

NAME OF THE PROJECT

MICRO CREDIT DEFAULTER MODEL

Submitted by:

V TARAK RAM SAI

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

This is regarding a telecom industry collaborating with Microfinance institution to provide micro-credit on mobile balances to be paid back in 5days. We need to develop a model which can be used to predict in terms of a probability for each loan transaction, whether the customer is paying back the loaned amount within 5 days of insurance of loan.

* Conceptual Background of the Domain Problem

This is a Classification problem. The classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset and then classifies new observations into a number of classes or groups.

* Review of Literature

This project reviews the literature on the traditional credit risk assessment methods, the machine learning algorithms that are increasingly being used to evaluate credit risk and the relevance of the literature to the problem statement.

The type of data used in traditional credit risk assessment are the past credit, records of late payment and all the credit history. Banks must put in a lot of effort to determine whether a customer can pay back the loan amount on time or not.

* Motivation for the Problem Undertaken

The purpose of this problem is to determine the nature, background, and credibility of the client who is seeking a loan. To cope with challenge of granting or rejecting a loan request, or in short loan prediction, we apply exploratory data analysis. The purpose of this model is to see if a loan made to a specific individual or organization will be approved.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Any machine learning model should follow the below steps while dealing a business problem. They are:

**i.)** **Business Understanding:** The first step is to comprehend the research’s background, the problem description, and how the proposed project will achieve the goals.

**ii.**) **Data Understanding:** The second stage requires collection of data listed in the project resources. This involves in determining the data requirements and exploring key data attributes.

**iii.**) **Data Preparation:** The third stage involves data cleaning and should the handle the missing values in the data.

**iv.**) **Modelling:** This involves determining the modelling technique and testing the design.

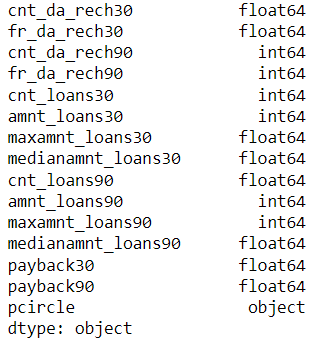
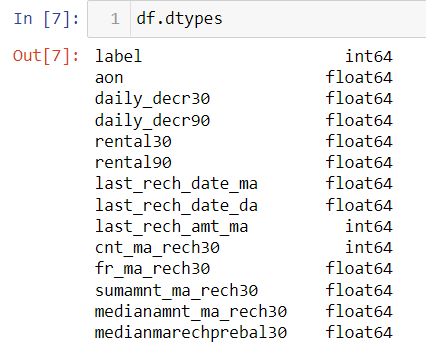
**v.**) **Evaluation:** Here, we should evaluate the achieved results and should determine the performance of the model with best accuracy.

**vi.**) **Deployment:** The last stage is implementation of the model.

* Data Sources and their formats

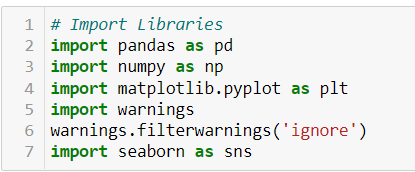
|  |  |
| --- | --- |
| label | Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure} |
| msisdn | mobile number of user |
| aon | age on cellular network in days |
| daily\_decr30 | Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah) |
| daily\_decr90 | Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah) |
| rental30 | Average main account balance over last 30 days |
| rental90 | Average main account balance over last 90 days |
| last\_rech\_date\_ma | Number of days till last recharge of main account |
| last\_rech\_date\_da | Number of days till last recharge of data account |
| last\_rech\_amt\_ma | Amount of last recharge of main account (in Indonesian Rupiah) |
| cnt\_ma\_rech30 | Number of times main account got recharged in last 30 days |
| fr\_ma\_rech30 | Frequency of main account recharged in last 30 days |
| sumamnt\_ma\_rech30 | Total amount of recharge in main account over last 30 days (in Indonesian Rupiah) |
|  |  |
| medianamnt\_ma\_rech30 | Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah) |
| medianmarechprebal30 | Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah) |
| cnt\_ma\_rech90 | Number of times main account got recharged in last 90 days |
| fr\_ma\_rech90 | Frequency of main account recharged in last 90 days |
| sumamnt\_ma\_rech90 | Total amount of recharge in main account over last 90 days (in Indonasian Rupiah) |
| medianamnt\_ma\_rech90 | Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah) |
| medianmarechprebal90 | Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah) |
| cnt\_da\_rech30 | Number of times data account got recharged in last 30 days |
| fr\_da\_rech30 | Frequency of data account recharged in last 30 days |
| cnt\_da\_rech90 | Number of times data account got recharged in last 90 days |
| fr\_da\_rech90 | Frequency of data account recharged in last 90 days |
| cnt\_loans30 | Number of loans taken by user in last 30 days |
| amnt\_loans30 | Total amount of loans taken by user in last 30 days |
| maxamnt\_loans30 | maximum amount of loan taken by the user in last 30 days |
| medianamnt\_loans30 | Median of amounts of loan taken by the user in last 30 days |
| cnt\_loans90 | Number of loans taken by user in last 90 days |
| amnt\_loans90 | Total amount of loans taken by user in last 90 days |
| maxamnt\_loans90 | maximum amount of loan taken by the user in last 90 days |
| medianamnt\_loans90 | Median of amounts of loan taken by the user in last 90 days |
| payback30 | Average payback time in days over last 30 days |
| payback90 | Average payback time in days over last 90 days |
| pcircle | telecom circle |
| pdate | Date |

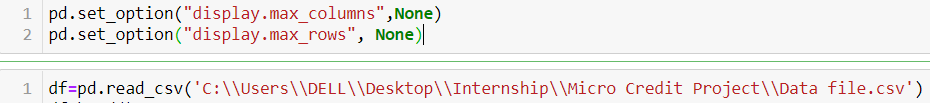
The format of data is as follows



* Data Pre-processing

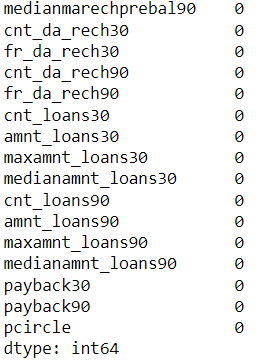
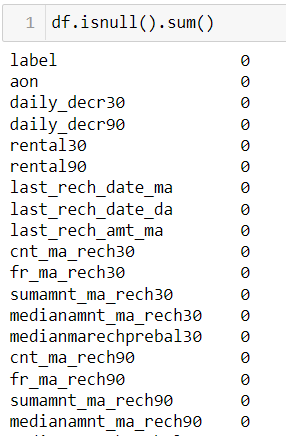
It entails converting raw data into comprehensible format that a machine learning model can understand. The data pre-processing involves data cleaning which involves handling missing values, transformation of data i.e. normalizing the data and data reduction which involves only required features.





**Data Cleaning**

The first step of pre-processing is data cleaning by checking and eliminating any missing values because they affect the accuracy of the model. This is achieved by either filling the missing values with a mean or mode function or by dropping all the missing values. In this case, there are no missing values.



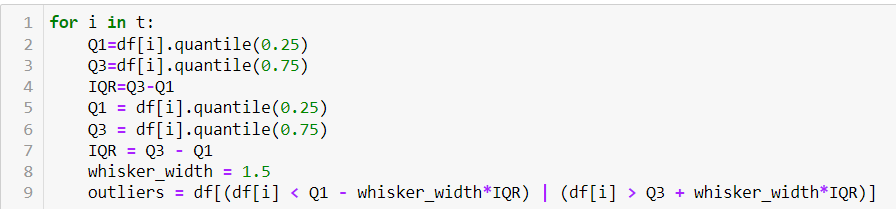
**Data Reduction**

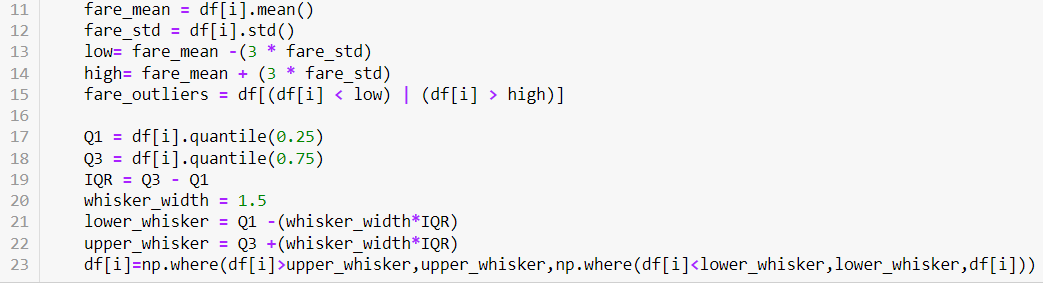
The next step of data processing is data reduction. This is used to remove duplicate features present in the data i.e. unwanted features for prediction.





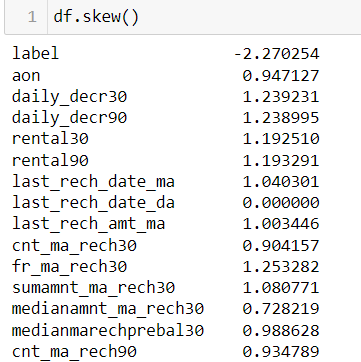
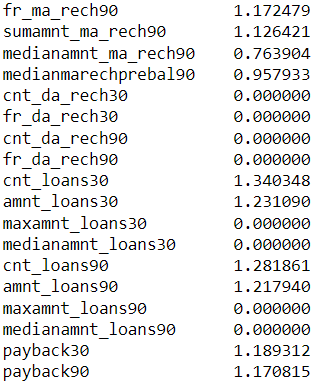
**Handling Outliers**





### Instead of eliminating the outliers, we have replaced the outliers higher than the upper whisker by the value of upper whisker and the outliers lower than the lower whisker by the value of lower whisker.

**Skewness Removal**

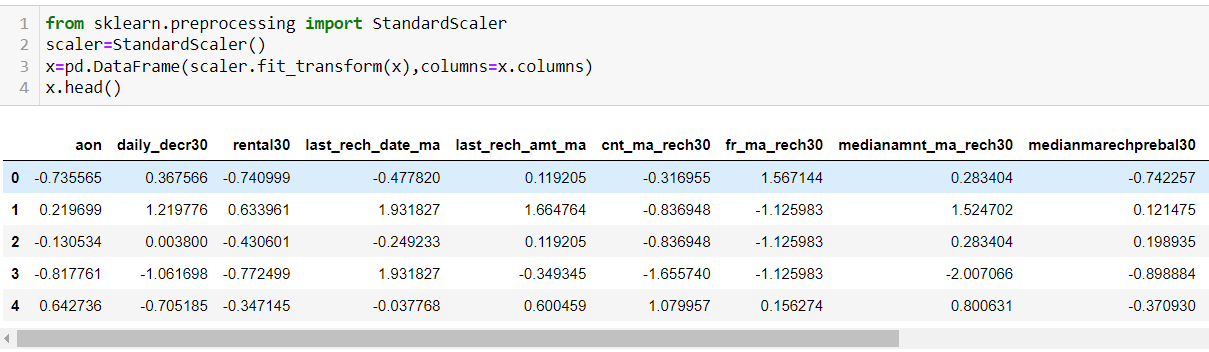
We have to remove the skewness from the columns whose skewness is not in the range of -0.5 to +0.5. Here, we are using a power transformer to remove the skewness as follows



This will remove skewness from the data.

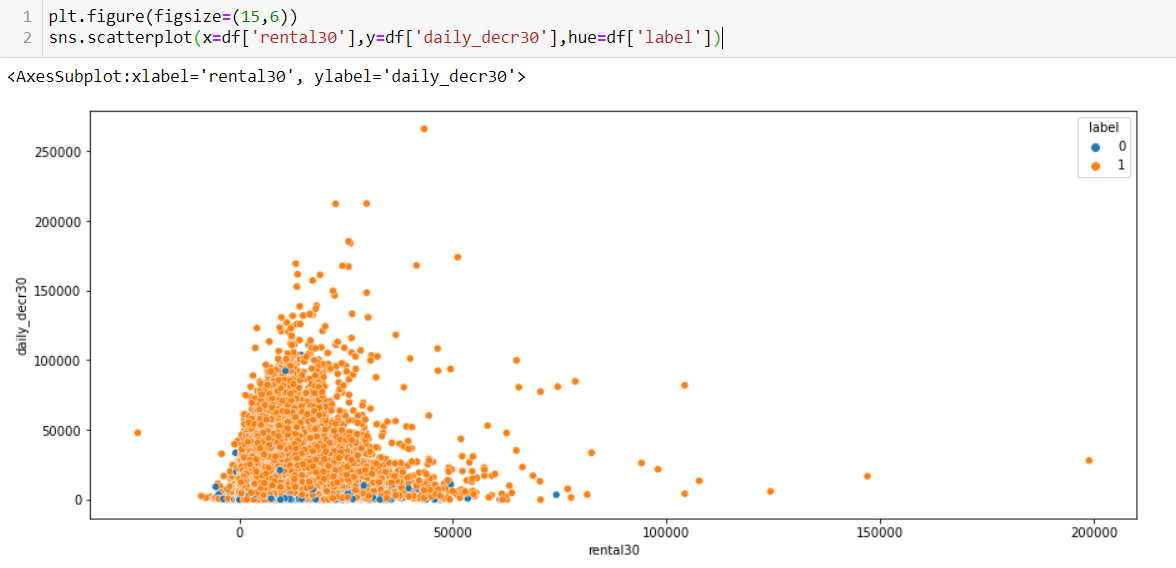
**Scaling**

Feature scaling is **a method used to normalize the range of independent variables or features of data**. To convert data into a distribution with a mean of 0 and standard deviation of 1, we will use a standard scalar.

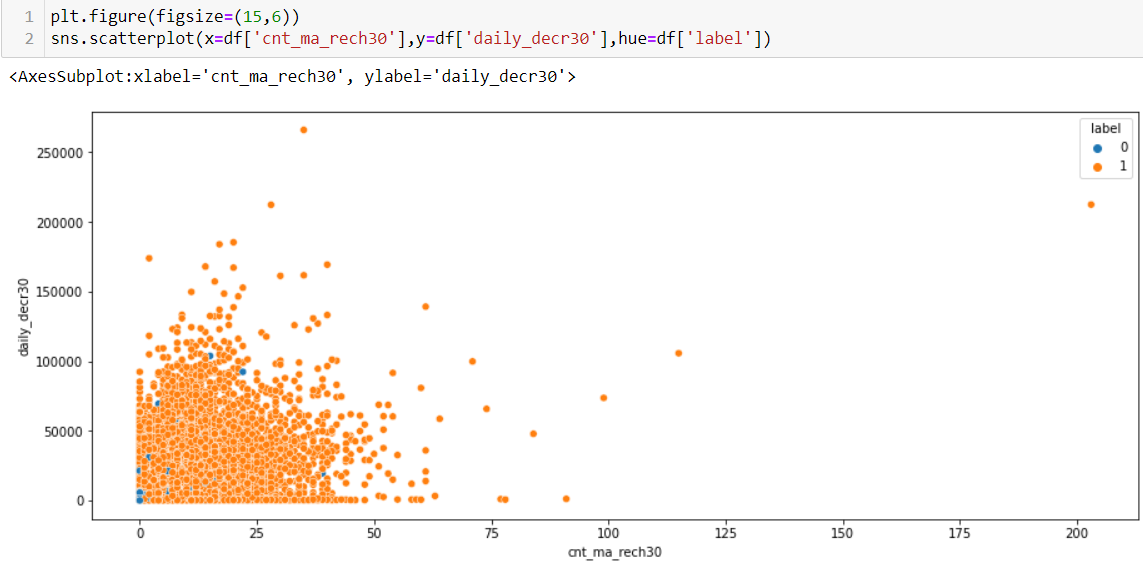


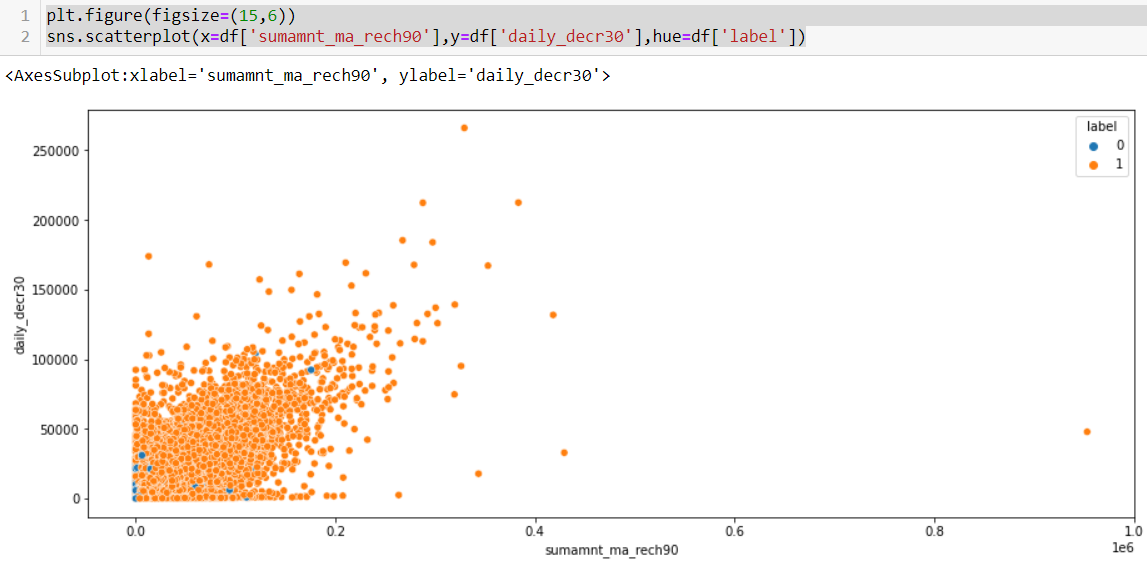
* Data Inputs- Logic- Output Visualization

The Relation between input and output variables are determined in the below graphs as follows

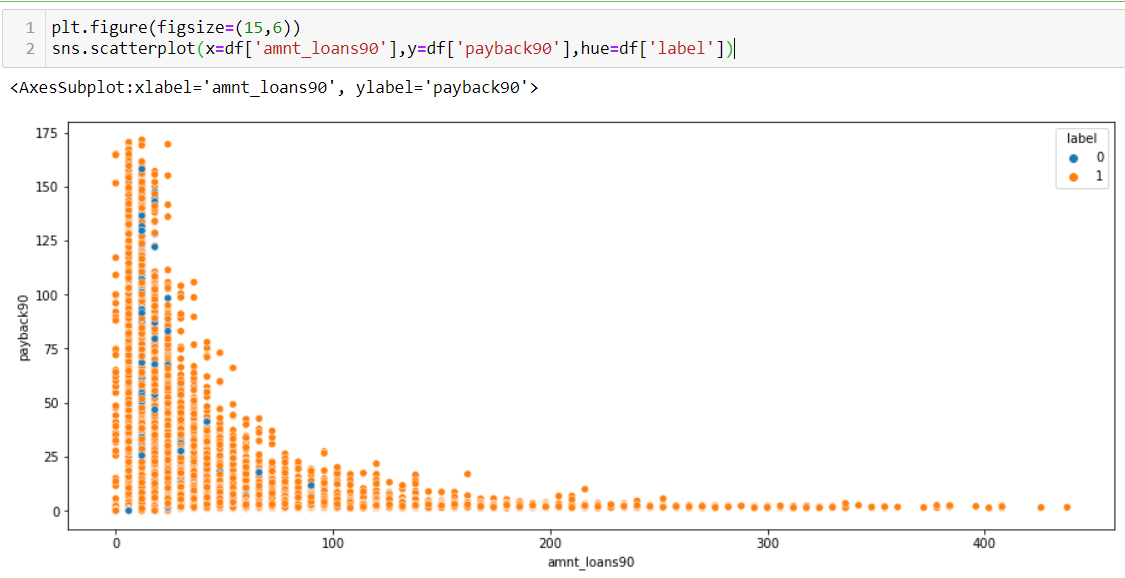


### The average main account balance is mostly in between 0 and 50000. Daily amount spent from main account, averaged over last 30 days is mostly between 0 and 100000.

Number of times main account got recharged in last 30 days is mostly between 0 and 40, If the no. of times the customer recharged is more, then the customer is a non-defaulter.



### Total amount of recharge in main account over last 90 days is positively related to daily\_decr30.



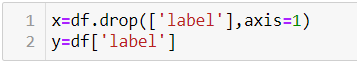
### If the Total amount of loans taken by user in last 90 days is more than the Average payback time in days over last 90 days is less.

* Hardware and Software Requirements and Tools Used

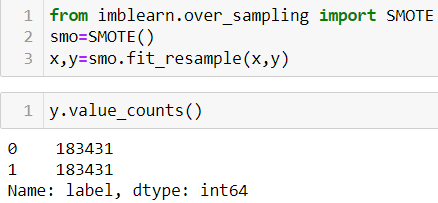
The hardware requirements for the project includes a laptop with atleast 4GB RAM. This project uses a Jupyter Notebook as a code editor. The Machine Learning models are implemented using python version 3.7 with libraries like numpy, pandas, matplotlib, seaborn and sklearn.

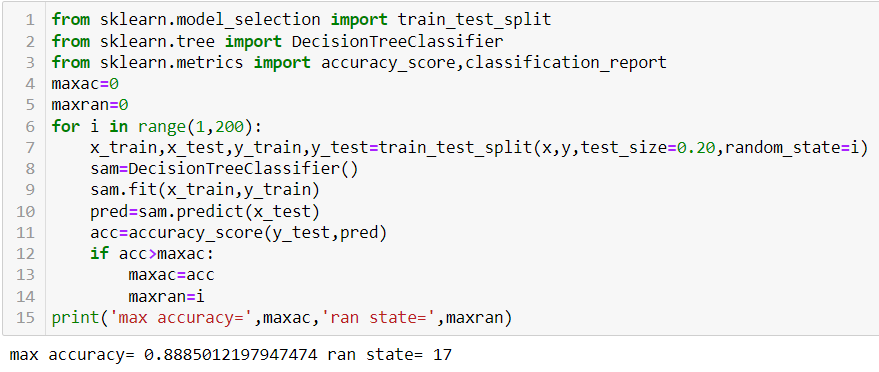
**Model/s Development and Evaluation**

The independent variables are declared in x and the dependent variable i.e. ‘label’ is declared in y as follows



Now sampling should be done, in order to eliminate priority on the output variable by the model.

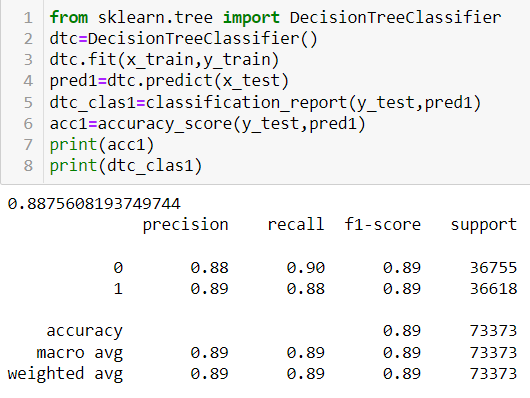




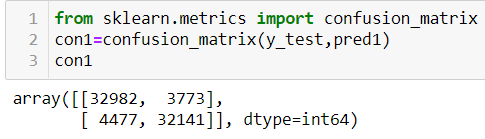


The above code is done for choosing the Random state variable. We should do testing by using any of the four classification algorithms.

**Decision Tree Classifier Algorithm**

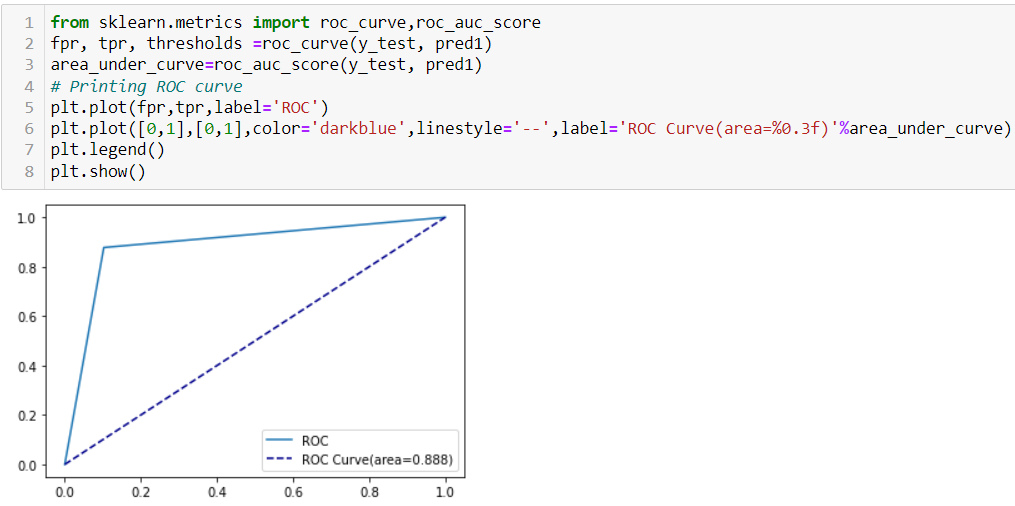


**Confusion Matrix**



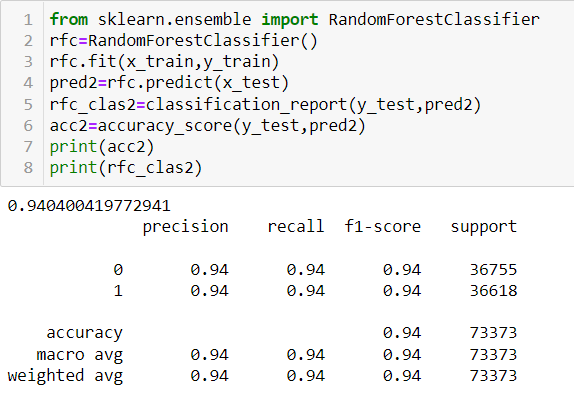
The right diagonal elements in the confusion matrix shows the correctly predicted values by the model.

**AUC-ROC Curve**

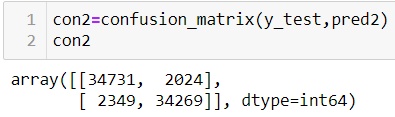


Higher the area under the curve, better the performance of the model.

**Random Forest Classifier Algorithm**

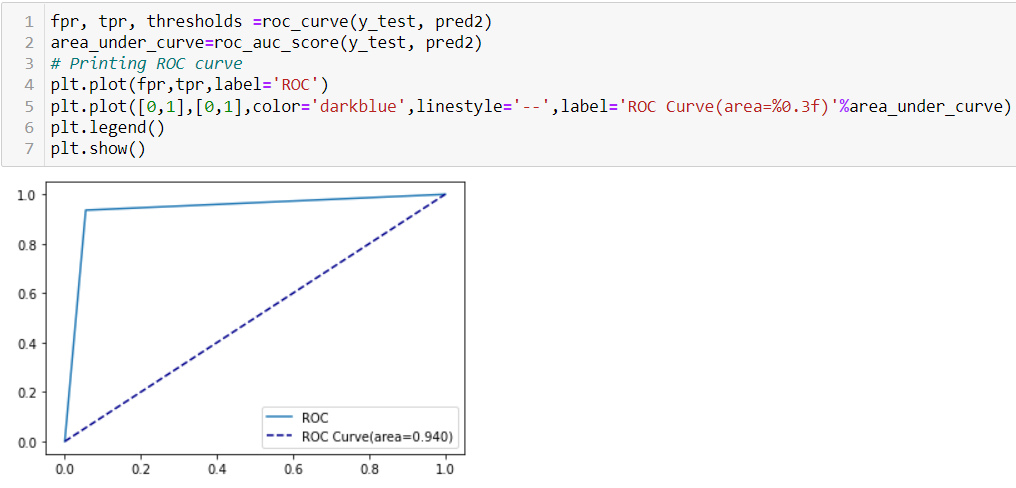


**Confusion Matrix**

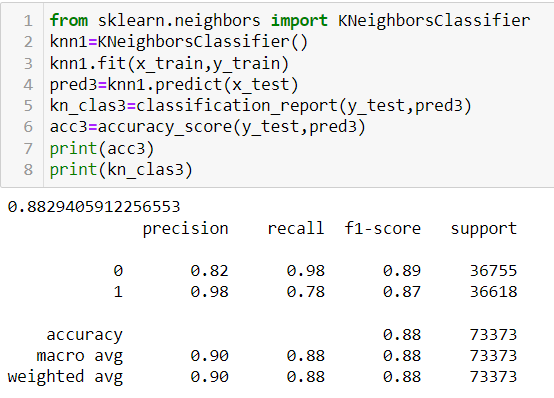


The right diagonal elements in the confusion matrix shows the correctly predicted values by the model.

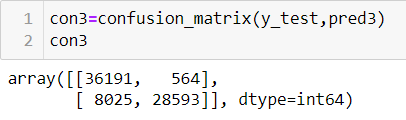
**AUC-ROC Curve**



Higher the area under the curve, better the performance of the model. **KNN Classifier Algorithm**

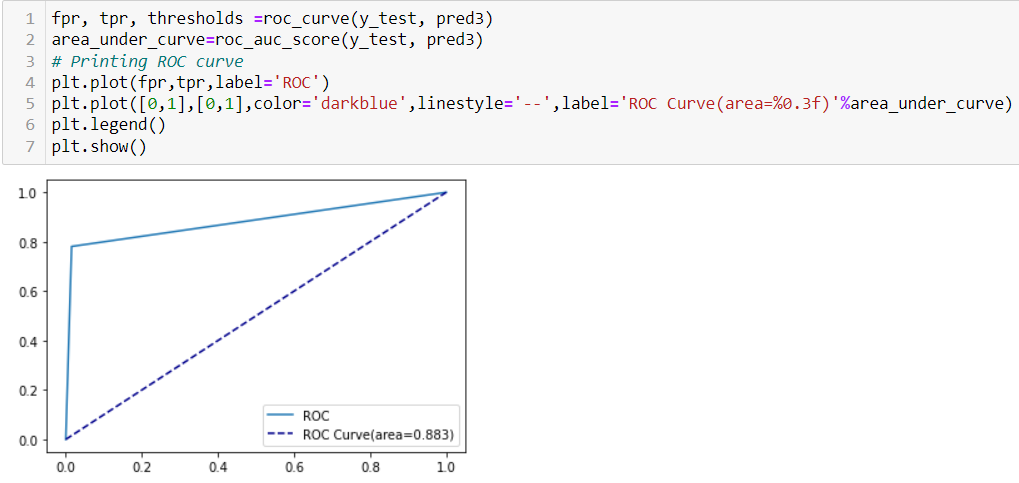


**Confusion Matrix**

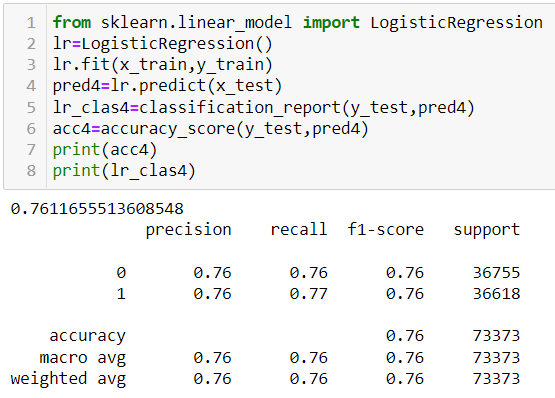


The right diagonal elements in the confusion matrix shows the correctly predicted values by the model.

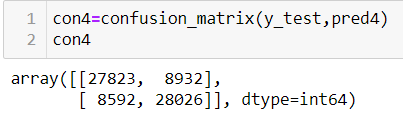
**AUC-ROC Curve**



Higher the area under the curve, better the performance of the model. **Logistic Regression Algorithm**

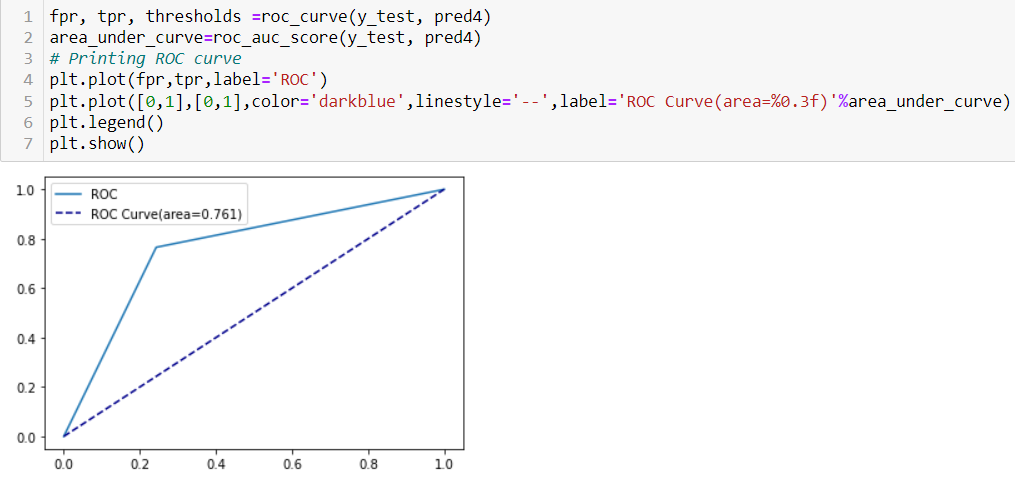


**Confusion Matrix**



The right diagonal elements in the confusion matrix shows the correctly predicted values by the model.

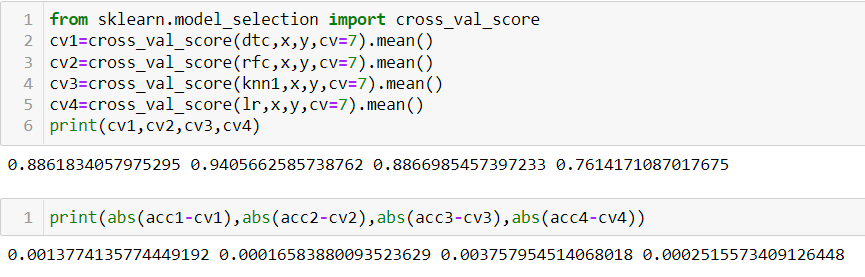
**AUC-ROC Curve**



Higher the area under the curve, better the performance of the model.

**Cross Validation**

Cross Validation is a **very useful technique for assessing the effectiveness of your model**, particularly in cases where you need to mitigate overfitting.



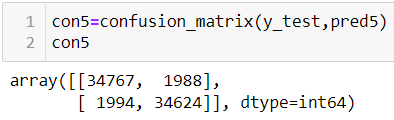
### We can choose Random Forest as our model since its cv\_score and accuracy score are almost similar.

**Hyper Parameter Tuning**

In machine learning, hyper parameter optimization or tuning is **the problem of choosing a set of optimal hyper parameters for a learning algorithm**. A hyper parameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

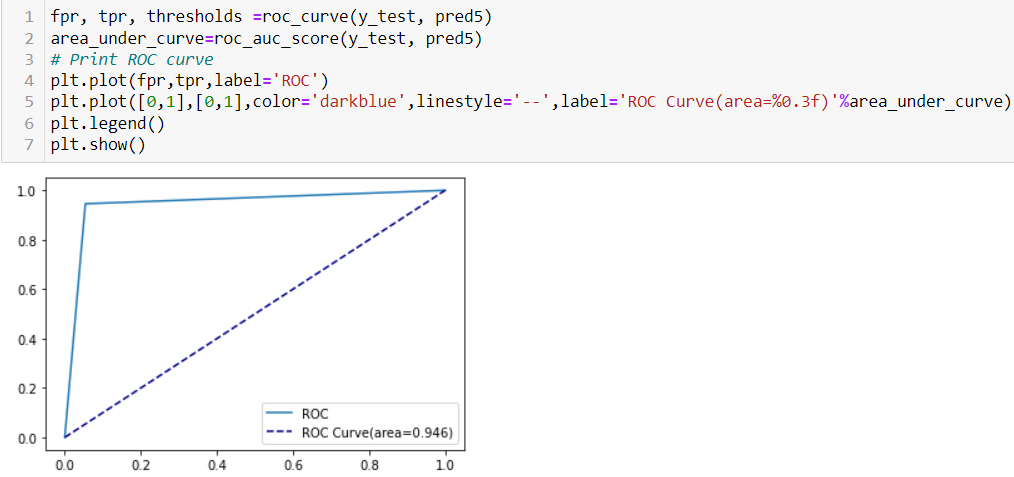


**Confusion Matrix of Final Model**



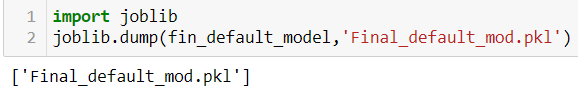
The right diagonal elements in the confusion matrix shows the correctly predicted values by the model.

**AUC-ROC Curve of Final Model**



Higher the area under the curve, better the performance of the model. By performing Hyper Parameter Tuning, we managed to increase the performance of the model.

**Saving the Model**



The above code will save the model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The applications of machine learning techniques in the financial sector with the goal of profit maximization has seen a rising interest over the last few years. There has been increasing number of research conducted in the areas of credit scoring, risk management and bankruptcy prediction using machine learning approaches.

* Learning Outcomes of the Study in respect of Data Science

This research proposes machine learning as a method to improve the accuracy of loan default predictions. This better understanding of customer behaviours to improve the prediction of loan default will contribute to tremendous financial benefit in the mobile lending sector. Exploratory data analysis shows the correlation of various features with loan default to select the most appropriate features to train the machine learning model. The train and test set are then applied to four machine learning algorithms to determine the one with most accurate results. Key performance metrics which include confusion matrix, accuracy, precision and recall are applied to evaluate the best learning technique in loan defaulter prediction.

* Limitations of this work and Scope for Future Work

Machine Learning algorithms are limited to the dataset used to train and test the model. This limits the generalization of the model as it specifies towards the dataset used to train and test the model. It would be beneficial to look comprehensively at the main features that are relevant to the characteristic that drive default and can be applied.

This research explores using machine learning algorithms to improve the accuracy of predicting loan default. This is a fair performance and can be further improved through different methods of parameter tuning and feature selection which may possibly yield improvements in the model performance. Since, the research is also limited to the probability of default state, further exploration may be made in determining the expected return of the loan based on borrower’s characteristics.