Sowmiya Kumar

STOCK MARKET ANALYSIS

Data Set

Dataset 1: S&P/ASX 200 data

The ASX 200 dataset provides historical data for the Australian stock market index, covering 200 listed companies. It includes monthly data from January 2000 to July 2024, comprising 295 rows and 7 columns. Key attributes include the index's opening and closing prices, highest and lowest prices reached, trading volume, and the percentage change from the previous day's closing price. This dataset offers valuable insights into the market's performance over more than two decades, enabling analysis of trends, volatility, and market behavior. https://au.investing.com/indices/aus-200-historical-data

Dataset 2: U.S.A Economic Indicators data

This U.S. Economic Indicators dataset is manually combined monthly data from January 2000 to July 2024, consisting of 295 rows and 7 columns. It has key indicators like GDP, Unemployment Rate, CPI, INDPRO, Housing Price Index, and Recession Data. These indicators provide comprehensive insights into the U.S. economy, covering aspects like market health, inflation, industrial output, and recession periods, making it essential for economic analysis and forecasting. https://fred.stlouisfed.org/series

Choice of Dataset:

I chose the U.S. economic indicators because the United States plays a major role in the global economy, and its economic conditions significantly influence international markets, including Australia[2]. By analyzing these indicators, I can better understand and predict trends in the Australian stock market, making it crucial for a data science role focused on financial analysis and market forecasting.

Data Analysis

Problem Statement:

The goal is to analyze the impact of U.S. economic indicators on Australian stock market performance and to predict stock price movements.

Hypothesis:

1. U.S. economic indicators have a direct impact on Australian stock market performance.
2. Market reactions during U.S. recessions mirror broader economic trends.

Analysis and Insights:

To investigate the relationship between U.S. economic indicators and Australian stock prices, I merged the datasets and conducted a correlation analysis (Figure 1). The analysis revealed strong correlations between U.S. GDP, the Housing Price Index, the Industrial Production Index (INDPRO), and the ASX 200 index, suggesting that these U.S. economic factors significantly influence Australian stock prices.

Further analysis during periods of U.S. recessions (Figure 2) showed that ASX 200 prices either declined or remained stagnant, aligning with broader economic expectations and highlighting the negative impact of U.S. economic downturns on the Australian market[1]. This proves hypothesis 2.

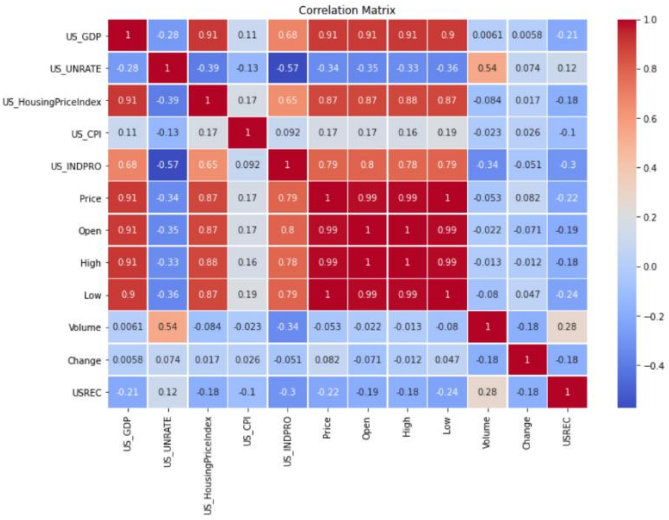


Figure 1: Correlation matrix

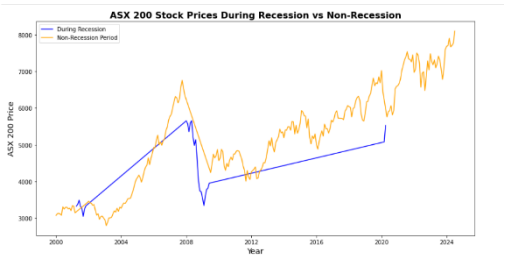


Figure 2 Stock prices vs recession

Modeling and Prediction:

To predict stock prices, Linear Regression and Random Forest Regressor[3][4] models were used. The evaluation metrics used were RMSE (Root Mean Square Error) and R^2 (coefficient of determination).

- RMSE was selected to measure the average magnitude of the prediction errors, with lower values indicating more accurate models.
- R^2 was chosen to assess the proportion of variance in stock prices explained by the models, with values closer to 1 indicating better performance.

The results (Figures 3 and 4) demonstrate that the Random Forest model outperformed Linear Regression, especially in capturing non-linear relationships, with an RMSE of 159 and an R^2 of 0.98, compared to 475 and 0.86 for Linear Regression. Feature importance analysis from the Random Forest model (Figure 5) reaffirmed that Housing Price Index, GDP, and INDPRO are the most significant predictors, consistent with the correlation findings. This proves hypothesis 1.

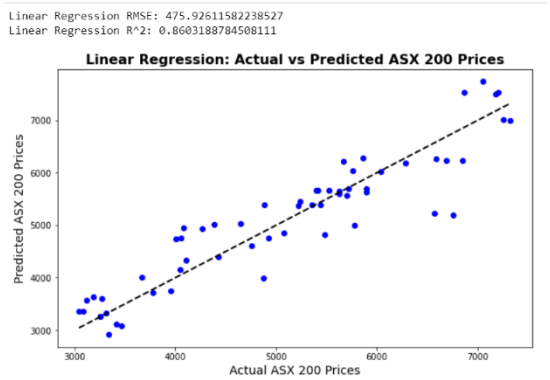


Figure 3: Linear regression results

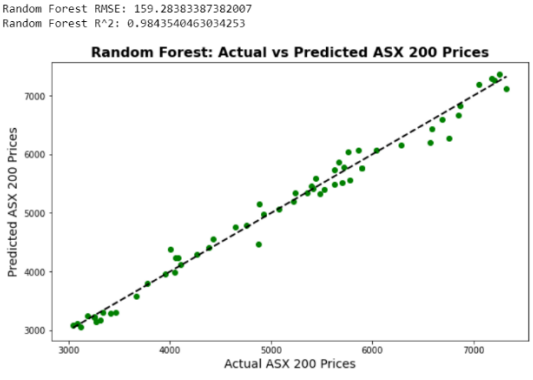


Figure 4: Random Forest results

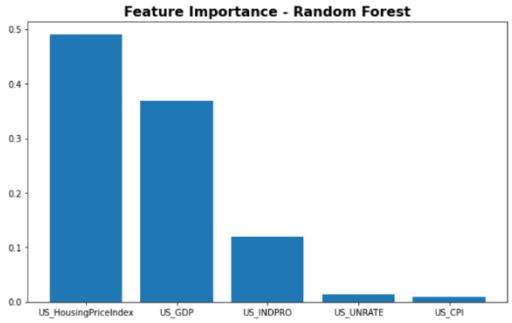


Figure 5: Feature importance

Stock Price Direction Classification Analysis:

To further explore whether stock prices would increase or decrease based on U.S. economic indicators, Logistic Regression and Random Forest Classifier models were used. Logistic Regression was chosen for its simplicity and interpretability in binary classification tasks.

- Precision was used to measure the accuracy of positive predictions (price increases), crucial for minimizing false positives.
- Recall was chosen to evaluate the model's ability to capture actual stock price increases, ensuring that most opportunities are identified.
- F1 Score was used to balance precision and recall, providing a comprehensive measure of model performance.
- ROC Curve and AUC were employed to assess the models' ability to distinguish between stock price increases and decreases, with AUC values closer to 0.5 indicating poor performance. (Figure 6)

Metric	Logistic Regression	Random Forest
Accuracy	0.57	0.51
Precision	0.58	0.56
Recall	0.92	0.63
F1 Score	0.71	0.59

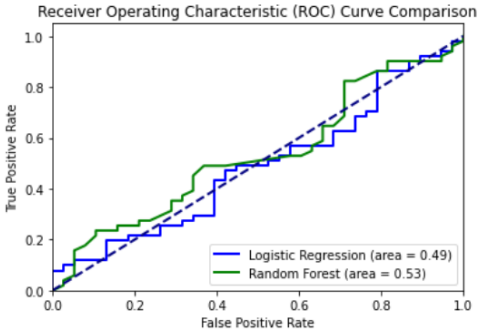


Figure 6: ROC Curve

The coefficients from the Logistic Regression model (Figure 8) revealed that factors such as the Unemployment Rate, Housing Price Index, and CPI positively influence stock price increases, while GDP, INDPRO, and U.S. Recession data negatively impact stock prices. Interestingly, GDP, typically a positive indicator[5], showed a slightly negative coefficient. This unexpected result might be attributed to the market's response during the 2020 pandemic, where economic growth did not necessarily lead to stock market gains due to broader uncertainties.(Figure 7)

Figure 7: US GDP vs AU stock price

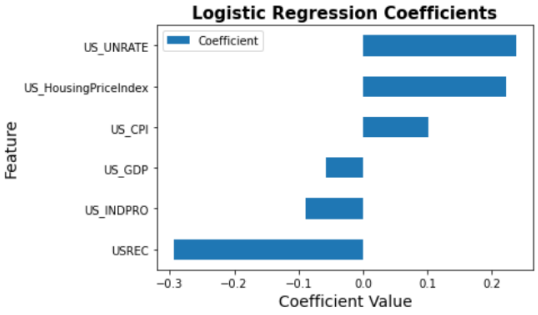


Figure 8: Coefficient value of features

The classification models, however, performed poorly, with Logistic Regression showing a high recall (0.92) but low precision, leading to many false positives, while Random Forest offered more balanced, but still suboptimal, results. The ROC curve (Figure 6) shows that both models had AUC values close to 0.5, indicating that neither model effectively distinguished between positive and negative classes.

Inference:

The analysis confirms that U.S. economic indicators, particularly during recession periods, significantly influence Australian stock prices. The insights from both regression and classification models are complementary. While the Random Forest Regressor effectively predicted the extent of stock price changes, the classification models attempted to determine the direction of these changes based on the same economic indicators. However, the classification models' poor performance highlights the need for further model refinement.

The high recall and low precision in the Logistic Regression model suggest potential class imbalance, which could be addressed by adjusting class weights or exploring more advanced models like Gradient Boosting, XGBoost, or SVM. These approaches may enhance the models' ability to accurately classify stock price movements based on U.S. economic indicators.

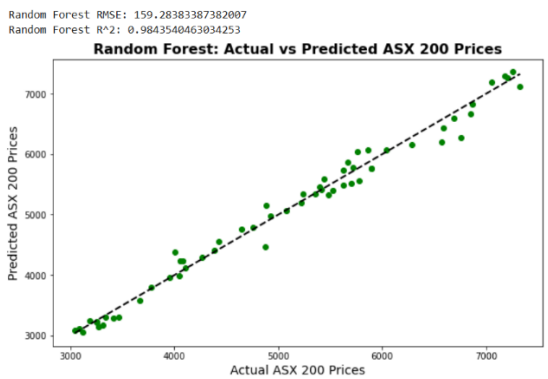
Evaluation of Bias in Models:

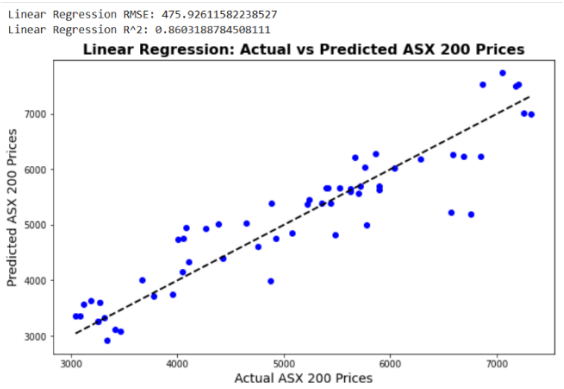
The chosen model performance was tested under two scenarios:

1. Model Performance analysis with simple Train-Test Split (without Cross-Validation)
2. Model Performance analysis with Cross-Validation and random sampling

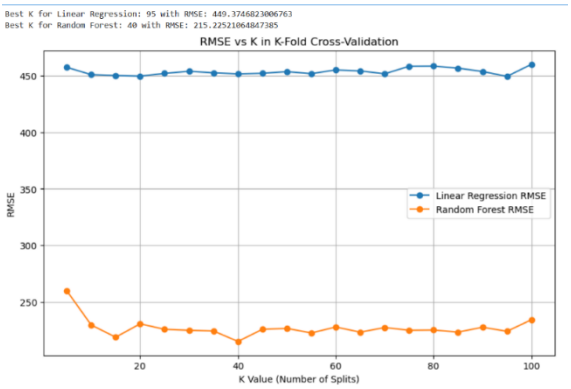
Train-test split:

In the first scenario, an 80-20 train-test split was applied to the dataset, and the data was standardized before training the model. The model was trained and tested using this single split. The results show that the Random Forest model significantly outperformed Linear Regression, achieving a much lower Root Mean Squared Error (RMSE) and a higher R^2 score, indicating better fit and predictive accuracy.





k-fold cross validation:



In the second scenario, the model's performance was evaluated across multiple splits, with k-values ranging from 5 to 100 using random sampling. The optimal k-value was determined through this process and used for final prediction.

The Random Forest model consistently outperformed Linear Regression across different k values. The best RMSE for Random Forest (215.22) was substantially lower than for Linear Regression (449.37). Notably, Random Forest's performance varied more at lower k values but stabilized as k increased.

Insights from two scenarios:

In both the train-test split and cross-validation setups, Random Forest outperformed Linear Regression.

In K-Fold Cross-Validation, the RMSE for Random Forest increased from 159.28 (train-test split) to 215.23. This is expected, as cross-validation provides a more reliable performance estimate by training and evaluating the model on multiple folds. The single train-test split may have been particularly favorable for Random Forest, explaining the slightly lower RMSE in that scenario.

Cross-validation marginally improved Linear Regression's RMSE (from 475.93 to 449.37), indicating that the train-test split might not have been fully representative, and cross-validation helped reduce variability in the dataset.

Bias Evaluation:

Cross-validation provided a more unbiased evaluation of model performance. By using K-Fold Cross-Validation, each data point was utilized for both training and testing, reducing the risk of overfitting to a specific subset of data, which can happen with a simple train-test split.

The slight performance advantage for Random Forest in the train-test split scenario may have been due to the randomness of the split. Cross-validation is a more robust method for assessing generalization ability and provides a more reliable estimate of model performance on unseen data. For Random Forest, the increase in RMSE during cross-validation reflects a more realistic evaluation of its predictive power. Similarly, Linear Regression's improvement in cross-validation suggests that the train-test split was less favorable for this model.

Learning Curve Analysis:

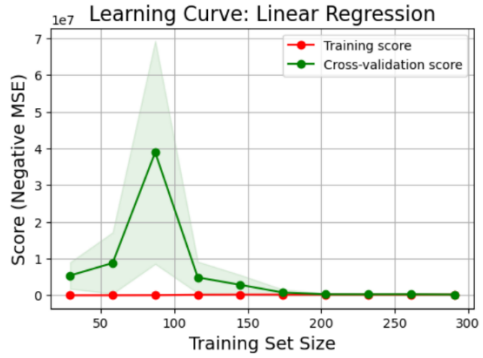
The learning curves for both models show important insights into how they handle varying amounts of data and generalize to new information.

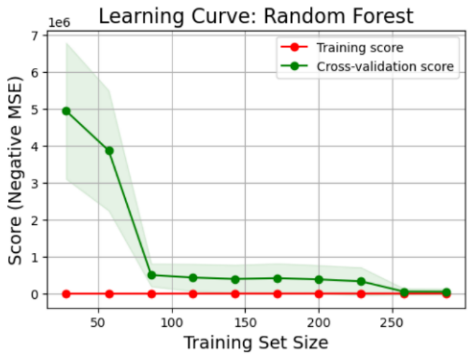
For Linear Regression, the training score consistently stays low, which suggests that the model is fitting the training data well. However, the cross-validation score fluctuates significantly, especially with smaller datasets, showing instability and a large gap compared to the training score. This gap indicates overfitting, as the model struggles to generalize to new data. As the dataset grows, the cross-validation score begins to stabilize, but it never fully converges with the training score. This persistent gap suggests that Linear Regression may be too simplistic for the complexity of the data. The cross-validation score levels off around 200 samples, showing that beyond this point, additional data offers diminishing improvements.

Random Forest, by contrast, displays a much smoother learning curve. While its training score remains low, the cross-validation score consistently improves as more data is introduced. The gap between the training and cross-validation scores narrows with increased dataset size, indicating that Random Forest is better at identifying the underlying patterns. As the dataset grows, the model's performance stabilizes, demonstrating that it handles data complexity more effectively than Linear Regression.

In terms of handling model complexity and the potential for overfitting, Linear Regression shows clear signs of overfitting with smaller datasets due to its simplicity. Random Forest, however, proves to be more robust, showing less overfitting and generalizing better as the data size increases.

Overall, this analysis demonstrates that Random Forest, particularly when combined with cross-validation, produces more unbiased and reliable results, making it the stronger model for this dataset.





Bias detection using tools:

Using IBM AI Fairness 360 (AIF360), I analyzed the dataset to detect potential biases, focusing on the model's treatment of individuals based on their unemployment status. AIF360 provides a variety of fairness metrics and bias mitigation tools, making it an ideal choice for this task.

The first fairness metric, disparate impact, examines whether the model treats individuals in the high unemployment group (unprivileged) less favorably than those in the low unemployment group (privileged). Ideally, a value of 1 suggests fairness, but in this case, the disparate impact score was 0.3947. This means that individuals in the unprivileged group were only 39% as likely to receive favorable outcomes as those in the privileged group, which indicates a significant bias.

Next, the statistical parity difference measures the gap in favorable outcomes between the privileged and unprivileged groups. Here, the result was -0.4306, showing that individuals in the high unemployment group were much less likely to receive favorable outcomes, further reinforcing the existence of bias.

The consistency score, which checks whether similar individuals receive similar predictions, was 0.9308. This suggests that the model is relatively consistent in its treatment of similar data points. However, even though the model is consistent, it can still be consistently biased, as seen in this case.

The results from the disparate impact and statistical parity difference metrics clearly show that the model exhibits bias against the unprivileged group. Although the model performs consistently, it does not treat the unprivileged group fairly, highlighting a significant issue.

To address this bias, I applied the reweighing technique, which adjusts the importance of data from different groups before training the model. After reweighing, the disparate impact improved to 1, indicating that the model now treats both groups equally and that the bias has been effectively mitigated.

Deploying on a large scale:

When deploying cross-validated models on a large scale, the following considerations apply:

1. **Security**: The model may be vulnerable to data poisoning or adversarial attacks, where malicious inputs can lead to distorted predictions. Implementing strong access controls and secure infrastructure is essential to mitigate these risks.

2. **Privacy**: Since data is sourced from publicly available official websites, privacy concerns are minimal. However, adhering to best practices for data handling and ensuring compliance with privacy regulations remain important.

3. **Ethical**: After applying bias mitigation using IBM AI Fairness 360's reweighing metrics, the model shows no signs of bias, ensuring fair and equitable predictions.

Cross-validation strengthens model reliability, but addressing these security, privacy, and ethical considerations is critical for responsible and secure deployment.

References:

1. Johanne R. Trippas, Damiano Spina, Falk Scholer (2024). *Adapting Generative Information Retrieval Systems to Users, Tasks, and ScenariosLinks to an external site.*. Information Access in the Era of Generative AI, Springer Nature Switzerland AG (In Press).
2. Hew, K. F., & Lo, C. K. (2018). Flipped classroom improves student learning in health professions education: a meta-analysis. BMC medical education, 18(1), 1-12. https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-018-1144-z
3. BMC Medical Informatics and Decision Making. (n.d.). Home page. https://bmcmedinformdecismak.biomedcentral.com
4. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. BMJ Quality & Safety, 28(3), 231-237. https://qualitysafety.bmj.com/content/28/3/231
5. Gao, Y., Reddy, M., & Niemelä, I. (2020). Using artificial intelligence to improve the quality and safety of radiation therapy. Journal of applied clinical medical physics, 21(1), 6-11. https://pubmed.ncbi.nlm.nih.gov/31492404/
6. Masters, K. (2019). Artificial intelligence in medical education. Medical teacher, 41(9), 976-980. https://pubmed.ncbi.nlm.nih.gov/31007106/
7. J. Kirsch, "Does The Recent Stock Market Dive Indicate A Recession In 2024?," *Forbes*, Aug. 06, 2024. Available: https://www.forbes.com/sites/investor-hub/article/does-recent-stock-market-crash-indicate-recession-2024/
8. K. Bredemeier and W. Gallo, "World stocks plunge on fears of US economic slowdown," *Voice of America*, Aug. 05, 2024. https://www.voanews.com/a/wall-street-headed-sharply-lower-again-after-japan-s-nikkei-index-tumbles-to-worst-loss-since-1987/7730241.html (accessed Aug. 17, 2024).
9. scikit-learn, "3.2.4.3.2. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.20.3 documentation," *Scikit-learn.org*, 2018. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
10. A. Dutta, "Random Forest Regression in Python - GeeksforGeeks," *GeeksforGeeks*, Jun. 14, 2019. https://www.geeksforgeeks.org/random-forest-regression-in-python/
11. E. Trovall, "How cooling U.S. GDP growth affects the global economy - Marketplace," *Marketplace*, May 30, 2024. https://www.marketplace.org/2024/05/30/how-cooling-u-s-gdp-growth-affects-the-global-economy/ (accessed Aug. 17, 2024).