Using fundamental data to help predict market movement

Author: Vladimir Zakov

Abstract: "Beating the market" has been of interest to researchers and investors for a long time. As early as 1900, Louis Bachelier (Bachelier L. 1900) notes that the dynamics of the stock exchange will never be an exact science, however it is possible to mathematically study the state of the market at a given moment and try to calculate probabilities of market movements. He concludes that past, present and future events often do not show relationship to price movements. Since then, different research has found different evidence about the predictability of market movements. This paper aims to explore financial statement items and which of them have the strongest predictive power in relation to stock price movements. The popular information gain metric is empirically calculated for items in the financial statements of a group of US companies and the results are presented on a sector basis.

Keywords: fundamental analysis

JEL: G10

1. Introduction

The possibility of predicting market movements has been of interest among researchers for a long time. In the 1960s the advancement of computer technology made easier calculating predictions of stock price movement and the actual movements of hundreds of stocks simultaneously. This made possible the testing of different hypothesis regarding the predictability of the market movements. In his famous paper (Fama, 1970) Eugene Fama provides review of theoretical and empirical research on the subject and provides a testable framework, introducing the Efficient Market Hypothesis. EMH has three form – strong, semi-strong and weak form. Those forms are defined with respect to the information set used to predict the price movements. The weak form contains only information contained in the historical prices – technical analysis, the semi-strong tests all publicly available information, such as financial statements and the strong form tests all available information including private. Through the years research has found both lack of predictability – (Fama, Fisher, Jensen, and Roll, 1969), (McLean and Pontiff, 2016) and predictability - (Shi & Zhou, 2017), (Nedev & Bogdanova, 2019). When it comes to the semi-strong EMH there are papers which show how the fundamentals of the companies can be used to predict stock returns - (Setiono & Strong, 1998), (Noma, 2010).

This paper focuses on fundamental analysis and particularly in showcasing a methodology which can be used to identify potentially strong predictors of stock prices movement. It outlines an easy way in which the standard financial statements published by the US security and exchange commission can be examined for items which can be used by investors in their fundamental analysis.

2. Methodology

The methodology outlined in this section resembles closely the exhibition in (Bogdanova & Stancheva-Todorova, 2020). The utilized algorithm finds the best predictors of the probability of price increase/decrease in the next period for a particular stock. For this purpose, two data streams are used.

The first data stream is historic stock price information from Yahoo finance (https://finance.yahoo.com), where the stock prices, adjusted for splits and dividends is recorded on daily frequency.

The second data stream consists of records on items reported in the quarterly financial statements of the company. This information is obtained from SimFin (https://simfin.com), which itself obtains it from the SEC data archives(https://www.sec.gov). The financial statements used are the original financial statements and not restated versions, if such were made. It encompasses more than 3000 US listed companies with information about the sector (technology, financial services etc.) the company is in.

The utilized approach consists of two stages – data preprocessing stage and calculation of information gain for each of the financial indicators.

First, the data on historic stock prices is cleaned and preprocessed. The raw data of the historic stock prices consist of daily adjusted closing prices $\left\{P_t, \Delta t = \frac{1}{252}\right\}$ of the stocks. Then the following two steps are performed:

1. Daily adjusted closing prices series $\left\{P_t, \Delta t = \frac{1}{252}\right\}$ are averaged on quarterly basis so as to obtain the average quarterly price series $\left\{P_t, \Delta t = \frac{1}{4}\right\}$ The quarter marker is later used as a merging identifier
2. A binary response variable is defined in the following way:

$$Direction = \begin{cases} Up, if\ P_t \geq P_{t-1} \\ Down, if\ P_t < P_{t-1} \end{cases}$$

The final output is a timeseries $\left\{Direction_t, \Delta t = \frac{1}{4}\right\}$

The next step is the preparation of the financial data. For each company three financial statements are combined – balance sheet, cash flow statement and profit and loss statement. Only companies which have all three of the financial statements are kept for further analysis.

Once the statements are combined, all items in the statements which contain missing values are removed. As this is done for each individual company, different companies might end up with different number of items in their final combined statements.

After that, the stock price data and the financial statements data is merged, in a such a way that the price movement for a certain quarter is matched with the financial statement for the quarter before it. For example, a financial statement released for Q2 2020, will be matched with the price movement of a stock between Q2 2020 and Q3 2020.

Finally, only records with at least 20 observations with both stock price movement and financial statements are kept, to ensure that results are more meaningful.

The second stage of the algorithm consists of calculating the information gain for each of the financial statement items against the price movement – up or down.

Information gain is a measure usually used in machine learning algorithms (usually decision trees) to decide which features to be used in the prediction model. It measures the decrease of entropy between two states, where entropy measures the decree of randomness in the data points. The higher information gain, the better the feature is. The information gain of the items is calculated using the attrEval function from the CORElearn R package.

Once calculated the top 3 items with highest information gain are kept and reported. In case of ties the algorithm uses a sporting/min method of determining ranking. For example, a numerical vector with the values 14, 14, 12, 12, 10 will be ranked as 1, 1, 3, 3, 5. industry.

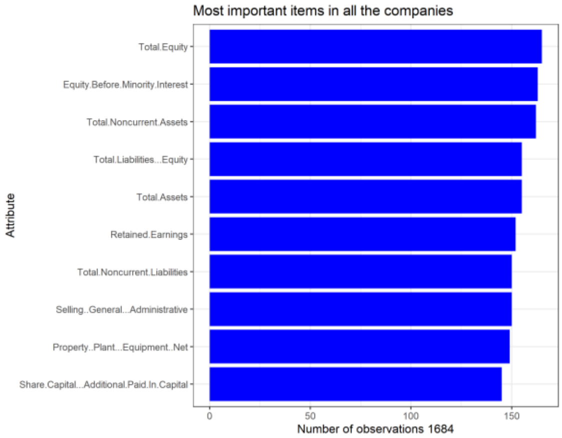
3. Empirical Findings

The application of the aforementioned methodology is presented in this section. After following the instructions outlined in the stage of the methodology section, the initial list of 3121 companies is reduced to 1684 companies which had at least 20 quarters of full financial data between 2009 Q1 and 2022 Q3. The sector representation of the companies is:

- Technology – 299
- Healthcare – 295
- Consumer cyclical – 294
- Industrials – 268
- Consumer defensive – 197
- Basic materials – 97
- Energy – 79
- Real estate – 78
- Financial services – 57
- Utilities – 52
- Vusiness services – 32
- No information – 23
- Other – 3

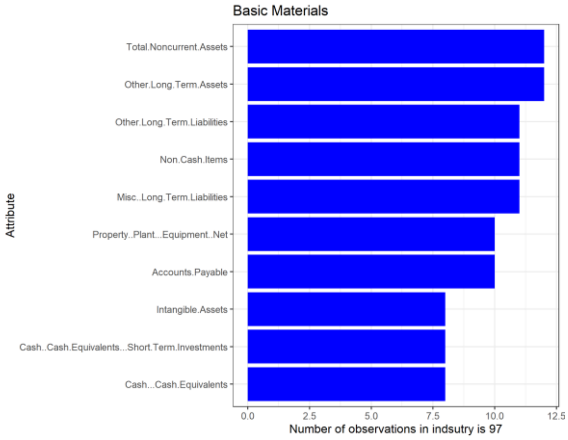
In the figures below, the ten most important features among all the companies for which data was available are presented. Then the same information, but for each of the sectors, excluding the “No information” and “Other” is shown after that.

Fig1. Most important financial state items in all the companies



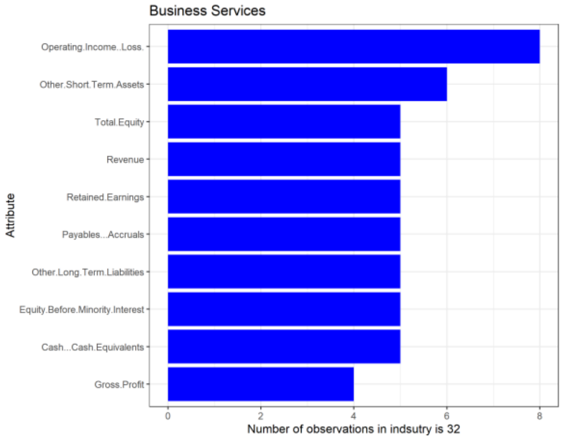
Source: Own data

Fig2. Most important financial state items in the basic materials sector



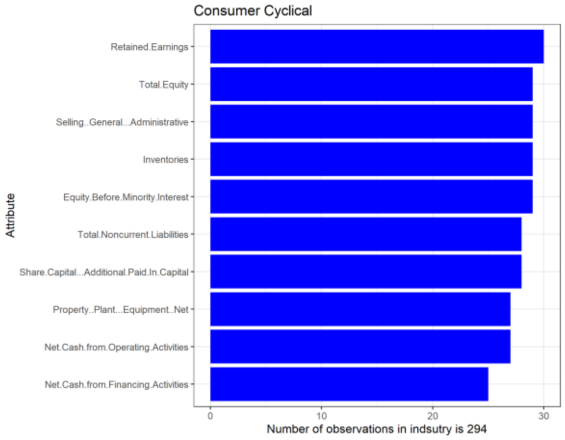
Source: Own data

Fig3. Most important financial state items in all the business services sector



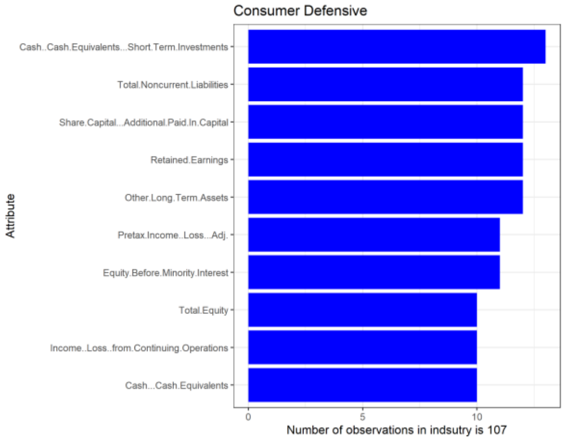
Source: Own data

Fig4. Most important financial state items in all the consumer cyclical sector



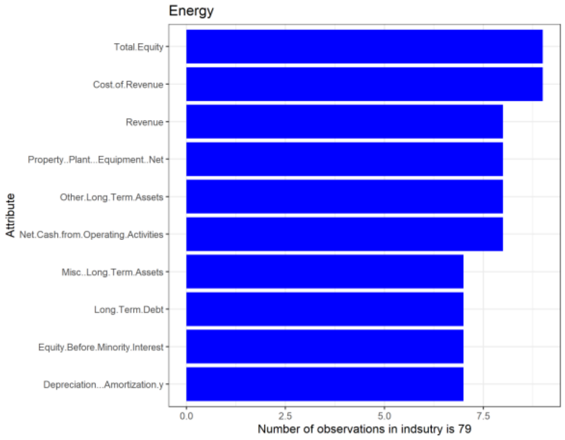
Source: Own data

Fig5. Most important financial state items in all the consumer defensive sector



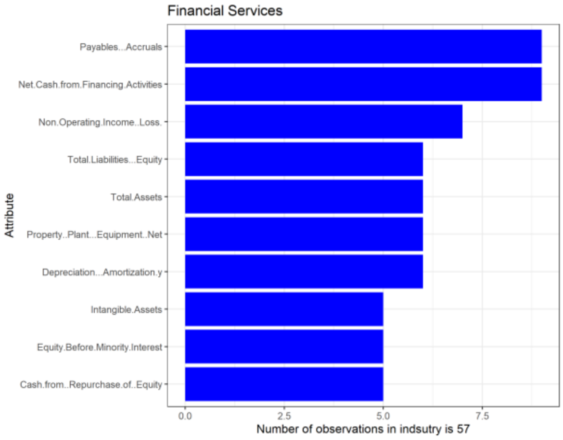
Source: Own data

Fig6. Most important financial state items in all the energy sector



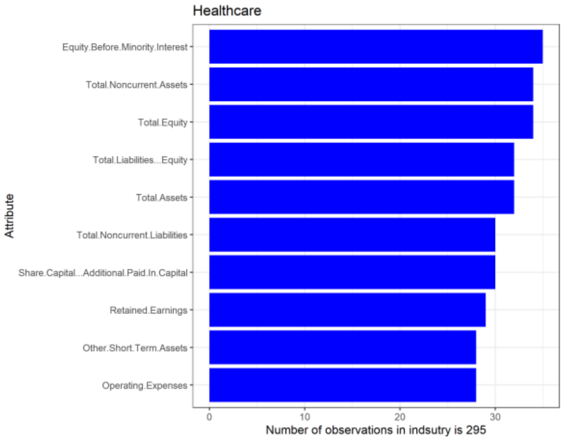
Source: Own data

Fig7. Most important financial state items in all the financial services sector



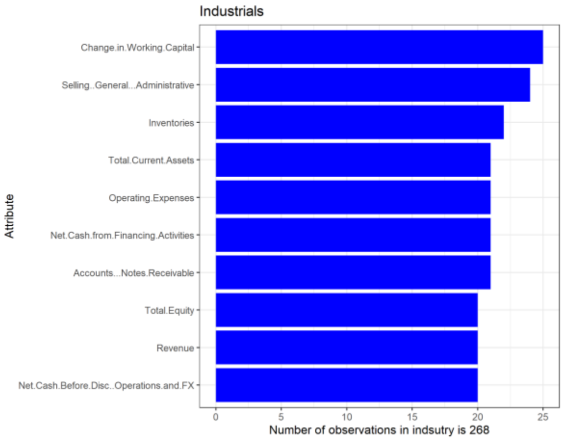
Source: Own data

Fig8. Most important financial state items in all the healthcare sector



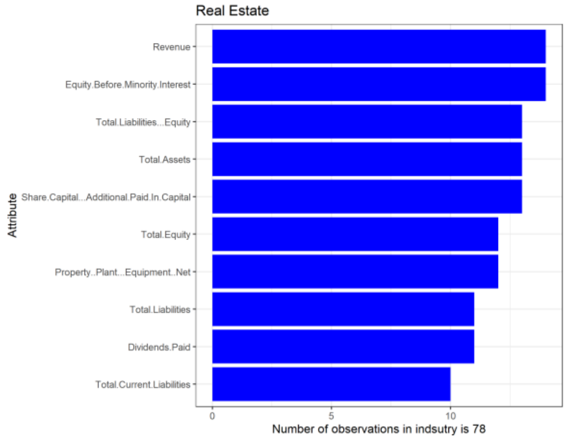
Source: Own data

Fig9. Most important financial state items in all the industrials sector



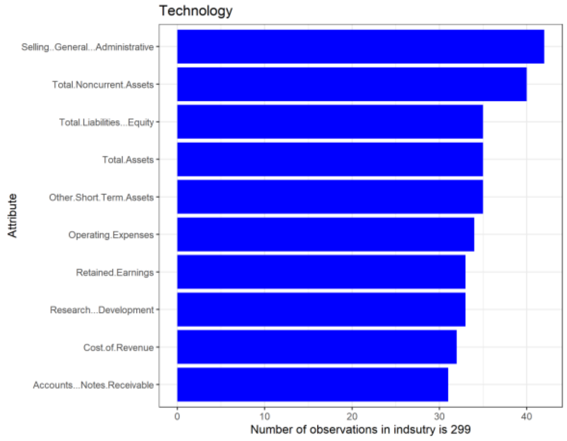
Source: Own data

Fig10. Most important financial state items in all the real estate sector



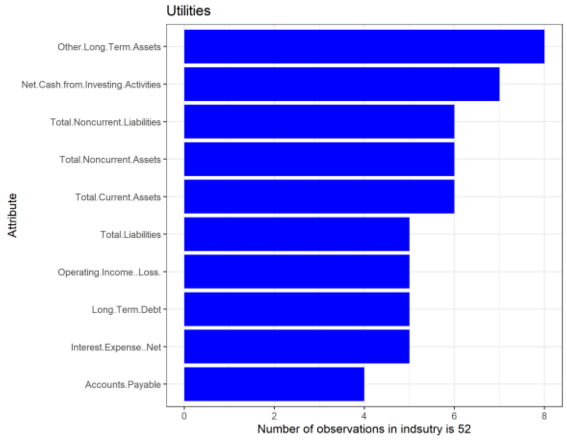
Source: Own data

Fig11. Most important financial state items in all the technology sector



Source: Own data

Fig12. Most important financial state items in all the utilities sector



Source: Own data

The data provides interesting observations and shows both items which are common among the industries and such that clearly have stronger impact on specific sectors. Total equity is consistently among the most important factors, which should come as no surprise. Interesting is that while the stock movement is calculated for the next quarter (short term) the long term assets/liabilities seem to have stronger impact than the short term ones. While SG&A (Selling, General and Administrative expenses) is often looked at in by managers, regardless of sector, it has biggest impact in the technology /consumer cyclical and industrials sectors.

Due to correlation between some of the financial items or different ways of including the same information in a financial statement, potentially predictive effects might not be properly shown. If there are two strongly correlated financial items, which are top predictors, they will be counted separately as top 3 predictors. However, the underlying reason, which drives the price change will be the same, while other predictors, which are strong might not be shown. This methodology also does not show the type of relationship between the financial items and the price movement.

Still the outlined methodology might be used a supplementary tool by investors in the process of their investment decision-making. It can identify relative importance of each financial item and the relative importance of them for each industry.

References

1. Bachelier, L. Annales scientifiques de l'École Normale Supérieure, Serie 3, Volume 17 (1900), pp. 21-86.

2. Fama, E. F., 1970. Efficient capital markets: a review of theory and empirical work. Journal of Finance, 25(2), pp. 383-417.

3. Fama, Eugene F. and Fisher, Lawrence and Jensen, Michael C. and Roll, Richard W., The Adjustment of Stock Prices to New Information (February 15, 1969). International Economic Review, Vol. 10, February 1969,

4. McLEAN, R. DAVID, and JEFFREY PONTIFF. “Does Academic Research Destroy Stock Return Predictability?” The Journal of Finance, vol. 71, no. 1, 2016, pp. 5–3

5. Shi, H.-L. & Zhou, W.-X., 2017. Time series momentum and contrarian effects in the Chinese stock market. Physica A, Volume 483, pp. 309-318.

6. Nedev, B. & Bogdanova, B., 2019. Comparative analysis of momentum effect on the NYSE and the SHSE from the perspective of cultural specifics. AIP Conference Proceedings, 2172(1), p. 080011.

7. Noma, M., 2010. Value investing and financial statement analysis. Hitotsubashi Journal of Commerce and Management, 44(1), pp. 29-46.

8. Setiono, B. & Strong, N., 1998. Predicting stock returns using financial statement information. Journal of Business Finance & Accounting, 25(5-6), pp. 631-657.

9. Bogdanova, B. & Stancheva-Todorova, E., 2020. ML-based preditive modelling of stock market returns. forthcoming in AIP Conference Proceedings.