Final Year Project

Credit Risk Assessment using Selected Machine Learning Algorithms

Debanjan Das

Student ID: 1702833

A thesis submitted in part fulfilment of the degree of

BSc. (Hons.) in Computer Science

Supervisor: Professor Mark Keane

UCD School of Computer Science

University College Dublin

May 25, 2021

Abstract

It is important for a bank to assess its credit risk and the extent of its exposure in the event of non-performing customers. Statistical approaches have been used to estimate this type of risk for decades, and with recent advances in the field of machine learning, there has been an interest in seeing whether machine learning algorithms can achieve better risk quantification. The aim of this study is to see which approach from a collection of machine learning techniques performs the best in default prediction when model evaluation parameters are chosen. Logistic Regression, Random Forest, XGBoost, and Artificial Neural Network were the techniques studied. To address the imbalance between classes for the response variable, an oversampling technique known as SMOTE was used. In order to enhance model accuracy, data was divided into clusters, followed by the deployment of Machine Learning models on each of these clusters. In terms of the chosen model evaluation metric, the results showed that XGBoost with hyperparameter tuning achieved the best result. An attempt has also been made to simulate data in a pandemic environment and the built-in model has been adjusted accordingly to assess credit risk.

Acknowledgement

I would like to take this opportunity to thank both my supervisor, Professor Mark Keane and my assigned mentor, Dr. Aonghus Lawlor of the University College Dublin (UCD) Insight centre for Data Analytics for their support and assistance in completing my project work.

Details

- Gitlab Credit Risk Assessment Using Selected Machine Learning Algorithms
- Website https://debanjandas-loanpredictor-rtbl5.ondigitalocean.app/

Table of Contents

Chapter 1: Project Specification

Credit Risk Assessment using Selected Machine Learning Algorithms

People

Academic Supervisor: Prof. Mark Keane

Project Mentor: Dr Aonghus Lawlor

Project Specification

Subject: Credit Risk Assessment in Banks/ Financial Institutions

Type: Machine Learning and Applied Research

Software Requirements: Python, Jupyter Notebook

Project Description

For financial institutions, credit risk assessment [1] is important as it directly affects business outcomes. While artificial intelligence (AI) and machine learning are not recent, in their credit risk assessment phase, microcredit organizations are shy about accepting these approaches and still use standard credit scoring methods based on the linear calculation of a small number of indicators.

Based on linear estimates of a limited number of indicators, traditional credit rating methods are used. Mixed and inconsistent results are provided by this scoring model. On the other hand, machine learning provides a much wider view of a client and can be used not only to handle credit risk but also to manage other business risks.

To find the best suited for the lending sector, several algorithms have been used. BigML's OptiML has also been used to peek into a few of the best-generated models with a higher accuracy rate.

The dataset used in this project is German Credit Data [2] which is the most influential dataset that has been used in the field of Credit Risk assessment. There are 1000 rows with 20 predictor variables (quantitative, ordinal and nominal variables) and one binary response variable. Information on data generation or context is missing, as was also recently criticized by an anonymous blogger on Reddit (Anonymous 2019). A small python script has been written to convert it to a readable CSV file.

In general, it is very difficult to obtain datasets on the credit scoring scenario since there are issues related to the maintenance of confidentiality of credit scoring databases.

Core Requirements

• In this project, our main objective is to develop a machine learning model to help predict if a bank would grant a loan to a client. Various machine learning algorithms like Neural Networks, Logistic Regression, Boosted Decision Trees and Random Forest would be used to correctly classify the outcome of a loan. The motivation behind using these algorithms can be understood in the 'Related Work' section.

• We also aim to identify the features that play a crucial role in determining the result of the loan process. Different feature selection methods (like Information Gain) would be implemented to find the most relevant features. A model with a certain number of features leads to better accuracy. Further Data Analysis would be performed to explore exciting hidden patterns that might exist.

• Another goal is to conduct customer segmentation [3] using suitable clustering methods. Customer segmentation is the process of dividing customers into groups based on common

characteristics so companies can market to each group effectively and appropriately. As a part of this project, we aim to find out which cluster of people tend to get a loan. A financial institution might segment customers according to a wide range of factors, which can be used for filtered marketing purposes.

Advanced Requirements

• The possibility of a global pandemic/market crash would also be taken into consideration. A new dataset would be simulated for the same people that would look like one affected by a pandemic/market crash. A credit score would be calculated for each customer. Following that, both the datasets would be compared and the model would be able to identify the pandemic affected people automatically based on the credit score. The most affected columns would tend to be Job, Income, etcetera. The built-in model would then be adjusted to predict for the pandemic affected people – where factors like deflation rate (last year vs current year) and decrease in salary might be taken into consideration to measure the risk of lending a loan.

• Finally, a User Interface would be designed where the user can fill in the attributes required for a loan and check eligibility.

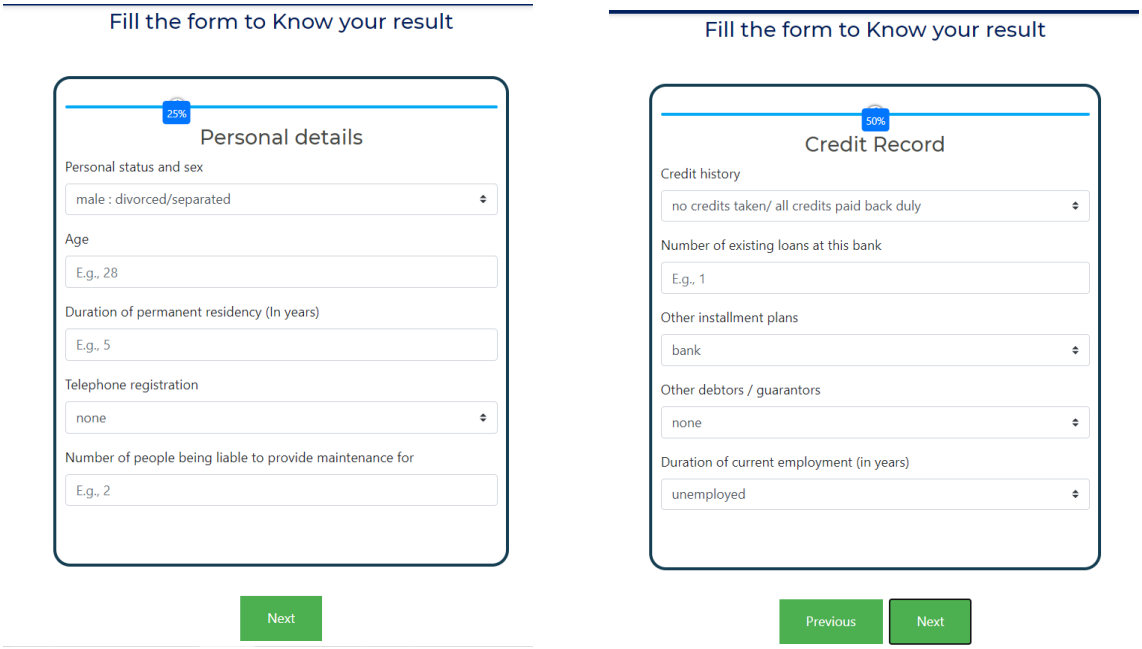

Figure 1.1: A Possible User Interface for Credit Risk assessment

Chapter 2: **Introduction**

The recent emergence of machine learning and data mining techniques has sparked interest in using these techniques in a variety of fields. The banking sector is no exception, and the growing pressure on financial institutions to provide robust risk management has spurred interest in improving current risk estimation methods. Machine learning techniques may potentially lead to a better quantification of the financial risks that banks are exposed to.

In this project, I will focus on assessing credit risk for individuals using various machine learning algorithms. The dataset used in this project is German Credit Data which is a standard imbalanced machine learning dataset. The dataset was used as part of the Statlog project, a European-based initiative in the 1990s to evaluate and compare a large number (at the time) of machine learning algorithms on a range of different classification tasks. The dataset is credited to Hans Hofmann.[2]

The data has been thoroughly analyzed and it would be discussed in details in the upcoming sections of this report. Following the data analysis, various Machine Learning models were deployed on the dataset to predict if a customer would be able to pay off a loan or not. The model was implemented on both the balanced (Using SMOTE) [4] and the original (imbalanced) version of the dataset. After the implementation of classification models following Baesens et. al's work [5], a new approach was taken into consideration in strive for increasing model accuracy levels. Inspired by Yiyun Liang et. al.'s paper [6], the dataset was divided into small clusters, and then the existing classification models were deployed again on each of these clusters.

Classification Method	Before Clustering	After Clustering
XGBoost	0.79	0.81
Logistic Regression	0.79	0.79
Neural Networks	0.50	0.58
Random Forest	0.79	0.81

Figure 2.1: Improvement in Mean ROC Curve after performing Clustering

A sharp increase was noticed in almost all the models as seen in Fig 4.1. XGBoost and Random Forest seemed to be the two most promising models showing high levels of Mean ROC Curve.

We have also taken into consideration, the possibility of a global pandemic or a market crash. A new dataset is simulated where few columns related to savings account balance, checking account balance, skills and employability have been tweaked according to a pandemic or a market crash scenario. The built-in model is then adjusted to predict for the pandemic affected people and it has been noticed that the number of loans getting approved have dropped considerably when compared to the number of loans getting approved before the pandemic.

Last but not the least, an attempt has been made to design an User Interface where users would be able to give in their details in a form and then, on clicking predict one can get to know if their loan would be approved or rejected.

It should be noted that the results might not be very accurate and it does not give any guarantee that a bank would make the same decision since this result is based on a data set which is very limited and also, fairly old (but highly credible).

Chapter 3: Related Work and Ideas

This section focusses on the research that have been conducted in the field of Bank Loans and Machine Learning Algorithms. There have been several studies which involve the analysis of loans in banks and other financial institutions, which have been discussed below. First, we would discuss the effectiveness of various classification methods in credit risk assessment and we will follow that up with the performance enhancement in a model with clustering methods. In the final summary section, the most relevant work to the current project would be mentioned explaining what it does and how my work will be advanced on it.

3.1 Effectiveness of Various Classification Methods in Credit Risk Assessment

3.1.1 Recent Developments

For financial institutions, credit risk management is important as it directly affects business outcomes. Although artificial intelligence (AI) and machine learning are not new, in the credit risk assessment phase, micro-credit organizations are shy about accepting these approaches and still use traditional credit scoring methods based on linear measurement of a limited number of indicators. Mixed and inconsistent results are provided by this scoring model. On the other hand, machine learning provides a much wider view of a client and can be used not only to handle credit risk but also to manage other business risks. Predictions made using machine learning are highly accurate, but it should be made sure that the data quality is good. The accuracy of models differs from one another. Hence, we will discuss various models and their performances in the field of credit risk assessment.

Baesens et. al.'s work [5] is essentially a modern update to the milestone benchmarking analysis of classification algorithms for credit scoring. In predictive modelling, several new techniques and algorithms have been developed since 2003. This paper discusses all the newer state-of-the-art techniques as well as those covered in the earlier study. This research paper has been chosen to be discussed as its purpose of comparing credit scoring classification algorithms is closely related to the problem that this project aims to solve. In this study, a vector X of M dimensions with each dimension is defined as an attribute characterizing an application for a credit product such as a loan. The dataset used in this project is in the same format, where an individual loan is represented by each row (vector). The study then goes on to address a binary response variable which demonstrates the existence or non-existence of a default event. In the research, the probability of a default event given X is the classification issue being discussed. Finally, a decision-maker will take this probability and the application will be approved if it falls below a given threshold, otherwise, it will be refused.

In this study, the efficiency of each algorithm is being tested in the classification of credit scoring using ROC Curve Area Under (AUC) [7]. According to the report, the top three most reliable classifiers are Random Forests, Bagged (MLP) Neural Networks and Bagged Decision Trees across all output measures. Building on Baesens et. al.'s work, there was another study [8],conducted by Rich Caruana, that compares the performance of eight machine learning algorithms. The study concludes that the best average output (before calibration) for all metrics and all problems is obtained by bagged trees, random forests and neural nets. The overall best performing algorithm is boosted decision trees when calibration is taken into

account. Random forest is close to second, followed by bagged decision trees (uncalibrated).

The classifiers discussed below have been chosen for use in my project. We will talk about these classifiers in detail and explore the factors that make them better than other classifiers.

3.1.2 Use of Random Forest Classification Algorithm in Credit Risk Assessment

Random forest is a supervised learning algorithm, an ensemble method for classification, regression and other tasks that work by constructing a variety of decision trees at the time of training and outputting the class that is the mode of the classes or the prediction of the individual trees.

The decision tree is in the form of a tree (which can be a binary tree or a non-binary tree). Each of its non-leaf nodes corresponds to a feature test, with each branch representing the feature attribute output over a range of values, and with each leaf node storing a category. The decision tree begins with the root node, checks the corresponding attributes of the function in the category to be categorized, and selects the output branches according to their values until the leaf node is reached. Eventually, the decision result is considered to be the category stored by the leaf node.

A random forest is a group of such decision trees in which there is no relationship among each of the decision trees. The Gini index is the selection metric that can be used to separate attributes in the decision tree, and the number of levels in each tree branch depends on the algorithm parameter d (depth).

They exploit the significant strengths of decision trees, including the handling of non-linear relationships, the robustness of noisy information and outliers, and the determination of predictor relevance. However, they do not need to be pruned, are less susceptible to overfitting, and yield aggregate results that appear to be more reliable, unlike single decision trees.

Random Forest has the following benefits, which is why we prefer the Random Forest algorithm over other machine learning algorithms:

• It operates efficiently on broad databases of data and among current algorithms, it is unexcelled in accuracy;

• It can fix well with errors in class population unbalanced data sets.

• It supports an efficient method of estimating missing data, which can preserve accuracy even though the data is missing on a wide scale.

Lin Zhu et. al.'s paper [9] shows the effectiveness of the Random Forest algorithm in credit risk assessment. The results show that the performance of random forest and decision tree is better than that of support vector machine and logistic regression. The random forest performs the best, with an accuracy of 98 percent, higher than the decision tree with an accuracy of 95 percent. The precision and recall of the prediction model based on the random forest are all above 0.95, indicating that the model has strong ability of generalization.

3.1.3 Use of Bagged Neural Networks (MLP) in Credit Risk Assessment

In R. Alejo et. al.'s paper [10], we study the popular Multi-layer Perceptron Neural Network using three misclassification cost functions. Before getting into depth, let's get to understand what MLP is.

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output

layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

The back-propagation algorithm, which uses a collection of training instances to modify the free parameters U, is the most common learning procedure for the MLP neural network. Several works have shown that during the training process, where the major contribution to the MSE (Mean Square Error) is made by the majority class, the class imbalance problem generates unequal contributions to the mean square error (MSE). In R. Alejo et. al.'s paper, three different cost functions were defined whose aim is to compensate for the unequal contribution of the MSE during the training phase. Four different MLP models were trained (one with each cost function and one without any.) It was discovered that MLP classifiers usually perform better than the original MLP (without any cost function). It can be noticed by either analysing the global AUC or comparing the AUC of each algorithm over each data set.

Each layer in MLP contains a given number of nodes with the activation function and nodes in neighbour layers are linked by weights. The optimal weights are obtained by optimizing objective or loss function using a backpropagation algorithm to build a model as defined:

$$\underset{\omega}{\operatorname{argmin}} \frac{1}{T} \sum_{t} l(f(\omega x + b); y) + \lambda \Omega(\omega)$$

where ω denotes the vector of weights, x is the vector of inputs, b is the bias and f($*$) is the activation function and $\lambda\Omega(\omega)$ is a regularizer. Several parameters need to be determined in advance for the training model, such as number of hidden layers, number of their nodes, learning rate, batch size and epoch number.

Some of the benchmark models require prior parameter initialization. For NN, with one hidden layer, a multilayer perceptron (MLP) model is constructed. A small learning rate of 0.01 is set in line with Ala'raj and Abbod (2016b) [11]. The maximum of epochs set by default is 1000. For the German dataset, the number of hidden neurons is tuned by GS based on the 10-fold cross-validation ACC and it is found out to be 5.

R. Alejo et. al.'s work showed that in general, the three cost-sensitive MLP classifiers usually perform better than the original MLP (without a cost function), which can be seen by either analysing the global AUC or comparing the AUC of each algorithm.

The selection of the optimization algorithm in a neural network has a major influence on the dynamics of training and task performance. There are several methods to enhance the gradient descent optimization and Adam [12] is one of the strongest optimizers [mentioned in Appendix]. From estimates of first and second gradient moments, Adam calculates adaptive learning rates for various parameters and understands the advantages of both the Adaptive Gradient Algorithm and Root Mean Square Propagation. In the field of deep learning, Adam is therefore considered one of the best gradient descent optimization algorithms.

3.1.4 Use of Logistic Regression in Credit Risk Assessment

The Logistic Regression approach has been considered a benchmark in the issue of credit scoring. Logistic Regression calculates the conditional risk of default of the creditor and describes the relationship between the creditworthiness of clients and explanatory variables. The method for Logistic Regression to create a model consists of estimating a linear combination of interpreter X and binary dependent variable Y and labelling using the logistic function to translate log-odds to probability. The LR formula is as:

To estimate regression coefficients, the maximum likelihood estimate is commonly used. We have an interpreter of x and a binary dependent variable of y for each data point. If y =

$$Y \approx P(X) = 1 / 1 + e^{-(\beta 0 + \beta X)}$$

1, the probability of dependent variable is p(x), or $1 - p(x)$, if y = 0. Then probability is written as:

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

The probability of occurrence of a categorical output can also be found by logistic regression model by fitting the features in the logistic curve as shown in the figures [5] below.

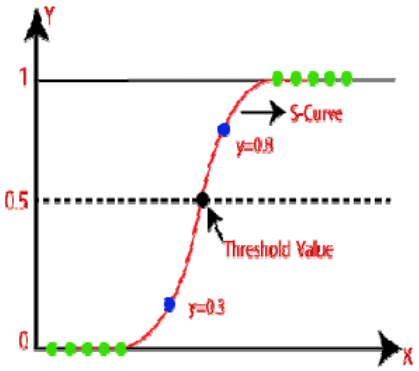

Figure 3.1: Logistic function [13]

In Logistic regression, rather than fitting a regression line, we fit an "S" shaped logistic function, which predicts two greatest values (0 or 1). It is a significant algorithm because it can provide probabilities and classify the use of different types of data and easily determines the most effective variables that are used for classification.

The sigmoid function is a numerical function used to outline predicted values to probabilities. It maps any real value to another value that is between 0 and 1. The S-structure curve is also known as the sigmoid function or the logistic function. In logistic regression, we utilize the concept of a threshold value, which characterizes the probability of either 0 or 1 and the value below the threshold values tends to 0.

In Mong'are et. Al's paper [14], a study was done on the analysis of individual loan defaults using Logistic Regression and this approach was found to have an accuracy of 0.7727 with the train data and 0.7333 with the test data. The model also showed a precision of 0.8440 and 0.8244 with the train and test data respectively. The data used in this study was obtained from the Equity Bank of Kenya for the period between 2006 to 2016. A random sample of 1000 loan applicants whose loans had been approved by equity bank of Kenya during this period was obtained.

3.1.5 Use of Boosted Decision Trees in Credit Risk Assessment

An improved variant of the simple decision tree classification algorithm is the Boosted Decision Tree. It is accomplished using boosting through an ensemble of decision trees. Boosting implies that, while using more than one decision tree as the classification algorithm during the training phase, each tree is dependent on the prior trees.

Boosted Decision Tree is a supervised learning technique. Its dataset must be labelled with columns containing numerical values. By fitting the residual of the trees preceding it,

the algorithm learns better. Therefore, by optimizing tree with arbitrary differentiable loss function, the boosted decision tree ensemble also increases the accuracy of the algorithm with less threat of coverage.

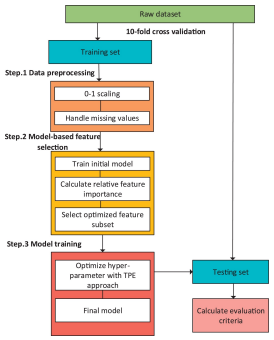

Figure 3.2: Flowchart of XGBoost-based credit scoring model

In a recent work of Xia et. al. [15], we have seen the use of XGBoost algorithm with Bayesian hyper-parameter optimization method to construct a credit scoring model. They achieved the classification performances compared to other machine-learning methods on the different benchmark credit scoring datasets. XGBoost is a boosting ensemble algorithm which optimizes the objective of function, size of the tree and the magnitude of the weights are controlled by standard regularization parameters. This model was proved to outperform several baseline techniques and was validated on five datasets over five performance metrics.

Xia et. al. explored the sequential methods for establishing credit scoring models. XGBoost was introduced as a novel variant of boosting technique, which adds a regularization term in the loss function and makes some engineering modifications based on GBDT (Gradient Boosting Decision Tree). However, XGBoost includes various hyper-parameters and tends to fail without careful tuning. To prevent this situation, a TPE, which is a Bayesian hyper-parameter optimization method, is employed to tune these hyper-parameters. His paper revealed that TPE outperforms RS, GS, and MS, even though the latter two are commonly used measures in parameter-tuning. The comparisons with baseline models showed the superiority of the XGBoost-based model in terms of predictive performance.

3.2 Effectiveness of clustering along with classification methods:

Following up research papers discussed above related to classification, we would now touch a little on various clustering methods that would help in improving the overall accuracy of a model. Yiyun Liang et. al.'s paper [6] talks about the inefficiency of a generic machine learning model that might not be appropriate to evaluate someone's repayment ability, such as students or people without credit histories. In such scenarios, we can potentially categorize the various types of loan applicants in our dataset and build different prediction models for different groups based on their distinct characteristics.

In this paper, we would first discuss the different models that were implemented for credit risk assessment and we would follow that up by the performance of K-means clustering along with classification.

In Liang et, Al.'s study, the dataset used was imbalanced due to which both up-sampling and down-sampling were performed. The best performance came from the down sampled dataset. The logistic regression classifier had the highest accuracy followed by random forest and MLP algorithms. In the figure given above, from Table III, we can notice that all

the models yield a good precision score on the positive class (second column of precision). This value shows us that the model is confident in its result in classifying as a borrower as trustworthy. This aligns with my goal, where we would want to provide another criterion to evaluate trustworthy borrowers who may have been affected by a global pandemic or a market crash resulting in not having enough credit scores.

Another metric, this paper focussed at is ROC curve. It tells us about the classifier's ability to distinguish between the two classes. Based on the figure below, MLP (Multi-Layer Perceptron) achieves the best area under the curve, followed by random forest and logistic regression.

Cluster	age	credit_amount	duration_in_month	classification
0	28.7	5593.7	32.2	60.0
1	27.9	1654.3	15.1	70.0
2	46.0	1551.7	10.7	80.0
3	47.5	4977.9	27.6	70.0

Figure 3.3: Performance of clustering on LightGBM

This study mainly focusses at the unsupervised learning part of the project, which is to get some meaningful insights into the structure of the data and to potentially categorize the various types of loan applicants in our dataset. This is in alignment with my vision, where I would want to see if there exist distinct characteristics among different groups of borrowers. If that is done right, we would build different prediction models for different groups.

The unsupervised learning technique tried in this paper is K-means clustering. The models achieved the best results when k=4. A prediction model was built for each one of the clusters. The results are shown in Table IV in fig. For each one of the four clusters, the LightGBM model performs the best out of all machine learning models. If we compare the model's performance on both Table III and Table IV, we can see that there has been a significant improvement in the model's accuracy from 57.47 percent to 71.57 percent. We could infer from the result that each cluster identified by the k-means algorithm exhibit characteristics that could be picked up by the model when trained separately, but not when the model is trained on the entire dataset.

Note: For classification, LightGBM is a highly efficient gradient boosting decision tree algorithm. It is an improved version, over 20 times faster, of the Gradient Boosting Decision Tree (GBDT) algorithm. GBDT is a machine learning algorithm commonly used in the prediction of clicks and multi-class classifications. As it is an augmented tree-structure classifier, to get a sense of performance differences between standard algorithms and more advanced algorithms, we can make comparisons between itself and the Random Forest classifier.

3.3 Summary, Limitations and Emergent Ideas

Above sections have demonstrated what the state-of-the-art for each approach looks like. In this final section, we will discuss the limitations of these studies that we can build upon and some general observations about each approach. According to various studies and research in this paper, we have noticed that ensemble approaches are usually the best performers from each of the studies. I was also pleasantly surprised to note that the top-performing algorithms were nearly similar in all studies. The top classifiers were Random Forests, Bagged (MLP) Neural Networks, and Bagged Decision Trees, according to Baesens, Van Gestel et. al. Besides, according to Rich Caruana, the top three chosen are Boosted Decision Trees, Random Forests and Bagged Decision Trees while another Stanford research paper discussed above claimed MLP, Random Forests and Logistic Regression to be the best ones. However, it is important to note that Caruana did not incorporate Bagged Neural Networks.

In a combined study of all research papers discussed above, Random Forests, Bagged Neural Networks (MLP), Boosted Decision Trees and Logistic Regression came out on top in all cases. Even Ereiz, Zoran. 's paper on OptiML produced similar results when it was fed with German credit dataset [16]. I have opted to use the above-mentioned models for this project given the agreement in these studies. Support Vector Machine would also be used for the sake of comparison (SVM).

In addition to these classification models, an attempt would be made to increase the overall accuracy by combining the work with clustering methods like k-means. In a global pandemic situation, it would be important to divide the customers into different segments and build a model for each of those segments so that a fair decision can be made on all potential customers. In Yiyun Liang's research paper, we noticed how there was an increase in accuracy after using K-means clustering with classification algorithms. In addition to the selection of classification algorithms, metrics are needed for the performance evaluation of those algorithms.

For this capstone thesis, this is essentially the direction I am following. In the German Credit data set that I am using for this project, over 20 features are characterizing the borrower, the vector X. The data set also includes a loan status feature with different possible values, each of which can easily be grouped into default or non-default status values. As defined in the report, this is essentially my binary response variable. Finally, given a set of borrower characteristics, I aim to estimate the likelihood of default and use that to assess if they are likely or unlikely to default. Again, the analysis is parallel to Baesens et. al.'s work and given the parallels between two, I have chosen to use some of its key performance indicators like Percentage Correctly Classified (PCC) and the area under the ROC curve (AUC) specifically.

To conclude this section, it can be said that to fulfil this project's objectives, we would improve the existing state-of-the-art credit scoring classification algorithm by combining it with certain clustering methods. To construct an excellent model, it would be significant to take certain factors into consideration that can affect the cost of a financial institution. Those factors would be measured by certain existing metrics like the accuracy, precision and recall of a model.

Chapter 4: Project Approach

The core aim of this project revolves around the prediction of credit risk using various machine learning classification algorithms. Various preprocessing and feature extraction techniques have been applied using Pandas to create data visualizations in order to get a better understanding of the dataset.The original dataset has been split into training and testing data, and K-fold cross validation [17] has been used on various models to determine those with the highest mean accuracy.

Four models have been explored in this project: Random Forest, XGBoost, Logistic regression and Multi Layer Perceptron. An attempt has also been made to improve the accuracy levels of the models by using certain clustering methods followed by classification. Last but not the least, we have also created a simulated environment for a pandemic where we try and see if there are any changes in the patterns observed earlier (before pandemic). All these factors would be touched upon as we move further in this report.

This project can be divided into five phases which would be discussed in detail in the upcoming subsections:

4.1 Data Design and Considerations

For this project, a imbalanced machine learning dataset known as the "German Credit" dataset would be used. The German Credit data set is a publically available data set downloaded from the UCI Machine Learning Repository.

The dataset was used as part of the Statlog project in the 1990s, which evaluated and compared a large number (at the time) of machine learning algorithms on a variety of classification tasks. The dataset is credited to Hans Hofmann.

The German credit dataset describes financial and banking details for customers and the task is to determine whether the customer would be able to pay back the loan or not. The dataset includes 1,000 examples and 20 input variables, 7 of which are numerical (integer) and 13 are categorical.

There are two classes, 1 for good customers and 2 for bad customers. Bad customers are the default or negative class where people are not able to pay back the credit amount, whereas good customers are the exception or positive class where people do pay back the loan on time. A total of 70 percent of the examples are good customers, whereas the remaining 30 percent of examples are bad customers. In this project, we have replaced 2 with 0 which basically denotes:

1 - Positive Class, 0 - Negative Class

After taking a close look at the data, we can see that the numerical variables have different scales, e.g. 6, 48, and 12 in column 2, and 1169, 5951, etc. in column 5. This suggests that scaling of the integer columns will be needed for those algorithms that are sensitive to scale.

The target variable or class is the last column and contains values of 1 and 2. These will need to be label encoded to 0 and 1, respectively, to meet the general expectation for imbalanced binary classification tasks where 0 represents the negative class and 1 represents the positive class.

One of the major limitations in this project has been the availability of large datasets. These datasets can be very sensitive as it revolves around one's financial records and hence, it can always lead to privacy and ethical concerns. German Credit Data is one of the very few datasets that gives us an accurate picture about how loans work but one of it's major limitation is that it has only 1000 examples to work around. Even then, it stands out the most among all other datasets because of it's integrity and completeness. Nonetheless,a small python script has been written to convert it into a more readable CSV file.

4.2 Data Collection and Analysis

The German Credit data set is a publically available data set downloaded from the UCI Machine Learning Repository. It is a highly credible dataset in this field and is known for it's completeness and data integrity.

In an attempt to make the data set more understandable, a small python script has been written. Owing to it's data quality, there were no data cleaning procedures that were implemented.

In Fig 4.1, we can see the relevance of each feature in this dataset.

Figure 4.1: Feature Importance

After acquiring knowledge in how important each attribute is, we would explore these on an individual basis.

First and foremost, we would discuss the effect 'Age' has on a loan outcome. In this project, we have created categorical groups based on the age column:

Student: Clients age ranges from (18 - 25)
Young Adults: Clients age ranges from (26-40)
Senior: Clients age ranges from (41-55)
Elder: Clients age is more than 55 years old

What we have wanted to accomplish is to create different age groups based on their age and observe the credit amounts borrowed by clients belonging to each age group. We have also determined which loans are high risk and seen if there are any patterns with regards to age groups.

The main observations which could be concluded are discussed hereby. The younger age group tended to ask slightly for higher loans compared to the older age groups. The student and elderly groups had the highest ratio of high risk loans. With 45.29 percent of all the clients that belong to the student age group being considered of high risk. The number of loans that were considered of high risk within the elderly group is 44.28 percent of the

total amount of people considered in the elderly group. One interesting fact to observe is that these are the groups that are most likely to be unemployed or working part-time, since the youngest group either don't have the experience to have a job or they are studying in a university so they don't have enough time to work in a full-time job. In the elderly group side, this is the group that are most likely receiving their money from their pensions, meaning the elderly group is most likely unemployed or working part-time as well.

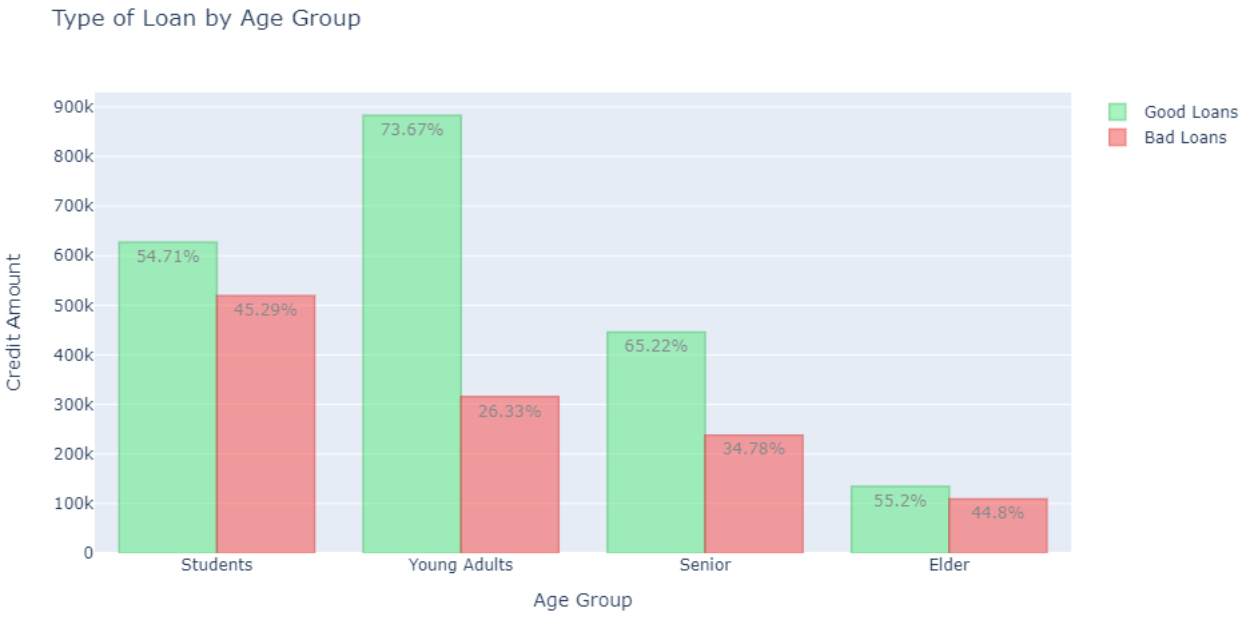

Figure 4.2: Type of Loan by Age Groups

Next, we look if a person's housing status affects his or her loan outcome. From Fig 4.3, we can clearly infer that people having their own houses and the chances of the loan getting approved have a very high correlation compared to others.

People having their own house has a 73 percent chance of getting a loan approved while the other two categories only stand 60 percent chance, keeping everything else constant.

Moving on to our next attribute 'Gender', we would try and see if a loan outcome can be biased based on a person's sex. First observation made is that there are 2x more males than females in our data set, which is usually the case too if observed carefully. Most females that applied for a credit loan were less than 30 while most men who applied for a loan ranged from their 20s-40s. Females were more likely to apply for a credit loan to buy radio/television. (10 percent more than males) On the other hand, males applied 2x more than females for a credit loan to invest in a business. Quite interestingly, 2x of females were unemployed non-resident compared to males.

In this project, we have also analyzed the amount of wealth our clients have by analyzing their checking accounts and whether the wealth status of our clients contribute to the risk of the loans being issued to customers.As expected, individuals belonging to the "little wealth" group, had a higher probability of being bad risk loans than other types of groups. In figure 4.4 , we can notice the higher the wealth, the lower the probability it became of being a high-risk loan.

Like other attributes, we also explore the purposes of loans. In this section, my main aim is to see what purposes were most likely to bring most risk, in other words which of these purposes are more likely to be considered high risk loans. Also, I would like to explore the operative side of the business, by determining which purposes where the ones that contributed the most towards loans issued. [18]

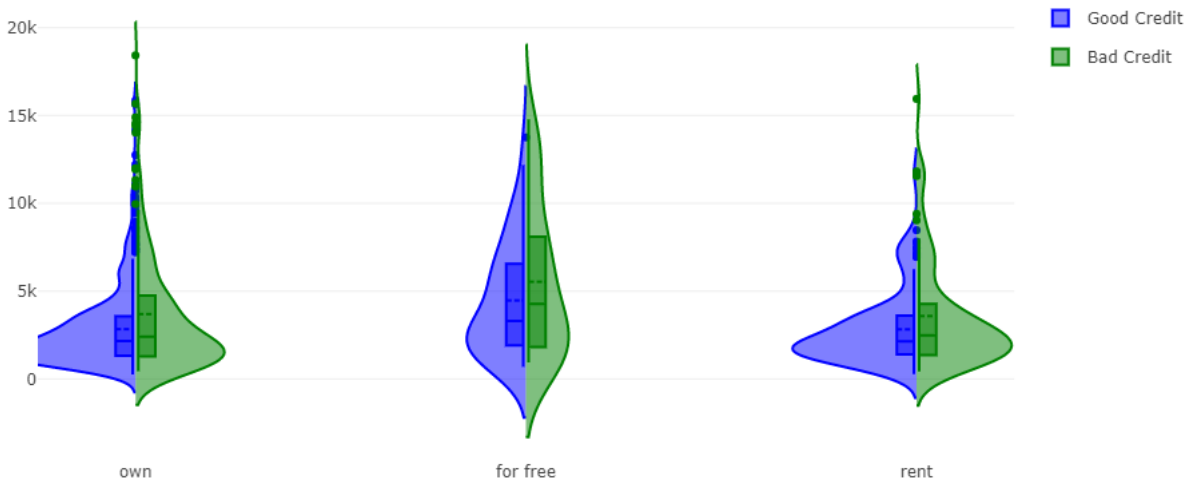

Figure 4.3: Risk association to Loan by one's Housing Status

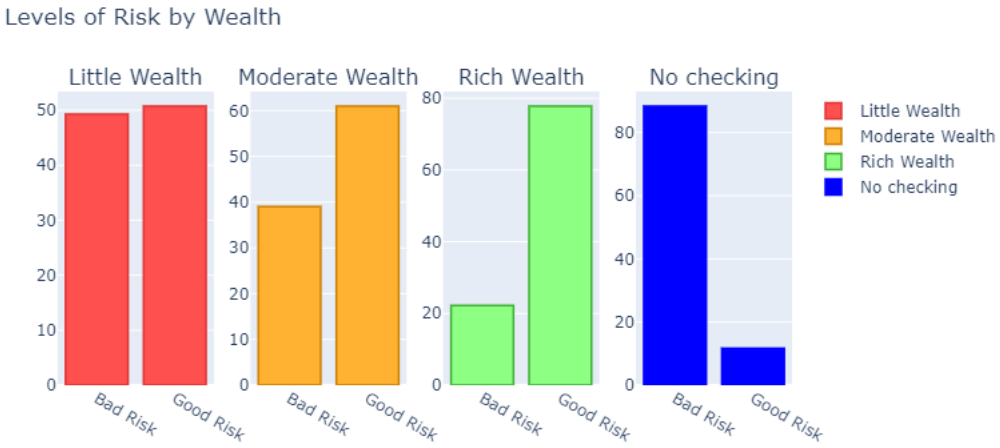

Figure 4.4: Levels of Risk by Wealth

Cars, repairs and domestic appliances made more than 50 percent of the total risk and has the highest distribution of credit issued. The rest of the purposes were not frequent purposes in applying for a loan. Used cars and domestic appliances were the less riskier purposes from the operative perspective since it had the widest gap between good and bad risk.

With this, we come to an end in the analysis section where all of our important attributes have been discussed and analyzed on an individual basis.

In our next subsection, we would talk about our various classification models that has been used for training and evaluation purposes.

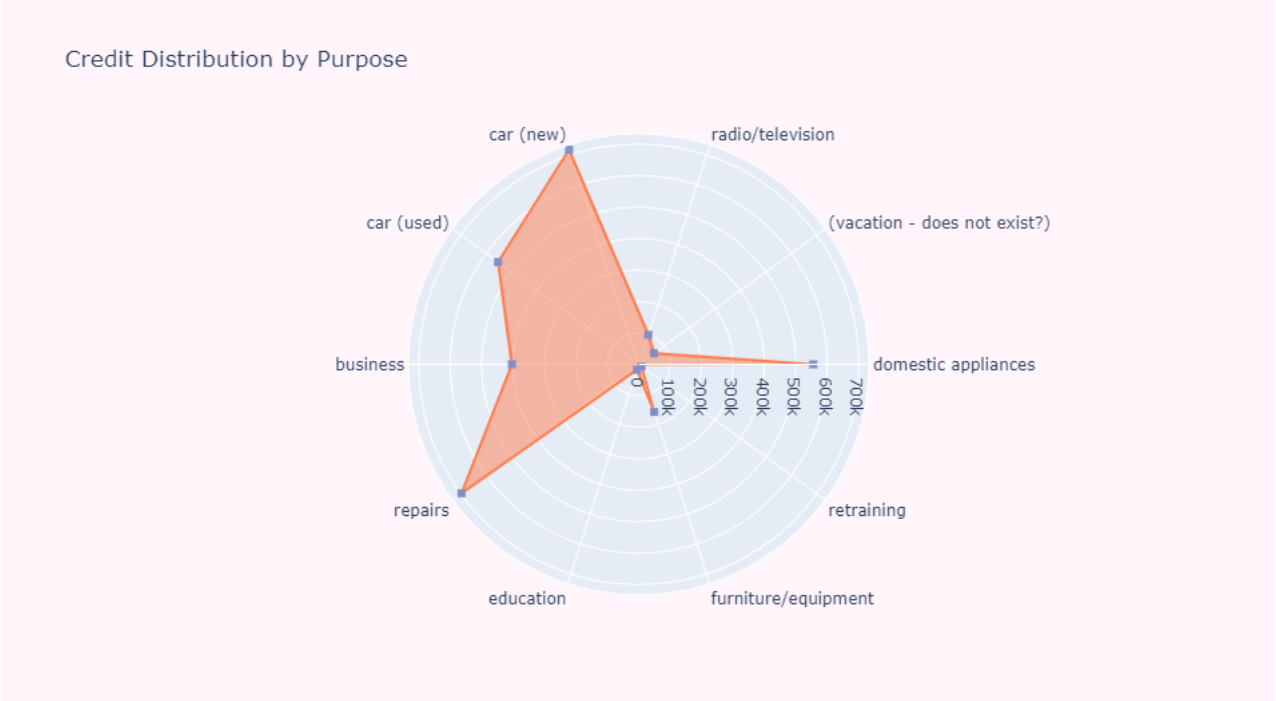

Figure 4.5: Credit Distribution by Purpose

Risk	bad	good
Purpose		
domestic appliances	1.33	1.14
vacation/others	1.67	1.00
repairs	2.67	2.00
education	7.67	5.14
business	11.33	9.00
furniture/equipment	19.33	17.57
radio/TV	20.67	31.14
car	35.33	33.00

Figure 4.6: Type of Credit by Purpose

4.3 Classification Models Training and Evaluation

In this project, we have used four Classification models: XGBoost, Random Forest, Logistic Regression and Neural Networks (MLP). These models have been implemented on two datasets - balanced and the imbalanced (original) one [19]. To address the imbalance between classes for the response variable, an oversampling technique known as SMOTE was used.

Before the implementation of each of these models, the data sets were split into two parts - the training set and the test set. We would train one existing data set and using that, build a model with the help of selected algorithms.

First and foremost, XGBoost model [20] was deployed. XGBoost is a boosting ensemble algorithm which optimizes the objective of function, size of the tree and the magnitude of the weights are controlled by standard regularization parameters. This model was proved to outperform several baseline techniques and was validated on five datasets over five performance

metrics.

Under normal circumstances, that is while using the imbalanced dataset, XGBoost model showed an ROC curve [21] of 0.76. After using hyper parameter tuning (setting a learning rate of 0.05), the ROC curve increased to 0.79 which can be considered as a significant improvement. Surprisingly, when we used the balanced dataset the model did not show any significant improvement. At the same time, the recall for a default loan using this model is 84 percent with a precision of 80 percent. For most classification problems, precision is usually less important than the recall. Predicting a loan that will be good as bad (false positive) is not as costly as predicting a loan that will be bad as good (false negative). In the first case some clients that would be good may be lost, but this has an opportunity cost associated that is likely lower than the cost of giving away a loan that will not be repaid or that will take a long time and effort to get paid back. [22]

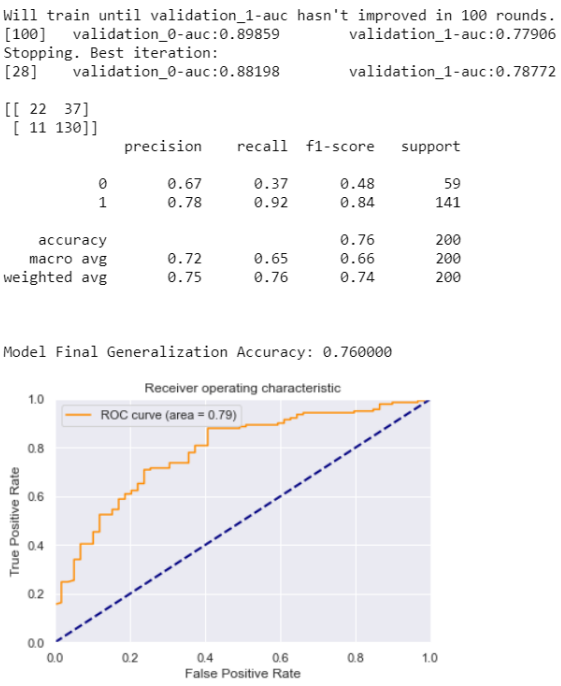

Figure 4.7: ROC Curve of XGBoost Model

Second model to be implemented was Logistic Regression [23]. The Logistic Regression approach has been considered a benchmark in the issue of credit scoring. Logistic Regression calculates the conditional risk of default of the creditor and describes the relationship between the creditworthiness of clients and explanatory variables. The method for Logistic Regression to create a model consists of estimating a linear combination of interpreter X and binary dependent variable Y and labelling using the logistic function to translate log-odds to probability. This model achieved the best accuracy of 80 percent and also showed an ROC curve of 0.79 while using the imbalanced dataset. Like XGBoost, even this model didn't show any major improvement when the balanced dataset was used.

One of the other well performing model was Random Forest [24] which excelled in terms of giving high accuracy and high ROC curve of around 0.80. Random forest is a supervised learning algorithm, an ensemble method for classification, regression and other tasks that work by constructing a variety of decision trees at the time of training and outputting the class that is the mode of the classes or the prediction of the individual trees. The decision tree is in the form of a tree (which can be a binary tree or a non-binary tree). Each of its non-leaf nodes corresponds to a feature test, with each branch representing the feature attribute output over a range of values, and with each leaf node storing a category. The decision tree begins with the root node, checks the corresponding attributes of the function in the category to be categorized, and selects the output branches according to their values

until the leaf node is reached. Eventually, the decision result is considered to be the category stored by the leaf node.A random forest is a group of such decision trees in which there is no relationship among each of the decision trees. The Gini index [25] is the selection metric that can be used to separate attributes in the decision tree, and the number of levels in each tree branch depends on the algorithm parameter d (depth). In this project, it is the only model that worked better with a a balanced dataset. After using the balanced dataset, the cross validation dataset rose from 0.75 to 0.82 which can be marked as a big improvement in the model performance.

Last not but the least, we had implemented a Neural Network model [26](MLP) that can be considered as the least performing model among the four models that are being used. A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer.Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. In this project, this model performed with an accuracy of 0.70 while achieving a ROC AUC score of only 0.5.

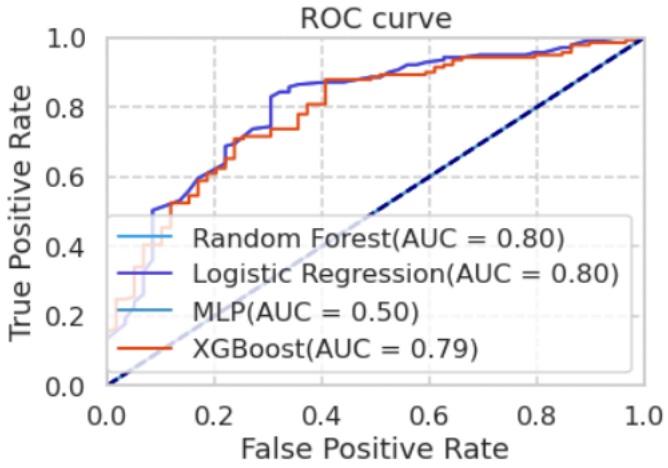

Figure 4.8: Model Comparison

In figure 4.8, a comparison of all models have been shown. We can observe that in our first attempt, Random Forest and Logistic Regression are the best performing ones.

4.4 Applying Clustering - an attempt to improve existing model and Bank customer segmentation

In this project, we have used clustering methods for two very important purposes -

1. Bank Customer Segmentation [27]:

In this section, we have mainly considered three important columns: Age, duration of loan and the credit amount. Before, we move on to clustering, we can have a look at their distribution in Fig 4.9.

We can see that distributions are right-skewed. To obtain better results the skewness was removed by logarithmic transformation. The next step was centering and scaling of variables

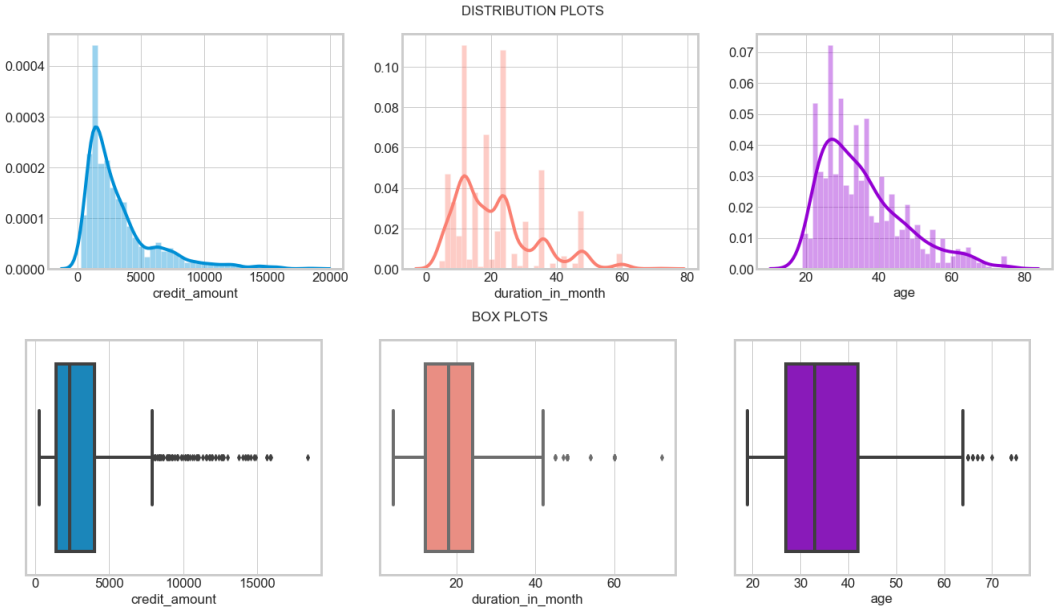

Figure 4.9: Feature Distribution

– which is required by KMeans algorithm. StandardScaler from sklearn library have been used for that purpose. The number of clusters chosen were according to the silhouette scores for various combinations of random state and number of clusters. The highest scores were for 2 and 3 clusters and they were relatively insensitive to seed.

The observations made were:

Cluster 0 dealt with older customers borrowing lower mean of credit amount in a short duration.

Cluster 1 dealt with middle aged customers borrowing high mean of credit amount in a long duration.

Cluster 2 dealt with young customers with lower mean of credit amount in a short duration.

There is one more method that have been implemented, which is, clustering with affinity propagation [28].In this algorithm there are two relevant parameters: preference and dumping. It means that we don't define upfront number of clusters, algorithm itself chooses their number. We fix dumping and check number of clusters in function of preference parameter. According to this method, the number of clusters chosen were 4.

From fig 4.10, we can see how four of the clusters react differently. To be precise, the observations have been listed below supporting it with Fig 4.11:

Cluster 0 contains of younger customers having a 60 percent chance of loan approval who borrows high mean of credit amount for a long duration.

Cluster 1 contains of younger customers having a 70 percent chance of loan approval who borrows low mean of credit amount for a short duration.

Cluster 2 contains of older customers having a 80 percent chance of approval who borrows low mean of credit amount for a short duration.

Last but not the least, cluster 3 contains of older customers having a 70 percent chance of loan approval who borrows high mean of credit amount for a middle-time duration.

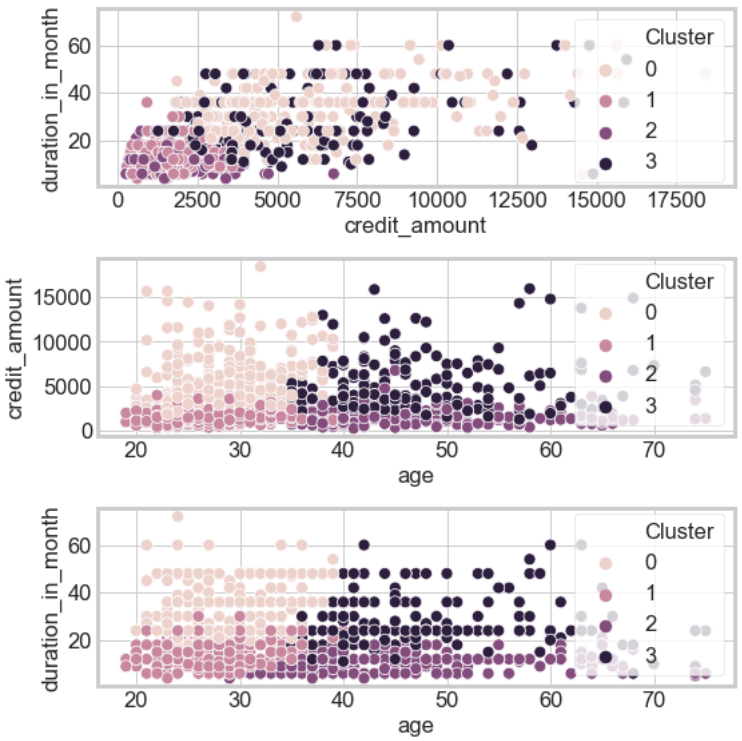

Figure 4.10: Clustering

Cluster	age	credit_amount	duration_in_month	classification
0	28.7	5593.7	32.2	60.0
1	27.9	1654.3	15.1	70.0
2	46.0	1551.7	10.7	80.0
3	47.5	4977.9	27.6	70.0

Figure 4.11: Cluster Behaviour

2. Improvement in Model Performance:

The main reason for clustering has been to improve the performance of our existing classifier models to predict a loan outcome. This part of the project has also been considered as an advanced aim.

In this section, we have used the elbow method to find the optimal k for our kNN clustering [29]. The number of clusters was chosen to be 3 which produced excellent results. The clusters can be seen given in Fig 4.12.

After we had our three clusters ready, we ran our 4 existing models on each of them. To our surprise, some of clusters delivered excellent results when implemented with a certain model. For example, our cluster 3 achieved a ROC Curve score as high as 0.86 with our Logistic Regression model. If we look at it in one picture, we can observe in Fig 4.13, that 3 of our 4 existing models showed an improvement in the mean ROC curve levels.

The goal of this unsupervised learning part of the project was to get some meaningful insights into the structure of data and to potentially categorize the various types of loan applicants in our dataset. We wanted to see if there exist distinct characteristics among different groups of borrowers, if so, we could build different prediction models for different groups. The unsupervised learning technique implemented in this project is K-means clustering. It was

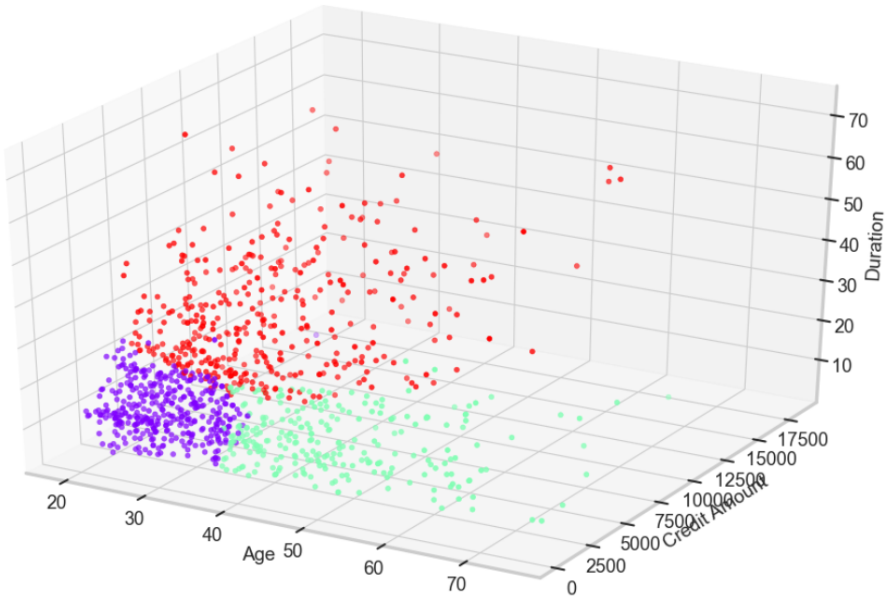

Figure 4.12: Clusters - 3D Model

Classification Method	Before Clustering	After Clustering
XGBoost	0.79	0.83
Logistic Regression	0.79	0.79
Neural Networks	0.50	0.53
Random Forest	0.79	0.82

Figure 4.13: Improvement in Mean AUC ROC Curve

first performed on the dataset where we had experimented with several values of k. Then a prediction model was built for each one of these clusters. The results are shown in Fig 4.13. The models achieved the best overall performance when k = 3. For each one of the four clusters, the XGBoost and the random forest model performed the best out of all the machine learning models.

4.5 Data Simulation and Model Development for a Global Pandemic

In this project, this section can be considered the most trickiest since we do not have real life datasets for this situation. Hence, datasets have been simulated based on research and keeping international news in mind.

For example, according to research we know that the section of people who have been affected the most due to this pandemic (Covid-19) are the people in the lower income category or those who are working in retail industry and restaurants [30]. Keeping that in mind, we have simulated our data accordingly. In the dataset, people who were in the unskilled category and also belonged to 'moderate wealth' category as seen in Fig 4.4, they were moved to the 'Little Wealth' category. Their employment status were also changed to being unemployed. From this simulation, we can clearly observe that the unskilled workers (or the potential low

income people) were affected the most. It should be kept in mind that this dataset is just a simulation and is used for project purpose. The simulated dataset has a lot of limitations since we do not have any real life dataset to compare it with and hence, it should not be used for any realistic purposes.

After the data simulation, our existing classification models have been used to predict the loan outcome on our simulated dataset. The most important observation that can be made is that the number of loans had dropped considerably after the pandemic. We can see the pandemic's effect on the number of loans approved in Fig 4.14. The number of loan approvals had drastically reduced from 700 to 602 out of 1000 samples.

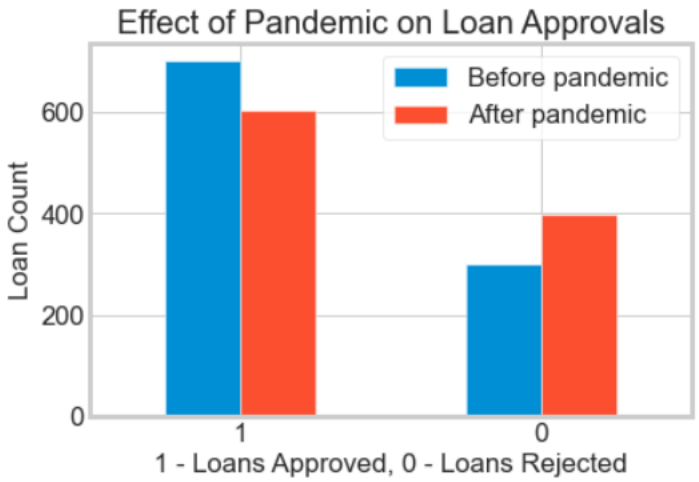

Figure 4.14: Effect of Pandemic on Loan Approvals

The XGBoost model had been deployed to this simulated dataset and it showed a accuracy of 0.81 with a very high ROC curve of 0.89.

An assumption has been made that all people pose a threat who belong to the low wealth and unemployed category, and they still have been approved a loan. The sum of credit amount borrowed by these people has been considered the potential risk. A comparison has been made for the potential risk of a bank before and after the pandemic where it can be observed that the risk grew from around 13000 euros to 65000 euros (based on 1000 people in the existing dataset) after the pandemic.

4.6 User Interface Design

In this section, we will talk about the design and the technicalities of the website which has been built in order to help people get an idea if they could get a loan approved in a bank. It should be kept in mind that the dataset used to train the machine learning model is fairly old and due to it's limited data, the result/prediction obtained may not always be accurate when compared to the actual decision offered by banks.

This section can further be divided into two sub sections where we would discuss about the frontend and the backend separately.

4.6.1 Front-end Web Design

This website has been designed keeping user experience as a priority. Efforts have been put to ensure that an user finds it easy to navigate through the website and also should be able

to understand the functionality of each page. An attempt has been made to keep the website fairly simple and user friendly. It is a responsive website which would allow an user to access it on any device such as PC or a mobile phone.

The structure of the website can divided into the following parts:

1. Home Page:

The Home page has been designed in a way to tell an user what the website is about and the functionality that it has. For example, in our website we have kept a heading (German Credit Data Analysis/Visualization/Prediction) and have also kept two buttons which can redirect an user to the visualization or the prediction page.

Underneath that, a section has also been kept where users have been given some information about the dataset, how the data analysis have been performed and how we have made predictions based on the data.

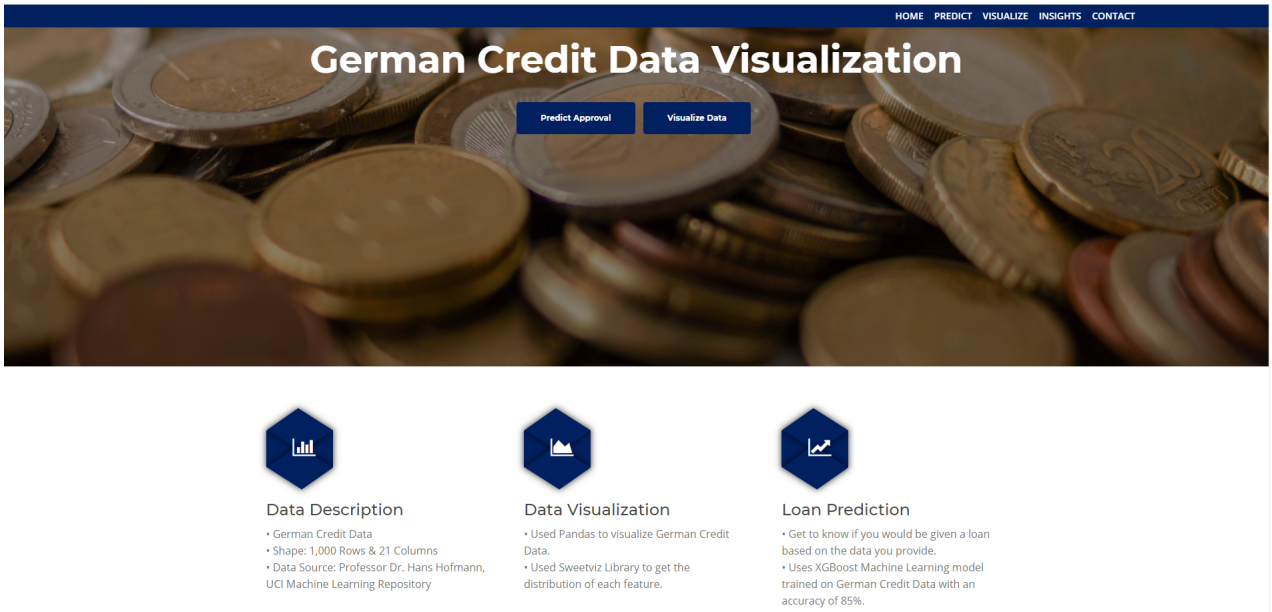

Figure 4.15: Home Page

2. Predict:

In this section, attempts have been made to keep things quite straight forward. We directly ask an user if they are interested in an individual loan or not. If they are, there's an 'Yes' button which can be clicked, that would redirect the user to the form page where they would be asked to enter personal information which are usually needed by banks to process a loan. Underneath than the form, we have provided a few figures as well that reveal information about our dataset and the model performance.

Once the form is filled, the user can click on the 'Predict' button which would take them to the result page that would be reveal whether their loan application has been accepted or rejected.

Keeping certain ethical considerations and privacy policies in mind, it has been ensured that no customer's information is stored in our database. The information entered by an user is used by a machine learning model to predict the loan outcome after which the data is completely removed from the system.

3. Visualize

In this section, the visualizations related to the most important observations have been shown. For example, we look into the age and sex distribution and the risk associated to it. The type of loans by age groups, loan application reasons, and levels of risk by wealth category has also been looked into.

Figure 4.16: Visualisations shown in Website

4. Insights

In the insights section, we have shown how our machine learning model is a little different from other since we have used clustering algorithms too in order to improve model accuracy. A 3D model has also been shown in order to give a clear idea about the data clusters in German credit data.

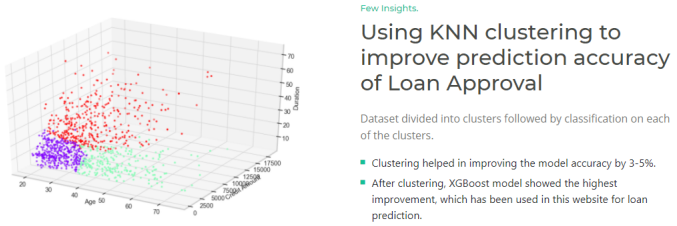

Figure 4.17: Insights

5. Contact

The contact section has been made the footer of the web page where one can get access to my personal contact information and can connect with me on my social media handles.

4.6.2 Back-end Web Design

This website can be launched by running app.py on the local terminal. App.py is a flask app [31] that is used for prediction. At first, we have imported the model using pickle library. The model which has been deployed in this website is XGBoost because of it's high accuracy and ROC curve area as observed in this project.

At the frontend, after the form is filled we receive the data in the predict function in app.py. This is followed by conversion of these user inputs into a pandas data frame which is used further to predict the loan outcome column using our machine learning model. After that, the result is being returned which gets displayed on the website.

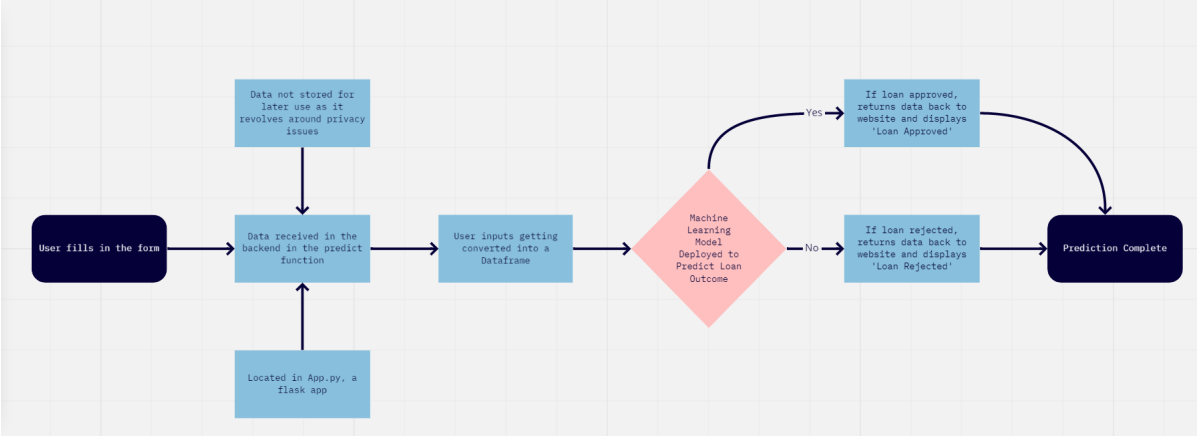

Figure 4.18: Flow chart of the Backend Process

Chapter 5: Project Workplan

5.1 Future Plan

Below is an outline of how this project will be approached by using a Gantt Chart. The pre-processing stage including environment setup and data collection would be conducted within the first 2 weeks. This phase will be crucial in the build-up to developing further code and training suitable predictive models. It also plays an important role in the performance of the classifier and clustering methods.

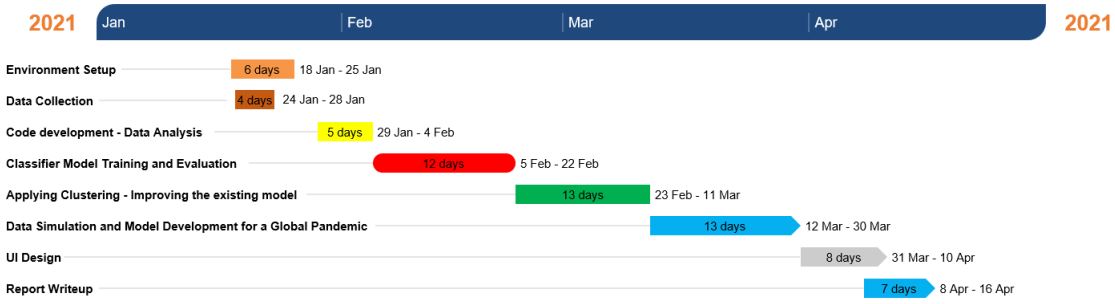

Figure 5.1: Gantt Chart

In the Data Analysis phase, we would dig in deep to identify any interesting pattern that would help us moving forward in the project. For example, we would try to find out how relevant each field is to our target column in the dataset – Risk Classification/Credibility (Good/ Bad). Using the correlation, we might form a risk function which would help in determining in the prediction of a loan. This phase would roughly range around 7 days and it would be expected to be completed by week 3 of college – 4th February.

Moving on to the classifier model, we would split the entire dataset into two parts – train and test. We would train one existing dataset and using that, build a model with the help of selected algorithms – MLP (Neural Networks), Boosted Decision Tree, Support Vector Machine, Logistic Regression, and Random Forest. After our model is ready, we would apply the trained classifier on the new test data and try to predict the loan outcome using the highest accuracy model (or pipeline of two good models). One of the evaluating measures would be the accuracy score which would reveal how good our model is. This phase would take around 2-3 weeks and would be expected to be completed by 22nd of February.

Classification would be followed by clustering that would take up around two more weeks. It would help us to find out some common behavioural patterns of customers. The clustering method would also be useful for predictive modelling and we would also use hyperparameter tuning [32] to find an optimal combination of hyperparameters that minimizes a predefined loss function to give better results.

Before moving on to UI design, we would also try to build a predictive model which would take an unforeseen event (like a global pandemic or a market crash) into account. As discussed earlier, we would have a risk function that would calculate a credit score for all customers. Based on the original dataset, a second dataset would be simulated keeping in mind that there's been a global pandemic. The model would automatically be able to identify the customers who have been affected by the pandemic. A separate model would be used to treat the affected customers so that a fairer decision can be made. A period of

13-18 days would be considered to build that model and it would be expected to be finished by 30th March.

After the skeleton work is ready, an attempt would be made to design a UI where customers can fill in a form that is usually required for a loan. On clicking the Submit button, it would call an API that would give the final result, that is, if the loan has been approved or rejected. We would devote somewhere around 7 days to build a UI which would help users to interact with the application. This part of the project would bring to an end and it is expected to be completed by 10th April.

The final stage of the project would involve writing a detailed report which would commence on 8th April and would be completed by 16th of April. The entire project would thus end a week earlier than the college term giving us some time to revise all our considerations.

5.2 Evaluation

To evaluate our classifier model, various metrics have been used such as the model accuracy, precision, recall and ROC AUC curve [33]. The easiest metric to understand would be the accuracy as it would reveal the number of times the prediction is being made correctly. We would also look at the confusion matrix that would reveal some of the metrics – true positive rate, true negative rate, false-positive rate and false-negative rate. Predicting a loan that will be bad as good (false negative) is very costly compared to a loan that will be good as bad (false positive).

Three important factors by which clustering can be evaluated is a) clustering tendency, b) Number of clusters, k and c) clustering quality. Sklearn package provides some metrics that help us to evaluate clustering models [34]. For example, we would be able to calculate the Silhouette Coefficient score which relates to a model with better-defined clusters. It is defined for each sample and is composed of two scores:

- a: The mean distance between a sample and all other points in the same class.
- b: The mean distance between a sample and all other points in the next nearest cluster.

There are several other metrics like the Davies-Bouldin Index [35] and Calinski-Harabasz Index [36] that would help us get an idea of how good is our clustering model.

Towards the end, the classifier models have been combined with clustering methods to improve overall accuracy. The goal of clustering is to get some meaningful insights into the structure of data and to potentially categorize the various types of loan applicants in our dataset.

The evaluation of this project can be divided into two phases:

A. Classification Approach - When we used normal classification methods, XGBoost and Random Forest performed the best among all models showing ROC curve levels of 0.79 and 0.80 respectively. Random forest was the only model that showed improvement when used with balanced dataset after the dataset was upsampled using SMOTE. Logistic Regression showed a ROC curve level of 0.78 while Neural Networks only had a score of 0.50. The performance of all our existing models have been compared as observed in Fig 5.2.

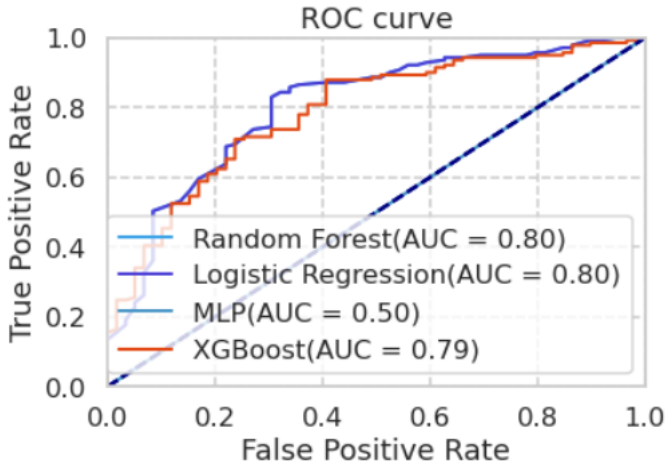

Figure 5.2: Comparison of Machine Learning models

B. Classification with Clustering Approach - The dataset was divided into small clusters using kNN clustering. The clusters showed a decent Silhouette Coefficient score [37] of 0.65. After the clusters were formed, our existing models were run on each of these clusters individually. Some clusters performed exceptionally well showing ROC curve as high as 0.89 while some of them performed at an average level. Overall, three out of our four machine learning models showed an improvement in the mean ROC curve levels as seen in figure 4.13. XGBoost and Random Forest remained two of our best performing models with an improved ROC score of 0.81.

Moving on to our next aim, we have simulated a dataset for a pandemic. There is no right way of evaluation for this as there is no real life dataset at present. But, all the assumptions have been made after thorough research which can be based on valid news sources.

To meet our last aim, a website has been designed keeping user experience in mind. The main functionality of the website is to allow an user check if he or she would be eligible for a loan.

Chapter 6: Conclusions and Future Work

The goal of each and every business is to make profit. For a lender, profit depends on whether or not the borrower repays the principal as well as interest. Without repayment, the lender would incur a loss and that loss can even potentially be greater than the initial loan amount when lawyer, court and collection fees are taken into consideration. Therefore, it is critically important for a lender to be able to identify whether a potential borrower can and will make all of his or her loan payments. The purpose of this project is to identify the characteristics (limited to those found in the dataset) of persons who are likely to default on their loans and provide a simple framework for borrowers to make such a distinction. In order to make predictions on which potential borrower is or is not likely to default, binary classification predictive models have been created using a variety of machine learning algorithms.

The two big takeaways from this project would be on how we have improved our prediction accuracy levels by clustering and how a pandemic has been dealt with. After clustering, we have been able to divide customers on the basis of age, sex and the credit amount which has allowed us to make highly accurate predictions for each of these groups specifically. In order to conclude the project, it can be said on a satisfactory note that a loan prediction system has successfully been formed which can make accurate predictions with a highly advanced novel method that has not been explored on a big scale before. The model is specially equipped to handle uncertain situations as well like a global market crash or a pandemic. The model can be explored and tested real-time through a user interface that has been created.

The potential future work for this project will be a further development of the model by deepening analysis on variables used in the models as well as creating new variables in order to make better predictions. Data available for the scope of this thesis has constraints in terms of many years are covered by the data presented as well as geographical breadth of Germany's clients. The majority of customers are Germany's clients, thus, it should be considered that the behaviour of German customers influence the results of this research. It means that the behaviour of clients outside Germany may or may not follow the same pattern and therefore one should make additional analysis and obtain a geographically-broader data set if the objective is to have a model unbiased of the geographical location.

An assumption can also be made that if there is data available for longer time span as well as broader geography of clients, there is an interest to implement marco- economical variables, which in turn might open some new insights about factors impacting default of a customer as well as what machine learning methods are more suitable for this type of a problem.

It would also be interesting to make a study concerning what metrics are the most relevant for this type of the problem. As mentioned previously, in this project the main metric all evaluations were analyzed by was AUC ROC curve and accuracy score. If a deeper analysis could be performed regarding the most relevant metric for this type of problem, then potentially a weight function could be implemented if one of the metrics explored turned out to be of more importance. With every day passing by, technology is improving at a very fast pace to make human life seamless. With this project, the goal is similar - to make financial lives easier for both individuals and financial institutions.

"Financial freedom is available to those who learn about it and work for it." - Robert Kiyosaki.

Chapter 7: Appendix

7.1 Sex Distribution: Sex Count and Credit Amount by Sex

Figure 7.1: Sex Distribution

In figure 7.1, it can be observed that there are twice the number of males than females in our dataset and if we look at how much credit one person takes, it can noticed that men tend to take a higher credit amount as compared to women.

7.2 Savings account exploration

Figure 7.2: Savings account exploration

From figure 7.2, it can be concluded that a large percentage of people who had a savings account status of $>$ 100 DM, their loans were most likely to get approved, taking the credit amount into consideration.

7.3 Duration Frequency for good and bad credit

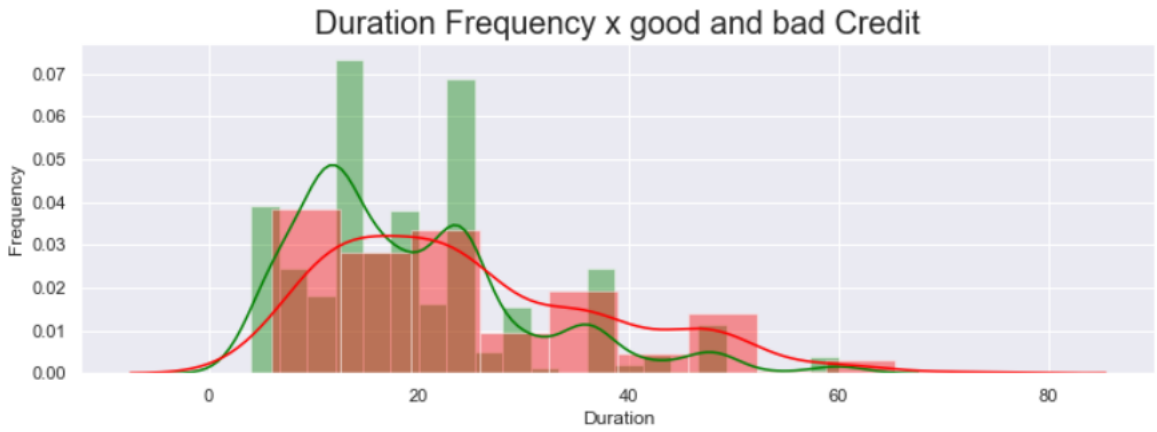

Figure 7.3: Duration Frequency

In figure 7.3, it can be seen that most of the people who apply for loans take a time duration of 12-24 months on an average to pay back the loan.

7.4 Adam Algorithm

Algorithm 4: Adam algorithm

Choose a stepsize γ
Determine decay rates for the moment estimates ψ_1 and $\psi_2 \in [0, 1)$
Define stochastic objective function with parameters θ
Initialize θ_0
Set first and second moment vectors m_0 and v_o to 0
Set timestep $t \leftarrow 0$
Set $\varepsilon \leftarrow 10^{-8}$
while θ_t *has not converged* **do**
    $t \leftarrow t + 1$
    Gradients with regards to stochastic objective is computed
    $\phi_t \leftarrow \nabla_\theta R_t(\theta_{t-1})$
    First moment estimate is updated $m_t \leftarrow \psi_1 \cdot m_{t-1} + (1 - beta_1) \cdot \phi_t$
    Second moment estimate is updated $v_t \leftarrow \psi_2 \cdot v_{t-1} + (1 - beta_1) \cdot \phi_t^2$
    Estimation of bias-corrected first raw moment estimate is performed
    $\hat{m}_t \leftarrow \frac{m_t}{(1-\psi_1^t)}$
    Estimation of bias-corrected second raw moment estimate is performed
    $\hat{v}_t \leftarrow \frac{v_t}{(1-\psi_2^t)}$
    $\theta_t \leftarrow \theta_{t-1} - \gamma \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\varepsilon}$
Output: θ_t

Figure 7.4: Adam algorithm

Bibliography

1. *Credit Risk Assessment* https://www.assetzcapital.co.uk/credit-risk-assessment.
2. Dua, D. & Graff, C. *UCI Machine Learning Repository* 2017. http://archive.ics.uci.edu/ml.
3. *Creating actionable customer segmentation models* https://looker.com/blog/creating-actionable-customer-segmentation-models.
4. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.* **16,** 321–357. ISSN: 1076-9757 (June 2002).
5. Lessmann, S., Baesens, B., Seow, H.-V. & Thomas, L. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* **(doi:10.1016/j.ejor.2015.05.030)** (May 2015).
6. Liang, Y. *Loanliness: Predicting Loan Repayment Ability by Using Machine Learning Methods* in (2019).
7. Narkhede, S. *Understanding AUC - ROC Curve* May 2019. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.
8. Caruana, R. & Niculescu-Mizil, A. *An Empirical Comparison of Supervised Learning Algorithms* in *Proceedings of the 23rd International Conference on Machine Learning* (Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, 2006), 161–168. ISBN: 1595933832. https://doi.org/10.1145/1143844.1143865.
9. Zhu, L., Qiu, D., Ergu, D., Ying, C. & Liu, K. A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science* **162.** 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence, 503–513. ISSN: 1877-0509. http://www.sciencedirect.com/science/article/pii/S1877050919320277 (2019).
10. Alejo, R., García, V., Marqués, A., Sánchez, J. & Antonio-Velázquez, J. in, 1–8 (Jan. 2013). ISBN: 9783319005683.
11. Alaraj, M. & Abbod, M. Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems* **104** (Apr. 2016).
12. Kingma, D. & Ba, J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (Dec. 2014).
13. Vangaveeti, S. A. May 2020. http://www.jctjournal.com/gallery/39-may-2020.pdf.
14. Mong'are, D., Njoroge, G. & Muraya, M. Analysis of Individual Loan Defaults Using Logit under Supervised Machine Learning Approach. *Asian Journal of Probability and Statistics* **3,** 1–12 (May 2019).
15. Xia, Y., Liu, C., Li, Y. & Liu, N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Syst. Appl.* **78,** 225–241 (2017).
16. Ereiz, Z. *Predicting Default Loans Using Machine Learning (OptiML)* in (Nov. 2019), 1–4.
17. Refaeilzadeh, P., Tang, L. & Liu, H. in *Encyclopedia of Database Systems* (eds LIU, L. & ÖZSU, M. T.) 532–538 (Springer US, Boston, MA, 2009). ISBN: 978-0-387-39940-9. https://doi.org/10.1007/978-0-387-39940-9_565.

18. Robson, B. *The Top 9 Reasons To Get A Personal Loan* Feb. 2021. https://www.bankrate.com/loans/personal-loans/top-reasons-to-apply-for-personal-loan/#:~:text=Personal%20loans%20are%20borrowed%20money,emergency%20expenses%20and%20much%20more..
19. Brownlee, J. *A Gentle Introduction to Imbalanced Classification* Jan. 2020. https://machinelearningmastery.com/what-is-imbalanced-classification/.
20. Brownlee, J. *A Gentle Introduction to XGBoost for Applied Machine Learning* Feb. 2021. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.
21. Melo, F. in *Encyclopedia of Systems Biology* (eds Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H.) 38–39 (Springer New York, New York, NY, 2013). ISBN: 978-1-4419-9863-7. https://doi.org/10.1007/978-1-4419-9863-7_209.
22. 22, J. & Johnson, J. *Precision, Recall amp; Confusion Matrices in Machine Learning* July 2020. https://www.bmc.com/blogs/confusion-precision-recall/.
23. Hoffman, J. I. *Logistic Regression Analysis* https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis.
24. Donges, N. *A complete guide to the random forest algorithm* June 2019. https://builtin.com/data-science/random-forest-algorithm.
25. Menze, B. H. *et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data* July 2009. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-213.
26. Franckepeixoto. *A Simple Overview of Multilayer Perceptron (MLP) Deep Learning* Dec. 2020. https://www.analyticsvidhya.com/blog/2020/12/mlp-multilayer-perceptron-simple-overview/.
27. *What Is Customer Segmentation in Banking?: Cognizant* https://www.cognizant.com/glossary/customer-segmentation-banking#:~:text=Customer%20segmentation%20is%20the%20approach,geography%2C%20income%20and%20spending%20habits..
28. Dueck, D. *Affinity propagation: clustering data by passing messages* (Citeseer, 2009).
29. Yuan, C. & Yang, H. Research on K-value selection method of K-means clustering algorithm. *J—Multidisciplinary Scientific Journal* **2,** 226–235 (2019).
30. Klein, A. & Smith, E. *Explaining the economic impact of COVID-19: Core industries and the Hispanic workforce* May 2021. https://www.brookings.edu/research/explaining-the-economic-impact-of-covid-19-core-industries-and-the-hispanic-workforce/.
31. comments, 0. A. 2. H.-W. *An introduction to the Flask Python web app framework* Apr. 2018. https://opensource.com/article/18/4/flask.
32. Jordan, J. *Hyperparameter tuning for machine learning models.* Dec. 2018. https://www.jeremyjordan.me/hyperparameter-tuning/.
33. Laken, P. v. d. *ROC, AUC, precision, and recall visually explained* May 2020. https://paulvanderlaken.com/2019/08/16/roc-auc-precision-and-recall-visually-explained/.
34. Science, O. O. D. *Assessment Metrics for Clustering Algorithms* Nov. 2018. https://medium.com/@ODSC/assessment-metrics-for-clustering-algorithms-4a902e00d92d.
35. Morgan, D. *Scoring metrics reference* https://code.kx.com/q/ml/toolkit/clustering/score/.

36. Wang, X. & Xu, Y. *An improved index for clustering validation based on silhouette index and Calinski-Harabasz index* in *IOP Conference Series: Materials Science and Engineering* **569** (2019), 052024.

37. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).