Goals:

This project is similar to an ongoing project that I have been involved with for years but using a different set of financial products and different software. The goal, initially, was to predict the price of an ETF based on different features created from that particular ETF (moving averages, percentage changes, etc…) but then evolved into:

Predicting the price of an ETF using the prices of other ETFs. The idea being, if I use data that is correlated as well as data that is uncorrelated to the ETF to be predicted, am I able to get better results than just having data that is correlated to a single ETF.

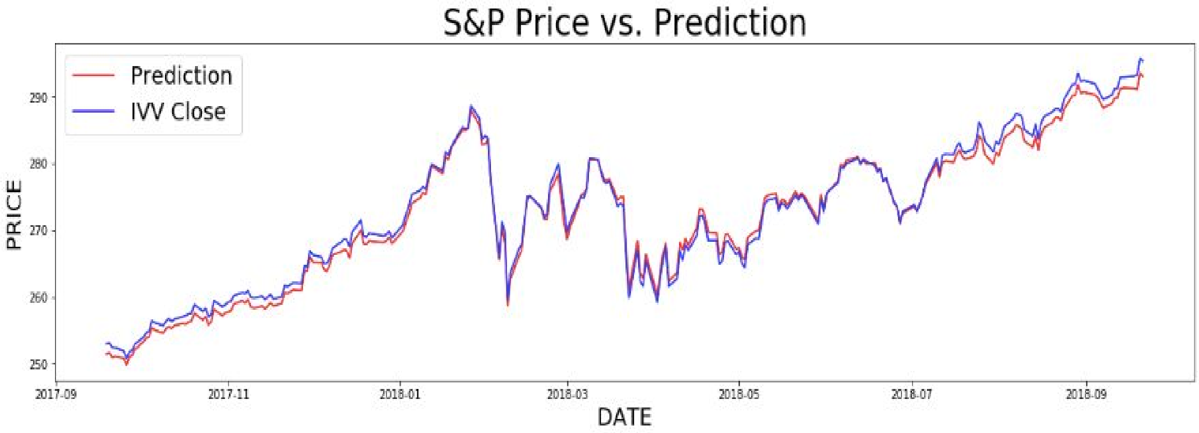
Data:

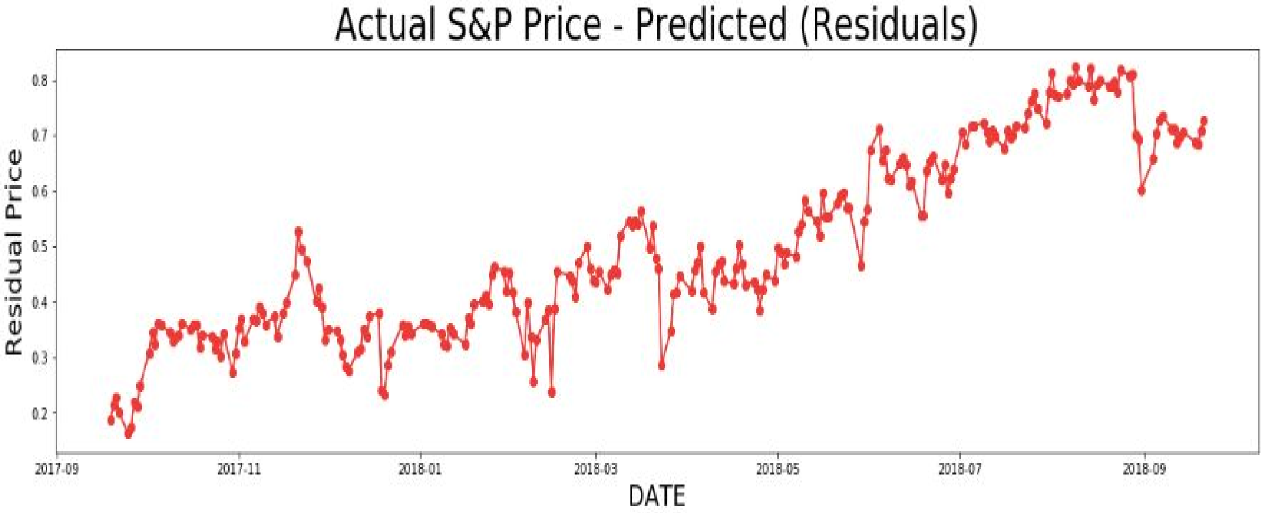
Pulling the data in from the Tradestation trading platform, I was able get reliable data in a relatively “friendly” construct. This platform is popular in the trading world as a low cost option for anyone looking to trade stocks, futures, and options (not an exhaustive list)

The EDA was a smaller part of the project being that the data was downloaded from a reputable site. Most of the cleaning and evaluation of the data revolved around merging data sets together and matching up the dates. For example, the IVV ETF didn’t start on the same date as IYR. From there I was able to do some basic manipulation to get the data ready for modeling. The features to be used are the daily closing prices of ETFs that fall into different assets classes (bond, equity, commodity, etc…) One thing to keep in mind, in addition to the features, are the different ways to train and test the data, which i’ll hit on in the “future steps”.

Modeling:

I focused mainly on different regression methods to be able to find a good r2 score while being able to utilize a balanced set of the features. In an effort to increase the r2 score, I started with a series of random forest regressions and tuned the hyperparameters using gridsearch but the scores on the testing test at one point were only slightly positive. I moved on to gradient boosting which produced negative numbers in the testing set. ADA boost worked really well when I used a linear regression base estimator. I tuned the hyperparameters using gridsearch again and the r2 score was over .95.





Findings:
Using adaboost with a linear regression as the base estimator with the closing prices of all the features resulted in a large r2 score of over .95. What this means is the model can explain a large portion of the moves in actual price. However with further adjustments to these models and bringing in some out side libraries and formulas, specifically a walk-forward analysis, which I came to as I was doing research on training and testing time series data, there should be room to tune the models further and hopefully minimize the residual number shown above.

Future Steps:
During my research on time series, there are some tools that I would like to implement such as time series split, which is an sklearn tool.
Another technique that looks interesting is the walk forward validation. The purpose here is to create a new model whenever new data makes itself available. In the case here, every day would have a new model associated with it.