RESEARCH PAPER

Efficient Credit Risk Analysis through Machine Learning and PySpark

T.P.S. Kumar Kusumanchi[1*], Boye Jyothi Mercy[2], Pasupuleti Sri Charan[3], Obulasetty Sai Kumar[4], Shaik Ahmadsaidulu[5]

[1]Department of IoT, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522501, India. Email: satishkumar8421@gmail.com

[2]Department of IoT, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522501, India. Email: 2200100014@kluniversity.in

[3]Department of IoT, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522501, India. Email: 2200100029@kluniversity.in

[4]Department of IoT, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522501, India. Email: 2200100051@kluniversity.in

[5]School of Electronics Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India. Email: Shaik.ahmadsaidulu@vit.ac.in

***Corresponding Author: T.P.S. Kumar Kusumanchi**
Department of IoT, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522501, India
Email: satishkumar8421@gmail.com
Received: 31st May, 2026; **Revised:** 8th June, 2026; **Accepted:** 10th June, 2026; **Available Online:** 13th June, 2026

ABSTRACT

Credit risk is one of the aspects which is very important in financial institutions like Commercial Banks, Credit Unions, Investment Firms, and Insurance Companies, etc to estimate the likelihood of borrower's default and improve the credit decision making with improved analysis. The Banks which intake a large data may find it difficult to manage customer interest that leads to poor customer satisfaction. This Research utilizes some big technologies through PySpark and incorporates Mongo DB as a scalable NoSQL database for efficient data storage, retrieval and operations. This Research mainly works on analyzing key features such as income, education, financial history, demographics, through applying various supervised machine learning models such as Random Forest, SVM, Gradient Boosting, XGBoost, MLP, Stacking Classifier with different ensemble methods to classify applications into risk categories. Here these models give some evaluation metrics as accuracy, precision, recall. These metrics help us decide the final accurate model with best output results. The integration of Mongo DB database with PySpark enhances the speed and model's performance on large datasets. Effectiveness of real time prediction also enhances with these models. The Final findings of this research aim to improve the reliability and accuracy of credit risk predictions, data driven risk-management, supporting robust, improve the stability across financial sectors.

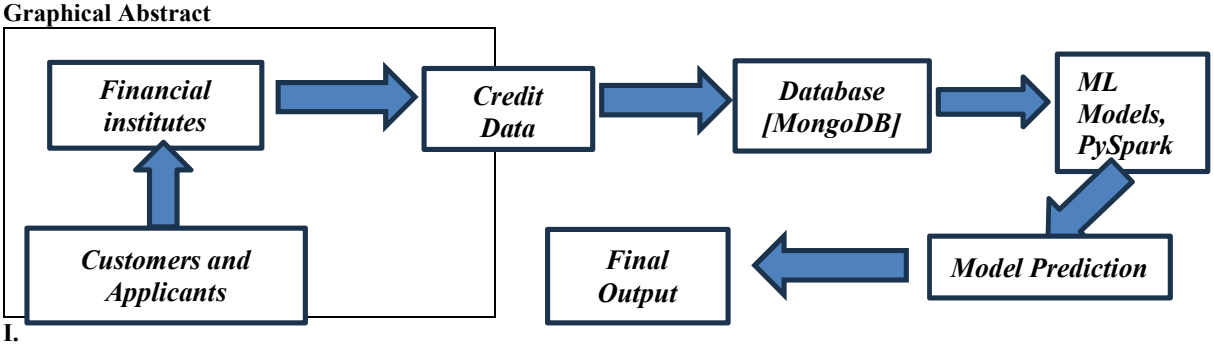
Keywords: Credit risk, Financial institutions, Ensemble methods, Credit risk prediction, Supervised machine learning.
How to cite this article: Kusumanchi TPSK, Mercy BJ, Charan PS, Kumar OS, Ahmadsaidulu S. Efficient Credit Risk Analysis through Machine Learning and PySpark. Int J Drug Deliv Technol. 2026;16(59s): 1478-1489. DOI: 10.25258/ijddt.16.59s.165

Source of support: Nil

Conflict of interest: None

(MLP) and Autoencoders to analyse the higher

Graphical Abstract



I. Introduction

Over the last few years, there has been a tendency in the financial industry to implement data analysis solutions for the enhancement of business decisions, including credit risk evaluation. In today's world credit cards and loans have become one of the people's necessities, without a credit or loan people find it difficult to buy goods and services to maintain their monthly expenses, etc. Initial Credit cards offer an easy method of borrowing money and function as replacement for cash.

The loans which are given by the banks/financial institutes have become one of their important income sources to maintain their stability and economic growth. Financial institutions such as banks, investment firms, credit unions need to decide whether to approve/reject an application submitted by the customers. So, there are Different factors involved to decide whether a certain customer is good for the repayment of the credit taken. For this reason, credit risk prediction is one of the important task that is very challenging to face in financial sector by all the financial institutions globally [1]. The credit risk analysis mainly helps the financial processors to reduce their losses with the customers who delay the payment.

Previously, credit scoring in applications approval was used to be a simple statistical model, as years passes this has been upgraded to advance techniques that analyses the credit scoring of the clients much better. These advanced techniques offer accurate, reliable, and efficient solutions for credit forecasting. Applications submitted by the clients are on a large scale that range from a few thousand to lakhs according to the financial institute. In respect to tackle this problem we have used PySpark an open source which is a Cross platform, distributed-computing, OS environment for Processing big data as well as training the Machine Learning Model at large scale.

PySpark is to extend its support for processing the data which would be ideal for working with big data and used for machine learning algorithms. This PySpark also works for both supervised and unsupervised machine learning models such as Deep Learning Models including Multi-Layer Perceptron dimensionality datasets which have a huge volume with analysis.

There are machine learning models like Random Forest, Gradient Boosting, XG Boost, and Voting classifier which unites several classifiers, these are able to take the best one out of the other without having any problem with other classifiers. There is a similar comparison for some of the models like K-Nearest Neighbors (KNN), Logistic Regression, Decision Trees, Support vector Machines (SVM) and Naive Bayes. Models like the Logistic regression offer fast and accurate analyzing results while maintaining simplicity, making it select one of the ideal models for insights and performance metrics for the research [2]. In the credit risk analysis not only, improved accuracy is done but also the factors leading to the default of the loan.

The models are used to enhance the borrower behavior forecasts and to help financial institutions to have more enhancement of risk management. Generally, the number of applicants who have defaults would be minority if we compared with applications without defaults, this creates a problem for models where it predicting the majority class and failing to identify the minority class [3].

So, such complications need to be improved with the enhanced models to work with large data sets and the prediction capabilities also need to be improved. The PySpark used in the research helps to fill the gaps by distributed-computing and various machine learning architecture with the ensembled classifiers. There may be many challenges faced by the financial institutions but the credit risk, which is bad customers, delays in payments, late monthly payments could hit them more and cost them more.

The credit risk has very dynamic and complex nature with different factors combined with in view of economic and market trend a very advanced strategies are required [4]. The model we have developed is not a whole sum advance model, but this is the best model built by us. The poor credit risk analysis can lead to a huge amount of financial losses which can impact financial institutes equally to economy.

The ability who are applicants are actually or more

likely the defaults of loans are essential for long term sustainability [5]. The banks also need to maintain the best track records of every customer possible to avoid and future risks to the banks. In our real time prediction tool, the better the data given the accurate the credit risk prediction is accurate. The preparation of each credit given to us not only gives financial institutions revenue but also growth toward the future. Finance is one such valuable asset that is very important for us humans, where the credit is one such source where we can fulfill our dreams.

II. Literature Survey

The credit risk assessment deals with different types of approach in working and research. One such model proposed from a Taiwanese banking data, where multiple data mining algorithm models with customer related problems used for credit default prediction [6]. There are different machine model which are investigated by the researchers with four machine learning algorithms were to give a best result with 82% in neural networks compared to other models [7].Some other research also present a customer's default prediction in which six data mining techniques were used for which results showed neural networks are best in predictions [8].One of noticing works were random forest, bagging, boosting with neural networks are applied for analyzing the defaulters payment data to obtain the most promising results among all [9]. There is an ensemble credit scoring model developed by integrating bagging and stacking model with a trainable fuser. For this they have used data from multiple datasets like German, Australian, We.com, Lending club. The results outperformed all the traditional ensemble models making a benchmark [10].

The very recent work on the credit risk is focused on the ensemble strategy where prediction from multiple classifiers is done to improve the robustness of individual models for credit scoring [11].Then there is a work in which deployed multi financial models to get the results showing stakeholders with all these predictions can assess more risk effectively with accurate decision making[12]. Through the years traditional models are quite helpful in their way yet as years pass the traditional methods exhibit limitations. A common scenario is that when dealing with imbalanced datasets for credit risk modelling [13]. The following class imbalance problem was addressed by some techniques which have been very useful like SMOTE and ENN. A model with similar problem of class imbalance was solved by SMOTE and neural networks resulting in an accuracy 98.6% that is very and best to be noted in credit risk prediction [14]. There are also some problems caused by the class imbalance, which is overfitting that would cause trouble in machine learning models like reputation management, criminal deception and much more [15].

The financial institutes must focus first on loan defaults so that they can deal with losses and reduce them a lot. There is a model which was integrated by support vector machine, a best output was obtained that has outperformed the logistic regression model with an accuracy: 82.12%, precision: 0.7831. This result has advised some of the financial institutions to check on loan defaults more [16]. For much better results, we can use backpropagation neural network (BPNN) that has specialty like learning, adapting, acquiring various knowledge, then train on uncertain knowledge on its own with the data. The research work also proves to us that the model integrated with backpropagation will reduce credit risk, eventually regulate the data, also with better decision making in customer operations [17].

So, as the following research works continue the ML techniques have demonstrated some of best results while dealing many risk analyzing problems with complex scenario, which has been easier for understanding to the stakeholders making them valuable [18]. Continuous research on this has received significant attention for using machine learning leading on more integrating with SHapley Addictive exPlanations [SHAP] for explaining on different applications to improve the credit prediction [19]. The predictions of ML models utilizing the SHAP values on student loan defaults are also applied. The following SHAP techniques gave the student loan default insights with factors such as exam scores, academic performance, and scholarships they have received which will influenced their applications approval [20].

The need of the credit risk prediction to financial institutions is more, but the development of the predictive models that also includes explainable artificial intelligence is very complex and need some time to improve in transparency with some challenging cases in the financial institutions [21]. According to a research work done, machine learning is growing more significantly every year in financial services. In UK the AI/ML environment is observed, where it is making financial operation more advance and convenient. The UK financial service domains like credit scoring, financial distress, robot-advising and algorithm trading proved to be their best domains with a significant influence in their services. They also showed that the ML applications demand in predicting the defaulters with the credit risk prediction is continuing to grow [22]. So, there are various research works done around the world that have helped financial institutions. This use of different ML models is quite challenging, but it is very necessary in predictions problems and decision making [23].

The development has led to possibility of borrower's evaluation to enrich the prediction model. They have also started the use of images, interviews, text, social information and mobile technology integration

which would soon give a multi-dimensional accessibility on credit risk prediction [24]. Finally, we should also note that the increase in demand for the credit risk also increases its risk and grows indeed. So, the development of these models should be done by keeping in mind factors like risk, loan amount, rate, and loan terms [25].

The research done by us has an overview of these problems individually which have the similar best solutions. The research kept progressing by ensemble classifiers and models that utilized in the previous research, but each research has solution to few in each where we tried to note and address as many as possible. The best feature of our research is use of PySpark which would advance the research with large datasets in financial institutions. The previous research works also inspired us to solve the class imbalance problem with SMOTE. The following research is done while maintaining significant advancements and solving the challenges. The following research progresses to the need for scalability, explainable and robust credit risk prediction on a large-scale data of the financial institutions.

III. Methodology

This research is mainly taken based on a secondary dataset from Kaggle and from this a project was built which obtained a to use a loan prediction and approval by a structured means of machine learning model. The building methodology in this project involves few key stages which are very important in building an accurate model for credit risk prediction. The stages which are involved are Data extraction from a NoSQL database, Preprocessing and Cleaning of the dataset, Feature engineering to obtain more informative features, Handling the class imbalance by using SMOTE, Model training using multiple classification algorithms, The performance evaluation, the pipeline is designed with PySpark for scalability and efficiency. The following research project is done through the few main key steps:

A. Dataset

The dataset of this project is from Kaggle name "CreditRisk.xlsx" which is used by many other researchers in the similar type of projects. The dataset file we name in the project as "credit risk.csv" to be handled by the model.

The dataset contains different variables of the customers in order. The factors included in columns are Loan ID, Gender, Married, Dependents, Education, Self Employed, Applicant Income, Co-applicant Income (if applicable), Loan Amount, Loan Amount Term, Credit History, Property Area. These factors are given with values to be noted which are for example Loan ID - LP001002, Gender - Male/Female, Married - Yes/no, dependents (if present like parents, children etc.) - 0/1, education - Graduate, Not Graduate, self-employed - Yes/No Applicant income (monthly) - 5000, Co-applicant (if applicable) - 6000, Loan amount - 15000, Loan amount term - 360 days, Property area - Urban/Rural (for security).

From these factors the final Loan status either approve or rejected will be predicted by model as 0/1 in a binary system value, where the 0 represents as No (applicant loan rejected), 1 represents as Yes (approved).

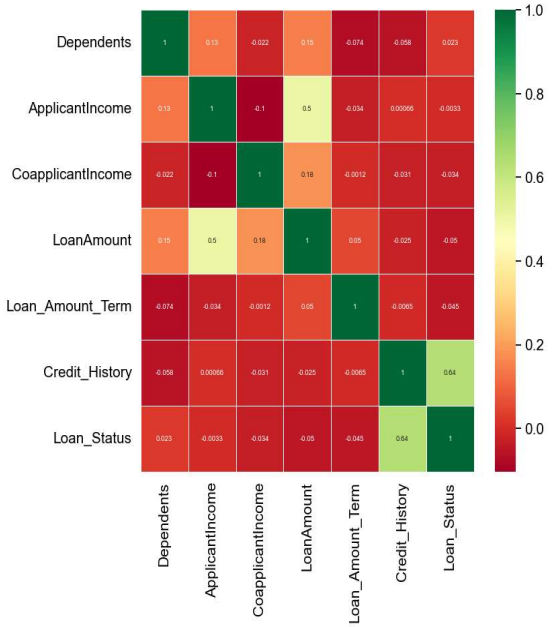


FIGURE 1. Heatmap for the variables of dataset.

B. Data Loading.

The total research project needs a proper place to store data without problems such as errors, data loss, data breach, etc. The data which is given must be easily processed without any problem of speed and performance. The new data given by the user has also needs to be stored in the following storage for further future use.

The MongoDB is the most suitable database for this so, we have taken MongoDB where the requirements are fulfilled for our project. The MongoDB is a NoSQL database known for its flexible document-based storage. The data is extracted with help of pymongo, used by establishing a connection to MongoDB instance and reading the collection of loan records. The data here is retrieved as a list of dictionaries, then it is converted to Pandas Data Frame for intermediate processing in Data Frame for intermediate processing in the data loading part. Here after a Spark Session is initialized using PySpark which is one of the main key features of this research project, In the following process the Pandas Data frame is converted into a PySpark Data Frame using the spark.createDataFrame() method. This conversion of the data enables the distributed data processing capabilities that are provided by PySpark.

This process is crucial for handling large datasets efficiently in the subsequent steps, in which the speed and performance will be improved for the large datasets used by the financial institutions.

C. Data Preprocessing

The Data preprocessing is used in this project mostly to improve quality and usability of the model training dataset. The missing values are identified and handled accordingly here. This is done as numerical columns with missing values are first imputed using mean values and categorical columns are filled with the most frequent category. The Categorical features here are Married, Education and Self Employed which are encoded using the String-Indexer and OneHotEncoder in the PySpark, this convert them into numerical vectors that are suitable for the following machine learning algorithms. The PySpark pipelines is where the transformations are mainly performed to maintain the efficiency and scalability of the model. After the following preprocessing is done the Spark Data Frame is converted into a Pandas Data Frame using the Pandas() method to adjust the model for better training and evaluation with use of libraries which would require the in-memory data, for example Scikit-learn.

D. Handling Class Imbalance

The dataset will be examined for a class imbalance in the target variables, as soon as the preprocessing is completed. In simple terms to explain the financial institution have the data of both the loan approved customers and rejected customers (who mostly haven't cleared their past credit loan would need to again reapply for loan mostly). So, both variation in the data where the approved customers would be minimum in number and the rejected would be maximum which are more unevenly distributed. This creates a negative impact on the model's performance.

To overcome this kind of problem Synthetic Minority Oversampling Technique (SMOTE) is used on the PandasDataFrame. The SMOTE will generate synthetic samples for minority class by interpolating between existing samples and balancing the class distribution for the data. This makes the model performance improve and handle the majority and minority data class distribution.

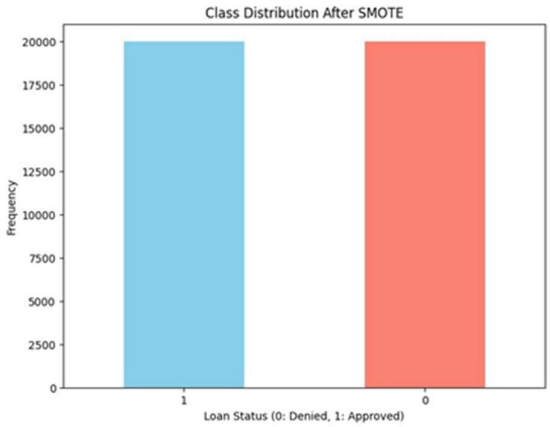


FIGURE 2. Class Distribution After SMOTE.

E. Model Training

In the following research project, we have taken Several machine learning classification models which are trained and evaluated to predict the loan approval according to the credit risk.

The main Goal for this to compare and analyze the performance of both the individual models and the ensemble approaches. Basically, the models were implemented using the Scikit and XGBoost libraries on the preprocessed and balanced dataset to have more convenience, better prediction without any errors to occur in the training process of models.

The XGBoost is a model algorithm that is efficient as well as scalable implementation of the gradient boosting that uses decision trees as a base learner. The XGBoost works in handling the missing values, data regularization, tree pruning, while making the model more robust and performing high on the structured data.

The Decision Tree is known for its simple yet powerful model that splits data with feature values to form a tree structure that is very used to interpret complex datasets. This solves the problem of overfitting in datasets. The decision tree helps a lot in model training to enhance the accuracy of the model.

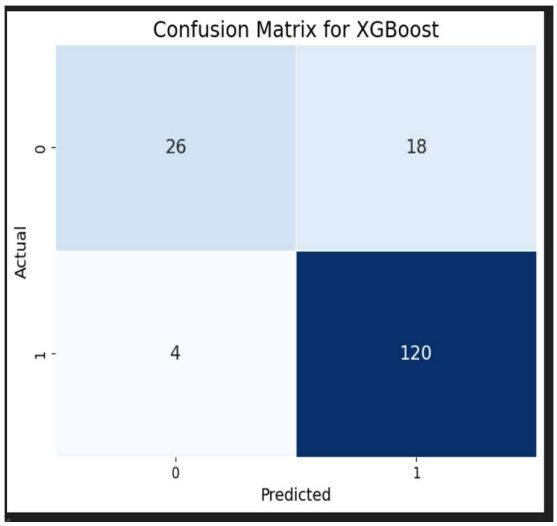


FIGURE 3. Confusion Matrix for XGBoost

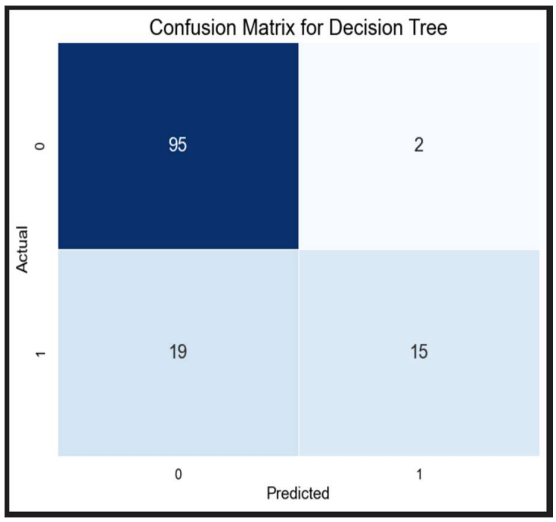


FIGURE 5. Confusion Matrix for Decision Tree.

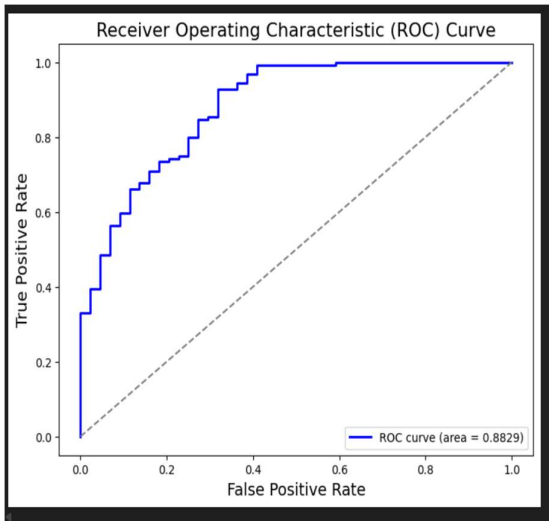


FIGURE 4. ROC Curve for XGBoost.

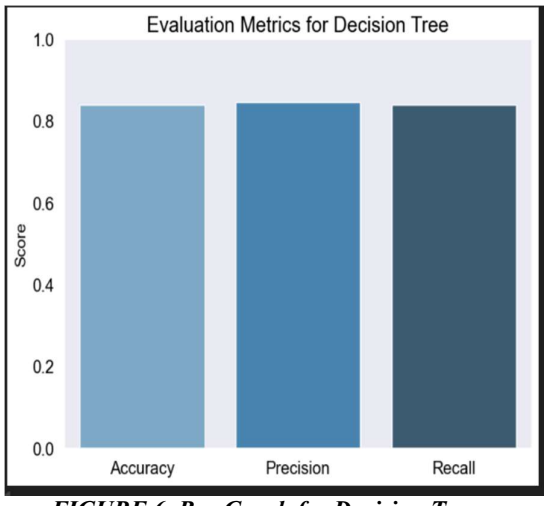


FIGURE 6. Bar Graph for Decision Tree.

Support Vector Machine, which is known as SVM in popular is a supervised machine learning model that is used to find the optimal hyperplane so it can separate data points of different classes in the datasets. The SVM is mostly used in high-dimensional spaces, yet this is very powerful and useful while working with large datasets such as financial datasets in our project.

The Random Forest is an ensemble method that is used for building and combining different decision trees with their outputs.

The Random Forest is very useful with a complex and large datasets in which the outputs applied under regression through averaging and the classification is done through voting. This method reduces the overfitting and improves the generalization. The Logistic regression is done by a linear model that is used for binary classification. Our project in which binary classification plays one of the important role. This performs very well on linear separable data. Here the model estimates the probability of a sample that belongs to a class using the logistic function.

The multi-layer perceptron is basically a deep learning model that is feed forward network which consists of multiple layers of neurons. This is very helpful for the model training of financial datasets which are linear in structure. This can also be used with complex nonlinear types of data.

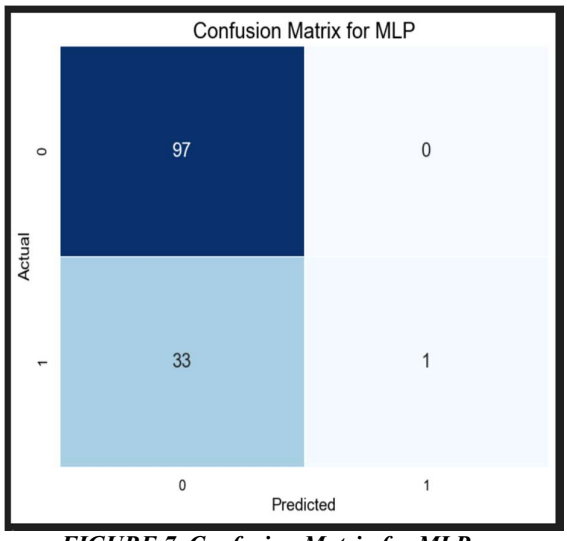


FIGURE 7. Confusion Matrix for MLP.

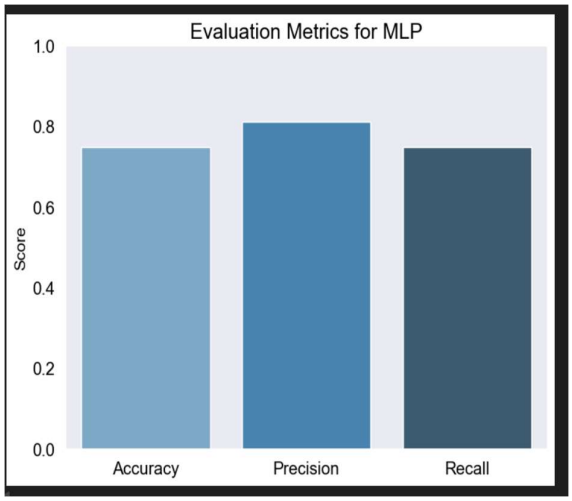


FIGURE 8. Bar Graph for MLP.

Gradient Boosting is one of the very useful ensemble techniques in model training and building the model, where each model works to specifically correct previous errors. This is very accurate in terms of output but very sensitive to the noise and overfitting of the large datasets.

The AdaBoost or often referred to as adaptive boosting algorithm in technical terms is an algorithm that takes multiple simple weak learns to combine and create a strong model that is much more accurate. It is simple and effective while working on a clean dataset.

The Voting classifier is an ensemble technique which takes the different model outputs to analyze and predict a stronger accurate output. In our project this model takes the prediction from Random Forest, Ada Boost, and Logistic Regression.

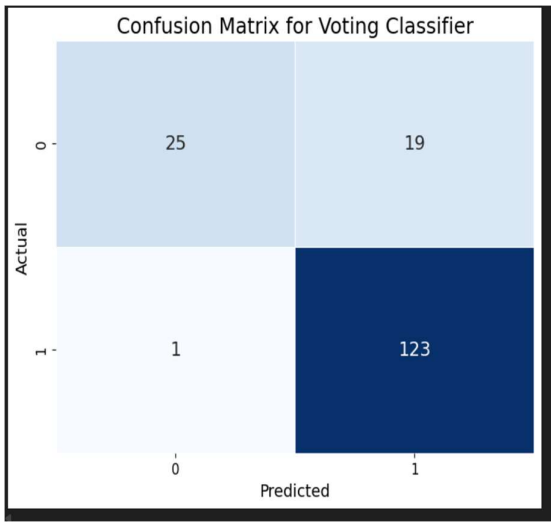


FIGURE 9. Confusion Matrix for Voting Classifier

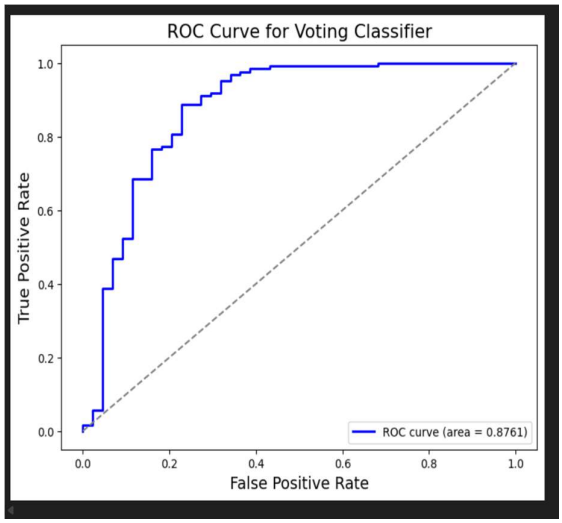


FIGURE 10. ROC Curve for Voting Classifier

The Stacking Classifier is an advanced ensemble method where multiple diverse models are combined to form a better prediction by training a meta-classifier. The models like AdaBoost, Random Forest, Logistic Regression are used as base learners in meta classifier and a final accurate prediction is given.

F. Model Evaluation

The Model evaluation is where the performance of the trained models is assessed. To check how each model actually working and performing, there are few evaluation metrics applied. These metrics give proper understanding of each model capability especially in context of the imbalanced data proportion of classified instances measured.

This is where the total model working is checked to be best or not to make changes and furthermore in the model performance. While it is very commonly used, but sometimes it can be misleading while working with imbalanced datasets.

The Precision metric is where the prediction of models is evaluated as if the proportions of positive predictions are correct or not. This is one very important metric in places like finance, security, governance, etc. where false predictions can be very costly to the institutions.

Recall is where the actual positive predictions are calculated to the false predictions. The measured proportion of the positive prediction is required for proper model working while High recall is necessary where the false/negative recalls must be minimized in the model.

This Confusion matrix is table summary where the summary of true vs predicted classifications, while showing the true positives, true negatives, false positives, and false negatives are described in a practical and easier way to be analyzed and understand the actual output of the models.

The ROC-AUC (Receiver Operating Characteristic – Area Under Curve) is graphical plot to represent the model performance across all the thresholds. This shows the tradeoff between true positive and false positive thresholds. So, this evaluation metric generally is used in binary classification models such as our project - Credit risk prediction. A higher AUC indicates that a particular model is performing very well.

The Models in our project were evaluated on a held-out test set, a cross validation also been applied on in the test to ensure the generalization and robustness of results are accurate and best of their performance. The models which are working need to compare visually to have better understanding, so a histogram was plotted showing the accuracy obtained from each model is used. This visualization of the models makes us notice the differences and

performances of the models, this also helps to select the best and most effective model furthermore, models can be added in future to check the model's performance and robustness to have much better predictions.

G. Real-Time Prediction Tool

The purpose of using different evaluation metrics is to check each model working is to get the best model which gives the best output. Based on these evaluations a real time loan approval prediction tool was built. This tool provides accurate prediction by providing a practical solution for faster loan approval decisions.

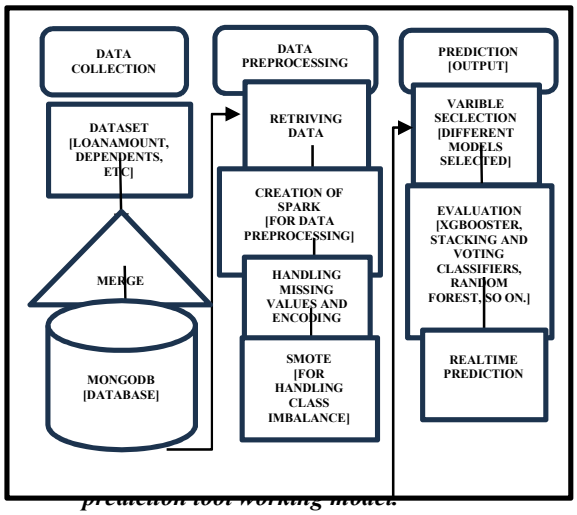
a) **User Input**: The tool takes user-provided data and stores data which are loan-related features such as marital status, education, income, loan amount and credit history from applications submitted by users.

b) **Data Preprocessing**: The input data given to tool, which is generally not balanced and is unclean data. Which is why data is preprocessed to handle the problems like missing values and encoding the categorical features that are needed for model training.

c) **Prediction Output**: The data which was given as input in the tool will analyze according to the training and outputs whether the loan application submitted by the user should be approved or rejected based on the model's prediction making quick and low risk decisions.

d) **Confidence Score**: A confidence score is best provided by the model based on representing the model's probability of approving the loan, indicating the certainty and accuracy of the prediction. Showing how much, it is sure of the output given according to the training done.

e) **Real-Time Decision Making**: The tool will make better and faster predictions, saving time and risk of finance by assisting the loan officers so that they can quickly make more informed decisions on loan applications, improving the efficiency.



prediction tool working model.

These following features make our model more accurate and user-friendly that can be utilized by financial institutions. This real time predictions tool by using can make quick decisions while making the data more scalable. Such models also make the financial predictions more accurate to make work quick while giving customers satisfactory service by the financial institutions. The model will be upgraded to improve analysis with advanced features and more robustness.

IV. Results

In this project, we have trained and tested the model and tool then have applied evaluation by using four key metrics: accuracy, precision, recall, and AUC (Area Under the Curve). With these metrics we can determine the working performance of each model with their strong points and weak points which eventually helps us to select properly the finest model that can help distinguish between applicants who are likely able to repay a loan versus those who are default (cannot repay the loan).

The results which are predicted by each model separately are presented in Table I

Table 1: Evaluation metrics of Different models

Model	Accuracy	Precision	Recall
XGBoost	0.8393	0.8819	0.9032
Decision Tree	0.7687	0.7565	0.9667
SVM	0.7985	0.7739	0.9889
Random Forest	0.7910	0.7719	0.9778
Logistic Regression	0.7985	0.7739	0.9889
Perceptron (MLP)	0.6716	0.6716	1.0000
Gradient Boosting	0.7761	0.7727	0.9444
AdaBoost	0.8679	0.8657	0.9748
Voting Classifier	0.8742	0.8613	0.9916
Stacking Classifier	0.8805	0.8731	0.9832

From the above data in Table I we have obtained that the Stacking Classifier is the best model to perform from overall all other models. So, the Stacking Classifier was selected as the real time prediction tool. It was saved as credit_risk_model.pkl and deployed in a real- time prediction tool for a finest prediction.

This tool allows users to enter relevant applicant information and receive an immediate prediction on loan approval or rejection very quickly, making it user friendly and more accurate for the officers in the financial institutions. The above Confusion matrix that shows us result of the Stacking Classifier model where the data has be trained through the model. Here we can observe the different metrics which are distributed

in different square blocks where the prediction is high which is showing us that the risk predictions are being adjusted with a the given from the actual input from different applicants dataset. Where the accuracy and precision of models at Where the accuracy and precision of models at very good level showing us that the model is working with better accuracy and precision. So, the model is used in a real-time prediction tool. This real-time prediction tool demonstrates the successful integration of data preprocessing, model training, and deployment with working to predict the output in a practical setting.

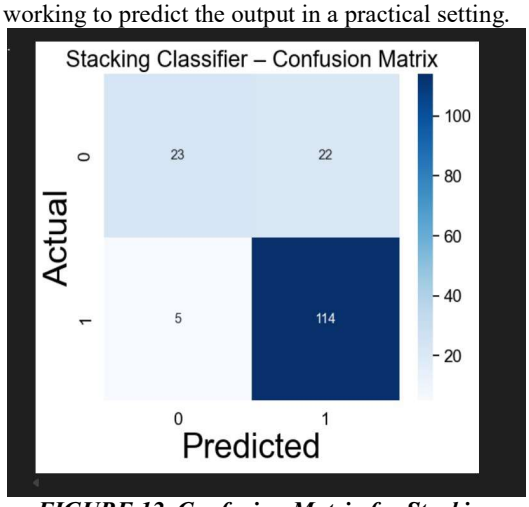


FIGURE 12. Confusion Matrix for Stacking Classifier

```
Please enter the following details for loan prediction:

You entered the following details:
Marital Status: Yes
Dependents: 3
Education: Graduate
Self-Employed: Yes
Applicant Income: 1456.0
Coapplicant Income: 1332.0
Loan Amount: 1688.0
Loan Amount Term: 360
Credit History: 0

Prediction Result:
Loan Denied
```

FIGURE: Output given by the real time prediction tool

So, by using the overall best-performing stacking classifier, the system ensures both robustness and quick response in predicting loan eligibility. This solution not only supports automated decision-making for loan eligibility in financial institutions but also gives access to the scalability of the approach for similar risk assessment tasks in other domains which work similarly.

V. Conclusion

a) Summary:

The proposed Generally payments from a bad customers causes very huge burden on financial institutions, and they also come with high expense. This is one of the huge problems faced by banks today which would lead them into instability, which is not good for the economy and people.

By applying features that are most required to the model with input dataset set that is cleaned and balanced. While selecting the right model Stacking Classifier has the best overall metrics with robustness in the real time prediction tool. PySpark, which is one of the important tools is a best distribution of

large data with preprocessing. While this application is only developed to ensure the risk of credit given is lower, financial institutions also should not give a loan that is enormous amounts of credit to which they may have risky or crashing of bank.

b) Limitations of the Research and Future Work

The research work can be explored more through different adaptive learning models, adjusting model parameters to new data source patterns. The future work for the research can be focused on deeper learning frameworks that would still reduce time and effort while giving the best result with robustness to the financial officers. The financial institutions need to review their decisions with certain customers comparing them to the model results, why they are approved/rejected. The cost of the economy is rising every year so the customers may not find the same credits to have in the future which does not fulfill their needs so according to the people's income earnings, this credit limit needs to be more advanced to deal with economic instability. This would make the research more advanced and active working for

the financial institutions for every new change occurring in the world.

VI. References

[1] X. M. R. Machado and S. Karray, ‘‘Assessing credit risk of commercial customers using hybrid machine learning algorithms,’ Expert Syst. Appl., vol. 200, Aug. 2022, Art. no. 116889.

[2] H. Sufriyana, A. Husnayain, Y.-L. Chen, C.-Y. Kuo, O. Singh, T.-Y. Yeh, Y. Wu, and E. C. Su, ‘‘Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: Systematic review and meta-analysis,’’ JMIR Med. Informat., vol. 8, no. 11, Nov. 2020, Art. no. e16503.

[3] Z.Zhao, T. Cui, S. Ding, J. Li, and A. G. Bellotti, Resampling techniques study on class imbalance problem in credit risk prediction,’ Mathematics, vol. 12, no. 5, p. 701, Feb. 2024.

[4] Y. Shi, Y. Qu, Z. Chen, Y. Mi, and Y. Wang, ‘‘Improved credit risk prediction based on an integrated graph representation learning approach with graph transformation,’’ Eur. J. Oper. Res., vol. 315, no. 2, pp. 786–801, Jun. 2024.

[5] C.-F. Wu, S.-C. Huang, C.-C. Chiou, and Y.-M. Wang, ‘‘A predictive intelligence system of credit scoring based on deep multiple kernel learning,’’ Appl. Soft Comput., vol. 111, Nov. 2021, Art. no. 107668.

[6] I.-C. Yeh and C.-H. Lien, ‘‘The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,’’ Expert Syst. Appl., vol. 36, no. 2, pp. 2473–2480, Mar. 2009.

[7] O. J. Leong and M. Jayabalan, “A comparative study on credit card default risk predictive model,’ J. Comput. Theor. Nanosci., vol. 16, no. 8, pp. 3591–3595, Aug. 2019.

[8] M. Pasha, M. Fatima, A. M. Dogar, and F. Shahzad, ‘‘Performance com parison of data mining algorithms for the predictive accuracy of credit card defaulters,’’ Int. J. Comput. Sci. Netw. Secur., vol. 17, pp. 178–183, 2017.

[9] S. Hamori, M. Kawai, T. Kume, Y. Murakami, and C. Watanabe, ‘‘Ensemble learning or Deep Learning? Application to default risk analysis,’’ J. Risk Financial Manage., vol. 11, no. 1, p. 12, Mar. 2018.

[10] Y. Xia, C. Liu, B. Da, and F. Xie, ‘‘A novel heterogeneous ensemble credit scoring model based on bstacking approach,’’ Expert Syst. Appl., vol. 93, pp. 182–199, Mar. 2018.

[11] Bao, Wang, Ning Lianju, and Kong Yue. 2019. Integration of unsupervised and supervised machine learning
algorithms for credit risk assessment. Expert Systems with Applications 128: 301–15.

[12] Chang, Victor, Raul Valverde, Muthu Ramachandran, and Chung-Sheng Li. 2020. Toward business integrity modeling and analysis framework for risk measurement and analysis. Applied Sciences 10: 3145.

[13] S. Birla, K. Kohli, and A. Dutta, Machine Learning on imbalanced data in credit risk,’ in Proc. IEEE 7th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON), Oct. 2016, pp. 1–6.

[14] M. Mahbobi, S. Kimiagari, and M. Vasudevan, Credit risk classification: An integrated predictive accuracy algorithm using artificial and deep neural networks,’ Ann. Oper. Res., vol. 330, nos. 1–2, pp. 609–637, Nov. 2023.

[15] Andrade Mancisidor, R.; Kampffmeyer, M.; Aas, K.; Jenssen, R. Deep generative models for inference in credit scoring. Knowl.-Based Syst. 2020, 196, 105758.

[16] Dm,Obare, and Muraya Mm. 2018. Comparison of Accuracy of Support Vector Machine Model and Logistic Regression Model in Predicting Individual Loan Defaults. American Journal of Applied Mathematics and Statistics 6: 266–71.

[17] Liu, Lulu. 2022. A Self-Learning BP Neural Network Assessment Algorithm for Credit Risk of Commercial Bank. Wireless Communications and Mobile Computing 2022: 9650934.2104

[18] M. A. M. Hassan, U. M. Mansur, R. Jha, F. H. Fahim, and T. Mahesh, ‘‘Interpretable machine learning models for credit risk assessment,” in Proc. 11th Int. Conf. Comput. Sustain. Global Develop. (INDIACom), Feb./Mar. 2024, pp. 361–365.

[19] Y. Gebreyesus, D. Dalton, S. Nixon, D. De Chiara, and M. Chinnici, ‘‘Machine learning for data center
optimizations: Feature selection using Shapley additive exPlanation (SHAP),’’ Future Internet, vol. 15, no. 3, p. 88, Feb. 2023.

[20] Y. Wang, Y. Zhang, M. Liang, R. Yuan, J. Feng, and J. Wu, ‘‘National student loans default risk prediction: A heterogeneous ensemble learning approach and the SHAP method,’’ Comput. Educ., Artif. Intell., vol. 5, Jan. 2023, Art. no. 100166.

[21] Moscato, V.; Picariello, A.; Sperlí, G. A benchmark of machine learning approaches for credit score prediction. Expert Syst. Appl. 2021, 165, 113986

[22] Buchanan, Bonnie G., and Danika Wright. 2021. The impact of machine learning on UK financial services. Oxford Review of Economic Policy 37: 537–63.

[23] Pławiak, P.; Abdar, M.; Acharya, U.R. Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. Appl. Soft Comput. 2019, 84, 105740.

[24] Fan, S.; Shen, Y.; Peng, S. Improved ML-based technique for credit card scoring in internet financial risk control. Complexity 2020, 2020, 8706285.

[25] Orlova, E.V. Decision-making techniques for credit resource management using machine learning and optimization. Information 2020, 11, 144.