NAME OF THE PROJECT

MICRO CREDIT

Submitted by:

ABHISHEK SINGH

BATCH - 1829

EMAIL - - - singh.abhishek030@gmail.com

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

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.

Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) 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. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

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

We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have 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.

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. 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).

The sample data is provided to us from our client database. It is hereby given to you for this exercise. 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.

  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.

- ## Motivation for the Problem Undertaken

  The objective of this project to provide the information that, whether the customer is goin to pay the loan amount with interest within 5 days or not. To predict the cases or Non-defaulter and defaulter.

  The motivation behind this project is focusing on providing their services and products to low income families and poor customers that can help them in the need of hour.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  We have use the following modelling done during

  this project :

  distplot, bar, scatterplot , Heatmap,pairplot,

  Heatmap (using the values) to analyze the data

  correlation bewteen target nd independent

  variables , Boxplot to find the outliers and found

  outliers in maximum columns

  The Statistical approach like mean , median, 75percentile,

max value and standard deviation.

1> Mean is greater than median in Almost all the columns

2> Standard deviations is also high in all the columns so, we have to check skewness by graph and remove the skewness

using methods and try to perform after removing outliers.

3>The difference between 75% and the max value is very high (Abnormal)

4> After considering all the 1,2 and 3 cases , We found that heavy outliers and skewness are present in the dataset.

In Heatmap , We observed : (Target-label column)

The lighter color : Highly correlated

The Darker color : Negative correlation

label column is having good correlated with daily_decr30,daily_decr90,cnt_ma_rech30,cnt_loans30,amnt_loans30 .

Label columns is having almost good correlation with every column. No negative correlation is there.

- Data Sources and their formats

The dataset "micro credit" which is a csv file has been taken

from flib robo company and the origins located in C drive

folder . The location is :

('C:/Users/vishu/Downloads/Micro-Credit-Project/Micro Credit Project/Data file.csv')

It has 209593 rows and 37 columns in which 5 useless

rows have been removed.

The label column which is a target variable is a integer

datatype namely 0 and 1 .

cnt_ma_rech90 int64

fr_ma_rech90 int64

sumamnt_ma_rech90 int64

cnt_da_rech90 int64

fr_da_rech90 int64

cnt_loans30 int64

amnt_loans30 int64

amnt_loans90 and maxamt_loans90 are integer type.

Apart from these columns , all the columns have float data

type. pcircle and pdate are of object type.

- Data Preprocessing Done

  Steps for Cleaning the data :

  1> There is no-null values in dataset.

  2> The few outliers have been removed (generally 98percentille data) using the IQR method.

  After removal of the data using IQR , out of 209593 rows ,186395 rows is present. We have lost and cleaned almost 9-10 percent of the data.

  3> The skewness have been removed by Power Transform method.

- Data Inputs- Logic- Output Relationships

The inputs are all the columns except the label column.

The output Label (0/1) is positively correlated to all the input variables and it has been check using correlation method.

No negative correlation is there.

- Hardware and Software Requirements and Tools Used

Important Libraries used :

```
import pandas as pd

import numpy as np

import random


import sklearn

import scipy


from sklearn.linear_model import LogisticRegression

from sklearn.metrics import
accuracy_score,confusion_matrix,classification_report

from sklearn.model_selection import KFold

from sklearn.preprocessing import StandardScaler


from sklearn.model_selection import GridSearchCV

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score

from sklearn.ensemble import IsolationForest
```

```
from sklearn.neighbors import LocalOutlierFactor

from sklearn.svm import OneClassSVM

from pylab import rcParams

rcParams['figure.figsize'] = 14, 8

RANDOM_SEED = 42

LABELS = ["success", "failure"]


import matplotlib.pyplot as plt

import seaborn as sns

import warnings warnings.filter('ignore')
```

The Tools used are

scikit-learn, Numpy ,Matplotlib, Seaborn and Pandas.

From scikit learn , we used almost 5 models to build and run the dataset. Like examples:

Logistic regression ,Random forest , Decison Forest,KNN clasisifier and SVC .

1>from sklearn.metrics import accuracy_score,confusion_matrix,classification_report = for calculation of accuracy score , report and confusion matrix.

2>from sklearn.model_selection import GridSearchCV = for running the hypertuning parameter.

3>from sklearn.preprocessing import StandardScaler = for Scaling of data upto a similar range

4>from imblearn.over_sampling import SMOTE = for balancing the input and target variable.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Steps to check correlation :

  1> Using .corr() function , we have check correlation.

  2> Using Seaborn (sns) , have created heatmaps to check correlation.

  Steps for Univariate analysis :

  3> Using distplots, bar plot ,countplot

  Steps for Bivariate analysis :

  4> Using scatter plot,pairplot

  Steps for Skewness and Correction :

  5> I observed using distplot , kdeplot and .skew() method that daily_decr30 ,daily_decr90,last_rech_date_ma , last_rech_date_da,fr_da_rech90 ,rental_30,rental_90,cnt_loans90,cnt_ma_rech30,last_rech_amt_ma have skewed data and we have applied power transform method to remove the skewness.

Steps for Outliers and Correction :

6> I observed using Boxplot , Z-Score method that daily_decr30 ,daily_decr90,last_rech_date_ma , last_rech_date_da,fr_da_rech90 ,rental_30,rental_90,cnt_loans90,cnt_ma_rech30,last_rech_amt_ma and almost every columns have Outliers and we have applied IQR(Interquartile range) method to remove the Outliers.

I have apllied Z-score also before , but using z-score , i was losing 22 percent of data.So, I removed using IQR.

7> Before removing skewness , I have divided target and independent variable into x and y and apply power transform method and again the correct data saved in x.

Steps for Handling Imbalance problem:

8>Now, I have balanced the data x(independent var.) and y(target var.) using over sampling SMOTE method.

Steps for Scaling:

9>After balancing , I have used Standard Scaler using StandardScaler() method so that all the columns is going to be of similar range.

10>Now, Using train_test_split method, I have split data into x and y and ready for machine learning.

11>Using 5 models , I have test down the accuracy using each model of the data.

- Testing of Identified Approaches (Algorithms)

Algorithms Used :

1> Logistic Regression : I got 76 prcnt accuracy.

2>Decision Tree Classifier : 64 percent accuracy.

3> Random Forest Classifier : 77 prcnt accuracy.

4> Support Vector Classifier : 76 prcnt accuracy.

5>KNN Classifier : 87 prcnt accuracy.

- Run and Evaluate selected models

  I have run all 5 models and after that by checking cross validation i obtain cross_val_score and after that selected the best model which was KNN and its accuracy was 87prcnt.

  but due to kernel problem , it was taking so much of time and i didnt able to finalize the hyper tuning of the model.

  I have mentioned and write Logistic Reg code there.

- Key Metrics for success in solving problem under consideration

  Key Metrics are Confusion matrix , accuracy_score and classification report

  LR :

  accuracy score --- 0.7572158127610304 -

  Confusion matrix

  [[37528 10880]

  [12605 35719]]

  DTC : accuracy score --- 0.6354877393210107

  Confusion matrix

  [[41899 6509]

  [28751 19573]]

  RFC : accuracy score ---- 0.7440247281147914

[[44851 3557]

[21204 27120]]

SVC :

accuracy - - - 0.7560579746102634

confusion matrix :

[[37285 11123]

[12474 35850]]

KNN :

accuracy score --- --- 0.8670553694744242

confusion matrix :

[[46026 2382]

[10478 37846]]

- Interpretation of the Results

  We achieved from KNN and logistic resp 87 and 76 prcnt accuracy and model is performing good.

CONCLUSION

- Key Findings and Conclusions of the Study

Model is running good with 87 prcnt accuracy so we can choose KNN model.