Capstone Project Proposal

Prepared for: Udacity Machine Learning Nano Degree Program

Prepared by: Sezer Bakay

4 June 2018

PREDICTING STOCK YIELD

Domain Background

Shares of companies provide partnership for the company's profits to its holder. For a shareholder, eventual aim is to yield a targeted return than its investment value. In 2017, 77.6 Trillion USD worth of shares were traded worldwide (The World Bank, World Federation of Exchanges Database). Investors, individuals, companies buy shares of publicly held companies with the expectation of increase in its worth and sell them with a expectation of decrease in its worth or to monetise its value. One profits if the value of the stock increases and exceeds its cost or make loss if not so. While it is easy to define the outcome of situations, the main challenge is to make right trading decisions which satisfy the targeted returns. In order to make profit from stock trading, investors have to give the right trading decisions.

Problem Statement

The aim is to make the right trading decision that will satisfy the expected return. The question is, how these decisions will be made. Investors benefit from various approaches, strategies, tools and techniques to build a solid foundation for their decisions. Due to the indeterministic structure of the stock market, there is no optimal solution for predicting the behaviour of stock prices. Depending on the trading strategy, some investors benefit from fundamental analysis of companies, some benefit from technical analysis of the historical prices and some benefit from both.

In this project, we will deal with technical analysis approach in order to predict the expected return of a stock share. Technical analysis relies on the acceptance that the price of a stock contains the required information to predict its future movement. It analyses the historical price and volume data of the stock, calculates various arithmetic equations and interpret the price behaviour of the stock. If we take account only the most used technical analysis techniques, there are still tens of different calculations, formulas and patterns to interpret the stock. Unfortunately, all of them are valid under some circumstances and we can't claim an optimal matching. It is an intensive and complex data analysis problem which can be portrayed as walking through a swamp. If you pick the right route, you will reach to your goal. Otherwise, you will end up with losing your money.

Datasets and Inputs

In this project, we are going to use the historical price and volume data of THYAO (Turkish Airlines) stock share which is traded at Istanbul Stock Exchange (ISE). Dataset contains daily data for the features from May 2013 to May 2018. The motivation for picking THYAO to practice is that, THYAO is one of the most traded stock in ISE and due to its trade volume, it is least affected from manipulative price changes.

Technical analysis calculations are done by using the functions from TA-Lib library (http://ta-lib.org/).
(Github repository for TA-Lib: https://github.com/mrjbq7/ta-lib)

Feature	Definition	Source
Open	Opening price of the day	Yahoo Finance
High	Highest price in the day	Yahoo Finance
Low	Lowest price in the day	Yahoo Finance
Close	Closing price of the day	Yahoo Finance
Volume	Volume of the trade during the day	Yahoo Finance
APO	Absolute Price Oscillator	Calculated by using TA-Lib Library
AROONOSC	Aroon Oscillator	Calculated by using TA-Lib Library
MACD	Moving Average Convergence/ Divergence	Calculated by using TA-Lib Library
MACD Sig	MACD Signal	Calculated by using TA-Lib Library
MACD Hist	MACD Histogram	Calculated by using TA-Lib Library
Momentum	Momentum	Calculated by using TA-Lib Library
RSI	Relative Strength Index	Calculated by using TA-Lib Library
SLowK	Stochastic Slow K	Calculated by using TA-Lib Library
SLowD	Stochastic Slow D	Calculated by using TA-Lib Library
Williams	Wlliams' %R	Calculated by using TA-Lib Library
UpBand	Bollinger Upper Band	Calculated by using TA-Lib Library
MidBand	Bollinger Middle Band	Calculated by using TA-Lib Library
LowBand	Bollinger Lower Band	Calculated by using TA-Lib Library
P-SAR	Parabolic SAR	Calculated by using TA-Lib Library
WMA	Weighted Moving Average	Calculated by using TA-Lib Library
Chaikin	Chaikin A/D Oscillator	Calculated by using TA-Lib Library
MA28	28 days Moving Average	Calculated by using TA-Lib Library
Hilbert	Hilbert Transform - Trend vs Cycle Mode	Calculated by using TA-Lib Library

Target Variable:

Target variable is calculated according to the upcoming 10 days of the stock compared to its current price. If the price of the stock exceeds 3% of today's price in any day during this 10 days window, target variable is labeled as 1. Otherwise it is labeled as 0.

Solution Statement

Aim of the project is to predict the yield of a stock's price in determined period (10 days). Therefore, our model needs to classify the stock if its price will increase more than 3% in 10 days or not. After classifying the future behaviour, our trading algorithm will decide to buy the stock or not. In order to do this classification, supervised learning will be used. The model will be trained by using the features that contain the price information, volume information and technical analysis calculations of the stock with corresponding target variable. For this project, THYAO share will be used to train and test our model. As a performance measure, recall of the test set will be used to measure the performance of the model (false negatives won't cause the investor to lose money but false positives that give you wrong buy signals will cause losing money). In order to measure the performance of the trading algorithm, a simulation will be run with test set and the profit will be measured.

Benchmark Model

Our benchmark will be the natural performance of the stock. With our trading algorithm based on the trained classification model, the aim will be to achieve more profit than the natural performance of the stock during the simulation period.

Evaluation Metrics

F0-Score (Recall): Recall score will be used to evaluate our classification model. The reason for choosing recall than accuracy is, while false negatives won't cause the trader lose money but false positives will cause so due to generating wrong buy signals.

Profit Ratio Against the Natural Performance of the Stock: To evaluate our algorithm, a trading simulation will be made during the period of test dataset. Generated profit by using the algorithm will be compared to the natural price change (if the trader buys the stock at first day and sells it at the last day) of the share.

Project Design

PART 1- Preparing the dataset:

- Data in a csv file, which is retrieved from Yahoo Finance, will be imported.
- Rows that contain N/A values will be dropped.
- Features for technical analysis will be calculated by using TA-Lib library.
- Target variable will be calculated according to the pre-determined price change ratio in a window of upcoming 10 days.
- Rows that contain N/A values will be dropped and index will be reset.
- Dataset will be split into training and testing sets in a ratio of 80/20.

PART 2 - Pre-processing the data:

- Datasets will be scaled by using StandardScaler function from Sklearn Preprocessing Library. The scaler will be fit to test data and applied to both datasets.
- Due to that we have included various technical analysis values without applying a solid filter, Principal Component Analysis (PCA) will be applied for dimensionality reduction in order to increase the performance of the model. Sklearn library will be used for PCA.

PART 3- Training the model and testing

- The dataset will be trained with different supervised learning classifiers (K-NN, SVM, Random Forrest Classifier, Adaboost, ANN) and their performances will be compared to each-other. After deciding the classifier, fine-tuning will be made by using Grid-Search technique. Keras will be used for ANN and Sklearn libraries will be used for the rest.
- Test data will be predicted with the trained model and recall of the result will be evaluated.

PART 4 - Trading Algorithm:

- A deterministic trading algorithm will be created which uses predicted target variable for to decide buying the stock and use a targeted yield or pre-defined stop-loss level for selling the stock.
- After running the simulation for the test dataset, the profit generated by the algorithm will be compared to the natural price performance of the stock.