

MICRO CREDIT DEFAULTER PROJECT – PREDICTING WHETHER A CUSTOMER WILL BE ABLE TO PAY BACK THE LOAN ON TIME OR NOT

Submitted by:

NATASHA PODDAR

. **INTRODUCTION**

* Business Problem Framing

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.

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients..

They understand 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 that can help them in the need of hour.

They 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.

In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

* Conceptual Background of the Domain Problem

Basic knowledge relating to Micro Finance Institution and how they function is needed. A basic understanding of the various terms in their day to day working is essential

* Motivation for the Problem Undertaken

This issue is very realistic and common in today’s world and one should know to deal with such situations in the future

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Firstly missing values were checked and

Correlation with all independent variables and wrt target were checked

Outliers were identified through zscore but they were not removed as the data had more than 20% outliers

Skewness was checked and tools were applied to control them and scale the data

Multi colinearity was checked and worked upon

Models were applied to train and test the model

* Data Sources and their formats

Detailed data was procured from Micro Credit Defaulter Case Study,

* Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

1. Duplicate values check
2. Unique & Count of all columns were checked
3. Missing values were imputed
4. Columns which had more than 40% data missing were removed
5. Catagorical data was Encoded
6. Skewness removal through Power Transform and scaling of the data
7. VIF Check -for multicollinearity
8. Correlation check
9. Graphical Univariate, Bivariate & Multivariate Analysis
10. Outliers check -ZSCORE

* Data Inputs- Logic- Output Relationships

Mostly all the columns were low – moderately correlated with the target column, both positive and negative in nature

* Hardware and Software Requirements and Tools Used

1. Pandas – For Data Reading and understanding
2. Label Encoder –(SK LEARN) – For Encoding the categorical data into numerical ones
3. Zscore(SCIPY)-For checking & removal of outliers
4. Power Transform ()- Skewness removal
5. Duplicate- To check for duplicate Values
6. CORR-To check Correlation
7. VIF -To check for multicollinearity
8. Numpy- For mathematical operations
9. LOGITSIC REGRESSION (SKLEARN) – Training & Testing the model
10. KNN CLASSIFIER (SKLEARN) – Training & Testing the model
11. DECISION TREE CLASSIFIER (SKLEARN) – Training & Testing the model
12. GAUSSIAN NB (SKLEARN) – Training & Testing the model
13. CROSS VAL SCORE – Regularizing the model
14. GRID SEARCH CV- Hyper Tuning the Model for higher accuracy
15. SEABORN- VISUALIZATION LIBRARY – HISTPLOTS, DISTPLOTS, SCATTERPLOTS, COUNTPLOTS, BOXPLOTS and other graphs
16. MATPLOTLIB.PY PLOT -Visualization tool

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Firstly missing values were checked .

Correlation with all independent variables and wrt target were checked

Outliers were identified but they were NOT removed through zscore as the data had more than 20% outliers present

Skewness was checked and tools were applied to control them and scale the data

Multi colinearity was checked and worked upon

Models were applied to train and test the model

* Testing of Identified Approaches (Algorithms)

1. LOGIATIC REGRESSION
2. KNN CLASSIFIER
3. DECISION TREE CLASSIFIER
4. GAUSSIAN NB CLASSIFIER

* Key Metrics for success in solving problem under consideration

1. ACCURACY SCORE
2. CONFUSION MATRIX
3. CLASSIFICATION REPORT
4. F1 SCORE
5. PRECISION
6. RECALL SCORE
7. AUC-ROC SCORE

* Visualizations

Seaborn Library was used along with matplotlib Library for visualizations

Histplots, bar plots, count plots, swarmplots, boxplots etc were made and analysed

* Interpretation of the Results

All the models predicted an accuracy in the range of 65-89 where as KNN CLASSIFIER had the highest accuracy and DTC had the least difference between CV MEAN SCORE & MODEL ACCURACY SCORE hence we had hyper tuned the said model and saved the same

**CONCLUSION**

* Key Findings and Conclusions of the Study

KNN CLASSIFIER had the highest accuracy and DTC had the least difference between CV MEAN SCORE & MODEL ACCURACY SCORE hence we had hyper tuned the said model and saved the same

* Learning Outcomes of the Study in respect of Data Science

KNN CLASSIFIER had the highest accuracy and DTC had the least difference between CV MEAN SCORE % MODEL ACCURACY SCORE hence we had hyper tuned the said model and saved the same

. With unique feature we realized the type of data all the columns had, The various visualization tools helped in understanding the different relationships between the variables,. VIF was a very powerful tool to detect multicollinearity and Cross Val score helped in regularizing the model