Santander Customer Transaction Prediction

April 21, 2019

1 Sahba Salarian

2 Introduction

Banking industry evolves at a rapid pace with new technology innovations, changing consumer dynamics, geopolitical movements and evolving demographics. Currently we are in the midst of and industrial revolution in banking industry which new models and technologies will change the way business is conducted.

Lots of new fintech firms or bank tech firms have erupted in recent times to target newer groups of consumers who prefer to do things virtually. Some banks have also developed solutions to target thin-file customers. For example, technology can help monitor various alternative sources of information for example on creditworthiness, like ensuring they are paying rent and utilities on time.

There are many areas in this evolving industry where Machine Learning can be beneficial. All these potensial areas, require advanced customer's data analysis with high accuracy and prediction capabilities. Among these areas are:

1. Manage portfolios with algorithms: AI and machine learning could streamline the process for developing a portfolio by assessing a customer's goals and risk tolerance to develop an individualized portfolio.
2. Conduct High-Frequency Trading: Trading has the potential to accelerate at a faster pace when AI and machine learning algorithms are utilized for trading decisions.
3. Detect Frauds And Threats to Financial Systems: Routine checks of risk factors that could affect customer information can provide an understanding of the potential threats. As a result, the response of initial invasion detection can be quicker, and the security at financial institutions can flag the unusual behavior for monitoring.

In this project, as a step toward fulfilling these ends, advanced data analysis tools are used to predict and identify which customers of Santander bank will make a specific transaction in the future, irrespective of the amount of money transacted.

The data has been provided by Santander bank for a competition for kagglers on kaggle website. The dataset has 200k observations and 200 column variables for each customer and a class observation defining whether the customer makes a transaction or not. The target is to predict a binary classification response with highest posible accuracy which will be evaluated by comparing the real data from Santander bank and the predicted values for a separate test data set.

The results from this project were among the 10 percent top accurate predictions and has been awarded the Bronze medal.