Masterthesis

Predicting the Probability of Default using the Random Forest Algorithm

Meikee PAGSINOHIN

Matrikelnummer: 01327477

Summary

Abstract

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 will then be used to decide if they want to decline or grant a credit application and among other things, for setting 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 afterwards. Corporate clients usually need to submit financial reports regularly, which is then analysed by their bank advisor and credit analyst, while for retail customers it is done automatically via behaviour scoring. The information used during the application scoring is limited because it is mainly provided by the applicant and generally covers variables about their financial health, e.g., income, outstanding debt. For the behaviour scoring model, internal historical data is used, for example the borrower's payment history and credit utilization. The behaviour model show a better predictive performance than the application model. If a decline of financial health or behaviour rating is detected, the bank may try to decrease the overdraft limit to regulate the credit risk. In the case of delayed payments, 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.

2.2 Default Rate and Probability of Default

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

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, a portrayal is visible in Fig. 2.1. The default rate per category (e.g., month, rating grade; given in Fig. **??**. and **??**.) is then calculated as the number of defaults divided by the total number of customers (Eq. 2.1).

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

where:

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

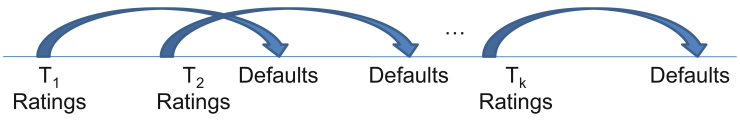

Fig. 2.1: Observation period after reference point

2.3 Regulatory Framework

2.3.1 History of regulatory framework

The regulatory framework is set by the Basel Committee on Banking Supervision, which consists 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 all over the world. The first Capital Accord was published in 1988 and since then was 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. In the end of 2010, a new reform called Basel III was approved.

2.3.2 Credit risk regulatory capital

The calculation of credit risk capital requirement was significantly improved compared to the first Capital Accord. Total loss of a bank can be split into expected and unexpected loss (Fig. 2.2). The former should be covered by revenue and for the latter a bank is obligated to allocate an appropriate level of capital. The formula is given in Fig. **??** In the original approach, each exposure has been assigned into one of four risk categories and then a multiplier ranging 0-100% was applied. Regulations now allow the Standardized (SA), Foundation or Advanced Internal

Rating Based (IRBF, IRBA) Approach (Fig. **??**). In the SA five risk buckets are defined to calculate the regulatory capital and the Standard Approach also allows the use of external ratings. For the IRB, internal models estimate input parameters of the regulatory formulas, which then result in risk weights for each exposure. The IRBF approach only allows the estimation of the PD, while for the IRBA the risk parameters Loss Given Default, Exposure at Default, Conversion Factor and Effective Maturity are additionally derived from internal models. The corporate segment permit both IRBF and IRBA approaches, but for the retail portfolio, only the IRBA is possible.

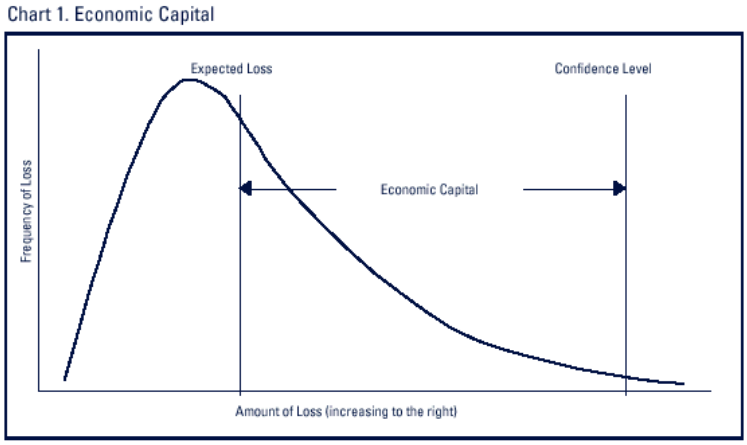

Fig. 2.2: Economical Capital, Expected and Unexpected Losses

2.3.3 Challenges and Limitations

Good data is of utmost importance for the credit risk assessment. While it will be used for modelling purposes, it is also important that already known negative information of customers is available and taken into consideration. Examples are internal information, for example a client already has a history of fraudulent activity during a credit application, or credit bureau information, where negative credit information is made available for all participants. In Austria, institutions are Kreditschutzverband (KSV) and CRIF.

Other possible challenges are the constant change in economic conditions and regulatory frameworks, which have an influence on the PD estimates. In addition, during the PD estimation the default event of individual borrowers are assumed to be independent, which does not adequately capture behavioural risks (e.g., strategic default) or systemic risks (e.g., market-wide shocks) that can affect multiple borrowers simultaneously. It is therefore important that PD models are continuously refined and adapted to accurately reflect current economic situation.

Chapter 3

Traditional models

3.1 Overview

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 analyse 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 type of PD models are only available for a limited portion of borrowers.

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

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

3.2 Logit and Probit regression

Logistic regression is one of the most commonly used statistical models in the banking industry. It is particularly useful 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 ranging between 0 and 1. A high model score means a lower probability to default and vice versa. For the link function the logistic function or standard normal cumulative distribution function can be used, resulting in the logit or probit model respectively (Fig. **??**, **??**). An advantage of the logit model is the heavier tails in the logistic distribution, which would therefore put higher weights to extreme events, visible in Fig. **??**.

3.3 Other models

3.3.1 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 both be continuous and categorical independent variables, for example income,

employment duration and profession (Fig. ??, Eq. ??). Unfortunately, due to the binary dependent variable, the residuals are heteroscedastic and therefore the estimation of the coefficients is inefficient. Additionally, the model may output non logical result, like negative values or a PD over 100%.

3.3.2 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, but also due to the type of information available. If the financial institution does not have enough resources to develop and maintain internal models, external ratings may be used. 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 years and additional information to derive a rating, which is then discussed in a rating committee. Afterwards, the corporation is informed about their rating, the corresponding factors, and given the opportunity to respond and finally the ratings will be published. A disadvantage of external ratings observed in the past is the conflict of interest, since the ratings are mainly paid by the company. 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.

3.3.3 Shadow Rating

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

Chapter 4

Machine Learning Models

4.1 Overview

4.2 Decision Trees

Classification trees are used to separate the categories (default, non-default) as best as possible using explanatory factors. The split is determine by maximising the homogeneity of the resulting subgroups (branch). For numeric variables, the algorithm calculates the measure for each possible threshold, while for categorical variables, it determines the value for each possible split. This step is repeated until a stopping condition is met and the final subgroup is called 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 Fig. 4.1. A Separate validation should then be performed on a third test data set, since the validation sample became part of the modelling process.

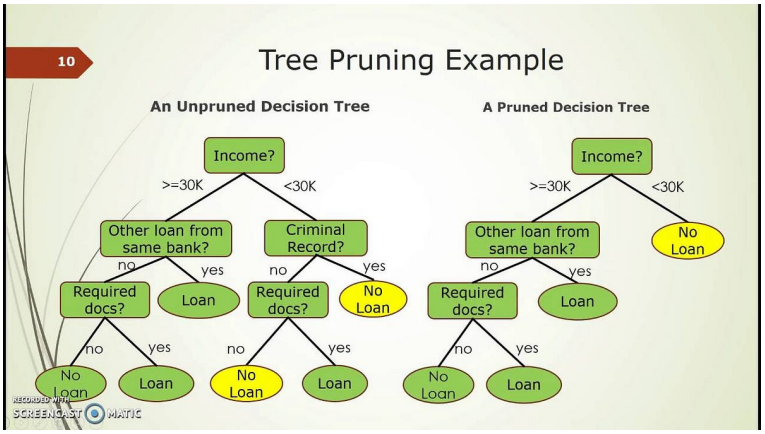

Fig. 4.1: Decision Tree incl. Pruning

Examples for stopping conditions are that the resulting subgroups are homogeneous, no significant improvement is detected, the minimum leafsize is reached, maximum splits are performed or maximum depth of the tree is achieved. The estimated PD is the average default rate per leaf and each terminal node can be classified into default or non-default using a defined threshold. Popular statistics to measure the homogeneity are Gini index, Kolmogorov-Smirnov statisic or 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.

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

where:

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

4.2.1 Boosted Decision Trees and Random Forests

Boosted Decision Trees (BDT), also known as Gradient Boosted Decision Trees combine decision trees with boosting techniques to achieve higher predictive performance. This algorithm iteratively build decision trees, placing more weight on misclassified observations in each iteration, resulting in a strong ensemble model, seen in Fig. 4.2. BDT are able to capture complex interactions and non-linear relationships in PD modelling.

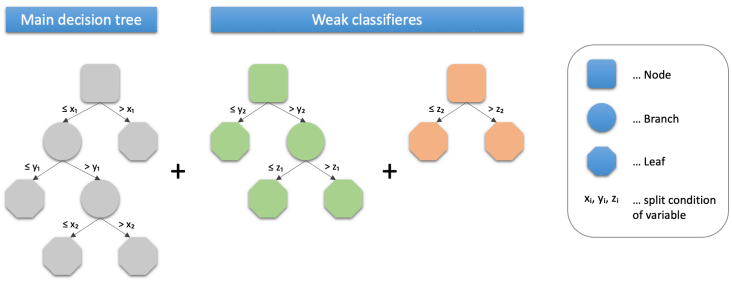

Fig. 4.2: Boosted Decision Tree

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 (Fig. 4.3). This approach reduce the risk of overfitting and the variance of predictions. Random forests are known for their robustness, scalability, and ability to handle high-dimensional data.

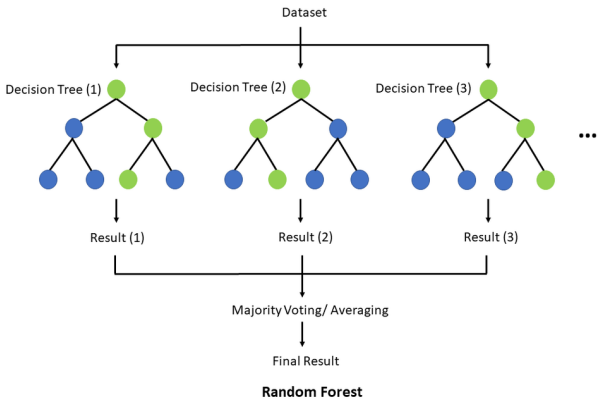

Fig. 4.3: Random Forest

4.3 Other models

4.3.1 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 use activation functions to pass along a value. The result of the model is a numerical or classification value. Commonly used activation functions are logistic function, threshold function or tangent hyperbolic function, listed in Eq. **??** - **??**.

The first and last layer is called the input and output layer respectively, and the layers in-between are 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 increasing number of hidden layers and nodes, also called deep neural networks. With the ongoing advancements in computational power and data collection capabilities, neural network algorithms have become highly successful in various AI domains.

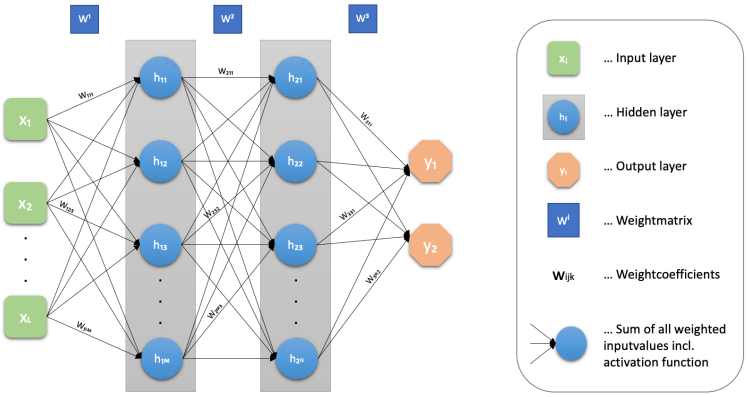

Fig. 4.4: Neural Network

4.3.2 k-Nearest Neighbour

Using a data set with explanatory factors and known default events, the unknown PD of a new data entry can be determined by taking the nearest data points determined via their risk factors and calculating the average default rate of the new data point. For the definition of distance the Euclidean metric or Manhattan Distance as defined in formula **??** and **??** can be used. The Euclidean distance measures the straight-line distance between two points, while the Manhattan distance is the sum of the absolute differences between two points in a space, where it is only allowed to move along coordinate axes. The number of nearest neighbours k is a hyper-parameter. Advantages of this model is the simple approach and the possibility to dynamically update new and outdated data entries and, if k is set as a low number, individual scorings can be viewed manually by a credit analyst.

4.3.3 Ensamble models

Chapter 5

Modelling process

5.1 Data Preparation

Explanatory factors can be split into numerical and categorical variables. They are mostly numerical variables in the corporate segment like financial ratios and macroeconomic conditions, while in the retail sector there are usually categorical factors, for example profession, marital status and residential status. Even numerical values like age or employment duration are often represented as categories after a binning process (e.g., "20-25 years", "25-25 years"). If there are too many categories or one category contains a very low portion of observations (rule of thumb: at least 5% per bucket), a merge of categories might be beneficial. One approach is to merge categories with a similar default rate or using measures like Weight of Evidence (WoE) or Information Value (IV). The formulas are given in Eq. 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. The IV indicates the ability of a variable to differentiate between default and non-default events - a higher IV relates to a higher discriminatory power and vice versa. As a final step, categorical variables need to be transformed into dummy variables to be used in the modelling process, meaning if the variable contains n distinct values, n-1 dummy variables will be created, illustrated in Fig. **??**. One dummy variable is omitted or else a linear combination is introduced during the modelling process.

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

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

where:

n = number of categories or buckets

5.1.1 Handling Missing Treatment

A common problem in real datasets are missing information, which need to be appropriately handled during the modelling process. One approach is to replace the missing values using statistical methods like mean, median or algorithm-based imputation, for example k-nearest neighbour Imputer. In this case the value is imputed with an average value of their k-nearest neighbours. The process of the kNN-algorithm is described in chapter 4.3.2. Data entries with missing information may also be removed from the data set, but this process can lead to information loss and bias. For categorical variables the missing information may also be viewed as separate category "Missing", therefore no adaption is necessary.

5.1.2 Erroneous Data Handling

Erroneous data, such as data entry mistakes or inconsistencies, can introduce noise and bias into the PD modelling process. Expert knowledge is crucial to detect erroneous data. The best way to reduce incorrect data are control procedures implemented in the data entry systems and data quality frameworks, where data validation rules are applied to identify inconsistent or illogical data. Those invalid information should then be corrected or removed by the associated expert using domain knowledge.

5.1.3 Outlier Detection and Treatment

Extreme values, also called Outliers, in the dataset, can significantly impact the estimated PD model. Expert knowledge is crucial in this domain as well. Variables resulting from ratio calculation are especially prone to outliers due to division with small numbers. A simple technique for outlier detection is the visual inspection, where the distribution is plotted and analysed (Fig. **??**), however, this would become strenuous with increasing number of variables. A quantitative approach is the utilisation of statistical measures such as Interquartile-range (IQR), box plots or Z-scores to detect outliers, depicted in Eq. **??** - **??**. A boxplot is also called box-and-whisker plot and it is a visual representation of the distribution and the spread of the variable. The box displays the IQR and the whiskers are the upper and lower limit. Z-Score is a measure of how many standard deviations a data point is from the mean of a dataset. After the detection, the problem can be handled using winsorisation, where all values above a certain threshold are set to the upper limit and all values below the lower limit are caped to that value to minimize the impact of outliers.

$$IQR = Q_3 - Q_1 \tag{5.3}$$

$$UpperLimit = Q_3 + 1.5 \times IQR \tag{5.4}$$

$$LowerLimit = Q_1 - 1.5 \times IQR \tag{5.5}$$

where:

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

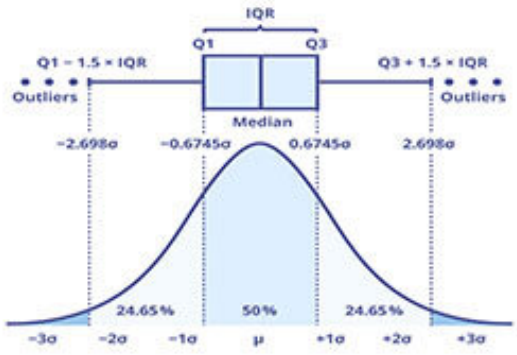

Fig. 5.1: Interquartile range and boxplot

5.2 Variable selection

During the modelling process, the goal is to estimate a model, which shows the best performance on in-sample as well as out-sample data. While using all available information in the scoring function will lead to a high discriminatory power on the trained sample, usually result to multiple variables showing insignificant coefficients, meaning the p-value is below the confidence level and the null hypothesis that the coefficient = 0 cannot be rejected. This would also mean, that the confidence level may be too wide and the sign of the coefficient might be incorrect. Therefore, the model will likely show a worse performance on a different data set, as well as on the newest data, leading to unstable predictions. This problem is handled by considerate selection of variables depending on their discriminatory power, correlation and expert judgement.

5.2.1 Univariate Analysis

All available variables should be considered for the modelling process. Then the missing rate, number of outliers and plausible values is assessed. If the variable complies with all data quality requirements, the discriminatory power will be determined. This can be performed using the univariate Gini coefficient or Information value, a more detailed explanation for both measures are in chapter 6.2.3 and 5.1. The remaining risk factors with satisfying discriminatory power make up the long list. 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

5.2.2 Multivariate Analysis

The model variables should not be highly correlated because this can lead to multicollinearity issues and therefore in unstable coefficient estimates in the modelling process. Correlation between two variables can be measured using the Pearson or Spearman correlation coefficient (Eq. 5.6, 5.7). The Pearson correlation coefficient is more appropriate if a linear relationship is present. Spearman correlation coefficient is able to detect non-linear, monotonic interactions and is also more robust against outliers. The coefficient ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. After calculating all coefficients between each variable pair, all values can be arranged to a correlation matrix. An example is visible in Fig. 5.2.

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

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

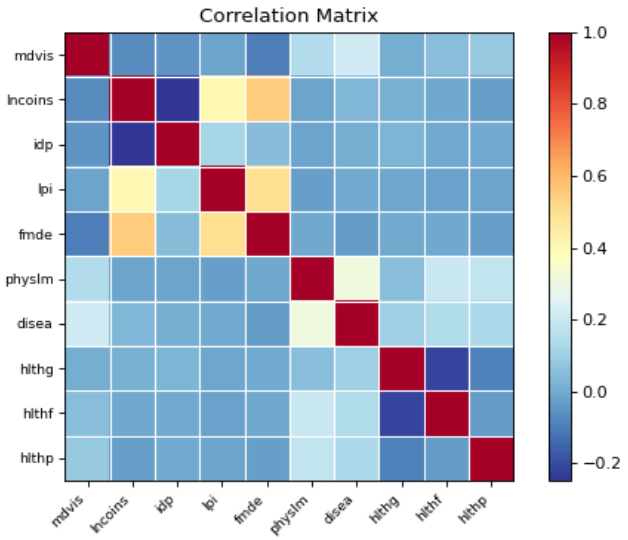

Fig. 5.2: Correlation matrix example

If two variables are highly correlated with a correlation coefficient above a certain threshold, the variable with the lower discriminatory power should be removed from the model list. For categorical variables, the correlation statistic Variance Inflation Factor (VIF) can be utilised (Eq. 5.8), which measures the collinearity in the regression analysis. Finally, adjustments based on expert judgement should be applied, where disqualified variables are forced into the list or variables, even if they meet all requirements, are removed. Especially the relationship between the explanatory factor and default rate should be analysed. If the behaviour contradicts the economic reasoning or experience, the variable should be eliminated from the model list, also called short list.

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

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

5.3 Modelling steps

After selecting the most promising key factors, there are different approaches to determine the final model variables - the forward, backward, forward stepwise and backward stepwise selection procedure. In the first approach there are no variables in the model at the beginning and step by step the variable with the highest discriminatory power is added if the coefficient is significant (p-value below threshold). This step is repeated until no variable meets the condition. In the backward selection procedure, all variables are put into the model and then removed one by one depending on the coefficients' p-value, until all coefficients left in the model are significant.

The forward stepwise procedure is a combination of both, where the model is first empty and variables are added successively, but after each addition, variables, which became insignificant, will be removed and the backward stepwise procedure is the opposite process. An procedures are illustrated in Figure **??** - **??**.

5.4 Rating grades

It is common to group the PDs into different rating bands such as investment and non-investment grade. The simplest approach is to use fixed ranges per grade. Another approach is to define PD intervals. An example is displayed in Fig. **??**.

Chapter 6

Validation

During the model validation process the model's performance and consistency is tested. This helps to detect any potential weaknesses or shortcomings in the model and provides an opportunity for improvement.

Important components of model validation are:

- Data Quality Assessment: Evaluating the quality, completeness, and reliability of the data used to build the PD model, including their data sources.
- Model Performance Evaluation: Assessing the model's predictive power and discrimination ability, which is usually analysed by the following metrics: Area Under the Receiver Operating Characteristic curve (AUC-ROC), accuracy, precision, recall, and F1-score.
- Sensitivity Analysis: Analysing the impact on the result of the model for different changes in the input variables. This helps identifying the most influential factors and potential vulnerabilities of the model.
- Robustness Testing: Examining the model's stability and performance under different scenarios and stress conditions. This involves stress testing the model with extreme cases or outlier data points to assess its resilience and reliability.

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

Out-of-sample validation involves testing the model's performance on a data set that was not used during model development. Out-of-time validation goes one step further, it includes new data which covers additionally a different, more current time period. An illustration is depicted in Fig. 6.1. The purpose is to assess the model's ability to make accurate predictions on new, unseen data. Because the model is estimated on the training sample, it is common that the model shows a better performance on in-sample data sets than on another sample. However, if the performance metrics differs significantly, this could be seen as a sign of over-fitting.

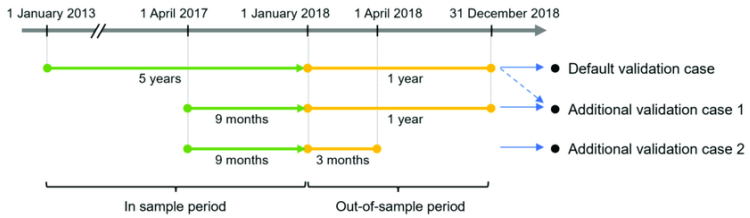

Fig. 6.1: Selection of out-of-sample and out-of-time sample

The full data set is usually split into 70% training and 30% testing sample, also called validation sample. Different ratios, e.g. 60/40, 80/20, are also popular and mainly depends on the size

of the data set and number of default events. During the splitting process it is important to keep the number of default events even in both samples to prevent situations, where one class is disproportionate to the other. This process is called stratification, visible in Fig. **??**. It ensures that there will be no biased model training and evaluation.

6.2 Model Performance Evaluation

Popular metrics to assess the discriminatory power of rating systems are Confusion Matrix, Area Under the Receiver Operating Characteristic Curve (AUC-ROC Curve), Gini coefficient, Cumulative Accuracy Profile (CAP) and its summary index, the Accuracy Ratio (AR).

6.2.1 Confusion matrix

The confusion matrix is a table with four elements (Fig. 6.2), it shows the number of observations which have been correctly (True Positive, True Negative) and incorrectly (False Positive, False Negative) identified as default or non-default.A False Positive, meaning a customer was predicted to default but survived, is also called Type I Error and aFalse Negative, thus a borrower is expected to survive but defaulted, is also known as Type II error. In practice, a Type II error is more severe, because the loss caused by a defaulted exposure is higher than the lost opportunity income due to a rejected non-defaulted application. For the transformation from PD to a default flag a cut-off has to be selected. To determine an ideal cut-off, the F1-Score can be utilized, where the cut-off value is set where the F1-score is maximised.

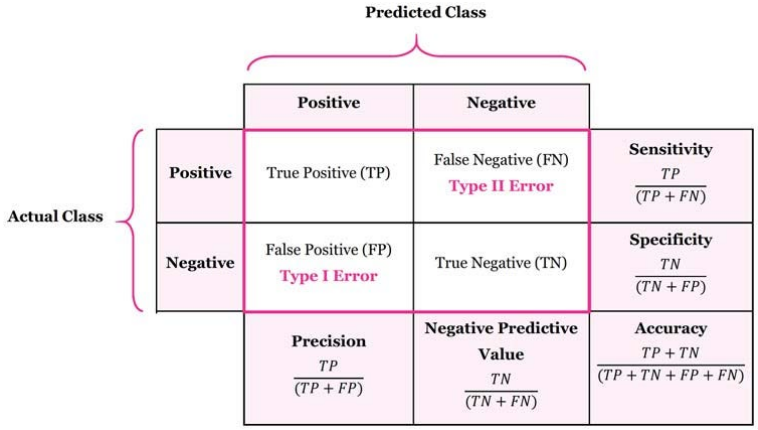

Fig. 6.2: Confusionmatrix

Using the elements of the confusion matrix, the measures Accuracy, Precision, Recall, F1-Score and others can be calculated, visible in Eq. 6.2 to 6.7. However, measures like Accuracy and Precision are not recommended for unbalanced data because they can provide misleading insights about model performance. In the case of unbalanced data, where one class is significantly larger, a model can achieve 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. Recall and F1-score provide a more accurate view.

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

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

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

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

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

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

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

6.2.2 Receiver Operating Characteristic Curve

The Receiver Operating Characteristic Curve (ROC Curve) 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 and the curve of a perfect model would be a step function that goes from (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. The areas are marked as A and B in Fig. 6.3 and the formula is seen in Eq. 6.8.

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

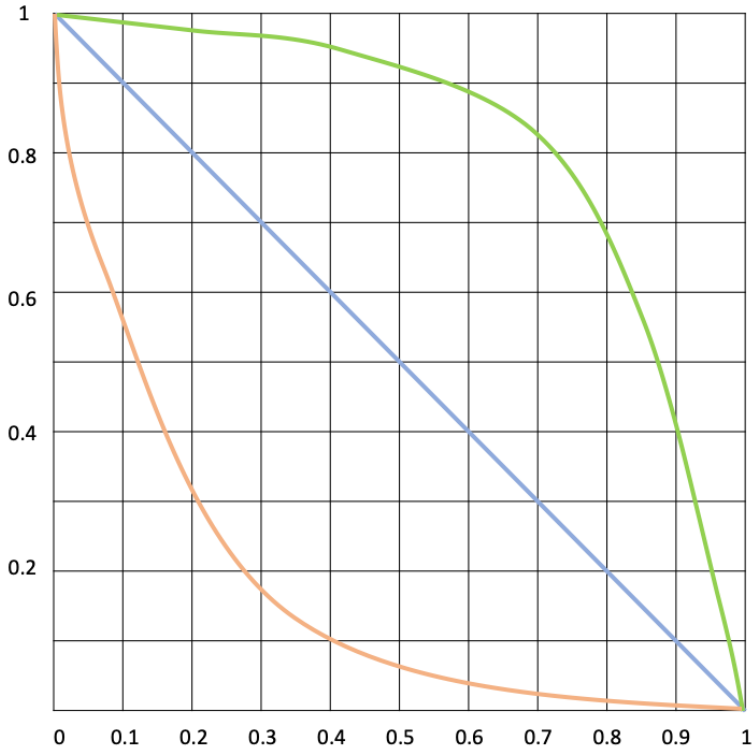

Fig. 6.3: ROC curve

6.2.3 GINI coefficient

The Gini coefficient and AUC are connected via the formula Eq. 6.4 and a visual presentation is visible in Fig. 6.4. Therefore, they relate the same information but are differently scaled. 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). Generally, the AUC can also take on a value below 0.5, but that would indicate, that the model's predictions is worse than the random model and therefore imply an issue in the model's ability to differentiate between the classes.

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

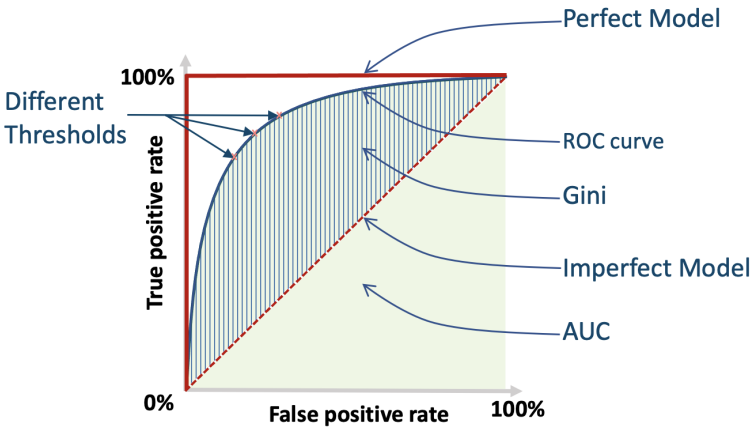

Fig. 6.4: GINI coefficient

6.2.4 Cumulative Accuracy Profile

In the CAP-graph the percentage of all borrowers are plotted along the x-axis and the percentage of defaulted borrowers are plotted along the y-axis (Fig. 6.5). The resulting curve is a way to assess how well the model differentiates between the two groups, comparing the performance to the perfect or random model. The Accuracy Ratio is the ratio of the area between the analysed and random model divided by the area between perfect and analysed model (Eq. 6.10).

$$AccuracyRatio = \frac{A}{A+B} \tag{6.10}$$

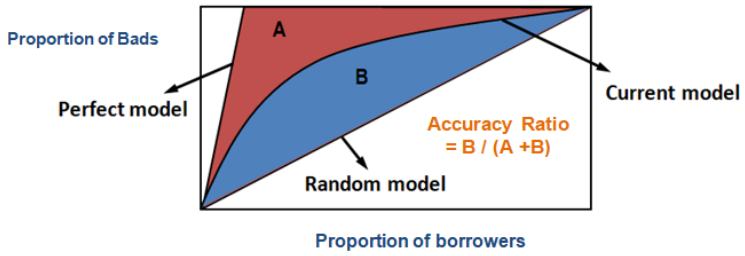

Fig. 6.5: Cumulative Accuracy Profile

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 to identify 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 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.

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 therefore require model developers and users to provide explanations for credit-related decisions to their customers. Modellers, internal and external audit are obligated to validate the model structure and their result, whether the model aligns with domain knowledge and expectations. Interpretability helps identify potential biases, data issues, or model limitations. Additionally, a model, which is unexplainable but is used in production, increases operational risk due to the difficulty to assess possible consequences (bias, fariness) and if the result was correctly calculated or contains a system error. To avoid the limitation of regulation requirements and consumer protection laws, machine learning models can be employed in areas, where the model structure and output are not of the highest priority, such as collection process or fraud detection.

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 minority groups should not be over- or underrepresented. If the data used during the training phase or in production are not corrected and validated, this can lead to unexpected results or to a biased model. Machine learning algorithms are able to amplify the errors, as a popular saying goes "Garbage In - Garbage Out".

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 applied on global or local level, which means on dataset or data observation level respectively. Depending on which aspect should be analysed, the techniques can be allocated into five categories: feature importance, input variable impact, specific prediction analysis, output analysis and robustness check.

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 while changing the value of a variable while holding other risk factors constant implies the importance of that feature. Relative feature importance compares the importance of features relative to each other w, which helps prioritize features based on their impact on the model's predictions. The ranges of each variable has to be normalized to the same scale, which allows a direct comparison of the impact.

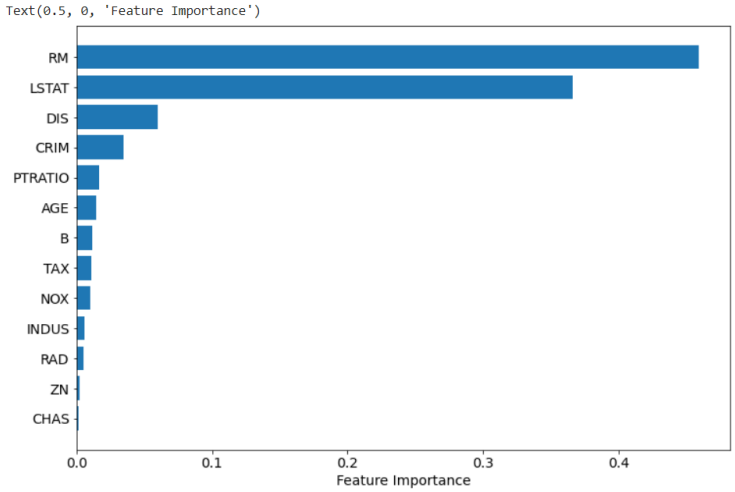

Fig. 7.1: Feature Importance

7.2.2 Input variable impact

Techniques to analyse the impact of different variables are Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) and saliency maps. An illustration is visible in Fig. 7.2 - 7.4. PDP visualises the relationship between a specific feature and the model's predictions while holding other variables constant. PDPs 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. Saliency maps highlight regions in the feature map, which contributes to the prediction.

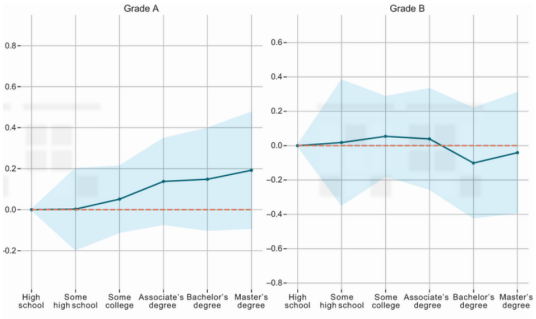

Fig. 7.2: Partial Dependence Plots

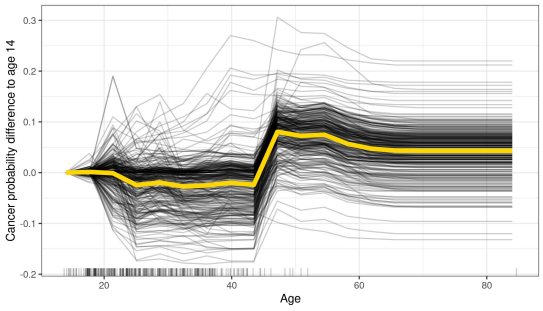

Fig. 7.3: Individual Conditional Expectation

Fig. 7.4: Saliency maps

7.2.3 Specific prediction analysis

To interpret specific predictions, Local Interpretable Model-Agnostic Explanations (LIME, Fig. 7.5) and Local rule-based explanations can be utilized. For the LIME process, a local interpretable surrogate model is estimated. A small sample with similar variable values is selected and used to create 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 as a surrogate model instead.

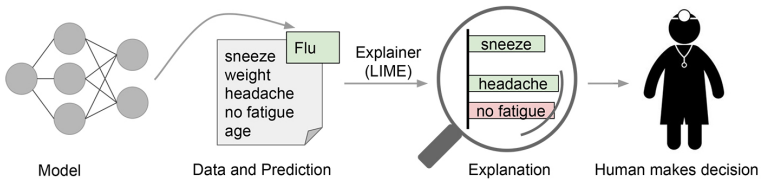

Fig. 7.5: Local Interpretable Model-Agnostic Explanations

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 analyse how the machine learning model reacts to adversial attacks. Those are input data deliberately selected with the goal 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 in the training phase.

During the sensitivity test, data with value ranges not captured by the training sample are used to analyse the model predictions and their performance.

Chapter 8

Used Data and Results

8.1 Freddie Mac's Single Family Loan-Level Dataset

Freddie Mac is one of the leading government-sponsored enterprises (GSEs) in the United States and plays a crucial role in the secondary mortgage market. Over the years, Freddie Mac has accumulated vast amounts of data related to single-family mortgages. Recognizing the value of this data, Freddie Mac has made significant steps in promoting transparency and made their Single Family Loan-Level Dataset available.

8.1.1 A Wealth of Information

The Freddie Mac Single Family Loan-Level Dataset is a comprehensive collection of loan-level data that provides detailed information about individual mortgages acquired by Freddie Mac. This dataset encompasses a wide range of attributes, including borrower characteristics, loan terms, property details, and performance metrics.

8.1.2 Key Features and Contents

The Single Family Loan-Level Dataset contains a vast array of variables and some of the essential features of this dataset include:

1. Loan-Level Attributes: This category includes borrower information such as credit score, income, employment status, and demographic details. It also encompasses loan-specific characteristics like the loan amount, interest rate, loan-to-value ratio (LTV), and occupancy status.

2. Property Details: The dataset includes information about the properties associated with the loans, such as property type (single-family, condominium, etc.), location, property value, and property condition.

3. Loan Performance: This section captures vital data related to loan performance over time. It includes details about payment history, delinquency status, modification history, and foreclosure outcomes.

4. Securitization and Pooling: Freddie Mac's securitization activities are reflected in the dataset, allowing researchers to examine the characteristics of mortgage-backed securities (MBS) and their underlying loans.

8.1.3 Data Quality and Limitations

While the Freddie Mac Single Family Loan-Level Dataset offers a wealth of information, it is essential to consider its quality and limitations. Freddie Mac maintains rigorous data quality

control processes to ensure the accuracy and reliability of the dataset. However, like any dataset, it may contain certain limitations and potential biases.

Firstly, the dataset primarily represents mortgages acquired by Freddie Mac, which may not be fully representative of the entire mortgage market. Therefore, caution should be exercised when generalizing findings from this dataset to a different population.

Secondly, the dataset is subject to data privacy and confidentiality regulations. Personally identifiable information is anonymised or removed to protect borrower privacy. While this is crucial for compliance, it can limit the depth of analysis in certain areas.

Lastly, the dataset is periodically updated, reflecting new acquisitions and loan performance. One should account for any changes or updates that may impact their analysis.

8.1.4 Access and Usage

Freddie Mac's Single Family Loan-Level Dataset is publicly available on their website (`https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset`). However, users are required to register and agree to the dataset's terms of use.

8.2 Dataset

The Freddie Mac Dataset consists of two parts, 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. During the data preparation and modelling process, the focus is on the origination data since an application scoring model will be created. 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 is continuously updated every quarter. On average, each month contains 157.000 mortgage loans and 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 had to be created. This information was derived from the performance data of the mortgage loan. As a first step, the number of months between the date of the first payment of interests 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 Article 178(1)(a) of the CRR (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 fulfil the definition of default stated in Article 178(1)(b) of the CRR, the variable "Zero Balance Code" was used. It contains the reason, why the loan balance was reduced to zero, displayed in table 8.2. Therefore, Balance Code 02, 03, 09, 15 indicate a negative financial health and was 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

For the modelling process, a time period of 12 months was selected. After identifying the default events, the default flag was set with 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 behaviour in the "Zero Balance Code" during the first 12 months in the books.

8.3 Sample creation

Data from January 2017 to December 2021 was selected for the development sample. It nicely covers a time period of different economic status from before and during the corona crisis, while also limiting the number of observations due to the limitation of computational power.

8.3.1 Data exclusions

The first two months of the whole data set showed an unusual low number of observations and default events and were therefore excluded. Because of the 12 month observation period for the default flag, the last 12 months of the data set were not considered. Additionally, accounts, which were prepaid before the 12 months observation period ended, were also removed. Lastly, mortgage loans without monthly performance data were deleted as well, because it was then not possible to approximate a default flag. The number of exclusions per reason is listed in table **??**.

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 and Test data

The data preparation and univariate analysis was performed on the whole development sample. The data set was then split into 70% training and 30% test sample stratified on the default flag and year to ensure a balanced data set for the multivariate analysis and the modelling process of both modelling approaches. The sample sizes and default rate are given in TABLE XX.

Fig. 8.1 shows the number of observation as well as the default rate per month for the whole sample before and after data exclusions and Fig. 8.2 is a separate display for the development sample. If the number of defaults wouldn't have been sufficient, an increase of the the observation might have been a solution. Both figures shows a satisfying number of defaults per month with an average default rate of xxx%. A plausibel development of the default rate is visible, it follows the expected increase of defaults during the Dot-Com crisis in the late 1990s, financial crisis in 2007/2008 and corona crisis in 2020/2021.

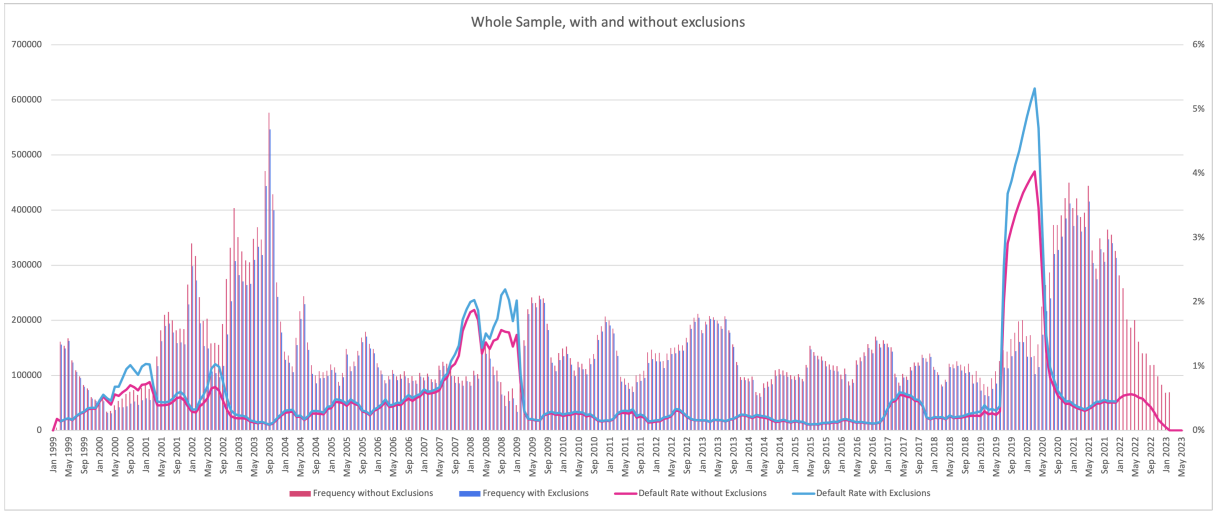

Fig. 8.1: Distribution and default rate of whole sample

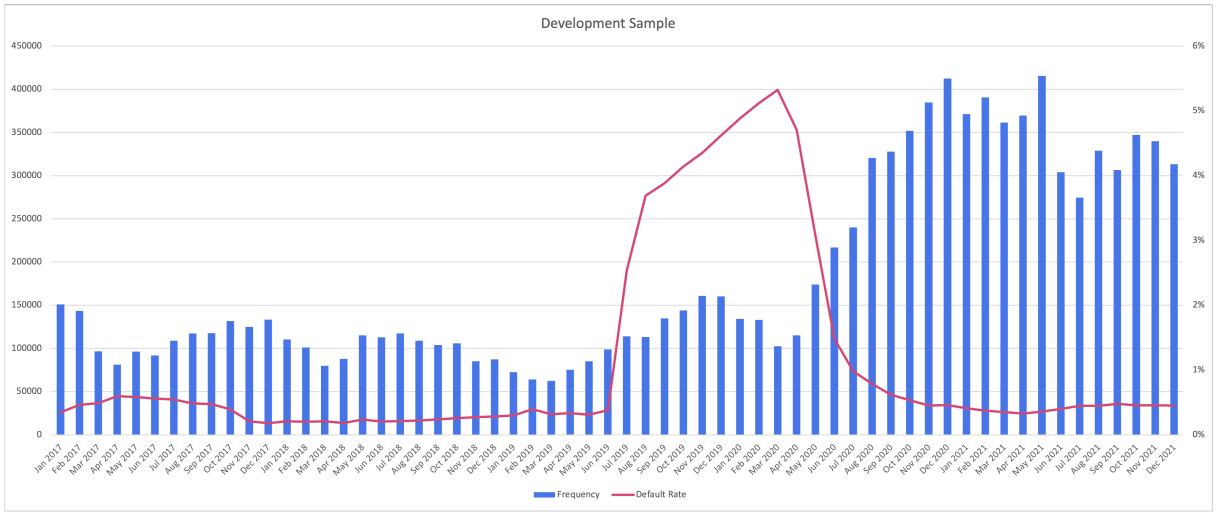

Fig. 8.2: Distribution and default rate of development sample

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 8.7. Program Indicator, HARP Indicator and Super Conforming Flag show a very high missing value proportion of over 90% and are therefore no candidates for the model. All other risk factors either don't have missing values or an acceptable amount of less than 20%. Due to the low number of missing values for the numerical variables DTI, Credit Score and other numerical variables with a missing rate below the 3rd decimal, simply the median was imputed (seen in table 8.4). For Property Valuation Method and other categorical risk factors, these data points were analysed as a separate category. Missing values of indicator variables were set to = 'Y'.

A treatment of erroneous data was not performed, because data entries outside of pre-defined ranges were already set as "Not Available" or "999" by Freddie Mac and therefore, additional analysis of possible error was not considered necessary.

8.4.2 Outlier Treatment

The outliers were detected using the interquartile approach explained in chapter XX and visually presented by creating boxplots, visible in Fig. XX - YY. The upper and lower boarders were therefore calculated using the forumals 5.4 and 5.4. The quartiles of all numerical risk factors are displayed in table 8.4 and the resulting limits are visible in table 8.5. The proportion of upper and lower outliers were calculated to analyse, if a significant amount of data entries are affected and could influence the modelling process. All variable show outliers - however, only the variables Mortgage Insurance Percentage and Original Loan Term show a concerning amount. New risk factors were derived to circumvent this issue. Multiple versions of indicator variables were created using the definition listed in table 8.6. Additionally, after analysing the distribution and box plots of LTV and CLTV and considering, that they are ratios, a winsorization was performed on both variables to test, if the outliers affect the modelling process negatively. Therefore, data points with values above the upper or below the lower limit, were capped at the respective limit.

Variable	Sum	Mean	Mode	StdDev	Min	P1	P5	P25	Median	P75	P95	P99	Max
CLTV	789.276.016	72	80	18	2	24	38	61	75	83	95	97	645
No Borrowers	16.196.343	1	1	1	1	1	1	1	1	2	2	2	6
No Units	11.234.497	1	1	0	1	1	1	1	1	1	1	2	4
DTI	370.260.245	34	45	10	1	12	17	27	35	42	48	50	65
Credit Score	8.221.074.894	752	801	44	300	636	669	722	761	789	809	817	850
LTV	786.544.248	72	80	18	2	24	38	60	75	82	95	97	611
MI Perc	67.423.997	6	-	11	-	-	-	-	-	6	30	30	55
Loan Term	3.513.344.140	321	360	72	60	156	180	360	360	360	360	360	506
UPB	2.905.914.537.000	265.889	200.000	137.670	6.000	56.000	88.000	162.000	240.000	346.000	518.000	689.000	1.582.000

Tab. 8.4: Descriptive statistics

Variable	Lower Boarder	Upper Boarder	# below Lower Boarder	# above Upper Boarder	% below Lower Boarder	% above Upper Boarder	# Outliers	% Outliers
CLTV	28	116	181.953	2.922	1,66%	0,03%	184.875	1,69%
No Borrowers	- 1	4	1	9.409	0,00%	0,09%	9.410	0,09%
No Units	1	1	-	218.561	0,00%	2,00%	218.561	2,00%
DTI	5	65	77.087	20	0,71%	0,00%	77.107	0,71%
Credit Score	622	890	24.181	-	0,22%	0,00%	24.181	0,22%
LTV	27	115	164.222	2.015	1,50%	0,02%	166.237	1,52%
MI Perc	- 9	15	21	2.252.233	0,00%	20,61%	2.252.254	20,61%
Loan Term	360	360	2.628.031	45	24,05%	0,00%	2.628.076	24,05%
UPB	- 114.000	622.000	-	203.564	0,00%	1,86%	203.564	1,86%

Tab. 8.5: 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 analysing the different distinct values of categorical variables, risk factor Property states was grouped to five US regions according to their geographical position: Northeast, outheast, Southwest, Midwest, West and other Regions, e.g. outside of the North American continent. A summary of the new variables is listed in table 8.6.

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 $\geq$ 360 Months Indicator	Indicator, that Original Loan Term $\geq$ 360 Months	Loan Term $\geq$ 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. 8.6: Description of new variables

8.5.1.2 Discriminatory power

To assess the discriminatory power, the distribution plots including the default rates and ROC-curves were created as well as AUC and GINI coefficients were calculated. A full list of all metrics is given in table 8.7 and plots of relevant variables are given in Fig. XX to Fig. YY. All plots are displayed in Annex xx. As expected, the external credit score provided by FICO shows the highest discriminatory power, followed by financial ratios, e.g., DTI, LTV and CLTV. Other numerical risk factors indicate good predictive power and were therefore seen as candidates for the modelling process, but instead of the raw variable, the indicator version Loan Term $\geq$ 360m and MI Flag were selected. The performance of the categorical and indicator risk factors were disappointing, with barely any GINI coefficients above 5%. A Gini threshold of 5% was therefore set for the selection process of the long list.

To summarize, all variables with a missing proportion below 20%, outlier proportion not succeeding 20% and GINI coefficient above 5% were selected for the long list given in table 8.9.

Variable	Value	% Missing	GINI	AUC
Amort Type	Fixed Rate Mortgage	0,00%	0,00%	0,5000
Prop Val Method	ACE Loans	2,78%	9,80%	0,5490
	Full Appraisal	2,78%	11,74%	0,5587
	Other Appraisals (Desktop, driveby, external, AVM)	2,78%	0,33%	0,5017
	Not Available	2,78%	1,61%	0,5081
Channel	Not Available	0,00%	0,00%	0,5000
	Broker	0,00%	1,05%	0,5052
	Correspondent	0,00%	7,21%	0,5361
	Retail	0,00%	8,26%	0,5413
	TPO Not Specified	0,00%	0,00%	0,5000
Prog Flag	Not Available or Not Applicable	94,03%	4,56%	0,5228
	HFA Advantage	94,03%	0,78%	0,5039
	Home Possible	94,03%	3,78%	0,5189
	Refi Possible	94,03%	0,00%	0,5000
Loan Purpose	Refinance - Cash Out	0,00%	1,05%	0,5053
	Refinance - No Cash Out	0,00%	5,26%	0,5263
	Purchase	0,00%	4,21%	0,5210
Occupancy	Investment Property	0,00%	0,20%	0,5010
	Primary Residence	0,00%	1,00%	0,5050
	Second Home	0,00%	1,20%	0,5060
Loan Term Group	= 360 Months	0,00%	10,23%	0,5512
	>. 360 Months	0,00%	0,01%	0,5000
	<360 Months	0,00%	10,24%	0,5512
Prop Type	Condo	0,00%	0,34%	0,5017
	Co-op	0,00%	0,04%	0,5002
	Manufactured Housing	0,00%	0,04%	0,5002
	PUD	0,00%	0,53%	0,5026
	Single-Family	0,00%	0,79%	0,5039
US Region	Midwest	0,00%	5,92%	0,5296
	Northeast	0,00%	1,79%	0,5090
	Other	0,00%	0,06%	0,5003
	South	0,00%	3,00%	0,5150
	West	0,00%	1,07%	0,5054
Homebuyer Flag		0,00%	5,15%	0,5258
Int Only Flag		0,00%	0,00%	0,5000
MI Flag		0,00%	12,93%	0,5647
Loan Term = 360m		0,00%	10,23%	0,5512
Loan Term ≥ 360m		0,00%	10,24%	0,5512
Loan Term >360m		0,00%	0,01%	0,5000
Sup Conf Flag		96,26%	0,00%	0,5000
HARP Flag		99,35%	0,00%	0,5000
PPM Flag		0,00%	0,00%	0,5000
CLTV		0,00%	21,04%	0,6052
No Borrowers		0,00%	11,33%	0,5566

No Units	0,00%	1,27%	0,5064
DTI	0,66%	26,60%	0,6330
Credit Score	0,02%	34,58%	0,6729
LTV	0,00%	20,75%	0,6038
MI Perc	0,00%	13,14%	0,5657
Loan Term	0,00%	10,51%	0,5526
UPB	0,00%	10,44%	0,5522

Tab. 8.7: Discriminatory power

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 selected and all risk factors with a correlation coefficient above +0.25 or below -0.25 are removed. This step should be repeated with the remaining variables until all features were analysed. Table 8.8 shows the correlation coefficient for all numeric variables in the long list. Possible interaction effects between categorical and indicator variables are considered using the Variance blabla (VIF) during the modelling process. A correlation coefficient of 100% between the original and winsorized version of ltv and ctlv is expected as well as a coefficient of 99% between cltv and ltv. Because the CLTV without outlier treatment shows the highest discriminatory power, all other correlated variables were not considered for the modelling process. The short list is given in table 8.9.

Variable	**Credit Score**	**CLTV**	**CLTV adj**	**DTI**	**LTV**	**LTV adj**	**No Borrowers**
Credit Score	1	-0,11	-0,11	-0,17	-0,11	-0,11	-0,05
CLTV	-0,11	1	1	0,12	0,99	0,99	-0,06
CLTV adj	-0,11	1	1	0,12	0,99	0,99	-0,06
DTI	-0,17	0,12	0,12	1	0,12	0,12	-0,08
LTV	-0,11	0,99	0,99	0,12	1	1	-0,06
LTV adj	-0,11	0,99	0,99	0,12	1	1	-0,06
No Borrowers	-0,05	-0,06	-0,06	-0,08	-0,06	-0,06	1

Tab. 8.8: Correlation matrix

Variable	**Long List**	**Short List**
Credit Score	Missing Treatment applied	
Homebuyer Flag	Missing Treatment applied	
MI Perc	Adapted Variable derived (MI Flag)	-
No Units	Low GINI	-
Occupancy	Low GINI	-
CLTV	Missing Treatment applied, additional Variable derived with Outlier Treatment applied	Variable without Outlier Treatment used due to higher discriminatory power
DTI	Missing Treatment applied	
UPB	Missing Treatment applied	

LTV	Missing Treatment applied, additional Variable derived with Outlier 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		
Loan Term	Adapted Variable derived (Loan Term $\geq$ 360m)	-
No Borrowers	Missing Treatment applied	
Sup Conf Flag	Low GINI	-
Prog Flag	High Missing Rate	-
HARP Flag	Low GINI	-
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 $\geq$ 360m		
Loan Term = 360m	Similar Variable used (Loan Term $\geq$ 360m)	-

Tab. 8.9: Long list and Short list

8.6 Modelling

8.6.1 Logistic regression

Using the stepwise selection algorithm, all short list variables were put into the modelling process. The resulting model is seen in table XX. The order of the risk factors indicate the significance in the model. Only the last variable shows a p-value > 0.05. To limit the number of explanatory variables in the model, the AIC and BIC was calculated for each modelling step. The relative change per step was determined and is visible in table XX. A change is visible in step 10 and therefore all other variables were removed. The final model is visible in table YY, the ROC-curve is displayed in Fig. XX and it shows a discriminatory power of GINI = bla.

8.6.2 Boosted Decision Trees

8.7 Comparison

Chapter 9

Summary and Conclusion

Appendix A

Dunno Man

A.1 Erster Section

A.1.1 Erster ding

blabla hier Text