Machine Learning Engineer Nanodegree

Vitor Bona de Faria

March 2019

1 Domain Background

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? Successful applications of machine learning to this type of tasks are found in the literature [1].

2 Problem Statement

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.

3 Dataset and Inputs

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;

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

test.csv. Both files contain 200k observations and 201 features, named as: var_0, var_1, ..., var_199, target.

4 Solution Statement and Benchmark

This is a binary classification problem, thus its solution will be based on classification algorithms such as Logistic Regression, Decision Trees, K-Nearest Neighbours (KNN), LightGBM and more. The evaluation metric proposed in the problem is the area under the ROC curve. The chosen benchmark for this project is the Complement Naive Bayes classifier. Since we look for a specific type of bank clients in this project, it's expected that our data will be somehow unbalanced, in which case the proposed benchmark is well suited.

5 Project Design

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

The data exploration, visualization and preprocessing will consist of the following steps:

- Peeking at first lines of the dataset
- Checking for normal with some features distribution graphics
- Checking if data is balanced or not
- Checking for any missing values in dataset
- Feature rescaling for better model performances
- Detecting outliers
- Checking features importances

If any missing values are found, one of the following three actions will be taken, taking into account features particularities : missing value will be replaced by mean value, missing value will be replaced by discrepant value or missing value will be replaced by the most common value.

The model decision will contain a pipeline where 3 supervised learning 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. Other supervised learning algorithms may be tested outside of the pipeline. Unsupervised learning algorithms will be restricted to KNN. Since we don't have much previous information about the anonymized dataset, unsupervised learning algorithms are expected to perform poorly.

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. Depending on the chosen algorithm, parameters may be tuned using grid search, randomized search or even manually. Finally the predictions for the testing data will be made and submitted to kaggle and results will be discussed concluding the project.

References

[1] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.