MS&E 448 FINAL REPORT
Trading Strategies based on Explained Cross-Sectional Returns

Group 3: Amay, Varin, Chor Seng, Yu Xin, Greg

Spring 2019

Abstract

Under market efficient conditions, cross-sectional returns variance (meaning that different stocks have different returns for the same time period) is explained by the different exposure to latent sources of systematic risks called risk factors. We employ various statistical techniques to identify these firm specific factors, rank firms according to them, and then build a zero-net-investment portfolio that longs the stocks with highest expected returns and shorts those with lowest expected returns.

1 Introduction of Strategy

Cross-sectional variance of expected returns are generally explained by exposure to systematic risk factors. Certain factors, such as value, low size, volatility, high yield, quality and momentum, have historically earned a risk-premium through exposures to systematic sources of risk. Factor investing, is therefore, harvesting these risk premia through exposure to the factors. Finding the "right" factors has thus become the central question of asset pricing as exposure to systematic risk factors which gives rise to risk premium are the sources of alpha.

For this strategy, we consider only the contemporary constituents of the S&P 500 Index for the purpose of our portfolio. We will select for stocks based on factors that we identify as having explanatory power on stock returns. Thereafter, we will long the stocks with highest expected returns (based on factors) and short those with the lowest expected returns. Our time horizon is one year, and the holding period of our portfolio ranges from a month to a year. We expect our portfolio to achieve a Sharpe ratio above 1.

2 Data and Investment Universe Selection

Our data universe consists of S&P 500 constituents from 2011 to 2018. We collected firm level characteristics from WRDS for these stocks, updating our S&P 500 constituents annually to account for survivorship bias in our selection. Using the S&P 500 constituents allows us to access significant amounts of firm level characteristics, that may not be as readily available for firms with smaller market capitalizations.

Some of the data was available daily, while others were only available monthly, quarterly or annually. We organized all the observations on a monthly basis; for quarterly data, we kept the same values for the 3 months of that quarter (Quarter 1 = January to March etc.), and annual data was copied for all the calendar months of that year. For the daily data (e.g. stock closing price), we only kept the data for the last day of each month. We ended up with over 40,000 observations and over 1,000 firm specific factors. We split this data into a training set (2011 - 2016), validation set (2017), and test set (2018).

During the data cleaning process, we observed that a large number of entries contained NA values. We removed all the factors that had more than 10% of values missing. Next, any rows/observations that still had missing values were deleted.

3 Models for Factor Selection

We used stepwise linear regression, principal component analysis and historical returns to select factors that are associated with higher excess returns.

3.1 Stepwise Linear Regression

From the 181 factors we downloaded, we plotted the correlation heat map of the factors.

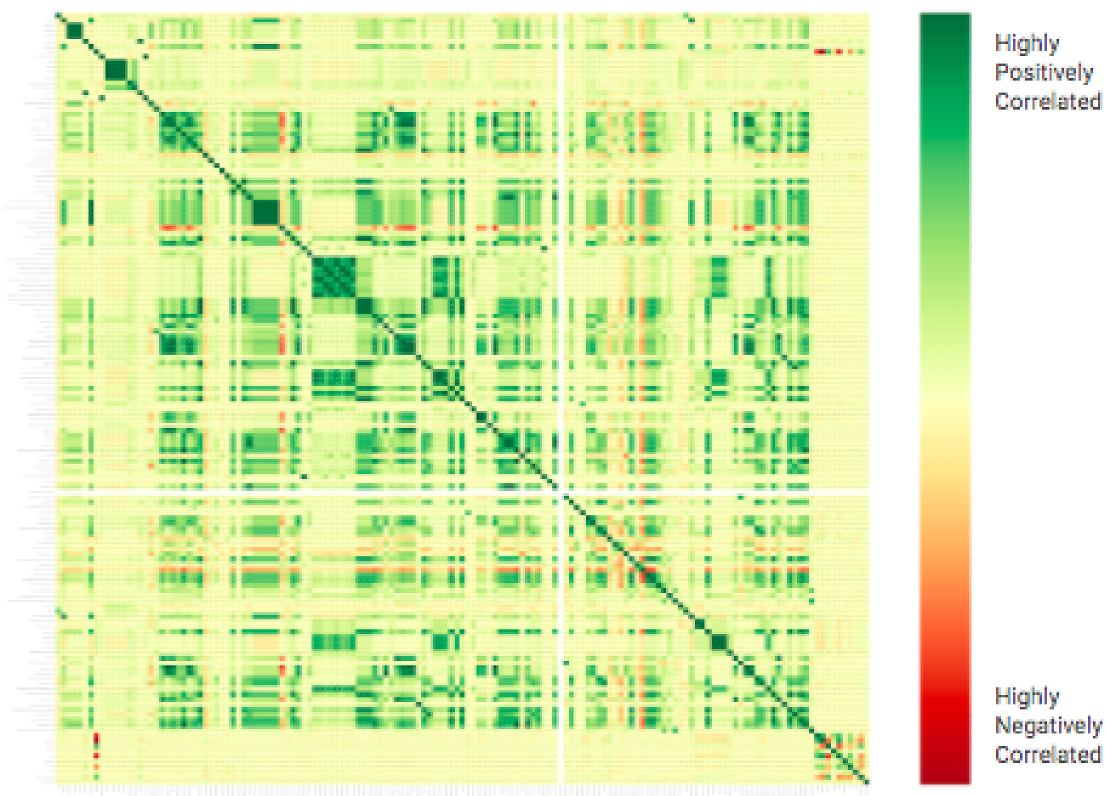

Figure 1: Heat map of correlation

We noted that there were clusters of strong correlation between different factors, rendering a direct linear regression impractical. Moreover, these factors failed to take into account the context, such as the size of the company. To prove this, we ran linear regression directly on 181 factors (Experiment A) to predict returns, as well as on a dimensionally-reduced feature vector (reduced using shallow and deep autoencoder). The conditions and results are shown in table 1.

(a) Conditions

	A	B	C
Feature size	181	30	30
Epoch	-	100	100
Deep	-	No	Yes

(b) Results (RMSE)

	A	B	C
Train	7.590	8.344	8.243
Validation	8.306	8.137	8.302
Test	8.812	8.263	8.845

Table 1: Linear regression on 181 factors or dimensionally-reduced factors

Experiments B and C were run using autoencoders to reduce the feature size from 180 to 30. Deep means that we utilized three hidden layers whereas shallow means that we utilized one hidden layer. For each experiments A, B and C, we also played around with the number of epochs, activation functions, feature reduction size and loss functions. The root mean square error ($RMSE$) did not change significantly with any of these variables changed and hence, the results are omitted from the report.

Due to the poor linear regression results, we decided to investigate the features in greater detail. By reading up on research literature, we decided to use the following systematic features and the firm-specific factors associated with them:

- Value (captured by Book Value to Market Value, CF Margin, CFO to Total Liabilities)
- Size (captured by Market Capitalization)
- Momentum (captured by Simple Moving Average (SMA) of Returns)
- Volatility (captured by Relative Trading Volume)
- Dividend Yield (captured by Dividend Yield)
- Quality (captured by Earnings to Price, Returns on Equity, Financial Leverage)

For this method, we first regress stock returns on market returns (represented by the $S\&P500$) to get residuals $\epsilon_{i,t}$. The linear regression equation for this step is given by,

$$r_{i,t} = \alpha_i + \beta_{m,i} r_{m,t} + \epsilon_{i,t}$$

where $r_{i,t}$ is the monthly total return of the asset, α_i is the intercept of asset i, $\beta_{m,i}$) are the factor loadings for asset i, and $\epsilon_{i,t}$ are the residuals.

As the second step, we regress the residuals $\epsilon_{i,t}$ against normalized factors (introduced in the bullet points above) to look at magnitude of significant factors at 95% confidence interval to select factors. The equation is given by,

$$\epsilon_{i,t} = \alpha_i' + \beta' f_{i,t} + v_{i,t},$$

where α_i' is the intercept of asset i, $\beta' = (\beta_1, \beta_2, ..., \beta_k)'$ are the factor loadings for k factors and f are the factor variables.

For the third step, we reran linear regression on unnormalized values of the selected factors and obtained the beta coefficients. Using the factors and beta coefficients, we predicted the returns on the validation and test set. The accuracy of the returns was measured using $RMSE$. We repeated this (experiment A) by grouping the data by industry (Experiment B) and grouping the data by company (Experiment C). If we group the data by industry, this means we obtained different $\beta_{m,i}$ values for different industries in the first step of the linear regression. The factors (chosen when p values are less than 0.05) and the corresponding $RMSE$ are shown in table 2.

(a) Factors

	A	B	C
Momentum	Yes	Yes	Yes
ROE	Yes	Yes	Yes
Trading Vol (Rel)	Yes	Yes	No
BV/MV	No	Yes	Yes
Earnings-Price	No	No	Yes
Other factors	No	No	No

(b) Results

	A	B	C
Train	5.590	5.575	5.532
Validation	5.306	5.339	5.791
Test	5.515	5.685	6.472

Table 2: Factors chosen from p-values and resulting errors

We note that as there is greater granularization of the data, there is greater over fit (observed when the test error is greater than the training error). Hence, we decided to focus our attention just on A (the experiment where we did not group the data). The overall linear regression equation obtained was,

$$Return = 0.979(SMA3) - 0.113(ROE) + 6.97e^{-6}(RTV) + 1.13(S\&P)$$

where Return is the monthly total return of a stock, SMA3 is the 3 month moving average of the stock's excess returns, ROE is the return of equity, RTV is the relative trading volume, and S&P is the monthly return of the S&P 500 Index.

We deduced from the coefficients that the two main factors were the simple moving average and return on equity (the relative trading volume had a small coefficient even when we took into account the relative magnitude of RTV values as compared to ROE and SMA_3 values).

3.2 Principal Components Analysis (PCA)

PCA was used as a dimension reduction technique to explain the majority of information in the sample covariance matrix. The principal components (PC) are constructed and ordered such that the first PC explains the largest variance, the second PC explains the second largest variance, and so on. Our aim was to find PC's that explain sufficient variance in the sample covariance matrix, and find correlations of these PC's with excess returns.

More specifically, 10 normalised factors were included in the sample covariance matrix: Book Value / Market Value, Market Capitalisation, Earnings/Price, Dividend Yield, Return on Equity, Financial Leverage, Price to Cash Flow, Cash Flow Margin, Operating Cash Flow / Total Liabilities, Relative Trading Volume, and Momentum represented by SMA 3-month excess returns.

We repeated the PCA analysis for the dataset containing all observations and the dataset only containing observations in the top and bottom decile of excess monthly returns. This is because the effect may be more pronounced for the top and bottom decile, or there may be specific traits for these top/bottom-performing stocks.

3.2.1 PCA: All Observations

From Figure 2 below, the first 4 principal components explain more than 85% of the variance. Therefore, we focus on the correlations between excess monthly returns and the first four PC's.

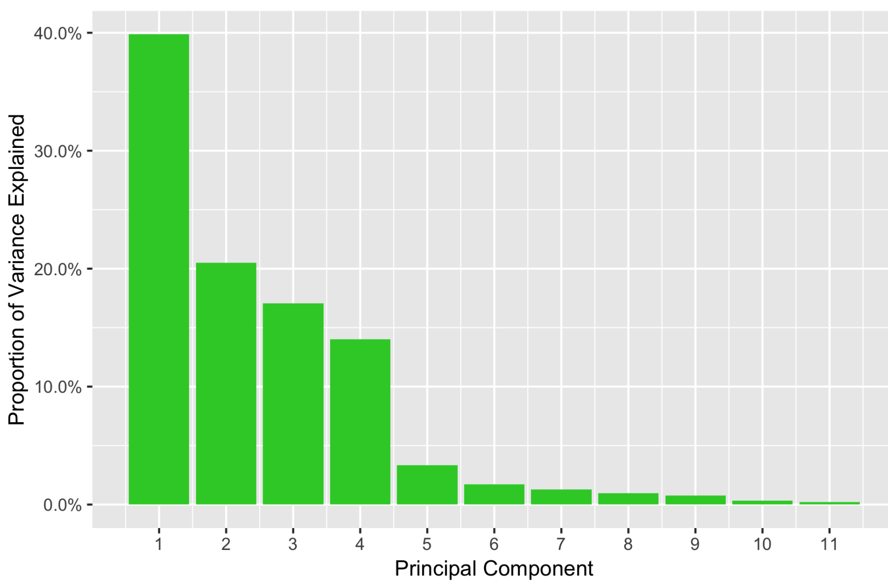

Figure 2: Proportion of Variance Explained By Each Principal Component

Figure 3 shows the relationships between excess returns and the principal components and Figure 4 shows the compositions of the respective PCs. Since excess returns are negatively correlated with PC2, this means that growth stocks and stocks with positive momentum have higher excess returns. Positively correlation between excess returns and PC3 implies that small caps do better. Furthermore, the positive correlation shown with PC4 shows that value stocks and positive momentum would likely have excess returns.

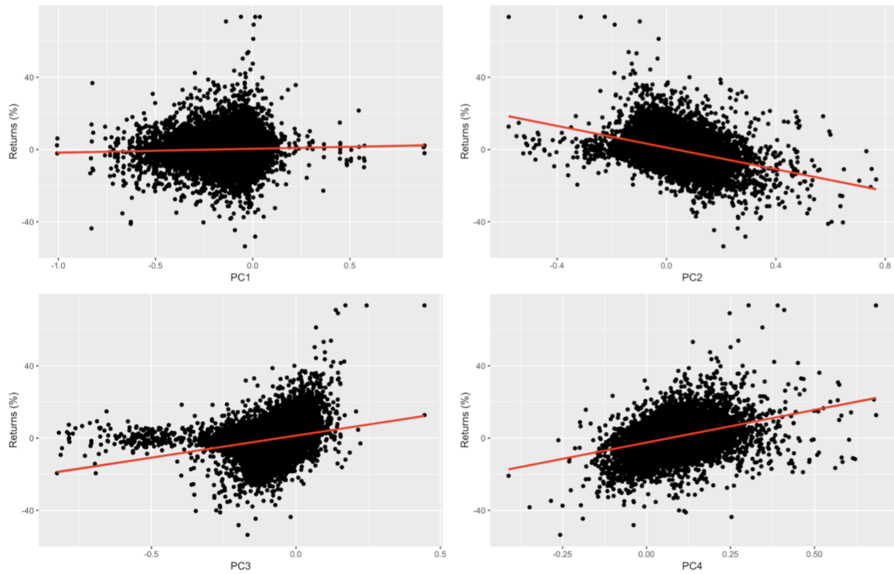

Figure 3: Correlation between Excess Monthly Returns and Respective Principal Components

	PC1	PC2	PC3	PC4
BVOverMV	-0.056	0.632	-0.251	0.723
Market Cap	-0.095	-0.51	-0.842	0.147
EarningsToPrice	0.006	-0.021	0.004	0.005
DivYield	-0.006	0.014	-0.007	0.002
ROE	0.002	-0.01	-0.001	-0.002
FinLeverage	0.001	-0.002	0.002	-0.001
PtoCFRatio	0.005	-0.001	-0.001	-0.003
CashFlowMargin	-0.983	0.001	0.108	-0.018
CFOtoTotalLiab	-0.141	-0.059	0.036	-0.089
TradingVolume_relative	0	-0.002	-0.005	0.002
ExcessMonthlyReturn_SMA3	0.036	-0.58	0.464	0.669

Figure 4: Weights of the First 4 Principal Components

3.2.2 PCA: Top and Bottom Decile

For PCA conducted on the top and bottom decile, the first 4 PC's also explain more than 85% of the variance. Excess returns are positively correlated with PC1 and PC3, and strongly negatively correlated with PC2. This implies that low cash flow margin, positive momentum and value stocks are correlated with excess returns.

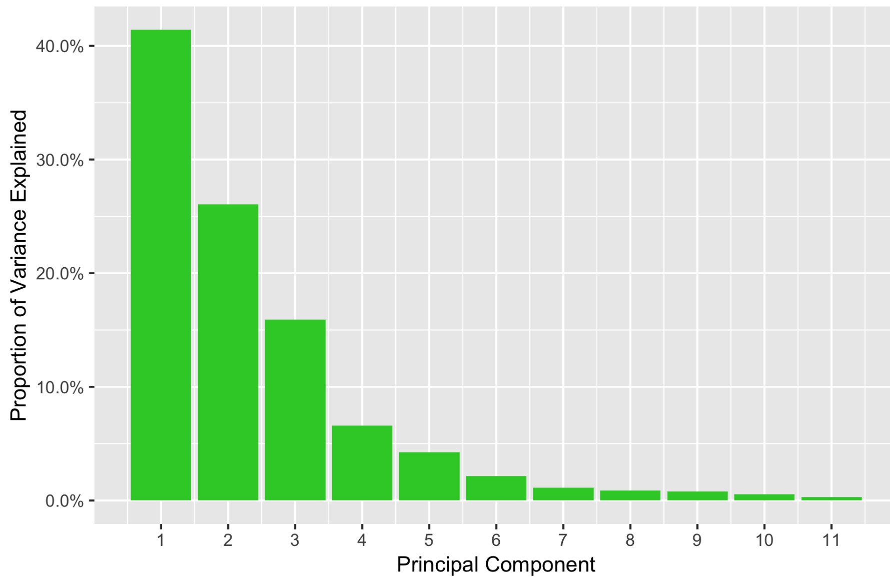

Figure 5: Proportion of Variance Explained by Respective Principal Components

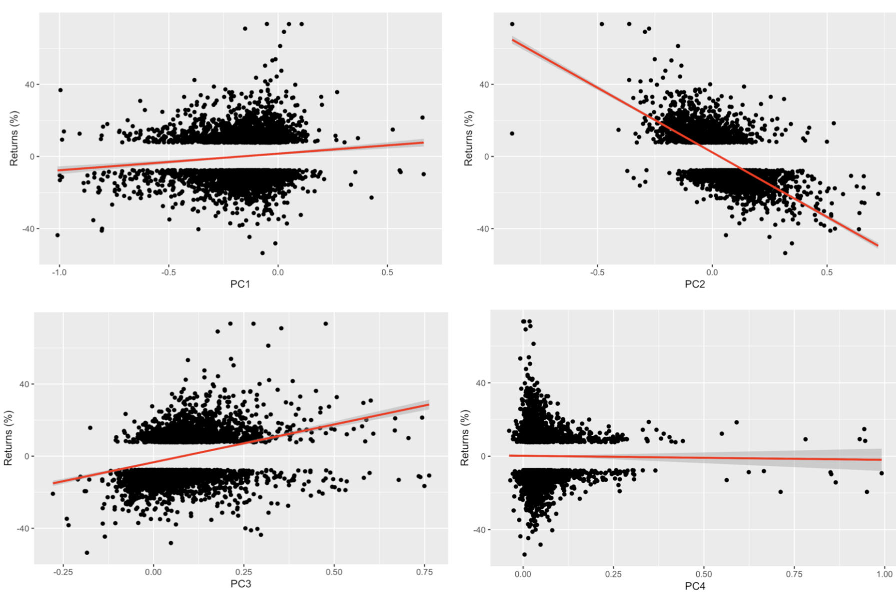

Figure 6: Correlation between Excess Monthly Returns and Principal Components

	PC1	PC2	PC3	PC4
BVOverMV	-0.122	0.441	0.867	0.136
MarketCapInverse	-0.030	-0.082	-0.107	0.988
EarningsToPrice	0.018	-0.028	-0.002	0.029
DivYield	-0.004	0.015	0.002	0.002
ROE	0.007	-0.005	-0.004	0.002
FinLeverageInverse	0.001	-0.002	-0.002	-0.004
PtoCFRatioInverse	0.002	0.001	-0.003	0.008
CashFlowMargin	-0.969	-0.134	-0.028	-0.055
CFOtoTotalLiab	-0.184	-0.087	-0.144	0.029
TradingVolume_relative	-0.002	0.000	-0.001	-0.004
ExcessMonthlyReturn_SMA3	0.107	-0.879	0.464	-0.020

Figure 7: Weights of the First 4 Principal Components

Overall, from the PCA conducted on all observations and the PCA conducted on just the top and bottom decile, one common factor that was positively correlated with excess returns was positive momentum.

3.3 Using Historical Returns

We examine each of the factors individually to investigate how they affect portfolio returns. To do this, we ranked all the stocks in our training set by each factor, and created portfolios by longing the top 50 stocks for each particular factor and shorting the bottom 50 stocks.

To evaluate each factor, we considered the average value of the factor over the past 12 months, and also the value of the factor from the previous month. We then selected the top and bottom stocks and rebalanced the portfolio annually (starting from January each year), quarterly (starting from Jan/Apr/Jul/Oct), as well as monthly. For the portfolios that were rebalanced quarterly and monthly, we only considered the factors from the previous month. The four strategies are summarized below:

- Rebalance annually, based on average value of factors from the past 12 months
- Rebalance annually, based on value of factors from the past month (December)
- Rebalance quarterly, based on the value of factors from the previous month
- Rebalance monthly, based on the value of factors from the past month

We repeat this for the top/bottom 25 and 10 stocks and compared all the portfolio returns at the end of 2016. We then identified factors that yielded a return that surpassed the risk-free return (defined to be 3.5% annually based on the national bank prime loan rate from the Federal Reserve Bank). The reason we used this as our comparison instead of the S&P 500 is because we are building net-zero portfolios.

The graphs 8 to 11 below show the results for each of the four strategies, and the overall results are summarized in Figure 12.

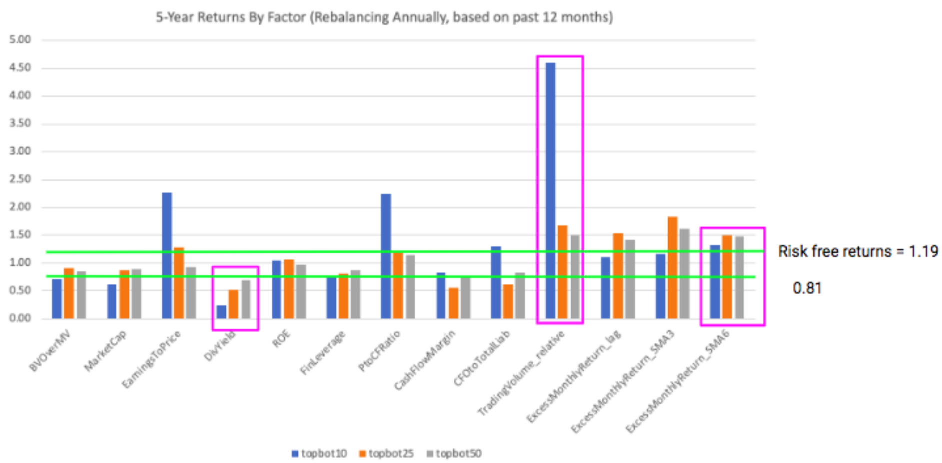

Figure 8: Returns by factor - rebalancing annually, taking average of past 12 months

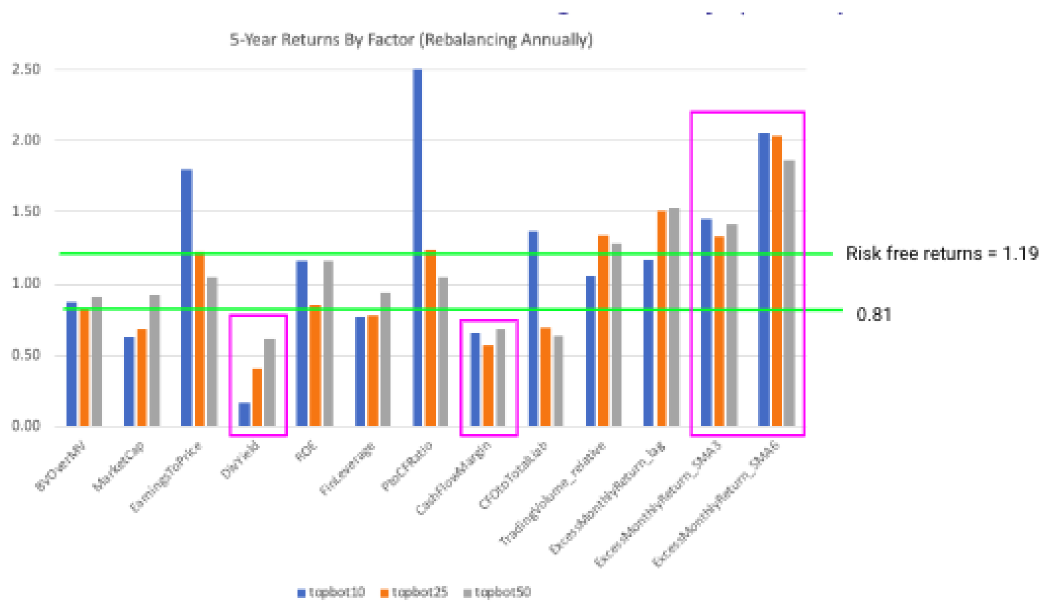

Figure 9: Returns by factor - rebalancing annually, based on previous month

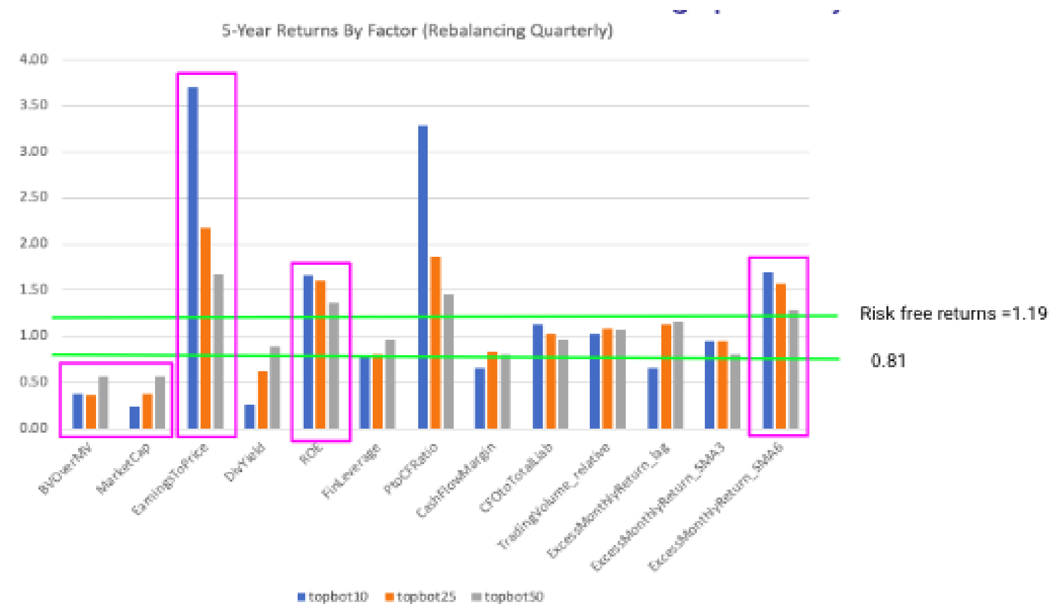

Figure 10: Returns by factor - rebalancing quarterly, based on previous month

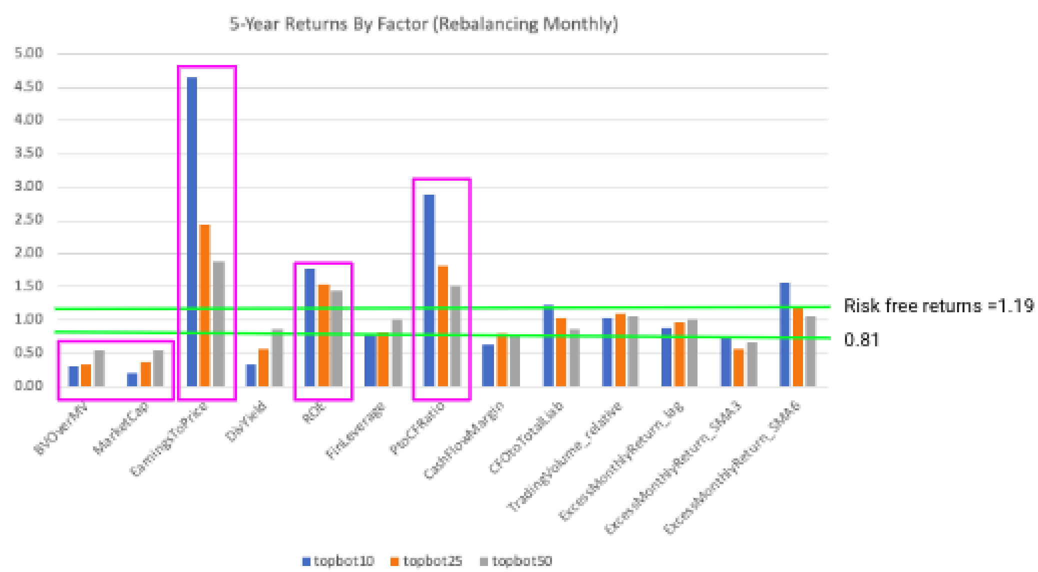

Figure 11: Returns by factor - rebalancing monthly, based on previous month

Factor	Rebalance Annually (using last 12m)	Rebalance Annually (using previous month)	Rebalance Quarterly	Rebalance Monthly
Low Book Value over Market Value (growth)			✓	✓
Large Market Cap			✓	✓
High Earnings to Price			✓	✓
Low Dividend Yield	✓	✓		
High ROE			✓	✓
Low Financial Leverage				
Low P/CF			✓	✓
High CF Margin		✓		
High CFO to Total Liab				
High Trading Volume	✓			
Lagged excess monthly returns (positive short-term momentum)				
SMA3 returns (positive momentum)		✓		
SMA6 returns (positive momentum)	✓	✓	✓	

Works for rebalancing monthly/quarterly strategy   Works for rebalancing annually strategy

Figure 12: Summary of results from building portfolios based on factors

We consider factors that had higher returns than the risk-free rate for top/bot 10, 25, and 50 strategies to be features that could generate good returns. In some cases, for example dividend yield for strategy 1, we longed stocks that had high dividend yield and shorted stocks with low dividend yield. Although this portfolio gave very poor returns (defined as losing more than the risk-free rate annually), we can build the reverse portfolio (i.e. long stocks with low dividend yield and short stocks with high dividend yield) and get a portfolio that beats the risk-free rate, hence these factors are also considered to generate good returns. In addition, we noticed that some of the factors seemed to only give returns above the risk-free rate when rebalanced monthly or quarterly (highlighted in yellow), some would give returns above the risk-free rate when balanced annually (highlighted in green), while some did not meet our criteria in any of the strategies.

Next, we used our 2017 validation set data to validate the features that we found. For the factors that worked for annual rebalancing, we found that both low dividend yield and high positive momentum (SMA 6 months) still gave returns that beat the risk-free rate. For the monthly and quarterly rebalancing data, we found that only large market capitalization and high earnings-to-price ratio still held.

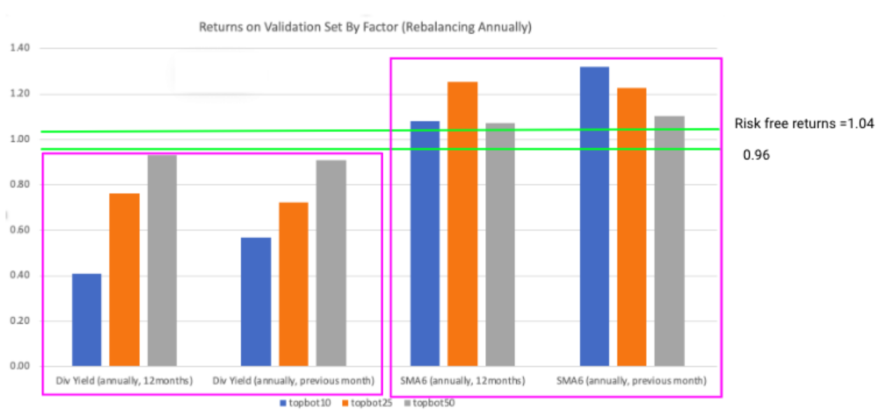

Figure 13: Validating factors - both dividend yield and momentum (SMA 6 month) still hold

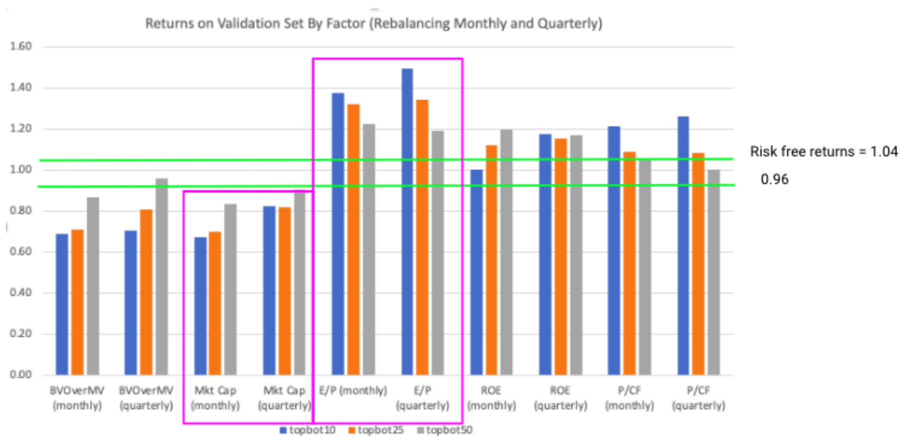

Figure 14: Validating factors - market capitalization and earnings-to-price still hold

3.4 Summary of factor selection

A summary of all the factors/features that were identified by stepwise regression, PCA, and historical return are summarized in the next table. The only factor that is identified in all three methods is the momentum factor, with the regression method also suggesting high return on equity as being an important feature, while the historical returns method suggests large market capitalization, high earnings-to-price, and low dividend yield as important factors.

Factor	**Linear Regression**	**PCA**	**Historical Returns**
Positive Momentum	✓	✓	✓
High ROE	✓		
Large Cap			✓
High E/P Ratio			✓
Low Dividend Yield			✓

Figure 15: Summary of results from the different factor selection methods

4 Validation of factors using Neural Networks

To validate our factor selection methods, we created a 7 layer fully connected neural network to see how well these selected factors explained excess returns. This neural network was trained on our dataset containing 13 firm-level factors for S&P 500 constituents.

We trained our model for 50 epochs, with a learning rate of $\frac{1e-2}{t}$ where t represents the current epoch. This resulted in a L_1 training loss of 4.7%, and a subsequent loss of 4.9% on our validation set.

Although we initially were able to get the training loss down to below 1%, this resulted in significant over-fitting to our training data, and consequently a much higher loss on our validation set. In order to avoid this, we implemented various methods to prevent over-fitting, including dropout regularization ($p = 0.4$) and early stopping. Our goal in training this model was to achieve similar losses on both the training and validation sets, to ensure we were not over-fitting to our training data. Furthermore, in order to reduce the loss on both sets, we reduced the number of features by filtering out irrelevant factors and normalizing the factors as ratios before inputting them into our model.

We used our model to predict excess returns on our test set, and sorted our stocks in descending order based on the prediction of our model. We can see the correlation between our network's predicted returns and the actual returns for our data set in the graphic below:

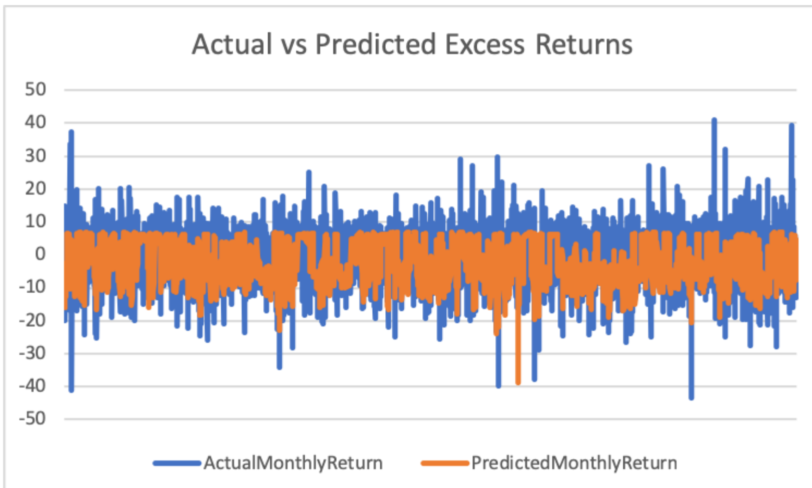

Figure 16: Plot of predicted and actual excess returns

Based off these predicted excess returns, we then constructed the following portfolios:

- Long top 10 stocks & short bottom 10 stocks:
  Annualized Return: 20.39%

- Long top 25 stocks & short bottom 25 stocks:
  Annualized Return: 12.94%

- Long top 50 stocks & short bottom 50 stocks
  Annualized Return: 9.46%

For each of the stocks in our portfolios above, we calculated the average value for the factors we predicted to affect returns the most. We can see these average values presented in the table below:

	Dividend Yield/%	ROE	Earnings to Price	Market Value/ billion $	Momentum (6 month SMA)/%
Top 10	0.41%	0.139252	0.010434	86.919	0.41%
Bottom 10	0.32%	-4.106257	-0.018279	8.102	-2.40%
Top 25	0.45%	0.083983	0.01228	86.648	0.64%
Bottom 25	0.37%	-1.672887	0.001149	9.563	-2.35%
Top 50	0.47%	0.066367	0.012473	76.554	0.49%
Bottom 50	0.56%	-0.821791	0.006562	12.854	-2.03%
All	0.50%	-0.057927	0.011955	40.815	0.12%

Figure 17: Average values for selected factors for portfolio constituents

From this, we can conclude that all of our selected factors except dividend yield demonstrated the expected correlation with returns, based off our factor selection models.

5 Portfolio Construction

Our current portfolio construction methods include ranking the stocks based on predicted excess returns, based off our selected factors. We then go long the top X number of stocks, and short the bottom X number of stocks (X = 10, 25, 50) to construct a zero-net investment portfolio.

Figure 18 shows that for each factor chosen, adopting a long-short portfolio strategy for the top/bottom 10 stocks rather than the top/bottom 50 stocks yielded higher returns. Thus, to aim for higher excess returns, we decided to use the top/bottom 10 number of stocks, instead of top/bottom 50 number of stocks.

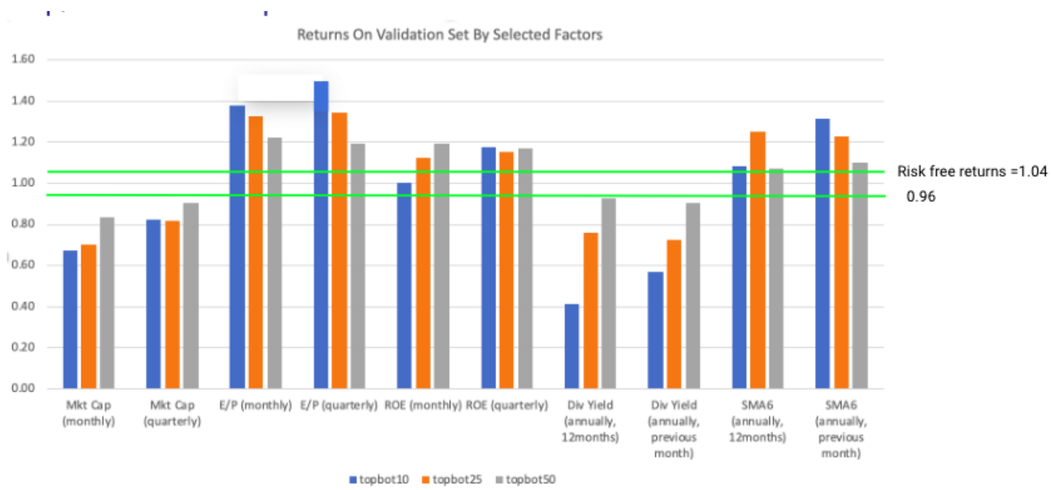

Figure 18: Returns on Validation Set By Selected Factors

Based on our investigation on the factors from Section 3, we found that high positive momentum was a feature that was identified by all three methods that we used, while other features such as high market capitalization, high earnings-to-price ratio, low dividend yield, and high return on equity were only identified by one method. As such, using out-of-sample data from our test set, we decided to allocate 50% of our portfolio to stocks with the highest 6 month SMA of excess returns, and 12.5% of our portfolio to each of the other four factors. We also shorted the stocks that performed the most poorly among the five factors, with an identical portfolio ratio. The results are shown in Figure 19.

The top/bot 50 portfolio achieved a return of 2.6%, which did not beat the risk-free rate (4.5% in 2018). However, the top/bot 10 portfolio achieved an impressive return of 19.4%. Interesting, the returns generated by each factor were quite different from each other, as well as for the top/bot 10 versus top/bot 50 investment strategies. Investing in high market capitalization companies did extremely well in both cases, and is resposible for most of the positive returns. However, the SMA 6 factor performed very poorly for the top/bot 50 strategy despite doing well for top/bot 50 strategy. The high market capitization factor had disappointing returns in the top/bot 10 case, and even generated a loss in the top/bot 50 case, while for the earnings-to-price factor the situation for the top/bot 10 and 50 cases are flipped. Finally, for the return on equity factor, the top/bot 50 case actually generated superior returns to the top/bot 10 case, which is contrary to our early findings. We discuss these observations further in the later sections.

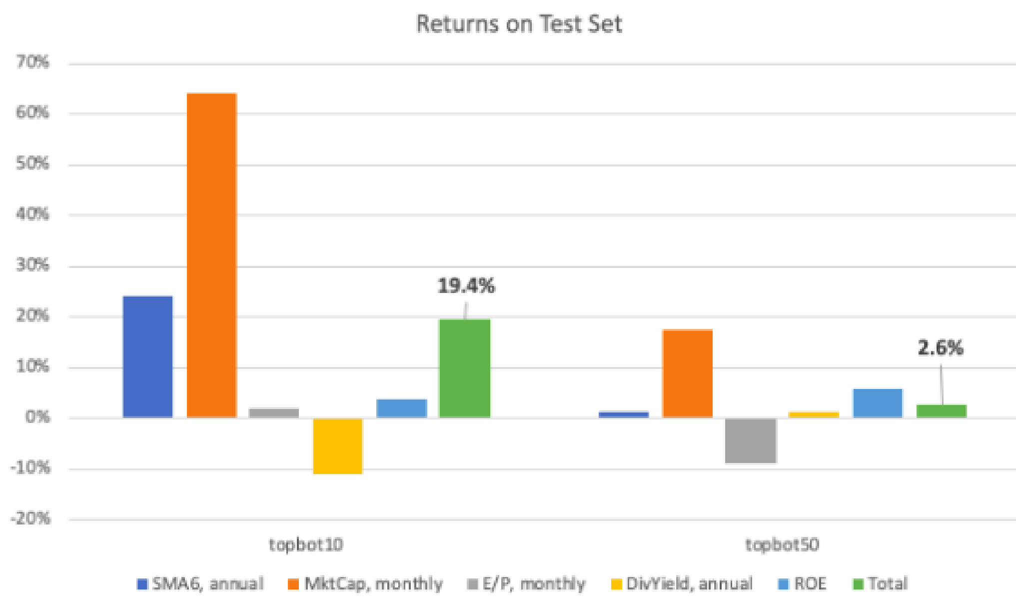

Figure 19: Chart of returns by strategy (top/bot 10 vs. 50), sorted by factor and also the total portfolio return

6 Risk Management Philosophy

By employing a zero-net investment portfolio, we greatly reduce our exposure to market risk. We also diversify our portfolio by selecting for multiple factors, which is an important risk management measure given the likelihood that the influence of factors in explaining returns is expected to change with time.

We calculated the Sharpe ratios for using a top/bot 10 versus top/bot 50 investment strategy for our validation set and test set, and the results are displayed in the table below. For the validation set, we got high Sharpe ratios for both cases, with that of the top/bot 50 strategy being higher, as we would expect since the standard deviation of the excess returns is lower. Hence we get a higher risk-adjusted returns. However, in the test set, we got a negative Sharpe ratio for the top/bot 50 strategy as that portfolio under-performed the risk-free rate. This suggests that it might be better to adopt a top/bot 10 strategy despite the slightly greater risks.

	Top/Bottom 10	Top/Bottom 50
Validation	7.49	8.73
Test	9.80	-2.37

Table 3: Sharpe ratio for validation and test set

7 Retrospective Discussion

7.1 Data Processing

Our current method entails splitting the training, validation and test set by year, but this may affect results due to time-related market conditions. By selecting factors using the training and validation set, we assume that these factors will continue to yield excess returns in later years. Therefore, it might be worth exploring a random split.

7.2 Factor Selection

In our strategy, we only selected factors that yielded in excess of risk-free returns when going long top/bottom 10, top/bottom 25 and top/bottom 50 stocks. While this ensured more consistency and reduced overfitting, we could have built a more specific factor selection strategy that matched with the x number of stocks we want to long/short. For example, some factors might not have worked for top/bottom 50, but worked very well for top/bottom 10 and top/bottom 25. In our model, we would have not included these factors, but in the future, we could look into pairing a number of stocks to long/short together with the factor.

Because our results suggest that the factors that can best explain better returns change with time, it will be worthwhile to investigate what the optimal length/duration of historical data to look at is to determine the most current factors, and how long these factors can be expected to continue to have strong explanatory power. It would also be interesting to try to study if the factors that matter vary with changing macroeconomic conditions, as it is rather plausible that there is some sort of correlation between the two.

Another idea is that since we are considering only stocks from the S&P 500 Index, which is a rather selective group of stocks in general, it might be beneficial to have a simpler model, perhaps just with just two instead of the five factors as we did in the project.

7.3 Portfolio Construction

Our current method consists of allocating equal amounts to each stock in our portfolio. In the future, we hope to allocate based on the strength of the signal and see whether in practice this produces higher overall returns for our portfolio. We could do it more quantitatively whereby we weight factors that yield consistently larger magnitude of returns more heavily, than factors that just beat the risk-free rate. A more quantitative approach can also be taken by optimizing the weights of the factors we have chosen over several time periods.

We currently do not account for transactions costs in our strategy, but given this is a low frequency trading strategy, we do not expect this to have a huge impact on the returns of our portfolio.