Path To Herd Immunity
Time Series Analysis-COVID19

Team Members and Contributions To Tasks

1. Jayarani Emekar -
   a. Major work contribution: ARIMA and LSTM Model ,Hot Exponential smoothing
   b. Minor work contribution: Prophet Model, Statical data prediction
2. Monika Gadage -
   a. Majorly work contribution: Prophet Model, Statical data prediction

    b. Minor work contribution: ARIMA and LSTM Model
3. Viraj Sonawane-
    a. Majority work contribution: Prophet Model, Plotting the Prediction,EDA analysis
    b. Minor work contribution: Data exploration

Major Goal

Many wonder when the World will be “normal” again after the pandemic. Resuming safe travel and trade between countries is imperative to improve the global economy. This project aims to calculate the time to reach herd immunity for countries and the world. How the COVID-19 vaccine can help with international travel and trade, that will return a somewhat familiar feeling of normalcy.

Below are the questions we are addressing:

1. Exploratory data analysis to find out vaccination prediction for each country
2. Statical Analysis of data
3. Time series analysis using Prophet, ARIMA, LSTM model
4. Comparing which model fits better to predict herd immunity?
5. When will the world hit herd immunity(Prediction date), making unencumbered global trade and travel a reality?

Tasks

1. Data Exploration
2. Data Cleaning
3. Exploratory Data Analysis
4. Time series analysis using prophet on country basis
5. Statistical Analysis prediction
6. Time series analysis using ARIMA and LSTM for world population
7. Compare different time series model and predict which one fits better

Methods

1. **Augmented Dickey Fuller test (ADF Test)**
    a. Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not.
    b. In ARIMA time series forecasting, the first step is to determine the number of differencing required to make the series stationary.
    c. There is a hypothesis testing involved with a null and alternate hypothesis and as a result a test statistic is computed and p-values get reported.
    d. It is from the test statistic and the p-value, you can make an inference as to whether a given series is stationary or not.
    e. This P-value calculated from test add significance in projecting the future time series analysis
2. **Exploratory Time Series Decomposition**
    a. Time series decomposition involves thinking of a series as a combination of level, trend, seasonality, and noise components.
    b. Decomposition provides a useful abstract model for thinking about time series generally and for better understanding problems during time series analysis and forecasting.

3. **Auto-Correlation Function**
    a. Autocorrelation refers to how correlated a time series is with its past values whereas the ACF is the plot used to see the correlation between the points, up to and including the lag unit.
    b. In ACF, the correlation coefficient is in the x-axis whereas the number of lags is shown in the y-axis.
4. **Partial autocorrelation Function**
    a. A partial autocorrelation is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.
    b. The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags.
5. **Hot Exponential smoothing**
    a. The Exponential Smoothing (ES) technique forecasts the next value using a weighted average of all previous values where the weights decay exponentially from the most recent to the oldest historical value.
    b. When you use ES, you are making the crucial assumption that recent values of the time series are much more important to you than older values. The ES technique has two big shortcomings: It cannot be used when your data exhibits a trend and/or seasonal variations.
    c. The Holt ES technique fixes one of the two shortcomings of the simple ES technique.
    d. Holt ES can be used to forecast time series data that has a trend. But Holt ES fails in the presence of seasonal variations in the time series.
    e. This method looks more accurate but the drawback of this method is we can calculate the forecast upto 90days ahead

Model Selection

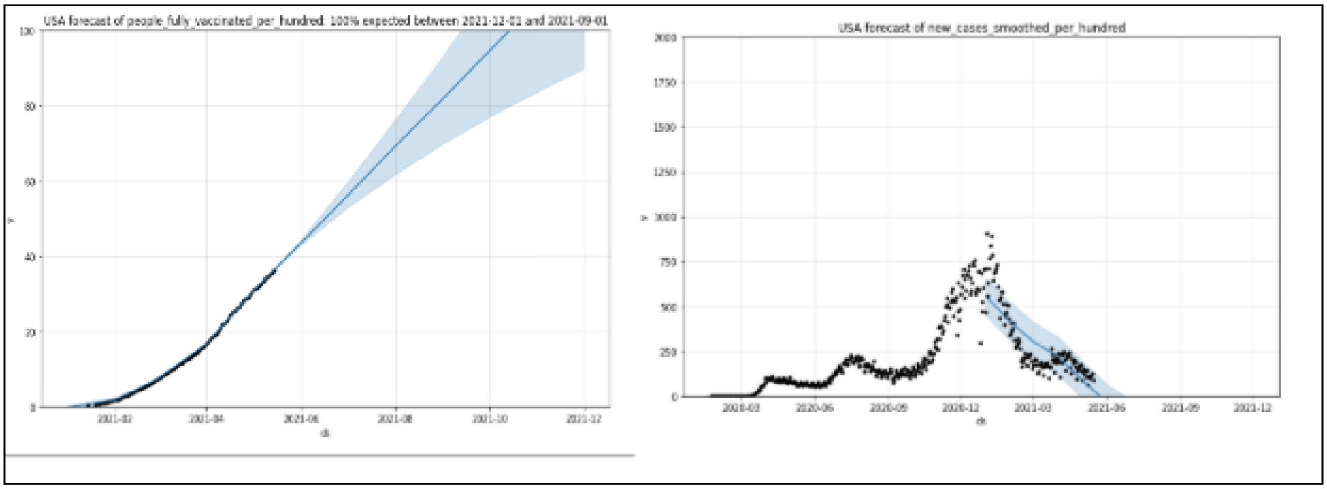
1. **Prophet**
    a. It is a procedure for forecasting time series data based on an additive model where nonlinear trends are fit with yearly, weekly and daily seasonality, plus holiday effects.
    b. Prophet is robust to missing data and shifts in the trend and typically handles outliers well.
    c. Prophet requires time series data to have a minimum of two columns—the timestamp and the values. With just a few lines
    d. We used this model to predict vaccination vs infection rate for various countries in the world
2. **ARIMA**
    a. ARIMA (Auto Regressive Integrated Moving Average) can be defined in below terms
    p — the number of autoregressive
    d — degree of differencing
    q — the number of moving average terms
    m — refers to the number of periods in each season
    (P, D, Q)— represents the (p,d,q) for the seasonal part of the time series
    b. In ARIMA analysis we used different types of ARIMA models like statistical, Grid Search, SARIMAX,ARIMAX.
    c. We found around same Root Mean Squared Error Error and Mean absolute Percentage Error and which is very low, Hence ARIMA looks best fit to find Herd Immunity
3. **LSTM**
    a. It is Long Short Term Memory networks It is a special kind of recurrent neural network that is capable of learning long term dependencies in data.
    b. This is achieved because the recurring module of the model has a combination of four layers interacting with each other.
    c. In our analysis we used SequentialLSTM model, which is having highRoot Mean Squared Error Error and Mean absolute Percentage Error, so this doesn't seems to be a good choice for this data

Project Outcomes

1. Prophet Model Predictions

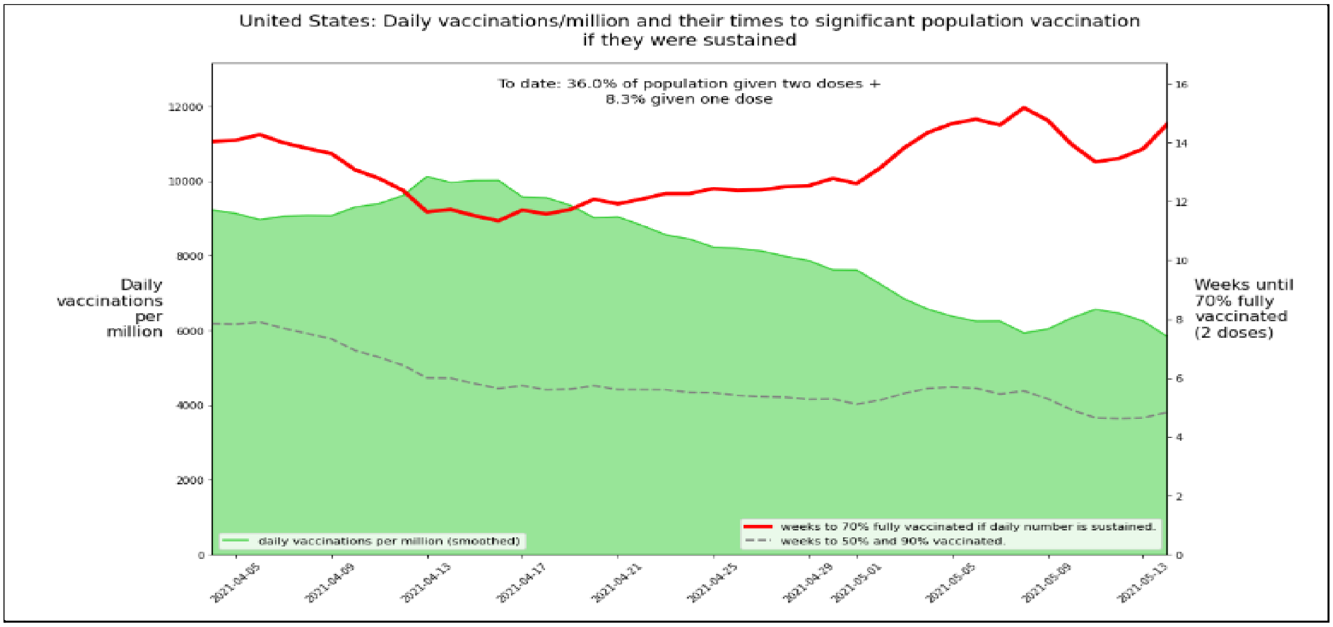
We used a prophet model to predict the vaccination rate for each country and plotted it.

Below is the prediction for USA with infection rate vs vaccination rate, we can clearly see that infection rate decreased as vaccination started increasing



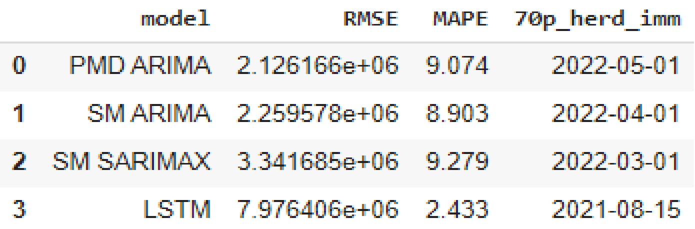
2. Statistical Prediction

We also find out the prediction for the USA using statistical calculation, values determined by prophet model and statistical calculation matches. For USA both prediction determine USA will achieve herd immunity by Jul 2021

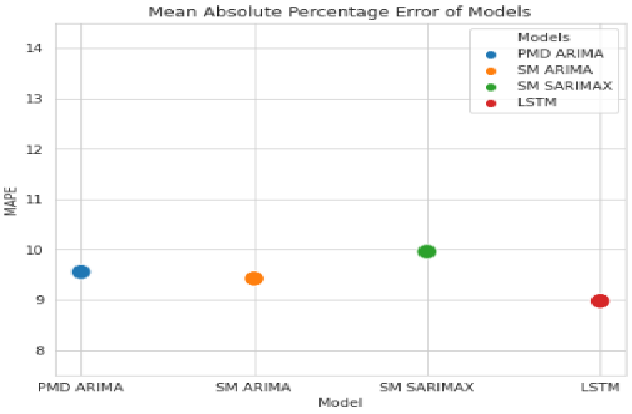


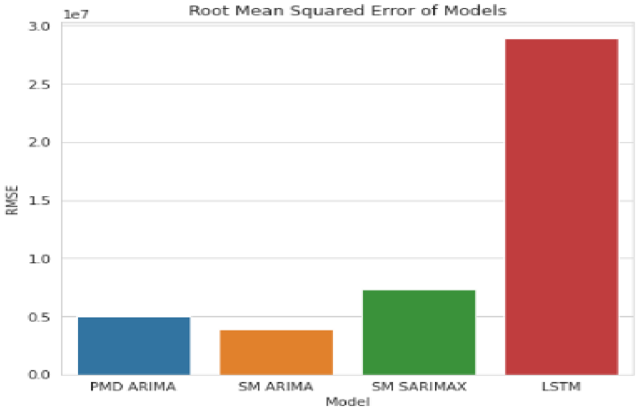
3. ARIMA and LSTM Herd Immunity dates for world population

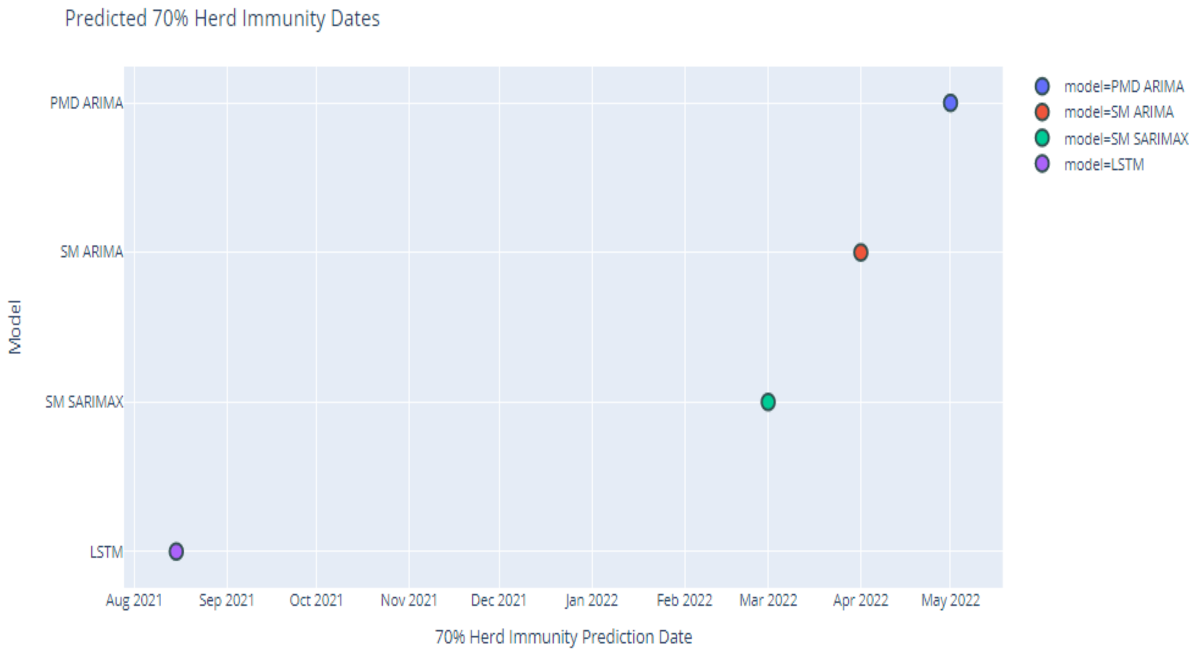
Below table and graph shows Mean Absolute Percentage Error(MAPE) and Root Mean Squared Error(RMSE) for models
All three of the ARIMA models were similar when looking at the mean absolute percentage error, while the statsmodels model was slightly higher.



	model	RMSE	MAPE	70p_herd_imm
0	PMD ARIMA	2.126166e+06	9.074	2022-05-01
1	SM ARIMA	2.259578e+06	8.903	2022-04-01
2	SM SARIMAX	3.341685e+06	9.279	2022-03-01
3	LSTM	7.976406e+06	2.433	2021-08-15







Challenges

1. Usage of dynamic COVID-19 data which is updated on daily basis in OurWorldInData github repository
2. As Dynamic COVID-19 data is used for all the predictions, sometimes there is a chance of receiving NaN values which can cause issues in predictions. This can be overcomed by using a CSV file containing data till last month.
3. Making unstatinary Covid 19 data to stationary format to predict the accuracy.

Conclusion

1. All three of the ARIMA models were similar when looking at the mean absolute percentage error, while the statsmodels model was slightly higher.
2. Looking at those points, the two statsmodels models were able to get closer to the actual data (lower RMSE).
3. When looking at the RMSE and MAPE, either of the ARIMA models would be a reasonable choice. rather than LSTM model
4. When you look at the forecasted herd immunity, it appears that the statsmodels models might potentially be more accurate.
5. It seems more likely that it would take approximately a year, rather than a few months or 4 years.
6. After doing statistical analysis of data and predicting through the Prophet model we got almost the same results. For many countries we got plus minus the same number of weeks for 70% herd Immunity.

Future Steps

1. In the LSTM model RMSE and MAPE values are very high, We need to work on the tuning of this model.
2. There are different types of LTSM model like Stacked,Bidirectional, Multivariate CNN to reach the minimum RMSE and MAPE value
3. Prophet Model : Prediction was done based on Vaccination rate data and hence Herd immunity predictions were achieved. Few countries are still facing uses after vaccinating 70% of the population hence type of vaccine plays a major role.
4. Data can be gathered related to different vaccine types used worldwide and predict the effectiveness of that vaccine which can be used to predict Herd Immunity graph.

Project Ethics

- Li mentions how important humanity is for technical people like us, especially in healthcare. We have this tendency to look at things through a technical lens. We approach things like scientists especially in hard science like computer science and often forget to reflect on things which a specialist from that domain might think about.
- People working in healthcare doctors, nurses are actually dealing with lives and healthcare vulnerabilities. It is very important to look at a project related to healthcare from their perspective as well. Going first hand and experiencing all behind the scene conversations and visiting patients' rooms and interacting with them is very important.
- This works as a way to humanise what we as scientists want to achieve. Hence it is always beneficial to collaborate with someone who is from healthcare if you're workin on healthcare projects or someone who gets to experience the issue first hand.

- We as scientists think about performance and availability in a project but doctors think about the patients and the lives involved. This is where there is a major difference of thinking which will benefit the project in a big way.
- Covid-19 emerged in China and spread extremely quickly along the modern-day Silk Roads: intercontinental flight paths. International lockdown and the effective suspension of civic and commercial activity across entire countries has thrust up a mirror on how our economic, social and political systems operate have been affected since last year.
- The COVID-19 outbreak has affected all segments of the population and is particularly detrimental to members of those social groups in the most vulnerable situations, continues to affect populations, including people living in poverty situations, older persons, persons with disabilities, youth, and indigenous peoples.
- Everyone has experienced Covid 19 first hand and there is a major part of humanity involved in this project. The project predicts when the world would be Covid-19 safe and ready to roll back in action.

References

1. Data
    a. Github link : https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations
    b. Website : https://ourworldindata.org/
2. Prophet Model
    a. Tutorial: https://www.tutorialspoint.com/time_series/time_series_prophet_model.htm
    b. Guide: https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-prophet-in-python-3
3. ARIMA Model
    a. Video Tutorial: https://www.youtube.com/watch?v=e8Yw4alG16Q
    b. Guide : https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
4. LSTM Model
    a. Video Tutorial: https://www.youtube.com/watch?v=xaIA83x5Icg
    b. Guide: https://www.tensorflow.org/tutorials/structured_data/time_series
5. Other References
    a. Exponential Smoothing: https://towardsdatascience.com/holt-winters-exponential-smoothing-d703072c0572