Covid-19 Cases in Ohio.

Aparna Palit[1] and Aditya Chilla[21]

***Abstract*— Covid-19 has been a major concern all around the world and Ohio in the United state was no exception. All the 88 counties were affected by the Covid-19 pandemic. The data is split into training and testing sets, and the training data is preprocessed to select only relevant features and remove rows with missing values. L1 regularization with logistic regression is applied to select features, and the selected features are used to train an XGBoost model to predict the cases in the test dataset and out of which the model has given the best R2 score out of the two.**

I. INTRODUCTION

The main task of the challenge was to predict the number of COVID-19 cases in all of the 88 counties of Ohio. The dataset provided for the challenge included millions of tweets, including around 46 million tweets and 91,000 users from Ohio. These tweets were collected using certain awareness hashtags, such as COVID19, coronavirus, and education. To analyze the data, we first created a bar plot to visualize the trend of COVID-19 cases in Ohio over time. We found that for about 70 days, there were no reported cases in the state. We used this observation to build our prediction model, where we excluded the rows where cases were zero until the "date index converted reached 70 As shown in Fig 1. This preprocessing step helped to remove unnecessary data and improved the accuracy of our model.

For our prediction model, we used XGBoost Regressor, which is a powerful algorithm for predicting numerical data. To create the model, we first split the dataset into training and testing sets and then applied L1 regularization with logistic regression to select the best features. We trained the model on the selected features and made predictions on the test set. We achieved an accuracy of 0.93, which is a very promising result.

We employed the R2 score statistic to assess our model's accuracy. The R2 score quantifies the proportion of variance in the dependent variable (cases) that can be predicted by the independent variables (features). Our model received a high R2 score of 0.93, indicating that it predicts the number of COVID-19 cases quite accurately. Overall, our prediction model and data analysis provide useful information about the COVID-19 situation in Ohio. The findings of our study can assist policymakers and healthcare professionals in making educated decisions about the pandemic. Furthermore, using similar datasets, our model can be used to predict the spread of the COVID-19 virus in other regions.

[2]University of Rochester

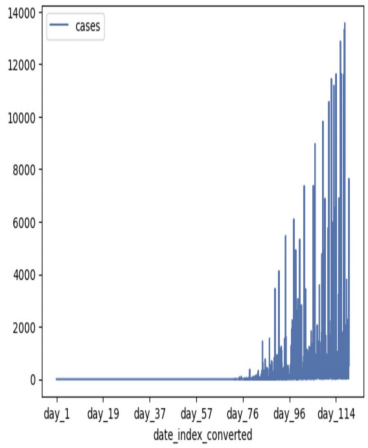

Fig. 1. Number of Cases per Day

II. DATA

The data used contains measurements on the level of awareness about Covid-19-related topics in all counties of Ohio during the pandemic, the number of Covid-19 cases, and the number of people infected with Covid-19. There are 3141 records in the training dataset and 7331 recordings in the test dataset. The awareness data has been extracted from over 46,000,000 tweets posted by over 91,000 users in Ohio. Different similarity measures (Jaccard, Cosine, Intersection) were used to detect the intensity of discussion on topics related to Covid-19. We performed initial Exploratory Data Analysis (EDA) using a bar plot of the number of cases over the 120 days of data collected, and observed that the number of cases is zero for the first 70 days as shown in Figure 1. We removed the first 70 days data and trained our model with the rest of the data which increased our accuracy. With the EDA we also observed that Delaware has the highest level of awareness of COVID-19 among all other counties. The highest number of per capita cases was observed to be in Pickaway, Marion and Lucas, and the highest number of per capita deaths were observed to be in Miami, Darke and Columbiana as indicated by Figure 2 and 3.

III. METHODS

To predict the number of cases in Ohio, we first utilized XGBoost. This powerful ensemble learning algorithm can combine the outputs of many individual models (called weak learners) to make more accurate predictions. XGBoost builds models iteratively, and at each iteration, it tries to correct the errors of the previous model. The final prediction is obtained by summing the predictions of all the models.

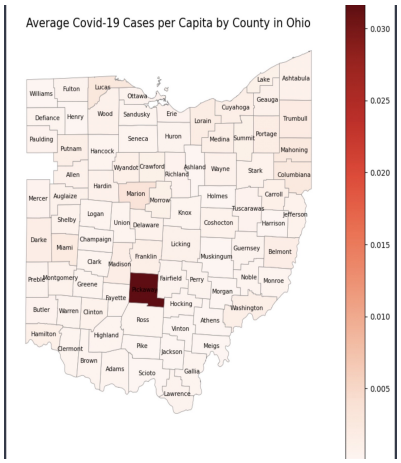

Fig. 2. Average Covid-19 Cases per Capita

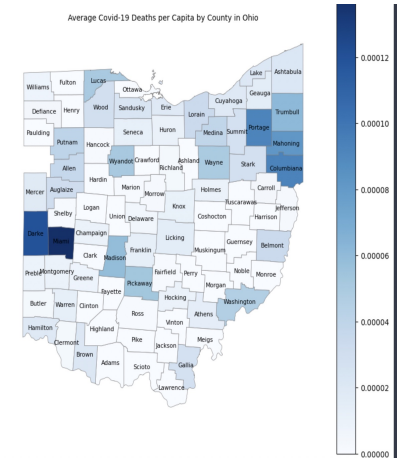

Fig. 3. Average Covid-19 Deaths per Capita

Before feeding the data to our model, we split the dataset into training and validation sets in an 80-20 ratio. This helps us to tune our model and evaluate its performance before testing it on unseen data. When we first trained the model, we obtained an R2 score of 0.86, which is a decent score but could still be improved.

To further enhance the performance of our model, we applied Lasso regularization to logistic regression. Lasso, also known as L1 regularization, is a linear regression regularization technique that can address the issue of overfitting in a model. Overfitting happens when the model fits the training data too well, so it struggles to generalize well on unseen data. Lasso adds a penalty term to the cost function of the linear regression model, which shrinks the coefficients of the less important features to zero, effectively removing them from the model. This helps to reduce the model's complexity and improve its generalization performance.

We also used the MinMax Scaler function to do 0-1 normalization on the features of our data. After normalizing the training data, we utilized Lasso regularization to choose the most essential features using the Lasso coefficients. The normalized dataset was then divided based on the selected attributes.

With the help of Lasso regularization, we raised our model's R2 score to 0.93. This greatly improved above our initial R2 score of 0.86 before applying Lasso regularization. With our model's enhanced performance, we could reduce the characteristics of the test data while properly predicting the number of cases.

Overall, we used XGBoost as the principal model and Lasso regularization as a strategy to improve its performance. With an R2 value of 0.93, we were able to forecast the number of cases in Ohio using a combination of models and methodologies.

IV. RESULTS

Overall in this project, we have used data preprocessing, feature selection, and regularization techniques to build an accurate and good-performing model. With the help of XGBoost, Lasso, and MinMaxScaler normalization techniques, we improved our R2 score from 0.86 to 0.93 in predicting the number of cases in Ohio.

REFERENCES

https://en.wikipedia.org/wiki/Ohio