Does Weather Impact Internet Search Behavior for Medical Symptoms?

Anthony Le ale@gatech.edu | Arti Ranjan aranjan39@gatech.edu | Benjamin Lott blott7@gatech.edu | Honya Elfayoumy helfayoumy3@gatech.edu | Keerthi Vishal Poludasu kpoludasu3@gatech.edu

Problem

Everyone is exposed to weather. Change in weather can lead to different responses within the human body that can trigger varying health symptoms. For example, COVID-19, a potential factor cited in some studies is cold and dry conditions - leading to a quicker spread.

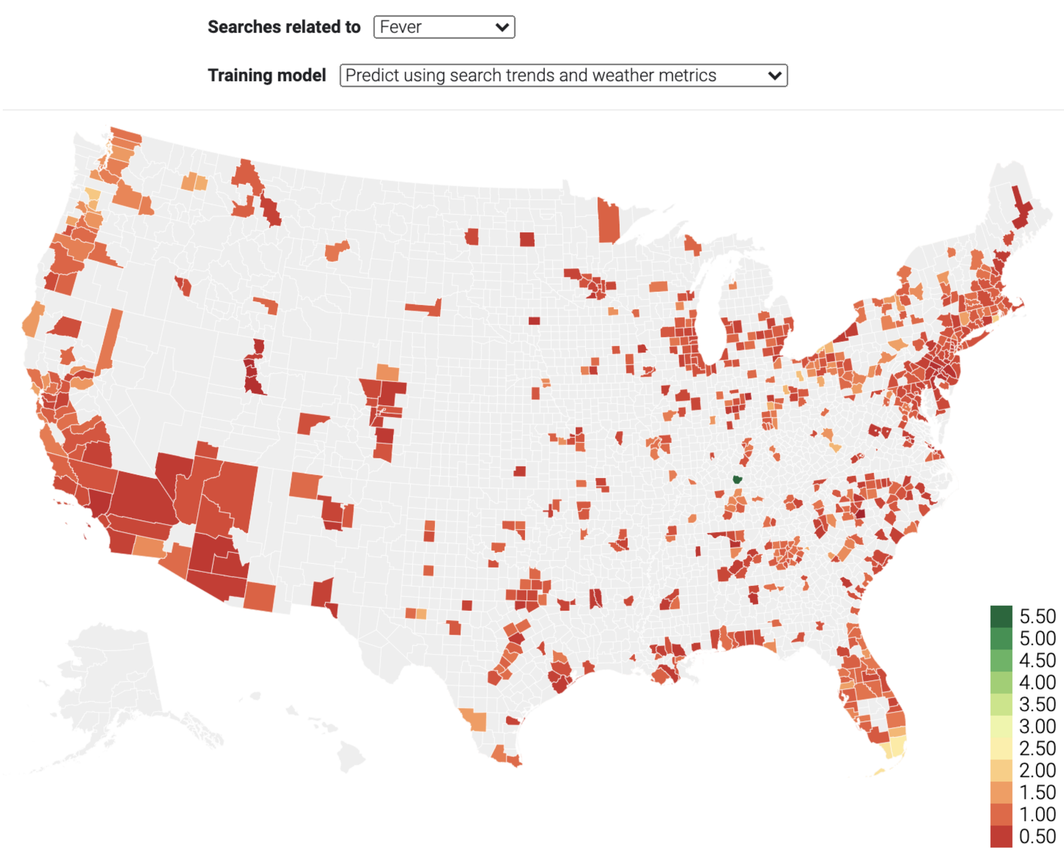
Data

Data characteristics?

Both csv files contain temporal data.
Google Search Symptoms 2017-2020: 10.1 GB, 3,081,476 rows, 436 variables
Weather Data 2019-2020:
61.3 MB, 1,149,960 rows, 12 columns

How did we get it?

We utilized data publicly available from Google Cloud by downloading csv files.



Approach

We utilized ARIMA (time series) forecasting. We also used interactive visualizations showing the RMSE and MAPE for each county for different symptoms.

What is new in our approach?

Our approach is new in that we are explored and analyzed the impact of temperature and humidity on the U.S population's internet search behavior for medical symptoms and provided interactive visualizations and reproducible methods - lacking in other studies.

Why is it important and why should you care?

Many people utilize the internet as a medical resource. With the modern digitization of information, potential trends and conditions could be analyzed. Attaining promising results could help the government and medical agencies better prepare for patient care by reducing health costs or creating specialized tools given the geolocation and climate.

Why do we think they can effectively solve our problem?

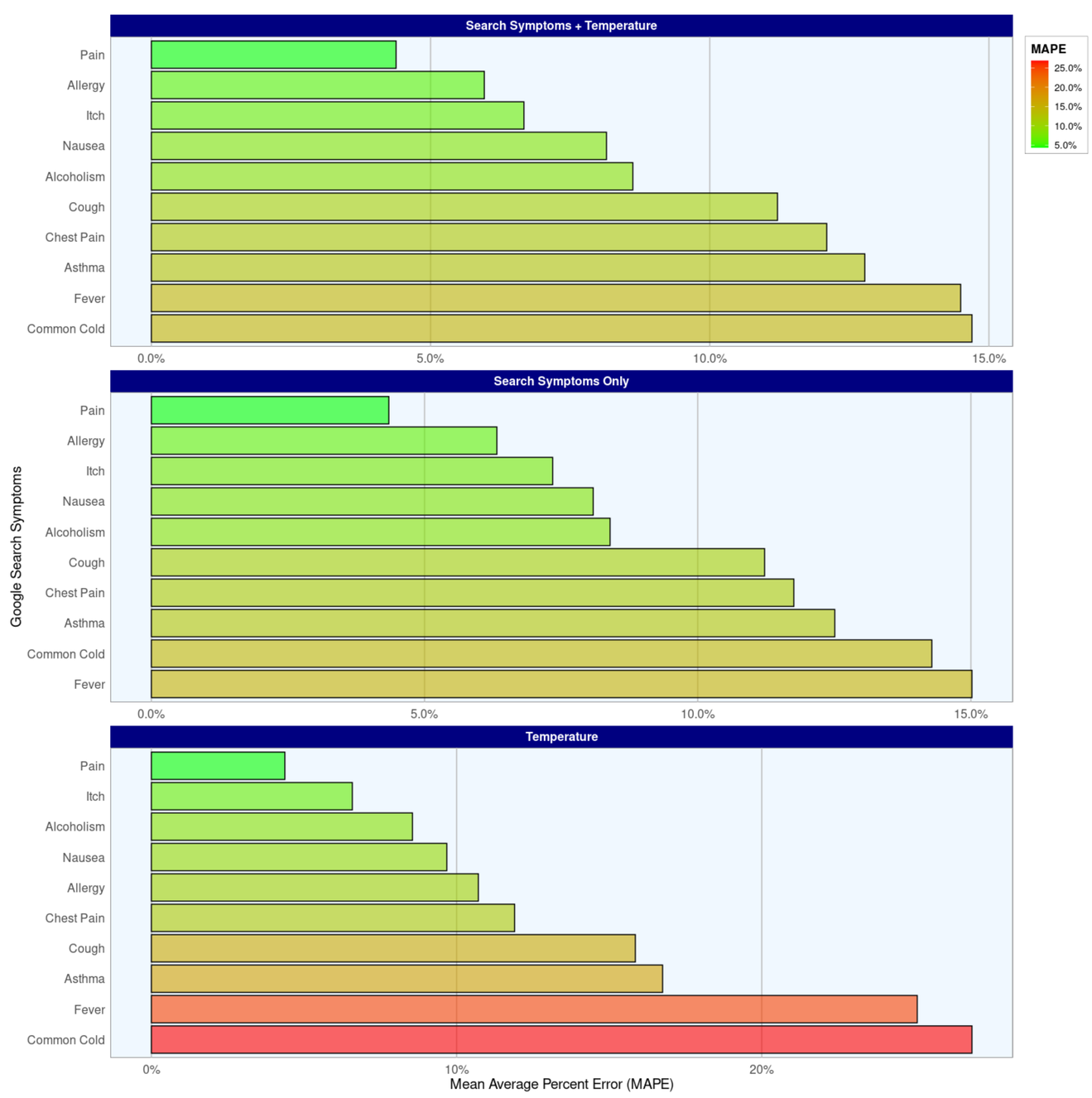
ARIMA allows us to use autoregressive and moving average components so it helps model the observations from previous time steps to predict the value at the next time step. It also helps capture smoothed trends in the data - allowing important patterns to stand out. RMSE and MAPE were both used to determine the accuracy of our ARIMA model.

How do they work?

Root mean square error (RMSE) is the squared distance measure between the predicted numeric target and the actual numeric answer - the smaller the RMSE, the more accurate the model.
Mean average percent error (MAPE) is the measure of predicition accuracy. It is used to calculate error in a statistical forecast.
Autoregressive Integrated Moving-Average (ARIMA) is a statistical analysis model that uses the time series data to better understand the dataset or predict future trends. It is a form of regression analysis that gauges one dependent variable's strength relative to other changing variables.

How did we evaluate our approaches?

We evaluated our approaches by analyzing the MAPE and RMSE which would let us know the accuracy of our models. For each symptom we used three different models of model variations by changing regressors:
Search Symptoms + Temperature + Humidity
Search Symptoms
Temperature + Humidity



Results

Across 27,103 models, 27% of our ARIMA forecasting models have less than 0.3 RMSE.

How do our methods compare to other models?

We looked at other models such as XGboost. In the example, XGBoost regression does not capture the trend well when predicting Google Search Index for Pain. However, ARIMA can take the trend into account for its prediction. We concluded that the average MAPE across predictions for all counties using XGBoost was 64.6%. Given these results, ARIMA was proven to be the most accurate.