ASSIGNMENT

Submitted By:

Name : Pawan Kumar

Roll no: 17PE10042

Petroleum Engineer

2017-2021

Indian Institute of Petroleum and Energy, Visakhapatnam

Email id: pawankumar@iipe.ac.in

Github link : https://github.com/Pawan17PE10042/Algo8ExcerciseProblem

Contact no: 8340519855

Title: *Delinquency Telecom Model*

Definition:

Delinquency is a condition that arises when an activity or situation does not occur at its scheduled (orexpected) date i.e., it occurs later than expected.

Use Case:

Many donors, experts, and microfinance institutions (MFI) have become convinced that using mobile financialservices (MFS) is more convenient and efficient, and less costly, than the traditional high-touch model for delivering microfinance services. MFS becomes especially useful when targeting the unbanked poor living in remote areas. The implementation of MFS, though, has been uneven with both significant challenges and successes.

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

One of our Client in Telecom collaborates with an MFI to provide micro-credit on mobile balances to be paidback in 5 days. The Consumer is believed to be delinquent if he deviates from the path of paying back the loaned amount within 5 days

Exercise:

Create a delinquency model which can 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 (label ‘1’ & ‘0’)
Find Enclosed the Data Description File and The Sample Data for the Modeling Exercise

Business objectives and constraints.

- No low-latency requirement for Paying Back loaned amount.
- Probability of a data-point belonging to each loan transaction is needed.

Performance Metric

- ➢ Log-loss (Since probabilities is our concern)
- ➢ Confusion matrix (Also want to check some precision and recalls)

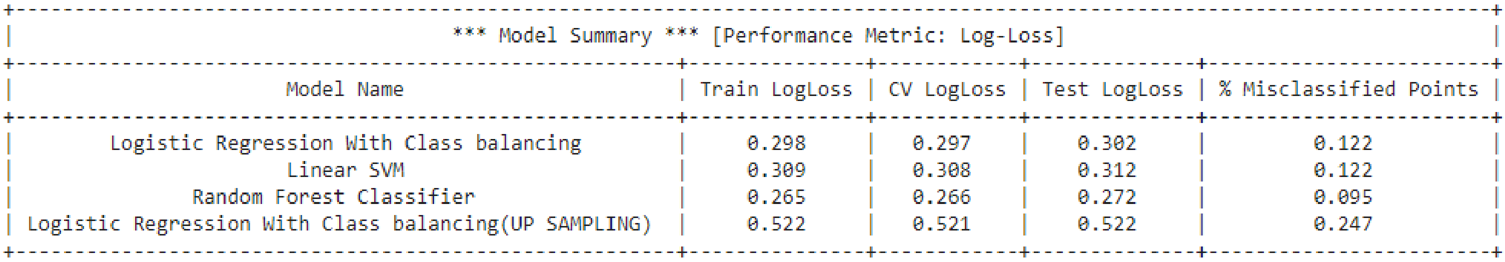
- **Why Log-loss Method for Solving this Problem ?**

Since we want a prediction probabilistic interpretation from the model under one of the two classes(1 or 0). so we will use **Log-loss** as the Metric here

Prediction Probability: The binary classification algorithms First predict probability for a recorded to be classified under class (1 or 0) based on whether the probability crossed a threshold value, which is usually set at 0.5 by default.

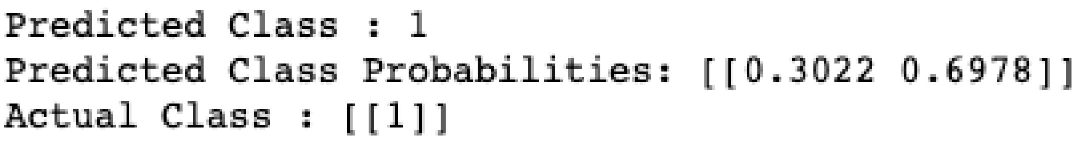
Steps Involved are :

1.Exploratory Data Analysis
2. Checking Missing values,NaN, Duplicates etc.
3. Checking Data Imbalances
4. Checking Correlations among features
5.Preprocessing the data
6. Train-Test Split
7. Random Model Design for comparing it's LogLoss with the ML models developed later on the dataset.
8. MODELS USED and their results



```
+---------------------------------------------------------------------------------------------------------------------------+
|                                    *** Model Summary *** [Performance Metric: Log-Loss]                                   |
+-------------------------------------------------------+---------------+------------+--------------+-------------------------+
|                      Model Name                       | Train LogLoss | CV LogLoss | Test LogLoss | % Misclassified Points |
+-------------------------------------------------------+---------------+------------+--------------+-------------------------+
|        Logistic Regression With Class balancing       |     0.298     |   0.297    |    0.302     |          0.122          |
|                      Linear SVM                       |     0.309     |   0.308    |    0.312     |          0.122          |
|                Random Forest Classifier               |     0.265     |   0.266    |    0.272     |          0.095          |
| Logistic Regression With Class balancing(UP SAMPLING) |     0.522     |   0.521    |    0.522     |          0.247          |
+-------------------------------------------------------+---------------+------------+--------------+-------------------------+
```

- ➢ Also we can see the probabilities for both classes, since probabilistic interpretation wasneeded. This is done for all models. Please do check in the ipynb file on given Github link For example :



```
Predicted Class : 1
Predicted Class Probabilities: [[0.3022 0.6978]]
Actual Class : [[1]]
```

CONCLUSION:

- In Log-loss method which have lower value of log-loss, that model will be the better performer.
- So, From the pretty table we can see that, **RandomForest** performed best here.
- Even the overfitting is not present if we check the train and test log-loss, they are very close
- Over sampling method was also applied on the training data to make the data more balanced, but it gave worse results.