**Stock Market Prediction**

**Introduction**

Stock market forecasting is a very important area in the financial sector. There has been a significant interest shown in this subject by academic researchers, investors and practitioners to aid in decision making. According to Bagheri et al., (2014), the ability to correctly predict future market trends is a prerequisite for successful financial market trading.

In the past stock selection relied heavily on the investors (or advisors) personal knowledge on the industry/company. Prediction methodologies were broadly classified into two categories. They are fundamental analysis and technical analysis. Fundamental analysis attempts to estimate a company’s intrinsic value by analysing the company’s financial statements. On the other hand, technical analysis analyses statistical trends gathered from past price movement and volumes. In the recent past, there has been a significant interest in using AI techniques to predict stock markets by examining patterns in massive amount of real time equity and economic data

**Literature Review**

Many researches have been conducted in the past to predict stock performance in order to generate profitable trading opportunities. There are numerous literatures on this subject that have been published and these were consulted to ensure the project has a good foundation by understanding the strengths and weakness of these researches. The analysis of research papers referenced are outlined below. The reference section at the end of this document contains full citations

[1] In the research “*Predicting Stock Prices Using Technical Analysis and Machine Learning*” submitted on June-2010, the author Jan Ivar Larsen(Norwegian University of Science and Technology)

The author applies a knowledge-intensive first layer of reasoning based on technical analysis before applying a second layer of reasoning based on machine learning. The first layer of reasoning thus performs a coarse-grained analysis of the price data that is subsequently forwarded to the second layer of reasoning for further analysis. For both layers the author has used decision trees extensively. The authors concluded that the developed prediction model using domain knowledge, machine learning and a money management strategy can create substantial profits when evaluated on the Oslo Stock Exchange. The two-layer reasoning architecture is also deemed successful due to it’s adaptivity and easy integration with new techniques. It was noted that the biggest drawback of this model is omission of transaction costs and the presence of inherent stochasticity in the model. Future works is required to include fundamental and technical analysis in the model.

[2] In the research “*Predicting the direction of stock market prices using Random Forest*”( ***(***[***arXiv:1605.00003***](https://arxiv.org/abs/1605.00003)*[cs.LG])* )submitted on May-2016, Luckyson Khaidem,Snehanshu Saha and Sudeepa Roy Dey

The authors explore Random Forest classifier to predict direction of stock prices to exploit ensemble learning methods. The research has claimed to have impressive results and the robustness were evaluated using various parameters like accuracy,recall,precision and specificity. The research has concluded that the model is able to achieve accuracy in the range of 85-95% for long term prediction.

[3] In the research *“Equity forecast: Predicting long term stock price movement using machine learning”*([arXiv:1603.00751](https://arxiv.org/abs/1603.00751)[cs.LG]) Nikola Milosevic

The author presents a machine learning aided methodology to predict long time equity movement. The author trained his models using C4.5 decision trees, Support Vector Machines with Sequential Minimal Optimization, JRip, Random Trees, Random Forest, Logistic regression, Naïve Bayes and Bayesian Networks. Firstly, he performed 10-fold cross validation on all these algorithms with all indicators and history price used as features. Then he performed manual feature selection by removing features and evaluating whether performance of the algorithm improved or decreased. He performed this process iteratively, until we didn’t get the optimal model with minimal number of features and the best performance. The results of his experiment indicate that algorithm that performed best was Random Forest. The limitation of this study is that the models are not created out of data that were not limited in time. More accurate way of generating financial machine learning models would be to limit training data until certain year, while test it on unseen future data. In this case, that has not been done and whole dataset is split into training and testing folds and tested against. Markets may change over time, and therefore n-fold cross-validation evaluation would not capture these changes and may introduce, what is in finance called “Look-Ahead Bias”.

[4] In the paper “*Which Artificial Intelligence Algorithm Better Predicts the Chinese Stock Market*”L. Chen, Z. Qiao, M. Wang, C. Wang, R. Du and H. E. Stanley, "Which Artificial Intelligence Algorithm Better Predicts the Chinese Stock Market?" in IEEE Access, vol. 6, pp. 48625-48633, 2018, doi: 10.1109/ACCESS.2018.2859809

The authors have used Deep Learning techniques to predict stock market behaviour and compare the performance of this approach with the performance of traditional Back Propagation network (BP), Extreme Learning Machines (ELM), Radial Basis Function (RBF). To show the effect of sample volume on network training and predicting, they divided the sample into three scale datasets and compare the stock price prediction of deep learning with three traditional artificial neural networks (BP, ELM, RBF). They compared their predictive fitting degree and directional predictive accuracy and proved that the predictive performance of deep learning is superior to that of BP, ELM and RBF. They also found that sample volume strongly affects stock prediction, and that deep learning performs well when applied to large data. Also, deep learning does not need prior predictive information to extract features from large datasets, and this increases its usefulness in predicting stock market behavior. They concluded that there is enough evidence to prove that DL is an effective method of predicting stock price

[5] In the research, *“PREDICTING AND BEATING THE STOCK MARKET WITH MACHINE LEARNING AND TECHNICAL ANALYSIS*” Anthony Macchiarulo.

The author conducts a short comparisonbetween machine learning and technical analysis to predict stocks market is provided. The machine learning algorithms used by the authors are Support vector Machines, Neural Network and Ensemble Learning. The authors have concluded that using machine learning as a trading strategy can positively impact the returns generated compared to using many technical indicators. It was concluded that in up markets ML outperforms technical analysis whereas in down markets technical analysis outperforms ML. However, the down market had very few observations (<50) so the results may not be usable, and more data will be required.

**Dataset**

The source for this dataset is Kaggle New York Stock Exchange data (<https://www.kaggle.com/dgawlik/nyse>). This dataset is primarily intended for Fundamental and Technical analysis.

There are four files in the dataset: prices.csv, price-split-adjusted.csv, securities.csv and fundamentals.csv.

**Prices.csv:** raw, as-is daily prices. Most of data spans from 2010 to the end 2016, for companies new on stock market date range is shorter. There have been approx. 140 stock splits in that time, this set doesn’t account for that. This file contains 7 columns and around 851k rows. Each stock has approximately 1800 rows of data, which will be filtered for our analysis. The attributes in this file are Date, symbol, open, close, high, low, volume. There are price data for 501 unique symbols.

**Table 1:**

|  |  |  |
| --- | --- | --- |
| S.no | Attribute | Description |
| 1 | Date | Stock trading date  Data Type: Datetime |
| 2. | Symbol | Ticker identifier for the stock of the company  Date Type: string |
| 3. | Open | NYSE opening trade price  Date Type: float |
| 4. | Close | NYSE closing trade price  Date Type: float |
| 5. | High | NYSE day highest trade price  Date Type: float |
| 6. | Low | NYSE day lowest trade price  Date Type: float |
| 7. | Volume | Total No. of the shares traded for the day  Date Type: Integer |

**Securities.csv:** general description of each company with division on sectors. This file contains 8 columns and around 505 rows. The attributes in this file are Ticker symbol, security, SEC filings, GICS sector, GICS Sub Industry, Address of Headquarters, Date first added, CIK. There are securities data for 504 unique securities. (Note: Highlighted columns are removed as part of data cleaning)

**Table 2:**

|  |  |  |
| --- | --- | --- |
| S.no | Attribute | Description |
| 1. | Ticker symbol | Ticker identifier for the stock of the company  Date Type: String |
| 2. | Security | Company name  Data Type: String |
| 3. | SEC filings | Financial statement filing for the company  Data Type: String |
| 4. | GICS Sector | Sector of the company based on MSCI & S&P classification  Date Type: String |
| 5. | GICS Sub Industry | Sub-Industry of the company’s sector  Data Type: String |
| 6. | Address of Headquarters | Address of the company  Data Type: String |
| 7. | Date first added | First filing date in SEC  Data Type: Date |
| 8. | CIK | Unique key to identify a corporation in SEC data base  Data Type: Integer |

*Note*: For the purpose of the project we will not use price-split-adjusted.csv but use the non adjusted prices in the prices.csv. We will also not use fundamentals.csv as we are not incorporating Fundamental analysis in our model and limit our scope to technical analysis.

**Approach:**

The files contain historical data for about 500 securities. For our illustrative purpose we will use a single stock (MSFT) but the program will be customisable so users can run the data preprocessing and model on any stock by passing a parameter. The following diagram outlines the approach we will use in our project

**Data Collection:**

Participants in the stock market would need to get a sense of the direction in which a stock will move before trading on it. The New York Stock Exchange data from Kaggle(<https://www.kaggle.com/dgawlik/nyse> )has been downloaded in order to build a model for stock market prediction.

**Research Question:**

The motivation for this project is to help users make better decisions when trading stocks by building a machine learning model for predicting the stock market and Classifying top performing industries. So, the research question for this project is How investors can use Machine learning techniques to analyse the stock market’s trend?

**Data Cleaning:**

* There were no missing values in the Prices.csv file.
* After new features were added, we removed all rows which has nan values for feature extraction.
* From the Securities.csv we removed SEC filling, Address of head quarters, Date first added, CIK (Data in these columns will not have any impact on the model)
* The data type for Date column is loaded into the Pandas data frame as objects by default. For this data to be compatible, we convert the data type to Datetime. Please refer Table 1 & 2 for updated data types.
* Outliers for the data were analysed for the attributes in the dataset and visualised using line chart (for closing price and volume). Stock market is heavily influenced by external factors such as economical factors, sector specific changes, market sentiments and various other factors. As a result, no outliers were removed but only observed.

**Data Preparation:**

* The following features are extracted by performing simple transformations on the price and volume attributes. These features are some of the important ones used by industry practitioners for technical analysis. For better understanding, I have placed them into subcategories below.
* After the features were extracted, all rows (about 60 rows) which had NAN values were removed to improve model efficiency. The NAN values were expected for the first few time periods as some of the columns calculated moving average which will be missing for the first few time periods. For eg: the 60d volatility will be missing for the first 60 rows.
* Once the features are extracted, Correlation matrix between all the features (including the below) has been prepared.

1. **Return features:**

* **1-day return**: Ratio of today’s close to yesterday’s close for a stock
* **close\_to\_open:** Ratio of close to open for a day
* **close\_to\_high:** Ratio of close to high for a day
* **close\_to\_low:** Ratio of close to low for a day

1. **Trend indicator Features:**

* **Ma\_50day:** Moving averages are used to smooth price data by calculating an average over a specific period. For our purpose we calculate 50 days simple moving average for each stock on a day
* **MACD\_diff:** Moving average convergence difference is an indicator that shows the relationship between two moving averages of a security’s price. It is calculated as the difference between 26 period exponentials moving average (EMA) and 12 period EMA.

1. **Momentum Indicator features:**

* **Stochastic\_Oscillator:** Stochastic Oscillator is a momentum indicator that shows the location of the close relative to the high-low range over a set number of periods.
* **CCI:** The commodity channel indicator(cci) is a versatile indicator that can be used to identify a new trend or warn of extreme conditions. In general, CCI measures the current price level relative to an average price level over a given period of time.
* **RSI:** Relative Strength Index (RSI) is an extremely popular momentum oscillator that measures the speed and change of price movements. RSI oscillates between zero and 100.

1. **Volatility Indicator features:**

* **5d\_volatility:** Calculated as standard deviation for 5-day(rolling) returns.
* **21d\_volatility:** Calculated as standard deviation for 21-day(rolling) returns.
* **60d\_volatility:** Calculated as standard deviation for 60-day(rolling) returns.
* **Bollinger\_bands:** Bollinger Bands® are volatility bands placed above, middle and below a moving average. Volatility is based on the standard deviation which changes as volatility increases and decreases. The bands automatically widen when volatility increases and contract when volatility decreases. Their dynamic nature allows them to be used on different securities with the standard settings.

1. **Volume features:**

* **On\_Balance\_Volume:** On Balance Volume (OBV) measures buying and selling pressure as a cumulative indicator, adding volume on up days and subtracting it on down days.

**Data Modelling and Evaluation**

The model will conceptually have two modules:

1. Technical analysis module: In this module technical indicators described above will be computed. The output from this module will feed the ML module (see below)
2. Machine learning module: This module will apply machine learning techniques for the data generated from the above module.

* We will use simple sampling technique for the data in order to avoid biases introduced by randomised sampling.
* 3 types of algorithms will be used on the resampled dataset to predict the stock market trend (Up or down).
* The model will be evaluated using split validation method

**Conclusions and Future Work**

Conclusion and Future recommendations will be provided based on the outcome of the research

**References**

[1] Jan Ivar Larsen *,”Predicting Stock Prices Using Technical Analysis and Machine Learning”*

[2] Luckyson Khaidem,Snehanshu Saha and Sudeepa Roy Dey, “*Predicting the direction of stock market prices using Random Forest”*

[3] Nikola Milosevic, “*Equity forecast: Predicting long term stock price movement using machine learning”*

[4] L. Chen, Z. Qiao, M. Wang, C. Wang, R. Du and H. E. Stanley, “*Which Artificial Intelligence Algorithm Better Predicts the Chinese Stock Market*”

[5] Anthony Macchiarulo,*“PREDICTING AND BEATING THE STOCK MARKET WITH MACHINE LEARNING AND TECHNICAL ANALYSIS”*