Long term portfolio building strategy

Abstract:

The change of stock price involves many factors, such as transaction number, the performance of company and the current news. The analysis to stock price can be separated into long term prediction and short-term prediction, which analyze different factors. In short-term prediction, the volatility and turnover rate measures how many investors buy or sell this stock and how much money they spend in the transaction. With more money being invested, the price of stock will rise. Otherwise, the price of stock will drop. In long-term prediction, healthy development is an important factor to the company. By analyzing the sharing feature of company with positive performance, we could portrait a pattern for healthy development.

1. Introduction:

In this project, we focus on build strategy to pick company with long term growth potential. The popular indicators in long term growth include ROE, Market Capability and so on. ROE is short term for Return on equity, is a measure of the profitability of a business in relation to the equity. It is considered as a return on assets minus liabilities.

$$ROE = \frac{Net\ Income}{Shareholder's\ Equity}$$

Market Capability involves integration of all marketing related activities of a firm using superior market knowledge from customers and competitions. A large company with more business is more likely to have higher market capability. Besides these two indicators, there are hundreds of financial indicators in annual or quarter report from companies. We would like to use machine learning algorithm or linear regression algorithm to filter the important factors. Then we use the important factors to select stocks and build our customized the portfolio. In the test process, we will compare the annual return of our portfolio with that of S&P500 index.

2. Method:

2.1 Dataset and preprocess:

Source: In our project, we use quarter financial report and daily historical price as resource dataset. The daily historical close price from 2020/11/16 to 2020/12/21 is exported by Alpha Vantage. The 2020 third quarter financial report is from Quandl. Quandl provides more than 100 core financial fundamental indicators for each company based on their quarter report. The indicators are in the following part.

Structure: The historical data is daily price in 28 transaction days after the quarter report is published. There are 475 companies from S&P 500 list. Then the dimension of historical price data is $x^{(t)} \in \mathbb{R}^{475*28}$. The financial indicators include 141 indicators for 475 companies in the historical price matrix. Then the input dimension is $\mathbb{R}^{475*141}$.

Preprocessing: Since not every financial indicator can be updated in quarter frequency, we need to remove the financial indicators which contain more than 30% missing data from the fundamental input matrix.

Training/Test split: The historical price is the output dataset, while the fundamental financial indicators are the input dataset. We separate 70% of them into training dataset and other 30% of them into testing dataset.

Normalization:

Revenues	Enterprise Value	Deferred Revenue	Dimension
Cost of Revenue	Invested Capital	Deposit Liabilities	Calendar Date
Selling General and Administrative Expense	Average Equity	Property Plant & Equipment Net	Date Key
Research and Development Expense	Average Assets	Inventory	Report Period
Operating Expenses	Invested Capital Average	Tax Assets	Last Updated Date
Interest Expense	Tangible Asset Value	Trade and Non-Trade Receivables	Ticker Symbol
Income Tax Expense	Return on Average Equity	Trade and Non-Trade Payables	Filing Date
Net Loss Income from Discontinued Operations	Return on Average Assets	Goodwill and Intangible Assets	Form Type
Consolidated Income	Free Cash Flow	Total Liabilities	Issuer Name
Net Income to Non-Controlling Interests	Return on Invested Capital	Shareholders Equity	Owner Name (Insider / Investor)
Net Income	Gross Profit	Accumulated Retained Earnings (Deficit)	Officer Title
Preferred Dividends Income Statement Impact	Operating Income	Accumulated Other Comprehensive Income	Is Director?
Net Income Common Stock	Gross Margin	Current Assets	Is Officer?
Earnings per Basic Share	Profit Margin	Assets Non-Current	Is Ten Percent Owner?
Earnings per Diluted Share	EBITDA Margin	Current Liabilities	Transaction Date
Weighted Average Shares	Return on Sales	Liabilities Non-Current	Security Acquired/Disposed Code
Weighted Average Shares Diluted	Asset Turnover	Tax Liabilities	Transaction Code
Capital Expenditure	Payout Ratio	Total Debt	Shares Owned Before Transaction
Net Cash Flow - Business Acquisitions and Disposals	Enterprise Value over EBITDA	Debt Current	Transaction Shares
Net Cash Flow - Investment Acquisitions and Disposals	Enterprise Value over EBIT	Debt Non-Current	Shares Owned Following Transaction
Net Cash Flow from Financing	Price Earnings (Damodaran Method)	Earnings before Tax	Transaction Price per Share
Issuance (Repayment) of Debt Securities	Price to Earnings Ratio	Earning Before Interest & Taxes (EBIT)	Transaction Value
Issuance (Purchase) of Equity Shares	Sales per Share	Earnings Before Interest Taxes & Depreciation Amortization (EBITDA)	Security Title
Payment of Dividends & Other Cash Distributions	Price to Sales Ratio	Foreign Currency to USD Exchange Rate	Direct or Indirect?
Net Cash Flow from Investing	Price Sales (Damodaran Method)	Shareholders Equity (USD)	Nature of Ownership
Net Cash Flow from Operations	Price to Book Value	Earnings per Basic Share (USD)	Date Exercisable
Effect of Exchange Rate Changes on Cash	Debt to Equity Ratio	Revenues (USD)	Price Exercisable
Net Cash Flow / Change in Cash & Cash Equivalents	Dividend Yield	Net Income Common Stock (USD)	Expiration Date
Share Based Compensation	Current Ratio	Cash and Equivalents (USD)	Row Number
Depreciation Amortization & Accretion	Working Capital	Total Debt (USD)	Ticker Symbol
Total Assets	Free Cash Flow per Share	Earning Before Interest & Taxes (USD)	Institutional Investor Name
Cash and Equivalents	Book Value per Share	Earnings Before Interest Taxes & Depreciation Amortization (USD)	Security Type
Investments	Tangible Assets Book Value per Share	Shares (Basic)	Calendar Date
Investments Current	Share Price (Adjusted Close)	Dividends per Basic Common Share	Value
Investments Non-Current	Ticker Symbol	Share Factor	Units

Figure1. Financial Indicators from Quarter Report.

2.2 Methods:

The methods we use to filter the feature is Random Forest Regression model, which evaluates the importance of each indicator. Random forest is a supervised learning algorithm which uses ensemble learning method. As its name, the algorithm builds multiple decision trees in the training process. Each decision tree is run in parallel, which prevents the interaction between trees. Each tree is built by the randomly selected sample from the input dataset. Every time the node begins to split, the features will be selected. The optimal feature node will be split in the next iteration. Each tree can growth without limitation, but the selected feature can be customized. The prediction is evaluated by the RMSE value of the model.

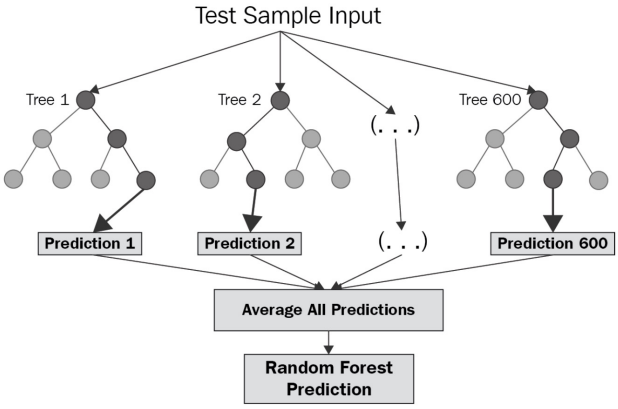

Figure2. Random forest structure

In random forest model, the coefficient or importance to each feature will be calculated. The more influential features will have higher coefficient. We will filter the features based on their coefficient from random forest model.

The advantage of random forest includes: the model can be parallel trained among trees. It has a good algorithm in processing the missing value. And it can evaluate the similarity between sample data, which can be further filter or clustered. The disadvantages include: The model is a black box, which is hard to explained with financial concepts. Overfitting is a problem to random forest model.

3. Experiment:

3.1 Sector classification:

According to Morgan Stanley Capital International and S&P Dow Jones Indices, two pioneer and leader in financial market, the companies can be assigned to a specific economic sector or industry groups based on their business operations. Global Industry Classification Standard aims to enhance the investment research and asset management process for financial professionals worldwide. It is the result of numerous discussions with asset owners, portfolio managers and investment analysts around the world. It was designed in response to the global financial community' s need for accurate, complete and standard industry definitions. The GICS structure consists of 11 sectors: Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Real Estate, Communication Services and Utilities.

In our project, the company information dataset, including historical price and fundamental financial indicators, will be separated into 11 groups based on their identity in GICS sector classification. To companies with similar business, their similarity in fundamental financial indicators can explain more about the pattern in their belonging sector.

3.2 Indicator selection by random forest model

Our random forest model is built with a max depth of trees set to 12, random state to 0, number of estimators to 600 and criterion to MSE. The input for the model is 142 financial indicators and the output is the 28-daily rate of change to the stock price. Then we use the feature importance function to calculate the coefficient to each indicator. The feature importance is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree. Then rank the feature importance from high to low and select the 15 largest features as the customized feature list for healthy-developed companies. According to the RMSE for our random forest model in 11 sectors, the random forest model can predict the price with fundamental indicator accurately.

Sector name	RMSE
Communication	0.050015367
Consumer discretionary	0.064205944
Consumer staples	0.042413060
Energy	0.152429037
Financials	0.043700900
Health Care	0.121584983
Industrials	0.046539706
Information Technology	0.117529704
Materials	0.045139761
Real Estate	0.087128283
Utilities	0.089783752

`

Figure4. RMSE for sector models

Trade and Non-Trade Receivables
Trade and Non-Trade Payables
Net Cash Flow from Financing
Interest Expense
Net Cash Flow - Investment Acquisitions and Disposals
EBITDA Margin
Operating Income
Consolidated Income
Net Income Common Stock (USD)
Current Liabilities
Depreciation Amortization & Accretion
Earnings per Basic Share (USD)
Current Ratio
Income Tax Expense
Free Cash Flow per Share

Figure 5. Selected 15 features to Communication sector.

After successfully build the fundamental indicator list for each sector, we filter the input fundamental dataset with the selected 15 features for the corresponding sector. Then for each indicator, we will use the following rubric to give score based on the performance of each company in this field.

Range	score
x > 75% population	10
50% population < x < 75% population	6
25% population < x < 50% population	3
X > 25% population	1

Figure 6. Grade rubric

We will use this rubric to do a prediction based on the 2021 first quarter report for these companies. We filter their fundamental financial dataset with the customized 15 features in the same sector from previous quarter. Then grade the performance of each companies in each feature, based on the previous rubric. The dimension of graded matrix is m^{n*15}, n is the number of companies in one sector and 15 is the number of selected features. Then, calculate the total score for each company, and compare them with the companies in the same sector.

ticker	eceivables	payables	ncff	intexp	ncfinv	bitdamarg	opinc	consolinc	tinccmnu	liabilitiesc	depamor	epsusd	urrentrati	taxexp	fcfps	sum
ZBH	10	3	3	6	10	6	10	10	10	10	10	6	10	1	6	111
XRAY	10	10	1	10	10	1	10	10	10	10	10	1	3	10	3	109
XYL	10	6	3	10	10	3	10	10	10	6	6	3	6	6	6	105
XRX	10	10	1	1	1	1	10	10	10	10	10	1	6	10	3	94
YUM	3	10	3	10	1	6	10	10	10	6	3	6	3	6	6	93
ZTS	6	3	3	6	1	10	10	10	10	1	6	3	1	10	6	86
ZION	1	1	10	1	1	10	10	6	6	1	1	10	1	10	1	70
XOM	10	10	1	10	1	1	1	1	1	10	10	1	1	1	1	60

Figure7. The grade matrix for communication sector

With the rank of total score to companies in each sector, we can customize our own portfolio based on the rank. In my test process, I select two top companies from each sector and build my portfolio. In the back test process, I use PORTFOLIO VISUALIZER to compare the return of my portfolio and that of S&P500 portfolio.

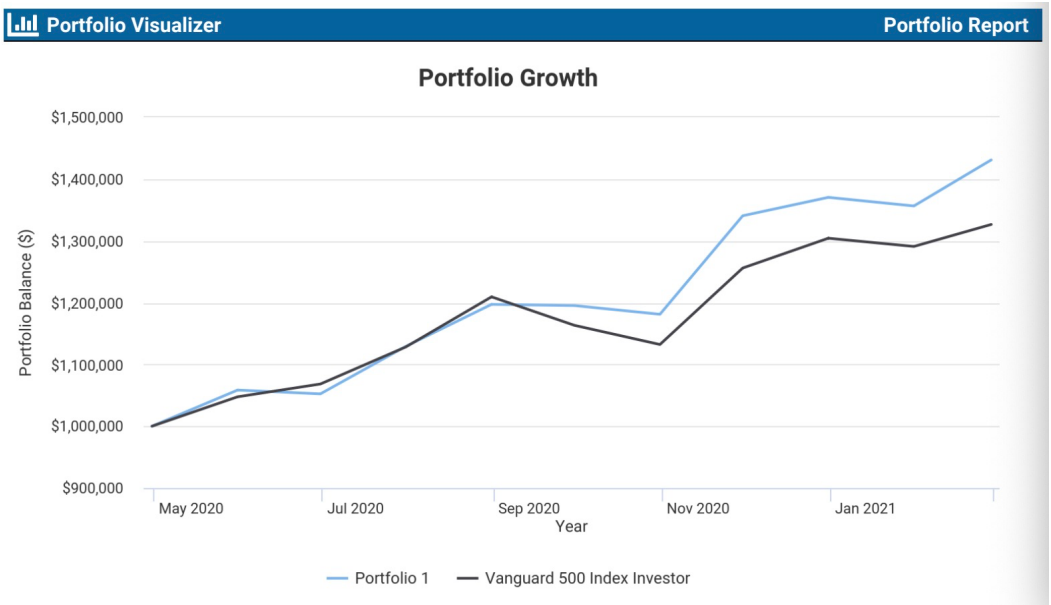

Figure8. The annual return graph for customized portfolio and Vanguard 500 index

From the test result, the return of our customized portfolio is 40% while that of Vanguard 500 index is 30%, which is 25% lower than the customized portfolio return.

4. Conclusion:

According to the experiment result, we conclude that the fundamental indicators from quarter or annual financial report is a good standard to evaluate the long-term development of a company. From the perspective of machine learning model, the Random Forest Regressor selects the efficient financial indicators accurately and increases the return of customized portfolio.

In the future, we could further improve our model by defining the cycle of each sector. To some certain sector, they have their own development cycle. For example, Cyclical industry, including tourism and raw material, reflects the fluctuations in different seasons. To determine a clear pattern of development and decay cycle, I propose to use cluster algorithm to analyze the rate of change to each sector.