Machine Learning Engineer Nanodegree

Capstone Proposal

Soham Mukherjee
April 12, 2018.

Santander Customer Satisfaction

Domain Background

This problem is taken from Kaggle – a platform for predictive modelling and analytics competitions. Companies and different groups post a problem statement on this platform along with a declaration of some sort of prize which can be in the terms of prize money , job interview or any other kind of prizes.

Santander Customer Satisfaction is one such competition posted by Santander Bank on Kaggle asking it’s user base for a possible solution to the problem declaring about $60,000 as prize money. As for any business, customer satisfaction is the key to its success. Similarly, Santander is asking Kagglers to find a possible machine learning algorithm to identify unhappy customers early in their relationship.

Problem Statement

Santander Bank asks Kagglers to help them identify dissatisfied customers early in their relationship. For any business, customer satisfaction is the key for its success. Santander wants to identify unhappy customer so that it can take appropriate measures in order to improve their satisfaction before it’s too late. Identifying unhappy customers don’t tend to stick around and the challenge in this situation is that, unhappy customers don’t tend to voice their dissatisfaction before leaving. Thus, the only way to find out is through the historical data containing hundreds of features which might help in predicting if a customer is satisfied or not.

Datasets and Inputs

Santander Bank provided Kagglers with 3 data sets. Train.csv , test.csv, and sample_submission.csv. The only challenge is , the dataset is an anonymized dataset which means, there is no way to tell which column corresponds to which feature and what it describes. Only a thorough research into the domain would reveal the description. Exploratory data analysis is crucial which would be included in the final report. The inputs are anonymized and the target variable describes if the customer is satisfied or not – 1 for unsatisfied and 0 for satisfied.

Solution Statement

The solution to this problem is to apply classification algorithms to the training data set to train the model then use the model to test it against the given test.csv. The challenge would be to find what the features describe and engineer the features and find out a suitable classification algorithm by evaluating some algorithms namely – Logistic regression ,Decision trees, Ensemble methods and also neural nets and to find out whether a single algorithm works well or a combination would work best.

Benchmark Model

The model would be able to correctly classify a customer from the test data set from as 1 for unsatisfied and 0 for satisfied. Most of the Kagglers have been using XGBoost as a gradient boosting decision tree to come up with a benchmark AUC score of 0.82. We will keep this score as the benchmark score and try to improve it.

Evaluation Metrics

According to the project description provided by the bank at Kaggle, the submission are to be evaluated on area under the ROC curve.

The ROC or Receiver Operating Characteristic curve plots the true positive rate against the false positive rate. This is used to diagnose the ability of a binary classifier system. It can be used to find possibly optimal models and to discard suboptimal ones from one another. The true positive rate is known as the recall and the false positive is known as the fall-out. Thus, the ROC curve can be thought of as a plot of Power as a function of Type 1 error of the decision rule.

Project Design

The general lifecycle of the project can be formulated as a series of steps as described below :

1. **Exploratory Data Analysis -** This involves going through the data through a series of statistical techniques to understand the data well enough to draw conclusion about the data's behavior and the to draw a hypothesis as to which model or models to use . This steps also includes a series a Data Visualization techniques to see visible patterns of the predictors and the target variable.
2. **Data Preprocessing –** Normalization and Scaling of the data is necessary to bring all of the features in the same range to make the models more sensitive to the data. This step also includes dividing the data into train, validation and test set. As the bank already has provided us with the train and test , we might be able to divide the training set into train and validation to test our models accuracy.
3. **Feature Engineering –** This step includes finding relevant features and engineer new features wherever possible.
4. **Model Selection –** This step includes experimenting with various algorithms to find the best possible model for our problem.
5. **Model Tuning –** This step includes finding the hyperparameters of the model we might use and tune it to increase our performance keeping our mind the problems of over and under fitting.
6. **Testing –** This steps includes testing the model on the test set provided by the bank.

∗Note – The steps 3-5 act as a cycle by evaluating the final model at each stage against the validation set.