Advances in Economics, Business and Management Research, volume 219
Proceedings of the 2022 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022)

Stock Return Prediction Using Machine Learning Classifiers

Shenghan Zhao

Arts and science
University Of Toronto
**Corresponding author. Email: shenghan.zhao@mail.utoronto.ca*

ABSTRACT
Machine learning is one of the major tools that could extract patterns from big data. The bigger the size of the data, the better the machine will learn. From the technical perspective, machine learning makes better predictions. In this paper, the major objective is to predict annual average stock returns using machine learning methods. To build the model, fundamental and financial data about each firm are used. Furthermore, news of each firm is webscrabbed and categorized as either positive or negative news using sentiment analysis methods. The conclusion is that adding sentiment analysis decreases the accuracy of the entire model. This result indicates that with influential incidents like COVID-19[1], models have more significant scores than usual (when considering only SP500 components). In 2018, adding sentiment variables decreased the model accuracy by 3% on average.

Keywords: *Stock return prediction, machine learning, sentiment analysis, random forest, decision trees, KNN*

1. INTRODUCTION

Since the first introduction of the stock market in the US in the 1700s, the stock return has always been the most popular topic among investors. Financial economists dedicate enormous time and resources to researching factors affecting stock returns. Previous research has proven that many factors could pivot future stock returns of a firm, including balance sheet components such as return on earning (ROE) and price to earnings ratio(P/E ratio). Unlike some pre-existing studies, qualitative variables are as crucial as quantitative ones in stock return analysis. Textual analysis of news [2], logistic regression [3], and sophisticated neural networks [4] have been shown to be reliable in explaining stock return variations and performing forecasts. With the previous success of machine learning algorithms, this study continues with these methods to conduct analysis. Specifically, this study aims to use K-Nearest neighbors, logistic regression, decision tree, and random forest to examine whether a stock could outperform the market. The research scope is the SP500 firms in the US stock market from 2018 to 2020. The fitted model has one binary response variable, and the independent variables consist of quantitative and qualitative variables. Apart from basic firm information, qualitative variables include additional variables from sentiment analysis. The analysis outcome shows that the K-Nearest neighbour model generally has the highest crossvalidation score for predicting stocks' performance. Besides, the variation of features is vital to model performance. Through this study, investors can more comprehensively understand which variables will affect stock returns and the impact of news. In addition, the model in this study can also help investors have a credible prediction of future stock returns. The following section describes the relevant information regarding the dataset used in the analysis and introduces the web scraping process for sentiment analysis. Then, K-Nearest neighbours, logistic regression, decision tree, and random forest methodologies are used to examine whether any stock return outperforms the market. The interpretation of the test results and summaries are presented in the result and conclusion sections, respectively.

2. DATA AND METHODOLOGY

2.1. Data

Firm-level financial statistics are derived from Wharton Research Data Services' database, 'compustat.'

'compustat' is a panel data collection of fundamentals statistics acquired from SP500 component businesses and structured chronologically across even intervals from 2016 to 2020. The dataset is comprehensive and contains 505 firms from January 2015 to January 2021. There are 3561 observations and 55 variables in the dataset. The response, 'btm,' is a binary qualitative variable, which equals one when the firm's annual return is greater than the SP500 annual return and equals 0 otherwise. The firms' rate of return is calculated as a percentage change of the annualized close price, given in the compustat database. The annual SP500 (ticker GSPC) returns are from yahoo finance. 45.69 percent of all the observations beat the market and made an alpha in the dataset. The key identifying information for firms in the dataset is the global company key (GVKEY), a unique identifier and primary key for each company defined in the database. Feature 'gind' represents the third level in the Global Industry Classification Standard (GICS) hierarchy, which is used as a basis for SP500. The industry is represented by the leftmost six digits of the total GICS code. For example, 2030 means the industry group of United Airlines is transportation, and a 'gind' of 203020 of United Airlines means its industry is airlines. Specific industry taxonomy is listed in the reference. Besides the company identification, the rest predictors come from the balance sheet, income statement and cash flow. Descriptive statistics are shown below. Other than the factors used by stock price calculation models, dividend discount model and the Gordon Growth model, relevant features like book value per share, retained earnings, and working capital, are included. Miscellaneous items like employment, which represents the number of company workers reported to shareholders, are considered an input in the model.

2.2. Sentiment Analysis

Web scraping is the preliminary procedure before conducting sentimental analysis. The 'Selenium' package is used for scraping news posted in 2018 by authoritative publishers (i.e. CNBC, Yahoo Finance). Articles for each SP500 firm are stored in a pre-built dictionary, with the firm name being the key. Since this analysis focuses on the relative percentage of positive and negative news within the one-year duration, multiple occurrences of the same report on the same day will only be considered once. The subsequent sentimental analysis uses the 'TextBlob' package to calculate the sentiment score for each news article. The sentiment score lies between -1 and +1, with -1 being strongly negative and +1 strongly positive. In this study, the threshold to distinguish the news sentiment is set to 0.05. Specifically, articles with a sentimental score higher or equal to 0.05 are considered 'positive,' while less or equal to -0.05 are considered 'negative,' and any score in between is considered 'neutral.' Then, new variables are introduced to represent the percentage of positive and negative news that each firm has and the average sentiment score in 2018 for the firm. These variables are used in the subsequent classification modelling.

2.3. Modelling

The classifiers used to analyze the data are logistic regression, decision tree, random forest, and K-nearest neighbors. Given the qualitative response variable, classification is the appropriate machine

learning method. In the logit model, the logistic function employed is:

$$p(X) = e^{(\beta 0 + \beta 1X1 + \cdots + \beta pXp)} / 1 + e^{(\beta 0 + \beta 1X1 + \cdots + \beta pXp)} \quad (1)$$

Let X = (X1,X2, . . . ,Xp), representing p features. The odd ratio is: P(Y = 1|X) / P(Y = 0|X) = p(X) /(1 − p(X)) value of the odds close to 0 and ∞ indicate very low and very high probability of beat the market, respectively. The estimation function of K-nearest neighborhood[5] is:

$$\hat{f}(x_0) = \frac{1}{K}\sum_{x \in N'} y_i \quad (2)$$

Given a value for K and a prediction point x0, KNN regression first identifies the K training observations that are closest to x0, represented by N′. It then estimates f(x0) using the average of all the training responses in N′. The choice of k is the optimal number of neighbours to be included in the classifying process, it's estimated by built-in function from sklearn. The goal of decision tree is to find boxes R1,R2, ... ,RJ that minimize the sum of square of residuals,

For the set of trees Tb, b = 1, 2…., b, the algorithm for random forest in regression [6] is:

$$\hat{f}(rf) = \frac{1}{B}\sum_b T_b(x) \quad (3)$$

In classification, it is most of the results of trees. To estimate the performance of the models, k-fold cross validation is used to calculate scores for each model. The logic is that observations are randomly divided into k groups with approximately equal size. Then, treat the first fold as validation set and fit the rest k − 1 folds with selected models. Then mean square error MSE1 is computed on the observations in the left-out fold. After repeating the process k times, MSE1, …, MSEk are calculated. In the model, k = 5 as conventional choice. By averaging the values, the 5-fold cross validation estimation is proceeded [7]

$$CV = \frac{1}{5}\sum_{i=1}^{5} MSE_i \quad (4)$$

As cross validation score is a number, a figure is more informative than just statistics. Receiving Operating Characteristics curve is employed, which simultaneously displays two types of errors (true positive rate as y-axis and false positive rate as x-axis) for all thresholds. Corresponding area under the curve are calculated, which increases in quality of fitting.

3. RESULTS

Since the Web Scraping process only uses several representative news publishers, some firms have no news or very little news available after scraping. These firms are excluded in the subsequent analysis. The summary of truncated data is presented in Table 1. In general, the average sentiment score for 502 SP500 companies in 2018 ranged from -0.1496 to 0.2767. The average sentiment score among 502 companies is approximately 0.074, and the average percentage of negative news is only approximately 9.4%. This result is strong evidence that firms only tend to publish news conducive to profits and construct a beneficial outlook. Moreover, a high level of the subjectivity of news means that the effect of this news on the stock price may be exaggerated, with all the fluctuations being transient. Therefore, a plausible conjecture is that news may not strongly correlate with the annualized return [8].

Table 1. Sentiment Summary

	% of positive news	% of negative news	Average sentiment score
Mean	57.46	9.42	0.074
Min	0	0	-0.1496
Max	90	75	0.2767

Before introducing sentiment scores to the model, the predictors are financial variables of firms. Precisely, the first 5 predictors are assets, book value per share, common/ordinary equity, cash and short-term investment, and debt. To clarify first, because of the randomly assigned observations to training and testing group, this part gives us slightly different outcome every run. According to Table 2, the logistic regression and KNN model using 2018 data have the highest and same cross-validation score, 0.68, with the best choice of k being 15. The logistic regression and random forest model using 2019 data have the highest cross-validation score,0.64, with the best choice of k being 14. The three methods fit equally well compared with each other, as the scores slightly fluctuate when observations are randomly assigned to the training and test groups. However, it is not convincing that the models predict stock performance well according to the ROC curves. The area under the curve ranges from 60 to 75, which is adequate for estimation [9].

Table 2. Results in 2018

Method	Best K	Cross-validation score(CV)	AUC
Logistic regression	N/A	0.68	0.75
Decision tree	N/A	0.62	0.643
Random forest	N/A	0.66	0.758
KNN	15	0.68	0.74

Table 3. Results in 2019

Method	Best K	Cross-validation score(CV)	AUC
Logistic regression	N/A	0.64	0.623
Decision tree	N/A	0.57	0.478
Random forest	N/A	0.64	0.705
KNN	14	0.62	0.626

The logistic regression and KNN model using 2020 data have the highest cross-validation score, 0.73, with the best choice of k being 33. For all three years, before introducing sentiment and news analysis, the logistic regression model has the best performance, and KNN performs equally best as logistic regression for 2018 data. Random forest is also the best model for 2020 data. The performance of models improves notably for 2020 data. The initial hypothesis is that the pandemic in 2020 may make it harder to capture an explanatory pattern. The unexpected result indicates that large-scale corporations, especially SP500 firms, have relatively stable returns and financial fundamentals.

Consequently, the models did not perform well in the previous two years because features have a slight variance for the models to capture or learn, increasing the standard error of estimation. Under the pandemic, financial statements vary much compared with before, and even large corporations are influenced to some extent.

This finding gives us confidence that the models, especially logistic regression and K-Nearest Neighbor, are likely to predict stock performance with higher scores if more sample firms of different sizes are included.

Table 4. Results in 2020

Method	Best K	Cross-validation score(CV)	AUC
Logistic regression	N/A	0.73	0.806
Decision tree	N/A	0.71	0.663
Random forest	N/A	0.71	0.870
KNN	33	0.73	0.816

After introducing sentiment as a new predictor to the model, it is shown that the validation score is decreased for logistic regression, random forest and decision tree for 0.02, 0.04 and 0.06 respectively. Yet the validation score for KNN has increased by 0.03 and best of K decreases to 7. Method Best K Cross-Validation score AUC.

Table 5. Results in 2018 with sentiment variable

Method	Best K	Cross-validation score(CV)	AUC
Logistic regression	N/A	0.66	0.735
Decision tree	N/A	0.56	0.588
Random forest	N/A	0.62	0.744
KNN	7	0.68	0.742

4. DISCUSSION

One limitation of the sentiment analysis is that each firm's number of news scraped differs. Leading enterprises such as Apple and Amazon have more market attention, causing their relevant news to far exceed that of other companies. Although percentage was used instead of the actual number of positive news to compensate for some parts of bias, there is still non-negligible variances difference in percentage form for each firm. As a result, the news predictor's effect on larger firms outweighs its effect on smaller firms. For example, if a firm has ten news articles, eight are positive. Then 80 percent of the news is positive. Meanwhile, consider a larger firm with 50,000 news, it is very hard for such firms to have 80 percent positive news, but it can easily be achieved for a firm with less news about itself. An alternative method is to assign firms into groups according to their market capitalization and perform analysis within each group. This would resolve the problem of different sample sizes and suggests a more fitted model. Logistic regression is always the best model before introducing news and sentiment to the model. Also, investors' sentiment will also have influence on the stock return[10] yet those information is not collected giving the technical difficulty. It is very hard to webscrap those information from all social media.

In addition, the random forest model has a higher score than the decision tree classifier in all three years. This could be explained by the nature of the random forest, which has combined decision trees. Increasing the number of decision trees and the model size is more likely to generate better prediction results. The cross-validation score of four models in 2020 was the highest in three years. Interestingly, the best-fitted model is the one using the 2020 data, which is significantly different from the usual market environment due to the pandemic. One reasonable cause is the problem of the restricted data. Since greater variation in predictors decreases the standard error of estimation, considering only the financial statements of SP500 firms in the sample, which are stable in annual returns and similar in size to each other. A wider range of companies with different sizes, market capitalization and structures should be included in further research. Besides, if stocks can be treated as independent time series and use appropriate machine learning methods for time series, there may have been a better result with much detailed variation in stock prices. After introducing news and sentiment scores to the model, a new model with slightly reduced accuracy is observed. It is assumed that stock price is strongly correlated with sentiment and news. Nevertheless, the model failed to show this strong correlation. The cause might be that recent negative news would cause irrational trading, which will affect stock liquidity. This implies that the effects of negative news are not reflected in stock price immediately and reduce the price. This contradicts the assumption that news will affect a stock's fair price and price traded. Thus, the model predicts the true stock price, which is not expressed in market price.

5. CONCLUSION

Although adding sentiment variables to the model could not improve the overall accuracy, random forest classification always outperforms other methods in this study. It has an accuracy of 77.8% on average which is reflected from area under the curve (AUC). Some limitations are that this study used data from 2018 to 2020, which are years with a huge influence of COVID-19 on the entire market. This results in unfavorable decline or gain on random industries of stocks, which will decrease the overall accuracy of the model. In addition, sentiment analysis only included news analysis but investors' sentiment. Investors' sentiment is difficult to collect but crucial. Media could be biased giving the corresponding magazine, that is, country of origin. Investors' sentiment represents the true market response to the companies' stock, but they are not always posted on social media or even on any online resources. Therefore, ideas like expanding the size of the dataset will drastically increase the accuracy of the model. This can also be reflected from the fundamental properties of machine learning, which is that the machines learn better with more data points. In future, adding investors' sentiment along with the news sentiment in the model would increase overall accuracy and will be an improved benchmark of stock market prediction.

REFERENCES

[1] Garcia, D., "Sentiment during recessions," Journal of Finance, 68(3), 2013, pp. 1267– 1300.

[2] Box, T. (2018). Comovement and the news. SSRN Electronic Journal, 2018. DOI: https://doi.org/10.2139/ssrn.2139708

[3] Imran, K., Prediction of stock performance by using logistic regression model: Evidence from Pakistan Stock Exchange (PSX). Asian Journal of Empirical Research, 8(7), 2018, pp. 247–258. DOI: https://doi.org/10.18488/journal.1007/2018.8.7/1007.7.247.258

[4] Heston, S. L., & Sinha, N. R., News versus sentiment: Predicting stock returns from news stories. Finance and Economics Discussion Series, (048), 2016, pp. 1–35. DOI: https://doi.org/10.17016/feds.2016.048

[5] K nearest neighbor: Knn algorithm: KNN in Python & R. Analytics Vidhya, 2020. Online: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[6] Random Forest: Introduction to random forest algorithm. Analytics Vidhya, 2021. Online: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[7] Brownlee, J., A gentle introduction to k-fold cross-validation. Machine Learning Mastery, 2020. Online: https://machinelearningmastery.com/k-fold-cross-validation/

[8] Cepoi, C.-O., Asymmetric dependence between stock market returns and news during COVID-19 financial turmoil. Finance Research Letters, 36, 2020, p. 101658. DOI: https://doi.org/10.1016/j.frl.2020.101658

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R., An introduction to statistical learning with applications in R. Springer, 2021.

[10] Tetlock, P. C., "Giving content to investor sentiment: The role of media in the stock market," Journal of Finance, 62(3), 2007, pp. 1139–1168.