FIT5149
APPLIED DATA ANALYSIS
Report :
Mining Knowledge from Data (Assignment 1)

Submitted by :

Madeleine J Gibellini (24195545)
Bhavna Balakrishnan (33954437)
Sukumar Dodda (34079408)

Contents :

1. Description of the Models Tried

To predict stock volatility, four different regression models were implemented and tested on the dataset:

1. **Linear Regression**: This is the baseline regression model that assumes a linear relationship between the independent variables (features) and the target variable (volatility). No regularization was applied, making it prone to overfitting, especially in cases with multicollinearity.
2. **Polynomial Regression**: Polynomial regression extends linear regression by introducing polynomial features, capturing non-linear relationships. In this case, we applied degree 2 polynomial features, which allowed the model to capture more complex interactions between features but at the cost of increased model complexity.
3. **Lasso Regression (L1 Regularization)**: Lasso is a type of linear regression that adds an L1 regularization term, which helps in both reducing overfitting and performing feature selection by shrinking coefficients of less important features to zero. This allows the model to focus on the most relevant variables affecting volatility.
4. **Ridge Regression (L2 Regularization)**: Ridge regression is another regularization technique, applying an L2 penalty to the coefficients. Unlike Lasso, Ridge shrinks coefficients but does not eliminate them entirely. It is particularly effective in dealing with multicollinearity, as it distributes the effect of correlated features.

---

2. Data Preprocessing and Feature Engineering Techniques

1. **Handling Missing Values**: In all models, missing values in the feature set were imputed using the mean strategy. This ensured no missing values were present in the dataset before training the models.
2. **Feature Selection**: Recursive Feature Elimination (RFE) was used to select the most relevant features for the models. This step ensured that only the most important features were included, reducing dimensionality and improving model performance.
3. **Feature Scaling**: For models such as Lasso and Ridge, feature scaling was applied to ensure that all features contributed equally. Standardization was performed using standard scaler, transforming the data to have a mean of 0 and a standard deviation of 1.
4. **Log Transformations**: Logarithmic transformations were applied to the `Volume` and `Amount` columns to correct for skewness in the data. This helped reduce the impact of extreme values on the model's performance.

---

3. Analysis of Model Predictions and Performance

The performance of each model was evaluated using **Root Mean Squared Error (RMSE)** and **R-squared (R^2)**, as shown in the table below:

Model	RMSE	R-squared (R²)
Linear Regression	2.2655	-0.9981
Polynomial Regression	0.7539	0.1007
Lasso Regression	0.5396	0.5392
Ridge Regression	1.5591	0.0537

Observations:

- **Linear Regression** performed poorly with a high RMSE and negative R^2, indicating that it failed to capture the underlying patterns in the data.
- **Polynomial Regression** improved the performance significantly by capturing non-linear relationships, resulting in a much lower RMSE and a positive R^2.
- **Lasso Regression** achieved the best overall performance, with the lowest RMSE (0.5396) and the highest R^2 (0.5392). The feature selection property of Lasso helped in focusing only on the most important variables, improving both predictive accuracy and model generalisation.
- **Ridge Regression** had a better RMSE than Linear Regression, but it did not outperform Lasso or Polynomial Regression. This indicates that while Ridge helped reduce overfitting, it was not as effective in this context as Lasso.

---

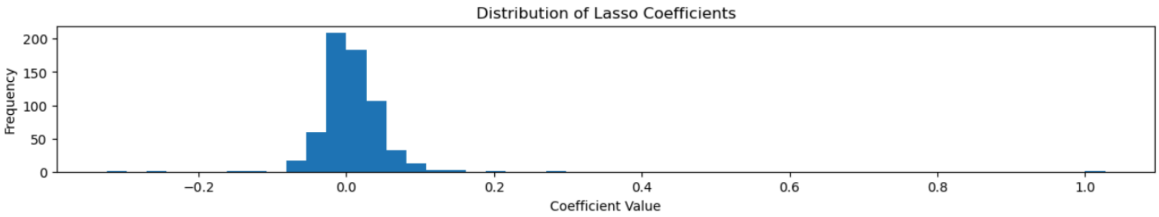
4. Evidence-Based Discussion of Feature Importance

Lasso Feature Importance: Lasso regression shrinks many feature coefficients to zero, effectively performing feature selection. The following features were identified as the most important in predicting stock volatility:

Key Features Identified:

- **Log_Volume** (Coefficient: 1.027): Trading volume has the highest positive impact on stock volatility, suggesting that as trading volume increases, volatility tends to rise significantly.
- **Return** (Coefficient: 0.275): Positive stock returns also contribute to increased volatility.
- **Stock-specific features** like **Stock_STAF** (0.206), **Stock_LSXMB** (0.161), and **Stock_DIT** (0.142) indicate that these specific stocks contribute to volatility, likely due to market behaviour or company-specific events.
- On the negative side, **Total_Assets** (-0.324), **Log_Amount** (-0.251), and **Log_Revenue** (-0.139) are associated with reduced volatility, suggesting that companies with more assets, revenue, or transaction amounts are more stable, thus less volatile.

This reinforces the fact that Lasso not only selected key features but also helped to interpret the relationship between these variables and stock volatility.



This distribution of Lasso coefficients, as shown in the provided graph, reflects that some coefficients were shrunk to zero, highlighting the model's effectiveness in eliminating irrelevant features while preserving the most influential ones.

5. Development of the Submitted Model (Lasso Regression)

Model Development Process:

Cross-Validation and Hyperparameter Tuning: The Lasso model was developed using LassoCV to perform feature selection. 5-fold cross-validation, with a grid search used to find the optimal value of the regularization parameter alpha. The best alpha value found was 0.000256.

Standardisation of features: Lasso requires standardisation of features in order to penalise all variables equally and assess feature importance of all variables relative to each other.

Model Justification:
Lasso was chosen as the final model due to its ability to balance bias and variance through regularization. The model's strong feature selection capability made it the most appropriate for predicting stock volatility, especially given the sparsity of important features in the dataset. The Lasso model demonstrated superior performance over the other models, achieving the lowest RMSE and the highest R-squared value, making it the best choice for this task.

The benefit of considering many coefficients in Lasso Regression, while shrinking only a select few to zero, lies in Lasso's ability to balance between overfitting and underfitting while preserving key predictive features. Here's a detailed explanation of why retaining many coefficients in Lasso can be advantageous:

1. Capturing Complex Relationships

By retaining many coefficients, Lasso allows the model to consider multiple features that may have a smaller but significant impact on the target variable (in this case, stock volatility). This is beneficial in financial data, where many variables can interact in complex ways to influence outcomes. Rather than eliminating features with moderate importance, Lasso shrinks their influence, ensuring that the model still captures these subtler relationships.

2. Avoiding Underfitting

While Lasso is known for its ability to reduce overfitting by eliminating less important features, retaining many coefficients helps avoid underfitting. If too many features are shrunk to zero, the model might lose valuable information, leading to poor predictions. By considering many features, the Lasso model maintains a more comprehensive understanding of the data, ensuring that it can still generalize well to unseen data.

3. Data-Driven Feature Selection

In the case of stock volatility, a wide range of factors — from financial metrics like **Log_Volume** and **Return** to stock-specific behaviours — contribute to price swings. Lasso carefully selects the features that have the most predictive power by shrinking coefficients based on their importance. Retaining many coefficients allows the model to remain flexible and dynamic, as volatility is often driven by various interacting factors, rather than a handful of dominant variables.

4. Incorporating Small but Relevant Effects

Some features may not have a large individual impact on stock volatility but could still contribute valuable predictive information in combination with other features. Lasso's ability to retain many features, while reducing the magnitude of their influence, ensures that the model considers these smaller yet relevant effects without letting them dominate or lead to overfitting.

5. Flexibility in Prediction

By considering a wide range of coefficients, Lasso Regression maintains flexibility in its predictions. Stock volatility is affected by numerous factors, from market sentiment to specific company events. Retaining many coefficients ensures that the model can adjust its predictions based on various factors, even if some have less pronounced but still relevant impacts on the outcome.

Conclusion

In conclusion, after testing and comparing multiple models, **Lasso Regression** was selected as the optimal model for predicting stock volatility. It demonstrated superior performance, achieving an RMSE of 0.5396 and an R^2 of 0.5392. Lasso's feature selection capability proved invaluable, focusing on the most important predictors, such as trading volume and specific stock features, while eliminating irrelevant ones. This model provides a robust, interpretable, and generalizable solution for predicting stock volatility.

---

6. Project Management :

- The group used WhatsApp as the primary form of communication, a Trello board was created to list all required tasks to be completed with a workflow of ‘To-do’, ‘In progress’ and ‘Complete.’
- Tasks were distributed evenly across group members, with each member being involved in discussing assignment specifications, reviewing both EDA and data preprocessing. Each team member worked on at least one model.
- A project manager was assigned to each week who had the role of ensuring group members were on track to complete assigned tasks.
- All code and report writing was proof-read by every group member.

Team Member	Week-1	Week-2	Week-3	Week-4
Bhavna	**L**, Understand and discuss specifications	Reviewed EDA and did data transformation	**L**, Continued Data preprocessing and feature selection	Did Simple linear and polynomial regression models
Maddi	Understand and discuss specifications	**L**, Started EDA	Added more to EDA and data processing	Did Lasso model
Sukumar	Understand and discuss specifications	Reviewed EDA and data preprocessing	Began with the report	**L**, Did Lasso and ridge regression

Table 1. Weekly roles and contributions of project team members. ‘L’ = team leader of the week.

7. Appendix :

"OpenAI. (2024). ChatGPT (Version GPT-4) [Large language model]. https://www.openai.com" - the assistance of OpenAI's ChatGPT for providing grammar correction and editorial support in the preparation of this report.

Matplotlib: Visualization with Python. (n.d.). *Matplotlib.* https://matplotlib.org/

Acknowledgement that Monash University FIT5149 applied class and lecture code and content was used for this assignment

Documentation History:

Trello board

https://trello.com/invite/b/66c690b4c43d292d362f619f/ATTI6e98a902db2e1ff2728e51419dfe170aDBDF5EA6/fit5149-ada-a1

Google colab

https://colab.research.google.com/drive/1vk1sZbL0Y-n3MLBrFetNhVoMLiKQLJWI?usp=sharing