LeMaur Kydd

Santander Customer Transaction Prediction

1. Introduction

Predictive modeling has proved to be one of the most helpful and insightful advances of modern data science. One application of predictive modeling that Santander Bank has employed is for the purpose of helping their customers understand their financial health and locate the products and services that may assist them in achieving their financial goals. The goal of this white paper is to layout the process from EDA to modeling on the anonymized dataset from Santander Bank.

1. Dataset

The dataset in question is from Santander Bank on their respective Kaggle challenge page. It has been anonymized for the safety of customer data and consists of roughly 200 features representing 200,000 customers.

1. Data Preparation

As with most other data science projects, the first major task is to prepare the data. Regarding this dataset, different types of data transformation will be needed e.g., normalization, one hot encoding. Checking for null values or major outliers will most likely be necessary as well.

1. Exploratory Data Analysis

Post data preparation, the next step is to complete varying forms of EDA or descriptive statistics. Plotting out histograms for each column of data would give insight to any major skewness or large standard deviations, and it would show if the data is primarily normally distributed or other types of distributions. Also, running the simple pandas function ‘pd.describe()’ on all columns would give valuable insight as to what type of data is present.

1. Modeling

Simply because this dataset has been heavily anonymized, there may have to be multiple modeling techniques used. There are a large number of columns, therefore dimension reduction by PCA, LDA, or Random Forest will have to be employed to subset the important variables. For the task of binary classification, deep leaning neural networks with a final layer that return a 1 or 0 should be a potent model with enough training. Perhaps Random Forest would also be a good model to make a binary classification.

1. Evaluation

The testing dataset will be used to test the effectiveness of the model as well as the Kaggle submission feature that grades the results. The ROC curve is also taken into account when evaluating/scoring a submission. The accuracy/precision/ of the testing/training of the neural network is another good metric to assess.