Machine Learning Engineer Nanodegree

Capstone Proposal

Abdulellah Alnumay
July 30th, 2018

Domain Background

From frontline support teams to C-suites, customer satisfaction is a key measure of success. Unhappy customers don't stick around. What's more, unhappy customers rarely voice their dissatisfaction before leaving. In this problem I'll try to identify dissatisfied customers[1] early in their relationship. Doing so would allow the bank to take proactive steps to improve a customer's happiness before it's too late. I'll work with hundreds of anonymized features to predict if a customer is satisfied or dissatisfied with their banking experience. This is Santander Customer Satisfaction competition on Kaggle[2].

Problem Statement

The problem is a binary classification problem[3], and the goal is to predict whether a customer is satisfied with the bank or not based on given data.

Datasets and Inputs

The dataset is provided by Santander Bank as part of the competition[4], and it’s available to download for everyone. It is a large anonymized dataset consisting of 76,000 instances, and 371 attributes, of which 2 can be understood:

- ID: Customer’s ID
- TARGET: 0 if customer is satisfied, 1 if not. This is the attribute to predict

The dataset suffers from an imbalance, where 96% of the dataset represents one class, and only 4% represents the other. For this, I hope that the evaluation metric that I will be using is enough to handle this imbalance. If not, I will try under-sampling the dominating class to reduce the imbalance. The last resort would be turning to semi-supervised learning and using the unlabeled dataset provided by Santander Bank in the competition.

I will be splitting the dataset into training, validation, and testing sets. I will use the training set to train the models, and the validation set to tune each model to better perform. And finally I will use the testing set to compare the models against each other to determine the better performing model.

Solution Statement

Since this a classification problem, most suitable solution would be using a classification algorithm. There are many classification algorithms, but I will be using SVMs, XGBoost, and Neural Networks, and comparing their performance. I will also perform feature selection to reduce the number of features.

Benchmark Model

The benchmark model for this problem would be a Random Forest model, and I don't think it would be an easy task to beat a powerful model as Random Forest. I will be using area under receiver operating characteristic curve to evaluate the performance of this model, and to compare my choices of models to.

Evaluation Metrics

The evaluation metric used in the competition is the area under receiver operating characteristic curve[5], which I will be using to evaluate the performance of the models trained.

Project Design

Data Exploration:

I would like to start by exploring the data, like determining the number and percentage of satisfied customers, and the mean values of the attributes, and other useful information about the data.

Data Preprocessing:

Preparing the data properly is critical for achieving good performing models, and in this step I will use what I learned from the data exploration to treat attributes with skewed values or abnormal distribution. Also, if there's any attributes that need normalizing, I will also perform that.

Feature Selection:

369 attribute is a large number of attributes. I will try to reduce the number of attributes to the most relevant and important attributes.

Model Selection:

In this step I will train the models with the algorithms I think would be proper for this sort of problem. Currently I think training models Naive Bayes, SVM, XGBoost, and Neural Network would be sufficient to get to a well performing model.

Model tuning:

One I've selected a model, I will tune it's hyper parameters to further improve the performance of the model.

References

1. http://www.sciedu.ca/journal/index.php/air/article/viewFile/10534/6572
2. https://www.kaggle.com/c/santander-customer-satisfaction/
3. https://pdfs.semanticscholar.org/55b0/b0478fad2cd6c70d851e2be594a2788910ff.pdf
4. https://www.kaggle.com/c/santander-customer-satisfaction/data
5. https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve