Machine Learning Approach to Predicting COVID-19 Cases in Ohio State

Harshal Talele
University of Rochester
Rochester, US
htalele@ur.rochester.edu

Harshalkumar Loya
University of Rochester
Rochester, US
hloya@ur.rochester.edu

***Abstract*—The paper proposes a machine learning approach to predict covid-19 cases by using a time series data of Ohio county in the initial weeks of outbreak to train the model. It also aims to explore relation between various social awareness features derived by the tweet data and case prediction. We compared the performance of myriad machine learning models such as Linear Regression, XGboost, Random Forest and Decision Tree and selected the best performed model on R2 metric. The feature engineering and modelling used in this can help policy makers and health officials to make informed decision and take early steps to tackle the pandemic.**

***Index Terms*—Covid-19, Time-Series Analysis, Random Forest, Decision Tree, XGBoost, Machine Learning, Prediction, Jaccard-Similarity, Cosine-Similarity, Intersection-Similarity**

I. Introduction

The once-in-a-lifetime pandemic of novel coronavirus has has sever health, social and economical impact on all cross the world. In this research, we focus our attention on the USA's Ohio state and deep dive into the counties to explore the severity of the case load, deaths, the awareness of various socioeconomic topic as features to predict the covid cases. The disease outbreak is an exceptional situation and it is very unexceptional for any state or county administration to be completely prepared for the same. The situation evolves rapidly to make the facts known with time cutting through the clutter of hoax and misinformation. Ohio has proven to be one of the early states to 'sound-the-alarm' as it's governor Mr. DeWine released a stay-at-home advisory within few days of the first case of the disease was detected. The state administration's response to the pandemic, highlighting the proactive measures taken and public opinion supported by news media. Ohio experienced the advantages of early intervention by its government and medical community, with fewer than 1/3 as many cases and significantly fewer deaths than nearby comparably-sized states like Illinois, Pennsylvania, and Michigan. The criticism of state policies is briefly mentioned, hinting that it is primarily politically motivated and comes from COVID deniers within the Republican party. The outcome of these measures is reflected in the state's high vaccination coverage.

The progression of cases and deaths in can be seen with the two major waves occurred in the two years following the pandemic's onset. The first wave peaked in November-December before a period of normalcy, and then a larger second wave peaked in January with 30,000 daily cases. Deaths followed the case peaks but were better managed in the second wave despite a threefold increase in cases compared to the first wave. In total, Ohio reported 553,461 cases, leading to 31,803 hospitalizations and 7,477 deaths.

In such cases, it is imperative for the data science community to step in and leverage the power of data and machine learning to build models to predict the cases to be able for medical practitioners and policy makes to move faster and have empirical evidence to paint a clearer picture of the situation. Many features can present themselves as helpful int he prediction. we analyze the relation between several social, political, economic and demographic features to gauge which could be used for a near perfect prediction. Also, we have used Jaccard Similarity, Intersection Similarity and cosine similarity and their normalized score derived from the twitter data of the Ohio county users. Twitter data opens up new horizons for scientists, both as a rich data source in its own right and also as a way of gathering information from the public. Twitter helps in gaining public opinion on various topics and it is interesting to see how the discourse change as the pandemic progresses and if that could help in deriving the insights. The Twitter data help to get the equitable and real-time public opinion.

From a machine learning perspective, this is a regression problem which leverages models like Random Forest Regressor, Gradient Boosting Regressor, Extra Trees Regressor, XGBoost Regressor and others. The results from multiple models were compared depending on the features selected as an output of feature engineering. The best performer in this case is the XGBoost Regressor.

II. Data and Methodology

In this section, we describe the data used, data manipulation and the methodologies used to analyze the data

A. Descriptive Analysis

I. The data is best described when it is visualized. We have plotted a time series graph of the data of make sense of the data and to see what type of pattern of data we are dealing with in Fig 1. As is evident by the plot, we are given a data of 3 months to make the predictions. The case load shoots up after the around 80th day and is increasing exponentially. The

death toll start to rise and see a upward trajectory as we reach the end of third month.

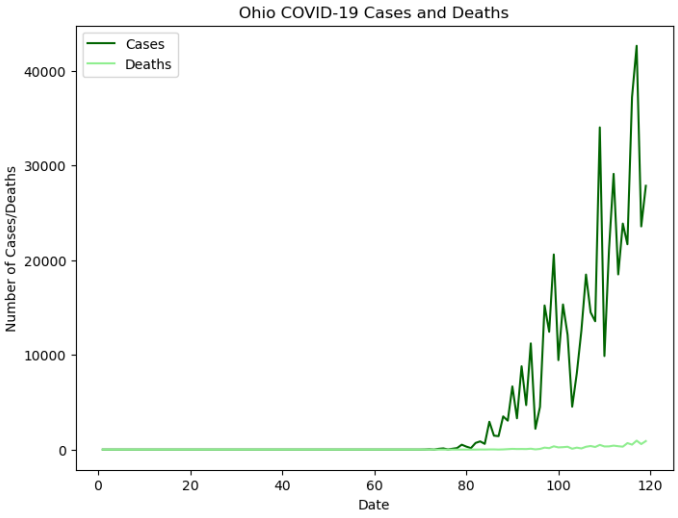

Fig. 1. Time-Series cases and deaths

II. The bar plot diagram in Fig 2 depicts the average normalized Jaccard similarity-based awareness values on different topics. Each bar represents a specific topic, and the height of the bar indicates the level of awareness among residents of the county about that particular topic. A higher Jaccard similarity-based awareness score for a topic indicates that the topic is more prevalent in the minds of the people in that county.

Interestingly, the sports, entertainment, and illness topics have outperformed the general COVID-19 awareness score, indicating that these topics have managed to grab people's attention and act as a unifier in a time of global pandemic. However, issues related to society, politics, religion, and the economy are not as prevalent in people's minds, with most of them having an awareness score of less than 0.01.

Surprisingly, discussions related to health and health technologies have the lowest awareness score among all topics. This finding may come as a surprise, given that one would expect people to be highly focused on the pandemic's nitty-gritty details. However, it is possible that Twitter lacks nuanced discussions, and people tend to view the world in black and white, which could explain this result. III. The bar plot diagram displays important information about the level of COVID-19 awareness among the counties in Ohio, based on their core_jaccard_normalized Jaccard similarity scores. The graph reveals that some counties, such as Delaware, Richland, and Perry, have achieved the highest awareness scores, while others like Lawrence, Champaign, Hocking, Holmes, and Paulding have a score of zero.

The higher scores may be attributed to factors such as better internet access, an educated population, and exposure to mainstream news outlets. These counties are generally located within or around major cities in Ohio. On the other hand, counties with low awareness scores tend to be more remote and isolated from current events.

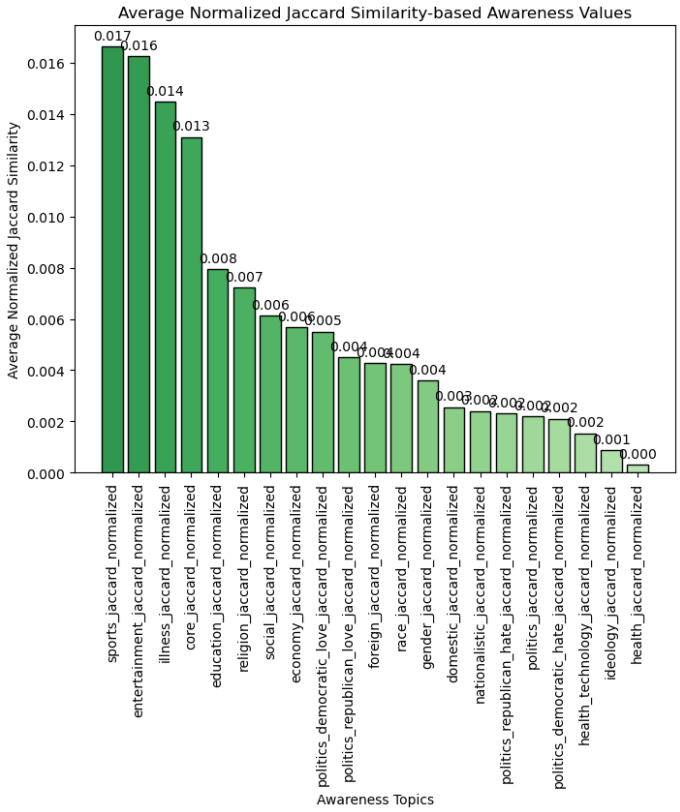

Fig. 2. Average Normalized Jaccard Similarity-based Awareness Values

This graph could serve as a useful tool for policymakers in Ohio to target counties with low awareness levels and increase their awareness campaigns. It is crucial to improve the awareness of citizens in all counties, as it can significantly impact the overall health and well-being of the state.

IV. The multi-line chart graphically represents the awareness of various topics related to COVID-19 over time. The topics are manually labeled based on Jaccard similarity measure, and they range from general COVID-19 awareness to social and political events, as seen in the legend table. This data provides a valuable commentary on how different topics gain traction and how a global pandemic can impact the societal conversation.

The initial period shows that topics related to Entertainment, Religion, and Sports were widely discussed and dominated social media conversations. General COVID-19 awareness was not a priority during this period. However, as the number of cases started to rise, the later part of the time series plot showed a significant concentration of topics related to health, health technology, and general COVID-19 awareness. Political discussions about support and dislike towards political parties remained highly volatile throughout the entire period.

The multi-line chart provides a unique view of the evolution of topics and how they are prioritized as the pandemic progresses. It can serve as a useful tool for policymakers and public health officials to understand the societal conversation

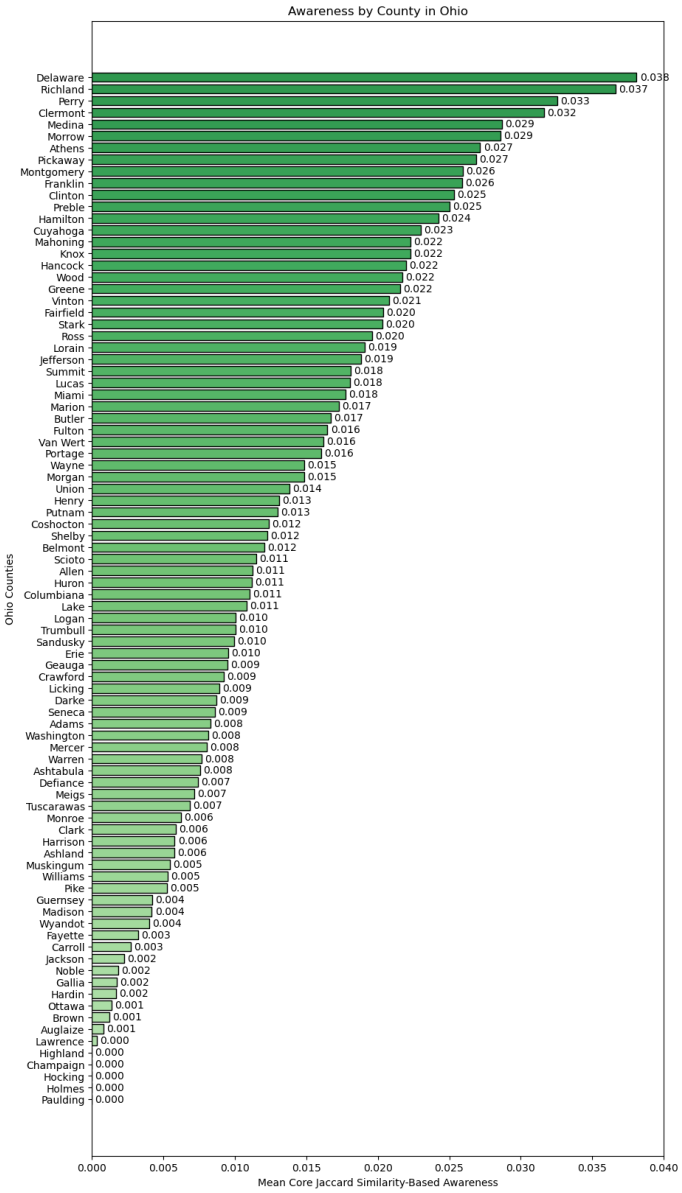

Fig. 3. Time

and tailor their messaging to reach a broader audience.

V. The Map has the cases and deaths per capita projected onto it for better visualization and it provides valuable information on the top counties in Ohio with the highest number of cases and deaths per capita. However, it is essential to note that various factors, such as population density, demographics, and healthcare resources, can significantly influence the number of cases and deaths per capita in a county.

The cases per capita show a considerable disparity among the counties, with Pickaway being an outlier, having seven times the number of cases compared to the next highest county. The northern and eastern parts of the state have a higher density of cases, while the southern part is relatively isolated from the high volume of cases.

Interestingly, the county with the highest number of cases

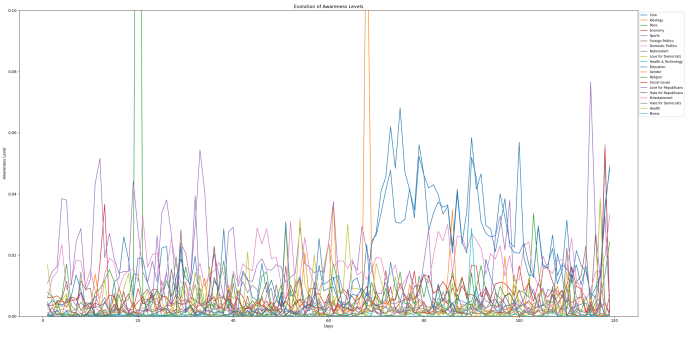

Fig. 4. Evolution of Awareness Levels

does not feature in the top 5 counties with the most deaths per capita. Miami County has the highest deaths per capita, with 539 deaths per 100,000 people, followed by Darke and Portage counties. It is worth noting that almost half of the counties in Ohio had virtually zero deaths per capita.

To further visualize the distribution of cases and deaths in Ohio, projecting the data onto a map of the state would provide a more comprehensive understanding. The map could highlight the high and low density areas of cases and deaths across the state, enabling policymakers and health officials to identify hotspots and prioritize resources accordingly.

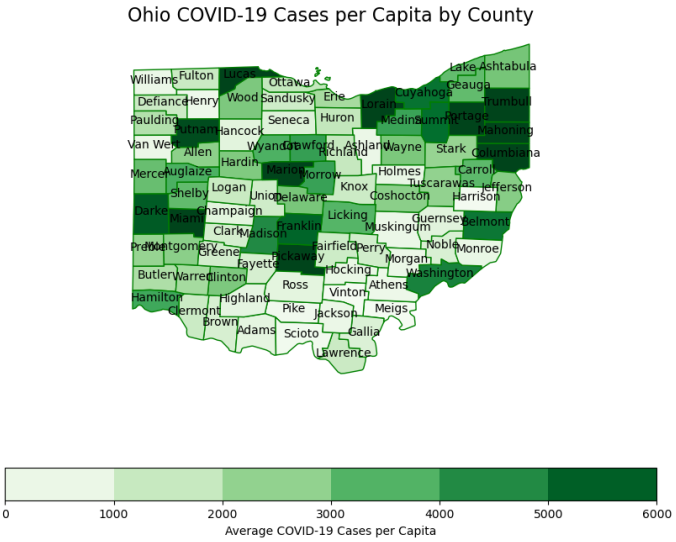

Fig. 5. Ohio COVID-19 Cases per Capita by County

B. *Data Pre-Processing*

The training data contains 147 columns which cannot be used directly for model building. Data pre-processing mainly takes care of the reduction of the features and making sure they are in the correct format and range. The categorical columns like *county* can be converted into numerical by using one-hot encoder. However, that would lead to a large increase in the number of features. Dimensionality reduction using PCA is implemented to the output of one-hot encoded values for

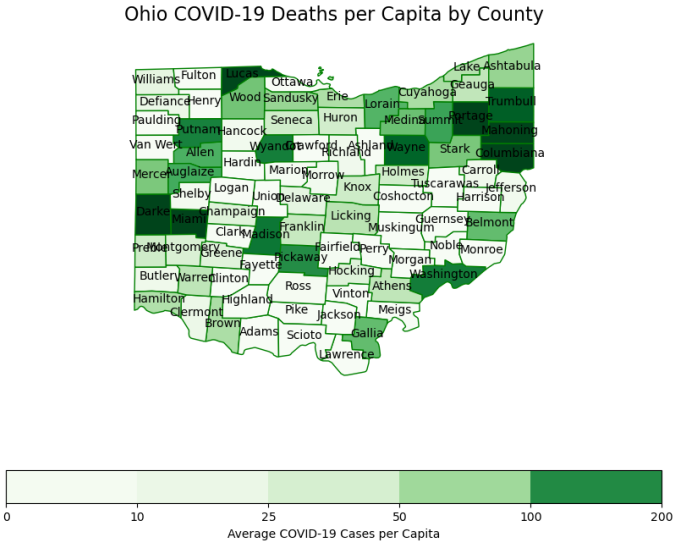

Fig. 6. Ohio COVID-19 Deaths per Capita by County

county to get a single output capturing the essence of the categorical variables. Similar was implemented and tested for the social awareness features that came for Twitter data.

The date feature is a potentially important one as it captures the trend of cases over time. However, because it is a string, it cannot be directly used for training. String transformation is implemented to strip the date of unnecessary text and ultimately converted into an integer making it a numerical feature.

For use of decision tree bases regressors, the trees can better fit on ordinal features and potentially give better results. Binning was implemented to convert features like total_pop, unemployment_rate, median_household_earnings, poverty_rate to convert them from numerical to ordinal features.

C. Machine Learning Modelling

Initially, the training data was further split into train-test groups to compare the regression models. Upon initial comparison, we decided to focus on Random Forest Regressor, Gradient Boosting Regressor, Extra Trees Regressor and XGBoost Regressor for they had the best r^2 scores. The feature set was divided into two groups, the county-based columns and the awareness-based columns from social media. Features from these two sets were first checked for any correlation. The ones having a high correlation between themselves like median_housing_cost, median_household_earnings, median_worker_earnings, median_property_value were filtered to keep one of the features for model training. The remaining features were iteratively tested in different combinations with the regression models in focus to give the best r^2 output. The $feature_importances$ attribute from sci-kit learn library of the model trained was used to check for the least important features that could potentially be discarded from the training process.

III. Results

Model used - XGBoost Regressor

R2 - 0.89

Features used - $date\_int$, $total\_pop$, $deaths$, $county\_data\_length$, $percent\_25\_34$, $labor\_force\_rate$, $unemployment\_rate$, $median\_household\_earnings$, $core\_intersection$, $health\_technology\_cosine\_normalized$, $politics\_democratic\_hate\_intersection$, $race\_cosine$

IV. Conclusion

In conclusion, The machine learning approach to predict the covid-19 cases works well given the accurate data pre-processing and feature engineering on the data. We could predict the cases with high R2 score of 89%. These results and research in this field can help to improve the accuracy and speed of COVID-19 diagnosis and improve disease outcomes. It was an interesting observation to note how a society deals with a once-in-a-lifetime disease and changes its discourse and discussion around it as things reveal. The results reveal that topics such as sports and entertainment can act as unifies during a global pandemic, while health-related discussions remain low. The findings provide insights into how policymakers can use machine learning to target awareness campaigns to improve public health outcomes.