Covid-19 Awareness and Covid-19 Cases in Ohio

Ayush Singla and Tanvi Shroff

***Abstract*— The Ohio Covid-19 dataset had various features including but not limited to county, day, total population, cases, deaths, etc. Awareness data has been extracted from tweets. These have been classified into various categories like core, sports, entertainment, etc. This paper walks you through various methods which can be used to predict the number of cases on a particular day. The evaluation metric is the R2 Score. We start with analyzing the dataset and understanding it through Exploratory Data Analysis. This is important since we could find meaningful patterns and trends in the data that acted as a good starting point for our modeling process. We then go ahead to compare the performance of models such as XGBoost, RandomForest, Linear Regression, Logistic Regression, SVM, and more. This was done since we are predicting continuous variable i.e. Covid-19-19 cases. We also talk about which features in the dataset had importance while building the models which is done using feature importance model parameters. This significantly helped in selecting the right parameters and boosting the overall R2 score. Overall, we identified numerous insights and trends in the data tha helped in finding the best regression model and parameters for the prediction results.**

I. INTRODUCTION

As part of our analysis and model development process, we expect the Covid-19 cases to rise exponentially in both train and test set. We confirmed this through visualizations in our code. Interestingly, we observed another feature 'deaths' to have a very similar pattern and it was identified as the most important feature for this analysis. We wanted to try and compare variety of regression models to predict the Covid-19-19 cases and therefore, used almost all popular models. It was expected that there would be good number of variables with high correlations which needs to be removed and a good approach to identify that is by plotting correlation matrix. Some of the significant findings from the overall analysis are: The county of Franklin had the highest number of cases i.e. 65,513, although the maximum number of cases per capita is in the county of Pickaway i.e. 0.89. This county had the 3rd highest number of cases i.e. 50,162. Cuyahoga had the highest number of deaths i.e. 873. Deaths although is undoubtedly the most important feature (56% feature importance) but in order to achieve high accuracy we had to create deathspercapita to include both deaths and total_pop features. This increased the importance of some of the other features and drastically improved results. Similarly, taking average of core normalized and illness normalized scores showed the same effect. Such strategies helped in getting a high R2 score (91.927%) for the regression model.

II. DATA

We have two sets of data - county level and topic awareness variable scores. The former focuses on cases, deaths, and other useful information for each county in Ohio for each day while the latter reflects the level of awareness on a given day in a county across multiple areas like sports, politics, illness etc. All the variables in the awareness data set are normalized using 0-1 normalization technique. The train and test datasets are separate so there was no need to split the data.

A. Covid-19 experience in Ohio

i. The intensity of the virus in Ohio was pretty similar to the other states in the country. The restrictions placed early on helped in reducing the number of cases. But, as the restrictions were lifted, there was eventually a surge of cases in the state. The cases were very high in December 2020 and January 2021.

ii. The state of Ohio is a Republican state. At the time of the pandemic, Mike DeWine was the Governor of Ohio. The Director of the Ohio Department of Health was Amy Acton. He took action against Covid-19 even before the state had any confirmed cases. In March 2020 numerous policies went into effect which is tabulated below.

Date	Event
March 9th	Universities moved to the online space
March 12th	Statewide schools closed
March 12th	Gatherings > 100 people were banned
March 15th	Closure of bars and restaurants
March 16th	Gatherings > 50 people were banned
March 17th	Elective surgeries were postponed
March 22nd	Stay at home order

iii. The 'views' of Wikipedia on the pandemic in Ohio are varied. The strategies implemented by the Governor were lauded by many. But the strategies were also criticized. The Governor, Mike DeWine canceled the Arnold Classic. The Washington Post called this a radical decision for that time. This was a multi-sport festival. Canceling this event cost the government 53 million dollars. When the rule to postpone all elective surgeries was announced, abortions were also a part of it. This received severe backlash. This decision was challenged in court by multiple people and organizations. In the end, the State of Ohio Attorney General Dave Yost said that all medical abortions were still allowed. From May 12th, 2021 to June 1st, 2021 most of the Covid-19 health

orders were rescinded. This included the mask mandate. These decisions were criticized by many as it was said that these decisions contributed to a surge in cases in the latter half of the year.

B. Average normalized Jaccard similarity-based awareness

Fig 1 is a bar plot of the mean of various jaccard normalized awareness values. It tells us that sports (0.016608) must have been the most discussed topic during the period when the data was collected. The topics after that were entertainment, illness, and then core. After these 4, there was a significant drop in the mean values. Surprisingly, the least discussed awareness topic was health. Health technology was also not very significant with a value of 0.001536.

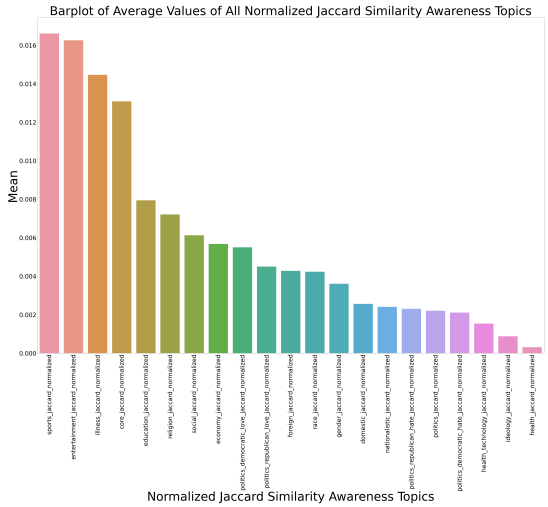

Fig. 1.

The values for
1. politics_republican_love_jaccard_normalized
2. foreign_jaccard_normalized
3. race_jaccard_normalized

are all around 0.004. The average of all these values is 0.0058941. 1/3rd of the variables are above this mean.

C. Aggregated mean Core Jaccard awareness value county-wise

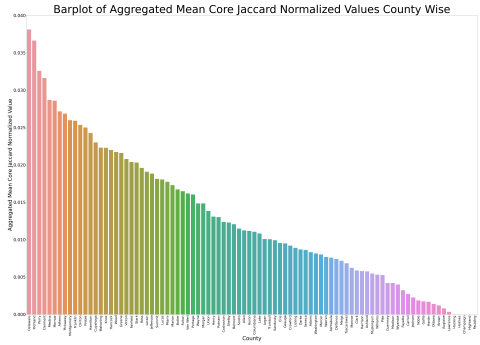

Fig. 2.

Fig 2, again a bar plot of all counties versus their aggregated mean core jaccard normalized values. All data in the core_jaccard_normalized column was grouped by the county and the mean of all values county-wise was calculated. We can see that the highest mean Core jaccard awareness value was for the county of Delaware. This was followed by Richland, Perry, Clemont, and Medina. The counties of Hocking, Holmes, Champaign, Highland, and Paulding had values equal to 0. The mean of all these plotted values is 0.0129. 43% of the variables are above this mean threshold.

D. County level maps of Ohio

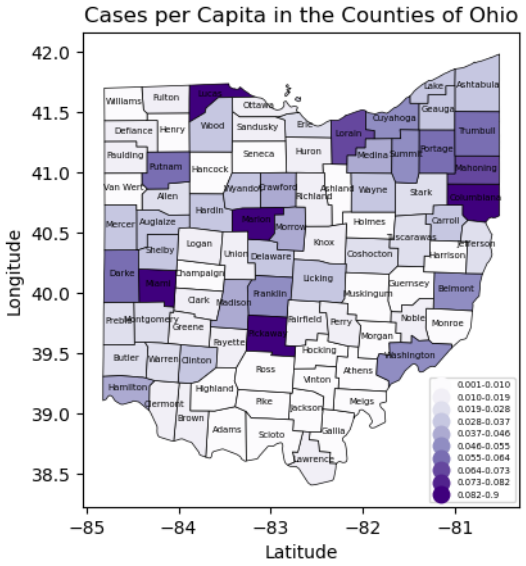

Fig. 3.

Fig 3 depicts the cases per capita of every county. The highest value is in the county of Pickaway represented in the darkest purple. It has 0.887587366 cases per capita. On the other hand, Meigs has the lowest number of cases per capita, 0.001491075. The top-5 counties with the highest number of per capita cases are:
1. Pickaway
2. Marion
3. Lucas
4. Columbiana
5. Miami

Fig 4 depicts the deaths per capita in every county. Miami has the highest deaths per capita, whereas multiple counties have the lowest deaths per capita, almost 0. The top-5 counties with the highest number of per capita deaths are:
1. Miami
2. Darke
3. Portage
4. Columbiana
5. Mahoning

The counties of Miami and Columbiana are in the top 5 for both, per capita cases as well as per capita deaths.

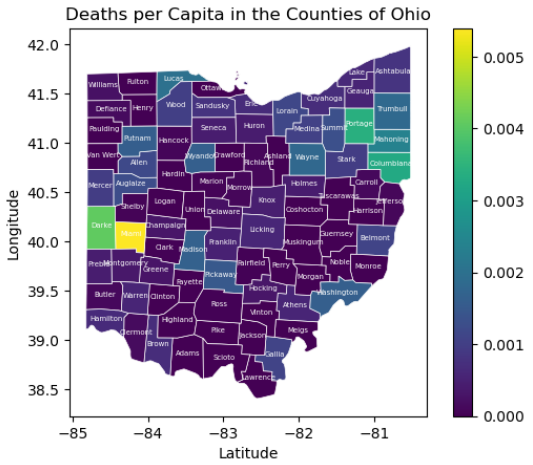

Fig. 4.

E. Average normalized Jaccard awareness scores day-wise

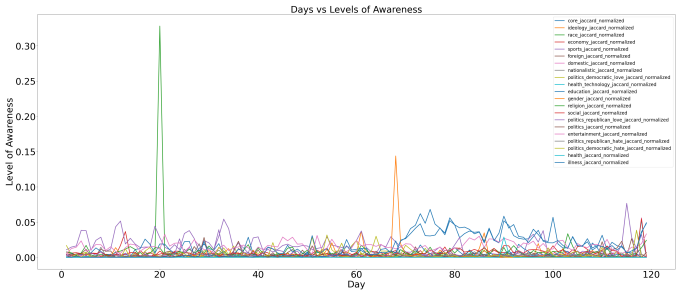

Fig. 5.

Fig 5 is a line chart. The X-axis represents the day number and the Y-axis is the average topic awareness variable score. A particular average awareness score for a day is calculated by grouping awareness scores of each county for that day. We can see that the Race_jaccard_normalized variable has a sudden high spike on the 20th day. Another significant spike is seen for the gender_jaccard_normalized variable at day 68. Core_jaccard_normalized has a peak at day 100.Illness_jaccard_normalized peaks at day 75. Most of the values lie low in the range of 0 to 0.05.

III. METHODS

In order to predict the number of Covid-19-19 cases based on given data sets, we have performed exploratory data analysis and trained numerous models using county-level data and topic awareness variables. For this, no external data sets were used.

A. Data Cleaning

Since the county variable is a categorical variable, one hot encoding was performed to create additional 88 variables. The awareness topic variables had three different similarity measure (Jaccard, Cosine, and Intersection), and therefore, a correlation matrix was built to check correlations among these variables. A high correlation was observed among all three similarity measures for each topic and therefore, only Jaccard normalized scores were selected for the model training.

The index column from the test data set was dropped as it was irrelevant to the model. Similarly, the cases column was dropped from the training data set and added as a target variable to the y_train data set. Further, the date_index_converted column in the training data set was converted into a numeric day column.

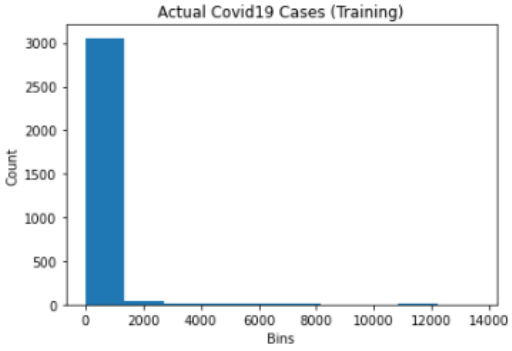

Fig. 6.

As shown in the graph above, most of the Covid-19-19 cases are between 0-1500 range with very few being very high (more than 6000)

B. Exploratory Data Analysis

The total number of Covid-19 cases and deaths per county was visualized to understand the trend over the period of time. It was observed that both cases and deaths follow a similar increasing trend post first 80 days of the pandemic.

i. Correlation Matrix

A correlation heatmap was created to identify pairwise highly correlated variables (threshold = 0.80). It was found that median_household_earnings, median_worker_earnings, median_property_value, and median_housing_cost are all highly correlated, and therefore, only median_property_value was retained.

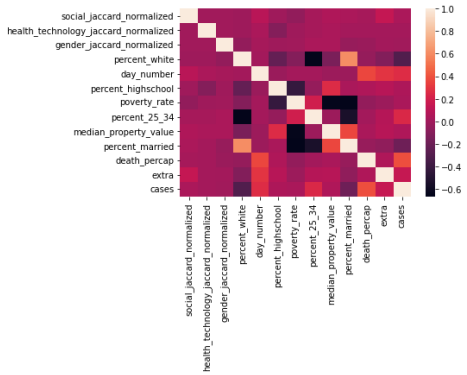

Fig. 7. Correlation Matrix

Similarly, unemployment_rate and poverty_rate, domestic_jaccard_normalized and social_jaccard_normalized showed high correlation, and therefore, unemployment_rate and domestic_jaccard_normalized were dropped respectively. Furthermore, some awareness variables had no impact on the target variable and therefore, were not included in the model development process, namely race, foreign, religion, politics, county_data_length etc.

ii. Feature Addition

core_jaccard_normalized and illness_jaccard_normalized showed a correlation value of 0.78 and therefore, an average of the two variables was used as an independent variable and these two columns were dropped.

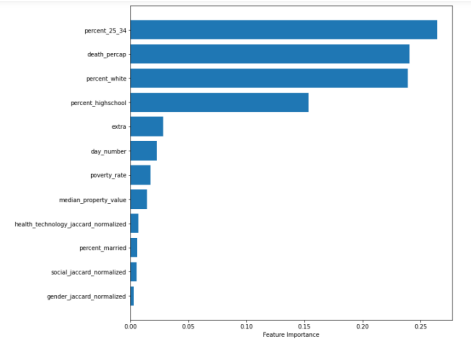

Fig. 8. Feature Importance

Since deaths alone was contributing more than 50% to the target variable prediction, another column death_percap was created by dividing deaths and total_pop column from the data set. This was done to increase feature importance for other variables that would help in better prediction results. With this percent_25_34, death_percap and percent_white became the top 3 most important features for the model.

iii. Normalization

In addition to the normalized awareness variables given in the data set, county-level numeric variables were normalized which did not have any impact on the prediction results, and therefore, normalization was not performed.

iv. Dimensionality Reduction

PCA as a dimensionality reduction technique was performed on the final set of features selected to further reduce the dimensions but it did not provide the desired prediction results and therefore, was not included.

C. Model Development

A variety of regression models were used to predict the Covid-19-19 cases, namely Linear, Logistic, Polynomial, RF, SVM, Gradient Boosting, Bagging, XGBoost, Lasso, and ExtraTreesRegressor. Further, the ensemble method of weighted average was used using Logistic, Decision Tree, and RF in order to get the best results.

Feature importance scores were checked for each model to identify the best possible combination of features for the model that would give the best prediction results. The results were validated using the R2 score.

IV. RESULTS

The best R2 score of 91.927% was achieved for XGBoost model with 100 estimators and therefore, XGBoost model was used to publish the final predicted values. This model significantly improved the accuracy score by 5% as compared to results obtained through Random Forest.

```
# Using cross validation to get mean absolute error which is pretty low
scores = cross_val_score(xgb_r, xtrain_v1, y_train, scoring='neg_mean_absolute_error', cv=cv, n_jobs=-1)
scores = abs(scores)
print('Mean MAE:',(scores.mean()) )
```

```
Mean MAE: 38.995093253463786
```

Fig. 9. Mean Absolute Error

Figure 9 shows the MAE to be around 38 which is low.

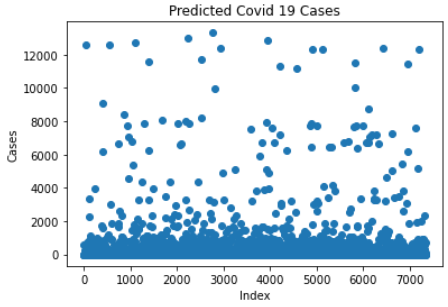

Fig. 10.

Figure 10 above shows that most of the predicted values are close to 0 which is as per expectation. This closely follows the pattern we observed in Covid-19-19 cases in the train data set

V. CONCLUSIONS

To conclude, XGBoost worked out to be the best regression model to predict the number of Covid-19-19 cases. We found deaths to be the most parameter that closely followed the pattern of Covid-19-19 cases which is followed by percent_25_34. Achieving an accuracy of almost 92% was driven by the careful selection of parameters and choosing the right regression model.

REFERENCES

[1] https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.annotate.html
[2] https://matplotlib.org/stable/api/cm_api.html
[3] https://seaborn.pydata.org/generated/seaborn.barplot.html
[4] https://drmattcrooks.medium.com/how-to-set-up-rcparams-in-matplotlib-355a0b9494ec
[5] https://machinelearningmastery.com/xgboost-for-regression/
[6] https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
[7] https://towardsdatascience.com/seaborn-heatmap-for-visualising-data-correlations-66cbef09c1fe
[8] https://www.hindawi.com/journals/sp/2021/5587188/