The German University in Cairo
Faculty of Management Technology

Department of Business Informatics
Artificial Intelligence - CSEN933
Winter 2024

Project Report

Markov Chain to Predict Price Fluctuations of Stocks

Submitted by:

Maya Mohamed Moneib	ID# 55-26782
Rawan Ahmed	ID# 55-13033
Seba Mostafa Said	ID# 55-19560
Jowayreyah Khaled Bassim	ID# 55-1184
Maryam Elgharraz	ID# 55-21792

Link to dataset
https://www.kaggle.com/datasets/vijayvvenkitesh/Microsoft-Stock-Time-Series-Analysis

Professor: Dr. Mohamed Karam Gabr
TA: Moustafa Ashraf
Date: 8/11/2024

1. **You should provide a discussion about the problem of choice, The challenges why this problem is important, and the methodology used in assessing such a problem.**

Business Problem:

This project's goal is to dive into stock market investments by predicting fluctuations in Microsoft's stock prices using a Markov Chain model. Stock market environments are known to follow perfect competition conditions, meaning they exist in highly dynamic environments where the price fluctuations tend to be rapid and unpredictable. The stock investors focus on enhancing their roles as sellers by selling stocks at peak prices and as buyers by buying stocks at lowest prices. This strategy maximises their investment gains and minimises their investment costs, leading to the increase of their profitability rates. Through the modelling of fluctuations' states and classifying them as 'Increasing', 'Decreasing', or 'Stable' prices, the investors are able to make the necessary decisions to maximise & optimise their return on investment. The model is implemented with the focus on Microsoft's stocks from April 1st, 2015 till March 31st, 2021.

Challenges:

As for why it is important to address such a problem, the prediction of stock prices is essential for the majority of financial stakeholders since the prices directly affect the risk management, profitability rates, and decisions taken within the financial markets.

The predictive insights help optimise risk management(RM) strategies since RM is about taking advantage of potential gains and minimising losses from investment choices. Our exploration of the dataset provides insights such as:

1. **Right-skewed data:** Meaning that the prices' means are higher than their medians, therefore the prices tend to have positive outliers. The distribution of stock prices (open, closed, low & high) indicates that there is a likelihood of high risk but also higher returns. So investors should keep track of the spiked prices.
2. **Volatile stock prices:** depends on relative standard deviation(RSD) which tells us the weight of STD of a column in the dataset compared to its average (mean). If the RSD is high then this indicates high price fluctuations and high risk.
   RSD= (STD/mean)*100

- The RSDs for open, high, low & closed are 52.78%, 52.91%, 52.65% and 52.78% respectively.
- The RSDs are relatively high indicating high volatility.
- As the volatility in prices increases over the six years, the risks increase. A RM strategy that can help mitigate risks can be a stop loss strategy which prevents getting stopped out during rapid fluctuations while also attempting to stabilise the investors' overall risk exposure.

Also, the trading volume can improve portfolio optimization performance since volume provides information about demand and supply. The standard deviation of volume is 14,252,660 which is very high, creating high portfolio risk as the Microsoft stocks are required to earn higher returns than bonds.

The methodology used in assessment:

- Firstly, the training dataset chosen from Kaggle is for Microsoft's stocks.
- The columns are firstly date which shows the daily recordings of the stock prices; however, the weekends are not included. Then open which shows the opening price, or the price that the stock starts with at the beginning of the day. Next is high which shows the highest price stocks managed to reach at any duration of the day. Moreover, low which shows the lowest price the stock managed to reach at any duration of the day. Also, close which shows the price that the stock managed to reach by the end of the day. The closing value indicates the stock's true value which makes it one of the most important values in our dataset. And lastly, volume which represents the amount of stocks or shares that were sold during the day. The higher the volume, the more people are interested. The less the volume, the less people are interested.
- Furthermore, data statistics such as mean, standard deviation, frequency, mode, etc. are shown for all columns
- The range of stocks' dates are shown from April 1st, 2015 till March 31st, 2021. The fluctuations are over 6 years.
- The entropy for open, close, and volume are shown.
- **Data visualisation** is performed using correlation matrix for all columns except the date.

- Then plotting graphs for all columns separately as Y axis while the date is the X axis.
- **Prediction and analysis** are done by creating two columns as trading_gap which subtracts the opening price of the current day from the closing price of the previous day, and gap_label which classifies each trading_gap as “Up”, “Down” or “Stable”
- Plotting graph to show trading gap fluctuations and count of trading gaps in bar chart
- **Markov Model building** is done using previous and current gap_label values and transition matrix is shown
- The model is evaluated through several techniques including confusion matrix, training and testing the data, and k-folds in order to determine if the results are satisfactory to us or not

2. You should provide a discussion about the reasoning behind the choices made for the data set of Kaggle, in order for it to be taken as the data set for this case study.

The reasons why this training dataset was chosen because:

1. Microsoft is one of the largest technology companies worldwide which has many business endeavours from personal computing products to cloud products. It is considered a good long-term investment as its revenue in 2015 was $93.58B and continued growing till it reached $168.02B in 2021. Therefore, it maintains its attractive track record with diversified product portfolio and its increasing revenue growth.
2. **Time-series forecasting:** For stock prices, time-series datasets are very useful as they identify patterns and market trends necessary for investors when making decisions. The methodology assists MM by providing previously observed stock values in order to predict the unknown prices. As the prices change over time, we are able to predict whether the new stock prices will fall under “Increasing”, “decreasing”, or “Stable”. The dataset helps with long-term investments as it focuses on stock values for six consecutive years, uncovering the different seasonal trends that occurred from April 1st, 2015 till March 31st, 2021. In our case, the impact of COVID-19 pandemic was

apparent in our training and testing datasets for the Markov modelling. For example, throughout the years, there has been an upward trend in prices except for 2020, and also the trading gap fluctuations were more rapid during Covid.

3. **You should provide a discussion about the importance of the attributes which are chosen as the attributes upon which the predictive model is based, alongside the attributes to be predicted.**

Firstly, our dataset includes the following attributes:

1. Date: this shows the daily recordings of the stock prices; however, the weekends are not included

2. Open: this shows the opening price, or the price that the stock starts with at the beginning of the day

3. High: this shows the highest price stocks managed to reach at any duration of the day

4. Low: this shows the lowest price the stock managed to reach at any duration of the day

5. Close: this shows the price that the stock managed to reach by the end of the day. The closing value indicates the stock's true value which makes it one of the most important values in our dataset

6. Volume: this value represents the amount of stocks or shares that were sold during the day. The higher the volume, the more people are interested. The less the amount of volume, the less people are interested

We could not choose the following attributes:

- Low or High (attributes that represent the lowest/highest price the stock reached throughout the day) as price fluctuations throughout the day does not determine the final price. The final price is represented by the Close.
- Volume: as shown in the data visualisations, volume had almost no correlation with the close price as volume is more indicative of activity and people's interest in the stock rather than what the price will be. Increase/decrease in volume is an outcome of factors such as customer expectations and other activities rather than an outcome due to price increase/decrease

- Open or Close: while we do need them to know the market fluctuations, they do not create discrete states. However, we do need them in order to create the attribute we want as our input and output.

There was no attribute that showed the price fluctuations from one day to the other, so we created two new attributes: Trading_Gap and Gap_Label. Trading Gap in the stock market is the difference between the open price of the new day and the close price of the previous day. Trading gap shows the fluctuation that happens in the stock price from day to day. The Gap Label discretizes the Trading Gap into 3 values:

- Up: meaning that the open price of the next day is greater than the close price of the current day. It signifies that the stock price has increased
- Down: meaning that the close price of the current day is greater than the open price of the next day. It signifies that the stock price has decreased
- Stable: meaning that there is no change. The open price of the next day is the same as the close price of the current day

Since our goal is to predict stock price fluctuations, we wanted to choose an attribute representing the stock price fluctuations today as the input, and the stock price fluctuations of the next day as the output.

Hence, our input attribute is the current Gap_Label, and our output attribute is the Gap_Label of the next day. This will be managed by using the same column but for the input (current), we will use it as it is; however, for the output attribute (next), we will Gap_Label but shift one down.

Our output will include the three states (Up, Down, Stable) along with a transition matrix that shows the probabilities of the stock price increasing, decreasing, or remaining stable the following day, based on whether the current day's price increased, decreased, or remained the same.

4. **In case a variation of the Markov Model is utilised (HMM, CTMC, or MDP for example), an introduction to the variation used should be provided, alongside an argument about the suitability of the variation to the problem nature.**

Our chosen model was a Discrete Time Markov Chain also just known as Markov chain which is the default type of markov model. It's a sequence of random variables, known as a stochastic process, in which the value of the next variable depends only on the value of the current variable, and not any variables in the past. So simply we move in steps or time intervals to the next state and we only care about the state we are currently in and the one we are going to next, any other information is irrelevant, and all states are observable and there are usually no hidden patterns. The reason as to why we chose to go with a markov chain and no other variation of markov models was because our dataset contained daily changes, changes that happen at a certain interval and not at any time. We found that no other model was applicable, as we had no hidden patterns so therefore a hidden markov wouldn't work, as well as since as mentioned change was in intervals a continuous model where transitions happen at any time is also inapplicable.

5. **You should provide a discussion about the conclusions documenting the end results and probabilities produced by the constructed Markov Model showcasing whether the achieved results are satisfactory or not.**

We used three different validation strategies which are the confusion matrix, 80% 20% Train Test split and K-folds cross validation. All of these strategies are different ways in measuring our model's accuracy so we will go through each in the following part:

1- Confusion Matrix:

- The confusion matrix is one of the classification tools that is used to evaluate the accuracy of the model by checking whether the predicted label is similar to the actual or not. The results are classified to four categories which are the True Positives (TP) where the actual is positive result while the output is positive too, True Negatives (TN) where the actual is negative and the predicted value is also negative, the False Positives (FP) where the actual is negative and the predicted is positive and lastly the False Negatives (FN) where the actual is positive but the predicted value was negative.

- Applying this to our model, the confusion matrix categories represented whether the model accurately predicted the transition from a state to another. The TP displays a correct prediction where the model predicts the correct transitions between the different states.TN shows how accurately the model predicted that no change will happen and that we will remain in the same state. FP displays how much the model predicted a transition that didn't occur and the FN displays the no change the model expected while actually there was a change and a transition occurred.

- Moving to the results, it has been observed that there is price fluctuation leading to a positive trend where the prices tend to increase more and more. The confusion matrix accuracy is 54.64% which is not really satisfactory as an accepted accuracy should range from 70% - 80% so this means the model needs improvement.

```
Confusion Matrix with TP, FP, FN, TN:
          Up  Down  Stable
Up       815    10       0
Down     656    10       0
Stable    19     0       0

Accuracy: 0.5463576158940397
```

- Comments: This shows that the Markov model predicted that the Markov chain predicted that the price will always increase, disregarding all the decreases in the market which shows that this Markov model is not able to accurately predict market place decrease/stable fluctuations. This is due to a weakness in our dataset that the prices tend to have a positive trend overall; hence, the decrease in prices was severely ignored. In addition to that, there are also other factors playing into price increases including pandemics, political events, etc... which are not included in this dataset. That is a weakness in our dataset.

2- Train /Test Split

- The Train/Test split technique involves splitting the data set into two parts; a part for training the model and the second part is to test the model on the rest of the data set. Usually, 80% of the data is used for training and 20% is used for testing.

- We used this approach to test the Markov model we implemented in order to know how accurate the model is in predicting unseen data and its generalizability.

- Regarding the results, the train – test technique accuracy is 55.30% which is slightly higher than that of the Confusion Matrix but it still needs improvement.

3- K-Fold Cross Validation

- K-Fold Cross Validation is one of the model evaluation techniques where the data is split to “K” subsets/folds, the model is trained on “K-1” and the testing is done on the remaining subset. This is repeated “K” times where each time the testing subset changes.

- Applying this to our project, we split the data set into 5 folds where every time a fold was tested based on the other 4 trained folds. We have chosen to test using this technique as it is more reliable than a single train-test split, we will test the model on different parts of the data in order to make sure it is accurate.

- The K- Fold Cross Validation’s accuracy is 54% which is quite similar to that of the confusion matrix showing that the Markov model still need improvements or another data set can be considered as the accuracy is still below average even after depending on a reliable technique like the K- Fold cross validation.

Conclusion:

In this experiment, we were able to conclude that the Markov Chain produced an accuracy of 54% due to it only predicting that the stock price was always increasing; it disregarded all the times that price stayed as it was or it decreased. For us, those are not satisfactory results at all.

This was due to that the data is right skewed and that is normal in finance as inflation would always lead to increase in price. The data set we have chosen was too simple as it didn't have any attributes that explain the stock behaviour or any market features that could also affect the stocks. If our dataset had such attribute, we could have used a more complex version of the Markov chain which is the Hidden Markov Model (HMM) , which could have improved the results by having hidden states that incorporate the other hidden factors that affect stock prices

In fact, stocks rely on all events; whether those happened in the past or in the present; hence, a markov chain might not be the best option as it does not take in mind the past given the present. Stock prices are in fact one of the hardest things to predict as prices rely on a lot of factors such as inflation, customer expectations, political events, pandemics, natural disasters, etc... Some of those factors are not able to be predicted; hence, it is not easy to predict with high accuracy. However, integrating several datasets that have the events of the world with the stock prices might help produce better results in the long run. As mentioned above, having other attributes from different datasets which represent different factors might help us find a hidden pattern and predict states based on them. In that case, a hidden markov might work better if integrated with other datasets.