Micro Credit Defaulter

Submitted by:

Mitesh Verma

ACKNOWLEDGMENT

I would like to acknowledge Mr. Shwetank Mishra (FlipRobo) for his vital cooperation and help in ensuring the successful completion of my assignment. He deserves the utmost credit for the assignment's outcome.

Finally, I would want to convey my sincere thanks Datatrained Academy and their guidance without them, the task would not have been accomplished.

The website that I referred are:

https://learning.datatrained.com

https://www.w3schools.com

https://www.freecodecamp.org

https://github.com

https://www.geeksforgeeks.org

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).

- Conceptual Background of the Domain Problem

  This project is all about predicting the Micro Credit Defaulter.

- Review of Literature

Features

1. msisdn: mobile number of user
2. aon: age on cellular network in days
3. daily_decr30: Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)
4. daily_decr90: Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)
5. rental30: Average main account balance over last 30 days
6. rental90: Average main account balance over last 90 days
7. last_rech_date_ma: Number of days till last recharge of main account
8. last_rech_date_da: Number of days till last recharge of data account
9. last_rech_amt_ma: Amount of last recharge of main account (in Indonesian Rupiah)
10. cnt_ma_rech30: Number of times main account got recharged in last 30 days
11. fr_ma_rech30: Frequency of main account recharged in last 30 days
12. sumamnt_ma_rech30: Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)
13. medianamnt_ma_rech30: Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)
14. medianmarechprebal30: Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)
15. cnt_ma_rech90: Number of times main account got recharged in last 90 days

16. fr_ma_rech90: Frequency of main account recharged in last 90 days
17. sumamnt_ma_rech90: Total amount of recharge in main account over last 90 days (in Indonasian Rupiah)
18. medianamnt_ma_rech90: Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah)
19. medianmarechprebal90: Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah)
20. cnt_da_rech30: Number of times data account got recharged in last 30 days
21. fr_da_rech30: Frequency of data account recharged in last 30 days
22. cnt_da_rech90: Number of times data account got recharged in last 90 days
23. fr_da_rech90: Frequency of data account recharged in last 90 days
24. cnt_loans30: Number of loans taken by user in last 30 days
25. amnt_loans30: Total amount of loans taken by user in last 30 days
26. maxamnt_loans30: maximum amount of loan taken by the user in last 30 days
27. medianamnt_loans30: Median of amounts of loan taken by the user in last 30 days
28. cnt_loans90: Number of loans taken by user in last 90 days
29. amnt_loans90: Total amount of loans taken by user in last 90 days
30. maxamnt_loans90: maximum amount of loan taken by the user in last 90 days
31. medianamnt_loans90: Median of amounts of loan taken by the user in last 90 days
32. payback30: Average payback time in days over last 30 days
33. payback90: Average payback time in days over last 90 days
34. pcircle: telecom circle
35. pdate: date

Target

label: Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}

Information of the Dataset.

- RangeIndex: 0 to 209592
- Data columns: 37
- dtypes: float64(21), int64(13), object(3)

- Motivation for the Problem Undertaken

This project is on the data science and machine learning model, build the model to predict the Micro Credit Defaulter based on some features.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	209593.0	104797.000000	60504.431823	1.000000	52399.000	104797.000000	157195.00	209593.000000
label	209593.0	0.875177	0.330519	0.000000	1.000	1.000000	1.00	1.000000
aon	209593.0	8112.343445	75696.082531	-48.000000	246.000	527.000000	982.00	999860.755168
daily_decr30	209593.0	5381.402289	9220.623400	-93.012667	42.440	1469.175667	7244.00	265926.000000
daily_decr90	209593.0	6082.515068	10918.812767	-93.012667	42.692	1500.000000	7802.79	320630.000000
rental30	209593.0	2692.581910	4308.586781	-23737.140000	280.420	1083.570000	3356.94	198926.110000
rental90	209593.0	3483.406534	5770.461279	-24720.580000	300.260	1334.000000	4201.79	200148.110000
last_rech_date_ma	209593.0	3755.847800	53905.892230	-29.000000	1.000	3.000000	7.00	998650.377733
last_rech_date_da	209593.0	3712.202921	53374.833430	-29.000000	0.000	0.000000	0.00	999171.809410
last_rech_amt_ma	209593.0	2064.452797	2370.786034	0.000000	770.000	1539.000000	2309.00	55000.000000
cnt_ma_rech30	209593.0	3.978057	4.256090	0.000000	1.000	3.000000	5.00	203.000000
fr_ma_rech30	209593.0	3737.355121	53643.625172	0.000000	0.000	2.000000	6.00	999606.368132
sumamnt_ma_rech30	209593.0	7704.501157	10139.621714	0.000000	1540.000	4628.000000	10010.00	810096.000000
medianamnt_ma_rech30	209593.0	1812.817952	2070.864620	0.000000	770.000	1539.000000	1924.00	55000.000000
medianmarechprebal30	209593.0	3851.927942	54006.374433	-200.000000	11.000	33.900000	83.00	999479.419319
cnt_ma_rech90	209593.0	6.315430	7.193470	0.000000	2.000	4.000000	8.00	336.000000
fr_ma_rech90	209593.0	7.716780	12.590251	0.000000	0.000	2.000000	8.00	88.000000
sumamnt_ma_rech90	209593.0	12396.218352	16857.793882	0.000000	2317.000	7226.000000	16000.00	953036.000000
medianamnt_ma_rech90	209593.0	1864.595821	2081.680664	0.000000	773.000	1539.000000	1924.00	55000.000000
medianmarechprebal90	209593.0	92.025541	369.215658	-200.000000	14.600	36.000000	79.31	41456.500000
cnt_da_rech30	209593.0	262.578110	4183.897978	0.000000	0.000	0.000000	0.00	99914.441420
fr_da_rech30	209593.0	3749.494447	53885.414979	0.000000	0.000	0.000000	0.00	999809.240107
cnt_da_rech90	209593.0	0.041495	0.397556	0.000000	0.000	0.000000	0.00	38.000000
fr_da_rech90	209593.0	0.045712	0.951386	0.000000	0.000	0.000000	0.00	64.000000
cnt_loans30	209593.0	2.758981	2.554502	0.000000	1.000	2.000000	4.00	50.000000
amnt_loans30	209593.0	17.952021	17.379741	0.000000	6.000	12.000000	24.00	306.000000
maxamnt_loans30	209593.0	274.658747	4245.264648	0.000000	6.000	6.000000	6.00	99864.560864
medianamnt_loans30	209593.0	0.054029	0.218039	0.000000	0.000	0.000000	0.00	3.000000
cnt_loans90	209593.0	18.520919	224.797423	0.000000	1.000	2.000000	5.00	4997.517944
amnt_loans90	209593.0	23.645398	26.469861	0.000000	6.000	12.000000	30.00	438.000000
maxamnt_loans90	209593.0	6.703134	2.103864	0.000000	6.000	6.000000	6.00	12.000000
medianamnt_loans90	209593.0	0.046077	0.200692	0.000000	0.000	0.000000	0.00	3.000000
payback30	209593.0	3.398826	8.813729	0.000000	0.000	0.000000	3.75	171.500000
payback90	209593.0	4.321485	10.308108	0.000000	0.000	1.666667	4.50	171.500000

Description of the dataset.

- Data Sources and their formats
    1. Information of the dataset.
    2. Description of the dataset.
    3. Visualization.
    4. Correlation present in the dataset.
    5. Skewness is present.
    6. Outliers are present in the dataset.
    7. Target column is Imbalanced.

- Data Preprocessing Done
    1. Shape of our dataset is 209593, 37
    2. Dropped unwanted columns.
    3. Dropped correlated columns.
    4. Dropped columns whose score with the target column is less than 1.
    5. Apply Transformation - PowerTransformer (method = “yoe-Johnson”)
    6. Apply Oversampling on target column.

- Data Inputs- Logic- Output Relationships
    1. Correlation:

Column 1	Correlation	Column 2
daily_decr90	98%	daily_decr30
rental30	96%	rental90
cnt_ma_rech90	89%	cnt_ma_rech30
sumamnt_ma_rech30	89%	sumamnt_ma_rech90
medianamnt_ma_rech90	86%	medianamnt_ma_rech30
cnt_loans30	96%	amnt_loans30
amnt_loans90	90%	amnt_loans30
medianamnt_loans90	91%	medianamnt_loans30
payback30	83%	payback90

2. Feature Selection

	Feature Name	Scores
6	cnt_ma_rech30	12510.083303
11	sumamnt_ma_rech90	9268.913603
17	amnt_loans30	8486.771736
1	daily_decr30	6109.541601
8	medianamnt_ma_rech30	4281.623215
5	last_rech_amt_ma	3705.420270
10	fr_ma_rech90	1503.150310
21	maxamnt_loans90	1494.523796
2	rental90	1202.229543
22	payback90	508.224020
19	medianamnt_loans30	417.538139
12	medianmarechprebal90	324.209116
16	fr_da_rech90	6.152367
9	medianmarechprebal30	4.887441
20	cnt_loans90	4.695565
13	cnt_da_rech30	3.070127
0	aon	3.002765
3	last_rech_date_ma	2.913269
15	cnt_da_rech90	1.884674
4	last_rech_date_da	0.613610
7	fr_ma_rech30	0.370856
18	maxamnt_loans30	0.012865
14	fr_da_rech30	0.000147

Selected all the Features with score more than 1.

3. Applied PowerTransformer to remove skewness.

4. Used Oversampling method to balance the target column.

- Hardware and Software Requirements and Tools Used

Anaconda-navigator

jupyter notebook

matplotlib-inline==0.1.6

numpy==1.23.2

packaging==21.3

pickleshare==0.7.5

platformdirs==2.5.2

prompt-toolkit==3.0.30

pyparsing==3.0.9

python-dateutil==2.8.2

scikit-learn==1.1.2

scipy==1.9.0

sklearn==0.05

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
  - EDA
  - Visualization
  - Correlation
  - Features Selection
  - Normal Distribution
  - Outliers
  - Imbalanced Target Column
  - Final Dataset
  - Model Building
  - Cross-Validation
  - Saving the model.

- Testing of Identified Approaches (Algorithms)

  Algorithms used for the training and testing:

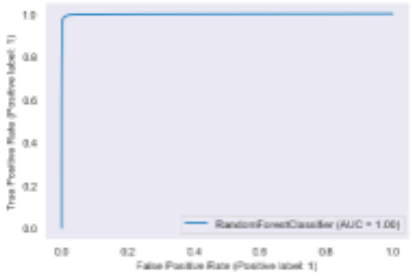
  - RandomForest Classifier.
  - AdaBoost Classifier.
  - KNeighbors Classifier.
  - GradientBoosting Classifier.
  - Logistic Regression.

- Run and Evaluate selected models

- RandomForest Classifier.

```
rf.fit(x_train,y_train)
score(rf, x_train,x_test,y_train,y_test,train = True)
score(rf, x_train,x_test,y_train,y_test,train = False)
```

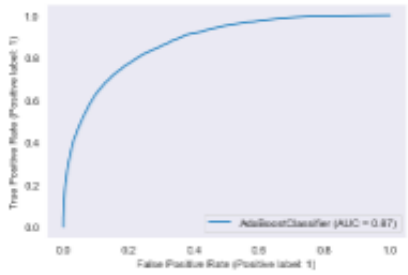
```
 ----- Train Result -----

Accuracy Score: 0.9997564929164879

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    137763
           1       1.00      1.00      1.00    137383

    accuracy                           1.00    275146
   macro avg       1.00      1.00      1.00    275146
weighted avg       1.00      1.00      1.00    275146


 ----- Confusion matrix -----
 [[137763      0]
 [    67 137316]]

 ----- Test Result -----

Accuracy Score: 0.9766998124645645

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       0.96      1.00      0.98     45668
           1       1.00      0.96      0.98     46048

    accuracy                           0.98     91716
   macro avg       0.98      0.98      0.98     91716
weighted avg       0.98      0.98      0.98     91716


 ----- Confusion matrix -----
 [[45551   117]
 [ 2020 44028]]

 ----- Roc Curve -----
```



- AdaBoost Classifier.

```
ada.fit(x_train,y_train)
score(ada, x_train,x_test,y_train,y_test,train = True)
score(ada, x_train,x_test,y_train,y_test,train = False)
```

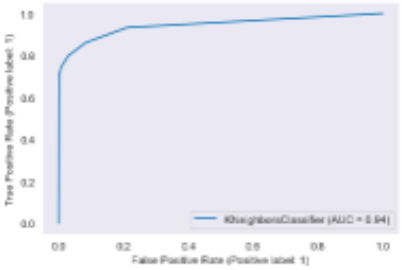
```
 ----- Train Result -----

Accuracy Score: 0.7866587193708068

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       0.78      0.80      0.79    137763
           1       0.79      0.77      0.78    137383

    accuracy                           0.79    275146
   macro avg       0.79      0.79      0.79    275146
weighted avg       0.79      0.79      0.79    275146


 ----- Confusion matrix -----
 [[110088  27675]
 [ 31025 106358]]

 ----- Test Result -----

Accuracy Score: 0.7856644424091761

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       0.78      0.80      0.79     45668
           1       0.79      0.78      0.78     46048

    accuracy                           0.79     91716
   macro avg       0.79      0.79      0.79     91716
weighted avg       0.79      0.79      0.79     91716


 ----- Confusion matrix -----
 [[36334  9334]
 [10324 35724]]

 ----- Roc Curve -----
```



- KNeighbors Classifier.

```
knn.fit(x_train,y_train)
score(knn, x_train,x_test,y_train,y_test,train = True)
score(knn, x_train,x_test,y_train,y_test,train = False)
```

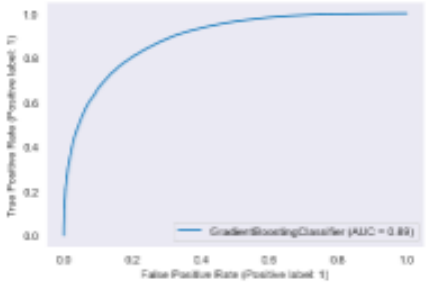
```
 ----- Train Result -----

Accuracy Score: 0.9199515893380241

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       0.87      0.99      0.93    137763
           1       0.99      0.85      0.91    137383

    accuracy                           0.92    275146
   macro avg       0.93      0.92      0.92    275146
weighted avg       0.93      0.92      0.92    275146


 ----- Confusion matrix -----
 [[136434   1329]
 [ 20696 116687]]

 ----- Test Result -----

Accuracy Score: 0.8837825461206332

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       0.83      0.97      0.89     45668
           1       0.96      0.80      0.87     46048

    accuracy                           0.88     91716
   macro avg       0.90      0.88      0.88     91716
weighted avg       0.90      0.88      0.88     91716


 ----- Confusion matrix -----
 [[44322  1346]
 [ 9313 36735]]

 ----- Roc Curve -----
```



- GradientBoosting Classifier.

```
gb.fit(x_train,y_train)
score(gb, x_train,x_test,y_train,y_test,train = True)
score(gb, x_train,x_test,y_train,y_test,train = False)
```

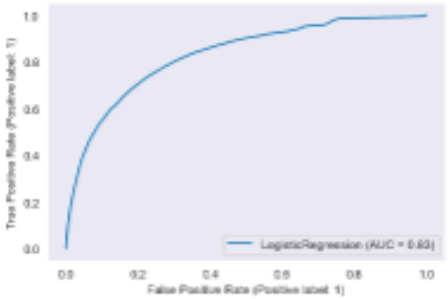
```
 ----- Train Result -----

Accuracy Score: 0.8034752458694656

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       0.80      0.81      0.80    137763
           1       0.81      0.80      0.80    137383

    accuracy                           0.80    275146
   macro avg       0.80      0.80      0.80    275146
weighted avg       0.80      0.80      0.80    275146


 ----- Confusion matrix -----
 [[111571  26192]
 [ 27881 109502]]

 ----- Test Result -----

Accuracy Score: 0.8014414060796371

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       0.80      0.81      0.80     45668
           1       0.81      0.80      0.80     46048

    accuracy                           0.80     91716
   macro avg       0.80      0.80      0.80     91716
weighted avg       0.80      0.80      0.80     91716


 ----- Confusion matrix -----
 [[36806  8862]
 [ 9349 36699]]

 ----- Roc Curve -----
```



- Logistic Regression.

```
lr.fit(x_train,y_train)
score(lr, x_train,x_test,y_train,y_test,train = True)
score(lr, x_train,x_test,y_train,y_test,train = False)
```

```
 ----- Train Result -----

Accuracy Score: 0.7537271121513669

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       0.75      0.75      0.75    137763
           1       0.75      0.75      0.75    137383

    accuracy                           0.75    275146
   macro avg       0.75      0.75      0.75    275146
weighted avg       0.75      0.75      0.75    275146


 ----- Confusion matrix -----
 [[103950  33813]
 [ 33948 103435]]

 ----- Test Result -----

Accuracy Score: 0.7538161280474508

 ----- Classification Report -----
              precision    recall  f1-score   support

           0       0.75      0.75      0.75     45668
           1       0.76      0.75      0.75     46048

    accuracy                           0.75     91716
   macro avg       0.75      0.75      0.75     91716
weighted avg       0.75      0.75      0.75     91716


 ----- Confusion matrix -----
 [[34419 11249]
 [11330 34718]]

 ----- Roc Curve -----
```



- Key Metrics for success in solving problem under consideration
  - Cross-Validation

    Score is approx. 97%
- Interpretation of the Results

  RandomForest Classifier, is giving the best score among all other models.

CONCLUSION

- Key Findings and Conclusions of the Study

  Cross Validation Score and model Accuracy score is very close, so we can say that our model is working well and doesn't having the overfitting/underfitting problem present.