Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 208 (2022) 471–477

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Intelligent, Interactive Systems and Applications

Preliminarily Explore the Steps of Financial Big Data Analytics

Jiaxin Li[a,*]

[a]*Haidian Foreign Language Academy, Beijing, 100195 P.R.China*

Abstract

Big data analytics is difficult for people since there are so many factors that can influence the ways stock market change. Big data analytics is different from weather forecasting, which can be made by applying the laws of atmospheric change and mass experience and other comprehensive research. However, stock markets do not have a fixed law that allows people to do analysis. They are influenced by governments decision, information on the website, citizens' behaviors, etc. Hence, this paper helps people learn how to do big data analytics and stock market prediction by three steps, such as sentiment analysis, information extraction and cleaning, and three models for stock market prediction. The implementation process of big data analytics generally includes several stages, such as data acquisition and recording, information extraction and cleaning, data integration and presentation, selection of modeling and analysis methods, interpretation of results, effectiveness of evaluation results and monitoring.[1],[2] The result of the prediction is that LSTM model is the best model in prediction of the trend of stock market.

Keywords: Big data analytics; sentiment analysis; prophet model; automatic ARIMA model; LSTM model

1. Literature Review

In order to predict the stock market, previous studies had established a lot of methods and models for people to get the most accurate prediction of the actual stock market.

Li Yanchun [3] mainly studies the process modeling and implementation methods of big data analysis, and proposes a domain-oriented big data analysis framework with domain business as the center and multi-mode collaboration.

Jianghui Cai, Yuqing Yang [4] provide the technical system of big data analysis and processing and study deeply in text big data analysis, network big data analysis, multimedia big data analysis, and mobile big data analysis.

* Corresponding author. Tel.: +86-137-1775-3562.
E-mail address: ljx20050129@163.com

1877-0509
10.1016/j.procs.2022.10.065

Ying Wang [5] uses the ARMA model to fit and analyze the historical data of the opening price of bank of China stock and predict the opening price of the next three days.

Yue Chongyuan [6] designs and implements a system to analyze whether the sentiment of shareholders on the Internet affects the real stock market with big data. The final results show that there is a positive correlation between the impact of stock sentiment on stock market returns. And negative sentiment affects the stock market more than positive sentiment.

Jiawa Zhong, Wei Liu, Sili Wang, Heng Yang [7] summarize the development trend and application scenerios of text sentiment analysis through literature review. They also summarize, analyze the main model methods and applicatioin scenerios of text sentiment analysis from the dimensions of time and subject. In addition, researchers find out both pros and cons for every method.

Ping Chen, Lin Feng [8] analyze and comment on the aspects of emotion analysis, providing reference for the research on aspects of emotion analysis.

Yuxin Yang [9] adds characteristics like enterprise market share to data set, making sales forecasting model to give consideration to both enterpise's interior and exterior by using game theory. He also uses gradient ascending tree to further fit model residual and improve the ability of the combined model to extract trend and feature relationship of data samples based on time series analysis.

Zou Cunzhu, Luo Jiping, Baijing Shengyuan, Wang Yuanze, Zhong Changfa, Cai Yi [10] propose RNN (LSTM) model for stock time series prediction. Moreover, researchers further analyze and optimize the application of RNN model in order to fit the change law in the stock market.

Generally, a part of previous researches only focused on one method or model for big data analytics and stock martet prediction. There are few studies focused on the whole process of big data analytics. Therefore, this paper discusses mainly on the steps that help people to do financial big data analytics, providing information of sentiment analysis and using Microsoft data as an example to make stock market predictions.

2. Sentiment analysis

Sentiment analysis is the most difficult step in analyzing big data. The stock market is influenced not only by the market, but also by politics and people's behavior. Thus, Sentiment analysis is important. In order to obtain people's comments on the stock market, crawlers need to obtain corresponding information, so as to better predict the stock. Crawler simply means that people enter the corresponding webpage for data acquisition by forging browsers. Headers is one of the solutions to request crawl. When obtaining the data, people usually use BeautifulSoup.

When people have successfully retrieved the data, they get the HTML (hypertext mark-up language).

However, when we obtain the information successfully, we may not successfully gain the true news. What is especially terrible is that if market rumors are deliberately created by people with good intentions, then there is no doubt that investors who listen to these market rumors will become the victims and objects of market rumors. Hence, Investors need to analyze stock market news intelligently, not only to distinguish between true and fake news, but also to overcome the various news traps.

Most information on the stock market comes from three main sources: the news media, market rumors, and insider leaks. Market rumors are a big source of stock market news. Since the publishers are not bound by liability, the reliability is low. It may be designed to confuse people who do not have the ability to do judgement. Of course, not all rumors in the market are groundless. Some also have certain reference value, which need specific case and specific analysis. For market rumors, the key is to judge the situation and determine how credible they are.

Moreover, there is the "inside information". If it's real inside information, it must be high credibility, according to this operation will generally have considerable profit. Nevertheless, this kind of information is hard for ordinary retail investors to get, and there is a suspicion of "insider trading". In this way, people need to learn to filter out unnecessary distracting information.

3. Information extraction and cleaning

Information extraction and cleaning will be the most workload in the construction of data warehouse or data mining. When we get the data from website, we need to preprocess the data. "Dirty" data includes data that is

missing data values, contains errors or outliers, and has inconsistent data. The data gets dirty because of the lack of proper values at the time of data collection, different considerations at the time of data collection and data analysis, problems with data collection tools, different data sources, etc.

The main task of information extraction and cleaning:

- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization

There are three classification of measures, such as distributive measure (count, sum, minimum, maximum), algebraic (average/mean), and holistic (median, mode, rank). To better understand the data, we need to measure the central trend of the data and the degree of dispersion of the data.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

$$median = L_1 + (\frac{n/2 - (\sum f)l}{f_{median}})c \tag{3}$$

It also includes trimmed mean, which is the mean obtained by removing the mean values of high and low extremes, and mode (The most frequently occurring value).

3.1 Data cleaning

Data is not always complete. In database tables, many records have no corresponding value for the corresponding field. In this way, we need to complete the empty value. We can use a global variable, like unknown or $-\infty$, the average value of the property, the average of all samples belonging to the same class as the given tuple, and Inference-based methods such as Bayesian formulas or decision trees to fill in the empty value.

Furthermore, we also have to deal with noisy data. Noisy data is a random error or deviation in a measurement variable. This is usually caused by problems with data collection tools, data entry error, data transmission error, etc. In order to solve these problems, we can use binning, which works by first sorting the data and dividing them into boxes of equal depth, then smooth by the average of boxes, smooth by the median of boxes, smooth by the boundaries of boxes, and so on. We can also use regression, which smooth the data by fitting it to a regression function. Other methods include cluster or a combination of computer and manual inspection all can process noisy data.

The steps for data cleaning are first deviation detection, and then data transformation, which is to correct the deviation.

3.2 Data integration and transformation

Data Integration is consolidating the data from multiple data sources into a consistent store. When integrating multiple databases, redundant data often occurs. Some redundancies can be detected by correlation analysis.

$$r_{A,B} = \frac{\sum (A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B} \tag{4}$$

Careful integration of data from multiple data sources can reduce or avoid redundancy and inconsistencies in the resulting data, and thus improving the speed and quality of mining. When doing correlation analysis of classified data, we should use chi-square test to analyse.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{5}$$

A higher x^2 value means a higher probability that the two variables are related, and the larger the difference between the expected value and the observed value, the larger the value will be. But correlation is not mean causation.

Data transformation is to transform or unify data into a form suitable for mining.

3.3 Data reduction

There are always a lot of data in data warehouse. It takes a long time to analyze and mine complex data on the whole data set. Data reduction can be used to produce reduced representations of data sets that are much smaller but yield the same analytical results.

Common data reduction strategies include data cube aggregation, dimensional reduction, numerical reduction, discretization and concept stratification.

3.4 Data discretization

Data discretization means the range of continuous attributes is divided into intervals, and then the valid specification data is discretized, and then the numerical value is used for further analysis. The conceptual layering of numerical attributes can be constructed automatically from the analysis of numerical distributions by means of binning, which is the same concept mentioned above.

4. Models

- Prophet
- ARIMA Model
- LSTM Model

4.1. Prophet

Prophet is one of the time series predictions and is easy to enforce. It mainly composed of trend, seasonality, and holidays.

The input of prophet is a two-column data box, including date and destination.

4.2. ARIMA model

ARIMA is the single product autoregressive moving average process. There are three basic principles, such as p, q, d. "p" means the past value used to predict the future value; "q" means errors in past predictions used to predict future values; "d" means the order of the difference.
The advantage of ARIMA model is that the model is very simple and requires only endogenous variables without recourse to other exogenous variables. However, the drawbacks are that time series data is required to be stable or stable after differencing. Also, it can only capture linear relationships, but not nonlinear relationships.

4.3. LSTM model

LSTM model is able to store information. There are three control gates
(1) Forget gate: it chooses to forget certain information from the past;
(2) Input gate: it remembers some information in the present;
(3) Output gate: it outputs the information.

5. Prediction results

The data need to be collected in order to predict the results. This essay selects data from Microsoft from Quandl data set. It contains the data from the closing prices between March 27, 2013 to March 27, 2018.
The date between March 27, 2017 to March 27, 2018 will be the part which this essay did the prediction.

Table 1.The import data of Microsoft.

Date	Open	High	Low	Close	Volume
2017-03-27	64.63	65.22	64.35	65.1	18614662
2017-04-27	68.15	68.38	67.58	68.27	32219234
2017-05-26	69.8	70.22	69.52	69.97	19644260
2017-06-27	70.11	70.18	69.18	69.21	24862560
2017-07-27	73.76	74.42	72.32	73.16	35518251
2017-08-28	73.06	73.09	72.55	72.83	14112777

In Table 1. , the data shows a part of the data from March 27, 2017 to March 27, 2018. The reason why there is no data for May 27, 2017, August 26, 2017, and August 27, 2017 is because on weekends and holidays, the stock markets will be closed. The, the models discussed previously need to be used for the prediction results. This essay mainly focuses on the prediction by three models:

- Prophet
- Automatic ARIMA model
- LSTM model.

5.1. Prophet prediction result

In Fig. 1. (a) shown below, the orange part is the real stock market price, and the green part is what prophet model predicts. The RSME value of the model below is 9.07. Hence, it can be concluded that the prophet model prediction on data about stock market is relatively bad.

5.2. Automatic ARIMA model

In Fig. 1. (b) shown below, the orange part is the real stock market price, and the green part is what auto-ARIMA model predicts. The RSME value of the model below is 10.41, which is still far from the real model. Further, this

value is greater than the value of RSME of prophet model. Therefore, it can be concluded that the auto-ARIMA model prediction on data about stock market is poor.

5.3. LSTM model

In Fig. 1. (c) shown below, the orange part is the real stock market price, and the green part is what LSTM model predicts. The RSME value of this model is 1.77. So, the prediction of LSTM model is much more closer to the real model than what the prophet model and automatic ARIMA model predict. Thus, it can be concluded that the LSTM model prediction on data about stock market is the best, compared with prophet and automatic ARIMA model.

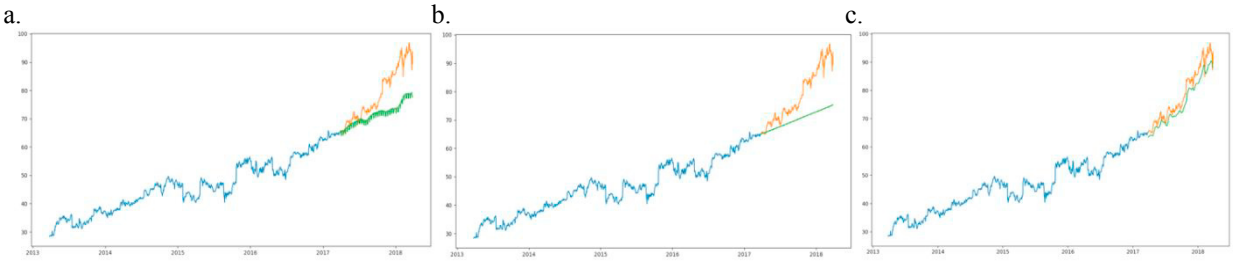

Fig. 1. (a) prophet model prediction; (b) auto-ARIMA model prediction; (c) LSTM model prediction.

6. Conclusion

To sum up, this essay mainly discusses the preliminary steps of big data analysis, showing how to do sentiment analysis. In the technical part, this essay, using prophet, automatic ARIMA model, and LSTM model, provides an example of the prediction of Microsoft stock market. It finally concludes that LSTM model is the best model in prediction of the trend of stock market. This is because the RSME for LSTM model is the lowest, as shown in Fig. 2.

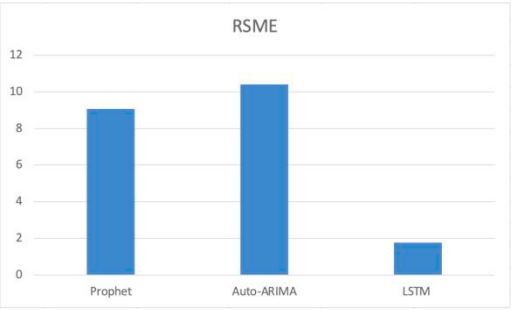

Fig. 2. bar graph for RSME.

References

[1] Xinyu Jiang, Bodi Wang. Preliminary Study on the Big Data. Analytics and Its Adaptability in Intelligence Studies. Department of Information management Beijing 100871.
[2] Mingjiang Li, Yi Zheng. Research on General Process Configuration and Analysis Method of Big Data. Cnooc Energy Development Co., LTD. Engineering Technology Branch.
[3] Li yanchun. Research on Key technologies of Big Data Analysis Process Modeling. June 2020.
[4] Jianghui Cai, Yuqing Yang. Overview of big data analysis and processing. College of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China.
[5] Ying Wang. Analysis and forecast of stock price based on ARMA model. School of Management, Shanghai University of Engineering Science, Shanghai 201620, China.

[6] Yue Chongyuan. The Design and Implementation for Analysis System on the Impact of the Investor's Sentiment on the Change of the Stock Market Based on Big Data. Hangzhou University of Science & Technology Wuhan 430074, P.R. China January, 2018.
[7] Jiawa Zhong, Wei Liu, Sili Wang, Heng Yang. A review of text sentiment analysis methods and applications. Northwest Institute of Eco-Environment and Resources, CAS Lanzhou 730000, University of Chinese Academy of Sciences Beijing 100190.
[8] Ping Chen, Lin Feng. Review of aspect extraction in sentiment analysis. College of Computer Science, Sichuan Normal University, Chengdu Sichuan 610101, China.
[9] Yuxin Yang. Detailed Market Description and Forecasting Based on Big Data Analysis. Beijing Jiaotong University.
[10] Zou Cunzhu, Luo Jiping, Baijing Shengyuan, Wang Yuanze, Zhong Changfa, Cai Yi. Stock Time Series Prediction Based on Deep Learning. 2019 2nd International Conference on Mechanical, Electronic and Engineering Technology (MEET 2019).