COMP 551 Mini-Project 1

Analyzing COVID-19 Search Trends and Hospitalization

Siddharth Raghavan, 260740094
Patrick Iskandar, 260761003
Carolyn Kolaczyk, 260791409

Faculty of Science, McGill University
817 Sherbrooke Street West, Montreal, QC, H3A 0C3

October 17, 2020

1 Abstract

In this project we investigated the performance of two regression models, namely K-nearest neighbours and decision trees, on predicting COVID-19 hospitalization cases from related symptoms search. We found that the K-nearest neighbour regression approach was a better model at predicting the hospitalization cases. Both models took approximately the same amount of time to train and compute their respective validation errors. Unfortunately, both models were unable to accurately predict COVID-19 hospitalization cases, as evident by their relatively high root-mean-squared-errors. This inaccuracy occurred regardless of whether a time-specific approach or a region-specific approach was taken to predict hospitalization cases. This was mainly attributed to the fact that the search trends data did not include obvious COVID-19 symptoms like fever, cold, or dry cough. Additionally, the search trends data set did not include the search trends of COVID-19 hot spots like Florida, California, Texas, and Wisconsin. Thus, we could not draw conclusions with our predictive models on this limited data set.

2 Introduction

The goal of the project was to predict ”new” hospitalization cases (i.e. new entries into hospitals) in different regions of the US with the search trends data. Thus, two data sets were used for this project:

1. Search Trends data set [1]: A region-specific, aggregated, and anonymized data set that shows trends in search patterns for symptoms. The data frame dimensions (for weekly resolution) were (624,125).
2. COVID hospitalization cases data set [2]: An open source data set that aggregates public COVID-19 data sources into a single data set. The data set includes time series data for COVID-19 hospitalizations. The data frame dimensions were (4483,2).

Thus, the search trends data set were used as the features to predict the new hospitalization cases (target variable). Since the target variable is continuous, regression models like K-nearest neighbours (KNN), decision trees, linear regression, and support vector machines were used for this project.

As seen in Figure 5 in the Appendix, after merging the two data sets under a common weekly time resolution and with common regions, the data frame was very sparse. This sparse data frame had dimensions (624,127). The cleaning and pre-processing of this merged data set will be discussed in detail in the Section 3. After the pre-processing steps, the resulting data frame had dimensions of (234,80). Principal Component Analysis (PCA) was applied to the data set to reduce the number of features to 20. After PCA, clustering methods were explored.

The PCA-reduced data (along with the pre-PCA-reduced data or ”raw” data was fed into the regression models; namely, KNN and decision trees. Additionally, a Support Vector Machine (SVM) model was also used to predict the hospitalization cases. A total of 4 approaches were taken: KNN and decision trees for both region-based splits and time-based splits. All validation errors are reported in Section 4.3. These validation errors were quite high, and indicate that the models poorly predict the hospitalization cases.

3 Data Sets

The weekly search trends symptoms data set was imported into a data frame. Additionally, the daily search trends symptom data set was used as a more effective tool for visualizing the search trends of symptoms. This data set involved a conversion of the time resolution from daily to weekly, which was done by setting the date as the index, grouping by date and the state codes, and then resampling and summing the data. The next data set that was imported into a pandas data frame was the COVID Hospitalization Cases data set. The data set was converted from daily to weekly resolution in the same way as the daily symptoms data set. The two data sets were then merged by the regions. The following steps were applied for data pre-processing:

1. Rows (or samples) with at least 50% of data missing in them were removed.
2. Columns (or features) with at least 25% of data missing in them were removed. Data was removed and then imputed to avoid introducing artificial biases. These threshold decisions for removing missing data were taken from literature [3]. Additionally, as seen in Figure 6 in the Appendix, the NA values are more or less randomly distributed at this stage.
3. The current data frame was region-normalized. This data frame had to be region-denormalized to be fed into a PCA model. Thus, the variance of a symptom across multiple regions had to be captured. This was done by computing the median of medians of all the symptoms in a region, reporting that as an ”overall state symptom” value, and then computing the percentage change of all the values in that region from the overall state symptom value.
4. The columns (or features) were then imputed in the order of most missing data to least missing data. Because the authors of this project are not physicians, the features lend no meaning to the overall goal. Thus, it was thought that imputation should be done by directly correlating with other features. Features with missing data were imputed via linear regression, where its 6 highest correlated features were taken as pseudo-features and the feature with missing data was a pseudo-target variable. It can be seen that the linear regression imputation was accurate, as seen in Figure 7 in the Appendix.

After these pre-processing steps, the filled data frame was fed into a PCA model and a cumulative variance limit was set (as 82%) to determine an appropriate number of principal components (20). The PCA inferences, subsequent clustering, and regression model application will be explored in Section 4.

4 Results

4.1 Visualizing Search Trends

Choropleth maps were used to visualize search trends for symptoms, as the data was split based on states. Both the daily and weekly symptoms data sets were visualized. Figure 8 in the Appendix shows one of the most popular symptoms from the weekly data set. As aforementioned, these symptoms were not relevant to COVID-19, and the data was missing for many states. Figure 1 shows one of the most popular symptoms from the daily data set. This was a more complete data set and hence its visualization was better.

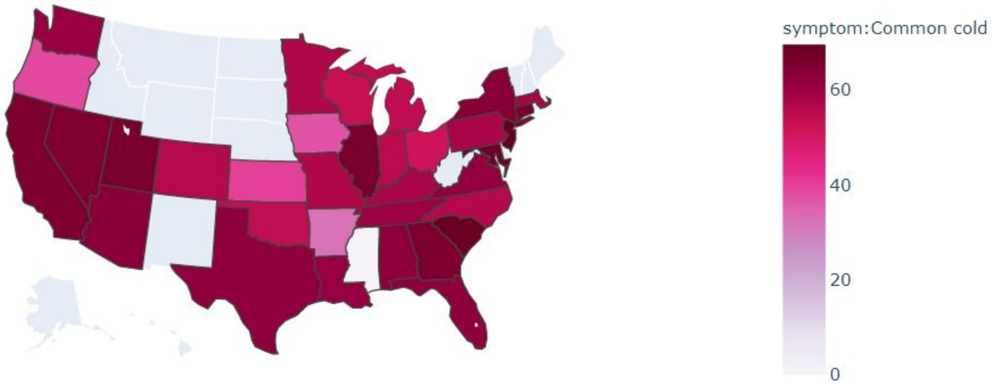

Figure 1: Heat map showing relative region-specific popularity of search trends for Common Cold - extracted for April 27, 2020

It was seen that Common cold had the highest search trends around the first two weeks of March 2020. This was approximately when COVID-19 started to get traction on media, and the corresponding lockdowns began toward the end of March 2020.

4.2 PCA Application and Clustering

As aforementioned, PCA was used to reduce the data dimensionality. Before doing so, the data set had 79 features. Using PCA, it was possible to reduce the number of features to 20 while still explaining 82% of the variance, as seen in Figure 9 in the Appendix. After PCA, the first two principal components can be visualized as a method of viewing the entirety of the Search Trends data set in a lower 2-D dimension (assuming PCA was applied correctly). This is seen in Figure 2.

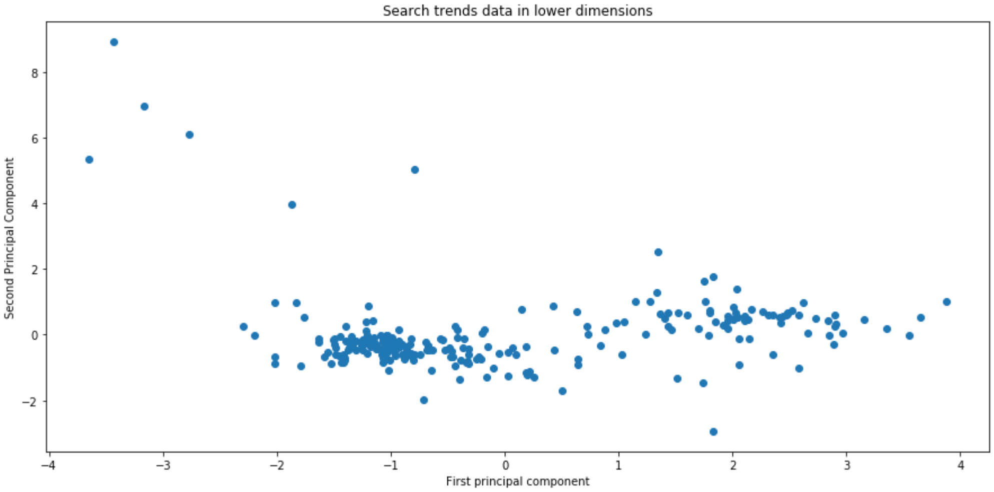

Figure 2: Search trends data in lower dimensions

After, the data can be clustered using K-means clustering. First, the elbow method to find that the optimal number of clusters is K = 3, as seen in Figure 10 in the Appendix. We can see the cluster labels resulting from K-means clustering applied to both the PCA-reduced data and the pre-PCA-reduced data in Figure 3.

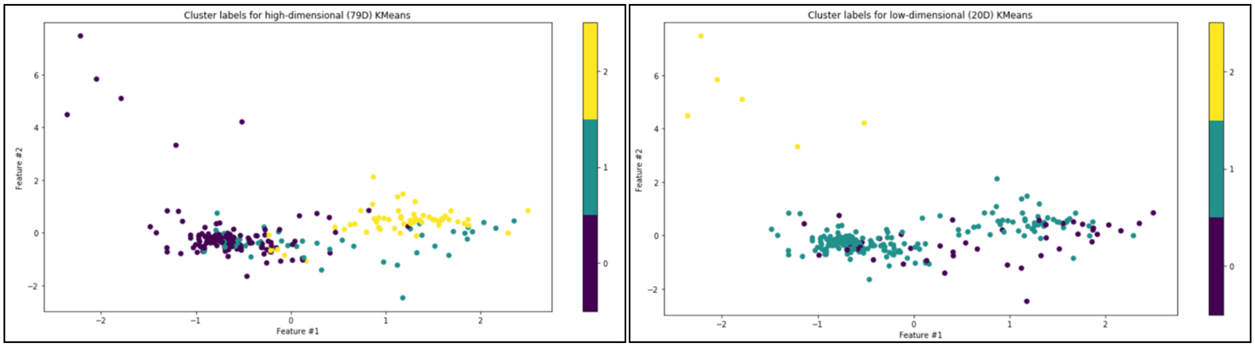

Figure 3: Clustering for the pre-PCA-reduced data (left) and PCA-reduced data (right)

As seen from both the plots, the cluster labels for both data sets are similar. However, the major difference is in classification of points in two of the clusters. Looking at the purple (C_1) and yellow (C_2) clusters in the pre-PCA-reduced data, there is a difference to its equivalent yellow ($C_{eq,1}$) and green clusters ($C_{eq,2}$) in the PCA-reduced data. Specifically, the majority of data points in the clump belong to C_1 in the pre-PCA-reduced data, while the majority of data points in the clump belong to $C_{eq,2}$ in the PCA-reduced data. The third cluster is the same for either data set. The similarity in clustering implies that the PCA-reduced data is a faithful representation of the raw data.

4.3 Regression Model Performance

Before applying regression models to the data, it was thought best to implement time lags into the target variable. Intuitively, it makes sense that the search trends are more greatly correlated to hospitalized cases in the next weeks rather than within the same week. This intuition was also confirmed by researching a recent study [4]. Thus, for each region, the best time lags out of 0, 1, 2, 3, or 4 week lags was chosen, and then the target variable was shifted by the appropriate time lags. In Figure 11 in the Appendix, the correlations for each time lag for a specific region was plotted. As can be seen, a 3 week time lag had the highest correlation, so the hospitalization cases for that region were shifted up by 3 weeks.

With the time-lagged data frames, KNN and decision trees were used to predict the hospitalization cases given the search trends data. The results are visualized in the Figure 4.

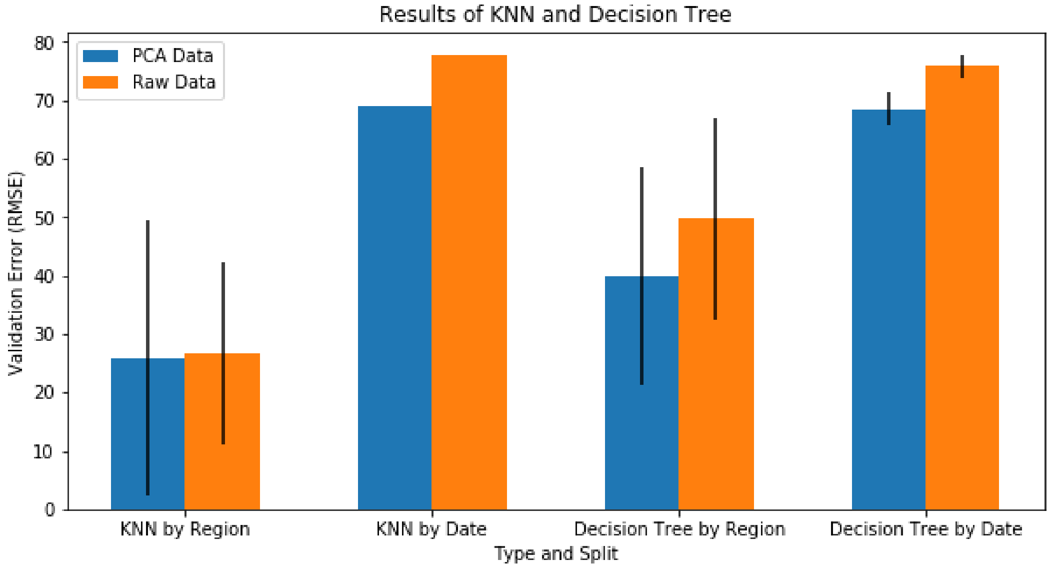

Figure 4: Results of the main regression models of this project

For both KNN and decision trees, using cross-validation with a region-based train-validation split performed better than a date-based train-validation split. Using the PCA-reduced data, KNN regression achieved an average RMSE validation error of 20.88 when splitting by region, but 48.71 when splitting by date. Decision trees achieved an average RMSE validation error of 29.53 when splitting by region, but 58.91 when spitting by date. Notice also that for both the region-based and date-based train-validation splits, KNN performs better than decision trees.

Comparing the regression using PCA-reduced data versus raw (or pre-PCA-reduced) data, KNN using a region-based split performs similarly for both data sets, although the PCA-reduced data's validation error is slightly better. However, for all other cases both classifiers perform significantly better using the PCA-reduced data. For KNN with a date-based split, PCA-reduced data achieves an average RMSE validation error of 48.71, while the raw data achieves an error of 57.29. For decision trees with a region-based split, PCA-reduced data achieves an error of 29.53, while the raw data achieves an error of 37.31. And lastly, for decision trees with a date-based split, PCA-reduced achieves an error of 58.91, while the raw data achieves an error of 62.99.

Additionally, an alternative model for predictions was explored: Support Vector Machines (SVM). This model was additionally used because of its ability to produce accurate results with less computation power. It is widely used for regression tasks. For this project, with a large number of features, there was a great chance a suitable hyper-plane could be computed. With this model, a comparable RMSE was obtained (around 20-30 depending on hyper-parameters gamma, C, and epsilon), solidifying the consistency of this mdoel compared to other models. With some specific values of the hyper-parameters, a RMSE of 26.7 was obtained. See Figure 12 in the Appendix to visualize the performance of this model.

5 Discussion and Conclusion

The resulting data set acquired after merging was a large sparse data set with only a handful regions of a common time resolution. This greatly limited the training samples which in turn made the regression models more unstable (i.e. every individual sample affected the model much more). The regions in the data set did not include search trends of COVID-19 hot spots like Florida, California, Texas, and Wisconsin. Additionally, the search trends data in the weekly resolution did not include obvious COVID-19 symptoms like fever, cold, or dry cough.

The daily search trends data set was used to visualize the most popular searched symptoms. For COVID-19 related symptoms like "Common Cold", there was a spike at approximately the time when COVID-19 first became a regular topic on media. This data set was chosen to be discarded for further analysis, because converting the daily resolution to a weekly resolution was not straightforward (considering that values reported were region-normalized and scaled considering all time periods).

The missing data of the merged sparse data set was imputed after appropriate cleaning of the data set. Linear regressions were chosen as the method for imputing missing data as the authors cannot make meaningful decisions about how the features interacted with each other. Applying the PCA model reduced the number of features by approximately 75% (from 79 to 20). The clustering thereafter showed consistency between the pre-PCA-reduced data and the PCA-reduced data. The PCA-reduced data set and the pre-PCA-reduced data set were then time lagged by finding the time lag at which the mean correlation of all features were highest with respect to the target variable. It was assumed that each region had their respective rates of transfer of information, so each region had their specific time lag applied.

The data was then fed into KNN and decision tree models. The mean root-mean-squared-error (RMSE) across the 4 approaches for the PCA-reduced data was 39.82. Meanwhile, the mean hospitalization cases value (i.e. average target value) after pre-processing was 25.8. This means the predictions had a greater error than the average value of hospitalizations itself! An alternative model of SVM was applied, and a comparable RMSE was obtained. Additionally, the validation errors using the PCA-reduced data set and the pre-PCA-reduced data set were both similar, demonstrating that PCA was applied correctly.

To conclude, it may be that the correlation between search trends and hospitalization cases is not obvious. Intuitively, it makes sense that if a spike in hospitalization cases occurs, a spike in search trends for particular COVID-19 symptoms also occurs (or occurred in the recent past). However, this was found not to be the case, and the authors could not draw conclusions with the predictive models used.

6 Statement of Contributions

- Siddharth Raghavan: Cleaned and imputed the data, and implemented PCA along with K-means clustering. Implemented the time lagging step. Additionally worked on the optional model - SVM. Compiled the report in LaTeX.
- Patrick Iskandar: Worked on loading (acquiring), cleaning, and merging the data sets as well as visualizing the search frequency of the most popular symptoms throughout the US over time.
- Carolyn Kolaczyk: Implemented both train-validation split strategies with KNN and decision trees regressions.

References

[1] everettk. COVID-19 Search Trends symptoms dataset. https://github.com/google-research/open-covid-19-data/blob/master/data/exports/searchtrendssymptomsdataset/README.md, October 2020.

[2] everettk. Open COVID-19 Data. https://github.com/google-research/open-covid-19-data, October 2020.

[3] Madley-Dowd, Paul, et al. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110:63–73, 2019.

[4] Ahmad, Imama, et al. Increased Internet Search Interest for GI Symptoms May Predict COVID-19 Cases in U.S. Hotspots. *Clinical Gastroenterology and Hepatology*, 2020.

Appendix

As seen in Figure 5, the merged data frame is very sparse. This necessitates data imputation later.

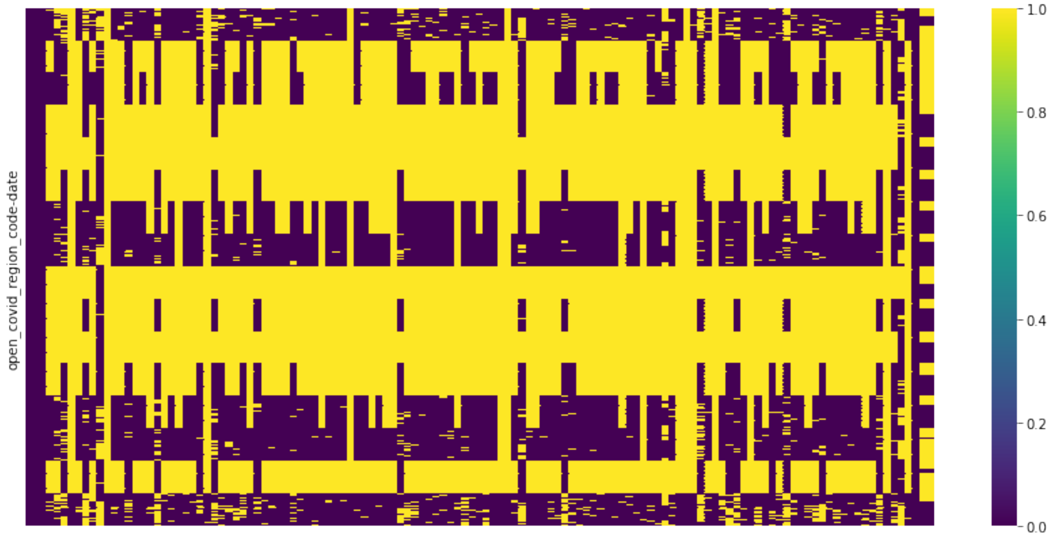

Figure 5: Initial heat map of merged data set - Yellow is NA, Purple is non-NA

As seen in Figure 6, after removing some data, the merged data frame looks much better. This is a data frame suitable for imputation, because the missing data is seemingly random. Also, the missing data does not constitute majority of the data set, thereby avoiding introducing biases when imputations occur.

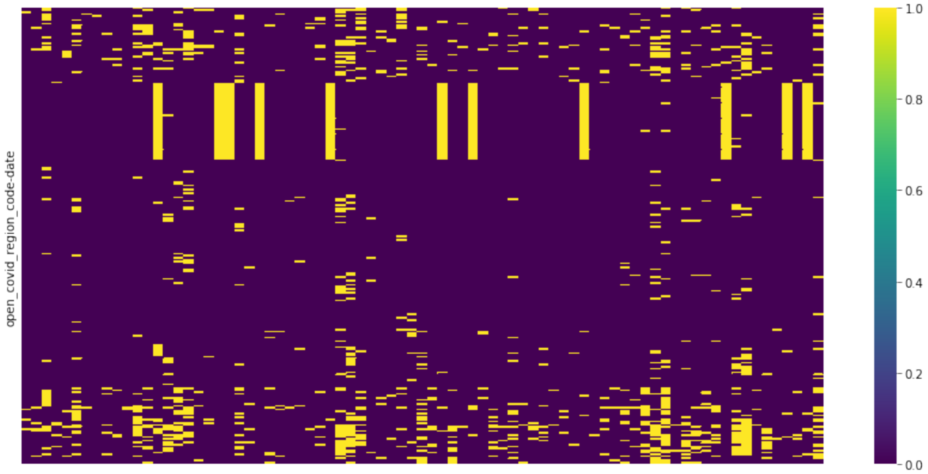

Figure 6: Heat map of merged data set after removing data - Yellow is NA, Purple is non-NA

As seen in Figure 7, the regression model is quite accurate. The scatter plot lies mostly along a straight diagonal line. This means that the incomplete data was imputed with good accuracy (i.e. linear regressions were applied properly).

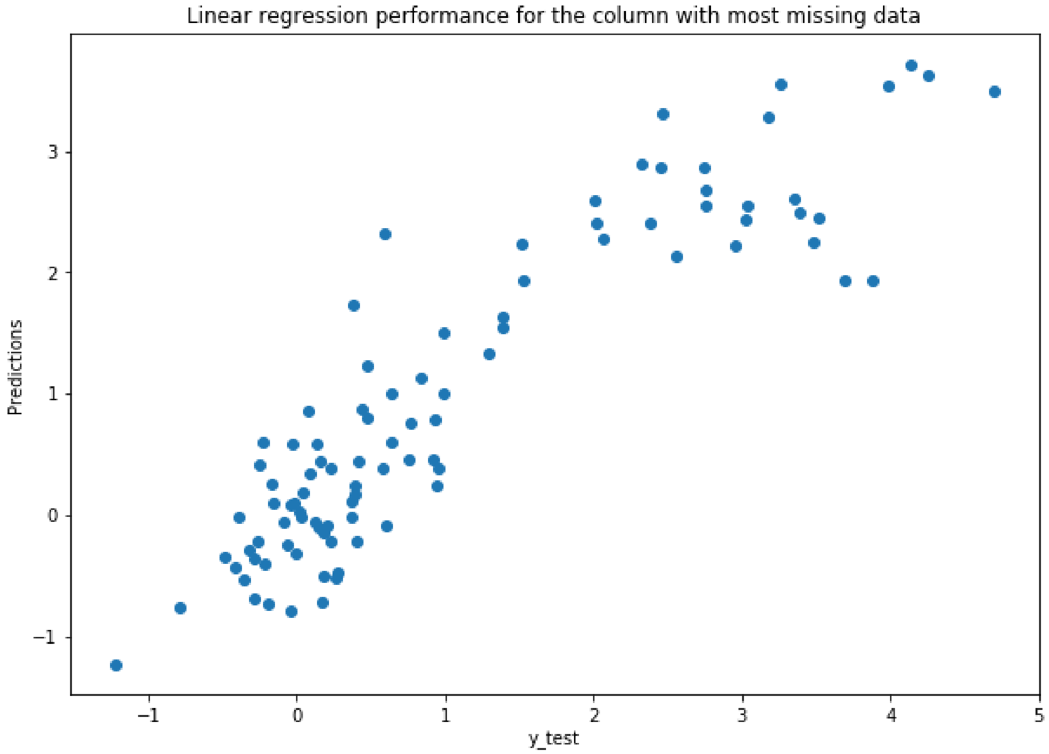

Figure 7: Scatter plot of the predictions versus true values for imputation via linear regression

As seen in Figure 8, the visualization is quite poor. The weekly search trends data set was quite incomplete, with irrelevant symptoms, and with data for only a few regions.

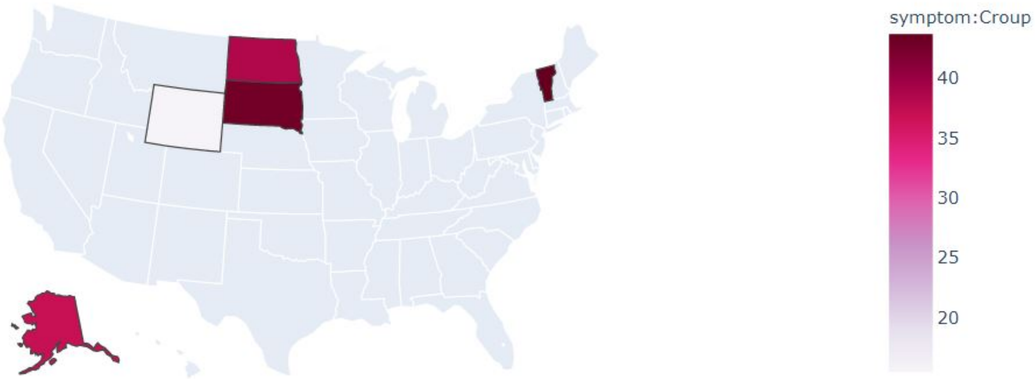

Figure 8: Heat map showing relative region-specific popularity of search trends for Croup - extracted for March 23, 2020

As seen in Figure 9, at 82% of cumulative variance explained, the number of principal components is 20. Thus, it is possible to represent the original features (of which there were 79) with only 20 new features. These 20 features are linear combinations of the original features and are computed to best represent the variance in the original data.

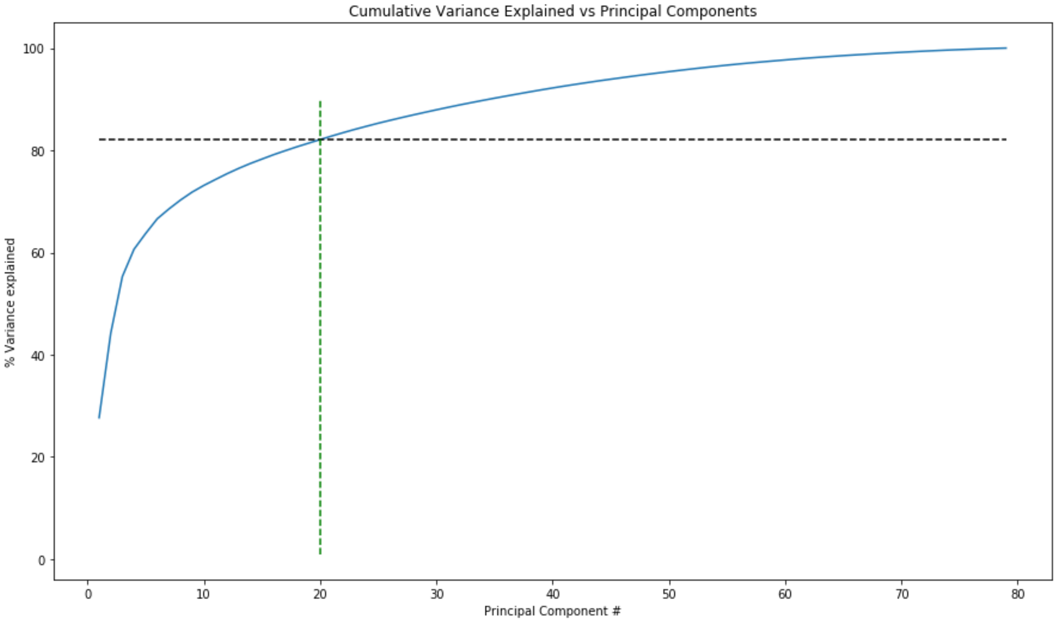

Figure 9: Cumulative variance explained versus number of principal components

As seen in Figure 10, the ”elbow” was found at $K = 3$. This represents the optimal number of clusters.

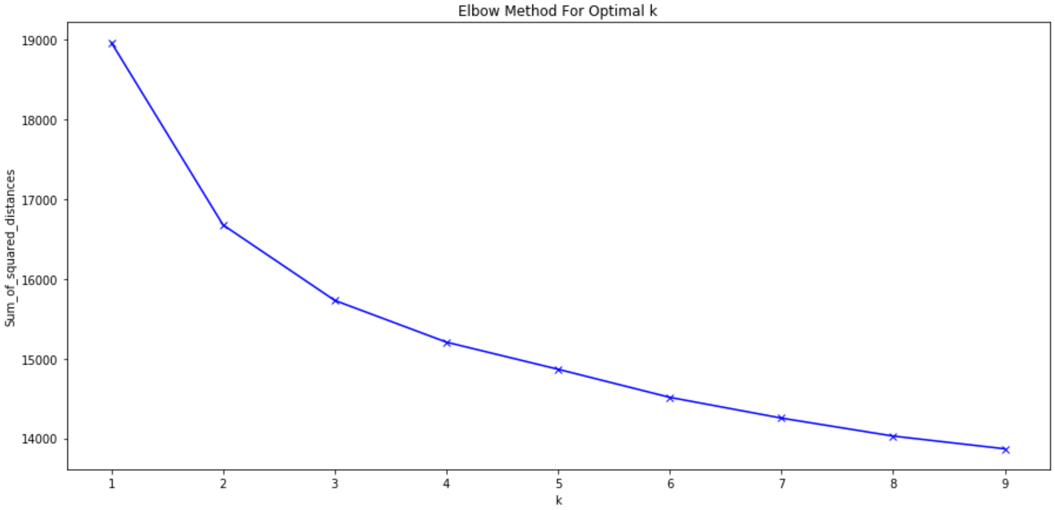

Figure 10: Elbow method for clustering the search trends data

As seen in Figure 11, the highest correlation for this region corresponds to a 3-week time lag. Thus, this time lag was implemented on the data for that region. This same process of finding time lags were implemented to each region (it is assumed that the rate of transfer of information of each region is different). Finally, the time-lagged data frame gets fed into the regression models.

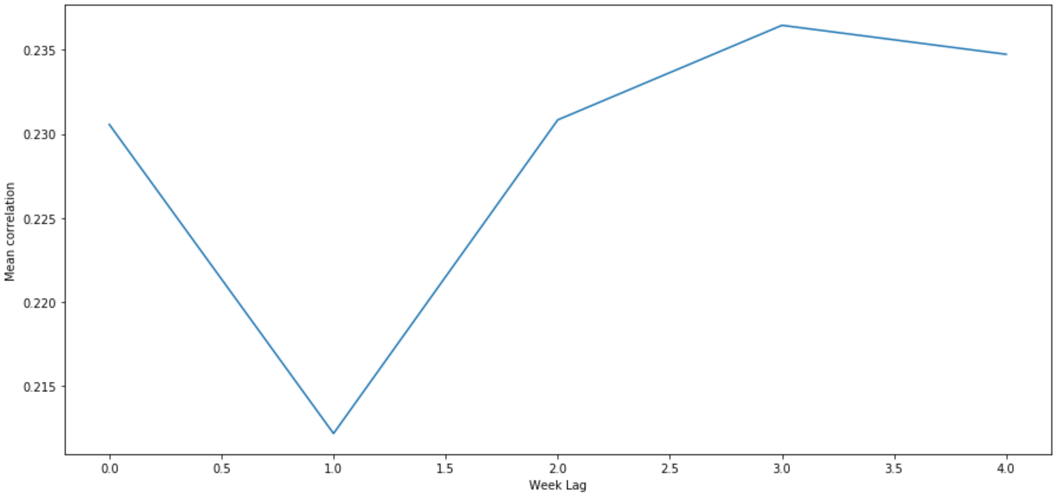

Figure 11: Correlations for different time (week) lags for a specific region

As seen in Figure 12, the scattered points do not lie on a straight diagonal line. The RMSE was found to be relatively high at 26.7. However, this model's performance is close to KNN model's performance, implying that all models perform equally poorly.

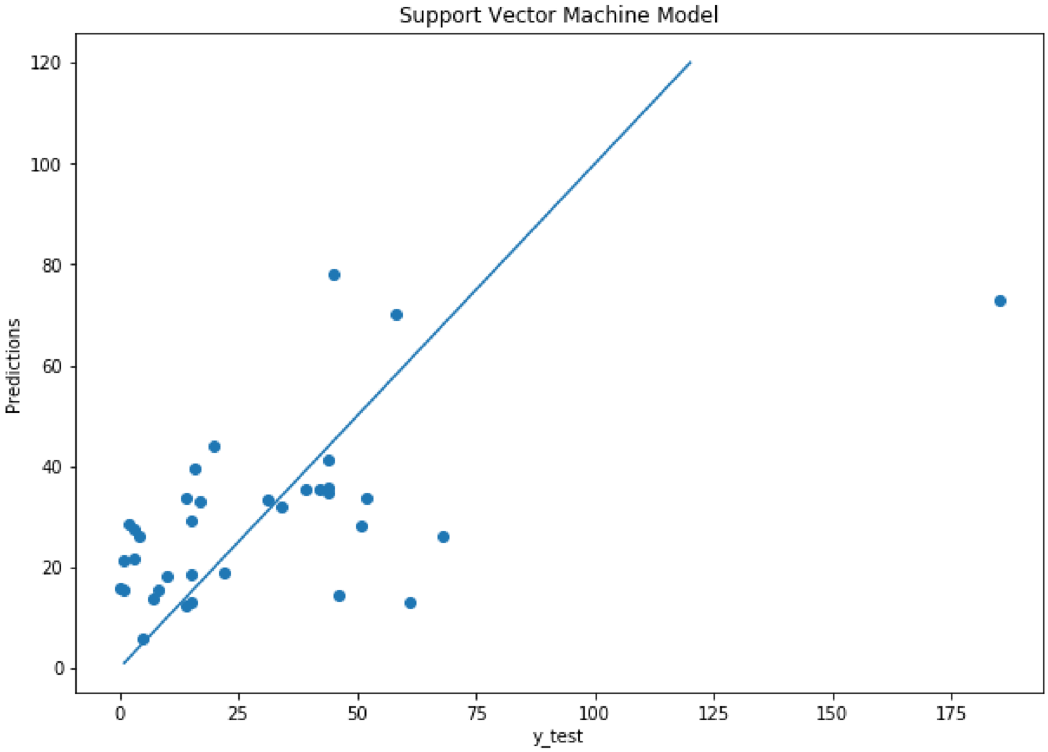

Figure 12: Performance of the SVM model