Machine Learning Engineer Nanodegree

Capstone Proposal

Scott Yao

07/31/2017

Proposal

I plan to work on the Santander Customer Satisfaction competition from Kaggle.

Domain Background

Santander Bank is asking Kagglers to help them identify dissatisfied customers early in their relationship. Doing so would allow Santander to take proactive steps to improve a customer's happiness before it's too late.

Problem Statement

In this competition, I'll work with hundreds of anonymized features to predict if a customer is satisfied or dissatisfied with their banking experience.

Datasets and Inputs

The competition will provide train and test dataset containing a large number of numeric variables. The "TARGET" column is the variable to predict. It equals one for unsatisfied customers and 0 for satisfied customers.

The competition is to predict the probability that each customer in the test set is an unsatisfied customer.

Solution Statement

The solution will contains a jupyter notebook which includes data analysis, python code, graphs and conclusion.

Benchmark Model

Thanks to the Kagglers in the discussion forum, they do provide some good examples that I can use as benchmark. This one using ExtraTreesClassifier to select 36 useful features and then apply XGBClassifier, got a 0.83 score. I like this solution because it's pretty straightforward and has a good score. I'd like to explore more on feature/model selection and parameter tuning to get a better result.

Evaluation Metrics

Results will be evaluated on area under the ROC curve between the predicted probability and the observed target.

Project Design

This is a typical supervised learning problem and falls into the classification category. My planned approach for this problem listed below:

1. Data exploration. Statistical analysis for features
2. Data cleansing
3. Feature selection. (Maybe adopt PCA)
4. Model comparison and selection. (Compare multiple sklearn classifier models)
5. Parameter tuning with grid search
6. Providing the final conclusion.