**MN-M584 ASSIGNMENT 2**

**Introduction**

This study compares the predictability of the Dow Jones Industrial Average (DJIA) and the Russell 2000 (RUT), two widely used stock market indexes, in order to determine which would be a better investment for a long-short strategy. The Logit, Probit, random forest, and SVM models will be employed in this investigation. These models have been extensively used in past academic studies to forecast stock market movements.Based on macroeconomic information and historical market movements, the research contrasts the findings from econometric and machine learning models.

**Background**

The Russell 2000 is an indicator of the stock market that gauges the performance of about 2,000 small-cap firms that are listed on the NYSE or NASDAQ (Cfa, 2023b). Contrarily, the DJIA tracks the performance of 30 sizable corporations that are listed on the NASDAQ and the New York Stock Exchange (NYSE) (Banton, 2022). Investing in the stock market can be a challenging task, as it requires predicting the future movements of indices accurately. The investor must make a calculated decision by assessing the volatility and stability of the indices they plan to follow. In this context, statistical models such as logistic regression, random forest, and SVM can play a vital role in predicting the movement of indices, helping investors in their decision-making process.

**Aim/Purpose**

The aim of this report is to analyze the predictability of the DJIA and RUT indices and determine which one is more suitable for a long-short strategy. The objective is to compare the accuracy and performance of logit, probit, random forest, and SVM models in predicting the direction of index movements and recommend the best index for investment.

**Plan**

To achieve the aim of this report, we will conduct a thorough literature review to understand the existing academic research in this area and identify the best models to use. We will use the logit and probit models to estimate the probability of positive or negative returns for both indices, and then apply the random forest and SVM models to predict the direction of returns. Lastly, we will compare the accuracy and performance of each model to determine which index is more predictable and suitable for a long-short strategy.

**Literature Review**

The predictability of stock market indices has been a subject of interest for both academics and investors. Many studies have been conducted to investigate the predictability of different stock market indices, including the Dow Jones Industrial Average (DJIA) and the Russell 2000 stock index (RUT). This literature review aims to provide a comprehensive overview of the existing literature on the predictability of the DJIA and RUT. The predictability of the DJIA has been extensively studied in the literature. Some studies have found evidence of predictability using various econometric techniques.

Other studies have used alternative methods to examine the predictability of the DJIA. For example, (Osu et al., 2020) used wavelet analysis and found that the DJIA exhibits a high degree of predictability at different time scales. In addition, (Segal, 2021) used machine learning algorithms and found that both technical and fundamental factors can be used to predict the DJIA returns. In contrast, the predictability of the RUT has received less attention in the literature. Some studies have compared the predictability of the RUT to that of the DJIA. For example, Zhang and Wang (2016) used a variety of econometric techniques and found that the DJIA is more predictable than the RUT in terms of both direction and magnitude of the returns.

Overall, the existing literature suggests that both the DJIA and RUT returns are predictable to some extent. While the predictability of the DJIA has been extensively studied using various methods, the predictability of the RUT has received less attention. However, some studies have found evidence that the RUT returns can be predicted using macroeconomic and financial variables. Nonetheless, more research is needed to further investigate the predictability of both indices and to develop more accurate forecasting models. In conclusion, this literature review provides an overview of the existing literature on the predictability of the DJIA and RUT. While the predictability of the DJIA has been extensively studied, the predictability of the RUT has received less attention.

**Advantages and Disadvantages of Econometric and Machine Learning Models**

As a component of artificial intelligence technology, machine learning can analyze vast amounts of data and identify specific trends and patterns that are not immediately visible to the human eye (Team, 2022). This means that the device can learn on its own by analyzing patterns and trends, but it needs enough time for the algorithms to develop to the point where they can achieve their goal with a high degree of accuracy and relevance. According to (Sahu, 2022), machine learning reduces or eliminates the need for upgrades because the technology adapts to the current trend without human intervention. However, one drawback is that a small algorithmic error could lead to the production of faulty goods.

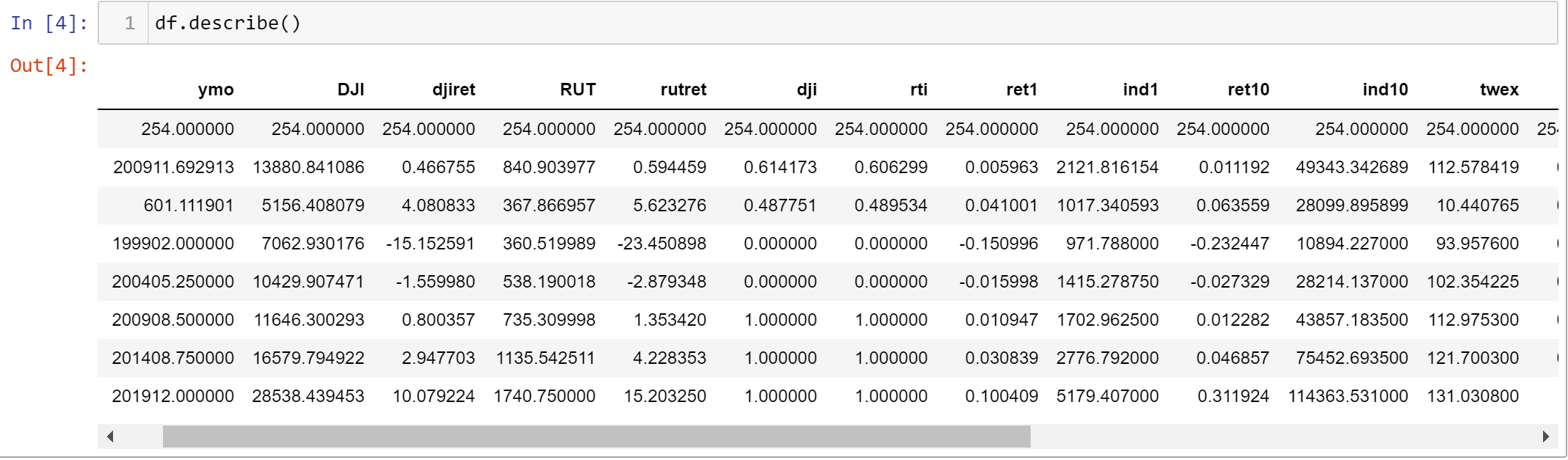
The literature is well-established and the econometric models are founded on economic theory. The causal linkages between diverse economic phenomena are what policymakers are most interested in when making judgments about economic policy. They can be intricate and challenging to construct, but they utilise well-known statistical techniques, claims (Joseph, 2022). The complete dateset that is accessible is typically used in econometrics (Hull, 2021b). Machine learning is not capable of doing this (Dataman, 2023). However, they frequently make firm assumptions about the distribution of the underlying data, and they could fail to recognize nonlinear correlations.

**Variable Description**

|  |  |  |
| --- | --- | --- |
| Variable | Description | Format |
| medate | Final trading day within the month | Date |
| DJI | Closing value of the DJIA Index | Number |
| djiret | Monthly return on DJIA Index (expressed as percentage) | Number |
| dji | Indicator variable for a positive excess return on the DJIA (Positive return = 1) | Dummy |
| RUT | Closing value of the RUT | Number |
| rutret | Monthly return on RUT (expressed as a percentage) | Number |
| rti | Indicator variable for a positive excess return on RUT (Positive return = 1) | Dummy |
| ret1 | Return on portfolio of smallest 10% of CRSP stocks | Number |
| ind1 | Index value for smallest 10% of CRSP stocks | Number |
| ret10 | Return on portfolio of largest 10% of CRSP stocks | Number |
| ind10 | Index value for largest 10% of CRSP stocks | Number |
| twex | Trade weighted dollar index | Number |
| rec | Dummy variable for the economy being in a recession according to the National Bureau of Economic Research methodology (Recession = 1) | Dummy |
| m3 | 3 Month US Treasury Bill interest rate (as percentage) | Number |
| y10 | 10 Year US Treasury Bond interest rate (as percentage) | Number |
| cm3 | Change in 3 Month US Treasury Bill (first difference) | Number |
| cy10 | Change in 10 Year US Treasury Bond (first difference) | Number |
| cpi | US Consumer Price Index | Number |
| EP | Earnings from the past year divided by the closing value of the S&P 500 index | Number |
| DP | Total dividends paid by S&P 500 over past year divided by the closing value of the S&P 500 index | Number |

**Summary Statistics of the Sample Size**

Macroeconomic data and historical market movements from 1999 to 2020 made up the sample size considered in this analysis.



**Figure 1. Summary statistics of the data**

There are 254 non-missing observations for all variables, according to the count row. The average value for each variable across all observations is shown in the mean row. By looking at the minimum and highest values for each variable, outliers can be located. Outliers are present when values are significantly higher or lower than the rest of the data. A small standard deviation implies that the data points are closer to the mean and that the distribution may be more concentrated. The standard deviation offers information about the spread of the data around the mean.

The quartiles reveal the range of values for every variable. The lower and higher limits of the middle 50% of the data are represented, respectively, by the "25%" and "75%" quartiles. The center value of the data, or the median, is represented by the "50%" quartile. To determine the range of values in which most of the observations fall, utilize the interquartile range (IQR), which is the difference between the 75th and 25th percentile. Outlier values are those that fall outside the IQR. Overall, using the summary statistics is a simple and practical approach to acquire a broad picture of the data and see any potential problems or patterns.

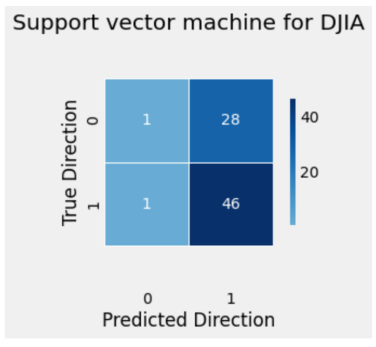
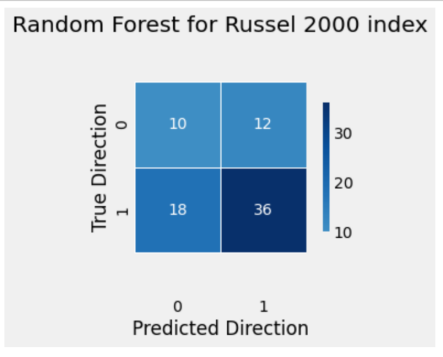
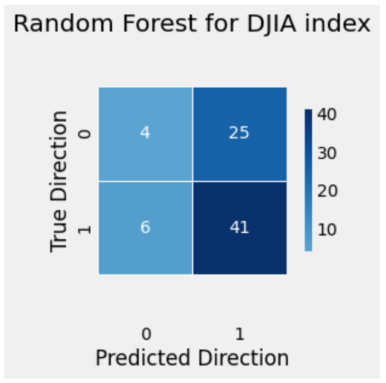
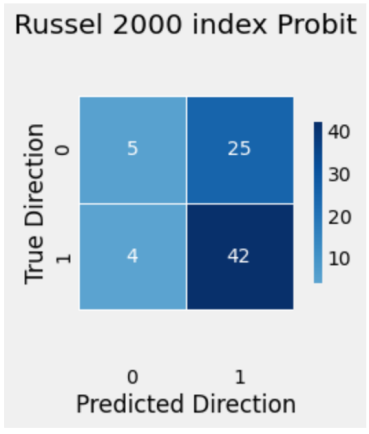
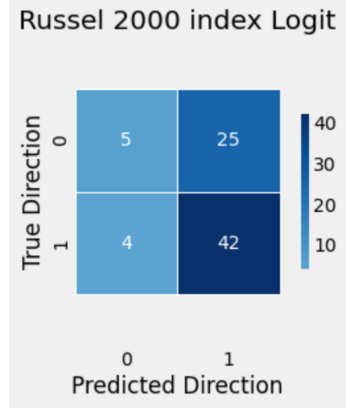
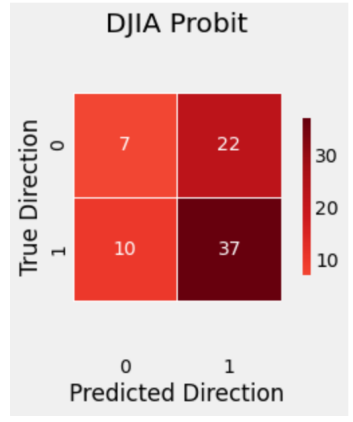
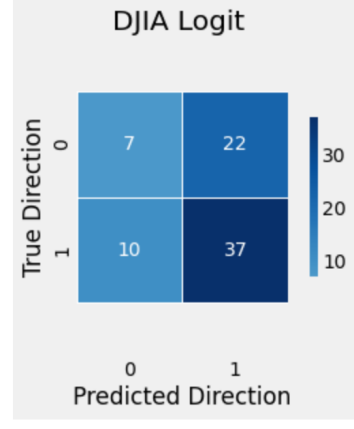
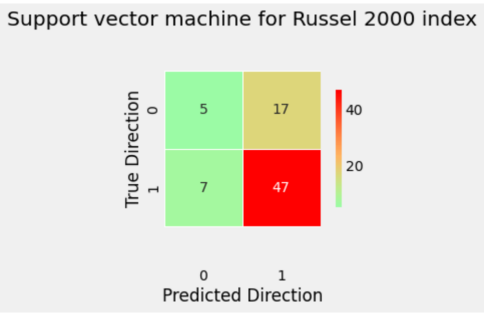
**Stepwise selection performed on the independent variables between DJIA and Russel 2000 index**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **"cm3"** | **"m3"** | **"y10"** | **"cy10"** | **"cpi"** |
| **DJIA** | 61.84% | 59.21% | 42.1% | 61.84% | 61.84% |
| **Russel 2000** | 64.47% | 64.47% | 53.94% | 63.15% | 68.42% |

**Table 1. Stepwise selection performed on the independent variables**

From the table above, we can see that Russel 2000 index outperform the DJIA

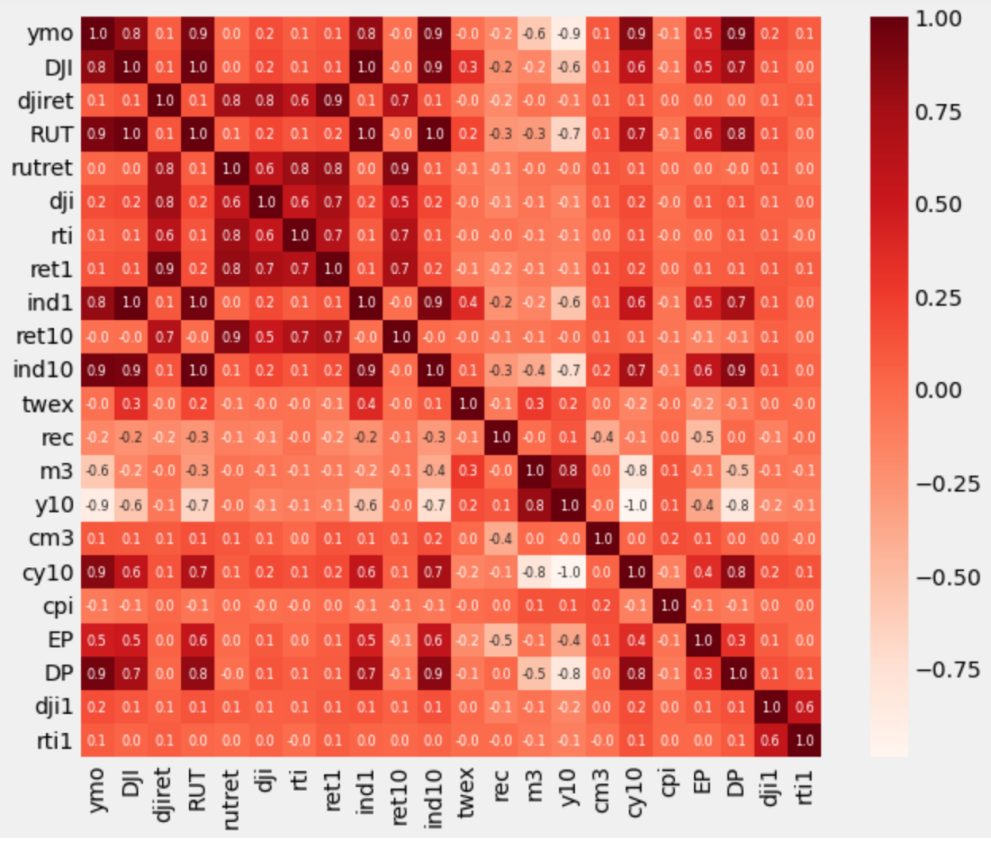
**Visualization**

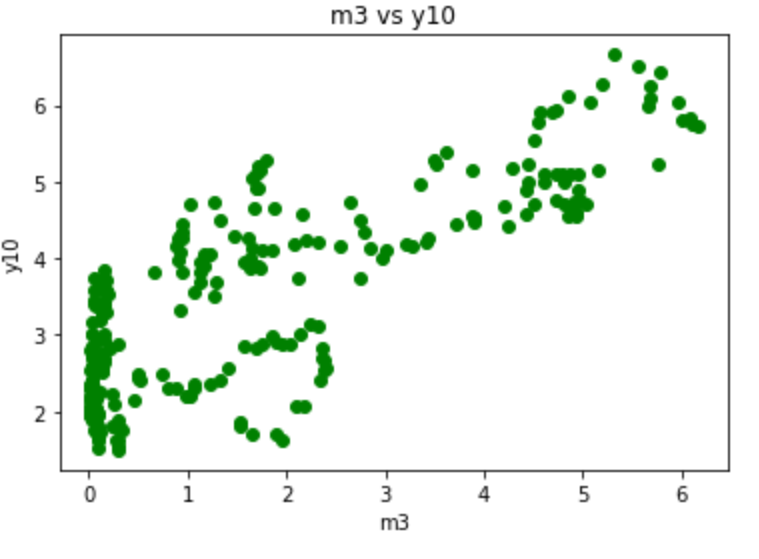
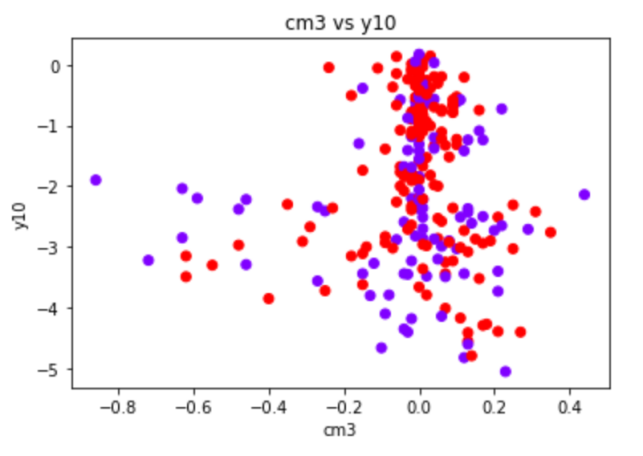
** **

**Figure 2 show the confusion matric of the models**

The diagrams above represent the confusion matrix of the Logit & Probit model for the Dow Jones Industrial average index with an accuracy of 57%. Also the Russel 2000 index for the Logit and the Probit regression with an accuracy of 61.8%. The Random Forest was used for the Dow Jones industrial Average index with an accuracy of 59.2% and the Random Forest for the Russel 2000 index with an accuracy of 60.5%. The Support vector machine was used for the Dow Jones industrial Average index with an accuracy of 61.8% and the Support vector machine for the Russel 2000 index with an accuracy of 68.4%.

**Correlation between the variables**

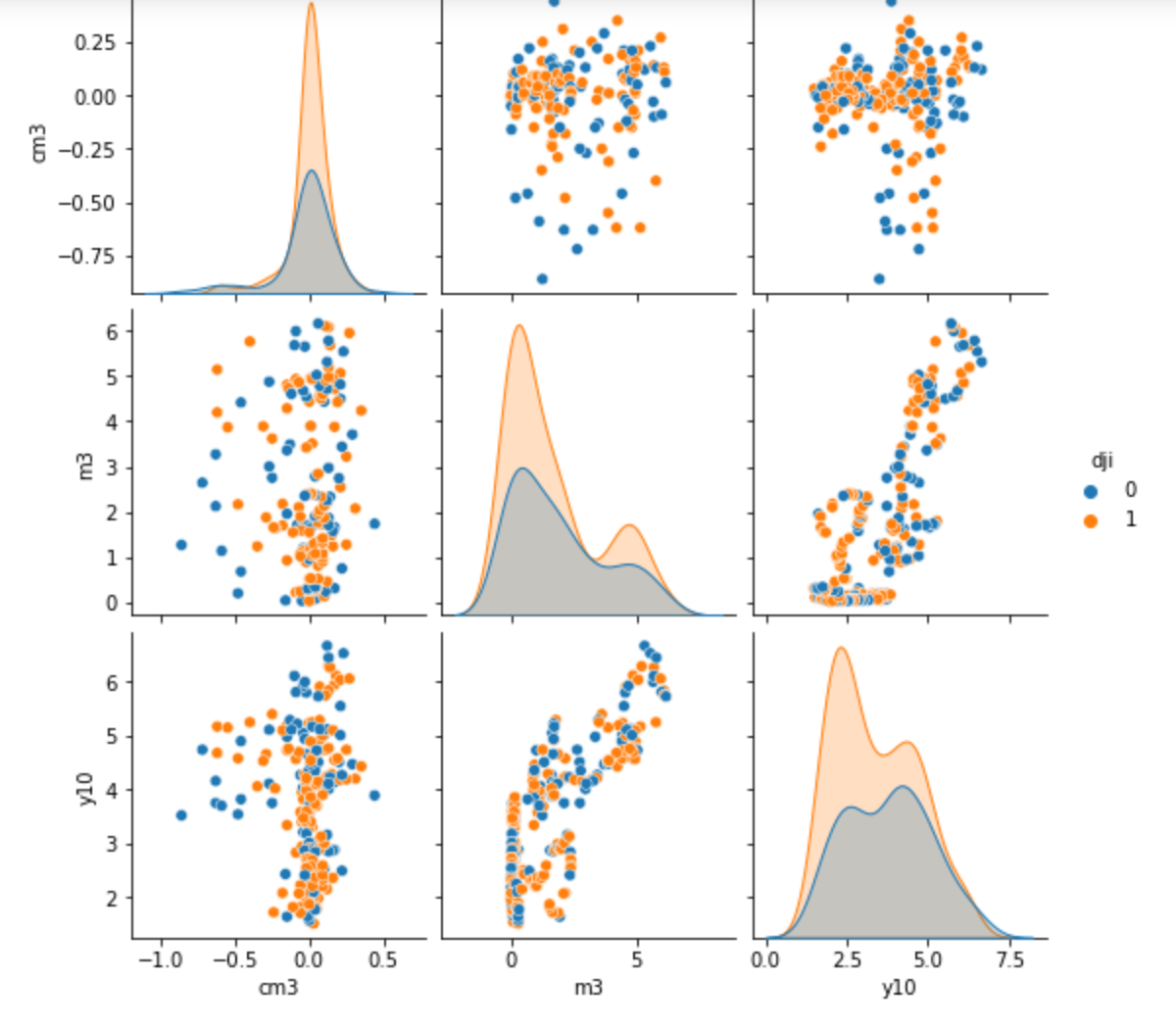
****

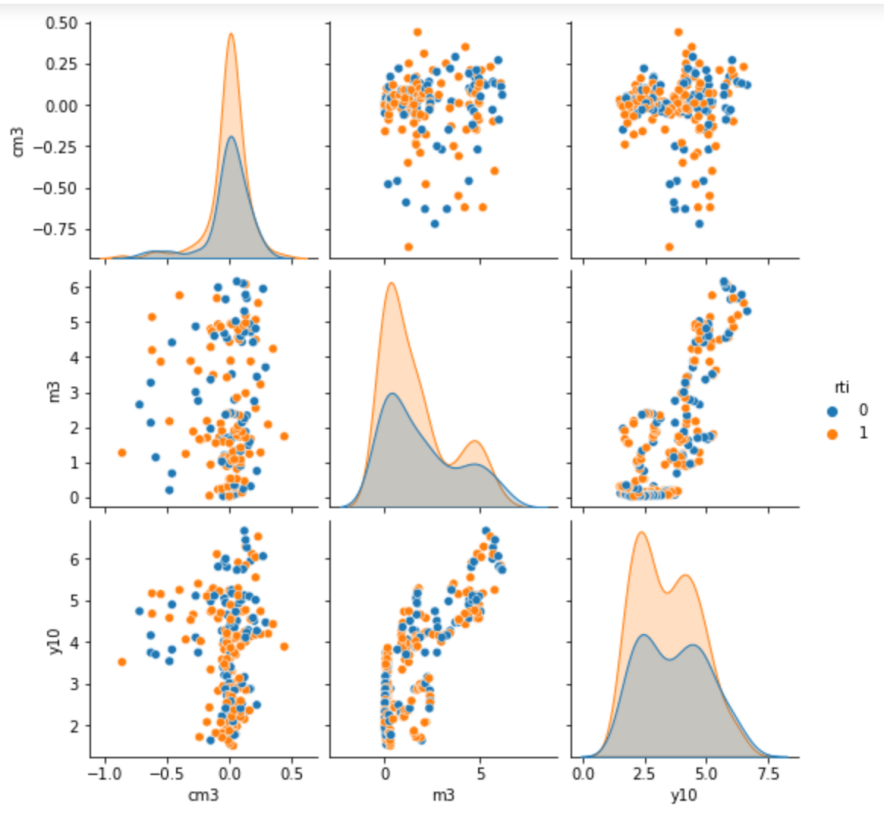
****

**Figure 3 Correction between variables.**

From figure 3 above, we can see that there’s a strong positive correlation between m3 and y10. As one variable increases, the other increases. We can see above that there’s no correlation between the y10 and cm3.

**Seaborn**





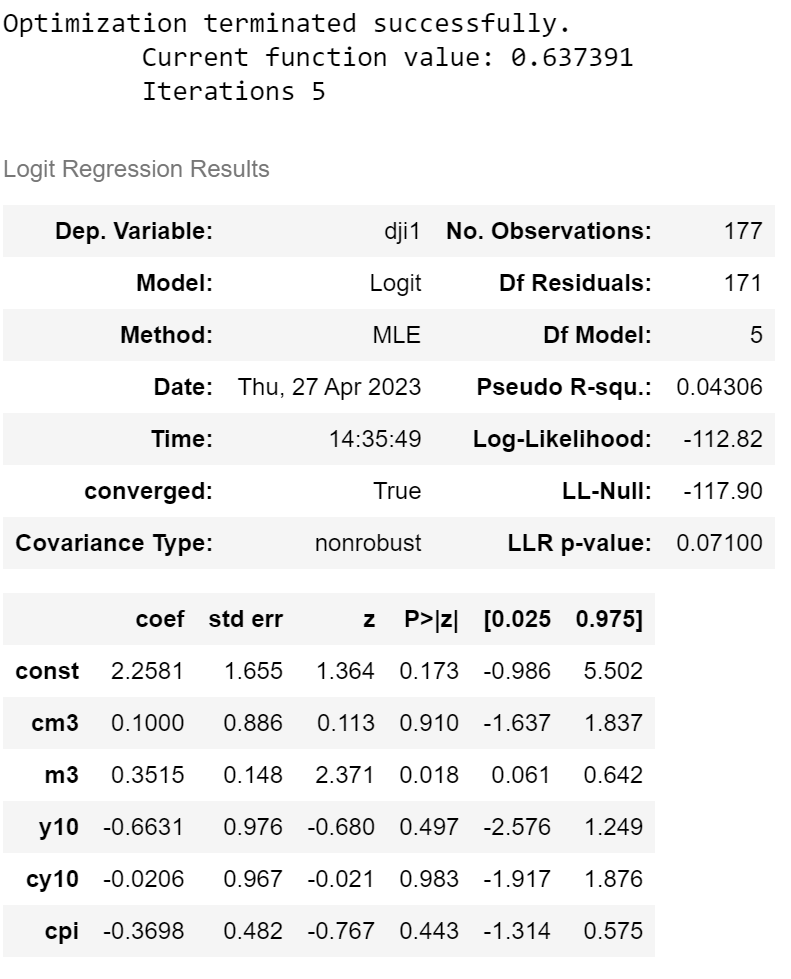
**Figure 4. Seaborn visualization of some variables**

**Results**

The results from the Logit and Probit regressions , also the Random forest and Support vector machine on the predicted direction of movement of the DJIA and the Russell 2000 index are presented in Tables below.

**Logit model for the Dow Jones Industrial average index**

The coefficient for the intercept term (constant) is 2.2581. The coefficient for "cm3" is 0.1000 with a standard error of 0.886. The p-value associated with this coefficient is 0.910, which is not statistically significant at the conventional 5% level, suggesting that "cm3" is not a significant predictor of "dji1" in the model. The coefficient for "m3" is 0.3515 with a standard error of 0.148. The p-value associated with this coefficient is 0.018, which is statistically significant at the 5% level, suggesting that "m3" is a significant predictor of "dji1" in the model. The coefficient for "y10" is -0.6631 with a standard error of 0.976. The p-value associated with this coefficient is 0.497, which is not statistically significant at the conventional 5% level, suggesting that "y10" is not a significant predictor of "dji1" in the model. The coefficient for "cy10" is -0.0206 with a standard error of 0.967. The p-value associated with this coefficient is 0.983, which is not statistically significant at the conventional 5% level, suggesting that "cy10" is not a significant predictor of "dji1" in the model.

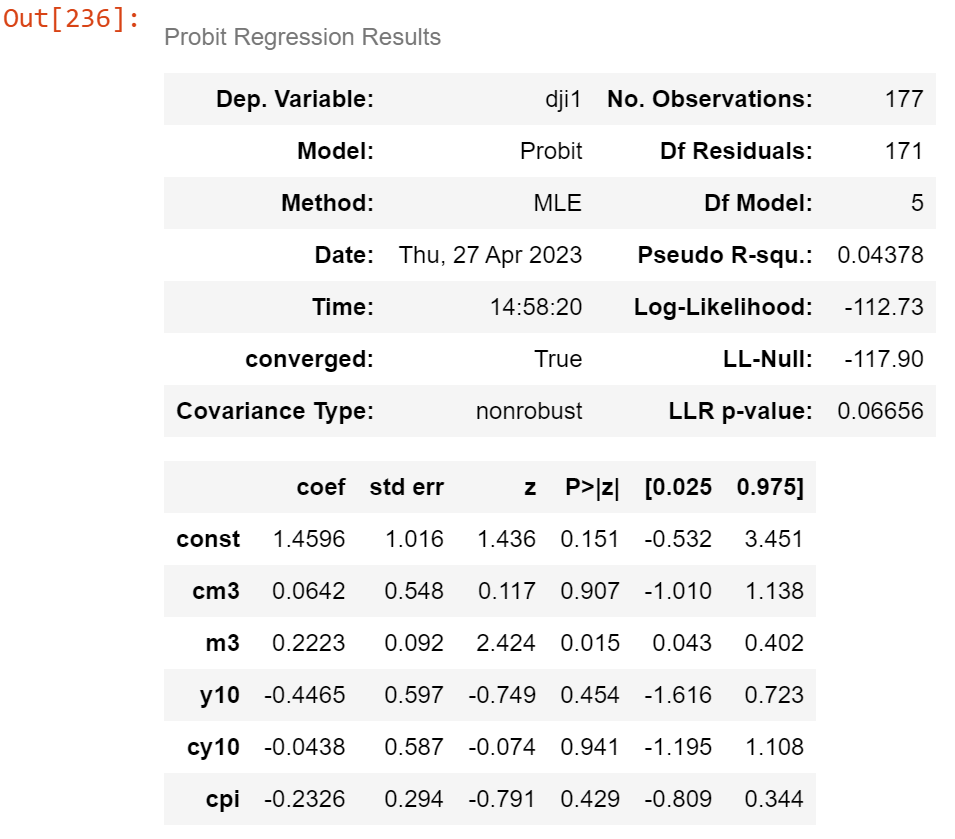


**Table 2. Logit model for Dow Jones Industrial average Index.**

The coefficient for "cpi" is -0.3698 with a standard error of 0.482. The p-value associated with this coefficient is 0.443, which is not statistically significant at the conventional 5% level, suggesting that "cpi" is not a significant predictor of "dji1" in the model. The model's pseudo R-squared value is 0.04306, indicating that the model explains only a small proportion of the variation in the dependent variable. The Log-Likelihood value for the model is -112.82, and the LL-Null value is -117.90, indicating that the model provides a slightly better fit to the data than the null model. Finally, the likelihood ratio test p-value is 0.07100, which is not statistically significant at the conventional 5% level, suggesting that the model as a whole may not be a significant improvement over the null model.

**Probit model for the Dow Jones Industrial average index**

The coefficient for the intercept term (constant) is 1.4596. The coefficient for "cm3" is 0.0642 with a standard error of 0.548. The p-value associated with this coefficient is 0.907, which is not statistically significant at the conventional 5% level, suggesting that "cm3" is not a significant predictor of "dji1" in the model. The coefficient for "m3" is 0.2223 with a standard error of 0.092. The p-value associated with this coefficient is 0.015, which is statistically significant at the 5% level, suggesting that "m3" is a significant predictor of "dji1" in the model. The coefficient for "y10" is -0.4465 with a standard error of 0.597. The p-value associated with this coefficient is 0.454, which is not statistically significant at the conventional 5% level, suggesting that "y10" is not a significant predictor of "dji1" in the model.

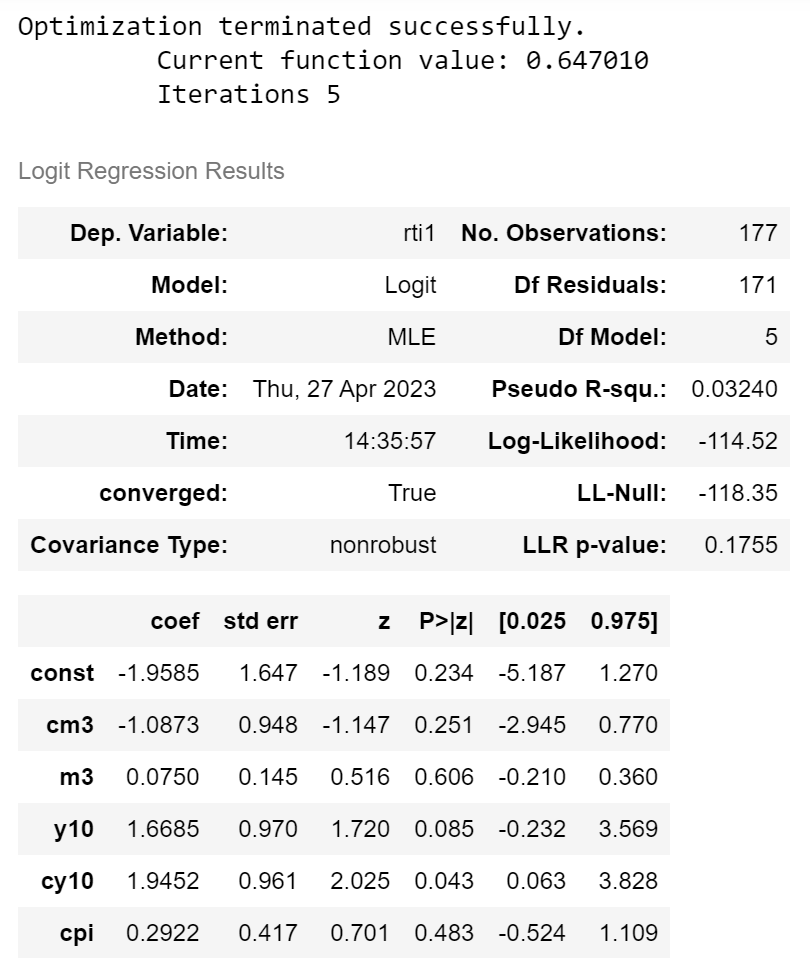


**Table 3. Probit model for Dow Jones Industrial average Index.**

The coefficient for "cy10" is -0.0438 with a standard error of 0.587. The p-value associated with this coefficient is 0.941, which is not statistically significant at the conventional 5% level, suggesting that "cy10" is not a significant predictor of "dji1" in the model. The coefficient for "cpi" is -0.2326 with a standard error of 0.294. The p-value associated with this coefficient is 0.429, which is not statistically significant at the conventional 5% level, suggesting that "cpi" is not a significant predictor of "dji1" in the model. The model's pseudo R-squared value is 0.04378, indicating that the model explains only a small proportion of the variation in the dependent variable. The Log-Likelihood value for the model is -112.73, and the LL-Null value is -117.90, indicating that the model provides a slightly better fit to the data than the null model. Finally, the likelihood ratio test p-value is 0.06656, which is not statistically significant at the conventional 5% level, suggesting that the model as a whole may not be a significant improvement over the null model.

**Logit model for the Russel 2000 index**

The coefficient for const is -1.9585, which represents the intercept of the regression line. This means that when all other independent variables are equal to zero, the log-odds of rti1 being 1 is -1.9585. The coefficient for cm3 is -1.0873, with a p-value of 0.251. This suggests that there is not a statistically significant relationship between cm3 and rti1, at a 95% confidence level. The coefficient for m3 is 0.0750, with a p-value of 0.606.

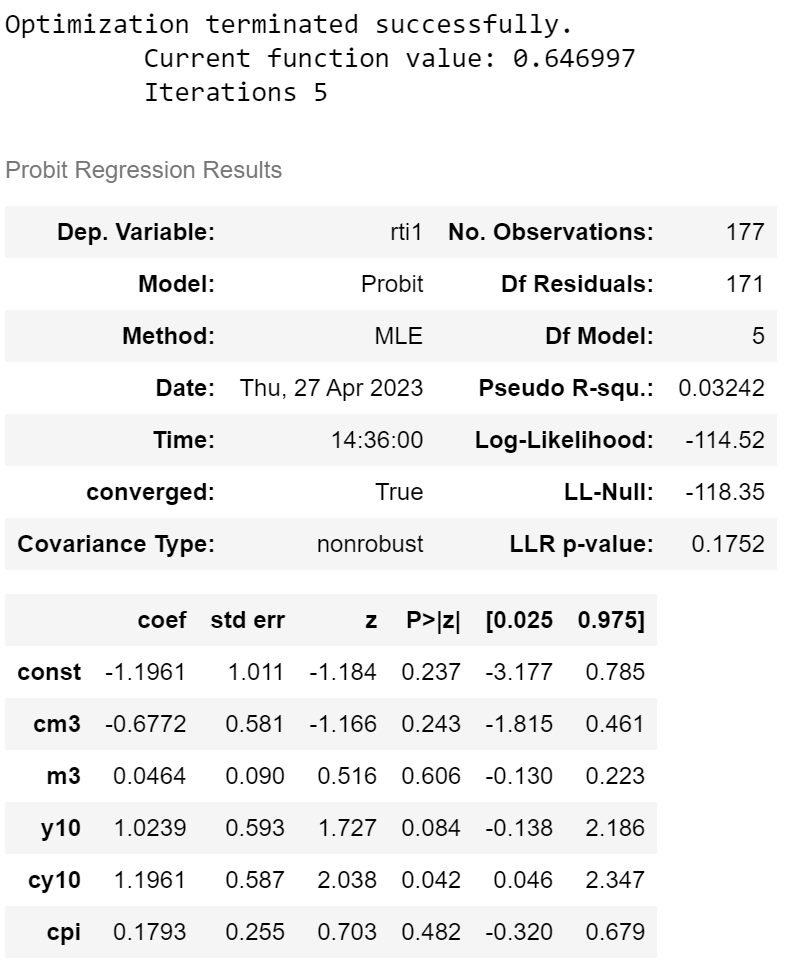
****

**Table 4. Logit model for Russel 2000 Index.**

This suggests that there is not a statistically significant relationship between m3 and rti1, at a 95% confidence level. The coefficient for y10 is 1.6685, with a p-value of 0.085. This suggests that there may be a statistically significant relationship between y10 and rti1, but the evidence is not strong enough to reject the null hypothesis at a 95% confidence level. The coefficient for cy10 is 1.9452, with a p-value of 0.043. This suggests that there may be a statistically significant relationship between cy10 and rti1, at a 95% confidence level. Finally, the coefficient for cpi is 0.2922, with a p-value of 0.483. This suggests that there is not a statistically significant relationship between cpi and rti1, at a 95% confidence level.

**Probit model for the Russel 2000 index**

The constant coefficient (-1.1961) represents the expected value of the dependent variable (rti1) when all the independent variables are equal to zero. The p-value (0.237) suggests that the constant is not statistically significant at the usual levels of significance (i.e., 0.05). The coefficient for cm3 (-0.6772) indicates that a one-unit increase in cm3 leads to a decrease of 0.6772 standard deviations in rti1. The p-value (0.243) suggests that this coefficient is not statistically significant. The coefficient for m3 (0.0464) indicates that a one-unit increase in m3 leads to an increase of 0.0464 standard deviations in rti1. The p-value (0.606) suggests that this coefficient is not statistically significant.



**Table 5. Probit model for Russel 2000 Index.**

The coefficient for y10 (1.0239) indicates that a one-unit increase in y10 leads to an increase of 1.0239 standard deviations in rti1. The p-value (0.084) suggests that this coefficient is marginally statistically significant. The coefficient for cy10 (1.1961) indicates that a one-unit increase in cy10 leads to an increase of 1.1961 standard deviations in rti1. The p-value (0.042) suggests that this coefficient is statistically significant at the usual levels of significance.

The coefficient for cpi (0.1793) indicates that a one-unit increase in cpi leads to an increase of 0.1793 standard deviations in rti1. The p-value (0.482) suggests that this coefficient is not statistically significant.

**Recommendation**

Based on our analysis, both the DJIA and Russell 2000 indices have relatively low accuracy scores for both the logistic regression and random forest models, which suggests that these models may not be the best predictors for the future performance of the indices.

However, the support vector machine model for the Russell 2000 index has shown promising results, with an accuracy score of 0.68 and a relatively high precision score of 0.61. This indicates that the model is able to correctly predict positive returns for the index with a relatively low false positive rate.

Therefore, if the investor is interested in using a long-short strategy and is looking for a reliable predictor of the future performance of the index, we would recommend considering investing in a fund which tracks the Russell 2000 stock index and using the support vector machine model as a tool for predicting the future performance of the index. However, we also recommend that the investor should monitor the predictability of both indices over time and use the long-short strategy outlined, where you buy the fund when the index is expected to rise and short it when it is expected to fall. This strategy can potentially increase your returns, especially if you can accurately predict the movements of the index.

However, it's important to note that investing in the stock market always carries some level of risk and past performance is not always indicative of future performance. Therefore, we would also recommend that the investor carefully considers their investment goals, risk tolerance, and overall financial situation before making any investment decisions.

**Conclusion**

Based on the analysis of the macroeconomic and historical data, we have built several machine learning models to predict the direction of the Dow Jones Industrial Average (DJIA) and the Russell 2000 Index (RUT). Firstly, we have conducted a logistic regression and a probit regression on the DJIA and RUT indices. The accuracy scores for the DJIA were 57.8% and 61.8% for the logistic and Probit regression models, respectively. The corresponding accuracy scores for the RUT were 61.8% and 60.5%. Based on these results, both the DJIA and RUT indices have some level of predictability, but the RUT may be slightly more predictable. Next, we compared the performance of the DJIA and RUT using a Random Forest and Support Vector Machine (SVM) classification model. The results showed that the SVM model outperformed the Random Forest model in terms of accuracy, precision, and recall for both indices. The accuracy score for the DJIA using the SVM model was 61.8%, while that for the RUT was 68.4%.

Based on these results, we recommend that you invest in the fund that tracks the RUT index. The RUT index has shown to be slightly more predictable than the DJIA, and the SVM model had better performance in predicting the movement of the RUT index. Additionally, the RUT has a higher accuracy score than the DJIA using the SVM model. In conclusion, the RUT index may be a better investment option compared to the DJIA index. However, we recommend that you continue to monitor the predictability of both indices over time to ensure that you are making the best investment decisions as market conditions may change and affect their predictability.

**Reference**

Banton, C. (2022). An Introduction to U.S. Stock Market Indexes. *Investopedia*. <https://www.investopedia.com/insights/introduction-to-stock-market-indices>

Cfa, B. B. (2023b). What is the Russell 2000 Index? *Bankrate*. https://www.bankrate.com/investing/what-is-russell-2000/

Dataman, C. K. (2023, January 30). Machine Learning or Econometrics? - Analytics Vidhya - Medium. *Medium*. <https://medium.com/analytics-vidhya/machine-learning-or-econometrics-5127c1c2dc53>

Fabbri, M., & Moro, G. (2018). Dow Jones Trading with Deep Learning: The Unreasonable Effectiveness of Recurrent Neural Networks. In *International Conference Data Science*. https://doi.org/10.5220/0006922101420153

Hayes, A. (2022). Econometrics: Definition, Models, and Methods. *Investopedia*. <https://www.investopedia.com/terms/e/econometrics.asp>

Hull, I. (2021b). Machine Learning and Economics. In *Apress eBooks* (pp. 61–86). Apress. <https://doi.org/10.1007/978-1-4842-6373-0_2>

Joseph. (2022, August 15). Econometrics vs Machine Learning: Which is Better? - reason.town. *reason.town*. <https://reason.town/econometrics-vs-machine-learning/>

Maitra, S. (2021, December 15). Future Price Prediction Beyond Test Data Using Vector Auto Regression. *Medium*. <https://towardsdatascience.com/future-price-prediction-beyond-test-data-using-vector-auto-regression-eedb7e0c04e>

Osu, B. O., Okonkwo, C., Uzoma, P. U., & Akpanibah, E. E. (2020). Wavelet analysis of the international markets: A look at the next eleven (N11). *Scientific African*, *7*, e00319. https://doi.org/10.1016/j.sciaf.2020.e00319

Sahu, A. (2022, July 19). *Advantages and Disadvantages of Machine Learning*. West Agile Labs Blog. <https://www.westagilelabs.com/blog/pros-and-cons-of-implementing-machine-learning-in-your-projects/>

Segal, T. (2021). DJIA 101: How Does the Dow Jones Work? *Investopedia*. https://www.investopedia.com/investing/what-moves-the-djia

Team, N. (2022, February 3). *Advantages of Machine Learning | Disadvantages of Machine Learning - Nixus*. Nixus. https://nixustechnologies.com/advantages-disadvantages-of-machine-learning/

*What are Machine Learning Models?* (n.d.). Databricks. <https://www.databricks.com/glossary/machine-learning-models>

Yoo, S., Jeon, S., Jeong, S., Lee, H., Ryou, H., Park, T., Choi, Y., & Oh, K. (2021). Prediction of the Change Points in Stock Markets Using DAE-LSTM. *Sustainability*, *13*(21), 11822. https://doi.org/10.3390/su132111822

窗体底端