Sapienza
Università di Roma

Fundamentals of Data Science and Laboratory

Home credit default risk

Autumn 2020

1 Summary

The goal of the challenge was to predict the probability that a customer will repay a loan granted by a credit institute. The project is composed by two files: **useful_functions.py** and **main_FDS.py**. The first .py file contains all the functions that have been used in order to accomplish the task, whereas the second one contains the code used to perform the analysis.The **main_FDS.py** is divided into 9 sections:

Section 1: Creates sub directories where store images

Section 2: Here data exploration has been performed on both application_train.csv and application_test.csv files. Moreover, the analysis to find out the most relevant categorical and numerical features (contained in these files) have been performed under this section.

Section 3: Here, bureau.csv, bureau_balance.csv, previous_application.csv, POS_CASH.csv, credit_card.csv and installments_payment.csv files have been imported and, from them, new features have been created.

Section 4: The new created features have been added to the train and test sets in order to create the final matrices to use in order to perform predictions.

Section 5: Here multicollinearity among features have been checked.

Section 6: The train and test sets have been alligned in order to have the same features.

Section 7: The train and test sets have been standardized and PCA has been computed. However, even if there was an improvement in terms of running time, PCA has not be applied, even maintaining the 90% of explained variance of dependent variable, due to the fact that there was a worsening on the performances.

Section 8: This is the section designed for Hyperparameter Tuning. Here the hyperparamiters of 4 classifiers have been tuned by using grid search cross validation. The classifiers were KNN, RandomForest, AdaBoosting and GradientBoosting. The returned best model was the AdaBoosting with the followinf settings:

- **algorithm:** SAMME.R
- **n_estimators:** 450
- **learning rate:** 0.1

Section 9: Here, the learned model has been passed through the test set and the predictions, in terms of probabilities, have been computed and stored inside the submission.csv file. The submission on Kaggle returned a public score of 0.74675.