

Machine Learning Engineer Nanodegree

Capstone Proposal

Seif El-Eslam Ibrahim Hegazy

seif.eleslam1990@gmail.com

May 15th, 2017

Proposal

Domain Background

Supervised learning is one of the most promising fields in Machine Learning. It is currently used in several real world applications. Customer is the focal point of any business, and his satisfaction is directly proportional to business growth. An important point that affects customers is the service they get and their level of satisfaction with this service.

Machine learning helps the business improving this issue. My personal motivation is investigating this area and provide the business with a tool that uses supervised learning and improves customer's satisfaction and push business forward.

Problem Statement

Santander Bank, started in Spain and it has been serving customers in the Northeast since 2013, cares about its customers satisfaction and work to take proactive steps to improve a customer's happiness before it's too late . From frontline support teams to C-suites, customer satisfaction is a key measure of success. Unhappy customers don't stick around, so Santander Bank posted a competition in **Kaggle.com** to identify dissatisfied customers.

As an input, Santander Bank provides hundreds of anonymized features to predict if a customer is satisfied or not based on their banking experience.

This task is **a classification problem** since customers should be classified as happy or not.

Problem link: <https://www.kaggle.com/c/santander-customer-satisfaction>

Datasets and Inputs

The dataset contains two csv files with anonymized features for classification.

- Data given is:

- ID: customer id.
- 369 numeric variables
- Target: 0/1 variable indicates customer happiness.

- Train.csv:

- Number of rows = 76020
- Number of columns = 371
- Highly relevant as this is the data for training.

- Test.csv:

- Number of rows = 75818
- Number of columns = 370
- Highly relevant as this is the data for testing.

It is a Kaggle competition. Dataset link: <https://www.kaggle.com/c/santander-customer-satisfaction/data>

Solution Statement

First, Some data exploration and visualization will be done to understand the relation among the features and the target label. There are 369 numeric features which, due to curse of dimensionality, may result in model over-fitting, accuracy reduction and increasing training time. So, principle component analysis (PCA) will be used to *reduce the dimensionality* and *feature selection*. Then, some models will be trained and tested to get the best performance. Finally, tuning the best model and get the best accuracy.

Models that will be used:

- 1- *LogisticRegression* as a base model.
- 2- *RandomForest* which is a great Bagging Algorithm.
- 3- *XGBoost* which is a great Boosting Algorithm.
- 4- *Deep neural network* which is a type of Deep Learning (if needed).

For tuning hyper-parameters, either *Randomized Search* or *Grid search* will be used based on a comparison.

Benchmark Model

As this is a Kaggle competition, a benchmark model would be the best Kaggle score for the test-set, which is 0.829072 of “*roc_auc_score*” accuracy of private leader-board. But, **for academic purposes**, a part of training data will be used as testing data. More precisely, 20% of training data will be set as a testing set.

A **LogisticRegression** model will be set as the benchmark model for this problem. Then, next models will be compared to it to see if they can beat it and by how much extent.

For satisfying the curiosity, The best model will be tested on the original test-set provided by Kaggle. The submission file will be uploaded on Kaggle website to check the score. A personal goal, The best model would have a score over 0.80 accuracy in Kaggle leader-board.

Evaluation Metrics

The official evaluation metric of this problem provided by Kaggle is area under the **ROC curve** between the predicted probability and the observed target. The same metric will be used for models evaluation.

Project Design

- Programming Language and Libraries:

- Python 3.
- Jupyter Notebook.
- Pandas.
- Numpy
- Scikit-learn: open-source machine learning library for Python.
- Xgboost: package for gradient boosted decision trees implementation.
- Keras: open-source neural network library written in Python.
- Other libraries will be added if needed.

- Work-flow:

The general sequence of steps are as follows:

Data Exploration and Visualization:

- Reading the data and providing some statistics.
- Provide some visual representation of data to find the degree of correlations between features and target variable
-

Data Preprocessing

- Scaling the data as preprocessing step as all features are numeric variables.
- Splitting the data to training and testing sets.

Set the Benchmark:

- Using LogisticRegression as a benchmark model.

Feature Engineering:

- Doing dimensionality reduction and feature selection using PCA.

Model Selection:

- Building several models (RandomForest, XGBoost and Deep neural network) and testing their accuracy against the benchmark model.

Model Tuning:

- Tuning the best model using either Randomized Search or Grid search.

Testing:

- Test the best model on testing dataset.

References

- Kaggle, Santander Customer Satisfaction competition:
<https://www.kaggle.com/c/santander-customer-satisfaction>
- Santander web site: <https://www.santanderbank.com/us/personal>
- Supervised learning Wikipedia page:
https://en.wikipedia.org/wiki/Supervised_learning