Predicting COVID Cases in Ohio*

Ajay Patel and Jared Coffey[1]

***Abstract*— The COVID-19 pandemic affected all of the United States in impactful ways. In this analysis, its effect on Ohio is measured through descriptive analysis, and then a predictive model is built to utilize demographic and social data to predict the number of cases on a given day. The descriptive analysis included researching the effect of the pandemic on Ohio, calculating Jaccard similarity-based awareness values by different topics and counties, graphing the number of cases and Deaths per Capita by county, and analyzing the awareness values over time. For our predictive model, we first tried many methods including time-series analysis, random forests, and more. Gradient-Boosted Regressors resulted in the best performance, reaching a 0.9364 R-squared on Kaggle. This lab demonstrates the importance of different reactions and topics to the awareness and protection against COVID-19, while honing in on how a specific state, Ohio, dealt with the crisis and was affected through stratified groups, such as counties.**

I. Introduction

The COVID pandemic led to the analysis of all types of factors and their influence on the spread of cases, from political sentiment to demographic breakdowns. The intent behind this assignment was to conduct descriptive analysis analyzing how the pandemic affected the state of Ohio, especially by county and over time. The second half included a modeling aspect, tasking one to utilize features relating to sentiment of domestic issues, gender, healthcare, politics and many more to predict the number of cases on a given day. Performance of the model was measured using the coefficient of determination, or R-squared.

Going into this lab, we expected, in our descriptive analysis, to see the number of cases and deaths to fluctuate over time and spike. We also thought that politics and healthcare especially would be divisive topics, given the public debate that occurred around them. Health's related Jaccard value turned out to be the lowest at 0.000, matching our expectations. There was a little more consensus in politics at 0.002, but our thoughts proved corrected.

In terms of modeling, at first we expected time-series analysis to work quite well, given that the number of cases was reliant on the day, as they rose over time. However, our initial results using time-series models were very poor, with reported R-squared values under 0.10, leading to a shift towards tree-based modeling, including Gradient Boosted Regressors and XGBoost, which we expected to perform well due to the Random-Forest base of these models. XGBoost performed quite well, resulting in a R-squared value around 0.87 on Kaggle, but our best iteration came using Gradient Boosted Regressors, which resulted in a 0.93 R-squared on Kaggle. Performance on the training and test sets before submitting predictions was just as high, if not better, leading us to feel comfortable with our final predictions on Kaggle.

*This work was not supported by any organization

[1]Ajay Patel and Jared Coffey are Undergraduate Students studying Data Science at the University of Rochester

II. Data

A.

Before diving into the data, it was necessary to get an understanding of how Ohio was affected and dealt with the pandemic. The COVID-19 pandemic first reached Ohio on March 9, 2020 with its first death coming on March 19, 2020. In terms of intensity, the state saw spikes similar to the rest of the country, by total number of cases and also deaths. The state's response to the pandemic included mask mandates, social restrictions on gathering sizes, and social distancing, again similar to the rest of the country. The state was actually precautionary with some of their measures, announcing the pandemic was a concern before a single case was detected and shutting down schools for multiple weeks as soon as the first case came. As cases dwindled, they slowly reopened key businesses or essential buildings, but were still very cautious. Similar to other states, they also had to scramble to increase hospital capacity and provide for essential workers who were working long hours in tough conditions. Vaccines were implemented as soon as possible, with natural push back from citizens through vaccine hesitancy along with supply issues to make sure everyone received one. While the tolls of the pandemic were certainly felt through cases and deaths, employment, and other impacts, Ohio did do relatively better handling the pandemic than a lot of states. Their preemptive action did help reduce the number of deaths, such as their aggressive closing of schools and unique vaccine rollout programs. It's hard to say a state dealt well with the pandemic given the holistic damage it caused, but Ohio was one of the seemingly well-equipped states thanks to their government and citizens' cooperation.

B.

The average values for all the topic awareness values in the dataset were next analyzed. These topics include common events or discussion points, such as sports, entertainment, race, politics, ideology, and more. The results were plotted in figure one, where one can see that the topics of sports, entertainment, and illness had the three highest average Jaccard similarity-based values (all plots

can be viewed in larger size in the Appendix). Topics such as health technology, ideology, and health had values on the opposite side, as they were the three lowest. This indicates that the former three topics and higher valued topics possibly had more of a consensus in awareness or agreement in discussion surrounding those topics, versus the latter three and lower valued topics, which could have been met with less awareness or disagreement. However, it is worth noting that the range of values is not too wide, as the highest is around 0.016 with the lowest around 0, as Jaccard values can range from 0 to 1. While there are differences that draw attention, this nuance is worth keeping in mind.

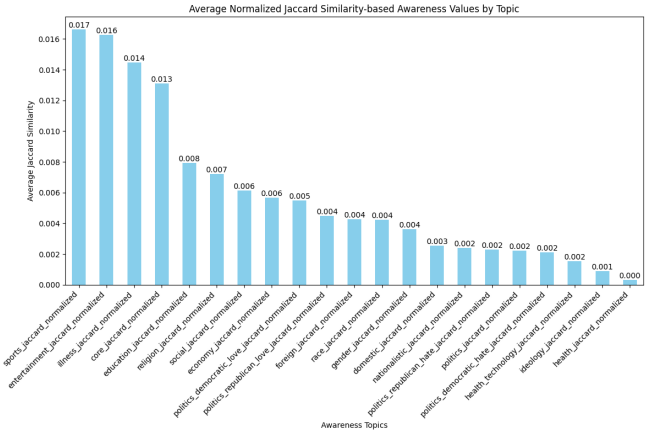

Fig. 1. Average Values For All Topic Awareness Values

C.

Next, the core Jaccard normalized feature was isolated and analyzed for the mean value by county to delve into which counties had the highest awareness values and which had the lowest. The results can be found in figure two, where it's evidenced that Delaware, Richland, and Perry counties had the highest values. On the other hand, counties including Champaign, Highland, and Paulding had values that were practically 0, indicating almost no similarity or agreement in discussion on topics that fell under the core denotation.

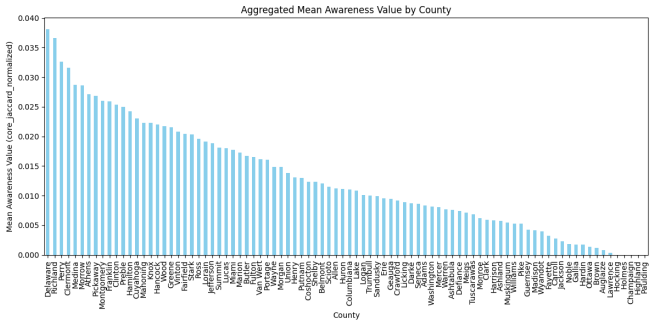

Fig. 2. Mean Awareness Value By County

Afterwards, the breakdown of cases and deaths per capita were analyzed by county to visualize which counties were affected by the pandemic the hardest, and which handled it relatively well. Analyzing figure three, one can see that Pickaway county was hit especially hard. They had about 88,759 cases per 100,000 people, an especially high rate when compared to the next highest, which was Marion county at 12,600 cases per 100,000.

D.

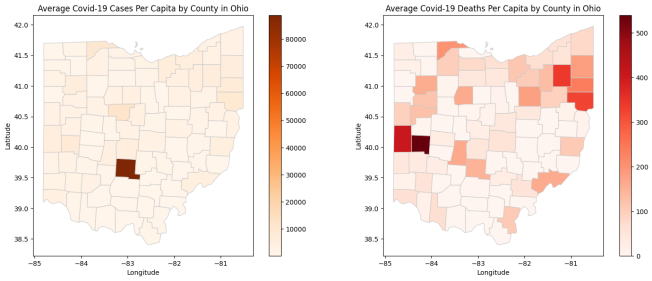

Fig. 3. Cases and Deaths Per Capita By County

It is fair to hypothesize that something went wrong in Pickaway county in terms of control or prevention to lead to such a bad outbreak. Somehow, they did not suffer the highest rate of death, as that was Miami and then Darke county, coming in at 539 and 406 deaths per 100,000 people, respectively. Most counties demonstrate similar rates of cases and deaths, shown by the color scaling in figure three, but there definitely were a few outliers that struggled with prevention.

E.

Lastly, all the Jaccard awareness scores related to key topics were analyzed over time to see the evolution of awareness for each topic. Taking a look at figure four shows that most of the Jaccard scores stayed relatively even over time, but there were a few key spikes that draw attention. The Jaccard score for race had a huge jump around day twenty, reaching a value past 0.30 that no other topic came close to, which indicates that something occurred around that time that led to an agreement in race-related discussion. Another spike of lesser magnitude is seen in gender around day seventy, coming close to 0.15, meaning again that something happened around this day that led to more consensus in gender related topics than usual.

III. Methods

Based upon the nature of this problem and the prediction of a continuous dependent variable, five regressor-based models were chosen. Linear Regression served as a benchmark standard, while other models such as XGBoost and Gradient Boost were used to test different theories on ways to optimize the prediction model. The results of each different

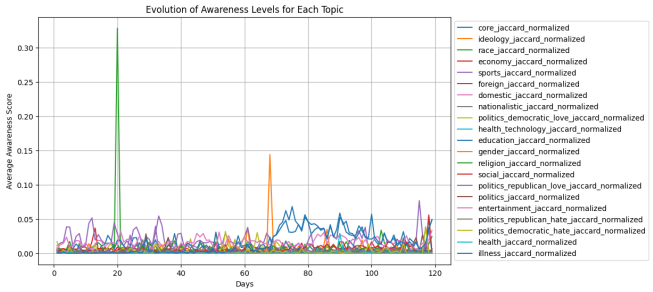

Fig. 4. Topic Awareness Values By Day

model are detailed in our Results section, utilizing R-squared as our evaluation metric.

For features selection, all numeric variables were initially used in all models. However, this greatly increased the runtime, but did not seem to result in a significantly high accuracy for any of the models. Thus, a list of county-related columns were compiled to be the most representative features out of the 145 different given variables and simplify the models in use. These were chosen off intuition as to what matters towards the number of cases but also what had strong correlations with that value, leading us to believe they would be strong predictors. The following were the columns chosen:

- 'date_int'
- 'total_pop'
- 'deaths'
- 'county_data_length'
- 'percent_25_34'
- 'labor_force_rate'
- 'unemployment_rate'
- 'median_household_earnings'

Surprisingly, it was found that predictions improved without using the date column provided at first. We expected it to matter given it's logical that cases would increase over time, but it seems any information from the date was already captured in other features, such as deaths, and not necessary. However, after experimentation, our best result came with the date column included. The strong performance of the model without the date included is something that raised our attention and could warrant further investigation.

The train/test split was originally the standard 80/20, but this was later changed to 90/10, as the amount of given training data was too little, especially compared to the actual test data needed for submission. Thus, 10% of the given dataset was allocated to model validation when testing different optimizations and regressors. Once optimal parameters were chosen, it was then refitted using the entirety of the given dataset to give as much data as possible to train on when predicting the submission data. These changes were found to significantly improve the accuracy scores of all models. PCA was considered, but ultimately not used, as it was felt that there were enough columns of significant value to put into our model, but not enough to necessitate any dimensionality reductions. We experimented with aggregating other features, such as the Jaccard values, but found no improvement in the model's performance, leading us to stick with what we had already done.

IV. Results

As shown, the Gradient Boost Regressor gave the most optimized results. Not only did it have the highest accuracy scores for all datasets, but it consistently provided an increase in accuracy as more and more parameters were manually optimized via grid search. On the contrary, the other models tended to plateau around 80% in terms of accuracy.

Since Gradient Boost Regressors are fairly robust and secure against overfitting, the number of model estimators in use was increased significantly to test the results, as was the depth of the trees and nodes within the regressor. Even though there was still a risk of overfitting, it was found that models utilizing higher numbers of estimators tended to perform well on *all* datasets. In general, Random-Forest-based regressors (i.e.: Random Forest, XGBoost, and Gradient Boosting Regression) tended to do extremely well with this form of regression, scoring in the 90% accuracy range for the training and testing sets. Nevertheless, there was a noticeable difference between model accuracy values when it came to their evaluation scores, with the Gradient Boosting Regressor being the most accurate and most consistent model.

TABLE I
Regression Models and R-Squared

Model Choice (Regressor)	Training Score	Testing Score	Evaluation Score
Linear Regression	0.6821	0.3498	0.5322
Random Forest Regressor	0.9860	0.9268	0.7851
XGBoost Regressor	0.9984	0.9499	0.8741
Gradient Boost Regressor	0.9999	0.9935	0.9364

V. Project Evaluation

Throughout the entire process - exploratory analysis, preprocessing, model selection, optimization - thorough testing was applied and each step was optimized to the best of the team's ability.

One of the biggest problems encountered during the project was the high dimenensionality of our feature set, combined with the relatively small dataset size. The evaluation dataset had 7,000 rows, but the entire training set had less than 1,000 rows which then was further split into train and validation sets. Not only this, but the large amount of features were naturally hard to navigate and choose between when considering model inputs. Given more time, it would be beneficial to deeply examine each variable and standardize each accordingly. Whether this was done through more rigorous feature selection or the employment of feature reduction to aggregate info

For future improvements for this model, it would be interesting to examine a possible mixture between a Random Forest-based Ensemble Model and a Time-Series Analysis. Further data augmentation to the time-based fields would

perhaps provide better context and significance for the other variable. If we had more time to experiment with these possible steps and tuning any time-series related models, our results could have been improved but what we used still performed well. The most successful models used in this project included little to no time-related data; however, additional data manipulation to these time-related fields could prove beneficial if used correctly and given enough time.

VI. Appendix

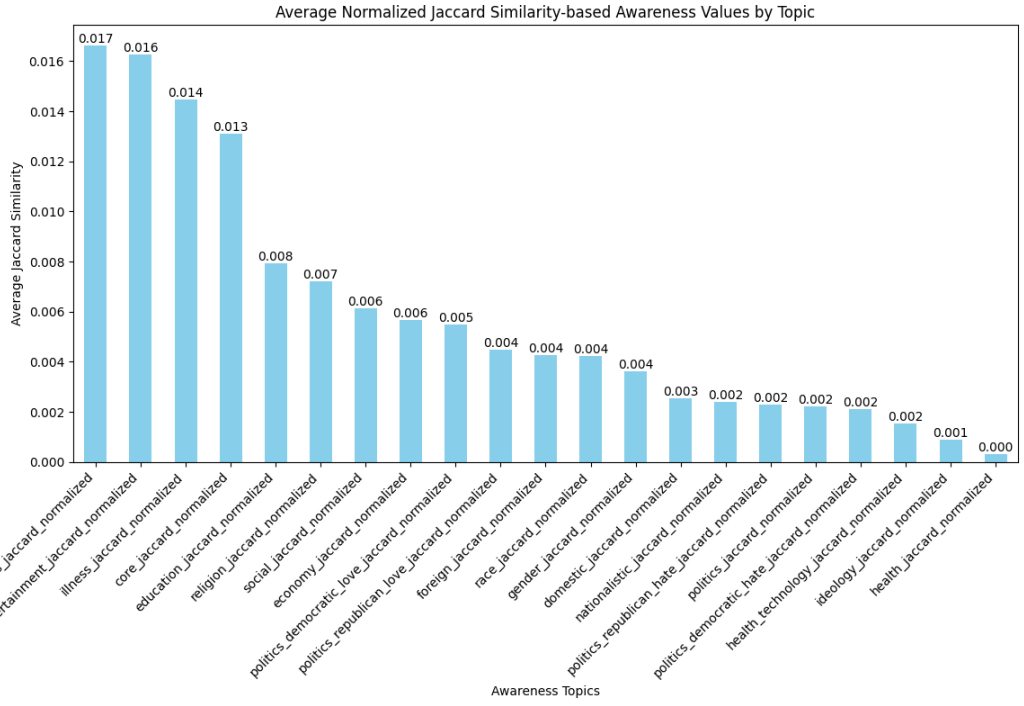

Fig. 5. Average Values For All Topic Awareness Values

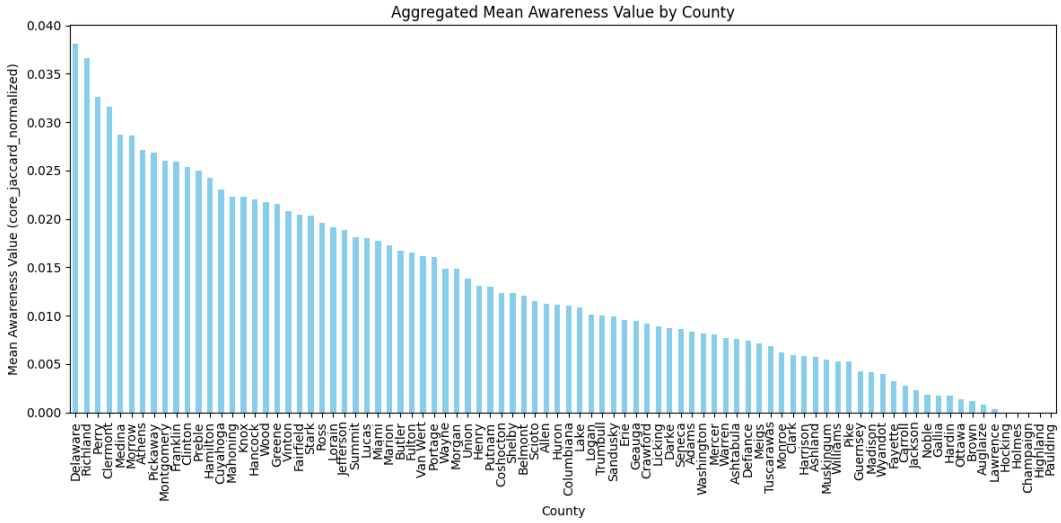

Fig. 6. Mean Awareness Value By County

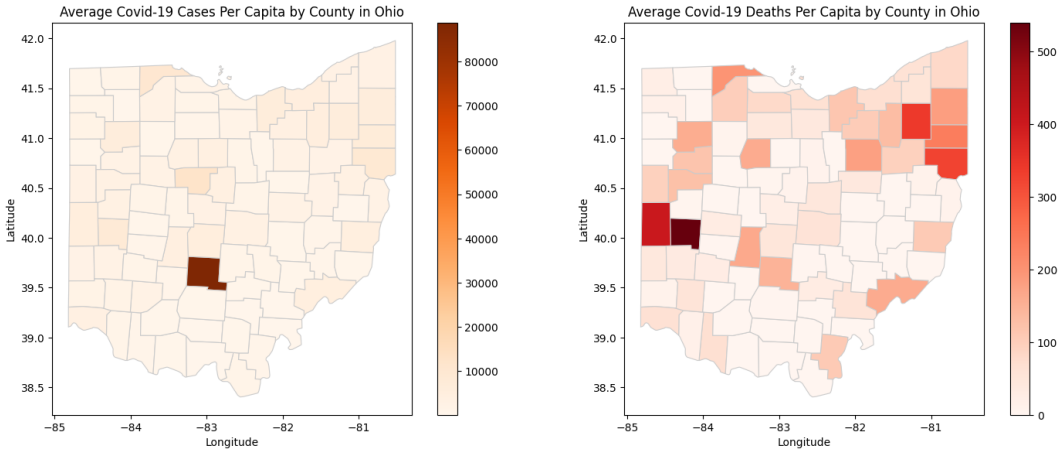

Fig. 7. Cases and Deaths Per Capita By County

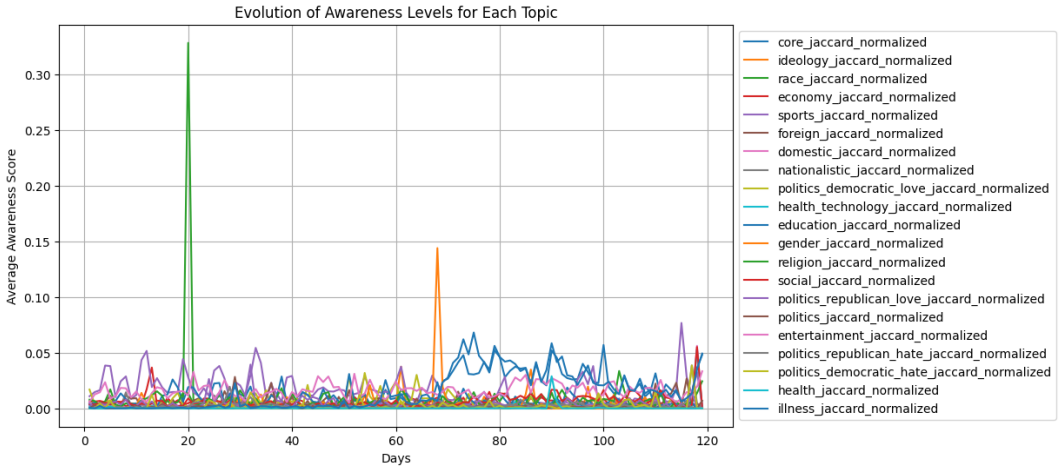

Fig. 8. Topic Awareness Values By Day