Machine Learning Engineer Nanodegree Capstone Proposal

Predicting the Directionality of Open price of SSE Index

Domain Background

Statistical models associated with machine learning algorithms are playing increasingly important roles in financial market analyses. In this study, the XGBoost classifier is used to predict daily directionality of the Open price of SSE composite index. An important benefit of XGBoost models is that there are several hyperparameters that can be tuned for improving accuracy. This study evaluates the performance of the XGBoost classifier on forecasting the daily directionality of the open price of SSE.

Problem Statement

The purpose of this project is to apply the XGBoost classifier into the China's stock market to determine if the XGboost classifier can be used to predict the directionality of Open Price of SSE Composite index (up or down). In this project, the label will be defined as "1" if the predicted Open price goes up relative to the Open price in the last day.

Datasets and Inputs

The data included in this study consist of daily directionality of open price of SSE as the output, along with 18 technical indicators as input features. The whole dataset consists of 700 trading days. All data used for this project comes from Yahoo Finance: The data is the historical data from 2000-06-08 to 2020-6-05. And the raw data is included in the file "000001.SS_new.csv". And it contains 6 indicators discussed below.

Date: calendar date for any given data row
Open: opening value (recorded at 9:30) for the SSE
High: highest value on any given day for the SSE
Low: lowest value on any given day for the SSE
Close: closing value (recorded at 16:00) for the SSE
Volume: number of shares of SSE components traded
The raw data consists of 4959 trading days.

Solution Statement

The solution will be predictions of either the Open price of SSE will go up or go down. This study provides a comprehensive study on predicting the directionality of daily Open price of the SSE. It predicts the daily adjusted Open price of the SSE over 2010 to 2015 based on 18 technical indicators as input features. Then, machine learning algorithms, XGBoost Classifier, are chosen for pattern recognition, with 18 technical indicators as input features and the directionality of open price of SSE Index 2010 and 2015 as the output. Also, the validation dataset is used for hyperparameter tuning.

Benchmark Model

The benchmark model for this problem is generally the model that output the guessing direction of the open price of SSE. In the testing data we introduced later, if all the guess is "goes up" (label as 1), the accuracy is 58%. And the XGBoost Classifier is built to beat its performance.

Evaluation Metrics

There are basically two metrics used to evaluate the performance of the model. One of them are Accuracy Score, define by Accuracy Score = (True Positives + True Negatives) / Total Population. Another one is Net Profit generated by trading simulation. The reason for choosing this metric is that sometimes although the accuracy score is high, the trading outcome doesn't suit people's expectation. Therefore, the trading simulation is employed. Because China's stock market doesn't allow shorting position and require all investors to hold the stock at least one trading days, long position will be created if the predicted open price of SSE goes up.

Project Design

Before training the model, I will do some statistics analysis of the dataset and perform ordinary data analysis. Then I will generate the 18 features based on the raw data for training the model. To train the model, XGboost classifier will be used. And then, the validation dataset will be used to perform hyperparameter tuning. After the best model is generated, I will use it to predict the label in the testing set.

I expect to spend 40% time to gather and clean the data and spend 60% of time working on the model and calculates the testing result. The final accuracy score will be then calculated based on the predictions.