Philip Demeri

Capstone Project 2: Project Proposal

What is the problem that you want to solve?

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.

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done 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 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.

What data are you using? How will you acquire the data?

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.

Briefly outline how you’ll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

The objective is to see if we can predict which stocks will outperform and which will underperform over a given future time horizon using stock-level fundamental data. So this is a supervised learning problem of the form (x1, x2, …, xn, **y**), where the x’s represent the features at time t and y represents the target at some future time t + i. The features are fundamental stock data that draw from widely used metrics including book value-to-price, earnings-per-share-to-price, earnings per share momentum, price momentum, and return on assets. Consistent with many studies, the features will be discretized from 1 to 5 (1 being worst-performing and 5 being best-performing); the target will also be discretized: 1 if the future return is greater than the median future return at time t, 0 otherwise.

Each time t will witness over 100 observations. So, this project is conducive to boosting and bagging. Further, as different models are trained and tested, predictions can be aggregated into ensemble voting, where different voting schemes can be trained and tested.

What are your deliverables? Typically, this includes code, a paper, or a slide deck.

Deliverables include a summary of the problem and the data; analysis of the data such as behavior of each feature over time and regressing stock-level returns on each feature; 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 scores that results from the confusion matrix.  Deliverables will include code, paper, and slide deck.