Philip Demeri

Capstone Project 2: Milestone Report 1

Problem statement: Why it's a useful question to answer and details about the client

Can I select stocks using fundamental data? I seek to predict stock performance using individual stocks' fundamental data (such as book-value-to-price, EPS-to-price, etc.) using an ensemble of classifiers – decision tree, random forest, neural network, etc. I can gather monthly stock-level data over 20 to 30 years and build classifiers that predict performance using unseen data. Each month (more specifically, at the end of the month), the regressors will be the fundamental data, and the target will be the stock's next-month return relative to the median stock's next-month return. I look to train my model on the first 50 to 60% of the data, validate on the next 20 to 30%, and then test on the final 10 to 20%. Considerations from the top-level (i.e., not at the unique model level) include (1) the nature of ensemble voting – should it be majority or weighted; (2) bootstrapping; and, (3) bagging. Performance will based on precision, recall, and accuracy. In this project, within the confusion matrix, "false positive" refers to buying losing stocks, and "false negative" refers to not buying winning stocks. More specifically, "false positive" refers to stocks that the model predicts will perform in the upper half out of all stocks in the next coming month, when in fact the stocks underperform; "false negative" refers to stocks that the model predicts will underperform the median, yet the stocks wind up outperforming.

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 stock-level fundamental variables (in industry, known as a "bottom-up" 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 which stocks that he or she thinks that will relatively outperform others over the next coming time horizon.

Description of the dataset, how you obtained, …

The stocks are drawn from the S&P 1500, an equity index that at any given time contains approximately 1,500 stocks ("approximately" due to the presence of splits, spinoffs, etc.). Within the S&P 1500, the financials sector has been chosen (i.e., banks, insurance companies, etc.). Monthly snapshots (more specifically, last trading day of each month) have been gathered from 1994 to 2015, and at each month, anywhere from 100 to 300 financials stocks are members of the S&P 1500.

The stock-level data are acquired from FactSet, a financial data vendor, via FactSet's Excel add-in. The stock-level data could also have been acquired from other vendors such as Bloomberg or Reuters. The data could also have been downloaded from vendors' UI's in the form of a flat file such as an XLS or CSV file.

… cleaned, and wrangled

The uncleaned data reflected all financials stocks from the S&P 1500. Financials stocks receive a classification of "40" by GICS (Global Industry Classification Standards), as opposed to other classifications for different sectors. Real estate, for example, receives a classification of "60." Note that the sector classifications are nominal; that is, that 60 is greater than 40 does not suggest that real estate supersedes financials in any manner.

The uncleaned data reflected monthly end-of-month constituents beginning on December 31, 1994, and ending on July 31, 2015. In addition, fundamental data were also provided for each stock: sales-to-price, book-value-to-price, last quarter's earnings-per-share-to-price, return on assets, current earnings-per-share-to-price, "price momentum," and the next month's total return, the latter being the target. Further description of each of these datapoints is provided below.

The first wrangling step was to only keep the necessary columns: effective date, company name, CUSIP, the features, and the target. Company name and CUSIP are not needed for any further steps; they are retained in order to uniquely reference a particular observation. CUSIP is an (alphanumeric, nine-character) identifier that maps to a particular security. For example, when a company undergoes a name change, the corporate structure might remain the same. CUSIP's provide an effective means of tracking and cataloguing securities.

When data types were checked, the next month's total return needed to be coerced to float and errors coerced to NaN.

A major consideration regarding stock-level fundamental data gathering is that not always is a given datapoint available at a given point in time. Financial vendors cannot always timely retrieve necessary data on account of, e.g., inconsistencies between corporate filings regarding the same data point, re-statements of a given data point, etc. So, for all of the features and also for the target, the number of nulls needed to be checked. The initial dataframe witnessed 54,435 observations, 50,233 of which were free of NaN's. It was decided not to impute any of the nulls and instead drop all rows containing at least one missing datapoint.

Only one feature needed to be engineered: EPS (earnings per share) momentum, which results when current EPS-to-price is divided by the last quarter's EPS-to-price. Given that EPS might equal zero, dividing by zero gives rise to inf (or -inf, if the numerator is negative). The decision was to coerce to 9999 or -9999. Practically speaking, given that EPS momentum typically never exceeds 1, and given that the continuous datapoints would eventually be discretized based on quantile, such extreme values of absolute 9999 would almost surely lie in the highest or (if -9999) lowest quantile.

At this point, the previous quarter's EPS-to-price would not be needed for any further analyses or machine learning, and hence the column was dropped.

Consistent with the methodology that this project aims to follow, the target and all of the continuous features were discretized. The target, the next month's total return for each stock, was coerced to 1 if, for each date, the stock's return lay above the median return on the same date, and 0 otherwise. So, 1 maps to stocks that "outperform" whereas 0 maps to those that "underperform."

For each feature, grouped by date, percentiles were demarcated at 20% intervals: 1 if less than 20%, 2 if less than 40%, …, and 5 otherwise.

So, the dataframe that would then be passed onto the EDA phase contains the following columns: date, company name, CUSIP, continuous features, continuous target, discrete features, and discrete target.

The features for this project are:

- Sales-to-price: a company's last twelve month's (TTM – trailing twelve months) sales divided by the stock price at time t
- Book-value-to-price: a company's TTM book value divided by the stock price at time t
- ROA: a company's TTM net income divided by TTM average total assets (average total assets is generally computed by averaging the value of a company's assets at time t and the value of a company's assets twelve months ago)
- Earnings-per-share-to-price: a company's TTM EPS (EPS is computed by dividing TTM net income by the number of common shares outstanding at time t) divided by the stock price at time t
- Price momentum: a stock's total return over the past month (total return includes not only share price but also any dividends per share that were paid)
- EPS momentum: a company's EPS at time t divided by EPS one quarter ago

As stated earlier, while all features inherently are continuous, consistent with the methodology that this project follows, the features are discretized by percentile based upon each date and are partitioned into five quantiles.

The target for this project is: a stock's total return over the next month.

Initial findings from exploratory analysis

The analysis focused on the relationships between and among the features as well as between each feature and the target.

Correlation matrices and heatmaps were constructed for the original features and then for the features when grouped by date. The original features witnessed close to zero correlation amongst each other, whereas the grouped (by date) features witnessed weakly negative or negative correlations. Note that the matrices and heatmap were applied to the continuous features.

Simple linear regressions were run for each feature versus the target, again in the continuous case. Three of the six features witnessed statistically significant P-values, yet none of the features attained a R^2 greater than 1.

Line plots of each of the features over time witnessed some disturbances around 2000-01 and then again around 2007-08: the former relates to the dot-com bubble, where many web and tech companies witnessed financial distress, whereas the latter relates to the financial crisis, a period of widespread systematic risk that permeated all aspects of the financial system and the broader economy. To illustrate, for sales-price, for the majority of the period under study, this variable does not exceed 200, yet amid the financial crisis, this variable precipitously rises to above 800. The rise in this variable is intuitive in that it reflects a greater and meteoric decline in stock price (the denominator) relative to periodic company sales (the numerator).

Histograms for four of the six feature witnessed right skew, as did the target. This finding is consistent with many other phenomena in finance, where purely normally

distributed data is rarely seen and instead, right-skewed (few positive outliers) variables dominate.