Machine Learning Capstone Project Proposal

Santander Customer Transaction Prediction

Name: Po-sheng Wang

Date: 03/10/2019

Domain Background:

Machine Learning techniques have been used widely on the financial field in terms of helping people understand their financial health and identify which products and services might help them achieve monetary goals. The prediction and classification algorithms could assist the financial organization in providing the proper solutions to the customers.

Some related research in this field have been done before. The regression algorithms are used to model the relationship between variables. Decision tree algorithms construct a model of decisions and are used in classification or regression problems. Here are some example and paper of how we can predict stock movements with the help of Decision Trees, SVM and Reinforcement Learning [1][2][3]

In this paper [4], it presented recent developments in stock market prediction models, and discuss their advantages and disadvantages. It includes Traditional time series prediction, Neural Networks and Support Vector Machine. The result shows that the NN offer the ability to predict market directions more accurately than other existing techniques.

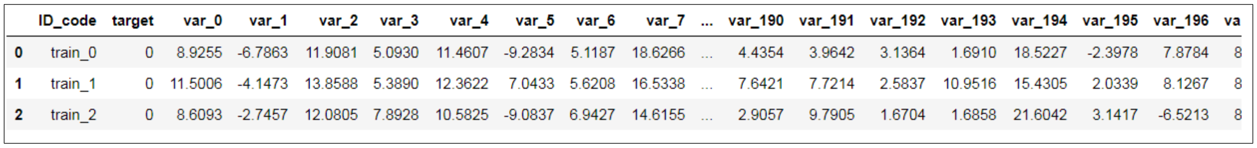
[1] https://www.quantinsti.com/blog/use-decision-trees-machine-learning-predict-stock-movements
[2] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8212859
[3] https://www.econstor.eu/bitstream/10419/183139/1/1032172355.pdf
[4] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1631572

Problem Statement:

This project is going to identify which customers will make a specific transaction in the future. This is a binary prediction. The result will be evaluated in multiple evaluation method. For example, area under the ROC.

Datasets and Inputs:

The dataset is obtained from Kaggle competition [5]. It is an anonymized dataset containing numeric feature variables and the binary target column. For the training and testing dataset, both contain 200000 data with 200 features. Instead of specifying the real meaning of each feature, the dataset only provides the alias name for the features (var_1, var_2, var_3…var_200). The training dataset also provides the target label for each data. There are 179902 data for label 0 and 20098 data or label 1. Here is some example of this dataset:



	ID_code	target	var_0	var_1	var_2	var_3	var_4	var_5	var_6	var_7	...	var_190	var_191	var_192	var_193	var_194	var_195	var_196	va
0	train_0	0	8.9255	-6.7863	11.9081	5.0930	11.4607	-9.2834	5.1187	18.6266	...	4.4354	3.9642	3.1364	1.6910	18.5227	-2.3978	7.8784	8
1	train_1	0	11.5006	-4.1473	13.8588	5.3890	12.3622	7.0433	5.6208	16.5338	...	7.6421	7.7214	2.5837	10.9516	15.4305	2.0339	8.1267	8
2	train_2	0	8.6093	-2.7457	12.0805	7.8928	10.5825	-9.0837	6.9427	14.6155	...	2.9057	9.7905	1.6704	1.6858	21.6042	3.1417	-6.5213	8

[5] https://www.kaggle.com/c/santander-customer-transaction-prediction#description

Solution Statement:

The most common solution to such problem is the method of supervised binary classification. Some of the possible algorithms are SVM with different kernels, Neural Network, Random Forest, and Gradient Boosting, which cover both linear and nonlinear classification. The Gradient Boosting is a machine learning algorithm for regression and classification, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Benchmark Model:

Since this is a Kaggle competition [6], the benchmark would be some of the best Kaggle score currently, which comes in at 0.85 to 0.9 for the ROC result. The classification model includes Light Gradient Boosting (LGB) with Bayesian parameters finding, Neural Network, and XGBoosting. The goal of this project is to optimize on feature engineering and training model with different technique combination and parameters setting. We will compare the result with those benchmark model to see if we are able to perform better with better ROC result. Since I will be using the exact same training dataset and testing dataset as those benchmark models do that provided by Kaggle, the result would be very clear to see if our model makes any difference and performs better.

[6] https://www.kaggle.com/c/santander-customer-transaction-prediction#description

Evaluation Metrics:

Some common evaluation metrics for binary classification is to evaluate on area under the ROC curve between the predicted probability and the observed target.

Project Design:

a. **Data Exploration**: In this stage, it is also called EDA (Exploratory Data Analysis), which means analytically exploring data for finding some hidden information of the data or simply getting more insights on the characteristic of the data. This can be done by visualization of the data.

b. **Statistical Tests:** Performing some statistical tests to confirm some of our hypotheses and get better intuition on how do we process the data for the next step. We will understand the scale, size, the distribution of the data from quantitative perspective.

c. **Data Preprocessing:** In most cases, we have to preprocess the dataset before constructing features and feeding into the classification algorithm. Such as outlier detection, encoding categorical variables if necessary, normalizing the dataset…

d. **Feature engineering:** once we have finalized our features in terms of the reliability from data preprocessing, this step is to work on the feature quality. We might use some feature selection or feature extraction algorithm to find the most significant features.

e. **Model Selection**: Try out different machine learning model to see which one works the best. The Gradient Boost, Random Forest, SVM and Neural Network are the ones I would like to try first.

f. **Model Training**: For each model, there are some hyper parameters we need to finalize by cross validation technique to build the training model.

g. **Testing the model:** Testing the model with testing dataset to ensure the quality of the model.