ORIE4741 Project Final Report

Bojun Li (bl755), Marshall Guan (jg2262), Rui Dai (rd576)

1 Introduction

One of the toughest challenges in investment management is to identify which aspects of a public company's fundamentals affect its stock price. Even for a seasoned stock analyst, distilling information from large swaths of operating performance data and news remains difficult. Furthermore, financial data embody the concepts of "big messy data" because of its notoriously noisy nature and high-dimensionality. We believe that by using computer science, statistics, and big data analytics, we could gain profound insights into the complex financial data environment, and identify correlations between operating and stock performances.

Specifically, we are interested in the following problems:

How fundamentals of a publicly traded company affect its stock price?

In practice, modeling techniques such as Discounted Cash Flow or Ratio Analysis were widely applied to forecast and evaluate stock price based on companies' fundamental data. In this project, we would explore predictive power of machine learning techniques and compare their effectiveness against DCF and Ratio analysis.

Can we build a profitable trading strategy based on the fundamental data of stocks?

If fundamentals data have predictive power of stock prices, it may be possible to find a trading strategy that allow us to profit from the stocks. Financial models can be categorized into theoretical models and statistical models, and both approaches worth exploration.

To address the above problems, we tried different models to analyze the data, including linear regression, lasso regression, ridge regression, logistic regression and multilayer perceptron, and combined them with financial theories to build two different portfolios to profit from the stocks.

2 Dataset Description, Data Manipulation, and EDA

Link: https://www.kaggle.com/dgawlik/nyse
In this project, 3 datasets are used that contains information of listed stocks in New York Stock Exchange (NYSE): the fundamentals dataset, the securites dataset, and the prices dataset.
The dataset is downloaded from Kaggle and the original sources are: Yahoo Finance (stock prices), Nasdaq Financials (fundamentals) and extended by some fields from EDGAR SEC databases.

2.1 Dataset Description

Fundamentals Dataset

The fundamentals dataset contains 77 explanatory variables which represent the fundamentals of 501 public listed stocks in NYSE according to their SEC 10K annual fillings. The time frame of this dataset is from 2012 to 2016. The number of observations is 1781.

Securities Dataset

The securities dataset contains 5 columns that gives general descriptions about the companies, but only the industry sectors information is used. The number of observations is 505, however, we will exclude the stocks which do not appear in the fundamentals dataset and prices dataset.

Prices Dataset

The prices dataset contains daily historical prices data of each stock from 2010 to 2016, including open, close, low, high and volume. The number of observations is 851,264. In this project, the closing prices are used to represent stock prices.

2.2 Initial Data Manipulation

1. Missing value: As shown in Figure 1, there are 6 features with missing values in the fundamentals dataset. Since the missing proportion is low and the features with missing values are not important from the financial modeling aspect, we can simply remove these features.

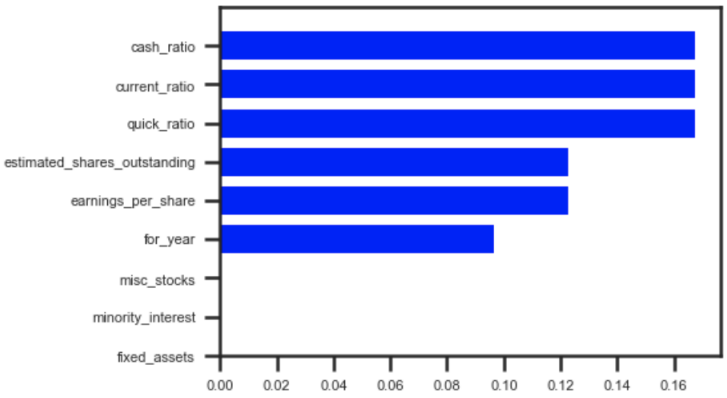

Figure 1: Features with missing values and missing proportion

2. Prices to return: It is difficult to compare prices between stocks, so so we are interested in their log returns:

$$r_t = \ln \frac{S_t}{S_{t-1}}$$

where S_t denotes the stock price of period t.
This transformation is under the commonly used assumption in the financial industry that the stock prices are log-normally distributed.

3. Market Capitalization Computation:
Since the value of stocks vary according to the shares of stocks outstanding and the size of the company, we need to develop a consistent measurement that factors in the quantity and the magnitude of the organization.

We assumed that a company's fundamental information has predictive power or causal effect over the next month following the date when the fundamental information was captured. Then we calculated the mean market capitalization next month as:

$$m_t = \frac{\sum_{\text{all business days } d \text{ next month}} S_d \times \text{number of shares outstanding}}{\text{number of business days next month}}$$

2.3 Exploratory Data Analysis

1. Industry sector breakdown: Figure 2 gives a visualization of the stock industry sector breakdown in our data.

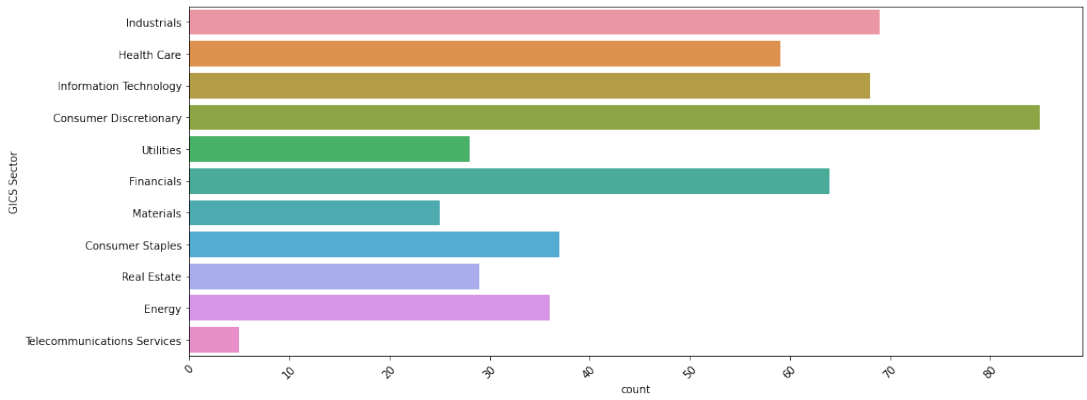

Figure 2: Industry Sector Breakdown

2. Fundamentals features correlations and importance by sectors:

   Public company in different sectors tend to have distinct revenue model, capital structure, and valuation schemes. As a result, we would decided to visualize the fundamental features and see if they vary by sectors.

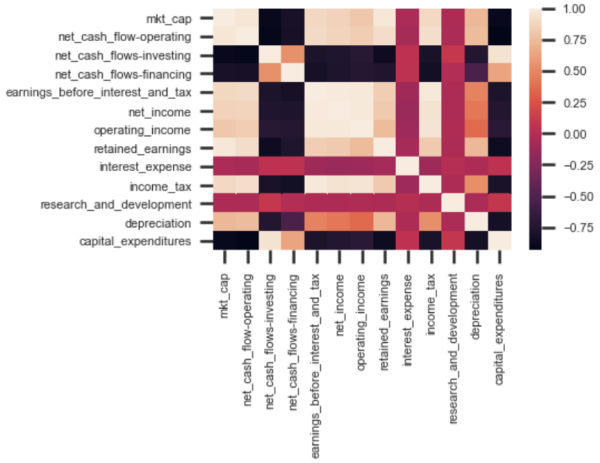

Figure 3: Correlation of Energy Industry

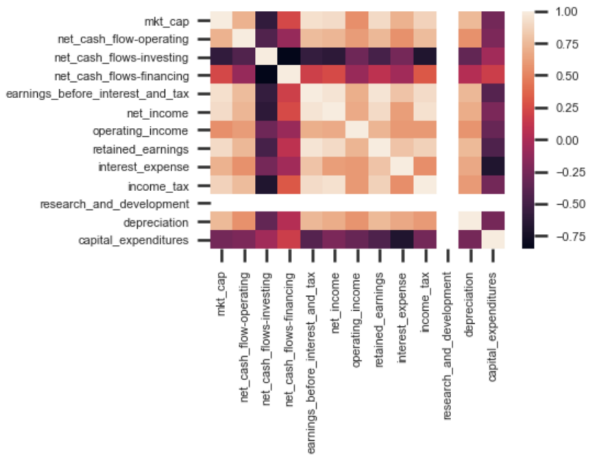

Figure 4: Correlation of Financial Industry

The first pair of sectors we would like to explore are traditional sectors, Energy and Financial. From the correlation plot, we can observe many differences. For example, net cash flow from operating is important to valuing company market capitalization in the Energy Industry, while its correlation with market cap is halved for companies in the Financial industry. At the same time, cash used for financing played an much more important role in bringing down Energy company valuation than for Financial company. These differences reflect the fact that Energy company rely heavily on making huge capital investment. As a result, the company's ability and need to finance, as reflected by the cash flow for financing, is more important to energy company than to Financial companies.

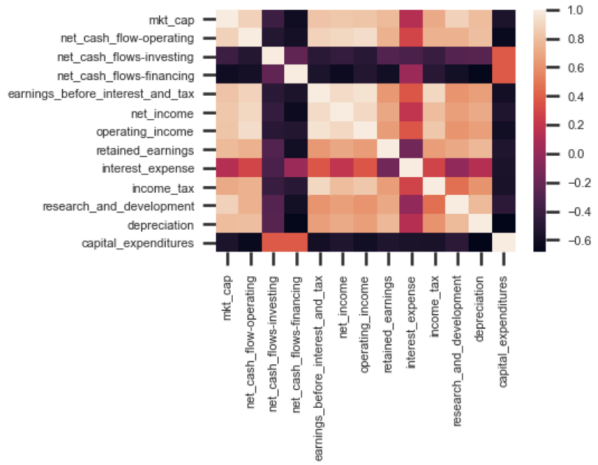

Figure 5: Correlation of Healthcare Industry

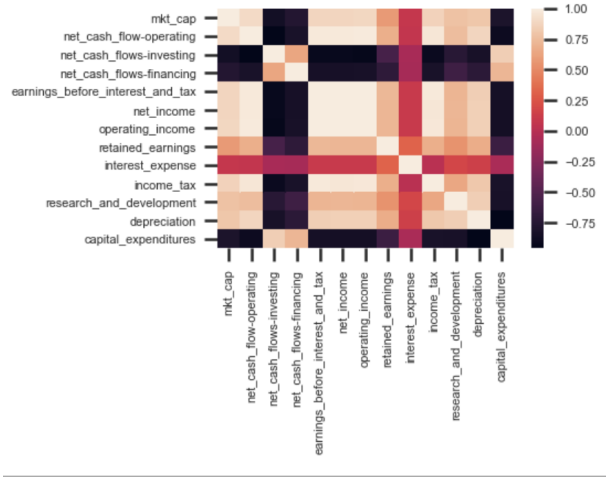

Figure 6: Correlation of IT Industry

Similar differences persist for growth sectors such as Healthcare and IT. For instance, earnings is more important for Healthcare companies as they need the cash to invest in new medical product research. While IT companies have less earning pressure given they are financed by venture capital or investors who do not press for short-term liquidity.

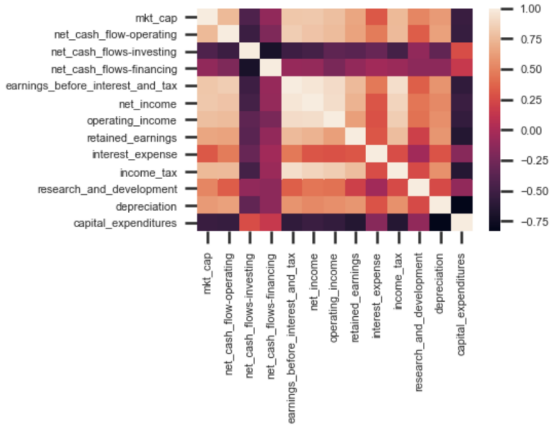

Figure 7: Correlation of All Industries

Eventually, when we look at features for companies in all sectors, capital expenditure would always brought down the valuation.

The following sections will discuss both the theoretical model and statistical model of analyzing the data and illustrate our portfolio strategies.

3 Portfolio 1: Value Investing - Linear Models with Fundamentals

3.1 Model Setup

Here we sought optimization of portfolio by finding the optimal equal-weight 50-stock portfolios that maximize the ex-post Sharpe Ratios. We assign each selected stock with weight 2%. To predict which stock should be selected, we also engineered features containing returns, volatility, beta, as well as fundamental data including key financial ratios and notably, **Piotroski F-score**. Calculation steps shown below:

Profitability

$$\begin{aligned}\text{Score}_1 = &\mathbf{1}(\text{Return on Assets} > 0) + \mathbf{1}(\text{Operating Cash Flow} > 0) \\ &+ \mathbf{1}(\text{Change in Return on Assets} > 0) + \mathbf{1}\left(\frac{\text{Operating Cash Flow}}{\text{Total Assets}} > \text{Return on Assets}\right)\end{aligned}$$

Leverage, Liquidity and Source of Funds

$$\begin{aligned}\text{Score}_2 = &\mathbf{1}(\text{Change in Leverage (long-term) ratio} < 0) + \mathbf{1}(\text{Change in Current ratio} > 0) \\ &+ \mathbf{1}(\text{Change in the number of shares} = 0)\end{aligned}$$

Operating Efficiency

$$\text{Score}_3 = \mathbf{1}(\text{Change in Gross Margin} > 0) + \mathbf{1}(\text{Change in Asset Turnover ratio} > 0)$$

Based on the above 9 criteria,

$$\text{Piotroski F-score} = \text{Score}_1 + \text{Score}_2 + \text{Score}_3$$

The score is an integer number between 0 and 9. However, to further explore how fundamentals of a company affect the portfolio construction, we created a new feature for each of the 9 scores.

In terms of linear models, we tried out the logistic regression model (selecting top 50 stocks when ranked by log likelihood values) and multilayer perceptron model (similarly, selecting 50 stocks with highest predicted probabilities).

3.2 Model Calibration

One challenge in calibrating the models is how to assign class labels to the dataset. Since constructing equal weight portfolios that maximize Sharpe Ratios involves integer programming, which is NP-hard, here we chose to approximate the result using the usual quadratic programming. The simulated result below shows our approach is sufficiently robust:

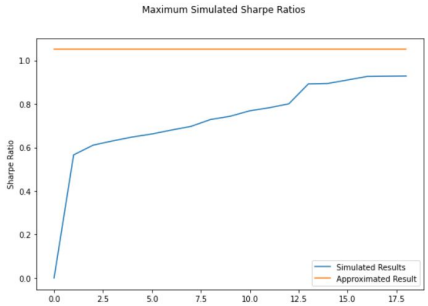

Figure 8: Quadratic Programming Approximated Portfolio VS Simulated Portfolios

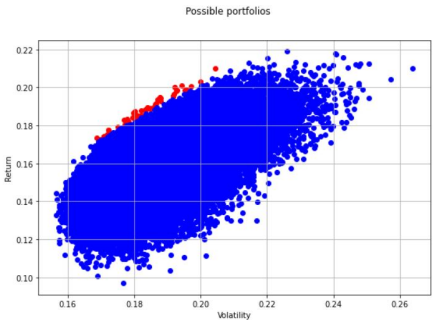

Figure 9: Red Dots are portfolios with maximum Sharpe Ratios

Overfitting is a common issue that often arises in models like neural networks. To minimize the impact of this problem, we chose to use the **5-Fold Cross Validation** method to select the proper number of neurons in each hidden layer, which control the model complexity and hence how likely the model is to overfit. Below shows a 3-D plot of the resulting Sharpe Ratios with respect to different model configurations.

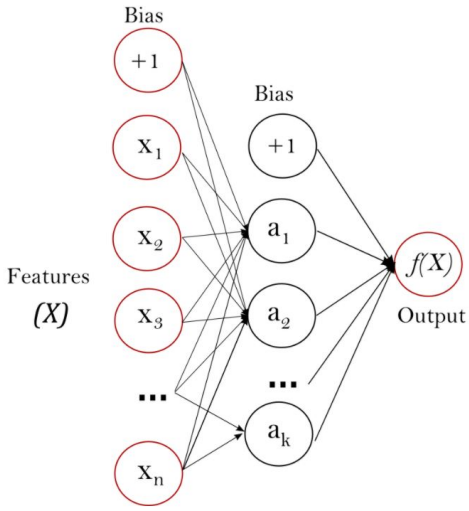

Figure 10: A Multilayer Perceptron Model with One Hidden Layer

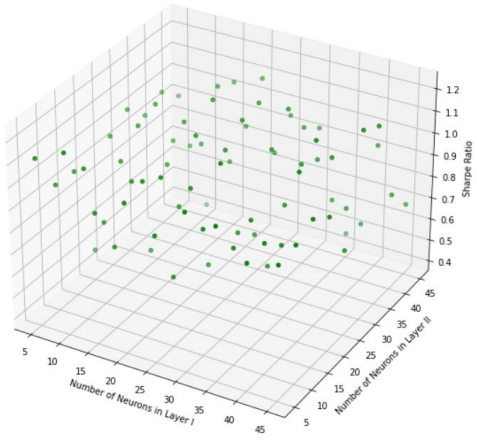

Figure 11: Results with different Hidden Layer Configurations

3.3 Result

To see how our models perform, we plot the backtested strategies' performances against the equal-weight (ie. money is evenly allocation to each stock) benchmark portfolio. The strategies are tested over the year of 2016 and as shown below, both strategies outperform the benchmark in terms of both absolute returns and risk-adjusted returns.

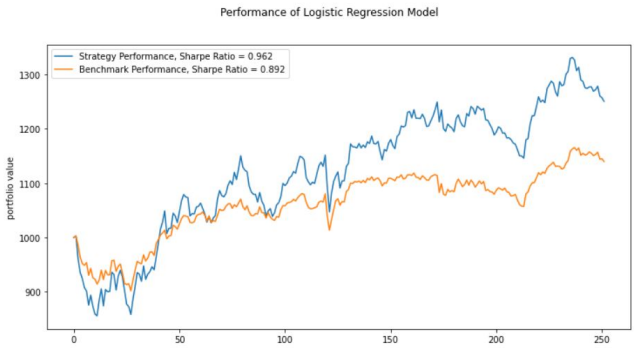

Figure 12: Backtesting Result of Logistic Regression Model

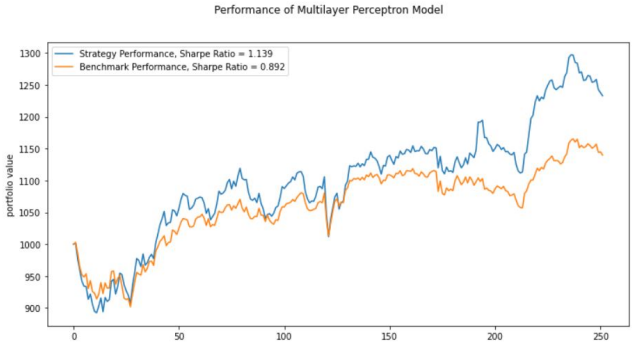

Figure 13: Backtesting Result of Multilayer Perceptron Model

4 Portfolio 2: Statistical Model with Historical Fundamentals

4.1 Model

We have already explored how to use economic theory behind company's financial figures to produce stock valuations. Here, we would also like to pursue a more statistical based techniques to directly use historical fundamental data to perform stock valuations without prior knowledge about finance.

Train test split: Since our data set span across the period between 2010 and 2016, we realized during this period the U.S. market was generally prosperous and expanding. Therefore, we assumed the fundamental and stock valuation would relate in the same way across this period. And we would use data from before

2016 to train the models while using data from 2016 to test the models.

Performance evaluation We decided to use R^2 as the performance evaluation matrix for our model. This is because in stock investment and valuation, variance and the volatility measurement it implied were vital. R^2 could describe the percentage of variance our model would be able to explain. In addition, the MSE measurement would run into unit problem for company of different sizes. The MSE measurement could easily run into 10^20 in magnitude due to the fact that most companies were valued in billions, and that makes R^2 more preferred over MSE.

We first performed a series of regression models over the whole data set. The weights and R^2 score for each measurement was reported in the figure below.

4.1.1 Linear Regression

The ordinary linear regression model achieved Training: $R^2 = 68.01\%$ and Testing $R^2 = 66.69\%$. The weights of the model reflected important fact that net income and R&D spending were positive signs for the growth potential of company.

4.1.2 Lasso Regression

Since our models include many features, we also applied Lasso regression. To our surprise, the result was identical in terms of weight and R^2. The result suggests sparsity might be the primary problem in our model.

4.1.3 Ridge Regression

To control for the situation our model had over-fitted, we performed Ridge regression with $\alpha = 0.05$. The results differed greatly from the that of the ordinary regression and the Lasso regression. The result flagged highly likely risks that our model over-fitted. Furthermore, given the significant decrease in R^2 score, the potential over-fitting problem could not be resolved by simply performing Ridge regression.

Result From Regressions			
Feature	**Linear Regression**	**Lasso Regression**	**Ridge Regression**
net_cash_flow-operating	-0.077	-0.077	0.610
net_cash_flows-investing	-0.374	-0.374	-0.217
net_cash_flows-financing	-0.515	-0.515	-0.161
earnings_before_interest_and_tax	2.278	2.278	1.131
net_income	4.072	4.072	2.092
operating_income	0.444	0.444	1.354
retained_earnings	0.040	0.040	0.142
interest_expense	2.160	2.160	2.328
income_tax	1.388	1.387	2.477
research_and_development	8.641	8.641	7.239
depreciation	4.160	4.160	2.180
capital_expenditures	1.100	1.100	-0.280
Test R-Squared	**66.69%**	**66.69%**	**58.31%**

Figure 14: Coefficients and R^2 from Regression on Full Data Set

4.2 Addressing Over-fitting by Meta Data Learning

Company's sectors would be an important Meta Data we need to consider, since we identified significant differences in feature distributions that were indicative of the distinct business models companies under different sector had. Therefore, we further addressed over-fitting by breakdown model training and testing by industry.

Regression results based on each sectors were displayed below. We noticed that the regression model had much more explanatory power for sector such as "Healthcare", "Consumer", "Financial", "Materials", "Energy", and "Telecom", while the model was less suitable for "Industrial" and "Utilities". Also, by applying Ridge Regression, the model gained more explanatory power in sector such as "IT", "Real Estate", and "Telecom".

By Sector Regression Outcomes						
Sectors	**Train R^2**	**Test R^2**	**L1 R^2 (α=0.05)**	**L2 R^2 (α=0.05)**	**Train Samples**	**Test Samples**
Industrials	35.78%	45.81%	45.67%	45.69%	157	62
Consumer Discretionary	74.88%	48.91%	48.91%	21.28%	190	69
Information Technology	82.81%	65.17%	65.34%	72.77%	158	58
Health Care	90.34%	90.78%	90.78%	89.74%	125	47
Consumer Staples	97.60%	90.54%	89.83%	87.29%	73	26
Utilities	-33.03%	14.74%	14.74%	13.67%	66	22
Financials	92.72%	93.33%	93.58%	95.07%	135	48
Real Estate	32.88%	63.79%	63.79%	71.97%	73	26
Materials	82.24%	70.31%	70.28%	66.43%	65	24
Energy	99.13%	91.58%	90.90%	86.81%	84	30
Telecommunications Services	99.97%	89.11%	89.75%	98.42%	15	5

Figure 15: Linear Regression Result by Sectors

5 Limits, Fairness, and Future Improvements Analysis

5.1 Limits

1. In 2010-2016, the economic was experiencing expansion, and the data may be biased in terms of the economics. As a result, our model may not able to capture market downturn.

2. Our model learns about the revenue/profit pattern of existing listed companies, but it is possible that the model does not work for companies or industry sectors with new business model such as Tesla or Pharmaceutical companies. Consequently, our model may miss opportunities for profitable investments in Growth Sectors.

5.2 Fairness and Weapon of Math Destruction Concern

Our models are built on data in 2010-2016 when economic was experiencing expansion, so it may be over optimistic about the profitability. As a result, investors using our models would be exposed to risk.

5.3 Future Improvement

1. Combination with Modern Portfolio Theory. Modern Portfolio Theory take risk(volatility) into consideration when optimizing the portfolio, and it also involves hedging and diversification. Therefore, it is a potential way to address the problems in our models discussed above.

2. Include technical features. In addition to fundamentals data, there are also a number of technical features such as trading volume and volatility, and include these features may also improve the performance of our portfolio.

6 Conclusion

In this project, we analyzed fundamental data of public listed stocks in New York Stock Exchange and derived a long horizon trading strategy based on Piotroski F-score as well as a short term trading strategy based on historical fundamentals. Both strategies can be considered effective in terms of profitability.
We also illustrated the limits and discussed the fairness and weapon of math destruction concerns in our models, and suggested future improvements. Overall, the fundamentals of a publicly traded company have significant influence on its stock price, and it is possible to develop trading strategies based on fundamental data.

7 References

X. Zhou, Active Equity Management, 2014

J. Piotroski, "Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers".

SKlearn API, https://scikit-learn.org/stable/index.html