732A92 Text Mining

CLASSIFYING STOCK PRICE MOVEMENTS BASED ON 8-K SEC FILINGS

November 28, 2019

Name: Julius Kittler
Student ID: julki092
Linköping University

Abstract

For over a century, fingerprints have been an undisputed personal identifier. Recent court rulings have sparked interest in verifying unique

Contents

1 Introduction 5
1.1 SEC Filings . . . 5
1.2 8-K Filings . . . 5
1.3 Research Questions . . . 5

2 Theory 7
2.1 Stock Price Forecasting with 8-K filings . . . 7
2.2 Stock Price Forecasting with Text Data . . . 7
2.3 Stock Price Forecasting in General . . . 7

3 Data 7
3.1 Sources . . . 7
3.2 Retrieval . . . 7
3.3 Processing . . . 8
3.4 Descriptive Statistics . . . 8
3.4.1 General . . . 8
3.4.2 Target Variable . . . 8
3.4.3 Feature Variables . . . 8

4 Method 8
4.1 Evaluation Metric . . . 8
4.2 Baseline Models . . . 8
4.3 Advanced Models . . . 8
4.4 Hyperparameter Tuning . . . 8
4.5 Feature Relevance . . . 9

5 Results 9

6 Discussion 9

7 Conclusion 9

8 References 9

List of Figures

1 A boat. 8

List of Tables

1 Overview of all 8-K sections and events 6
2 The pros and cons of Scala's optional classes 9

1 Introduction

Forecasting stock prices has been a relevant problem since the existence of publicly traded companies. Today, it is a more relevant problem than ever before because we have the technological infrastructure to build automatic trading systems and thereby put our forecasts to practice. We can not only obtain massive amounts of data relevant for forecasts via APIs but also execute trades via APIs. With commission free trading having become the industry standard in the U.S., the latter services are even offered for free, for instance by the commission free algorithmic trading API Alpaca [1].

This text mining project aims to explore whether filings of the U.S. Securities and Exchange Commission (SEC) can be used to classify if the price of a stock will decrease, remain unchanged or increase. The results will be evaluated critically from the perspective of a trader, with the question in mind whether the classifications could actually be put to practice in an automatic trading system.

1.1 SEC Filings

The SEC is a government agency in the United States with the mission to ”protect investors, maintain fair, orderly, and efficient markets, and facilitate capital formation” [2]. An important task of the SEC is to ensure that publicly traded companies inform their shareholders and the public about their business.

For instance, the SEC requires companies to publish their quarterly and annual results, and inform shareholders about certain relevant events. For each of these purposes, companies have to file specific documents. For instance, the annual report corresponds to the 10-K, the quarterly report corresponds to the 10-Q and another report for specific relevant events corresponds to the 8-K filing.

Importantly, SEC filings are actively used by traders when making investment decisions. Many trading platforms such as Webull and thinkorswim also provide traders with the recent SEC filings of any tradable company (along with other information such as fundamental data, news data and historical prices). SEC filings are interesting for stock price forecasting because they are standardized, publicly accessible for free and because they contain relevant, objective and generally accurate information.

1.2 8-K Filings

Companies need to publish an 8-K filing for major events relevant for their business. For instance, such events might be a change in the board of directors, a potential delisting from a stock exchange or a merger. To be precise, there are 31 different 8-K filing events from 9 different sections. One 8-K filing may contain information for several such events. Every 8-K filings clearly states for which events it contains information. A complete list of all events and sections, taken from the official SEC website [3], is shown in the table 1.

In general, 8-K filings are due within four business days after the event [4], a relatively short time period. Because 8-K filing correspond to major events for the company and because they need to be published shortly after an event occurred, 8-K filings seem interesting for predicting short-term volatility in the stock market. Moreover, the important information in 8-K filings is generally represented in form of text data, whereas other filings such as the annual report often focus on numerical data represented in tabular form. Text data is relatively simple to extract from HTML documents in order to generate features for training machine learning models (compared to tabular and graphical data).

To see some examples of 8-K filings, one may go to the official SEC website. In particular, three examples can be found here: example 1, example 2, example 3.

1.3 Research Questions

There are mainly three research questions that this report aims to address. The focus is not only on the stock price classification itself but importantly also on understanding how the model works. Assuming that we do not know much about SEC filings, we would like the model to give insights about how it is using the text data of the SEC filings for the classification and in which cases the model is more or less reliable.

1. Can we successfully forecast stock prices based on 8-K filings? If so, how well?
2. Which text features are most important when forecasting stock prices based on 8-K filings?
3. In which scenario do the forecasts perform best (e.g. type of the 8-K filing, industry of the company, other metrics of the company)?

Table 1: Overview of all 8-K sections and events

Section	Item	Event
Registrant's Business and Operations	1.01	Entry into a Material Definitive Agreement
	1.02	Termination of a Material Definitive Agreement
	1.03	Bankruptcy or Receivership
	1.04	Mine Safety - Reporting of Shutdowns and Patterns of Violations
Financial Information	2.01	Completion of Acquisition or Disposition of Assets
	2.02	Results of Operations and Financial Condition
	2.03	Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Registrant
	2.04	Triggering Events That Accelerate or Increase a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement
	2.05	Costs Associated with Exit or Disposal Activities
	2.06	Material Impairments
Securities and Trading Markets	3.01	Notice of Delisting or Failure to Satisfy a Continued Listing Rule or Standard; Transfer of Listing
	3.02	Unregistered Sales of Equity Securities
	3.03	Material Modification to Rights of Security Holders
Matters Related to Accountants and Financial Statements	4.01	Changes in Registrant's Certifying Accountant
	4.02	Non-Reliance on Previously Issued Financial Statements or a Related Audit Report or Completed Interim Review
Corporate Governance and Management	5.01	Changes in Control of Registrant
	5.02	Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensatory Arrangements of Certain Officers
	5.03	Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year
	5.04	Temporary Suspension of Trading Under Registrant's Employee Benefit Plans
	5.05	Amendment to Registrant's Code of Ethics, or Waiver of a Provision of the Code of Ethics
	5.06	Change in Shell Company Status
	5.07	Submission of Matters to a Vote of Security Holders
	5.08	Shareholder Director Nominations
Asset-Backed Securities	6.01	ABS Informational and Computational Material
	6.02	Change of Servicer or Trustee
	6.03	Change in Credit Enhancement or Other External Support
	6.04	Failure to Make a Required Distribution
	6.05	Securities Act Updating Disclosure
Regulation FD	7.01	Regulation FD Disclosure
Other Events	8.01	Other Events
Financial Statements and Exhibits	9.01	Financial Statements and Exhibits

2 Theory

This section gives a short overview of previous work about forecasting stock prices with 8K filings, with other text data and with any type of data. Considering previous work will help to put the results of this research project into perspective in the discussion section.

2.1 Stock Price Forecasting with 8-K filings

There are few public research papers that made use of 8-K filings for stock price forecasts. However, the papers that are publicly available show promising results. For instance, Lee et al. could achieve an increase in accuracy by 10 percent when including text data from 8-K filings into a baseline model that only used financial metrics [5]. This study was also solving a classification task: predicting if a price will increase, decrease or stay almost the same. Since there were three possible classes, a random classification would have given a 33 percent accuracy. However, the best model of the study could achieve a 55 percent accuracy (on the test data). The main model used for this study was a random forest classifier because it outperformed other models such as multi layer perceptron and logistic regression.

Another study by Saleh et al. extended the research by Lee et al. [6]. In addition to the data from the 8-K filings, the researchers used text data from Twitter. Furthermore, they used CNNs and LSTMs instead of random forests. Again, with three possible classes, a random classification would have given a 33 percent accuracy. The best test accuracies were 51 percent without and 52.6 percent the Twitter data.

2.2 Stock Price Forecasting with Text Data

2.3 Stock Price Forecasting in General

3 Data

3.1 Sources

The data used for this text mining project comes from a variety of sources. The stock price data (with daily resolution) was retrieved with the financial API Tiingo. The SEC filings were downloaded from EDGAR, the official archive for SEC filings. The overview of companies by CIK, necessary to merge the SEC filings with the stock prices, was taken from the service Ranked and Filed. The industry categorization (SIC domain) for the companies was taken from SICCODE. Lastly, the overview of the 8-K events, used to extract the 8-K events from each filing, was taken from the SEC documentation.

3.2 Retrieval

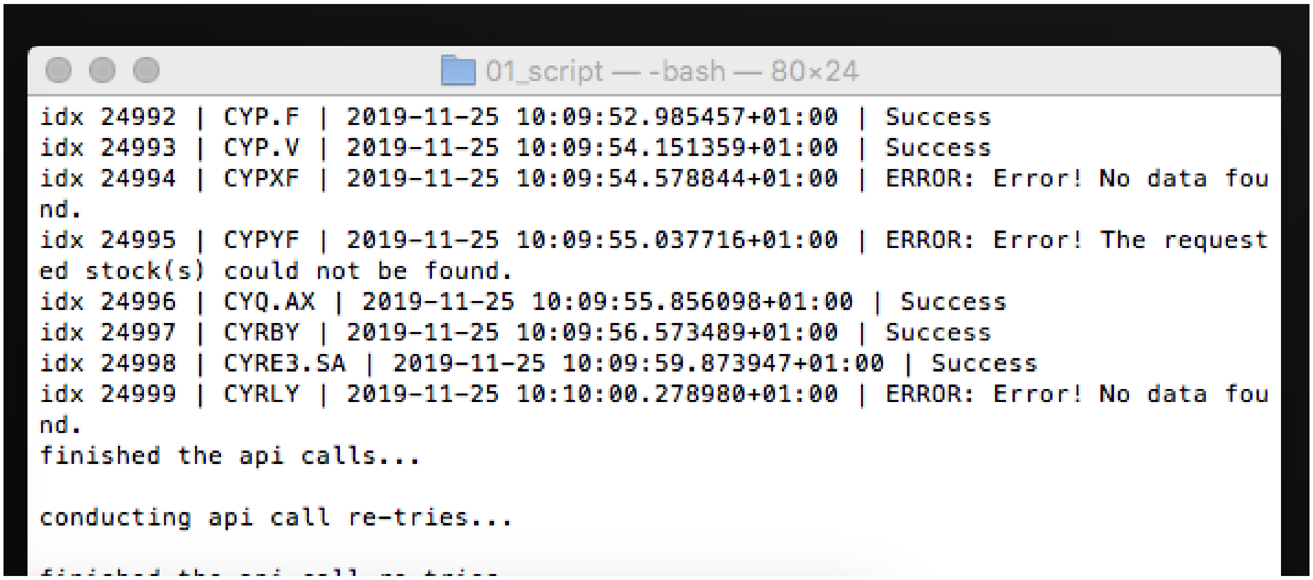
The dataset used for training the models was created with the following steps:

1. A list of all 8-K filings for the first quarter of 2019 was retrieved from EDGAR. This list contained a total of 16449 8-K filings, including the CIK number (to identify the company) and the filing date.
2. The list was merged with the data from Ranked and Filed to get the ticker symbol, exchange market and SIC number (representing the industry) of the companies corresponding to each 8-K filing.
3. All 8-K filings from companies that were not listed on the NASDAQ, the NYSE or the AMEX (according to Ranked and Filed) were removed. In particular, the removed 8-K filings corresponded to companies listed on NYSE ARCA, OTC or OTCBB. The resulting list contained a total of 9318 8-K filings.
4. All 8-K filings, for which no stock price data was available from Tiingo were removed from the list. The resulting shortlist contained a total of 8030 8-K filings.
5. For the 8030 8-K filings, the stock price data was extracted from the Tiingo data, making use of the filing date from EDGAR. For each filing, the percentage change from the closing price of the day before the filing date and the open price of the day after the filing date was computed as target variable.
6. The 8030 8-K filings from the shortlist were downloaded from EDGAR. When processing the data (see below) 8-K filings with more than 1 Mio. characters were removed due to file size limitations of the libraries that were used for processing. This left a total of 7975 8-K filings for training the model.

3.3 Processing

Each raw 8-K filing, a text file containing HTML code, was processed as follows. First, graphics and embedded PDFs were removed and HTML tags were removed as well. Second, the resulting text data was tokenized with the natural language processing library spaCy, using the English language model en_core_web_sm. Third, stop words, non-alphabetical tokens and tokens with only one character were removed. Fourth, the remaining tokens were lemmatized with spaCy. Later on, very rare and frequent tokens were removed as well. This will be covered in the section about hyperparameter tuning.

3.4 Descriptive Statistics



```
idx 24992 | CYP.F | 2019-11-25 10:09:52.985457+01:00 | Success
idx 24993 | CYP.V | 2019-11-25 10:09:54.151359+01:00 | Success
idx 24994 | CYPXF | 2019-11-25 10:09:54.578844+01:00 | ERROR: Error! No data fou
nd.
idx 24995 | CYPYF | 2019-11-25 10:09:55.037716+01:00 | ERROR: Error! The request
ed stock(s) could not be found.
idx 24996 | CYQ.AX | 2019-11-25 10:09:55.856098+01:00 | Success
idx 24997 | CYRBY | 2019-11-25 10:09:56.573489+01:00 | Success
idx 24998 | CYRE3.SA | 2019-11-25 10:09:59.873947+01:00 | Success
idx 24999 | CYRLY | 2019-11-25 10:10:00.278980+01:00 | ERROR: Error! No data fou
nd.
finished the api calls...

conducting api call re-tries...

finished the api call re-tries
```

Figure 1: A boat.

Figure 1 shows a boat.

3.4.1 General

3.4.2 Target Variable

3.4.3 Feature Variables

4 Method

4.1 Evaluation Metric

4.2 Baseline Models

4.3 Advanced Models

4.4 Hyperparameter Tuning

Train vs. test

4.5 Feature Relevance

5 Results

Table 2 summarizes the benefits and drawbacks ("Pros and Cons") of each approach.

Table 2: The pros and cons of Scala's optional classes

		Predicted				
		lg_dec	sm_dec	zero	sm_inc	lg_inc
True	lg_dec	107	46	53	32	78
	sm_dec	39	56	117	49	48
	zero	34	54	146	46	33
	sm_inc	55	59	102	65	43
	lg_inc	82	42	41	40	128

6 Discussion

7 Conclusion

8 References

References

[1] *Alpaca - Commission-Free API First Stock Brokerage*, en. [Online]. Available: `https://alpaca.markets` (visited on 11/23/2019).

[2] *SEC.gov | What We Do.* [Online]. Available: `https://www.sec.gov/Article/whatwedo.html` (visited on 11/23/2019).

[3] *SEC.gov | Form 8-K.* [Online]. Available: `https://www.sec.gov/fast-answers/answersform8khtm.html` (visited on 11/23/2019).

[4] W. Kenton, *8-K (Form 8k)*, en. [Online]. Available: `https://www.investopedia.com/terms/1/8-k.asp` (visited on 11/23/2019).

[5] H. Lee, M. Surdeanu, B. MacCartney, and D. Jurafsky, "On the Importance of Text Analysis for Stock Price Prediction.," in *LREC*, 2014, pp. 1170–1175.

[6] M. Saleh and S. Nair, "Neural based event-driven stock rally prediction using SEC filings and Twitter data," en, p. 8,

[7] J. Lee and H. Lee, "Predicting Corporate 8-K Content Using Machine Learning Techniques," *Graduate School of Business Stanford University*, 2008.

[8] C. Gleason, Z. Ling, and R. Zhao, "Selective Disclosure and the Role of Form 8-K in the Post-Reg FD Era," en, p. 62,