Data Analytics and Machine Learning Final Project

Predicting Power and Arbitrage of Financial Ratios at Industry Level

Introduction

In the financial field, most scholars conduct their researches on stocks or indices for trading strategy or return prediction, while few dives in the level of industries to observe how industries differentiate from each other on fundamentals or other perspectives. Such differences can possibly generate profitable trading strategies. In my project, I fit 49 Fama French industry portfolio returns and fundamental ratios into LASSO model and obtain suitable ones as predicting factors for each industry. Then I apply Elastic net method on selected variables using cross validation to create the mean-variance efficient portfolio as my industry portfolio monthly frequency trading strategy for hedge funds. In this project, I archived two goals. First, I look inside prediction power of the financial ratios on each industry and find the best set of the variables that reduce the prediction error. Second, based on the set of the best variables and using cross sectional regression, I select the variables for the further step under the best lambda and create the MVE portfolio at industry level which has the better performance than the value weighted market portfolio.

Data

In this project, I mainly use three datasets, Fama French industry portfolios, the corresponding Financial ratios and the risk-free rate. The industry portfolios consist 49 industries covering most stocks' segments in the market as Table1 shows. The raw industry data is divided into single industries, in which the value weighted returns of each industry's stock are used as dependent variable. I also download 72 Financial mean accounting characteristic ratios at industry level from Compustat-North America database as independent variable shown in table 2. In order to predict industrial excess returns with these accounting ratios, I also obtain risk free rate data from Kenneth French website. All data ranges from 1970 to 2019 on a monthly basis. The industry portfolio returns has already been lagged.

Table 1. 49 Fama French Industry Portfolios

AERO	BOXES	DRUGS	GUNS	MEALS	RLEST	STEEL
AGRIC	BUSSV	ELCEQ	HARDW	MEDEQ	RTAIL	TELCM
AUTOS	CHEMS	FABPR	HLTH	MINES	RUBBR	TOYS
BANKS	CHIPS	FIN	HSHLD	OIL	SHIPS	TRANS
BEER	CLTHS	FOOD	INSUR	OTHER	SMOKE	TXTLS
BLDMT	CNSTR	FUN	LABEQ	PAPER	SODA	UTIL
BOOKS	COAL	GOLD	MACH	PERSV	SOFTW	WHLSL

Table 2. Financial Ratio

Accruals/ Average Assets	Capitalization Ratio	Total Debt/ Equity	Dividend Payout Ratio	Interest/Average Total Debt	Net Profit Margin	Price/Operating Earnings (Basic, Excl. EI)	Price/Sales	Return on Capital Employed
Avertising Expenses/Sales	Cash Conversion Cycle (Days)	Total Debt/Total Assets	Effective Tax Rate	After-tax Interest Coverage	Operating CF/ Current Liabilities	Price/Operating Earnings (Diluted, Excl. EI)	Price/Book	Return on Equity
After-tax Return on Average Common Equity	Cash Flow/Total Debt	Total Debt/Total Assets	Common Equity/Invested Capital	Interest Coverage Ratio	Operating Profit Margin After Depreciation	Forward P/E to 1-year Growth (PEG) ratio	Pre-tax Profit Margin	Sales/ Stockholders Equity
After-tax Return on Total Stockholders Equity	Cash Balance/ Total Liabilities	Total Debt/ Capital	Enterprise Value Multiple	Inventory Turnover	Operating Profit Margin Before Depreciation	Forward P/E to Long-term Growth (PEG) ratio	Quick Ratio (Acid Test)	Sales/Invested Capital
After-tax Return on Invested Capital	Cash Ratio	Total Debt/ EBITDA	Free Cash Flow/ Operating Cash Flow	Inventory/ Current Assets	Payables Turnover	Trailing P/E to Growth (PEG) ratio	Research and Development/ Sales	Sales/Working Capital
Asset Turnover	Cash Flow Margin	Long-term Debt/Invested Capital	Gross Profit Margin	Long-term Debt/Total Liabilities	Price/Cash flow	Pre-tax Return on Total Earning Assets	Receivables/ Current Assets	Short-Term Debt/Total Debt
Book/Market	Current Liabilities/Total Liabilities	Dividend Yield	Gross Profit/ Total Assets	Total Liabilities/ Total Tangible Assets	P/E (Diluted, Excl. EI)	Pre-tax return on Net Operating Assets	Receivables Turnover	Labor Expenses/Sales
Shillers Cyclically Adjusted P/E Ratio	Current Ratio	Long-term Debt/Book Equity	Interest/Average Long-term Debt	The number of companies that belongs to the industry classification	P/E (Diluted, Incl. EI)	Profit Before Depreciation/ Current Liabilities	Return on Assets	Total Debt/ Invested Capital

Model

1. LASSO

In order to find the most fittable variables to predict return, I run LASSO model with window size of 80 on each industry's time series return. For each training sample, a LASSO model selects several variables for return prediction. The whole monthly dataset requires 509 times of resampling, therefore generating 509 sets of best variables for the purpose of return prediction. After 509 times of LASSO selection, I can obtain the frequency of variable which are selected as best predicting factors, which indicates the importance of each accounting characteristic in predicting returns for an industry. For 49 industries and 72 financial mean ratios on industry level, the frequency is shown below (Figure 1) .

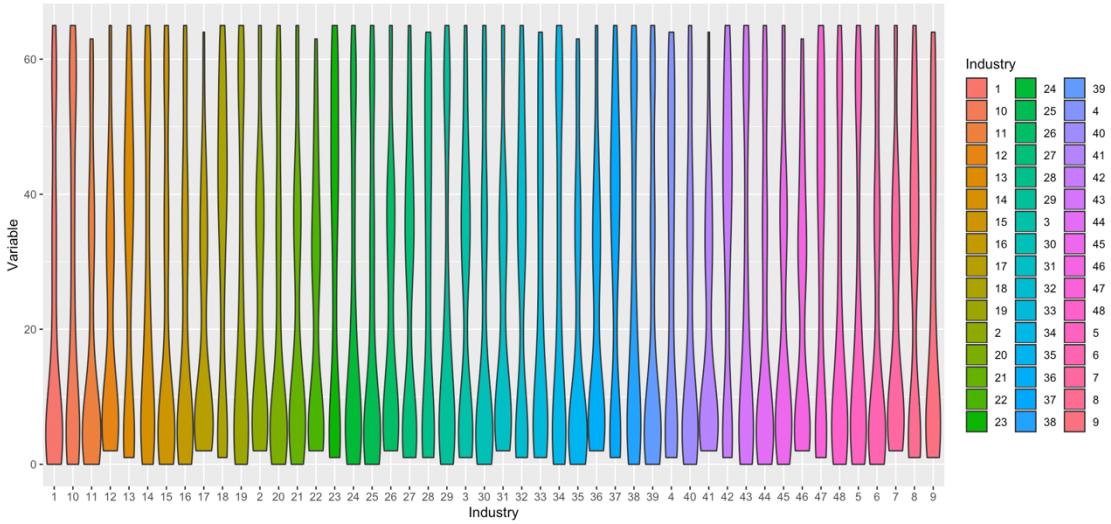

Figure 1.Ratio selected frequency

For each industry on x-axis, the frequency of ratios (denoted as variable) used are recorded as the width of the violin figure. The wider the violin, the larger frequency it is used as the best predictor, the better effect it has on predicting returns. We can observe from the figure that for all the industries, the first 15 characteristics are the most important ones. The difference comes from the rest 40 ones, where we can see the widths and narrowness location are not the same.

Using LASSO model on each industry for accounting characteristics selection, the first main take-away is that I can find out how different industries rely on different characteristics.

Table 3: Examples of important accounting ratios for single industries

Industry	Characteristic I	Characteristic II	Chacrateristic III
AERO	Return on Capital Employed	Dividend Yield	Shillers Cyclically Adjusted P/E Ratio
BANKS	Dividend Yield	Labor Expenses Sale	Debt/Ebitda
FOOD	Dividend Yield	Capitalization Ratio	Pre-tax Return on Net Operating Assets
OIL	Shillers Cyclically Adjusted P/E Ratio	Inventory/Current Asset	Payables Turnover

For a glance of these four industries, we observe that dividend yield, Shiller P/E ratio, Return on different perspectives and liquidity are all significant factors for return prediction. It is economically intuitive that, capital employed is more important in AEROPLANE industry than in other industries. For BANKS, employee expenses versus their sales revenue is a large part of banks' income, and therefore matters a lot. For OIL industry, inventory are significant for its normal operation and payables turnover, as a liquidity indicator, shows the healthiness of the industry.

On the other hand, from the perspective of each accounting ratio, we can also observe the frequency of it being selected, which indicates its importance in predicting an industry's return. The selection frequency of those financial ratio that are top five for each industry is shown below using block plot in Figure 2.

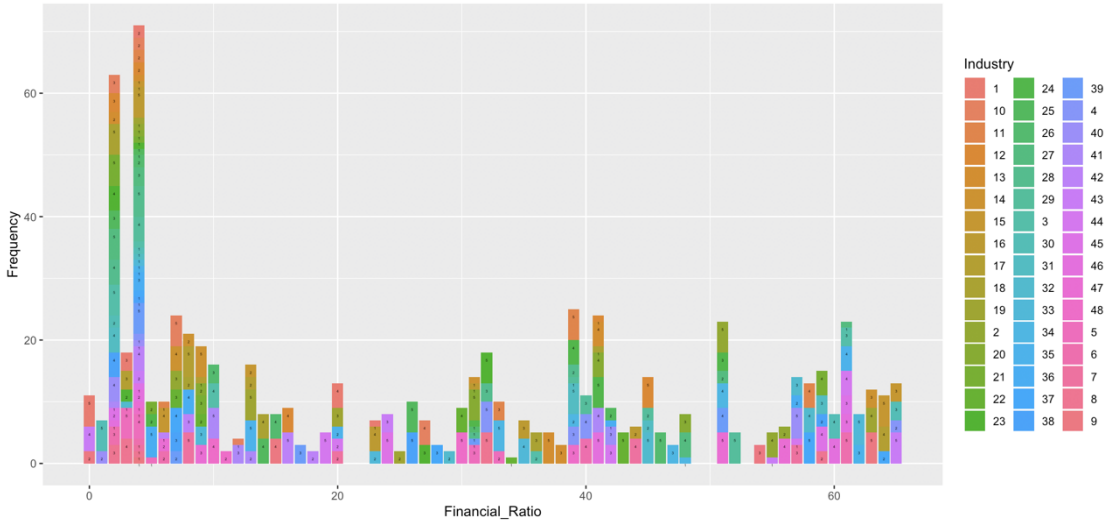

Figure 2: Financial Ratio's selected frequency

From the figure, in comparison, we can observe that some ratios are extremely frequently selected and some are rarely selected. Out of 509 times of LASSO selection for each industry and I add up all industries, the top five selected variables are Book/Market, Dividend Yield Mean, Debt Ebitda Mean, Int Debt Mean, PE Exi Mean. Intuitively, for most companies, these accounting characteristics reflect the healthy of a firm or an industry's profitability, operation healthiness, ability to reward shareholders or pay debt from different perspectives. Although for some particular industries, there are other characteristics more important, such as Staff Sale, Pay Turn and etc, on industry level, the basics of accounting characteristics are still the most valuable ones for return prediction.

After LASSO selection, I take the combination of these 49 sets of best accounting characteristics as a new pool for next step's cross-sectional regression.

2. Elastic Net and Cross-validation

In the next step, I try to use cross-sectional regression to create the MVE portfolio based on the variable set I selected. Since there are 72 financial ratio ready to use in my high dimensional data set, I need to adopt the proper method to reduce the dimension. To select the best set of my variables, I choose Elastic Net technique (alpha=0.5) in machine learning on my cross-sectional regression based on the research of Nagel et al ("Shrinking the Cross-Section"). Since in the asset-pricing setting, elastic net methods is known to have better performance compared with Ridge and Lasso when my variables are highly correlated which is shown in the Figure 3.

Since I have a large set of variables, to avoid overfitting problem, I apply cross validation method with 5 folds to select the best lambda for the Elastic Net method in my regression. The training data ranges from 1970 to 2004. I divided the training data in to 5 subset and apply cross validation technique on it. The best lambda is selected from cross validation results which leads to the minimum MSE under the Elastic net style regression. Then I run the Elastic net style regression on the whole training data with the best lambda and get the b_vector of the financial ratio. With the final b_vector in hand, I calculate the out-of-sample estimated MVE portfolio with return

b_vector*Factors_i,t in the period 2005-2019. The Factors_i,t is constructed by the mean Financial Ratio of each industry and their corresponding excess returns.

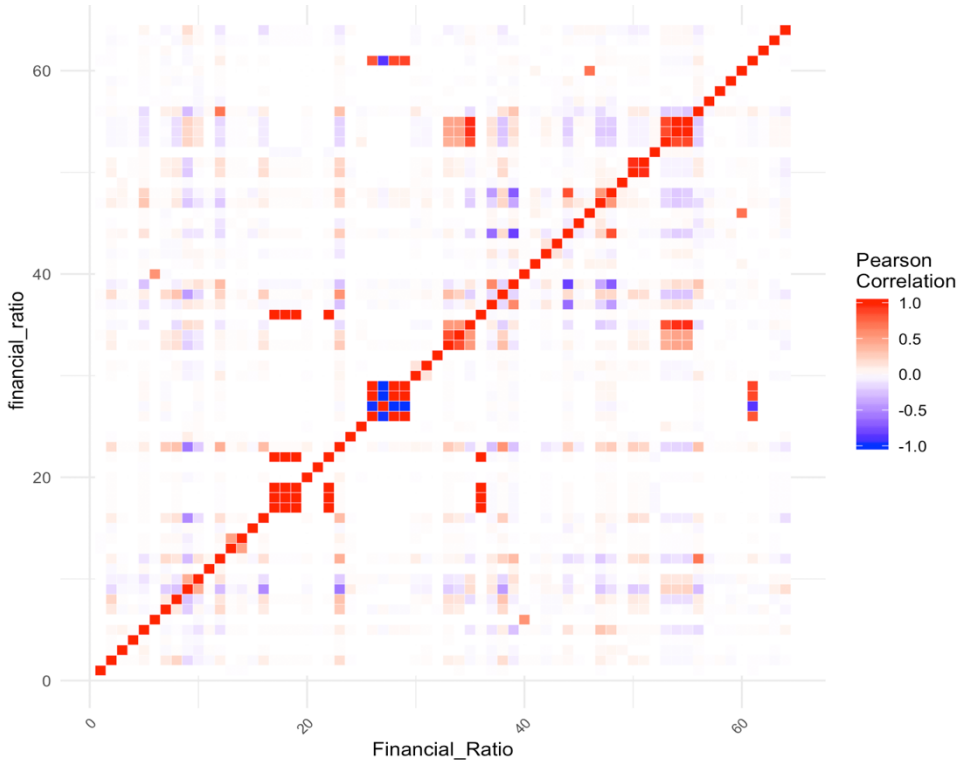

Figure 3. Correlation Heatmap of Financial Ratio

Performance

1. LASSO variable selection performance

In the phase of initial variable selection, I use LASSO and train data set of window size 80 to select best predictors. After each training, I use the predictor and coefficients to do out-of-sample testing. The cumulative sum of squared error for each industry are shown as below (Figure 4). I need to compare it with OLS model to see the performance. The SSE of a full model of OLS with all accounting ratios as predictors are shown in Figure 5. By comparison, we can easily observe that LASSO significantly reduce the SSE of out of sample tests.

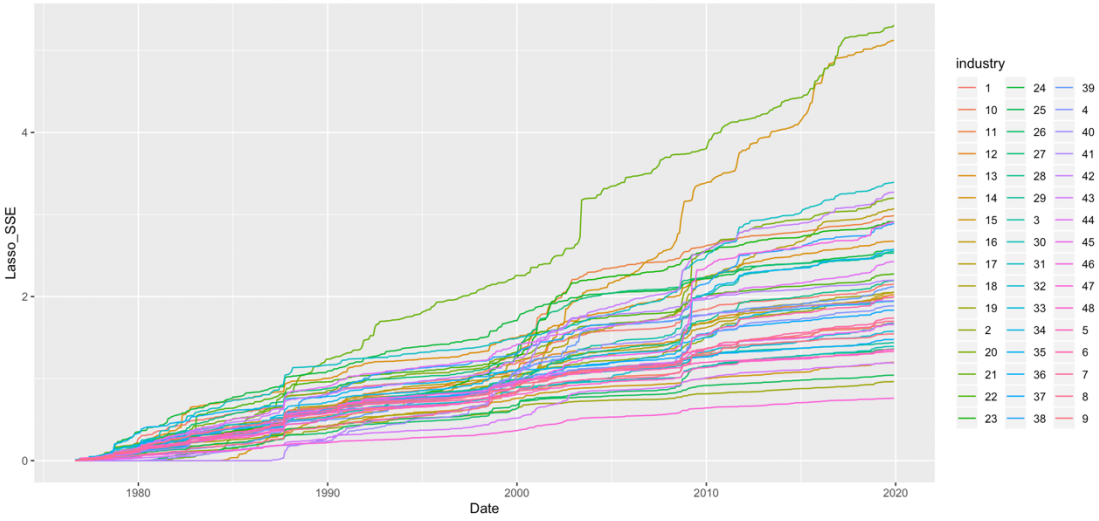

Figure 4: SSE of LASSO model for all industries

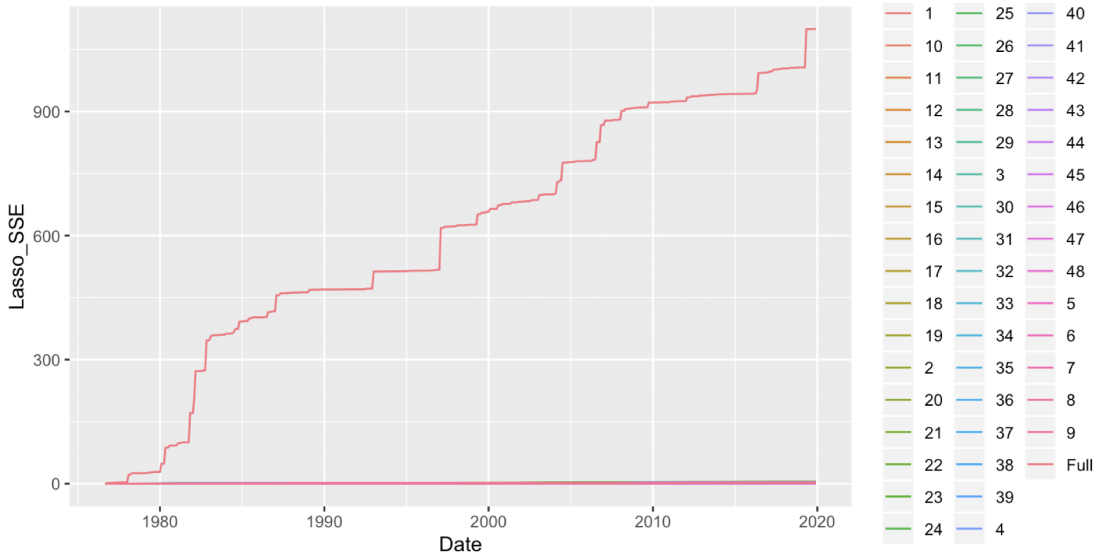

Figure 5: SSE of OLS and LASSO model

Therefore, I can conclude that LASSO model is efficient for variables selection compared to full OLS model.

2. MVE portfolio monthly frequency trading strategy performance

After the initial selection in the previous section, I have 34 selected variables shown in Table 4. I will use these selected variables to run the cross sectional regression. The final set of the variables will be selected by the elastic net method under the best lambda. There are only 7 variables shown in the Table 5. They are Capitalization Ratio, ROE, ROA, Common Equity / Invested Capital, Operating Profit Margin after Depreciation, Long term debt / Invested Capital, Total debt / Invested Capital. Capitalization ratios are indicators that measure the proportion of debt in a company's capital structure. They are among the more meaningful debt ratios used to assess a company's financial health. Return on equity (ROE) is a measure of financial performance. It is considered a measure of how effectively management is using a company's assets to create profits. Return on assets (ROA) is an indicator of how profitable a company is relative to its total assets. ROA gives a manager, investor, or analyst an idea as to how efficient a company's management is at using its assets to generate earnings. The operating margin measures how much profit a company makes on a dollar of sales, after paying for variable costs of production, such as wages and raw materials, but before paying interest or tax. The debt to capitalization ratio shows the financial leverage of a firm. Investors compare the financial leverage of firms to analyze the associated investment risk. High ratios indicate riskier investments, as debt is the primary source of financing and introduces a greater risk of insolvency. The selected financial ratios are quite consistent with the research of S. Kheradyar et al(2011) on the Stock Return Predictability with Financial Ratios. I also run the elastic net style regression on the whole variable set for comparison. There are 10 variables selected without the first selection. The result is shown in the Table 6.

Table 4. Financial Ratio after initial selection

Accruals/ Average Assets	Capitalization Ratio	Total Debt/ Equity	Dividend Payout Ratio	Interest/Average Total Debt	Net Profit Margin	Price/Operating Earnings (Basic, Excl. EI)	Price/Sales	Return on Capital Employed
Avertising Expenses/Sales	Cash Conversion Cycle (Days)	Total Debt/Total Assets	Effective Tax Rate	After-tax Interest Coverage	Operating CF/ Current Liabilities	Price/Operating Earnings (Diluted, Excl. EI)	Price/Book	Return on Equity
After-tax Return on Average Common Equity	Cash Flow/Total Debt	Total Debt/Total Assets	Common Equity/Invested Capital	Interest Coverage Ratio	Operating Profit Margin After Depreciation	Forward P/E to 1-year Growth (PEG) ratio	Pre-tax Profit Margin	Sales/ Stockholders Equity
After-tax Return on Total Stockholders Equity	Cash Balance/ Total Liabilities	Total Debt/ Capital	Enterprise Value Multiple	Inventory Turnover	Operating Profit Margin Before Depreciation	Forward P/E to Long-term Growth (PEG) ratio	Quick Ratio (Acid Test)	Sales/Invested Capital
After-tax Return on Invested Capital	Cash Ratio	Total Debt/ EBITDA	Free Cash Flow/ Operating Cash Flow	Inventory/ Current Assets	Payables Turnover	Trailing P/E to Growth (PEG) ratio	Research and Development/ Sales	Sales/Working Capital
Asset Turnover	Cash Flow Margin	Long-term Debt/Invested Capital	Gross Profit Margin	Long-term Debt/Total Liabilities	Price/Cash flow	Pre-tax Return on Total Earning Assets	Receivables/ Current Assets	Short-Term Debt/Total Debt
Book/Market	Current Liabilities/Total Liabilities	Dividend Yield	Gross Profit/ Total Assets	Total Liabilities/ Total Tangible Assets	P/E (Diluted, Excl. EI)	Pre-tax return on Net Operating Assets	Receivables Turnover	Labor Expenses/Sales
Shillers Cyclically Adjusted P/E Ratio	Current Ratio	Long-term Debt/Book Equity	Interest/Average Long-term Debt	The number of companies that belongs to the industry classification	P/E (Diluted, Incl. EI)	Profit Before Depreciation/ Current Liabilities	Return on Assets	Total Debt/ Invested Capital

Table 5. Financial Ratio after final selection

Accruals/ Average Assets	Capitalization Ratio	Total Debt/ Equity	Dividend Payout Ratio	Interest/Average Total Debt	Net Profit Margin	Price/Operating Earnings (Basic, Excl. EI)	Price/Sales	Return on Capital Employed
Avertising Expenses/Sales	Cash Conversion Cycle (Days)	Total Debt/Total Assets	Effective Tax Rate	After-tax Interest Coverage	Operating CF/ Current Liabilities	Price/Operating Earnings (Diluted, Excl. EI)	Price/Book	Return on Equity
After-tax Return on Average Common Equity	Cash Flow/Total Debt	Total Debt/Total Assets	Common Equity/Invested Capital	Interest Coverage Ratio	Operating Profit Margin After Depreciation	Forward P/E to 1-year Growth (PEG) ratio	Pre-tax Profit Margin	Sales/ Stockholders Equity
After-tax Return on Total Stockholders Equity	Cash Balance/ Total Liabilities	Total Debt/ Capital	Enterprise Value Multiple	Inventory Turnover	Operating Profit Margin Before Depreciation	Forward P/E to Long-term Growth (PEG) ratio	Quick Ratio (Acid Test)	Sales/Invested Capital
After-tax Return on Invested Capital	Cash Ratio	Total Debt/ EBITDA	Free Cash Flow/ Operating Cash Flow	Inventory/ Current Assets	Payables Turnover	Trailing P/E to Growth (PEG) ratio	Research and Development/ Sales	Sales/Working Capital
Asset Turnover	Cash Flow Margin	Long-term Debt/Invested Capital	Gross Profit Margin	Long-term Debt/Total Liabilities	Price/Cash flow	Pre-tax Return on Total Earning Assets	Receivables/ Current Assets	Short-Term Debt/Total Debt
Book/Market	Current Liabilities/Total Liabilities	Dividend Yield	Gross Profit/ Total Assets	Total Liabilities/ Total Tangible Assets	P/E (Diluted, Excl. EI)	Pre-tax return on Net Operating Assets	Receivables Turnover	Labor Expenses/Sales
Shillers Cyclically Adjusted P/E Ratio	Current Ratio	Long-term Debt/Book Equity	Interest/Average Long-term Debt	The number of companies that belongs to the industry classification	P/E (Diluted, Incl. EI)	Profit Before Depreciation/ Current Liabilities	Return on Assets	Total Debt/ Invested Capital

As we can see, most of the variables in the selection are related to debt. By comparing these two tables, we find that the result of variables' selection are very different. The diversification of the variables increased a lot if I adopt the first step selection in the previous selection. Moreover, the sign of some common variables in these two table becomes opposite such as Capitalization Ratio, Long-term Debt/Invested Capital, Total Debt/Invested Capital and Common Equity/Invested Capital. The sign of the variables decide the action I need to take when the mean financial ratio of one industry increase or decrease a lot. For examples, I need to long more when the Capitalization Ratio or the ROA ratio of a industry increased a lot. I need to short more when the Operating Profit Margin After Depreciation increase a lot. By diversifying the set of the selected financial ratio, I will estimate the performance from a more comprehensive perspective to make a better strategy.

Table 6. Coefficients after Initial Selection

	Return on Equity	Capitalization Ratio	Long-term Debt/Invested Capital	Common Equity/Invested Capital	Return on Assets	Operating Profit Margin After Depreciation	Total Debt/ Invested Capital	Constant
Coefficient	0.046718968	0.231734363	0.008080940	-0.001349481	0.643971644	-0.792903656	1.227548899	-0.010833226

Table 7. Coefficients without Initial Selection

	Long-term Debt/Invested Capital	Capitalization Ratio	Interest/Average Total Debt	Common Equity/Invested Capital	Short term Debt/Total Debt	Long term Debt/ Total debt	Total Debt/ Invested Capital	Total Debt/ Equity	Total Debt/ Capital	Total Debt/Total Assets	Constant
Coefficient	-0.578114525	-0.444544681	0.007381904	0.578750282	0.503364682	-0.221592360	-0.165759918	0.003487316	0.320065477	-2.147059886	-0.010833226

Table 8 shows the statistics of my MVE Portfolio. I also construct the value weighted industry portfolio for comparison. The weight of each industry portfolio is calculated by the total market value of each industry. Weight_i = mean market value at industry level*number of firms in each industry / total market value of 49 industries. From the result we can see that the mean of the return of three portfolios are all very low. It is reasonable because industry portfolio is not as volatile as the single stock. The lower risk leads to the lower expected return. I use a bunch of steady but low-yielding portfolios to make long and short strategy. However, the sharpe ratio is not disappointing, which suggests the return of the investment compared to its risk is pretty good. From the results below, we also find that the MVE portfolio without the initial selection is not improved significantly. After the first-step selection, the MVE portfolio is improved a lot. The sharpe ratio increased 27.27%. The cumulative returns during 2005-2019 of three portfolios are displayed in Figure 6. From the graph we can see that there is no obvious strength of my MVE portfolio before 2015. However, after 2015 the cumulative returns increases dramatically especially during 2019. During the recession time in 2008, the value weighted portfolio dropped a lot, however, my portfolio performed good and the MVE portfolio without initial selection even performed better. This is because in the MVE portfolio without initial selection, there are more variables related to debt. As what we discussed in the previous section, the financial ratio that include debt compares the financial leverage of firms to analyze the associated investment risk as debt is the primary source of financing and introduces a greater risk of insolvency. They also reveal whether or not it has loans and how its credit financing compares to its assets. Hence the MVE portfolio without initial selection performs better in the recession time.

Table 8. Excess Return Statistics

	Mean	**Standard Deviaiton**	**Sharpe Ratio**
Value-weighted Industry portfolio	0.00756853	0.01690906	0.447602054756444
MVE portfolio without initial selection	0.007447327	0.016059868	0.463722802702986
MVE portfolio after initial selection	0.00732455	0.01285738	0.569676714851704

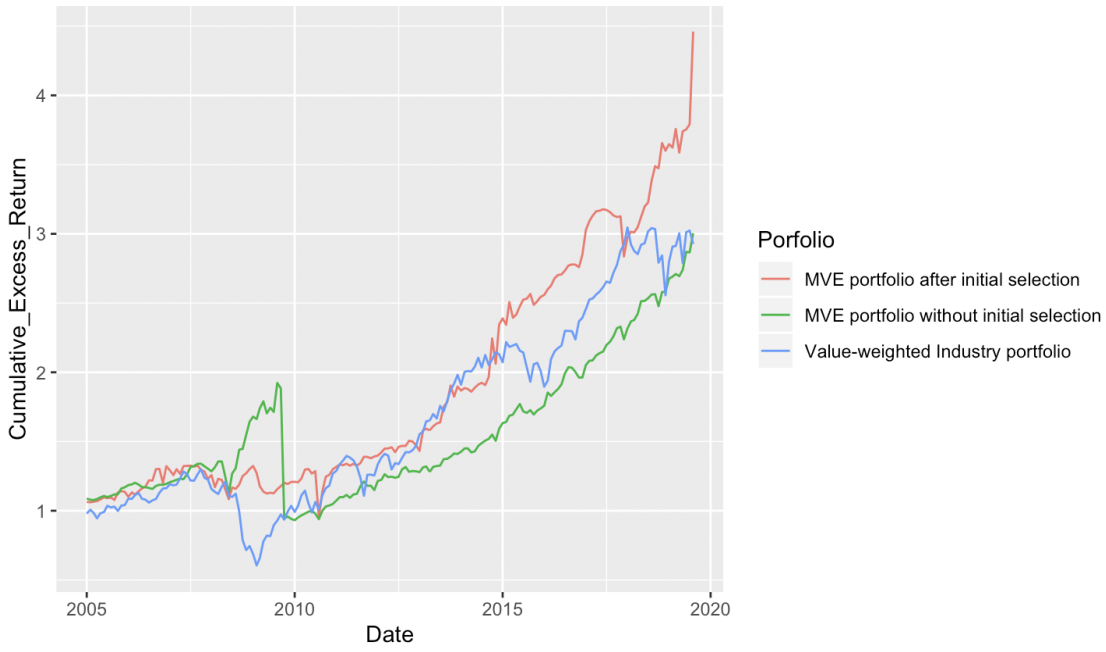

Figure 6. Performance of the MVE Portfolio

Conclusion

1. Summary

In initial selection using LASSO, I obtain a combination of 34 financial ratios from each industry's top five fittable predictors. The LASSO selection not only builds the basis for my next step's cross-sectional selection and trading strategy, it also provides some economic hints that how each industry depends on different financial ratios because of their different nature. I can use such difference to explore deeper on particular industries for more precise return prediction.

Using the selected 34 financial ratios, I construct the MVE portfolio by run the cross sectional regression. To reduce the dimension of my data in the further step, I apply elastic net and cross-validation with 5 folds methods. In the final selection, there are 7 financial ratios which suggested significant prediction power on the performance of the industry excess return. my industry investment strategy is based on these 7 financial ratios. The weight of each industry is decided by the performance of mean financial ratio at industry level of each month. I long more when their performance is good if the sign of the coefficient of the financial ratio is positive. The MVE portfolio performed significantly well in the recent 4 years. During the recession period, the MVE portfolio wasn't as volatile as the Market value weighted portfolio. The cumulative return still went up in 2008. In other words, the strategy may give us some inspirations of risk reduction during the recession time and the selection of industry when I want to make investment.

2. Improvement

LASSO's computation is very time-consuming and space costed, especially when I need to run LASSO over five hundred times per industry and for 49 industries. I may need to improve on the machine learning technique to reduce the cost while on the same time remain the model's

accuracy. A possible solution is to do another initial selection using a simpler non-machine-learning approach, such as economic examination, to first exclude some irrelevant or similar accounting characteristics.

In the forward step, I may try other method to select the variables such as XGBoost technique. More variables at industry level beyond financial ratio such as the wages will be contributed to my model. In the other hand, I can also improve my strategy by ranking analysis. Since there are 49 industry in my MVE portfolio, it is hard for implementation. By ranking the predict returns, I can only long the industries in the top quantile and short the industries in the last quantile which may also give us a good result. Another way to improve the model is to vary the train and test set. I may create a rolling window. The data in the rolling window will be used to train the model. The b_vector acquired will be contribute to the construction of the MVE portfolio for the next period.