Stage IV (Basic Machine Learning)

Priyanka Budavi

Team Task: Develop Linear and Non-Linear (polynomial) regression models for predicting cases and deaths in US

The linear Regression model for was developed using the dataset generated in Stage I. For this task my responsibility was to code for the Linear Regression and help my other team members to finish the team task.

- In the team task I created a linear model considering the new cases and deaths against the first occurrence of cases and deaths.
- Also checked whether the model was predicting new values based on the model trained.

Inferences:

- Thus, we see the regression line fitting through the datapoints in all the Linear and Non-Linear Model.
- The trend seems increasing for new cases the deaths. For few days the cases and deaths increasing, and the number goes down drastically.

Apart from the modelling I also contributed to find last week prediction along with the trend line and helped my other team member.

Note: For more details view the report and the notebook uploaded on GitHub.

Member Task:

Task1: Linear and Non-Linear (polynomial) regression models to compare trends for a single state and its counties.

Linear and Non-linear Regression for Alabama State

Inference for Alabama State:

- The cases and deaths seem to be increasing in both the models.
- For Linear Regression model, the regression line seems to fit the data points in an increasing manner.
- The confidence intervals can also be viewed along the regression line.
- The prediction line and trends also seem increasing for cases and deaths.
- For Non-Linear, the regression line seems to be increasing for both the cases and deaths.
- The regression line with polynomial degree 4 seems to best fit the line compared to other degrees.

Comparison of Alabama state with the counties

Inferences:

- For this task I used the dataset generated in Stage 1 and filtered the data by state and try to perform Linear and Non-Linear Model on the dataset.
- I performed the same by sorting the data by top affected counties both new cases and new deaths. Further I performed the regression on the dataset for both the categories and calculated the RMSE values.
- From the graphs we observe that the trends are increasing. So, when the cases were high in the counties the overall increase was seen in the state.
- The pattern for new cases was found increasing and deaths were high during some intervals and low during the other.
- Also, the polynomial of degree 4 was best fitting the regression line when compared to the other degree.
- In all the state and counties, the polynomial degree of 4 is found best fitting the regression line.

Task2: Utilize the hospital data to calculate the point of no return for a state. Use percentage occupancy / utilization to see which states are close and what their trend looks like.

To calculate the point of no return I used columns from the merged dataset with beds and total beds used. So based on these values if the graph line crosses the total utilized bed, then there is no point of return. I tried to apply similar techniques by taking the mean of the values and plotting the graph against the new cases / deaths with the no. of days column. I selected two variables from the data frame with which I can compare. The reason covid and hospital is related because when the cases increase, and the beds are totally accommodated then there will be no point of return. So, I calculate the sum of the columns to perform this task and check if utilized bed cross the total beds then there is no point of return.

Later I compare the data with another state, I chose North Carolina to check the trends of the two states.

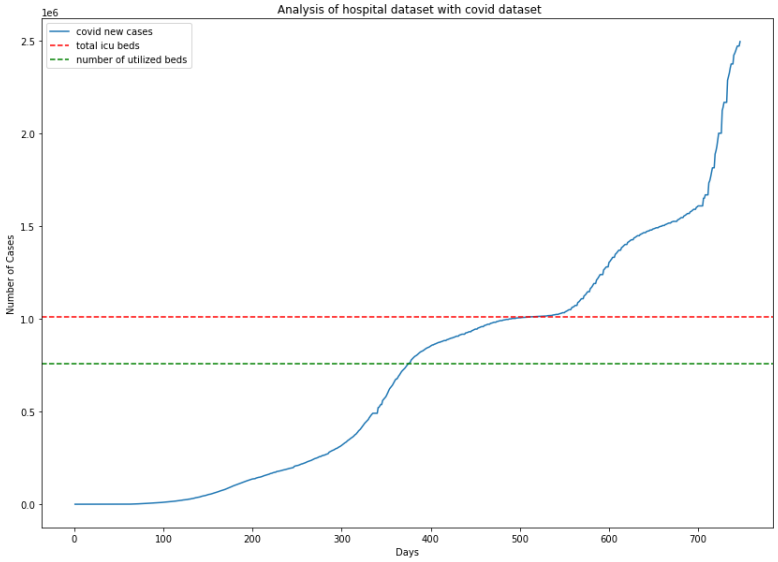
Inferences:

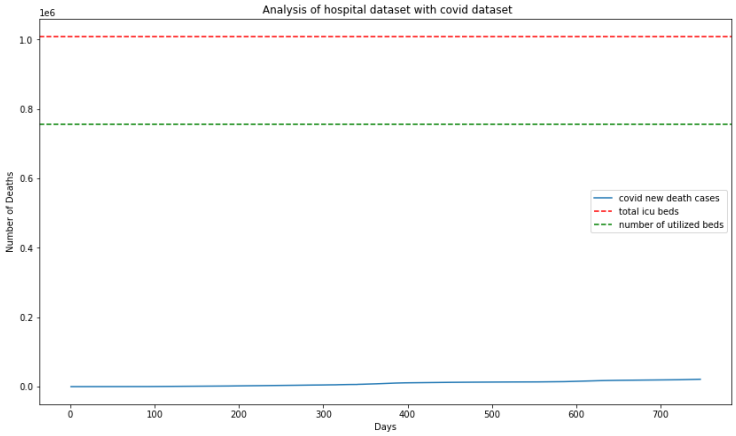
- In the graphs we observe that the total utilized beds have not crossed the total ICU beds which means that there is no point of return if it crosses beyond that line.

Task 3: Propose the Null hypothesis and calculate the chi squared value.

Null Hypothesis (H_0):

For this hypothesis I tried to find the mean of the variables of the hospital dataset i.e., total bed and the total beds used. By this we will be able to draw a hypothesis that if the beds utilized is more then the covid cases is rising and there is increase in deaths. To prove this hypothesis, we calculate the p-value and chi square to show the proposed null hypothesis is accepted or rejected. Below is the screenshot of cases and deaths with ICU beds used and ICU beds utilized.





Alternate Hypothesis (H_a):

Alternate Hypothesis will be the negation of the null hypothesis. Here we must prove that the increase in hospital beds utilized does not increase the covid cases and deaths

Inference:

My Hypothesis is that with the increase in the beds utilization the number of new cases and new deaths are increasing. To state the Null hypothesis and Alternate hypothesis I would define it as,

* Null Hypothesis (Ho): The beds utilization has no effect on increase in cases.

* Alternate Hypothesis (Ha): The bed utilization has an effect with the increase in cases.

I will try to to approve or reject a hypothesis by calculating the mean of the datapoints. As shown in the above plots we can observe the mean of the datapoints for both the variables and conclude that the proposed experiment is true. The mean for bed utilized is beyond the line of the mean of cases. We observe that the beds utilized is more than the mean of the cases which proves that the null hypothesis is correct. The two variables, beds and cases are increasing and dependent on each other.

Also, the dataset has a negative value which means that there were no beds available and the people who wanted a bed were added to the waitlist. Thus, in the graph the beds bin values are hidden as it goes negative. The p-value is 0.04213047 which is lesser than the threshold hence the null hypothesis is rejected.

Note: For the graphs and more explanation view the notebook.