Alec Engel
May 14, 2021

Introduction

The question at hand here is to determine what if any financial market variables play a role in the prediction of oil price changes in the United States. We want to run several cross-validation models to see what variables can be strongly correlated with our response variable of continually compounded price returns on US oil prices. This cross-validation method will allow us to split both the pre and during covid data sets into training, validation, and test portions. This split allows for better predictions of future outcomes by a reduction in potential overfitting and 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 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 US oil prices (RUSO); along with predictor variables consisting of return on energy sector index (RXLE), return on S&P 500 index (RSPY), one week lag of return on S&P 500 index (LSPY), continually compounded returns of US treasury bonds (REIE) and inflation measures on treasuries (RTIP). As we will elaborate on later, both energy sector index and the US primary market of S&P 500 companies will have highly unique information as it pertains to the correlation with the US oil market.

In order to narrow down our predictor variables, we ran a few statistical methods and judged our outcomes. Statistical methods include cross-validation tactics of ordinary least squares, stepwise forward, and stepwise backward.

Analysis and Model Comparison

Ordinary least squares involves the inclusion of all variables which can often lead 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. The training set of data may produce a high-quality R squared value but often times this doesn't translate to having strong R squared values on the test set which is prediction based and most important. We tend

to lean more heavily on stepwise forward which adds variables until our model can no longer increase its R squared value or stepwise backward which starts with all variables and removes them until R squared value can no longer be improved. As mentioned above, we will be implementing a cross-validation technique as well. We manually create a new column named "Holdback" and split each data set 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 three methods described on both data sets. Each method was then added as a data set column so that we could easily perform model comparisons once all were created.

Pre-covid model comparison results can be seen below. First with the zero "Holdback" notation, we see all three models having relatively high performance. All three are above .655 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 on to one and two "Holdback" notations, we see the same patterns in regards to relatively strong R squared values across the board with forward stepwise being our strongest model followed by backwards stepwise and the not so far behind least squares model. Although I will dig into each model to see what our strongest prediction profiler variables are, I want to choose the strongest model of forward stepwise with a testing set R squared value of .477 as my predominant model.

Model Comparison

Predictors

Measures of Fit for RUSO

Holdback	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
0	Pred Formula RUSO OLS	Fit Least Squares		0.6784	0.0134	0.0107	797
0	Pred Formula RUSO FWD Stepwise	Fit Least Squares		0.6552	0.0139	0.0111	797
0	Pred Formula RUSO BWD Stepwise	Fit Least Squares		0.6775	0.0135	0.0107	797
1	Pred Formula RUSO OLS	Fit Least Squares		0.4246	0.0129	0.0101	266
1	Pred Formula RUSO FWD Stepwise	Fit Least Squares		0.4752	0.0123	0.0097	266
1	Pred Formula RUSO BWD Stepwise	Fit Least Squares		0.4371	0.0128	0.0100	266
2	Pred Formula RUSO OLS	Fit Least Squares		0.4610	0.0151	0.0116	267
2	Pred Formula RUSO FWD Stepwise	Fit Least Squares		0.4777	0.0148	0.0114	267
2	Pred Formula RUSO BWD Stepwise	Fit Least Squares		0.4628	0.0151	0.0116	267

Covid model comparison results can be seen below. First with the zero "Holdback" notation, we see relatively high performance on ordinary least squares (.706) and backward stepwise (.706), but a lower performance with forward stepwise (.313). Again, all three have 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 give more weight towards our validation and testing metrics. Moving on to one and two

"Holdback" notations, we see new patterns emerge. There is still a moderately strong R squared value amongst forward stepwise, but now there are extremely weak validation and testing R squared values amongst our ordinary least squares and backward stepwise models. The clear choice here on the covid dataset with the only strong test R squared value is the model of forward stepwise with a testing set R squared value of .331.

Model Comparison

Predictors

Measures of Fit for RUSO

Holdback	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
0	Pred Formula RUSO OLS	Fit Least Squares		0.7064	0.0280	0.0201	193
0	Pred Formula RUSO FWD Stepwise	Fit Least Squares		0.3134	0.0427	0.0259	194
0	Pred Formula RUSO BWD Stepwise	Fit Least Squares		0.7064	0.0280	0.0200	193
1	Pred Formula RUSO OLS	Fit Least Squares		-1.244	0.0295	0.0242	65
1	Pred Formula RUSO FWD Stepwise	Fit Least Squares		0.1112	0.0185	0.0148	65
1	Pred Formula RUSO BWD Stepwise	Fit Least Squares		-1.234	0.0294	0.0242	65
2	Pred Formula RUSO OLS	Fit Least Squares		-0.908	0.0311	0.0240	67
2	Pred Formula RUSO FWD Stepwise	Fit Least Squares		0.3316	0.0184	0.0150	67
2	Pred Formula RUSO BWD Stepwise	Fit Least Squares		-0.906	0.0311	0.0240	67

Interpretation

For the pre-covid dataset, I've chosen the forward stepwise model for interpretation. Per below, we can see that ten predictor variables were used to maximize our overall R squared value. The strongest parameter estimates belong to RSPY, RXLE, RIEI and RTIP. We can also see that all parameter estimates are statistically significant with p-values less than 5%. I'd also like to point out here the lag variable of S&P 500 that was included which would be our strongest lag predictor and only usable lag variable for predictions of future US oil prices.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.000259	0.0005	-0.52	0.6042
RSPY	-1.593254	0.195643	-8.14	<.0001*
RXLK	0.329419	0.138198	2.38	0.0174*
RXLE	1.2603603	0.059376	21.23	<.0001*
RHYG	0.4003996	0.187095	2.14	0.0327*
REMB	0.3701584	0.171321	2.16	0.0310*
RTLT	-0.428153	0.128952	-3.32	0.0009*
RIEI	-1.936714	0.545528	-3.55	0.0004*
RTIP	1.3917158	0.322543	4.31	<.0001*
RFXC	0.9965531	0.116036	8.59	<.0001*
LRSLY	-0.139471	0.049734	-2.80	0.0052*

RSPY and RXLE will be our most important parameter estimates as they show both high estimate numbers as well as low standard errors. As oil prices increase by 1%, we have a negative correlation with return on S&P 500 index and that index drops by 1.59%. On the other hand, a positive correlation with return on energy sector index exists and increases by 1.26% while oil prices increase by 1%.

While RIEI and RTIP have strong parameter estimates, their standard errors are high which has a reduction in our prediction ability and makes these predictor variables have a much riskier range of outcomes. RIEI with a strong negative correlation means reduction of US treasury bond investments as US oil prices increase and RTIP shows a positive correlation of inflation measures on treasuries increasing alongside US oil price increases.

For the covid dataset, forward stepwise was by far the best model for interpretation. Per below, we can see that only one predictor variable was chosen which maximized our overall R squared value. RXLE shows a very high estimate number along with low standard error and it's statistically significant with a p-value far below 5%. With this positive correlation estimate, we see that as oil prices increase by 1%, energy sector index increases by .71%.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.004065	0.003093	-1.31	0.1904
RXLE	0.7098763	0.075827	9.36	<.0001*

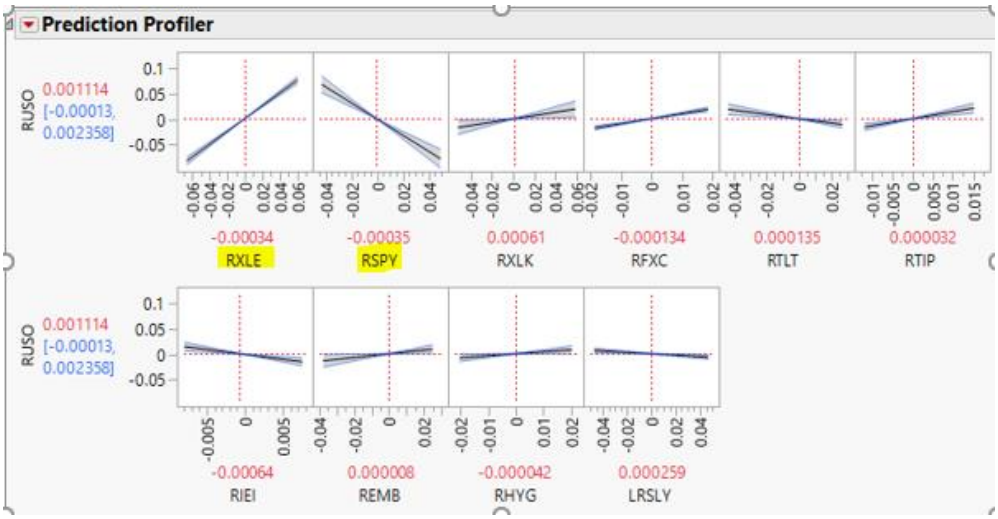
In comparison between pre and during the covid pandemic, we can see one similarity amongst many differences in our US oil predictor variables. RXLE is our strongest profiler and variable. Although the estimate declined by nearly 50% during covid times, it still exists as a relatively strong positive correlation predictor of US oil prices. All other nine variables however no longer provided strong estimates from pre-covid to during-covid era. So, although those nine were significant the previous five years, we most likely can no longer depend on their estimates until we are out of the covid World as we know it today.

When looking at our ten pre-covid variable profilers ranked by importance, we notice the two previously discussed variables of RXLE and RSPY having the most effect while our other eight variables had minimal effect. Again, RXLE is our strongest predictor variable as its positive relationship explains nearly 50% of variations in the US oil market. Some consideration should also be given to RSPY with nearly a 40% variation regarding its negative relationship alongside the US oil market.

Variable Importance: Independent Uniform Inputs

Summary Report

Column	Main Effect	Total Effect
RXLE	0.464	0.478
RSPY	0.382	0.397
RXLK	0.016	0.027
RFXC	0.018	0.026
RTLT	0.012	0.017
RTIP	0.01	0.017
RIEI	0.01	0.016
REMB	0.006	0.01
RHYG	0.002	0.005
LRSLY	0.002	0.004



When looking at our one covid variable profiler of RXLE, we notice again the positive relationship with US Oil prices. It has the highest/only effect being 100% and proves that if we want to understand events in the US oil market, the most important variable to investigate will be the energy sector index.