Department of Informatics,

University of Leicester

CO3015 Computer Science Project

Interim Report

for

Stock Market Prediction

using Machine Learning

Author: Osato Osagie (1690 11647)

Date of submission: 21/11/2020

Table of Contents

[Declaration 3](#_Toc56894600)

[1. Introduction 4](#_Toc56894601)

[1.1 Aims 5](#_Toc56894602)

[1.2 Objectives 5](#_Toc56894603)

[2. Survey of Literature/Information Sources 6](#_Toc56894604)

[2.1 Programming Language 7](#_Toc56894605)

[2.2 Jupyter Notebook 7](#_Toc56894606)

[2.3 Alpha Vantage Stock API 7](#_Toc56894607)

[3. Requirements 8](#_Toc56894608)

[3.1 Functional Requirements 8](#_Toc56894609)

[3.2 Non-functional Requirements 8](#_Toc56894610)

[4. Outline of Specification and Design 9](#_Toc56894611)

[4.1 Metrics to evaluate ML Algorithm 11](#_Toc56894612)

[4.1.1 RMSE 11](#_Toc56894613)

[4.1.2 Confusion Matrix 11](#_Toc56894614)

[4.1.3 Confusion Matrix 13](#_Toc56894615)

[5. Planning and Timescales 15](#_Toc56894616)

[6. References 16](#_Toc56894617)

# A screenshot of a cell phone Description automatically generatedDeclaration

# Introduction

Artificial Intelligence (AI) is the ability of machines to behave like humans and learn autonomously. For instance, a machine might display learning and problem-solving abilities without the use of hard-coded software containing detailed instructions [1].

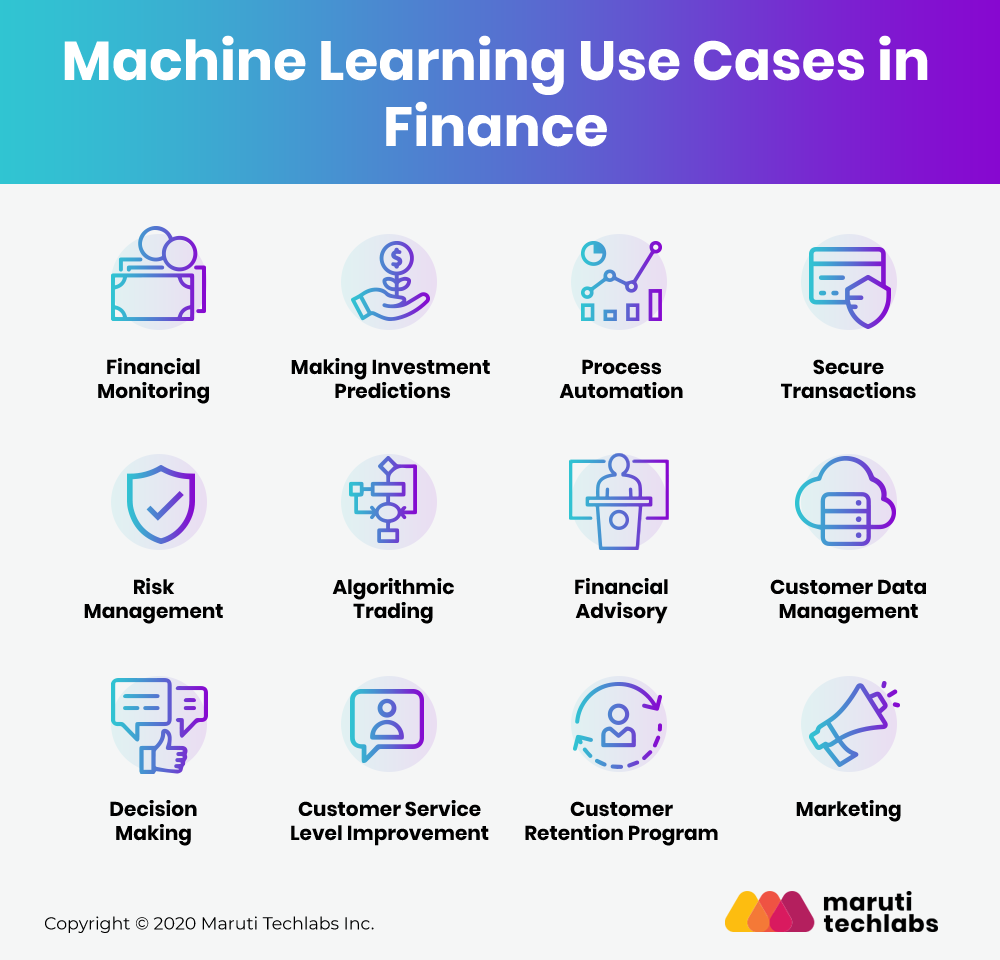


Figure 1- An image showing the main use cases of Machine Learning in Finance [2]

Machine Learning (ML) is a subfield of AI, that enables machines to learn from historic data or experiences without being explicitly programmed. Figure 1 shows the different use cases for Machine Learning in finance. The project I will be building focuses on using ML to make investment predictions. Using ML to make investment predictions is advantageous as it can lead to better predictions of stock prices, fewer errors, and greater efficiency for the investor. To do this, ML algorithms extract key insights from the dataset, learns from it then apply several techniques to accurately predict the result.

However, it is important to take into consideration the other factors that might affect the price of a company’s stock. The stock market is very volatile, thus meaning no system can accurately predict it.

## 1.1 Aims

Predicting markets has become an increasing priority for investors. The primary goal of an investor is to buy a stock when its value is low and sell when the value of the stock is high. However, this can be daunting for financial investors as they are unaware of the stocks that will return maximum profits. Using Machine learning to predict the long-term value of a stock makes this process somewhat easier. This project aims to predict the stock market price of a company using supervised machine learning algorithms.

## 

## 1.2 Objectives

The models will be implemented in python using the Alpha Vantage API for real-time and historical equity data. The models built in this project need to be able to predict the price of a company’s stock by using historical data. Also, the project must be able to produce visualizations that illustrate patterns in price movement.

In this proposed project, I will be implementing two prediction systems; one based on Decision Trees (DT) and the latter using Support Vector Machines (SVM). Doing this will enable me to compare the forecasting ability of both systems. Each will be evaluated and tested with the same data to determine their prediction accuracy.

There have been several other approaches to this problem. For example, Lei in [3] exploited the Wavelet Neural Network (WNN) to predict stock price trends. However, Neural networks (NN) have several drawbacks. One drawback of using NN is that it has a slow conversion rate, thus meaning that it can be time-consuming to train.

NN also uses an optimization algorithm for finding local optima called Gradient descent. However, this often gets suck in the local maxima, thus making it a challenge to find global minima and maxima.

Another approach to this problem was Piramuthu’s [4]. A study was conducted which thoroughly evaluated the different methods of feature selection for data mining applications. However, this study can be criticized because it only used decision trees. This, therefore, makes it impossible to trust the reliability of the results, because we are unsure if we can replicate the result of this model on a more complex model.

My approach to this problem will be different as it uses a combination of SVM and Decision Trees. Decision Trees offer several benefits over using Neural Networks. For example, they perform faster than Neural Networks after training and provide a visual representation of the data. SVM also offers its benefits over using NN. Unlike NN it is guaranteed to find a global optimum and less memory is required to store the predictive model. By comparing the results of both models, I will also be able to increase the validity of my results.

# Survey of Literature/Information Sources

To get better results I have decided to use Technical Indicators. Technical indicators are mathematical calculations based on the price, volume, or open interest of a security [5]. Whilst these indicators are designed to analyze short-term price movements, they are also useful to long-term investors who want to identify entry and exit points.

According to Investopedia [5], the following are some of the most commonly used technical indicators used by traders:

1. **Simple Moving Average (SMA)-** This is useful for eliminating noise and identify trends. It creates a regularly updated average price which smooths out price data. The SMA can be taken in various time frames i.e., 10 mins or 52 weeks. It applies an equal weight to all observations in a period, thus making it suitable for both long- and short-term investors.
2. **Exponential Moving Average (EMA)-** this technical indicator is similar to the SMA; however, it places more emphasis on the most recent data points, which means it will react more than a simple moving average to recent price changes.
3. **Moving Average Convergence Divergence (MACD)-** MACD describes the relationship between the two aforementioned moving averages, it helps the investors to understand whether the price movement is strengthening or weakening.

The periods for the MACD indicator are presented as 26 and 12. The MACD is calculated by taking away the 26-day EMA from the 12-day EMA. a ‘signal line’ represents a 9-day EMA of the MACD. It acts as a trigger for buy and sell signals and is plotted on top of the MACD.

1. **Stochastic Oscillator (SO)**- Stochastic Oscillator gives investors an insight into overbought and oversold signals. It is also referred to as a momentum indicator that compares a particular closing price of a security to a range of its prices over a certain period [5]. By taking the Moving Average of the result or by altering the period, we can reduce its sensitivity to market changes.
2. **Accumulation/Distribution Line (A/D)**- A/D uses volume and price to determine whether a stock is being accumulated or distributed. This indicator aims to spot divergences between the stock price and volume flow, thus providing insight into how strong a trend is. For example, if the price of a stock is on the rise but the indicator is falling this suggests that buying volume might not be sufficient to support the price rise and a price decline is forthcoming.
3. **Bollinger Bands (BB)**- Bollinger bands are used to identify if an asset is overbought or oversold. Therefore, this helps the trader by signaling changes in volatility. It consists of three bands: the middle band (a simple moving average of the typical price), the lower band, and the upper band (standard deviations above and below the middle band). These bands increase and shrink when the volatility of the stock price is higher or lower.
4. **On Balance Volume (OBV)**- OBV is a technical indicator of momentum, it uses volume flow to predict changes in stock price. For example, when the closing price of a stock is higher than the previous closing price, the volume of the stock is added to the running total. On the other hand, when the closing price of a stock is lower than the previous closing price, the volume is taken away from the running total.

## 2.1 Programming Language

For statistical computing, the two widely used programming languages used in the industry are Python and R. Between both, I have chosen to use python because it is faster and scalable. It is also one of the most supported languages nowadays, supported by many robust libraries which are constantly growing. Also, python has a diverse pack of visualization options available which makes it ideal for creating graphs and charts. To learn python and understand machine learning, I have taken a course on Udemy [6].

## 2.2 Jupyter Notebook

The programming environment I am using for this project is Jupyter notebook [7] is the programming environment I will be using in this project. It is a web application that allows users to create documents containing live code and visualizations.

To access Jupyter Notebook, I use the Jupyter Notebook App. This is a server-client application that allows editing and running of a notebook document through a web browser. This app has a dashboard that emulates the features of a file manager, for example, a user can rename and delete files. Also, it has a notebook kernel that acts as the computer engine. It executes the code stored in the notebook documents.

## 2.3 Alpha Vantage Stock API

Alpha Vantage Stock API [8] provides free real-time stock quotes, historical data, cryptocurrencies, technical indicators, FX rates, and more. This is the source I will be relying on to get my data to build the models. I will be collecting the last 2 years of a company’s stock data.

# Requirements

For this project I have split my requirements into two types:

1. Functional requirements- these types of requirements define the different components of the system.
2. Non- functional requirements- these requirements specify the quality attribute of the system. For example, how accurately a model predicts the price of a stock.

## 3.1 Functional Requirements

* The system must be able to Obtain 2 years of a company’s stock data from the Alpha Vantage Stock API.
* The system must be able to create functions that calculate the value of the aforementioned technical indicators.
* Split data into two parts, training and testing. 75% of the data will be allocated for training and 25% of the data will be used for testing.
* The system must be able to train the model using the selected training data
* The system must be able to test the model using testing data.
* The system must be able to create visualization of the market index.
* The system must be able to build a stock prediction model using SVM.
* The system must be able to build a stock prediction model using decision trees**.**

## 3.2 Non-functional Requirements

* **Accuracy & performance** 
  + The system must be correct when compared to reality.
  + The system must maintain the same performance on multiple datasets.
* **Fairness**
  + Sensitive features (e.g., race, gender) must be removed from the dataset to ensure fair results**.**
* **Transparency**
  + The steps taken to obtain results must be clearly outlined.

# Outline of Specification and Design

A picture containing graphical user interface

Description automatically generated

Fig. 2. A simplified breakdown of the project design.

In figure 2 we use layers to separate different ML concepts. The available algorithms are decomposed by the type of problem, then by algorithm characteristics, and then by the type of algorithm. Although this diagram is incomplete, it gives a high-level overview of the architecture of the system.

## 4.1 Support Vector Machine Algorithm (SVM)

## Diagram Description automatically generated

Fig. 3. SVM Classifier [11]

SVM is one of the two machine learning algorithms used in this project. It is a classification model for classification and regression problems. This algorithm is capable of solving linear and non-linear problems and works well with many practical problems such as the one we aim to solve in this project. The Idea of SVM is straightforward: The algorithm creates a line or a hyperplane that separates the data into classes [11].

SVM then finds the closes points to the line from each class. These points are known as Support Vectors. Then it calculates the distance between the line and the Support Vectors, this is known as the margin. SVM’s goal is to maximise the margin. The hyperplane that has the maximum margin is the optimal hyperplane.

## 4.2 Decision Tree (DT)Diagram Description automatically generated

Fig.4. Sample illustration of a Decision Tree [12]

The second machine learning algorithm used in this project is Decision Tree. It is used for modelling both regression and classification problems. DT makes sequential and hierarchical decisions regarding the target variable based on the predictor data.

A Decision tree model behaves like a protocol in a series of if/else conditions that produce a specific result from input data. The decision tree aims to make the optimal choice at the end of each node. [13]

A picture containing graphical user interface

Description automatically generated

Fig. 5. Program flowchart

In figure 5 we can see a flowchart of how my program executes. Firstly, real-time data is obtained from the Alpha Vantage API. Putting the stock’s ticker symbol into Alpha Vantage API will allow me to obtain the last 2-3 years of data regarding the stock price of any company. I will only consider companies that are part of the Standard and Poor’s 500 index [9]. This is a stock index is made up of 505 common stocks issued by 500 large-cap companies and traded on American Stock exchanges. This covers approximately 80 percent of the American stock market by capitalisation.

Secondly, the program will calculate the dependant variables, which are the aforementioned technical indicators (SMA, EMA, MACD, SO, A/D, BB, OBV). The next step in the program is to create a dummy variable. This variable will be used as an indicator of the price change for each day, it will be obtained by subtracting the closing price from the opening price. If the outcome of this subtraction is more than 0, then ‘Up’ will be stored in the dummy variable, else ‘DOWN’.

Furthermore, a new input data frame will be created containing just the dummy variable column and the technical indicators. I will then divide this data into two parts. Part one will be the data used to build the model (training set); this will comprise 75% of the data. The remaining 25% will be used for testing the model (test set). The test set will be used to assess the likely performance of a model. For example, if a model fits the training set better than the test set, then this is a sign of overfitting. The output of the data is then displayed, and the program ends.

Diagram

Description automatically generated

Fig. 6. Root Mean Squared Error (RMS) Error formula

## 4.1 Metrics to evaluate ML Algorithm

### 4.1.1 RMSE

The accuracy of our prediction model refers to the fraction of predictions our model got right. To measure this, I will be using a metric called Root Mean Squared Error (RMSE). RMSE measures the average error performed by the model in predicting the outcome for an observation.

RMSE is measured by first determining the residuals. Residuals refers to the difference between the actual values and predicted values. RMSE can be either negative or positive because the value predicted by our model can either be under or over estimated. A lower RMSE indicates a better model. RMSE is calculated as with the following formula: sqrt (Mean ((observed(s) – predicted(s)) ^2)).

### 4.1.2 Confusion Matrix

Another evaluation metric I will be using in this project is a Confusion Matrix. Its output is a matrix which is used to describe the complete performance of a model. For example, in a binary classification problem with classes belonging to UP or DOWN concerning the stock market price of a company. Also, we have a classifier that predicts a class for a given input sample. On testing the model with 165 samples, we get the following result.

|  |  |  |
| --- | --- | --- |
| **n=165** | **Predicted:**  **DOWN** | **Predicted:**  **UP** |
| **Actual:**  **DOWN** | 50 | 10 |
| **Actual:**  **UP** | 5 | 100 |

When using classification matrix, there are four terms we must be familiar with:

Fig. 7. Classification Matrix

* True positives- these are the positive predictions made by our model, were the output was positive.
* True Negatives- these cases were predicted to be negative by our model and the actual output was negative.
* False Positives- these cases were predicted to be negative by our model, but the actual output was negative.
* False Negatives- these cases were predicted to be negative by our model, but the actual output was positive.

Table

Description automatically generated

Fig. 8. A formula for calculating the accuracy of a model using the Classification Matrix

Figure 8 shows how we calculate the accuracy of a model using the Classification Matrix. It is calculated by taking the average of the values lying across the “main diagonal”.

Chart, line chart

Description automatically generated

Fig. 9. Area Under Curve [10]

### 4.1.3 Confusion Matrix

Area Under Curve (AUC) is a widely used metric for evaluation. I will also be making use of this in my project. It is used for binary classification problems. Two basic terms need to be understood:

* True Positive Rate (Sensitivity): this refers to the number of cases that were correctly predicted positive out of all positive data points. It is calculated using the following formula shown in figure 10.

A picture containing text

Description automatically generated

Fig. 10. A formula to calculate True Positive Rate [10].

* True Negative Rate (Specificity): this refers to the number of cases that were correctly considered negative out of all the negative data points. It is calculated using the formula shown in figure 11.

Text, letter

Description automatically generated

Fig. 11. A formula to calculate True Negative Rate [10].

* False Positive Rate: This refers to the number of negative data points that are mistakenly considered as positive, concerning all negative data points. The formula for calculating the False Positive Rate is as follows.

A picture containing text, letter

Description automatically generated

Fig. 12. A formula to calculate False Positive Rate [10].

As shown in figure 9, AUC refers to the area underneath the curve of

is the area under the curve of plot False Positive Rate vs True Positive Rate at different points in [0,1]. AUC has a range of [0,1]. A higher value indicates a more accurate prediction our model [10].

# A screenshot of a computer screen Description automatically generatedPlanning and Timescales

Fig. 13. A Gantt Chart showing my work schedule for semester 1

A screenshot of a computer screen

Description automatically generated

As shown in figure 13, semester 1 of this project has mostly been focused on research. Due to the nature of this project and the fact that I am not familiar with Python, I made it a priority to fully understand Machine Learning and python before I started building the project.

Fig. 14. A Gantt Chart showing my work schedule for semester 2

The harder tasks in this project will be encountered in semester 2 as shown in figure 14. I have created a task to add additional features to my project if time permits, this may be building a 3rd neural network model, to fully test the reliability of my predictions.

# References

[1] Deloitte.com (-) Artificial intelligence in investment management [ONLINE]. Available at: <https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/fsi-artificial-intelligence-investment-mgmt.pdf>

[2] Marutitech.com (-) 12 Use Cases of AI and Machine Learning In Finance [ONLINE]. Available at: <https://marutitech.com/ai-and-ml-in-finance/>

[3] Science Direct (-) Wavelet Neural Network Prediction Method of Stock Price Trend Based on Rough set Attribute Reduction [ONLINE]. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S1568494617305689>

[4] Science Direct (-) Evaluating Feature Selection Methods for Machine Learning in data mining applications [ONLINE]. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0377221702009116?via%3Dihub>

[5] Investopedia (2020) 7 Technical Indicators to Build a Trading Toolkit [ONLINE]. Available at: <https://www.investopedia.com/top-7-technical-analysis-tools-4773275>

[6] Udemy () Machine Learning A-Z: Hands-On Python & R [ONLINE] Available at <https://www.udemy.com/share/101WciA0oSdFhSRXg=/>

[7] Jupyter/Ipython Notebook Quick Start Guide (-) What is the Jupyter Notebook [ONLINE]. Available at: <https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html>

[8] Alpha Vantage (-) Alpha Vantage [ONLINE]. Available at: <https://www.alphavantage.co>

[9] List of S&P 500 Companies (-) Business Insider [ONLINE]. Available at: <https://markets.businessinsider.com/index/components/s&p_500>

[10] Metrics to Evaluate your Machine Learning Algorithm (2018, Feb) towards data science [ONLINE]. Available at: <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>

[11] Support Vector Machines (SVM) (2018, June) towards data science [ONLINE]. Available at: <https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989>

[12] Use Decision Trees in Machine Learning to Predict Stock Movements (2017, Oct) quaninsti [ONLINE]. Available at: <https://blog.quantinsti.com/use-decision-trees-machine-learning-predict-stock-movements/>

[13] What is a Decision Tree (2017, Jul) towards data science [ONLINE]. Available at: <https://towardsdatascience.com/what-is-a-decision-tree-22975f00f3e1>