Title: CS6510 Hackathon In-class Kaggle challenge

Method we tried:

Logistic regression

XGBOOST

GradientBoostingClassifier

Random Forest

Naive Bayes

Data Preparation 1:

For a taxi\_type we consider maximum occurring taxi type for NAN

For a months\_of\_activity we consider a mean value of column to fill NAN

For a customer score, we consider mean value of a column to fill NAN

For customer\_score\_confidence maximum occurring score is consider for NAN

We have dropped the column anon\_var\_1 as it is having more missing values.

We have created dummies for the following column

customer\_score\_confidence

drop\_location\_type

taxi\_type

We have dropped the column sex as it is having less important according to corelation matrix.

Without Normalization accuracy:

Random Forest ->0.67758

XGBOOST -> 0.68824

Gradient Boosting-> 0.68560

We also did normalization on final data and we have a similar result as above.

Data Preparation 2:

Here I am predicting missing value using machine learning models.

customer\_score\_confidence -> RandomForestClassifier

months\_of\_activity -> linear regression

customer\_score -> linear regressiondrop\_location\_type -> RandomForestClassifier

taxi\_type -> random forest

Again, we have created dummies for the following column

customer\_score\_confidencedrop\_location\_type

taxi\_type

We got very less accuracy from above data preprocessing. It is overfitting.