Capstone Project 1: Milestone Report

Problem statement: Why it’s a useful question to answer and for whom

Given macroeconomic data, and given the performance of two underlying investable indices, I wish to devise a predictive model that beats the performance of merely buying and holding either of the two underlying indices.  More specifically, I wish to allocate between the S&P 600 Growth Index, an index of U.S. small cap growth stocks, and the S&P 600 Value Index, an index of U.S. small cap value stocks, based on macroeconomic variables that include inflation, interest rates, and economic activity.  Predictive modeling will be accomplished by two models: (1) logistic regression, and (2) SVM classification. Each model will use the macroeconomic and index-level variables as its features, and the target will be the returns spread between the two indices, coerced to Boolean (1 if small cap outperforms large cap, 0 otherwise).

The range of potential clients includes: (1) an asset manager who’s looking to launch an investable portfolio that seeks a rules-based allocation scheme involving macroeconomic variables (in industry, known as a “top-down” model); (2) a sell-side firm that sells research to clients that include banks, insurance companies, and asset managers; and, (3) a retail investor who in his or her own brokerage account is looking to allocate based on the broader performance of the economy in order to beat major market indices.

Devising a predictive model that outperforms simple buy-and-hold strategies would be incredibly coveted by numerous financial participants. That is, supposing one were to simply buy one of the portfolios that the predictive model must choose between and hold it indefinitely, if the predictive model bests such passive (“do-nothing”) strategy, then the model would be more attractive to investors.

The data for this project can be found from publicly available sources such as FRED (Federal Reserve Economic Data) or OECD (Organization of Economic Development).  Data is also readily available from financial data vendors such as Bloomberg, FactSet, and Reuters. The data is available for download in the form of a flat file such as XLS or CSV. The data can be retrieved from each source’s websites or through an API.

The overall objective is to predict whether the S&P 600 Growth Index (“growth”) outperformed the S&P 600 Value Index (“value”) or vice versa.  This objective is conducive to a classification problem where the target is “1” if portfolio A outperforms portfolio B and “0” otherwise. The features include macroeconomic and index-level variables that research has shown has influenced equity markets to differing extents depending on the underlying stocks’ attributes.  For example, research has shown that small cap stocks outperform large cap stocks during bouts of high inflation. Can a logistic regressor or SVM classifier accurately predict which of large cap and small cap will outperform?

Deliverables include a summary of the problem and the data; analysis of the data, such as regressing the return spread on each of the individual features; the steps taken to wrangle the data; discussions on how the models are trained, validated, and tested, and the hyperparameter tuning therein; and, performance as measured by the accuracy score that results from the confusion matrix.  Deliverables will include code, paper, and slide deck.

\*\*\*\*\*

Description of the dataset: how you obtained, cleaned, and wrangled it

It’s desired that the data be indexed monthly (though some features or the targets might witness a quarterly or annual percentage change).  Not all data series observe monthly periodicity – some are daily or weekly. So, they needed to be downsampled. Some of the data series are raw figures and need to be converted to percentage changes.  The choice of whether to convert to monthly, quarterly, or annual change rested on intuition and troubleshooting against the target.

The data are available over a range of time intervals.  The time interval that was selected for further analysis is that which contains no missing values for any features or the target.  After all of the individual series (imported into Python as dataframes) were concatenated, the relevant (or, said correctly, irrelevant) observations were dropped.  Additionally, all indices were converted to datetime and, as stated, indexed monthly, which were considerations for appropriate concatenation.

One of the features, the fixed income “butterfly spread,” did contain missing observations.  Not every day is a trading day for all financial instruments, even for those instruments that map to the same country.  Quasi-holidays like Columbus Day or Veteran’s Day witness this occurrence. So, those nulls were forward-filled. Interpolation wouldn’t have made sense, given that no price change occurs on a non-trading day.

The target and a few of the features needed to be computed based off of the raw data.  For example, the target is the upcoming quarterly returns spread between two portfolios whose constituents tend to behave differently based on macroeconomic and fundamental data.  And after engineering the features, the original features were discarded.

The computed target was continuous, yet for the eventual logistic regression and SVM classifier, this target needed to be converted to Boolean.  The Boolean target was mapped to 1 if the returns spread was positive and 0 otherwise.

Additional cleaning involved pivoting, re-indexing, and parsing dates.

The presence of outliers would be addressed in later stages of the project, during the in-depth analysis and machine learning implementation.  A possible measure that would be taken is to winsorize a standardized series at, for example, +/- 3 standard deviations, so as to remove the effect of outliers.  But most likely, outliers wouldn’t call for the deletion of the entire observation.

\*\*\*\*\*

Initial findings from exploratory 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.