Alec Engel
May 23, 2021

Introduction

The question at hand here is to determine what if any financial market variables play a role in the prediction of our clean energy exchange traded fund (ETF). We want to run several cross-validation models to see what variables can be strongly correlated with our response variable of continually compounded returns on clean energy ETF stocks. This cross-validation method will allow us to split our dataset into training, validation, and test portions. The split allows for better predictions of future outcomes with reduction in potential overfitting and by removing unnecessary model noise amongst highly correlated independent variables.

Predictor variables consist of both continually compounded ETF weekly returns and one week lag prices for different currencies, bonds, crypto currencies and stock markets. Each variable at hand is converted into percentage returns and then turned into natural log prices and one week lag prices. The most important variables in this data set are the response variable of return on clean energy ETF stocks (RPBW); along with predictor variables of S&P 600 index (RSLY), return on technology sector index (RXLK), return on international stock index (RACWX) and continually compounded returns of US oil prices (RUSO). As we will elaborate later, the US smaller markets of S&P 600 companies will have highly unique information as it pertains to the correlation with clean energy price changes.

In order to narrow down our predictor variables, we ran several statistical method models and judged our outcomes. Statistical methods include cross-validation tactics of Ordinary Least Squares (OLS), Elastic Net, and a few Lasso regressions with varying distributions. The latter two methods involve assessing penalties to our variables in order to place more emphasis on important variables while less emphasis or in some cases eliminating the unimportant predictor variables.

Analysis and Model Comparison

Ordinary least squares involved the inclusion of all variables which often leads to overfitting or the misuse of highly correlated predictor variables. While using all variables, there are often noisy and unnecessary data variables that are accounted for. For a clearer picture, we then move on to penalized regressions of Lasso and Elastic Net models. Lasso will eliminate

uninformative or highly correlated variables while Elastic Net may still include such variables. We will also run adaptive versions of each method. Adaptive models will take into account OLS estimates of all variable importance and implement smaller penalties against our pre-determined most important variables. As previously mentioned, we will also be taking distribution techniques of Lasso Couchy and Lasso t(5) into account while all other models will be normally distributed. These additional model variations will allow us to account for the possibility of extreme outliers within our dataset.

The dataset includes a cross-validation column which I've renamed "Holdback" as it's time series data. This splits our data into training (60%), validation (20%) and testing (20%) sets to allow our model to gain predictive attributes and score our R squared values for how we believe future data may be scored. After cross-validation, we created predictor formulas with each of the above seven methods described. Each method was then added as a dataset column so that we can easily perform model comparisons.

Model comparison results can be seen below. First with the zero "Holdback" notation, we see all seven models having moderately high performance. All seven are above 0.55 R squared with very small route average standard error as well as low absolute average error. These however are our training sub sets of data and we want to focus more towards our validation and testing metrics. Moving to one and two "Holdback" notations, we see the same patterns in regards to relatively strong R squared values across the board with normal distribution models of Adaptive Lasso and Adaptive Elastic Net appearing to be our two strongest models. Although I will dig into each model to see what our strongest prediction profiler variables are, I want to choose the strongest model of Adaptive Elastic Net with a testing set R squared value of 0.62602 as my predominant model. Adaptive Elastic Net explains nearly 63% on our unbiased test data.

Model Comparison

Predictors

Measures of Fit for RPBW

Holdback	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
0	Pred Formula RPBW OLS	Fit Least Squares		0.5860	0.0092	0.0068	993
0	RPBW Prediction Formula ALasso	Fit Generalized Adaptive Lasso		0.5553	0.0095	0.0071	994
0	RPBW Prediction Formula Lasso	Fit Generalized Lasso		0.5758	0.0093	0.0069	993
0	RPBW Prediction Formula ALasso Couchy	Fit Generalized Adaptive Lasso		0.5534	0.0095	0.0070	993
0	RPBW Prediction Formula ALasso t(5)	Fit Generalized Adaptive Lasso		0.5504	0.0096	0.0071	994
0	RPBW Prediction Formula Elastic Net	Fit Generalized Elastic Net		0.5758	0.0093	0.0069	993
0	RPBW Prediction Formula AElastic Net	Fit Generalized Adaptive Elastic Net		0.5552	0.0095	0.0071	994
1	Pred Formula RPBW OLS	Fit Least Squares		0.6668	0.0084	0.0065	332
1	RPBW Prediction Formula ALasso	Fit Generalized Adaptive Lasso		0.6749	0.0083	0.0064	332
1	RPBW Prediction Formula Lasso	Fit Generalized Lasso		0.6701	0.0084	0.0065	332
1	RPBW Prediction Formula ALasso Couchy	Fit Generalized Adaptive Lasso		0.6688	0.0084	0.0065	332
1	RPBW Prediction Formula ALasso t(5)	Fit Generalized Adaptive Lasso		0.6728	0.0083	0.0064	332
1	RPBW Prediction Formula Elastic Net	Fit Generalized Elastic Net		0.6701	0.0084	0.0065	332
1	RPBW Prediction Formula AElastic Net	Fit Generalized Adaptive Elastic Net		0.6750	0.0083	0.0064	332
2	Pred Formula RPBW OLS	Fit Least Squares		0.6203	0.0219	0.0169	332
2	RPBW Prediction Formula ALasso	Fit Generalized Adaptive Lasso		0.6260	0.0218	0.0168	332
2	RPBW Prediction Formula Lasso	Fit Generalized Lasso		0.6121	0.0222	0.0172	332
2	RPBW Prediction Formula ALasso Couchy	Fit Generalized Adaptive Lasso		0.6173	0.0220	0.0170	332
2	RPBW Prediction Formula ALasso t(5)	Fit Generalized Adaptive Lasso		0.6229	0.0218	0.0169	332
2	RPBW Prediction Formula Elastic Net	Fit Generalized Elastic Net		0.6121	0.0222	0.0172	332
2	RPBW Prediction Formula AElastic Net	Fit Generalized Adaptive Elastic Net		0.6260	0.0218	0.0168	332

Normal Adaptive Lasso with Validation Colum

Model Summary

Response	RPBW
Distribution	Normal
Estimation Method	Adaptive Lasso
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

Measure	Training	Validation	Test
Number of rows	993	332	332
Sum of Frequencies	993	332	332
-LogLikelihood	-3214.711	-1114.11	-370.4485
Number of Parameters	6	6	6
BIC	-6388.018	-2193.389	-706.0663
AICc	-6417.337	-2215.961	-728.6386
RSquare	0.5556831	0.674917	0.625983
RASE	0.0095014	0.0083008	0.0217563
Lambda Penalty	0.0024794	.	.

Normal Adaptive Elastic Net with Validation C

Model Summary

Response	RPBW
Distribution	Normal
Estimation Method	Adaptive Elastic Net
Validation Method	Validation Column
Mean Model Link	Identity
Scale Model Link	Identity

Measure	Training	Validation	Test
Number of rows	993	332	332
Sum of Frequencies	993	332	332
-LogLikelihood	-3214.586	-1114.115	-370.7147
Number of Parameters	6	6	6
BIC	-6387.768	-2193.399	-706.5985
AICc	-6417.087	-2215.971	-729.1709
RSquare	0.5555712	0.674955	0.6260212
RASE	0.0095026	0.0083003	0.0217552
Lambda Penalty	0.0025044	.	.

Interpretation

Per below, we can see that four predictor variables were used to maximize our overall R squared value. Uninformative variable selection was used with our Adaptive Elastic Net model to remove unimportant variables which are denoted with zeroes. The parameter estimates used in our model are RSLY, RXLK, RACWX and RUSO. We can also see that all four parameter estimates used are statistically significant with p-values less than 5%.

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-0.000666	0.0003067	4.7120786	0.0300*	-0.001267	-6.463e-5
RBTC	0	0	0	1.0000	0	0
RUSO	0.047056	0.0193155	5.9349329	0.0148*	0.0091982	0.0849137
RGLD	0	0	0	1.0000	0	0
RSPY	0	0	0	1.0000	0	0
RXLK	0.2180654	0.0564691	14.912542	0.0001*	0.107388	0.3287428
RACWX	0.1921981	0.0657966	8.5328046	0.0035*	0.0632392	0.321157
RTLH	0	0	0	1.0000	0	0
RHYG	0	0	0	1.0000	0	0
RFXE	0	0	0	1.0000	0	0
RFXY	0	0	0	1.0000	0	0
RFXB	0	0	0	1.0000	0	0
RVIX	0	0	0	1.0000	0	0
RFXF	0	0	0	1.0000	0	0
RLQD	0	0	0	1.0000	0	0
RSHY	0	0	0	1.0000	0	0
RSLY	0.6816127	0.0656419	107.82354	<.0001*	0.552957	0.8102684

RSLY and RXLK will be our most important parameter estimates as they show both high estimate numbers as well as low standard errors. As clean energy ETF prices increase by 1%, we have a positive correlation with return on S&P 600 index and that index increases by 0.68%. We can also see a positive correlation with return on technology sector index as it increases by 0.22% while clean energy ETF prices increase by 1%. RACWX and RUSO also have strong enough parameter estimates to be used in our model. Both of these variables do have positive correlations meaning increases of US oil prices and international stock index as clean energy ETF prices increase.

In comparison between US oil and clean energy predictor variables we see several similarities. US oil prices is in fact one of the predictors of clean energy with a positive relationship and I would assume if given this clean energy as a predictor variable in the covid dataset, we would see this relationship vice versa. Interestingly enough, we also see the variable of RXLK as relatively strong positive correlation predictors of both US oil and clean energy ETF prices. We also see a correlation with the larger market of S&P 500 having negative correlation with US oil prices while the smaller market of S&P 600 shows a positive correlation with clean energy ETF prices.

RSLY is our strongest profiler and variable. When looking at our four variable profilers ranked by importance, we notice the three previously discussed variables of RXLK, RACWX and RUSO having an effect while all other variables have been removed and noted as

insignificant/unimportant to our model. Some consideration can be given to RXLK with just over 12% variation regarding its slight positive relationship alongside the clean energy market. But again, RSLY is our strongest predictor variable as its positive relationship explains over 78% of variations in the clean energy stock ETF prices.