Name: Muhammad Bilal Elahi **Applied Session: 04**

Monash ID: 34200223 **Group: 54**

Background:

In the rapidly evolving financial markets, accurately predicting stock volatility is crucial for investors, portfolio managers, and financial institutions. Volatility, which measures the rate at which the price of a security increases or decreases for a given set of returns, is a key indicator of market sentiment and risk. Understanding and forecasting volatility allows for better risk management, informed decision-making, and the development of trading strategies that can mitigate potential losses. This assignment focuses on the development and evaluation of predictive models aimed at forecasting stock volatility using a combination of fundamental financial indicators and trading data. By leveraging various machine learning techniques, including Ridge Regression, Lasso Regression, and Principal Component Regression (PCR), the goal is to build a robust model that can generalize well across different stocks and market conditions.

Objectives:

Data Exploration and Preprocessing:

- To thoroughly explore the given dataset, identifying key features, patterns, and any potential data quality issues such as missing values or outliers.
- To perform feature engineering by creating lagged variables, rolling statistics, and other derived features that capture the temporal nature of the data and enhance the predictive power of the model.
- To preprocess the data by handling categorical variables, scaling numerical features, and preparing the data for model training.

Model Development:

- To develop and train multiple predictive models.
- To evaluate the models based on their ability to predict stock volatility, using appropriate metrics such as Root Mean Squared Error (RMSE) to quantify model accuracy.

Model Comparison and Selection:

- To compare the performance of the different models on both training and validation datasets, selecting the best-performing model based on RMSE and generalization ability.
- To analyze the importance of different features in the final model and interpret the results in the context of financial market predictions.

Submission and Deployment:

- To apply the selected model to a separate submission dataset, generating predictions for stock volatility and preparing a submission file.
- To ensure that the final model is robust, scalable, and capable of being deployed in a real-world financial setting for ongoing volatility prediction and risk management.

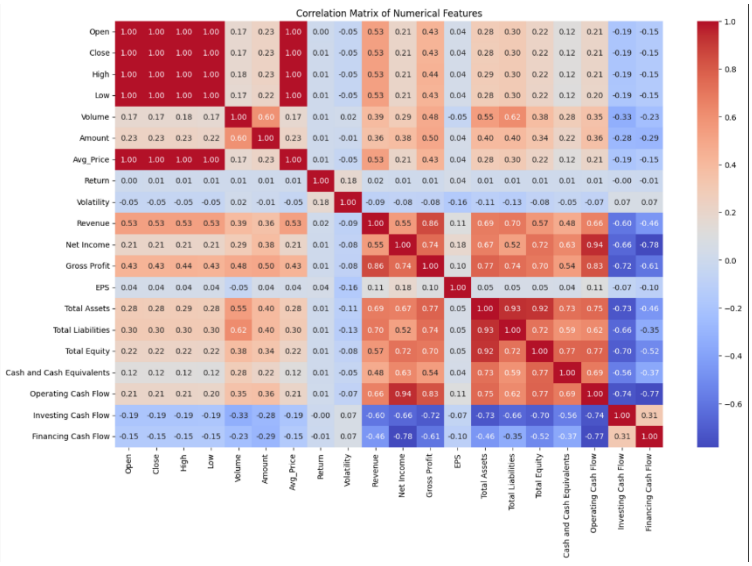
Data Exploration and Preprocessing:

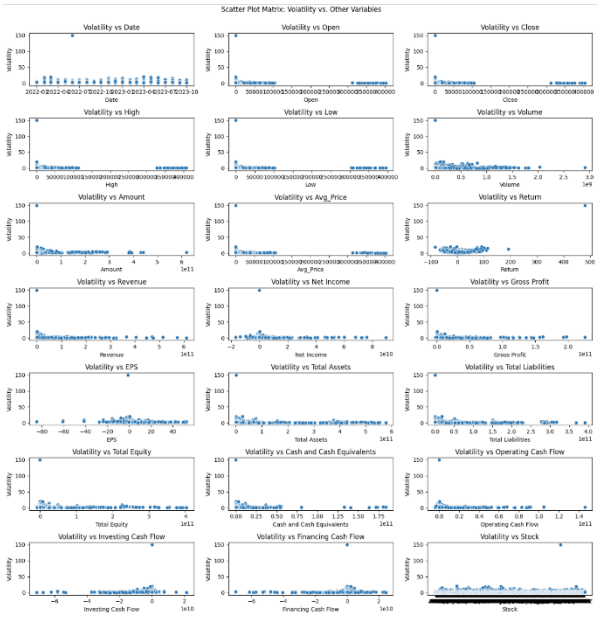
The dataset contains 22 columns and 13,486 rows. Date and Stock are categorical variables. Other columns are primarily numerical and represent various financial metrics and trading data.

No missing values are detected in the dataset.

Most columns are float64, representing continuous numerical data. Date and Stock are object (categorical) types.

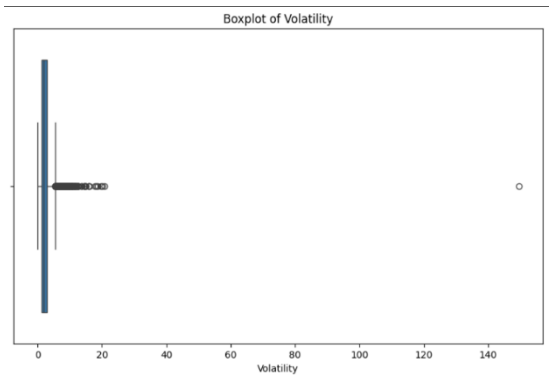
- The heatmap of the correlation matrix reveals relationships between the various features.
- Strong correlations among financial metrics such as Assets, Liabilities, and Total Equity.
- Trading metrics such as Open, Close, High, Low, and Average Prices are also strongly correlated.





- To fully understand the relationship between our features and target (Variability), we will be plotting scatter plots of each variable against variability.
- Features such as Open, Close, Revenue, Assets, Liabilities, etc., seemed to be highly skewed and would require transformations such as log to un-skew them and get a better prediction result.
- A common trend of a single outlier for `volatility` can be noticed, having an extremely high value compared to its mean value

- We will be filtering our dataset to include rows with volatility less than 140 to reduce data skewness.



Feature Extraction:

Training Set:

We will generate lagged variables from the features Open, Close, High, Low, Volume, Amount, Avg price, and Return to account for temporal dependencies. Specifically, the lagged variables will represent data from two months prior, such as Open_lag2 containing the Open values from two months ago.

Additionally, ordinal features like month, quarter, and year will be extracted from the Date column. To account for the cyclical nature of time (e.g., months of the year), we will transform the month into cyclical sine and cosine features. This approach helps capture periodic patterns in the data.

Quarter ends, typically in March, June, September, and December, will be indicated through a binary feature, alongside another binary feature marking the end of the fiscal year. The company's size will be categorized based on Assets into Small, Medium, and Large categories. Similarly, Revenue will be classified as Low, Medium, or High, and Net Income as Loss, Low Profit, or High Profit.

The original values of Open, Close, High, Low, Volume, Amount, Avg price, and Return will be excluded, as these features may rely on future data that is not available at the prediction time. The Date column will also be dropped, as it is no longer needed.

To reduce skewness and improve model accuracy, log transformations will be applied to features such as Revenue, Total Assets, Total Liabilities, Amount_lag2, High_lag2, Low_lag2, Close_lag2, Open_lag2, and Avg_Price_lag2. Additionally, squared and cubic features of variables like Return_lag2, EPS, and Avg_Price_lag2 will be created to better capture non-linear relationships.

All the numerical features will be scaled using standard scalar and all the non-numerical columns will be one-hot-encoded in the model's pipeline. We will be left with 670 features.

Submission Set:

The quarterly-updated features like Revenue, Net Income, Gross Profit, EPS, and balance sheet variables (Total Assets, Total Liabilities, Total Equity, etc.) will remain the same in both, the labelled training set for October and the submission set, since November and December fall within the same quarter.

All necessary features from training set, including lagged, cyclic, binned and transformed variables, will also be created for the submission set to ensure both datasets have consistent feature sets.

Feature Importance:

We found the correlation of each predictor variable with Variability.

Low_lag2_log	0.432791
Close_lag2_log	0.425673
Avg_Price_lag2_log	0.422502
Open_lag2_log	0.418746
High_lag2_log	0.412641
Profitability_Category_Loss	0.406780
Revenue_log	0.331457
Total Assets_log	0.303862
Company_Size_Small	0.290660

Stock_DK	0.000537
Stock_UFPT	0.000488
Stock_TBPH	0.000484
Stock_ZTO	0.000383
Stock_QLYS	0.000304
Stock_QRVO	0.000299
Stock_RCL	0.000187
Stock_CNNE	0.000107
Stock_INTU	0.000073
Stock_YELP	0.000031

The top features include the log transformed features as they are highly correlated with our data. The log transformations un-skewed the relationships and now are highly correlated with our predictor.

The worst performing variables include some of the stock dummy variables as they must have negligible impact on the variance. This can be interpreted as some dummy variables have good correlation with the target while in others, it doesn't matter from which stock do they belong to, hence the low correlation and impact.

Models Selected:

Based on the correlation matrix, several features exhibit high multicollinearity. To address this, I will focus on models that penalize these dependencies, minimizing their influence on the overall model's performance.

Given the presence of categorical variables that require one-hot encoding, the total number of features will become 670. Consequently, the models I've selected are tailored to handle high-dimensional datasets efficiently but panelising low impact features.

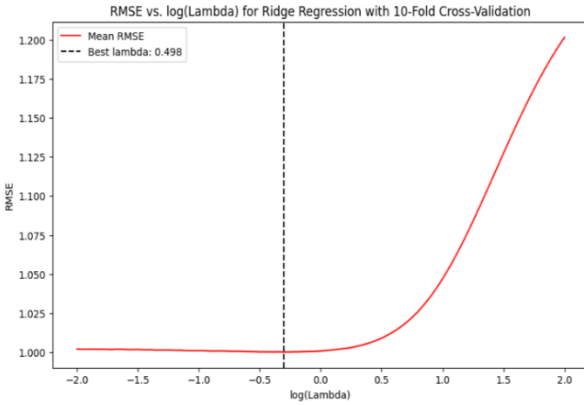
The models I will be evaluating are Ridge Regression, Lasso Regression, and Principal Component Analysis (PCA).

Feature Selection and Model Evaluation:

Ridge Regression:

First, we will split the data into training and test sets and then perform cross-validation within the training set to get the best alpha hyper parameter for ridge. Then we can use the test set to get our RMSE score. We will also use cross validation to get an estimated RMSE score.

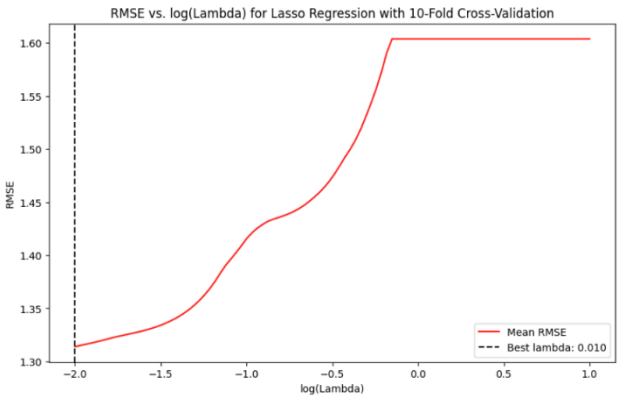
- The best alpha parameter with a search radius of, 100 logarithmically spaced values between 10^{-2} and 10^{2}, that we obtained through a 10-fold CV is 0.4977.
- The corresponding mean RMSE is 0.985.
- The top performing features in our regression model include lagged High, Low, and different dummy variables for Stock, depicting different trends for each stock.



Lasso Regression:

Like Ridge, Lasso will also split the labelled dataset and then use cross validation to determine the best alpha. But, as my laptop was taking too much time determining the best alpha, I used CV to figure out the trend and then used expert search to get the best alpha value given the number of parameters and mean RMSE.

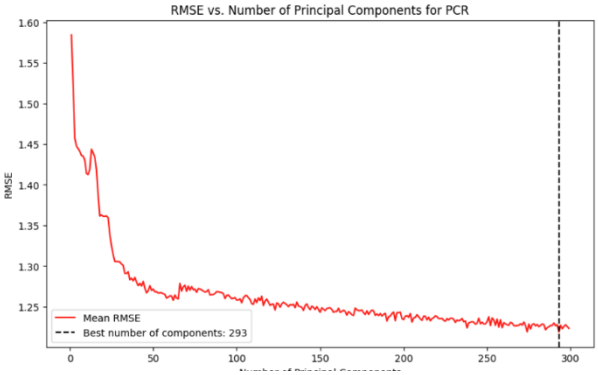
- There's a decreasing trend in RMSE for 100 logarithmically spaced values between 10^{-2} and 10^{2} using 10-fold CV.
- By expert search, the best approximate alpha for lasso was found to be 0.0005 having a mean RMSE score of 1.093 and 650 features.



Principal Component Analysis (PCA):

PCA is primarily used to decrease the dimensionality of a dataset by accounting for variances in each dimension. We will be using cross validation to determine the best number of components.

- The best number of components is 293.
- Best RMSE for PCA is 1.225.
- The graph has highly fluctuating RMSE score for different components. This can be imagined as PCA utilizes the best variances and not the actual features. Adding features that offer little to no variation can only decrease the RMSE score.

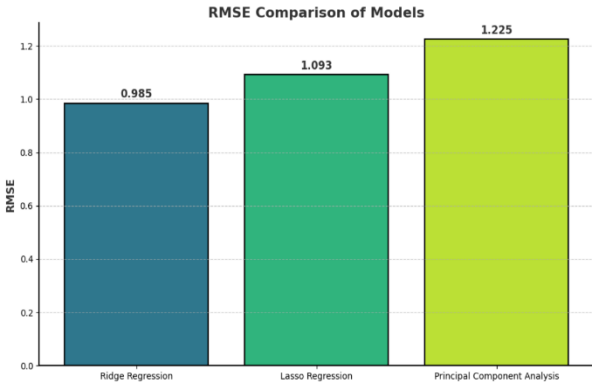


Statistical Comparison of Models:

Ridge Regression*:*

Handles multicollinearity: Ridge penalizes large coefficients, which is useful for high-dimensional datasets like ours where features may be highly correlated.

Best RMSE: The lowest RMSE of 0.985 indicates it is the best model for prediction performance.



Lasso Regression:

Feature selection: Lasso performs automatic feature selection by setting some coefficients to zero, resulting in a simpler model with only 650 features.

Trade-off between accuracy and complexity: While Lasso's RMSE of 1.093 is higher than Ridge's, it has a lower dimensionality.

Less accurate than Ridge: It may not capture the full complexity of the data as effectively.

Principal Component Analysis:

Lower accuracy: The RMSE of 1.225 indicates that PCR is the least accurate model among the three. While it reduces the dimensionality, it may discard useful information that is important for predicting volatility, leading to weaker performance.

Interpretability issues: Since PCR relies on principal components rather than the original features, it is more challenging to interpret the impact of individual features on stock volatility.

Conclusion:

Ridge Regression is the best model with the lowest RMSE of 0.985, making it the most accurate and effective for predicting volatility. While Lasso Regression simplifies the model by reducing the number of features, its RMSE of 1.093 shows a trade-off between accuracy and complexity. PCA has the highest RMSE of 1.225, indicating a loss of predictive power. Overall, Ridge Regression is chosen for its superior accuracy and ability to handle multicollinearity.

Progress Track:

This was a solo assignment, and my progress can be tracked at https://github.com/Billz942/Mining-Knowledge-from-Data .