ZHAW
Zürcher Hochschule für Angewandte Wissenschaften
School of Management and Law

Master of Science in Banking and Finance (PiE)
Capital Markets & Data Science

Master Thesis

Credit Risk Assessment via Machine Learning: Impact of Pandemic and Macroeconomic Variables on Mortgage Loan Default Prediction

Submitted by:

Sugandhita

Supervisor:
Dr. Bledar Fazlija

Winterthur, 22 Nov 2022

Management Summary

Credit risk forms an integral part of any financial institution, and the onset of COVID-19 pandemic warrants assessment of credit risk on priority. Applicability of machine learning methods has been increasingly extended to assess credit risk. Assessing the borrower's capability of repaying the loan or categorizing the borrower as high or low risk, is one of the many instances where machine learning is applied. However, when the pandemic disrupted and halted the economies, most machine learning models tended to fail in predicting the defaulting borrowers correctly.

The primary aim of this master's thesis is to demonstrate the impact of the pandemic on a simple decision tree model which is one of the baseline models used by many institutions and researchers to assess borrower risk due to its interpretability. Moreover, in the light of highly imbalanced nature of the data with majority towards non-defaulters, the impact of pandemic can be assessed on various sampling techniques which could be used to produce better results. As consumer behavior is affected by their surrounding environment and the state of the economy, macroeconomic indicators could provide signals or better predictions while predicting borrower defaults. The aim of the thesis extends to incorporating a few relevant macroeconomic indicators as independent variables in the models to assess if the same increase the predictive performance of the model.

To enable this, mortgage loan performance data from Fannie Mae is utilized from the years 2015 until 2020 coupled with macroeconomic variables for predicting mortgage loan default. Decision tree models are trained and fine-tuned for the loan performance data from 2015 to 2017 using sampling techniques such as undersampling, and oversampling, and evaluated on the following years including the pandemic year 2020. The models are again tested in combination with the macroeconomic variables to assess if the predictive performance of the models is affected.

The results indicate that although the use of macroeconomic indicators does return better results than models without macroeconomic data, when the models are evaluated on out-of-time data pertaining to the year 2020, the predictive performance declines substantially. Based on this, it could be demonstrated that pandemic indeed impacted the performance of a model used for mortgage loan default prediction but adding the macroeconomic variables as extra independent variables slightly improved the predictions.

Further research in this domain can be extended to using superior machine learning methods such as ensemble or deep leaning models for an improvement in predicting defaults. In practice, the macroeconomic indicators could be complemented with historical loan performance data to improve not only the credit scoring models but also models used for predicting probability of default or loss given default.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADA	AdaBoost
API	Application Programming Interface
AUC	Area Under the Curve
ANN	Artificial Neural Network
CPI	Consumer Price Index
CV	Cross-Validation
DTI	Debt to Income
DNN	Deep Neural Network
XGB	Extreme Gradient Boosting
FICO	Fair Isaac Corporation
FRED	Federal Reserve Bank of St. Louis
FN	False Negatives
FP	False Positives
GB	Gradient tree Boost
GDP	Gross Domestic Product
HPI	Housing Price Index
LTV	Loan to Value
NB	Naïve Bayes
NN	Neural Networks
P2P	peer-to-peer
RF	Random Forest
ROC	Receiver Operating Characteristic
SATO	Spread-at-origination
SVM	Support Vector Models
SMOTE	Synthetic Minority Oversampling Technique
TN	True Negatives
TP	True Positives
UPB	Unpaid Balance
US	United States

1. Introduction

Credit risk forms an integral part of lending activities undertaken by traditional banks, FinTech firms or peer-to-peer (P2P) lending platforms. Credit risk is a major risk encountered by many banks and has become a top priority upon the onset of COVID-19 pandemic (BIS, 2022). The applicability of machine learning models encompasses credit risk assessment through credit scoring models for better evaluating the possibility of default by the borrower (Laborda & Ryoo, 2021, p. 1). Lending institutions gather borrower specific information to obtain credit scores and establish the risk associated with the individual (Turjo ,Rahman, Karim, Biswas, Dewan & Hossain, 2021, p. 125) by using statistical and machine learning methods. Consumer behavior is affected by surrounding economic environment and hence macroeconomic indicators play an important role for financial institution customers (Tang, Thomas L, Thomas S & Bozetto, 2007, pp. 22–38) whilst utilizing econometric models.

As COVID-19 pandemic disrupted economies and financial institutions, this master thesis aims to demonstrate how COVID-19 affected performance of a machine learning model, used for consumer credit risk evaluation. Large datasets can also hinder the predictive performance of a machine learning model. Decision tree is the chosen model due to its interpretability and easy to understand nature. In addition to this, an interplay of macroeconomic indicators with traditional information on loan is utilized to assess consumer credit risk. Previously, some research incorporating impact of COVID-19 and macroeconomic factors has been done for SME's and sovereign credit risk but has not yet been explored for predicting default of mortgage loans. Macroeconomic factors are integrated with historical data of publicly available Fannie Mae single-family loan performance data on mortgage credit to study whether the predictive power of machine learning model is improved.

1.1. Motivation

The aim of this thesis is to demonstrate the impact of COVID-19 on the performance of credit risk model. In the era of big data, different resampling techniques are also used to train and evaluate large datasets. Additionally, macroeconomic indicators are integrated together with historical data of mortgage loans to study the impact on performance of credit risk model. It also becomes imperative to understand which factors affect the most in predicting mortgage loan default. The thesis lays a foundation on topics of machine learning, credit risk assessment, and relevant macroeconomic factors. The single-family loan performance data from Fannie Mae is utilized, models are built, trained, and tested through Python programming language.

1.2. Research Question

Disruption by the COVID-19 pandemic in the financial sector has been noteworthy. As lending is a key part in majority of the financial institutions (BIS, 2022), more so in the time of crisis such as the COVID-19 pandemic, appropriate evaluation of borrowers becomes crucial. Increasing amount of consumer data has also been available for financial institutions to leverage for precisely predicting borrower defaults. However, in the time of crisis, it is necessary to study if the conventional machine learning models can help achieve the desired results as they are originally tuned for non-crisis. Built on this, the foremost research question answered by this study is:

Has COVID-19 pandemic impacted the performance of machine learning models for predicting mortgage loan default?

With growing data pertaining to not only consumers but also external economic environment, it becomes increasingly important to amalgamate data for better understanding and evaluating credit risk. Most statistical and economic models consider macroeconomic indicators, but machine learning models featuring these indicators for mortgage credit risk assessment are less explored. Considering this, this thesis also aims to answer the following research question:

Could incorporating macroeconomic indicators into machine learning methods improve the performance of credit risk models?

1.3. Scope

The all-encompassing nature of the topic and limited resources available to the author enforces limitations to the work as defined below:

- The thesis focuses on studying the impact of the pandemic and macroeconomic factors on the performance of machine learning models. The machine learning method evaluated in this thesis is limited to decision trees.

- The datasets perused in the thesis are publicly available and limited by the resources available to the author.
- The macroeconomic data retrieved from various official websites of the U.S departments are assumed to be correct and comprehensive as of writing.
- Open-source Python packages and libraries, none of which are developed by the author, are utilized for implementing the algorithms.
- The thesis is part of author's master's degree in banking and finance with a specialization in capital markets and data science. Hence, the thesis is presented in such a way that the same can be grasped by the students of the aforementioned master's degree.

1.4. Structure of the Research

The thesis comprises of five sections. The introductory section is followed by the second section, which provides an overview of theoretical background and current state of research on the topic in discussion. The data assessed in the thesis are described in the third section. The fourth section presents the empirical research which includes model implementation and model evaluation. The fifth section discusses the results alongside stating the answer to the research question. It also presents the limitations of the methodology, recommendations for future direction of research and the practical implications.

2. Literature Review

This chapter provides a theoretical background of machine learning methods used by financial institutions for credit risk assessment. The theoretical framework allows for understanding of terminology and concepts of machine learning perused in this research in the context of credit risk. Initially, a brief introduction to machine learning concepts and models is laid out. Subsequently, existing research relevant to the study is reviewed, which provides an overview of current practices and state of research in credit risk assessment through machine and deep learning. Further, for thorough explanations and detailed information, the literature referenced in this thesis can be consulted.

2.1. Credit Risk Assessment

According to BIS (2000, p. 1), credit risk arises when borrowers may potentially fail to honor the debt they owe to a financial institution. As credit risk forms a major part of financial risk faced by many banks (BIS, 2022), it becomes crucial to manage credit risk for an extensive approach to risk management (BIS, 2000, p. 1). Credit risk assessment is a continuous process which uses historical data on loans to predict if a borrower (individual/company) may default or not, belongs to high/low risk category or becomes insolvent (Chen N, Ribeiro, & Chen A, 2016, p. 2). While consumer credit comprises of loans such as mortgage, automobile, personal or credit cards (S. Chen, Guo & Zhao, 2021, p. 358), loans given to companies are categorized under corporate credit. Usually, the criterion for identification of default ranges between 90 days to 180 days past due (DPD) for different types of exposures (BIS; BCBS, 2002). However, the more general 90 days past due trigger for default of loan is utilized in the thesis.

Lending decisions are made by financial institutions based on credit scoring models used for evaluating consumer creditworthiness (S. Chen et al., 2021, p. 358). Model inputs comprise of diverse financial and non-financial factors considering institution’s credit policies, legal outline and economic situation (E.Saygili, T Saygili & Isik, 2019, p. 161) in addition to being consumer relevant. Hence, in pursuit of identification of relevant factors for assessing credit risk, research to ascertain the factors influencing credit defaults has been swelling after the financial crisis of 2008 (Barbaglia et al., 2021, p. 1). Amongst this, with rising use of data and technology particularly machine learning in the financial sector, quest for high performing models continues. Proper classification of consumers and assessment of credit risk has led to development of different models in financial institutions (Leo, Sharma & Maddulety, 2019, p. 8). Traditionally used models in financial institutions include logistic regression (S. Chen et al., 2021, p. 358) and Linear Discriminant Analysis (Shi, Tse, Lua, D'addona & Pau., 2022, p. 14327). Since these methods are unable to handle sizeable datasets (Shi et al., 2022, p. 14327), usage of big credit datasets has made provisions for flexible traditional machine learning and deep learning methods (Shi et al., 2022, p. 14328). In research pertaining to credit risk management, models such as Support Vector Models (SVM), Random Forest (RF) and Neural Networks (NN) have been the most examined algorithms (Leo et al., 2019, p. 11). More light is shed upon the current state of research on the drivers of credit default and machine learning models in the sub-section 2.4.

2.2. Machine Learning

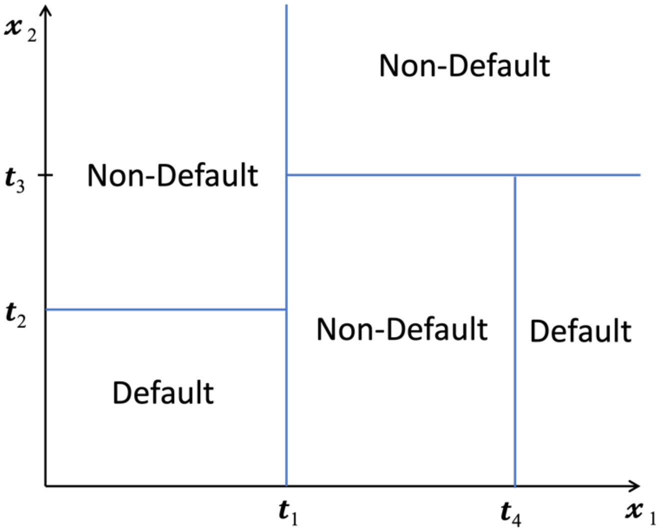
A collection of techniques used to automatically find patterns in data and using those patterns for forecasting or carrying out uncertain decision-making is referred to as machine learning (Murphy, 2012, p. 1). In short, machine learning is about knowledge extraction from data (Müller & Guido, 2016, p. 1). Machine learning can be segregated into three categories: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the algorithm learns from sample data and related target variables to further predict the correct response when modeled with new data. When the target variable is numeric, it is known as regression problem, whereas, when the target is a tag or a class, it is deemed to be a classification problem (Mueller & Massaron, 2021, p. 140). The second category of ML i.e., unsupervised learning deals with data which does not contain a target response and the patterns in data are decided by the algorithm itself (Mueller & Massaron, 2021, p. 141). Some of the techniques used in unsupervised learning comprise of clustering analysis and transformation of data through dimensionality reduction (Müller & Guido, 2016, p. 131). While the third category namely reinforcement learning happens when the algorithms are presented with data lacking labels similar to unsupervised learning and is forced to learn from its failures and make decisions to eventually succeed (Mueller & Massaron, 2021, p. 141).

In the case of credit risk assessment, particularly predicting loan default, the relevant machine learning category for this thesis is supervised learning. The task of predicting whether a loan will default or not is a binary classification problem, and so, various methods used under supervised learning can be deployed to solve this binary classification problem. Decision tree is one of the few methods serving as a baseline model over the years for credit scoring models while advanced ensemble methods, and neural networks have appeared to stand out as ‘challenging’ models delivering superior results (Markov, Seleznyova, & Lapshin, 2022, p. 191). Focusing on the conventional models, this thesis uses decision trees for the classification problem and demonstrating the impact of pandemic and macroeconomic factors on predicting loan default.

2.2.1. Decision Trees

Decision trees, also known as Classification and Regression Trees (CART) models, are a method used in supervised learning that can be used for regression and classification problems. Besides being a simple model, decision trees allow for interpretable decisions particularly in classification tasks related to finance or medicine field (Fazlija, 2022). Decisions arrived at by a decision tree are through a learning process of sequential if/else questions (Müller & Guido, 2016, p. 70). Binary decision trees wherein at each step there are only two possible answers such as yes/no, true/false and good/bad are most widely applied (Fazlija, 2022). In such a tree-based method the feature (variable/input) space is divided into a set of rectangles through recursive binary splitting and a simple model is fit into each region (Hastie, Tibshirani & Friedman, 2017, p. 305; Christopher M. Bishop, 2009, p. 663).In the context of default prediction, the same is depicted in Figure 1.

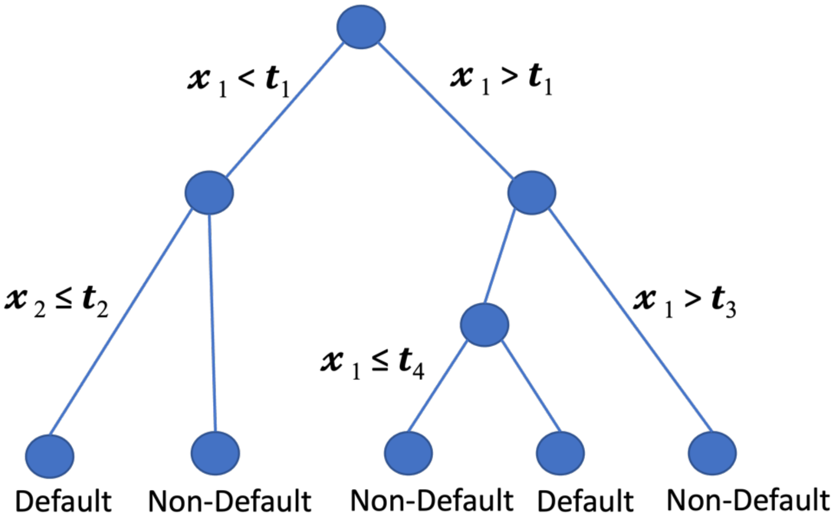
Figure 1- *Partitioning of Input Space in a Decision Tree*



Note: Two-dimensional division of input space in a tree-based method. Own representation based on Christopher M. Bishop (2009, p. 663).

The whole input space is divided into two regions i.e., default and non-default, in the first step, based on whether $\boldsymbol{x}_1 > \boldsymbol{t}_1$ or $\boldsymbol{x}_1 \leq \boldsymbol{t}_1$ where $\boldsymbol{t}_1$ is a parameter of the model. The resultant sub-regions can be further subdivided independently based on more parameters to achieve the best fit. For any new input variable $\boldsymbol{x}$, the region/section it falls into is identified by beginning at the root node (top of the tree) and following the way down to specific leaf nodes in accordance with the decision criteria at each node respectively (Christopher M. Bishop, 2009, p. 664). The leaf or terminal nodes correspond to the regions denoted by Non-Default and Default in Figure 2. A decision tree is learned through a greedy algorithm which maximizes the results in each step of the optimization process i.e., selecting the best choice at each node (Fazlija, 2022; Mueller & Massaron, 2021, p. 181).

Figure 2 - *Binary Tree Structure*


$\boldsymbol{x}_1 < \boldsymbol{t}_1$
$\boldsymbol{x}_1 > \boldsymbol{t}_1$
$\boldsymbol{x}_2 \leq \boldsymbol{t}_2$
$\boldsymbol{x}_1 > \boldsymbol{t}_3$
$\boldsymbol{x}_1 \leq \boldsymbol{t}_4$
Default
Non-Default
Non-Default
Default
Non-Default

Note: Binary tree in correspondence with the partitioning depicted in Figure 1. Own illustration based on Christopher M. Bishop (2009, p. 664).

In order to grow the tree, measures such as entropy or Gini index are used (Hastie et al., 2017, p. 310) to find the features and splitting point at each node to split the given classes or tags in best possible way (Fazlija, 2022).Additionally, information gain explains how a decision tree can easily identify a way to increase predictive ability at a certain split (Mueller & Massaron, 2021, p. 181). Entropy, on which the information gain formula is based on, describes the expected value from the information in a message:

$$Entropy\ (p) = H(p) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Also known as cross-entropy, $H(p)$ is the entropy of set of probabilities $p = \{p_1, p_2, \dots . p_n\}$ summing up to 1 in class i, representing percentage of each class where $\log_2$ is the base 2 logarithm. In a set of features, entropy measures the amount of impurity (Fazlija, 2022) or disorder or uncertainty. Information gain determines the importance of an attribute in a feature vector to proceed with the classification. The feature with the highest information i.e., highest entropy is chosen at every node for each consecutive question (Fazlija, 2022). For instance, information gain for a feature V in the dataset S is calculated as:

$$Information\ Gain\ (S, V) = Entropy\ (S) - \sum_{v \in values(V)} \frac{|S_v|}{|S|} Entropy\ (S_v)$$

Where $|S_v|$ is the number of elements in S which have the value v for feature V and $Entropy\ (S_v)$ is calculated as the entropy S with datapoints for which the feature V has values v (Fazlija, 2022). The information gain measures the reduction in entropy or uncertainty once the dataset is split. Higher the reduction in uncertainty, more information is gained.

Gini index also known as Gini impurity can also be used instead of cross-entropy for learning of Decision Trees. It is computed as:

$$G(p) = \sum_{i=1}^{n} p_i(1 - p_i)$$

Gini index is the expected error rate wherein p_i is the probability that an element in the leaf node belongs to class i and $(1 - p_i)$ is the probability of being misclassified (Murphy, 2012, p. 548). The node which contains datapoints with the same target value or class is known to be pure. Either Gini index or cross-entropy can be used for splitting the data and this partitioning of the data is continued until each region contains a single class value, i.e., the terminal node is pure. For completeness both the measures have been presented but for consistency purposes, entropy will be used in the analysis.

A question always remains that how large the tree should be grown? If the tree is grown to its full size, then it may become deep i.e., complex, and prone to poor generalization performance on unseen data and eventual overfitting. To avoid overfitting, the growth of trees is halted by stopping addition of nodes when the reduction in cross-entropy/Gini index (for a classification tree) and residual error (for a regression tree) falls below some threshold (Christopher M. Bishop, 2009, p. 665). This is called pruning the tree. In practice, the tree is grown to be deeper, and pruning is applied until a good generalization performance is achieved (Fazlija, 2022). According to Christopher M. Bishop (2009, p. 665), for pruning, denoting the starting tree by T_0, a subtree $T \subset T_0$ can be defined which can be obtained by pruning nodes from T_0. The leaf nodes indexed by τ = 1,….,$|T|$ representing a region $\mathcal{R}_\tau$ of the input space contain N_T data points with $|T|$ being the total number of leaf nodes. Then the ideal prediction for $\mathcal{R}_\tau$ is provided by:

$$y_\tau = \frac{1}{N_T} \sum_{x_n \in \mathcal{R}_\tau} t_n$$

Where it is assumed that the starting point was with labeled data of form (x_n, t_n) (Fazlija, 2022). In the case of a regression problem, the resulting contribution to the residual sum of squares would be

$$Q_\tau(T) = \sum_{x_n \in \mathcal{R}_\tau} (t_n - y_\tau)^2$$

For a classification problem the resulting contribution $Q_\tau(T)$ would be to $Entropy\ (p)$ or Gini index $G(p)$ already defined above. The pruning criteria then goes on to be:

$$C(T) = \sum_{\tau=1}^{|T|} Q_\tau(T) + \lambda|T|$$

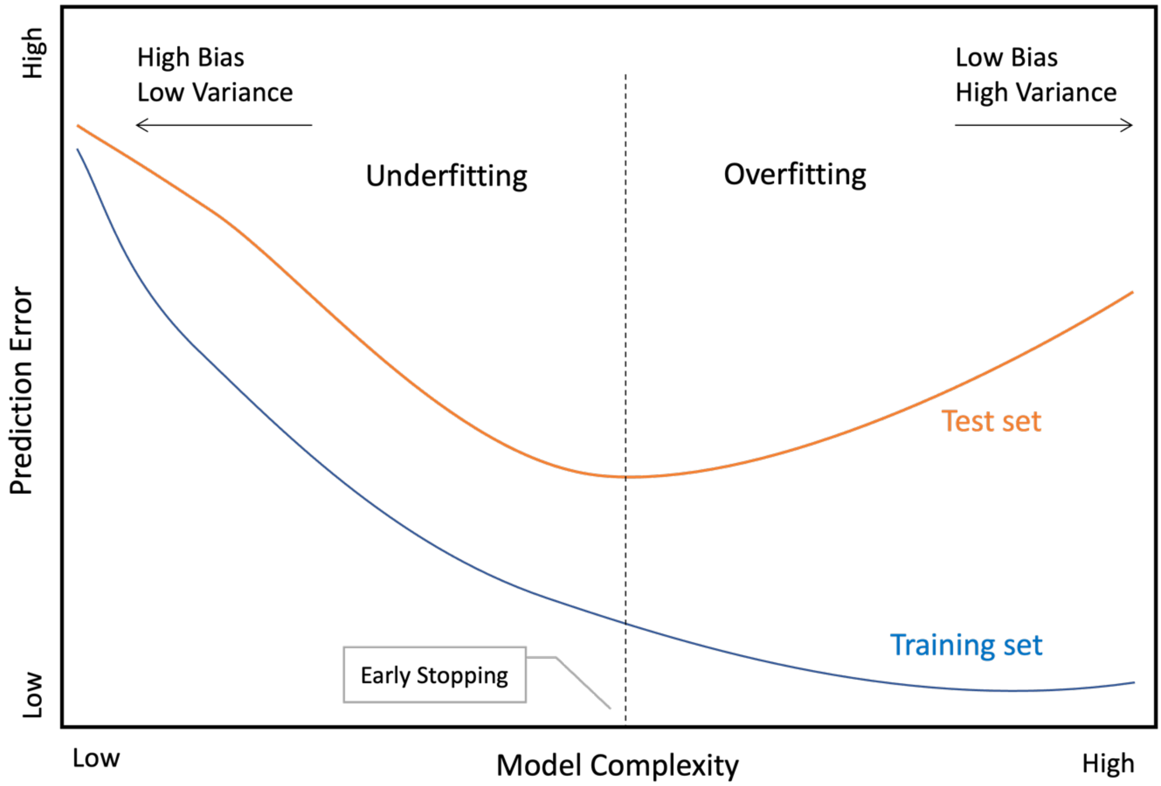
The optimal tradeoff between the residual loss/cross entropy/Gini index and model complexity i.e., number of leaf nodes is determined by the regularization parameter λ whose value is chosen through cross-validation. The depth of the tree influences the generalization performance of the model on unseen data and hence the max depth of the tree can be chosen through hyperparameter tuning. The process for cross-validation is presented in the section

2.2.2. The Bias Variance Trade-off and Cross-Validation

When a model is learned, the prediction capability of that model i.e., the generalization performance relates to an independent test dataset (Hastie et al., 2017, p. 219). This means that the data is split between training and test set, where the model is learned on the former. During the process of learning, the parameters of the model are optimized based on a performance metric score and then the model is fed with test data. The generalization performance i.e., out of sample performance is evaluated by comparing the ground truth with the predicted output. The assessment of this performance guides the model choice and the key method includes the trade-off between bias, variance and model complexity (Hastie et al., 2017, p. 219).

For instance, during the learning of a model by the algorithm, as the complexity if the model grows, it is possible for it to memorize the data and fit the in-sample data well with low errors and low bias (Hastie et al., 2017, p. 221). However, when the model is tested out of sample (test set), the prediction or errors alter significantly as compared to when relearning from the same data through a distinctive approach resulting in overfitting (Mueller & Massaron, 2021, p. 161). In this scenario, any minor changes in the training data would produce an erratic prediction on the unseen test data. In contrast, when the model becomes too simple, they are unable to map the appropriate relationship between the input variables and target variable inducing high bias errors and lower variance (Mueller & Massaron, 2021, p. 160). This is the classic case of underfitting.

Figure 3- Overview of *Bias Variance Trade-off with Model Complexity*

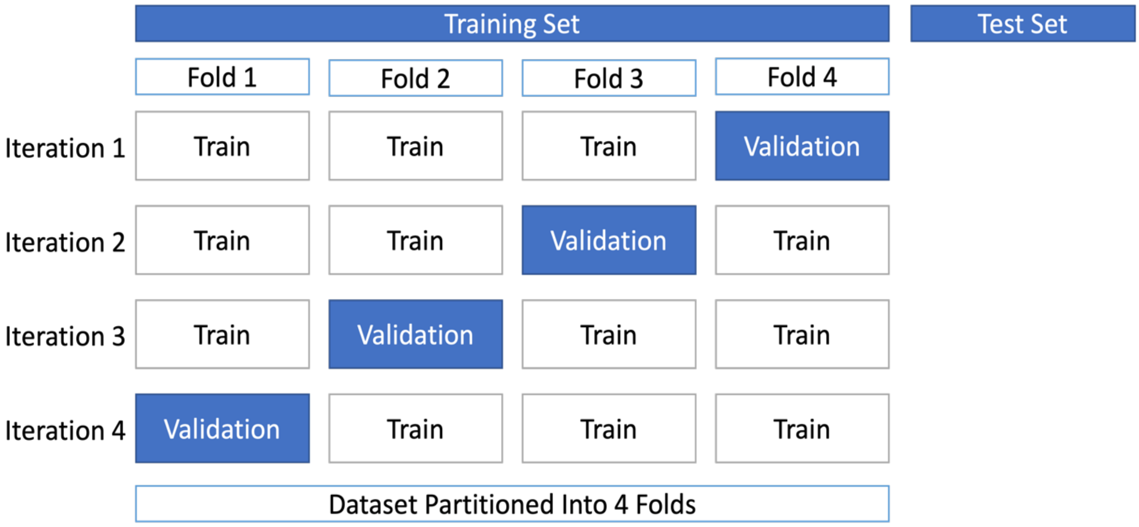


Note: Own illustration. Adapted from Hastie et al. (2017, p. 220)

Figure 3 depicts that as complexity of the model increases, even though the training error declines, the test error increases. Hence the ideal spot where the model generalizes well on both the training and out of sample test set, is determined to be at an intermediate point, which provides minimum expected test error. This is the point where the early stopping criterion on a decision tree is applied through pruning.

Therefore, in order to choose the optimal model or assess the model for its generalization capabilities, Cross-Validation (CV) technique can be applied. K-fold cross-validation relies on randomly splitting the dataset into *k* number of distinct folds of equal size and using each fold as a test set while the others as training set (Mueller & Massaron, 2021, p. 166). An error estimate is produced at each iteration using different folds as test beside the others used for training. The resulting *k* number of errors are averaged to compute the prediction error or the cross-validation sore (Mueller & Massaron, 2021, p. 166; Fazlija, 2022). The same process can be done for calculating the average performance metric instead of prediction errors (Fazlija ,2022). The main advantage of this method is that each observation is tested, and a mean score can provide a probabilistic approximation of the predictive performance. Figure 4 describes a 4-fold cross-validation as an example.

Figure 4 - *4-Fold Cross-Validation Framework*



	Training Set				Test Set
	Fold 1	Fold 2	Fold 3	Fold 4	
Iteration 1	Train	Train	Train	Validation	
Iteration 2	Train	Train	Validation	Train	
Iteration 3	Train	Validation	Train	Train	
Iteration 4	Validation	Train	Train	Train	
	Dataset Partitioned Into 4 Folds				

Note: Own illustration. Adapted from Mueller & Massaron (2021, p. 166)

2.2.3. *Handling Imbalanced Dataset*

Usually, in the credit datasets, the ratio of defaults vis-à-vis non-defaults remains low which poses problems for model training and estimations. As the minority class may not be recognized by the learning algorithms, chances of predicting of all loans classified as non-default remain high. To address the issue of imbalanced data, many resampling methods such as under sampling, oversampling, and Synthetic Minority Oversampling Technique (SMOTE) can be deployed. For instance, when the minority class is oversampled using replacement i.e., causing duplicate records of minority class, it is known as oversampling as opposed to under sampling where the majority class is downsized to match the minority class. SMOTE has been one of the most generally used approach to address the problem of imbalanced dataset (Shi et al., 2022, p. 14333). SMOTE is an oversampling technique which utilizes the k-nearest neighbors to give new records based on the distance between the rare class and randomly selected nearest neighbors (Moscato et al., 2021, p. 4). All these techniques provide a balanced dataset, but under-sampling and oversampling could lead to loss of valuable information and overfitting respectively.

2.2.4. Performance Measurement Methods

Numerous evaluation metrics are present depending on the business use case. The most widely used evaluation metric in literature has been accuracy i.e., the fraction of correctly classified sample. However, an appropriate measure capturing the expected business impact of choosing a model over another should be employed (Müller & Guido, 2016, p. 276). Accuracy may not be a good measure of predictive performance when the dataset is highly imbalanced (Müller & Guido, 2016, p. 279). The algorithm might just be predicting the more often represented class while the business use case focus might lie on predicting the minority class. Hence, alternative metrics become helpful in measuring the predictive performance of a model. For binary classification, which is the case at hand for the thesis, the most comprehensive way to exemplify the result of evaluation is a confusion matrix.

Table 1- *Confusion Matrix*

		Predicted	
		Positives	Negatives
Actual	Positives	True Positives (TP)	False Negatives (FN)
	Negatives	False Positives (FP)	True Negatives (TN)

Note: Elaboration of confusion matrix based on Dastile et al. (2020, p. 9)

Since reporting a single performance measure may not capture enough information and portray a wrong picture of how the model performs in a binary classification, confusion matrix becomes a vital instrument in analyzing the performance of a classification machine learning algorithm (Fazlija, 2022). Table 1 depicts a confusion matrix consisting of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) from which different metrics can be calculated. In the context of credit scoring classification, TP and TN are the number of borrowers correctly classified as defaults and non-defaults respectively. FP is the number of non-defaulted borrowers incorrectly classified as defaults, whereas, FN is the number of defaulters incorrectly classified as non-defaults (Dastile et al., 2020, p. 9).

Recall

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

Sensitivity or True positive rate or Recall measures the correctly classified positive samples out of all the positive samples (Tharwat, 2020, p. 172). This measure is important when the need to identify all positive samples arises i.e., avoiding false negatives (Müller & Guido, 2016, p. 283. In the context of loan default prediction, this measure is the most important as the cost of classifying defaulters as non-defaulters is more than classifying non-defaulters as defaulters.

Specificity

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

Specificity, also known as True Negative Rate evaluates the accuracy of the negative classes in the dataset (Moscato et al., 2021, p. 5).

AUC ROC and Balanced Accuracy

In the binary classification cases, the Area Under the Curve (AUC) for Receiver Operating Characteristic (ROC) curve can be measured for classification performance. In case of probabilities or score predictions, the ROC curve takes into consideration all possible thresholds for a classifier and plots the true positive rate and false positive rate (Müller & Guido, 2016, p. 293). It measures how well the classifier can distinguish correctly between the classes and is a probability curve. Closer the curve to 1, better is the quality of the classifier. To summarize the ROC in a single number, it can be referred to as AUC. Higher the AUC, the better is the model in classifying.

In the case of binary predictions, the AUC mentioned above is measured as the arithmetic mean of sensitivity or recall (true positive rate) and specificity (true negative rate) (Scikit-learn, n.d.). In Scikit-learn library of Python, it is equivalent to balanced accuracy score.

$$AUC = \frac{\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)}{2}$$

According to Scikit-learn, this measure avoids performance estimates which could be inflated due to imbalanced datasets. Hence, when the datasets are balanced, this term reduces to accuracy.

Other Metrics:

Precision

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

Also known as positive predictive value, precision measures how many samples predicted as positive are truly positive. This metric is used when the end goal it to limit the number of False Positives (Müller & Guido, 2016, p. 283).

F-measure

$$\frac{2 \times Recall \times Precision}{Precision + Recall}$$

F1-score summarizes the precision and recall together calculated by their harmonic mean. High values of F-measure indicate a high classification performance (Tharwat, 2020, p. 172). F1 score can be utilized if a balance must be stricken between the precision and recall score.

Accuracy

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

Accuracy is the ratio between the number of correctly classified records and all the samples. As mentioned before, it is an unreliable measure if the dataset is unbalanced. This research presents accuracy but does not consider it in evaluation of the model due to class imbalance.

2.3. Macroeconomic Indicators

Representing the tendencies of economic movements between expansion to contraction, macroeconomic variables are closely related to the economic cycle (Xia, Li, He, Xu, & Meng, 2021, p. 3). This section presents a general description of four specific macroeconomic variables utilized in the study in addition to a theoretical overview of their role in the economy, interdependencies, and impact on credit risk.

Gross Domestic Product (GDP)

The volume of production within a country's geographical boundaries i.e., value of goods and services produced in an economy is measured by GDP (Krugman et al., 2018, p. 39). This output comprises of consumption expenditure, investment expenditure, government expenditure and current account (Krugman et al., 2018, p. 488). Hence, changes in these elements would impact the output in the country. For instance, during a downturn in the economy, a reduced demand for goods and services would be followed by a phase of less production of goods and services, leading to less labor requirements (decreasing employment) and a reduced level of output in the economy, making it difficult for people to repay their debt. Most of the researchers have a consensus that both household and firms are more inclined to meet their financial obligations during favorable economic conditions (Naili & Lahrichi, 2022, p. 338). The real estate non-performing loans tend to decrease with inflation-adjusted GDP i.e., real GDP growth (Ghosh, 2017, p. 35), thereby decreasing credit risk.

Consumer Price Index

Rising price levels in an economy is termed as inflation (Krugman et al., 2018, p. 434). It is measured by changes in Consumer Price Index (CPI) of a country. Theoretically, if the demand for goods and services is higher, the increased production of goods and services would increase the demand for labor and eventually raise the wages with a rise in production costs. Hence, if prices were to rise more for basic commodities than a borrower's income, then a borrower must be left with less income to repay their debt obligations. According to Naili & Lahrichi (2022)'s review, few researchers argue that higher inflation puts pressure on the level of non-performing loans, as the withering real value of debtor's income leads to non-repayment. Whereas, some researches point out a negative relation between inflation and credit risk citing reasons such as an increase in income because of high inflation leads to better repayment capacity (Naili & Lahrichi, 2022, p. 338). This was also corroborated by the results of Ghosh (2017) in the single-family residential loan context. Hence, the impact of inflation seems to be ambiguous.

Unemployment Rate

The number of unemployed people i.e., people not having a job but available to work, represented as a percentage of labor force is known as unemployment rate (U.S. Bureau of Labor Statistics, n.d.-b). The impact of unemployment rate on credit risk has been found to have a positive relationship (Naili & Lahrichi, 2022, p. 338). Referring to *Lawrence (1995),* Naili & Lahrichi (2022, p. 338) indicated that low income earners are prone to risks of unemployment and subsequent difficulties in repayment of debt. In the real estate mortgage context, the level of defaulting loans rises with an upturn in unemployment rates (Ghosh, 2017, p. 37).

Housing Price Index

The Housing Price Index (HPI) built by the Federal Housing Finance Agency captures the price movements of US single-family houses as it measures the "average price changes in sales or refinancing on the same properties" (FHFA, n.d.). A rise in housing price index improves the value of collateral thereby reducing default in real estate and individual loans (Ghosh, 2017, p. 35). Hence, a rise in housing price index indicates a rise in financial wealth of the homeowners and vice versa. A decline in financial wealth with declining housing prices might render a housing loan borrower to default on loans due to the risk of not fulfilling loan obligations.

Interest/Lending Rates

The interest rates are a prime instrument of any central banks' monetary policy (Krugman et al., 2018, p. 441). Any increase or decrease in interest rates affect the lending rates set by the lending institutions in the same direction. For instance, a decrease in interest rate would make loans less expensive and more attractive to borrowers. As people would consume more, GDP growth is stimulated with eventual rise in inflation (Krugman et al., 2018, p. 441). In the context of credit risk, probabilities of loan default increase as loan repayments become costly due to rise in interest rates of the loans (Ghosh, 2017, p. 32). However, as interest rate fluctuations do not affect fixed rate loans, repayment capacity of borrowers maintains status quo (Naili & Lahrichi, 2022, p. 341). Since the data perused in this thesis pertains to U.S mortgage data, to consider changes in the interest rate, Spread-at-origination (SATO), the difference between the original interest rate on the mortgage loans and 30/15-year mortgage rates in the U.S are utilized. The SATO can be representative of credit quality, as an increased spread would indicate a risky borrower and vice versa. If the spread tends to be high, the borrower may find it difficult to repay their obligations due to high burden of loan repayments. Hence this variable may not be directly treated as a macroeconomic indicator.

2.4. State of Research on Credit Risk Assessment through Machine Learning Methods

Ample studies have been conducted for evaluating credit risk through machine learning methods recently. The most extensive analysis was conducted by Chen, Guo, and Zhao (2021) wherein the performance of 13 machine learning models were investigated for predicting mortgage early delinquency probabilities. They used the Fannie Mae public dataset of mortgage loans together with macroeconomic variables over many post-crisis periods between 2009 to 2016 with the focus on model predictive accuracy and out-of-time analysis instead of scrutinizing factors influencing the early delinquency/default in mortgages. For risk classification, ensemble methods and Neural Network (NN) outperformed the other methods including decision trees but predictive accuracy remained a challenge for mortgage portfolios as none of the machine learning models could seize predictive accuracy precisely (S. Chen et al., 2021, p. 370). Prior to this Sirignano, Sadhwani and Giesecke (2018) deployed a deep neural network to model multi period-mortgage delinquency for 120 million mortgages over a period of 20 years (1995-2014) along with examination of loan, borrower-specific, and macroeconomic factors influencing mortgage delinquency. The non-linear dependencies on borrower behaviour are captured by their work which had not been addressed by previous research (Sirignano et al., 2018, p. 3). Incorporating these non-linear effects allowed the authors to improve accuracy of out-of-sample mortgage risk forecasts.

Mamonov, and Benbunan-Fich (2017) also used the Fannie Mae public mortgage dataset for 4th quarter of 2007 as the default rate was the highest during the time. To capture the pattern of defaults, 6 models namely logistic regression, decision tree, RF, SVM, boosted trees and Artificial Neural Network (ANN) were applied wherein ANN performed better in predicting delinquencies but at the expense of low precision, i.e., positive predicted values (Mamonov & Benbunan-Fich, 2017, p. 246). For the european mortgage market, Barbaglia, Manzan and Tosetti (2021) applied boosted tree-based algorithms such as Gradient tree Boost (GB) and Extreme Gradient Boosting (XGB) in addition to logistic regression, NN, RF and Naïve Bayes (NB). The comparison by authors resulted in XGB and GB outperforming other methods irrespective of performance metrics considered, while NN and NB performed weakly.

Studies on application of ML models for consumer credit risk assessment are prevelant and not just limited to the mortgage market. Alonso, and Carbó (2021) perused anonymized consumer credit data from a major Spanish bank to compare predictive performance of logistic regression with ML models such as lasso penalized logistic regression, CART or decision tree, RF, XGB, and Deep Neural Network (DNN) for predicting credit defaults. XGB and RF outperformed the other models whereas CART and DNN had performances alike (Alonso & Carbó, 2021, p. 22). Another comparison of 5 ML models with a logistic regression was conducted by Aniceto, Barboza, and Kimura (2020) examining consumer loans of a major Brazilian bank's credit portfolio where AdaBoost (ADA) outperformed the other tested models viz., SVM, RF, decision tree, and bagging. Research by Turjo, Rahman, Karim, Biswas, Dewan, and Hossain (2021) employed bank loan status data from Kaggle to 6 ML models viz., k-nearest neighbors, logistic regression, GB, XGB, ADA, and ANN, to find out that GB method gave the highest accuracy to predict if an individual is eligible for credit. Hamori, Kawai, Kume, Murakami and Watanabe (2018) applied NN with different activation functions in comparison with boosting methods on credit card default data from Taiwan only to discover that boosting method outperformed NN with respect to prediction accuracy, AUC and F-score. Credit risk has also become relevant for peer-to-peer lending (P2P) platforms owing to rising FinTech companies for predicting loan default. While examining machine learning methods such as RF, GB, XGB, and ANN on data from Renrendai.com, RF exceeded performance in predicting occurrence of default while NN performed weakly (Xu, Lu, & Xie, 2021, p. 16). The superior prediction performance of RF method was also supported by results of Liu, Yang, Wang, Li, Xiong, and Li (2022). Hence, in different application of credit risk assessment, no single ML model seemed to outperform. However, the most frequently used method has been the logistic regression.

2.4.1. Impact of COVID-19 Pandemic on Credit Risk Assessment

For corporate credit default risk prediction, Nehrebecka (2021) estimated various machine learning models using Polish non-financial enterprise data spanning from 2015-2020, which includes COVID-19 pandemic transition period. Bankruptcy prediction for select US firms has also been examined for the period of COVID-19 pandemic using machine learning models such as RF, XGB and SVM by Narvekar and Guha (2021), where XGB outperformed the rest of the tested models. A research paper was recently published by Saudi Central Bank examining the impact of COVID-10 pandemic on consumer credit scoring covering the period between 2018 and 2020. Decision tree was deployed by the authors to determine different characteristics of borrowers before and after the pandemic and that the default rate had substantially increased in the pandemic year 2020 (Bouaguel et al., n.d., pp. 33–34). There is still dearth of research which considers the impact of COVID-19 pandemic on the performance of machine learning methods for consumer credit risk assessment. The same is pointed out by Markov, Seleznyova, and Lapshin (2022, p. 193) in their comprehensive review of research pertaining to credit scoring in the past 5 years from 2016-2021.

2.4.2. Macroeconomic Factors and Credit Risk in a Machine Learning Environment

The economic environment is influenced by consumer behavior which can ultimately have implications on credit risk assessment. For instance, using varying logistic regression models, Carvalho, Curto and Primor (2022) evaluated non-financial firm data from Eurozone along with macroeconomic variables to study its influence on probability of default. Incorporating macroeconomic variables, particularly the GDP variable, was found to be bolstering the accuracy of models forecasting credit default, (Carvalho et al., 2022, p. 2070). Moreover, unemployment rate suggestively contributed to probability of default risk (Carvalho et al., 2022, p. 2065). The study by Sirignano et al.(2018) perusing the US mortgage dataset from CoreLogic combined with various macroeconomic variables to model a DNN and analyze borrower behavior, likewise, found that unemployment rate had the highest explanatory power among other variables. Another research, incorporating unemployment rates alongside housing price index and credit spreads to the Fannie Mae U.S mortgage dataset, Chen, Guo, and Zhao (2021) suggested that it was necessary to use all the three macroeconomic variables else the results had sizable prediction errors in ML models.

In the European context, a sharp decline in GDP or unemployment rates in some countries suggested that the economic conditions and its interactions with loan-specific characteristics had an indirect impact on loan defaults rather than a direct one (Barbaglia et al., 2021, p. 22). On the contrary, macroeconomic variables did not seem to enhance predictions on data for Italian SME’s to predict bankruptcy (Corazza et al., 2021, p. 331) which used an ANN model. While in the Chinese context, online consumer lending data together with multilevel macroeconomic variables, estimated through RF, gradient boosted decision trees, and linear regressions, achieved better predictive performance than standalone consumer lending data in the study by Xia, Li, He, Xu, and Meng (2021).

Even though the reviewed studies for effect of macroeconomic variables are recent, none of the studies have deployed data covering the COVID-19 pandemic period. Moreover, during this period, the area of consumer credit risk assessment is also under-researched with respect to integration of macroeconomic variables with loan and borrower specific information. Prior to this period, exclusion of macroeconomic variables in credit scoring literature has also been pointed out by Dastile, Celik and Potsane (2020, p. 13).

2.5. Libraries

The analysis of data and programming for this master thesis was carried out in the open-source Python programming language. The packages necessary for the empirical research are enumerated in this sub-section.

2.5.1. Pandas

An open-source package of Python, Pandas, became open-sourced in 2009 after its development by AQR Capital Management (Pandas Developers, 2022). The library allows for reading and writing data with data manipulation. In short, it is used for data analysis and modeling (Python, 2022).

2.5.2. NumPy

NumPy library offers scientific computing tools such as mathematical functions and random number generators whilst being accessible and productive for people with diverse backgrounds (NumPy, 2022).

2.5.3. Matplotlib and Seaborn

Matplotlib library supports the visualization of data in Python through quality plots, layouts and customizable visual style (Matplotlib, 2022). Seaborn, based on Matplotlib, also provides exploration of data through a high-level interface for statistical graphics (Seaborn Pydata, 2022).

2.5.4. Scikit-learn

Scikit-learn is quite a popular tool and one of the prominent libraries in Python containing state-of-the-art machine learning algorithms (Müller & Guido, 2016, p. 5). It is available as an open-source project with comprehensive documentation for all the algorithms (Müller & Guido, 2016, p. 6).

3. Data

This section describes the datasets used in the thesis for conducting the assessment and evaluation of the research problem. The data sources and data preparation steps deployed are presented along with a focus on the selection of the features and exploratory data analysis.

3.1. Single-Family Loan Performance Data

Fannie Mae publishes monthly performance data of a portion of loans they acquire through various mortgage sellers. The dataset consists of Fannie Mae's 30-year and less, fully amortizing, full documentation, single-family, conventional fixed-rate mortgages (Fannie Mae®). New fixed-rate mortgage loan acquisitions and latest performance data at loan level are publicly available and published quarterly with a four-month lag. For instance, the data published under July 2020 would reflect acquisitions and performance through Q1 2020. The data published is anonymized by Fannie Mae to prevent identification of individual borrowers. For the thesis, loans originated in and after year 2015 until the year 2020, having at least 12-month performance period, are analyzed. It is pertinent to mention that this period encompasses the pre-pandemic (2015 to 2019) and pandemic period (2020 to 2021). The original dataset contains 108 features, of which 38 features are not applicable to the performance dataset according to the Fannie Mae loan glossary. The same are eliminated in first step. Out of the remaining 70 features, relevant static and dynamic variables chosen for data preparation and further analysis are described under Table 2. The features chosen for modelling are specified later in section 3.4 Table 5.

Table 2- *Description of Relevant Variables of Fannie Mae Dataset*

Variable Name	**Description**
Loan Identifier	Unique ID for a mortgage loan.
Monthly Reporting Period	The as-of month and year (MMYYY) for loan information in the record.
Original Interest Rate	The original interest rate as per the mortgage note.
Current Interest Rate	The rate of interest in effect for the periodic installment due.

Variable Name	Description
Original UPB (Un-paid Balance)	The dollar amount of loan stated on the note at the time of loan origination.
Current Actual UPB	The current outstanding unpaid principal balance of the loan.
Original Loan Term	The number of months in which the monthly borrower payments are due since the loan origination.
Origination Date	The date of each individual note in MMYYYY format.
First Payment Date	The date of the first scheduled loan payment to be made by the borrower in MMYYYY format.
Loan Age	The number of calendar months since the loan's origination date. It is also calculated using the reporting period minus the first payment date.
Remaining Months To Maturity	The number of calendar months remaining until the outstanding balance of the loan amortizes to zero balance.
Original Loan to Value Ratio (LTV)	Amount of loan at origination divided by the value of property; expressed as a percentage.
Number of Borrowers	The number of individuals obligated to repay the loan.
Debt-To-Income (DTI)	The ratio of borrower's total monthly debt expense to the total monthly income at the time of loan origination.
Borrower Credit Score at Origination	Credit score in terms of a numerical value assigned to evaluate quality of borrower's credit.
Co-Borrower Credit Score at Origination	Co-borrower's credit score as per definition mentioned above.
First Time Home Buyer Indicator	An indicator which denotes if the borrower or co-borrower qualifies as a first-time homebuyer. (Y-Yes, N-No)
Loan Purpose	An indicator that reflects whether the mortgage loan is either a refinance mortgage or a purchase money mortgage.
Property Type	An indicator that reflects whether the property type is a condominium, co-operative, planned urban development (PUD), manufactured home, or single-family home.
Number of Units	The number of dwelling units comprising the related mortgage property.

Variable Name	Description
Occupancy Status	The classification reflecting whether the property occupancy status at the time of loan origination was principal, second , investor or unknown.
Property State	Two-letter abbreviation indicating the state within which the property is located.
Mortgage Insurance Percentage	The original percentage of mortgage insurance coverage for the loan.
Current Loan Delinquency Status	The number of months the obligor is delinquent as determined by terms of loan. (00=Current, 01= 30-59 days, 02= 60-89 days, 03= 90-119 days, 04= 120-149 days, XX= unknown)
Loan Payment History	The coded string of values that reflects the payment performance of the loan over the most recent 24 months from right to left.
Modification Flag	Indicator denoting if the mortgage loan has been modified. (Y/N)
Zero Balance Code	A code which indicates the reason the loan's balance was reduced to zero or experienced a credit event, if applicable.
Zero Balance Effective Date	The date on which the loan balance reduced to zero.
UPB at the Time of Removal	The unpaid principal balance amount at the time of loan hitting the zero-balance code or is liquidated.
Foreclosure Date	The date on which the legal action of foreclosure was completed. Also referred to as the liquidation or sale date.

Note: Own representation, Source: Fannie Mae® Single-Family Loan Performance Data Glossary and File Layout

3.2. Macroeconomic Data

In addition to the historical credit information on loans, five macroeconomic variables are also analyzed in the thesis. These additional macroeconomic variables differ in terms of the periodicity and granularity. Real GDP, Unemployment rates and HPI are available at a more refined level i.e., at state (50 states and District of Columbia) level, while CPI is available at regional (West, South, Midwest, and Northeast) level. The 30-year and 15-year mortgage interest rates are available at a national level. The time-varying macroeconomic variables are retrieved for 7-year period starting from 2014 up to 2021 to supplement the loan performance dataset for predicting loan defaults and are described in Table 3. The data has been retrieved from Federal Reserve Bank of St. Louis (FRED) through pandas DataReader package in python and an Application Programming Interface (API) key from FRED. Only macroeconomic data pertinent to the years of reporting of loans are used.

Table 3 - *Description of Macroeconomic Variables*

Variable	Description	Level	Frequency	Source
GDP	Real Gross Domestic Product measuring state of economic performance	State	Quarterly	U.S. Bureau of Economic Analysis
UR	Unemployment Rate measured by number of unemployed as a % of labor force	State	Monthly	U.S. Bureau of Labor Statistics
CPI	Consumer Price Index measuring change in prices of a basket of goods and services	Region	Monthly	U.S. Bureau of Labor Statistics
HPI	All-Transactions Housing Price Index measuring movement of house prices	State	Quarterly	U.S. Federal Housing Finance Agency
Interest Rates	30-year and 15-year fixed rate mortgage interest rates	National	Weekly	Freddie Mac

Note: Own representation

3.3. Data Preparation

Since the size and dimensionality of the data is a substantial challenge, preparation of data is deemed to be necessary before the machine learning model is implemented. The necessary steps taken to clean and organize the data are presented in this sub-section. The historical loan performance data had 70 features out of which 30 relevant features as per Table 2 were singled out to aid the analysis. The dataset consists of a mix of static variables available at origination of the loan and dynamic variables that change monthly. In the quarterly published acquired loan data files, each loan's monthly performance data from the origination date until mortgage liquidation or maturity with the cut-off date of March 2022 is tracked. The data preparation would ensure that all the monthly observations through the year 2015 till 2021 are compressed into one record for one loan format.

Initially, the loans are filtered based on the single-family home indicator of ‘Property Type’ feature as the focus lies on single-family mortgage loans. Thereafter, the next step is to only incorporate loans originated in and after 2015 through 2020 having a performance period of 12 months (ending 2021), and are, hence, filtered based on the ‘Origination Date’, ‘First Payment Date’, ‘Monthly Reporting Period’ and ‘Loan Age’ columns. The dataset’s index is fixed by the monthly reporting period i.e., the status date and other columns mentioned above are not used for modelling. The dataset consists of various loan terms, out of which, loans with original term of 360 months (30 year) and 180 months (15 year) are filtered. The data for each quarter from 2015 and 2020 having approximately 273 million monthly observations is merged and a cross-sectional dataset is obtained with one observation for each loan. The resultant dataset after checking for duplicate values based on the Loan Identifier feature, had 7,825,919 unique records each representing one mortgage loan which either defaulted or not within 12 months of performance period. The dataset has missing values in the feature columns as depicted in Table 4 .

Table 4- *Percentage of Missing Values*

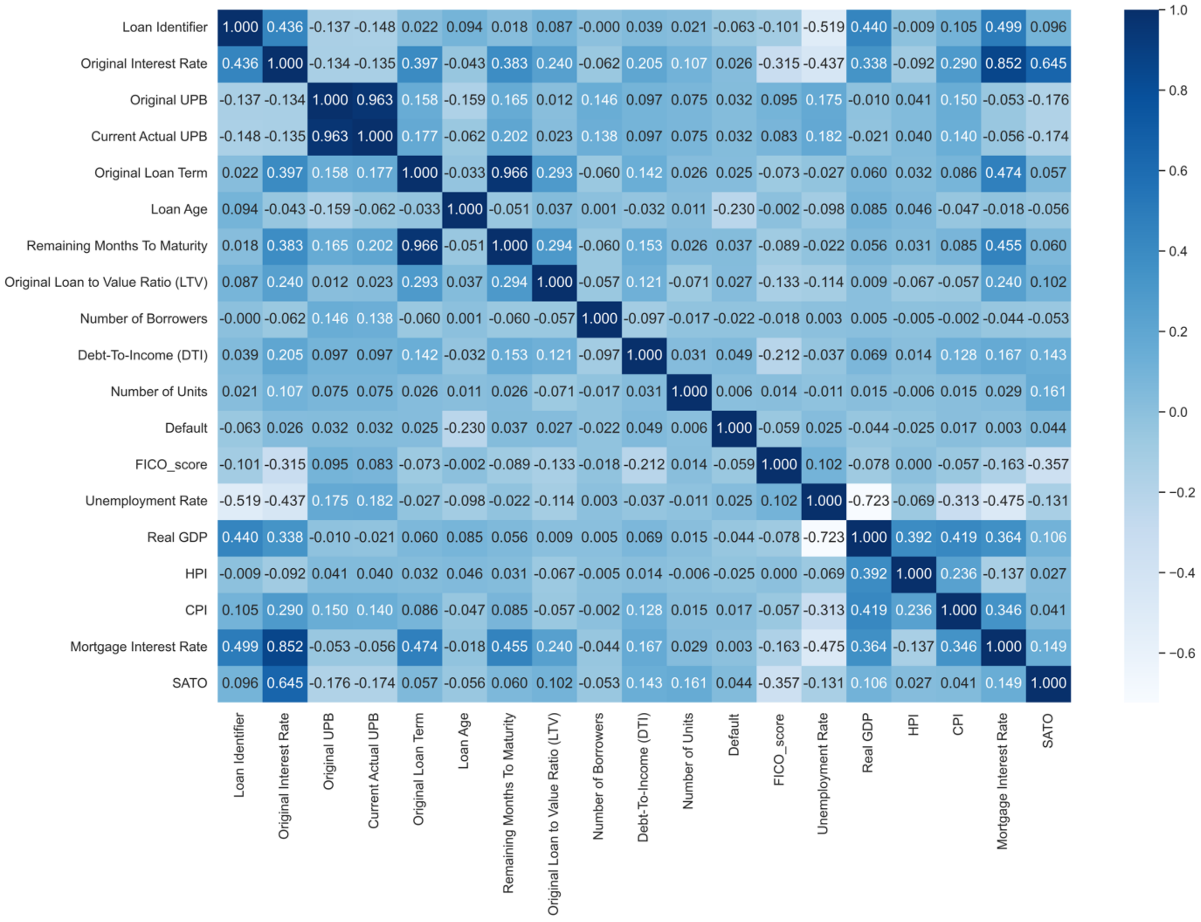
Features	Percentage Missing
Foreclosure Date	100.00
UPB at the Time of Removal	100.00
Zero Balance Effective Date	100.00
Zero Balance Code	100.00
Mortgage Insurance Percentage	71.47
Loan Payment History	55.60
Co-Borrower Credit Score at Origination	51.12
HPI	0.17
Real GDP	0.17
Unemployment Rate	0.17
Borrower Credit Score at Origination	0.062
Debt-To-Income (DTI)	0.018

Features	Percentage Missing
CPI	0.004
Region	0.004
Remaining Months to Maturity	0.001
First Time Home Buyer Indicator	0.00002

Note: Own calculations using python.

The columns with more than 50% missing values are eliminated as they would not serve any purpose in predictions. Since 51.12% loans did not have more than 1 borrower, the borrower and co-borrower credit score features are transformed into 1 feature column ‘FICO score’ by determining the minimum of the two credit scores. After merging the two columns, the Fair Isaac Corporation (FICO) score contained only 0.04% missing values.

Figure 5 - *Correlation Heatmap of Numerical Features*



Note: Own illustration.

Features such as ‘Current Interest Rate’, ‘Current Actual UPB’ and ‘Remaining months to Maturity’ are heavily related to ‘Original Interest Rate’, ‘Original UPB’ and ‘Original Loan

Term' respectively, and are, therefore, dropped. Loan Identifier would also be dropped as it is not relevant for default prediction.

The macroeconomic data as per Table 3 is mapped to the loan level data based on states, and regions, by the date of scheduled first payment date of the loan. The Real GDP and All-transactions HPI were available quarterly while CPI and Unemployment Rates were available monthly. The mortgage interest rates were available weekly and down sampled to get average monthly interest rates. The Real GDP and HPI were linearly interpolated and up sampled to have monthly observations. Every macroeconomic variable is represented in year-over-year percent changes except the mortgage interest rates. The national 30 and 15-year mortgage interest rates are used to derive SATO by deducting the market mortgage interest rate on the mortgage loan from the actual interest rate of the mortgage loan by the date of loan origination. To accommodate the geographic effect on defaults, the United States (US) states are already present in the credit dataset. However, having more than 50 states as a categorical variable for input in machine learning model, this high cardinality might make interpretations more complicated and therefore are grouped into four regions West, North-East, Mid-West, and South in the United States, as per the classification of US Census Bureau.

The target variable 'Default' in the dataset is derived from the 'Current Loan Delinquency Status' variable. The dependent or target variable is labeled '1' (i.e., 'default') if the current delinquency status is equal to 90 days or more within the first 12 months of loan repayment, and labeled '0' otherwise (i.e., no default). Loans which have curated later in their running time have not been considered as non-default.

The instances or loan records pertaining to the territories not contained in the region classification or for which macroeconomic data is not available accounted for 0.17% of the records according to Table 4. As the DecisionTreeClassifier() function of scikit learn library in python, does not intake null values even while predicting in the test set and that the rows with 1 or higher null values accounted for only 0.22% of 7,825,919 records, dropping them all was considered the best option to train an unbiased model. Subsequently, the final dataset consists of 7,808,584 mortgage loans sorted based on the last reporting period and 19 features (including target variable) as per Table 5. The descriptive statistics are presented in the next section.

Table 5- *Features Selected for Modelling*

Variable Name	Characteristic	Values
Original Interest Rate	Loan-specific	Continuous
Original UPB	Loan-specific	Continuous
Original Loan Term	Loan-specific	180, 360
Original Loan to Value Ratio (LTV)	Loan-specific	Continuous
Number of Borrowers	Borrower-specific	1, 2, 3, 4, 5, 6, 8
Debt-To-Income (DTI)	Borrower-specific	Continuous
FICO Score	Borrower-specific	Continuous
First Time Home Buyer Indicator	Borrower-specific	(Y-Yes, N-No)
Loan Purpose	Borrower-specific	C=Cash-out Refinance, R= Refinance, P=Purchase
Number of Units	Loan-specific	1, 2, 3, 4
Occupancy Status	Borrower-specific	P= Principal, S=Second, I= Investor
Region	Loan-specific	West, North-East, Mid-West, and South
Default	Loan-specific	0=No Default, 1=Default
Modification Flag	Loan-specific	(Y-Yes, N-No)
ΔHousing Price Index	Macroeconomic	Continuous
ΔReal GDP	Macroeconomic	Continuous
ΔConsumer Price Index	Macroeconomic	Continuous
ΔUnemployment Rate	Macroeconomic	Continuous
Spread-at-origination (SATO)	Macroeconomic	Continuous

Note: Own representation.

The decision tree has an advantage when it comes to handling data with outliers (Breeden, 2021, p. 17). The trees usually also have another advantage that the dataset does not require standardization or normalization, hence the same have not been applied. Additionally, before the variables are fed as input into the algorithm, the categorical variables namely ‘First Time Home Buyer Indicator’, ‘Loan Purpose’, ‘Occupancy Status’, ‘Modification Flag’ and ‘Region are one-hot encoded. The one-hot encoding creates new binary columns, indicating the presence of each possible element from the original categorical column. The same is done by pd.get_dummies() function of pandas library in Python. The function is performed separately on the training set while training the model and on the test sets while validating or predicting the results.

3.4. Exploratory Analysis

The summary statistics for the full dataset is presented under Table 6. The descriptive statistics are presented for the pre-pandemic period (2015-2019) and the year 2020 marked as the pandemic year to show the variations in the data pre and post crisis. Due to the high number of loans in 2021 and an offsetting effect on the data from 2020, to evaluate the models out-of-time on the pandemic period, loans for the year 2020 will be utilized.

Table 6- *Summary Statistics of Numeric Variables based on Full Dataset*

A) Pre-pandemic (2015-19) **Features**	**Mean**	**Std Dev**	**25th Pctl**	**50th Pctl**	**75th Pctl**
Original Interest Rate	4.16	0.61	3.75	4.12	4.56
Original UPB	229337.78	124350.01	134000	204000	304000
Original Loan Term	329.47	67.55	360	360	360
Original (LTV)	75.18	17.04	67	80	90
Number of Borrowers	1.5	0.52	1	1	2
Debt-To-Income (DTI)	34.41	9.26	28	36	42
Number of Units	1.05	0.29	1	1	1
FICO score	743.11	48.15	708	751	784
ΔUnemployment Rate	-9.96	7.38	-15.22	-10.81	-5.71
ΔReal GDP	2.43	1.74	1.21	2.38	3.66
ΔHPI	5.59	2.29	3.77	5.54	7.05
ΔCPI	1.61	1.05	1.03	1.62	2.38
SATO	0.32	0.37	0.06	0.25	0.53

B) Pandemic (2020)

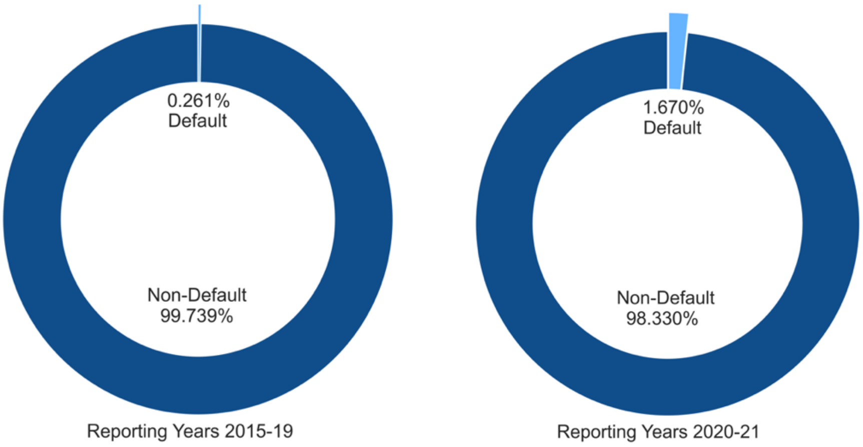
Features	**Mean**	**Std Dev**	**25th Pctl**	**50th Pctl**	**75th Pctl**
Original Interest Rate	4.15	0.64	3.75	4	4.5
Original UPB	266440.18	139862.11	158000	241000	352000
Original Loan Term	336.6	60.53	360	360	360

Features	Mean	Std Dev	25th Pctl	50th Pctl	75th Pctl
Original (LTV)	74.75	17.42	65	79	90
Number of Borrowers	1.48	0.52	1	1	2
Debt-To-Income (DTI)	35.3	9.48	29	36	43
Number of Units	1.03	0.25	1	1	1
FICO score	746	45.26	714	753	783
ΔUnemployment Rate	15.88	73.69	-9.52	-4.65	2.22
ΔReal GDP	1.31	2.74	0.51	2.07	2.99
ΔHPI	4.68	1.19	3.8	4.7	5.37
ΔCPI	1.91	0.69	1.41	1.8	2.64
SATO	0.39	0.45	0.11	0.3	0.61

Note: Own calculations in Python.

The descriptive statistics conveys that the average loan unpaid balance in pre-pandemic period stood at $ 229,338 whereas increased to $ 266,440 in the pandemic year. On the contrary, the mean Debt to Income (DTI) ratio declined while an increase in the average FICO score.

Figure 6- *Loan Defaults in Pre and Post Pandemic Years*



Note: Cumulative default rate for the years 2015 to 2019 and 2020 to 2021.

The number of loan records in the performance period of 2015 till 2019 stood at 4,176,549, while in the years 2020 and 2021 the number of loans stood at 1,254,020 and 2,378,015 respectively. The percentage of default in these periods is depicted in Figure 6 whereas yearly individual loan observations are mentioned under Table 7 and illustrated in Figure 7.

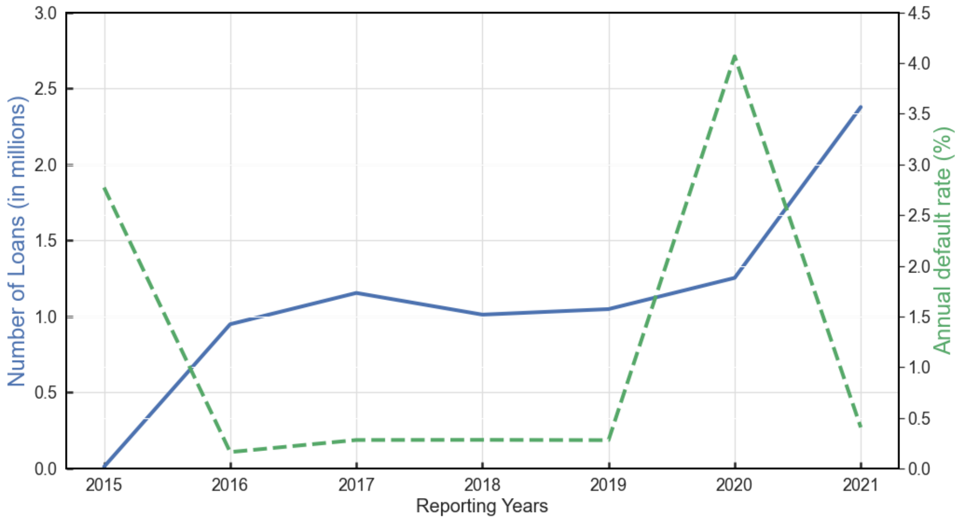
Table 7- *Number of Loans and Defaults*

Year	Observation Count	Number of Defaults
2015	11,686	324
2016	949,210	1538
2017	1,154,864	3247
2018	1,012,195	2861
2019	1,048,594	2933
2020	1,254,020	51015

Note: Own calculations.

Figure 7 depicts that even though the number of loans increased in the year of 2020, the number of defaults increased simultaneously, accelerating the percentage of default in the years 2020-21. But 1.67% of default in the post pandemic years is lower than expected due high number of loans in 2021 and less defaults in that year.

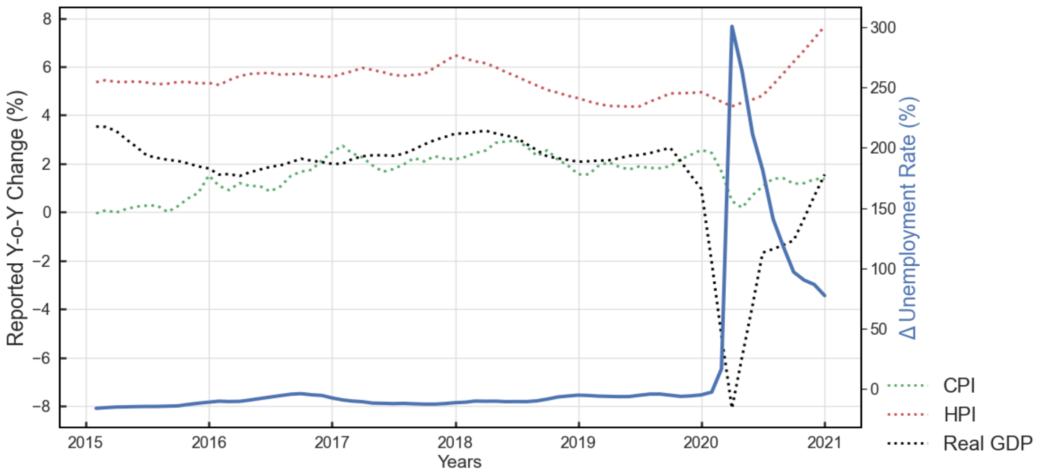
Figure 7- *Number of Loans and Default rate*



Note: Own calculations and representation. Source: Fannie Mae single-family loan performance data.

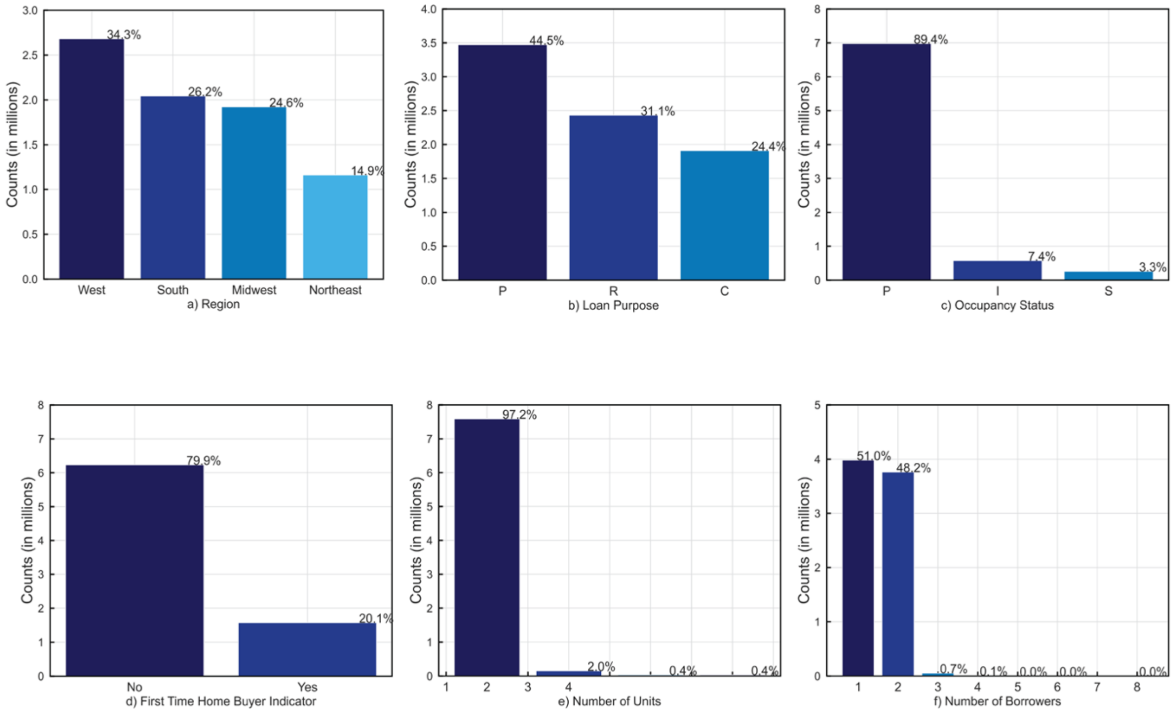
The highest annual default rate peaked at approximately 4% for the year 2020 evidently due to the pandemic. However, despite the pandemic, the number of loans originated in 2020 increased and, therefore, the number of loans in the reporting period of 2021 are at elevated levels. As the default levels in the year 2020 were affected by the pandemic, so was the state of economy as can be seen in Figure 8. The real GDP dropped more than 8% while unemployment rate increased manyfold. The CPI also declined but increased within months after the pandemic struck. The increasing loans in the year 2020 and afterwards could be attributed to the growing housing price index which reflects changes in the housing prices. It is necessary to mention that the CPI reported here is based on four regions of the U.S while the unemployment rate, HPI and real GDP are collected on a state level.

Figure 8 - *Overview of Macroeconomic Indicators*



Note: Own representation. Source: FRED.

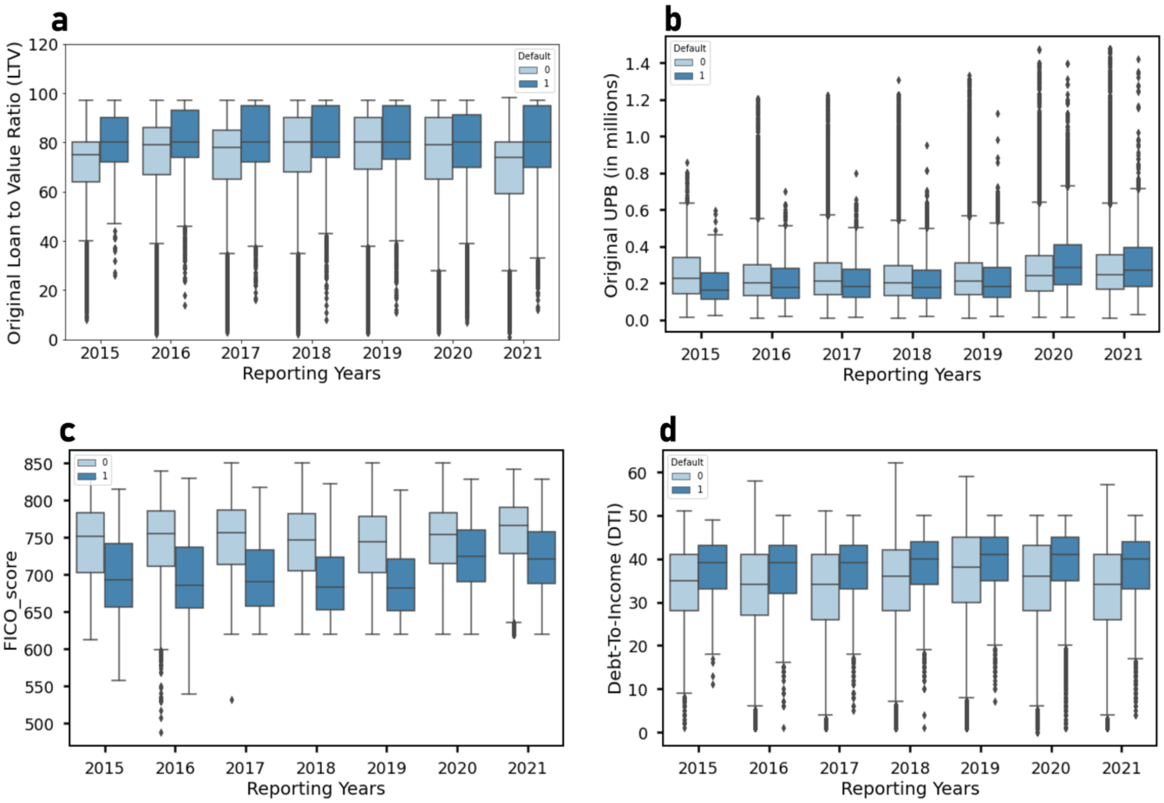
Figure 9 - *Bar Plots of Categorical and Numerical Variables*



Note: Own representation.

Since the number of states was fifty, they are clustered together based on the region and the categorical variable of 'Region' was derived. Figure 9 (a) shows that majority of the properties mortgaged for loans are situated in the West region with 34.3% of the loans followed by the South and Midwest region. The mortgage loans differ over the purpose for which the loan was availed. The same is depicted by 'Loan Purpose' in Figure 9 (b) where close to 45% loans are availed for purchase money mortgage, 31% are refinance mortgages and 24% are cash-out refinances. When the loan originates, the 'Occupancy Status' is classified as the principal residence, second home, or an investment property. Approximately 89% of the loans in the dataset has occupancy status as principal. The loan dataset also states whether the buyer is purchasing the home for the first time, which is the case in approximately 80% of the loans as depicted in Figure 9 (d). The number of borrowers is almost balanced between 1 and 2, while more than 3 borrowers are very rare according to Figure 9 (f). Lastly, the number of dwelling units in the property are depicted in Figure 9 (e). Some other illustrative examples which may explain the relationship between specific variables and loan defaults are also presented in Figure 10 and 11 to cultivate some intuition for the data and the algorithms.

Figure 10 - *Box Plots of Continuous Variables*

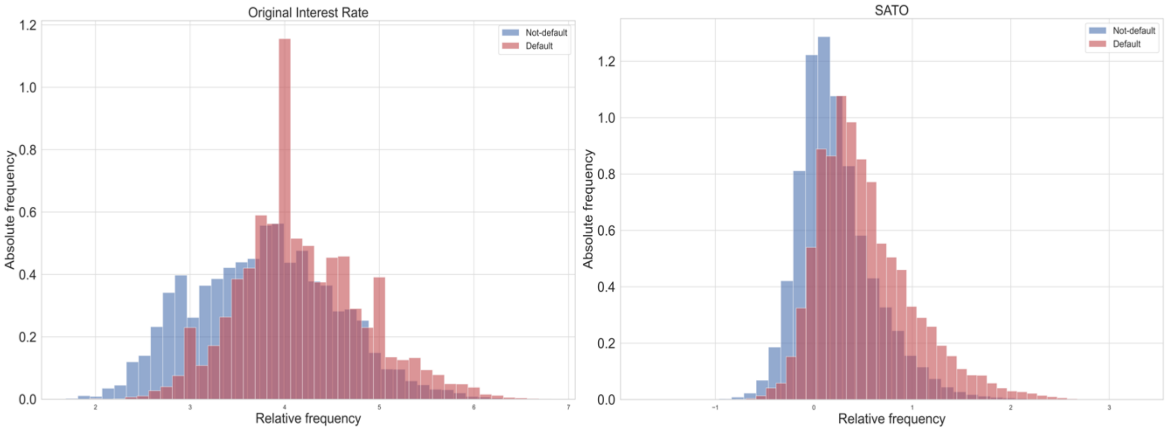


Note: Own representation.

Figure 10 (a) shows that the average loan to value ratio at the time of loan origination has remained stable over the years, but the defaulting loans have had a higher ratio of loan to value of the property on an average. The average Loan to Value (LTV) ratio has been approximately 75% before the pandemic and improved slightly to 74% in the pandemic year. The Original UPB of the loan has been higher for non-defaulters until 2019 after which the defaulters had a higher unpaid balance on an average as evident from Figure 10 (b). The FICO score, a scoring which is highly indicative of defaults, has been intuitively lower throughout all years for the defaulters. But the level of FICO scores increased in the reporting years 2020 onwards. For non-defaulters, the average FICO scores have been stable across all years as depicted in Figure 10 (c). The average FICO score before the pandemic stood at 743 and ranged from anything between 488 (high risk) to 850 (low risk). Subsequently, the DTI ratio, indicating borrower risk which represents the proportion of monthly income used to pay the debt, has been on average higher for defaulters rather than non-defaulters for all the years as per Figure 10 (d). The same ranged between 28 to 58, with an average of 34.41.

The individual box plots and distributions of the variables in Figures 8 and 10 are displayed in Appendix I. Lastly, Figure 11 depicts the interest rate and the spreads at the time of loan origination. Both the interest rates and the spreads at origination have been lower for non-default loans as opposed to defaulting loans. This is instinctive as an increase in interest rates raises the debt burden on borrowers.

Figure 11 - *Histogram for Interest Rates and SATO*



Note: Own representation.

Even though a few categorical variables such as Occupancy Status, First Time Home Buyer Indicator, and Modification Flag and numerical variables such as Number of Units are highly inclined towards a single category, the features are still used as inputs into the machine learning algorithm. Later, the resultant feature importance of the decision tree can be used to verify whether these variables are important in loan default prediction. The next section describes the methods and implementation of the machine learning model using the data described in this section.

4. Empirical Research

This chapter provides the necessary steps taken to implement and evaluate the decision tree models to assess the model performance pre and post pandemic along with impact of macroeconomic factors. The analysis is performed using Python programming language. The aim of the research is to show the impact of pandemic on the performance of a machine learning model namely decision tree. Moreover, traditional credit performance data is integrated with macroeconomic variables to test for any improvement in the machine learning model. In order to achieve this, decision tree models are tested using different scenarios over different time periods to capture the impact of pandemic and the macroeconomic variables. Different sampling techniques are employed due to the dataset consisting of extremely imbalanced classes. The decision tree models are implemented using no resampling, under-sampling, and oversampling techniques both with and without macroeconomic variables. Thereafter, based on the results, the importance of the features is discussed. The sub-section 4.1 provides the overview on splitting of the dataset into training and test set based on the reporting years of the loan. In sub-section 4.2 the implementation of decision tree models using different sampling techniques is discussed which peruse loan performance data with and without macroeconomic variables. In sub-section 4.3, the resulting important features are presented. The results are consolidated in section 5.

4.1. Train-Test Split

The model building and testing through different sampling techniques is done using data from 2015 till 2020 and split according to Table 8. Due to splitting the data based on the reporting years, the in-sample and out-of-sample split results in a ratio of 51:49. Moreover, as the pandemic hit in the year 2020, the same is used as out of time test set to study the model performance when predicting out of time. Many financial institutions build their credit scoring models to implement out-of-time in actual business activities. Additionally, any consumer related model built on historical data would be sensitive to any drastic changes in the future consumer behavior. Hence, due to this reason, along with the pandemic, the emphasis lies on the out-of-time analysis. The models built are henceforth tested on out-of-sample data and out-of-time data. This technique can also serve as an indicator about the stability of a model over time. The training, test and out of time split in the Table 8 would be used in the scenarios where the model is evaluated using different sampling techniques with and without macroeconomic data.

Table 8 - *In-sample, Out-of-sample, and Out-of-time Period for Different Sampling Techniques*

	Period	Number of Observations
In-sample Period	01-01-2015 to 01-12-2017	2,115,760
Out-of-sample Period	01-12-2018 to 01-12-2018	2,060,789
Out-of-time Period	01-01-2020 to 01-12-2020	1,254,020

To assess the influence of macroeconomic factors along with loan performance data, the models are evaluated using two datasets. The dataset simply created with the historical loan performance data is named Ensemble I, and second dataset created with the historical loan performance data complemented with macroeconomic data is named Ensemble II.

4.2. Decision Tree Model with Different Sampling Techniques

Various resampling techniques are utilized to address the problems posed by imbalanced classes in predictions. Moreover, the size of dataset also influences the learning process of algorithms. In this section, the model setting for training of a decision tree using various sampling techniques are presented. The models are then evaluated on out of sample and out of time data as exhibited in Table 8 of sub-section 4.1.

If the parameters of a decision tree are not optimized, the tree can grow to its full size i.e., depth and lead to overfitting. Hence, the hyperparameter of max_depth for the decision tree is fine-tuned during fitting of the models on the training data. The default setting in the DecisionTreeClassifier is kept as ‘entropy’ for criterion of node splits and the class weights as ‘balanced’ to address the data imbalance issues by the model internally assigning the weights to classes inversely proportional to their corresponding frequencies. When tested with class weights as not balanced, all the ratios were at a minimal level and accuracy was the highest. The value range for max_depth is provided in Table 9 for different sampling techniques used. The range is different for SMOTE technique due to excessively high amount of data.

Table 9 - *Range of Tested Hyperparameters*

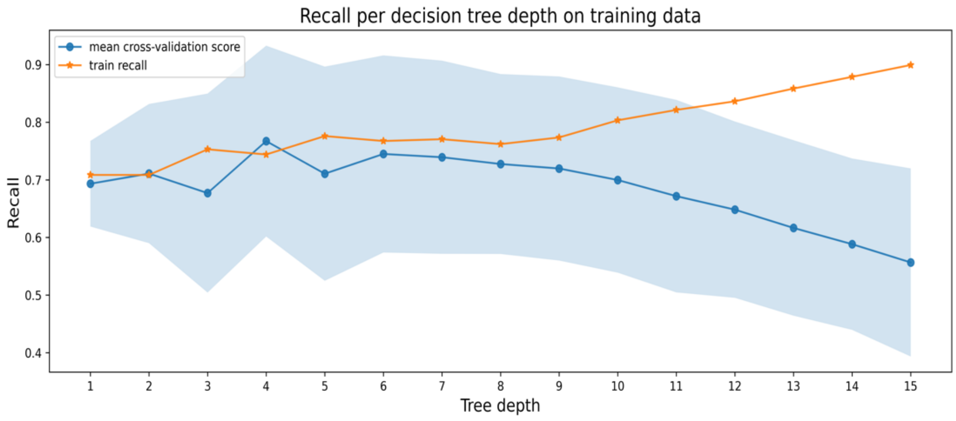
Sampling Technique	Value/Range for max_depth parameter
No resampling	1 to 15
Undersampling	1 to 15
Oversampling	2 to 10

For implementing SMOTE and hyperparameter tuning together, the GridSearchCV method of Scikit-learn is used with 10-fold cross-validation. The model simulates different decision trees with the distinct range of values for the hyperparameter according to Table 9, simultaneously cross-validating them and in the end outputs the decision tree model with the best cross-validated score. This outputted model with the highest cross-validated score is then fitted on the training data for subsequent predictions. Similarly, for no resampling and undersampling techniques, the cross validated scores are computed at each iteration of a 10-fold cross validation on the training set for the values of the specified hyperparameters and the hyperparameter with the best average score is selected to further fit the model on the training set and predict on out-of-sample and out-of-time data. The score on which the models are optimized can be one of many performance metrics mentioned in section 2.2.4. For the analysis, recall score is used as the performance metric to optimize the decision tree models. All these sampling techniques are used on Ensemble I i.e., simple loan performance data, as well as on Ensemble II i.e., loan performance data complemented with macroeconomic variables. The out of sample and out of time dataset are not resampled and used as they are. The same steps were also followed for optimizing decision tree based on balanced accuracy score and the summarized results are present in section 5.1

4.2.1. No Resampling

The training dataset of Ensemble I from the year 2015 till 2017 is used to optimize the decision tree based on methods described in section 4.2. In order to avoid over or under fitting, the cross validated recall score is calculated corresponding to the hyperparameter maximum depth of the decision tree, and the highest mean cv-recall score is obtained with the hyperparameter.

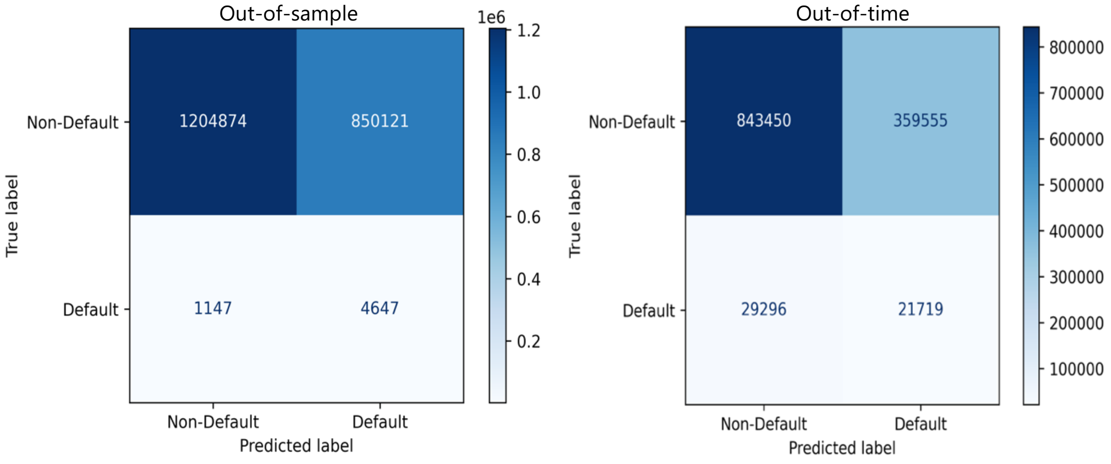
Figure 12 - *Cross-validated Recall Scores and Hyperparameter*



Note: Own illustration

For instance, Figure 12 displays the mean cv-scores across a range of 1 to 15 depth of the tree and the optimal max depth is obtained where the mean cv-recall score is the highest. In this case, the highest cv-recall score corresponds to the max depth of 4. The same method is true for all the models and hence only an example in this section is shown.

Figure 13- *Out-of-sample and Out-of-time Confusion Matrix for No Resampling Technique on Ensemble I*



Parameters	*In-sample*	*Out-of-sample*	*Out-of-time*
Recall	*0.746*	*0.802*	*0.426*
Balanced Accuracy	*0.746*	*0.694*	*0.563*
Specificity	*0.746*	*0.586*	*0.701*
Description	*Value*		
Max_depth	*4*		

CV-Recall Score	*0.760*
Standard Deviation	*0.07*

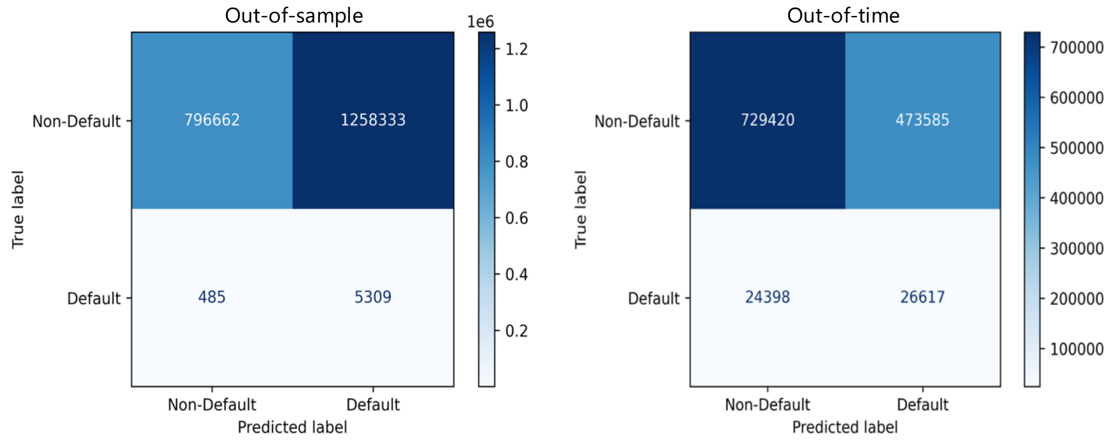
Note: Own illustration.

The Figure 13 consists of recall, specificity, and balanced accuracy scores reported for out-of-sample and out-of-time evaluation of the model after the fitting of best model on the training data. To have a base comparison of how the model is performing out-of-sample and out-of-time, the performance metrics for the training set are also provided. The corresponding confusion matrices are also illustrated in Figure 13.

It is observed that the out of sample recall score slightly gets better when evaluating on out-of-sample data, while the balanced accuracy only slightly declines. This is attributed to the fact that the in-sample and out-of-sample data both belong to the same pre-pandemic period and the model is able to generalize well on the out-of-sample data for the year 2018-2019. As soon as the model is evaluated on out-of-time data from the pandemic year 2020, the recall declines by 50% to a mere 0.423. This means that when predicted out of time during the pandemic, the model is not able to predict loan defaults as well as it did it for the out-of-sample period. It is imperative to mention that as the test size increases, even though the ability to predict defaults was better in out-of-sample period, this came at a cost of drastically increasing false positives which can be captured by the specificity measure. This also implies that the precision score and F1 score suffered significantly. Since the focus lies on detecting the defaults correctly, the scores reported are recall and balanced accuracy. However, the precision and F1 scores can be easily calculated using the classification matrix and formulae as per section 2.2.4.

To assess whether incorporating macroeconomic variables with the loan performance data can enhance the predictive power of the decision tree both out-of-sample and out-of-time, Ensemble II is used to build the model, train, and evaluate out-of-sample and out-of-time. The same procedure as listed in the previous section of hyperparameter tuning is applied. Figure 14 depicts the confusion matrices corresponding to the out-of-sample and out-of-time evaluation of the model.

Figure 14- *Out-of-sample and Out-of-time Confusion Matrix for No Resampling Technique on Ensemble II*

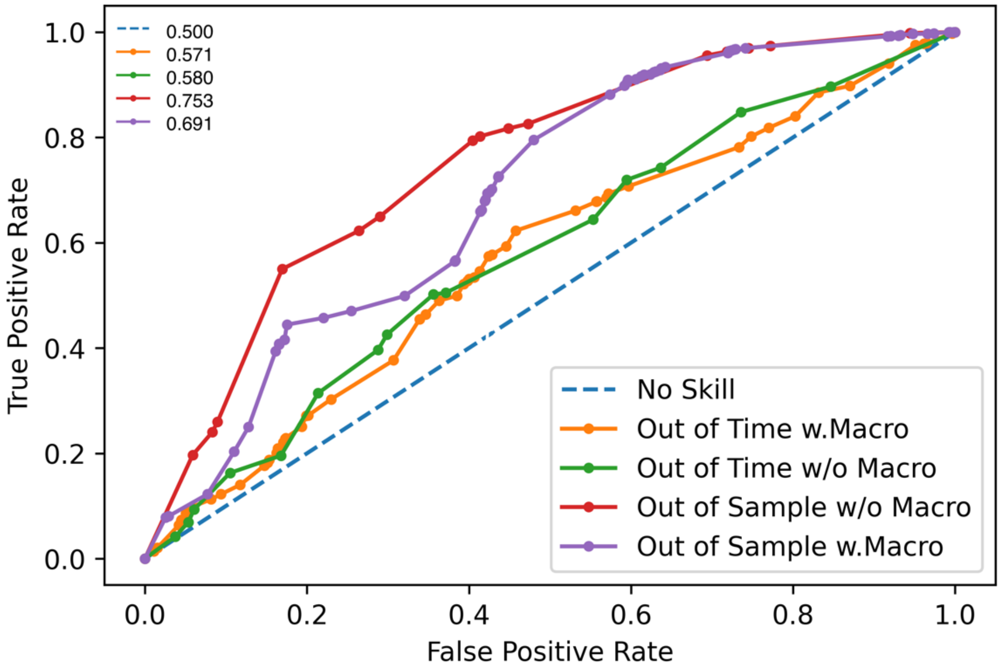


Parameters	*In-sample*	*Out-of-sample*	*Out-of-time*
Recall	*0.815*	*0.916*	*0.522*
Balanced Accuracy	*0.787*	*0.651*	*0.564*
Specificity	*0.758*	*0.388*	*0.606*
Max_depth	*6*		
CV-Recall Score	*0.752*		
Standard Deviation	*0.15*		

Note: Own illustration

In contrast to the Ensemble I, out-of-sample evaluation of this model gives a higher recall with classifying correctly 91.6% of the loans that defaulted. The recall score was also better on evaluating out-of-time as compared to Ensemble I, but the model was able to capture the decline in predictive performance due to pandemic. The “max_depth” parameter increased from 4 in Ensemble I to 6 in Ensemble II. This indicates that a deeper tree was formed as more data was fed into the model. Even though the models were not optimized based on the AUC-ROC curve but based on the recall score, the comparison with respect to the AUC-ROC score among both the models for out-of-sample and out-of-time performance is depicted in Figure 15. Despite the recall score of the Ensemble II being better, Figure 15 depicts that the balance between the true positive rate and false positive rate is better for no resampling technique with Ensemble I i.e., the quality of model’s predictions is better.

Figure 15 - *Receiver Operating Characteristic Curve- No Resampling*

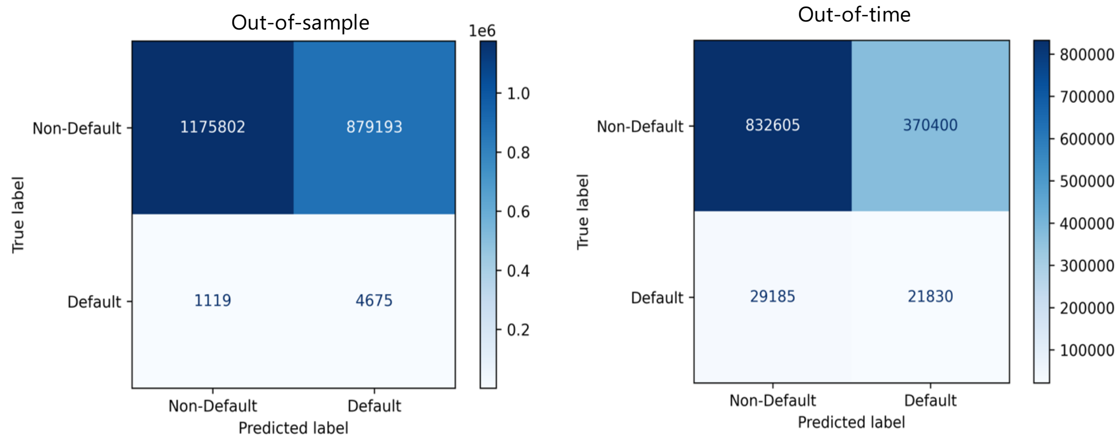


Note: Own illustration.

4.2.2. Under sampling

Due to the imbalanced classes, the training set from Ensemble I is under sampled to have 1% of the majority class observations without replacement and 100% of the minority class. Due to this the training data is reduced by a considerable size without touching the out-of-sample and out-of-time set. Choosing 1% of the majority class still does not equate the majority and minority class numbers but reduces the proportion of majority to minority class substantially. Again, the hyperparameters are tuned for the decision tree based on methods described in section 4.2. Figure 16 depicts the confusion matrices along with the parameters of the model and performance measurement metrics.

Figure 16 - *Out-of-sample and Out-of-time Confusion Matrix for Undersampling Technique on Ensemble* I

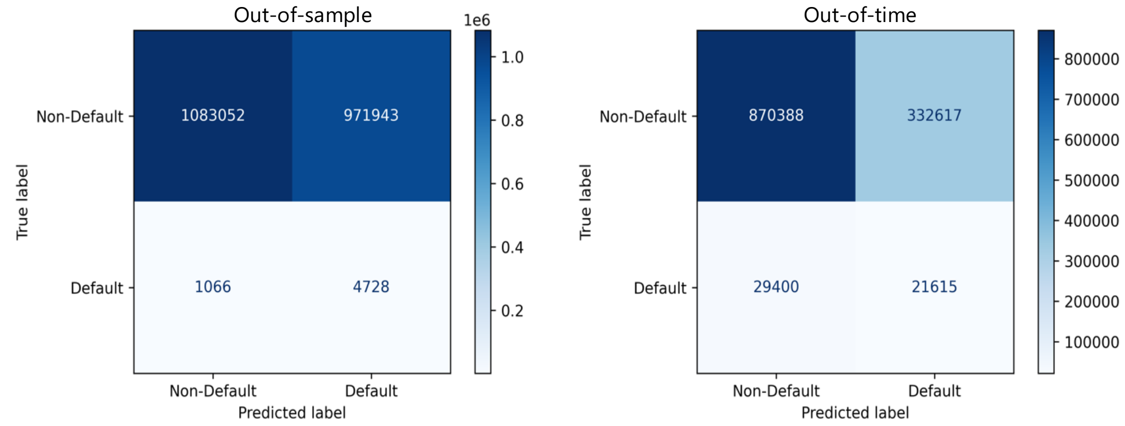


Parameters	*In-sample*	*Out-of-sample*	*Out-of-time*
Recall	*0.752*	*0.807*	*0.428*
Balanced Accuracy	*0.747*	*0.690*	*0.560*
Specificity	*0.742*	*0.572*	*0.692*
Max_depth	*4*		
CV-Recall Score	*0.767*		
Standard Deviation	*0.08*		

Note: Own illustration.

It is observed that the scores for Ensemble I in undersampling almost mimic the results from no resampling technique with Ensemble I. The out of sample and out-of-time performance both are similar along with the recall and accuracy scores. However, with Ensemble II, the recall score is lower than the no resampling technique with Ensemble II. Moreover, addition of the macroeconomic variables only slightly adds value than with no macroeconomic variables when compared within the undersampling techniques. This is observed both in out-of-sample and out-of-time performance. The impact of pandemic remains consistent on the performance of the model wherein the score declines by approximately 50%. The difference in the features deemed important by the decision tree using different sampling styles are discussed in the section 4.3. Figure 17 depicts the confusion matrices corresponding to the out-of-sample and out-of-time evaluation of the model on the macroeconomic data or Ensemble II. Alongside, the balance accuracy, specificity, and recall scores are also mentioned in Figure 17.

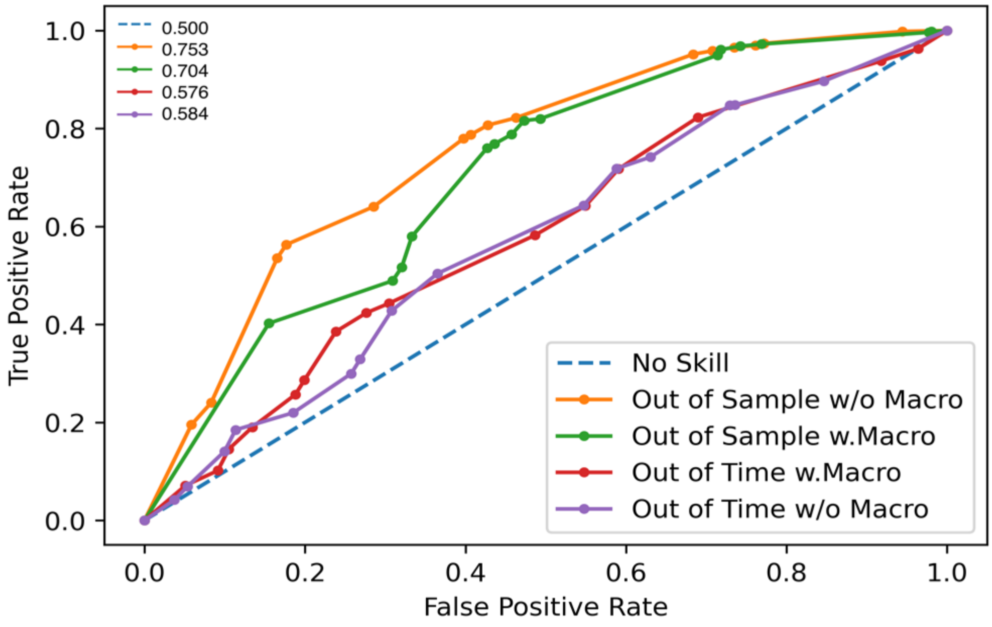
Figure 17 *- Out-of-sample and Out-of-time Confusion Matrix for Undersampling Technique on Ensemble II*



Parameters	*In-sample*	*Out-of-sample*	*Out-of-time*
Recall	*0.764*	*0.816*	*0.428*
Balanced Accuracy	*0.765*	*0.672*	*0.573*
Specificity	*0.766*	*0.527*	*0.724*
Max_depth	*4*		
CV-Recall Score	*0.745*		
Standard Deviation	*0.12*		

Note: Own illustration.

Figure 18 *- Receiver Operating Characteristic Curve- Undersampling*



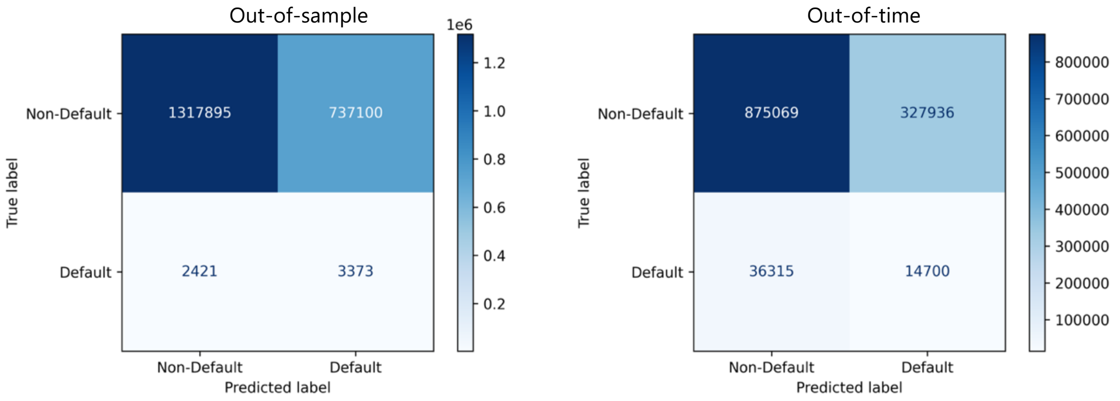
Note: Own illustration

Both the models using undersampling technique were able to correctly predict approximately 81% of the loans defaulting out of the total defaults in the out-of-sample performance. Likewise, the out-of-time performance was alike, though underwhelming, for both the Ensemble I and Ensemble II in undersampling. The ROC curve is depicted for comparison between the models in Figure 18.

4.2.3. Oversampling (SMOTE)

Scikit-learn has a feature of SMOTE that can be applied to data which can synthetically replicate the minority class based on *k*-nearest neighbors. The default value of *k*=5 is utilized, and the data is oversampled. The technique is combined with the GridSearchCV using pipeline in Python. This allows for oversampling within the cross-validation folds and not on the whole training set. The best cross-validation score and the corresponding hyperparameter is derived and described in Figure 19.

Figure 19 - *Out-of-sample and Out-of-time Confusion Matrix for SMOTE Technique on Ensemble I*

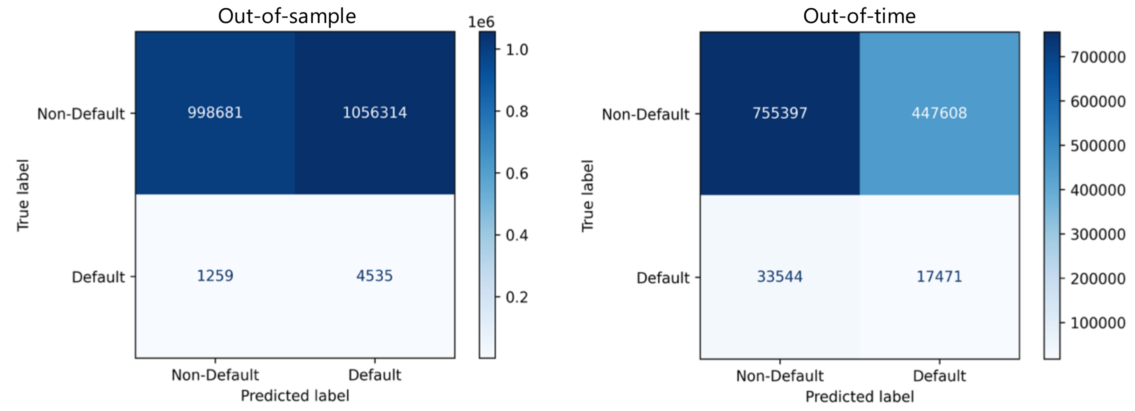


Parameters	*In-sample*	*Out-of-sample*	*Out-of-time*
Recall	*0.560*	*0.582*	*0.288*
Balanced Accuracy	*0.68*	*0.612*	*0.508*
Specificity	*0.804*	*0.641*	*0.727*
Max_depth	*4*		
CV-Recall Score	*0.550*		
Standard Deviation	*0.03*		

Note: Own illustration

The oversampling method did worse than the other sampling techniques. Upon evaluating on the out-of-sample Ensemble I data, only 58% of the loans were correctly classified as defaulting meaning the ratio of false negatives was extremely high. Similarly, it performed even worse on out-of-time data. The cross-validated recall score is only 0.55 when compared to other sampling techniques with the score of more than 0.70 with approximately the same depth of tree. Figure 20 depicts the results of oversampling with the macroeconomic data Ensemble II.

Figure 20 - *Out-of-sample and Out-of-time Confusion Matrix for SMOTE Technique on Ensemble II*

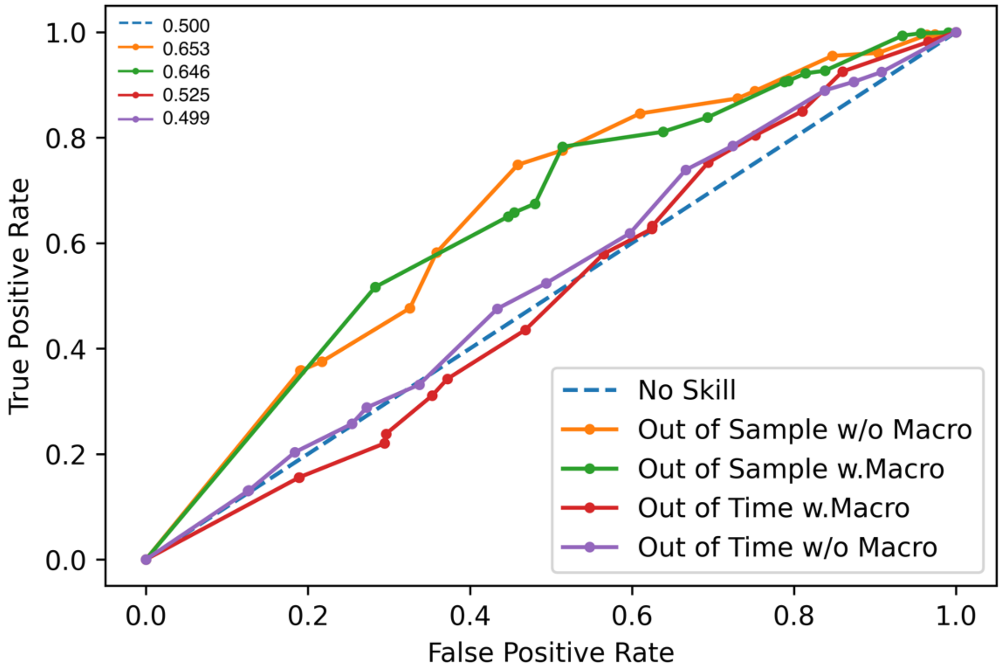


Parameters	*In-sample*	*Out-of-sample*	*Out-of-time*
Recall	*0.613*	*0.783*	*0.342*
Balanced Accuracy	*0.703*	*0.634*	*0.485*
Specificity	*0.789*	*0.486*	*0.628*
Max_depth	*4*		
CV-Recall Score	*0.584*		
Standard Deviation	*0.14*		

Note: Own illustration

Despite the cross-validated recall score of 0.58, the out-of-sample performance with macroeconomic data returned a recall score of 0.78. The varying nature of macroeconomic data could induce variance in the data and therefore the scores fluctuate immensely. Compared to other techniques, this technique with and without macroeconomic data performed the poorest. This could also be attributed that due to oversampling on already large data, the algorithm may have not been able to learn properly upon the training and validation of the model. The balanced accuracy scores were also not up to the mark for both the oversampling models with and without macroeconomic data. The ROC curve depicted in Figure 21 represents the trade-off between the true and false positive rates at different thresholds for the probabilities predicted for loan default.

Figure 21- *Receiver Operating Characteristic Curve- Oversampling (SMOTE)*



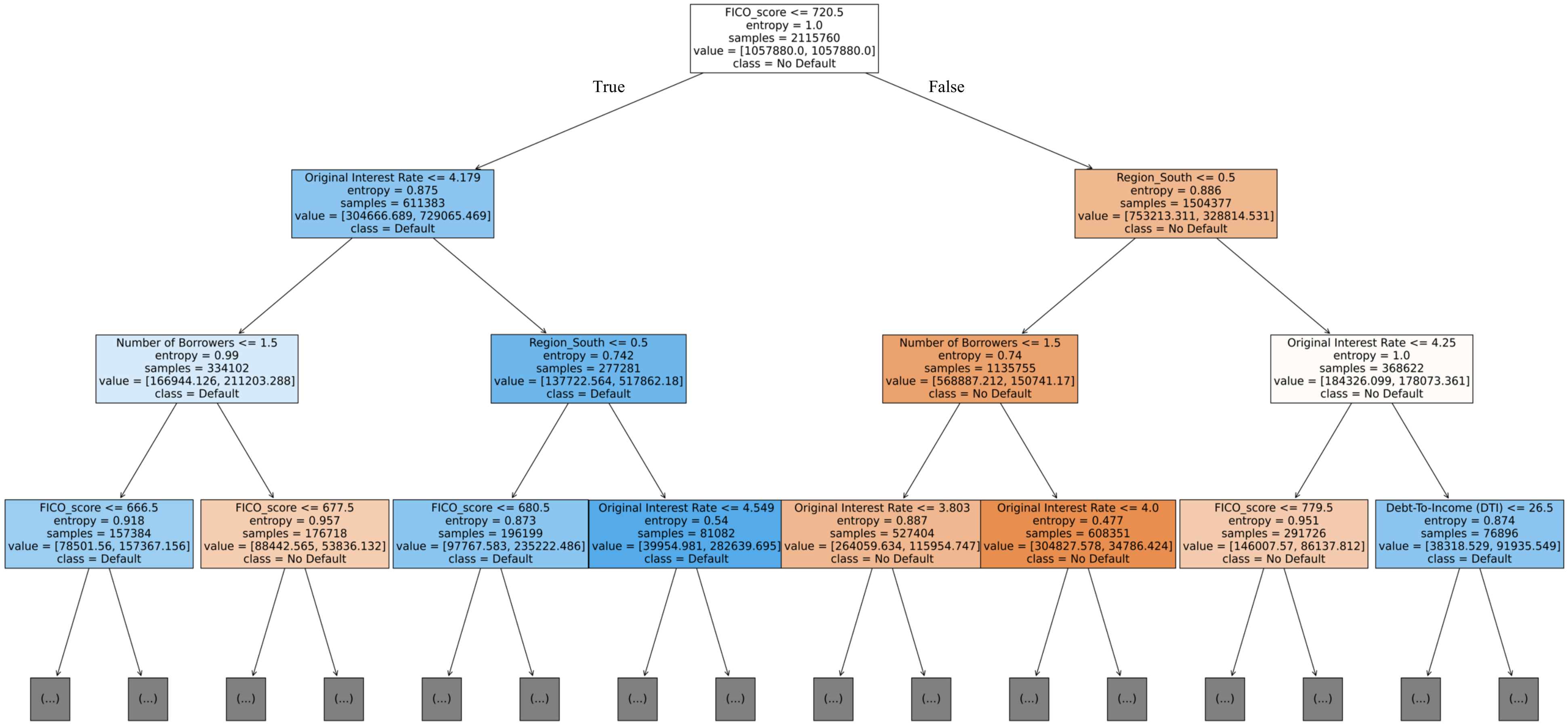
Note: Own illustration

It is worth mentioning that as the focus of the models was on recall scores, due to large amount of data, the F1 and precision scores of all the models ranged between 0.004 to 0.10 only when evaluated on out-of-sample and out-of-time data.

4.3. Feature Importance and Interpretations

The decision tree can capture the non-linear relations in the data and can be easily visualized after the model if fitted. Figure 22 shows the Decision Tree which was trained and fine-tuned with no resampling technique with Ensemble I i.e., only loan performance data. Due to space constraints the maximum depth of 3 is displayed. At the first node, the best feature is chosen by the fitted decision tree model and the quality of the split is determined by entropy. The best result is derived by maximizing the information gain after every split.

Figure 22 - *Decision Tree of Ensemble I Data- No resampling*



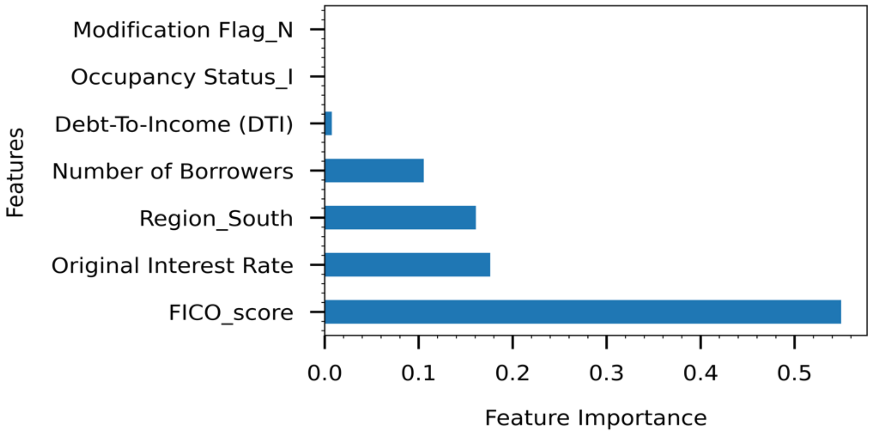
Note: Own illustration.

For instance, in the decision tree diagram of Figure 22, FICO score is chosen as the best feature and the sample splitting cutoff chosen by the algorithm stands at 720.5 FICO score. The line of samples shows the number of samples at that node whereas the value provides the number of samples in each class.

It is noteworthy that equal number of classes are displayed in the first node despite imbalanced dataset, as in python function, the class weights were chosen as balanced while fitting the model. Further, if FICO score is less than or equal to 720.5, i.e., satisfies the condition, the path towards the left is chosen and further split upon the Original Interest Rate feature. The corresponding entropy and the majority class is displayed on the nodes. The darker the shade of the nodes, more the purity in the nodes. Hence as the depth of the tree increases, more refined split criterions are decided by the algorithm. The tree in Figure 22 has a maximum depth of 4 and features marked as important by the algorithm can be derived from the feature importance function in scikit-learn.

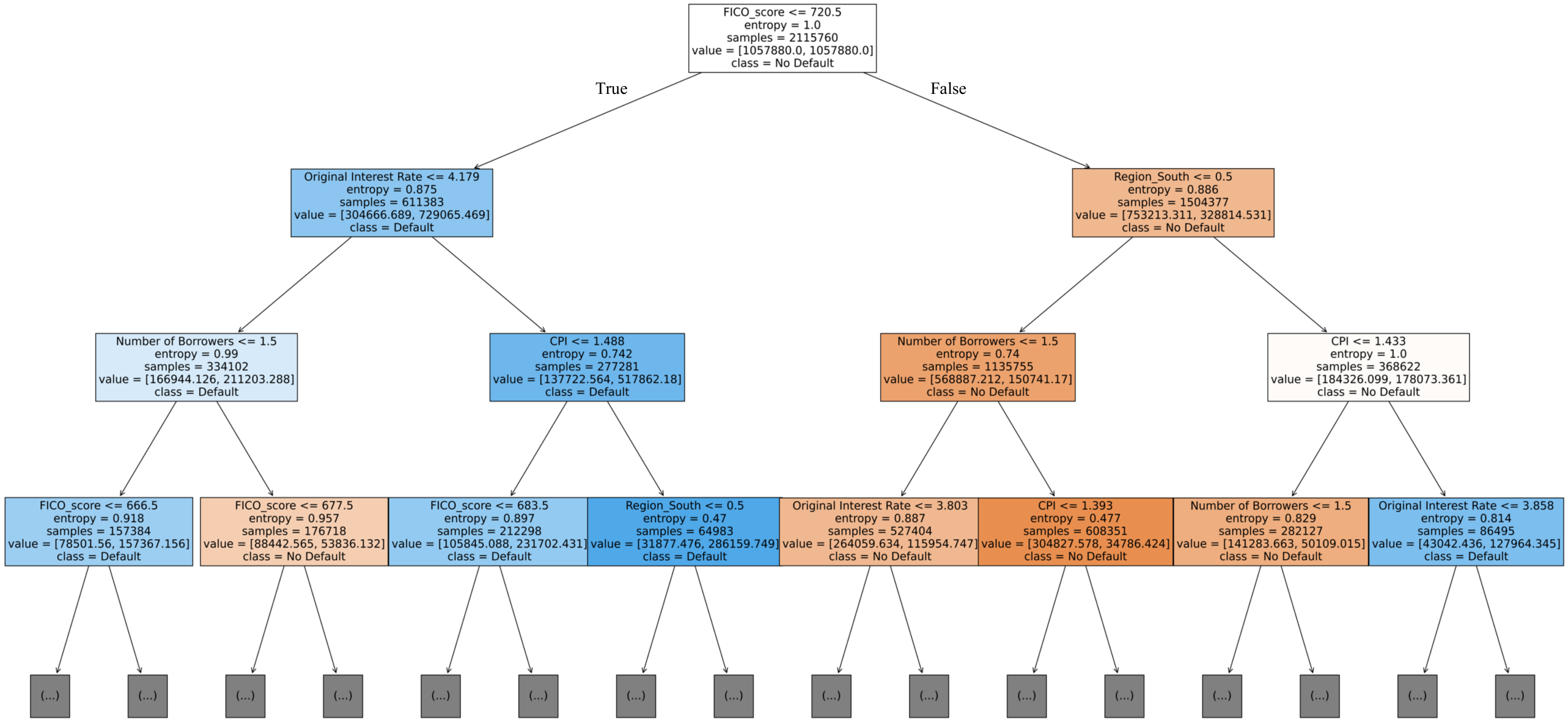
The importance of a feature is measured as a fall in node impurity (entropy) brought about by the feature. A score is assigned to the input variables based on how useful they are at predicting a target variable. Higher the value, higher is the importance. Thus, feature importance can provide insights into the data. Features which were deemed to be important for the decision tree model using no resampling technique are plotted in Figure 23. These correspond to the tree displayed in Figure 22 using Ensemble I data.

Figure 23 - *Feature Importance corresponding to Decision Tree in Figure 22*



Note: Own illustration.

***Figure 24** - Decision Tree of Ensemble II Data- No resampling*

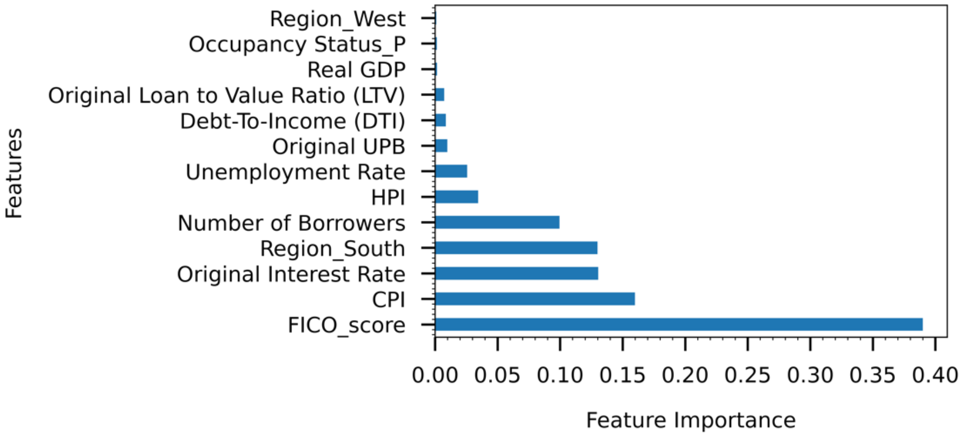


Note: Own illustration.

Out of 22 variables (including encoded variables) in Ensemble I, the decision tree considered only 5 variables to be important as the feature importance coefficient was zero for the other variables. The highest importance was placed to FICO scores followed by the interest rates of the mortgages both of which are rather intuitive. Thereafter, the South region in which the property is situated was considered for node splitting criterion as can be seen on the right path of the decision tree at the first depth in Figure 22. Finally, the number of borrowers and DTI were given the penultimate and last importance.

Similarly, the decision tree built using the no resampling technique on Ensemble II is displayed in Figure 24 and corresponding important features are displayed in Figure 25. Incorporating the macroeconomic variables seemed to have an influence on the decision-making criteria of the tree model. The depth of the optimized tree model was 6. As can be seen in Figure 24, the first node is split at FICO score which has the highest importance. CPI which has the second highest score can be seen at the second depth of the tree. According to Figure 25, the HPI and Unemployment rate are also ranked at 6th and 7th in feature importance. Although with a miniscule coefficient, Real GDP is also ranked at 11th position.

Figure 25 - *Feature Importance corresponding to Decision Tree in Figure 24 using Ensemble II*



Note: Own illustration.

Hence in a tree with a higher depth, more features are considered important at nodes for splitting, but majority of these features had a minor coefficient of importance. However, all the four macroeconomic variables were present in the feature importance ranking. In addition to the FICO score and interest rates, most of the features as depicted by the feature importance variables are intuitive in prediction loan defaults. For example, a higher debt to income ratio or a higher loan to value ratio does indicate more risky behavior than the loans having lower ratios. If the original loan amount i.e., the unpaid balance is higher, the burden on the consumer is higher which is riskier if coupled with a higher interest rate.

Likewise for the macroeconomic indicators, increase in inflation puts pressure on the prices of all other goods which could impact the ability to honor debt obligations. The same is the case if a person gets unemployed. These impact the debt-to-income ratio indirectly. Unemployment rate was also found to be important in other research of Sirignano, Sadhwani, and Giesecke (2018); Carvalho, Curto, and Primor (2022); Chen, Guo, and Zhao (2021). While the HPI has a positive impact as growing prices would increase the value of the houses and indirectly improve the loan to value ratios and vice versa. According to Chen et al., (2021), HPI was important to be included in their assessment of credit risk as well. It was rather surprising that the SATO was not considered by the algorithm to be important for predictions even though intuitively, higher the SATO, higher is the riskiness of a borrower.

As envisaged, the variables ‘Occupancy Status’, ‘First Time Home Buyer Indicator’, ‘Modification Flag’ and ‘Number of Units’ being more inclined towards one category in the dataset were not deemed to be important in predicting defaults. The model with undersampling or a model with macroeconomic data with a smaller depth returned lesser features to be important, however, inflation was still present as the second most important feature after FICO score and before original interest rate. This implies that incorporating macroeconomic features can help in relevant and better predictions in line with the existing literature. The feature importance figure for the undersampling techniques with both Ensemble I and Ensemble II are present in the Appendix II.

5. Conclusion

The consolidated results of the analysis are presented in this section for comprehensive comparison and discussion. Different decision tree models were tested based on different sampling techniques. Moreover, macroeconomic data was also used to assess its impact on loan default prediction. The different sampling techniques were not only evaluated on out of sample data, but also were used to predict on out-of-time data. The out-of-time data comprised of the year of pandemic i.e., 2020. While predicting default, the most important performance metric is recall which aims at accurately classifying all the positive classes, i.e., the defaults. The results derived from the models optimized according to the recall score are presented in sub-section 5.1 with a brief overview on the results. Furthermore, the limitations of the study are acknowledged followed by recommendations for future research and practical implications of the findings

5.1. Summary of Results

Table 10 displays the consolidated results based on optimized recall scores for prediction of loan defaults using various techniques with and without macroeconomic data as discussed in section 4.2.

The different simulations done using all the data from 2015 to 2020 through various resampling techniques exhibited that the no-resampling and undersampling techniques displayed similar results. The large size of the dataset particularly, out-of-sample and out-of-time affected the performance of the model with respect to minimizing the false positives. It is observed that in each of the sampling techniques used, the recall score enhances when the respective model is evaluated on out-of-sample data but comes at a cost of classifying false positives i.e., non-defaults as defaults. The undersampling technique displays similar results as a no-resampling technique when evaluated for Ensemble I data i.e., loan data without macroeconomic data. A decent recall score of approximately 81% is achieved by both these techniques upon testing out-of-sample. This implies that 81% of the loans that defaulted were classified correctly. On the contrary, SMOTE is the worst performer among the sampling strategies. Adding macroeconomic data does increase the recall score in all the sampling techniques when evaluated on out-of-sample data but it comes at a cost of decreased balanced accuracy in no-resampling and undersampling strategies. This indicates that even though the classifier can majorly correctly predict loan default i.e., true positives, the true negatives decrease with a simultaneous increase in the false positives.

Table 10 - *Summary of Results based on Decision Trees Optimized for Recall Scores*

	Method	No Resampling			Undersampling			SMOTE		
Metric	Data	In-Sample (2015-17)	Out-of-Sample (2018-19)	Out-of-time (2020)	In-Sample (2015-17)	Out-of-Sample (2018-19)	Out-of-time (2020)	In-Sample (2015-17)	Out-of-Sample (2018-19)	Out-of-time (2020)
Recall*	Ensemble I	0.746	0.802	0.426	0.752	0.807	0.428	0.561	0.582	0.288
Specificity	Ensemble I	0.746	0.586	0.701	0.742	0.572	0.692	0.804	0.641	0.727
Balanced Accuracy	Ensemble I	0.746	0.694	0.563	0.747	0.690	0.560	0.682	0.612	0.508
Recall*	Ensemble II	0.815	0.916	0.522	0.764	0.816	0.424	0.618	0.783	0.342
Specificity	Ensemble II	0.758	0.388	0.606	0.766	0.527	0.724	0.789	0.486	0.628
Balanced Accuracy	Ensemble II	0.787	0.652	0.564	0.765	0.672	0.574	0.704	0.634	0.485

Note: (*) denotes that the recall score was optimized during fine tuning of the model. The highest out of sample scores within a row for one type of data across the 3 different sampling schemes is highlighted in blue, while the highest out-of-time score is highlighted in yellow. The highest recall scores between the two data types for the respective sampling technique represented by a grid is highlighted in green

Table 11- *Summary of Results based on Decision Trees Optimized for Balanced Accuracy Scores*

	Method	No Resampling			Undersampling			SMOTE		
Metric	Data	In-Sample (2015-17)	Out-of-Sample (2018-19)	Out-of-time (2020)	In-Sample (2015-17)	Out-of-Sample (2018-19)	Out-of-time (2020)	In-Sample (2015-17)	Out-of-Sample (2018-19)	Out-of-time (2020)
Recall	Ensemble I	0.809	0.871	0.477	0.785	0.798	0.442	0.561	0.582	0.288
Specificity	Ensemble I	0.717	0.530	0.649	0.744	0.598	0.694	0.804	0.641	0.727
Balanced Accuracy*	Ensemble I	0.763	0.701	0.563	0.764	0.698	0.568	0.682	0.612	0.508
Recall	Ensemble II	0.815	0.916	0.522	0.758	0.834	0.420	0.617	0.780	0.342
Specificity	Ensemble II	0.758	0.388	0.606	0.840	0.483	0.701	0.790	0.487	0.628
Balanced Accuracy*	Ensemble II	0.787	0.652	0.564	0.799	0.659	0.561	0.703	0.634	0.485

Note: (*) denotes that the recall score was optimized during fine tuning of the model. The highest out of sample scores within a row for one type of data across the 3 different sampling schemes is highlighted in blue, while the highest out-of-time score is highlighted in yellow. The highest recall scores between the two data types for the respective sampling technique represented by a grid is highlighted in green

The scores optimized based on balanced accuracy score are displayed in Table 11. Although, yielding better recall scores when optimized by balanced accuracy score, it came at a cost of classifying more false positives and were prone to overfitting returning longer depth of the tree while tuning the hyperparameter.

To the extent that the impact of COVID-19 pandemic is envisaged, all the models with different sampling techniques, whether incorporating macroeconomic data or not, did not perform well when evaluated on out-of-time pandemic period. Although, generalizing well on the data from years 2018 and 2019, these data- driven pattern perception methods tend to fail when the past patterns in data are not extrapolative to future behavior, as did the changing distributions within the data after the lapse of 2019 and beginning 2020 induced by the pandemic. Testing the model on out-of-time unseen data concerning the pandemic period demonstrated the adverse impact of pandemic on the predictive performance of the model answering the first research question positively that, pandemic did impact the performance of the machine learning model affecting the ability to correctly predict mortgage loan default.

Including macroeconomic indicators, particularly the CPI, did increase the recall scores while testing out-of-sample and out-of-time in comparison to data without macroeconomic indicators. Although very small, the contribution of Unemployment rate, HPI and real GDP cannot be ruled out while the model was tested with no resampling technique. This benefit of including the macroeconomic factors falls in line with the related literature. This answers the second research question, *could incorporating macroeconomic indicators into machine learning methods improve the performance of credit risk models?* as yes. However, this answer comes with a reservation that the non-defaulters were not predicted as accurately as the loans that actually defaulted were predicted. The balance between the true negatives and true positives was slightly lacking.

Other findings suggest that handling a large dataset with a huge imbalance in the classes was a difficult task which affected the ability of decision tree to perform as accurately as it could have been with other machine learning algorithms.

5.2. Acknowledgement of Limitations

The thesis aimed to demonstrate the impact of pandemic on the predictive performance of widely used decision tree models. The selection of model was limited due to computational resources required to process large datasets and its ability to provide interpretable results. However, due to limited ability of decision tree to handle big data, the evaluation of pandemic's impact and the macroeconomic factors can be further performed using more advanced and robust ensemble or deep learning methods. Additionally, the macroeconomic variables used for the predictions were subject to availability on a regular basis to better integrate into the loan performance data. More macroeconomic variables, capturing the behavior of the consumer can be obtained to integrate into the loan performance data for better prediction of mortgage loan defaults. The focus on predicting loan defaults correctly on a huge dataset comes at a cost of predicting non-defaulters as defaulters.

5.3. Recommendations for Further Research

As acknowledged in the limitations, more stable models could be built with ample computational resources whilst being able to provide better explanations on drivers of mortgage default. The duration of periods tested are long and may be further evaluated with shorter time windows. More refined loan performance data with relevant borrower-specific features and macroeconomic data addressing changes in consumer behavior can be used to improve default predictions. The evaluations can also be extended to other countries or different products such as credit cards, vehicle loans, agricultural loans which could provide validity to the methods externally.

5.4. Implications of Findings

The conclusions of the thesis point towards the inability of machine learning model such as the decision tree to embrace the impact of a crisis such as the pandemic. However, the addition of macroeconomic factors does lead to better default prediction than just using traditional loan performance data. Hence, supplementing historical data with relevant additional data can help in better credit risk assessment especially for enhancing loan prediction models. Further, cost-benefit analysis can be conducted based on the classification performance when predicting a loan defaulter correctly is more important than predicting non-defaulters as defaulters. The macroeconomic data inclusion can also be extended beyond credit scoring models to other credit risk assessment methods.

6. List of References

Alonso, A., & Carbó, J. M. (2021). *Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation* (SSRN Scholarly Paper No. 3774075). Social Science Research Network. https://doi.org/10.2139/ssrn.3774075

Aniceto, M. C., Barboza, F., & Kimura, H. (2020). Machine learning predictivity applied to consumer creditworthiness. *Future Business Journal*, *6*(1), 37. https://doi.org/10.1186/s43093-020-00041-w

Barbaglia, L., Manzan, S., & Tosetti, E. (2021). Forecasting Loan Default in Europe with Machine Learning*. *Journal of Financial Econometrics*, nbab010. https://doi.org/10.1093/jjfinec/nbab010

BIS. (2000). *Principles for the Management of Credit Risk*. https://www.bis.org/publ/bcbs75.htm

BIS. (2022). *Newsletter on Covid-19 related credit risk issues*. https://www.bis.org/publ/bcbs_nl26.htm

BIS, BCBS. (2002). *QIS 3 FAQ: F. Definition of Default/Loss*. https://www.bis.org/bcbs/qis/qis3qa_f.htm

Bouaguel, W., AlSulimani, T., & Alarfaj, O. (n.d.). *The Impact of the COVID-19 Pandemic on the Saudi Credit Industry: An Empirical Analysis Using Machine Learning Techniques to Focus on the Factors Affecting Consumer Credit Scoring*. 38.

Breeden, J. (2021). A survey of machine learning in credit risk. *The Journal of Credit Risk*. https://doi.org/10.21314/JCR.2021.008

Carvalho, P. V., Curto, J. D., & Primor, R. (2022). Macroeconomic determinants of credit risk: Evidence from the Eurozone. *International Journal of Finance & Economics*, *27*(2), 2054–2072. https://doi.org/10.1002/ijfe.2259

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: A recent review. *Artificial Intelligence Review*, *45*(1), 1–23. https://doi.org/10.1007/s10462-015-9434-x

Chen, S., Guo, Z., & Zhao, X. (2021). Predicting mortgage early delinquency with machine learning methods. *European Journal of Operational Research*, *290*(1), 358–372. https://doi.org/10.1016/j.ejor.2020.07.058

Christopher M. Bishop. (2009). *Pattern recognition and machine learning*. New York : Springer.

Corazza, M., De March, D., & di Tollo, G. (2021). Design of adaptive Elman networks for credit risk assessment. *Quantitative Finance*, *21*(2), 323–340. https://doi.org/10.1080/14697688.2020.1778175

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

Fannie Mae®. (n.d.). *Fannie Mae Single-Family Loan Performance Data | Fannie Mae*. Retrieved October 24, 2022, from https://capitalmarkets.fanniemae.com/credit-risk-transfer/single-family-credit-risk-transfer/fannie-mae-single-family-loan-performance-data

Fazlija, B. (2022). *Class material from Machine Learning and Deep Learning course: Fall semester 2022/23, Winterthur: School of Management and Law, Zurich University of Applied Sciences*.

FHFA. (n.d.). *Housing Price Index Frequently Asked Questions*. Retrieved September 21, 2022, from https://www.fhfa.gov/Media/PublicAffairs/Pages/House-Price-Index-Frequently-Asked-Questions.aspx

Freddie Mac. (1971, April 2). *30-Year Fixed Rate Mortgage Average in the United States*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/MORTGAGE30US

Ghosh, A. (2017). Sector-specific analysis of non-performing loans in the US banking system and their macroeconomic impact. *Journal of Economics and Business*, *93*, 29–45. https://doi.org/10.1016/j.jeconbus.2017.06.002

Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C. (2018). Ensemble Learning or Deep Learning? Application to Default Risk Analysis. *Journal of Risk and Financial Management*, *11*(1), Article 1. https://doi.org/10.3390/jrfm11010012

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York : Springer.

Krugman, P. R., Obstfeld, M., & Melitz, M. J. (2018). *International Economics: Theory & Policy* (Eleventh, Global edition.-[2018]). Pearson Education.

Laborda, J., & Ryoo, S. (2021). Feature Selection in a Credit Scoring Model. *Mathematics*, *9*(7), Article 7. https://doi.org/10.3390/math9070746

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine Learning in Banking Risk Management: A Literature Review. *Risks*, *7*(1), Article 1. https://doi.org/10.3390/risks7010029

Liu, Y., Yang, M., Wang, Y., Li, Y., Xiong, T., & Li, A. (2022). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China. *International Review of Financial Analysis*, *79*, 101971. https://doi.org/10.1016/j.irfa.2021.101971

Mamonov, S., & Benbunan-Fich, R. (2017). What Can We Learn from Past Mistakes? Lessons from Data Mining the Fannie Mae Mortgage Portfolio. *Journal of Real Estate Research*, *39*(2), 235–262. https://doi.org/10.1080/10835547.2017.12091471

Markov, A., Seleznyova, Z., & Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science*, *8*, 180–201. https://doi.org/10.1016/j.jfds.2022.07.002

Matplotlib. (n.d.). *Matplotlib—Visualization with Python*. Retrieved November 16, 2022, from https://matplotlib.org/

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, *165*, 113986. https://doi.org/10.1016/j.eswa.2020.113986

Mueller, J. P., & Massaron, L. (2021). *Machine Learning for Dummies* (2nd ed.). John Wiley & Sons, Inc.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python, A guide for Data Scientists*. O'Reilly Media, Inc.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, Massachusetts : The MIT Press.

Naili, M., & Lahrichi, Y. (2022). The determinants of banks' credit risk: Review of the literature and future research agenda. *International Journal of Finance & Economics*, *27*(1), 334–360. https://doi.org/10.1002/ijfe.2156

Narvekar, A., Guha, D., Narvekar, A., & Guha, D. (2021). Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession. *Data Science in Finance and Economics*, *1*(2), Article DSFE-01-02-010. https://doi.org/10.3934/DSFE.2021010

Nehrebecka, N. (2021). Internal Credit Risk Models and Digital Transformation: What to Prepare for? An Application to Poland. *European Research Studies*, *XXIV*(Special 3), 719–736.

NumPy. (n.d.). *NumPy*. Retrieved November 16, 2022, from https://numpy.org/

Pandas Developers. (n.d.). *pandas—Python Data Analysis Library*. Retrieved November 16, 2022, from https://pandas.pydata.org/about/

Python. (n.d.). *Applications for Python*. Python.Org. Retrieved November 16, 2022, from https://www.python.org/about/apps/

Saygili, E., Saygili, A. T., & Isik, G. (2019). An Analysis of Factors Affecting Credit Scoring Performance in SMEs. *Ege Academic Review*, *19*(2), Article 2. https://doi.org/10.21121/eab.526093

Scikit-learn. (n.d.). *3.3. Metrics and scoring: Quantifying the quality of predictions*. Scikit-Learn. Retrieved November 18, 2022, from https://scikit-learn/stable/modules/model_evaluation.html

Seaborn Pydata. (n.d.). *An introduction to seaborn—Seaborn 0.12.1 documentation*. Retrieved November 16, 2022, from https://seaborn.pydata.org/tutorial/introduction.html

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. *Neural Computing and Applications*, *34*(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Sirignano, J., Sadhwani, A., & Giesecke, K. (2018). *Deep Learning for Mortgage Risk* (SSRN Scholarly Paper No. 2799443). Social Science Research Network. https://doi.org/10.2139/ssrn.2799443

Tang, L., Thomas, L. C., Thomas, S., & Bozzetto, J. (2007). It's the economy stupid: Modelling financial product purchases. *International Journal of Bank Marketing*, *25*(1), 22–38. https://doi.org/10.1108/02652320710722597

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Turjo, A. A., Rahman, Y., Karim, S. M. M., Biswas, T. H., Dewan, I., & Hossain, M. I. (2021). CRAM: A Credit Risk Assessment Model by Analyzing Different Machine

Learning Algorithms. *2021 4th International Conference on Information and Communications Technology (ICOIACT)*, 125–130. https://doi.org/10.1109/ICOIACT53268.2021.9563995

U.S. Bureau of Economic Analysis. (2005, January 1). *Real Gross Domestic Product: All Industry Total in California*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/CARQGSP

U.S. Bureau of Labor Statistics. (n.d.-a). *Consumer Price Index, Regional Resources*. Retrieved October 25, 2022, from https://www.bls.gov/cpi/regional-resources.htm

U.S. Bureau of Labor Statistics. (n.d.-b). *Labor Force Characteristics (CPS)*. Retrieved September 20, 2022, from https://www.bls.gov/cps/lfcharacteristics.htm#unemp

U.S. Bureau of Labor Statistics. (1976, January 1). *Unemployment Rate in the District of Columbia*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DCUR

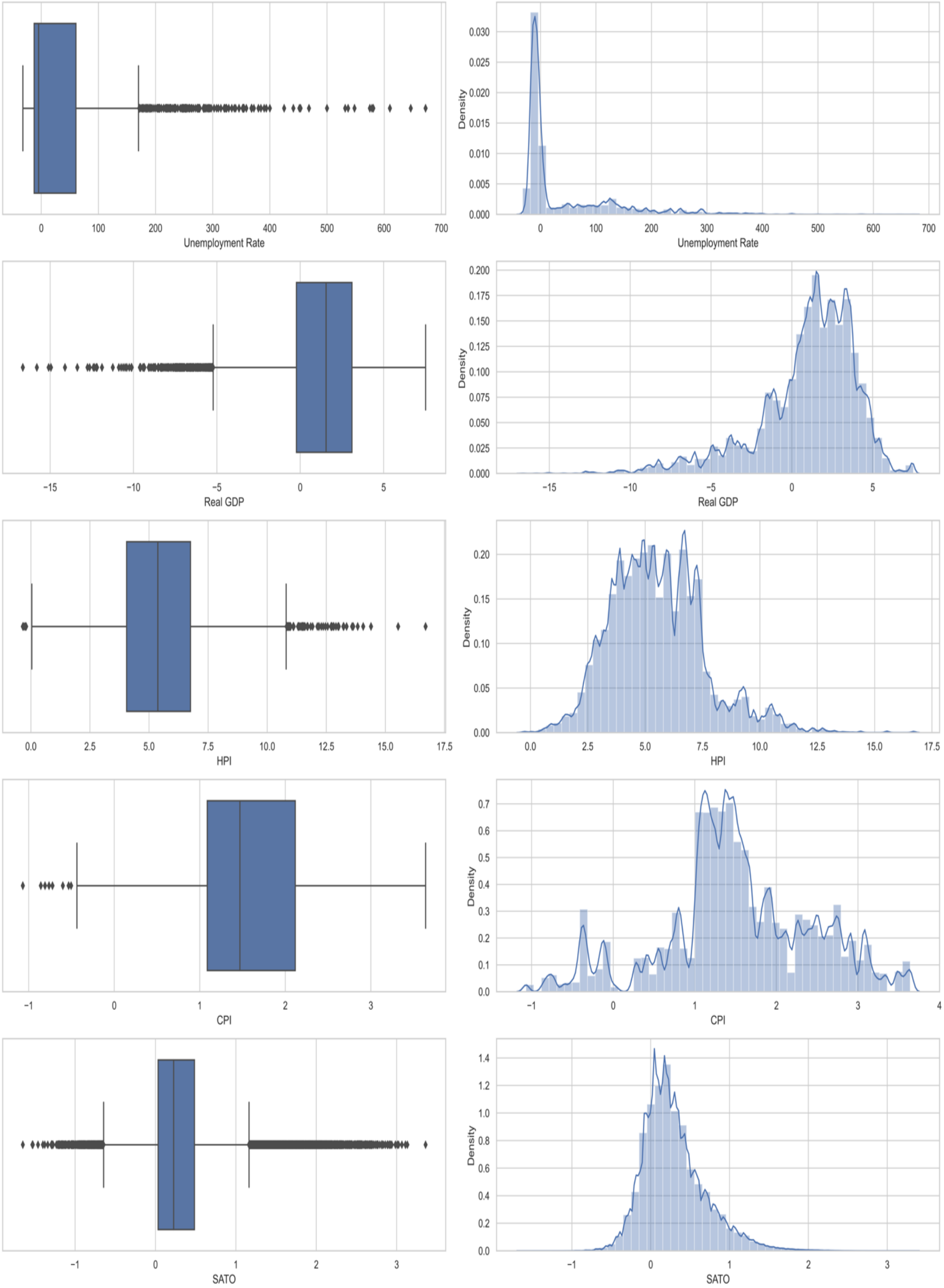
U.S. Federal Housing Finance Agency. (1975, January 1). *All-Transactions House Price Index for the District of Columbia*. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DCSTHPI

Xia, Y., Li, Y., He, L., Xu, Y., & Meng, Y. (2021). Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending. *Electronic Commerce Research and Applications*, *49*, 101095. https://doi.org/10.1016/j.elerap.2021.101095

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: A machine learning methodology. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-98361-6

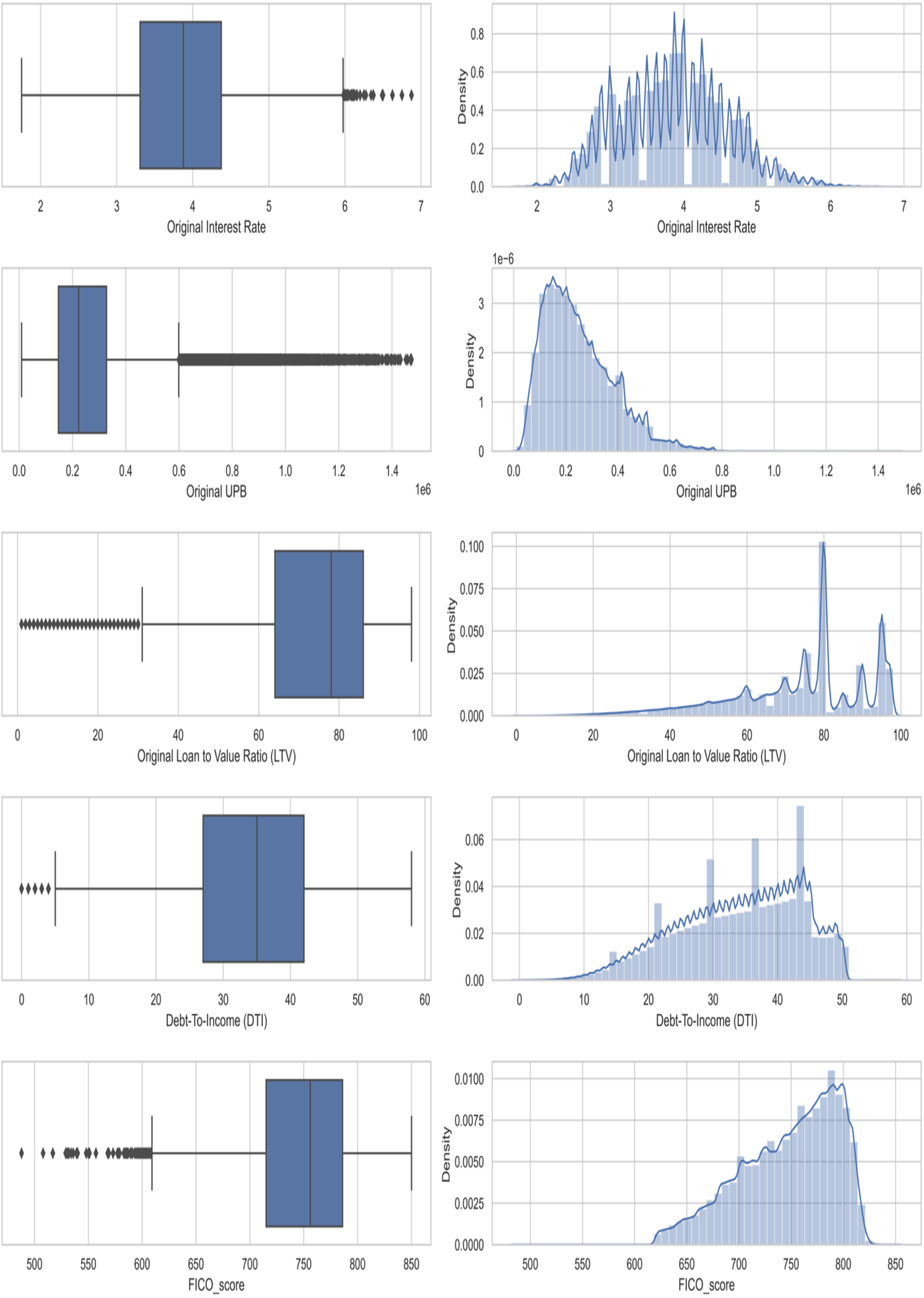
Appendix I: Graphs as a Part of Exploratory Data Analysis

Figure A 1 - *Distribution and Box Plot for Macroeconomic Variables*



Note: Own illustration.

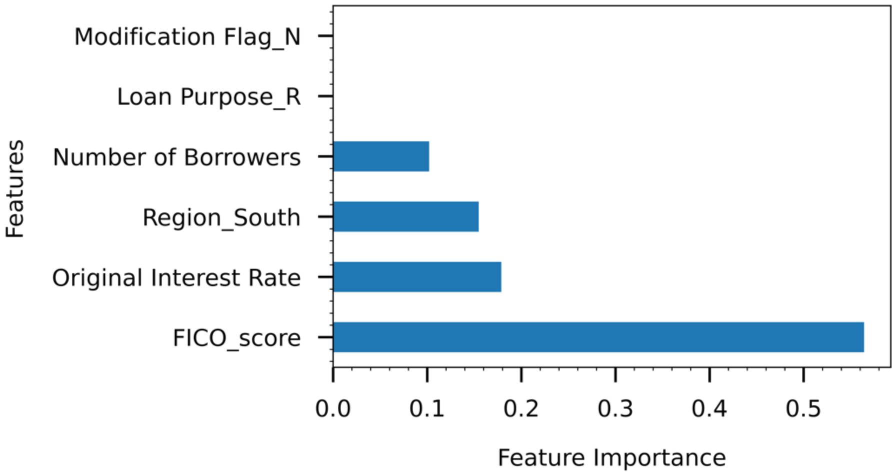
Figure A 2- *Distribution and Box Plots for Continuous Variables*



Note: Own illustration.

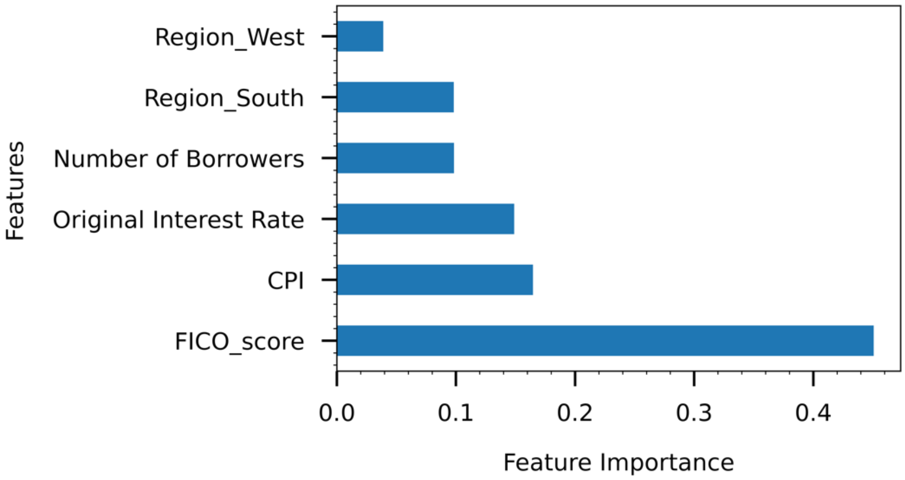
Appendix II: Feature Importance for Undersampling Techniques for Section 4.3

Figure A 3 - *Feature Importance for Undersampling Technique on Ensemble I*



Note: Own illustration.

Figure A 4- *Feature Importance for Undersampling Technique on Ensemble II*



Note: Own illustration.