

MICRO CREDIT PROJECT

Submitted by:

Shreya Jain

**ACKNOWLEDGMENT**

It had been a great learning experience to work in this Project.

Any achievement is incomplete without the guidance of our mentors. I would sincerely thanks to **Swati Rustagi and Data Trained Institute** who had been the excellent guide throughout my training period.

I thank them for their patience and their ability to help me see how different things in data science are actually related to each other. They taught me and supported me like a true mentor who not only helped me to understand each concept of each topic but who also provide me the resources necessary to complete this work.

Last but not the least I am very thankful to my parents who believed in me and encouraged me.

I sincerely acknowledge Flip Robo Technologies for providing me with such platform where I could get an exposure to the world of Data Science and enhance my capabilities.

# INTRODUCTION

## Business Problem Framing

The problem is to build a model which can be used to predict each loan transaction in terms of probability and whether the customer will be paying back the loaned amount within 5 days of insurance of loan or not.

## Conceptual Background of the Domain Problem

For better understanding of the project, I have learned and understood all the concepts of Statistics, Python, EDA, Data Cleaning, Machine Learning, Algorithms of Machine Learning, Accuracy/Metrics, Cross- Validation, Hyper Parameter Tuning and AUC ROC Curve.

## Review of Literature

In this problem, there is an MFI (Micro Finance Institution) which provides financial services to low income population. It becomes very useful when targeting especially the unbanked poor families. The services provided by MFI are Group Loans, Agricultural Loans, and Individual Business Loans and so on. It is representing $70 Billion in outstanding loans and a global statement of 200 million clients. One of their client is Telecom Industry. Telecom Industry is collaborating with MFI and provide micro credit on mobile balance to consumers that need to pay back within 5 days. If consumer does not payback within 5 days then he/she is considered as a Defaulter which is denoted by 0 in data and non – defaulters are denoted as 1.

## Motivation for the Problem Undertaken

I have completed this Micro Credit Project in partial fulfilment of requirement for the award of certificate of 6 month Internship from Flip Robo Technologies.

# Analytical Problem Framing

## Data Sources and their formats

The data source is in the form CSV (Comma Separated file) which we have downloaded from the portal of flipnwork.com. As the column names are present in abbreviation, so, I have referred the Data\_Description.csv for more understanding about the columns.

## Data Pre-processing Done

Following steps are used for data cleaning:-

1. Checked the Summary Statistics of data
2. Explored the target variable
3. Checked the missing values in data
4. Completed the Univariate, Bivariate and Multivariate Analysis
5. Find the correlation between variables
6. Plotted Outliers
7. Used Label Encoders
8. Dropped three columns
9. Removed Outliers using Z-Score 10.Plotted Skewness

11. Removed Skewness using log and cube root methods

## Data Inputs- Logic- Output Relationships

There are 36 input columns and one output column are present in dataset. Either input columns are of integer or float datatype and only two input columns are in object datatype. Data type of output column is integer. Output variable (label) is highly impacted by two input columns,

i.e. cnt\_ma\_rech30 and fr\_ma\_rech90.

## Software Requirements and Tools Used

I have used Python Jupyter Notebook for all the steps required in building the model. Data was present in .csv file, so, I have used Microsoft excel also. I have used many libraries of python, like, numpy, pandas, matplotlib, seaborn, scipy, imblearn and sklearn e.t.c.

# Model/s Development and Evaluation

1. Firstly, I have separated target and training data by forming x and y variable. 33 columns are used as training variable ‘x’ and 1 column is set as target variable ‘y’ and then I have solved the class imbalance problem for target variable.
2. Secondly, I have found the best random state and it turns out to be 108.
3. Then, I have used Standard Scaler to scale all the columns in same range.
4. Fourthly, I have used algorithms, like, Logistic Regression, Decision Tree, Naïve Bayes and Support Vector Machines to calculate accuracy scores.
5. Now, I have used some ensemble techniques, like, Random Forest Classifier and Decision Tree Classifier to calculate accuracy scores.
6. Then, I have checked cross validation score of the best 2 algorithms.
7. Finally, I have calculated the difference between accuracy score and cross validation score of both models and the algorithm with least difference turned out to be the best model.
8. Now, I have tried to improve the accuracy score of best algorithm by hyper - parameter tuning.
9. I have plotted the auc-roc curve for the best algorithm with best parameters from the last step.
10. Finally, I have saved the best algorithm using joblib.

# CONCLUSION

1. Using fourth and fifth steps of model development, Random Forest Classifier has become the best algorithm with accuracy score of 93.930%.
2. Using seventh step of model development, the difference between accuracy score and cross validation score of Random Forest found to be least, i.e, 0.221.
3. Using hyper parameter tuning, n\_estimators=100, max\_depth=None, max\_features=sqrt and criterion=gini come out to be the best parameters for Random Forest classifier and has changed the accuracy score from 93.930% to 93.923%