1-2023

Automated Loan Approval System for Banks

Abdulrahman Saeed Almheiri
asa6920@rit.edu

Automated loan approval system for banks

by

Abdulrahman Saeed Almheiri

A Capstone Submitted in Partial Fulfillment of the Requirements for the Degree

of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology, Dubai

2023 – January

RIT

Master of Science in Professional Studies:

Data Analytics

Student Name**: Abdulrahman Saeed Almheiri**

Cohort: 7

Graduate Capstone Title**: Automated loan approval system for banks**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date: 8/2/2023**

Chair of committee

Name: Dr. Hammou Messatfa **Date: 8/2/2023**

Member of committe

Acknowledgements

I want to sincerely thank Drs. Sanjay Modak and Hammou Messatfa for their contributions to the completion of my project, "Automated Loan Approval System for Banks."

I want to give a particular thank you to our mentor Dr. Hammou Messatfa for all his help over the last year. Your insightful counsel and recommendations were quite beneficial to me as I finished the assignment. I will always be grateful to you for this.

I would like to acknowledge that this project was completed entirely by me and not by someone else.

Table of Contents

Abstract

Machine learning and deep learning have revolutionized industries. Businesses have seen a boost in revenue through the use of these techniques, including the multibillion-dollar loan and credit industry. This study aims to develop a machine learning model to approve or disapprove loan applications based on customer characteristics, using algorithms such as statistical models, tree-based models, boosting models, and voting/stacking models. An exploratory analysis was conducted to understand customer characteristics, and the decision tree model proved to have the best performance with high precision in loan approval prediction.

Keywords: loan approval, machine learning, data analytics, financial institution, banks, automated system

Chapter 1 Introduction

1.1 Background of the Problem

The interconnected network of networks, including the internet, has created a highly connected world where services can be accessed and delivered online in real time. This has greatly impacted our daily lives, making internet an integral part of the modern society. The internet has also brought about the extinction of some technologies and the advancement of others, and opened up new possibilities for technology, information sharing, and information manipulation.

One of the most crucial needs in the modern world is loans. Banks play a significant role in the loan market, generating a significant portion of their earnings from this. Loans allow people to purchase luxury items such as homes and vehicles, and also help students manage their educational and living expenses.

The rise of technology has also led to the automation of many jobs, including the loan approval and management system. Historically, loan approval required input, effort, and resources from both the applicant, lender, and manager. However, with the help of machine learning algorithms, a more efficient and autonomous loan management system can be created.

The loan approval process involves evaluating an applicant's background data, including credit score, loan amount, lifestyle, career, and assets, to determine if a loan should be approved. Machine learning algorithms can be utilized to create a data-driven solution that predicts the loan status of a new applicant using comparable criteria. This study aims to develop an autonomous loan management system that can approve or disapprove loan applications efficiently and accurately.

1.2 Business Understanding

Identification of the opportunity in the industry to higher profits

The loan and credit approval industry is a multi-billion dollar business. Processing loan applications is a complex and time-consuming task for banking institutions, as individuals and businesses alike seek financial support for their ventures. In 2022, over 20 million Americans had open loans, totaling 178 billion USD in debt. Despite this, over 20% of loan applications were

rejected. The approval or denial of a loan has significant consequences for both the applicant and the bank, potentially causing missed opportunities for both parties. Thus, efficient and accurate loan processing is crucial for maximizing profits in this industry. The primary challenge in loan approval is identifying the right candidates for loans. Machine learning provides a valuable advantage in this regard, enabling the identification of suitable loan applicants. (Statista, 2022)

Efficiency in giving the state of the loan application to the customer

When a person seeks a loan, they are often in urgent need of funds. If a loan company takes too long to decide whether to approve the loan, the customer may turn to a more efficient competitor. Time is a critical factor in this industry, as there are many competitors offering loans. Even if the interest rates are high, customers will choose a loan company that processes their application quickly. By using machine learning to predict loan approval, the time needed for manual tasks can be reduced, making the loan company more competitive and attractive to potential customers.

How much to lend and under what conditions

Determining the appropriate loan amount and terms poses another challenge in loan approval. Borrowers' income and credit scores are often used to make this decision. There are several factors that might influence the loan terms and amount, such as the purpose of the loan (e.g. buying a car vs starting a business) and the borrower's employment history. By leveraging machine learning, the selection of loan terms and amount can be automated. Machine learning algorithms can analyze data from past loan applications to identify patterns that indicate the likelihood of loan default, allowing for more informed and efficient loan decisions.

1.3 Statement of the Problem

This research is being conducted to propose and develop an autonomous smart loan approval system which will approve or disapprove loan applications using machine learning techniques. Pre-approvals are also made simpler as a result of this streamlined procedure. Customer information can be simply entered in for pre-approval opportunities by having an automated system. This is especially true for repeat consumers whose card information has already been provided for earlier offers.

In recent years, many have tried to build loan approval systems but still there are many ways this process can be improved. This study will address those issues and will be able to overcome that. This binary classification system can be used to cater the fully functional autonomous loan approval system.

1.4 Project Goals

This study aims to develop a machine learning model to predict loan approvals for clients. The integration of machine learning in this system enables it to process parameters and produce results independently. A variety of classification techniques will be evaluated to identify the optimal method that minimizes overfitting and maximizes performance.

The study will also analyze the key factors affecting loan approvals and their interrelated relationships with the response variable. A comparative analysis will be conducted to assess the difference between this proposed solution and previous methods, and to identify the most significant features and correlations among the variables in the study.ine learning is that

1.5 Limitations

This study will implement a Smart Loan Approval system which will be put into practice in two different phases and requires different algorithms. As it is a complex piece of software it comes with a lot of constraints, some of which are as follows:

1. The study focuses on mainly one type of loans and if this needs to be generalized need to get more attributes and data
2. The main challenge with the current approach is that it assigns equal importance to all factors in determining loan approval. However, in reality, loan approval can sometimes be based on a single dominant factor, which cannot be captured using this approach.

Chapter 2 Literature Review

Machine learning applications are being used in practice more and more frequently. However, due to low data quality, many ML initiatives in production fail. Data has to be preprocessed to improve quality. There are hundreds of data preparation techniques, which are manually chosen based on use-case requirements. Due to these factors, data preprocessing presently takes 80% of ML-projects' time and is executed in an unstructured manner. So, by comparing discovered data preprocessing methods to the performance of ML-models on five data sets, (Frye et al, 2021) offered a structured data preprocessing approach in which data preprocessing methods are recommended depending on production use-case requirements. In order to recognize how sensitive certain data points' results are to different preprocessing methods and determine whether a preprocessing step is causing bias to grow or shrink for particularly weak population groups, (Zelaya, 2019) conducted research to examine the way data preprocessed for ML processing has noticeable effects, which can be measured and analyzed.

The loan approval process is crucial for financial companies, especially in today's market where a large number of individuals and businesses are seeking financial backing. However, banks are limited by their available assets and therefore must carefully evaluate applicants in order to minimize risk and make the best use of their resources. To address this challenge, researchers have attempted to develop more accurate predictive models to streamline the loan approval process and reduce the risk involved in lending decisions.

The study by Asare-Frempong et al, 2017, aimed to analyze the effectiveness of using different classifiers for predicting customer responses to a bank's direct marketing campaign. The study used a dataset from a typical bank and evaluated four classifiers - Multilayer Perceptron Neural Network, Decision Tree, Logistic Regression, and Random Forest. The findings showed that the Random Forest Classifier was the most effective with an accuracy rate of 87% as evaluated using classification accuracy and ROC. The second goal of the study was to identify the key characteristics of clients who had already subscribed and were most likely to do so again for term deposits. This was achieved through a cluster analysis.

(Aslam et al, 2019) analyzed the three most commonly used and reliable machine learning algorithms and neural networks for predicting loan defaults. The study discussed the benefits and limitations of using specific models, providing valuable insights for others in the field. The authors noted that while a few studies have focused on the impact of false negatives, which can be highly detrimental to lending organizations, most studies have primarily focused on accuracy in predicting loan default. As a result, the authors recommended that future research in this area should place greater emphasis on reducing false negatives in loan lending predictions.

Banks in the financial system have various alternatives to generate income, but the main source of revenue is from their credit lines and the interest on the loans they provide. Predicting loan defaults and lowering Non-performing Assets is crucial for the bank's profitability. Several studies have been conducted to forecast loan approvals, using various machine learning models including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Neural Networks, AdaBoost, and K-Nearest Neighbors (KNN). The results from these studies show that the accuracy of the models varies, with Decision Tree often emerging as the most accurate.

In (J. Tejaswini et al., 2020) DT was found to be more accurate than LR and RF, while (K. Arun et al, 2016) reported higher accuracy using DT and RF. (M. A. Sheikh et al, 2020) reported an accuracy of 0.811, but didn't mention the specific model that achieved this result. (Hemachandran et al, 2022) used various classification algorithms and found the best performance with KNN, which was evaluated using metrics such as Jaccard index, F1-score, and Log Loss. (Dosalwar et al, 2021) obtained the highest accuracy of 0.78 using Logistic Regression among the seven machine learning models they tested.

A prediction model based on the logistic regression technique was used by Shinde et al (2022) in their research. Over 600 sample data were gathered and assessed to develop a logistic classification model that predicts loan status, achieving an accuracy rate of 82%. The study found that credit history and balanced income were the most important features.

In another study by Aphale & Shinde (2020), real bank credit data was analyzed and numerous machine learning algorithms were applied to the data, including Linear regression, Naive Bayes, and KNN. These algorithms all had accuracy rates ranging from 76% to over 80%.

Meshref (2020) used various ensemble machine learning approaches, including Bagging and Boosting, in their research on loan approval prediction. Their findings indicated that the loan approval prediction model had an accuracy of 83.97% for the AdaBoost model.

Abakarim et al (2018) suggest a Real-Time Binary classification model for loan approval in their study. Their methodology, based on a deep neural network, enables classification of loan applicants as either excellent or bad risks. According to experimental findings, the Real-Time model based on deep neural networks performs better in terms of precision, recall, and accuracy than typical binary classifiers.

Eletter et al (2010) aimed to construct a proposed model that recognizes artificial neural networks as a tool that may be used by Jordanian commercial banks to evaluate credit applications and assist loan decisions. The proposed model was constructed using a multi-layer feed-forward neural network and a backpropagation learning technique, achieving 95% accuracy for the test set. In a comparison study by Kar Yan Tam & Melody (1992), the neural network approach was contrasted with Linear regression, Logistic Regression, KNN, and ID3 using bank default data. The research found that neural networks are a potential tool for assessing bank conditions due to their prediction accuracy, flexibility, and durability, but also discussed the drawbacks of employing neural networks as a broad modeling tool.

The main issue with imbalanced datasets is that machine learning algorithms have a tendency to favor the majority classes and disregard the minority classes. Islam et al (2019) suggest the SMOTE technique to overcome the imbalance dataset for forecasting bank telemarketing success, obtaining an accuracy rate of 88.86% for the Gaussian Naive Bayes algorithm. The technique of stacking allows for the combination of different regression or classification models. Bagging enables the averaging of several comparable models to reduce variance, while Boosting creates numerous incremental models to reduce bias while minimizing variance. In the research by B. Pavlyuchenko (2018), the stacking strategy was investigated for the construction of machine learning model ensembles, considering cases for logistic regression and time series forecasting.

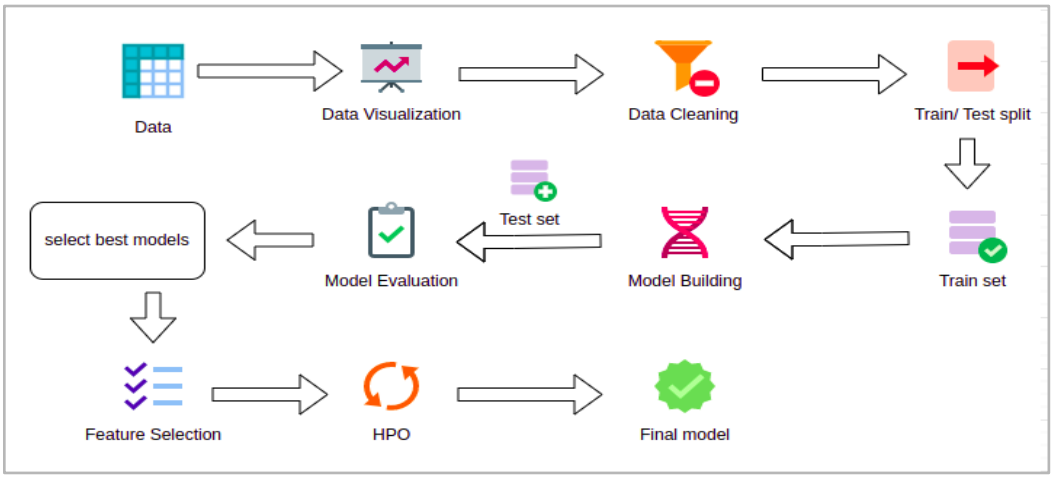
2.1 Key takeaways can be listed as follows

- Frye et al, 2021, offered a structured data preprocessing approach that recommends data preprocessing methods based on production use-case requirements.
- Zelaya, 2019, conducted research to examine the noticeable effects of data preprocessing for ML processing, which can be measured and analyzed to determine if preprocessing is causing bias.
- Asare-Frempong et al, 2017, found that the Random Forest Classifier was the most effective for predicting customer responses to a bank's direct marketing campaign with an accuracy rate of 87%.
- Aslam et al, 2019, analyzed the three most commonly used ML algorithms and neural networks for predicting loan defaults and recommended future research to place greater emphasis on reducing false negatives.
- The accuracy of loan approval prediction models varies among studies, with Decision Tree, Random Forest, and Logistic Regression being the most commonly used models. KNN and AdaBoost were also used in some studies.
- Shinde et al, 2022, found credit history and balanced income to be the most important features in a logistic classification model that predicts loan status with an accuracy rate of 82%.
- Aphale & Shinde, 2020, applied Linear Regression, Naive Bayes, and KNN to real bank credit data and found accuracy rates ranging from 76% to over 80%.
- Meshref, 2020, used ensemble ML approaches (Bagging and Boosting) and found an accuracy of 83.97% for the AdaBoost model in loan approval prediction.
- Abakarim et al, 2018, proposed a Real-Time Binary classification model based on a deep neural network that performs better in terms of precision, recall, and accuracy than typical binary classifiers.
- Eletter et al, 2010, proposed an artificial neural network model that recognizes credit applications and assists loan decisions, achieving 95% accuracy for the test set.

Chapter 3 Methodology

3.1 Research Process

Data Analytics Model in loan approval prediction



The final goal of this project is to automate the loan approval prediction process completely. The dataset needs to be analyzed comprehensively. The research process of this project is shown by the diagram below. Model based feature selection will be used from the best models that were chosen from the initial model evaluation process.

3.2 Environment

In this project, we have used SPSS Statistics and SPSS Modeler to determine key insights and perform different modeling steps with the given datasets. We have used SPSS Statistics to perform Exploratory Data Analysis and then SPSS Modeler to perform the machine learning modeling steps. The Statistics tool has helped us derive different insights through visualization and data statistics summary as well.

3.3 CRISP-DM Methodology Process

The CRISP-DM is a process model that outlines a data science process, consisting of six sequential phases. They are shown in figure 3.4.

Business Understanding

In this stage of the process, the industry is thoroughly understood and relevant scenarios regarding loan processing and application in the banking industry are studied to acquire a comprehensive understanding of the topic and approach the problem with well-informed solution strategies..

Data Understanding

In this stage, the dataset is obtained and subjected to exploratory data analysis and statistical modeling to gain deeper insight into the data. The initial step involves exploring and gaining an overview of the dataset, including evaluating the mean values and standard deviation of continuous features.

Data Preparation

After an overview exploration of the dataset, it is necessary to clean it of any missing values. This step involves identifying missing values and data types for all features, which helps ensure the accuracy of the analyses. For instance, if the response column contains missing values due to typing errors, it can result in biased results, so cleaning the data of these issues is crucial.

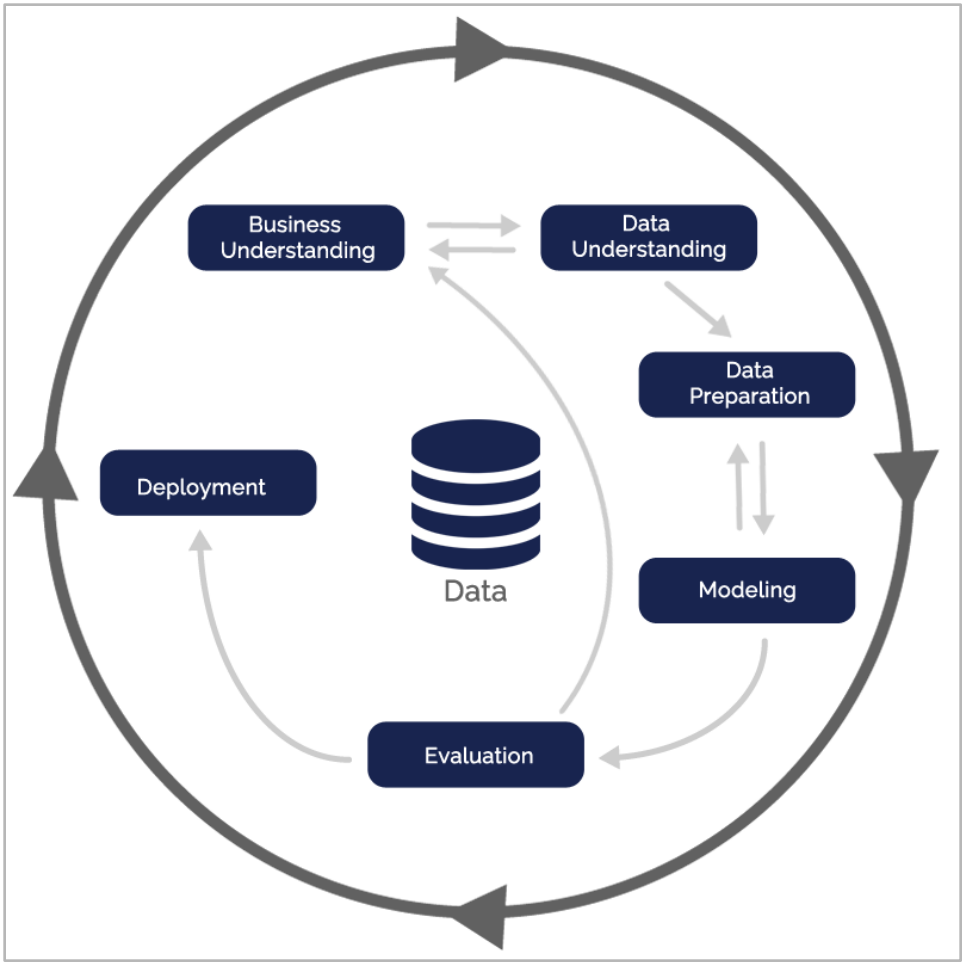
Data Modelling

After cleaning the dataset, it is used to build a machine learning model and train the data. This results in the ability to predict outcome values in the final step. Understanding the features in the dataset is crucial for preparing the train and test sets, which allow for training and testing the accuracy of the clustering results.

Evaluation

The final step involves validating the developed model and evaluating the quality of the formed clusters. It's important to assess if the values in the clusters are meaningful and aligned with expectations, to avoid any misinterpretation of the results.

Adhering to the sequential steps outlined in the CRISP-DM methodology provides a systematic approach to solving the problem and yields accurate outcomes for stakeholders and recommendations. Such a process assures a well-planned project and is aimed to achieve the desired output..



In this project, the steps of the CRISP-DM process were followed to analyze and determine the best model for loan processing and application in the banking industry. The dataset was obtained from Kaggle and then prepped for analysis and machine learning by removing missing values and outliers. Finally, the model performance was evaluated using metrics such as accuracy, gain, and AUC, and the best-fit model was selected for the prediction of loan applications based on creditworthiness and other factors.

3.4 Classification Methods

A classification is a data-mining method that divides a group of data into categories to help with analysis and prediction. Classification models will make an effort to predict the value of one or more outputs for a given one or more inputs. The results of this study are labels of Loan status 1 or 0. Here, Accept or reject loan applications using machine learning models.

3.4.1 Machine Learning models

Supervised and unsupervised learning are the two main categories of machine learning.. Supervised learning enhances algorithms' ability to recognize data or accurately predict results using labeled datasets. The concept of supervised learning is shown in Figure 3.3.

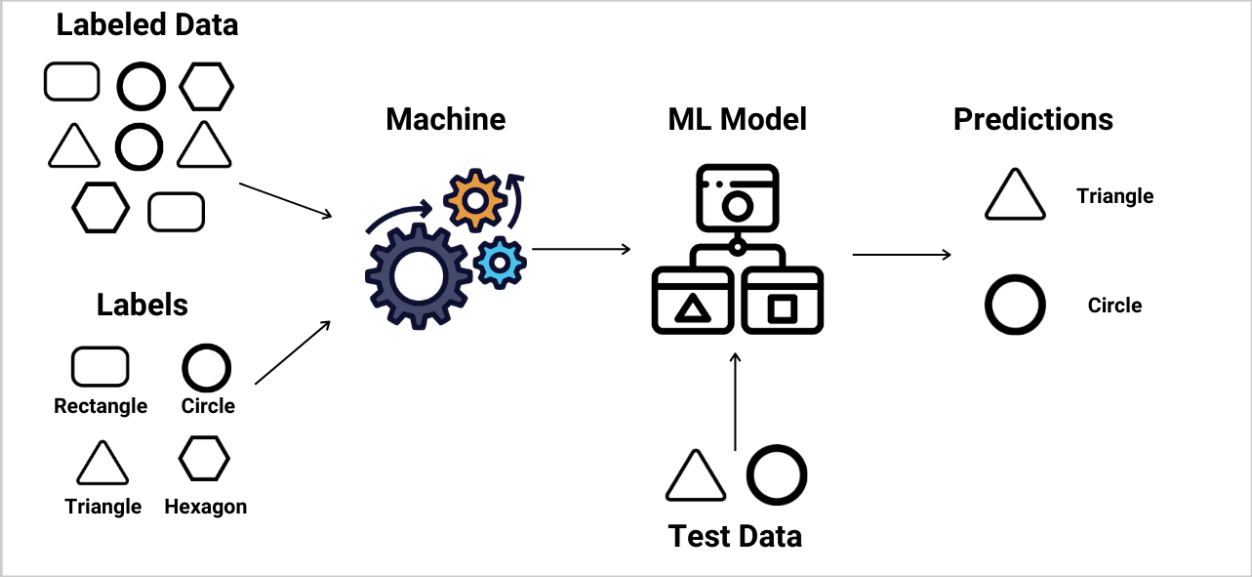


Figure 3.3 - Supervised Learning Process (Raj, 2021)

When there is only input data and no related output variables or labels, unsupervised learning is necessary. Figure 3.4 shows how unsupervised learning is conceptualized.

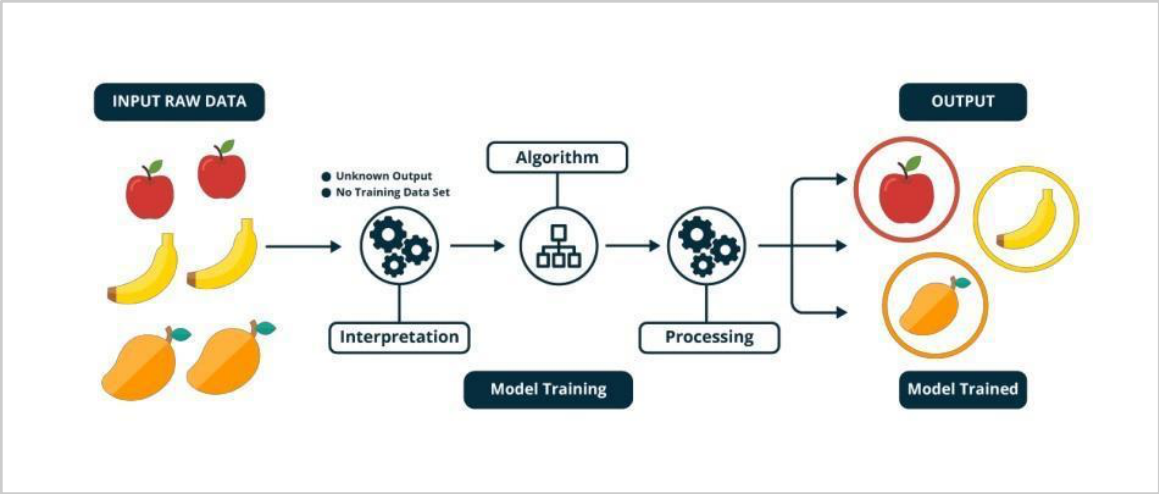


Figure 3.4 - Unsupervised Learning process (Damron, 2018)

In this study, the model was trained using labeled datasets. As a result, supervised learning models were employed as opposed to unsupervised learning models. Support Vector Machine, Multinomial Naive Bayes, Decision Tree, Random Forest, Logistic Regression, and XGBoost are a few examples of supervised learning classification models. These models on approve or disapprove loan applications are implemented using the open-source Python sklearn framework.

Logistic Regression.

A commonly employed and basic approach for binary classification in Machine Learning is Logistic Regression. The relationship between one dependent binary variable and independent variables is described and approximated using logistic regression. This type of linear regression specifically occurs when the target variable has a categorical structure. (Analytics Vidhya, 2015) The logistic sigmoid function uses the results of logistic regression to provide a probability value that may be applied to two or more discrete classes. The sigmoid function can change any real number into a value between 0 and 1.

Linear regression equation:

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \quad \text{..... (1)}$$

Sigmoid Function:

$$p = \frac{1}{1+e^{-y}} \quad \text{..... (2)}$$

Apply Sigmoid function on linear regression:

$$p = \frac{1}{1+e^{-\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}} \quad \text{..... (3)}$$

Logistic Regression Equation:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \quad \text{..... (4)}$$

Where, y : Response variable

p : Probability of response class

βi : Regression coefficients

Xn : Predictor variables

The values of these coefficients are estimated using the maximum likelihood technique. The dependent variable for binary logistic regression should be dichotomous, the data should be independent, and there shouldn't be any outliers.

When the assumption of conditional independence is true, this classifier performs better than some of the more intricate machine learning models. With categorical data rather than numerical data, it is performing comparably well. The scikit-learn package has a variety of Naive Bayes models with names like MultinomialNB, GaussianNB, ComplementNB, BernoulliNB, and CategoricalNB. Although Naive Bayes is an effective classifier, it is a poor estimator, which causes the model's probability outputs to be inaccurate. (Murthy, 2020)

This section is dedicated to explaining the models that we plan to use in the subsequent sections of the modeling phase. It is essential to understand a bit more about the candidate models along with their related performance KPIs in detail before we proceed with the practicalities of the task. In this experiment, we use quite a lot of models on SPSS Modeler like Random Trees, Logistic Regression, Decision Trees, Neural Network and C5. A brief explanation about each of the models are listed below.

Random Trees.

Random Forest, an ensemble learning approach, is a powerful method for solving both classification and regression problems. It utilizes a group of decision trees to make predictions and is known for its robust scoring techniques. This approach has been validated through numerous research studies and has demonstrated consistently good performance results. (Le Gall, 2005)

C5.

This uses either a decision tree or rule set to derive a predictive approach. This approach of splitting the tree again and again is followed in an iterative manner. (Ahmad, 2021). This modeling process is an extended version of the decision tree modeling process which adds a layer of rules to perform the process in a structured manner.

Neural Network 1.

Artificial Neural Networks (ANNs) are a type of machine learning technique modeled after the structure and function of the human brain. This approach utilizes a network of interconnected nodes, including input, hidden, and output layers, to process input features and make predictions based on their relationships. The ANN model is capable of making complex decisions by considering the interplay between the input nodes, hidden layers, and output nodes, ultimately providing results that are representative of the information passed through the network. (Dash, 1997)

3.5 Model Evaluation Matrices

Performance metrics are a part of every machine learning pipeline and they report on progress and assign a numerical value to it. Performance metrics also can be divided into regression and classification. These metrics are mainly used to evaluate the generalization ability of the model that has been trained and to select the best model (M & M.N, 2015).Since this study is based on a classification problem, the performance metrics used for classification are discussed here. Classification models provide discrete outputs. Hence, a measure that in some sense compares distinct classes is required. Classification metrics assess a model performance and tell whether the classification is good or bad, but each one does so in a unique way.

Confusion Matrix

A confusion matrix is a table that compares the actual classifications or labels of a dataset with the predictions made by a model. It is not a direct measure of model performance, but it provides a basis for other evaluation metrics that assess the accuracy of the model's predictions. (Kulkarni, 2020)

Each cell in the confusion matrix represents an evaluation factor.

TP (true positives) – Indicates how many positive class samples have been predicted correctly by the model.

FP (false positives) – Indicates how many negative class samples have been predicted as positive by the model.

FN (false negatives) – Indicates how many positive class samples have been predicted as negative by the model.

TN (true negatives) – Indicates how many negative class samples have been predicted correctly by the classifier.

Accuracy

The simplest metric to assess the accuracy of a classification model is the Classification Accuracy, which is calculated by dividing the number of correct predictions made by the model by the total number of predictions made. This statistic is straightforward to calculate and implement, but its performance is highly dependent on the distribution of classes in the dataset. It works best when the class distribution is evenly balanced. (M & M.N, 2015).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

AUC ROC

A Receiver Operating Characteristic (ROC) curve displays the True Positive rate vs. False Positive rate at various threshold settings. The area under the ROC curve, ROC AUC, is a single-valued

measure used to evaluate classifier performance. The greater the ROC AUC, the better the classifier. This metric is commonly used in imbalanced classification problems. (M & M.N, 2015).

As we learned above, there are various Machine Learning models that SPSS Modeler has to offer based on the requirements of the problem statement as well as the shape of the data. In our experiment, the tool will help us determine the best modeling approach in order to have the best

results for the final results and output. There are certain evaluation metrics that we will also use in this modeling phase like ROC, AUC, Gain etc. Some of the description about each of them are -

ROC (Receiver Operating Characteristic) is a graph used to show the performance of various classification models at different thresholds. The plot includes two main components: True Positive Rate (TP) and False Positive Rate (FP).

The Area Under the Curve (AUC) is a popular metric used to evaluate the performance of a binary classifier model. It provides an aggregated view of performance measures across all classification thresholds. Another key performance indicator is Gain, which measures the reduction in entropy that results from making a decision based on the input data. It is often used in the context of decision tree models, where the information gain of each variable is calculated to help determine the best split in the training set. (Sarang, 2018)

Chapter 4 Project Description

4.1 Dataset Description

This study uses a dataset of 5584 loan applicants from a private bank in Singapore. The dataset includes information such as the applicants' id, age, gender, marital status, education, employment, income, loan amount, loan term, and credit history. Each record in the dataset represents the details of a single applicant, with a total of 13 features. The dataset was obtained from Kaggle and serves as a binary classification problem, with a response variable consisting of two classes. Access the original loan approval dataset at the following Kaggle link: Partial Bank Loan Dataset | Kaggle

Attribute description: There are 13 attributes including the loan status and the details of the attributes are given below.

Table 4.1 - Description of the dataset

Variable	Type	Description
Loan_ID	Integer | Quantitative	Unique loan ID
Gender	String | Qualitative	Male/Female
Married	String | Qualitative	Marriage status (TRUE/FALSE)
Dependent_No	Integer | Qualitative	Number of dependent/s
Education	String | Qualitative	Education status (Graduate/Not Graduate)
Self_Employed	String | Qualitative	Employment status (TRUE/FALSE)
Applicant_Income	Integer | Quantitative	Amount of applicant's income

CoApplicant_Income	Integer | Quantitative	Amount of co applicant's income
Loan_Amount	Integer | Quantitative	Amount of loan requested
LoanAmountTerm	Integer | Quantitative	Term of loan in months
Credit_History	Boolean | Qualitative	Applicant's credit history (0 for no /1 for yes)
Property_District	String | Qualitative	Applicant's district of property owned
Loan_Status	Boolean | Qualitative	Loan approval status (0 for no /1 for yes)

Univariate Statistics

	N	Mean	Std. Deviation	Missing Count	Missing Percent	No. of [a]... Low
Loan_ID	5584	25793.50	1612.106	0	.0	0
Applicant_Income	5584	50658.55	28580.614	0	.0	0
CoApplicant_Income	4999	25366.45	14181.280	585	10.5	0
Loan_Amount	5584	35333.09	17147.360	0	.0	0
Loan_Amount_Term	5584	212.64	102.964	0	.0	0
Self_Employed	5584			0	.0	
Property_district	5578			6	.1	
Education	5218			366	6.6	
Dependant	5584			0	.0	
Married	5328			256	4.6	
Gender	5425			159	2.8	
Loan_Status	5584			0	.0	
CreditHistory	5584			0	.0	
Dependent_No	5584			0	.0	

Univariate Statistics

	No. of [a]... High
Loan_ID	0
Applicant_Income	0
CoApplicant_Income	0
Loan_Amount	0
Loan_Amount_Term	0
Self_Employed	
Property_district	
Education	
Dependant	
Married	
Gender	
Loan_Status	
CreditHistory	
Dependent_No	

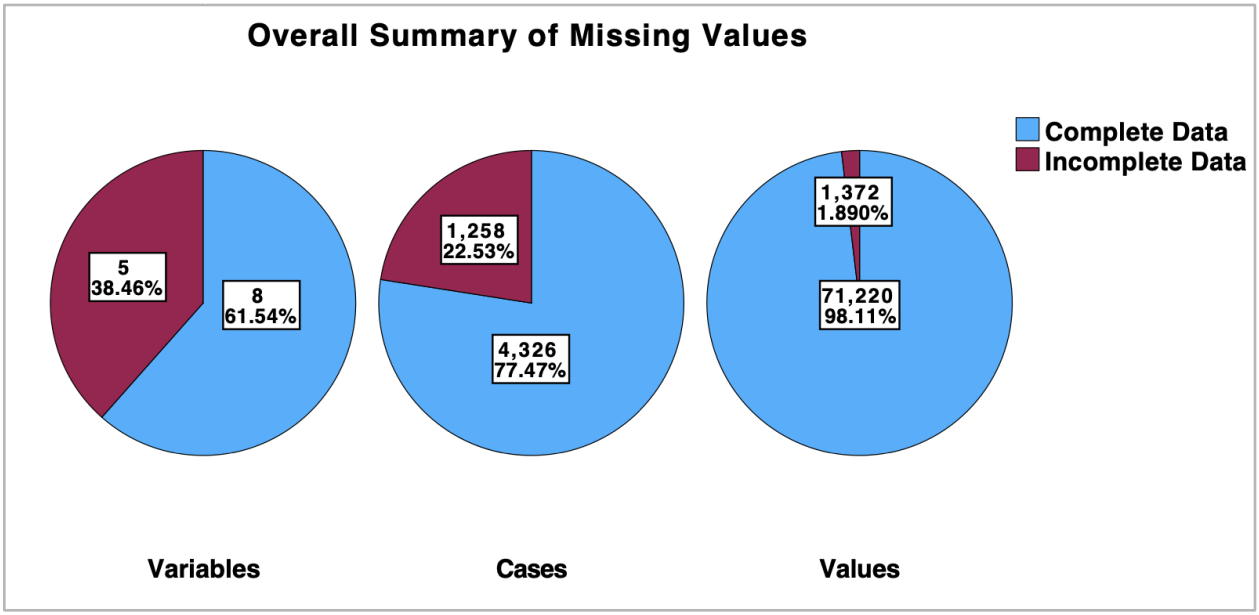
a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

The above dataset description helps us understand the summary statistics of all the features available in the data. We have 5584 observations with 13 features to explain a loan application. The average application income is 50,658 units while the loan amount applied for is on an average 35,333 units as per the data description shown above. On an average, customers opt for 212 units shown above.

4.2 Exploratory Data Analysis

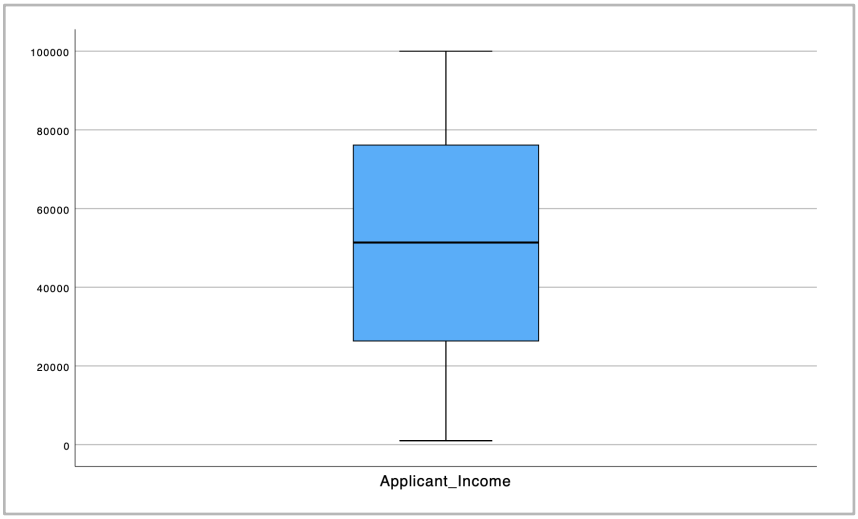
This section is dedicated to understand the data better through visualizations as well as determining patterns for effective data cleaning steps. This will help us derive a standardized dataset to be able to derive optimal results.

In the given dataset, we have cases of data inconsistencies and we would like to visualize the same and then be able to clean to have better data.

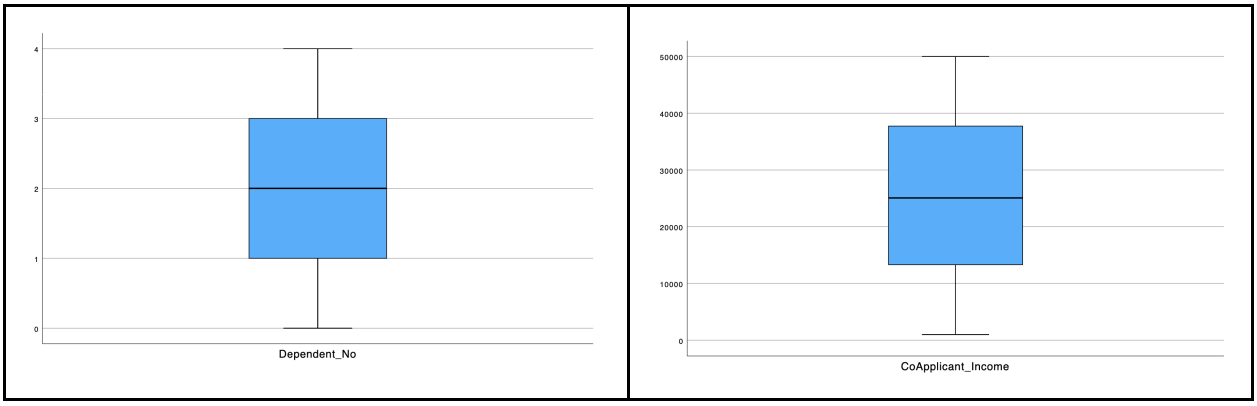


In the above section we see that we have 38% of the features that have some or the other form of missing values in them. On the other hand we have 22% cases within the data that are missing also which can be filled with different processes like forward/backward filling, missing value imputation with processes like mean, median or mode imputation. The last pie chart is the %

missing values for all values in the data and we see that only ~2% of them are missing while the majority is available to us.

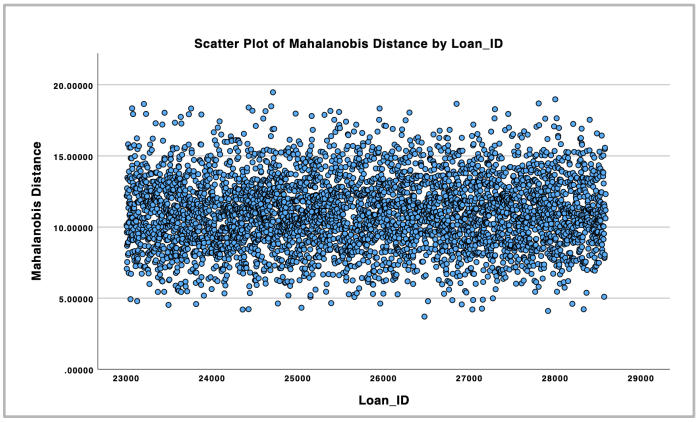


In the above section, we will use some univariate and multivariate analysis to understand the different features within the dataset. (Vedansh, 2022) SPSS Statistics would be used as a means to explore the dataset and determine some key insights from the same. In the above figure, we have the box plot for the applicant income distribution and observe that on an average, the income is 50,000 units with the upper and lower bounds for the customers in terms of their income.



We see from the above distributions for number of dependents and the co applicant incomes and understand that the average number of dependents for the loan applicants is 2. On the other hand, the average income of the co applicants is somewhere around 24,000 units which is ~50% lesser than the applicant income. This makes sense because the co applicant might be a partner or spouse

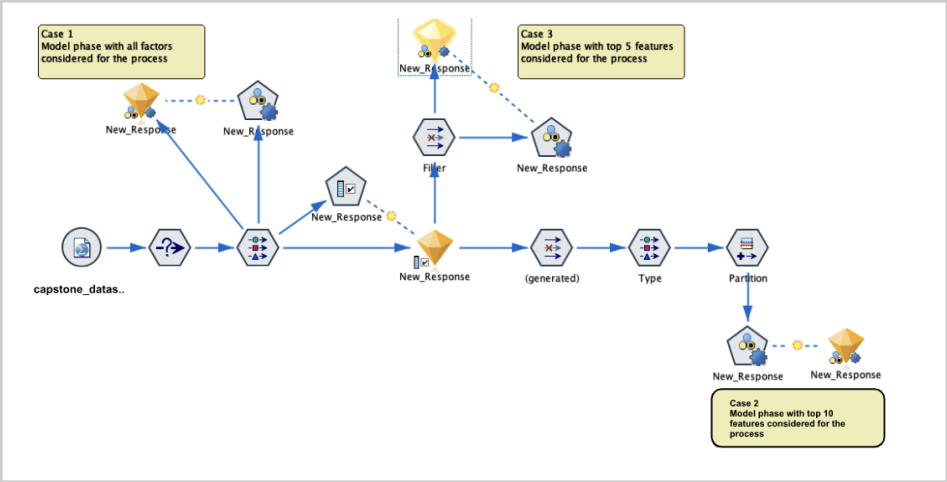
with a lower income band compared to the applicant who might act as a supporting side of the entire transaction.



From the above chart we are able to interpret the Mahalanobis Distance of the data points in our selected dataset. The Mahalanobis distance is used to determine the outlier values which may not be in line with the data standards. These outlier values might skew the output results of the visualizations as well as the machine learning model implementations. Hence, we use Mahalanobis Distance along with p-value factor reduction steps to remove unwanted values from the data to standardize the same. (Stephanie, 2021)

4.4 Modeling

This section of the report is assigned to the implementation of statistical analyses and modeling techniques to solve the problem statement. The cleaned dataset will be used in SPSS Modeler for modeling. The modeling phase involves determining important factors from the dataset, building machine learning models to predict loan applicant status, and selecting the best model based on its accuracy.



In the above provided model steps, we first import the cleaned dataset along with the trimmed features from the data. Then we perform the modeling based on different scenarios in the above modeling phases. We consider three different scenarios while implementing the modeling phase. The first scenario would be considering all the features in the dataset to perform a holistic modeling approach and to create a benchmark for the model scoring performances. In the second scenario, the top 10 features from the dataset will be selected and their impact on the model scores will be analyzed. Similarly, in the third scenario, only the top 5 features will be considered and the results of the model scores will be evaluated. As shown in the modeling phase above in the figure, there are three different networks/branches that we use to showcase the different scenarios that we plan to perform our modeling process. The first one takes into consideration all the factors that we have in the dataset to build a set of models and compare their performance KPIs (marked as Case 1). In this particular scenario, we obtain different model values and performance metrics that tell us which model is the best out of all. We then use a different set of features to perform another scenario of the modeling process as shown in Case 2, with only top 10 features out of all and then again we compare the model performances to pick the best result out of all. (Prasad, 2019) This step is again repeated in the final scenario (as shown as Case 3) with only 5 features to understand the modeling metrics and then we compare the model performance metrics. We provide the detailed analysis of each scenario in the below section based on different scenario and modeling phases. To summarize, below are the steps that we plan to perform -

- Scenario 1: In this case, we use all the features that are contained in the dataset to build our models in which we primarily use Random Trees, C5, Logistic Regression and Neural Networks
- Scenario 2: Here, we used feature selection method to use only top 7 features based on relevance and the models that we use here are C5, SVM, Neural Networks and Logistic Regression
- Scenario 3: In this case, we used the Random Trees and the top 5 features were determined based on feature importance scoring. The models used in this case are C5, Neural Networks, Logistic Regression and Decision List.

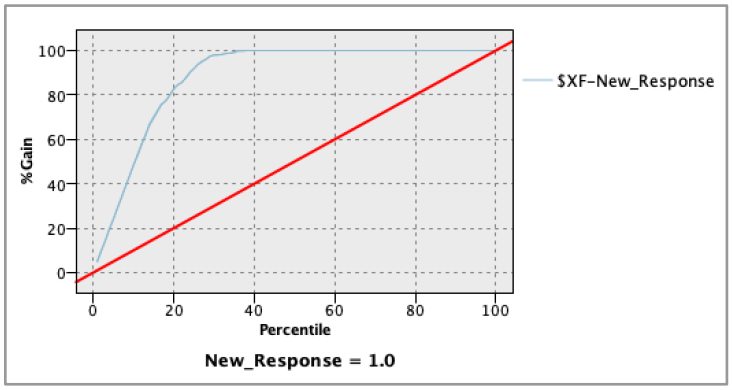
Scenario 1

In this case, we used some models like Random Trees, C5, Logistic Regression and Neural Networks to determine the accuracy scores of the models and perform a comparative study of the classification capability. Below are the findings and the table with their respective scores.

Model Graph Summary Settings Annotations

Sort by: Use Ascending Descending Delete Unused Models View: Training set

Use?	Graph	Model	Build Time (mins)	No. Fields Used	Overall Accuracy (%)	Accumulated Accuracy (%)
☑		Random Trees 1	< 1	13	92.420	93.814
☑		C5 1	< 1	7	87.192	86.245
☑		Logistic regression 1	< 1	4	79.317	78.015
☑		Neural Net 1	< 1	9	81.921	83.107

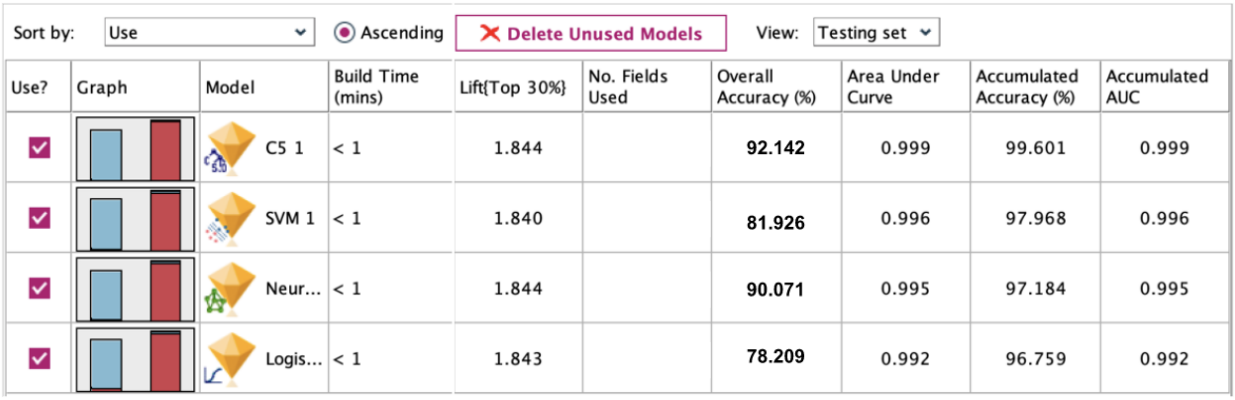


In the above results matrix, we can see that the best performing model is Random Trees with an overall accuracy of 92% while the second in the category is for C5 with 87% overall accuracy.

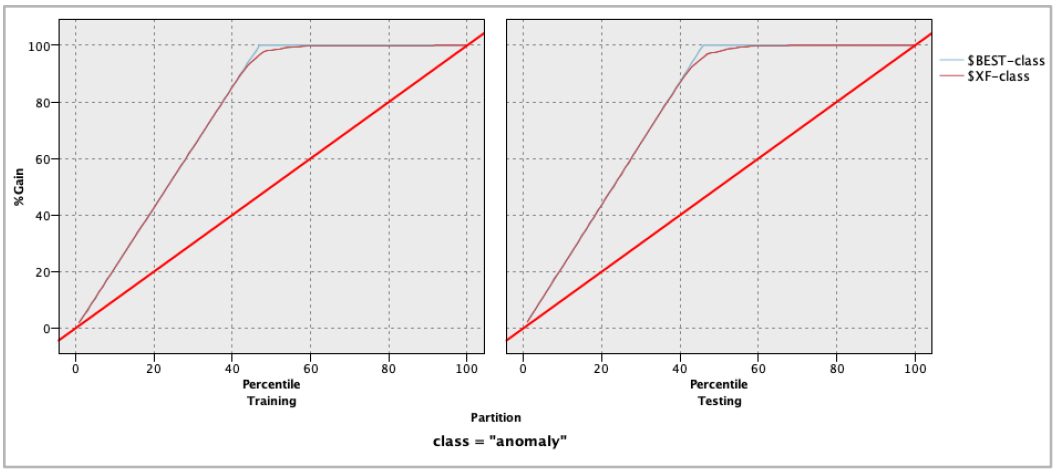
Similarly, in the list we have Neural Network and Logistic Regression in the other ranks based on the overall accuracy. In this case, we chose all the features and saw that Random Trees recommended all the features with a good Accuracy score. This case has chosen all features because we want to create a baseline of the model performances using all the fields in the dataset. We have seen that the gain % for both the train and test data is good, as the gain % peaks towards 95 and this proves that the model is a good performing model.

Scenario 2

In the next case we are going to use some similar approach as the previous one but will only move ahead with the top 7 features from the entire dataset by using feature selection method. After creating the baseline modeling process in Scenario 1, it is imperative that we create a feature selection process using models and pick the most relevant features out of all to improve the model accuracy and scores. We will again perform a comparative study of all the model performances and see which is the best performing model of all.



Use?	Graph	Model	Build Time (mins)	Lift(Top 30%)	No. Fields Used	Overall Accuracy (%)	Area Under Curve	Accumulated Accuracy (%)	Accumulated AUC
☑		C5 1	< 1	1.844		**92.142**	0.999	99.601	0.999
☑		SVM 1	< 1	1.840		**81.926**	0.996	97.968	0.996
☑		Neur...	< 1	1.844		**90.071**	0.995	97.184	0.995
☑		Logis...	< 1	1.843		**78.209**	0.992	96.759	0.992

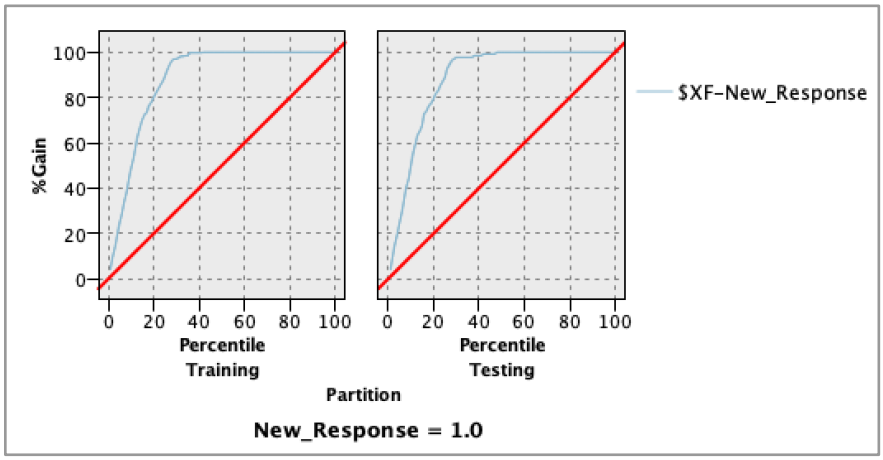


In the above result from the modeling we can see that C5 performs the best with an overall accuracy of 92% and second is the Neural Network model with 90% accuracy. Based on the feature importance scoring process, we determined that Credit History, Loan Amount Term, Loan Amount, Applicant Income, Co Applicant Income, Gender and Education (Enes, 2022) were the good features to be selected for this scenario of the modeling phase. In this case we observe that C5 performed the best out of all the models. The gain percentage of the model in this case is 92, which is considered a good score to use its outputs for further evaluation.

Scenario 3

In this case, we will consider the top 5 features from the dataset based on their performance scores and then check if the results differ in terms of the performance metrics. This is because we want to have the relevant features from the entire dataset, because the higher the relevance score for the features, the better the model is able to explain the results and the accuracy improves. In the feature selection process using the modeling technique, it was shown that the following features had higher relevance score as per the Random Trees modeling. This will also help us choose one of the robust models in terms of their accuracy scores and then present the same. In this process of the modeling and scenario chosen in SPSS Modeler, Credit History, Loan Amount, Loan Amount Term, Applicant Income, Education played a key role in building the model and hence we chose the same to determine the model accuracy eventually. The model results and performance have been described in the below figures in-depth for better understanding.

Use?	Graph	Model	Build Time (mins)	No. Fields Used	Overall Accuracy (%)	Area Under Curve	Accumulated Accuracy (%)	Accumulated AUC
☑		C5 1	< 1	5	93.035	0.995	97.078	0.995
☑		Neural Net 1	< 1	4	85.276	0.989	95.51	0.989
☑		Logistic regression 1	< 1	4	88.102	0.949	88.868	0.949
☑		Decision List 1	< 1	5	63.091	0.81	33.595	0.81



In the above case of the model implementation, we again notice that the best performing model here is C5 with an overall accuracy of 93% while second in line is Logistic Regression with accuracy of 88%. This shows that C5 performed well in two cases of the model implementation which helps us choose the best fit model for our use case.

4.5 Evaluation

Here we would like to summarize the study of the different models and combinations that we performed above based on which we will choose the good performing model for our process. The different KPIs are considered in the table below to perform a comparative analysis of all the models and the scenarios considered (while also considering the number of features for each of the models).

Case #	Algorithm	Features	Accuracy	Gain %
Scenario 1	Random Trees	13	92.4%	95
Scenario 2	C5	7	92%	92

Scenario 3	C5	5	93.1%	93

It has been observed that C5 performed with the best score in two scenarios, i.e., with 92% and 93% accuracy respectively. The set of features that are used in the modeling phase has a great contributing factor in determining the model dependencies and performance valuation scores in the end. There are different methods of feature analysis, selection and transformation which can help us improve model scores further.

Chapter 5 General Discussion and Conclusion

5.1 Discussion

The process of loan approval is essential for financial institutions. Banks are in need of an improved predictive modeling system due to various challenges. With the growth of the banking industry, a large number of loan applications are submitted, but banks have limited assets to lend, so they can only approve a limited number of them. To ensure safety, banks must carefully choose who to lend to. This study aims to minimize the risk associated with loan approval and help banks save time and resources by selecting safe loan applicants.

The dataset used for this study includes details on 5584 people who have applied for bank loans. They have to provide information on their id, age, gender, marital status, number of dependents, education level, employment history, income, loan request, loan term in months, and credit history. A single record contains the details of 13 attributes for a single application. The Kaggle Datasets are where this loan dataset was taken. Several scenarios of modeling were considered to derive a robust modeling approach. We were able to build models using different combinations of the feature sets in order to arrive at the optimal result. This helps us ensure as a validation step that the process that we follow is accurate and is not biased towards any partial information. We observe that different scenarios provided different results and performances based on the feature combination as well as the model chosen. This helped us ensure that we provided the right result for our model phase.

5.2 Conclusion

The main research question is to develop an autonomous smart loan approval system which will approve or disapprove loan applications using machine learning techniques. The machine learning models with different levels of complexity had been used for the study. This range includes linear models, tree-based models, boosting models, voting models, and stacking models. The correlations among predictor variables and the response variable are very low. This has led to a decrease in performance overall. There were some significant patterns in the dataset although the correlation is low. This was able to be identified from the thorough explanatory analysis process. There were no outliers and most of the numerical variables have been distributed normally. The imbalance of the dataset also had affected the overall performance as seen in Chapter 4. This problem was overcome by applying resampling techniques. Although there are undersampling and oversampling techniques to balance the dataset. This study used the SMOTE oversampling method to reduce the loss of information. This oversampling technique had better performances and that can be clearly identified from the confusion matrices and classification reports in chapter 5. We have made the following observations in the given project -

- Based on the loan approval data provided, we were able to derive different insights from the data at customer level and also the distribution of various customer properties
- We performed data cleaning procedure to arrive at the optimal data format to be able to process and model better
- There are different risks associated with different cohorts of customers based on various properties like their income group, their co applicant income, their loan history, number of credit or loans taken and many other demographic factors. These different properties help us understand the credibility of the customer which we can use to decide the loan application status of the customer
- We implemented different machine learning modeling techniques using SPSS modeler to arrive at an optimal model selection which were based on different performance evaluation metrics like gain, AUC, ROC etc.

The purpose of this research is to determine the key factors that impact loan approval, examine the relationships between these predictor variables and the loan approval outcome, and build a

machine learning model to predict loan approval for clients. This will benefit both customers and the company by providing insights before the model is used. The study uses the Cross Industry Standard Process for Data Mining process model to identify business values and the overall research methodology. The core value and details about the methodology has been described in chapter 3. The theory behind machine learning models and other metrics that can be used in this classification problem have been covered in chapter 3. The dataset description, exploratory data analysis, and data pre-processing steps have been discussed in chapter 4.

According to the objectives and the business value, identifying the correct clients to give loans is the most important thing. Hence the final model should have low false positive predictions. If not, the model would predict some clients who are not actually eligible for the loan amount, which will lead to a costly outcome for the company. Hence the final model was the model with low false positives. Also, a comprehensive feature selection was done to reduce the complexity of the model. This showed around 2% accuracy improvement in the models. This feature selection was conducted using the inbuilt feature importance method which utilizes information gain. After that tuned the hyper parameters of the best models and this also improved the performance.

In conclusion, the project on loan attribution using machine learning achieved success in finding the best model for predicting loan default. The decision tree model was found to be the most accurate and efficient, offering a high degree of explainability. The decision tree model provides clear visualization of the decision-making process, making it easily understandable for stakeholders. This level of transparency is essential in the financial industry, where loan attribution decisions can significantly affect a person's financial stability. In conclusion, the decision tree model demonstrated its worth as a valuable tool for loan attribution, delivering precise and explicable predictions for loan default risk.

5.3 Future Works

The next step for this project would be to build an application based on this model, which can be used by financial institutions to predict loan defaults and make more informed lending decisions. This application would enable financial institutions to quickly and easily access the benefits of the decision tree model and make use of its improved efficiency and effectiveness in loan management. The application can also be further enhanced by incorporating more features, or by

fine-tuning the model using more data to improve its performance. Based on the features that were determined in the above processes, they can be recommended to the banks to consider as important features to check while evaluating loan applications of different customers which are Loan Amount, Loan Term, Education, Credit History and Applicant Income. This helps the bank determine what to look into for the applications that the customers send for loans. On the other hand, the data science teams can also implement robust models based on our findings and report to better evaluate the model results and the application statuses. The proposed cloud-based application can be utilized by banks and other financial institutions to streamline and automate the entire loan process for improved tracking and processing.

Bibliography

1. Asare-Frempong, Justice & Jayabalan, Manoj. (2017). Predicting Customer Response to Bank Direct Telemarketing Campaign. 10.1109/ICE2T.2017.8215961.
2. Tejaswini1 ,T. Mohana Kavya2 , R. Devi Naga Ramya3 , P. Sai Triveni4 Venkata Rao Maddumala, ACCURATE LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH J.
3. Arun K., Ishan, G., & Sanmeet, K. (2016). Loan Approval Prediction based on Machine Learning Approach.
4. M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.
5. Aslam, Uzair & Aziz, Hafiz Ilyas Tariq & Sohail, Asim & Batcha, Nowshath. (2019). An Empirical Study on Loan Default Prediction Models. Journal of Computational and Theoretical Nanoscience. 16. 3483-3488. 10.1166/jctn.2019.8312.
6. Dosalwar, Sharayu & Kinkar, Ketki & Sannat, Rahul & Pise, Nitin. (2021). Analysis of Loan Availability using Machine Learning Techniques. International Journal of Advanced Research in Science, Communication and Technology. 15-20. 10.48175/IJARSCT-1895.
7. Abakarim, Youness & Lahby, Mohamed & Attioui, Abdelbaki. (2018). Towards An Efficient Real-time Approach To Loan Credit Approval Using Deep Learning. 306-313. 10.1109/ISIVC.2018.8709173.
8. Hemachandran, Kannan & Rodriguez, Raul & Toshniwal, Rajat & Junaid, Mohammed & Shaw, Laxmi. (2022). Performance Analysıs of Different Classıfıcatıon Algorıthms for Bank Loan Sectors. 10.1007/978-981-16-2422-3_16.
9. Islam, Md & Arifuzzaman, Mohammad & Islam, Md. (2019). SMOTE Approach for Predicting the Success of Bank Telemarketing. 1-5. 10.1109/TIMES-iCON47539.2019.9024630.
10. Hemachandran, Kannan & George, Preetha & Rodriguez, Raul & Kulkarni, Raviraj & Roy, Sourav. (2021). Performance Analysis of K-Nearest Neighbor Classification Algorithms for Bank Loan Sectors. 10.3233/APC210004.
11. Shinde, Anant & Patil, Yash & Kotian, Ishan & Shinde, Abhinav & Gulwani, Reshma. (2022). Loan Prediction System Using Machine Learning. ITM Web of Conferences. 44. 03019. 10.1051/itmconf/20224403019.

12. Eletter, Shorouq & Yaseen, Saad & Elrefae, Ghaleb. (2010). Neuro-Based Artificial Intelligence Model for Loan Decisions. American Journal of Economics and Business Administration. 2. 27-34. 10.3844/ajebasp.2010.27.34.
13. Aphale, A.S., & Shinde, S.R. (2020). Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval. *International journal of engineering research and technology, 9*.
14. Meshref, H., 2020. Predicting Loan Approval of Bank Direct Marketing Data Using Ensemble Machine Learning Algorithms. International Journal of Circuits, Systems and Signal Processing, 14, pp.914-922.
15. Kar Yan Tam, Melody Y. Kiang, (1992) Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. Management Science 38(7):926-947.
16. Frye, M., Mohren, J., & Schmitt, R. H. (2021). Benchmarking of Data Preprocessing Methods for Machine Learning-Applications in Production. *Procedia CIRP*, *104*, 50-55.
17. Zelaya, C. V. G. (2019, April). Towards explaining the effects of data preprocessing on machine learning. In *2019 IEEE 35th international conference on data engineering (ICDE)* (pp. 2086-2090). IEEE.
18. B. Pavlyuchenko, "Using Stacking Approaches for Machine Learning Models," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018, pp. 255-258, doi: 10.1109/DSMP.2018.8478522.
19. Industries Insight (2022), Banking, Finance & Insurances - United States, Statista, https://www.statista.com/outlook/io/banking-finance-insurances/united-states.
20. Google Developers (2022), Classification: ROC Curve and AUC, Foundational Courses, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.
21. Sarang Narkhede (June 26, 2018), Understanding AUC - ROC Curve, Towards Data Science, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
22. Ahmad Abujaber, Adam Fadlalla, Diala Gammoh, Hassan Al-Thani & Ayman El-Menyar (2021) Machine Learning Model to Predict Ventilator Associated Pneumonia in patients with Traumatic Brain Injury: The C.5 Decision Tree Approach, Brain Injury, 35:9, 1095-1102, DOI: 10.1080/02699052.2021.1959060

23. Stephanie Glen (2021). "Mahalanobis Distance: Simple Definition, Examples" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/mahalanobis-distance/.

24. Enes Zvornicanin (4 Nov 2022), What is Feature Importance in Machine Learning?, Baeldung, https://www.baeldung.com/cs/ml-feature-importance

25. Vedansh Shrivastava (4 Feb 2022), Loan Approval Prediction Machine Learning, https://www.analyticsvidhya.com/blog/2022/02/loan-approval-prediction-machine-learning/

26. P.L. Murthy, (2020), Loan Approval Prediction System Using Machine Learning, Journal of Innovation in Information Technology, https://innovation-journals.org/IV4i1-5.pdf

27. K. Gana Sai Prasad, P. Vamsi Sai Chidvilas, V. Vijay Kumar (November 2019), Customer Loan Approval Classification by Supervised Learning Model, International Journal of Recent Technology and Engineering (IJRTE), https://www.ijrte.org/wp-content/uploads/papers/v8i4/D9275118419.pdf

28. Analytics Vidhya (1 November 2015), Simple Guide to Logistic Regression in R and Python, https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/

29. Ajay Kulkarni, Deri Chong, Feras A. Batarseh (2020), Foundations of data imbalance and solutions for a data democracy, https://doi.org/10.1016/B978-0-12-818366-3.00005-8

30. Jean-François Le Gall (2005), Random trees and applications, 2005, https://doi.org/10.1214/154957805100000140

31. M. Dash (1997), Feature Selection for Classification, 21 March 1997, Intelligent Data Analysis , Elsevier, http://machine-learning.martinsewell.com/feature-selection/DashLiu1997.pdf

32. Stacia Damron (2018), AI Academy: What is Machine Learning?, 8. November 2018, One Model, https://www.onemodel.co/blog/ai-academy-what-is-machine-learning

33. Ravish Raj (2022), Supervised, Unsupervised and Semi-Supervised Learning with Real-life use case, Enjoy Algorithms, https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning