Capstone Spring 2023:JPX Tokyo Stock Exchange Prediction

Assignment: Project Introduction

February 2023

1 Background

Success in any financial market requires one to identify solid investments. When a stock or derivative is undervalued, it makes sense to buy. If it's overvalued, perhaps it's time to sell. While these finance decisions were historically made manually by professionals, technology has ushered in new opportunities for retail investors. Data scientists, specifically, may be interested to explore quantitative trading, where decisions are executed programmatically based on predictions from trained models.

2 What is the problem?

There are plenty of existing quantitative trading efforts used to analyze financial markets and formulate investment strategies. To create and execute such a strategy requires both historical and real-time data, which is difficult to obtain especially for retail investors. This project will provide financial data for the Japanese market, allowing retail investors to analyze the market to the fullest extent.

3 Why is it interesting and important?

There are plenty of reasons for predicting stock exchanges. This is an important field because it has the potential to drive financial gains, mitigate risks, stabilize the economy and advance technological deviance gains.

One of the primary reasons people are interested in predicting stock prices is for financial gain. Accurately predicting the movement of stocks can help investors make better decisions and increase their profits.

It also helps with risk management. Accurately predicting the stock market can also help investors manage risk by allowing them to adjust their portfolios based on market conditions. This can help prevent losses and stabilize investments.

It is needed for economic stability. The stock market is often seen as an indicator of the overall health of the economy. Accurately predicting market movements can help policymakers and economists make better decisions and take actions to prevent economic instability.

Predicting the stock market requires the use of advanced technologies and techniques, such as machine learning and artificial intelligence. This has led to significant advancements in these fields, which have potential applications beyond stock market prediction.

4 Why is it hard?

The stock market is a complex system with many variables, including economic indicators, political events, and company-specific factors. It can be difficult to accurately capture and model all of these factors, making it challenging to predict future market movements.

The stock market is also inherently unpredictable, with a large degree of randomness and volatility. Even small changes in market sentiment or investor behavior can lead to significant price swings, making it challenging to predict future price movements.

Historical data can be a valuable tool for predicting future stock prices, but the data can be limited, and it may not accurately reflect future market conditions. Additionally, unexpected events can occur that can't be predicted based on past data.

Stock market movements are also influenced by human behavior, including emotions such as fear, greed, and panic. These factors can be challenging to predict and can cause prices to deviate from what might be expected based on fundamental analysis or historical trends.

Overall, stock market prediction is a challenging task due to the complexity of the market, randomness, limited data, and human behavior. While there have been advances in technology and techniques for predicting the stock market, accurate predictions remain difficult to achieve.

5 What is your plan?

With the given datasets, my goal is to predict the direction of the stock exchange. I will then choose a machine learning algorithm either linear regression, decision trees, random forests, or neural networks. I also have to split data into training and validation sets. Use the training set to train the model and the validation set to evaluate its performance. And then train and validate the model, fine-tune the model and then test the model.

Stock exchange prediction is a difficult task, and no model can predict the stock market with 100 percent accuracy. However, with careful data preprocessing, algorithm selection, and model training, I can create a model that can make reasonably accurate predictions.

Here are the datasets provided by Kaggle: stock_prices.csv The core file of interest. Includes the daily closing price for each stock and the target column.

options.csv Data on the status of a variety of options based on the broader market. Many options include implicit predictions of the future price of the stock market and so may be of interest even though the options are not scored directly.

secondary_stock_prices.csv The core dataset contains on the 2,000 most commonly traded equities but many fewer liquid securities are also traded on the Tokyo market. This file contains data for those securities, which aren't scored but may be of interest for assessing the market as a whole.

trades.csv Aggregated summary of trading volumes from the previous business week.

financials.csv Results from quarterly earnings reports.

stock_list.csv - Mapping between the SecuritiesCode and company names, plus general information about which industry the company is in.

https://www.kaggle.com/competitions/jpx-tokyo-stock-exchange-prediction/data

6 What are the performance metrics?

I don't have a decided performance metric, but I will choose from some common performance metrics that are used in stock market prediction.

Mean Squared Error (MSE): This metric measures the average squared difference between the predicted values and the actual values. It is a common metric for regression problems, where the goal is to predict a continuous value, such as stock prices.

Root Mean Squared Error (RMSE): This metric is like MSE, but the square root is taken to make the units of the error the same as the original values. Mean Absolute Error (MAE): This metric measures the average absolute difference between the predicted values and the actual values. It is another common metric for regression problems.

Directional Accuracy (DA): This metric measures the percentage of predictions that are correct in terms of the direction of the market movement (up or down).

7 GitHub Repository

https://github.com/ssarowar/Stock-Exchange-Prediction

8 Literature Review

8.1 Table to compare ML methods

PaperML Methods	RA	ANNs	SVMS	DT	En	CNN	RNN	LSTM
hegazy2014machine	yes	yes	yes	yes	yes	no	no	no
CHONG2017187	no	no	no	no	no	yes	yes	no
JIANG2021115537	no	no	no	no	no	yes	no	yes
8541310	no	no	no	no	no	yes	no	no
arXiv:1605.00003	no	no	no	yes	no	no	no	no
This model	no	no	no	yes	no	no	no	no

8.2 hegazy2014machine

The paper provides a comprehensive review of the state-of-the-art machine learning techniques that have been used for stock prediction. It discusses the importance of accurate stock prediction for investors and traders and highlights the potential benefits of using machine learning techniques to improve the accuracy of these predictions.

The paper presents an overview of several machine learning techniques, including artificial neural networks, support vector machines, decision trees, and genetic algorithms, that have been applied to stock prediction. It provides a detailed explanation of each technique, its advantages and limitations, and its application to stock prediction.

The authors also discuss the factors that affect stock prices, such as market trends, company financials, and macroeconomic indicators. They highlight the importance of incorporating these factors into machine learning models for accurate stock prediction.

The paper concludes with a discussion of the challenges and future directions of machine learning for stock prediction. It notes that accurate stock prediction is a complex problem due to the unpredictability of the stock market, and that further research is needed to develop more accurate and robust machine learning models for stock prediction.

Overall, the paper is a valuable resource for researchers, practitioners, and investors interested in using machine learning techniques for stock prediction. It provides a comprehensive survey of the different techniques available, and highlights the potential benefits and challenges of using these techniques for stock prediction.

8.3 CHONG2017187

The paper ”Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies” provides a comprehensive overview of the application of deep learning networks for stock market analysis and prediction. The authors discuss the challenges and opportunities of using deep learning networks in this field, including data representation, feature extraction, and model selection.

The paper presents several case studies that apply deep learning networks to various aspects of stock market analysis, such as stock price prediction, trend analysis, and sentiment analysis. The authors also discuss different types of deep learning networks, including feedforward neural networks, convolutional neural networks, and recurrent neural networks.

The authors emphasize the importance of selecting appropriate input data and designing effective network architectures for achieving accurate and reliable stock market predictions. They also highlight the potential benefits of combining multiple deep learning models and integrating external data sources, such as news and social media feeds, into the analysis.

Overall, the paper provides a valuable resource for researchers and practitioners interested in applying deep learning networks to stock market analysis and prediction.

8.4 JIANG2021115537

The paper "Applications of deep learning in stock market prediction: Recent progress" provides an overview of the recent advancements in using deep learning techniques for predicting stock prices. The author discusses the challenges and opportunities of using deep learning algorithms in stock market prediction, including feature selection, model selection, and the impact of external factors.

The paper covers a range of deep learning models, including feedforward neural networks, recurrent neural networks, convolutional neural networks, and deep belief networks. The author also discusses different approaches to data preprocessing and feature engineering, as well as the use of transfer learning and ensemble methods for improving prediction accuracy.

The author highlights several case studies that demonstrate the effectiveness of deep learning models in predicting stock prices, including studies that use technical indicators, news articles, and social media data as input features. The paper also identifies some of the limitations and challenges of using deep learning models in this field, including the difficulty of interpreting the results and the potential risks associated with relying on machine learning models for financial decision-making.

Overall, the paper provides a valuable overview of the recent progress and future directions of applying deep learning techniques for stock market prediction.

8.5 8541310

The paper "Deep Learning for Stock Market Prediction Using Event Embedding and Technical Indicators" proposes a deep learning framework for stock market prediction that incorporates event embedding and technical indicators. The authors highlight the limitations of traditional technical analysis methods and argue that incorporating news events can provide additional insights into stock price movements.

The proposed framework uses a combination of long short-term memory (LSTM) and convolutional neural network (CNN) models to extract features from the input data. The input data includes historical stock prices, technical indicators, and news events represented as word embeddings. The authors also use feature selection techniques to identify the most important features for predicting stock prices.

The authors evaluate the performance of the proposed framework on several real-world datasets and compare it to other machine learning models, such as random forest and support vector regression. The results show that the proposed framework outperforms the other models in terms of prediction accuracy, especially when incorporating news events.

Overall, the paper presents a promising approach for using deep learning techniques to incorporate news events into stock market prediction models, which can provide additional insights and improve prediction accuracy.

8.6 arXiv:1605.00003

The article discusses the use of a machine learning algorithm called random forest to predict the direction of stock market prices. The authors explain the concept of random forest and its advantages over other machine learning algorithms. They also describe the dataset used in the study and the various features that were considered for the prediction.

The article goes on to explain the methodology used to train the random forest model and the metrics used to evaluate its performance. The results of the study show that the random forest model outperforms other traditional methods in predicting stock prices.

The authors conclude that the use of machine learning algorithms such as random forest can be an effective tool for predicting stock market trends, and can help investors make informed decisions.

8.7 Project Work

A decision tree is a type of machine learning model that uses a tree-like structure to represent decisions and their possible consequences. The tree is constructed by recursively splitting the data into subsets based on the values of input features. At each node of the tree, a decision is made based on the value of a specific feature, and the data is split into subsets accordingly. This process is repeated until a stopping criterion is met, such as a maximum tree depth or a minimum number of data points at a leaf node.

Once we have our labeled data-set and input features, we can use a decision tree algorithm to train a model to predict the target variable. During training, the decision tree algorithm would recursively split the data into subsets based on the input features, with the goal of maximizing the information gain or minimizing the impurity of the subsets.

9 Back Ground

9.1 JPX Tokyo Stock Exchange Prediction competition

The JPX Tokyo Stock Exchange Prediction competition on Kaggle was a data science competition that took place from May 2021 to August 2021. The competition was organized by the Tokyo Stock Exchange and JPX Nikkei Index 400, and the goal was to predict the closing prices of the Nikkei 225 index for the following day.

The competition provided participants with historical data of the Nikkei 225 index, including the opening price, closing price, highest price, lowest price, and trading volume. The challenge was to use this data to create a predictive model that would accurately forecast the closing price for the next day.

The winning models used a variety of techniques, including machine learning algorithms, deep learning models, and ensemble methods.

9.2 JPX

The JPX (Japan Exchange Group, Inc.) is a financial services company based in Tokyo, Japan. It operates the Tokyo Stock Exchange (TSE), which is the main stock exchange in Japan and one of the largest stock exchanges in the world.

The Tokyo Stock Exchange was established in 1878 and has a long history of supporting the growth and development of the Japanese economy. It lists a wide range of securities, including stocks, bonds, exchange-traded funds (ETFs), real estate investment trusts (REITs), and other financial products.

In addition to the Tokyo Stock Exchange, JPX operates other exchanges in Japan, including the Osaka Exchange, the Nagoya Stock Exchange, and the Fukuoka Stock Exchange. JPX also offers a range of financial services, including clearing and settlement, information services, and market data services.

JPX has a strong commitment to innovation and technology, and it has been actively promoting the use of financial technology (FinTech) to drive growth and development in the Japanese financial industry. The JPX Tokyo Stock Exchange Prediction competition on Kaggle was one example of JPX's efforts to promote innovation and collaboration in the field of data science and predictive modeling.

9.3 Stock exchange prediction

Stock exchange predictions refer to the process of using statistical and machine learning models to forecast the future prices of stocks, securities, or other financial assets traded on an exchange. These predictions are based on the historical prices and other market data of the asset, as well as various other economic and financial indicators.

The prediction of stock prices is an important field in finance and investment, as accurate forecasts can help investors make informed decisions about buying and selling securities. Stock exchange predictions are used by investors, traders, and financial institutions to identify potential opportunities for profit, manage risk, and make more informed investment decisions.

Predictive modeling techniques, such as linear regression, decision trees, random forests, gradient boosting, and neural networks, are commonly used in stock exchange predictions. These models analyze past market data to identify patterns and relationships between variables, which can be used to make predictions about future market trends.

However, it's worth noting that stock exchange predictions are subject to various sources of uncertainty, including unexpected market events, changes in economic and political conditions, and other factors that may not be captured by historical data alone. As a result, stock exchange predictions should be used in conjunction with other financial and economic analysis, as well as careful risk management strategies, to make informed investment decisions.

10 Infrastructure

Infrastructure is critical for a project involving machine learning for stock exchange prediction using decision trees. This is because the project involves processing large data-sets, training and deploying machine learning models, and managing network resources, all of which require significant computational resources.

10.1 Cloud Deployment

There are several benefits of cloud deployment for a machine learning model used in stock exchange prediction, including scalability, flexibility, cost-effectiveness, reliability and security. Cloud deployment enables you to scale your infrastructure up or down as needed, allowing you to handle sudden changes in traffic or data volume. This can be particularly useful in the stock market, where there can be sudden spikes in trading activity.

Cloud deployment enables you to quickly provision resources and experiment with different configurations, without the need for upfront capital investment. This can be particularly useful in the early stages of model development, where you may need to test different algorithms or data sources.

It also enables you to pay only for the resources you use, allowing you to optimize costs while ensuring high performance. Additionally, many cloud providers offer cost-saving options such as reserved instances or spot instances, which can help you further reduce costs.

And cloud providers typically offer high levels of availability and reliability, with built-in redundancy and fail over mechanisms. This can be particularly useful in the stock market, where downtime can result in significant financial losses. Cloud providers typically offer robust security measures, including access controls, encryption, and compliance certifications. This can help ensure the security and privacy of sensitive financial data.

10.2 Cloud Provider

For cloud providers there are three major ones, Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP).

We will be using Amazon Web Services (AWS) because it is widely adopted and established, with a large community and extensive documentation. Also offers a wide range of services and features, including machine learning tools such as Amazon SageMaker. Is provides reliable and scalable infrastructure, with strong security and compliance measures. And most importantly offers flexible pricing options, including pay-as-you-go and reserved instances. On the other hand, it can a little complex and Some services are tightly coupled, making it difficult to switch to a different provide.

10.3 Computational Resources

For the computational resources, we will use personal computers and move to Virtual machines provided by the cs department at Kent State, which are software emulations of physical computers, running on shared hardware infrastructure in a data center. They provide a high level of control and flexibility, enabling you to install and configure software and operating systems as needed. There is also AWS Lambda, it is a serverless computing service provided by Amazon Web Services (AWS) that enables you to run code without the need for managing servers or infrastructure. With Lambda, you simply upload your code and the cloud provider takes care of running and scaling it as needed. It is cost-effective and supports multiple languages.

10.4 QoS

QoS is Quality of Service, and it refers to the ability to provide predictable and consistent network performance by prioritizing certain types of traffic over others. QoS is important in networks that carry different types of traffic with varying requirements for bandwidth, latency, jitters, and packet loss.

10.5 GPU

Using a GPU for a project involving machine learning for stock exchange prediction using decision trees can provide significant benefits in terms of performance and accuracy. GPUs are designed to perform parallel computations, which can accelerate the training and prediction processes of decision trees and other machine learning models.

GPUs can perform parallel computations on large data sets, which can significantly reduce the time required for training and prediction of decision trees and other machine learning models. GPUs can perform computations with higher precision and accuracy than CPUs, which can lead to more accurate predictions and better performance of decision trees and other machine learning models. GPUs can be easily scaled to accommodate large data sets and complex models, making them well-suited for machine learning projects that require high levels of computational resources. However, it's important to note that using a GPU for your project will require specialized hardware and software, and you may need to make modifications to your existing code to take advantage of GPU acceleration. Additionally, not all machine learning algorithms and models are well-suited for GPU acceleration, so you should evaluate your specific project requirements and determine if a GPU is necessary.

A cloud-based GPU resource is Amazon Web Services (AWS). This cloud providers offer virtual machines and containers with GPU acceleration capabilities that you can use to run your machine learning algorithms.

11 System Architecture

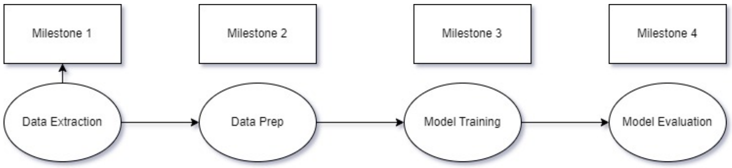

Figure 1: Example Image

12 Milestones

12.1 Milestone 1 : Data extraction

Website: Had to create a Kaggle account and down the data after agreeing to the competition rules Extract them from the zip file Stock-list: A csv file. It has over 4000 list of stocks. It has information of the stock names, their products, trade date, sectors and most importantly market capitalization. There are total of 4 folders for the data. Data-specifications: This folder contains 5 csv files. And they are all giving out the specifications of the data. There are options-spec, stock-fin-spec, stock-list-spec, stock-price-spec and trades-spec. Example-test-files: Contains 6 csv files with example of using the data for testing. Supplemental-files: along with the stock list, this folder has more data with other variables of the stock list. This has financials, options, stock-prices, trades and secondary-stock-prices. There are around 100,000- 1 million records in each file. Trains-files; this folder is the same as supplemental-files but it has less records compared. This has financials, options, stock-prices, trades and secondary-stock-prices. The first step was to create conditions in excel sheet and remove ay records that had information missing. Lastly, created amazon AWS to start with the next milestone.

12.2 Milestone 2 : Data prep

Decided on the building a machine learning model for stock exchange prediction using the Long Short-Term Memory (LSTM) algorithm. The model will be trained on historical data and will be used to predict the future prices of The LSTM algorithm is a type of recurrent neural network (RNN) that is particularly well-suited for processing sequential data. The model consists of multiple LSTM layers, followed by one or more fully connected layers, and a final output layer. Each LSTM layer contains several memory cells that can retain information over time, and a set of gates that control the flow of information through the layer. The fully connected layers are used to perform feature extraction and reduce the dimensionality of the data. Finally, the output layer produces a single value that represents the predicted stock price of a stock. Imported the data into Jupiter notebook and cleaned it to remove any missing values. Then the data had to be split from supplemental files. We have taken around 80 percent from each file for training and 20 percent for testing. And using example test files for validation. Split the data into input and output variables.

12.3 Milestone 3 : Model Training

We Scaled and reshaped the stockprices files. We first load the dataset and define the lookback parameter. We then extract the input and output data, normalize the input data using StandardScaler, and reshape the input data to a 3D tensor. Next, we define the LSTM model using Keras Sequential API. The model consists of a single LSTM layer with 50 memory cells, followed by a Dense output layer with one unit. We compile the model using the mean squared error loss function and the Adam optimizer. Lastly, we train the model using the fit() method, with a batch size of 64 and a validation split of 20 percent. The training process is printed to the console using the verbose argument. After training, the model can be used to make predictions on new data using the predict() method.

12.4 Milestone 4 : Model Evaluation

Using the Mean Squared Error (MSE) we create evaluation cases. This measures the average squared difference between the predicted and actual values. Lower values indicate better performance. Then visualized the predictions against the actual values using a line plot.

13 References

[?]

[?]

[?]

[?]

[?]

[?] JPX Tokyo Stock Exchange Prediction. Kaggle. (n.d.). Retrieved February 16, 2023, from https://www.kaggle.com/competitions/jpx-tokyo-stock-exchange-prediction/data

[?] Biswal, A. (2023, February 16). Stock price prediction using machine learning: An easy guide: Simplilearn. Simplilearn.com. Retrieved February 16, 2023, from https://www.simplilearn.com/tutorials/machine-learning-tutorial/stock-price-prediction-using-machine-learning

[?] Yates, T. (2022, July 13). 4 ways to predict market performance. Investopedia. Retrieved February 16, 2023, from

https://www.investopedia.com/articles/07/mean$_{r}eversion_{m}artingale.asp$

[?] Agrawal, R. (2022, July 21). Evaluation metrics for your regression model. Analytics Vidhya. Retrieved February 16, 2023, from

https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/