**Credit Risk Prediction Using Machine Learning Algorithms**

*Agrim Jain & Aryaman Bansal*

*BTech (Hons) in CS*

*Birla Institute of Technology and Science, Pilani Dubai Campus*

*DIAC*

*Dubai, United Arab Emirates*

**ABSTRACT**

Credit risk prediction is essential for financial institutions such as banks because it helps them to make informed decisions regarding lending money. Since lending money is a risky business, financial institutions need to ensure that they lend money to individuals or firms with a high probability of paying back the money. Banks and financial institutions who can’t assess credit risk accurately lead towards financial uncertainty and eventually, bankruptcy. Credit risk prediction is vital because it helps the financial institutions mitigate the risk of default by analyzing different factors that influence a borrower’s ability to pay back the amount. The traditional credit scoring model only rely on a few variables such as age, income, financial history, credit history; however, these methods aren’t enough to provide the complete information regarding the borrower’s creditworthiness. To get an improved accuracy than traditional credit scoring models, machine learning (ML) techniques are used to analyze large amount of data to make better credit decisions. The ML models help to identify patterns and relationships in big data that humans may miss doing it manually. This allows the lenders to make informed decisions with respect to credit decisions and it reduces the probability of default.

Lending money is the main type of revenue for financial institutions such as banks; however, it comes with the risk of default which results in big losses. Accuracy credit risk prediction is, thus, crucial for financial institutions to minimize the risk that they expose and ensure that they only give money to the borrowers who will repay their loans. Supervised learning models such as random forests, decision trees, and logistic regression are the frequently used models in credit risk prediction. Each of these models require past data to train the model for predicting the likelihood of a borrower defaulting on a loan. This model incorporates a wide range of variables such as credit history, income, age, to make accurate credit decisions. Logistic regression models are used to predict credit risk because of their simplicity, high interpretability, and effectiveness. Decision trees work by recursively splitting the data based on the most significant variables used in the dataset, which creates tree structures representing the decision making process. This tree can be interpreted and avail to make credit decisions based on the attributes of the borrower. We’ll be using decision trees as there are multiple parameters and the relationships between variables are not linear. Random Forest , which is a ML algorithm, is applied in credit risk prediction. It works by the creation of multiple decision trees to make the most accurate predictions. This model is, usually, the most accurate because they analyze a large number of variables and are able to identify complex relationships between those variables. Performance metrics for example F1-score, accuracy, and will be used to compare various ML models.

**Keywords: Random Forest(RF), Decision Tree, Credit Risk, Logistic Regression**

**INTRODUCTION**

Financial Organizations often use credit scores, which are a numerical representation of person’s creditworthiness. Financial organizations frequently use credit scores, which are numerical representations of a person's creditworthiness, to assess and analyze the risk that might be involved in extending credit to that person. A person's credit score is determined by a number of attributes, including their credit history, duration of credit history, credit usage, payment history, and additional relevant variables. A high credit score suggests better creditworthiness and a reduced risk of failure, whereas a low credit score suggests worse creditworthiness and a greater risk of failure.

In the past few years, both researchers and professionals have focused a lot of attention on the crucial issue of credit score prediction in the financial sectors. The conventional approach to credit score projection depends on statistical models that project an individual's creditworthiness based on past credit data. However, machine learning-based credit score projection has become a viable option to the conventional strategy thanks to recent developments in big data and ML algorithms.

ML algorithms have valuable and effective tools that can extract relationships and patterns from massive databases and use those patterns to generate precise forecasts that help in better decision making . Machine learning algorithms can accurately forecast a person's creditworthiness based on learned trends in the context of credit score prediction by learning from a large collection of past credit data.The paper aims to solve the problem of Credit Score Prediction using ML Algorithms precisely. First, a general summary of the credit score prediction issue and its significance in the financial sector. The constraints of the conventional method to credit score prediction are covered in the second section. Finally, we discuss machine learning techniques and how they can be used to forecast credit scores. In the fourth section, we compare and contrast various machine learning methods for credit score projection. The paper is concluded with a summary of the major discoveries and upcoming study paths.Credit score prediction is the process of assessing and predicting a person's creditworthiness based on a number of variables, including credit history, payment history, credit usage, credit history duration, and other relevant variables. A greater score indicates a lower risk of failure and better trustworthiness, and a lower score indicates a higher risk of default and poorer creditworthiness. The credit score is usually expressed as a number value varying from 300 to 850. The credit score prediction issue is crucial to the financial sector because it enables lenders to decide whether to give money to a particular person and at what interest rate. People who have high credit scores are more likely to settle debt promptly and have a low risk of defaulting; whereas, people with lower credit risk are more likely to fail in paying back the loan.

The conventional method of predicting credit scores depends on statistical models that project an individual's trustworthiness based on past credit data. Logistic regression, a binary classification method that forecasts the probability of a person defaulting is on the basis of their past credit and other relevant variables, is the most widely used statistical model for credit score prediction. As a result of its ease of use and readability, logistic regression has been extensively used to forecast credit scores. Logistic regression, however, has a number of drawbacks. First, it implies a linear relationship between the predictors and the answer, which may not hold true in real-world situations. Second, non-linear relationships between the predictors and the answer cannot be captured. Last but not least, it poorly manages missing data and anomalies, which can result in skewed forecasts.

Popular ML algorithm Random Forest is applied to predict the credit scores. A prediction is made using an ensemble learning technique that blends different decision trees. Using various random subsets of the data and feature sets, numerous decision trees are constructed using the Random Forest method. The predictions from these decision trees are then combined to produce the end estimate.

Random Forest can be used to forecast a person's creditworthiness based on a variety of characteristics, including credit history, payment history, credit utilization, duration of credit history, and other relevant variables. Both qualitative and numerical data can be handled by Random Forest, and it also does a good job of handling lost data and anomalies.Random Forest can handle non-linear relationships between the predictors and the response, which logistic regression cannot handle. Random Forest can handle missing data and outliers which are not handled in logistic regression.

Extreme Gradient Boosting, also known as XGBoost, is a potent machine learning method that has been applied to the forecast of credit scores. In this kind of gradient boosting method, several weak learners are combined to produce a projection. The fundamental principle of XGBoost is to incrementally add new poor learners to the model while each one tries to correct the flaws of the prior ones. The creditworthiness of a person can be predicted using XGBoost in the context of credit score prediction based on a variety of characteristics including credit history, payment history, credit usage, duration of credit history, and other relevant variables. Both categorical and numerical data can be handled by XGBoost, and it also does a good job of handling incomplete data and anomalies.

Another well-liked ML method that could be applied to credit score forecasts is the decision tree. It is a straightforward but effective algorithm that operates by repeatedly dividing the data into groups according to the characteristic that is most informative at each stage. The end product is a tree-like structure, with the expected outcomes represented by the leaf nodes. Decision trees can be used to forecast a person's creditworthiness based on a variety of characteristics, including credit history, payment history, credit utilization, duration of credit history, and other relevant variables. Decision trees are effective at handling absent data and anomalies, as well as numerical and qualitative data. The decision tree is an effective instrument for describing the decision-making process to stakeholders because it is simple to read and comprehend. Given that credit scoring databases frequently contain a mixture of both numerical and categorical data, the decision tree's ability to manage both kinds of data makes it helpful for credit score prediction.in the context of predicting credit scores, where there may be millions of events, decision trees are essential because they are numerically efficient and can manage big datasets effectively.

Machine learning algorithms are effective tools that can extract relationships and patterns from massive databases and use those patterns to generate precise forecasts. Machine learning algorithms can make precise predictions in the context of predicting credit scores by learning from a sizable dataset of past credit data.

**LITERATURE SURVEY**

Many articles and journals have been published pertaining to the topic of credit risk prediction. These research papers provide the recent advances in ML for credit risk assessment. Summary of the different research papers and its methodologies have been presented below.

Paper [1] talks about predicting bank credit worthiness. Majorly, three different types of classifiers were used: KNN, Naive Bayes, and Decision Trees to rank the customer credit worthiness. Authors of this paper mention that creditworthiness is an important aspect to minimize risk and avoid defaults. The dataset was taken from a UCL machine repository. ML algorithms such as SVM, Decision Tree, KNN, RF, and Naive Bayes were used. While CART and KNN only had an accuracy of 0.73 and 0.76, respectively, RF and AdaBoost had the highest accuracy in this study of 0.82 and 0.83.

Paper [2], applied ML algorithms such as XGBoost, RF, and logistic regression to predict risk in online lending. The results of [2] indicate that random forest and XGBoost significantly outperform logistic regression in terms of accuracy. [2]’s results showed the Support Vector Machine had the high recall value followed by Random Forest with a recall value of 0.795. Both Paper [1] and Paper [2] used machine learning techniques to predict credit risk. Yu, the author of Paper [2] focused on online lending while Turkson, author of Paper[1], focused on bank creditworthiness.

Li, the author of Paper [3] uses XGBoost, a ML algorithm, to identify people who don’t pay back the loans. The paper compares the XGBoost algorithm and Logistic Regression. Li collected a dataset for a peer-to-peer lending platform in China, and he used a separate testing dataset to ensure generalizability of the performance. The results of his study prove that the commonly used Logistic Regression model has lower model discrimination and stability than XGBoost algorithm, which improves personal fast credit risk identification. The performance measure was measured on the basis of PSI value. For logistic regression it was 0.0456 and for XGBoost it was 0.0565.

Paper [4], written by Juneja, predicts credit risk using ML algorithms such as SVM, RF, and gradient boosting. Juneja uses a credit card dataset which consists of transactions and customer information to train various classification models. The algorithms that Juneja used in her study were gradient boosting, logistic regression, decision trees, and SVM. Juneja concluded that the use of ML algorithms to enhance credit risk evaluation procedures at financial organizations, lowers the risk of defaults, and increases the total profitability. While using the different machine learning models, Juneja concluded that the gradient boosting algorithm had the best accuracy, precision, recall F1-score, AUC-ROC of 0.8667, 0.88, 0.86, 0.87, 0.92, respectively. Both Paper [3] and Paper [4]’s results concluded that gradient boosting has the highest accuracy. However, the datasets used in both were limited and lacked diversity. Moreover, the performance metrics of Paper [3] were measured in terms of PSI value, whereas in Paper [4] it was in terms of precision, F1-score, recall, and accuracy.

Parvin and Saleena, authors of Paper [5], were the ones to uniquely use an ensemble classifier ML technique to forecast credit scoring and compare the performance with other machine learning models including decision trees and logistic regression. The authors used a dataset from bank customers’ credit information in India. In their study, they used different ML algorithms such as RF, SVM, Decision Tree, and Naive Bayes. The author’s goal was to use ensemble learning models to enhance the results of credit risk prediction. Not only did they evaluate the performance for the ensemble model, they also did an in-depth comparative analysis of the different classifiers. Furthermore, they used feature selection methods such as correlation analysis and PCA to offer practical and effective approaches to credit risk prediction which will benefit financial institutions.

Trivedi, author of Paper [6], talks about credit score prediction models using the different ML techniques and feature selection. The author of this paper uses algorithms such as principal component analysis (PCA), decision trees, and neural networks. Trivedi looked at how feature selection affects classification performance and offers advice on how to create precise credit scoring models using machine learning. The dataset that the author used incorporated real-world information from an India bank, which increases the research’s applicability. However, the dataset is only limited to the Bank of India, the paper doesn’t provide information about why the algorithms that have been selected are considered for the research. In conclusion of the research conducted by Trivedi, it is noted that KNN achieved the best result of 0.816, whereas SVM had the highest AUC-ROC of 0.899. Logistic regression had an accuracy of 0.797 and an AUC-ROC of 0.747.

Paper [7] presents a basis of ML techniques for credit score forecasting. Moscato, the author of Paper [7] uses a dataset of credit risk and compares the outcome of various ML algorithms like RF, KNN, SVM, ANN, Logistic Regression, and Gradient Boosting. Author used around 10 different machine learning algorithms, which meant that there was a comprehensive evaluation which led to more practical and accurate insights and results. The study that was conducted by Moscato, provided empirical implication for financial institutions that use credit scoring models. Through the results, it was noted that Random Forest and gradient boosting were the 2 algorithms that had the highest accuracy. Paper [5], written by Parvin and Saleena, proposed the ensemble classifier model whereas Paper [7], by Moscato, focused on benchmarking the different ML algorithms for credit score forecasting.

Dumitrescu, author of Paper [8], implied that traditional credit scoring models have very limited predictive power. This paper addresses that problem and the author poses that the use of decision tree algorithms can solve this problem. Author used a dataset from the European financial institution, and used logistic regression, XGBoost, and Decision trees as the different ML algorithms to anticipate the credit risk. Author quickly identified that using a non-linear decision tree, there was improved accuracy and the results were robust. With respect to the kaggle dataset from the study, accuracy for linear logistic regression was 0.6983 whereas the accuracy for non-linear logistic regression was 0.776. However, random forest was the model which had the highest accuracy of 0.8529.

Paper [9] focuses on the practical applications of machine learning models for credit scoring used by financial institutions. AUthors note that traditional credit scoring models have certain limitations in capturing complex relationships between variables which, in turn, lead to biased predictions. This paper offers the use of random forest and logistic regression to forecast the credit scores. Authors suggest that the use of feature engineering techniques to select relevant features for the model is also vital. The method proposed in Paper [9] consisted of a dataset of loan applications from a Chinese bank and outperforms the traditional credit scoring models. While logistic regression had an outcome of 0.5921, the RF’s accuracy was 0.5377. While Paper [8] focuses on improving logistic regression models by incorporating decision trees, Paper [9] uses a machine learning model using RF and Logistic regression techniques which overcome the limitations of traditional credit scoring models. Paper [9] builds on the limitations of the conventional credit scoring models, while Paper [8] aims to improve them. Both papers, in common, highlight the importance of the fact that machine learning methods in credit scoring will improve prediction accuracy.

Amato, author of Paper [10], focuses on developing a model to accurately predict creditworthiness based on features relating to an individual’s financial history. They propose an ensemble model combining ML algorithms including SVM, Decision Trees, and logistic regression to improve the performance. Authors, moreover, performed a comparative study with the traditional scoring models to show that their approach outperforms them. The accuracy for the RF algorithm was 0.913 whereas the other algorithms’ accuracy ranged from 0.875 to 0.897.

As in Paper [9], Wu and Pan addressed the problem of credit scoring in financial institutions and analyzed different machine learning applications to evaluate credit worthiness. In Paper [10], Amato also focuses on credit scoring prediction using machine learning techniques. Authors compare performance of different algorithms including SVM, Random Forest, KNN, and XGBoost. While Paper [9] provided only a general overview of the practical use of ML in credit scoring, Paper [10] highlights the specific problem of comparing the performance of different ML algorithms for credit scoring prediction through presenting an empirical study to evaluate the performance in terms of accuracy of each ML algorithm.

Paper [11], written by Therese, proposed a credit card approval system using different supervised learning algorithms. They evaluated performance of different ML techniques such as SVM, Decision Trees, KNN, and random forest. This study suggested that a machine learned based method for detecting credit card fraud can aid financial institutions in quickly and accurately identifying fraudulent transactions. It compared various algorithms based on evaluation metrics including accuracy and F1-score. Credit card transaction data is clean and prepared using preprocessing methods, which increases the precision of the results. The random forest algorithm received the best accuracy, precision, recall, F1-score of 0,9999, 0.9155, 0.7182, and 0.843, respectively.

Kozodoi discusses the potential challenges and implications of fairness within the realm of credit scoring models. The author of Paper [12] presented a framework for assessing fairness within the credit scoring model and proposed an algorithm which optimizes fairness and profitability, as well. This paper provided the readers with a thorough empirical analysis of fairness constraints on performance of credit scoring models. It provides the readers with an organized set of guidelines for implementing fair credit scoring. Moreover, the paper provides practical implications for borrowers who want to use the fair credit scoring models. The paper doesn’t give experimental analysis; however, it compares different fairness techniques for credit scoring and only provides literature review and conceptual framework.

Paper [13] proposes a unique technique of a hybrid machine learning approach for assessing credit risk of commercial customers. They used a combination of clustering and decision trees to identify the most important features used in credit risk assessment. The hybrid model used by the authors, means that the paper was flexible. It could handle non linear relationships between variables and handle different types of data. It was generalizable even though the dataset was picked up from a Brazilian bank, the approach used in the hybrid model is generalizable to other contexts as well. The hybrid model’s performance measure value was measured in R^2. They were in combinations, since it was a hybrid model. K-mean + Adaboost (0.84009), K-means + Gradient Boosting (0.99642), K-means + Decision Tree (0.99706), K-means + Random Forest (0.99744), and K-mean + SVM (0.55380).

While Paper [11] aimed to develop a credit card approval system using supervised learning techniques, they also wanted to improve efficiency of the credit card approval process by reducing time taken to evaluate an application and minimize changes of human error. Paper [13] identified a similar problem but it was mainly for commercial customers. The objective of this research was to build an accurate model which assists financial institutions to make better credit decisions. What Paper [13] solved from Paper [11] was that hybrid models are one way to outperform any other model in terms of accuracy and efficiency and that is proven by the results of each study.

Zhenge, the author of Paper [14], proposes a default prediction method which will accurately predict default risk of borrowers using credit data. To solve these problems, author uses XGBoost and LightGbm which handle complex nonlinear relationships within data. Author aims to improve the accuracy and efficiency to help financial institutions. Using these two algorithms, Zhenge knew that there was robustness to the asymmetric data and the approach received results including high accuracy and efficiency. However, their interpretability is very limited since both XGBoost and LightGbm are two complex machine learning algorithms. The accuracy metric for both XGboost and LightGbm were similar, 0.798 and 0.799 respectively.

Anand, in Paper [15], talked about loan behavior prediction and he evaluated performance of decision trees, random forest and SVM’s on a dataset of loan applications. The authors picked up the dataset from Kaggle and developed, evaluated their model using k-fold cross validation. They measured the performance using F1-score, accuracy, and precision. The algorithms they used included RF, Decision Tree, Gradient Boosting, and Logistic Regression. The paper included accurate predictions of credit risk, improved efficiency, and also low cost. Most importantly, the paper mentioned enhancing security within credit risk and how machine learning models can help to improve the security of the lending processes for the financial institutions. The accuracy of random forest was 0.8555, CatBoost was 0.8492, and for XGBoost it was 0.8387. All three types of gradient boosting algorithms resulted in similar accurate metrics.

**REFERENCES**

(1) R. E. Turkson, E. Y. Baagyere and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness," 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, Poland, 2016, pp. 1-7, doi: 10.1109/ICAIPR.2016.7585216.

(2) Yu, Xiaojiao. "Machine learning application in online lending risk prediction." arXiv preprint arXiv:1707.04831 (2017).

(3) Li, Y. (2019). Credit Risk Prediction Based on Machine Learning Methods. 2019 14th International Conference on Computer Science & Education (ICCSE), 1011-1013.

(4) Juneja, Simran. "Defaulter Prediction for Assessment of Credit Risks using Machine Learning Algorithms." 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020.

(5) A. Safiya Parvin and B. Saleena, "An Ensemble Classifier Model to Predict Credit Scoring - Comparative Analysis," 2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Chennai, India, 2020, pp. 27-30, doi: 10.1109/iSES50453.2020.00017.

(6) "Shrawan Kumar Trivedi,A study on credit scoring modeling with different feature selection and machine learning approaches,Technology in Society,Volume 63,2020,101413,ISS 0160-791X,"

(7) Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. Expert Syst. Appl., 165, 113986.

(8) Dumitrescu, E.I., Hué, S., Hurlin, C., & Tokpavi, S. (2021). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. Eur. J. Oper. Res., 297, 1178-1192.

(9) Wu, Y., &amp; Pan, Y. (2021). Application analysis of credit scoring of financial institutions based on machine learning model. Complexity, 2021, 1–12. https://doi.org/10.1155/2021/9222617

(10) Amato, F., Ferraro, A., Galli, A., Moscato, F., Moscato, V., & Sperlí, G. (2022). Credit Score Prediction Relying on Machine Learning.

(11) M. J. Therese, A. Devi, R. Gurulakshmi, R. Sandhya and P. Dharanyadevi, "Credit Card Assent Using Supervised Learning," 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), Villupuram, India, 2022, pp. 1-6, doi: 10.1109/ICSTSN53084.2022.9761307.

(12) Kozodoi, Nikita, Johannes Jacob, and Stefan Lessmann. "Fairness in credit scoring: Assessment, implementation and profit implications." European Journal of Operational Research 297.3 (2022): 1083-1094.

(13) Machado, M. R., &amp; Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. Expert Systems with Applications, 200, 116889. https://doi.org/10.1016/j.eswa.2022.116889

(14) Y. Zheng, "A Default Prediction Method using XGBoost and LightGBM," 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Xi’an, China, 2022, pp. 210-213, doi: 10.1109/ICICML57342.2022.10009823.

(15) Anand, M., Velu, A., & Whig, P. (2022). Prediction of Loan Behaviour with Machine Learning Models for Secure Banking. Journal of Computer Science and Engineering (JCSE)

**Credit Risk Prediction Using Machine Learning Algorithms - Literature Survey**

| **NO** | **Reference** | **Objective** | **Problem Statement** | **Methodology** | **Dataset** | **Algorithm** | **Advantage** | **Disadvantage** | **Performance measure value** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | R. E. Turkson, E. Y. Baagyere and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness," 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, Poland, 2016, pp. 1-7, doi: 10.1109/ICAIPR.2016.7585216. | Develope machine learning model to predict the credit worthiness of application who wish to take bank loan. Authors aim is to identify the important variables in predicting the credit worthiness and compare the different machine learning algorithms. | Authors mention that credit worthiness is an important feature for financial institutions to minimize the risks and avoid defaults. Authors feel that machine learning algorithms will provide accurate and efficient way to assess credit worthiness. | Dataset was from UCI machine repository. VData preprocessing followed by feature collection was done. Furthermore, machine learning algorithms included SVM, Decision Tree, KNN, Random Forest, and Naive Bayes. The model was evaluate using measures such as accuracy, precision, F1-score, and AUC-ROC curve. | Dataset was obtained from the UCL machine repository. The dataset used in this paper is labeled data and is suitable for doing analysis. Data included 1000 loan applications and variables such as age, income, loan amount, loan duration, credit history. | Decision Tree, Naive Bayes, Random Forest, KNN, SVM | The algorithms used can be applied to other datasets so it is generalized. The bias is reduced since the dataset is so large such that factors such as race and gender can be considered neglected. | There's overfitting of data as the training data is unable to generalize the new data. The disadvantages are that the algorithms are expensive and time consuming for data scientists and software engineers. | Accuracy is Random Forest (0.82), AdaBoost (0.83), SVM(0.78), KNN(0.76), CART(0.73) |
| 2 | Yu, Xiaojiao. "Machine learning application in online lending risk prediction." arXiv preprint arXiv:1707.04831 (2017). | In order to forecast credit risk for online lending sites, study compares the performance of various machine learning algorithms. It analyzes actual data from a site for online lending and compares the performance of various models using measures like accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). The final objective is to offer information about the use of machine learning in credit risk prediction and to contribute to increasing the precision and effectiveness of credit risk evaluation. | This study looks at how well machine learning algorithms forecast credit risk for online lending platforms using a variety of factors, including the borrower's credit score, loan sum, loan duration, etc. It emphasizes the need to use machine learning methods to increase the precision and effectiveness of credit risk evaluation.  The issue is how to use a machine learning algorithm to forecast credit risk for online money lending sites. | In this paper, we evaluate the efficacy of four machine learning methods in predicting credit risk using real-world data from an online lending site. The approach includes gathering data, preparing it, choosing features, choosing models, assessing models, and discussing results. The most pertinent characteristics for credit risk projection are chosen using methods like the chi-squared test and correlation analysis in this study. The preprocessed data is used to train and test the models, and measures like accuracy, precision, recall, F1-score, and AUC-ROC are used to assess their performance. The article examines the effectiveness of various models while also presenting the experiment's findings. The study's limitations are also discussed in the report, which offers details on how machine learning is being used to forecast credit risk for | The dataset includes information on borrower's credit score, loan amount,loan purpose, loan term, l employment status, etc | Logistic Regression Decision Tree Random Forest and ANN(Artificial Neural Network) | The research done was based on real time dataset and different Machine Learning algorithms were compared for efficiency. Multiple evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess the performance of the prediction models comprehensively. | There is no economic analysis , Lack of external validation Limited features are used for credit score prediction . | The Random Forest algorithm had the highest accuracy and precision, followed by Gradient Boosting with the highest precision and recall values. The study also found that Support Vector Machine had the highest recall value, followed by Random Forest with a recall value of 0.795 and F1-score of 0.783. AUC-ROC was found to be the highest, with Random Forest having the highest AUC and Recall-ROC values. |
| 3 | Li, Y. (2019). Credit Risk Prediction Based on Machine Learning Methods. 2019 14th International Conference on Computer Science & Education (ICCSE), 1011-1013. | 1.Using the XGBoost algorithm to identify the bad customers who do not pay moneyback from the good customers. 2. The Paper shows a comparative study between XG Boost algorithm and Logistic Regression | 1. machine learning-based credit risk prediction model that can accurately predict the creditworthiness of loan applicants using a set of input features. 2. the paper aims to compare the performance of different machine learning algorithms and identify the most accurate algorithm for credit risk prediction. | The authors collected a dataset from a peer-to-peer lending platform in China and preprocessed it into a suitable format for machine learning algorithms. They trained six different machine learning algorithms on the preprocessed dataset and selected features. They evaluated the performance of each algorithm using various metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. They also tested the selected machine learning algorithm on a separate testing dataset to measure its generalization performance. | peer-to-peer lending platform in China. The dataset contained 1000 loan applications, and each loan application had 20 attributes, including personal and financial information of the loan applicants, loan amount, loan term, and loan status | 1.Logistic Regression Algorithm 2.XG Boost Algorithm | XGBoost model has higher model discrimination and model stability than logistic regression model, improving personal fast credit risk identification. | The paper only used a limited set of evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. While these metrics are commonly used for binary classification problems, they may not fully capture the performance of the machine learning algorithms for credit risk prediction. The paper only considered a limited number of features . | The PSI value obtained using Logistic Regression was 0.0458 PSI value for XG Boost was 0.0565 |
| 4 | Juneja, Simran. "Defaulter Prediction for Assessment of Credit Risks using Machine Learning Algorithms." 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020. | In order to enhance the evaluation of credit risks, it is important to look into the use of machine learning algorithms for forecasting credit default risk. The goal is to evaluate how well a collection of previous credit transactions can be used to forecast credit failure risk for a number of machine learning methods, including Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forest, and Support Vector Machines and to deliver a trustworthy and accurate credit risk evaluation instrument that will help banks and other financial organizations make wise loan choices. | Traditional credit risk evaluation techniques frequently rely on incomplete data and arbitrary judgments, which can result in unreliable forecasts and high default rates. In order to enhance credit risk evaluation, this study will look into the use of machine learning algorithms that analyze historical credit transaction data to forecast default probabilities. In order to assist banks and other financial organizations in making educated loan choices, it is necessary to evaluate the effectiveness of various machine learning algorithms. | 1. The data was collected and processed by the authors . 2. The most relevant features were selected using corelation technique. 3. The model was then trained using 5 different machine learning algorithms . 4. Then the model was evaluated using , precision, recall, F1 score, and area under the ROC curve. | 1.The dataset was obtained from a financial organization which comprised of like gender education etc. and his loan amount, loan duration, credit history, property,area. 2. History of payment ,credit data, default payments and bill statements of credit card clients residing in Taiwan from the duration April 2005 to September 2005 | Gradient Boosting, Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Artificial Neural Network (ANN) and Random Forest (RF). | The report tackles a real issue that financial organizations have when determining credit risks. The use of machine learning algorithms to enhance credit risk evaluation procedures at financial organizations, lower the risk of defaults, and increase total profitability is explored in this study. It offers a comparison study of various machine learning algorithms, including Gradient Boosting, Logistic Regression, and Random Forest, in addition to an experimental assessment of the various algorithms using actual data. This assessment offers information on how well the algorithms perform at forecasting defaulters and aids in demonstrating the efficacy of the suggested strategy. | The dataset is limited and lacks diversity . The paper uses limited set of machine learning algorithms which may not capture the best predicted outcome. Machine learning algorithms are only as unbiased as the data they are trained on. If the data used to train the algorithms is biased, the resulting predictions may also be biased. This could lead to unintended discrimination against certain groups of borrowers. | Gradient Boosting algorithm had an accuracy of 86.67%, precision of 0.88, recall of 0.86, F1-score of 0.87, and an AUC-ROC of 0.92. |
| 5 | A. Safiya Parvin and B. Saleena, "An Ensemble Classifier Model to Predict Credit Scoring - Comparative Analysis," 2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Chennai, India, 2020, pp. 27-30, doi: 10.1109/iSES50453.2020.00017. | Objective is to compare the performance of Decision Tree, Naive Bayes, RF, and SVM to predict credit scoring. Authors evaluate performance of ensemble model which combines the predictions of multiple classifiers to improve accuracy of the model. | Credit scoring methods manually is very time consuming and is subjective. The need for more accurate and efficient credit scoring models help financial institutions to evaluate the credit worthiness of loan application more effectively. | The data was collected from a financial institution in India. Dataset was preprocessed as well to handle the values, outliers. The techniques that author implemented included the mean imputation, z-score, and encoding. Authors compared performance of four classifiers Decision Tree, Naive Bayes, RF, and SVM. The models were evaluated based on the accuracy, F1, and ROC-AUC. | The dataset was obtained from a financial institution in India. Contains demographic information including financial history, credit worthiness. Authors preprocessed and handled the missing values, outliers and the categorical variables. | Random Forest, Naive Bayes, SVM, and Decision Tree | The ensemble classifier model combines prediction of multiple ML algorithms, which improves the accuracy. The study also does an in-dept comparative analysis of the different classifiers. Moreover, the feature selection methods used such as the correlation analysis, PCA, and mutual information. The paper offered a practical and effective approach to credit risk prediction which benefits financial institutions. | Only uses one single dataset from a financial institution in India, so the data can be generalized to other regions. It only evaluates the performance of ensemble classifier against a set of individual classifiers. It is not the optimal method as additional research needs to be conducted in order to evaluate. | Random Forest algorithm achieves the best performance, with an accuracy of 0.8841, Logistic Regression (0.833), SVM (0.8768), Naive Bayes (0.8043). |
| 6 | *Shrawan Kumar Trivedi,*  *A study on credit scoring modeling with different feature selection and machine learning approaches,*  *Technology in Society,*  *Volume 63,*  *2020,*  *101413,*  *ISSN 0160-791X,* | Using a real-world dataset from an Indian bank, compare the performance of various machine learning methods and feature selection approaches in forecasting borrower creditworthiness. Also to evaluate the effects of feature selection on classifier performance and determine the key characteristics that influence credit scoring and to offer recommendations on the most effective methods for creating precise credit rating models using machine learning. | The Problem statement is that to identify the best ML algorithm to predict credit score most effectively, to address this a real life data set is considered by the authors. | Data preprocessing is used to encode classified factors, scale the data, and eliminate missing values. The most pertinent features are found using feature selection techniques like reciprocal information feature selection, correlation-based feature selection, and recursive feature removal. Models are created using model-building techniques like logistic regression, decision trees, random forests, k-nearest neighbors, and support vector machines. Several assessment measures, including accuracy, precision, recall, F1-score, and AUC-ROC curve, are used to assess the effectiveness of a model.  To find the model that performs the best, models are compared, and the effect of feature selection on classification performance is also examined. | Paper Uses a real life dataset of Bank Of India where the missing values where added. | Logistic regression, decision tree, random forest, k-nearest neighbor, and support vector machine. | The efficacy of various machine learning methods for predicting borrowers' creditworthiness is compared and evaluated in this study. Additionally, it looks at how feature selection affects classification performance and offers advice on how to create precise credit scoring models using machine learning. The study incorporates a real-world information from an Indian bank, which increases the research's applicability. The research contrasts the effects of feature selection on classifier accuracy. | The dataset is only limited to Bank Of India. The paper does not provide information about why the algorithms that are selected are considered for the study. | K nearest achieved the highest value 0.817 , support vector had highest AU ROC 0.899. Logistic Regression had an accuracy of 0.797n and AU ROC of 0.747. |
| 7 | Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. Expert Syst. Appl., 165, 113986. | Compare performances of multiple machine learning algorithms to predict credit scores. Moscato V Picariello and Sperli G, the authors, collected data from real world credit scoring applications and then application different machine learning algorithms including logistic regression, decision tree, random forest, neural networks. After this, they evaluated the performance of each algorithm using metrics such as accuracy, precision, F1- score and ROC curve. Results of the study can be used to improve credit scoring models and create possibilities of new models | The problem of this research is to notify the readers the challenges of predicting credit scores using the different algorithms. Authors of the research paper identified the need to evaluate performance of different algorithms and identifying the most effective one. Then, provide insights into the strengths and weaknesses of each algorithm which will help financial institutions to improve the credit scoring models | The dataset that the authors used was from a financial institution in Italy. Authors created new features from existing data to improve performance of the different machine learning algorithms. After that, authors selected 10 different models that they can use to predict credit scores. Furthermore, they trained the models using the preprocessed data. Additionally, performance was evaluated using accuracy, precision, F1-score, and ROC curve. The wilcoxon test was used to compare the performance between the different machine learning algorithms. Moreover, author provided insights to the strengths and weakness of each machine learning algorithm and identified the most effective machine learning algorithm. | The Dataset used was a real world credit scoring dataset which the authors picked up from a financial institution in Italy. It contains a 5000 credit applications along with 15 variables. | Logistic Regression, Random Forest, Gradient Boosting, KNN, SVM, ANN, AdaBoost | Evaluation of 10 different machine learning algorithms meant that the comprehensive evaluation leads to more practical insights which means more effectiveness of algorithms in credit score prediction. Study conducted in this research provided practical implication for financial institutions that use credit scoring models. The results of the study helps improve credit risk assessment. This study's results also outperformed the state of the art models. | The dataset that the author picked up was only limited to the financial institutions in Italy. The models used in the study were black box models, which are hard to interpret and understand. The measures to evaluate the study aren't enough to make insights. Moreover, optimization of algorithms could improve the performance. | Accuracy of the different models of Random Forest (0.773), MLP (0.771), Logistic Regression (0.770). RF - RUS had least accuracy of 0.640. |
| 8 | Dumitrescu, E.I., Hué, S., Hurlin, C., & Tokpavi, S. (2021). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. Eur. J. Oper. Res., 297, 1178-1192. | The author uses ML methods which incorporate non-linear decision tree effects to improve the accuracy of logistic regression models. Author argues that the traditional way of finding logistic regression models have limited predicting power since they assume the relationships between the predictors and the outcome variables. Author's aim is to use the predictive power of credit scoring models. | Traditional credit scoring models have limited predictive power. This paper addresses that problem and the author poses that the use of decision tree based ML techniques can solve this problem | Traditional credit scoring models have limited predictive power. This paper addresses that problem and the author poses that the use of decision tree based ML techniques can solve this problem in order to capture high order interactions between variables that traditional models can't capture. | The author used a dataset from a European financial institution. The dataset comprised of 18 variables including information about applicant: age, income, education level, credit history, etc. Dataset was split 80% into training and 20% into testing. | Logistic Regression, XGBoost | The paper has improved accuracy used the non-linear decision tree. The sensitive analysis meant that the results were quite robust. The paper, moreover, maintained interpretability by providing valuable insights to the decision making processes. | There were limited variables used in the study, which doesn't capture all the information required for credit scoring. The author, too, did not compare his study with other studies to make comparative analysis. | With respect to kaggle dataset, Accuracy for Linear LR (0.6983), Non-linear LR (0.776), Random Forest (0.8529), PLTR(0.8568) |
| 9 | Wu, Y., &amp; Pan, Y. (2021). Application analysis of credit scoring of financial institutions based on machine learning model. Complexity, 2021, 1–12. https://doi.org/10.1155/2021/9222617 | Investigation different data preprocessing techniques, feature selection methods. Aim is to evaluate performance of logistic regression, decision trees, random forest, svm, and neural networks. | The problem statement is that financial institutions may not predict credit risk. Therefore, authors explore application of ML models to improve the accuracy of credit risk predictions. Other challenges that they faced included limited data, changing market, etc. | Authors collected a credit dataset from a financial institution in China. They preprocesses the data, trained and evaluated performance using 5 different ML algorithms.Conducted evaluation and analysis, and uses different performance metrics. | The dataset was obtained from a financial institution in China which had 1.5 million records and 30 variables relating to borrower and loan applications. | Logistic Regression, Decision Trees, Random Forest, SVM, and Neural networks | Authors did a comprehensive and sensitive analysis. The practical implication improves the risk management and decision making process of financial institutions. | The data cannot be generalized as it is specific to China. There is limited feature engineering, and limited interpretability. | Logistic Regression (0.5921), Random Forest (0.5377), SVM (0.5524) |
| 10 | Amato, F., Ferraro, A., Galli, A., Moscato, F., Moscato, V., & Sperlí, G. (2022). Credit Score Prediction Relying on Machine Learning. | 1.To develop and compare machine learning models for credit score prediction 2. To identify the most effective machine learning algorithm and feature set to predict credit scores accurately. The study also aimed to provide insights into the importance of different features in credit score prediction and explore the potential of machine learning in improving credit risk assessment. | 1.To identify the creditworthiness of loan applicants based on historical data, including personal, financial, and employment information. 2.To use the best machine learning algorithms to optimize Credit Score Prediction | The data was collected from an Italian bank and preprocessed it to prepare it for machine learning algorithms. They used a correlation matrix to select the most relevant features for credit score prediction and principal component analysis (PCA) to reduce the dimensionality of the data. They used four different machine learning algorithms to build models and evaluated their performance using metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). They also compared their results with a baseline model to validate the effectiveness of their proposed approach. Results and analysis showed that the Gradient Boosting algorithm outperformed the other three models in terms of accuracy. The authors concluded that their proposed machine learning-based approach can effectively predict credit scores and can be used by financial institutions for risk assessment and decision-making. | The Data was collected from an Italian Bank and contained information about loan applicants' personal and financial data. | 1.Decision Tree (DT)  2.Random Forest (RF)  3.Gradient Boosting (GB)  4.AdaBoost (AB)  5.Extreme Gradient Boosting (XGB)  6.Light Gradient Boosting Machine (LGBM) | Decision trees: Decision trees are easy to interpret and provide a clear insight into the decision-making process. They can handle both categorical and numerical data and are less prone to overfitting.  Random forests: Random forests use multiple decision trees to generate a more accurate and stable prediction. They can handle missing data, noisy data, and can estimate the importance of each feature in the model.  Support vector machines: SVMs are known for their robustness and ability to handle high-dimensional data. They work well with both linear and non-linear data and can handle noise in the data. | 1. Some machine learning algorithms used in the paper, such as neural networks, are highly complex and difficult to interpret. 2.The dataset used in the study may have had class imbalance,This can lead to biased models that perform well on the overrepresented class but poorly on the other class. | The accuracy of the Random Forest algorithm was 91.3%, while the accuracy of the other algorithms ranged from 87.5% to 89.7%. The F1 score of the Random Forest algorithm was reported to be 0.91, while the F1 score of the other algorithms ranged from 0.86 to 0.89. |
| 11 | M. J. Therese, A. Devi, R. Gurulakshmi, R. Sandhya and P. Dharanyadevi, "Credit Card Assent Using Supervised Learning," 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), Villupuram, India, 2022, pp. 1-6, doi: 10.1109/ICSTSN53084.2022.9761307. | The objective is to suggest a supervised learning-based method for detecting credit card theft. The goal is to compare the results of various machine learning algorithms for detecting fraudulent credit card transactions in order to determine which algorithm performs the job the best. The paper also provides a dataset of credit card transactions that can be applied to additional studies on financial sector fraud detection. | 1. The paper aims to solve the problem of credit card fraud. 2.It aims to identify the best machine learning algorithm to detect credit card fraud. 3. Using Machine Learning approach financial institutions can minimize the fraud in transactions. | A supervised learning-based strategy that includes data preprocessing, feature selection, model selection, and assessment for detecting credit card fraud. Encoding categorical variables and removing missing values and anomalies are both parts of data preprocessing. From the preprocessed data, the most significant characteristics are chosen through feature selection. Comparing the efficacy of various machine learning algorithms, such as Decision Trees, Random Forest, K-Nearest Neighbors, Logistic Regression, and Support Vector Machines, is necessary for model selection. Model evaluation entails using various evaluation metrics to assess the effectiveness of the chosen machine learning algorithms. Comparing the outcomes of various algorithms to determine which one is the most successful is what results analysis entails. | The dataset was collected from kaggle which consisted of name, gender, age, email. Aadhar no, PAN card no. They collected credit card transactions from financial institutions | The algorithms used are Logistic Regression, Random Forest, Decision Trees, K-Nearest Neighbors,and Support Vector Machines | This study suggests a machine learning-based method for detecting credit card fraud that can aid financial institutions in quickly and accurately identifying fraudulent transactions. It compares the performance of various algorithms by evaluating their performance using various evaluation metrics, including recall, accuracy, and F1-score. The credit card transaction data is also cleaned and prepared using preprocessing methods, which increases the precision of the outcomes. Overall, by assisting financial organizations in minimizing financial losses brought on by fraud, the article makes a significant addition to the field of fraud identification. | The PaperMakes a valuable addition to the field of credit card scam detection, some drawbacks and would benefit from more in-depth data, a wider range of analysis, and more talk on how to understand and apply the findings.It does not compare the suggested method to other approaches that have already been published in the literature, discuss other forms of fraud, or discuss unsupervised learning techniques for fraud detection. The article offers a thorough analysis of the findings, but it does not address how to understand them or their application to the financial sector. | Random Forest algorithm achieves the best performance, with an accuracy of 99.99%, precision of 91.55%, recall of 78.12%, and F1-score of 84.3%. |
| 12 | Kozodoi, Nikita, Johannes Jacob, and Stefan Lessmann. "Fairness in credit scoring: Assessment, implementation and profit implications." European Journal of Operational Research 297.3 (2022): 1083-1094. | In addition to discussing equity and discrimination in credit scoring, this study gives a summary of various fairness metrics.It covers a variety of justice mitigation strategies, including algorithmic modifications, data pre-processing, and post-hoc fixes. The study examines the profit consequences of fairness in credit scoring by analyzing how the implementation of fair credit scoring models can impact the profitability of lending organizations. Fairness can ultimately result in a more long-lasting and profitable loan industry. | The issue paper is trying to solve is of fairness and discrimination in various credit scoring models. The paper aims to identify and use different fairness techniques to achieve this goal | The article examines the existing research on fairness in credit scoring and catalogs the various fairness metrics that have been put forth. It then goes over various methods, such as algorithmic alterations, data pre-processing, and post-hoc solutions, that can be employed to lessen unfairness issues in credit scoring. The writers then investigate the effects of the adoption of a fair credit scoring model on the profitability of a lending organization in a case study. They discover that using equitable loan practices can eventually result in a more resilient and successful lending industry. | The paper does not uses a specific dataset but mentions previously used datasets from Consumer Financial Protection Bureau (CFPB) and the Federal Reserve Bank of New York. Instead they used a case study | Data Pre processing Post Hoc technique Various algorithmic adjustments to make sure model is fair. | Paper provides with a thorough empirical analysis of fairness constraints on performance of credit scoring models. It provides the readers with an organized set of guidelines for implementing fair credit scoring. In addition to that, the paper provides practical implication for borrowers who want to use fair credit scoring models. | One disadvantage of this paper is that the dataset cannot be generalized to other contexts. The problem that they chose to identify and solve is based on binary classification which cannot be applied to complex credit scoring models. Paper could've discussed more regarding the topic of ethical and social implications of using credit scoring models. | The paper given is not an experimental analysis but compares different fairness techniques for credit score.The paper ony gives literature review and Conceptual analysis. |
| 13 | Machado, M. R., &amp; Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. Expert Systems with Applications, 200, 116889. https://doi.org/10.1016/j.eswa.2022.116889 | Develop a hybrid machine learning model to assess credit risk of commercial customers. Author's aim is to combine different machine learning techniques to overcome the challenge of predicting credit risk models | To overcome the limitations of credit scoring models and improve the accuracy of credit risk assessment | Data was collected from a brazilian bank and the model was developed combined random forest and XGBoost to assess credit risk. Performance was evaluated on the basis of accuracy, precision, recall and F1-score. The authors interpreted the results and discussed the limitation and future works of the study they proposed. | Dataset was used from a Brazilian bank. it contains the demographic and financial information of commercial customers. Model was split into training and testing sets. | Random Forest and XGBoost | The advantage of this paper was that the hybrid model meant that the paper was flexible. It could handle non-linear relationships between variables and handle different types of data. It was generalizable even though the dataset was picked up from a Brazilian bank, the approach using the hybrid model is generalizable to other regions and contexts as well. | The limitations to this study include lack of interpretability as the hybrid model is accuray, but less interpretable due to the complexity within the model. The author, too, didn't compare his study with the traditional credit scoring model. | Hybrid Models, values are in measured in R^2, kmeans + Adaboost (0.84009), kmeans + gradient boosting (0.99642), kmeans + decision tree (0.99706), kmeans + random forest (0.99744), kmeans + SVM (0.55380) |
| 14 | Y. Zheng, "A Default Prediction Method using XGBoost and LightGBM," 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Xi’an, China, 2022, pp. 210-213, doi: 10.1109/ICICML57342.2022.10009823. | Propose a default prediction method that accurately predicts default risk of borrowers using the credit data. Authors addresses the various different limitations of statistical models for credit risk assessment. | The problem of this research is that financial institutions such as banks need to assess the credit worthiness of borrowers to make smart decision and manage their credit risk. The models have limitations which can lean to inaccurate prediction of default risk. | To solve these problems, the author aims to predict the credit risk using XGBoost and LightGbm, two ML algorithms which handle complex nonlinear relationships within the data. Author aims to improve the accuracy and efficiency of default prediction to help financial institutions and make better decisions. | The dataset used was a large and an asymmetric dataset. They applied various data preprocessing techniques to improve the accuracy and efficiency of predicting credit risks. | Mainly two algorithms are used XGBoost and LightGbm | There is robustness to the asymmetric data, the approach received high accuracy while at the same time maintaining high efficiency. | There is limited generalizability as dataset cannot be applied to other regions. While XGBoost and LightGbm provide the author with high accuracy, their interpretability is very limited. The research paper also doesn't explain the feature selection. | The accuracy value metric of Xgboost was 0.798 whereas for Lightgbm it was 0.799. |
| 15 | Anand, M., Velu, A., & Whig, P. (2022). Prediction of Loan Behaviour with Machine Learning Models for Secure Banking. Journal of Computer Science and Engineering (JCSE). | The objective is to investigate application of ML models when predicting credit score. Author's primary goal is to identify factors which affect loan repayment and evaluate performance of the different ML models in predicting credit risk | There is need for more accurate credit risk models to help banks and financial institutions to make better and wise decisions. Authors identify the issue as a challenge faced by these institutions which end up in big financial losses for them. | Authors used dataset from kaggle. They developed their model and evaluated the model using k-fold cross validation and measured the performance using precision, accuracy, and F1-score. | The dataset was picked up from kaggle, called Loan Prediction Dataset. It contains 614 observations and 13 variables. Authors preprocess and cleaned dataset to handle values, outliers, and the categorical variables before training and evaluating their model. | Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Gradient Boosting | The advantages of this paper include accurate prediction of the credit risk, improved efficiency which reduces time and cost, enhanced security, explained how ML models can help to improve security of the lending processes for the financial institutions | The disadvantages of this paper are the bias in the data, since some historical data might contain bias, model inherits this bias as well which impacts the group of borrowers. The models aren't interpretable as it is a major task for the lenders to explain why his/her loan application was approved/denied which leads to a sense of dissatisfaction amongst the lenders. | Accuracy of Random Forest (85.55), CatBoost(84.92), XGBoost(83.87) |

**IMPLEMENTATION**

1. **DATASET**

The dataset is stored in a CSV file and we obtained the dataset from Kaggle. The Kaggle dataset included a separate train and test csv file. The train dataset includes 100,000 rows and 28 columns; whereas, the test dataset contains 50,000 rows and 27 columns. The significant attributes in the dataset are described below:

* **ID:** Each entry in the dataset has a specific identification number, which is represented by this attribute. Each row of data may be uniquely identified using it.
* **Customer\_ID:** Each individual in the dataset is given a special identification number, which is represented by this property. Data may be grouped by specific clients using this technique.
* **Month:** The month of the year in which the data was gathered or entered is represented by this feature.
* **Name:** The name of a person is represented by this characteristic. It may be used to locate a specific person in the dataset.
* **Age:** A person's age is indicated by this characteristic. It may be applied to age-based comparisons and analysis.
* **SSN:** The social security number of a person is represented by this characteristic. It can be used to specifically identify people.
* **Occupation: T**his characteristic denotes a person's line of work. It may be used to aggregate data by occupation and to examine income or bank account trends or patterns according to occupation.
* **Annual\_Income:** Based on income, it may be used for comparisons and analysis.
* **Monthly\_Inhand\_Salary:** This characteristic denotes a person's monthly base pay. It may be applied to salary-based comparisons and analysis.
* **The Num\_Bank\_Accounts:** shows how many bank accounts a person has. Based on the total number of accounts owned, it may be used to study banking activity and trends.

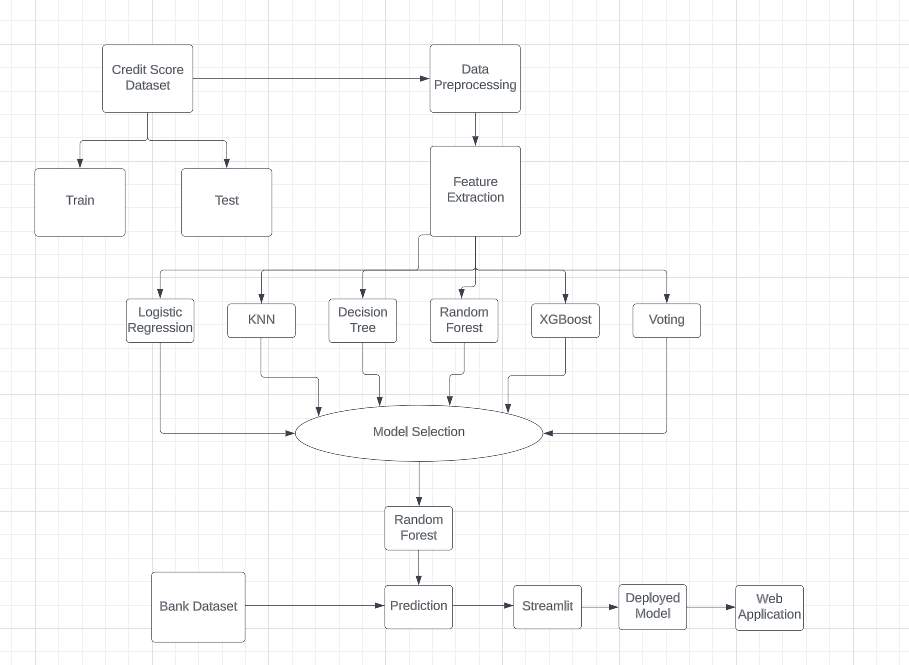
**Figure 1.1** *Dataset*

**Importing a Bank Dataset**

The dataset that we used for training and testing purposes was only on the basis of credit card information. There was a need for importing a bank dataset as well to obtain information regarding the individual’s card and loan history, so that the individual.

1. **ARCHITECTURE DIAGRAM**

In order to analyze the data and provide predictions, an architectural diagram for credit score prediction normally consists of several levels and components. The data component, which is the foundation of the architecture, is where the credit score dataset and bank dataset are located. The preprocessing layer includes operations like feature scaling, normalization, and data cleaning. The feature engineering layer is the following layer, where new features may be developed from the available data to increase the models' precision. The machine learning models itself, which contains the different ML algorithms such as the LR, DT, RF, XGBoost.

**Figure 2.1** *Architecture Diagram* 

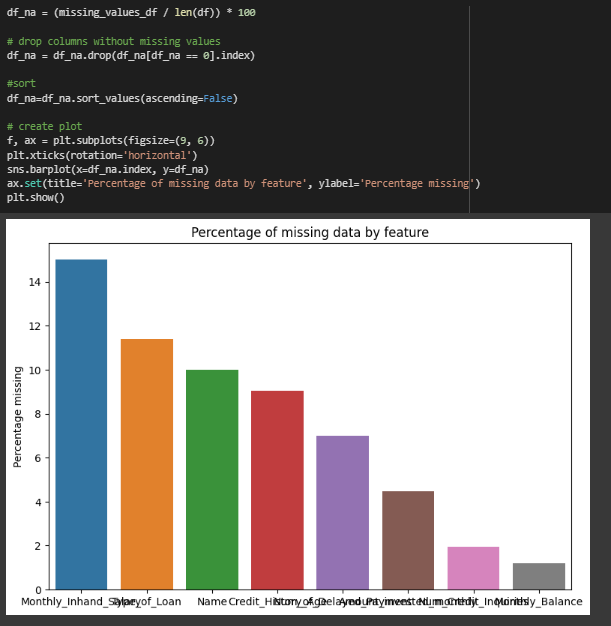
1. **DATASET PREPROCESSING**

A crucial phase of every machine learning research is data preparation. To make sure the raw data is prepared for analysis by a machine learning algorithm, it entails cleaning and preparing it. For example, controlling outliers, scaling the data, and transforming categorical variables into numerical ones are all examples of data preparation chores. In order for the machine learning algorithm to successfully recognize patterns and produce reliable predictions, the data must be preprocessed. Additionally, it can help the model perform better and lower the danger of overfitting. Data preparation is frequently an iterative process, thus it is essential to carefully assess the effects of each step on the data and the effectiveness of the machine learning model. Any machine learning project must have proper data preparation in order to be successful, and the machine learning report must include all of the pretreatment processes that were completed.

* 1. **Missing Values and Type Casting**

Type casting and missing values are two problems that frequently occur in machine learning applications. Missing values are values that are absent from the dataset, whereas type casting is the process of changing variables' data types. Data preparation procedures like handling missing values and type casting are essential because they have a big impact on how accurate the machine learning model is.

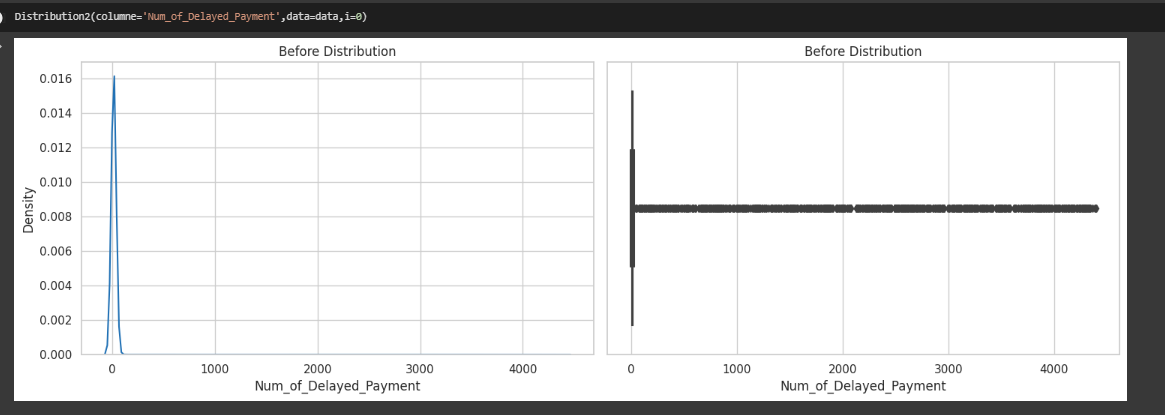
Another crucial task in machine learning programs is type casting. Different data formats are needed for different algorithms. As an illustration, although certain algorithms may need numerical data, others may need categorical data. Because of this, it's crucial to accurately identify the data types of the variables in the dataset and convert them to the relevant data type.



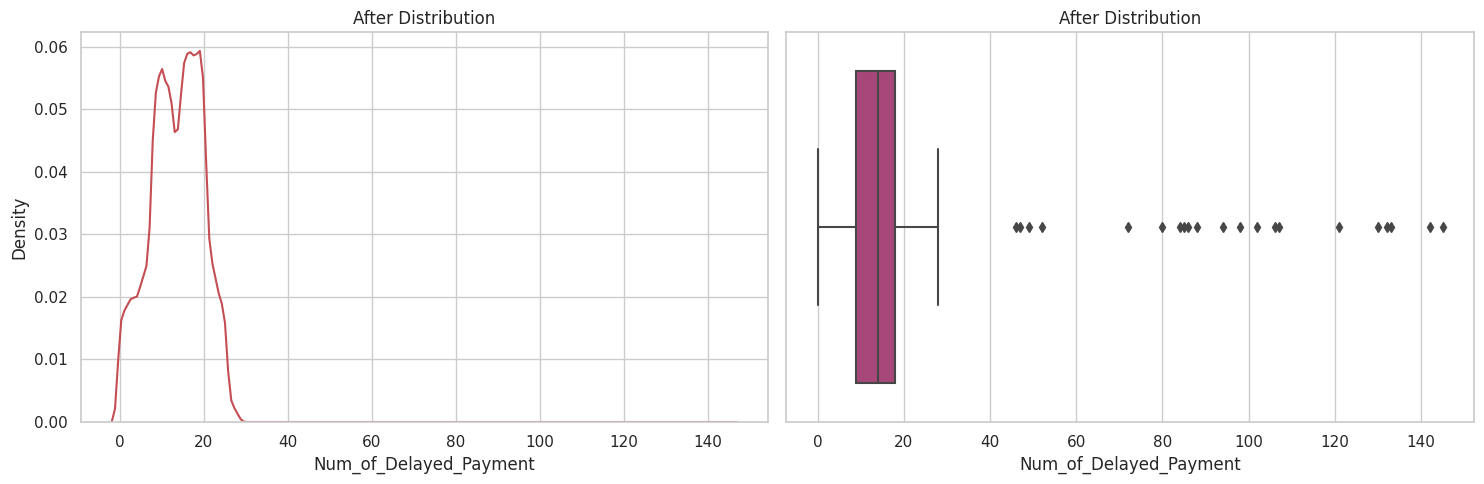
**Figure 2.1** *Identifying missing values* **Figure 2.2** *Type Casting of attributes*

* 1. **Handling Outliers**

Handling outliers is crucial since outliers can greatly reduce the model's accuracy. Outliers are extreme numbers that deviate considerably from the other data points. They might be the result of measurement mistakes, data gathering faults, or uncommon occurrences. A machine learning algorithm may overfit the model as a result of outliers, which may result in subpar generalization performance. Outliers can be dealt with in a number of ways, such as by deleting them from the dataset, altering the data to make it more regularly distributed, or applying powerful statistical techniques. The strategy chosen relies on the type of data being utilized and the particular machine learning algorithm being applied.

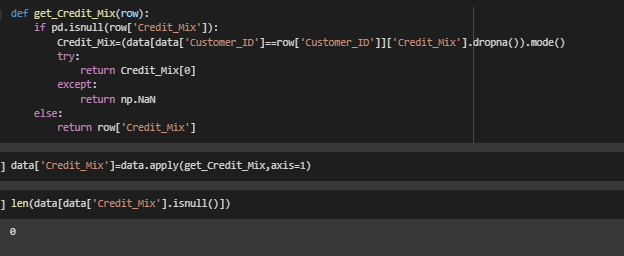
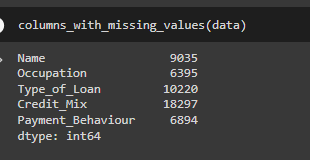


**Figure 2.3** *Before handling the outliers in Num\_Of\_Delayed\_Payment*

**Figure 2.4** *After handling the outliers in Num\_Of\_Delayed\_Payment*

* 1. **Filling NaN in place of Missing Values**

Missing values can frequently be found when dealing with a dataset. One frequent method for filling in missing numbers in data preparation is to use NaN. It is crucial to remember that replacing missing values with NaN should only be done with caution and only after carefully weighing the potential effects on the data's correctness and the effectiveness of the machine learning model.

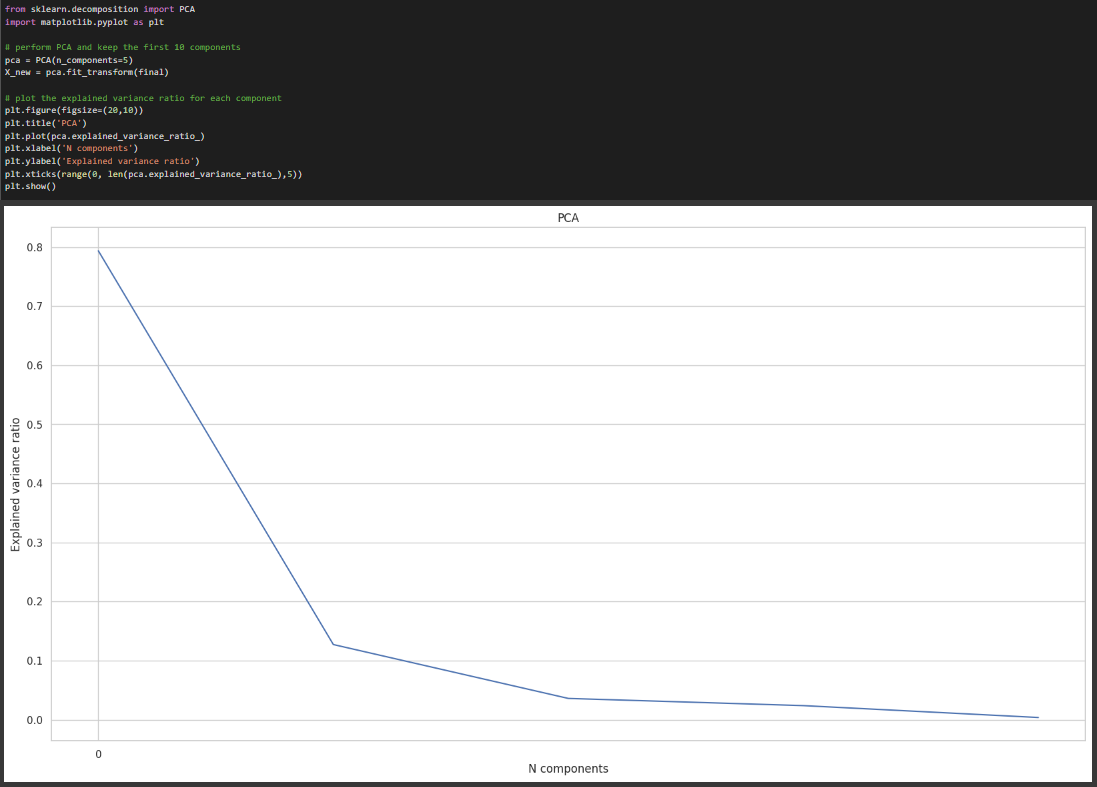
**Figure 2.5** *Count of Missing Values* **Figure 2.6** *Replacing missing values with NaN*

1. **FEATURE EXTRACTION USING PCA**

A vital stage in machine learning projects is feature extraction, which entails locating and choosing the most significant characteristics from the dataset. Principal Component Analysis (PCA) is a well-liked feature extraction method. By locating the principle components (PCs), which are linear combinations of the original characteristics that account for the majority of the variation in the data, PCA is a statistical technique that decreases the dimensionality of the dataset. PCA can assist in enhancing the effectiveness of the machine learning algorithm, lowering the danger of overfitting, and accelerating the training process by reducing the dimensionality of the data.

Standardizing the data by removing the mean and dividing by the standard deviation is the first step in doing PCA. The covariance matrix of the standardized data's eigenvectors and eigenvalues are then calculated via PCA. The primary components are represented by the eigenvectors, and the variance explained by each component is represented by the eigenvalues. The top n principle components are then chosen after the principal components are rated according to how much variance they explain.

Finally, utilizing the chosen primary components, the original features are changed into a new feature space. The machine learning algorithm can then utilize this modified feature space as input. In machine learning, PCA is an effective method for feature extraction.

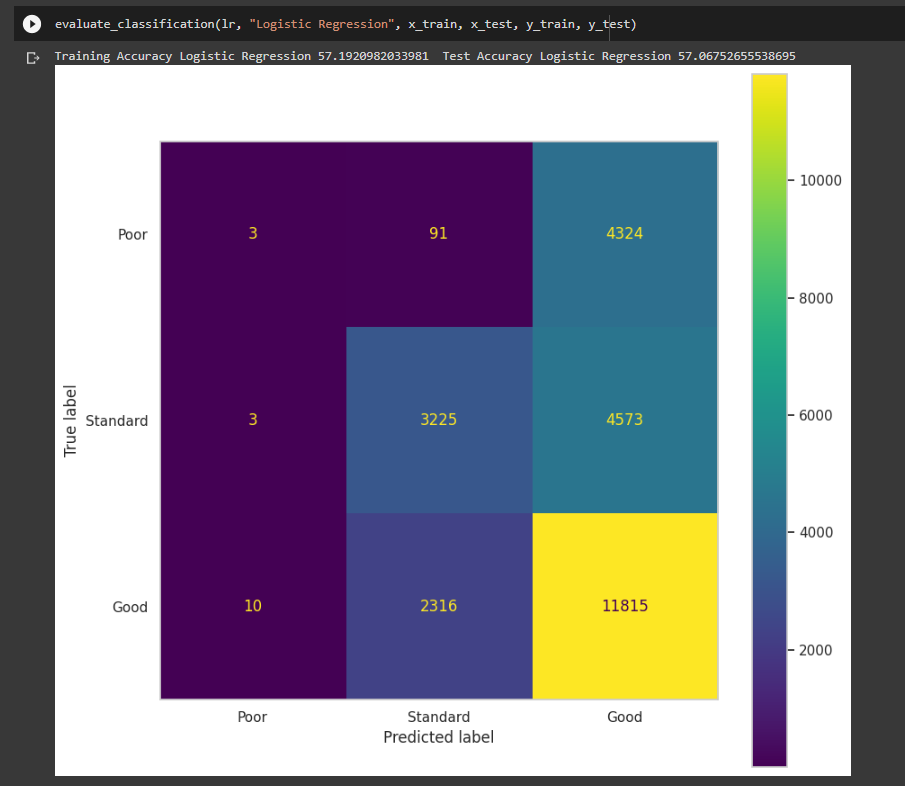
****

**Figure 3.1** *Application of PCA*

1. **ALGORITHMS**
2. **LOGISTIC REGRESSION**

Logistic Regression is a Machine Learning Technique that uses statistical methods that analyze connections between a dependent variable and one or more independent variables.When there are only two potential values for the dependent variable, it is frequently used to forecast binary outcomes. Based on the values of the independent variables, logistic regression estimates the likelihood that the dependent variable will fall into one of the two categories.

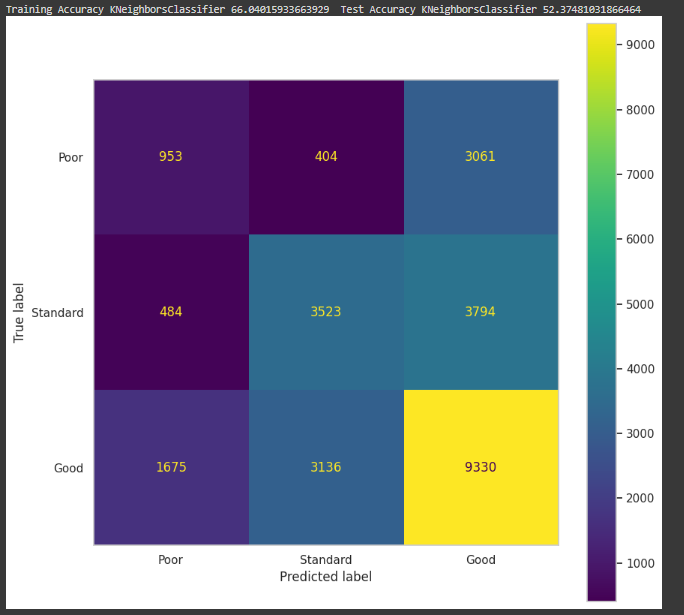
When it comes to determining the relationship between a borrower's credit history and their chance of defaulting on a loan, Logistic Regression may be an effective tool. A borrower's likelihood of defaulting on a loan may be predicted using logistic regression based on a variety of variables, including their credit score, payment history, debt-to-income ratio, etc. Lenders can use the model to determine whether or not to accept a loan application and, in that case, what interest rate to provide. In general, logistic regression can be a useful method for estimating credit risk and making data-driven lending choices.

**Figure 4.1** *Actual and Predicted Labels heatmap for Logistic Regression*

1. **K- NEAREST NEIGHBOR (KNN)**

Machine learning algorithms for classification and regression analysis include K-Nearest Neighbours (KNN). Being non-parametric, it makes no assumptions on the distribution of the data. In KNN, the closest data points in the training set are used to forecast the outcome of a new observation. The number of closest data points to take into account is represented by the value of k.

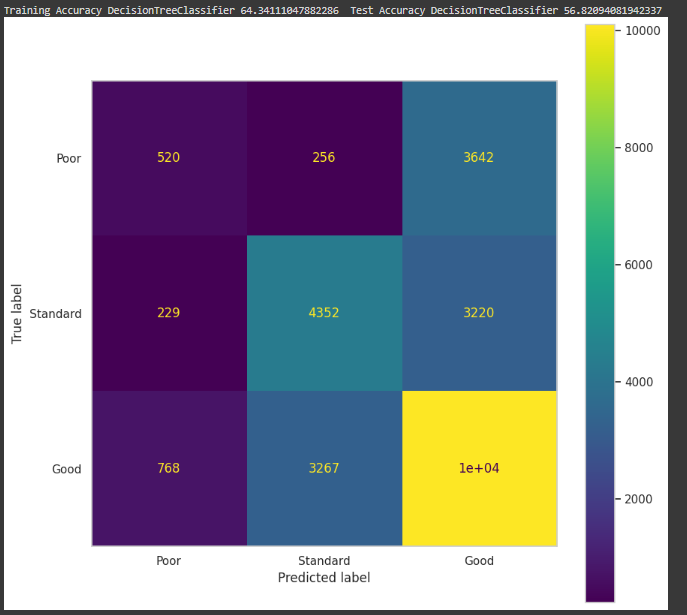
KNN may be used to examine the relationship that exists between a borrower's credit history and their chance of defaulting on a loan when it comes to credit score prediction. Based on the credit ratings of their neighbors in the training set, KNN can potentially be used to estimate a borrower's credit score. Lenders can use this model to determine which borrowers are most likely to fail on a loan and then take the necessary action, such raising the interest rate or rejecting the loan application. KNN may be used to increase the precision of credit score prediction models, either alone or in combination with other machine learning methods. KNN is a useful tool for forecasting credit risk and making wise lending decisions overall.

**Figure 4.2** *Actual and Predicted Labels heatmap for KNN*

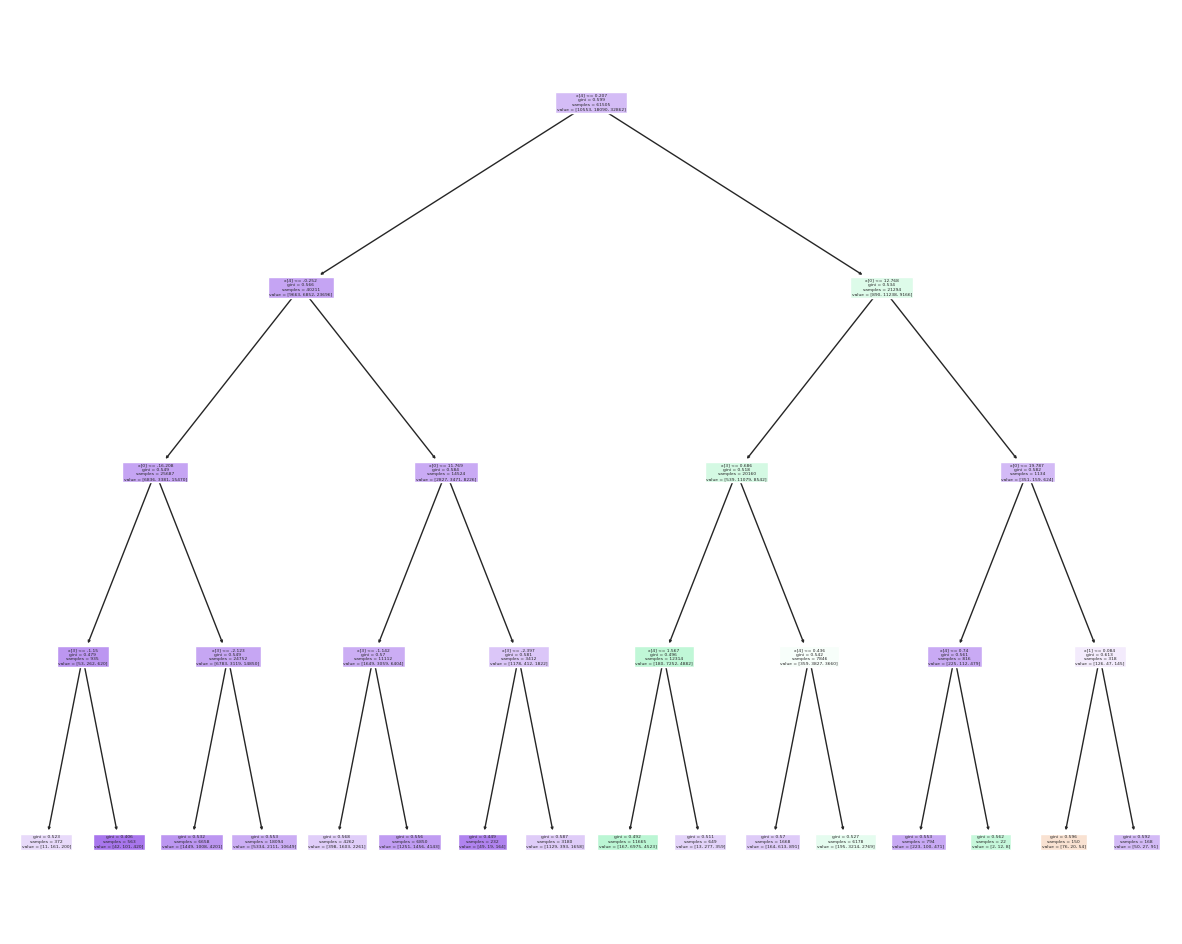
1. **DECISION TREE**

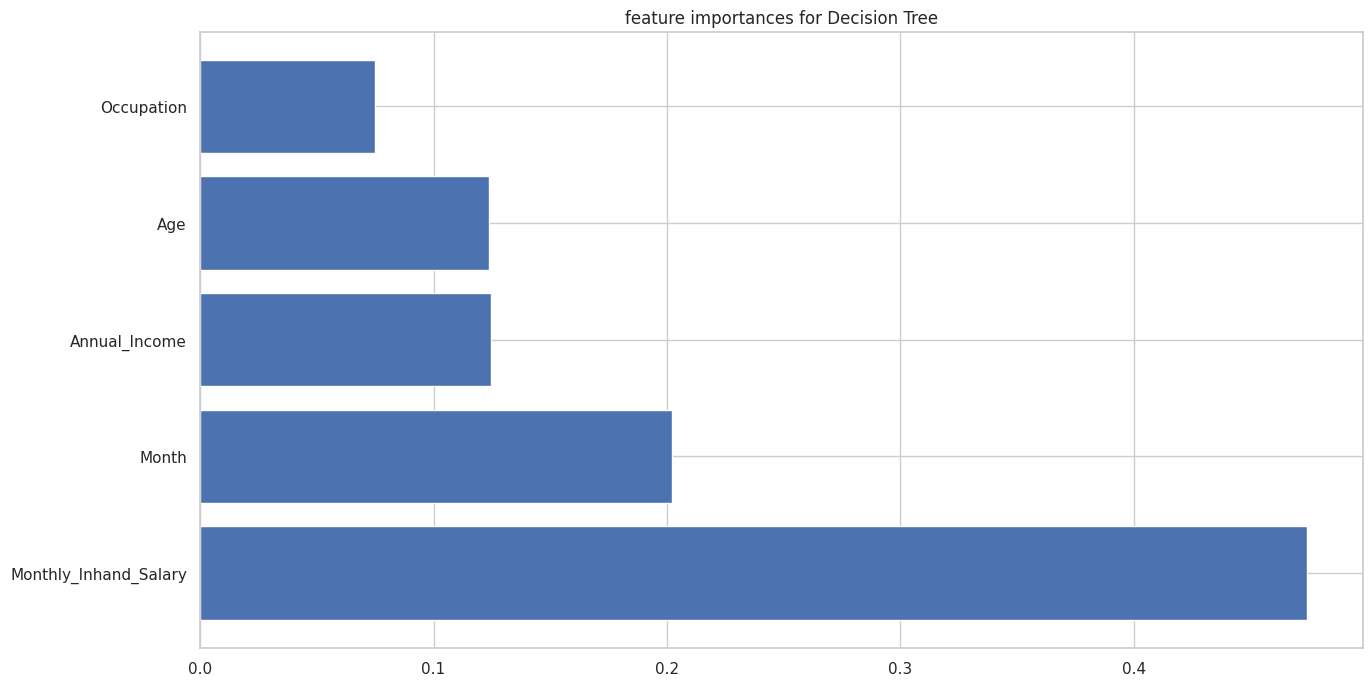
A popular machine learning approach for classification and regression analysis is called a decision tree. In order to create predictions, they use a non-parametric method that divides the data into smaller groups depending on the most crucial attributes. Decision trees are a popular option for many applications because they are simple to understand and visualize.

Decision Trees are often used to examine the connection between a borrower's credit history and their propensity to default on a loan in the context of Credit Score Prediction. A dataset with several elements, including credit score, payment history, debt-to-income ratio, and other pertinent variables, may be used to train the Decision Tree model. A new borrower's credit risk may then be predicted using the model based on their credit history. Decision trees may identify borrowers who are most likely to fail on a loan, which can assist lenders in making wise lending decisions. The accuracy of credit score prediction models can also be increased by combining Decision Trees with other machine learning methods.



**Figure 4.3** *Actual and Predicted labels heatmap for DT*

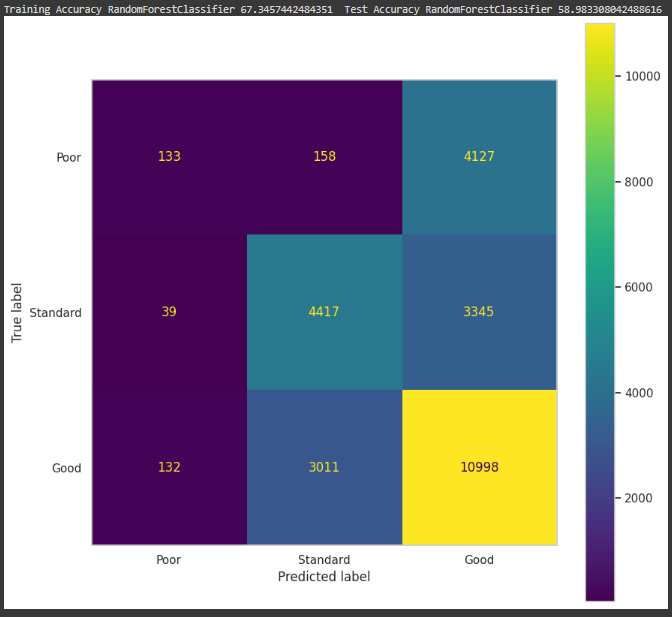
**Figure 4.4** *Decision Tree with max depth of 4*

**Figure 4.5** *Feature Importance for Decision Tree*

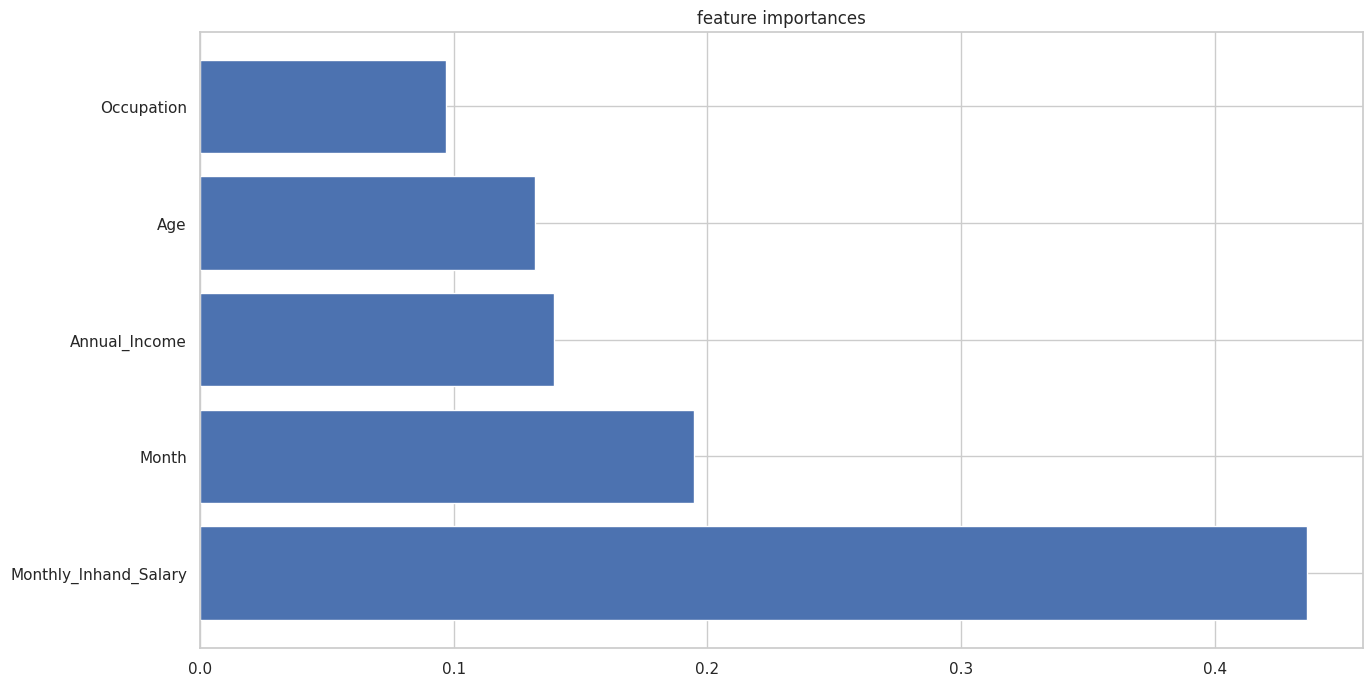
1. **RANDOM FOREST**

A sophisticated machine learning technique called Random Forest creates a group of decision trees to generate predictions. Each decision tree is constructed by the algorithm using a subset of features and data points from the training set that are chosen at random. The final forecast is then created by combining the predictions from each decision tree. The primary advantages of Random Forest are its excellent accuracy, resilience, and capacity for handling numerous input variables.

Random Forest may be used to examine the connection between a borrower's credit history and their chance of defaulting on a loan when it comes to credit score prediction. A dataset with several elements, including credit score, payment history, debt-to-income ratio, and other pertinent variables, may be used to train the Random Forest model. A new borrower's credit risk may then be predicted using the model based on their credit history. By identifying borrowers who are most likely to fail on a loan, Random Forest can assist lenders in making wise lending decisions. The accuracy of credit score prediction models can also be increased by combining Random Forest with other machine learning methods.



**Figure 4.6** *Actual and Predicted labels heatmap for RF*

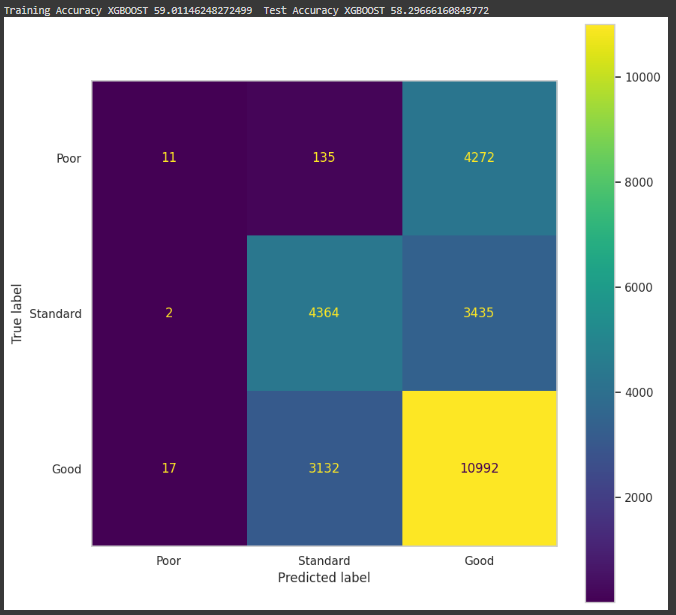


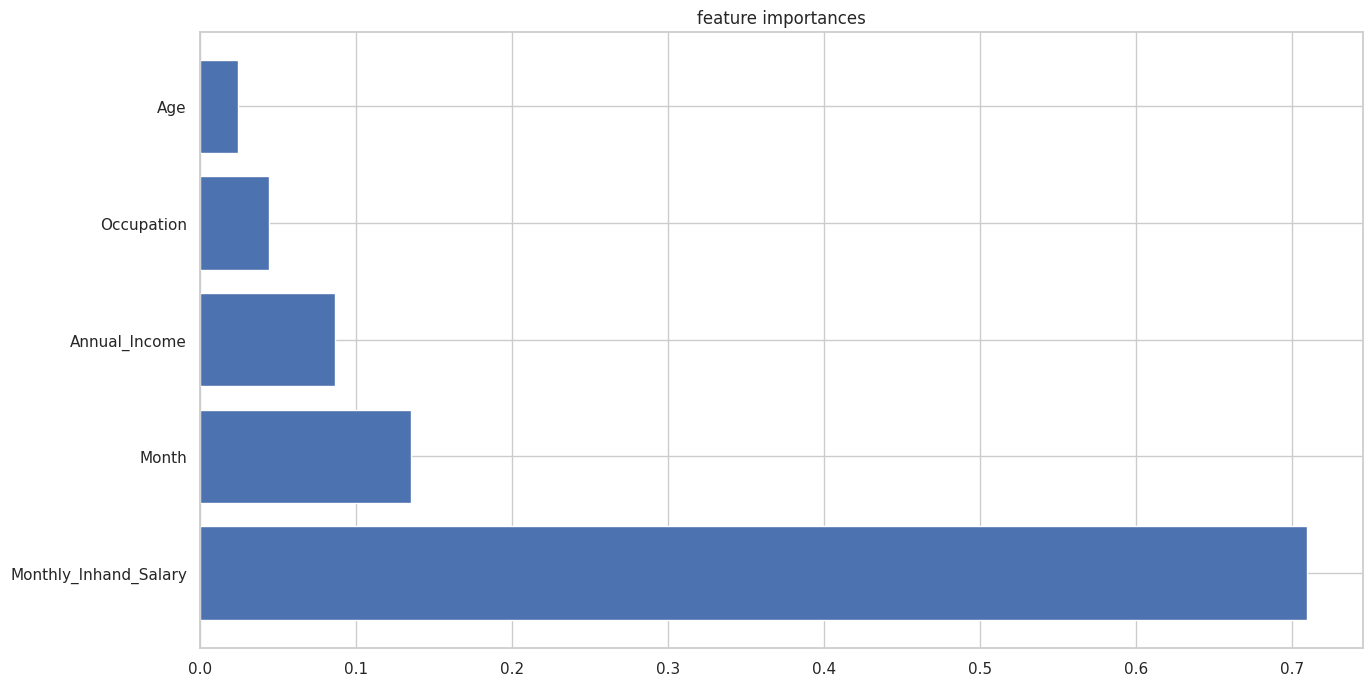
**Figure 4.7** *Feature Importances for RF*

1. **EXTREME GRADIENT BOOSTING (XG Boost)**

A well-liked machine learning approach called XGBoost (Extreme Gradient Boosting) is utilized for both classification and regression analysis. It is a modification of the Gradient Boosting technique that controls overfitting by using a more regularized model and a cutting-edge algorithm to handle missing variables. XGBoost is a well-liked option for many applications because of its great accuracy and efficiency.

XGBoost may be used to analyze the correlation between a borrower's credit history and their chance of defaulting on a loan in the context of credit score prediction. A dataset with different elements, including credit score, payment history, debt-to-income ratio, and other pertinent variables, may be used to train the XGBoost model. A new borrower's credit risk may then be predicted using the model based on their credit history. By identifying borrowers who are most likely to default on a loan, XGBoost can assist lenders in making wise lending decisions. The accuracy of credit score prediction models may also be increased by combining XGBoost with other machine learning methods.

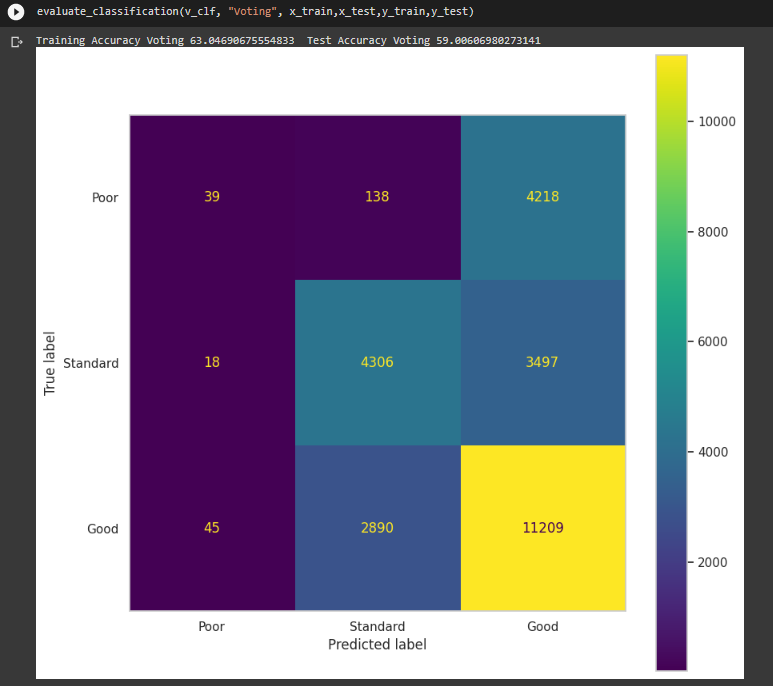
**Figure 4.8** *Actual and Predicted labels heatmap for XGBoost*

**Figure 4.9** *Feature importances for XGBoost*

1. **VOTING CLASSIFIER**

The final prediction is formed by aggregating the predictions of all the models using a voting mechanism in a voting classifier, a sort of ensemble learning in machine learning when numerous models are trained on the same dataset. Either "hard voting" or "soft voting" may be used as the voting method. In a hard vote, the majority predicted class is chosen as the final forecast, but in a soft vote, the final prediction is made using the average predicted probability of all models.

By integrating the capabilities of many models, voting classifiers are used in machine learning to increase the accuracy and resilience of models. When the separate models have diverse strengths and limitations, this strategy might be extremely helpful because the combination could potentially overcome the limitations of individual models.

**Figure 4.10** *Predicted and Actual labels heatmap for the Voting Classifier*

1. **RESULTS & EVALUATION**

Results illustrate how six different machine learning models performed on a dataset when tested against the training data. The Random Forest model fits the training data well and generalizes to new data well, as seen by its best train score of 0.770775 and highest test score of 0.691009.

In comparison to the Random Forest model, the Voting model has a comparable test score of 0.690061 but a slightly lower train score of 0.730469. This shows that the Voting model, which has comparable generalization performance to the Random Forest model, would be an acceptable substitute.

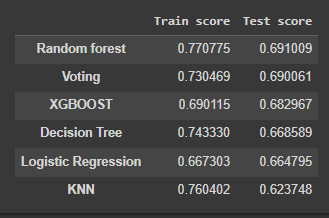
A train score of 0.690115 and a test score of 0.682967 for the XGBOOST model show that it satisfactorily matches the training data and may perform well in terms of generalization.

The Decision Tree model's train and test scores of 0.668589 and 0.743330, respectively, indicate that it may be overfitting the training data and may not generalize well to new data.

The Logistic Regression model has a train score of 0.664795 and a test score of 0.67303, indicating that it performs similarly on both the training and test sets of data and may have high generalization capabilities.

The KNN model, which has the greatest train score of 0.760402 but the lowest test score of 0.623748, may be overfitting the training set of data and not generalizing effectively to fresh data.

In conclusion, the models with the greatest test results—Random Forest and Voting—might be suitable candidates for additional testing and optimization. Before drawing any firm conclusions, it is crucial to thoroughly assess the models' performance on various datasets and test their generalizability.

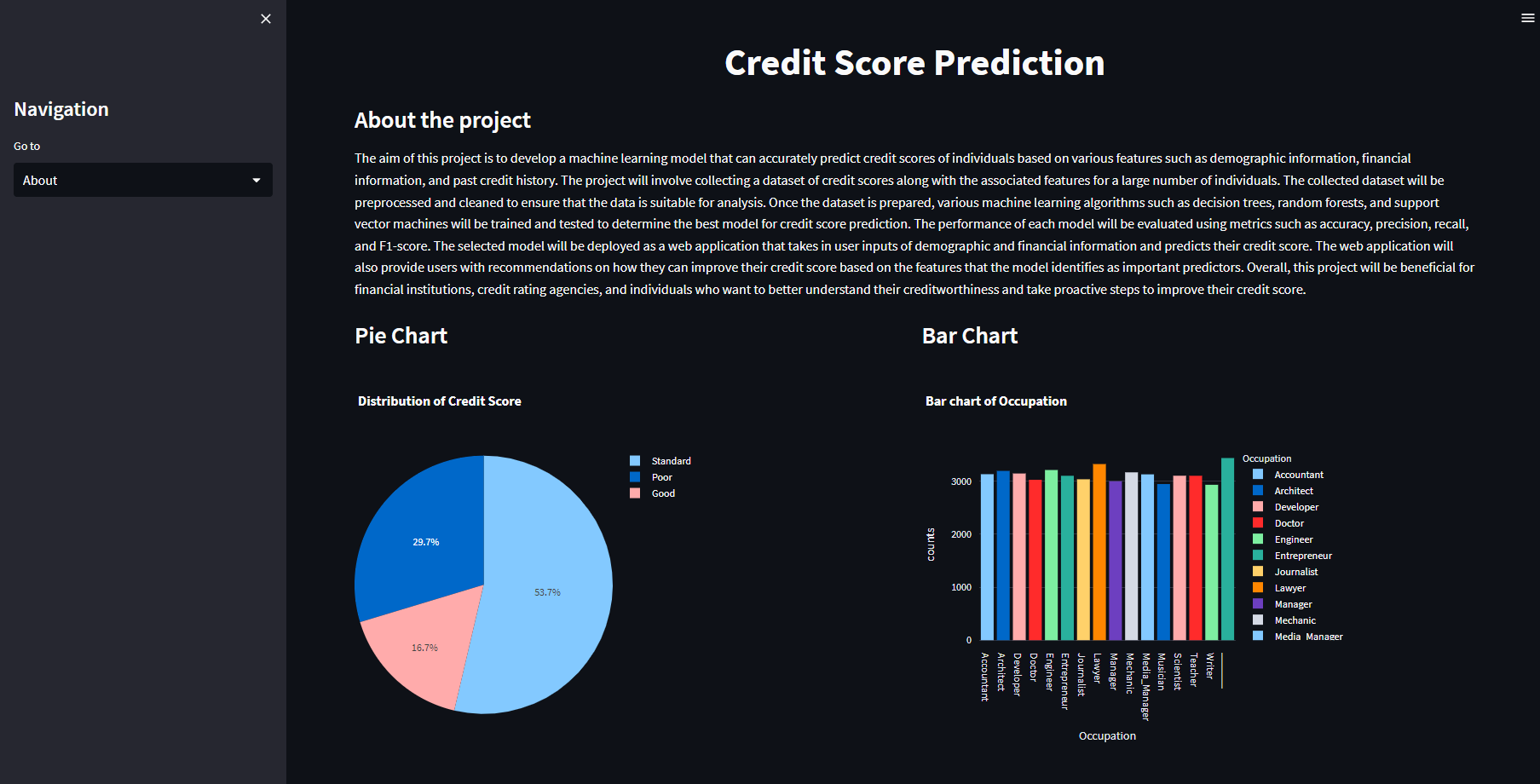
**Figure 5.1** *Comparison Between Algorithms*

1. **FRONTEND**

Deploying a machine learning model as a GUI using Streamlit can greatly improve its usability and accessibility. In the case of credit score prediction using Random Forest, a Streamlit application can allow users to input their financial information and receive an instant credit score prediction. The deployment process typically involves preparing the model for deployment, creating a Streamlit application, and then hosting the application on a server. Streamlit provides a user-friendly interface for building and deploying web applications, and the Random Forest model can be easily integrated into the application using Python. Once the application is deployed, users can interact with the model through the GUI, providing an intuitive and easy-to-use interface for credit score prediction. With this deployment, financial institutions or credit agencies can use the model to provide quick and efficient credit score predictions to their customers.

* 1. **ABOUT PAGE**

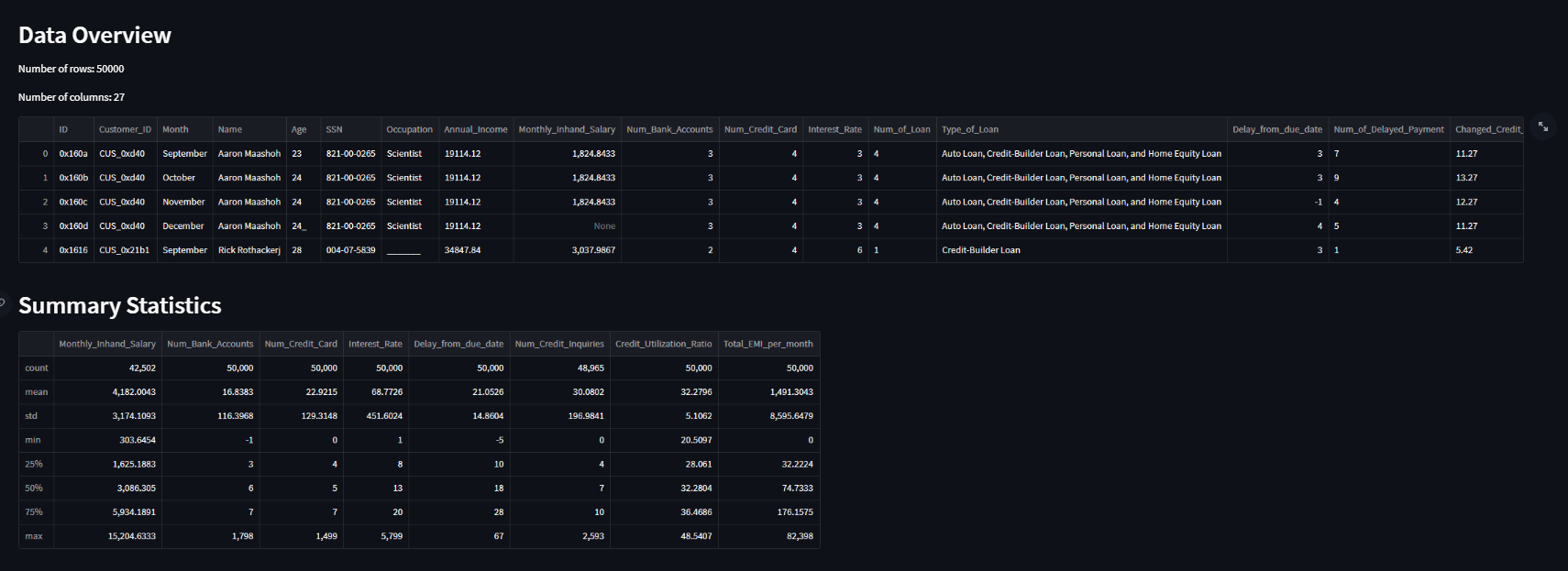
The aim of this project is to develop a machine learning model that can accurately predict credit scores of individuals based on various features such as demographic information, financial information, and past credit history. The project will involve collecting a dataset of credit scores along with the associated features for a large number of individuals. The collected dataset will be preprocessed and cleaned to ensure that the data is suitable for analysis. Once the dataset is prepared, various machine learning algorithms such as decision trees, random forests, and support vector machines will be trained and tested to determine the best model for credit score prediction. The performance of each model will be evaluated using metrics such as accuracy, precision, recall, and F1-score. The selected model will be deployed as a web application that takes in user inputs of demographic and financial information and predicts their credit score. The web application will also provide users with recommendations on how they can improve their credit score based on the features that the model identifies as important predictors. Overall, this project will be beneficial for financial institutions, credit rating agencies, and individuals who want to better understand their creditworthiness and take proactive steps to improve their credit score.



**Figure 6.1** *About Page in Streamlit*

* 1. **DATA EXPLORATION PAGE**

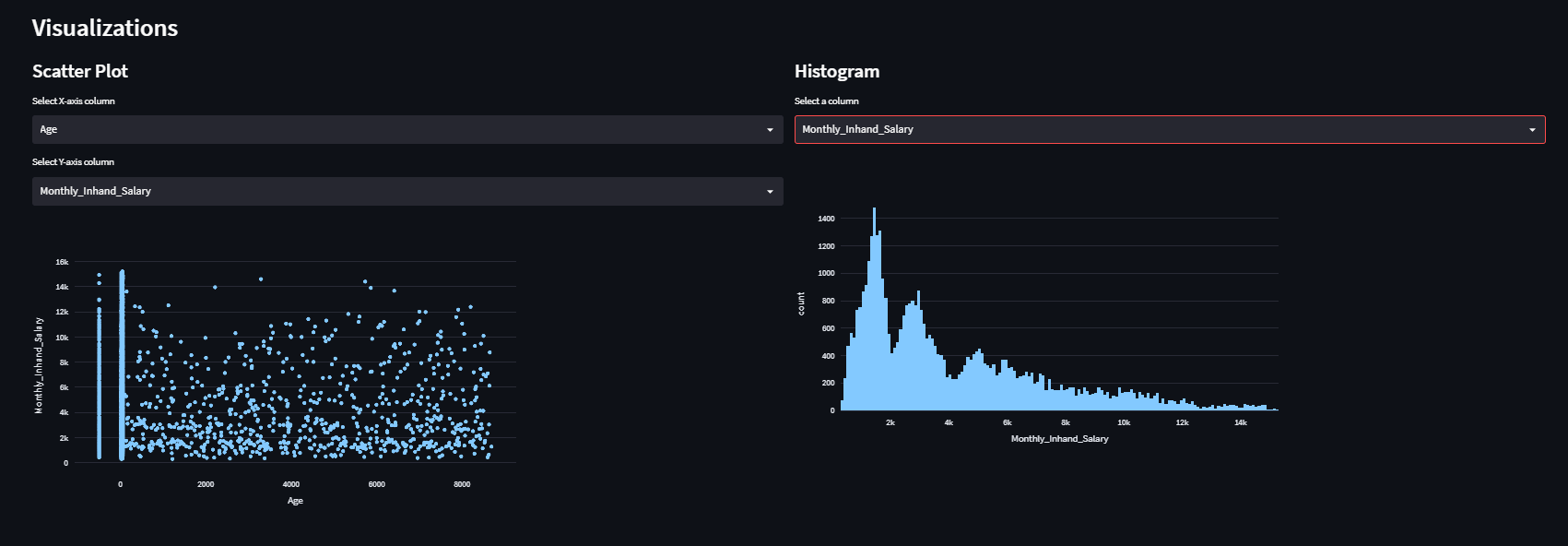
A data overview is presented, which includes 50,000 rows and 27 columns. This dataset is from the testing dataset. As that is the one that is used for the prediction. A summary of the dataset is also provided wherein the count, mean, standard deviation, minimum and maximum value, 1st quartile, and 3rd quartile values are displayed.

**Figure 6.2** *Testing dataset description*

Then, a visualization option is presented to the user wherein he can view two different plots, scatter plot and histogram.

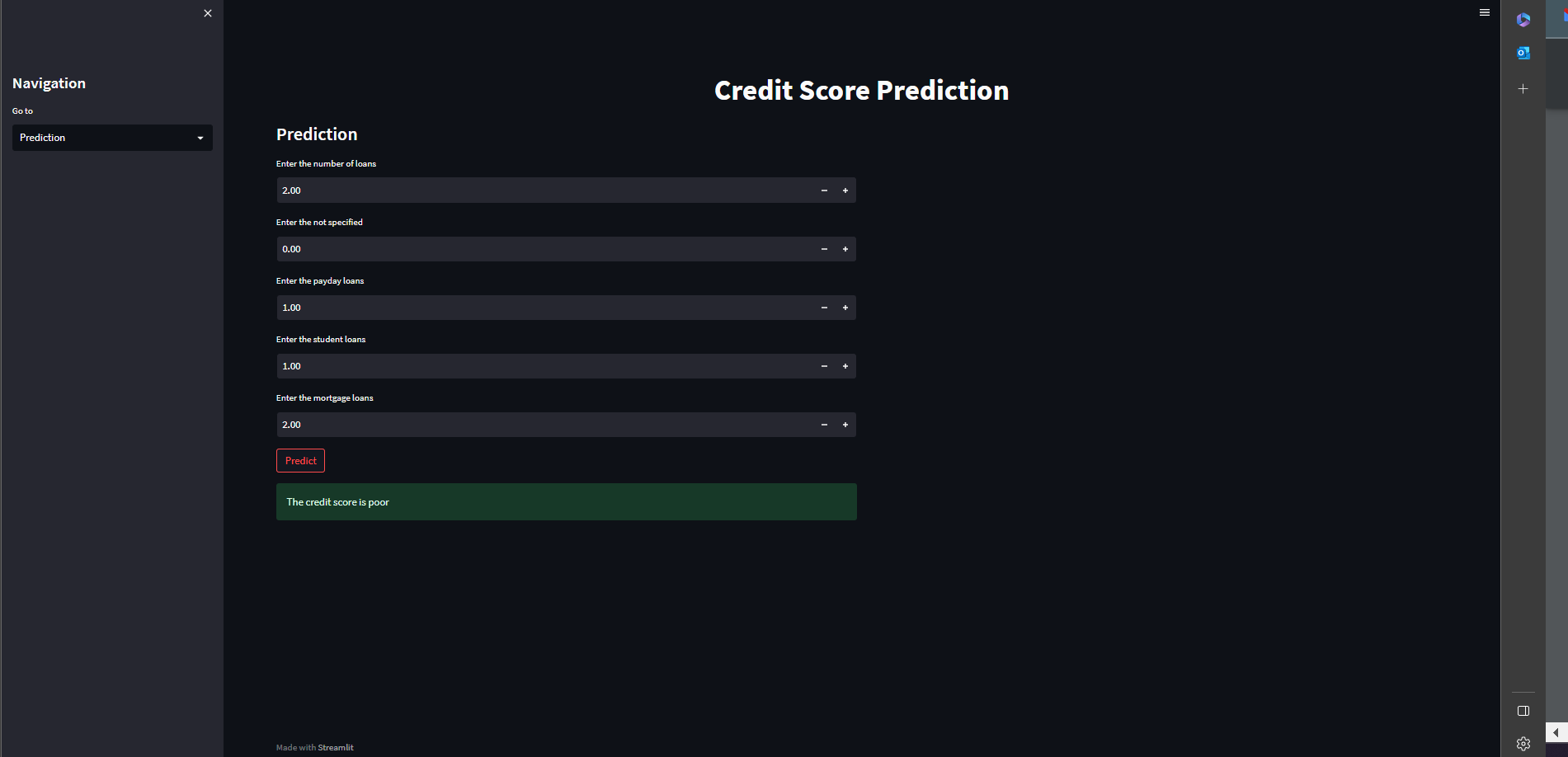
When predicting credit scores using ML, the scatter plot aids in visualizing the link between two factors. It enables us to spot correlations between the variables, such as whether they are positively or negatively associated, and patterns and trends between them. We may learn more about how these factors are impacting the credit score by analyzing the scatter plot, and we may be able to spot any outliers or abnormalities that require attention.

When using machine learning to predict credit scores, a histogram is a useful tool for visualizing the distribution of data for two variables. It displays the frequency with which various values of the variables occur and might draw attention to any patterns or trends in the data. We may decide how to preprocess the data or select the best machine learning models by looking at the histogram to acquire insights on the distribution of the variables, such as whether they are regularly distributed or skewed.

**Figure 6.3** *Visualization using Scatter Plot and Histogram*

* 1. **PREDICTION PAGE**

Here, we chose the best features that were presented during the feature selection of the random forest classifier. The variables were the number of loans, ‘not specified’, payday loans, student loans, and mortgage loans. After the user inputs the values for the following attributes, it will either output poor, standard, or a good credit score.



**Figure 6.4** *Prediction Page*

**Contribution:**

**Agrim:** Literature Survey, Feature Extraction and Selection using PCA, Random Forest, XGBoost, Voting Classifier

**Aryaman:** Abstract, Introduction, Dataset, Data Cleaning and Preprocessing, Logistic Regression, Decision Tree, KNN

Both contributed towards the backend and the frontend of the code as well as the designing and execution of the project and report.