P1 (3 mins) Nicholas

* This video should start by clearly displaying your Lab Group, Team number, and names of Team Members.
* Motivation
* Problem definition
* Data Collection (Collected(NLP))
* Data Preparation (clean, preparation)
* What tool and why we use (reason on using arima, lstm, nlp and using classification accuracy)
* Exploratory Data Analysis (Explain Moving Average and Exponential Moving Average)

P2 (4 mins) Long

* Explanation of ARIMA
* Further Optimization (Using LSTM & News as Predictor)

P3 (3 mins) Hendy

* What we apply outside n inside of course that we used
* How effective the solution
* Closing (insight : what we can learn through project)
* This video MUST end with a clear description of which Team Member contributed in what part of the work.

P1:

Slide 1: Hello Prof, I am Nicholas and here with me today is Zhan Long and Hendy. We are from Lab SC13 and Group 10. Our mini project would be about Stock Prediction.

Slide 2: Our motivation for doing this mini project, would be due to the rise of accessible trading platforms such as Tiger Broker, MooMoo and many more. The Stock market has seen an increase in popularity and is attracting more and more traders, as it brings benefits such as potential for high returns, provides passive income which is also taxless and also protects your wealth from inflation.

Slide 3: Our team’s problem definition would be, are we able to accurately predict the future stock price using the history of data from the last 10 years?

Slide 4: Starting off with the collection of our data, we extracted 40 years worth APPLE stock price dated back from 1980 to 2021 from Yahoo Finance.

Slide 5: Continuing on from Data Collection, we learnt how to use RapidAPI to pull news data through HTTP Request via keywords such as Apple, Ipad and Iphone. We then converted the data from JSON to Dataframe and saved it as a CSV file.

Slide 6: Moving on to the preparation of our dataset, firstly we made the index column our date to increase the readability. Next, we checked our dataset for any NULL values and dropped them if they exist. Then, we filtered out to 10 years worth of data due to the low fluctuation at the beginning, making our prediction more accurate. Finally, we would be predicting our values based on the Adjusted closing price hence all other columns are dropped. Now we are done preparing the dataset for our ARIMA model.

Slide 7: Continuing on from our data preparation, we had to add in an additional step to transform our dataset into a 3D dataset in order for it to fit into the LSTM model.

Slide 8: The tools and techniques we use in our mini project were mostly not taught in this course which includes ARIMA, GARCH, LSTM, NLP and VADER. However, we did make use of the Classification Tree which was taught in this course.

Slide 9: Moving on to our Exploratory Data Analysis, we used the Moving Average which is used to help smoothen out the price action by filtering out “noise” from the random price fluctuation happening over time. By plotting a moving average graph of 90 days, we are able to see a smoother graph as compared to the others and judging from the upwards slope towards the end of the graph, it tells us that our data is trending towards an increase in the near future.

Slide 10: Next we use the Exponential Moving Average as it emphasises more weight on recent data points as compared to the Simple Moving Average, as mentioned in the last slide, which applies equal weight to all the data points throughout the period. From the Exponential Moving Average graph we also are able to see an upwards trend towards the end of the graph which also depicts that our data is trending towards an increase in the near future. Now I will be passing over my time to Zhan Long.

P2:

Slide 12:

Here i will explain about the main techniques that we have applied to attain our objective, to predict stock price.

Slide 13:

Before delving into our method used, we have to understand the basics of time series.

Time series can be decomposed into three factors:

* Trend
* Seasonality
* Residual

Firstly, Trend is the movement of the stock, whether a stock is moving upwards or downwards

Afterwards, Seasonality is where predictable and regular changes recur on a certain period

Lastly, residual can be simplified into the error between a predicted value and observed value

Slide 14:

Hence, the issue of SMA/EMA as mentioned above is that it is only able to predict the trend with certain accuracy. It does not consider the seasonality of the data or residuals of the data.

Therefore, we utilized ARIMA, AutoRegressive Integrated Moving Average as the first technique.

Slide 15:

ARIMA is a stochastic time series model used to capture complex relationships relying on three parts.

* AutoRegression, a process of regressing a variable on past values of itself, estimating the degree of a the time series as compared to white noise
* Integrated, used to eliminate seasonality from the time series
* Moving Average, used to reduce and eliminate random movement from time series.

Thus, we have applied ARIMA into the time series, with the result indicating the three factors of time series.

With the diagram on the bottom-left, the quantile-quantile (QQ) graph, it is able to tell the trend of the time series, pointing up.

The autocorrelation graph (ACF), it have indicated that there are no clear seasonality after applying ARIMA.

Lastly, the standardized residual shows that there are some outliers, with the rest of the error margin below 3

Slide 16:

Although there is a higher margin of error, it still satisfies the conditions for factors of time series, thus allowing a prediction with moderate accuracy.

Slide 17 & 18:

However, with the standardized residual, it can be shown that there are a higher variance with the recent data, and might lead to a less accurate prediction.

Hence, the model could be optimized when paired with model that can measure and model after the change in variance over time.

18:

GARCH or (Generalized AutoRegressive Conditional Heteroskedasticity) incorporates ARIMA, modelling the conditional change in variance and changes in time-dependent variance as well, measuring the volatility of the stock price movement, giving a range of the price movement.

Nevertheless, due to a huge dataset, the prediction of ARIMA-GARCH has a moderate accuracy, it could overestimate the prediction with history data.

Slide 19 & 20:

Therefore, with the use of neural network, specifically LSTM (Long Short-Term Network), it could predict future values based on past data.

LSTM, a chain-like structure with four layers of neural network, each with different functionalities

* A cell state, where information are being stored
* A forget layer, where information will be completed gotten rid of
* An input layer,where information will be mapped and new information will be updated to the state
* Output layer, where the stored information in cell will be filtered and output

By stacking LSTM, it is able to make the model deeper, recombining the learned representation from prior layers and create new representation at high levels of abstraction. Hence it allow further optimization of the Neural Network.

Slide 21:Predicted values of LSTM & Segue to further optimization

With ample dataset, LSTM is able to predict the stock with high accuracy, with a margin of errors. However, these models are unable to predict a drastic sudden movement similar to that of 2008 market crash or more recently the sudden increase in GME and AMC.

Slide 22:

From the graph, it is able to tell that there is some relationship between the quarter report of apple and the increase in stock price. So, we have assume that the news can be one of the predictor for price movement.

Using News as Predictor

Slide 23:

By using Natural Language Processing (NLP), we are able to convert an unstructured data into a structured data through a third-party application, classifying the news into two categories, positive and negative. Using daily returns as the target, it is able to derive whether there is a correlation between a positive news and a positive return.

With this, I will pass off to Hendy to continue explaining about the tools and effectiveness of the solution.

P3

In order to compare and check the accuracy of Arima and LSTM, we need to bring in news variables as 2 of the models do not consider news data as predictor to predict the stock price.

But in reality news affects stock value.

So we want check how well these model perform to predict bad and good news

How do we do it?  
We apply things outside of the scope of what we learn in this course which is Natural Language Processing.

We use sentimental analysis because reader who read news or company info through social media or any other platform affect the buy and sell action of a stock which affects the market trend.

VADER is the model we use for sentimental analysis in our project

→

We use NLTK which is a toolkit built for working with NLP in Python.

to convert news article to a scoring points system and then categorised it into either positive and negative news.

→

Compound is the scoring points and comp\_score is the category

In this case we filter out neutral news which has compound value of 0 as it doesn’t benefit in telling us how the stock will move, we output the result as a new column

→

We then use classification tree through what we have learnt in our course.

to check how accurately our Arima and LSTM model in predicts the bad and good news.

We set the sample size of the train and test sets to be 80% to 20% ratio

And use minmax to check the stock price increase up or down per day and transform the data to a suitable dataset to fit the algo

And use RTX which will be the daily difference as our X component

The result shows that ARIMA is the better model than LSTM as

Our training and test sets show 60% and 58% accuracy

While LSTM model shows lower accuracy in the test case which is around 41%

→

Suggestion

Use other predictors

Increase the window of target for classification to improve accuracy

→

In conclusion,

In the case of our project LSTM perform better overall

what we have learn through this project are even with the best model available rgt now, to predict the stock price, there are still a lot of unknown factor that we need to consider to adjust, As ARIMA and LSTM are well known method to predict stock price, but there are still information such as news and other information that might not be available to us for example a company finance reports.

Through this project we are able to practise how to handle different formats of data and information.

We are also able to implement the existence theory and check the accuracy of them and compare them.

Thank you for listening.