COVID-19
New Cases Prediction for San Diego County

Yousif Jabbo, Steven Nguyen, Jonathan Lee
Computer Science Department
San Diego State University

Introduce Problem

- COVID-19 has been a mysterious disease that has affected the entire world since early 2020.
- People can not only be asymptomatic but can also suffer brain damage, lung complications, loss of taste, loss of smell, or a combination of these side effects.
- People of San Diego do not know the future behavior of COVID-19 in the form of more infections or less infections.

Background

- Limitations of our data: since we only used the number of new cases to train the model, the model could not see the number of vaccination and how that impacts the number of cases. To prevent skewing in our results, we had to remove the volatile data of 2020.
- Limitations of Linear Regression: We found that using Linear Regression can help us understand the current trend of the data but since it is a single line it keeps moving in a single direction. It does not take non-linearity into consideration. Thus, the Linear Regression Model was very limited to predicting the data as it was only able to predict the next 1-2 weeks accurately. Since the Linear Regression Model forms a line, if we gave this model the full data, the model ended up with outliers, which will skew the results. Also, we had to represent weeks by numbers instead of dates.
- Limitations of ARIMA: The ARIMA model can take the non-linearity into consideration however, it thinks that what happened last time will happen again.

Problem Motivation

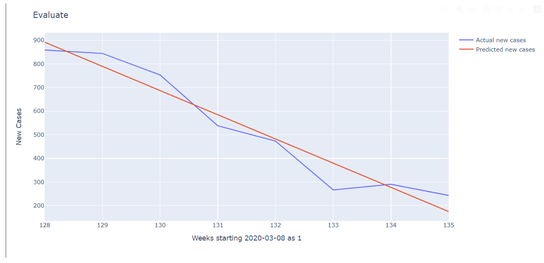
- Without a vaccine that provides 100% protection against COVID-19, the citizens of San Diego should be provided with an estimation as to how many future COVID-19 cases there will be in San Diego. This well help the public make decisions on when to wear masks, avoid large social gatherings, or receive an additional booster shot when available.

- Having an accurate prediction of COVID-19 positive cases for the coming weeks will greatly help businesses, health product manufacturers, and hospitals be well prepared and not be overwhelmed by any sudden increases in positive cases.

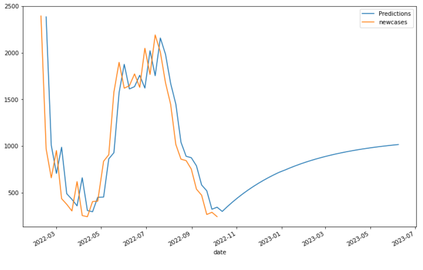
Approach

- To attempt to predict future positive COVID-19 cases in San Diego County, our team initially utilized the **Linear Regression Model**. However, we soon realized that Linear Regression did not produce an accurate prediction (discussed more in *Background*). We needed to find a different model that would consider trends of seasons as well.
- Upon recommendation from Dr. Homayouni, we then switched and implemented the **Auto-Regressive Integrated Moving Average (ARIMA) Model**. This model, which does account for seasons, produced an accurate prediction of the training data as well as providing predictive values that is consistent with the trends of the positive cases of the same month of previous years.

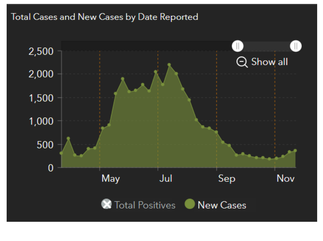
Results



- Using the Linear Regression model the equation of the line is **y = 14008.87 + -102.48 * week**
- **RMSE** is: 59.58
- **Week after 10/5/22** is : 72.10



Using the ARIMA model
RMSE is: 36.39
Week after 10/5/22 is: 351



Since we started this project in October, we had data for the first week only. Now that some weeks have passed the real data show that the next week's true data value is 332, so, our ARIMA model is only off by 19 cases which are within our expected error rate. The 3rd week's prediction using ARIMA is 399 and the real data is 354.

Results Evaluation

- The SARIMAX (Seasonal ARIMA Model) is more optimal than the usage of Linear Regression when attempting to predict the future number of COVID-19 cases within San Diego County.
- Linear Regression faces the limitation such that once the slope is defined to be positive, predictions will always increase, and when the slope is negative, predictions will always decrease. This becomes an issue as the Linear Regression Model may eventually try to predict a negative number of cases when the slope is negative when trying to predict far into the future.
- The SARIMAX Model is more accurate than the Linear Regression Model as the SARIMAX Model captures the seasonality of the data, allowing it to fluctuate up or down in the prediction values instead of strictly increasing or decreasing. This can be seen when comparing the RMSE (Root Mean Squared Error) metric for the Linear Regression Model and the SARIMAX Model.
- The value of our RMSE represents how much our predicted value of future COVID-19 cases are off by. For example, if our RMSE is 60, our predicted value of future COVID-19 cases is off by +/- 60 cases.

Conclusions

- The Linear Regression Model is limited when it comes to non-linear data predictions. The ARIMA Model is great for time series data prediction and non-linear data predictions.

- COVID-19 may slowly increase as the year 2022 ends and the year 2023 begins due to colder weather, the normal flu season, and the large social gatherings people may have for Thanksgiving, Christmas, and New Year's Eve.

References

[1] Alabdulrazzaq et al., 2021, "On the accuracy of Arima based prediction of COVID-19 spread", *Science Direct*, https://www.sciencedirect.com/science/article/pii/S2211379721006197

[2] Chyon et al., 2022, "Time Series Analysis and predicting COVID-10 affected patients by ARIMA model using machine learning", *National Center for Biotechnology Information*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8669956/

[3] Rath et al., 2020, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model", *National Center for Biotechnology Information*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7395225/