Mod 1 – Risk Factors for Oil Prices: Before & After Covid-19

BAN 525

Khayrayyah Haamid – Day

This analysis seeks to understand the relationship between oil price changes (RUSO) and other variables, such as stocks, bond prices, and exchange rates. The potential risk factors to oil price changes will also be analyzed. In this instance, the risk factor is Covid-19 which has caused an immense amount of disruption in the financial markets. The results from the pre and post Covid datasets can provide further insight into what factors impact oil prices.

The pre-Covid dataset contains data collected from 2014 to 2019 and there are 1332 observations. The post-Covid dataset contains data collected from January 2020 to April 2021 and there are 326 observations. Both datasets contain the same variables that will be used as predictors for this analysis. The response variable that will be used is RUSO while the predictor variables range from stocks, bonds, exchanges rates, crypto, and ETFs. The analysis will be done by creating three different models, the first will be a standard linear regression and the last two will be stepwise forward and backward regression. The three models will be compared in order to determine the most successful predictive model for RUSO.

The first model created is using the method standard least squares. This model tends to produce a higher RSquare on the training set while producing a low and negative RSquare on the validation and test set. This is due to the fact that this method is not capable of performing model variable selection so it will model everything including random noise. The second model created uses the forward stepwise method, this method begins with no variables and will start adding variables until the stopping rule is satisfied. For this model the stopping rule used was max validation RSquare. The third model created uses the backward stepwise method, this method works opposite of the forward method by beginning with a full model that removes the least significant variables until the max validation RSquare is reached. The most obvious advantage to using the forward method over the backward method is that forward stepwise works well with a larger number of variables simply because it begins with an empty model. Though the backward stepwise method is still great for its ability reduce collinearity which will also reduce the likelihood of the model being overfit. Overall, both stepwise methods produce easily interpretable and objective models, but they can also be biased models with RSquare values being too high. The pre and post Covid datasets will be using all three methods to determine the best predictive model.

Each model will use cross-validation in order to reduce random noise that could result in an overfit model. The cross-validation procedure involves creating a training, validation, and testing set that will be split up as 60/20/20. The training set is the portion of data that will be used for building the model, the validation set is the data held out, and the testing set is withheld during the model building. The testing set will provide an unbiased evaluation of the model’s performance.

Graphical user interface, text

Description automatically generated Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application

Description automatically generated

The post Covid-19 dataset is used with this cross-validation and three models are created with the first model using the standard least squares method. This model returned a high RSquare value for the training set but very poor, and negative, values on the validation and test set. The second model employed the forward stepwise method, this model returned decent values on the training and test set but a low value on the validation set. The third model used the backward stepwise method which returned similar values as the first method, very high training set with negative validation and test sets. Due to these values, I would estimate that the final model, that is the model with the best performance, would likely be the model that used the forward stepwise method.

Graphical user interface, table

Description automatically generated

The model comparison returned expected RSquare values for each method used. The standard and backward method showed an identical high value of 84% on the training set, while the forward method showed a decent value of 31%. The validation set was negative for both standard and backward method, while the forward stepwise method has a low and positive value of 11%. The testing set for the standard and backward method was, once again, negative values. The forward stepwise method testing set is 33% and a few values higher than the training set. The test set for the forward stepwise method also had the lowest RASE and AAE values. The results from the model comparison shows that the forward stepwise method is the best performing model for the post Covid-19 dataset.

Graphical user interface, text, application

Description automatically generated Text

Description automatically generated Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

The pre Covid-19 dataset will use the same cross-validation method and same model methods. The first model uses the standard least squares method, this model returned a high RSquare value of 70% on the training set and negative values for the validation and test set. The second model uses the forward stepwise method, this model had good RSquare value of 65% for the training set that decreased by more to 47% on the validation set and 48% on the test set. Despite this, these values are still high. The final model uses the backward stepwise method that has high RSquare value of 70% on the training set but negative values for the validation and test set. The forward stepwise test set also shows a lower RASE and AAE value when compared to the standard and backward method. Based on the RSquare value of the test set, the model that uses the forward stepwise method is the best performing model.

Graphical user interface, text, application

Description automatically generated Chart, line chart

Description automatically generated

The final models for the pre and post Covid-19 dataset uses the forward stepwise method as it returned the highest RSquare values on the test set as well as the lowest RASE and AAE values. The post Covid-19 final model only returned a single variable, RXLE, that has a significant relationship with the response variables, with a p-value of <.0001. Due to this, there is no variable importance analysis, RXLE is the most important variable. This model’s profile does show that RUSO and RXLE have a positive relationship, when returns for this sector fund increase so does the returns for oil.

Table

Description automatically generated Table

Description automatically generated

Chart, line chart

Description automatically generated

The pre Covid-19 final model returns eleven variables that show significance. The parameter estimates highlight that while each variable is considered significant, there are a few that are <.0001 while others are between <.0001 and <0.05. The four variables that show a p-value of <.0001 are RSPY, RXLE, RTIP, and RFXC. The variable importance analysis shows that the two most important variables are RXLE and RSPY. The models profile highlights that, much like with the final model for the post Covid-19 dataset, as returns on RXLE increases so does RUSO. On the other hand, as returns decrease for RSPY, RUSO increases which shows that the two have an inverse relationship.

These final models and their most important variables are easy to investigate in real time, as oil prices are currently increasing. In the case of RXLE, these shares are currently increasing as oil is increasing. This relationship makes sense as these are energy sector stocks from companies that produce and supply energy products. RSPY is an ETF that tracks the S&P 500 index, and the S&P 500 consists of 500 of the top publicly traded companies in the U.S. The inverse relationship that RSPY has with RUSO could be speculated on but there is no concrete correlation though there is a historical relationship between the two. This analysis also sought to consider the effect of Covid-19 and while the test set for the pre Covid has a higher RSquare value, it is not that much higher than the post Covid model. Furthermore, both datasets return RXLE as a variable of significance to the response variable, though the pre Covid model does also include RSPY which is not present in the post Covid model. Due to this, it could mean that Covid was not much of a risk factor for the response variable. In conclusion, the final models appear to accurately predict the response variable RUSO as this is currently being shown in the stock market today.