Data Science Project Stage III

Pratik Devkota

CSC 605

State of choice - North Carolina

Like before, I chose North Carolina for Stage III of this project. Initially, I loaded the super dataset, pulled out NC data, and created two separate dataframes for new daily cases and deaths. Additionally, I removed all columns with 0 recordings in NC for both new cases and deaths.

Linear and non-linear regression

To create linear regression and polynomial regression model for NC cases and deaths, I used statsmodels's OLS function with number of days since first covid case as independent variable and the number of new case (cases or deaths) as dependent variable. Once the data is fitted into the models, an array with next 7 days was provided as input to the models to get the predictions. With RMSE between actual number of cases and predicted number of cases, I found out that RMSE was very high for linear regression models.

RMSE for different models

To reduce RMSE, I added complexity to the model by transforming the independent variable to higher degree features. With added complexity, polynomial regression fitted our data much better. RMSE was then reduced from 2433 to 2030. There were some outliers that contributed to such a large value of RMSE. After ignoring two of the outliers, RMSE dropped down to 1436.

Trends of top 5 infected counties

To describe the trends of the top 5 infected counties, I pulled 5 counties with the highest number of cases till the last date. A function was defined to build linear and polynomial regression models for a county. Using the function inside a for loop for the five counties, their respective plots were generated. RMSE for both linear and non-linear models were calculated for all the counties. Using the polynomial regression model, the number of new cases for the next 7 days was predicted and the trend was displayed in a continuous line graph.

Point of no return

To calculate the point of no return, I used the hospital enrichment dataset. Using the number of **ICU beds** and **bed utilization**, I calculated the total number of **available beds** across NC. Deaths per covid case was calculated using the cases and deaths values for the last 100 days. For every new prediction of cases in the upcoming days, the **product** of **predicted cases** and **deaths per case** gives us the probable number of

new deaths for that day. If the rate of new covid cases per day keeps on increasing, it is highly likely that one day, the number of deaths will be more than the number of ICU beds available. That day will be our point of no return. To calculate that, I checked the number of probable deaths for the next 100 days. Till the 71st day, the number of covid patients with higher health risk was 915, lower than 918 (number of available ICU beds). On the 72nd day, patients with higher health risk reached 966 and so, **72nd day from the last day of cases is the point of no return** (if all other contributing factors remain unchanged).

Hypothesis testing

My hypothesis in stage II was: **Higher number of population between 35 and 44 years of age contributes in increasing the daily covid cases.** Initially, I created two groups (high and low): one with daily cases for 6 counties with highest population between 35-44 and another with daily cases for 6 counties with lowest population between 35-44. To start with two tail two sample t - test, the null hypothesis is that the number of covid cases in the higher group is similar to the number of covid cases in the lower group. Alternative hypothesis, on the other hand, is that the number of covid cases in the higher group is not similar to that of the lower group. Using statsmodels's independent t-test method (ttest_ind), t-statistic and p-value were found to be 2.3933 and 0.0167. Since the p-value is smaller than 0.05 (with 95% significance level), we reject the null hypothesis and accept the alternative hypothesis. To understand if the difference is higher or lower, I performed one tail two sample t-test. Since p/2 was less than 0.5 and t-statistic was greater than 0, I deduced that the number of covid cases in group 'high' is higher than that of group 'low'. Hence, our hypothesis that counties with a higher population between 35 - 44 years of age are more likely to see a higher number of covid cases per day was accepted.