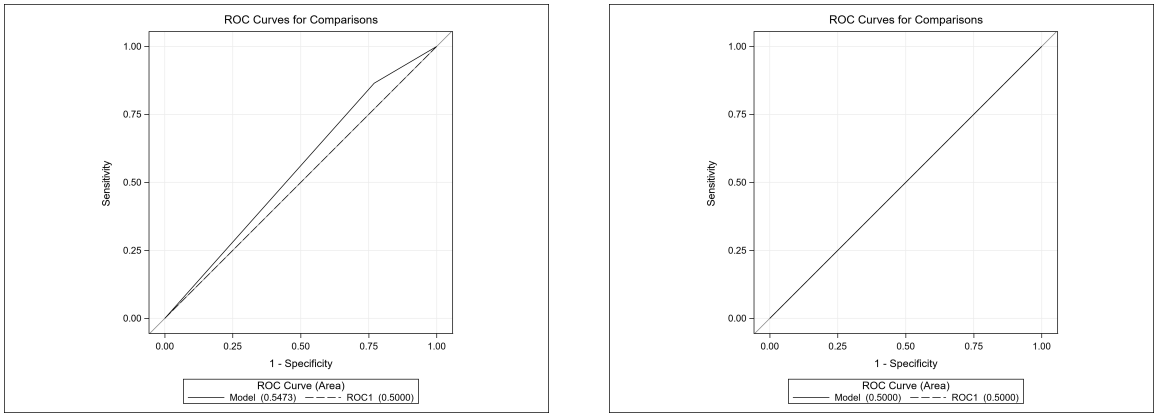

Fig. C.40: ROC-curve of Loan Term $\geq$ 360m and Flag Number of units

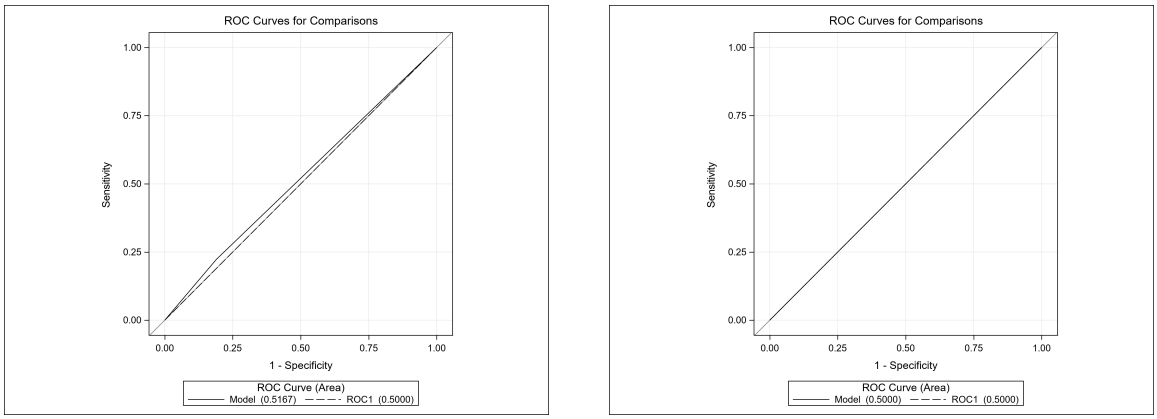

Fig. C.41: ROC-curve of Homebuyer Flag and Int Only Flag

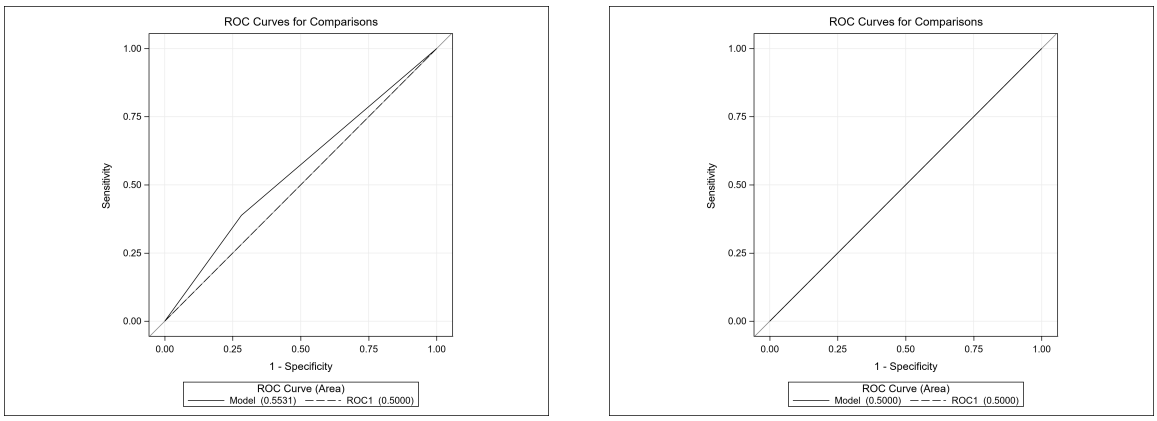

Fig. C.42: ROC-curve of MI Flag and Sup Conf Flag

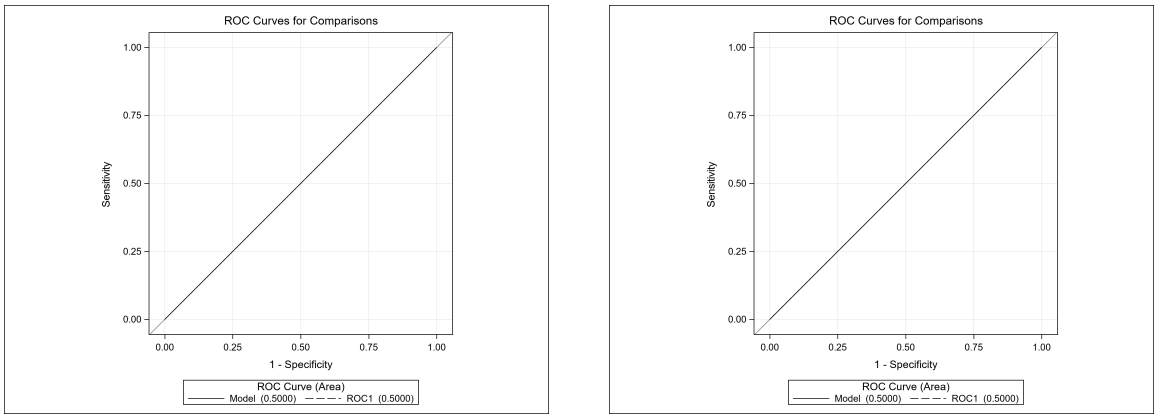

Fig. C.43: ROC-curve of HARP Flag and PPM Flag

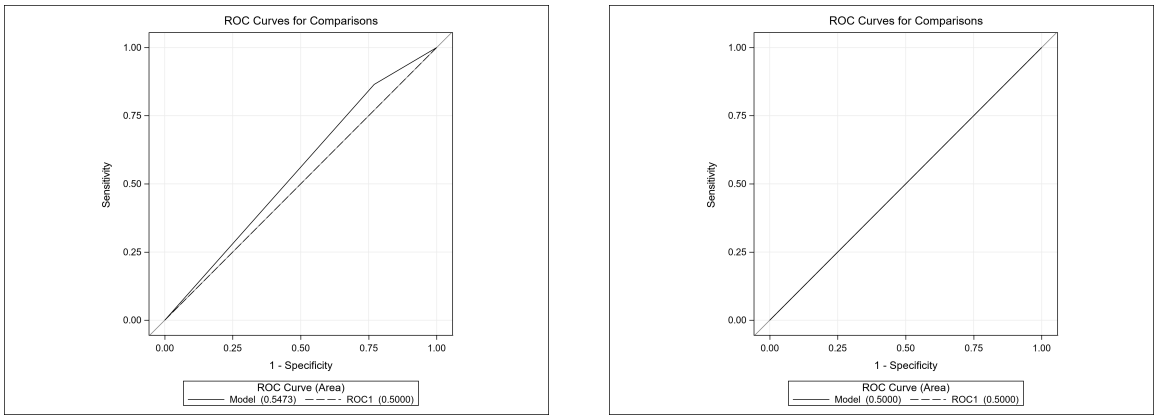

Fig. C.44: ROC-curve of Flag Loan Term = 360m and Loan Term > 360m

C.4 Partial Dependence Display of all model variables

C.4.1 Numerical variables

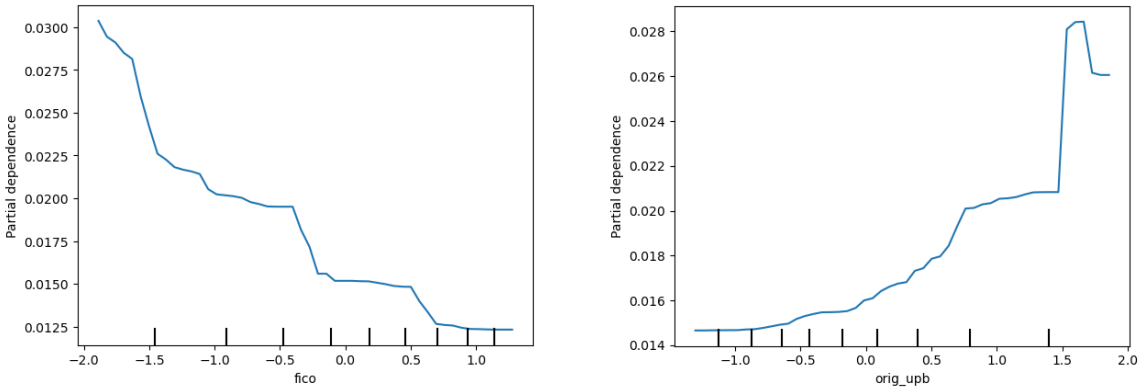

Fig. C.45: PDP of Credit Score and UPB

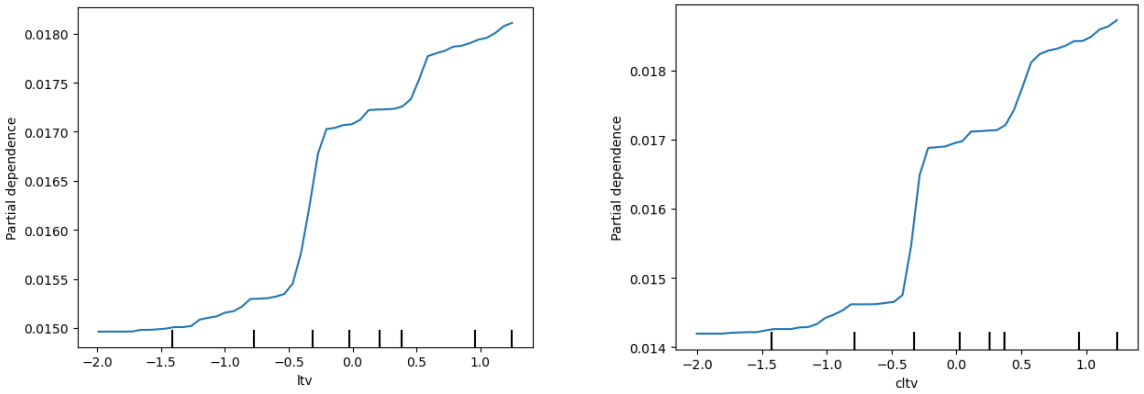

Fig. C.46: PDP of LTV and CLTV

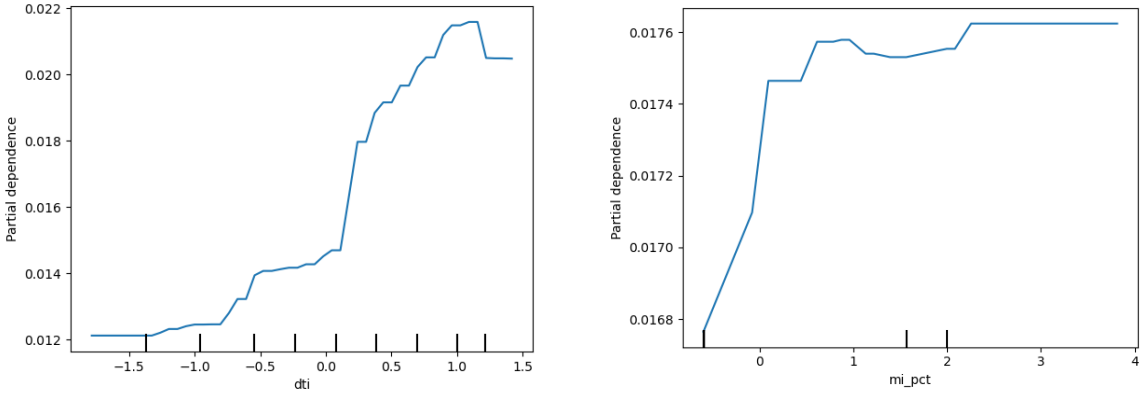

Fig. C.47: PDP of DTI and MI Perc

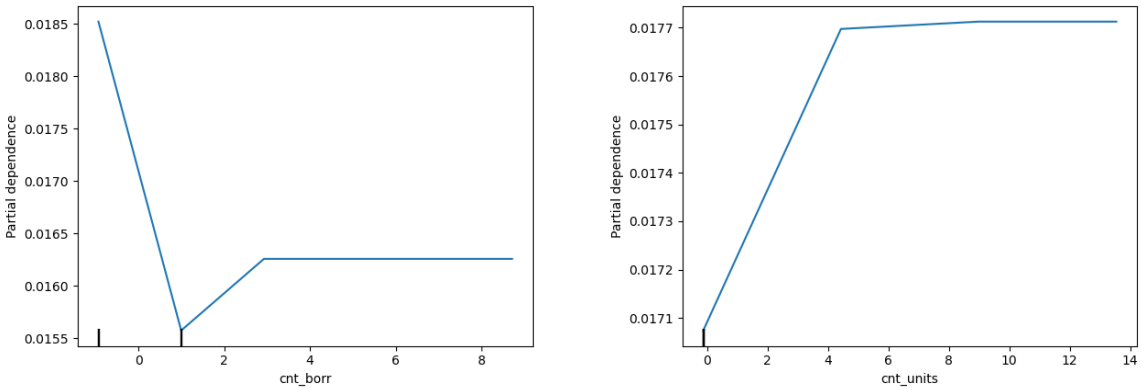

Fig. C.48: PDP of No Borrowers and No Units

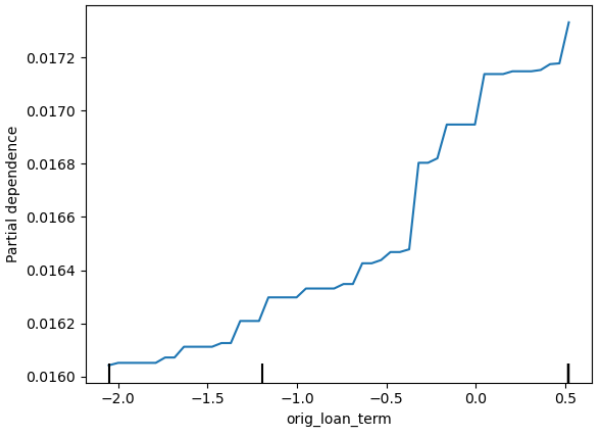

Fig. C.49: PDP of Loan Term

C.4.2 Categorical variables

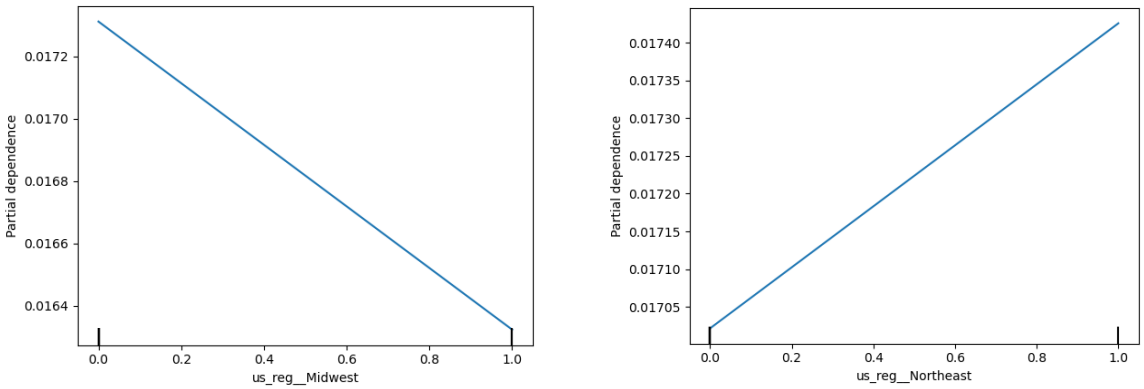

Fig. C.50: PDP of US Region = Midwest, Northeast

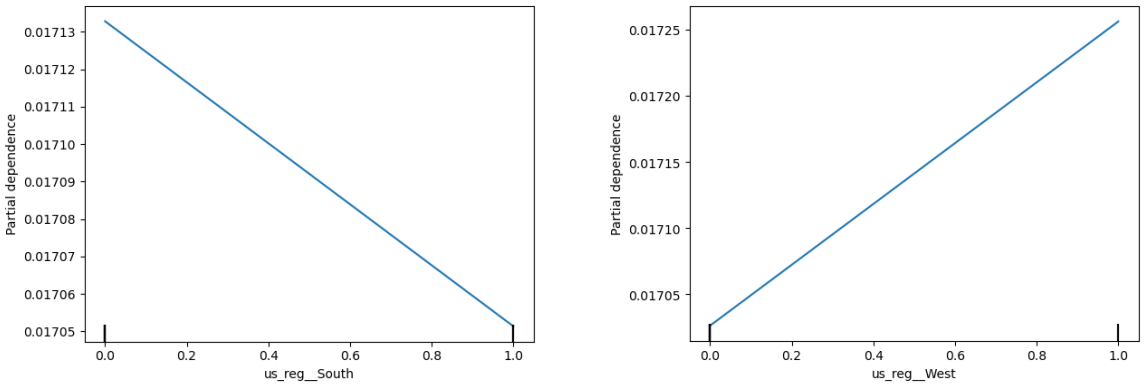

Fig. C.51: PDP of US Region = South, West

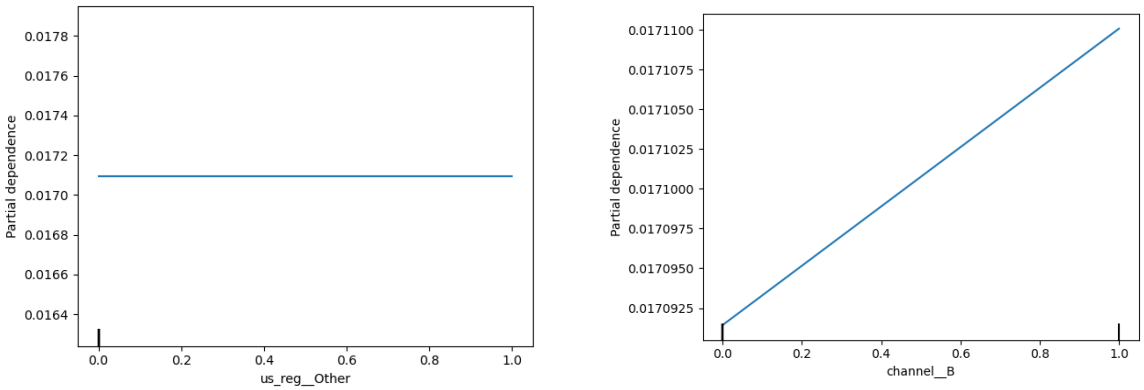

Fig. C.52: PDP of US Region = Other and Channel = B

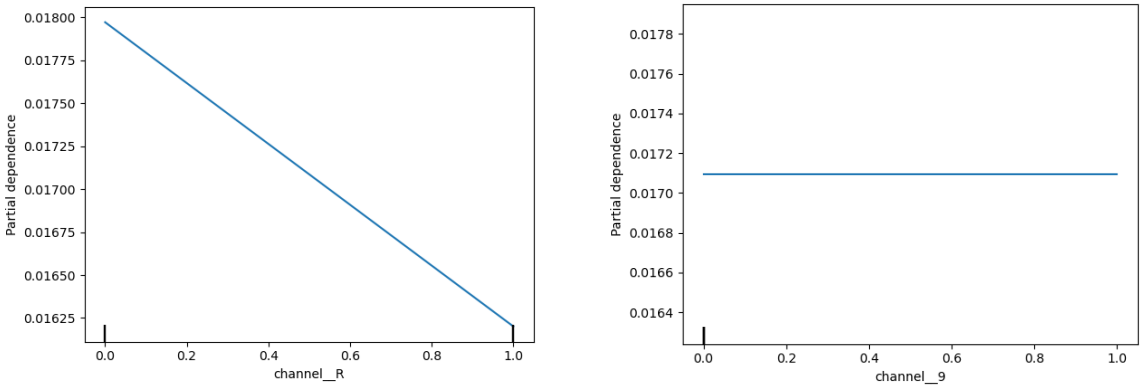

Fig. C.53: PDP of Channel = R, Missing

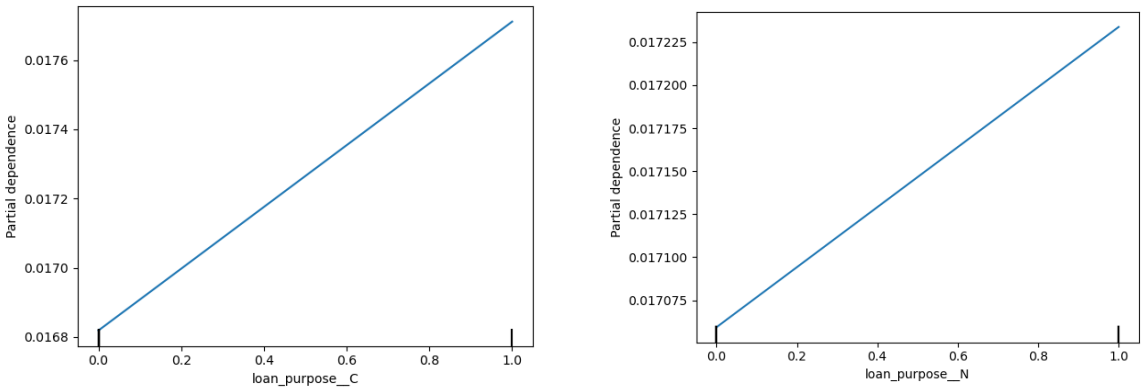

Fig. C.54: PDP of Loan Purpose = P = C, N

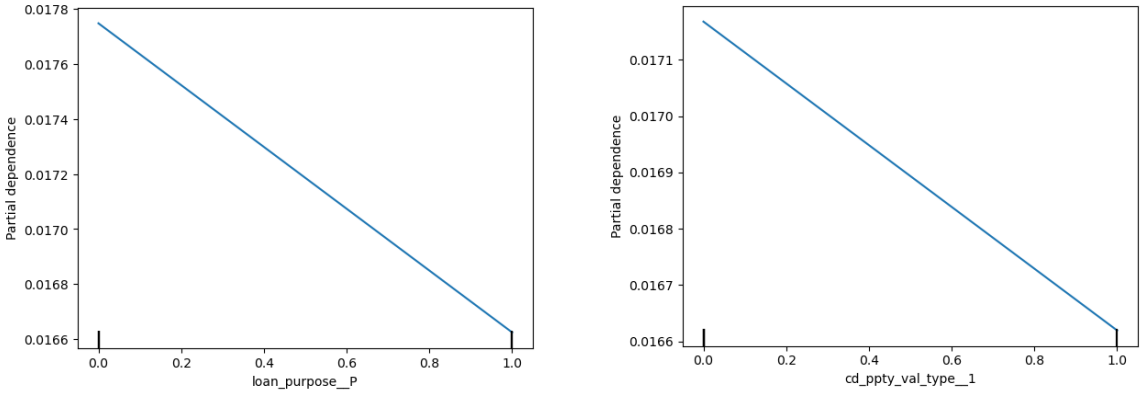

Fig. C.55: PDP of Loan Purpose = P and Prop Val Method = 1

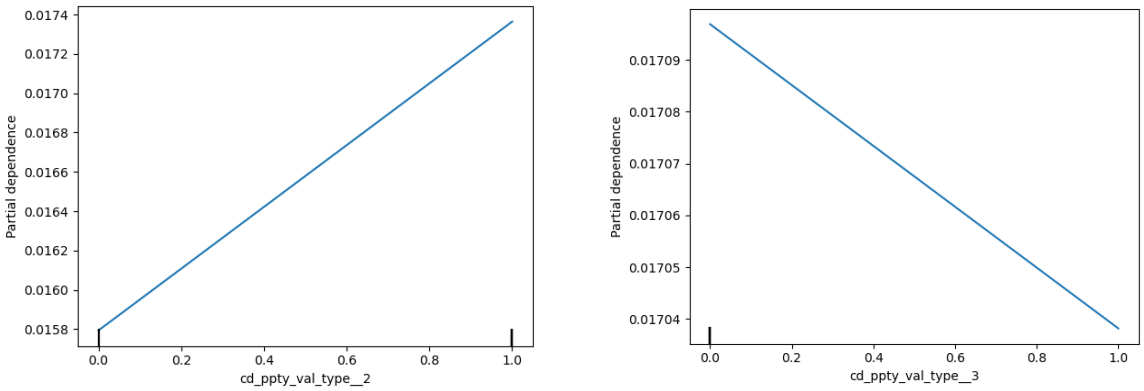

Fig. C.56: PDP of Prop Val Method = 2,3

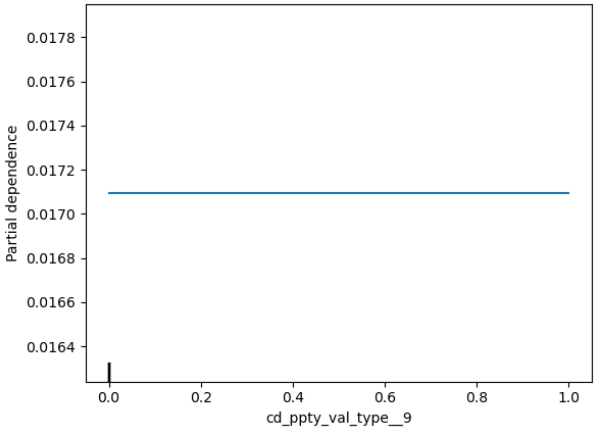

Fig. C.57: PDP of Prop Val Method = Not Available

C.4.3 Indicator variables

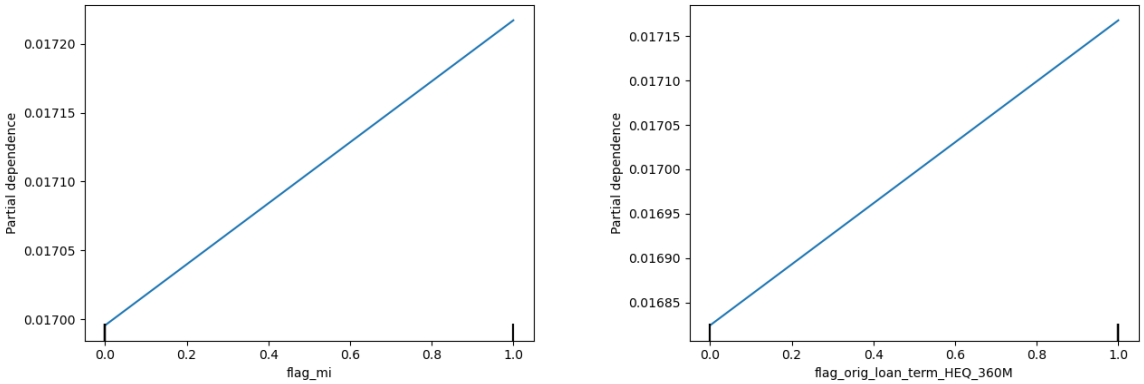

Fig. C.58: PDP of MI Flag and Loan Term $\geq$ 360m

Bibliography

[1] J. Witzany. *Credit Risk Management: Pricing, Measurement, and Modeling*. Springer International Publishing, 2017. URL https://books.google.at/books?id=YQ5cvgAACAAJ.

[2] Basel Committee on Banking Supervision. *International Convergence of Capital Measurement and Capital Standards*. URL https://www.bis.org/publ/bcbs107.pdf.

[3] Federal Deposit Insurance Corporation. *Economic Capital and the Assessment of Capital Adequacy*. URL https://www.fdic.gov/regulations/examinations/supervisory/insights/siwin04/siwinter2004-article01.html. Accessed: 2021-11-13.

[4] European Banking Authority. *Follow-Up Report from the Consultation on the Discussion Paper on Machine Learning for IRB Models*. URL https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf.

[5] T. Roberts and S. Tonna. *Risk Modeling: Practical Applications of Artificial Intelligence, Machine Learning, and Deep Learning*. Wiley and SAS Business Series. Wiley, 2022. URL https://books.google.at/books?id=tqdvzgEACAAJ.

[6] T. Atwan. *Time Series Analysis with Python Cookbook: Practical Recipes for Exploratory Data Analysis, Data Preparation, Forecasting, and Model Evaluation*. Packt Publishing, 2022. URL https://books.google.at/books?id=MmE6zwEACAAJ.

[7] M. Galarnyk. *Understanding Boxplots*. URL https://www.kdnuggets.com/2019/11/understanding-boxplots.html. Accessed: 2021-11-13.

[8] Y. Coadou. *Boosted Decision Trees and Applications*. EPJ Web of Conferences, 55:02004–, 07 2013. doi:10.1051/epjconf/20135502004.

[9] Simplilearn. *Random Forest Algorithm*. URL https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm. Accessed: 2021-11-13.

[10] Y. E. Khal. *Confusion matrix, AUC and ROC curve and Gini clearly explained*. URL https://yassineelkhal.medium.com/confusion-matrix-auc-and-roc-curve-and-gini-clearly-explained-221788618eb2. Accessed: 2021-11-13.

[11] P. Chism. *Freddie Mac: What Is The Federal Home Loan Mortgage Corporation (FHLMC)?* URL https://www.rocketmortgage.com/learn/freddie-mac. Accessed: 2021-11-13.

[12] A. Dehan. *What Is The Secondary Mortgage Market And How Does It Work?* URL https://www.rocketmortgage.com/learn/secondary-mortgage-market. Accessed: 2021-11-13.

[13] Federal Home Loan Mortgage Corporation. *Single Family Loan-Level Dataset*. URL https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset.

[14] Federal Home Loan Mortgage Corporation. *Single-Family Loan Level Dataset General User Guide*. URL https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset.