Machine Learning Engineer Nanodegree

Capstone Proposal

Santander Customer Transaction Prediction

Éderson André de Souza
September 13, 2019

Proposal

Domain Background

It is not unusual to hear a company's management speak about forecasts: "Our sales did not meet the forecasted numbers," or "we feel confident in the forecasted economic growth and expect to exceed our targets." You cannot predict the future of your business, but you can reduce risk by eliminating the guesswork. With accurate forecasting, you can make a systematic attempt to understand future performance. This will allow you to make better informed decisions and become more resistant to unforeseen financial requirements. Without correctly estimating financial requirements and understanding changing markets, your business decisions will be guess work which can result in insufferable damage.

So, with that in mind, without doubt, it is very important to help business forecasting future products and services demands.

And because of that, I chose the Santander Customer Transaction Prediction dataset to try building a model that can consistently handle this task.

Problem Statement

Banco Santander, S.A., doing business as Santander Group, is a Spanish multinational commercial bank and financial services company based in Madrid and Santander in Spain. Additionally, Santander maintains a presence in all global financial centres as the 16th-largest banking institution in the world. Although known for its European banking operations, it has extended operations across North and South America, and more recently in continental Asia. Wikipedia

In their Kaggle competition, Santander provided an anonymized dataset containing numeric feature variables, the binary target column, and a string ID_code column; the goal is to build a model that predicts the probability of a customer make a specific transaction in the future.

The model is evaluated on AUC (Area Under the ROC Curve) between the predicted probability and the observed target.

Datasets and Inputs

The dataset provided by Santander onKaggle competition includes approximately 400k costumers, split into training and testing sets. The training set contains more than 200k rows of data and 200 features, the binary

target column, and a string ID_code column. The testing set contains about same 200k costumers, 200 features and the string ID_code column.

There is no description of the features. They are just numeric and contains both positive and negative values.

The training dataset is very unbalanced. Only ~10% of the customers made a transaction in the past, which means we have about 20k "ones" in the target column.

Solution Statement

The solution is a classification model capable of predicting whether a customer will make a transaction or not in the future. First, I will use Pandas and Numpy to gain some understanding of the data and cleaning it, if necessary. For the model, I am inclined towards XGBoost, a powerful Gradient Boosting framework that has proven itself in many past Kaggle competitions while being versatile and work with other frameworks such as scikit-learn.

Benchmark Model

For the baseline benchmark, I have randomly predicted with 10% probability (the distribution of the training set) that a customer will make a transaction. This method yields an AUC score of ~ 0.50 on the submission to Kaggle. This is equivalent to guess that all customers will not make the transaction, which is very bad and naive.

So, if the final model results in an AUC score better than the 0.50, we have succeeded.

Evaluation Metrics

A model in this competition is evaluated on AUC (Area Under the ROC Curve) between the predicted probability and the observed target, measured on the test data. Since the test data is not labeled, grading is done by uploading the file containing the probability of a customer making a transaction to Kaggle.

Project Design

- **Programming language:** Python 3.7
- **Library:** Pandas, Numpy, Scikit-learn, XGBoost
- **Workflow:**
    - Establish basic statistics and understanding of the dataset; perform basic cleaning and processing if needed.
    - Train a base classification model on the given data as-is to gauge the performance.
    - Fine tune the model's hyperparameters.
    - Perform training.

Prospects

I think that with all the learning far, I may have the ability to obtain a good result in this task. I think the biggest challenge will be the feature engineering because there is no explanation what so ever about the features and they are meaningless numbers.