Machine Learning Engineer Nanodegree

Vitor Bona de Faria

February 2019

1 Proposal

Banks pose as important actors in any person's life. Through them, with the correct guidance, someone may earn enough through investments to fulfill monetary goals, which could lead to some personal life goals too.

The amount of data available to banks is incredible, from customers income to consume habits and their posture when it comes to investments. Common challenges banks face may turn into binary classification problems: Is a customer satisfied? Will a customer invest in such stock? Will a customer be able to pay a loan?

This project will build a solution for the "Santander Customer Transaction Prediction"[1] problem. In this problem what is being predicted is the probability of a customer making a specific transaction in the future, regardless of the amount of money transacted.

All data is available in Kaggle and consists of anonymized data containing only numeric feature variables, the target binary column and a string column corresponding to the customer's ID. The data is composed of 2 files: train.csv; test.csv. Both files contain 200k observations and 201 features, named as: var_0, var_1, ..., var_199, target.

This is a binary classification problem, thus its solution will be based on classification algorithms such as Logistic Regression, Decision Trees and K-Nearest Neighbours (KNN). The evaluation metric proposed in the problem is the area under the ROC curve. A good model, as stated in some of this course's videos, would have at least a value of 0.8 for the area under the ROC curve, therefore this will be the reference value. For further acceptance

[1] https://www.kaggle.com/c/santander-customer-transaction-prediction

of the proposed model the Kaggle competition leaderboard will be used for benchmark.

The solution will comply to the following sequence: data exploration and visualization, model decision, evaluation and validation.

The data exploration and visualization will consist of peeking the data by checking the first lines of our dataset, some graphics displaying features distributions and checking if the data is balanced or not. Along with that outlier detection will be employed in order to decide whether to remove discrepant data or not.

The model decision will consist of a pipeline where 3 classification algorithms will be compared using 3 different metrics: time, area under the ROC curve and f2-score. At this point, the classification algorithms will be run without any parameter tuning and against 3 dataset sizes, consisting of 1%, 10% and 100% of the training data available.

The model evaluation and validation will consist of using cross-validation to tune the selected model parameters and checking its performance based on the area under the ROC curve. Finally the predictions for the testing data will be made and submitted to kaggle and results will be discussed concluding the project.