WGU C951

Task 3

STOCK PRICE PREDICTION PROPOSAL

Armondo Dobbs Jr

Student ID # 010111115

12/27/22

Table of Contents

A. Project Overview

This proposal describes the implementation of a stock price prediction program. Predicting stock prices can be a complex task, as stock prices are influenced by a variety of factors including economic conditions, company performance, and market trends. This is where machine learning will come into play.

A.1. Organizational Need

There are several shortcomings of stock price prediction without machine learning.

Limited data - Stock price prediction without machine learning is often based on a limited amount of data, such as past stock prices or financial statements. This can make it difficult to accurately predict future stock prices, as these predictions may not take into account other important factors that could impact the stock price.

Time-consumption - Stock price prediction without machine learning can also be time-consuming, as it often requires manual analysis and interpretation of data.

The use of machine learning in stock price prediction can lead to improved accuracy, increased efficiency, and automation, making it a valuable tool for businesses and individuals interested in making informed decisions about the stock market.

A.2. Context and Background

Initially, stock price predictions were based on relatively simple methods, such as manually analyzing past stock prices or financial statements. With the advent of computers, more sophisticated methods were developed, such as technical analysis, which uses statistical tools to identify patterns in stock prices and make predictions based on those patterns.

The use of machine learning in stock price prediction has become more relevant in recent years as the amount of data available has increased and the computational power required to analyze that data has become more widely available. Machine learning algorithms can process large amounts of data and identify patterns and trends that may not be apparent to humans, making them a valuable tool for predicting stock prices.

Stock price prediction with machine learning involves the use of special algorithms to predict the future value of a company's stock. Machine learning algorithms can analyze large amounts of data and learn from that data to make predictions about future events. In this context, these algorithms are trained on data such as past stock prices, financial statements, and other market data, and then used to make predictions about future stock prices.

A.3. Outside Works Review

In review of several articles which describe the applications of machine and deep learning to make stock price predictions, there were several findings that prove the worth of implementing these techniques into more wide scale scenarios.

In the first review, researchers affiliated with the University of Tunisia conducted a machine learning study using supervised learning techniques to forecast historical data-based stock market prices and patterns as well as provide useful historical price analysis. The algorithms and techniques used in this study included linear regression, support vector machine (SVM), random forest, nearest neighbor, and decision trees. The results from this study proved that these machine learning principles can be successfully applied to business cases as the program performed very well in terms of one-day prediction patterns (Fathali et al., 2022). With more development, this kind of prediction system will help investors and individuals make smarter stock market decisions.

In a second study by Jingyi Shen & M. Omair Shafiq, a deep learning approach was used to make the prediction model for stocks. The architecture of their proposed solution was separated into three parts. The first was the feature selection part, to guarantee that the selected features were effective. Second was data analysis to perform the dimensionality reduction. And the final part was to build the prediction model of target stocks. Recursive feature elimination (RFE) was implemented to ensure all the selected features were effective to prevent any unwanted side effects. This study also focused on the short term prediction approach (one day and up to 2 weeks) for simplicity. Of their proposed prediction models, the Long Short Term Memory (LSTM) model achieved a binary accuracy of 93.25%, which is a high precision of predicting the bi-weekly price trend (Shen & Shafiq, 2020). This type of research will apply to our proposal as it proves that there can be a positive impact in stock market decisions when the predictive model is properly implemented.

In a final machine learning study on stock price prediction, researchers proposed the use of a class of machine learning frameworks known as Generative Adversarial Networks (GAN) to effectively sort and manage data that is both labeled and unlabeled. Without labeled data, a GAN can quickly learn from internal representations of data, generate data, learn density distributions and use a trained discriminator as classifier. This allows for more swift and effective processing of data to make predictions (Polamuri et al., 2022). Like the previously mentioned study, this program also implemented the use of an LSTM model paired with a GAN to form a hybrid prediction algorithm. Hybrid deep learning models have been found to be more efficient in processing. When compared with other predictive learning models, it was concluded that with the incorporation of valid and relevant data for processing, a GAN-HPA is proven to be adept in using reinforced learning to better predict stock market prices over time. This also proves to be a good business implementation because if the program can learn faster on its own, it will be able to make better predictions for stockholders to base their inputs on.

A.3.A. Relation to project

The first review of stock price machine learning relates to our proposal as the listed algorithms used in the study are the key focus of the design of our program. These algorithms are great for implementing large data sets.

In the second review, we learn that using relevant feature sets in the prediction model can have a profound impact on the prediction results. This idea will need to be implemented in our project to achieve an accurate outcome.

The final review relates to our proposal because we will be using large data sets to parse through to make predictions. It goes into detail on a hybrid prediction model which has a high success rate and would be very useful to implement.

A.4. Solution Summary

The stock price prediction program is a machine learning model which allows users to see stock forecasts for specific stocks and varying time intervals. The solution will have access to relevant company stock data and will then make calculated forecasts with the data on what price the stock will be at on a selected date. Users will be able to see projected prices and dates which will allow for easier business planning and better stakeholder communication.

A.5. Machine Learning Benefits

Machine learning can benefit stock price prediction in several ways:

Automation: Machine learning algorithms can analyze large amounts of data and make predictions automatically, without the need for human intervention. This can save time and reduce the risk of human error.

Reduced cost: Automating stock price prediction using machine learning can save money by reducing the need for human analysts and allowing for faster decision-making.

Increased scalability: As more data becomes available, machine learning algorithms can adapt and continue to improve their predictions, allowing traders to stay up-to-date with the latest market trends.

B. Machine Learning Project Design

B.1. Scope

Steps included in the scope are as follows:

- Data collection: A primary step in creating a machine learning model is to gather a sufficient amount of data to train the model. This data might include historical stock prices, financial statements, economic indicators, and news articles.
- Data preprocessing: Once the data has been collected, it will need to be prepared for the machine learning model.
- Model selection: There are many different types of machine learning algorithms that can be used for stock price prediction.
- Model training: Once the machine learning algorithm has been selected, it will need to be trained on the preprocessed data.
- Model evaluation: After the model has been trained, it will need to be evaluated to determine its accuracy.

Steps not included in the scope are as follows:

- Model optimization: If the model's performance is not satisfactory, there are a number of techniques that can be used to improve it. These might include adjusting the hyperparameters of the model, adding or removing features, or using different types of data.
- Non-financial factors: Factors such as weather, political events, or natural disasters are not typically relevant to stock price prediction and would be considered out of scope.
- Non-public information: The use of insider information or other non-public data would be illegal and out of scope

B.2. Goals, Objectives, and Deliverables

Goals

- The program should be able to make accurate predictions about future stock prices.
- The program should be able to handle large amounts of data and make predictions in real-time.
- The program should have a simple, easy to implement, and scalable architecture.

Objectives

- The main objective is to develop a machine learning model that can accurately predict future stock prices.
- The project should aim to improve the performance of the predictive model over time
- The model should be able to run in a timely manner and train on large datasets. The time required for training and testing the model can be measured in seconds or up to minutes.

Deliverables

- The complete machine learning system program

- A report detailing how the model was developed, including the choice of algorithms, features, and parameters. This report will also include an evaluation of the model's performance using metrics such as accuracy, precision, and recall.
- A user manual that explains how to use and interact with the predictive model, including how to input data and interpret the results.
- The source code for the predictive model, including any necessary scripts or libraries for training and evaluating the model.

B.3. Standard Methodology

Development will follow the SEMMA methodology. This methodology is a process that can be used to develop and evaluate machine learning models, including those for stock price prediction.

• Sample: A representative sample of the data that will be used to train and evaluate the predictive model. This can include financial data on stocks, market indicators, and other relevant information.

• Explore: Here, we will explore the data to gain a better understanding of its characteristics, such as patterns, outliers, and missing values. Included in this process will be data cleaning, transformation, and feature set processing.

• Modify: Based on the insights gained from exploration, we will make modifications to the data, such as handling such missing values, transforming variables, and creating new feature sets to be incorporated into the data.

• Model: Begin developing the predictive model. This step will include selecting the appropriate algorithm, tuning its parameters and evaluating the model performance. We will also implement the custom data sets for the model to use.

• Assess: Evaluate the performance of the predictive model using a variety of metrics such as prediction accuracy, precision, recall, etc. In this case we will be measuring the predictions given by the program.

B.4. Projected Timeline

2/1/23 – 2/13/23 - The proposal is accepted and model design begins. Data sets on stocks are collected.

2-14-23 – 2/21/23 - A mockup concept of the program is presented. Data is filtered and configured to be used by the application. Small scale version of the project.

2/22/23 – 3/7/23 - Project is up and running with limited functionality. Debugging process begins. Data can be processed by the application and goes through formal review.

3/8/23 – 3/30/23 - The application is complete and functional. Testing begins. The program will be able to use the given data sets to make accurate stock predictions. If not accurate at first it will be able to learn and self-adjust to make more precise predictions.

4/1/23 – 4/15/23 - Program completed and deployed on company network for business application.

Sprint Schedule

Sprint	Start	End	Tasks
1	2/1/23	2/13/23	Collect relevant data Design prediction model
2	2-14-23	2/21/23	Present mockup of program Configure filtered data
3	2/22/23	3/7/23	Testing/debugging phase Implement data learning
4	3/8/23	3/30/23	Final testing/Adjustments to data
5	4/1/23	4/15/23	Deployment to company network

B.5. Resources and Costs

Resource	Description	Cost
Application development	program development and debugging. (Man hours)	x3 employees @ $4,000/mo 1 months
Data analysis	Research and filtering of relevant data for the program to use. (Man hours)	x2 employees @ $2,500/mo 1 month
Hardware	Computers and equipment for working	up to $1,500
Software	machine learning framework/libraries	up to $200 per user
	Total	$16,000 startup cost $600/mo software subscription

B.6. Evaluation Criteria

Describe the criteria used to evaluate and measure the success of the completed project.

Objective	Success Criteria
Ease of Use	Prospective users will be able to successfully run applications with relative ease.
Prediction accuracy	The results will be within +-15% the expected parameters.
Algorithm Efficiency	The prediction will be generated within the 8-10 second maximum.

C. Machine Learning Solution Design

C.1. Hypothesis

We hypothesize that while using historical stock prices, financial news, and other relevant data, a machine learning model can accurately predict future stock prices with a high degree of accuracy. This will in turn solve the business need for more man hours as parsing and filtering data will be automatic and help companies make better business decisions.

C.2. Selected Algorithm

A simple linear regression algorithm will be used in our application as it can model the relationship between an arbitrary dependent variable (stock price in this case) and one or more independent variables (financial indicators or economic indicators).

C.2.a Algorithm Justification

The use of linear regression in stock price prediction with machine learning can be justified for several reasons:

- Linear regression is a simple and well-understood algorithm, which makes it easy to implement and interpret the results. It can provide an initial baseline or comparison model, or even a good starting point.
- This algorithm is more computationally efficient than some others and can handle large datasets with many independent variables. This is beneficial for us because that is a primary goal for our project.

C.2.a.i. Algorithm Advantage

This algorithm will work well with this project because linear regression assumes that there is a linear relationship between the independent variables and the dependent variable. These assumptions will then allow the use of traditional statistical methods to form conclusions, this is important for explainability in financial models.

C.2.a.ii. Algorithm Limitation

Linear regression can have some limitations due to its simplicity. The algorithm is sensitive to outliers and extreme values in data can have a big impact on the results. While the algorithm's simplicity is an advantage, it can also be a limitation because it will not have as much flexibility when it comes to complex or nonlinear data.

C.3. Tools and Environment

To implement our prediction application, we will of course need hardware to develop which will be desktop computers. A development environment such as Jupyter Notebook, PyCharm, or RStudio will be needed to write, test, and debug the code. TensorFlow will be the primary machine learning framework to test the model. Third party code will not be necessary in this case.

C.4. Performance Measurement

The performance of this model will be measured by its accuracy and the correlation coefficient. Accuracy in this case is classified as the proportion of correct predictions made by the model. The correlation coefficient, however, measures the strength and direction of the linear relationship between the predicted and actual stock prices. A coefficient of 1 indicates a perfect positive correlation, while a coefficient of -1 indicates a perfect negative correlation, which shows the fit between the predictions and actual values.

D. Description of Data Sets

D.1. Data Source

The selected stock data for this program has been extracted from a public domain known as Kaggle. This site hosts plenty of data sets ranging from all sizes which will be perfect for this application. Our chosen data set consists of the last 20 years of US historical stock prices with earnings data.

D.2. Data Collection Method

The data set in this case will be acquired directly through a one time download from the Kaggle site.

D.2.a.i. Data Collection Method Advantage

An advantage of collecting the data in this format is the simplicity of it. This will be a single download which will save time and money when compared to data collection through subscriptions or licenses. Our data set also has the added benefit of having a surplus of data for many use cases so we will not need to outsource more data after the extraction.

D.2.a.ii. Data Collection Method Limitation

Data quality can be a limitation when it comes to how our data is gathered. Data may be incomplete, inaccurate, or inconsistent, which can introduce bias and lead to poor predictions. The data format can also be a limitation as data may be in different formats, which can require additional preprocessing and cleaning before it can be used in the model.

D.3. Quality and Completeness of Data

The data will be prepared for the algorithms by the following steps:

- Data cleaning will be done. This involves removing or correcting any errors or inconsistencies in the data, such as missing values, duplicate records, or outliers.
- Transformation will take place after the cleaning. This means transforming the data into a format that is suitable for the algorithms.
- Feature selection techniques will be used to identify and select the most relevant features for the model.

Once these primary steps are completed, the data will then be imported into our IDE to be implemented.

D.4. Precautions for Sensitive Data

Sensitive data in an application such as this can include personal information about company employees or customers, as well as confidential financial information about the company. This data must be handled carefully to maintain compliance with legal regulations and protect individual privacy. To accomplish this, data minimization can be used to minimize the amount of data necessary to achieve the goals of the model. Data segregation can also be used to separate remaining sensitive data from other data and store it in a secure location, with restricted access and monitoring. In our case, there will not be any incriminating sensitive data to worry about.

References

Polamuri, S. R., Srinivas, D. K., & Krishna Mohan, D. A. (2022). Multi-model generative adversarial network hybrid prediction algorithm (MMGAN-HPA) for stock market prices prediction. *Journal of King Saud University - Computer and Information Sciences*, *34*(9), 7433–7444. https://doi.org/10.1016/j.jksuci.2021.07.001

Shen, J., & Shafiq, M. O. (2020). Short-term stock market price trend prediction using a comprehensive deep learning system. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00333-6

Fathali, Z., Kodia, Z., & Ben Said, L. (2022). Stock market prediction of nifty 50 index applying machine learning techniques. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2022.2111134