**Business Problem Framing:** The problem depicts the case study of loans being offered by Micro Finance Institution for people with low-income ranges. This particular problem statement speaks of a case related to Telecom Industry who are a fixed wireless telecommunications network provider. This industry understands the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low-income families and poor customers. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days.

**Conceptual Background of the Domain Problem:** Based on the above problem statement, we have to build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label ‘1’ indicates that the loan has been payed. i.e., Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e., defaulter.

**Review of Literature:** With the idea of supporting people with low-income range, MFI experts and donors are supporting the idea of using mobile financial services which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. For this purpose, the client which is a Telecom Industry who are a fixed wireless telecommunications network provider has collaborated with MFI for providing loans to their customer. We have to build a model to predict loan default where Label ‘1’ indicates that the loan has been payed. i.e., Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e., defaulter.

**Motivation for the Problem Undertaken:** With the idea of improving its business and developing importance of communication, the Telecom client has developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber. So in order to support them, the client has collaborated with MFIs to provide the loans

**Analytical Problem Framing:**

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

We have the concept of Central Tendency like Mean to find the average of the dataset, Median to find the middle value of this ordered dataset, Mode to find the most frequent value in the dataset, Standard Deviation which is the difference between each data point and the mean and the square root of variance and Skewness which is a measure of symmetry.

Under variability, we have used Zscore method to remove outliers

We have used Correlation to measure the relationship between two variables and ranges from -1 to 1, the normalized version of covariance.

We have used count plot to visualize the count of the target variable (label in this case)

We have used dist plot to visualize the distribution of the curves

We have used box plot to identify the outliers

**Data Sources and their formats:**

* label
* msisdn
* aon
* daily\_decr30
* daily\_decr90
* rental30
* rental90
* last\_rech\_date\_ma
* last\_rech\_date\_da
* last\_rech\_amt\_ma
* cnt\_ma\_rech30
* fr\_ma\_rech30
* sumamnt\_ma\_rech30
* medianamnt\_ma\_rech30
* medianmarechprebal30
* cnt\_ma\_rech90
* fr\_ma\_rech90
* sumamnt\_ma\_rech90
* medianamnt\_ma\_rech90
* medianmarechprebal90
* cnt\_da\_rech30
* fr\_da\_rech30
* cnt\_da\_rech90
* fr\_da\_rech90
* cnt\_loans30
* amnt\_loans30
* maxamnt\_loans30
* medianamnt\_loans30
* cnt\_loans90
* amnt\_loans90
* maxamnt\_loans90
* medianamnt\_loans90
* payback30
* payback90
* pcircle
* pdate

Above mentioned are the data sources or different variables in the dataset. Based on these inputs passed, the predictions are made

**Data Preprocessing Done:** Since dataset was already provided, data preparation part was done. Based on dataset provided, we could load the data to read the dataset, find out the number of rows and columns using shape, identify the type of columns, find out count, min, max, median, SD and percentiles. We could also identify the info of the data, datatypes and find out if there are any null/missing values in the dataset

**Data Inputs- Logic- Output Relationships:** Detailed execution is in the notebook (Eg: Loading data, shape, describe, columns, dtypes, EDA, visualization, encoding, outliers, train test split, algorithms, CV and hyper parameter tuning

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

S/W: Python Jupyter Notebook in .ipynb extension format

Packages: Numpy for numerical analysis, pandas for data analysis and data preprocessing, matplotlib and seaborn for visualization

H/W: 8 GB Ram, Windows OS

**Model/s Development and Evaluation**

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

**Testing of Identified Approaches (Algorithms):** Since there was a target variable and this was classified as 1 and 0 in the given dataset, we used Classification Algorithm. For the purpose of modeling, we have used Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier and depending on the highest accuracy, we have chosen one particular algorithm for deploying the model

**Run and evaluate selected models**



**Key Metrics for success in solving problem under consideration:** Applied metrics like precision\_score, recall\_score, f1\_score, roc\_auc\_score, accuracy\_score, classification\_report,confusion\_matrix etc.

**Visualizations**

We have used histogram for visualization, countplot for plotting the target variable and showing class imbalance, distplot for identifying the distribution in the curves and boxplot for identifying outlier. Respective visualizations for all these are displayed on Jupyter Notebook

**CONCLUSION**

**Key Findings and Conclusions of the Study**

We have applied Classification technique and have finalized RFC based on highest accuracy. However later part of the code is not getting executed in the system

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

There was a class imbalance in the target variable when count plot was applied

The column medianamnt\_ma\_rech90 had outliers and outliers were later removed using Zscore

As part of EDA while performing Bivariate Analysis between daily\_decr30 and daily\_decr90, we could find that 0

Also, as part of data cleaning, columns Pdate and Pcircle were dropped

We have tested using DTC, RFC and SVC algorithms (PS: The code was not getting executed in my system)

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

N/A