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" problem. In this problem what is being predicted is the possibility 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 numeric feature variables, the target binary column and a string column corresponding to the ID. The data is composed of 2 files: train.csv; test.csv. Both files contain 200k observations and 201 features.

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. The best submission available in Kaggle has a score of 0.906, therefore we'll lower the bar just a little and use 0.89 as the benchmark value.

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