Miniproject 1 Report - Group 63

Ismail Faruk, Arya Akkus, Rozerin Akkus

21 October 2020

Abstract:In this project, we explored the performance of two regression models, namely K-nearest Neighbors Regression and Decision Trees on estimating the weekly hospitalization cases from related symptoms searched. Our findings are as follows: We found that the k-nearest neighbour regression approach achieved better accuracy than decision trees and was significantly faster to train."

1 Introduction

The task we were aiming to complete in this project was to come up with two regression models to predict the weekly hospitalization number by learning from the related symptoms searches.Therefore we used the two datasets given, namely **2020 US Weekly Symptoms Search Trends Dataset** and **Covid Hospitalization Cases Dataset**. The first data set is obtained by mapping the Google search queries to various symptoms and organizing them by region. The time resolution of this dataset is weekly. For each region, symptoms are displayed by their relative frequency for a given week. The second dataset is aggregated by a open pipeline that is connected to many data sources. The pipeline, later, standardizes the data. We used the new hospitalizations from this dataset to construct our data. Through our projects, we were able to detect the accuracy of each model, visualize the most popular symptoms and varying frequency of symptoms over time and by region.

2 Datasets

To preprocess our two datasets, we first filtered out unnecessary or invalid data from the each of the datasets separately. We removed columns with too little or no data from Search Trends Dataset. We later, filled NAN values with 0 and sorted the dataset by region and date. As the second dataset has global daily data, we converted this dataset to weekly and filtered US region specific data. We only kept region, date and new hospitalizations in this dataset since the new hospitalization cases is our target value. We sorted the latter dataset in the same way we did for the former. Lastly, we merged these two by region and date.We started visualizing our data by plotting most popular symptoms over time by region. Some examples can be seen in figures. We also provided popularity heat maps for some of the most popular symptoms over time.

3 Results

3.1 PCA and Clustering

We applied PCA algorithm and observed the variance ratio and cumulative variance ratio with respected to the number of principal components.

We applied K-means clustering with and compared to see if the clustering would be affected due to a PCA reduction.

3.2 KNN Regression

We applied KNN Regression algorithm on the data by splitting based on region and time. For the regression model generated based on splitting the data by region, we split the data for training, validation: 80%, 20%.

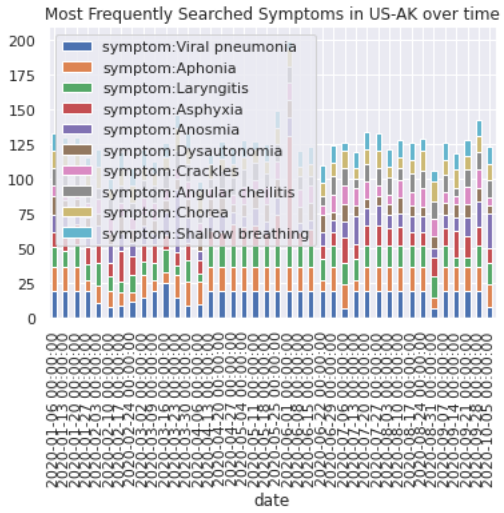

Figure 1: Most Popular Symptoms over time in Alaska (US-AK)

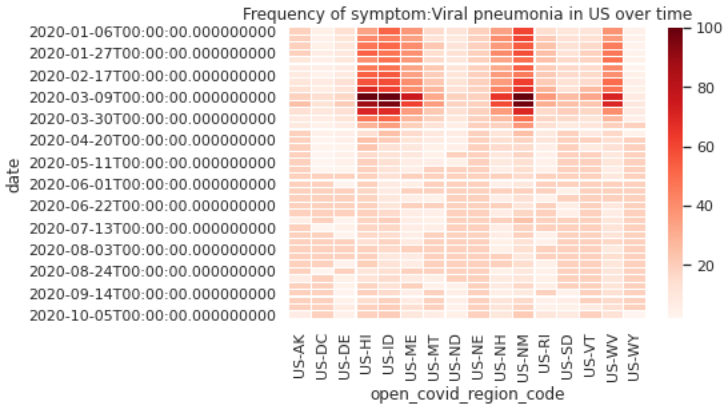

Figure 2: Popularity of Viral Pneumonia over time in US

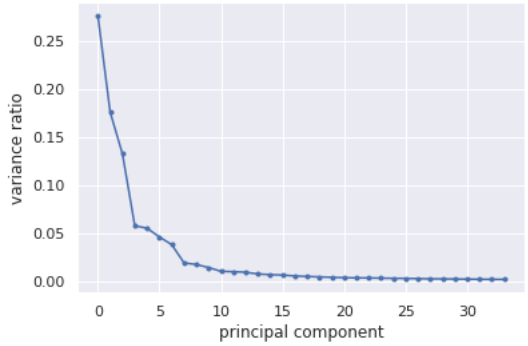

Figure 3: Fig: PCA - Variance ratio vs n-PC

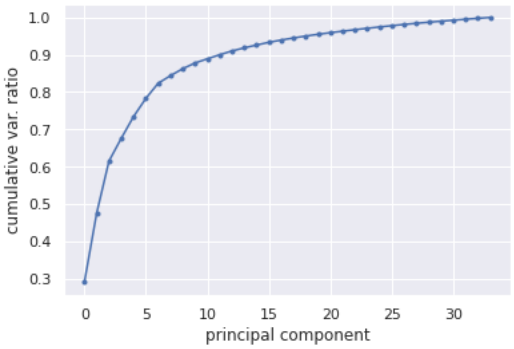

Figure 4: Fig: PCA - Cumulative Var. ratio vs n-PC

Figure 5: Fig: Cluster - without PCA

Figure 6: Fig: Cluster - with PCA

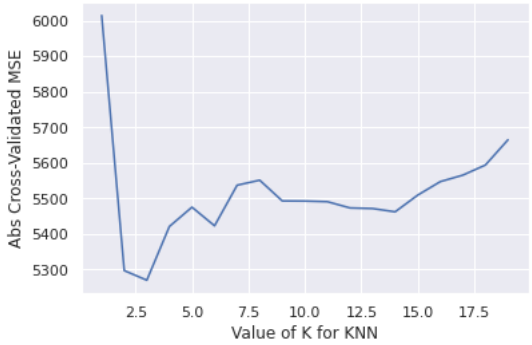

Figure 7: KNNRegres. Region Split: Abs. Cross-validation MSE vs K-NN

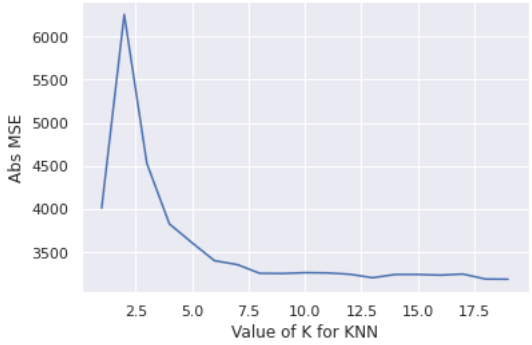

Figure 8: DecisionTreeRegress.Time Split: Abs. MSE vs K-NN

Then we performed 5-fold cross validation on the model. We performed this process for values of K: 1 to 20 and calculated their Cost: *Mean Squared Error* after cross-validation. The following is the plot of the *Absolute Mean Squared Error* vs K-Neighbors:

For the regression model generated based on splitting the data by time, we observed the data split with respect to the data of 2020-08-10, had split the data for training, validation: 75%, 25%. Then we calculated the predicted label and compared it with the validation label. We performed this process for values of K: 1 to 20 and calculated their Cost: *Mean Squared Error* using the predicted label and validation label. The following is the plot of the *Absolute Mean Squared Error* vs K-Neighbors:

3.3 Decision Tree Regression

For our Decision Tree Regression, we applied two different splits to the data similar to what we did in KNN regression. For the region based split, we split the training and validation data by 80%20, 20%. Then we performed five fold cross validation on the model. We performed this model for tree depth varying between 1-40 and calculated their cost.*Mean Squared Error* after cross-validation. The following is the plot of the *Absolute Mean Squared Error* vs Tree Depth.

For the regression model generated based on splitting the data by time, we observed the data split with respect to the data of 2020-08-10, had split the data for training, validation: 75%, 25%. Then we calculated the predicted label and compared it with the validation label. We performed this process for values of depth: 1 to 40 and calculated their Cost: *Mean Squared Error* using the predicted label and validation label. The following is the plot of the *Absolute Mean Squared Error* vs Tree Depth:

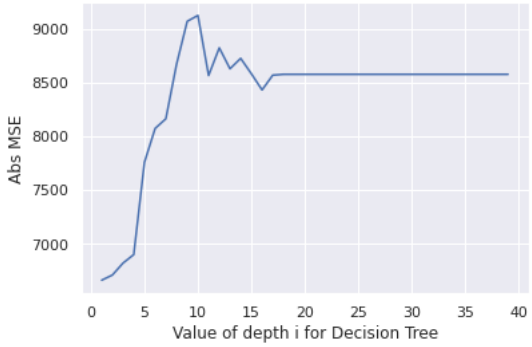

Figure 9: DecisionTreeRegres.Region Split: Abs.Cross Validation MSE vs Tree Depth

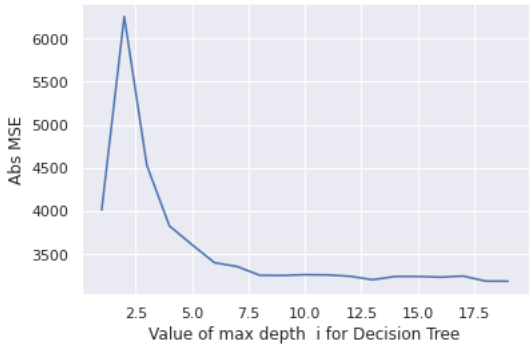

Figure 10: KNNRegres.Time Split: Abs. MSE vs Tree Depth

4 Discussion and Conclusion

The most apparent fact of all the models generated is that, there is a strong possibility of hidden/latent variables. At best, the models generated can be used to find some correlation between new cases and symptom searches but nothing more. Because the primary conclusion is that, symptom search has no causal bearing on new cases. Whereas, there are other data points in the hospitalization dataset that could contribute to a better model to predict new cases, as they are supported by health experts, epidemiologists, CDC of USA, UN and other scientific bodies. We will refer to different data points that could contribute as we discuss further.

4.1 PCA and Clustering

In raw clustering, we can see that the data is more spread out, whereas, PCA reduced clustering we can see that the data points tend to cluster more around certain cluster centers. This behavior is in alignment with PCA algorithm to maximize the variance and contain most information in the primary columns with respect to the latter. Clustering is done to observe if the data scatter has changed due to the PCA reduction. And we were able to observe better clustering affect on the reduced PCA data points than on the graph before with no PCA. This was bound to happen as more informed data would be present in the initial columns of the PCA vector than on the latter, which would make the variance maximized for the initial columns, having better clusters.

4.2 KNN Regression

In the case of the KNN Regression for region split, we can observe that for K=3, the cross validation error is lowest and it increases as K increases. In the case of the KNN Regression for time split, we can observe

that for K=8 and on wards, the cost stabilizes to a minimum. There can be reasons why this is the case. In the case of region splitting, we do not take into consideration the timeline. In different region, factors like timeline, cumulative case numbers, population density, counter disease policy, etc play a crucial role in contributing to the data of new cases for a particular region. But they were not taken into consideration as we did not perform KNN for individual regions and try to analyse what the model would predict. But utilization of cross-validation-error allowed us to better plot the outcome of the validation dataset, than doing a regular cost prediction, as it is the average of the cost of all folds. KNN Regression for time split is relatively stable, after K=8. The cost calculation policy was to use the predicted label with the test label and find their MSE. The model is timeline specific. This allows the model to show more stable behavior. The split allowed usage of 75% of data to be used for training and 25% data to be used for validation.

4.3 Decision Tree Regression

Our Decision Tree Regressor didn't perform as well as the KNN Regressor. Our accuracy for one fold was around 25-30%. This low accuracy may be caused by many factors, namely errors in our data cleaning and filling for invalid values. Although we scaled and trained our training data and tried different implementations the accuracy still remained low. However we observed that the error was relatively stable when the tree depth was i = 15 in regional splitting and the lowest when tree depth i = 8. We noticed that the decision tree performed much better on the time-split data as the error cost function was much lower. The cost calculation policy was to use the predicted label with the test label and find their MSE.

4.4 Proposed Ideas

We could utilize the cumulative case numbers and use them as wights when calculating the KNN, and perform weighted KNN. We could utilize parameters that directly affect the number of new cases to perform modelling, such as population density and school closure policies etc, but unfortunately, they are not recorded into the dataset.

5 Statement of Contributions

Although we initially discussed the procedure all together, eventually Arya conducted the preprocessing of the datasets where Rozerin and Ismail were responsible for visualization of the merged dataset. Additionally, they contributed to the PCA and k-means clustering. Later on, Arya and Rozerin focused on the decision tree regression while Ismail completed the KNN regression. The write-up is a collective effort of all the team members.