Random Forest & Forecasting Stock Prices After Covid-19

BAN 525

Khayrayyah Haamid-Day

This analysis will be various modeling methods to examine the possible predictors of weekly stock returns with a focus on returns post Covid-19. There will be two separate models developed to predict the response variables of RSPY and Sign of RSPY. The Sign of RSPY is a nominal variable with values of “1” or “0”, the former indicating positive returns while the latter indicates otherwise. The predictor variables to be used consist of a variety of economic and financial variables that measure the returns from currency, interest rates, and commodity markets stocks. The statistical methods that will employed are Standard Least Squares, nominal logistic regression, and random forest. Only random forest will be used for both response variables while the Standard will be used for RSPY and logistic regression for Sign of RSPY.

Each statistical method has their advantages and disadvantages and exploring different methods will provide a clear comparison of which models are the strongest. The first method that will be used is the Standard Least Squares method, this is primarily employed as a baseline model. It should be noted that this method will only be used for the response variable RSPY, and this is due to this method fitting models where the response is continuous. This method is great when used as a benchmark and is not difficult to understand and explain. Though this method does tend to be sensitive to outliers and create overfit data. Due to this, it is expected to return a high RSquare on the training set but a lower RSquare on the test set. This is unsurprising as this method may model random noise.

The next method that will be employed is the nominal logistic regression which will be used for the response variable Sign of RSPY. This method is used for Sign of RSPY because the response variable has two levels, a “1” and a “0”. This method is simple to understand but will fit the entire method as variable selection will not be performed. In this way, it is like Standard Least Squares and may return similar values. The last method that will be used is random forest which will be used for both response variables as this method can be used for classification problems as well. Random forest splits the data, creating randomly constructed decision trees from only a part of the sample of the data. A great aspect about each decision tree is each is uncorrelated to each other and is not high influenced by outliers. The downside with this method is that it may not be as easily interpretable as the other methods and can be computationally intensive with larger datasets. A cross validation will be implemented in order to reduce random noise that could cause overfit models. This cross validation is split into a training, validation, and test set as 60/20/20. The training set will be used initially in the building of the model, the validation set is the data that is held out, and the test set is complexly withheld during the model building. This will ensure that the test set provides an unbiased evaluation of the model’s performance.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, table

Description automatically generated with medium confidenceTable

Description automatically generatedText, table

Description automatically generated with medium confidence

The following images are results from the Standard Least Squares, nominal logistic regression, and random forest methods. As predicted previously, the Least Squares method returned a high training set and much lower values for the validation and test set. The nominal logistic regression method returned a perfect RSquare on the training set, for both the entropy and generalized RSquare, while the test set is negative. The random forest method for the response variable RSPY returned a high training set value while the validation and test set were lower. The random forest method for the response variable Sign of RSPY returned values that were much steadier between the training, validation, and test set. This brief overview indicates that the best models for both the response variables will likely be the one that uses the random forest method.

Graphical user interface

Description automatically generated  
Table

Description automatically generated

The first model comparison is for the response variable of RPSY, and as predicted, the random forest method has performed better in comparison to the OLS method. This is evident with the test set RSquare value of 0.55 in comparison to 0.44 as well as lower values of RASE and AAE. The second model comparison is for the response variable of Sign of RSPY = 1 and with this model the random forest appears to have outperformed the nominal logistic regression in the validation and test set. In fact, the nominal logistic regression only returned a perfect RSquare for the training set but all negative values for the remaining sets. In comparison, random forest returned a 0.45 on the entropy test set and a 0.61 on the generalized test set, RASE was also lower for the random forest method. The remaining columns, except for Mean Abs Dev, all contain smaller values for random forest compared to logistic regression, in this case it is the smaller the better. Due to the results from this model comparison, it is clear that the random forest method has produced the best fit model for both response variables.

Chart

Description automatically generated

Graphical user interface, text, table

Description automatically generated

Since the final model utilizes the random forest method, there is the addition of the column contribution along with the variable importance analysis. The column contribution is a report that shows similar information as the variable importance report does, though it is calculated with a different algorithm. With this model, for the response variable RSPY, the column contribution states the top 3 predictor variables RHYG, REMB, and RVIX contribute the most to the fit. RHYG accounts for almost 50% of the fit which indicates a strong relationship with response variable. The variable importance report returned the same order of importance.

Table

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

The final model for the response variable Sign of RSPY = 1, “1” meaning positive stock returns, returned many of the same variables except for RFXC. The implication of this is that RVIX, RHYG, RFXC, and REMB provide positive stock returns for the stock RSPY. In the case of this model, predictor variable RVIX accounts for almost half of the fit. The variable RHYG is a stock that tracks high yield corporate bonds, otherwise known as junk bonds, which is a highly volatile debt instrument. RVIX is, on the other hand, is a measure of stock market volatility. REMB tracks international emerging market bonds and RFXC tracks the Canadian dollar against the US dollar. The response variable RSPY tracks the performance of the S&P 500. The impact that bonds have the stock market is an inverse one, as the interest rates decrease, the stock prices rise. Exchange rate stocks can impact the stock market due to the relationship between interest rates and the value of the US dollar. The prediction profiler further elaborates on the type of relationship between the response variables and these predictor variables.

Chart, box and whisker chart

Description automatically generated

Chart

Description automatically generated

These images indicate that there is a positive relationship between RSPY, RHYG, and REMB; there is also a positive relationship between Sign of RSPY and RVIX. While there is a negative relationship with RVIX and RSPY as well as RHYG, RFXC, and Sign of RSPY. This can be interpreted to mean that if RHYG and REMB is increasing, so is RSPY and if RVIX is increasing than RSPY is decreasing. For Sign of RSPY, the opposite is true for RVIX and RHYG as well as RFXC. The returns of these securities were collected during the time the Federal Reserve implemented a series of policies because of the impact Covid-19 had on the economy. One of the policies involved lowering the interest rate which makes borrowing money less expensive, this also encourages spending and investing. This also encourages investors to move money from the bond market to the equity market that results in a rise in the equity market.

This analysis provides clarity on the financial market and the impact particular stock have on RSPY. The main objective is to determine when an investor should sell and/or buy stocks, due to this it is important to understand how stocks like RHYG and RVIX are performing in the market. For instance, if an investor wants to know the best time to sell RSPY then looking at how RHYG is performing in the market can help them determine this. Though the actions of the Federal Reserve should be considered when looking at debt instruments. In conclusion, both models returned similar results which indicate there is in fact a relationship between RSPY and these predictor variables.