Long-Short Equity Strategy Leveraging Machine Learning

Abstract

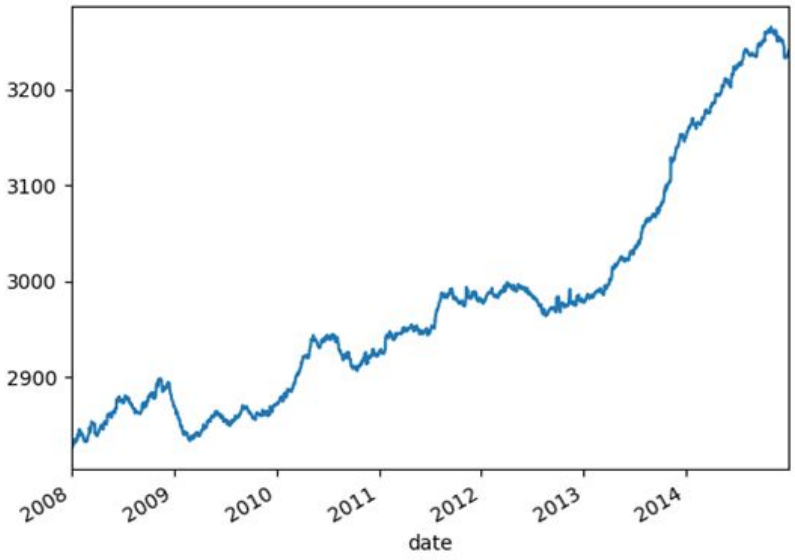
Machine Learning algorithms have been demonstrating promising results for investment signal generation and security selection (Myr). This analysis demonstrate the performance of a long/short equity investment strategy in which security selection for both the long and short arms of the strategy is performed by training a random forest to predict the securities from the population of publicly traded equities that are most likely to outperform or underperform the median return of the market. The optimized model selects securities with sufficient accuracy to outperform the market over the backtesting period in several variations of the long/short strategy.

Security Selection/Data

The Data for the Stock Selection Model come from CompuStat. The dataset starts in 01-Jan-2004 and is used until 31-Dec-2014. The remainder of the data is for further Out-Of-Sample Analysis. the following filters are applied to the dataset:

1. only Common Stock used
2. The Stock needs to trade for at least 1 year
3. The Company should have no missing or data errors during this period

This leads to an average of around 3000 Stocks in the investment sample.
All stocks are adjusted for dividends and corporate events.
Below are stocks available over time:



Create Features and Targets

1. Feature Creation

The model uses the following Features on a High Level:

	Nr of Features
Return Feature	31
Price Feature	32
Time std Feature	30
Cross Ranked Return Feature	31
Calendar Feature	31
Cross Std Feature	31
Cross Time Std Ranked Feature	30
Technical Features	53
Technical Features Normalized	51
Technical Features Normalized Ranked	51

1. Return Feature = Historical Returns of the Stocks over different Days
2. Price Feature = Historical Average of the Stocks Close Price over different Days
3. Time std Feature = Historical volatility for Stocks

4. Cross Ranked Return Feature = Ranked Returns of Stocks
5. Calendar Feature = Days and Months of the Week and Year.
6. Cross Std Feature = volatility over historical stocks
7. Cross Ranked Time Std = Ranked volatility for Stocks
8. Technical Features = Most Common Technical Features

the Features are created every day for every Stock for the 11 Years. That roughly leads to billion Feature data points.

Number of Feature Data Points = 11 years*252 days* 3000 stocks * 377 Features = 3.14 Bn

2. Targets

The Prediction Problem is a binary Classification. The Classification is based on outperforming or underperforming the weekly median of all stocks returns. Binary Classification from a Machine Learning sample is a lot simpler than a regression. The Goal is to show that one can add value on the simpler question before moving to a harder regression problem.
Roughly 8.4 million target data points are created.

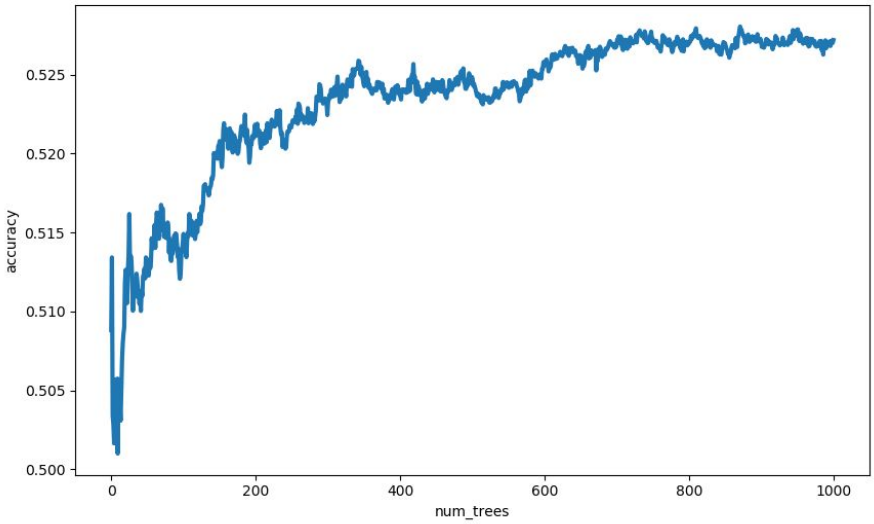
Number of Target Data Points = 11 years*252 days*3000 stocks *1 Target = 8.4mil

Investment Signal Generation

Signal Construction

1. Different Strategies
    a. Base Model

The Base Model is the core of our prediction model and uses the most data to predicted the cross section of stocks. The First Model is trained on Data from 2005 to 2006 and uses 2007 as Test period and every data point after 01-Jan-2008 is Validation. A Large Random Forest is used. Below is the Out of Sample Accuracy for 2007 based on the Number of Trees:



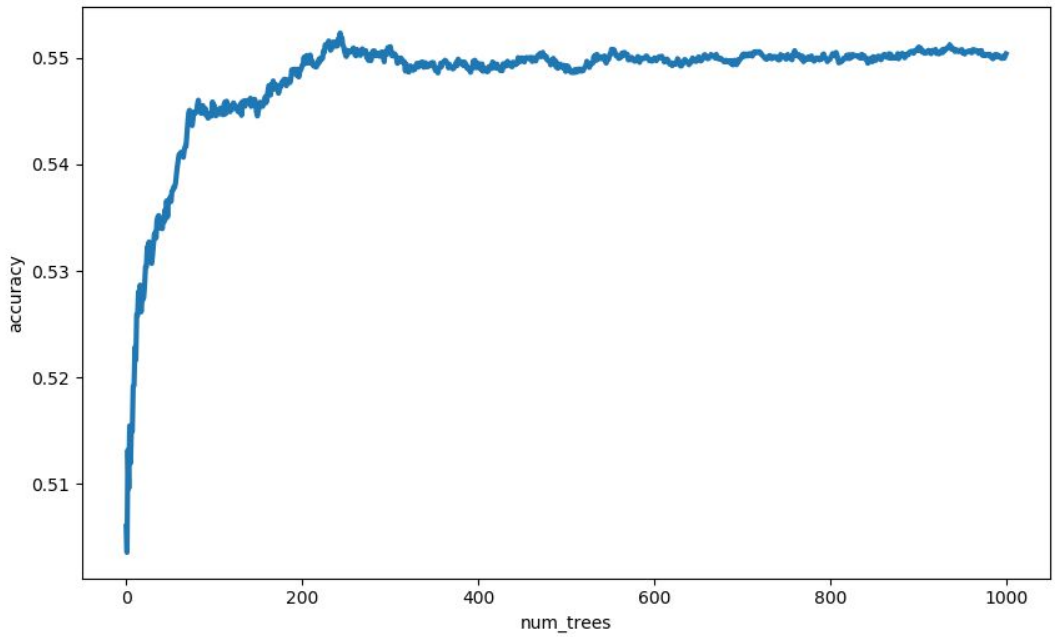
As one can see the Accuracy stabilizes after 750 Trees. Because of the heavy Regularization, the Random Forest converges against an Out-Of-Sample Accuracy. The model is run with 1000 Trees. This ensures that the model will have converges in the Out-Of-Sample Accuracy. The model is refitted on expending window every year, which makes it the slowest model to adjust to changes in Market structure.

b. Monthly model

As it is well know that monthly pattern exist in Financial markets, that's why another monthly Model is trained. This Monthly Model for example in January will only look at data points which occurred during January of the previous years. The Model is retrained every month. The same random forest parameter in the monthly Model as in the Base Model are applied.

c. Short-Dated Model

Both the Base and Monthly Model are slower moving model and are not that fast in adjusting to new market regimes. That is the reason another third adding Short-Dated Model is added. The Short-Dated Model is only trained on the Data from the previous month. The Test Accuracy can be found below. The Accuracy converges much faster than for example in the Base Model to a more constant Out-Of-Sample Accuracy. Model parameters are the same as in the Base Model of 1000 Trees.



Summary

There are three different Prediction Models which will predict for each stock each day the probability of out- or under-performing the median of all Stocks over the next 5 Business Days.

Portfolio Construction

1. Portfolio Construction for Base, Monthly and Short-Dated Portfolio

The Construction method for each of the Signals is the same. Each Signal holds 5 Sub Portfolios. The Trading frequency and rebalancing for each of the sub portfolios is 5 Business Days. For example; sub portfolio 1 will hold Stocks from Monday to next Monday whereas sub portfolio 2 will hold stocks from Tuesday to next Tuesday and so.
Each day the Model picks the 50 stocks with the highest and lowest probability of outperforming and goes long and short the Stocks accordingly.
The Rebalancing between the Sub Portfolios happens at the end of Month. All weights are equally weighted.

2. The Main Portfolio

The Main Portofolio holds an equal weight between the Base, Monthly and Short-dated Portfolio. The rebalancing happens at the end of each Month.

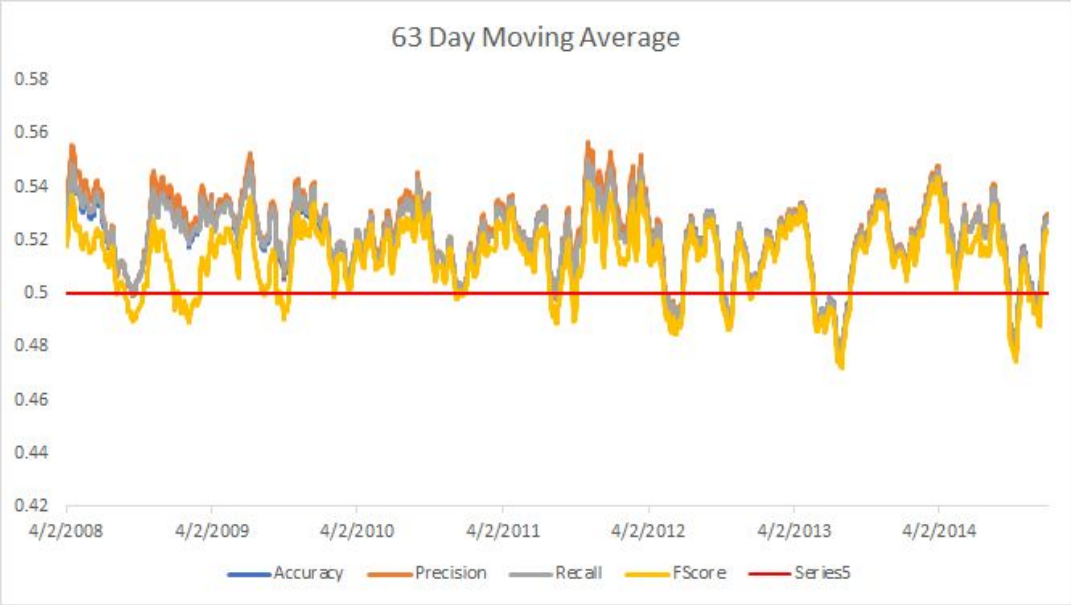
Transaction Costs

A constant spread of 0.20% is used on all Trades This is equivalent to 1-2% of the Market for Small Cap Stocks. This seems reasonable in first order.

Algorithm Diagnostics

Model Accuracy,Precision & Recall

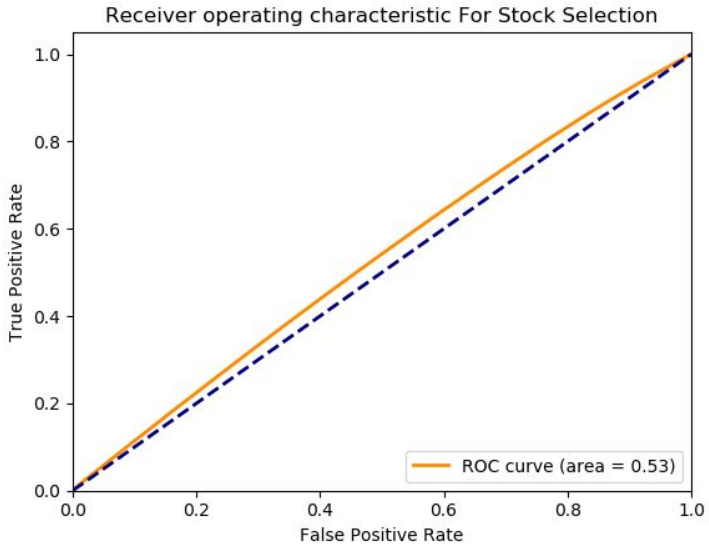
	Accuracy	Precision	Recall	Fscore
Base Model	0.520981	0.520979	0.520979	0.520976



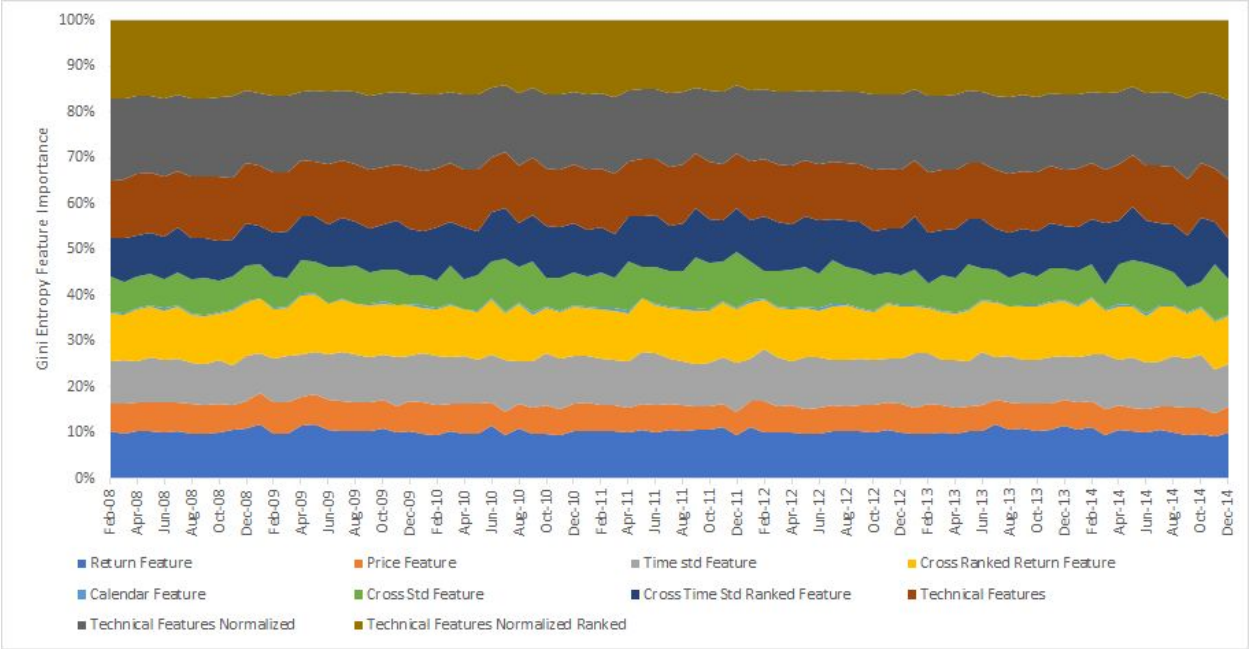
As One can see the Accuracy, Precision and Recall are very closely moving which indicates that the problem has a very low signal to noise ratio. Also, there are certain periods where the Model

drops below 50% Accuracy, which indicates a failure of the Model. This could be fixed by adding a hedge or timing the model.

The low signal to noise ratio can also be further seen in the ROC curve. This shows that One can add value but it is very small and hard



Feature Importance

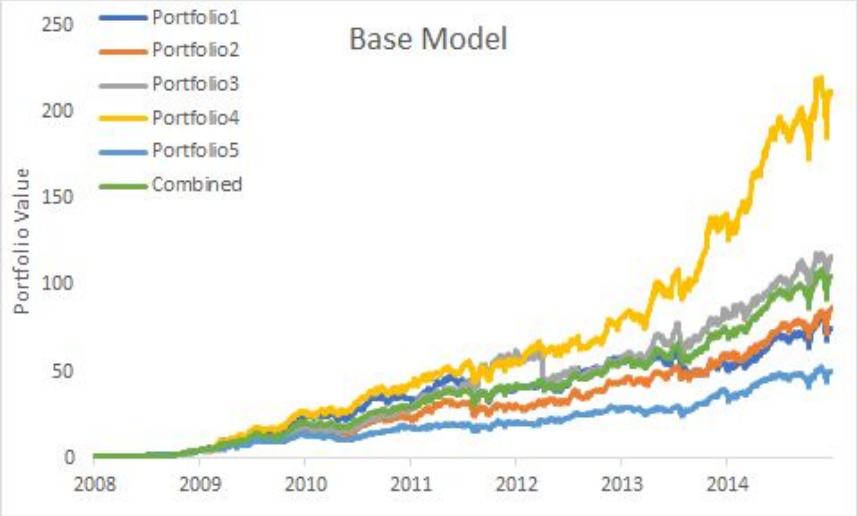


This chart shows the Feature Importance for the Validation Period of the Short-Dated Model. Technical Indicators are as expected the most important source of Information for the Model. However, The Return Features are in general also important but calendar effects do not matter. Another remarkable point is that the Feature importance is relatively stable over time.

Investment Strategy Backtest Results

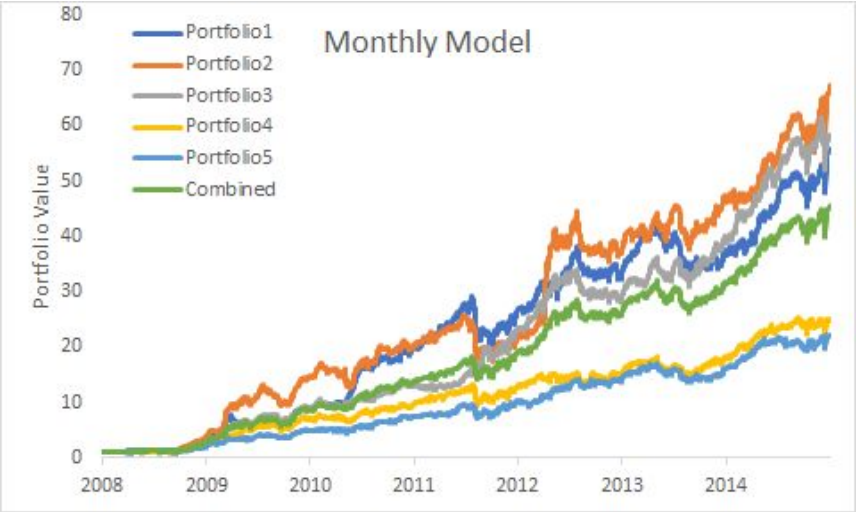
Strategy Performance

1. Base Model



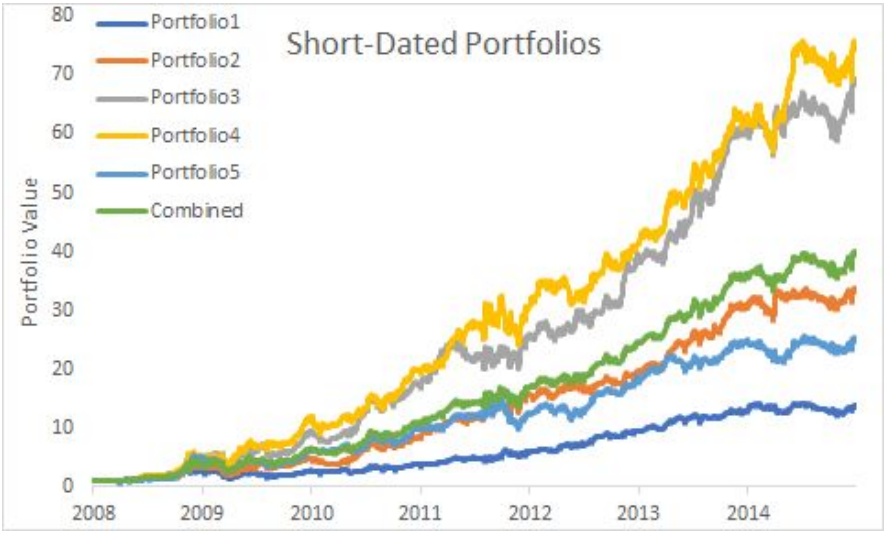
	Base Model					
	Portfolio1	Portfolio2	Portfolio3	Portfolio4	Portfolio5	Combined
Return	67%	70%	74%	82%	61%	71%
Vol	34%	33%	35%	34%	32%	28%
Sharpe	2.01	2.08	2.10	2.43	1.90	2.52

2. Monthly Model



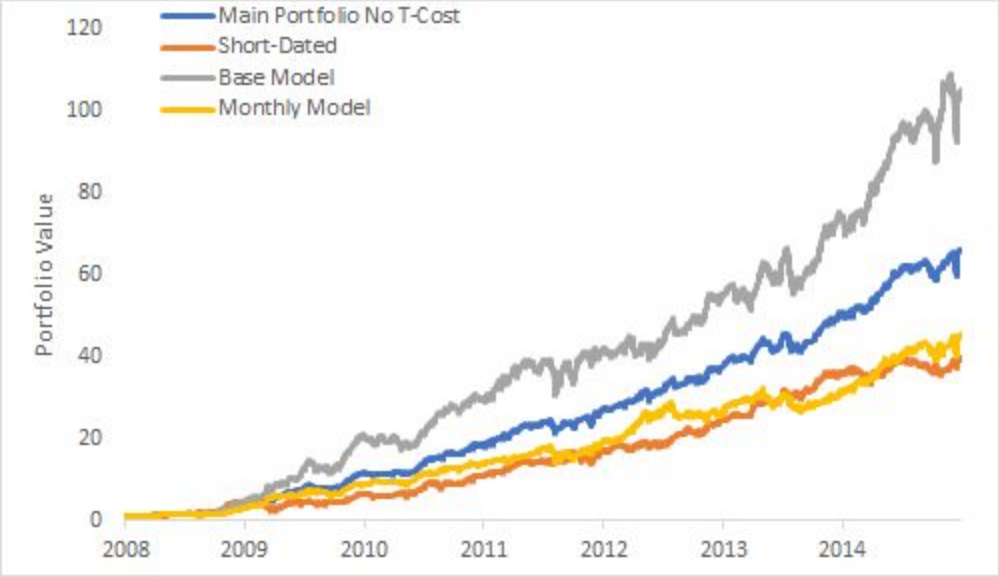
	Monthly Model					
	Portfolio1	Portfolio2	Portfolio3	Portfolio4	Portfolio5	Combined
Return	62%	65%	62%	50%	48%	58%
Vol	29%	30%	28%	29%	28%	24%
Sharpe	2.12	2.16	2.23	1.76	1.73	2.41

3. Short-Dated Model



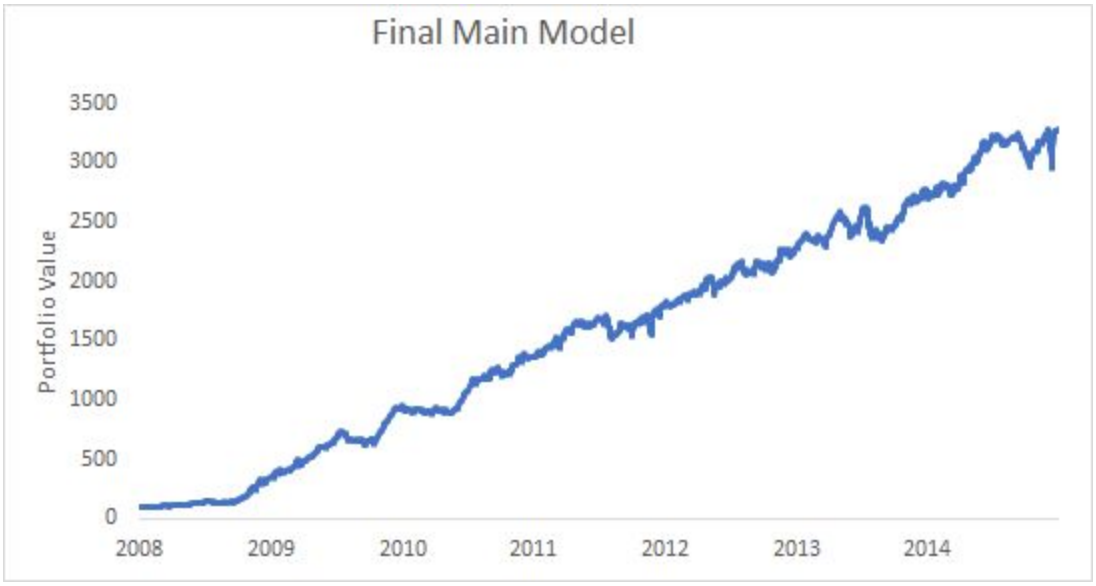
Short Dated Model						
	Portfolio1	Portfolio2	Portfolio3	Portfolio4	Portfolio5	Combined
Return	43%	56%	67%	68%	52%	57%
Vol	35%	35%	36%	36%	36%	31%
Sharpe	1.24	1.61	1.87	1.89	1.47	1.85

4. Main Model

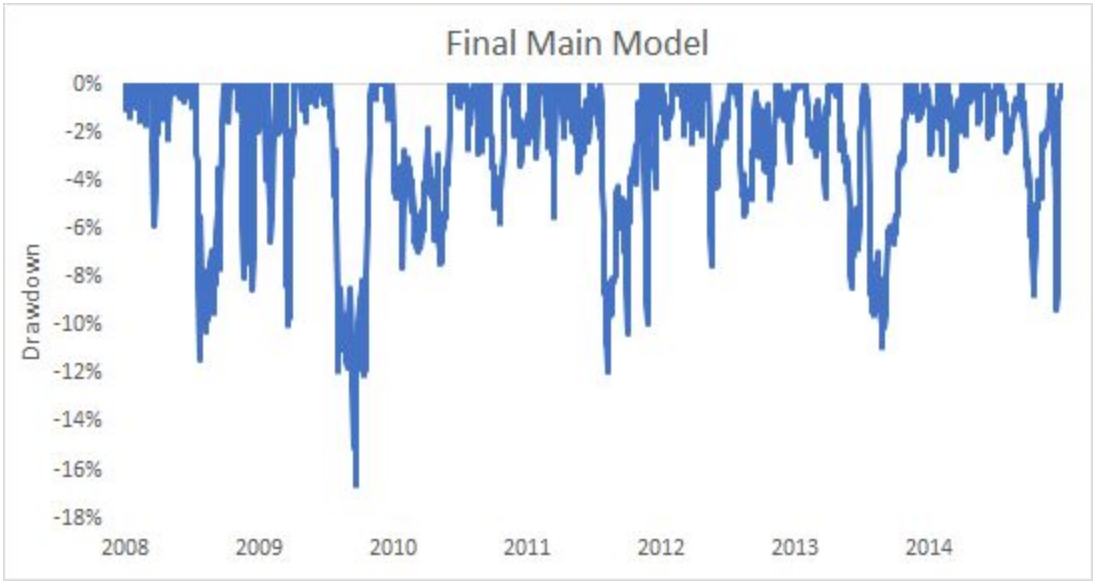


Main Model Summary				
	Short-Dated	Base Model	Monthly Model	Main Portfolio No T-Cost
Return	57.4%	70.5%	57.5%	61.8%
Vol	31.0%	28.0%	23.9%	18.9%
Sharpe	1.85	2.52	2.41	3.27
Max DD	-43.1%	-21.6%	-22.3%	-15.0%

Final Main Model



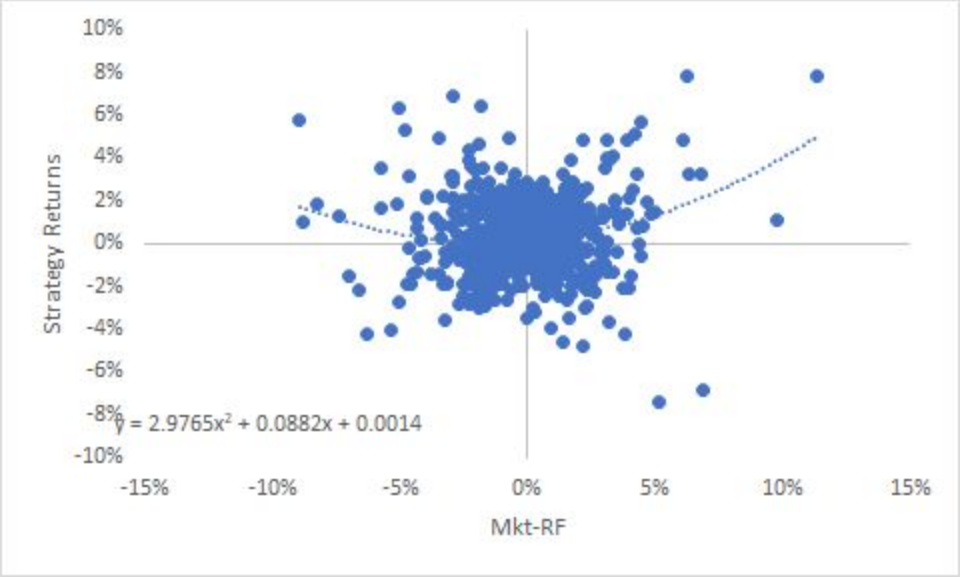
Main Model Summary		
	Main Portfolio No T-Cost	Main Portfolio T-Cost
Return	61.8%	51.7%
Vol	18.9%	18.9%
Sharpe	3.27	2.74
Max DD	-15.0%	-16.5%



Strategy Analysis

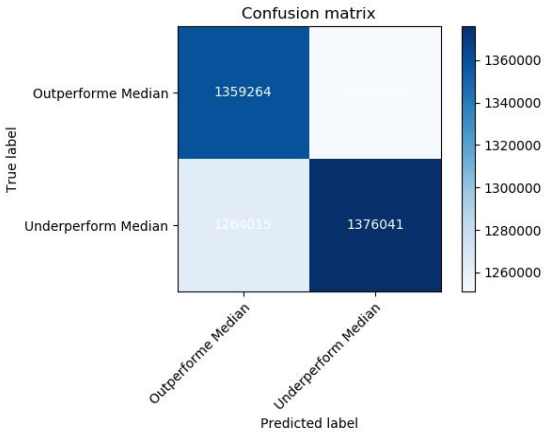
	Coefficients	t Stat
Intercept	0.20%	7.48
Mkt-RF	0.18	8.01
SMB	-0.13	-2.87
HML	-0.48	-10.76
RMW	-0.31	-3.52
CMA	0.03	0.35
Annualized Alpha	51.5%	

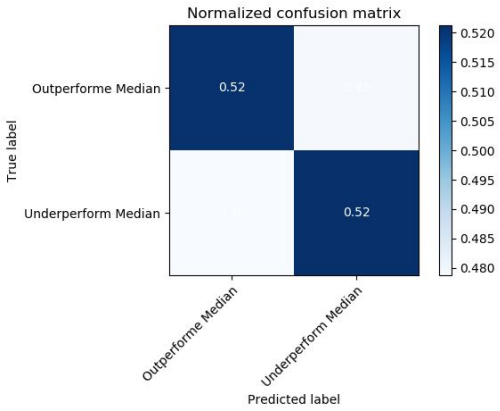
The Strategies exposure is tested to the five Fama French Factors. The Regression is run on daily data. As on can see the Strategy has a beta of 0.18 to the Market which is highly significant. The Strategy goes long Large Cap Firms and Short Small Caps as one can see in SMB with a roughly similar beta. HML is also very significant negative. However, all of the exposures are not explaining away the Alpha. This means the strategy picks up other drivers for the return besides the most common factors. This results suggest one should add layer to hedge out other Market Exposures.



This scatter plot shows that the Strategy is able to have a negative correlation during negative stock market days which is great feature. Moreover, the Strategy's scatter plot against the market looks like a long straddle performance

Appendix


Confusion matrix
Outperforme Median
Underperform Median
True label
1359264
1264015
1376041
Outperforme Median
Underperform Median
Predicted label
1360000
1340000
1320000
1300000
1280000
1260000


Normalized confusion matrix
Outperforme Median
Underperform Median
True label
0.52
0.52
Outperforme Median
Underperform Median
Predicted label
0.520
0.515
0.510
0.505
0.500
0.495
0.490
0.485
0.480