Google Flu Trends: An Analysis on Its Lapses and What We Learned from It

GROUP 6

10/03/2019

BAX 401 Business Analytics

Zunfeng Huang
Mitesh Jain
John Elmar Loretizo
Mayank Mani
Shalini Mishra
Prakriti Rastogi

Table of Contents

EXECUTIVE SUMMARY

In 2008, Google aimed to reduce lag in reporting flu cases so as to facilitate early detection and rapid response. With that intention, they developed a prediction model for possible flu cases based on queries from their search engine. Furthermore, if implemented correctly, this initiative had the potential to make resource allocation easier for different institutions by providing actionable insights well ahead of time. However, in 2013, the model doubled the projection for weekly flu cases against that of the U.S. Center for Disease Control and Prevention (CDC). The incident was initially attributed to the change in search query trends due to growing media coverage about creating flu awareness. To understand the root cause, we have come up with the following theories which can be classified into three main points:

- ***GFT only considered one variable.*** During the creation of Google Flu Trend (GFT), the team only used search queries as a metric without considering other key variables such as demographics.
- ***GFT was not regularly updated.*** This is particularly evident when the model underestimated the number of cases when H1N1 was becoming a pandemic in the U.S. Several other instances have also been discussed concerning this.
- ***There was an issue of overfitting.*** Even queries without direct relation to flu had been included in the creation of GFT since there was no filtering done.

While it's becoming the norm for us to use data in making decisions, the fiasco by GFT reminds us to take any model with a grain of salt and expect that it will fail in the future. With this comes the understanding that we, as the creator of these models, must interfere at certain times to constantly check, validate, and correct the direction of these models that we are using.

INTRODUCTION

Influenza, commonly known as the flu, a contagious respiratory infection caused by influenza viruses results in 8.5% of the US hospitalizations per year. Traditionally, the U.S. Center for Disease Control and Prevention (CDC) relies on both virological and clinical data to survey influenza and publish weekly national and regional reports (Ginsberg et. al, 2009). However, there is a lag of 1-2 weeks. Google launched the GFT Project with the idea of leveraging its 500 million+ search queries to estimate weekly influenza activity with only one to two-day lag.

Data

With the data from 2003 to 2008, 50 million most common ILI-related search queries in the US were selected by the GFT team. For 9 regions, the average percentage of all ILI related patient visits to physicians on a weekly basis is publicly posted by CDC. No data was available for weeks out-side of the influenza season. CDC does not provide weekly data for states to public.

Model

Each of the 50 million queries were regressed against percentage of ILI visits for the 9 regions separately, and then sorted on highest correlation across locations. The top-n queries were taken in combination and tested to arrive at an appropriate value of 'n' that produced the best-fit model.

Result

Model was trained on the data points between 2003 and 2007, obtaining a mean correlation of 0.90 (min = 0.80, max = 0.96, n = 9). Out-of-time testing was performed for 2007-08 data resulting in a mean correlation of 0.97 (min = 0.92, max = 0.99, n = 9).

What went wrong?

After extensive testing, Google published its first estimates of flu cases in 2009. Unfortunately for Google, that was the year when a different strain - the H1N1 or more popularly known as swine flu - started spreading in various parts of the world. As people got more curious about the swine flu, trends that affected Google's model went down, thus resulting in a gross underestimation of possible number of flu cases at the start of the swine flu pandemic.

Two years after Google fixed its algorithm because of the swine flu, in 2011, GFT started overestimating ILI related cases. As 2013 hit, the company realised that their once precise model had been overpredicting numbers now for 100 out of 108 weeks. In 2013, at the peak of the post-Christmas-flu season, GFT missed the mark by as much as 140%.

PROBLEMS

1. Millions of search queries used to explain a much smaller and limited ILI dataset
    - CDC ILI data not available for non-influenza season was in-turn predicted by the model itself without any feature to account for the impact of seasonality
    - GFT used state-reported UTAH ILI releases (42 data points) to validate the prediction for 50 states.

2. The answer to these questions kept on changing over the years: *How Google users search? How google search engine works?* These modifications and improvements were never factored into GFT algorithm
3. Steadfast way of query processing restricted the movement of new type of search queries into the dataset
    - Fixed 'n' top queries: GFT team considered top 45 search queries for prediction of flu trends because among all n-combinations, it fitted the best with ILI data (2003 -2008) used for validation. They never revisited this number
    - Order over Semantics: *'Flu symptoms', 'the Flu symptoms', 'symptoms of Flu'.* All the above web search queries mean the same, but the algorithm would consider each of them as a separate search query
4. Excessive reliance on one variable, query fraction
    - GFT was convinced that web-search queries were enough to explain the flu trends. The data used lack dimensionality
    - Aggregating z scores at regional level and then finding out top 45 queries dilute regional variability. Due to the wide variability in regional level data, it is not appropriate to apply the national baseline to regional data as per CDC
    - Error from week 1 carried over to week 2 and so on. As a result, error term snowballed and as time progressed, prediction became more unreliable

RECOMMENDATIONS

Model with a grain of salt: The model is as good as the people making them.

1. Adding checkpoints in the algorithm: As a precursor to implementation of any model which considers millions of data points through complicated techniques, google should have monitored the model on a regular basis. Another feasible way could have been to include a variable which could act as a penalizing factor if the value of prediction went above or below a set limit based on past trends
2. Refining the dataset: The possibility that one user or multiple users from the same IP address could have searched for flu symptoms multiple times was overlooked. A thorough analysis of clickstream data could have streamlined the query fractions better.
3. Avoiding dilution of regional variability: Instead of identifying top queries at national level, identifying region-specific queries using z scores would have been more efficient
4. More robust dataset: The data produced by google for the trends combined with the CDC reports or other up-to-date data can produce results.

CONCLUSION

The data produced by Google for the trends, though incorrect but was useful and if the data was combined with the CDC reports or other up-to-date data, then it could have produced results better than any of them individually as suggested by the article as well.

REFERENCES

Butler, Declan (2013). When Google got flu wrong, Nature Journal, Vol. 494, Iss. 14th February, pp. 155-156. Retrieved from https://www.nature.com/news/when-google-got-flu-wrong-1.12413.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., Brilliant, L. (2009). Detecting influenza epidemics using search engine query data, Nature Journal, Vol. 457, Iss. 19 February, pp. 1012-1014. Retrieved from https://www.nature.com/articles/nature07634#online-methods.

Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014, March 14), The Parable of Google Flu: Traps in Big Data Analysis, Science*, Vol. 343, Iss. pp. 1203-1205. Retrieved from http://science.sciencemag.org/content/343/6176/1203

Overview of Influenza Surveillance in the United States (2018). Retrieved from https://www.cdc.gov/surveillancepractice/index.html