**Micro Finance credit project**

Submitted By:

Srivani Ravuri

**ACKNOWLEDGMENT**

I have done the project in Jupyter\_Notebook and not used any scikit\_learn for the Model selection.

DataTrained classes are verymuch helpfull to me in completion of the project.

**INTRODUCTION**

* **Business Problem Framing:**

Telecom industry are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah).

* **Conceptual Background of the Domain Problem:**

There is a same concept in India happened long back and think this was very helpfull in an emergency situations as well but there it was like the telecom industry used to deduct the amount directly from their next Recharge. But here the situation seems to be different.

* **Motivation for the Problem Undertaken:**

I have understood that MFI is trying to do some help for the poor and if there is a problem in repaying the loan the trust will be gone and If we prove that max loan has been repayed then MFI will help with the more services to people.Thats my motivation in undertaking the problem.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem:**

I have first checked for the datatypes as our model can not undestand the object values I found phone number is in object and I have used lable encoding to convert it into integer. Then I have splited date into month and date as the complete data is of one year I have not mentioned the year.

* **Data Preprocessing Done:**

checked for the null values and found some outliers have verified with skewness and correlation, removed the skewness and outliers using Z score (np.where(z>3)).

* **Data Inputs- Logic- Output Relationships:**

*The best prediction of Y at any point X=x is the expectation given all information available.*

*The input out relationship tells us how much impact the particular input or column or features showing on the output or target. This can be obeserved from heat map.*

* **State the set of assumptions (if any) related to the problem under consideration:**

I have retained the few columns even correlation is very less as the data loss is becoming more.

* **Hardware and Software Requirements and Tools Used:**

I have used numpy and pandas libraries, matplotlib and seaborn for the visualization, scikit learn for model selection and parameters from it, python for basic functions and also to select the best random state.

**Model/s Development and Evaluation**

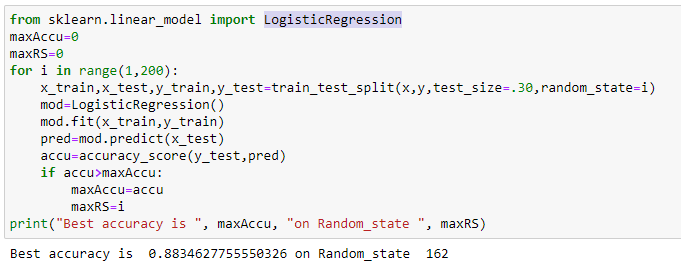
* **Identification of possible problem-solving approaches (methods):**

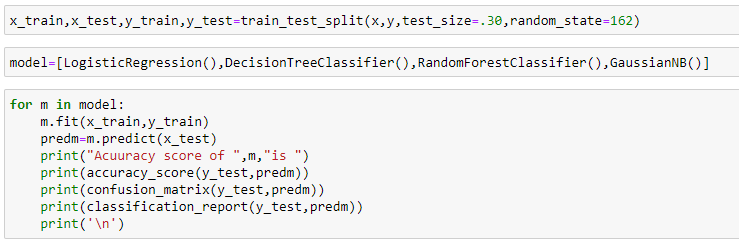
Checked for outlires, skewness to be in the range or not, z score value, correlation, heat map, power transform, label encoder by using all these methoda the data has been cleaned.

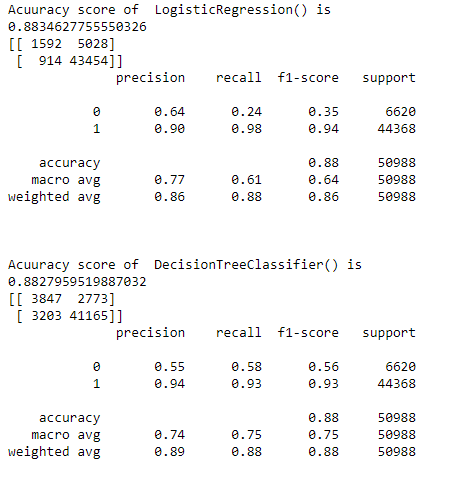
* **Testing of Identified Approaches (Algorithms):**

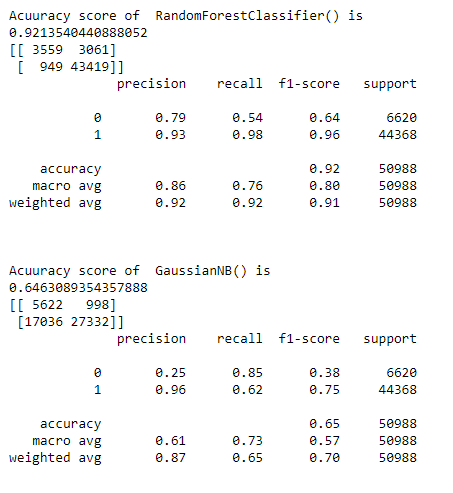
LogisticRegression, DecisionTreeClassifier, GaussianNB, RandomForestClassifier.

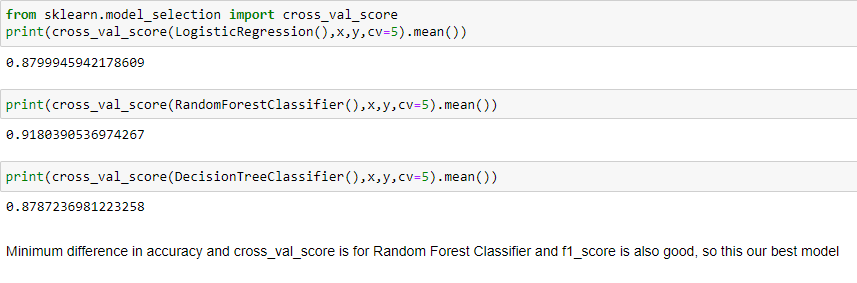
* **Run and Evaluate selected models:**











After this I have done hyperparameter tuning for the RandomForestClassifier using GridSearchCV.

* **Visualizations:**

I have used only heatmap for the visualization to undestand the correlation between the input and output.

* **Interpretation of the Results:**

I undestood the only few of the features are not highly correlated.

I have reduced the skewness and removed the outliers using z\_score in preprocessing.

RandomForestClassifier is the best model suitable it has given 91.80 accuracy and after the hyperparameter tuning it has been increased to 92.26.

**CONCLUSION**

* **Key Findings and Conclusions of the Study:**

I found that spilting of input and output should be done before we apply methods on so that our output will not get disturbed. We should also check if the model is overfitted or not as it should not be overfitted.

* **Learning Outcomes of the Study in respect of Data Science:**

I have used Cross\_val\_score in order to check my best model and also f1 score has been considered here I have tested on four different models and due to huge data it ha taken so much time for me.

* **Limitations of this work and Scope for Future Work:**

I thought all the features is not required in this data few of them can be neglected from the data source itself so that we don’t feel we are dealing with large data. Scope for the future work is good according to me.