**ECOM198 – Midterm Coursework**

**Question N°1.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ticker | Mean Close Price | Mean Volume | Mean Return | Mean Volatility |
| APTA | 41.4 | 3,375,932 | -0.00733 | 3,400,753 |
| BHL | 6.14 | 1,066,990 | -0.00180 | 1,373,314 |
| GLEN | 465 | 44,029,628 | 0.000381 | 21,426,080 |
| OCDO | 682 | 3,288,344 | -0.000980 | 2,200,238 |
| VOD | 92.2 | 111,982,829 | -0.000586 | 77,032,514 |

* Glencore (GLEN) is a commodity trading and mining company, making it highly dependent on commodity price fluctuations. With a high average trading volume (44M), it is one of the most liquid stocks, meaning investors can buy and sell shares easily.
* Vodafone (VOD) is a telecommunications giant, providing mobile and broadband services globally. With the highest trading volume (112M) among the stocks, it is highly liquid, though its low share price suggests limited recent price appreciation.
* Ocado (OCDO) is a technology-driven online grocery retailer. Its higher share price (682) reflects its growth aspect, but with a lower trading volume (3.28M), it is not as frequently traded as GLEN or VOD.
* Bradda Head Lithium (BHL) operates in lithium mining, a sector linked to battery demand and EV market growth. It has a low share price (6.14) and low trading volume (1.06M), indicating lower liquidity and higher investment risk.
* Aptamer (APTA) is a biotechnology company focused on diagnostics and drug discovery. It has a low share price (41.4) and moderate volume (3.37M), typical of early-stage biotech firms that are still in R&D phases.

**Question N°2**

A graph of different colored lines

AI-generated content may be incorrect.

This graph illustrates the closing price of stocks over time. GLEN remains relatively stable with moderate fluctuations while VOD experiences a gradual decline. OCDO and BHL show sharp downward trends, indicating a sustained decrease in investor confidence or sector-related downturns. APTA also experiences a steep decline, which may reflect financial instability or external pressures affecting its valuation.

Thes second graph illustrates the trading volume of each stock over time. Glencore (GLEN) and Vodafone (VOD) consistently exhibit high trading volumes, indicating strong liquidity. VOD displays extreme spikes which could be attributed to company’s key financial events or macroeconomic news. Ocado (OCDO) and Bradda Head Lithium (BHL) show moderate trading activity with occasional spikes, suggesting that their trading volume is influenced by specific market events. Aptamer (APTA), on the other hand, experiences a significant peak in trading volume toward the end of the period, which could be attributed to a corporate event, A screenshot of a graph

AI-generated content may be incorrect.increased market attention, or investor speculation.

A diagram of a box plot

AI-generated content may be incorrect.

The final graph presents boxplots of log-transformed volumes for each stock, highlighting their distribution and the presence of outliers. APTA has a highly skewed distribution with extreme outliers, reinforcing the idea that its trading activity is driven by occasional, irregular spikes rather than steady investor interest. BHL also displays substantial variability, indicative of an illiquid stock that occasionally attracts increased trading activity. In contrast, GLEN, VOD, and OCDO have more stable distributions, with VOD still showing a high number of extreme values.

**Question N°3**

I applied two different linear models to attempt to predict the daily volume of the stocks.

1. Ordinary Least Squares (OLS): The objective is to minimise the sum of squared errors. In this assignment, I have chosen to use RMSE over MSE for evaluation because RMSE is easier to interpret, given the scale of our target variable. My reasons for choosing this model are the following. First, it is the most interpretable models, as the coefficients provide a clear understanding of variable relationships. Additionally, when predictor variables are not highly correlated, OLS provides reliable estimates. Finally, OLS serves as a foundation for applying regularisation techniques, making it easier to extend the model with Ridge or Lasso regression.
2. Ridge Regression (L2): Ridge regression is a regularised version of OLS that introduces an L2 penalty, shrinking regression coefficients to prevent overfitting and multicollinearity. This is particularly useful when predictors are highly correlated, as it stabilises the solution and reduces sensitivity to small variations in the training data. I opted for Ridge over Lasso because we have a limited number of parameters and want to avoid forcing coefficients to shrink to exactly zero. The regularization parameter, lambda, controls the penalty strength, balancing model fit and coefficient shrinkage.
3. *See the R code*
4. To evaluate my model, I choose three specific metrics to assess model performance.

First, the R-squared which measures how much variance in trading volume is explained by the model. Second, the RMSE because it is easier to interpret than MSE in this context and because it had the same unit as volume so easier to interpret and make actions based on that. Also, RMSE makes sure large errors get penalised more heavily than small ones. The RMSE measures the average prediction error in actual volume units. Finally, I also considered the VIF which identifies multicollinearity and helped refine predictor selection. My reason for including this metric is that, Close, Low, High and Open are very likely to be extremely correlated.

OLS performance

|  |  |  |  |
| --- | --- | --- | --- |
| Ticker | RMSE (Train OLS) | RMSE (Test OLS) | R² (OLS) |
| GLEN | 23,466,018 | 21,439,142 | 0.217 |
| VOD | 79,308,416 | 107,576,928 | 0.139 |
| OCDO | 3,451,403 | 3,220,358 | 0.345 |
| BHL | 2,783,352 | 1,018,713 | 0.276 |
| APTA | 713,077,001 | 1,012,910,035 | 0.773 |

From what we can observe, we have relatively low R-squared values across most stocks, meaning the model explains little variance in trading volume. VOD and APTA have extremely high test RMSE values, suggesting poor generalization to unseen data. APTA's RMSE is particularly extreme, meaning the model struggles with large variations in trading volume. Additionally, I computed the VIF values (See R code) for each stock and found that “Open”, “High”, “Low” and “Close” all exhibited strong multicollinearity which can affect the model performance. Indeed, these predictors displayed a VIF well superior to 100.

Ridge performance

|  |  |  |  |
| --- | --- | --- | --- |
| Ticker | RMSE (Train Ridge) | RMSE (Test Ridge) | R² (Ridge) |
| GLEN | 22,670,028 | 21,556,290 | 0.184 |
| VOD | 76,830,961 | 103,418,093 | 0.103 |
| OCDO | 3,341,362 | 3,108,104 | 0.343 |
| BHL | 3,264,163 | 3,707,463 | 0.358 |
| APTA | 19,367,770,939 | 26,813,762,121 | 0.718 |

Ridge Regression did not significantly improve RMSE compared to OLS. R² remained low, meaning Ridge was not able to explain more variance than OLS. RMSE for APTA exploded, suggesting Ridge over-penalized coefficients for this stock.

1. The model performance is still limited by the available dataset. Trading volume is likely influenced by external factors such as macroeconomic indicators (interest rate, GDP) or company specific events (financial reports, dividends). Because, we were not given these data, I decided to compute additional feature engineering. Initially, I had computed return and volatility which seemed to be the variables with the highest correlation with volume, however it was not sufficient to obtain a relevant model. Stock volume often exhibits autocorrelation, implying that the current volume is influenced by yesterday’s volume. Therefore, I computed the one-lag version of volume and created an interaction term between volume and closing price. The interaction term captures the relationship between volume and price movements. A high volume combined with price movements may indicated strong market sentiment or momentum.

OLS after feature selection

|  |  |  |  |
| --- | --- | --- | --- |
| Ticker | RMSE (Train OLS) | RMSE (Test OLS) | R² (OLS) |
| GLEN | 31,677,403 | 7,890,216 | 0.846 |
| VOD | 187,465,942 | 1,819,511,532 | 0.744 |
| OCDO | 17,336,645 | 6,980,772 | 0.759 |
| BHL | 11,590,792 | 11,463,997 | 0.420 |
| APTA | 459,930,461 | 94,802,802 | 0.792 |

R-squared improved significantly, showing better variance explanation. Additionally, the Test RMSE decreased, except for WOD and APTA, which still showed extreme errors. Since VOD and APTA exhibited larges distributions spikes, I removed extreme outliers and refitted OLS.

OLS improvement for VOD and APTA

|  |  |  |  |
| --- | --- | --- | --- |
| Ticker | RMSE (Train OLS) | RMSE (Test OLS) | R² (OLS) |
| VOD | 22,514,375 | 28,101,355 | 0.833 |
| APTA | 13,920,979 | 5,282,311 | 0.791 |

Significant improvement in test RMSE for both VOD and APTA. Additionally, R-squared increased a lot for VOD and remained stable for APTA, meaning that removing extreme outliers was an effective technique.

**Question N°4**

The use of OLS and Ridge Regression provided a straightforward, easy and interpretable, and approach to predicting trading volume. Throughout the analysis, I was able to improve model performance by refining feature selection and improving R-squared while reducing RMSE. One key improvement was identifying and addressing multicollinearity by removing highly correlated predictors such as Open, High, and Low to stabilise the model. Additionally, I incorporated Lagged Volume and an Interaction Term (Volume × Close), allowing the model to capture time dependencies and volume-price relationships, which improved predictive power.

Despite these improvements, OLS and Ridge Regression assume linear relationships, which is a major limitation. Stock trading volume is often influenced by nonlinear factors, such as market shocks, investor sentiment, and algorithmic trading patterns, which are not well captured by linear models. Ridge regression, while helping to control variance, did not significantly improve RMSE. Additionally, the dataset lacked key external factors, such as macroeconomic indicators or news sentiment, which could have further improved predictions. Future models could benefit from nonlinear techniques like decision trees or neural networks for more robust forecasting.