HDFC Bank Machine Learning Hiring Challenge

Approach and Solution

By-

Vincy Sagar

Data Set Observations

- In the dataset of AggregateData_train there were features from V1 to V53, in which only 23 features had enough data to be used. Other columns contained ‘?’ (null value) greater than 90%.
- There were rows in the AggregateData_test that are same as that of the train data, and other rows with different data.
- In the dataset TransactionData_train feature C2 has same Ids that are in the V2 feature of AggregateData_train data set.
- The feature C8 in the TransactionData_train has same data as the V1 feature of AggregateData_train.
- Those features that should be in datetime format are in object format eg: V1, C4, C8 and C9.
- The above features have date out of range of month so we have to replace them with correct end date of month.

Approach

- First step was to separate the AggregateData_Test into 2 data sets.
- One data set that is same as that of the train data.
- Another data set is the data that is different from that of the train data.
- For the presentation point of view let us name the different data set as ‘testdiff’.
- Now let us name the same dataset as ‘testsame’.
- Create 2 models for both the datasets: Model1 for testdiff and Model2 for testsame.

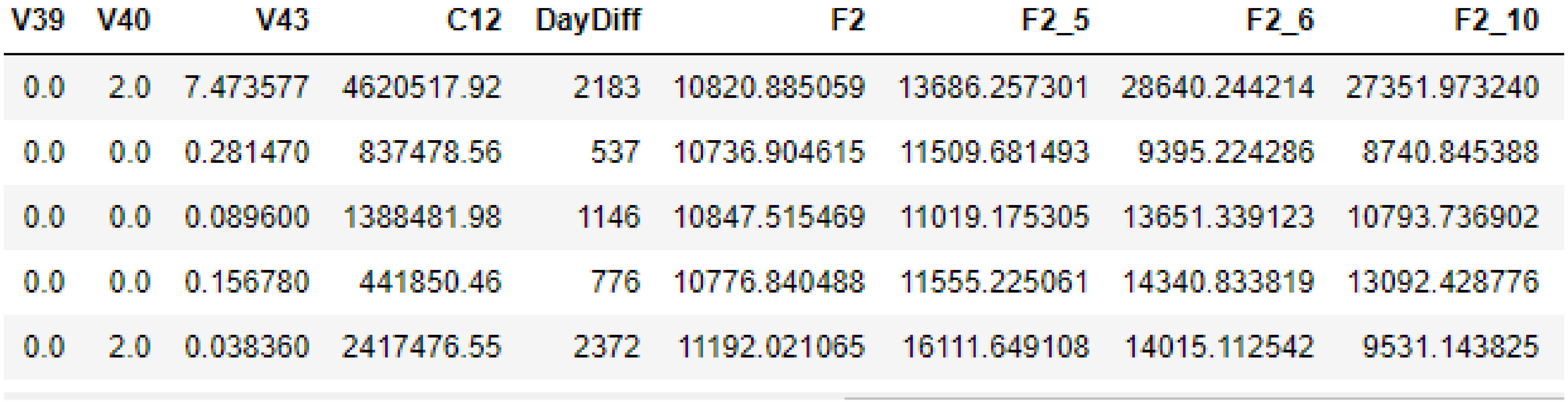
Data Preprocessing

- Firstly in the AggregateData_Train, the features that have ‘?’ in a large number are separated from the other features.
- The separated data set contains 23 features that can be used properly. Name it as ‘AggTrain’ data.
- Replace the ‘?’ in AggTrain data with NaN. It will convert to Null data type.
- Now, fill the Null values with the mean of the values of that particular feature.
- Change the data type of features with date and time values to ‘datetime’ format. For eg: V1(AggTrain) ,C4, C8, C9(Transaction Data)
- Followed the same process for test data.

Feature Extraction

- In TransactionData_Train create the feature ‘daydiff’ by subtracting the value of C4 from C9. This is the duration of the transaction.
- Merge daydiff feature to Aggtrain. This will help us in knowing the duration of transaction.
- Next feature is created by taking sum of C12 for every unique C2(id) and merge this feature C12 in AggTrain.
- Next feature is created by taking sum of C12 for every unique combination of C2 and C5. This will tell us the sum of C12 according to C5 i.e ‘C’(Credit) or ‘D’(Debit) for every unique C2 id.

- For next feature the above step is carried out on the C2 and C6 combination.
- For next feature the above step is carried out on C2 and C10 combination. In C10 the mode of transaction is given eg: ATM , so it will be helpful in making out the amount of C12 according to each mode in C10.

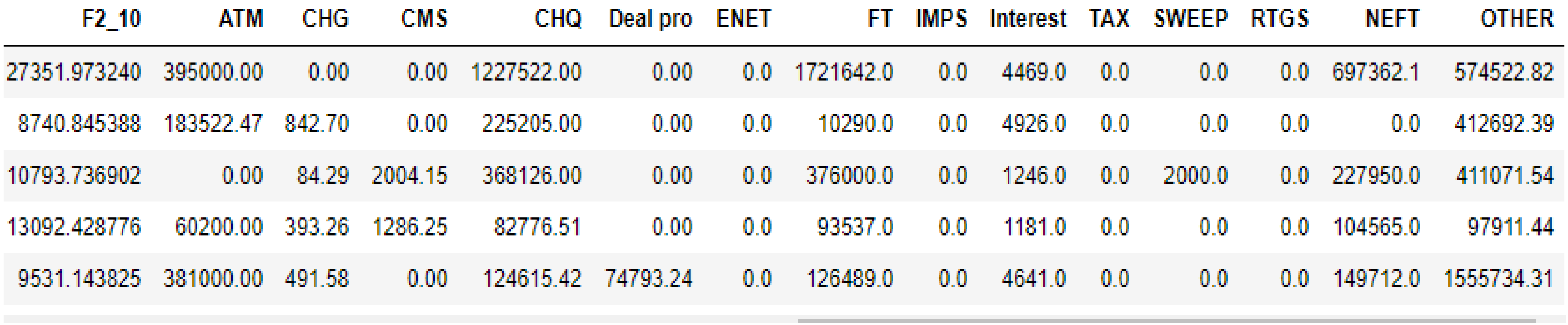


V39	V40	V43	C12	DayDiff	F2	F2_5	F2_6	F2_10
0.0	2.0	7.473577	4620517.92	2183	10820.885059	13686.257301	28640.244214	27351.973240
0.0	0.0	0.281470	837478.56	537	10736.904615	11509.681493	9395.224286	8740.845388
0.0	0.0	0.089600	1388481.98	1146	10847.515469	11019.175305	13651.339123	10793.736902
0.0	0.0	0.156780	441850.46	776	10776.840488	11555.225061	14340.833819	13092.428776
0.0	2.0	0.038360	2417476.55	2372	11192.021065	16111.649108	14015.112542	9531.143825

- Next features are created by taking the sum of C12 for each unique value of C5(i.e for D and C). It will help us in determining the value of transaction for Debit (D) and Credit(C) for every C2 id.

- Next features were created by taking the sum of C12 for each unique value of C10 (i.e the mode of transaction). These features will help us in determining how much transaction is done by client through every type of mode(for eg: ATM, CHQ,etc).

- In this way the unique values of C10 will become our features and will merge it into AggTrain.



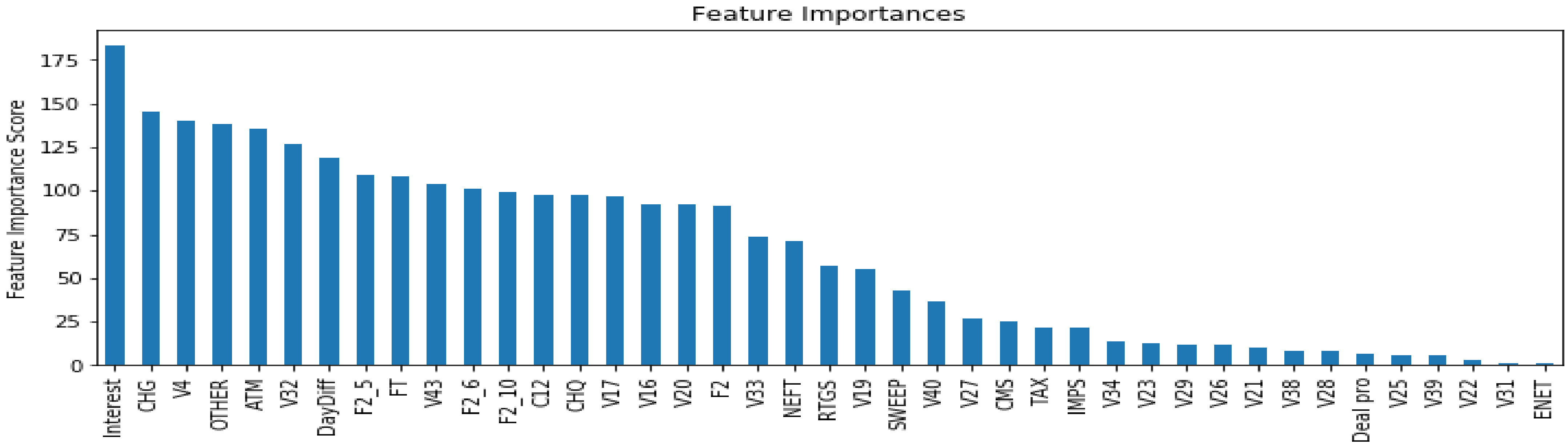
F2_10	ATM	CHG	CMS	CHQ	Deal pro	ENET	FT	IMPS	Interest	TAX	SWEEP	RTGS	NEFT	OTHER
27351.973240	395000.00	0.00	0.00	1227522.00	0.00	0.0	1721642.0	0.0	4469.0	0.0	0.0	0.0	697362.1	574522.82
8740.845388	183522.47	842.70	0.00	225205.00	0.00	0.0	10290.0	0.0	4926.0	0.0	0.0	0.0	0.0	412692.39
10793.736902	0.00	84.29	2004.15	368126.00	0.00	0.0	376000.0	0.0	1246.0	0.0	2000.0	0.0	227950.0	411071.54
13092.428776	60200.00	393.26	1286.25	82776.51	0.00	0.0	93537.0	0.0	1181.0	0.0	0.0	0.0	104565.0	97911.44
9531.143825	381000.00	491.58	0.00	124615.42	74793.24	0.0	126489.0	0.0	4641.0	0.0	0.0	0.0	149712.0	1555734.31

- These same process of creating features should be applied on test data also.

Model Building

Model 1: For the dataset ‘testdiff’

- ➢Apply XGBoost on AggTrain, use Grid Search CV for parameter tuning.
- ➢For getting good knowledge of which features are useful , I had used RFE and also plot the graph of feature score.
- ➢Predict the BadFlag probability for ‘testdiff’ using the same model1.



Model 2: For the dataset ‘testsame’

- Drop the columns ‘daydiff and C12 from AggTrain.
- Apply XGBoost on AggTrain and allow overfitting by setting the value of n_estimators and max_depth.
- Predict the BadFlag probability for ‘testsame’ using the same model2.
- Alternatively, we can replace the values of bad_flag with original values from the Aggtrain data.
- In the end we will obtain the dataset with UID and bad_flag for both, testsame and testdiff.

Combine the predicted results from both the models and sort the rows according to the UID to keep its original sequence of rows.

THANK YOU