Mod 2 – Penalized Regressions and Clean Energy Stocks

BAN 525

Khayrayyah Haamid-Day

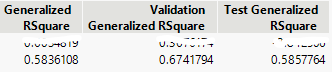
This analysis aims to understand renewable energy stock prices and potential determinants of these stocks returns. For instance, investigating the relationship between oil and clean energy and how prices could impact the attractiveness of clean energy. Technology stocks are another important consideration as clean energy requires advanced technology to be created and work. This analysis will be using a large dataset that contains values related to stocks, currencies, bonds, market sentiment measures, and cryptocurrency. These values were collected between September 17, 2014, and April 23, 2021, with a total of 1,659 observations. This analysis will be focusing on examining the variable RPBW, the response variable, and all other variables will be the predictor variables. The variable RPBW shows the returns of the clean energy ETF that it tracks. The statistical methods that will be used are, OLS as a benchmark, Lasso, Adaptive Lasso, Elastic Net, and Adaptive Lasso with a t(5) and Cauchy distribution.

The first model created is using the method standard least squares. This model is expected to produce a higher RSquare on the training set while producing a low and negative RSquare on the validation and test set. This is because this method is not capable of performing model variable selection so it will model everything including random noise. One model will be created using Lasso. Lasso is a penalized regression method that will shrink the coefficients towards zero while also selecting variables that will build a better, and easy to interpret, prediction model. The disadvantage to this is Lasso may ignore insignificant variables that could still be of interest. Another model will be created using Adaptive Lasso which is simply an modification of Lasso. This method adds weight to the parameters, and this will attempt to counteract the issue of Lasso estimates being biased. Two other models will be created using Adaptive Lasso with t(5) and Cauchy distributions. The t(5) distribution is most useful with a small sample size and when the population standard deviation is not assumed. Cauchy distribution is useful for data that has large number of outliers. The use of these distributions is done in the hopes of producing better quality models. The next model will be created with Elastic Net and this model applies both penalties. One penalty ensures that variable selection occurs, and the second penalty shrinks the coefficient. The final model to be created will be the Adaptive Elastic Net model, which is a similar modification as Adaptive Lasso.

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. 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.

Text

Description automatically generatedGraphical user interface, text, application

Description automatically generated Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated Graphical user interface, text

Description automatically generated

Text

Description automatically generated

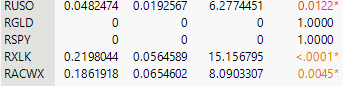
The first model created is the standard least squares model and this model returned a RSquare of 60% on the training set, a 36% on the validation set but a negative test set. The Lasso model had more stable RSquare values, the training set was at 58%, the validation set increased to 67%, and remained at 58% on the testing set. The Adaptive Lasso model returned similar RSquare values of 55%, 68%, and 62%. The Elastic Net model returned the same RSquare value as the Lasso model, 58%, 67%, and 58%. The Adaptive Elastic Net model had values of 56%, 67%, and 62% on the testing set. The Adaptive Lasso model that utilized the t(5) distribution began with a good training and validation score, 58% and 67%, but became negative on the test set. The final model was the Adaptive Lasso model with a Cauchy distribution and this model returned the exact same RSquare on the training and validation set, while the test set was not negative, it was very low.

Table

Description automatically generated

The model comparison returned all the RSquare values for each method used. This comparison highlighted that many of the models shared similar values with the training and validation set. The test set was also very similar among the various models, with the exception of the least square method which returned a negative value. Focusing on the test set, there are three models that share very similar RSquare values. The Adaptive Lasso, Adaptive Elastic Net, and Adaptive Lasso with the t(5) distribution. By the slightest amount, Adaptive Lasso method returned the highest RSquare value along with the lowest RASE and AAE. These results show that the Adaptive Lasso method is the best performing model for this dataset.

Graphical user interface, text, table

Description automatically generated



The Adaptive Lasso final model returned test set value of 62% and revealed there were 5 variables found to have a relationship with the response variable RPBW. Of these 5 there 2 that showed the most significance with a p-value of <.0001, RXLK and RSLY. The summary report reinforced this significance and highlighted RSLY as having the most importance to the response variable. This variable, RSLY, also was shown to have a positive relationship to RPBW, as the returns increase for one so does the returns increase for the other. It should be noted that the other 4 variables, though not as important as RPBW, also showed a positive relationship with RPBW, though SPY did show any relationship with RPBW.

Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated

The stock RSLY is a small cap 600 ETF, which is a index fund that tracks a range of small sized companies. Due this information, it would be difficult to determine which industries are positively impacted by renewable energy. It can be speculated that perhaps this relationship exists due to investment that companies are making with the creation and use of renewable energy. This same is not true for the stock RXLK, this is a technology select sector SPDR fund that tracks an index of S&P technology stocks. This type of relationship was assumed to exist with renewable energy before this analysis was conducted and appears to be proven with this final model. The third variable is RACWI and this index stock tracks investments returns from large to mid capitalization non-US equities. This stock is similar to RSLY in that perhaps this relationship exists due to investments companies included in this stock track. The fund RUSO, the U.S. oil ETF, was predicted earlier to have some kind of relationship with renewable energy and though this is not a strong relationship, it is interesting to see another prediction make an appearance in this final model. The final variable included in this model is SPY, but this relationship is not strong and does not appear to be a good predictor of RPBW.

In conclusion, it is clear to see that this model provided valuable insight into the potential predictors of a clean energy ETF. I believe this model also highlighted that the investment into renewable energy is being made by numerous industries, including internationally.