Exploratory Data Analysis

The EDA portion of the project was largely concerned with adjusting the features across time in order to find the appropriate lags relative to the target. To elaborate, it is often the case that macroeconomy doesn’t immediately exert influence on financial assets. Sometimes a few months, if not longer, pass before the effects are felt. Up to this point, the features and the target are contemporaneous; that is, all datapoints are as of time t. The only manipulations so far have been that some of the features reflect monthly or annual percentage changes and that the target reflects the forward-looking (to avoid lookahead bias) three-month returns spread. For example, the dataset as of end-of-December 2018 would reflect, e.g., the percentage change in inflation from December 2017 to December 2018 and the forward-looking quarterly returns spread between the value and growth portfolios from December 2018 to March 2019.

The predictive models are classification algorithms that seek to correctly predict in which of two portfolios to invest each month. The more highly correlated the features are with the target, while maintaining low pairwise correlation among themselves so as to reduce the effects of multicollinearity, the more accurate the model’s predictions should be.

How to lag the features was based on which lag produced the highest correlation between each feature and the target over the entire time interval. Given the repetitiveness of finding correlations at various lags, a custom function was written that, when a series or list for the feature was passed through, returned the correlation between the feature and the target at no lag, a one-month lag, a two-month lag, and a three-month lag. Almost all of the features witnessed higher correlations with the target when lagged.

The fixed income feature, the “butterfly spread,” is known to more substantially lag equity portfolios than macroeconomic variables do. No custom function was written for this feature. Lags were incrementally tested as far as 12 months, and ultimately the eight-month lag was settled on.

The final dataset (that is, the one that is to be used for predictive modeling) contains the following features at the following lags:

* Activity: two months
* Inflation: two months
* Money supply: two months
* Manufacturing: zero months
* Butterfly spread: eight months
* Dividend yield spread: two months

The heatmap for the features showed that the strongest correlation in terms of magnitude was 0.53, which indicates that the features don’t witness multicollinearity.

While the predictive models are classification algorithms, EDA focused on each feature’s linear relationship with the continuous target, and the joint effects of the features on the target. After a simple linear regression was run using each of the features, the output revealed that all of the features are significant at the 5% level and that all but one are significant at the 1% level. A multivariate linear regression revealed that four of the features are significant at the 5% level, two are significant at the 1% level, and the model’s adjusted R-squared is 21.4%.

Also of interest is the behavior of the dataset during the financial crisis. The period from June 2007 to June 2009 has been subset and proxies for the crisis. Whereas the heatmap over the entire period shows no significant correlation among the features, during the crisis, the correlations increase in magnitude precipitously: of the 15 pairwise correlations, only four are less than 0.5, while six are above 0.7. Further, the multivariate regression provides additional evidence of multicollinearity: all features lose their statistical significance (only one is just slightly above 5% while the rest are largely insignificant), whereas the model’s adjusted R-squared rises tremendously to 76.9%. It can be expected that during this period, more accurate predictions should arise. And if so, this would be welcome: especially during times of financial crises, investors tend to sell risky assets and fly to the safety of bonds and cash instruments. A model that can predict with high accuracy between stock portfolios would be highly sought after.