Christopher Dillard 9/10/2023

BAN 525

Module 2: Assignment 1 – Penalized Regression Models and Determinants of Clean Energy Stocks during Covid-19

The Covid-19 pandemic disrupted and transformed global market dynamics across many industries. In this assignment, my focus is on the determinants of Clean Energy Stocks during the pandemic, in particular the PBW stock and whether the results align with traditional assumptions. I explored these effects through empirical analysis of daily stock prices from January 2, 2020 to June 3, 2022, giving a total of 611 observations.

Using penalized regression methods, referred to as generalized regression methods in JMP, I evaluated which method is most effective at predicting the response variable in the test data. In penalized regression models, coefficients (parameters) are subjected to a penalty term, which serves to minimize the sum of squared errors and ideally prevent overfitting while also narrows the variables to only the ones that are deemed to be most relevant. For comparative purposes, I also ran a standard linear regression model to evaluate whether a complex model is actually necessary for this data and more importantly showcase the power of more robust models when handling data that does not follow a normal distribution given that economic and financial data tends to fit this narrative and extreme values are not uncommon.

In total, I ran 6 different models on the data, being: Standard Linear Regression (Ordinary Least Squares), Lasso, Adaptive Lasso, Elastic Net, Adaptive Elastic Net, Adaptive Lasso with Student's T distribution (called t(5) in JMP), and finally Adaptive Lasso with Cauchy distribution.

Ordinary Least Squares (OLS) is the simplest of all the methods that I used in this assignment. Where this method falls short compared to the others, is in how it handles large number of variables relative to observations. While the dataset has a total of 611 rows, only 60% (366 rows) will be used in the training set of the data. This aspect of the data could lead to an OLS model overfitting the data. Furthermore, OLS models assume linear relationships between predictor variables and the response variable, which is where it might not be suitable for real world data since fluctuations are prone to happen due to how many sectors follow business cycles throughout the year and, probably the most significant reason, is how extreme values cause by the random events and human interaction reduce the likelihood of perfectly linear relationships between variables.

Lasso (Least Absolute Shrinkage and Selection Operator) handles variable selection by shrinking irrelevant variables to 0, effectively eliminating them altogether from the model. This provides for a better suited model to handle newer data by only focusing on variables that are empirically shown to be the best at predicting the response variable in the test data.

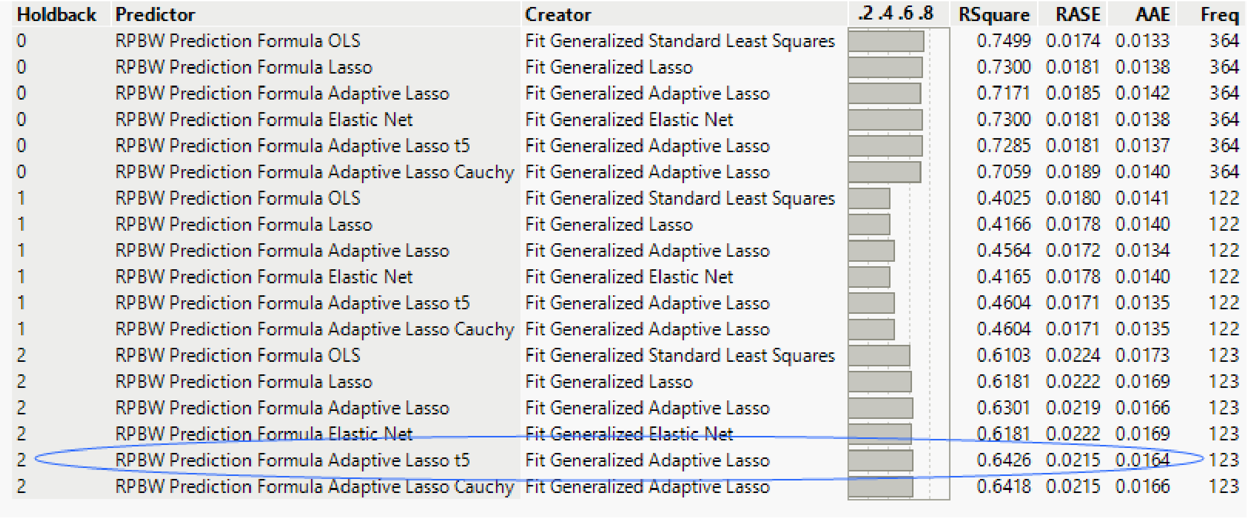
Elastic Net is a blend of the Lasso method and the Ridge method. This combination results in a model that has the enhanced variable selection capabilities of Lasso, with the addition of still considering how less important variable can have an impact on the predictive capabilities of the model, albeit to a small degree. So, the Elastic Net method can eliminate variables that it absolutely determines to be irrelevant and maintain some variables that would have otherwise been eliminated by Lasso. In this way, it is up to the analyst to determine which variables to keep after running and comparing the performance of the two models.

The ”adaptive” elements in these other methods’ names refers to how penalties are applied to coefficients on a weighted scale in such a way that less relevant variables are subject to more shrinking towards zero, if not completely eliminated. The adaptive versions of Lasso and Elastic Net are helpful to use when variables of interest are already known. Using initial estimates from the Ordinary Least Squares model, analysts can gain an idea of the weights of each predictor variable in terms of its impact on the response variable. One potential problem with adaptive methods is where predictors are highly correlated, otherwise known as multicollinearity, since this phenomenon could lead to uneven penalty applications to variables that in theory should be weighted equally, in which case uneven penalty applications can introduce bias in the model and greatly inflate the importance of one variable over another one of actually similar importance.

The last two methods I used were both adaptive Lasso models, with the difference that one was set to assume the Student’s T (t5) distribution while the other one assumed a Cauchy distribution. Both of these distributions imply the presence of heavy tails where the mean and standard deviation are less defined than in a normal distribution that would follow a bell curve with a single peak and no skewness. As I’ve previously mentioned, extreme values are not uncommon in financial datasets, especially when it pertains to stock prices. This reason alone would make it wise to test both of these methods in addition to the already very robust, adaptive lasso method to see which distribution is the best fit for the dataset. The primary difference between Student’s T and Cauchy distributions is just the level of heaviness assumed in the tails, with the Cauchy distribution being more spread out away from the mean than a Student’s T distribution. So, in short, we can run both of the tests to determine how much variance is observed around the mean in the dataset.

As in the previous assignment, the data is split 60-20-20 into training, validation, and testing sets, respectively. My interest when it comes to model selection, is directed towards each model’s ability to predict the changes observed in the testing set (i.e., The last 20% of the time series data from December 8, 2021 to June 3, 2022).

To determine which model was the best fit for this assignment, I first went to the Analyze drop-down box, then selected the Predictive Modeling option, and lastly chose the Model Comparison option, which resulted in the table shown below:



Holdback	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
0	RPBW Prediction Formula OLS	Fit Generalized Standard Least Squares		0.7499	0.0174	0.0133	364
0	RPBW Prediction Formula Lasso	Fit Generalized Lasso		0.7300	0.0181	0.0138	364
0	RPBW Prediction Formula Adaptive Lasso	Fit Generalized Adaptive Lasso		0.7171	0.0185	0.0142	364
0	RPBW Prediction Formula Elastic Net	Fit Generalized Elastic Net		0.7300	0.0181	0.0138	364
0	RPBW Prediction Formula Adaptive Lasso t5	Fit Generalized Adaptive Lasso		0.7285	0.0181	0.0137	364
0	RPBW Prediction Formula Adaptive Lasso Cauchy	Fit Generalized Adaptive Lasso		0.7059	0.0189	0.0140	364
1	RPBW Prediction Formula OLS	Fit Generalized Standard Least Squares		0.4025	0.0180	0.0141	122
1	RPBW Prediction Formula Lasso	Fit Generalized Lasso		0.4166	0.0178	0.0140	122
1	RPBW Prediction Formula Adaptive Lasso	Fit Generalized Adaptive Lasso		0.4564	0.0172	0.0134	122
1	RPBW Prediction Formula Elastic Net	Fit Generalized Elastic Net		0.4165	0.0178	0.0140	122
1	RPBW Prediction Formula Adaptive Lasso t5	Fit Generalized Adaptive Lasso		0.4604	0.0171	0.0135	122
1	RPBW Prediction Formula Adaptive Lasso Cauchy	Fit Generalized Adaptive Lasso		0.4604	0.0171	0.0135	122
2	RPBW Prediction Formula OLS	Fit Generalized Standard Least Squares		0.6103	0.0224	0.0173	123
2	RPBW Prediction Formula Lasso	Fit Generalized Lasso		0.6181	0.0222	0.0169	123
2	RPBW Prediction Formula Adaptive Lasso	Fit Generalized Adaptive Lasso		0.6301	0.0219	0.0166	123
2	RPBW Prediction Formula Elastic Net	Fit Generalized Elastic Net		0.6181	0.0222	0.0169	123
2	RPBW Prediction Formula Adaptive Lasso t5	Fit Generalized Adaptive Lasso		0.6426	0.0215	0.0164	123
2	RPBW Prediction Formula Adaptive Lasso Cauchy	Fit Generalized Adaptive Lasso		0.6418	0.0215	0.0166	123

Here, I determined that the Adaptive Lasso method with the Student's T (t5) distribution was the best fit by looking at its R Square value for the test data. In the cross-validation column, titled "Holdback", the test data was encoded by a 2, so the approximately 64% explanative capacity of the Adaptive Lasso t5 method indicates that this method is the most appropriate of the 6 methods used in this assignment.

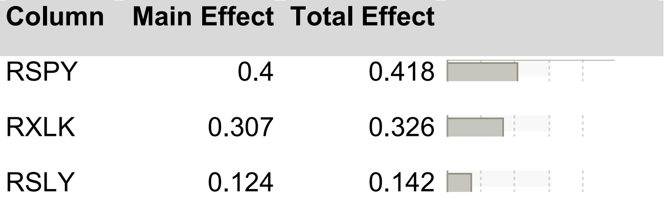
After determining that the Adaptive Lasso with Student's T distribution was the selected model, I further analyzed the model's performance by examining the parameter estimates of each predictor variable, which are shown in the following table:

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	0.001126	0.0010005	1.2666679	0.2604	-0.000835	0.0030868
RTIP	0.0975448	0.227494	0.183852	0.6681	-0.348335	0.5434249
RLQD	0	0	0	1.0000	0	0
RHYG	-0.012382	0.2007211	0.0038054	0.9508	-0.405788	0.3810241
RSPY	-1.896856	0.3909565	23.540282	<.0001*	-2.663117	-1.130595
RXLK	1.2926804	0.222175	33.852589	<.0001*	0.8572254	1.7281354
RSLY	1.0989329	0.1140335	92.870423	<.0001*	0.8754314	1.3224345
RUSO	-0.004479	0.0296013	0.0228943	0.8797	-0.062496	0.0535386
RGLD	0	0	0	1.0000	0	0
ROVX	-0.014509	0.0094635	2.3504831	0.1252	-0.033057	0.0040394
RFXE	-0.001829	0.2497378	5.3615e-5	0.9942	-0.491306	0.4876484
REMB	0	0	0	1.0000	0	0
RTLH	0	0	0	1.0000	0	0
RVIX	0	0	0	1.0000	0	0
RVEU	-0.161417	0.3876029	0.1734302	0.6771	-0.921105	0.5982706
RVSS	0.317222	0.3096442	1.0495438	0.3056	-0.28967	0.9241135
RVWO	0.6955936	0.2046562	11.552108	0.0007*	0.2944748	1.0967124
RVNQL	-0.201545	0.2031417	0.9843431	0.3211	-0.599696	0.1966053

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
RVNQ	0	0	0	1.0000	0	0
RBWX	0	0	0	1.0000	0	0
RIBND	0	0	0	1.0000	0	0
LRPBW	0	0	0	1.0000	0	0
LRTIP	0	0	0	1.0000	0	0
LRLQD	0.3106175	0.2289207	1.8411186	0.1748	-0.138059	0.7592937
LRHYG	-0.214836	0.2184633	0.9670643	0.3254	-0.643016	0.2133446
LRSPY	0.3176472	0.1986729	2.556305	0.1099	-0.071745	0.7070389
LRXLK	0	0	0	1.0000	0	0
LRSLY	0.0191947	0.1023956	0.0351398	0.8513	-0.181497	0.2198864
LRUSO	0	0	0	1.0000	0	0
LRGLD	0	0	0	1.0000	0	0
LROVX	0.0042796	0.0100315	0.1820023	0.6697	-0.015382	0.0239411
LRFXE	0	0	0	1.0000	0	0
LREMB	-0.054571	0.2105677	0.0671645	0.7955	-0.467276	0.3581342
LRTLH	0	0	0	1.0000	0	0
LRVIX	0.022017	0.0185773	1.4045995	0.2360	-0.014394	0.0584278
LRVEU	0	0	0	1.0000	0	0
LRVSS	-0.251876	0.331178	0.5784277	0.4469	-0.900972	0.3972214
LRVWO	0.1891533	0.153465	1.5191798	0.2177	-0.111633	0.4899392
LRVNQL	0.1380881	0.1659542	0.6923667	0.4054	-0.187176	0.4633524
LRVNQ	0	0	0	1.0000	0	0
LRBWX	0.4657478	0.2036636	5.2296772	0.0222*	0.0665745	0.864921
LRIBND	-0.520547	0.2549466	4.1689077	0.0412*	-1.020234	-0.020861

As shown by the rows that only contained zeros, the model eliminated many variables that it deemed to be irrelevant, leaving 24 predictor variables. A priori this may still seem like a lot of variables left in the model, but the most important thing to look for at this step is the signage of the estimates and their absolute distance from zero, which helps us gauge which predictor variables have the greatest impact on the response variable.

Below, I’ve included an extract from the table that shows each variable’s importance for the top three, as measured by their independent uniform inputs, which have a combined importance of roughly 89% in terms of their relative contribution to the model’s predictions.



Column	Main Effect	Total Effect
RSPY	0.4	0.418
RXLK	0.307	0.326
RSLY	0.124	0.142

The three most impactful predictor variables that this model identified were RSPY, RXLK, RSLY. Now, looking back at their signage in the parameter estimates, we can see whether growth in these stocks has a positive or negative effect on the prices of clean energy stocks (RPBW in the dataset). Indeed, RSPY had a large, negative estimate and is the most reliable predictor for clean energy stock prices.

SPY (S&P 500 index ETF), is generally regarded as a benchmark of the overall stock market. In other words, fluctuations in this stock can reveal the health of the economy in general. With this in mind, the fact that this stock would have such a large, negative relationship with the PBW stock reveals that clean energy might not be widely adopted in many of the largest companies that drive the price of this stock. By this logic, I can assume that investors relatively pour more money into companies that do not fully implement clean energy sources, if at all, since the opposite notion would have been supported by a positive relationship between this predictor variable, RSPY, and the response variable, RPBW, in the model.

XLK (Technology Select SPDR Fund) being the second most impactful stock was of little surprise given the assumption that technological innovations in terms of energy sources have been geared towards cleaner, renewable energy sources in recent years. This fund is a representation of the technology sector within the S&P 500 index (i.e., the SPY stock that had a negative impact on the response variable in the model). Global policies, like the 2015 Paris Agreement, incentivize companies to adopt cleaner energy sources and shows that public sentiment on topics like global warming and other environmental issues would drive some investors to direct funds into stocks like PBW and XLK that reflect growth in companies that advance this cause.