Celal Yanpar
0062448

CSSM 502 Final Project Report

Project: Modeling Late Payments for Credit Card Bills

Project Scope

In this project, I developed a machine learning solution in Python with sklearn packages for real-life classification problems from finance industry. My solution needs to predict whether a customer will delay his/her credit card bill payment using the information given about each customer.

Data

The training and test sets contain 8800 and 2200 (80% to 20%) data instances, respectively, where each data instance has 162 features. While splitting data I used, sklearn's StratifiedShuffleSplit – cross validator method which returns stratified randomized folds. The folds are made by preserving the percentage of samples for each class.

PS: Since the case is originated from real data, feature name are not provided from the author of the data, so its not interpretable.

Data Preparation

At the beginning of the project, since there are large number of features, I decided to dimensionality reduction. After that, I focus on two main technique. One of them is Principal Component Analysis and the other one is Linear Discriminant Analysis. First, I tried to use PCA. However, when I compare results between PCA and LDA, I found that LDA' s performance is better than PCA in that specific data set. Since each part of homework is binary classification, number of features used in LDA is 1. So, its highly preferable in terms of complexity.

Before I implement LDA I needed to do couple of things as pre-processing

- Encode categorical variables – In each training set there are couple of features which have string variables and I encode those via label encoder method
- Then I fill the NA values with string of NaN, so I can also encode those
- For numeric NA values I used mean of specified columns.

Model Development

After that, I try 3 main algorithms to come up result.

- Gradient Boosting (boosting)
- Random Forest (bagging)
- Multi-Layer Perceptron

I train my test data with all the trained algorithms. Then I concatenate predicted score functions into one data frame. After that I take the mean variables of those predictions and write it as final_pred variable.

Model Evaluation

In my opinion, it can be more flexible to predict probabilities of an observation belonging to each class in a classification problem rather than predicting classes directly. One of the best diagnostic tools that help in the interpretation of probabilistic forecast for binary (two-class) classification predictive modeling problems is ROC Curves. I evaluate the model performance with ROC (Area Under Curve) score which is **0.88**