Buğra Sipahioğlu

January 4, 2020

Mehmet Gönen, ENGR 421

Homework 07: Modeling Late Payments for Credit Card Bills

In this assignment, I've chosen to implement AdaBoost Classifier, with the base learner as the decision tree classifier, in Python language. Here is my implementation divided into parts:

Importing Libraries

I used sklearn's libraries for data processing, classification, calculations of ROC-AUC values. In order to plot the ROC AUC curves, I've used matplotlib.

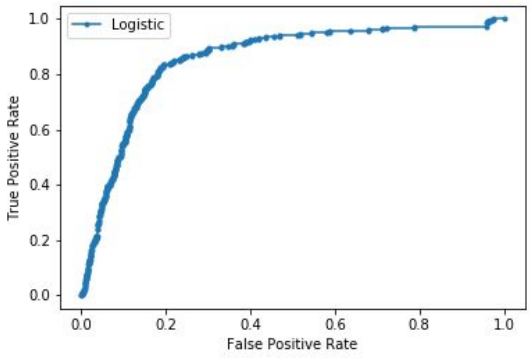
Preprocessing

First of all, I couldn't observe the joint probability distributions of features since number of features in the data is quite large. First observation about the data was the ID column. It is not necessary to include ID column into machine learning algorithm, therefore I've removed from training and test input datasets. Then, I've added dummy variables for categorical columns. This increased the dimension of the data. In order to avoid imbalance feature set between training and test datasets, I've concatenated train and test datasets for each target, get dummy variables and divided the datasets to test and train sets. Subsequently, I've replaced missing values with the means of other non-missing values in dataset. Final observation was that the numerical difference between data points was too high; both column wise and row wise. In order to make the data internally consistent, I've standardized the train and test datasets.

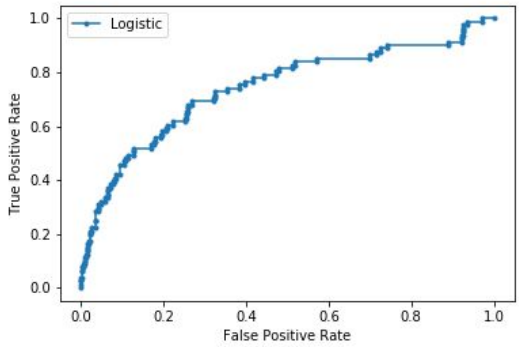
Machine Learning

The truth labels for test data was not given. In order to make sure that model works and improve the model by tweaking the parameters, I've split the train dataset into training and validation sets. Since the number of data is quite large, I've separated twenty percent of the training data for validation. In order to observe the results, I've plotted the ROC AUC curves. Here are the validation results for three targets:

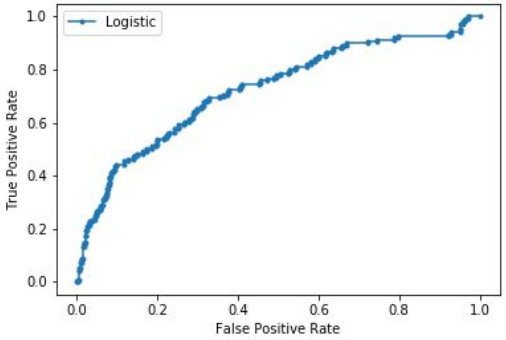
Target #1: ROC AUC = 0.855



Target #2: ROC AUC = 0.742



Target #3: ROC AUC = 0.724



Results

Disclaimer: I've indexed the test predictions with respect to given sample results, i.e., first target prediction index is starting from 11000, second is starting from 9000 and so on.

I've written the test predictions into csv files as described in the assignment description. Like in the given samples, there are two columns: "ID" and "TARGET". Target values are the probabilities of labels belonging to positive class labels.