JOB-A-THON May 2021

Pruthvi Vadhirajarao Sreenivasan

Student – University of New South Wales, Sydney, AUS

1) Findings in the Provided Data
   - The train data was of dimension 245725 rows × 11 columns.
   - The test data was of dimension 105312 rows × 10 columns.
   - There were missing values in column Credit_Product (12552 in test data and 29325 in train data)
   - The propotion of Occupation Entreprenour was bit less compared to other occupation as follows.

```
Other              0.349755
Self_Employed      0.324413
Salaried           0.318330
Entrepreneur       0.007502
```

   - The propotion of Channel_Code X4 was less than the other classes.

```
X1     102295
X3      60013
X2      55764
X4       4681
```

   - UpSampling the above to classes did not improve the performance of any model, hence, did not UpSample.

2) Encoding Data
   - Variables with Binary classes in them was encoded to 0 and 1 with sim ple lambda function.
   - Variables with more than 2 was encoded with Dummy encoder.
   - Variables with lot of classes like Region_Code was encoded with Target Encoding.

3) Imputing Missing Values
   - As the Imputing algorithm requires data in Numeric Format, Encoding was done before handling Missing Values.
   - The Missing Values was Handled with MissForest Algorithm.

4) Tried Polynomial Features
   - The output of polynomial features were around 680 features.
   - The top 40% of features which had high feature importance were used to build the model.
   - But the model performance was not good compared to simple catboost

5) Model Selection
   - I tried Multiple Models.

   - 1) First a Gradientboostingclassifer was tried, but the results in the JOB-A-Thon was not satisfying even after achieving an accuracy of 82%
   - 2) Second the CatBoost Model was Tried and the result for the model in Job-A-Thon as well as local model performance was good. With little Parameter Tuning the Model gave out My highest score of 0.872379832576796
   - 3) Tried Stacking GradientBoostingClassifier with CatBoost on top of LogisticRegression
   - Overall, The catboost model performed the best.