Cardiff University School of Computer Science and Informatics

Author: Saud Alharbi

Degree: BSc Computer Science

Supervisor: Oktay

Date: 23/01/24

Abstract

This project explores a machine learning approach to predicting stock market movements. I will be more focused on the analytical stance of the project and how stock prices of specific companies' stock prices change due to other companies' stock prices increasing or decreasing. Through testing, I will demonstrate the effectiveness of the model in forecasting stock prices and how they compare to other companies. I will also make the model interpretable, ensuring that users can understand the rationale behind predictions. Practical considerations such as scalability and risk management are also addressed. Overall, this research contributes to understanding machine learning and how stock prices of other companies fluctuate due to external factors.

Acknowledgements

As I am a religious man, I would like to first thank God for even having the opportunity to write this paper and to even get to this point. I would like to thank the Saudi Gov for funding my education fully as it is a blessing. I would also like to thank my family and friends for supporting me up to this point and my lovey supervisor for always encouraging me to continue even when I had no hope.

Table Of Contents

1 Introduction
- 1.1 Motivation for project
- 1.2 Aim
- 1.3 Objectives
- 1.4 Summary

2 Background
- 2.1 History on machine learning
- 2.2 Overview of machine learning algorithms
  - 2.2.1 Supervised learning
  - 2.2.2 Unsupervised learning
  - 2.2.3 Semi-supervised learning
  - 2.2.4 Reinforced learning
- 2.3 Data processing and feature engineering
  - 2.3.1 Data processing
  - 2.3.2 Feature engineering
- 2.4 Introduction to stock market
- 2.5 Evaluation metrics for stock price predictions
- 2.6 Theoretical foundations of stock price predictions
- 2.7 Challenges and limitations of stock price prediction
- 2.8 Ethical Considerations

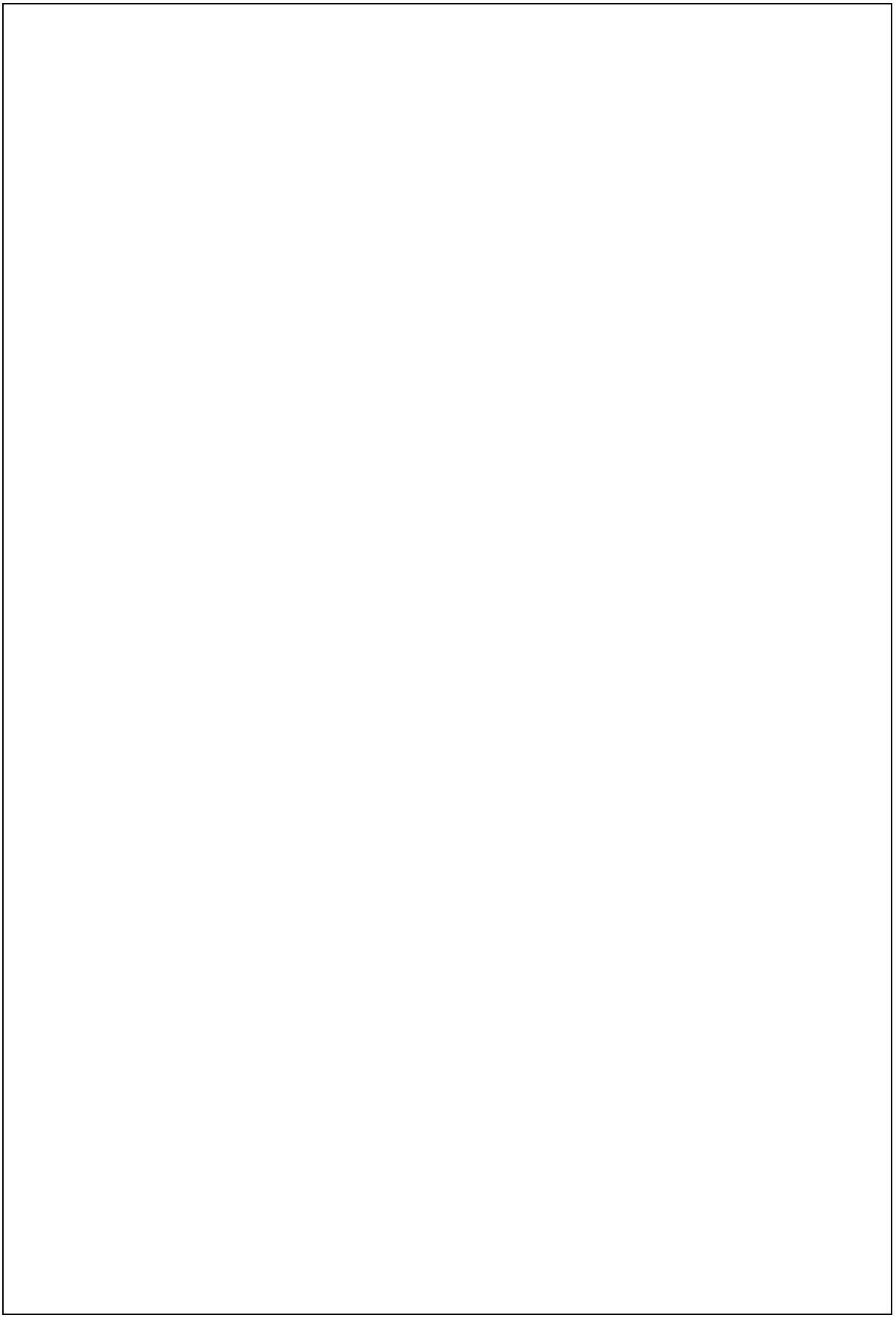


Table Of Figures

1.0 Introduction

1.1 Motivation for project

My love for stocks and trading dates back years ago to when I first stumbled across the concept of investing. The idea of getting wealthier just by making calculated moves in the stock market fascinated me. It's as if you are learning the mechanics of a certain game to master it, but in this case become wealthier. The stock market seemed like a wonderful opportunity for me to acquire knowledge while also benefiting from it financially, so I eventually took a deep interest in the topic.

Machine learning on the other hand sparked my interest due to my liking for automation. In today's world you can create an algorithm that can predict an outcome automatically just by giving it data, and that to me is extremely fascinating and intriguing. It's as if you are teaching a computer to think like a human, but with the efficiency and precision of a machine. The possibilities are endless and that is why there will always be a growing interest in the topic.

Combining these two interests of mine and making turning them into a project will not only be a good idea on a educational perspective, but will be a major advantage to amplify my resume when applying for jobs in the future. Having excellent abilities in machine learning and statistical analysis along with skills in finance and investment would help me stand out as a candidate in industries that greatly appreciate data-driven decision-making expertise.

1.2 Aim

The aim of this project is to fully understand the concept of machine learning and how it works, while using it to predict or forecast future stocks, and how other companies may affect a certain company's stock price. We will also analyse and test the effectiveness of the model in predicting the stock price, along with the stock price comparisons to other companies. We will be examining stock price changes and how other companies may affect a certain stock price, therefore this study aims to improve the potential of machine learning when it comes to accuracy in the financial market.

1.3 Objectives

In order to create a successful prediction and achieve a precise analytic, then the following objectives should be achieved:

1. Construct a solid machine learning algorithm
    - Capable of predicting future stock precisely
    - Capable of handling data without highly hindering performance
    - Accessible for anyone to use
    - Easy to navigate through (Clear and concise)

Potential risks involved: Might not be able to handle large datasets

2. Investigating / Testing other companies influence on the stock price
    - Testing the influence of companies to a certain company's stock price
    - Experimenting with different data to gather results
    - Comparing results to get a deeper analysis
    - Sharing the results

Potential risks involved:

3. Sharing my experience in using a machine learning algorithm
    - Advantages and disadvantages of using a machine learning algorithm
    - Examining how machine learning algorithms can capture complex patterns in historical data
    - Identifying relevant factors affecting stock prices
    - How machine learning algorithms can predict better than traditional methods

Potential risks involved:

By achieving these objectives, this project aims to contribute valuable knowledge to the ongoing exploration of machine learning in the context of stock market analysis.

1.4 Summary

Generally, this project will not only portray how the algorithm works and how machine learning is used, but it will also focus on the statistical analysis of the project and how some companies may affect the price of other companies. We will use the algorithm to help us figure out if some companies affect the price of stock positively or negatively on other companies.

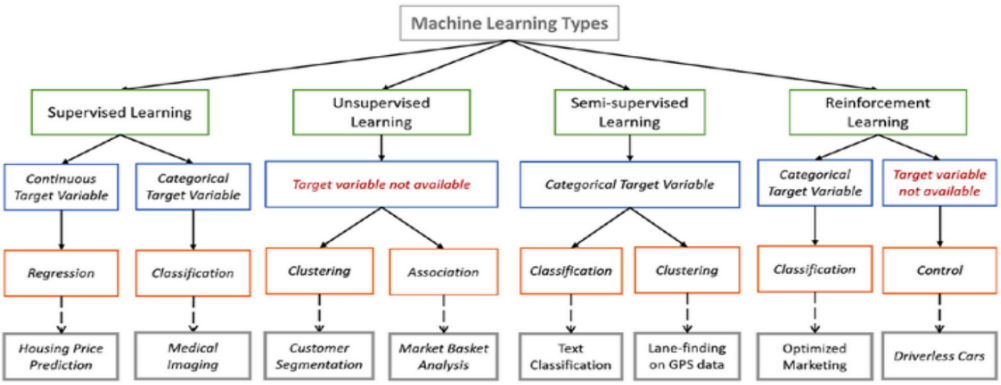
2.0 Background

2.1 History on machine learning

The history of machine learning is a wonderful story that has been studied for decades that has made significant advancements throughout the years. Historically, the roots of machine learning extend back to the seventeenth century with Pascal and Leibniz's creation of machines capable of copying human abilities. In more recent times, Arthur Samuel, while at IBM, came up with the term "machine learning" and showed the computers potential to learn tasks such as playing checkers which was revolutionary (El Naqa, I. and Murphy, M.J., 2015). Since then, there were many innovations of machine learning and after Arthur Samuel's concept came Rosenblatt and his development of the perceptron (Rosenblatt, F., 1958). Although his innovation was groundbreaking at the time there was one person by the name of Marvin Minsky a American mathematician and computer scientist, who disliked Rosenblatt's perceptron, because he thought that the innovation is limited to only linearly separable problems (Minsky, M. and Papert, S., 1969). About 6 years after that an American social scientist and machine learning pioneer named Paul John Werbos had taken the perceptron with its flaws and created the multiplayer perceptron (MLP) (Werbos, P., 1974). Approximately 12 years later came in the development of decision tress by John Ross Quinlan (Quinlan, J.R., 1986), and vector machines by Cortes and Vapnik (Cortes, C. and Vapnik, V., 1995). A lot of innovation had been developed after these years bettering machine learning algorithms to what it is today. Since then, machine learning has come a long way with the means of artificial intelligence with companies like OpenAI founded by Sam Altman, Greg Brockman, Peter Thiel, and Elon Musk creating a revolutionary artificial intelligence system called ChatGPT where it is widespread amongst the population used for all sorts of tasks.

2.2 Overview of Machine Learning Algorithms

Machine learning algorithms are specifically programmed to receive and analyse the given data to output a predicted value. Since the stone age humans have always come up with innovations to simplify their daily tasks. It's fascinating to know that algorithms like this are able to learn just by feeding it data or analysing its environment whichever one it might be. It would be able to recognize the patterns and clues or even solve upcoming problems, and that is what makes this type of algorithm fascinating.



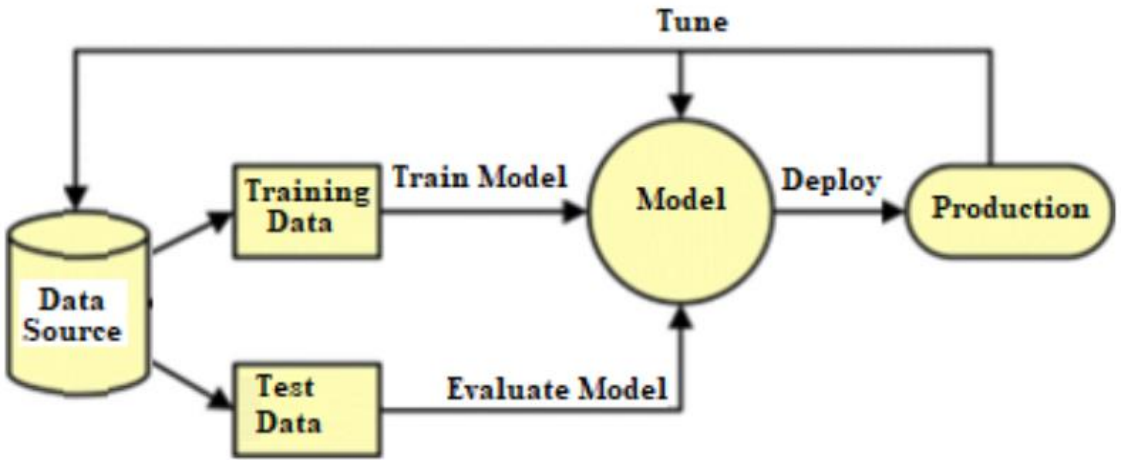
As we can see from the table above there are four major types of machine learning algorithms that are important. Supervised learning algorithms, unsupervised learning algorithms, semi-supervised learning algorithms, and reinforcement learning algorithms.

2.2.1 Supervised Learning

Supervised learning utilizes linear regression, which is a method that seeks to model the relationship between an input variable and an output variable. With a supervised learning you show the algorithm the input and output, so it can recognize the patterns and learn from it. It is like showing the algorithm a question and providing it with the answer to that specific question as if it was a puzzle. It would try to understand these patterns and learn from them. With this specific algorithm you would provide it with a input which is a set of training data and a set of testing data, which will train it enough to give a solid prediction. The bigger the set of data is the better prediction it might have.

2.2.2 Unsupervised Learning

On the other hand, unsupervised learning does the complete opposite, to where instead of having example pairs to train the algorithm it is instead provided with a large dataset, and it will try to find similarities and clues or important things within that dataset to learn and try predicting the future with it. Unsupervised learning employs K-means clustering, an algorithm designed to separate a dataset into K distinct clusters based on similarities among data points. Unlike a supervised learning it does not have a training and testing dataset to learn from, but instead independently learns from clues within a dataset. It is like providing a student with a picture of a city and asking them to find similarities and patterns within that city. In the figure bellow best explains the flow of training and testing data (Mahesh, B., 2020).



2.2.3 Semi-Supervised Learning

As for semi-supervised learning it is as if you had a mix of both supervised learning and unsupervised learning working together. So for example if the algorithm were to be given a dataset, then it would try to find information that are related to one another, but not exactly identical. What this does is it will try to find relations in the dataset to better understand and learn and in turn give out a better prediction. For a better understanding it says in the article by (El Naqa, I. and Murphy, M.J., 2015) that "semi-supervised learning is where part of the data is labelled, and other parts are unlabelled. In such a scenario, the labelled part can be used to aid the learning of the unlabelled part". What this will do is gather information about every task and use that knowledge to improve its understanding and improve on its learning. For example, its like learning about cats and dogs, so you end up understanding both animals better. The algorithm would try to find relations between these two animals and therefore would better understand each animal individually.

2.2.4 Reinforced Learning

Reinforcement learning uses Q-learning, a technique for training agents to make sequential decisions in an environment to maximize cumulative rewards. The best way to explain a reinforcement learning algorithm is that it uses trial and error to better understand it's environment. The algorithm would take a decision and will try to see if that specific decision is right or wrong, and based on the outcome of that decision the machine will learn from the action and make a better decision next time. What this algorithm will try to do is maximize its rewards every time it takes a action and through time it will become better and better, so it can gain as much as it can (Sutton, R.S. and Barto, A.G., 1998).

2.3 Data Preprocessing and Feature Engineering (CHECK)

Data preprocessing is basically the process of preparing raw data for analysis and modelling. It does that by cleaning and altering the data and what this will do is cover operations such as encoding categorical variables, scaling numerical features, and handling missing data. Feature engineering is the complete opposite where it will enhance the functionality of machine learning models by either producing new features or changing already existing ones. This may lead to selecting the most related features for the job and using methods like dimensionality reduction, extracting temporal or text-related features, or creating new features based on domain expertise. If we want to prepare the data for machine learning algorithms, which eventually create more reliable and accurate models, both feature engineering and data pretreatment are very important tasks (Zelaya, C.V.G., 2019).

2.3.1 Data Processing (CHECK)

Data preprocessing is a very important step in machine learning, because it focuses more on cleaning and modifying the raw data, so it can get it ready for analysis and modeling. One essential task in data preprocessing is data cleaning, which involves handling missing values and outliers. We can use mean, median and mode replacement as a imputation method to deal with missing values or we could delete the missing values to solve the same problem. On the other hand outliners have the ability to change the outcome, so they have to be dealt with (Zelaya, C.V.G., 2019).

Another important aspect of data preprocessing is data transformation, which includes scaling and encoding categorical variables. Scaling ensures that features are on a similar scale, preventing certain features from dominating others during model training. Common scaling techniques include standardization (z-score normalization) and Min-Max scaling. Categorical variables, on the other hand, need to be encoded into numerical values for machine learning algorithms to process them effectively. This can be achieved through techniques like one-hot encoding or label encoding.

Additionally, feature selection is often performed during data preprocessing to reduce the dimensionality of the feature space and improve model performance. Irrelevant or redundant features can be identified and removed using correlation analysis or feature importance ranking techniques. Dimensionality reduction methods such as Principal Component Analysis (PCA) can also be employed to capture the most important aspects of the data while reducing its complexity.

2.3.2 Feature Engineering (CHECK)

Feature engineering involves creating new features or transforming existing ones to enhance the predictive power of machine learning models. One common strategy in feature engineering is creating new features based on existing ones. This may involve generating polynomial features to capture nonlinear relationships or creating interaction features that represent the product or ratio of two existing features. Domain-specific knowledge can also be leveraged to engineer features that are particularly relevant to the problem at hand.

Temporal features play a significant role in many machine learning tasks, especially those involving time-series data. Extracting time-related features such as day of the week, month, or year can provide valuable insights into seasonal patterns or trends. Time since last event features can also be created to capture the recency of certain occurrences, which may be relevant in predicting future events.

In addition to numerical and temporal features, feature engineering techniques are also applied to text and image data. For text data, features can be extracted using techniques such as Bag of Words (BoW), TF-IDF, or word embeddings like Word2Vec. Image data, on the other hand, can

be transformed into features using methods such as color histograms, edge detection, or by extracting features from pre-trained Convolutional Neural Networks (CNNs). By carefully crafting and selecting features, practitioners can greatly improve the performance of machine learning models, making them more accurate and robust in handling real-world data.

2.4 Introduction to Stock Market

What is the stock market? In a simple explanation the stock market is where people can buy or sell a share of a company of their choosing. The word stock usually means when a person has ownership or equity in a company of his/her choosing. Generally, with this project we will be focusing on the closing price of a company's stock as that is what the machine learning prediction algorithm will try to determine. A closing stock price of a company is the last stock price at which a specific stock had been traded in then day. This is extremely important for investors to look at as it determines weather they want to buy or sell a share in the future. As we look to analyse the closing stock price of a certain company compared to other company's performances it's important to compare them precisely for a accurate analysis.

2.5 Evaluation Metrics for Stock Price Prediction

What are the metrics that will help us determine whether the prediction model will be successful or not? We have to measure the precision, accuracy, and recall of a prediction to determine its success. According to (Juba, B. and Le, H.S., 2019) accuracy is the simplest and most widely used metric to measure the performance of a classifier, but it may not be the best metric to use if the data is not balanced accordingly. What accuracy does is it takes the number of correct predictions it had made and compares it to all the predictions made in total (De Fortuny, E.J., De Smedt, 2014). If you wanted to attain high accuracy, you would attain it by predicting the dominant class and that would cause a low precision. Speaking about precision it is basically measuring the quality of the positive prediction made by the machine. This metric is crucial in machine translation tasks to accurately predict which words should be used in the translated sentence. Some important applications like the alarm systems in hospitals that have low precision is extremely unwanted by them. The amount of information needed for a trustworthy translation is going to be affected by the rhythm of the rarest word in the wanted sentence, which shows how important accuracy is in translation. As for recall it is a metric used to show how many times it chooses positive examples or true positives out of all the existing positive examples in the dataset. Recall can be calculated by dividing the total number of positive cases by the number of true positives existing in the dataset (Juba, B. and Le, H.S., 2019). All these metrics together are really important to determine an accurate stock price and looking at these metrics individually can give us a good idea of how well the algorithm is running.

2.6 Theoretical Foundations of Stock Price Prediction

Looking through some articles highlighting the theoretical foundations of stock price prediction I have found many important points highlighting the topic especially looking over the Efficient Market Hypothesis (EMH), Random Walk Hypothesis, and many other statistical approaches. As for the Efficient Market Hypothesis I looked through an article by **(Awiagah, R. and Choi, S.S.B., 2018)** on the Ghana stock exchange highlighting the hypothesis and what it does. It basically examines the predictability of the index returns on Ghana's stock market within weak-form efficient market hypothesis framework. The historical data provided within the research spans 28 years from 1990 to 2017 analysing the daily, weekly, monthly, and quarterly returns of investors. The statistics given show a large inconsistency between mean and standard deviation, which tells us that there is a large market risk that exists. As for the Random Walk Hypothesis, it was tested using four statistical tests. The article mentions it uses Ljung-Box autocorrelation test, unit root tests, runs test, and variance ratio tests. Unfortunately, the empirical results reject RWH in all four returns which are daily, weekly, monthly, quarterly. What this tells us is that the stock prices on the Ghana Stock Exchange do not follow the same walk patterns as they use different ones. If the walk behaviour is not random it tells us that the stock market may not exactly show us all the available information, which is the complete opposite of what RWH stands for. What this article tells us about the Theoretical Foundations of the stock price in Ghana is that the Ghana stock exchange may not comply or adhere to EMH, because the statistical tests tell us that stock returns are predictable and do not follow a random walk pattern as expected by the hypothesis.

2.7 Challenges and Limitations of Stock Price Prediction

Successfully predicting a price on the stock market is not going to be an easy challenge to go through as there are a lot of challenges and limitations, we may face throughout our journey to creating an efficient and working machine learning system. The first challenge we may face is our prediction being influenced by external factors like economic conditions, news, and an investors action. That is what this research will be focused on prominently to see whether external companies have an affect on a stock price or not.

The second challenge we may face is how unpredictable a stock market can be. Stock prices can be very unpredictable, like for example a volcano erupting on New Year's Eve. This unpredictability makes it hard to find clear patterns in the data given to us. Sometimes, the models get tricked into thinking they've found patterns when they're really just picking up on random patterns that are not related at all, like thinking a tsunami will happen, because you may have seen a few waves come by. The model we use might catch onto the unrelated patterns rather than genuine related pattens in the stock market (Bau, H.H. and Shachmurove, Y., 2009).

The third challenge we may face is when predicting stock prices, most of the time we have to look at many different things, like how prices have changed over time, how much trading is happening, important economic factors, etc. Since we have a lot to pick from it is going to be hard to choose which one to use or which ones are really important. If I ended up using all the options available to determine the prediction, then the algorithm might make relations within the dataset given that do not exist, which in turn gives us an unreliable prediction. If we stick with one or two variables it might give us a better prediction, so sticking to a simpler plan may be better for the algorithm.

The fourth challenge is we must have a good and reliable financial dataset, because it is very important for the training model to give accurate predictions for us. The problem with some of the datasets available is that they might be either limited, inconsistent, or contain a lot of errors which is not ideal when creating a functional accurate machine learning algorithm. That is why having a quality dataset can have a crucial positive impact on the outcome of a prediction system.

2.7.1 Solutions to These Challenges

2.8 Ethical Considerations

As we use this prediction model, we need to take into consideration the ethical use and implementation of this algorithm. First of all, we would need to follow the rules and regulations related to this topic of the country or state in which the person using or implementing this model resides. Another consideration we need to take into account is manipulating the market, because some people may use the prediction models to control markets unfairly, which will benefit them, but will also harm others at the same time. Also, if people use these models for their trading it might affect the stability of the market which is extremely risky. The third consideration will be privacy and data protection as collecting personal data for these models could raise concerns. Also, the use of resources to create these prediction models needs to be used responsibly. The fifth consideration that came to mind is that if people started relying excessively on these models, then it might change how markets work and will in turn affect society in the future (Ayling, J. and Chapman, A., 2022).

2.9 Summary

There are a lot of components that contribute to how a machine learning algorithm works as it is not a simple algorithm to create whatsoever, but all these components together even if we did not cover them all help aid in creating a functioning machine learning algorithm. As we covered the background of our project, we will later be touching on the methodology of it and how everything works together.

3.0 Methodology

3.1 Overview (CHECK)

This project will follow a structured method for predicting stock prices using machine learning techniques. After delving into the research methodology for this project we will begins by gathering relevant data, such as historical stock prices of companies, and carefully preparing the dataset to make sure of its quality. Then, it uses feature engineering methods to extract useful information from the data, helping to build predictive models and this particular point has been talked about in the "Background" section of this dissertation. To select a particular model, we have to consider different algorithms, which will focus on trying to understand the dataset we provide it with. After a lot of training and validation, the chosen models are fine-tuned for optimal predictive accuracy. Finally, the performance of these models are going to be tested through a lot of testing, giving insights into how well they can be applied in real-world situations. This summary lays the groundwork for a detailed explanation of each step in the following sections of the methodology.

3.2 Research Methodology

"Application papers are essential in order for Machine Learning to remain a viable science" (Provost, F. and Kohavi, R., 1998).

As there are plenty of papers that are being used to further enhance the information on machine learning algorithms, then it will be pretty simple to give a clear understanding on the topic and how we will use it for our statistical analysis. This quote above refers to research papers on machine learning algorithms becoming a source of financial gain rather than a study of knowledge, and hopefully my aim is to keep this quote alive.

3.3 Data Collection and Preprocessing

As I needed a complete and quality dataset to use, I was looking through countless options that did not really match my needs. That is when is stumbled across Yahoo Finance which had met all the requirements I had needed and also held a reputation of being a trusted place to gather datasets on companies' stock prices. I ended up opting to choose Tesla owned by Elon Musk as the main company that was going to be tested and I would be comparing Elon's Tesla to his other company's datasets like SpaceX, Neuralink and The Boring Company to see if his other companies have an effect on Tesla's stock price.

3.4 Model Selection and Training

3.5 Validation and Testing

3.6 Summary