Capstone Final Report – COVID-19 Cases in the US

I. Executive Summary

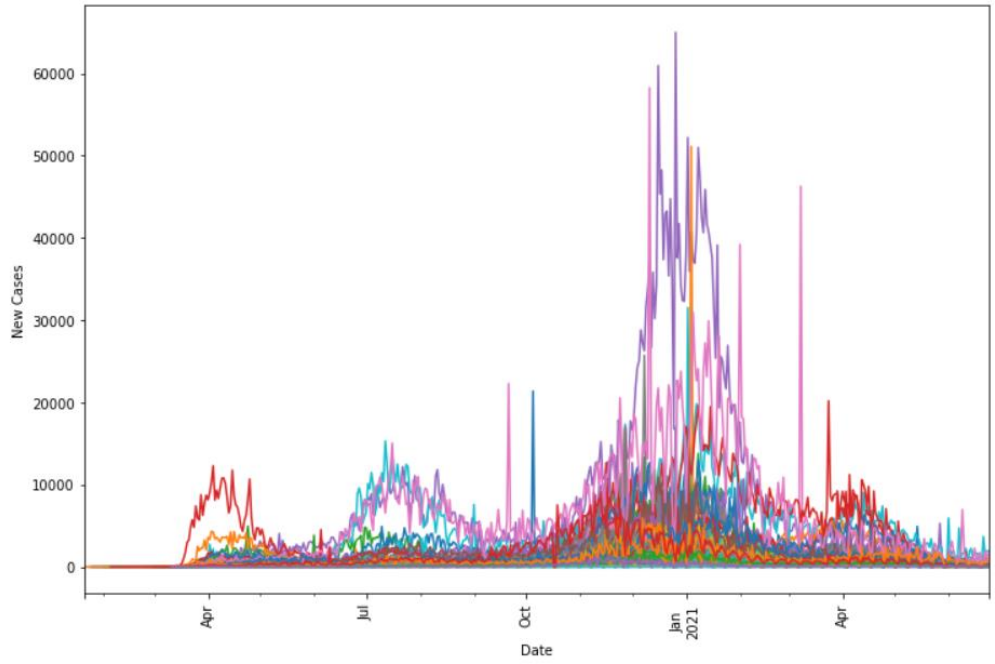
The surge of COVID-19 cases in the United States has caused widespread fear and uncertainty. Over the last 1.5 years, the pandemic has affected millions of lives and is still at the forefront of the nation's problems. To get ahead of the issue, it is extremely important to determine the main causes of the spread and analyze viable solutions for their efficacy. This project aims to focus on the effect that vaccinations have on the number of new cases in the US. Of course, there are many other factors that affect the number of cases, like mask mandates, lockdown policies, and more recently, Delta and other variants. But if we can show that vaccinations have some effect on the number of COVID-19 cases, we can determine that this is indeed a viable solution to stopping the pandemic.

II. Data

The datasets for this project were taken from Kaggle – one is a list of daily total cases per state, and the other is a list of daily vaccinations per state.[1] After some cleaning and wrangling, the two datasets were combined into one and null values interpolated. Two additional columns, number of new cases and population were calculated from the existing data.[2]

A quick EDA of our data shows the trend of new cases over the past year had a steep incline at the end of 2020, and then a slow decline early 2021.

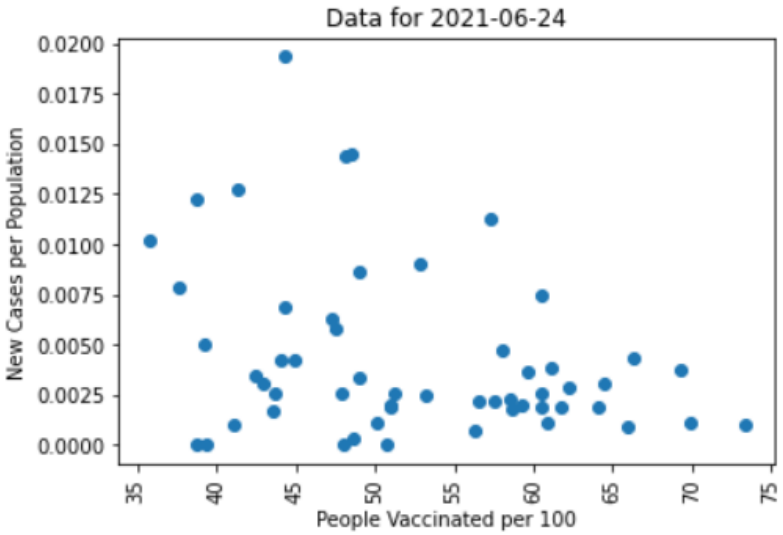
Figure 1 – Trend of Number of New Cases per State



This is roughly when vaccination data begin to emerge, so we will build a model that tries to show a correlation between the decline in new cases and the increase in vaccinations.

Because of the varying populations per state, we will normalize the number of cases and vaccinations by dividing by the population of the state. Exploring the data for one day, we plot people vaccinated per 100 against new cases per population:

Figure 2 – Scatterplot of People Vaccinated to New Cases for One Day



Although the scatterplot is quite random, there is an overall slight negative correlation between number of people vaccinated and new cases.

To prepare the DataFrame for modeling, we extract only 4 columns: date, state, people_vaccinated_per_hundred, and new_cases_per_hundred.[3] Now, we are ready to build a time series model using these features.

III. Model

This project takes a multivariate time series and converts it into a supervised learning problem to predict the number of new cases based on vaccinations, with a 7-day lead time. The data used is the daily number of new cases per state, as well as the daily total number of people vaccinated per state. This dataset is transformed into a supervised learning model with input features[4] of:

1. Number of new cases 7 days behind
2. 7-day rolling average of new cases 14 to 7 days behind
3. 3-day rolling average of new cases 10 to 7 days behind
4. Total number of people vaccinated 7 days behind

With the shifted data, we can remove the date and state identifiers from the model data since they have no effect on model performance.

We begin by splitting the data into training and testing sets, with the training set starting from the first datapoint until the 100th day of vaccination data, and the testing set as the remaining data. These features are then fit to various regression models[5] to find the one that results in the lowest MAE (mean absolute error).

IV. Results

We found that the Gradient Boosting Regressor model resulted in the lowest MAE when predicting the number of new cases at time t with the above features. We further decreased this value with hyperparameter tuning[6] to create our best model.

In order to evaluate performance, we created 2 baseline models, one dummy model that assumed new cases at time t is equal to new cases at time (t-7), and another ML model that had all the same parameters as our best GBR model except without the vaccination data[7]. Interestingly, the MAE for the machine learning model was larger than the MAE of the dummy model, causing us to go back and revisit our approach.

We realized that there is a significant difference between the training and testing set (most likely due to the absence of vaccination data prior to 2021, and the large fluctuation of new cases towards the end of 2020 and beginning of 2021), that we don't see in the testing data. For this reason, we decided to combine the training and test sets into one dataset and fit and predict on the same dataset.

In order to prevent overfitting, we use the same hyperparameters that were determined earlier in our best GBR model. The new best GBR model using the full dataset results in a MAE of 0.00609 cases/hundred. The feature importance of the number of people vaccinated is 1.9%. Although this percentage is low, it still has some significance when predicting the number of new cases.

Our first dummy model results in a MAE of 0.0067, meaning that our machine learning model reduced the error by approximately 9.1%.

Our second baseline model that excludes the people_vaccinated_per_hundred feature results in a MAE of 0.00644, which shows that vaccination data reduces our error by 5.4%.

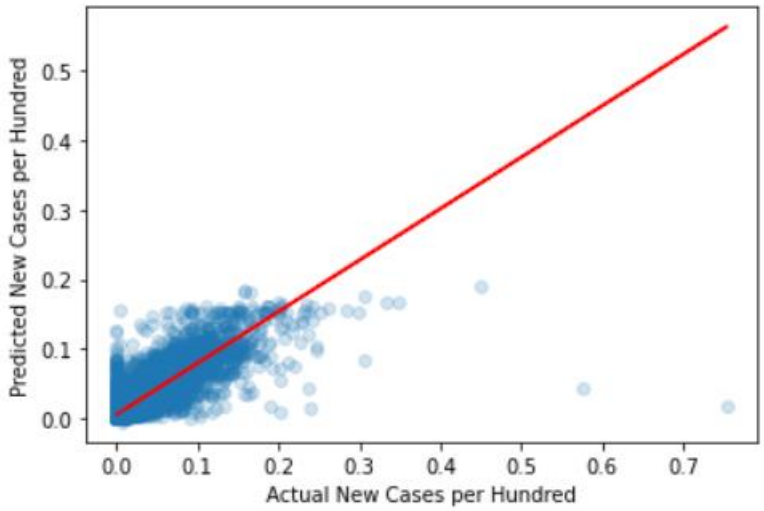
Table 1 – MAE of Baseline vs Best Models

	MAE	**% Reduction in MAE of Best Model compared to Baseline**
Baseline 1 - Dummy Model	0.0067	9.1%
Baseline 2 - GBR Model without Vaccination Data	0.00644	5.4%
Best GBR Model	0.00609	-

As a frame of reference, the average new cases per hundred in the test dataset is 0.0201, so error rates are relatively high with even our best model. This can be due to the fact that there are many other factors that affect the number of daily cases that have not been accounted for in the model.

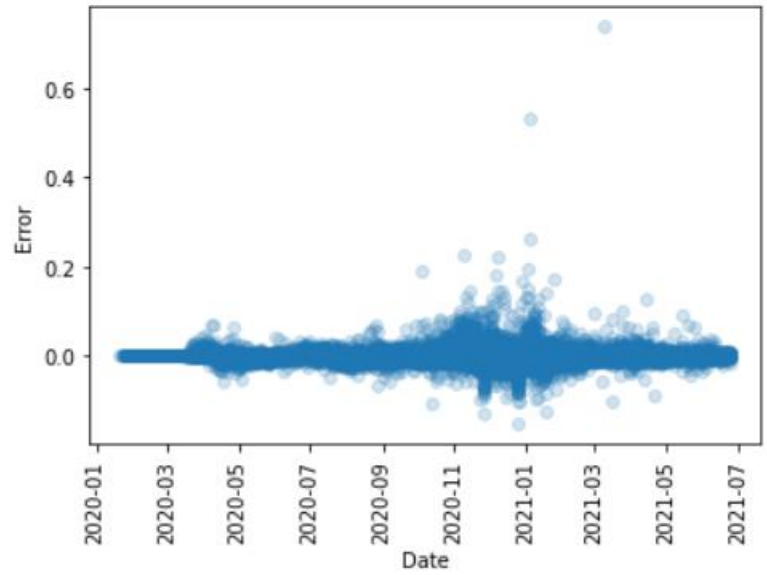
Plotting the predicted vs actual values of new cases using our model, we obtain a slope of 0.74, showing that the relationship between the actual and predicted values has a positive correlation.

Figure 3 – Actual to Predicted New Cases with Regression Line of GBR Model



When we plot the distribution of the error between actual and predicted values of new cases, we see that it is centered around 0, with some outliers on the positive x-axis.

Figure 4 – Distribution of Error (Actual – Predicted) of GBR Model



V. Conclusion

From our Gradient Boosting Regressor model, we see that there is in fact a correlation between the number of people vaccinated and the number of new cases in the United States. Comparing the two baseline models to our best performing model, we see a significant improvement when vaccination data is added. Although overall error values are high, this can be due to the various factors that determine the number of new cases that we did not account for in this project. Our findings conclude that vaccinations do have an effect on the number of Covid-19 cases in the United States.

VI. Appendix

1. Datasets from Kaggle:

   Coronavirus (Covid-19) Data of United States (USA):
   https://www.kaggle.com/joelhanson/coronavirus-covid19-data-in-the-united-states

   USA COVID-19 Vaccinations: https://www.kaggle.com/paultimothymooney/usa-covid19-vaccinations

2. The 2 new columns are calculated as follows:

   Number of new cases [new_cases] = total cases at time t minus total cases at time t-1. For any negative value, we set new cases of time t equal to new cases of time t-1.

   Population [population] = total number of people vaccinated on a particular day / people vaccinated per hundred on the same day * 100. This is done once for each state and added to a dictionary, then filled into the DataFrame based on state.

3. The four columns used in the modeling portion of the project are:

   Date: the date of the datapoint
   State: the state that the datapoint belongs to
   People_vaccinated_per_hundred: total number of people vaccinated in that state on that date divided by population multiplied by 100
   New_cases_per_hundred: the number of new cases that day in that state divided by population multiplied by 100

4. The engineered features that our final model takes in are:

   Number of new cases 7 days behind: shifted new_cases_per_hundred column by 7 days (t-7)
   7-day rolling average of new cases 14 to 7 days behind: average of new_cases_per_hundred column from (t-14) to (t-7)
   3-day rolling average of new cases 10 to 7 days behind: average of new_cases_per_hundred column from (t-10) to (t-7)
   Total number of people vaccinated 7 days behind: shifted people_vaccinated_per_hundred column by 7 days (t-7)

5. Regression models from sklearn that were tested in this project:
   - KNeighborsRegressor
   - DecisionTreeRegressor
   - ExtraTreeRegressor
   - SVR
   - AdaBoostRegressor
   - BaggingRegressor
   - RandomForestRegressor
   - GradientBoostingRegressor

6. The following hyperparameters were tested for the Gradient Boosting Regressor model, with the parameter in **bold** as the best performing:
   - Max_Depth = [**3**, 10, 1000]
   - Min_samples_leaf = [1, 50, **100**]
   - N_estimators = [**500**]

7. The 2 baseline models used as evaluation metrics are:

   1. Dummy model
      - Number of new cases at time t = number of new cases at time (t-7)
   2. GBR Model without vaccination data
      - Features used:
        - *Number of new cases 7 days behind*: shifted new_cases_per_hundred column by 7 days (t-7)
        - *7-day rolling average of new cases 14 to 7 days behind*: average of new_cases_per_hundred column from (t-14) to (t-7)
        - *3-day rolling average of new cases 10 to 7 days behind*: average of new_cases_per_hundred column from (t-10) to (t-7)