Predicting Market Returns

Matt Kwong and Winson Zeng

Introduction to Machine Learning - Spring 2016

1 Introduction

Predicting stock market and particular single name stock return had long been an attractive field to both scholars and practitioners. Among thousands of different approaches that were studied, factor modeling can be seen as the root of them all, with capital pricing model (CAPM) in particular being the fundamental root. CAPM can be summarized in the following form:

$$r_s = \beta r_m + r_e \tag{1}$$

In the equation, r_s is the expected return of a stock, β is the correlation between the stock and the general market, typically the S&P 500 stock index, and r_e is the excess or idiosyncratic return of the stock. The CAPM can be seen as a simple one factor return prediction model, where the factors are the expected return of the general market and the expected return by taking on the idiosyncratic risk of a stock. This model implies that the return of a particular stock can separated into two parts: non-diversifiable risk (market risk) and the diversifiable risk (stock specific risk). Later, more factors were introduced by Fama and French (1992, 1993). They argued that the return of the stock can be explained in three different factors: the market return (traditional CAPM defined), the size factor (large vs. small market capitalization), and the value factor (high vs. low market price to book value). The formula can be therefore written in the following form:

$$r_s = \beta_m r_m + \beta_{lvs} r_{lvs} + \beta_{hvl} r_{hvl} + r_e \tag{2}$$

where r_{lvs} and β_{lvs} are return and factor weight relating market capitalization sizes, r_{hvl} and β_{hvl} are return and factor weight relating to value. This way, the Fama-French model introduced two additional factors, and numerous academic papers had proven the explanatory power of the two additional factors.

In the four graphs in Figure 1, the upper left graph represent the Fama-French factors excess returns, which are all meaningfully above zero for a sustainable amount of time.

As computational powers picked up over the years, people started to pay more attention to factors that were not traditionally considered as seen in Figure 2.

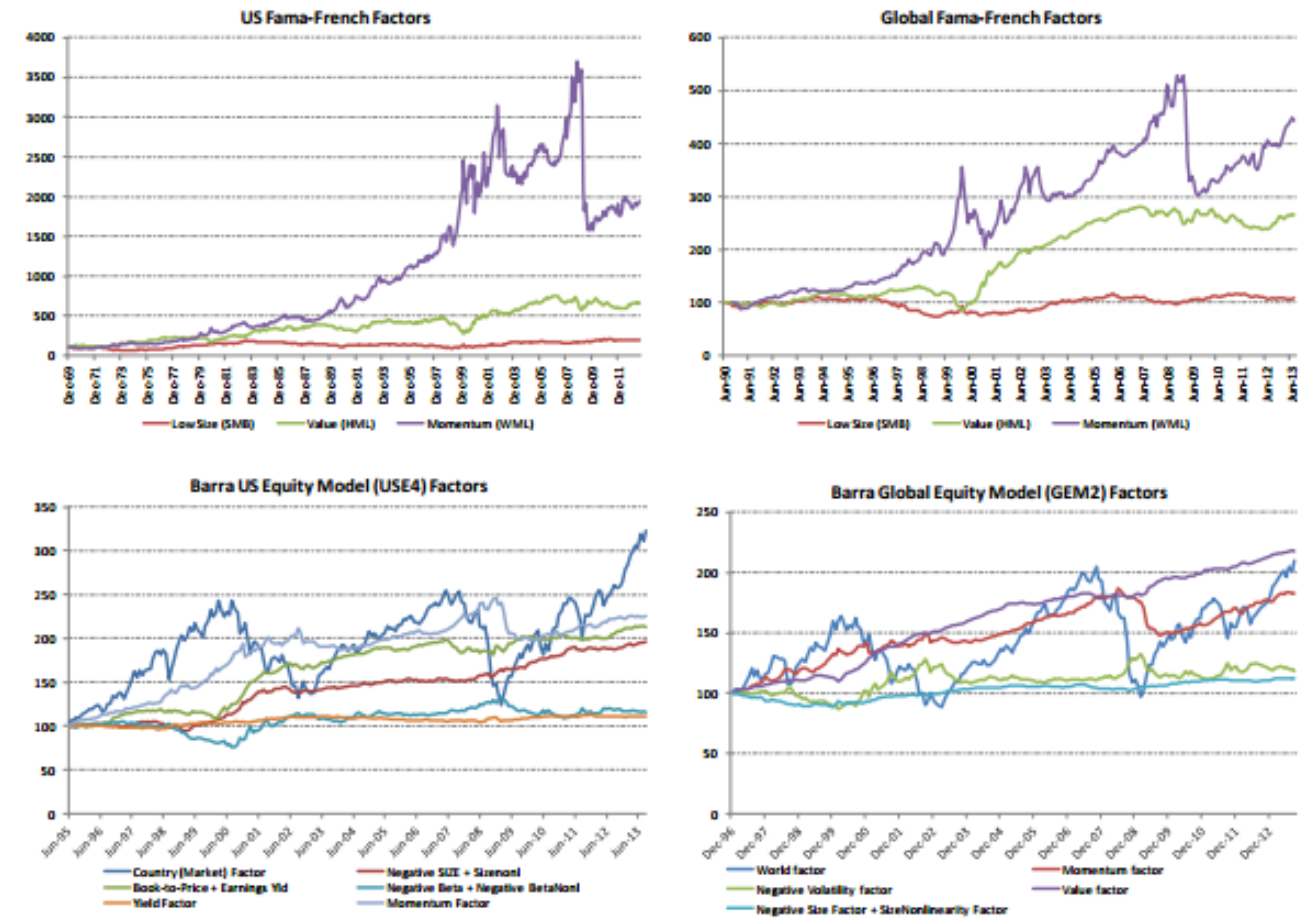

Figure 1: Market factors

In this report, we will examine the predictive power of the quality factors as described in Figure 2, which are much less understood than the traditional factors. Please note that in the following report, feature and factor are referring to the same concept, which is a particular feature of a data point.

2 Data

The data we are using is aggregated from Bloomberg terminal. We chose 1,000 of the most actively traded U.S. stocks, 42 quality features, which will be defined later, and the price series of the 1,000 different stocks. As mentioned in section I., quality features includes: ROE (return on equity), earnings stability, strength of balance sheets, etc. These can be viewed as the fundamental features of a particular stock. We initially chose roughly 110 different features, and narrowed down to 42 features that we think are most relevant to the performance of a stock. We later further reduce the number of features after reevaluating

Systematic Factors	What It is	Commonly Captured by
Value	➢ Captures excess returns to stocks that have low prices relative to their fundamental value	➢ Book to price, earnings to price, book value, sales, earnings, cash earnings, net profit, dividends, cash flow
Low Size (Small Cap)	➢ Captures excess returns of smaller firms (by market capitalization) relative to their larger counterparts	➢ Market capitalization (full or free float)
Momentum	➢ Reflects excess returns to stocks with stronger past performance	➢ Relative returns (3-mth, 6-mth, 12-mth, sometimes with last 1 mth excluded), historical alpha
Low Volatility	➢ Captures excess returns to stocks with lower than average volatility, beta, and/or idiosyncratic risk	➢ Standard deviation (1-yr, 2-yrs, 3-yrs), Downside standard deviation, standard deviation of idiosyncratic returns, Beta
Dividend Yield	➢ Captures excess returns to stocks that have higher-than-average dividend yields	➢ Dividend yield
Quality	➢ Captures excess returns to stocks that are characterized by low debt, stable earnings growth, and other "quality" metrics	➢ ROE, earnings stability, dividend growth stability, strength of balance sheet, financial leverage, accounting policies, strength of management, accruals, cash flows

Figure 2: Recently considered market factors

our approach of our goal.

We used data from 2010 to 2015. All the feature values were the value at the beginning of the particular fiscal year. For instance, we would have all the features values of all 1,000 stocks on January 1st, 2010, and we would have the price time series of the same 1,000 stocks from January 1st, 2010 to December 31st, 2010. We would label the stocks according to their annual return quartile: if the annual return is within the top 25Our intention is to split our 2010 2014 datasets into (800 training data points + 200 validation data points) and we use 2015 data as the test set.

3 Initial Failure

We initially hypothesized k-means clustering would be an effective tool to model and predict annual returns. We thought this because k-means attempts to cluster observations by minimizing the squared sums of the differences in features between observations or as illustrated in the following objective function: $\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$ where the set of observations is $(x_1, ...x_n)$, k is the number of clusters, S is the set of k clusters, and μ_i is the mean of points in S_i.

Features	Selected Stock names
Earnings/Price	EXXON MOBIL CORP
Forward Earnings/Price	APPLE INC
Price / Sales	INTL BUSINESS MACHINES CORP
FCF Yield	CHEVRON CORP
Dividend Yield	MICROSOFT CORP
Price / Tangible BV	JOHNSON & JOHNSON
Enterprise/Share	PROCTER & GAMBLE CO/THE
EV / EBITDA	GENERAL ELECTRIC CO
ROIC	AT&T INC
Cost of Capital	COCA-COLA CO/THE
ROE	PFIZER INC
ROA	GOOGLE INC-CL A
5-year (P/E)/(Current P/E)	PHILIP MORRIS INTERNATIONAL
YoY Sales Growth	ORACLE CORP
3-year Avg. Sales Growth	INTEL CORP
5-year Avg. Sales Growth	SCHLUMBERGER LTD
YoY Change in Cash Flow	VERIZON COMMUNICATIONS INC
5-year Avg. FCF growth	MERCK & CO. INC.
3-year Avg. Earnings Growth	PEPSICO INC
YoY% Change in NI	WAL-MART STORES INC
Sales/SGA	CONOCOPHILLIPS
Vol 10	MCDONALD'S CORP
Vol 20	CISCO SYSTEMS INC
Vol 30	QUALCOMM INC
Vol 60	ABBOTT LABORATORIES
Vol 90	AMAZON.COM INC
Vol 180	OCCIDENTAL PETROLEUM CORP
Vol 360	UNITED TECHNOLOGIES CORP
Market Cap	UNITED PARCEL SERVICE-CL B
CFFO	WALT DISNEY CO/THE
Long Term Debt	COMCAST CORP-CLASS A
Current Ratio	3M CO
Return on Working Capital	CATERPILLAR INC
Equity Shares Outstanding	ALTRIA GROUP INC
Gross Margin	HEWLETT-PACKARD CO
Asset Turnover	HOME DEPOT INC
CFFO > CFFO Last Year	UNITEDHEALTH GROUP INC
Borrow > Last Year	AMGEN INC
Curren Ratio > Last Year	BRISTOL-MYERS SQUIBB CO
Equity Share Outstanding > Last Year	BOEING CO/THE
Gross Margin > Last Year	CVS CAREMARK CORP

Figure 3: Our initial selection of features and some companies chosen

Our belief that k-means would be effective is rooted in the notion that stocks that exhibit similar values in our 42 quality features will also yield similar annual returns. K-mean clustering results are shown on the table below. We can see that whenever our clusters are generated, regardless of number of clusters, we see little difference in the Rand index, which is a measure of how well a clustering compares to our target results of separating our data labelled -1, 0, and 1, compared to randomly generated clusterings.

Cluster	Rand Index
3 clusters of random assignment	.5106
3 clusters formed with our data	.5201
4 clusters on random assignment	.5089
4 clusters formed with our data	.5192
64 clusters on random assignment	.6220
64 clusters formed with our data	.6436
128 clusters on random assignment	.6238
128 clusters formed with our data	.6431

Despite the poor results, we looked at the dendrogram of all of our generated clusterings to see if perhaps one of the clusterings was formed from mostly one label. Unfortunately, all of the clusters appeared to be random samplings of our data. We went further to see if any of the clusters were formed by mostly an industry, which might be a useful discovery. But again, we could not find any clusters that suggested clusters were being formed based on industry. Given this, we omit an illustration of the dendrogram because it is difficult to fit given our 1,000 data points and it does not show anything interesting.

We believe these poor results stem from these major reasons:

1. Too much noise
   Of our initial selection of quality features, many of these features are meaningless when we do not analyze the context of a company. For example, asset turnover ratio and current ratio of an arbitrarily chosen company from our 1,000 companies are meaningless even when compared to the other 999 companies because these numbers only hold meaning when comparing to a similar company in the same industry. Therefore, many of our quality features become noise and inhibit our clusters from being formed on our industry-agnostic features. And as we know about the objective function for k-means clustering, differences in features that indicate two companies belong to two different industries will be included in predicting differences in annual return. Furthermore, features with high standard devation have higher weight in determining clusters. Going back to the objective function, data with high standard devation and normalized will still retain that higher standard deviation, and this variance contributes more to the calculated distances between stocks. Also, these values with high standard deviation tend to be less significant when predicting annual return, with equity shares outstanding as an example.

2. Poor labeling of our data
   We separated our data into 3 values that indicated if their annual return was in the bottom 25, middle 50, or top 25 quartiles. There could a few issues with this. The first is that we used too few labels. If we were to increase our number of labels, say 10 labels for every 10 percent, then our model may prove to have more predictive power when we increase the sensitivity of our labeling. Another issue is that perhaps it is too optimistic to predict a stocks performance in a given year without any of its data

on previous years. Given this, we looked to change our labeling to be a function of a stocks annual return compared to a previous year.

After this failure, we made a few different changes to try and generate some positive results.

4 Using principle component analysis to reduce dimensionality

Since we identified too much noise as an issue for our clustering, we decided to use principle component analysis (PCA) to further trim our quality features. PCA calculates how much each features contributes to our labeling, so that we can set a threshold to remove noise. It does so by calculating the eigenvalue of each feature projected on the principal components and features with large eigenvalues contribute more to their labels. The formula to calculate this is: $\frac{\lambda_j}{\sum_{l=1}^{n} \lambda_l}$. We performed PCA on our data and found that only 9 of our features contributed to the variance in our labelling. Thus, we were able to further reduce our quality features to these 9:

1. Earnings / Price
2. FCF Yield
3. Price / Tangible BV
4. EV / EBITDA
5. Return on Invested Capital
6. Return on Assets
7. YoY Sales Growth
8. Growth Margin
9. YoY% Change in Net Income

After these changes, we found that the top 3 percentages of features contributing to variances are 64.1%, 13.5%, and 5.7%. Excited about our reduction in noise, we ran the same clustering tests as before and get the following results:

Cluster	Rand Index
3 clusters of random assignment	.5416
3 clusters formed with our data	.5912
4 clusters on random assignment	.5626
4 clusters formed with our data	.6012
64 clusters on random assignment	.6218
64 clusters formed with our data	.6841
128 clusters on random assignment	.6240
128 clusters formed with our data	.6560

As seen, reducing dimensionality has cut a lot of noise out of our data, so that clustering will not be influenced by arbitrary data. Currently, all of the features contribute at least 1% of variance to the labels, so we are satisfied with this current model. We kept the same labeling assignment, so that our results can be compared to our previous results with only one change. When examining the dendrogram, we noticed that we did see better clustering and that the clusters seemed less like a random sampling, but only by a little bit. This change alone is not enough to produce a profitable model, but this is definitely a step in the right direction.

5 Changing our labels

Separate from the previous section's changes, we decided to change the way we label our data. Keeping our previous 42 feature vectors, we only changed the labels to have 10 different values by dividing the annual return into tenths. Despite this change, we were unable to see significant improvements in our Rand index. The results follow:

Cluster	Rand Index
10 clusters of random assignment	.7985
10 clusters formed with our data	.8011
64 clusters on random assignment	.7971
64 clusters formed with our data	.7992
128 clusters on random assignment	.7682
128 clusters formed with our data	.7781

We concluded that this model still suffers from the issue of too much noise in the feature vectors, and that the noise is a much bigger contributing factor to our poor results compared to poor labeling.

6 Overhauling our features

Based off of feedback, we decided to overhaul our data representation to change our 1,000 stocks into a vector with 5 features: each feature being a comparison of annual return in years 2011 to 2015 compared to the previous year. This simplistic data representation was created to cluster stocks together to see if their annual returns correlate with each other. Using the 10 labels mentioned earlier, we were unable to compute a good Rand index. However, based off of the properties of a Rand index, we believed this to not be an issue. A Rand index determines how well clusters are formed compared to the true values, such that a cluster is most correct when all data from a cluster belongs to a single label. However, since we are using continuous values, clustering data labeled 1 and 0 together is much better than clustering data labeled 0 and 9. The Rand index does not account for this, so both are equally detrimental to the results.

Therefore, we looked into clusters themselves to see if we could find anything promising. We found that one cluster when using cluster size 128 contained mostly tech companies, which included Apple, Microsoft, Google, Amazon, Intel, Yahoo!, Oracle, Netflix, and Urban Outfitters. Aside from the last company, there seemed to be a clear pattern of tech companies being included in this cluster. Despite our clustering (what we consider) successfully recognizing companies that tend to have similar annual returns, our Rand index was worse than randomly generated labels for clusters of sizes 10, 64, and 128. We decided that this is likely because our label was generated on performance differences between 2015 and 2014, and the features were generated only on the 4 years of performance differences before this. Perhaps if we had data in decades of performance differences, we could better cluster companies, however we focused in depth on fewer years because of our initial focus to analyze quality factors. Furthermore, markets dont behave the same each year, much less each day, so a yearly performance difference might be too big of a time period gap to make meaningful inferences. Overall, we felt taking this simplistic approach offered valuable insight to combine with our previous testing.

7 Conclusions and Future Work

Utilizing machine learning to generate profits in the market is no easy task. If possible, funds globally would use these techniques to the point where the markets evolve to factor in machine learning, thus nullifying the ability of these techniques to generate profits. We understood this when deciding to pursue this topic, and we had a plethora of papers to reinforce this notion. Our main goal was to develop an understanding of the limitations of the impact of machine learning in markets. We tackled the problem with a few different angles, even overhauling our whole data representation, and were able to gain a better understanding of where human judgment of how to evaluate markets is necessary.

To continue on with this, we would like to further develop our 9 feature model that was able to generate clustering with a Rand index significantly better than the 42 feature model. Our PCA on the 9 feature model indicated that 4 features still contributed to less than 4% of the

variance in the label, so perhaps further stripping the features will yield even better results. Furthermore, adding additional data we did initially extract from the Bloomberg terminal or using data that spans decades could also yield better results. Even with these changes, we are doubtful to come up with a profitable model that relies mainly on machine learning, especially after factoring the overhead cost of trading or investing. Despite not coming up with a way to make money, we deepened our grasp of machine learning and markets, and this was good enough for us.

8 Bibliography

Bender, Jennifer, Remy Briand, Dimitris Melas, and Raman Subramanian. "Foundations of Factor Investing." (n.d.): n. pag. MSCI - Research Insight. MSCI, Dec. 2013. Web. May 2016.

Brown, Noam, Robert Mundkowsky, and Sam Shiu. "Predicting Intraday Price Movements in the Foreign Exchange Market." (n.d.): n. pag.Stanford.edu. Stanford. Web. May 2016.

Israel, Ronen. "Measuring Factor Exposures: Uses and Abuses." (n.d.): n. pag. AQR. AQR, Oct. 2015. Web. May 2016.

Kao, ChihChi. "Supervised Learning - Stock Trend Classifier." (n.d.): n. pag.Stanford.edu. Stanford, 16 Dec. 2011. Web. May 2016.

And of course, Introduction to Machine Learning Spring 2016 taught by David Sontag.