

**MICRO-CREDIT DEFAULTER MODEL**

**Submitted by:**

**SUSHREE SMITA NAYAK**

**ACKNOWLEDGMENT**

I would like to thank Fliprobo and my mentor Ms Swati Mahaseth who gave me a golden opportunity to work on this project. I’d also like to express my gratitude to the whole team of Datatrained wholeheartedly.

I must also thank my parents and friends for the immense support and help during this project. Without their help, completing this project would have been very difficult.

**INTRODUCTION**

* **Business Problem Framing**

In this micro finance institution, the problem we are facing is if customer who has availed loan from the institution will repay the loan amount within 5 days of insurance or not. This is a very important issue the world is facing. From this one problem, we can understand and mitigate other problems like retuning capacity, unemployment, economic parity, revenue generation etc. this also generates scope of new businesses and chance to grow old businesses. This might also reflect on the causes of some business failures and might lead to be a learning lesson and increase efficiency for new businesses to set up.

* **Conceptual Background of the Domain Problem**

This belongs to the banking and finance industry. A Microfinance Institution (MFI) is an organization that offers financial services to low-income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on. The MFI in collaboration with a client(business) from telecom industry are providing micro-credit on mobile balances to their end customers where the amount credited is 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.

* **Review of Literature**

This is a comprehensive summary of the research done on the micro credit defaulter. The MFI is providing micro loans especially targeting low-income groups. They collaborated with a telecom client and used their customer base to provide micro loans to mobile number users and increase their reach. Summarizing the problem statement, it is about predicting if the consumers of the telecom company who have availed micro loans from the MFI will return the amount within 5 days of cycle. The prediction will provide a better selection method of customers for giving credit, which could help them in further investment and improvement in selection of customers.

* **Motivation for the Problem Undertaken**

Objective behind this project is to be providing a model to predict if the customer will pay within 5 days. This will be the basis of the selection procedure. After deployment, before giving credit to the person, we can run it on the website to analyse if the person will repay within the limited timeframe or not. My motivation for this is my esteemed mentor Ms Swati Mahaseth, my datatrained family and mentor, my family and friends.

**ANALYTICAL PROBLEM FRAMING**

* **Mathematical/ Analytical Modeling of the Problem**
  + - Shape of the dataset: 37 columns and 209593 rows
    - 3 Object type columns, 13 integer type, 21 float type column.
    - Basic stats of the dataset indicates that the standard deviation will have a larger spread by forming peaks when plotted for most of the cases.
    - The box plot indicates high presence of outliers.
    - The histogram indicates most of the data is right sided skewed.
* **Data Sources and their formats**

The source of the data is client database. The data is about the consumers who have availed the micro loan and are telecom customers of our client. The format is it provides an immediate monetary aid to people in need.

* **Data Preprocessing Done**
  + - For data cleaning, first we checked if there are any null values. 0 missing values.
    - Then we checked for duplicates. There were no duplicates.
    - Then we analysed the data for those columns which are unnecessary and dropped it. Such as ‘unnamed: 0’ which is index number, msisdn which is mobile number similar to index number.
    - For the date column we separated the day and moth to get a better understanding. And dropped the pdate column.
    - Then we checked for multicollinearity and observed that 'amnt\_loans30','daily\_decr90','cnt\_ma\_rech90','maxamnt\_loans90','amnt\_loans90','rental30' has high collinearity and less contribution. Thus we dropped them to reduce covariance.
    - As the EDA already mentioned presence of high skewness and outliers, we reduced it by accepting those with z score value less than 3.
    - Then as the no. of columns was high we performed PCA and reduced the train part of dataset into 10 columns.
    - Data Inputs- Logic- Output Relationships
    - The input data greatly affects the cycle of input and output. Such as the payment cycle as well as the recharge cycle. It shows the pattern about how needy they are of the micro loan. It also indicates how much amount is generally borrowed.
* **Hardware and Software Requirements and Tools Used**

**Listing down the hardware and software requirements along with the tools, libraries and packages used.**

**Tools used:**

* Python
* Jupyter notebook

**Libraries and packages used:**

* pandas
* numpy
* matplotlib
* seaborn
* klib library
* from statsmodels.stats.outliers\_influence we used variance\_inflation\_factor
* from scipy.stats import skew and zscore
* from sklearn library we used train\_test\_split, StandardScaler, cross\_val\_score, RandomForestClassifier,DecisionTreeClassifier,KNeighborsClassifier, LogisticRegression, metrics, GridSearchCV, RepeatedStratifiedKFold, roc\_curve, roc\_auc\_score, confusion\_matrix
* joblib to save the model

**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**
  + - We observed that this is a classification problem. The target column describes 1 as the loan is paid within required timeframe. 0 indicates that is not paid.
    - We cleaned the dataset to increase the accuracy of our predicted results.
    - After separating the dataset into train data and test data, we performed standardized scaling.
    - We also performed PCA and reduced column size to 10.
    - We trained and fitted the dataset into 4 classification models which are:
      1. Decision tree model
      2. Logistic regression model
      3. Random forest model
      4. KNN neighbors model
* **Testing of Identified Approaches (Algorithms)**
  + Decision tree model
  + Logistic regression model
  + Random forest model
  + KNN neighbors model
* **Run and Evaluate selected models**
* **Decision tree model:** Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Text, letter

Description automatically generated

The accuracy score is 84.5%. This means our model is working good for 84.5%.

* **Logistic regression model:** Logistic regression model takes a linear equation as input and use logistic function and log odds to perform a binary classification task.

Text, letter

Description automatically generated

The accuracy score is 86.4%. This means our model is working good for 86.4%.

* **Random forest model**: Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

Text, letter

Description automatically generated

The accuracy score is 89.5%. This means our model is working good for 89.5%.

* **K-Nearest neighbors model:** K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.

Text, letter

Description automatically generated

The accuracy score is 86.5%. This means our model is working good for 86.5%.

* **Key Metrics for success in solving problem under consideration**
  + 1. **Accuracy score:** It is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.
    2. **Cross val score**: Cross-validation is a statistical method used to estimate the skill of machine learning models. The difference between cross val score and accuracy score of different models will tell which model is working best.
    3. **Roc auc curve:** The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.
    4. **Confusion matrix:** Confusion matrices are used to visualize important predictive analytics like recall, specificity, accuracy, and precision. Confusion matrices are useful because they give direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives.
* **Visualizations**

Chart, bar chart

Description automatically generated

Most of the consumers pay in time.

A picture containing calendar

Description automatically generated

A picture containing calendar

Description automatically generated

A picture containing calendar

Description automatically generated

Chart

Description automatically generated

Above distplots shows that the data is not normally distributed and has high standard deviation.

This also indicates that the data right sided skewed.

Timeline

Description automatically generated with medium confidence

Diagram

Description automatically generated

Diagram

Description automatically generated with medium confidence

A picture containing timeline

Description automatically generated

This is the klib display of each columns stats and distribution graph. It gives mean, std deviation, skew count, kurtosis, and total count.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated Graphical user interface, chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated Chart

Description automatically generated

All the columns have outliers.

* **Interpretation of the Results**
  1. Few columns were sort of index columns and were unnecessary.
  2. High presence of outliers and skewness
  3. Data not normally distributed and almost right sided skewed.
  4. Presence of high variance and multi collinearity
  5. Model is working with high accuracy.

**CONCLUSION**

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

Random forest fits best and is hyper tuned. The accuracy that our model is working well is approx. 90%. Thus, the model can be deployed. Key findings include that people of LIGs are quite particular in repayment of their micro- loans. This particular facility is quite handy to them. But based on their usage, recharge amount and cycle, and behavioural traits in terms of spending it can be predicted beforehand that if the person will pay in time or not and keep his end of bargain. the client can provide loans as per the prediction. This will reduce delinquency, bad debts, and help the client to create a strong customer base.

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

Limitations include presence large dataset. Acquiring knowledge and utilising it proficiently. Limitations of Timeline.