COMP 551 Assignment 1

Javier Tobar, Maja Milinkovic, Zoha Khan
Group 16
October 21, 2020

Abstract

In this project, we investigated whether spikes in COVID-19 related hospitalizations in the United States could be predicted from Google search data using two supervised learning approaches: k-nearest neighbors and decision trees. The k-nearest neighbors model outperformed the decision tree on this problem. Additionally, we examined whether regional and/or temporal trends in the data exist. We found that certain symptoms were searched more frequently in certain states, and within broader regions of the country. In general, COVID-19 hospitalizations seem to have decreased over the summer months compared to March and August, and are now rising again.

Introduction

Prior studies have shown that information about populations can be extracted from Google search trends (Ayers et al. 2013). Previous attempts made by Google to predict seasonal flus from search trends were abandoned, but recent studies suggest this project may merit revisiting (Kandula and Shaman 2019). Given the severe global impact of COVID-19, the ability to predict potential spikes in hospitalizations using predictive metrics such as what people are searching would be a valuable way to adequately allocate resources to communities/hospital systems in need.

In this project, we investigated whether spikes in COVID-19 related hospitalizations in the United States could be predicted from Google search data using two supervised learning approaches. We compared the performance of K-Nearest neighbors and decision trees on this problem. Additionally, we looked at regional and temporal trends in search terms in 16 states across 29 weeks.

When people are sick they often search the symptoms they are experiencing prior to seeing a doctor or going to the hospital. Therefore, we predicted that periods of high symptom searches will temporally precede high hospitalizations. Additionally, because people of air travel restrictions, people are more likely to travel by car to nearby states. Therefore, we expected that geographically close states will have similar trends in search terms.

Datasets

Data was obtained from publicly available datasets collecting information on Google searches performed in the United States, and COVID-19 related hospitalizations. (LLC 2020) We restricted the data to regions in the United States present in both datasets, and looked at the period of time between March 3, 2020 and September 21, 2020.

Search Dataset The dataset of searches contains data from the United States on the frequency of various health-related terms (including symptoms and conditions. We thresholded the data at 67%, considering only those features containing $\geq$67% valid values.

Hospitalization Dataset The dataset initially contained global hospitalization data including features such as number of people on ventilators, economic support index, and whether public transit was closed in that region. We restricted the dataset to only the US regions also present in the Search dataset, and chose to look at only the new and cumulative hospitalizations, at the weekly resolution. We thresholded this data at 67% as well, leaving us with 16 regions containing valid data.

To account for region- and date-specific normalization constants applied by the authors of the dataset, we standardized the data to have a mean of 0, and unit variance, using the StandardScaler module from Python's sklearn package (Pedregosa et al. 2011).

Results

Dimensionality reduction

To reduce the dimensionality of the data for better visualization, we performed principal component analysis. Using the first two principal components, we were able to capture 83.5% of the variance in the data. We first looked at trends in the data when grouping the components by state. Some states (HI, NE) show clear separation from the others (2a).

Additionally, we looked at the reduced-dimension data grouped by week. No clear linear separability can be seen in the data grouped by Date, however it does appear that the dates pre-pandemic were more spread out compared to those post-pandemic. As we can see, the more recent dates (warmer tones) of the graph are clustered towards the lower left quadrant while the earlier dates are more spread out. As

expected, there does seem to be a positive relationship between the onset of the pandemic and the tendency to search for symptoms related to COVID-19.
Adding a third component to the data would increase the explained variance to 94.9% (1).

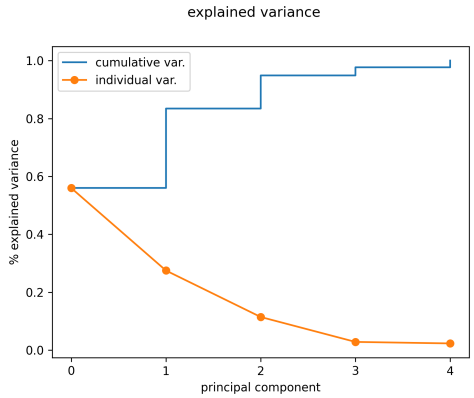


Figure 1

K-Means Clustering

We performed clustering of the search data on both the raw data as well as the PCA-reduced components. To determine the optimal number of clusters,we used the elbow method by taking the sum of squared distances of samples to their closest cluster center. From this, we determined the optimal k-value was 4 for both of the datasets (3).

Not much information is lost from the high dimension to the low dimension therefore, the clusters remain consistent. KMeans works similarly well for both the dimensions.

Regression performance: Date-based split

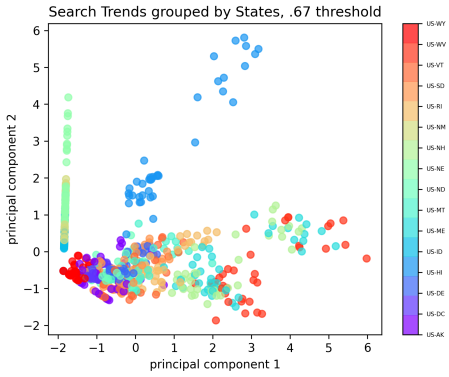
K-Nearest Neighbors We used the elbow method to determine the optimal value of K in the KNN Classifier. During training, we tried values of K from 1-50 to predict the optimal value. By adding more neighbors our model improves, but at some point there will be overfitting which can be seen in the curve. Therefore the value of K for KNN based on dates was found to be 5.

Decision Tree For tuning the max depth, we can look at the mean squared error for different depths of our tree. We used values from 1-50 to determine the optimal value. Plotting the error for the validation set, we found optimal tree depth to be 3 when data was split based on dates.

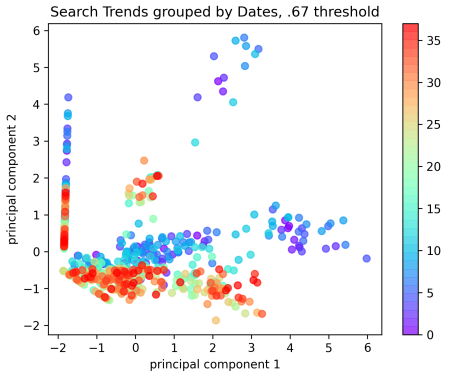
When the data was split based on dates, we obtained an optimal value of k=5 for the KNN with an error of 0.56, and an optimal depth of 3 for the decision tree with an error of 0.78.

Regression performance: Region-based split

To obtain the best results from our data, we performed 5-fold cross validation. For picking the best model and tuning the hyperparameters, we pick the simplest model that is within one standard deviation of the best performing model. This

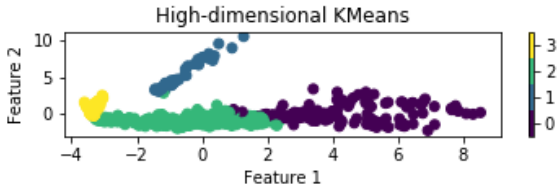


(a)

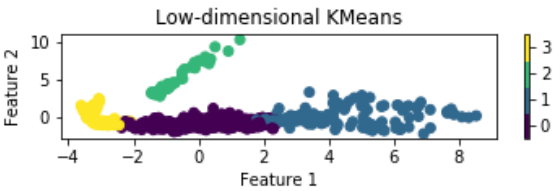


(b)

Figure 2: Principal component analysis of the search trends dataset. Data was thresholded at 67% non-nan values in both cases. (a) Grouping of the dataset by state. (b) Grouping the dataset by search date.



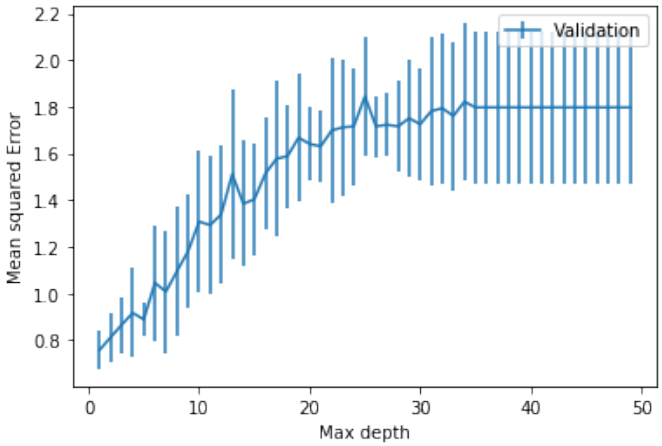
(a)



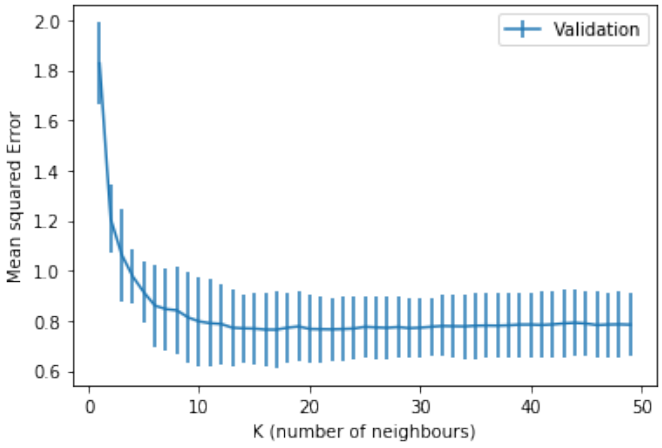
(b)

Figure 3: Clusters obtained by performing K Means on (a) the full dataset and (b), the first two principal components.

Figure 4: Cross-Validation Error



(a) Cross-validation MSE of Decision Tree



(b) Cross-validation MSE of KNN

may make our model simpler and the decision boundaries will also be smoother. For the KNN classifier, we obtained an optimal value of k=15. For the decision tree model, the optimal depth was 10. With these hyperparameters, we obtained a mean squared error of 0.60 for the KNN classifier, and 0.90 for the decision tree (4).

Therefore, the KNN model performed better on our dataset, and was best able to predict COVID-19 hospitalizations from Google search data grouped by date.

This could be due to decision trees being generally unstable. They can also lose valuable information while handling continuous data, such as number of hospitalizations and search frequency. Therefore, decision trees are not completely adequate for applying regression. There can also be overfitting in our training data due to which we get suboptimal results.

Visualizing the evolution of search trends and hospital cases

We decided to pick our symptoms based on their respective available data. With a threshold of 0.67, we managed to filter out all symptoms that contained less than 67% non-Nan values, namely only aphonia, dysautonomia, shallow breathing and ventricular fibrillation satisfied the constraint.

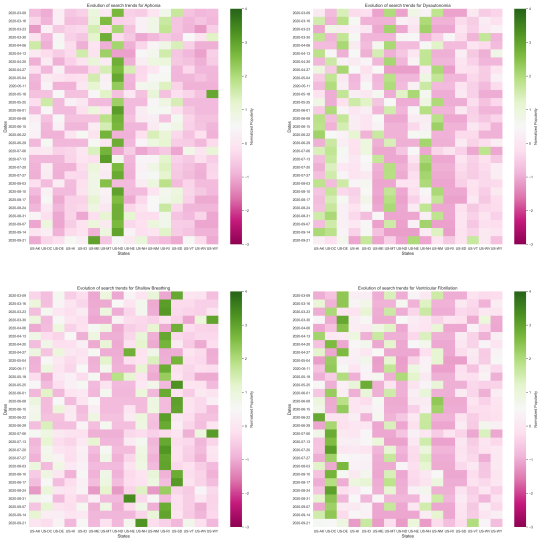
Upon inspecting our graphs when data is normalized by date, we notice a clear dominance from North Dakota and Montana for aphonia. It's also worth noting that these two states are geographically adjacent (5a).

Rhode Island takes the lead for shallow breathing and D.C. takes over ventricular fibrillation (6a). There is no clear winner for dysautonomia. When looking at the hospital data, we note that North Dakota and Montana have more cases than most of the other states which was to be expected. However, the most surprising result is that Idaho, which is geographically next to Montana, and Maine are the two states with the most hospital cases even though these states were both relatively under the radar for the 4 symptoms. It is obvious that looking at 4 symptoms is not indicative enough to determine which state will have the most COVID-19 hospitalizations.

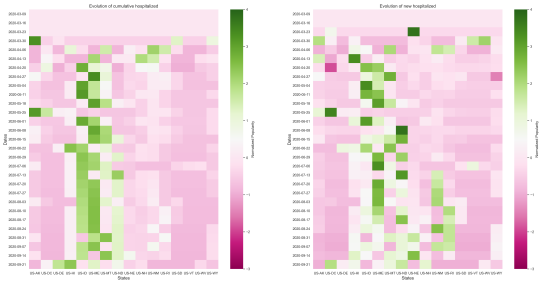
On the other hand, when observing the graphs in which data has been normalized by regions, we can see that most states have "good weeks" and "bad weeks" in terms of hospital cases. Taking a closer look at North Dakota, we notice that the first 11 weeks have had a higher z-score than the other weeks. However, the data provided by the government of North Dakota states that there has been a progressive increase of active positives (of Health 2020).

This is quite fascinating because prior to this revelation we assumed that the number of active positives was strongly correlated to the number of hospital cases. A plausible hypothesis for this phenomenon would be that there has been an increase in testing as well as not all cases are considered bad enough to be hospitalized. Nevertheless, it is harder to draw conclusions. This is because there are too many factors to take in consideration when trying to explain why a given week could've been worse or better than another week.

Figure 5: Evolution of data by date

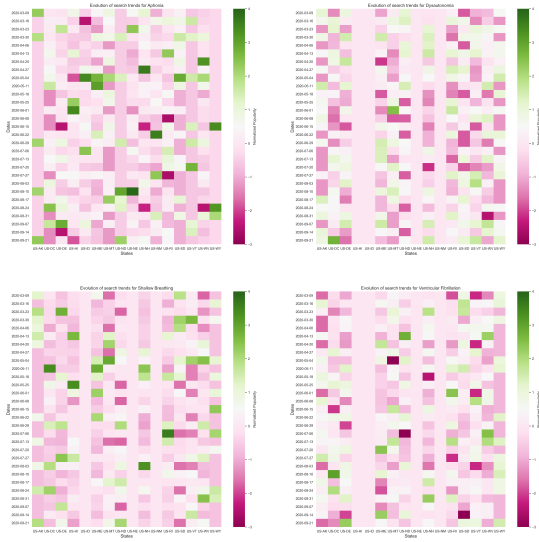


(a) Evolution of search trends when data is normalized by date

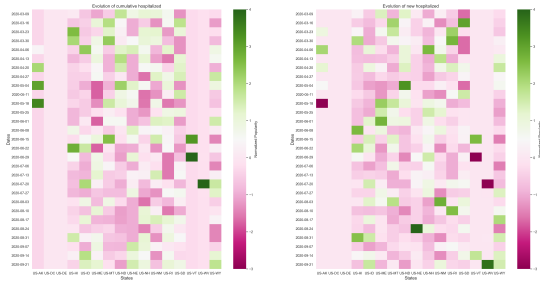


(b) Evolution of hospital cases when data is normalized by date

Figure 6: Evolution of data by region



(a) Evolution of search trends when data is normalized by region



(b) Evolution of hospital cases when data is normalized by region

Discussion and Conclusion

Overall, we show that predicting COVID-19 hospitalizations from Google search data is more reliable from temporal patterns than regional patterns. This is expected from the visualizations of the data, where symptoms seemed to maintain frequently searched over time. Using k-nearest neighbors to predict hospitalizations from the search data gave better performance than a decision tree on this dataset.

Comparison of symptom types

Looking at (5a), it appears there are two different categories of popular symptoms: either the symptom remains highly popular over a long period of time (aphonia, shallow breathing), or it is more widely popular but with less intensity (v. fibrillation, dysautonomia). Because COVID-19 is novel and new effects on the body are being found every day, it may be that a symptom that arises in searches suddenly with high frequency is more predictive of an outbreak than a symptom whose frequency remains constant in that area.

Limitations

One major limitation of our analysis is that good hospitalization data was unavailable for most US states. The data that is available only covers limited regions of the country, primarily parts of the Upper Midwest, Pacific Northwest, and Northeast. California, Texas, Florida, which contributed to the majority of COVID-19 cases in the country, did not have sufficient data available due to either privacy or quality concerns (Times 2020). With good quality data from these critical states, our model would have likely been more robust.

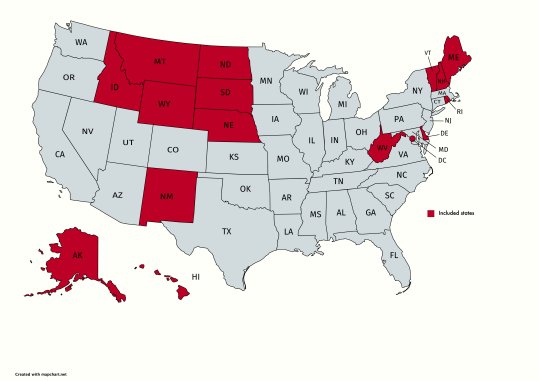


Figure 7: Map of states included in the analysis.

Statement of Contributions

Javier was responsible for the visualization of popularity of various symptoms, the visualization of the evolution of hospital cases and writing its corresponding section, as well as code reviewing.

Maja was responsible for writing the abstract, introduction, data cleaning, PCA, assisted with K-nearest neighbors, and contributed to writing the discussion.

Zoha was responsible for performing K-means clustering of the data, running the KNN and Decision Tree algorithms, generating analysis plots and summarizing the findings for the corresponding section, and contributed to the discussion.

References

Ayers, J. W.; Althouse, B. M.; Allem, J.-P.; Rosenquist, J. N.; and Ford, D. E. 2013. Seasonality in seeking mental health information on google. *American Journal of Preventive Medicine* 44(5):520 – 525.

Kandula, S., and Shaman, J. 2019. Reappraising the utility of google flu trends. *PLOS Computational Biology* 15(8):1–16.

LLC, G. 2020. Google covid-19 search trends symptoms dataset. http://goo.gle/covid19symptomdataset. accessed 03 October 2020.

of Health, N. D. D. 2020. Coronavirus cases. https://www.health.nd.gov/diseases-conditions/coronavirus/north-dakota-coronavirus-cases. accessed 18 October 2020.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Times, T. N. Y. 2020. Covid in the u.s.: Latest map and case count. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.htmlstates. accessed 18 October 2020.