Predicting JSE Stock trends

David Hagumuwumva
Deji Adebayo
Moayad Elamin
Muhammad Omer

December 2021

Abstract

Trading with securities and stock indices is becoming increasingly popular among investors. Stock exchange trading allows investors to hedge their market risk while also providing a good investment opportunity, and in some cases, they can take advantage of arbitration conditions. The Johannesburg Securities Exchange (JSE) is the world's 16th largest stock exchange and the largest of Africa's 22 stock exchanges. Market capitalization stood at R4 029 billion at the end of December 2003, up from R1 160 billion five years earlier. In this project, we aim to investigate methods to model and predict exchange value sequences for the JSE stock. In doing so, we employ Markov Chains, which is commonly used in this context, and benchmark our model with other popular machine learning techniques. We find that Markov Chains are very effective in this context, given its capacity to model time series data.

1 Introduction

Portfolio managers must have a thorough understanding of how equities are traded on the stock exchange, and they employ a variety of strategies for trading securities, the most common of which are: Active trading during day, trading through usage of a technical analysis, dollar-cost average, and Buy & Hold.

The Johannesburg Stock Exchange (JSE) describes itself as the ”engine room” of the South African economy, providing an orderly market for trading securities. Its primary function is raising of capital by re-channeling cash resources into productive economic activity and building the economy while improving employment opportunities and wealth creation.

This project aims to analyze and predict the stock option returns for the Johannesburg stock exchange (JSE) using Markov chain and machine learning prediction algorithms. It will compare the performance of Markov Chains with commonly used Machine Learning techniques, Linear Regression, Support Vector Machines and Logistic Regression.

In the next section the **Problem Statement** is formalized and research questions and objectives are presented. The section after it a **Literature Review** will be conducted on stock performance prediction and on Markov chains, Linear Regression, Support Vector Machines and Logistic Regression. The **Methodology** will give a detailed theoretical overview of the concepts that will be used in the project. We will discuss the design choices and outcomes of this the project in the **Results Discussion** section, explaining data pre-processing, the baseline models we investigated, and our Markov Chains model. The **Conclusion Future Work** reiterates what was detailed in this report and gives a statement on potential future work.

2 Problem Statement & Research Objective

The health of a country's stock market is an essential indicator of that country's economic well-being. Starting companies can register in the market, raising significant funds through initial public offerings IPOs that would only be raised through finding investors and borrowing from banks. These funds will not cause a worry to these companies as they can be returned naturally with trading and they will gain more money as they perform well providing their services or product. Individuals are more inclined to invest their savings in activities that will mobilize the economy, as investing in the stock market will create opportunities for more companies to emerge and for those individuals to use their savings in mobile methods other than assets or keeping the money stationary in local safes. Foreign direct investment is also more probable to be focused on the companies that are performing well in the market. The transparency that the stock market will bring to the country's investment environment will encourage more foreign investors to put their money into the country's economy creating mobility and jobs that will not be there if the investment is only handled with banks and venture funders.

The problem of predicting the stock market has been historically divided into two camps. The fundamental camp uses non-structured information like financial reports, country policies and earning reports to predict stock market mobility. The technical camp on the other hand uses historical market prices and movement to predict market mobility. The methods for predicting the stock market vary very much and their importance has been introduced in previous research [1]. Statistical methods have been introduced to estimate the stock market that includes regression analysis, Markov chains and random walks.

The Johannesburg Stock Exchange JSE is the largest stock exchange in Africa with a market capitalization estimation of over US$1,000 Billion. It was formed in 1887, joined the World Federation of Exchange in 1963 and implemented an electronic trading system in the early 1990s. Some research has been conducted in predicting the JSE stock market that included sentiment analysis [2], Price to earnings ratios [3] and neural nets [4].

Yahoo Finance is one of the leading websites that provide historical stock market data. JSE historical data can be obtained that include opening, closing, high, low and adjusted closing prices as well as stock exchange volume at a daily rate for many years.

There are four major methods that an investor uses to make decisions on the stock market. The first is the stock momentum, the speed at which the stock is changing. This can help in identifying stocks that are entering a stage of the rapid increase in price to then capitalize on this movement to buy at a relatively low point and sell when the momentum starts to decline to avoid losses. Mean reversion, the tendency of a specific variable to converge to a certain value is another important method that can be used by investors and it can be found in currency exchange rates, gross national product, and interest rates. This method is used for stocks with very high or low prices that are skewing from the average historical price with some distance and investors can make decisions on those stocks to avoid losses and increase gains. Martingales, the tendency of a stock to only depend on the current price "present in random walks and MDPs" is another method that some investors can make use of as previous prices tend to not affect the future price of the stock. The final method is searching for low-priced stocks and depending on mis-pricing adjustments to kick in on these stocks and investors to make profits from these abnormally low stocks.

This research has the objective of answering the following questions:

1. Can momentum be coupled with Markov Chains analysis to get accurate market behaviour prediction.

2. Are Markov Chains comparable in performance to Linear Regression, Support Vector Machines and Logistic regression as examples of Machine Learning methods.

3 Literature Review

All current advancement in stock market prediction is a direct descendent from Burton Malkiel's book "A Random Walk Down Wallstreet" in which he claims the impossibility of predicting stock prices using only historical data, comparing it to the statistical Random Walk process. Following his work, two distinct camps emerged for predicting stock mobility, fundamentalists, and technicians. The fundamental analysis revolves on using predicting the stock behaviour using information about the company itself, not the stock. This information varies from previous fiscal performance, credibility, news reporting and others. The most famous fundamental analyst is Warren Buffet with his Buffet Indicator which uses market capitalization to GDP ratio to predict the company performance. The information about the company is used to predict the actual value of the stock which in return will be used in comparison with the current value to predict market mobility. The technical analysis took many twists and turns in the history of stock market prediction from time series analysis [5], fuzzy logic and expert systems [6] and recently with combining different statistical methods to get better prediction accuracy [7].

Markov chains [8] are the fundamental random process that can be used to model a wide range of random processes successfully. In a Markov chain, the next state of an object traversing a state-space depends only on the current state that it is in. Modelling a stock market predictor as a Markov chain helps us to comply with the rules that Malkiel's book outlined for the prediction process. Work in applying Markov chains to stock prediction has been carried out to various degrees of success. Svoboda M. et.al [9] worked on predicting the stock market index for Prague stock exchange PX and analyzed different investment strategies based on their proposed model. Kostadinova, V et.al

[10] Implemented complex Markov chains to predict financial-economic time-series activity on global stock market indices. Different Machine learning techniques were applied in predicting stock market mobility and performance. Upadhyay A. et.al [11] Used multinomial logistic regression to predict the Indian market performance and introduce analysis of the factors that affect a stock's performance. Lin Y. et.al. [12] Analyzed the stock features using correlation filters then carried out a prediction process using support vector machines (SVM). In this research, we will be comparing Markov Chains, Support Vector Machines and Logistic regression in the task of predicting stock performance.

4 Methodology

Starting with data pre-processing, feature design, and labeling, we propose to design a multi-class problem of five classes: very high, high, low and very low, to represent the stock's momentum, in addition to "stationary" label which shows that there's no change. The data, extracted from Yahoo finance, represents stock prices of the Johannesburg Stock Exchange LTD in the global financial market. We design our features to be in a form of time series, with the weights being the daily difference between the stock's prices at the close of the market. We then proceed to create the major four classes using this data by ranking the weights into four quantiles, 0 to 25 percent, corresponding to very low, 26 to 50 percent, corresponding to low, 51 to 75 percent, corresponding to high, and 76 to 100 percent, corresponding to very high. These classes represent the stock's momentum as described above in the problem statement and literature review.

We propose to tackle this problem using Markov Chains. Markov Chains are stochastic processes that have the Markov property, which could be explained as: "The future is independent from the past given the present".

We aim to do that by representing returns as state instances and their respective momentum as the transition probability of the Markov Chain. We follow a similar approach to [13], as we use a finite, stationary, discrete-time Markov Chain, which can be written as follows:

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0\} = P_{ij}$$

Where X_i is the return value at state i, and j in P_{ij} is the next state after i.

For this problem, we construct a transition matrix of P_{ij} values $P \in \mathcal{R}^{mxm}$:

$$P = \begin{pmatrix} P_{11} & P_{12} & ... & P_{1m} \\ P_{21} & P_{22} & ... & P_{2m} \\ .. & .. & .. & .. \\ P_{m1} & P_{m2} & ... & P_{mm} \end{pmatrix}$$

Next we construct a state vector Q using the Markov property:

$$S(n) = S(n-1)P$$

$$Q = \begin{pmatrix} S(1) \\ S(2) \\ S(3) \\ ... \\ S(m) \end{pmatrix}$$

Taking the limit of Q, we can obtain the transition probability matrix which is the momentum of the stock return that we need to obtain:

$$\lim_{n\to\infty} Q^n = \begin{pmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1m} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2m} \\ .. & .. & .. & .. \\ \pi_{m1} & \pi_{m2} & \cdots & \pi_{mm} \end{pmatrix}$$

We will benchmark the model we create with Markov Chains on this data and compare it to the performance of other three popular machine learning models, Linear Regression, Multi-class Logistic Regression, and Multi-class Support Vector Machines. Our reasoning is that this will allow us to better understand the benefits of using Markov Chains, and whether it performs better on time series problems than regression models which have been commonly in use in the industry for decades.

Linear regression is a machine learning technique that solves a means least squares problem formed using the features and labels of given data.

Having a feature matrix from data X, labels y and weights w, linear regression using gradient descent to iteratively solve for optimal weights w^{LMS} that predicts the labels accurately [14]:

$$w^{LMS} = (X^T X)^{-1} X y$$

Logistic regression is a machine learning model using regression analysis to predict boolean pre-defined labeled classes.

Logistic regression is made possible due to the representational capabilities of the sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Building on it, Logistic Regression is made possible by optimizing the objective [14]:

$$L(w) = -\sum_{n} y_n log(\sigma(w^T x_n)) + (1 - y_i) log(1 - \sigma(-w^T x_n))$$

This is called the cross-entropy error function, where y_n refers to the class labels, x_n refers to the features, and w is the weights vector we try to get optimal values for in order to get good predictions.

For multi-class Logistic Regression, known as multinomial regression, we simply optimize for different classes using a similar objective, with the only difference being that we use a different function for classification called the softmax function:

$$\mu(x) = \frac{e^{w_c^T x}}{\sum_c e^{w_c^T x}}$$

Where c refers to the class we're trying to predict the likelihood for.

Support Vector Machines are an improvement on this logistic regression, as it optimizes over logistic regression to ensure the decision boundaries are situated optimally. This allows us to create a model that offers balance between bias and variance.

Support Vector Machines can solve for problems where is not linearly separable, and can be easily extended to solve for multi-class problems, to do this we need to solve the following problem [14]:

$$\min_{w,b,\delta} \frac{1}{2} ||w||_2^2 + C \sum_n \delta_n$$

$$y_n[w^T x_n + b] \geq 1 - \delta_n$$

$$\delta_n \geq 0$$

Where δ, the slack variable, is a parameter that insures the first terms is not out of bounds. y,x,and$w$ follow similar definition as to the Logistic

Regression case. And finally, C is a hyperparameter that we're free to choose to tweak our model. This problem is generally solved using what's known as a quadratic program, it is also convex over x and y and thus has a unique solution.

5 Results and Discussion

5.1 Data Pre-processing

We start by obtaining the JSE data from $YahooFinance$. The data was cleaned, removing zero-values from the data and storing in respective dataframe. The data contains daily historical data on the JSE stock from September 19th 2007 untill the present day. There are 7 features in the data, *Date*, *Open*, *High*, *Low*, *Close*, *AdjustedClose* and $Volume$.

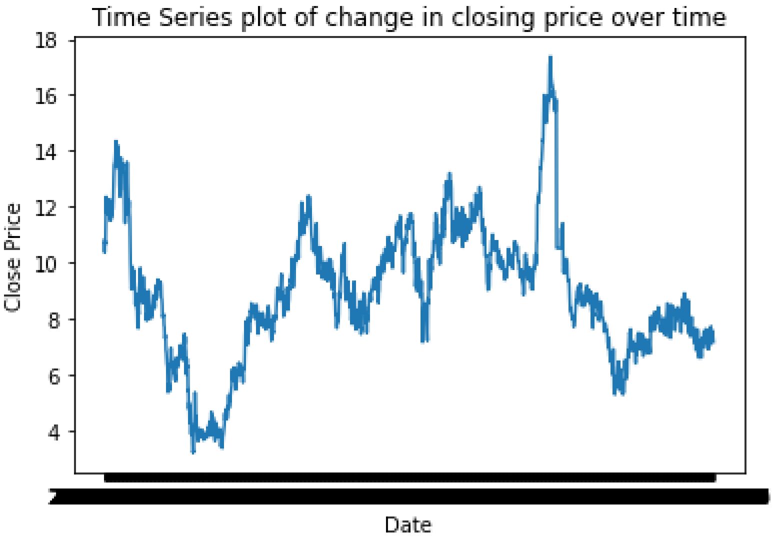

Figure 1: Time series of the *Close* feature

The datapreprocessing stage also included the creation of Categorical data from the numerical data. For each day a difference value was calculated by subtracting that day's close price from the previous closing price to get that day's change value. Using the %25, %50 and %75 percentiles of the data, four categories were created to describe the calculated change, $VeryLow$

describing the lowest %25 of the data, *Low* which describes the next %25, $High$ describing the section of the data that is between the $\%50^{th}$ and $\%75^{th}$ percentiles, and the top %25 were assigned a $VeryHigh$ Label, and finally, a "stationary" label was created when there's no change in the value. A comparison was carried out from the labels obtained from using the *Close* feature and those obtained when calculating difference with *AdjustedClose* feature. The comparison shows that the two resulting label vectors are similar which allows for the usage of *Close* based labels as target.

The dataset contained 3000 datapoints with 4 features chosen for training, *Open*, $High$, *Low*, and $Volume$. The *Close* and $DifferenceinClose$ features were obtained from the data, with $DifferenceinClose$ being used for the classification based models and *Close* being used for the regression model.

5.2 Baseline Calculation on the Labels

3 models were used to create a baseline for the Markov Chain model, a Logistic Regression model, a Support Vector Machine model and a Linear Regression transformed model.

5.2.1 Logistic Regression & SVM

Using *sklearn*, a Logistic Regression classifier as well as an SVM classifier were trained on the data, which was split into training and testing features using *sklearn*'s *train_test_split* method. Those models didn't perform on the task, reverting to labeling everything as "Stationary" and getting a zero accuracy.

5.2.2 Linear Regression

Using the original *Close* data, we use a regression model to predict the numerical data, then the difference label transformation is applied on the resulting predictions to get the final classification labels for each test point.

Regression accuracy of the model was gauged as %97.3, which was promising for great performance on the classification labels. After transforming the values into labels and calculating the accuracy vs. the true labels, a some what disappointing accuracy of approximately %60 was calculated from the full system. Although the accuracy is quite low, the model is a hard classifier, if there was some relaxation in the model, the accuracy would have increased drastically. This was also a major increase on the previous models and has great predictive ability than the other two models.

5.3 Markov Chains Model

Finally, we use the labels we've created earlier to create sequence labels for the Markov Chains models.

First, we use one step back Markov Chain, which performed well, with a %44 accuracy, then, we use two step back Markov Chain, which slightly improved the performance to a %45 accuracy. We then use three step back and four step back models that scored %50 and %55 accuracy respectively.

While the four step look back Markov Chain gave the better performance, it does not generalize well on test data. In all cases we trained the transition matrix and tested on a 80-20 train test split. For the four step back, there were sequences in the test data that were not in the training data, which means this is an overfitted performance, that need more data and is very context specific. We strive for generalization and less data usage.

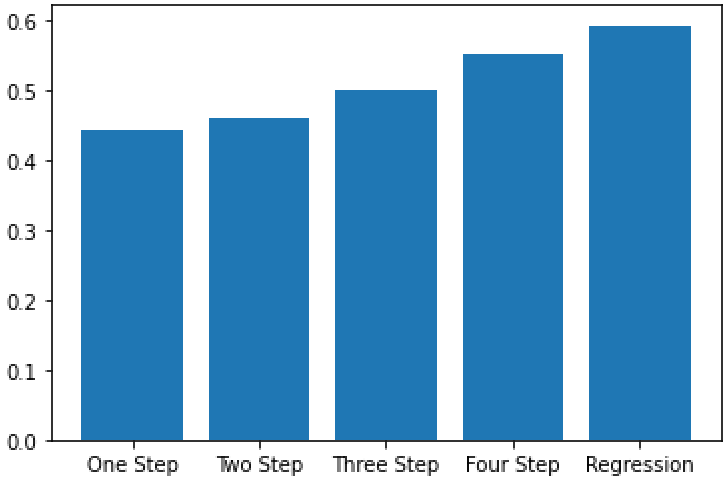

Figure 2: Performance of the the three Markov Models compared to Linear Regression

6 Conclusion & Future Work

In conclusion, we use Markov Chains to predict and analyze stock performance for the Johannesburg Stock Exchange. We elected to use the momentum model, which performs analysis on the performance of the stock by categorizing that performance into four distinct classes. These classes are, very high, high, low, and very low. For generalization and accuracy, we add a fifth category, "stationary", which refers to the instances where the stock won't change. In our data preparation, we use the difference between the stock prices on a daily basis at the close of the market. Finally, we benchmark this dataset on the model we've created along with three popular machine learning models, Linear Regression, Multi-class Logistic Regression and Multi-class Support Vector Machines SVMs.

Our results show that popular machine learning classification models fail to predict stock momentum, and that Linear Regression can be used to predict the prices, but also performs relatively poorly on momentum labels. Markov Chains however, succeed in modeling the momentum sequences and generalizes better than Linear Regression. This being said, it was found that Markov Chains requires more data.

Going forward we suggest the following:

1. Changing the Sequence Instead of just using the change in close as feature, we can add a collection of changes in all features as one label, and create sequence from them.

2. Randomizing the Sequence Length We can select for each data point a random length for the sequences "bounded to 3 of course" to test for robustness on different data sources.

3. Trying other statistical models Random Walks and Hidden Markov Models are models that can have provided better performance hat Markov Chains in similar problems, we can explore how well they solve the problems we faced.

7 Work Division

The workload was divided into pairs, the (Muhammad-Moayad) pair worked on the MC model design and data preprocessing. The (David-Deji) pair worked on the rest of the code including the benchmark models and experiments, however, the four of us were all involved in all parts of the code through revision, brainstorming, and recommendation.

1. Moayad Elamin Put together Problem Statement and Literature Review, designed presentation. Markov Chain model code. Data Preprocessing.

2. Deji Adebayo Put together Abstract and Introduction. Benchmark models code.

3. David Hagumuwumva Put together Results and Discussion. Benchmark models code.

4. Muhammad Omer Put together methodology, and revise the reports (writing, structure, etc). Markov Chain model code. Data pre-processing.

References

[1] A. Krause, "An overview of asset pricing models," *University of Bath School of Management. UK*, 2001.

[2] S. Bogle and W. Potter, "Sentamal-a sentiment analysis machine learning stock predictive model," in *Proceedings on the international conference on artificial intelligence (ICAI)*, p. 610, The Steering Committee of The World Congress in Computer Science, Computer . . . , 2015.

[3] T. I'Ons and M. Ward, "The use of price-to-earnings-to-growth (peg) ratios to predict share performance on the jse," *South African Journal of Business Management*, vol. 43, no. 2, pp. 1–10, 2012.

[4] K. Ayankoya, A. P. Calitz, and J. H. Greyling, "Real-time grain commodities price predictions in south africa: a big data and neural networks approach," *Agrekon*, vol. 55, no. 4, pp. 483–508, 2016.

[5] M. C. Angadi and A. P. Kulkarni, "Time series data analysis for stock market prediction using data mining techniques with r.," *International Journal of Advanced Research in Computer Science*, vol. 6, no. 6, 2015.

[6] M. A. Boyacioglu and D. Avci, "An adaptive network-based fuzzy inference system (anfis) for the prediction of stock market return: the case of the istanbul stock exchange," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7908–7912, 2010.

[7] M. R. Hossain, M. T. Ismail, and S. A. A. Karim, "Improving stock price prediction using combining forecasts methods," *IEEE Access*, 2021.

[8] K. L. Chung, "Markov chains," *Springer-Verlag, New York*, 1967.

[9] M. Svoboda and L. Lukas, "Application of markov chain analysis to trend prediction of stock indices," in *Proceedings of 30th international conference mathematical methods in economics. Karviná: Silesian University, School of Business Administration*, pp. 848–853, 2012.

[10] V. Kostadinova, I. Georgiev, V. Mihova, and V. Pavlov, "An application of markov chains in stock price prediction and risk portfolio optimization," in *AIP Conference Proceedings*, vol. 2321, p. 030018, AIP Publishing LLC, 2021.

[11] A. Upadhyay, G. Bandyopadhyay, and A. Dutta, "Forecasting stock performance in indian market using multinomial logistic regression," *Journal of Business Studies Quarterly*, vol. 3, no. 3, p. 16, 2012.

[12] Y. Lin, H. Guo, and J. Hu, "An svm-based approach for stock market trend prediction," in *The 2013 international joint conference on neural networks (IJCNN)*, pp. 1–7, IEEE, 2013.

[13] A. Fitriyanto and T. Lestari, "Application of markov chain to stock trend: A study of pt hm sampoerna, tbk.," in *IOP Conference Series: Materials Science and Engineering*, vol. 434, p. 012007, IOP Publishing, 2018.

[14] K. P. Murphy, *Machine learning : a probabilistic perspective.* Cambridge, Mass. [u.a.]: MIT Press, 2013.