Predict the Pharmaceutical Stock Price with Google Trends Data

Zijun Cui, Junhong Li, Yiwen Ma
CS-401 Applied Data Analysis, EPFL, Switzerland, Project Report

Abstract—**Inspired by the prediction method using Google Trends data, we want to apply the same method to the prediction of pharmaceutical stock price. We collect the data include 3 keywords and 2 categories related to the pharmaceutical industry from Google Trends. Using AR-model, we obtained the prediction results without Google Trends data and with Google Trends data as the eigenvalues respectively. By comparing the prediction results, we find that at some periods these Trends data can improve the pharmaceutical stock price prediction.**

I. Introduction

If we are to choose a list of keywords for 2020, some medical-related keywords seem to have no reason not to be on the list. This year, affected by the COVID-19 pandemic, medical-related topics exploded at a rate that has never been seen in previous two years. Discussions about the ”pandemic” and ”vaccine” have continued for several months and have not stopped. Even the discussion about some rarely seen drugs such as ”Remdesivir” has begun to enter the public's field of vision. This is a phenomenon that is hard to be seen in normal times. Therefore, the huge impact the pandemic has brought to the world is undeniable. This is not only reflected in the pharmaceutical industry, but also in financial fluctuations.

In Predicting Present with Google Trends[1], the authors show how to use search engine data to forecast near-term values of economic indicators. The basic idea is to add some Google Trends data to the features of the prediction model to explore whether the addition of these data can improve the predicted results. In the paper, the author mainly used the auto-regressive model, which is a method specially used to process continuous time series.

Based on the background, since the auto-regressive model is a very suitable model for stock price prediction, as an extension, we want to combine Google Trends data to explore the impact of relevant search keywords on the stock price of the pharmaceutical industry.

II. Data Collection and Processing

There are two types of data required for the experiment: historical stock price and Google Trends data.

We selected the five most well-known stocks from the pharmaceutical industry, respectively GlaxoSmithKline, Johnson & Johnson, Merck & Company, Inc., Pfizer, Inc. and Sanofi. The stock price was collected from ”Yahoo! Finance” weekly from 2019-12-01 to 2020-12-01. Besides ”Date”, only ”Adj Close” was kept as the close stock price, and the rest columns (such as ”Start”, ”Close”, etc) were deleted in order to get a cleaner dataset.

With the same period and frequency, the non-real-time relevant Google Trends data were collected from the official website. As Google Lab News introduces, Trends data is an unbiased sample of our Google search data. It is anatomized, categorized, and aggregated. This allows us to measure interest in a particular topic across search, from around the globe, right down to city-level geography. [2] In our case, since the research objects are the stocks in the pharmaceutical industry, it is better to select global as the geographical range of the data. Also, it is easy to connect the pharmaceutical industry with keywords like ”virus”, ”flu”, and ”vaccine”. To get a broader perspective, we also download two categories of ”health” and ”finance” from Google Trends.

III. Dataset Description

With a weekly frequency, we collected 53 data for each stock from 2019-12-01 to 2020-12-01.

Table I
Summary of stock price

Stock	Mean	Median	Std	Max	Min
GSK	3.675	3.678	0.078	3.819	3.471
JNJ	4.959	4.975	0.050	5.022	4.766
MRK	4.379	4.379	0.056	4.486	4.245
PFE	3.523	3.539	0.070	3.646	3.285
SNY	3.881	3.899	0.070	3.975	3.602

We visualize the trends of the five stock price. Also, we plot the score of three Google Keywords and two Categories that vary during the whole period to get an initial glance of the correlation between them.

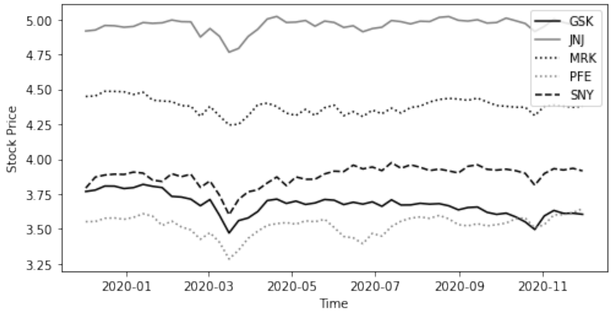

Figure 1. Trends of the five stock prices the pharmaceutical industry in 2020

According to the plot, the five stocks which all come from the pharmaceutical industry have similar trends over the year.

There are two obvious recession periods that could correspond to the two waves of the pandemic, respectively from March to April and from October to November.

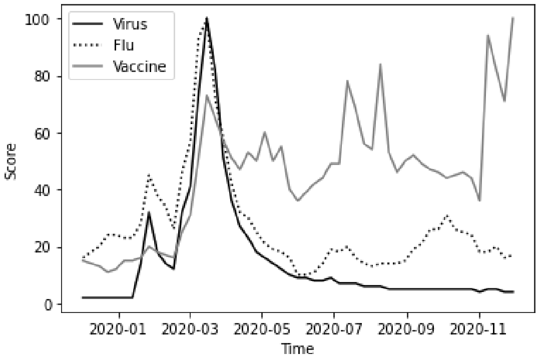

Figure 2. Google Trends Keywords 'Virus', 'Flu' and 'Vaccine' in 2020

For both of the keywords and categories from Google Trends data, it could be observed that the corresponding trends data during these two recession periods have also had a large wave, but with a negative correlation in March. During the first period, all three keywords we used to have a burst growth, while during the second period, only 'Flu' and 'Vaccine' show a significant wave in the score.

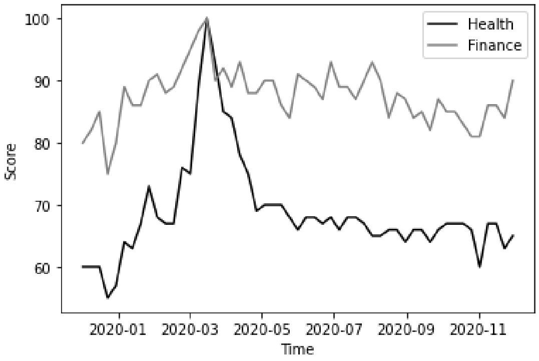

Figure 3. Google Trends Categories 'Health' and 'Finance' in 2020

IV. Model Selection

Auto-regressive model (AR Model) is the process of using itself as a regression variable, that is, using the linear combination of random variables at a certain time in the previous period to describe the linear regression model of random variables at a certain time after.[3]

$$Y_t = c + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \varepsilon_t \qquad (1)$$

where $\varphi_1, \ldots, \varphi_p$ are the parameters of the model, c is a constant, and ε_t is white noise.

In our project, since it is uncertain whether there is a cyclical effect in the forecast, we try with two form of models. The first model is $y_t = b_1 y_{t-1} + e_t$, i.e. the prediction is based on the week before. The second the model is $y_t = b_1 y_{t-1} + b_4 y_{t-4} + e_t$ because we consider the whole month as a cycle which includes four weeks.

To compare the performance of these two models, we do some experience with the stock GSK. The performance of the prediction result for the two models are reported as mean square error. It can be found that no matter whether Google Trends data is added as eigenvalues, the MAE of the model that only uses y_{t-1} for prediction is smaller. Therefore, we choose the first model for prediction.

Table II
Comparison between the MAE of two models for GSK

Model	Base MAE	Trends MAE
$y_t = b_1 y_{t-1} + e_t$	0.03423	0.03773
$y_t = b_1 y_{t-1} + b_4 y_{t-4} + e_t$	0.03615	0.07762

V. Implementation and Comparison

In this project, we implement three prediction models for the stock price: Base AR-1 model, Model with Searchers Data (Keywords), and Model with Category Data.

A. Base prediction

First of all, we try to do a base prediction without Google Trends variables in order to find out whether these variables have the potential to improve the prediction of the present or near future. As mentioned in Model Selection, we simply select the AR-1 model on the log of adjusted close price as our baseline regression. To check the prediction result, we calculate the mean absolute error for each stock.

Table III
Mean Absolute Error of Base Prediction

Stock	Base MAE
GSK	0.03423
JNJ	0.02968
MRK	0.03699
PFE	0.04745
SNY	0.0339

B. Searchers prediction

Then, we add the Google Trends Keywords to the model: 'Flu', 'Virus', 'Vaccine' and the combination of the variables. Similarly, we calculate the MAE of all the prediction results.

Table IV
Mean Absolute Error of Trends Prediction

Stock	Virus MAE	Flu MAE	Vaccine MAE	Trends* MAE
GSK	0.03766	0.03852	0.03794	0.03773
JNJ	0.02736	0.02927	0.0348	0.03284
MRK	0.03462	0.0325	0.03398	0.03726
PFE	0.04155	0.03521	0.04127	0.03565
SNY	0.02609	0.027	0.03805	0.02932

* Trends = [Virus & Vaccine & Flu]

We can see that the Google Trends variables somehow perform well on the prediction of MRK, PFE and SNY.

However, the prediction of GSK and JNJ show no signs of improvement. To look deeper into the result, we try to visualize our prediction to have a subjective justice that what Google Trends variables may contribute to the prediction.

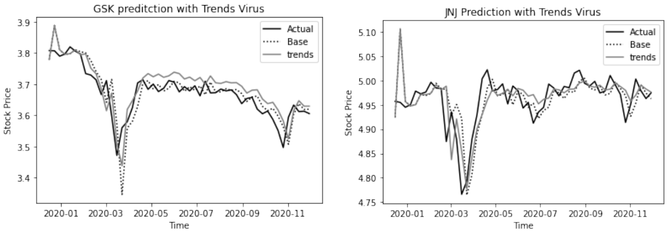

Figure 4. Virus prediction GSK and JNJ

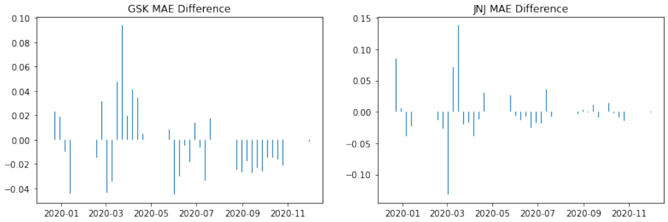

Figure 5. Absolute Error difference GSK and JNJ

Figure 5 illustrate the difference of absolute error between base prediction and trends prediction. A positive value indicates that trends prediction perform better than base prediction. It can be seen that, in terms of GSK and JNJ, Google Trends data may improve the results during certain periods.

As mentioned in the dataset description, there are two obvious recession periods of the stock price around March and November. According to figure 4 and figure 5, we can see that the base predictions for GSK and JNJ do not perform very well during the two periods. That's because turning points in economic series are usually much harder to forecast[1]. Thus, we want to zoom in the two periods to research the influence of the Google Trends data on the stock price prediction. So we choose the two time periods from 2020.03.01 to 2020.05.01 and from 2020.09.15 to 2020.11.01 based on our observation. In the first period, we use all three keywords as the trends data to forecast, while for the second period, we only use 'Flu' and 'Vaccine' as the trends variable because the score of 'Virus' seems to be stable during this period.

Table V
MEAN ABSOLUTE ERROR FOR THE FIRST RECESSION PERIOD, 2020

Stock	Base MAE	Trends MAE
GSK	0.12119	0.10255
JNJ	0.11749	0.11579
MRK	0.07502	0.07451
PFE	0.11998	0.10765
SNY	0.11982	0.10512

Table VI
MEAN ABSOLUTE ERROR FOR THE SECOND RECESSION PERIOD, 2020

Stock	Base MAE	Trends MAE
GSK	0.05967	0.07988
JNJ	0.0235	0.02424
MRK	0.03138	0.02612
PFE	0.02679	0.02754
SNY	0.03081	0.02684

Table V and Table VI show the MAE of base prediction and trends prediction around the two turning points. Around the first turning point, trends variables can improve the MAE for both GSK and JNJ. However, around the second turning point, we cannot see an improvement in MAE for the two stocks while there is an improvement for MRK and SNY.

C. Category prediction

What we discussed above is how searchers(keywords) Google Trends may affect our prediction. Besides keywords, Google also offers trends data for a category such as Science, Shopping, etc. Here we choose 'Finance' and 'Heath' as our new eigenvalues, which seem to be related to the pharmaceutical stock price fluctuations.

Table VII
MEAN ABSOLUTE ERROR FOR THE CATEGORY PREDICTION

Stock	Base MAE	Health MAE	Finance MAE	Trends MAE*
GSK	0.03423	0.03902	0.03736	0.04299
JNJ	0.02968	0.03035	0.0301	0.03234
MRK	0.03699	0.0351	0.03578	0.03319
PFE	0.02679	0.05146	0.03834	0.04018
SNY	0.0339	0.03004	0.03388	0.03075

* Trends = [Health & Finance]

Table VII presents the MAE of category trends prediction. It can be seen that the category trends data can improve the prediction for MRK and SNY but still cannot make an improvement for GSK, PFE and JNJ.

VI. CONCLUSION

According to the prediction results, these Google Trends data can improve some of the prediction of pharmaceutical stock price. But for some certain stocks(like GSK and JNJ), it does not perform well globally. This also indicates that the factors affecting stock prices are complex, which may contain policy, finance and many other elements.

In addition, when looking into prediction around some turning points, Google Trends data may contribute to improve the prediction results. During the two wave of pandemic, using proper combinations of 'Flu', 'Vaccine' and 'Virus' as eigenvalues can effectively reduce the MAE of the prediction. This also reflects the fact that the public's attention to the pandemic during the COVID period has a subtle connection with the fluctuation of pharmaceutical stock price.

Therefore, this extension research provides us with a new perspective to study stock price fluctuations. That is using search engine data to study the link between some hot topics and the stock price. And for this purpose, the selection of search keywords is very critical to obtain enlightening results.

References

[1] Hyunyoung Choi and Hal Varian. “Predicting the present with Google Trends”. In: *Economic record* 88 (2012), pp. 2–9.

[2] Simon Rogers. *What is Google Trends data — and what does it mean?* https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8. 2016.

[3] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. 3rd ed. 2011. Springer, Nov. 2010. ISBN: 9781441978646.