IJARSCT

ISSN (Online) 2581-9429

IJARSCT
Impact Factor: 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 1, December 2022

Credit Card Defaulter Prediction

Mr. Vikas Singh[1], Mr. Hassan Rahim[2], Mr. Robin Rai[3], Mr. Aditya Suple[4], Mr. Ashwin Tijare[5], Mr. Dewam Katole[6], Ms. Alisha Badhel[7]

Assistant Professor, Department of Computer Science & Engineering[1]
Students, Department of Computer Science & Engineering[2,3,4,5,6,7]
G.H Raisoni Institute of Engineering and Technology, Nagpur, Maharashtra, India

Abstract: *Our aim is to develop a Machine learning model and testing the model by using the data in relating to previous 6 months payment behaviour which is behavioural data and personal information which is demographic data as input of a client is used for this study. The research study is conducted using Random Forest Algorithm. Our aim is to identify that credit card customer is likely to default in the coming month.*

Keywords: Random Forest classifier, Hyperparameter tuning, Grid search CV, Support Vector Machine, Logistic Regression, Flask

I. INTRODUCTION

In past decade Credit cards become the major consumer lending product following the personal loans in banking sector. Also, credit card is a most flexible and easiest way to use banks money in short term basis. Using credit cards people pay utility bills, purchase goods and services, implement instalment plans and do international transactions easily. But if the customer does not pay the monthly instalments of the credit cards, the credit card will be defaulted and it will impact negatively to the Banks. These defaulted credit cards will write offs at some extend and it will reduce the profit of the Banks drastically. These write-offs will result to significant financial losses to the Bank on top.

Even seemingly manageable debt like credit cards can get out of hand. Unemployment, a medical crisis, or a failed business are just some of the reasons that can affect your finances. It gets out of control first in this situation because of the penalty. Many of us can relate to this scenario where we have missed one or two credit card payments due to forgetting due dates or due to financial issues.

Credit risk plays a major role in the banking industry. Banking's main activities include granting loans, credit cards, investments, mortgages, etc. Credit cards are one of the fastest growing financial services offered by banks in recent years. However, as the number of credit card users increases, banks are facing rising credit card failure rates. Therefore, data analytics can provide solutions to address current phenomena and manage credit risk.

Due to the large number of banks selling credit cards, the phenomenon of credit card failure is gradually emerging. It is very important for banks to effectively identify credit card users who are at high risk of default. In general, you will have fewer delinquent payments compared to credit card customers who do not pay their loans in arrears.

II. METHODOLOGY

2.1 Random Forest Classifier

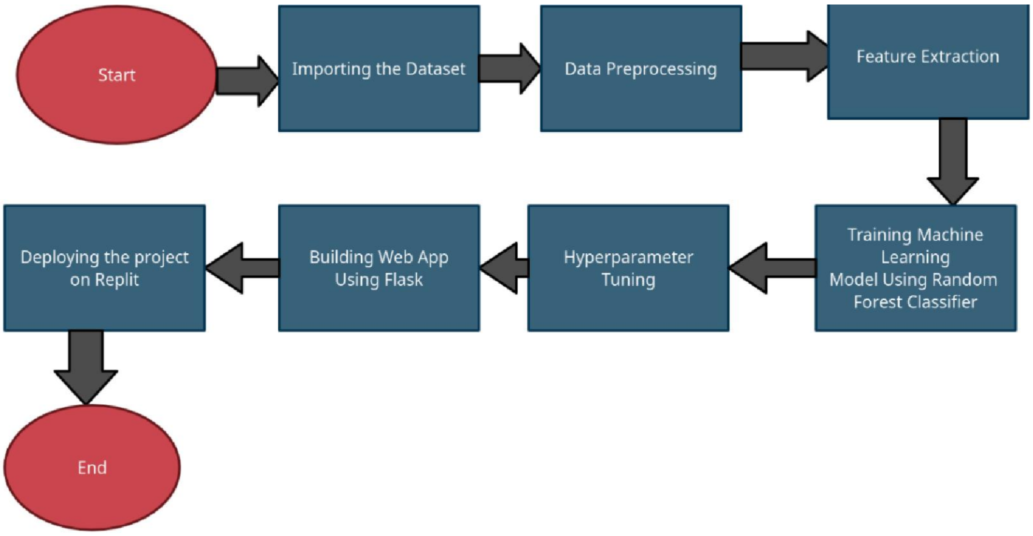
Random Forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

2.2 Hyperparameter Tuning

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is

DOI: 10.48175/IJARSCT-7705

hyperparameter tuning. Parameters which define the model architecture are referred to as hyperparameters and thus this process of searching for the ideal model architecture is referred to as hyperparameter tuning.


Start
Importing the Dataset
Data Preprocessing
Feature Extraction
Training Machine Learning Model Using Random Forest Classifier
Hyperparameter Tuning
Building Web App Using Flask
Deploying the project on Replit
End

First we import data set. We have imported data set from UCI machine learning algorithm. The second step of the methodology is data pre-processing in which we remove redundant data, null values etc. Then we did feature extraction in which we only select relevant data. We had total 25 different parameters out of which we removed 2 parameters which were Id's and result. Then the most important phase of building our machine learning model which is training and testing. We used 80% of our data set for training and 20% for testing. In our machine learning model we used Random Forest Classifier for training and testing purpose.

The purpose of using this algorithm is because it is most efficient algorithm as compared to other algorithms especially with the type of data set which our model is dealing with. This algorithm adds additional randomness to the model while growing trees. It searches for the best feature among the subset of features instead of looking for the most important feature. Hence overfitting problem is resolved, similarly it also takes average of the outputs of all the subtrees which improves the accuracy.

The next phase in the proposed methodology is hyperparameter tuning. Parameters that define the model architecture are hyperparameters and the process of finding the ideal model architecture is called hyperparameter tuning. This is done to improve the accuracy of our machine learning model. The next step is building a web API, this will be the main channel of interaction. The front end is divided into two sections- Demographic data and Behavioural data.

The demographic data section will contain input fields such as gender, educational qualification, age, sex, marital status and limit balance. The behavioural data section contains past six months payment history of the applicant. This data is then converted into a list using flask. Then the content of the list is rearranged according to requirement of the model. After organizing the input, it is fed to the machine learning model for prediction. After processing the input the model will predict either 1 or 0 (1 stating the user will be a defaulter in the next month and 0 stating that the user will not be a defaulter in the next month). The final step of the methodology is deployment. The deployment is done using Replit.

III. WORKING AND IMPLEMENTATION

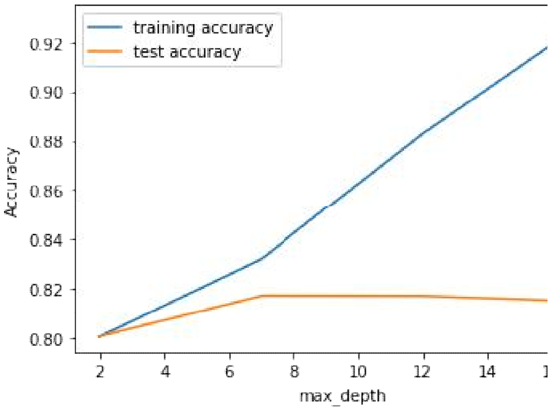
3.1 Hyperparameter Tuning

The following hyperparameters are present in a random forest classifier. We will tune each parameter: -

- n_estimators
- criterion
- max_features
- max_depth

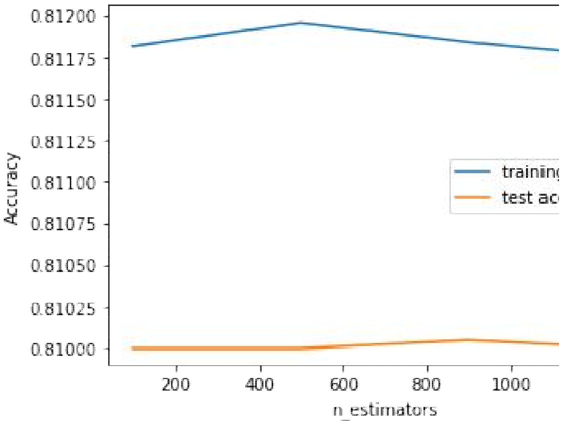
DOI: 10.48175/IJARSCT-7705

- min_samples_split
- min_samples_leaf
- min_weight_fraction_leaf
- max_leaf_nodes
- Min_impurity_split



As we increase the value of max_depth, both train and test scores increase till a point, but after that test score starts to decrease. The ensemble tries to overfit as we increase the max_depth. Thus, controlling the depth of the constituent trees will help reduce overfitting in the forest.

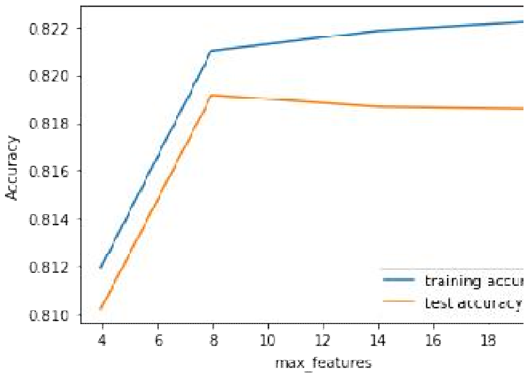
We'll find the optimum values for n_estimators and understand how the value of n_estimators impact the overall accuracy. We'll specify an appropriately low value of max_depth, so that the trees do not overfit.



3.2 Tuning max_features

Let's see how the model performance varies with max_features, which is the maximum number of features considered for splitting at a node

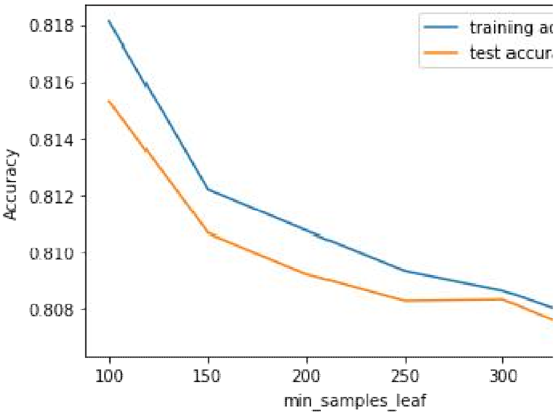
DOI: 10.48175/IJARSCT-7705



3.3 Tuning min_samples_leaf

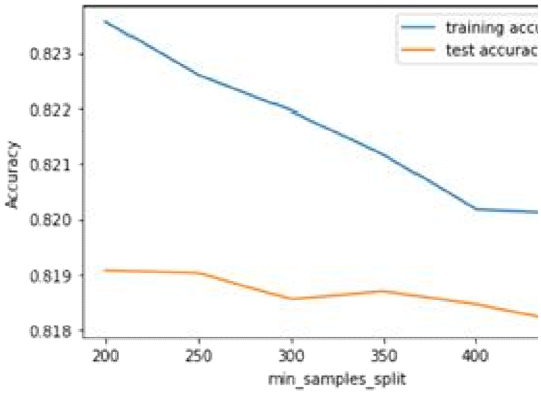
The hyperparameter min_samples_leaf is the minimum number of samples required to be at a leaf node:

- If int, then consider min_samples_leaf as the minimum number.
- If float, then min_samples_leaf is a percentage and ceil(min_samples_leaf * n_samples) are the minimum number of samples for each node.



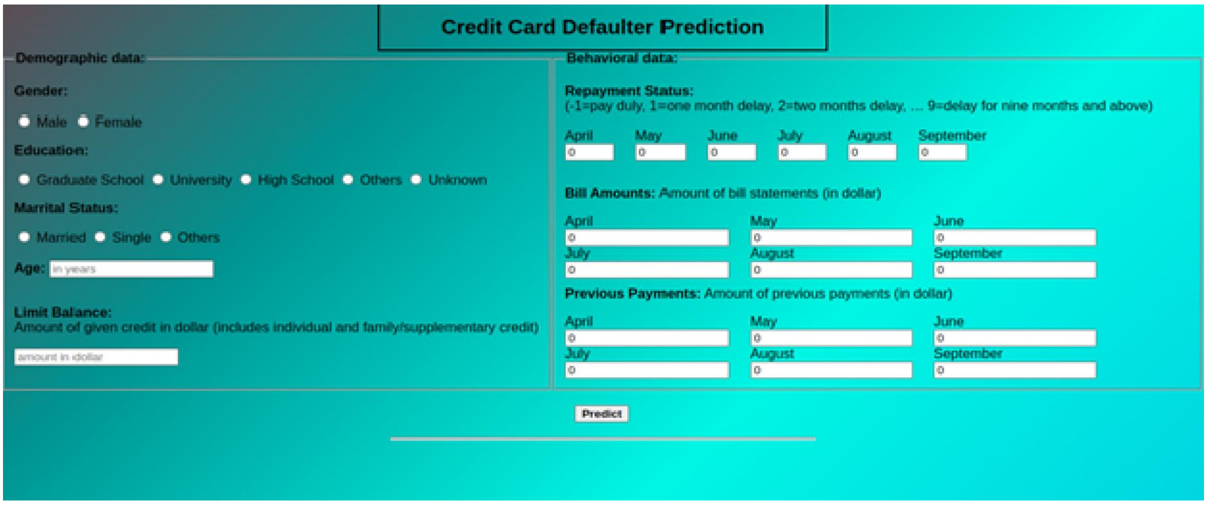
3.4 Tuning min_samples_split

- Let's now look at the performance of the ensemble as we vary min_samples_split.



DOI: 10.48175/IJARSCT-7705

3.5 Screenshot of Front End of the Project



Credit Card Defaulter Prediction

Demographic data:

Gender:

○ Male ○ Female

Education:

○ Graduate School ○ University ○ High School ○ Others ○ Unknown

Marital Status:

○ Married ○ Single ○ Others

Age: in years

Limit Balance:
Amount of given credit in dollar (includes individual and family/supplementary credit)

amount in dollar

Behavioral data:

Repayment Status:
(-1=pay duly, 1=one month delay, 2=two months delay, ... 9=delay for nine months and above)

April	May	June	July	August	September
0	0	0	0	0	0

Bill Amounts: Amount of bill statements (in dollar)

April	May	June
0	0	0
July	August	September
0	0	0

Previous Payments: Amount of previous payments (in dollar)

April	May	June
0	0	0
July	August	September
0	0	0

The above figure shows the front end of our project. This is where all the input are taken from the user to perform calculations. The input data is considered in two sections. i) Demographic data and ii) Behavioural data.

The first section in the left-hand side of the front end collects data as gender, education details, marital status, age, and limit balance which is used to understand the personal aspect of any client. These attributes prove a lot of certainty in prediction of results accurately.

The second section in the right-hand side of the front end also the most important part takes input of users or clients past payment history. Collecting such data is where our calculation depends upon. We use this information to feed our algorithm and at last on the basis of this it predicts where the applicant is a defaulter or not.

After filling both the section we press on predict button, this leads to starting of the calculation process and which few sections it displays the result.

IV. RESULT

The final outcome of the project is to predict whether the person will be a defaulter in the next month or not. The accuracy of the model according to various papers which we've gone through was around 80%, and our project's accuracy of the model is more than 90 %. We were able to build a machine learning model using Random Forest algorithm which gives us more accurate results.

This project can help banks in predicting the future of Credit Card non performing or not and its status and depends on that they can take action in initial stage of approving credit card

According to the results obtained, it can be seen that the model outperformed various existing models which had an overall accuracy of mostly 80% in predicting the default rate.

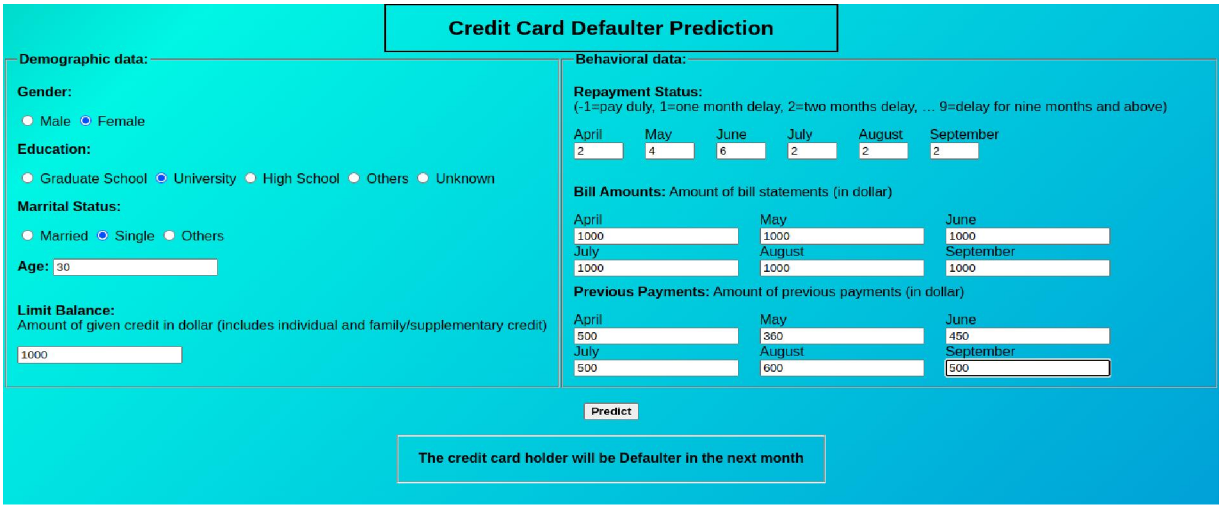
The project has a Web API which takes the input of the customer to predict their defaulter probability.

When compared to a model created from a separate study, the model's performance was evaluated using the criteria of accuracy, recall, and precision, indicating greater performance.

The system provided a user functionality and interfaced with the default prediction model to perform prediction. A web application provided the user interface for the loan officer to upload credit cards data and to obtain predictions of loans.

The goal of this study was to learn from credit card account data and identify patterns in the data that could be used to forecast the likelihood of loan default which was achieved efficiently to provide us with accurate predictions.

4.1 Screenshot of Final Output



Credit Card Defaulter Prediction

Demographic data:

Gender:

○ Male ● Female

Education:

○ Graduate School ● University ○ High School ○ Others ○ Unknown

Marrital Status:

○ Married ● Single ○ Others

Age: 30

Limit Balance:
Amount of given credit in dollar (includes individual and family/supplementary credit)

1000

Behavioral data:

Repayment Status:
(-1=pay duly, 1=one month delay, 2=two months delay, ... 9=delay for nine months and above)

April	May	June	July	August	September
2	4	6	2	2	2

Bill Amounts: Amount of bill statements (in dollar)

April	May	June
1000	1000	1000
July	August	September
1000	1000	1000

Previous Payments: Amount of previous payments (in dollar)

April	May	June
500	360	450
July	August	September
500	600	500

The credit card holder will be Defaulter in the next month

V. CONCLUSION

- In future we can become a middleware software for banks and ecommerce sites. In future, we are hoping that more state-of-the art machine learning models will perform well than our existing model.
- By mapping relationship between a person's bank statement and behavioural data we can provide him with targeted advertisements and suggestions which will save huge amount on spent on irrelevant audience. Similarly it can also enhance prediction rate of banks for credit card default.
- The findings and this model have emphasized that the proper prediction of nonperforming loans and advances that can be arising in the future by using modern technologies (machine learning approaches) and then bankers can understand customer properly and provide accurate decision whether it is approving or not the applied credit card easily.
- The use of machine learning techniques for the prediction of credit card defaulters is essential for the identification of credit risk. This can help the financial institutions in designing their future strategies. The proposed system uses Random Forest Algorithm on credit scoring data set in relating to Credit card applications of Bank
- Analysis of cardholders' borrowing and repayment patterns can be used to predict delinquency, making it crucial to take into account while managing risk portfolios for financial institutions. While elements connected to customers' transaction activities are less effective predictors of eventual loan status than cardholder attributes, it is important to remember that information about these aspects accumulates over time and is more dynamic than personal qualities. This indicates that it will be possible to improve performance on default prediction models given enough time and data about the trends of these variables.
- This research helped to successfully develop a model for default prediction and a system that uses that model to give users access to the model's prediction functionality.
- Customers may be classified as defaulters or non-defaulters according to different standards at another bank. The model created for this study can be expanded to forecast default events regardless of how the prerequisites for someone to be labelled a defaulter have been set up.

REFERENCES

[1]. Talha Mahboob Alam, Kamran Shaukat, Ibrahim A. Hameed, Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li, MatloobKhushi(2020). An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. DOI: 10.1109/ACCESS.2020.3033784.

Link:-https://ieeexplore.ieee.org/document/9239944

[2]. Ying Chen, Ruirui Zhang (2021). Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network. DOI:10.1155/2021/6618841.
Link:-https://www.hindawi.com/journals/complexity/2021/6618841/

[3]. Abdulhamit Subasi, SelcukCankurt (2019). Prediction of default payment of credit card clients using Data Mining Techniques. DOI: 10.1109/IEC47844.2019.8950597.
Link:-https://ieeexplore.ieee.org/document/8950597

[4]. Mohammad Aman Ullah, Mohammad ManjurAlam, Shamima Sultana, Rehana Sultana Toma(2018).Predicting Default Payment of Credit Card Users: Applying Data Mining Techniques.
Link:-https://ieeexplore.ieee.org/document/8745571

[5]. Pu Xu, Zhijun Ding, MeiQin Pan (2017). An improved credit card users default prediction model based on RIPPER. DOI: 10.1109/FSKD.2017.8393037.
Link:-https://ieeexplore.ieee.org/document/8393037

[6]. **Dataset:** Default of credit card clients Data Set.
Link:-https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

DOI: 10.48175/IJARSCT-7705