Machine Learning Engineer Nanodegree
Capstone Proposal

Norah Alhomaimidi

January 12st, 2019

Domain Background: customer churn

Since customer satisfaction is an essential factor to success in any service industry, identify dissatisfied customers early and maintain them before leaving is the most challenging task. Especially that Unhappy customers rarely voice their dissatisfaction. Therefore, the longer a client stays with an organization, the more value he creates.

By focusing on the operating environment of the banking market, it observed that it is very challenging and competitive because of its nature for the profit growing needs, against the clients' variable demands. Thus, banks' Customer Relationship Management increasingly focused on identifying customer segments, needs, and satisfaction to avoid customer churn.

However, because I am interested in looking for a job in Customer Value Management Analytics field, I found this problem appropriate to discover, investigate and solve a real case by apply what I have learned in this course.

Problem Statement:

Santander Bank, which is a large corporation focusing principally on the market in the northeast United States, have asking Kagglers to help them identify dissatisfied customers early in their relationship through a Kaggle competition (Santander, 2015). This competition had aimed to predict whether a client will be dissatisfied in the future or not based on specific characteristics, to help Santander Bank to take proactive steps regarding improving a customer's happiness before it's too late and customers already left the bank.

Datasets and Inputs:

This competition has been provided by anonymized dataset containing a large number of numeric variables. This Data has been split into training and target sets. I decided to work on the training set because it contains the predicted variable TARGET, while the target set does not. The training set includes 76,020 rows of data and 371 features, where the TARGET column equals 1 for unsatisfied customers and 0 for satisfied customers.

Solution Statement:

Since the data is labeled by satisfaction information, the solution is obtaining a capable classification model that can predict whether the client is satisfied or not. First, I have to conduct a PCA technique on the data to reduce dimensionality space of features into a smaller number of predictor variables that can represent the data very well. Second, I will split the data into train and test datasets, then train and test the data on different classification algorithms in order to pick the model that has the highest accuracy and the best performance to predicting the customer satisfaction information.

Benchmark Model:

Since this problem is a classification problem, so the objective of the machine learning modeling is to identify whether the customer is satisfied or no, and the benchmark will be the highest performance model among the following:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Naive Bayes

Evaluation Metrics:

The model in this competition evaluated on the area under the ROC curve between the predicted probability and the observed target.

Project Design:

- Programming language:
  Python 3.6
- Library:
  Pandas, Numpy, Scikit-learn
- Workflow:
  1. Exploring the data by conduct some plots and descriptive statistics to understand the dataset.
  2. Perform some necessary cleaning if needed.
  3. Conduct a PCA technique to reduce dimensionality space of features into a smaller number.
  4. Train/Test different classification models on the data and measure the performance for each one.
  5. Pick the model with the best performance.