

Micro-Credit Defaulter Model

Submitted by:

VAISHNAVI BALAJI

**ACKNOWLEDGMENT**

1. Predicting Credit Default among Micro Borrowers in Ghana.

https://www.researchgate.net/publication/265161200\_Predicting\_Credit\_Default\_among\_Micro\_Borrowers\_in\_Ghana#:~:text=We%20found%20the%20following%20variables,younger%20generation%20and%20in%20males.

1. Rural Micro Credit Assessment using Machine Learning in a Peruvian microfinance institution. https://www.sciencedirect.com/science/article/pii/S1877050921009297
2. Strategies for Reducing Microfinance Loan Default in Low-Income Markets.

https://scholarworks.waldenu.edu/dissertations/4391/

**INTRODUCTION**

* Business Problem Framing

A microfinance institution (MFI) is a company that provides financial services to those with limited resources. Targeting impoverished, unbanked families in rural locations with few sources of income makes MFS highly effective. The MFI offers group loans, agricultural loans, individual business loans, and other microfinance services (MFS).

The usage of mobile financial services (MFS), which many microfinance institutions (MFI), experts, and donors believe to be more practical, effective, and cost-effective than the conventional high-touch strategy employed for many years to deliver microfinance services, is becoming increasingly popular. Although low-income households are the MFI industry's primary emphasis, and they are very helpful in these regions, the implementation of MFS has been inconsistent, with both major challenges and successes.

* Conceptual Background of the Domain Problem

For financial organisations, credit risk management is crucial since it has a direct impact on financial performance. Microcredit organisations still use traditional credit scoring methods based on linear calculation of a limited number of indicators, even though artificial intelligence (AI) and machine learning are not new. Instead, these organisations are reluctant to accept these techniques in their credit risk assessment process. On the other hand, machine learning nonetheless, provides a far larger perspective of a client and can be used to handle all business risks, not just credit risk.

One of the most significant risks that a financial organisation must handle is credit risk. Since there is no profit without loan repayment, the issue of credit risk management affects all financial organisations that lend to both individuals and businesses. This is even more true for microcredit businesses as they only offer loans. Despite the fact that banks have a diversified portfolio, credit risk is still the most crucial to control. Credit risk is a financial loss that results from a counterparty's inability to meet their contractual commitments (such as the timely payment of interest or principal) or from a higher chance of default over the course of the transaction. Financial companies have historically used traditional linear, logit and probit regressions to model credit risk.

* Review of Literature

Khandani et al. (2010) have employed machine learning techniques to build nonlinear, non-parametric forecasting approaches to measure consumer credit risk. To identify credit cardholder’s defaults, the authors used a credit office data set and commercial bank customer transactions to establish a forecast estimation. Their results indicate cost savings from 6% to 25% of total losses when machine learning forecasting techniques are employed to estimate the delinquency rates. Besides, their study opens up questions of whether aggregated customer credit risk analytics may improve systematic risk estimation.

Yap et al. (2011) used historical payment data from a recreational club and established credit scoring techniques to identify potential club member subscription defaulters. The study results demonstrated that no model outperforms the others among a credit scorecard model, logistic regression, and a decision tree model. Each model generated almost identical accuracy figures.

Zhao et al. (2015) examined a multi-layer perceptron (MLP) neural network’s accuracy regarding estimating credit scores efficiently. The authors used a German credit dataset to train and estimate the model’s accuracy. Their results indicated an MLP model containing nine hidden units achieved a classification accuracy of 87%, higher than other similar experiments. Their study results proved the trend of MLP models’ scoring accuracy by increasing the number of hidden units.

In Addo et al. (2018) the authors examined credit risk scoring by employing various machine and deep learning techniques. The authors used binary classifiers in modelling loan default probability (DP) estimations by incorporating ten key features to test the classifier’s stability by evaluating performance on separate data. Their results indicated that the models such as the logistic regression, random forest, and gradient boosting modelling generated more accurate results than the models based on the neural network approach incorporating various technicalities.

Petropoulos et al. (2019) studied a dataset of loan-level data of the Greek economy of examining credit quality performance and quantification of probability default for an evaluating period of 10 years. The authors used an extended example of classifications of the incorporated machine learning models against traditional methods, such as logistic regression. Their results identified that machine learning models had demonstrated superior performance and forecasting accuracy through the financial credit rating cycle.

Provenzano et al. (2020) introduced machine learning models to compose credit rating and default prediction estimation. They used financial instruments, such as historical balance sheets, bankruptcy statutes, and macroeconomic variables of a Moody’s dataset. Using machine learning models, the authors observed excellent out-of-sample performance results to reduce the bankruptcy probability or improve credit rating.

* Motivation for the Problem Undertaken

With a strategy of disruptive innovation that prioritises the subscriber, a telecom service provider has introduced a number of products and built its company and organisation around the budget operator model, giving superior products to all price-conscious clients at lower prices.

They focus on offering their services and goods to low-income families and underprivileged consumers in order to aid them when they are in need because they recognise the value of communication and how it influences a person's life. They are working with an MFI to offer microcredit on cell phone balances with a 5-day repayment period. If the Consumer does not repay the lent money within the allotted five days, he is considered to be in default.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

We use accuracy score, classification report, and confusion matrix as our evaluation metrics because the case study calls for the classification of Labels "1" and "0," where "1" denotes a loan that has been paid off, or a non-defaulter, and "0" denotes a loan that has not been paid off, or a defaulter. For our complete model, we additionally display the AUC score and plot the AUC ROC curve.

* Data Sources and their formats

The offered sample data is taken from the client database. The client needs some predictions that might aid them in future investments and better customer selection in order to increase the selection of clients for credit.

There are 209593 entries and 36 columns. There are no null values in the dataset. There are 3 categorical columns, 12 numerical columns and 21 float columns. The description of the features are as follows:

1. label :Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}
2. msisdn :mobile number of user
3. aon :age on cellular network in days
4. daily\_decr30 :Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)
5. daily\_decr90 :Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)
6. rental30 :Average main account balance over last 30 days
7. rental90 :Average main account balance over last 90 days
8. last\_rech\_date\_ma :Number of days till last recharge of main account
9. last\_rech\_date\_da : Number of days till last recharge of data account
10. last\_rech\_amt\_ma : Amount of last recharge of main account (in Indonesian Rupiah)
11. cnt\_ma\_rech30 : Number of times main account got recharged in last 30 days
12. fr\_ma\_rech30 : Frequency of main account recharged in last 30 days
13. sumamnt\_ma\_rech30 : Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)
14. medianamnt\_ma\_rech30 : Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)
15. medianmarechprebal30 : Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)
16. cnt\_ma\_rech90 : Number of times main account got recharged in last 90 days
17. fr\_ma\_rech90 : Frequency of main account recharged in last 90 days
18. sumamnt\_ma\_rech90: Total amount of recharge in main account over last 90 days (in Indian Rupee)
19. medianamnt\_ma\_rech90: Median of amount of recharges done in main account over last 90 days at user level (in Indian Rupee)
20. medianmarechprebal90: Median of main account balance just before recharge in last 90 days at user level (in Indian Rupee)
21. cnt\_da\_rech30: Number of times data account got recharged in last 30 days
22. fr\_da\_rech30: Frequency of data account recharged in last 30 days
23. cnt\_da\_rech90: Number of times data account got recharged in last 90 days
24. fr\_da\_rech90: Frequency of data account recharged in last 90 days
25. cnt\_loans30 : Number of loans taken by user in last 30 days
26. amnt\_loans30 : Total amount of loans taken by user in last 30 days
27. maxamnt\_loans30 : maximum amount of loan taken by the user in last 30 days
28. medianamnt\_loans30 : Median of amounts of loan taken by the user in last 30 days
29. cnt\_loans90: Number of loans taken by user in last 90 days
30. amnt\_loans90 :Total amount of loans taken by user in last 90 days
31. maxamnt\_loans90 : maximum amount of loan taken by the user in last 90 days
32. medianamnt\_loans90: Median of amounts of loan taken by the user in last 90 days
33. payback30 :Average payback time in days over last 30 days
34. payback90: Average payback time in days over last 90 days
35. pcircle: telecom circle
36. pdate :date

* Data Pre-processing Done

The summary statistics shows all the statistics of our dataset i.e. mean, median and other calculation. Mean is greater than median in all the columns so aur data is right skewed. The difference between 75% and maximum is higher that's why outliers are high which needs to be removed. The pdate column tells the date when the data is collected. It contains only three-month data. The columns- msisdn column has mobile numbers of customers which are unique and entries are not available for every customer since there are 209593 entries and for msisdn there are only 186243 values, hence eliminating the column. The features- pdate\_year, pcircle has only one value for the complete dataset hence eliminating the same.

The features which are highly correlated with each other are daily\_decr30, daily\_decr90, rental30, rental90, cnt\_loans30, amount\_loans30, amount\_loans30, amount\_loans90 column medianamnt\_loans30, medianamnt\_loans90. We have to drop one of the features which are highly correlated with other features.

There are many outliers in the data which are removed using the z-score and the target variable is imbalanced and resampling algorithm is applied on the target column.

The features and target were split from the database and Chi square test was utilised in determining the feature importance and 21 features are required to retain all the information for model development.

* Hardware and Software Requirements and Tools Used

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® Distribution that allows you to launch applications and manage conda packages, environments, and channels without using command line interface (CLI) commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository. It is available for Windows, macOS, and Linux.

The Jupyter Notebook application allows you to create and edit documents that display the input and output of a Python language script.

**Model/s Development and Evaluation**

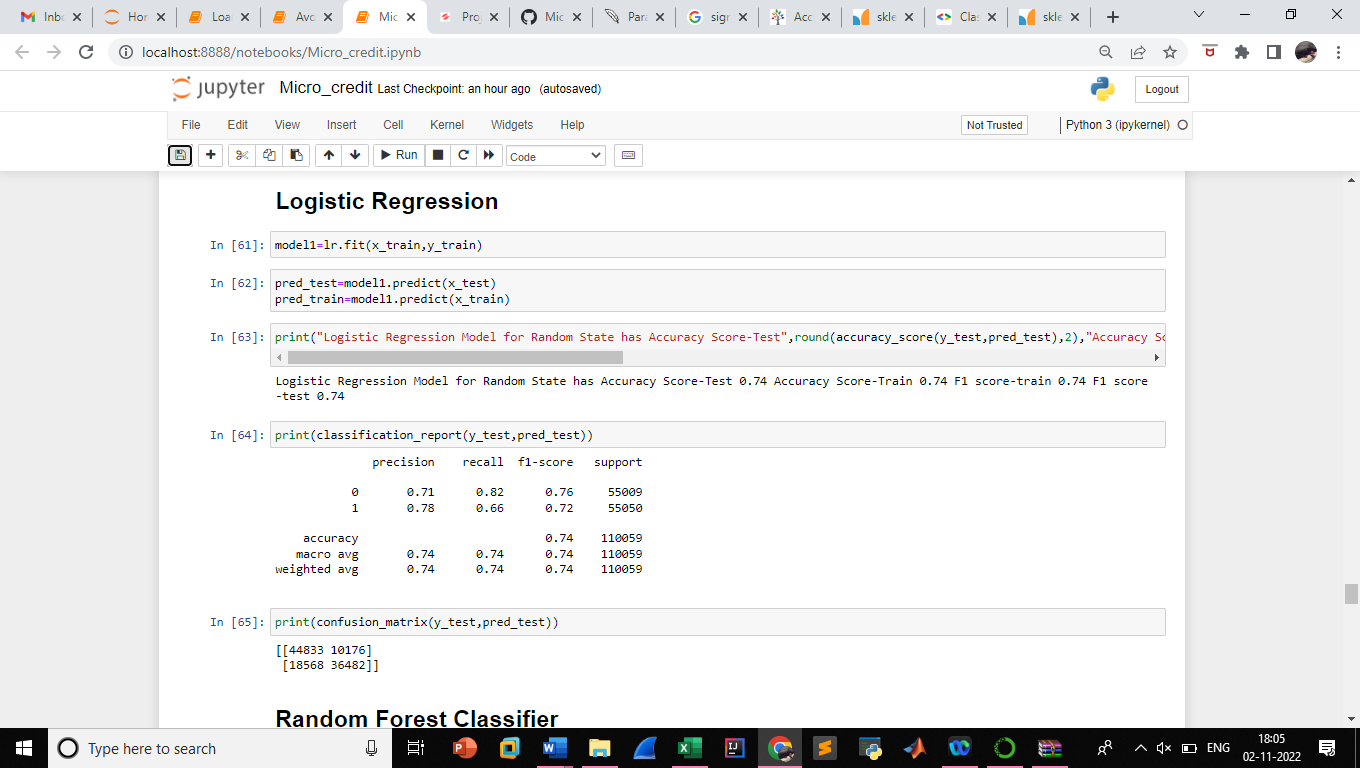
* Testing of Identified Approaches (Algorithms)

The features and target are split into train and test sections using the train\_test\_split function for random state-4 and test-size-0.30. The algorithms used for model training and testing are-

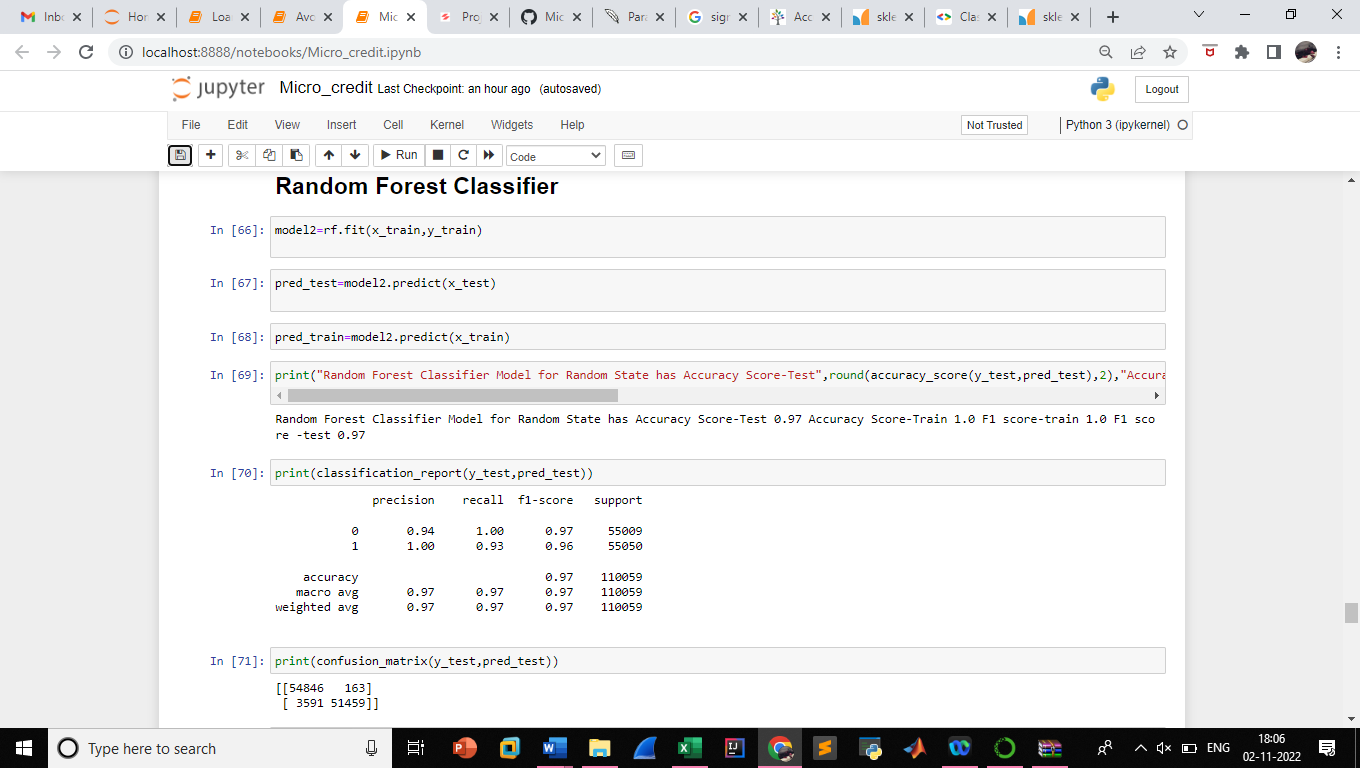
* 1. Logistic Regression
  2. Random Forest Classifier
  3. Decision Tree Classifier
  4. Gaussian NB Classifier
* Run and evaluate selected models

When the train and test data were used for developing models for the above algorithms the following observations were made:

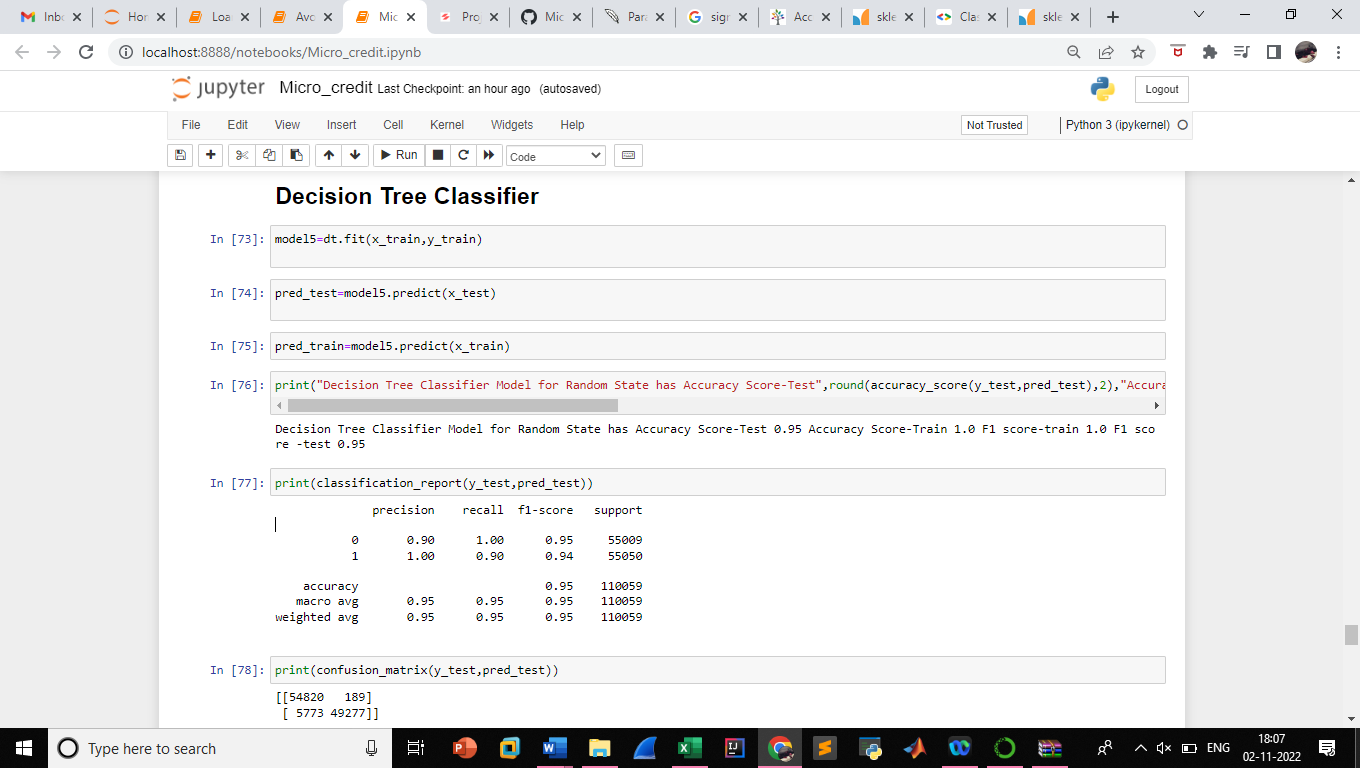
* Linear Regression models had no difference in accuracy score of training and testing data which is beneficial but the training score is only 74%.



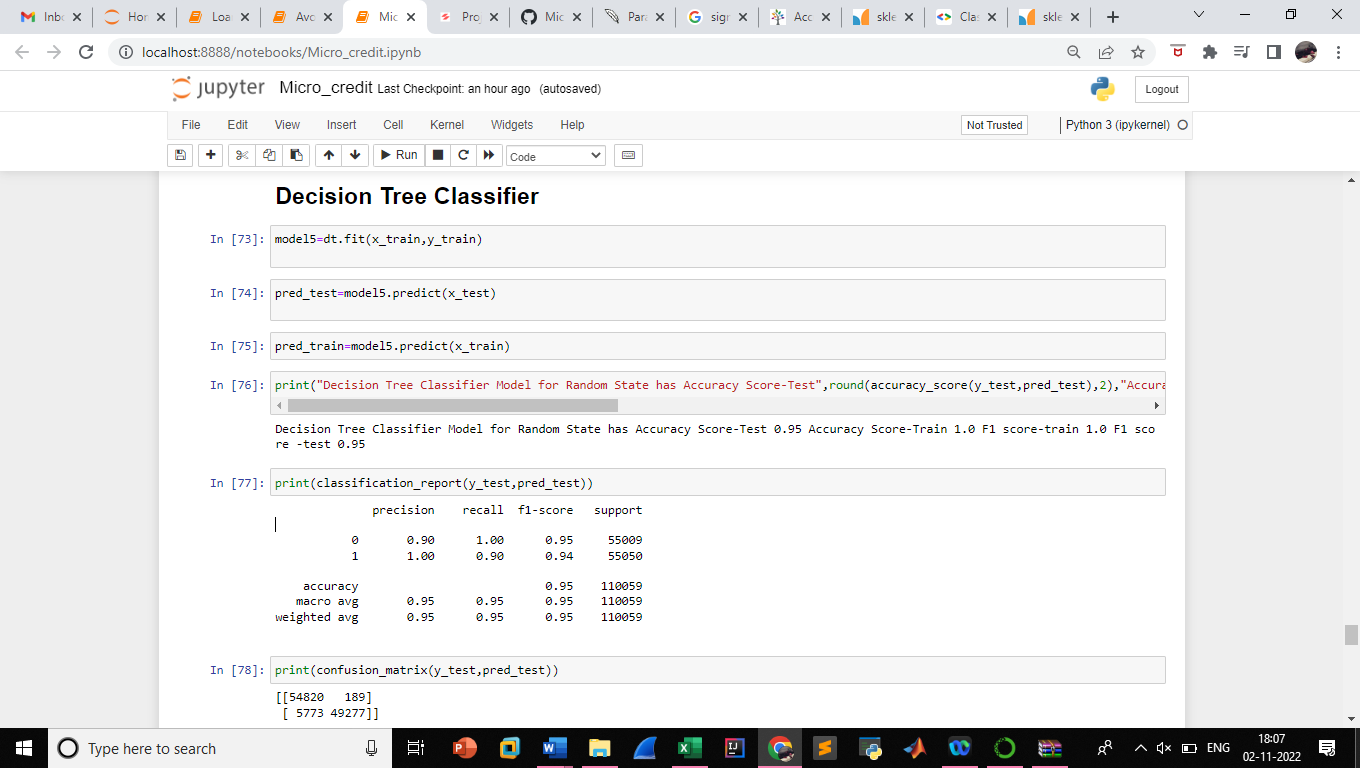
* Random Forest Classifier had difference of 3 units between accuracy score of training and test data.



* Decision Tree Classifier had difference of 5 units between accuracy score of training and test data.



* Gaussian NB Classifier had no difference of accuracy score between training and test data.



The four algorithms were employed in model development and the model was selected based on the accuracy score and f1 score on training and testing data and the results are as below,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy Score (%) | | F1-Score (%) | |
| Train | Test | Train | Test |
| Logistic Regression | 74 | 74 | 74 | 74 |
| Random Forest Classifier | 100 | 97 | 100 | 97 |
| Decision Tree Classifier | 100 | 95 | 100 | 95 |
| Gaussian NB Classifier | 66 | 66 | 64 | 64 |

* Key Metrics for success in solving problem under consideration

To forecast the target class of the data sample in classification issues, classification models are used. The likelihood that each occurrence belongs to a certain class is predicted by the categorization model. To effectively employ classifications models in production for resolving practical issues, it is critical to assess their effectiveness. Machine learning classification models' performance metrics are used to evaluate how well they perform in a specific situation. Accuracy, precision, recall, and F1-score are some of these performance indicators. Model performance is vital to machine learning since it enables us to comprehend the advantages and disadvantages of these models while making predictions in novel circumstances.

**CONCLUSION**

Random Forest Model is the effective model for the case study since it has a good training score and the difference between training and testing score is less compared to Decision Tree Classifier.

The model was further analysed using the hyperparameter tuning and from the results the False Negative reduced from 163 to 158.