IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

NON-PAYMENT RISK AUTOMATION USING MACHINE LEARNING AND ITS DEPLOYMENT ON ANDROID APPLICATION

Shashank Jagtap[1], Parth Apamarjane[2], Sourav Chitnis[3], Akshay Patil[4], Prof. Naved Raza Q. Ali[5].

[1,2,3,4,5]Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, India.

Abstract

One of the most significant and well-known elements of research in the banking and insurance industries is loan prediction. Non-payment risk is a significant concern for financial institutions as it can lead to financial losses and impact their overall stability. Processes for making decisions can become much more effective and accurate by automating the prediction of nonpayment risk. In the proposed study, we automate the nonpayment risk using the well-known machine learning technique XGBoost. The ensemble learning algorithm XGBoost is renowned for its superior performance with structured data and classification issues. XGBoost can efficiently analyses historical data and discover significant patterns related to nonpayment risk by utilizing its strong features. The XGBoost model can be used to forecast nonpayment risk for fresh, unforeseen cases after training. The model creates a probability score that indicates the chance of nonpayment by taking into account pertinent case-specific data, such as client demographics, transactional information, and credit history. To evaluate the performance of the XGBoost model, various metrics such as accuracy, precision, recall, and F1 score can be utilized. The proposed system will contribute to the growing body of literature on the use of machine learning in financial risk management and highlight its potential for improving efficiency and reducing risk and also provide recommendations for future research.

Keywords: Loan Repayment, Machine Learning, Random Forest algorithm**,** XG Boost algorithm, Heroku.

I. Introduction

Non-payment risk is a significant concern for financial institutions, as it can lead to financial losses and impact their overall stability. Traditional methods of non-payment risk analysis relyon manual processes and can be time-consuming and error-prone. However, recent advancements in machine learning have opened up new possibilities for automating non- payment risk analysis and improving risk management practices. It is a challenging operation to be performed where machine learning plays a significant role.

Our main aim is to investigate the use of machine learning algorithms for automating non- payment risk analysis in the context of financial transactions. The paper will explore how machine learning can be used to automatically identify and assess non-payment risk factors, such as customer creditworthiness and payment history, to improve risk management practices.The proposed study will use a dataset of financial transactions to train and evaluate the performance of the machine learning model. Most important thing is to check whether the particular customer is capable for paying the loan back or not.

Non-payment risk automation system is used for prediction according to the attributes considered from the dataset. The purpose of study is to test and train the dataset using differentmachine learning algorithms and understanding how the features are relevant. Briefly having study on how to train the model and predict the features from the model. Understanding the concepts of loan prediction and having knowledge on validation of accuracy and model loss while training and testing data. Then selection of the model is much important according to theaccuracy obtained from each model.

The primary objectives of this project are given below:

1. To understand the machine learning algorithms and implement them.
2. To understand dataset to implement in the project
3. To understand the accuracy of the models to implement in the project.
4. To train and test the given dataset.
5. To build an android application using a model.

The algorithms we used to compare with our XGBoost Machine Learning model are:

1. Logistic Regression: Logistic regression is a type of statistical analysis that is used to model the relationship between a categorical dependent variable and one or more independentvariables. The goal of logistic regression is to predict the probability of an event occurring, based on the values of the independent variables. The output of the logistic regression model is a binary outcome, either a 0 or 1. The 0 represents the event not occurring, while the 1 represents the event occurring. By using Logistic regression, the projected probabilities are guaranteed to range from 0 to 1. Logistic regression can be used for both binary classificationproblems and multi-class classification problems.

2. SVM: Support vector machines (SVMs) are a form of machine learning technique used for classification and regression analysis. The main objective of SVMs is to find the best possible boundary (i.e., hyperplane) that can separate the data points of different classes in a high- dimensional space. In SVMs, each data point is represented as a feature vector in a high- dimensional space. The algorithm then tries to find the hyperplane that maximizes the margin.If the data is not linearly separable in the original space, SVMs use a technique called the kerneltrick to map the data into a higher-dimensional space where it is linearly separable.

3. Decision Tree: Decision tree is a type of machine learning algorithm used for both classification and regression analysis. It functions by building a tree-like representation of decisions and their potential outcomes. Each node in the tree represents a decision based on one of the input features, and each leaf node represents a predicted output value. Each split in the algorithm's learning process maximizes information gain by recursively dividing the data into subsets according to the values of the input features.

4. Random Forest: Random Forest is an ensemble learning method for classification, regression,and other tasks that operates by constructing a multitude of decision trees at training time andoutputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The idea behind random forest is to create multiple decisiontrees using a randomly selected subset of input features and samples from the training dataset.Each tree is constructed using a subset of the training data and a random subset of the input features, and the final output is obtained by aggregating the outputs of all the individual trees.

XGBoost: XGBoost stands for eXtreme Gradient Boosting, and it is a popular machine learning algorithm used for classification, regression, and ranking problems. It is a tree-based ensemble method that uses a gradient boosting framework, which sequentially trains weak learners to correct the errors made by the previous learners. The algorithm works by iteratively adding decision trees to a model, where each new tree tries to correct the mistakes of the previous trees. During each iteration, the algorithm computes the gradient of the loss functionwith respect to the predicted values and uses this gradient to update the parameters of the model, making it more accurate. XGBoost is known for its scalability, speed, and accuracy, and it has been used to win several machine learning competitions on Kaggle.

Mobile applications have become an essential part of our daily lives, with millions of people around the world relying on them. Android, being the most popular mobile operating system globally. In the proposed system, we will also focus the development of an Android application,which is a complex process that requires a deep understanding of the platform's architecture and various tools and technologies used to build Android apps. Android offers a marketplace for the sale of apps. Android functions as a mobile app ecosystem.

II. Literature Survey

Anant Shinde et al. [1] used algorithms like Logistic regression using stratified k-folds cross validation and Random Forest. Various features that are taken into the dataset are compared with each other to know their relation with each other. From the comparison done, Credit_History is the most major point. The dataset consists of 600 sample data points. This model obtains a maximum accuracy of about 82 percent.

Ritika Purswani et al. [2] proposed a methodology on loan approval prediction method used by banks. Initially the data is collected then EDA is performed on the data. Outliers are removed and the data is structured. Then model is built using various classification algorithm like logistic regression, Decision Tress, Random Forest, Support Vector Classification. The accuracy obtained by logistic regression was 81 percent whereas the accuracy by Decision tree was 82 percent. Using random forest there is less chances of overfitting and higher the precision.

Sudhamathy G. [4] proposed a study using random forest algorithm for credit risk analysis. Dataset is taken from UCL Repository having 1000 records and 21 attributes. Initially data normalization is performed on the dataset. Dataset is divided into 4:1 ratio. 80% training dataset and 20% test dataset. The training dataset is balanced using SMOTE function. A threshold value is fixed for selecting number of features based on their rankings.

Kumar Arun et al. [5] proposed a model to reduce the risk factor for selecting the right person to get the loan from the banks. Data is collected first and then comparison is done between different machine learning models. Training is done on the most appropriate model and then testing is performed. Every time a new user fills the data it is treated like testing data. Algorithms used for building model are Decision tree, Random Forest, SVM, Linear model, Neural network, Adaboost.

Pidikiti Supriya et al. [6] proposed a model which reduces bank's efforts for selecting a candidate for loan who will not default on loan. Data mining technique is used to understand the underlying trends and patterns. Decision tree is used to build model. The data collected is divided into the ratio 4:1 of training set and testing set. The data collected is cleaned and correlation among the attributes is considered. Decision tree brings out an accuracy of 81%.

Saqib Aziz et al. [8] proposed a system based on SVM and Decision tree algorithm with neural network. However, it explored how artificial intelligence (AI) and machine learning solutions are transforming risk management. This paper explained the application to credit risk, application to market risk, application to operational risk, application to Reg Tech. Moreover, after taking the dataset for reducing to appropriate dimensions, they used Principal component analysis (PCA).

Majid Bazarbash et al. [9] elaborated what is Machine Learning. Machine learning models as nonparametric models, Prominent Machine Learning Models and proposed decision tree and random forest algorithm. They Achieving widespread financial inclusion is a major step in achieving sustainable development goals in many countries, and FinTech credit appears as a promising solution and a potential leapfrog for countries with low financial inclusion.

Mehul Madaan et al. [12] has studied on Decision tree random forest algorithm. The paper automates the process of borrower's likelihood of repaying the loan. Machine Learning concepts are used to solve this problem. Both the algorithms have been used on the same dataset and the conclusions have been made with results showing that the Random Forest algorithm outperformed the Decision Tree algorithm with much higher accuracy. Dataset used was used from Lending Club in Kaggle. Random Forest showed accuracy of 80% while Decision Tree provided 73% accuracy. The paper showed comparative study and did not consist of an actual user interface.

Anshika Gupta et al. [13] has studied on Logistic Regression and Random Forest. The paper was published and presented in 9th International Conference on System Modelling & Advancement in Research Trends organized by IEEE. Date of presentation was 4th-5th December,2020. to train the models was taken from Kaggle competition which belong to different age group and gender of the applicants having 13 attributes in all. Moreover, the proposed system has user interface developed in HTML and CSS and connected to local server via Django. The paper lacked in amount of accuracy and user interface.

Shamsa Khalid et al. [14] researched on decision trees and random forests was conducted in the year 2022. The goal of this research study was to analyze, assess, and create machine learning algorithms that could precisely and error-free estimate risk for nonfinancial enterprises listed on PSX. These kinds of models, particularly those that focus on quantitative investment, could be used in nonfinancial organization's investment management since they provide a thorough description of the most recent technology, their potential uses, and the possibility that they would be successfully applied.

Sivasree M. S. et al. [16] assisted underperforming banks improve their operations and turn a profit. Two key goals are to identify pertinent attributes and choose the appropriate models to assess credit risk. There are 17 attributes and 4520 records in the collection. The remaining 34% of the dataset is used to test the model, and 66% of it is utilized to train the model. 'Job' is the most significant attribute taken into account here.

Kaoutar Erramy et al. [20] proposed an international trade strategy. The following article appeared in the peer-reviewed journal "Journal of Advanced Research in Accounting & Financial Management." This paper's main goal is to outline the acceptable strategies used by a multinational corporation to address scenarios including the risk of international non-payment. This study shows that a number of variables affect the international non-payment risk management strategy, which makes it necessary to assess the experiences of all Moroccan exporting enterprises.

X.Francis Jency et al. [21] assisted in understanding the nature of customer who applies for a loan. EDA is done on a given dataset. The dataset initially undergoes the process of normalisation, missing values are taken care, essential columns are filtered, deriving new columns, identifying the target variables and visualizing the data in graphical format. Pandas' library available in python is used to derive information from given dataset. Various graphs are used to visualize and from the graphs it can be concluded that short term loan was preferred by the majority of the loan applicants and customers mainly apply loan for debt consolidation.

Aditi Kacheria et al. [22] proposed a suggestion using Naive Bayesian classification, K-NN, and Binning algorithms to keep the bank safe. The following article appeared in the 2016 issue of the International Journal of Soft Computing and Engineering (IJSCE). Using K-NN and Binning algorithms, the quality of the data is improved prior to classification in order to increase classification accuracy. Using pre-processing methods is necessary since there are many situations when the data set is inconsistent as a result of missing values and abnormalities. They utilise the K-NN method to handle missing values. K-NN is a straightforward algorithm that categorises new data based on a similarity metric and stores all accessible data.

Aboobyda Jafar Hamid et al. [23] developed a classification system that employs BayseNet, NaiveBayes, and J48 to forecast loan defaulters. The used dataset is split into 20% of testing data and 80% of training data. naveBayes accuracy is 73%, bayesNet accuracy is 77%, and j48 accuracy is 78%. Weka application is used to put the model into action. As a result of its high accuracy and low mean absolute error, the J48 algorithm is determined to be the best.

III. Proposed Method

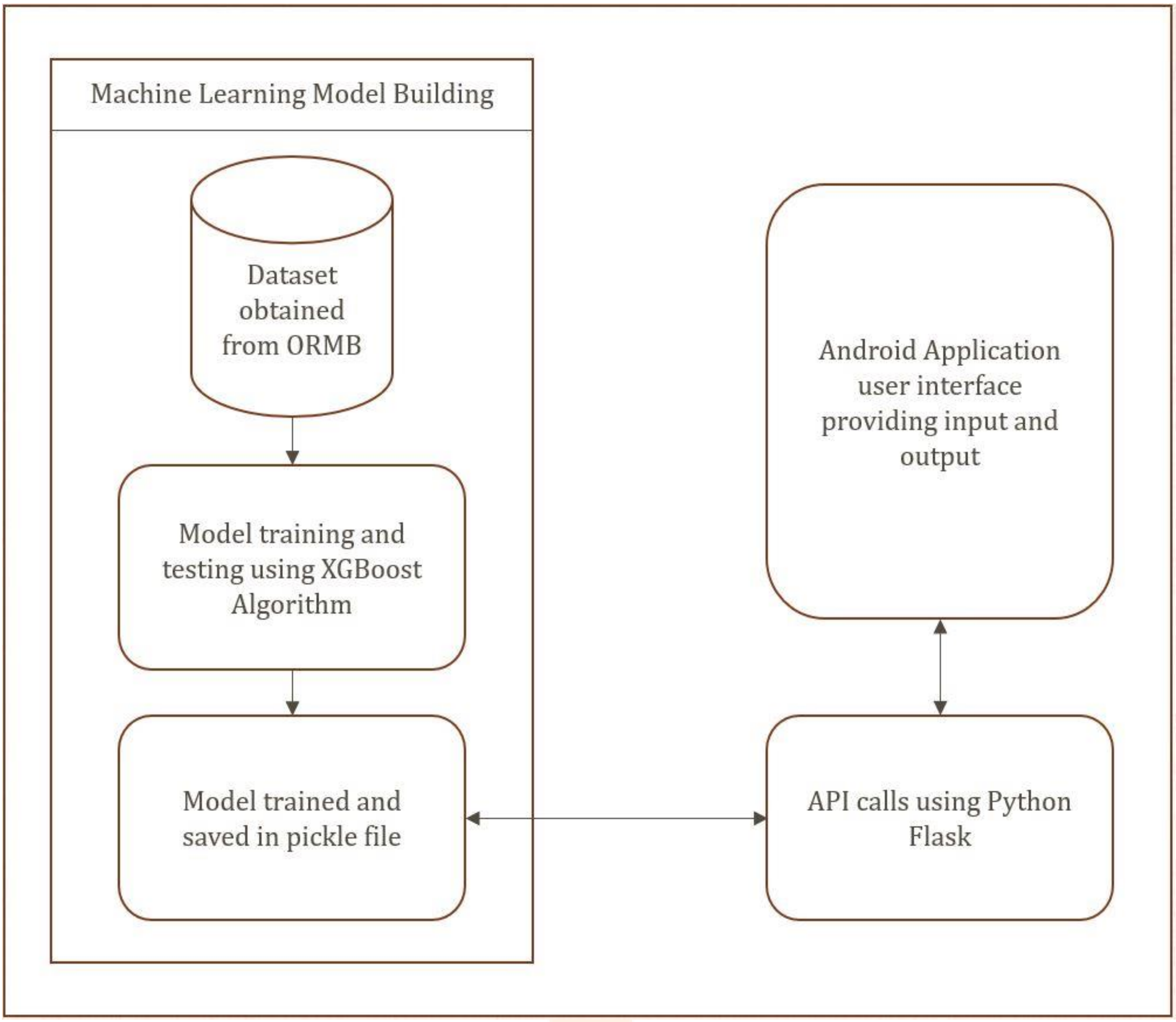


Figure 1: Proposed system architecture.

There are three parts majorly used in the proposed method:

A. Model Generation
B. Building android application
C. Deploying the application

A. Model Generation:

The dataset used in the proposed work consists of 32581 rows and 12 attributes. The dataset is a combination of different categorical variables as well as numerical variables. Our target variable is 'Loan Status' which is a dependent variable. It is a binary variable which consists of 2 categories 0 & 1. 0 stands for non-default (means the person will return the loan and will not do default on loan) and 1 stand for default (the person will default on loan). The categorical attributes are Person_Home_Ownership which gives home ownership status (Rent, own, mortgage, other), Loan_Intent which tells us the purpose of the loan, Loan_Grade (A, B, C, D, E,F, G), cb_person_default_on_file (Yes or No) describes historical default. Numerical variables are Person_age, Person_income, Person_Emp_length (employment length), Loan_amount, Loan_interest_rate, Loan_percent_income, cb_person_cred_hist_len.

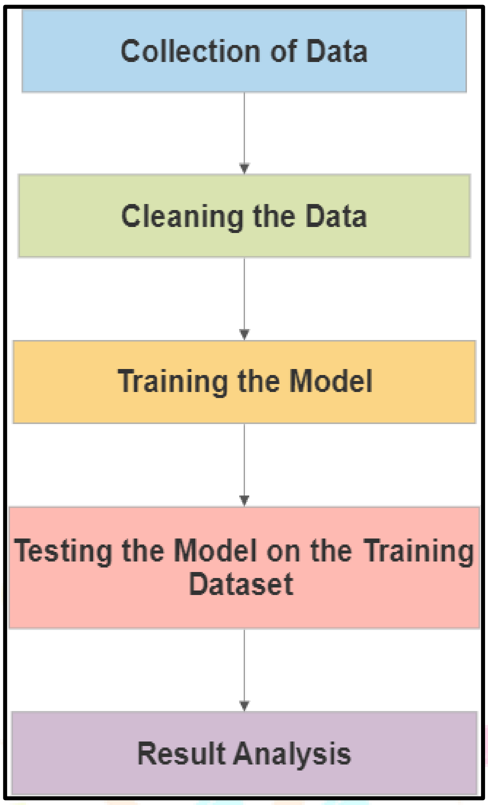


Figure 2: Steps performed to create ML model.

Figure 2 consists of the whole process of model building right from start where the data is loaded from drive in .csv format. Various libraries like pandas, seaborn,matplotlib, plotly, numpy, are used. All the machine learning algorithms are imported from sklearn library. Once the data is collected, the next step is to pre-process and clean the data. This step involves handling missing values, encoding categorical variables, scaling numerical features, and splitting the data into training and testing sets. All the duplicated values are dropped as it is seen as a noise in model building process. The total null values are removed. All the outliers are explicitly removed otherwise it can cause overfitting in the model. EDA (exploratory data analysis) is performed tounderstand the underlying pattern in the dataset. In EDA, univariate, bivariate and multivariate analysis is performed to check relationships between the attributes. Feature engineering and feature selection is done, as the best possible features which highly affect our target variable are selected. Features like person_age and person_gender aredropped. Fig. 2 shows importance of features.

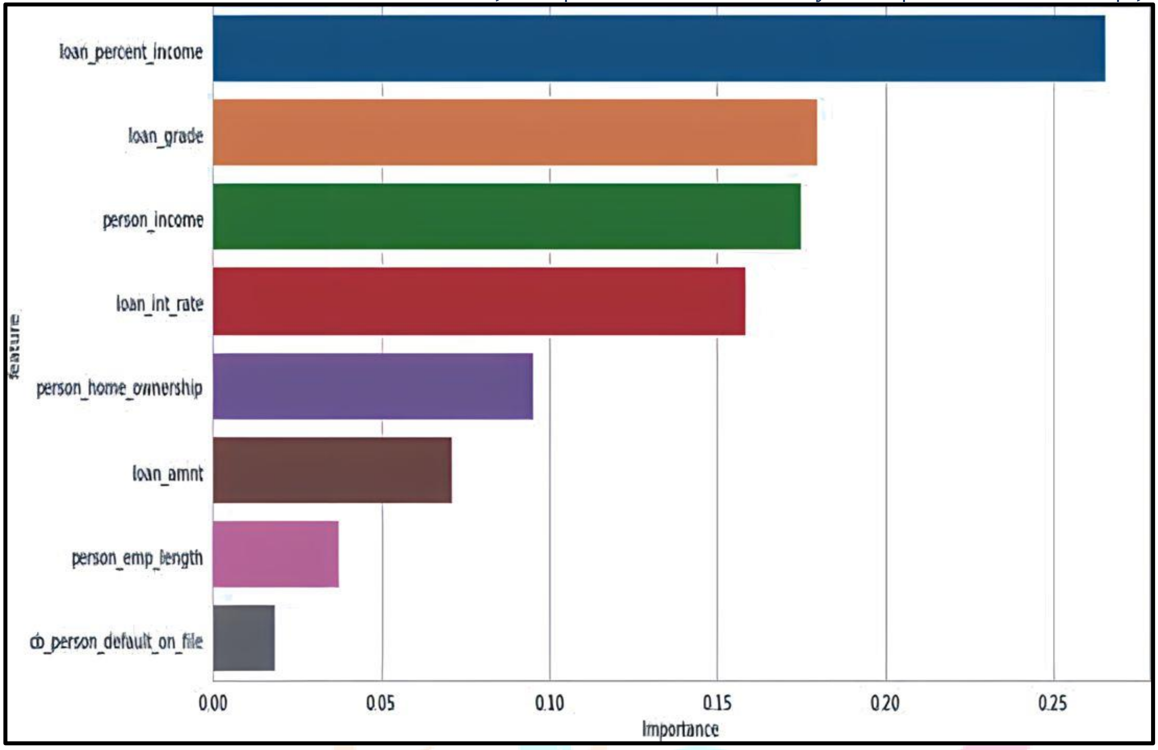


Figure 3: Feature Importance compared between all attributes in the dataset.

To balance our imbalanced target variable, we apply random sampling method is applied. For training our model we have used 75% of our data and rest 25% of data is used for testing of our model. This will help in comparing the predicted values and theoriginal values from the dataset. Scikit Learn is used to import Logistic regression decision tree classifier, K Nearest Neighbor and Random Forest Classifier algorithms to implement. XGBClassier is used to import XGBoost algorithm. The next step is to train the model using each algorithm and check the accuracies. Confusion matrix, Receiver Optimistic curves for each algorithm are created to check the performances appropriately Depending on the performance and accuracy the best suited algorithm ischosen, and the model is generated.

B. Building Android application.

Android application is built using java language. Python Flask is a communication toolbetween machine learning model and the android application. Setting up the communication between ml model and android app is done using API calls in flask. Input and output of flask is in JSON format which helps in communication. The definition of flask is Flux Advanced Security Kernel. Machine learning model is loadedusing pickle. The input is taken and saved in respective attribute variables. These variables are given as parameters to the model for prediction. The prediction is out andis printed as output. One of the most common ways to connect an Android app with a machine learning model is by using a REST API. With this approach, the machine learning model is hosted on a server, and the Android app makes HTTP requests to theserver to get predictions. The server can be built using a web framework like Flask. Fig.3 shows the flow of our application.

C. Deploying the Android Application

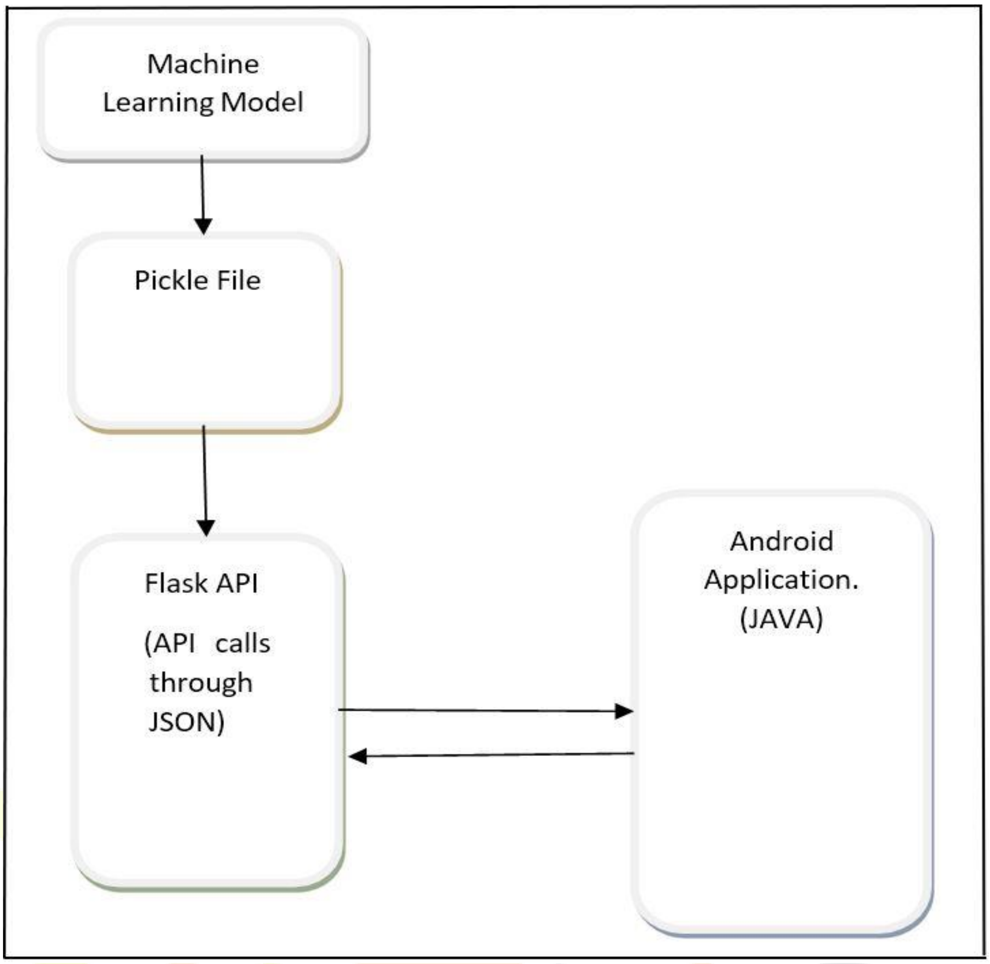


Figure 4: Communication model between Android Application and ML model.

Heroku app is a container-based cloud platform. Heroku app is used by developers to deploy, manage and scale modern apps. It is integrated with data services and powerfulecosystems. This android app is deployed on the Heroku server that uses gUnicorn WSGI HTTP Server. Source code and dependencies files provides information of the Heroku platform. A Procfile is a text file that accompanies the source code. Git is usedto manage the source code. Moreover, requirements.txt file contains dependencies required for the source code.

IV. Experiments and Results

In this section, the results are obtained based on training and validation of data of both accuracyand loss of the model obtained while performing the algorithms. The dataset we're working onis imbalanced dataset so we need to make sure we're utilizing the right assessment criteria in this situation. Consider that accuracy measures the ratio of all predicted values to all input samples. As a result, our model would achieve high accuracy by predicting the majority class but would, by default, fall short of capturing the minority class, which is not good. Because ofthis, Precision, Recall, and F1 score will be the assessment metrics we use to evaluate the classification performance of our models. Precision provides us with the proportion of real positives to all the positives predicted by a classifier, where positives in our context refer to default circumstances. We can see that our models do an excellent job of accurately predictingthose minor events given that they make up the minority class in our dataset. Recall, also knownas the true positive rate, also provides us with the proportion of components that genuinely fallinto the positive class to the overall number of true positives. Given that we care more about false negatives—when our model predicts that someone won't default but does—than false positives—when it predicts that someone will default but doesn't—Recall is a more crucial parameter in our situation than precision.

A. Machine Learning Algorithms based results

a. Heatmap for dataset: A heatmap is a graphical representation of data that uses colorsto indicate the relative intensity of values in a matrix or table. It is commonly usedto visualize data. The intensity of the color indicates the strength or magnitude of the corresponding value. Heatmaps can be used to identify patterns and relationships in data, or identifying correlations between different variables in financial data. Fig shows pictorial representation using heatmap that shows the interrelation between the features that is plotted.

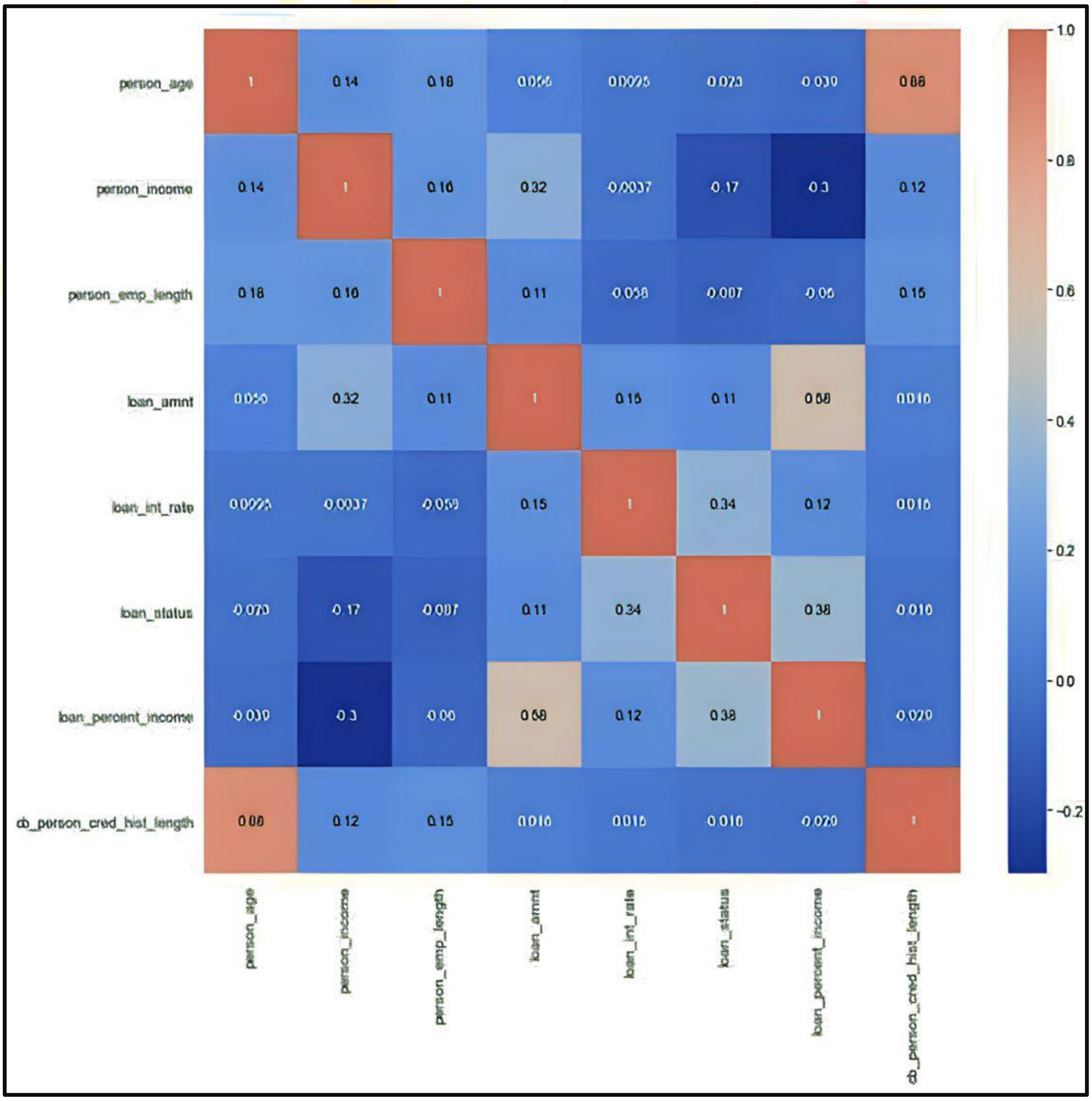


Figure 5: Correlation map between attributes/features.

b. Testing accuracies and confusion matrix: A confusion matrix is a table that is usedto evaluate the performance of a classification model by comparing the actual and predicted values of the target variable. It is a useful tool in machine learning and data analytics to measure the accuracy and effectiveness of a classification algorithm. The matrix is constructed with the actual values of the target variable onone axis and the predicted values on the other axis. The cells of the matrix represent the number of observations that fall into each combination of actual and predicted values. The four possible outcomes in a binary classification problem are true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These values are used to calculate various metrics such as accuracy, precision, recall, F1 score, and others. True positives are the number of outputs where the model correctly predicts the positive class. True negatives are the number of outputswhere the model accurately predicts the negative class. False positives are the total outputs where the model incorrectly predicts the positive class. Similarly false negatives are when a model incorrectly predicts the negative class. The confusion matrix provides a visual representation of how well a model is performing and canbe used to adjust and improve the model's performance.

i. Testing with Logistic Regression: Logistic regression is a type of statistical analysis that is used to model the relationship between a categorical dependent variable and one or more independent variables. The accuracy obtained is 78.3% and the recall is 77.8%. Fig 6 (a) depictsthe confusion matrix.

ii. Testing with Decision Tree: The Decision tree is a non-parametric supervised learning algorithm used for classification and regression. Theaccuracy acquired is 85.42% and the recall is 82.58%. Fig 6 (b) depictsthe confusion matrix.

iii. Testing with K Nearest neighbors: K Nearest neighbors’ algorithm is used to predict the values of the dataset points. It resembles the point that is based on the values of the dataset. The accuracy occurred is 83.25% and the recall is 80.72%. Fig. 6 (c) depicts the confusion matrix.

iv. Testing with Random Forest Algorithm: Fig. 6 (d) shows the testing with the Random Forest algorithm where the prediction is based on outof bag predictions. Firstly, predict and test the data using accuracy_score() and then classification is done based on reports that occurred and the accuracy occurred is 86.9% and the recall is 81.3%.

v. Testing with Support Vector Classifier (SVC): Fig. 6 (e) shows the testing with the SVC algorithm where the SVC algorithm is based on tofind a hyperplane in a high-dimensional space that best separates the different classes. The accuracy occurred is 82.27% and the recall is 76.09%.

vi. Testing with XGBoost: Fig. 7 shows the testing with the XGBoost algorithm where it uses a gradient boosting framework that allows it to learn from previous errors and improve its predictions iteratively. The accuracy occurred is 92.28% and the recall is 91.69%.

Among all the testing performed, the accuracy and recall of XGBoost is highest. Hence it will generate a model with highest accuracy. Below mentioned are all theconfusion matrices of Algorithms tested except XGBoost.

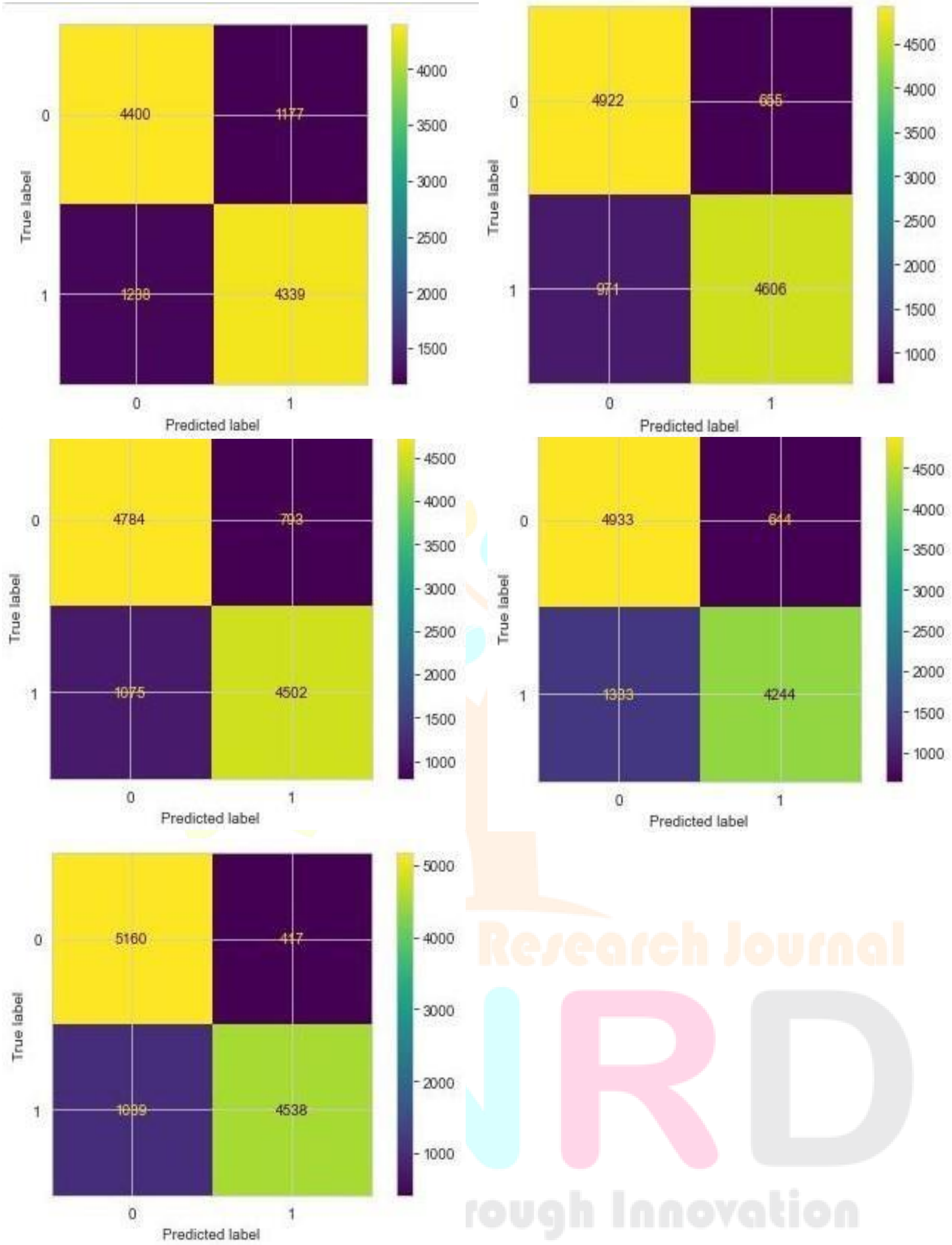
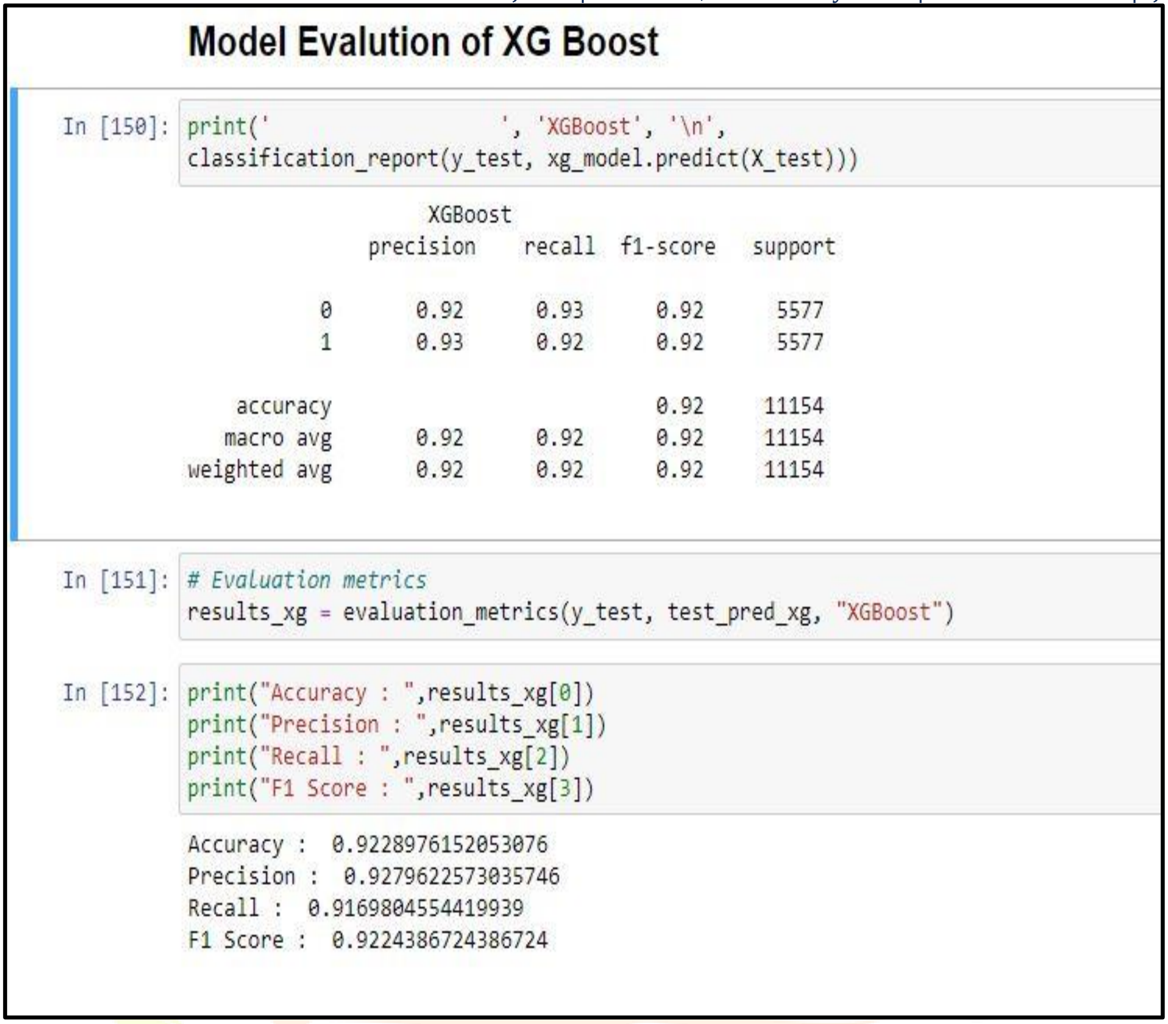


Figure 6: (a) Confusion matrix of Logistic Regression, (b) Confusion matrix of Decision tree, (c) Confusion matrix ofKNN, (d) Confusion matrix of Random-forest, (e) Confusion matrix of SVC.

Below mentioned are the performance measures of implemented XGBoost Algorithm. XGBoost outperformed all the previously explored algorithms and hence we used XGBoost machine learning model in our Android Application.



```
Model Evalution of XG Boost

In [150]: print('                    ', 'XGBoost', '\n',
          classification_report(y_test, xg_model.predict(X_test)))

                      XGBoost
               precision    recall  f1-score   support

           0       0.92      0.93      0.92      5577
           1       0.93      0.92      0.92      5577

    accuracy                           0.92     11154
   macro avg       0.92      0.92      0.92     11154
weighted avg       0.92      0.92      0.92     11154

In [151]: # Evaluation metrics
          results_xg = evaluation_metrics(y_test, test_pred_xg, "XGBoost")

In [152]: print("Accuracy : ",results_xg[0])
          print("Precision : ",results_xg[1])
          print("Recall : ",results_xg[2])
          print("F1 Score : ",results_xg[3])

          Accuracy :  0.9228976152053076
          Precision :  0.9279622573035746
          Recall :  0.9169804554419939
          F1 Score :  0.9224386724386724
```

Figure 7: Performance measures of implemented XGBoost Algorithm.

Figure 8 gives comparison between all evaluation performance metrics of all the algorithms used in building the model. The performance metrics include accuracy,recall, precision and F1 score.

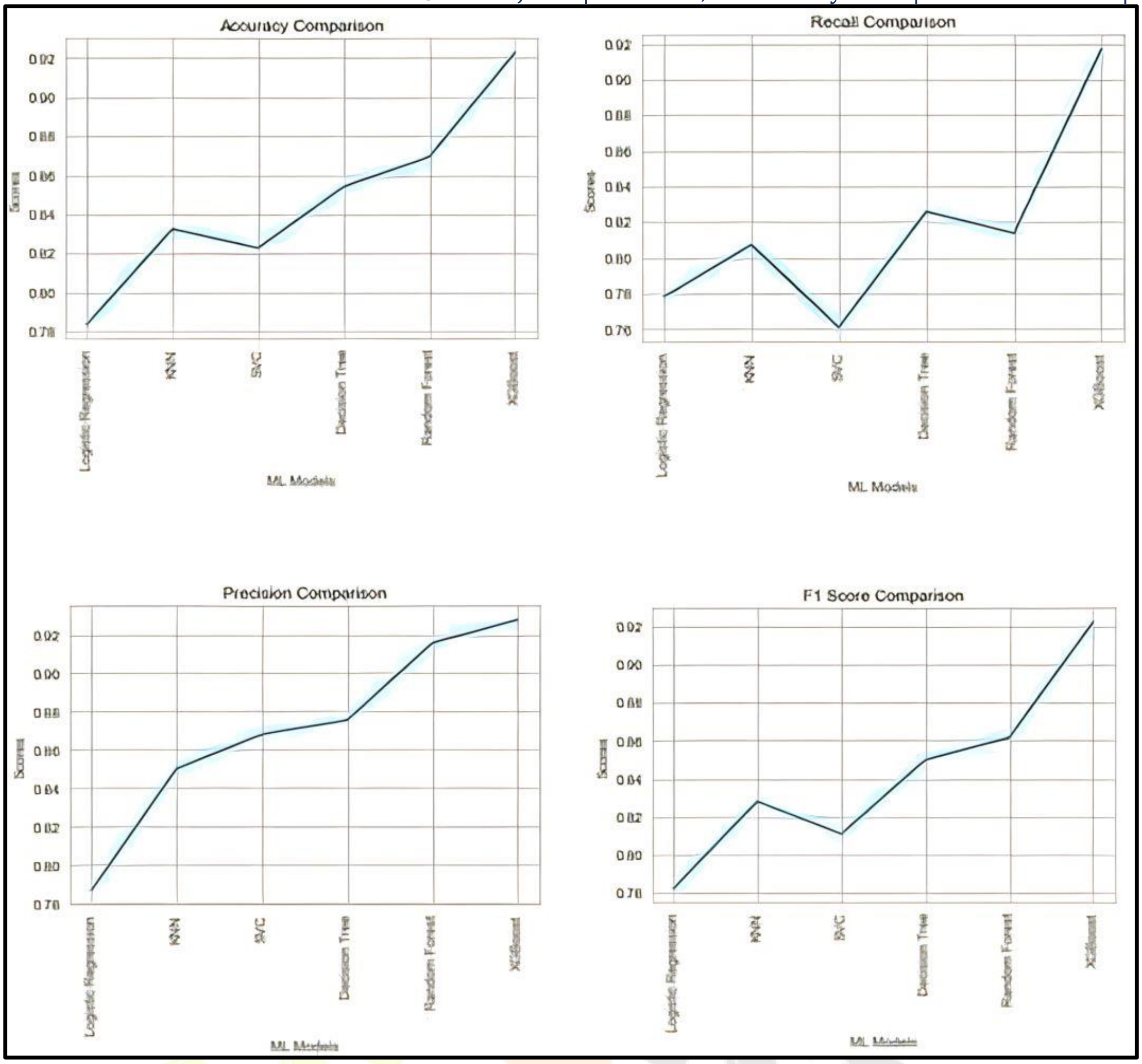


Figure 8: Graphical comparisons of XGBoost with remaining algorithms.

B. Deployment of model:

The app is developed and has a clean and neat user interface. The home page isseen in the figure 9 (a). All the attributes information is to be filled by the customer. Customer must fill this information accurately and precisely as shown in figure 9 (b). Upon filling the details and clicking on predict button, output will be displayed whether the customer will return the loan or won't return the loan. It will be helpful while working as filter to the banks.

The dataset used in this work covers majority of the attributes possible such aseducation, dependents, property area, credit history, loan amount etc. also the algorithmused proves to give a better accuracy when compared to the existing models which useloan amount, interest rate as their prime attributes in their loan calculators. This androidapp feature can be used by individual banks and can be implemented in their existing applications which can act as an extra level of security.

Figure 9 (a): Icon of Android App.

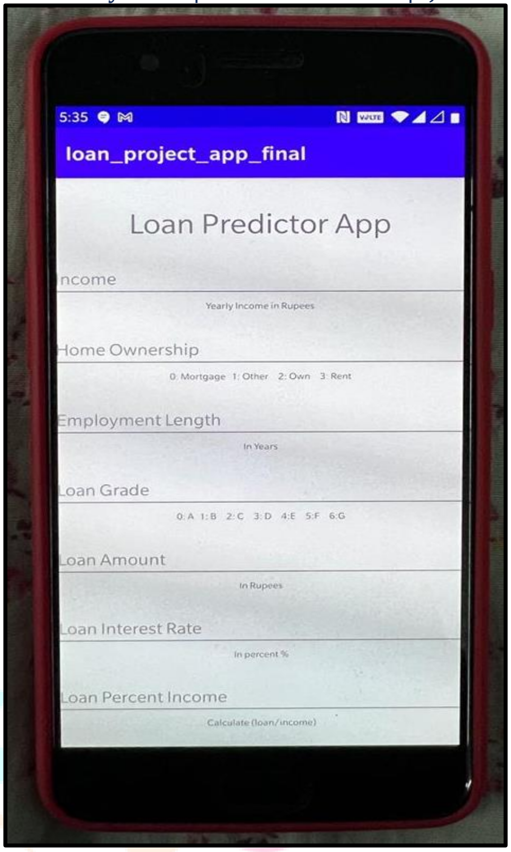


Figure 9 (b): Home page of Android App.

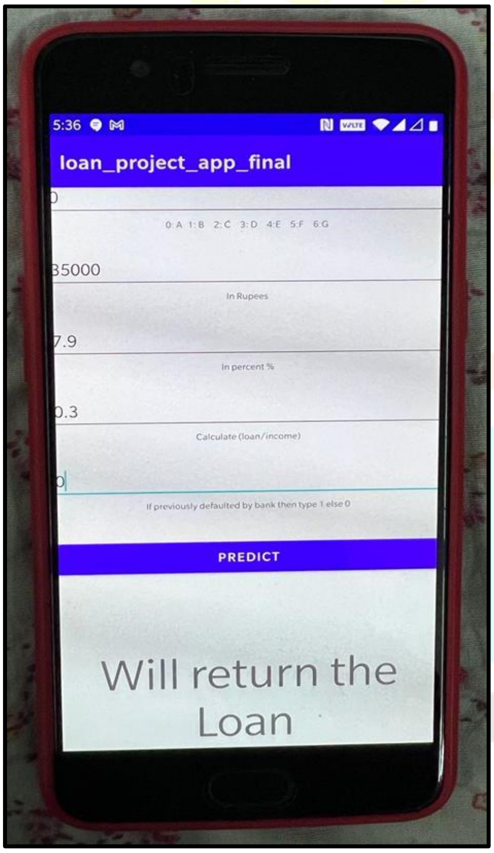


Figure 9 (c): Test case 1.

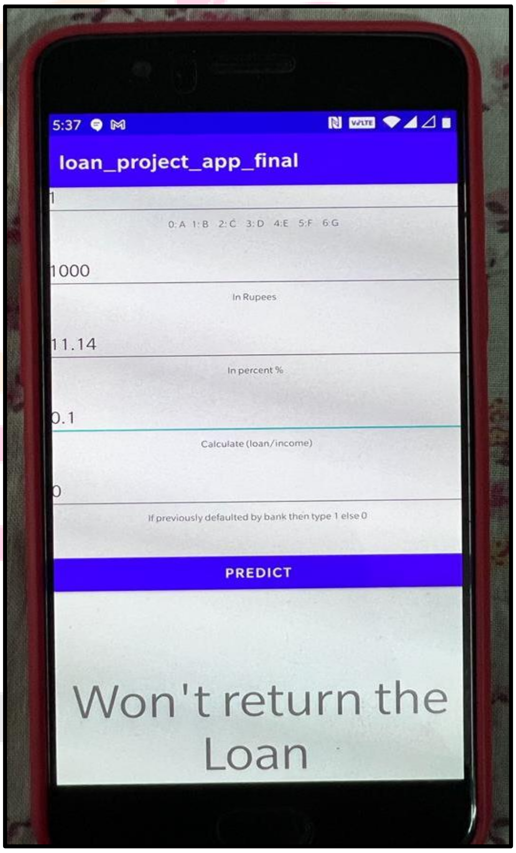


Figure 9 (d): Test case 2.

The above figures display the possible outputs that the application could provide. Figure 9 (c) illustrates first test case where the output to the given input is positive. It means that the following user will return the loan as predicted by the Machine Learning model. On the other hand, figure 9 (d) showcases second test case where the user won't repay the loan.

V. Conclusion

In this work, the loan prediction is performed by obtaining the dataset from various sources in .CSV file format. We have noticed that loan repayment utilizing machine learning has effectively demonstrated the usage of several algorithms and strategies to reliably forecast loan repayment probability. Through the use of Python language and various libraries, we were able to clean and preprocess our dataset by removing outliers and perform exploratory data analysis where we got to know how the attributes are inter-related, and feature engineering is useful for understanding the features which contribute more in predicting the output while building our model. We deployed our application using Flask and Heroku, making it easily accessible to users. Our research also revealed that XGBoost was the best performing algorithm, with the highest accuracy and recall factor. Anyone looking to apply for a loan or the banking system will use this model. It will be very beneficial for managing banks. It is abundantly obvious from the data analysis that it lessens all fraud committed at the time of loan acceptance. Everyone values their time highly, thus by doing this, not only the bank but also the applicant's wait time will be shortened.

VI. Future Work

Future research could explore the application of machine learning algorithms to otherareas of financial risk management and investigate the practical challenges and limitations of implementing these solutions in real-world settings. According to the dataset used, the future scope of this work can be applied to anticipate loans.Different countries prioritize different features, and those attributes might be used in the datasetwhile being trained based on the bank's preference. In the future, real-time data combined witha larger sample size will be deemed to have a high accuracy rate and minimal loss. It will alsobe projected to what extent the loan can be approved.

VII. References

[1] A. Shinde, Y. Patil, I. Kotian, A. Shinde, and R. Gulwani, "Loan prediction system using machine learning," in ITM Web of Conferences, vol. 44, 2022, pp. 1-5, doi: 10.1051/itmconf/20224400001.

[2] R. Purswani, S. Verma, Y. Jaiswal and Prof. S. M., "Loan approval prediction using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 08, no. 06, pp. 01-05, Jun. 2021.

[3] L. Zhua, D. Qiua, D. Ergua, C. Yinga and K. Liub, "A study on predicting loan default based on the random forest," *Innovations in Power and Advanced Computing Technologies (i-PACT)*, pp. 236-239, Jun. 2019.

[4] S. G., "Credit risk analysis and prediction modelling of bank loans using R," *International Journal of Engineering and Technology (IJET*), vol. 8, no. 6, pp. 2631-2635, 2016.

[5] K. Arun, G. Ishan and S. Kaur, "Loan approval prediction based on machine learning approach," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 18, no. 4, pp. 56-61, 2016.

[6] P. Supriya, M. Pavani, N. Saisushma, N. V. Kumari and K. Vikas, "Loan prediction by using machine learning models," *International Journal of Engineering and Techniques*, vol. 5, no. 2, pp. 32-37, 2019.

[7] Ms. K. R. Pramod, Ms. P. S. Dattatray and Ms. A. P. Prakash, "An approach for prediction of loan approval

using machine learning algorithm," *International Journal of Creative Research Thoughts (IJCRT*), vol. 9, no. 2, pp. 2357-2362, 2021.

[8] S. Aziz and M. Dowling, "AI and machine learning for risk management," SSRN Electronic Journal, Jan. 2018, doi: 10.2139/ssrn.3101253.

[9] M. Bazarbash, "FinTech in financial inclusion: machine learning applications in assessing credit risk," *International Monetary Fund*, 2019, doi: 10.5089/9781484393044.086.

[10] C. K. Gomathy et al., "The loan prediction using machine learning,*" Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University*, 2021.

[11] C. N. Sujatha et al., "Loan prediction using machine learning and its deployment on web application," *Innovations in Power and Advanced Computing Technologies (i-PACT*), 2021.

[12] M. Madaan et al., "Loan default prediction using decision trees and random forest: A comparative study," *IOP Conf. Series: Materials Science and Engineering*, 2021.

[13] A. Gupta et al., "Bank loan predicting system using machine learning," *9th International Conference System Modeling and Advancement in Research Trends (SMART*), 2020.

[14] S. Khalid et al., "Predicting risk through artificial intelligence based on machine learning algorithms: *A Case of Pakistani Nonfinancial Firms*," 2022.

[15] P. M. Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Maison des Sciences Économiques*, 2018.

[16] S. M. S and R. S. T, "Loan credibility prediction system based on decision tree algorithm," *International Journal of Engineering Research & Technology (IJERT*), 2015.

[17] N. G. Kiliswa and M. S. Bayat, "Determinants of loan repayment in small scale enterprises in developing countries," *Management Studies and Economic Systems (MSES*), 2014.

[18] V. Ramanujam and K. A. Vidya, "A study on the credit repayment behaviour of borrowers," *International Research Journal of Business and Management (IRJBM*), 2017.

[19] M. Sterbak, P. Segec, and J. Jurc, "Automation of risk management processes," *University of Zilina, Faculty of Management Science and Informatics*, Žilina, Slovakia, 2021.

[20] S. Ahrouch and K. Erramy, "Exportations non-payment risk management," *Journal of Advanced Research in Accounting & Finance Management Peer Reviewed Journal*, 2018.

[21] X. F. Jency, V. P. Sumathi, and J. S. Sr, "An exploratory data analysis for loan prediction based on nature of the clients," *International Journal of Recent Technology and Engineering (IJRTE)*, 2018.

[22] A. Kacheria, N. Shivakumar, S. Sawkar, and A. Gupta, "Loan sanctioning prediction system," *International Journal of Soft Computing and Engineering (IJSCE)*, 2016.

[23] A. J. Hamid and T. M. Ahmed, "Developing prediction model of loan risk in banks using data mining machine learning and applications," *An International Journal (MLAIJ*), vol. 6, no. 2, pp. 55-63, 2016.