Determining Feature Importance in Securities:
Why did an Individual Security's Return Surpass the S& P500's

BUS 346 - Advanced Business Analytics with R
Script and Analysis by Iván Sepulveda
Faculty Advisor: Shivani Shukla
March 22nd, 2018

Using various classification methods, I sought to determine the most relevant variables in a stock's performance; securities were evaluated in a binary manner to reflect if they had surpassed the Standard and Poor 500's (S&P-500) performance for the equivalent time period.

Introduction

Using the classification methods of Logistic Regression (LogR), Support Vector Machine (SVM), k-Nearest Neighbor (k-NN) on a financial dataset of ten features, I arrived at the conclusion that not only does variable importance change by classsification method, but also that financial data behaves unlike any other data due to it's inversed k-NN pattern. In addition, for the time period of April 1st, 2018 to August 24th, 2018, features of prominent importance were select sectors, price to book ratio, EBITDA, and earnings per share (as of April 1st, 2018).

Data Aquisition and Cleaning

The initial April 1st data was aquired from datahub.io [1] and provided the (I assume closing) stock price along with thirteen other features. Of these fourteen, Symbol, Name, and SEC Filings were not relevant and therefore deleted. The Sector column was provided in string for i.e. "Industrials", "Consumer Discretionary", "Materials". These values were encoded in integers from one to eleven in the order listed by ETF Database [3]. Stock prices for August 24th, 2018 were aquired from kaggle.com [2]. However, this dataset was comparatively larger in size, as it held roughtly 48 years of market data. A seperate, availailable-on-request script was written to extract our August 24th data. In addition, unlike our datahub.io dataset, the kaggle.com data did not include financial features such as Price to Book ratio or Market Cap. Therefore, these two datasets were combined into one. It should be noted that the kaggle.com dataset included stock prices such as the daily volume, high, low, etc., but I only extracted the stock's close price as it was the most relevant. Combining these two datasets had it's setbacks. It appears to me that the company's list on our original S&P-500 had changed from the beginning to end of the time period measured. Although the first dataset provided data for 505 companies, when the datasets were matched, only 490 remained. In addition, the former dataset was slightly incomplete, so ten rows were deleted for having NA values, leaving me to work with 480 companies. To finish the data integration, the stock price increase from start to end then was calculated and compared to the return of the S&P-500 for period of April 2018 to September 2018 as calculated by About Don't Quit Your Day Job's online S&P 500 Return Calculator. (Input parameters were in months, so I assumed that the return calculated pulled data from the first of the month(s) and rounded August 24th to September 1st. Acccording to this platform, the 'Total S&P 500 Return' for this time frame was 9.341%. So I added a new column to my dataset with a binary value indicating whether or not the individual stock's percent increase surpassed this return, then deleted stock prices for the beginning and end dates, as they were now irrelevant and could potentially hurt accuracies.

Results

This section will interprest my results of the classification methods used displayed in Table 3. LogR - Although it was the least accurate, I contend that 70% accuracy is a strong initial score for this simplistic method. SVM - This technique performed better than k-NN at k=1 and LogR, but only by a small margin. After variable importance analysis, I found that although LogR has k-Nearest Neighbors - These results were the most surprising: in most cases, k-NN becomes more accurate as the k parameter is decreased, as with more 'neighbors' a result has a wider classification range and therefore more room for error. However, because my our k-NN results behaved like the inverse, I would theorize that securities significantly dependent one another in the context of classification. We see this in the real world: when many companies within a certain sector of the economy are performing well, other companies within that sector tend to follow. This could be due to increasing public faith in that sector or because more than often companies within the same industry are either collaborators or competitors. When the former, one can see how when Company A, who much work to Company B, dues well, it will probably outsorce more work to Company B. When the latter, one can see how Company C, who is in constant competition with Company D, does well, Company D take measures to be more appealing to their shared customer base, driving revenue for Company D. The importance of sector is verified by a Variable Importance analysis, whose results are displayed in Table 1 (Note: these are variable importance measures when classifying by LogR). One minor note: Financials was manually added since it did not appear on the Variable Importance results and was therefore assumed to be zero. The Financials sector is the most volatile of the eleven, so I can safely make this assumption due to volatile sectors being unable to give much information other than noise.

Indicator	Value
Utilities	10.9%
Energy	9.7%
Real Estate	9.5%
Information Technology	9.5%
Health Care	8.0%
Market Cap	7.7%
Industrials	7.3%
EBITDA	6.3%
52 Week High	5.4%
52 Week Low	4.1%
Price to Sales Ratio	4.0%
Price to Book Ratio	3.8%
Telecommunication Services	3.7%
Consumer Discretionary	3.0%
Earnings per Share	2.8%
Consumer Staples	2.1%
Materials	1.7%
Price to Earnings Ratio	0.2%
Divident Yield	0.2%
Financials	0.0%

Table 1: Variable Importance Logistic Regression Model

Indicator	Value
Price to Book Ratio	22.2%
EBITDA	17.7%
Earnings per Share	13.9%
52 Week High	13.0%
52 Week Low	9.1%
Market Cap	8.6%
Price Earnings	7.0%
Divident Yield	3.1%
Sector	2.9%
Price to Sales Ratio	2.4%

Table 2: Variable Importance Suport Vector Machine Model

Method	Accuracy
k-Nearest Neighbor: k = 15	89.6%
Support Vector Machine	71.0%
k-Nearest Neighbor: k = 1	70.6%
Logistic Regression	69.8%

Table 3: Accuracy of Classification Methods Used

Conclusion

Based off these results, I've come to the following conclusions. First, for the specific time period analyzed, wheather or not a company belong to the utilities, energy, real estate, information technology sectors tended to be a strong indicator of wheather or not their share price growth surpassed that of the S&P-500. Second, for the specific time period analyzed, the price to book ratio, EBITDA, and earnings per share features also tended to be strong indicators of our performance benchmark. Third, financial data does not behaves uniquely; a stock's performace more strongly resembled the average of it's fifteen 'nearest neighboors' more so than a single stock almost identical to itself. Fourth and lastly, multiple methods of classfication may end up placing respective weights of importance on different features.

Acknowledgements

This work was supported by the University of San Francisco.

References

[1] S&P 500 Companies with Financial Information (2018).
https://datahub.io/core/s-and-p-500-companies#data-cli

[2] Daily Historical Stock Prices (1970 - 2018) (2018).
https://www.kaggle.com/ehallmar/daily-historical-stock-prices-1970-2018/metadata

[3] The 11 Sectors of the Stock Market (2015)
https://etfdb.com/etf-education/the-10-sectors-of-the-stock-market/

[4] Investopedia (2019) https://www.investopedia.com/

[5] Paulo Cortez, Mark J. Embrechts, Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models, Information Sciences, Vol. 225, 2013, Pgs. 1-17
ISSN 0020-0255, https://doi.org/10.1016/j.ins.2012.10.039.

[6] Lawrence Hamtil, Financials: The Market's Most Volatile Sector (2016)
https://www.fortunefinancialadvisors.com/blog/financials-the-markets-most-volatile-sector/

Appendix: Relevant Financial Terms

EBITDA - or earnings before interest, taxes, depreciation and amortization, is a measure of a company's overall financial performance and is used as an alternative to simple earnings or net income in some circumstances. EBITDA, however, can be misleading because it strips out the cost of capital investments like property, plant, and equipment. This metric also excludes expenses associated with debt by adding back interest expense and taxes to earnings. Nonetheless, it is a more precise measure of corporate performance since it is able to show earnings before the influence of accounting and financial deductions.[3]

Price-to-Book Ratio - Companies use the price-to-book ratio to compare a firm's market to book value by dividing price per share by book value per share (BVPS). An asset's book value is equal to its carrying value on the balance sheet, and companies calculate it netting the asset against its accumulated depreciation. Book value is also the net asset value of a company calculated as total assets minus intangible assets (patents, goodwill) and liabilities. For the initial outlay of an investment, book value may be net or gross of expenses, such as trading costs, sales taxes, and service charges. Some people may know this ratio by its less common name, price-equity ratio.[3]

Utilities - The utilities sector consists of electric, gas and water companies as well as integrated providers. In general, the sector generates consistent recurring income by charging consumers and businesses that provide higher-than-average dividend yields.[4]

Energy - The energy sector consists of oil and gas exploration and production companies, as well as integrated power firms, refineries and other operations. In general, these companies generate revenue that's tied to the price of crude oil, natural gas and other commodities.[4]

Earnings Per Share - Earnings per share (EPS) is the portion of a company's profit allocated to each share of common stock. Earnings per share serve as an indicator of a company's profitability. It is common for a company to report EPS that is adjusted for extraordinary items and potential share dilution. [4]