Report - Credit Score Classifier

Anonymous

November 8, 2018

1 Introduction

Given a dataset with personal informations about clients, the goal is to predict the probability of default , which is identified by the boolean variable *default* in the training dataset. Then, given a series of clients with undefined value for *default*, we submit the predictions for their corresponding IDs.

2 Requirements

The model was developed using the following libraries and resources, which can be jointly installed through the Anaconda package for Python 3.6.6 [1]:

- Python 3.6.6
- Jupyter Notebook
- Numpy 1.15.2
- Pandas 0.23.4
- Matplotlib 3.0.0
- Scikit-Learn 0.20.0

The code is written in a Notebook (.ipynb) file, opened on Firefox Browser (jupyter notebook –browser=firefox). It can be run through sequentially pressing the button *"run"* for each cell.

3 Design and Implementation

The full procedure consisted of:

1. Analyzing the data and the type of each feature;
2. Filling the NaN values in the data using different strategies for float, string and discrete variables;
3. Preprocessing the data, converting the categorical variables in their numerical equivalent;

[1] www.anaconda.com/download

4. Using the 0.7 train and 0.3 test ratios to find the best classifier among DecisionTreeClassifier, RandomForestClassifier and GradientBoostingClassifier.

5. Performing a grid search over the GradientBoostingClassifier to finally arrive at a reliable classification model;

6. Performing predictions at the unlabelled instance from the file *test_dataset.csv* and storing them at a file for submission.

4 Results and Performance Evaluation

For the final model, a GradientBoostingClassifier with *n_estimators* equal to 100 and *max_depth* of 5, the expected accuracy is 85.9273%.

5 Conclusions

The goal of the present project was, given a dataset with personal informations about clients, predicting their probability of default. After treatment of the data, a robust implementation of a Gradient Boosting Classifier was used, achieving satisfactory results.