Blake Hernandez

Final Report:
Covid-19 New Case Prediction in CA

Problem Statement

Covid has had many impacts on society, but one that is near and dear to my heart is the impact it has had on hospitals, and how that impact has further affected many communities. My mother works at a hospital, and though her job is not one that directly interacts with patients, she has commented numerous times on how hard the pandemic has been on some of her coworkers. Many doctors and nurses find themselves working even more hours than they usually do. This exhaustion has led to burn out and lower quality of life for hospital employees.

In addition to the exhaustion hospital workers have faced, the hospitals themselves have at times had difficulty meeting the needs of their communities in terms of available hospital beds. As covid patients fill beds and take up hospital staff time, other patients with serious, but non life threatening conditions or injuries have been unable to get the treatment they need. On multiple occasions throughout the pandemic the hospital network my mother works at has put a freeze on all non-essential medical operations.

The goal of this project is to use covid-19 data to create a time series model capable of predicting covid infections a full month in advance in the state of California. This can then be used by hospitals to allow them to prepare for changes in patient demand. These preparations can be in the form of scheduling, allowing workers to get a much needed rest during slumps and giving them forewarning when things are about to ramp up, and in the form of setting up and staffing overflow beds when they are needed.

Data Wrangling

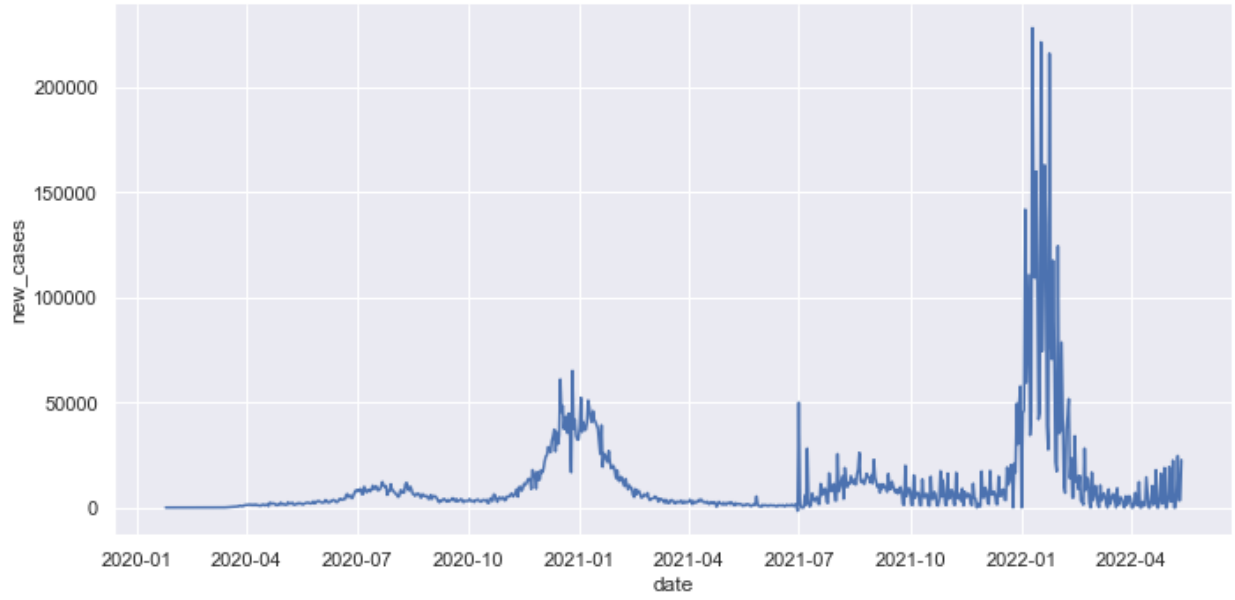
The primary data source for this project comes from the New York Times, which has been collecting daily covid data on a county basis across the United States. The data includes information on total daily cases and daily deaths beginning in 2020-01-21 to the most current date at the time of model creation, 2022-05-13. The data also contains fields for county name, county id, and state name.

There were few missing values in this data, the only ones present were for county id and daily death count. The missing values for deaths all came from the territory of Puerto Rico, and occurred before the first covid related death occurred in that area. As for the missing county id values, they all came from three specific counties, none of which are in California, so there was no action to take.

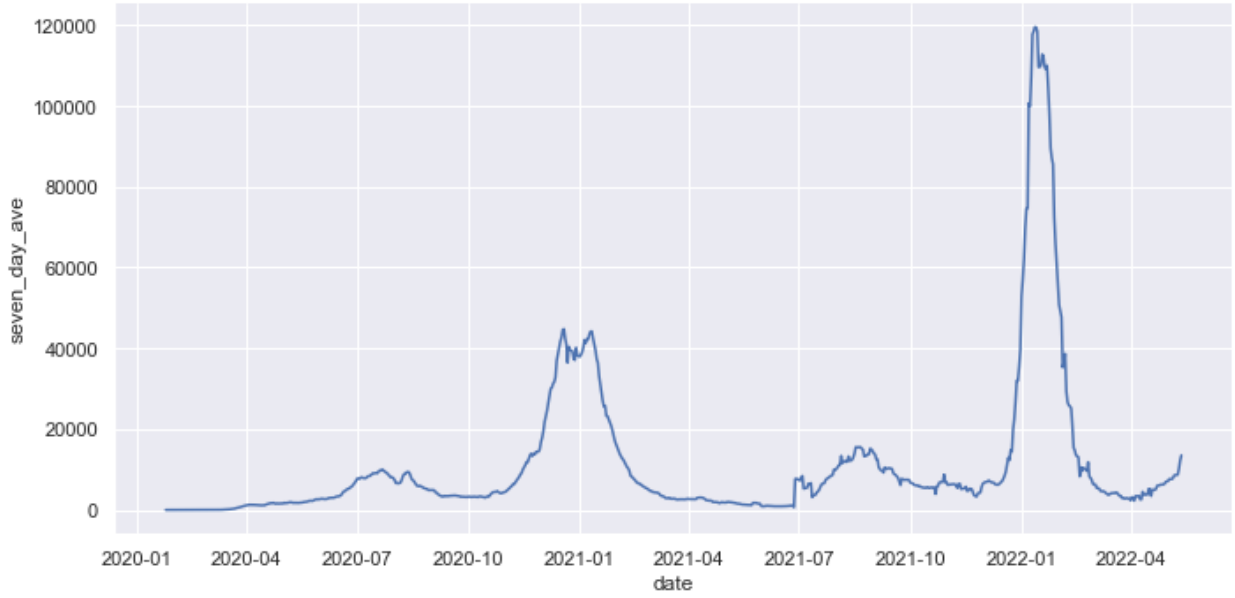
After looking into missing values, I grouped the data by state. Aggregating the county values for cases and deaths, the data was left with a single daily observation for each state. I then isolated the data for California.

At this point I created a few new features to be used for exploratory analysis and modeling. By taking the difference between the daily total cases data I created a new_cases feature for each day, and expanded upon that by creating a feature for daily percentage change in new cases. Finally I smoothed the daily new cases data using a rolling window to calculate the seven day average for each day, removing some of the noise from the data. This seven day average data is what was ultimately used as the target variable.

Daily New Cases in CA



Smoothed seven day average of daily cases



As part of the EDA for this project I also wanted to confirm/back up the claim that covid has had a real impact on hospital occupancy. For this I used government hospital data that included the total average inpatient beds filled by patients in a given hospital, collected on a weekly basis. This data was grouped and aggregated by state similar to how I had previously grouped and aggregated the covid data from the NY Times. For this comparison, I had to resample the covid data on a weekly basis.

EDA

With the covid and hospital data ready to be explored, I decided to use a simple linear regression to confirm that covid infections have been having an impact on hospitalizations. The results are as follows.

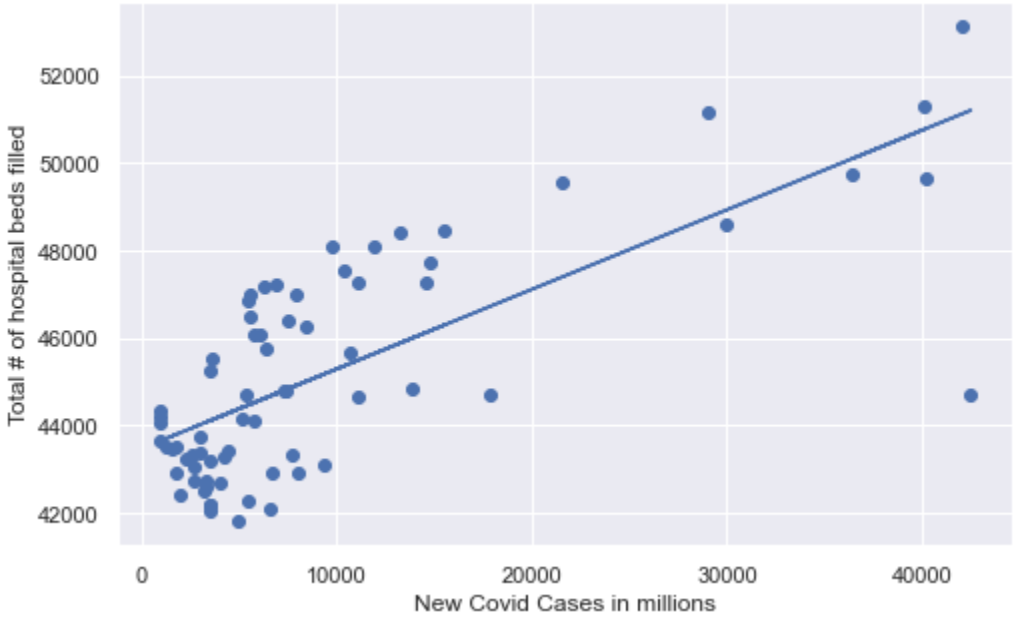
$$Yi = 43{,}470 + 0.1819*Xi$$

OLS Regression Results

Dep. Variable:	total_occupancy	R-squared:	0.539
Model:	OLS	Adj. R-squared:	0.532
Method:	Least Squares	F-statistic:	79.40
Date:	Fri, 20 May 2022	Prob (F-statistic):	4.90e-13
Time:	12:46:00	Log-Likelihood:	-621.45
No. Observations:	70	AIC:	1247.
Df Residuals:	68	BIC:	1251.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
new_cases	0.1819	0.020	8.910	0.000	0.141	0.223
const	4.347e+04	285.090	152.466	0.000	4.29e+04	4.4e+04

Omnibus:	6.231	Durbin-Watson:	0.886
Prob(Omnibus):	0.044	Jarque-Bera (JB):	5.781
Skew:	-0.505	Prob(JB):	0.0556
Kurtosis:	3.981	Cond. No.	1.89e+04



The results indicate that new covid infections do have a significant impact on hospital occupancy. To be specific, on average for every 0.1819 increase to the weekly average in new covid case, the average number of hospital beds occupied in that week will increase by one. This roughly means that a 5.5 increase in the weekly average of new covid cases leads to an additional hospital bed filled. These results confirm what I already knew from anecdotal evidence.

The remainder of the EDA was focussed on preparing the data for an ARIMA time series model. This involved making the data stationary by finding the appropriate amount of differencing. This was done using the Dicky-Fuller test and plots of the differenced data and its autocorrelation function. It quickly became clear that a differencing of one was appropriate for making the data stationary.

Model Selection

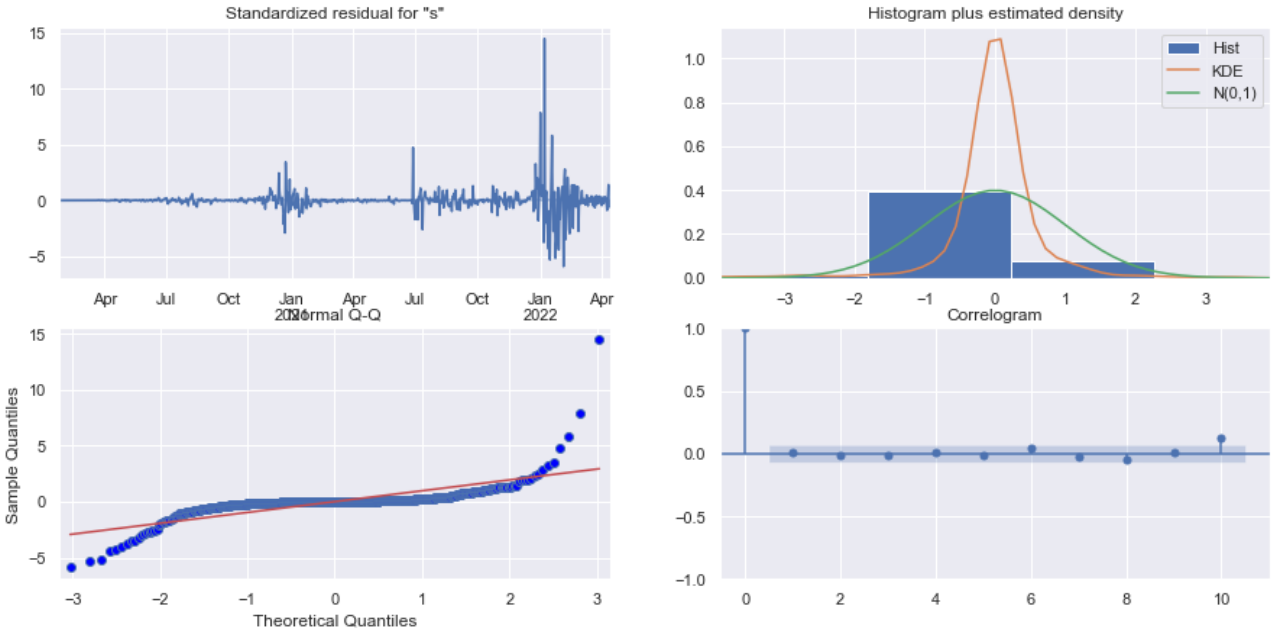
For the initial modeling I used an ARIMA model, which is a combination of an auto regressive and moving average model.

$$Y_t = (\beta_1 + \beta_2) + (\Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p}) + (\omega_1 \varepsilon_{t-1} + \ldots + \omega_q \varepsilon_{t-q} + \varepsilon_t)$$

Image Source: https://towardsdatascience.com/arima-simplified-b63315f27cbc

The auto regressive portion of the model uses previous values of the dependent variable to determine its future value. The moving average portion uses the errors of previous predictions to make an estimate for the future period.

I first attempted a simple ARIMA model, using the ACF (autocorrelation function) and PACF (partial autocorrelation function) plots to select AR and MA terms for the model. This model performed well on the training set, but did not do well with the testing data. While evaluating this model I noticed that while the model's residuals did appear to be uncorrelated, they were not normally distributed. This was a problem since a good ARIMA model's residuals should resemble white noise, meaning they are uncorrelated and normally distributed. Additionally this model seemed to overfit the training data.



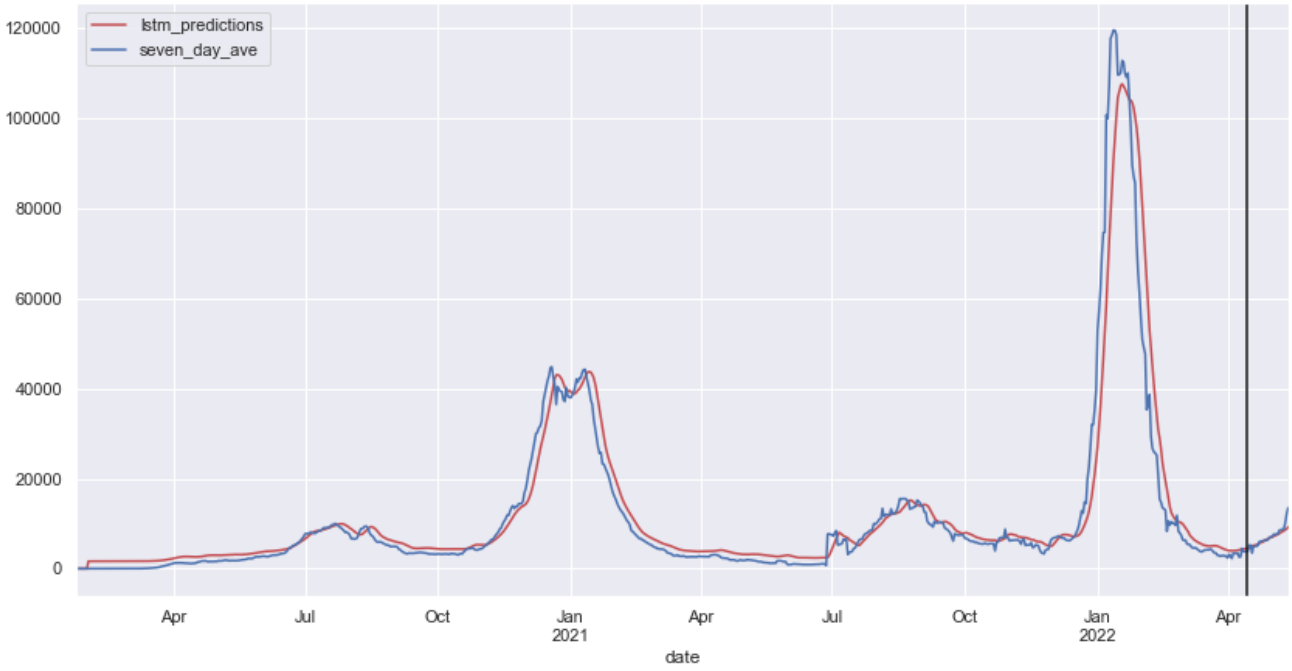
Here you can see from the upper right histogram and the lower left Q-Q plot that the residuals are not normally distributed.

To improve upon the first model I switched over to an auto ARIMA model, which would do a better job selecting hyper parameters than I did using the ACF/PACF plots, by doing a grid search. This model selected different parameters than I did for my initial ARIMA, and yielded better results, capturing the upward trend in the testing data much better than the initial model. However, this model still suffered from some of the same issues that the first model did. The residuals were not normally distributed, and the model seemed to be overfitting.

The model ultimately selected was an LSTM (Long Short Term Memory) model, a popular type of recurrent neural network that allows for previous outputs to be considered as inputs. This lets the model pick up on long term dependencies in the data. For the LSTM model I created a function, using tensor flow, that takes in a list of parameters and train/test data as inputs, then scales the data, trains, tests, scores, and returns the model along with relevant metrics. I then performed a gridsearch by running this function in a loop with different combinations of parameters, saving the results in a dataframe.

The LSTM model performed the best by far, scoring higher than either ARIMA models in all metrics for the test set. One thing I observed is that the LSTM did not outperform the ARIMA models when predicting the training data, indicating that the LSTM did not suffer from the same overfitting, and was a better generalization of the data.

Plot of LSTM predictions against true seven day average values



Model Metrics

Model	MAPE	MRSE	R-Squared
ARIMA1	0.2964	3135.65	-0.9433
ARIMA2	0.3673	2885.83	-0.646
LSTM	0.0885	1126.78	0.749

Takeaways

In the future it would be nice to directly connect hospital data and covid infections using time series, rather than just doing a univariate analysis using covid data. It could

also be worthwhile to explore other industries that might find this type of information useful. Some ideas that came to mind would be public transit, retail stores, airlines, and even home entertainment/streaming services. All of these industries could potentially benefit from having a future notice about when covid is likely to ramp up or down. For example, retailers could use the information when scheduling employees or purchasing inventory, since spikes in covid will likely result in less foot traffic. Similarly public transit and airliners could use this data to adjust their expected demand for transportation. Lastly, home entertainment services, such as streaming platforms, could use this information to adjust their marketing, and maybe even run promotions when they know that a large portion of the population will be housebound.