

**Micro-Credit Defaulter**

Submitted by:- Ayush Pathak

**ACKNOWLEDGMENT**

I would like to express my gratitude towards Flip-Robo for giving me this opportunity to show case my talent and also for their constant support and guidance. Also, I would like to thank data-trained for providing me internship. I express my deepest thanks to **Sajid Choudhary**, for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make my project easier. I choose this moment to acknowledge his contribution gratefully. I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

Thanks

Ayush Pathak

**INTRODUCTION**

* Business Problem Framing
* Loans of value 6 & 12 rupees are provided by our client (telecom operator) in collaboration with a Microfinance Institute (MFI).
* High return (20% return within 5 days) as well as High risk venture (high risk of default as loan is being provided to low income populations)
* The dataset contains default status of users along with associated features.
* Let’s also look into what the distribution of a feature means as well as how these features relates with loan defaults
* Objective is to study the behaviour of defaulters as well as prepare a machine learning model to classify all defaulters using the sample dataset provided by the client
* Conceptual Background of the Domain Problem

We are going to test multiple classifier algorithms with multiple evaluation metrics and select the best model based on proper metric, perform RandomSearchCV for best parameter settings.

* Review of Literature

First of all data is loaded and then it can be observed that, this is a classification based problem. It is clearly visible that, there is imbalance target feature’s(Label) and also the dataset needs cleaning as there are a lot of garbage & outliers values, luckily there’s no missing value. Many feature’s which are not having any relation with default status or have highly skewed data needs to be dropped before data analysis & modelling.

* Motivation for the Problem Undertaken

As, after looking into the dataset, it is clearly visible that this complex but handling such datasets brings your perception, judgemental ideas and curiosity to the next level.And, after working on this assignment many challenges I have faced, that boosten up my confidence too and a lot more things I have to learn also.

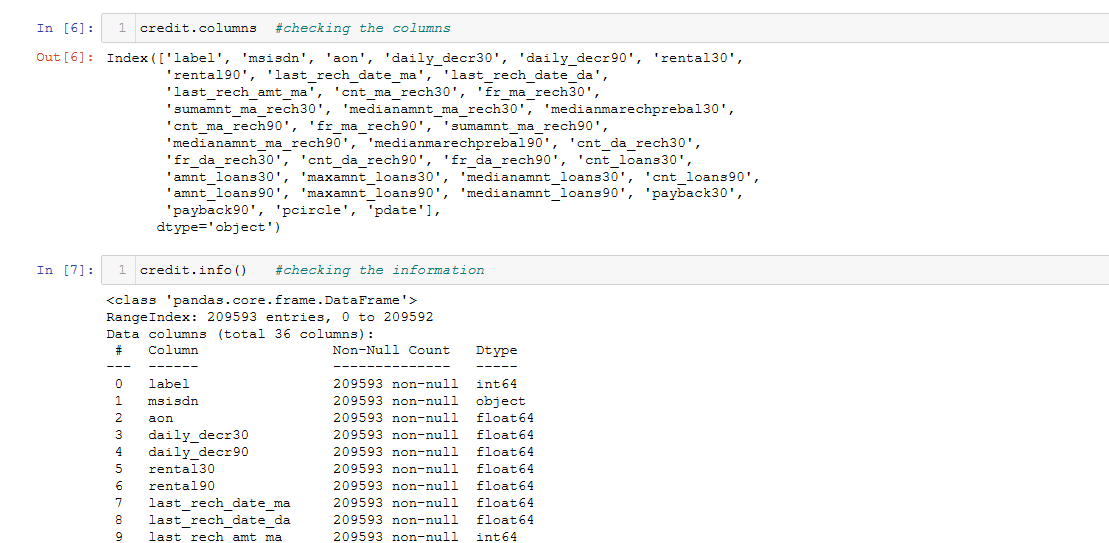
**Analytical Problem Framing**

The sample dataset is provided to us from our client database.

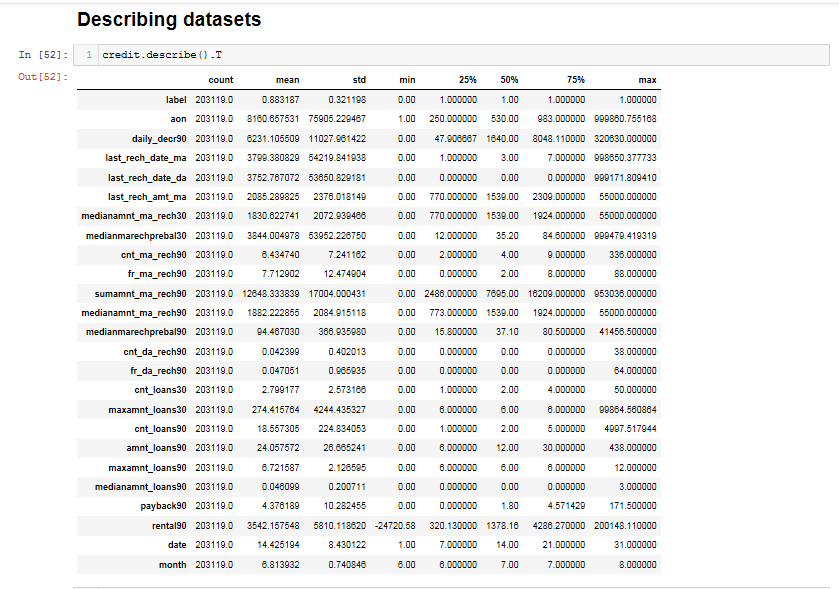
1) Firstly data is being loaded and previewing the first 5 rows.



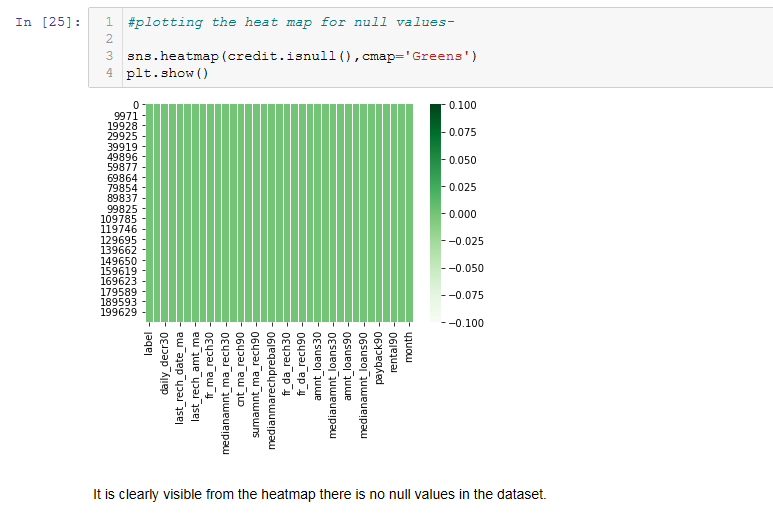
2) Analysing the columns present in the dataset and checking the basic information(if null values are presnt or not).



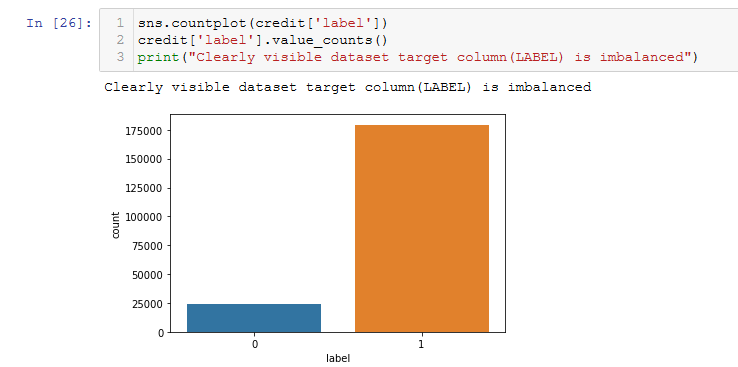
3)Checking the statistical description of the dataset. This report’s helps us to decide which transformation will improve the model’s accuracy,outlier and range of particular column.



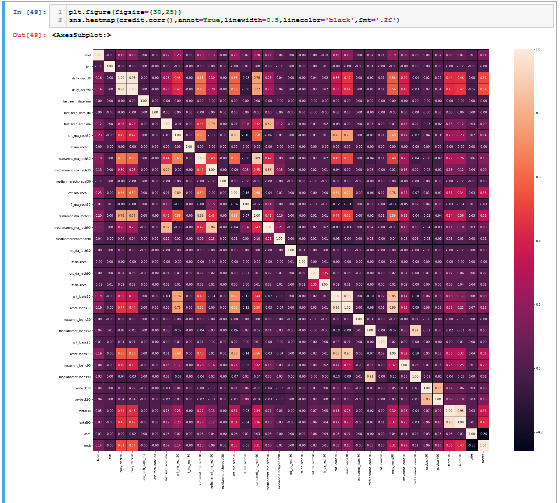
4) Plotting the heatmap for checking if null values are present.



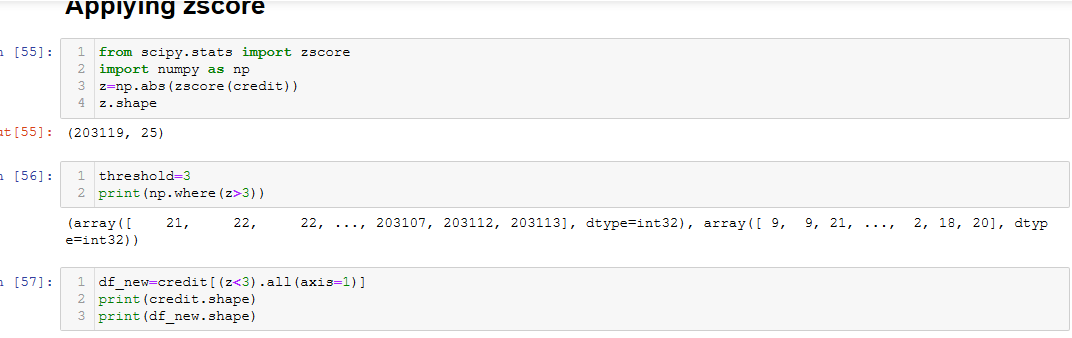
5) Now target distribution i.e “label” column.



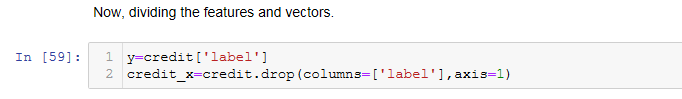
6) Now we are checking the correlation by using heatmap from which we can figure out some columns that we need to drop on the basis of correlation factor.



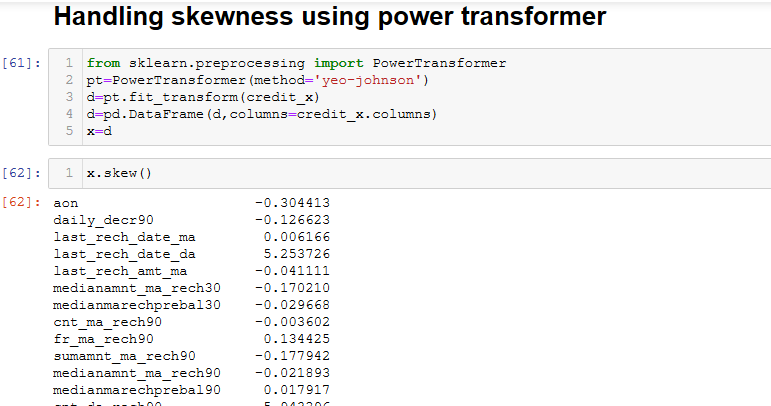
7) Applying zscore(factor-3) for treating outliers.



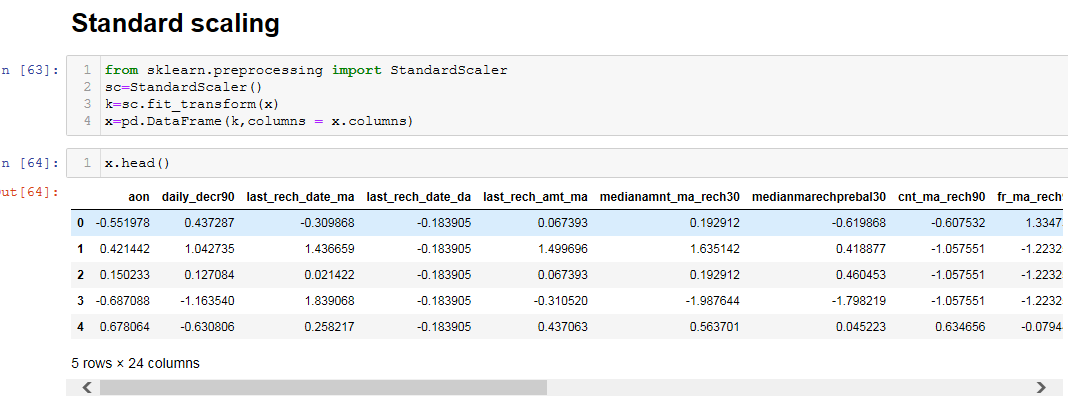
8) Now dividing the feature and target, separating the input and output.



9)Now handling skewness using power transformation to minimise the skewness present in the datatset.

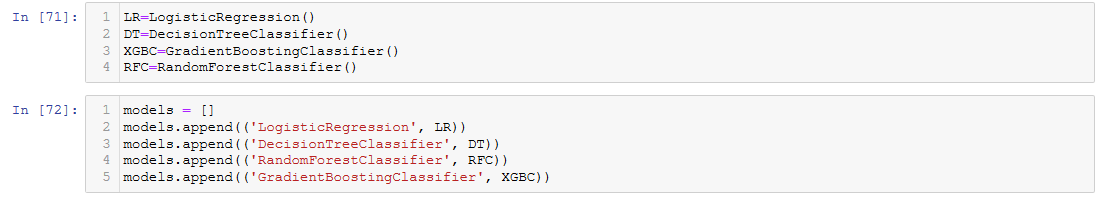


10)Standardizing the input dataset(Standard scaler makes the mean of the distribution 0).

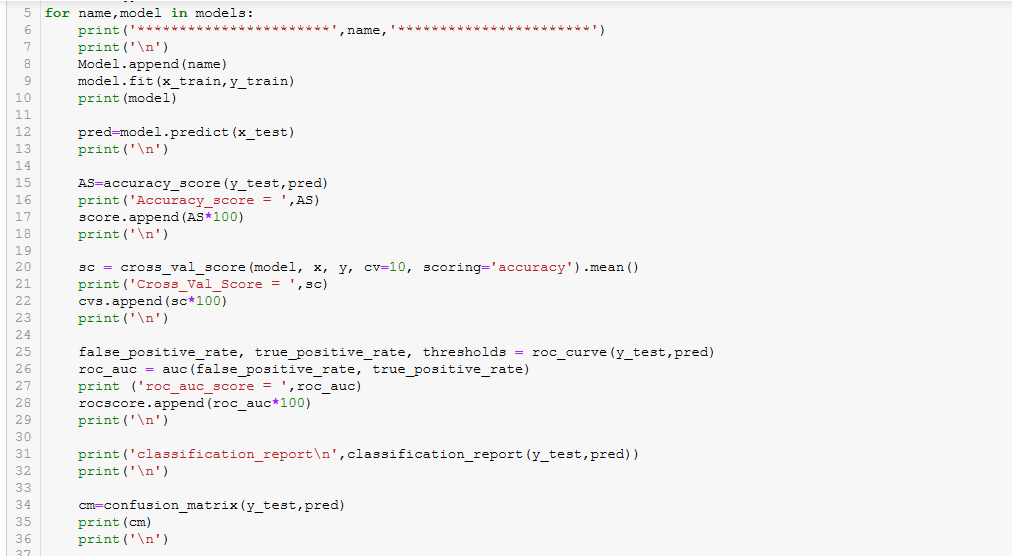


**Model/s Development and Evaluation**

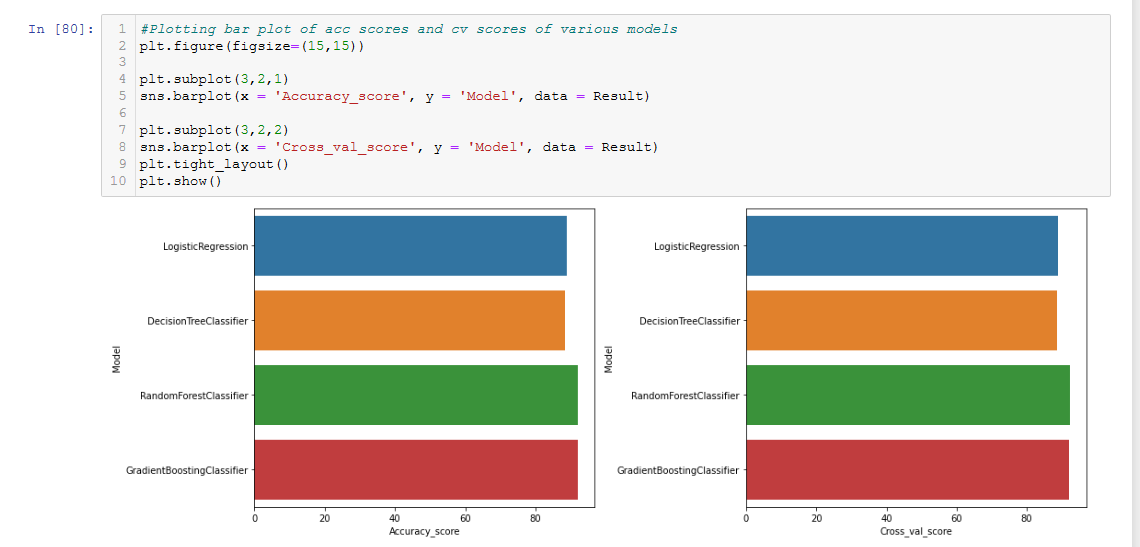
1) Basic models were chosen to check the best performing model.

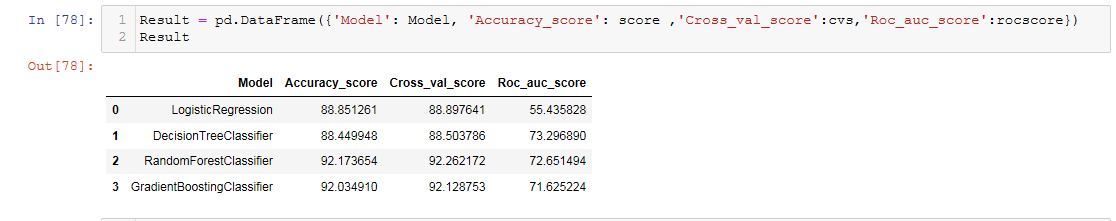


2) Already training & testing data splitted, models were trained and accuracy score, confusion matrix, classification report , roc auc curve and auc score were evaluated.

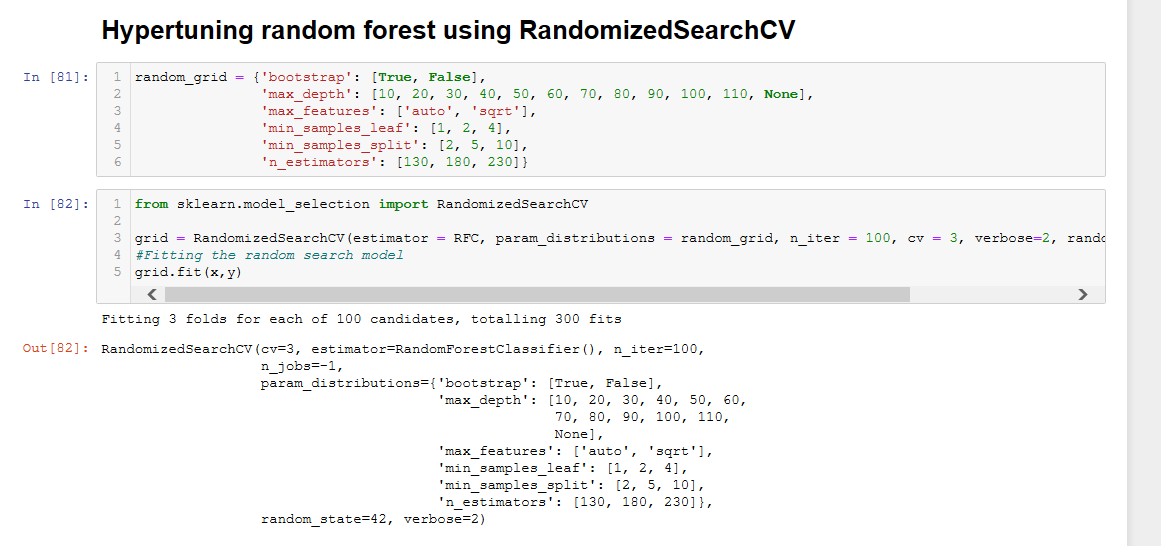


3) Now plotting the bar plot for accuracy score & cross validation score of all the models for better understand and difference can be seen from the dataframe observation.



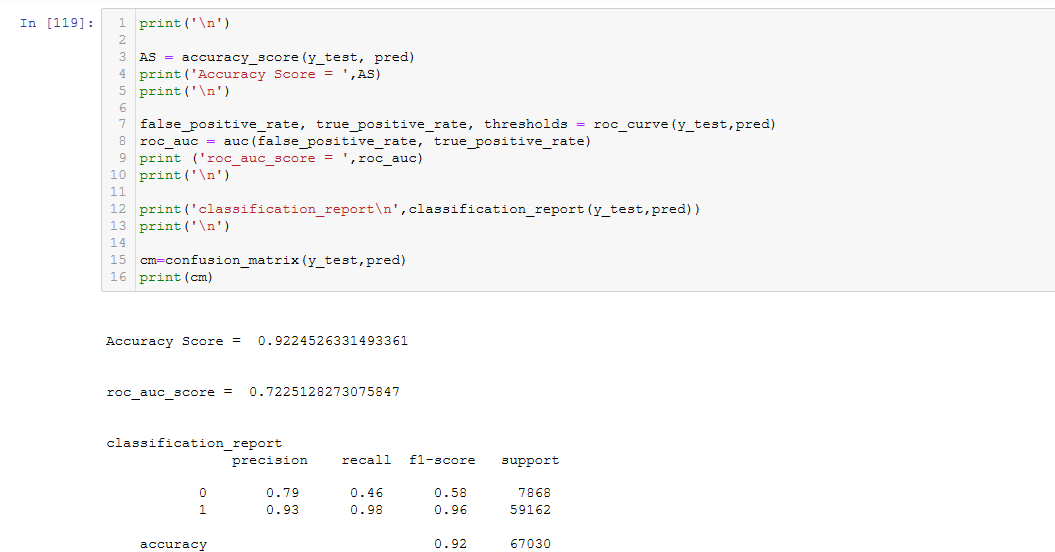


4) Best model I have chosen for hyper tuning i.e random forest classifier as highest accuracy score and cross\_val\_score .We can oberve that the difference between cross\_val\_score 7 accuracy \_score for gradient boosting & random forest is same but auc\_roc\_score is slightly greater in random forest classifier. So, I am choosing this for hyper tuning.

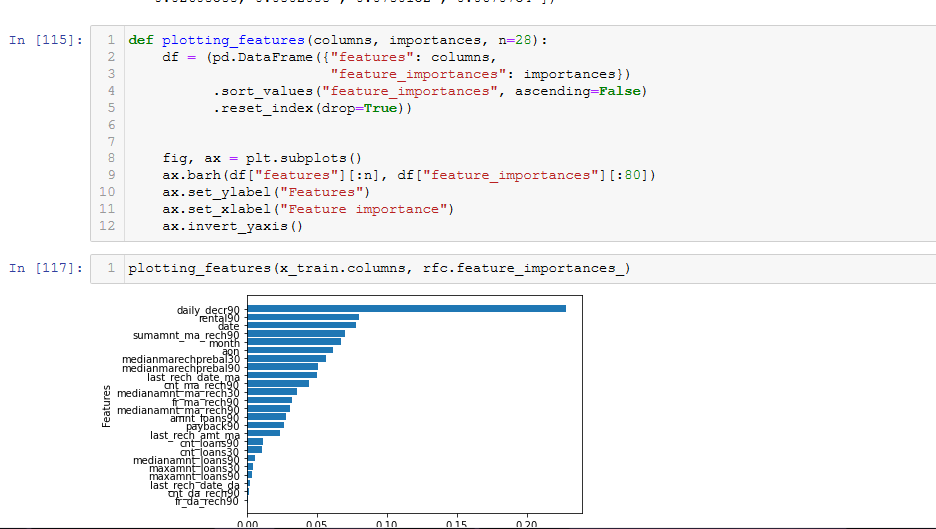


**CONCLUSION**

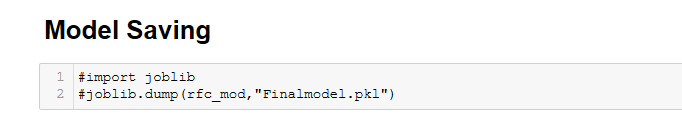
1)Metrics evaluated



2) As, after modelling and evaluation we can see the feature importance of our columns.

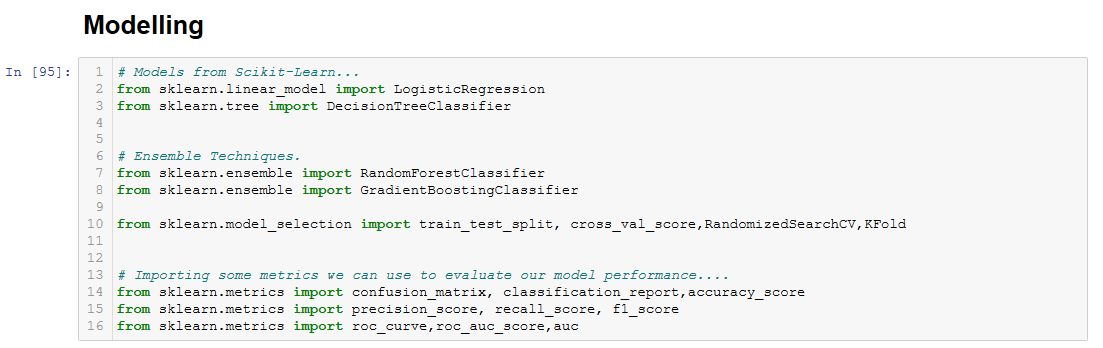


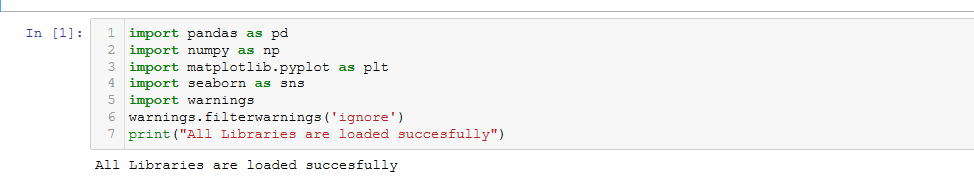
3) Saving the model at last(just commenting it).



4)As random forest classifier performs good over the dataset, hence it can be used for predictions.

Note- Libraries that I used for training & testing the dataset is enlisted below-



****

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

As, I don’t have high end processor laptop, unable to do label sampling because of memory reference value error. This is the major limitation I faced as my chipsets are not compatible for such large datasets. Apart from that, more learning , alayzing & exploration is needed in Data science field that I am continuing.

For more information, please visit:-

https://github.com/ayushpathak0912/Flip-Robo