Capitec stock prices modeling using regression methods

Sifiso Rimana

November, 2024

Abstract

In this paper, we discuss Capitec business overview and its recent financial achievements. We then analyze and model Capitec closing prices using its stock price data from Yahoo! Finance. After the analysis, we present the final model, and a formula which investors and traders can use to predict Capitec closing price.

Note

This article is for educational purposes only, and should not be used as a financial or investment advise, the author wrote the article as a way to practice what he learned in his Linear Algebra and Statistics at the University of Johannesburg during his second year studies (2024), and he combined the knowledge with what he is currently learning from Stanford's CS229 Machine Learning (can be obtained freely from YouTube). The models presented here are in no way perfect, and are for educational purposes only, please read with educational interest.

Introduction

Stock price prediction involves estimating the future value of a company's stock or the overall stock market. This process is crucial for investors and financial institutions as it aids in making informed investment decisions, managing risks, and ensuring financial stability. Accurate stock predictions can help investors maximize returns and minimize losses [1].

The methods used for stock price prediction include fundamental analysis, technical analysis, machine learning, and quantitative trading. Fundamental analysis evaluates a company's financial health by analyzing financial statements, management quality, industry conditions, and economic factors to determine the stock's intrinsic value. Technical analysis studies historical price movements and trading volumes to identify patterns and trends that might indicate future price movements. Machine learning and quantitative trading use advanced algorithms to analyze large financial datasets, including historical prices, financial news, and social media sentiment, to predict stock movements [3].

Capitec Business Overview

Attribute	**Value**
Company	Capitec Bank Holdings Limited
Ticker	CPI (Johannesburg Stock Exchange)
Current Stock Price	326.858 ZAR
Market Capitalization	376.19 billion (ZAR)
Beta	1.03, indicating that the stock's volatility is like the overall market.

Table 1: Summary of Capitec Bank Holdings Limited

Capitec Bank, established in 2001, is a prominent South African retail bank headquartered in Stellenbosch. It has rapidly grown to become the largest retail bank in the country by customer numbers, serving over 23 million clients as of October 2024 [4]. The Bank offers a wide range of financial services including transactional banking, savings & investments, credit facilities, and insurance products [5].

In 2019, Capitec expanded into business banking by acquiring Mercantile Bank, enhancing its services to small and medium-sized enterprises (SMEs). This acquisition enabled Capitec to offer tailored business banking solutions, including transactional accounts, credit facilities, and merchant services, supporting the growth and development of businesses in South Africa.

Capitec has demonstrated robust financial performance. In the six months leading up to August 31, 2024, the bank reported a 36% increase in interim profit, reaching 6.394 billion rand. This growth was attributed to reduced loan losses and an increase in transactions and commission income. This can be evidenced by the continued rise in its stock price.

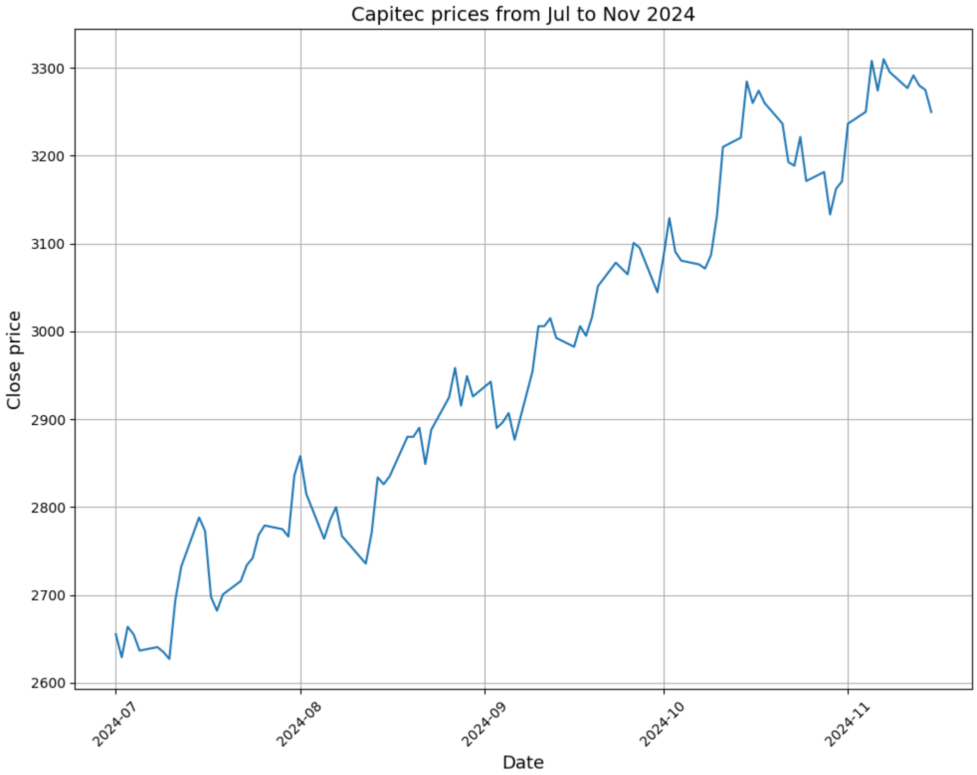

Figure 1: Capitec's stock price rising steadily from August to November

Capitec's success is largely due to its focus on customer needs, offering user-friendly digital platforms and maintaining extended branch hours to enhance accessibility. The bank's commitment to innovation and efficiency has solidified its position as a leader in the South African banking sector [6].

Data and Methods

In this paper, we attempt to estimate Capitec's closing stock price for the day using the classical linear regression method. Closing price prediction is import and useful for investors and traders, especially short-term traders, swing traders and position traders.[2]

We will also compare results obtained using linear regression with those obtained using a random forest regressor. The models will be judged by their train and test metrics, namely R^2 and root mean squared error (RMSE). The data was obtained from the package `yfinance`, which is free and open source, and the data is used as per Yahoo Finance's terms [7]. The data downloaded ranges from January 1, 2018, to November 18, 2024. Table 2 explains the variables and their meanings.

Since we will be predicting the closing price of Capitec for a given date, all the variables, except the closing price, will be treated as independent (also called predictors), and the closing price will be treated as the dependent variable (also called the response).

Variable	Explanation
date	The date in the format of Year-Month-Date. It tells us about the date for which the opening price, closing price, highest price, lowest price, and volume were recorded.
close	The closing price of Capitec stock at a particular date, measured in South African Rand (ZAR). It represents the value of a share at the end of the day when the stock exchange closes.
open	The opening price of Capitec stock on a particular day, measured in ZAR. It is the value of a share when the stock exchange opens for trade and may differ from the previous day's closing price.
high	The highest price of a Capitec share on a particular day, measured in ZAR. It is the maximum price at which the stock was traded during the day and helps assess the general movement of the stock.
low	The lowest price of a Capitec share on a particular day, measured in ZAR. It is the minimum price at which the stock was traded on that day and is also used to understand the stock's overall movement.
volume	The volume of Capitec stock traded on a particular day, measured in the number of shares traded per day. It represents the total number of shares traded during the day and is a crucial metric for assessing market activity.
adjusted close	The adjusted closing price of Capitec stock, measured in ZAR. This value accounts for events such as stock splits, dividends, and rights offerings that might affect the stock's price. It provides a more accurate reflection of the stock's value over time compared to the regular closing price. We will not be using this variable for simplicity.

Table 2: Explanation of the dataset variables used in the analysis.

Mathematical Formulation of Linear Regression

Linear regression is a statistical model that estimates a linear relationship between a real valued response and one or more explanatory variables. The explanatory variables are independent, and are also known as the predictors, while the response variable is dependent. If the predictor is a single variable, then we have a **simple linear regression**, and a **multiple linear regression** when the predictors are two or more.

Linear regression falls into the Supervised Learning category of learning algorithms since it 'learns' from the labeled datasets. That is, given the labeled dataset, it learns the linear relationship that exists between explanatory variables and the response variable, if there is such a meaningful relationship.

We now discuss the mathematical formulation of linear regression. Given a dataset $\left\{\left(\mathbf{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}$ of m observations, where $\mathbf{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$, under the assumption of linearity, we want to find a linear (not necessarily deterministic) function $h : \mathbb{R}^n \to \mathbb{R}$, such that

$$h_{\mathbf{w}}(\mathbf{x}^{(i)}) = w_0\mathbf{x}_0^{(i)} + w_1\mathbf{x}_1^{(i)} + \cdots + w_n\mathbf{x}_n^{(i)} \qquad \text{for } i = 1, \dots, m \tag{1}$$

where, if we let our feature matrix to be $\mathbf{X} \in \mathbb{R}^{m\times n}$ matrix containing m observations of n independent explanatory variables used to predict $y \in \mathbb{R}$, then $\mathbf{x}^{(i)}$ is the i^{th} observation vector; $\mathbf{w} \in \mathbb{R}^n$ is our parameter vector (also called weights). So, compactly, equation (1) becomes

$$h_{\mathbf{w}}(\mathbf{X}) = \mathbf{w}^T\mathbf{X} = \mathbf{X}\mathbf{w} \tag{2}$$

We of course want to choose $\mathbf{w}$ so that equation 2 is as close to the true observed y as possible. A **cost function** is a function that measures how close the hypothesis function is to the corresponding true y's:

$$J(\mathbf{w}) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_{\mathbf{w}}(\mathbf{x}^{(i)}) - y^{(i)}\right)^2 \tag{3}$$

or in matrix form

$$J(\mathbf{w}) = \frac{1}{2m}\left[\mathbf{w}^T\mathbf{X} - \mathbf{y}\right]^T\left[\mathbf{w}^T\mathbf{X} - \mathbf{y}\right] \tag{4}$$

Learning Methods

The primary goal in most machine learning problems is to minimize the objective function at hand, by this, the algorithm learns the patterns hidden in the dataset. There are numerous ways or methods for our machine algorithm to learn the patterns in our data. In this paper, we discuss the Gradient Descent, Newton-Raphson's Method, and the normal equations.

Gradient Descent Method

Gradient Descent is a method that learns the parameters **w** by iteratively updating **w** according to this rule

$$\mathbf{w} := \mathbf{w} - \alpha\nabla_{\mathbf{w}}J(\mathbf{w})$$

where α is the learning rate, and $\nabla_{\mathbf{w}}J(\mathbf{w})$ is

$$\frac{\partial}{\partial\mathbf{w}}J(\mathbf{w}) = \nabla_{\mathbf{w}}J(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}\left(h_{\mathbf{w}}(\mathbf{x}^{(i)}) - y^{(i)}\right)\mathbf{x}^{(i)} \tag{5}$$

whose vectorized form is

$$\nabla_{\mathbf{w}}J(\mathbf{w}) = \frac{1}{m}\mathbf{X}^T\left[\mathbf{w}^T\mathbf{X} - \mathbf{y}\right] \tag{6}$$

The gradient descent algorithm is then summarized below:

Algorithm 5.1 (Gradient Descent)**.** The gradient descent method can be run until convergence (i.e., until the updated **w** is negligible): Run until convergence { $\mathbf{w} := \mathbf{w} - \alpha\nabla_{\mathbf{w}}J(\mathbf{w})$ } or we can choose the number of iterations to run the algorithm:

```
Initialize w at random
for i until the number of iterations:
    update w according to the gradient descent updating rule
return w
```

□

One major benefit of gradient descent is its ability to handle large datasets efficiently, and can also converge to a global minimum for convex functions, ensuring optimal solutions for problems like linear regression.

However, the algorithm is sensitive to the learning rate, which requires careful tuning—too high a rate can cause divergence, while too low can result in very slow convergence. There are many studies done for how to choose the learning rate, also, variants of gradient have been developed to improve the original one, and they include the stochastic and mini-batch forms, where only subsets of data are used per iteration, making it scalable and computationally less expensive.

Newton-Raphson's Method

Another method for learning is the Newton-Raphson's Method, which has the following updating rule:

$$\mathbf{w} := \mathbf{w} - \mathbf{H}^{-1}\nabla_{\mathbf{w}}J(\mathbf{w})$$

where **H** is the Hessian of $J(\mathbf{w})$, which is found to be

$$\mathbf{H} = \nabla^2_{\mathbf{w}}J(\mathbf{w}) = \frac{1}{m}\mathbf{X}^T\mathbf{X} \tag{7}$$

It's algorithm is given below:

Algorithm 5.2 (Newton-Raphson's Method)**.** The Newton-Raphson method can also be run until convergence (i.e., until the updated **w** is negligible): Run until convergence { $\mathbf{w} := \mathbf{w} - \mathbf{H}^{-1}\nabla_{\mathbf{w}}J(\mathbf{w})$ } or we can choose the number of iterations to run the algorithm:

```
Initialize w at random
for i until the number of iterations:
    compute the gradient and Hessian
    update w using Newton-Raphson updating rule
return w
```

□

A significant benefit of Newton-Raphson is its fast convergence rate near an optimal solution, since it makes use of both gradient and second-order information (the Hessian). This means it often converges much faster than gradient descent, especially for well-behaved convex functions.

The drawbacks are notable too, calculating the Hessian matrix is computationally expensive, especially for high-dimensional datasets, making the method less practical for large-scale machine learning problems. Also, the method may struggle in situations where the Hessian is not invertible or near *saddle points*, which can lead to divergence or unpredictable behavior.

Normal equations

We define design matrix $\mathbf{X}$ as an $m \times (n+1)$ matrix (where m is the number of training examples and n is the number of features, plus an additional column for the intercept term). Each row of $\mathbf{X}$ represents a single training example, including its features:

$$\mathbf{X} = \begin{bmatrix} --(\mathbf{x}^{(1)})^T -- \\ --(\mathbf{x}^{(2)})^T -- \\ \vdots \\ --(\mathbf{x}^{(m)})^T -- \end{bmatrix}$$

where $\mathbf{x}^{(i)}$ is a row vector containing the features of the i-th training example. The normal equations are derived to find the optimal parameters $\mathbf{w}$ that minimize the cost function $J(\mathbf{w})$ in linear regression. To minimize $J(\mathbf{w})$ (equation 4), we take the derivative with respect to $\mathbf{w}$ and set it to zero:

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{y} = 0$$

Solving for $\mathbf{w}$, we obtain the normal equations:

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

This provides a closed-form solution for finding the optimal parameters $\mathbf{w}$.

The normal equations offer some key benefits and limitations when compared to iterative methods such as gradient descent since they provide a direct solution without requiring iterations. This makes the method easier to implement for small datasets, where computational resources are not a concern. Additionally, the solution obtained from the normal equations is deterministic, as there is no dependency on hyperparameters like learning rate or initialization values.

However, the normal equations also have some significant limitations. Computing $(\mathbf{X}^T\mathbf{X})^{-1}$ has a time complexity of $O(n^3)$, which can become computationally expensive for datasets with a large number of features (n). Furthermore, if the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible, such as when features are *linearly dependent* or *highly correlated*, the normal equations cannot be directly used. This can be partially addressed by using regularization techniques or computing the **pseudo-inverse**.

Making use of ready made libraries or packages

Implementing algorithms from scratch is intensive, and may result in poor performing algorithms than those that are already built and packaged, in this paper, we utilize ready made packages to streamline the process, in particular, we use `sklearn` and `statsmodels`.

Regression Analysis

A linear regression model was fitted using the `statsmodels`'s ordinary least squares API. Recall the ordinary least squares formula:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{8}$$

Where $\mathbf{w}$ is the ordinary least squares estimator, $\mathbf{X}$ is our features matrix, and $\mathbf{y}$ is our response variable. The following regression table was obtained

	Coefficient	**Standard Error**	***t*-Statistic**	***p*-value**
const	0.4828	1.880	0.257	0.797
open	0.3722	0.043	8.705	0.000
previous close	0.7683	0.060	12.720	0.000
previous open	-0.0573	0.039	-1.461	0.144
previous high	-0.0399	0.049	-0.806	0.420
previous low	-0.0428	0.051	-0.842	0.400
previous volume	-2.564 $\times 10^{-7}$	3.71 $\times 10^{-6}$	-0.069	0.945

Table 3: Coefficients for the initial regression model.

The constant is estimated to be 0.4828, and the coefficient of open is estimated to be 0.7683, and so on. To summarize, if we denote close price as Y, we get the following formula:

$$Y = 0.4828 + 0.3722 \cdot X_1 - 0.0573 \cdot X_2 - 0.0399 \cdot X_3 - 0.0428 \cdot X_4 - 2.564 \times 10^{-7} \cdot X_5 \tag{9}$$

Where $X_1, \ldots, X_5$ represent open, previous close, previous open, previous high, previous low, and previous volume, respectively.

A key advantage of linear regression is its interpretability. For example, a unit increase in open price results in an increase of 0.3722 in the close price, whereas a unit increase in the previous open price results in a decrease of 0.0573 in the close price.

Statistically, a variable is considered a significant predictor if its p-value is less than a significance level α. Here, we choose $\alpha = 0.05$ and conduct the following hypothesis tests:

- H_0: X_i is not a good predictor for Y
- H_1: X_i is a good predictor for Y

We reject the null hypothesis (H_0) whenever the p-value is less than α. Based on this criterion, only open and previous close are significant predictors since their p-values are approximately 0.

The residuals plot (Figure 2) shows no apparent relationship in the residuals, implying a good linear model.

Model Improvement

To improve the predictive power of the model, the insignificant variables were removed, and the ordinary least squares model was refitted. The updated coefficients are summarized in the following table:

	Coefficient	**Standard Error**	***t*-Statistic**	***p*-value**
open	0.3474	0.042	8.186	0.000
previous close	0.6536	0.042	15.413	0.000

Table 4: Coefficients for the improved regression model.

The resulting simplified model is:

$$Y = 0.3474 \cdot \text{Today's Open} + 0.6536 \cdot \text{Yesterday's Close} \tag{10}$$

The residuals plot remains similar to the initial model, indicating that we have achieved similar predictive power with only two variables.

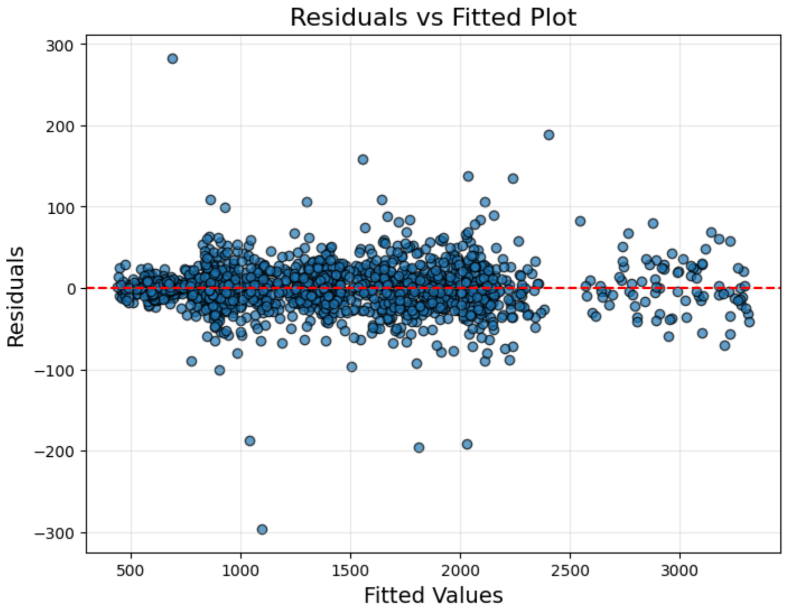

Figure 2: Residuals versus Fitted showing no apparent pattern in the residuals

Final Model Presentation

We now present the final models evaluated by their validation metrics, namely R^2 and RMSE.

Equations for R^2 and RMSE

The Root Mean Squared Error (RMSE) is a measure of the differences between values predicted by a model and the values actually observed. It is given by the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

Where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, and n is the number of observations. RMSE provides a measure of how well the model predicts the dependent variable, with lower values indicating better performance.

The R^2 (R-squared) metric, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

Where $\bar{y}$ represents the mean of the actual values. An R^2 value close to 1 indicates that the model explains a large proportion of the variance in the dependent variable, while a value close to 0 indicates that the model does not explain the variance well.

The metrics calculated using the training and validation datasets are summarized below:

	Training Data		**Validation Data**	
	RMSE	R^2	**RMSE**	R^2
Best OLS model	28.14	0.99785	25.22	0.99826
Best Random Forest Regressor	25.99	0.99817	29.07	0.99769

Table 5: Performance metrics for the final models.

Both OLS model and the Random Forest Regressor have relatively good predicting power as summarized by the Table 5. R^2 for both of them is approximately 99.8%, meaning that 80% of variation in closing price can be explained by the opening and yesterday's closing price. Morever, the RMSE of 25.22 means that the OLS model is wrong, on average, by approximately R25.22 in predicting the closing price. Whether the presented RMSE is good depends on the investor, and may be subjective. We plot the actual closing price versus those predicted by the OLS model in Figure 3.

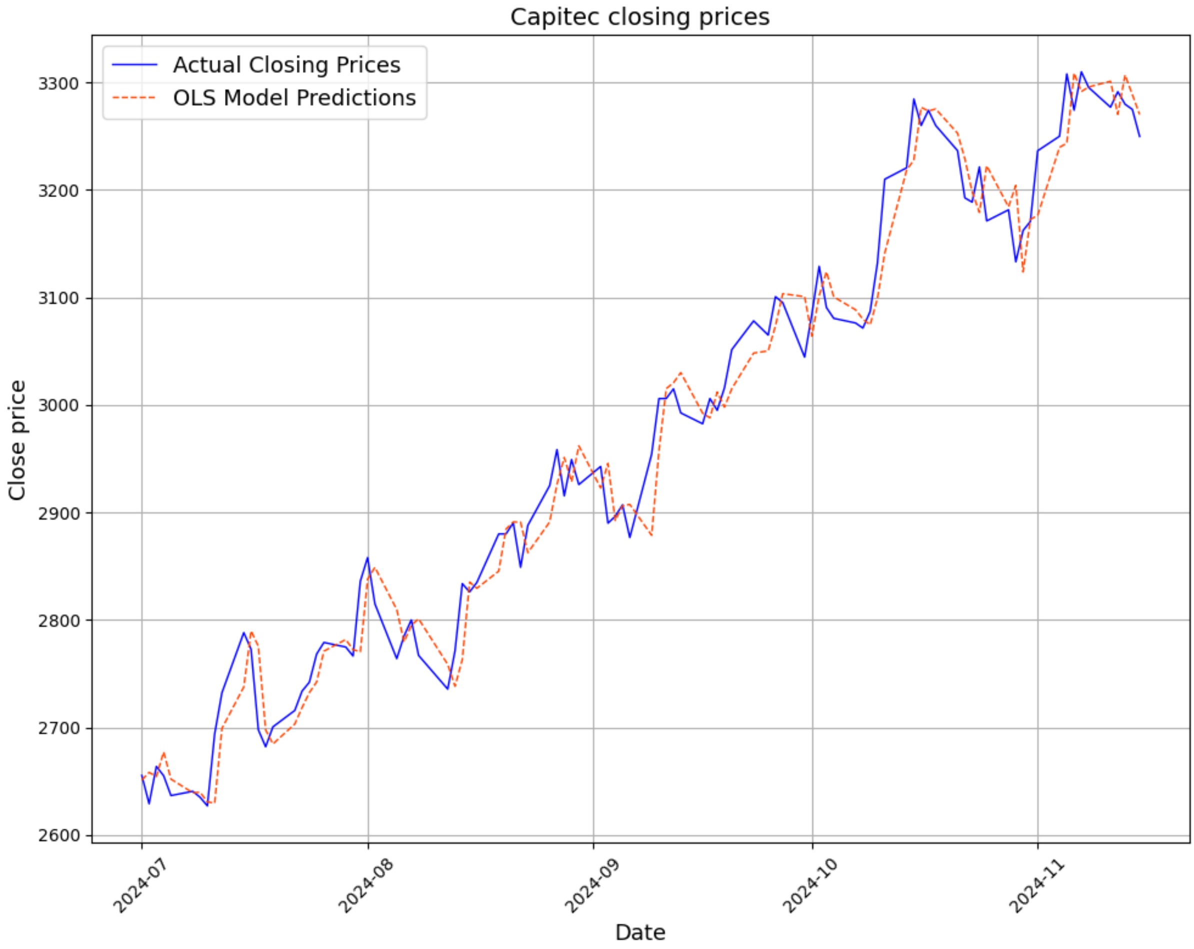

Figure 3: Capitec actual closing prices (blue line) versus those predicted by the best OLS model (red dashed line).

Using the Final OLS Model

We now present the final model for usability. If an investor, or a trader, wishes to predict tomorrow's closing price for Capitec, it is sufficient for them to obtain the yesderday's closing price and today's open price, and carry out the computation:

$$\text{Today's Close} = 0.3474 \times \text{Today's Open} + 0.6536 \times \text{Yesterday's Close} \tag{13}$$

as summarized in Table 4

Conclusion and Future Work

In this paper, we discussed Capitec Bank business overview and its recent success. We then obtained Capitec stock price data from Yahoo! Finance and fitted the ordinary least squares model using `statsmodels`, analyzed the statistically significant variables, and removed the insignificant ones to improve the model. We also trained `sklearn`'s Random Forest Regressor to compare it with the OLS model. Finally, we presented the model summary and a formula to predict the closing price.

There are many things that come into play when training models which were not discussed in this paper, such as regularization techniques, data dimensionality reduction using Principal Components. In the future, we plan to incorporate such and also perform Variance Inflation Factor analysis, correlation analysis, and other linear model assumption analysis. Furthermore, we plan to use incoporate classification models such as Logistic Regression, Gaussian Discriminant Analysis (GDA) and Support Vector Machines (SVMs).

References

[1] Shah, A. et al., 2023. *Identifying Trades Using Technical Analysis*, Mumbai, Maharashtra, India: VJTI College.

[2] R. Seethalakshmi, 2018. *Analysis of stock market predictor variables using Linear Regression*, School of Humanities and Sciences, SASTRA Deemed to be University, India.

[3] Wikipedia, 2024. *Stock market prediction.* [Online] Available at: `https://en.wikipedia.org/wiki/Stock_market_prediction` [Accessed 19 November 2024].

[4] Reuters, 2024. *South Africa's Capitec Bank reports 36% jump in half-year profit.* [Online] Available at: `https://www.reuters.com/business/finance/south-africas-capitec-bank-reports-36-jump-half-year-profit-2024-10-01` [Accessed 19 November 2024].

[5] Bank, C., n.d. *Business banking.* [Online] Available at: `https://www.capitecbank.co.za/business/` [Accessed 17 November 2024].

[6] Capitec, n.d. *About us.* [Online] Available at: `https://www.capitecbank.co.za/about-us/` [Accessed 18 November 2024].

[7] Yahoo, 2017. *Yahoo Terms of Service.* [Online] Available at: `https://legal.yahoo.com/us/en/yahoo/terms/otos/index.html` [Accessed 18 November 2024].

[8] Analysis, S., 2024. *Capitec Bank Holdings Limited (JSE: CPI).* [Online] Available at: `https://stockanalysis.com/quote/jse/CPI/` [Accessed 19 November 2024].