“COVID”: FORECASTING CASES & DEATHS

Dr. Jongwook Woo jwoo5@exchange.calstatela.edu
Mr. Jay Joshi jjoshi6@calstatela.edu
Miss. Shradha Shinde sshinde6@calstatela.edu
Miss. Sowmya Mareedu smareed@calstatela.edu
Department of Information Systems, California State University, Los Angeles

Abstract: Corona is a deadly disease that started spreading throughout the world starting in December 2019. It was first diagnosed in Wuhan, China. Since then, it has spread worldwide and has affected 2,475,440 people and 170,069 people have died due to this infection. Scientists have since then dedicated their time to the research of this disease and here we attempt to take a deeper look into this problem. In this paper, we have tried to forecast the number of deaths and cases in the major cities of United States like Los Angeles, and New York using Times Series, ARIMA, Linear Regression, and Depth Forest regression algorithm in Azure ML, Oracle CLI, and Databricks

Keywords: COVID-19, Time Series algorithm, ARIMA, Linear Regression, Azure ML, Oracle CLI, Databricks

1. INTRODUCTION

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illnesses The best way to prevent and slow down transmission is to be well informed about the COVID-19 virus, the disease it causes, and how it spreads by protecting yourself and others from infection by washing your hands or using an alcohol-based rub frequently and not touching your face. It was first diagnosed in Wuhan, China. Since then, it has spread worldwide and has affected 2,475,440 people and 170,069 people have died due to this infection. Scientists have since then dedicated their time to the research of this disease and here we attempt to take a deeper look into this problem. The dataset that used is not big in gigabytes but still provides results with efficient standing. Our dataset is of size 82 MB of CSV file format. It has 13 columns with a total 8 files having 540 rows in each.

2. DATA DESCRIPTION

Data is in CSV format and updated daily. It is sourced from this upstream repository maintained by the amazing team at Johns Hopkins University Center for Systems Science and Engineering (CSSE) who have been doing a great public service from an early point by collating data from around the world. We have cleaned and normalized that data, for example tidying dates and consolidating several files into normalized time series. We have also added some metadata such as column descriptions and data packaged it.

The collected dataset refers to the cumulative confirmed cases and deaths of COVID-19 that occurred in major cities of the United States from April 24, 2020, to May 4, 2020. This dataset includes time-series data tracking the number of people affected by COVID-19 worldwide, including:

- Confirmed tested cases of Coronavirus infection.
- The number of people who have reportedly died while sick with Coronavirus.
- The number of people who have reportedly recovered from it.

3. HARDWARE SPECIFICATIONS

For this project, we have used Microsoft Azure Machine Learning Studio and Databricks community edition to implement Spark ML. We have also used the Hadoop spark cluster on the Oracle CLI platform for the rating prediction. The specification is given below:

Azure	Databricks
• Memory – 10 GB • Nodes – 1 • Free Workspace • R Language	• Memory - 6 GB • Nodes- 1 • Driver (0.88 cores, 1 DBU),

Table 1. Hardware specifications

4. METHODOLOGY

The first step to getting the project started was to understand the approach we want to adopt to this project. Since a lot of papers are not available on this topic, we have researched papers that have predictions on other viral diseases like Ebola and the Zika virus. We saw how different factors affect different diseases. So, it becomes important to first understand all the features clearly as well as to understand the relationship between these features to select our target and descriptive features to get the maximum accuracy. At the end of this, we will also show the reasoning and analysis in the final part of our paper. In our project we have, implemented several algorithms and models in Azure ML, Databricks, and Oracle CLI.

4.1 Time Series Forecasting

Time series data is an important source for information and strategy used in various businesses. Time series forecasting is the machine learning modeling for Time Series data (years, days, hours…etc.) for predicting future values using Time Series modeling. We have used python Facebook prophet in Databricks for prediction of the number of cases and deaths in Los Angeles and New York. Databricks provides Databricks Runtime for Machine Learning a ready-to-go environment for machine learning and data science. The following figure depicts the use of time series forecasting in Azure Machine Learning.

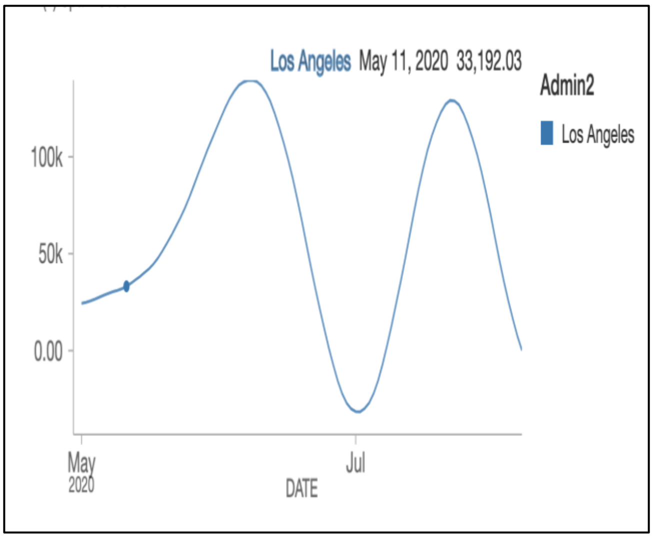


Fig 1: Time Series Forcast for Cases in Los Angeles Dated May 16, 2020

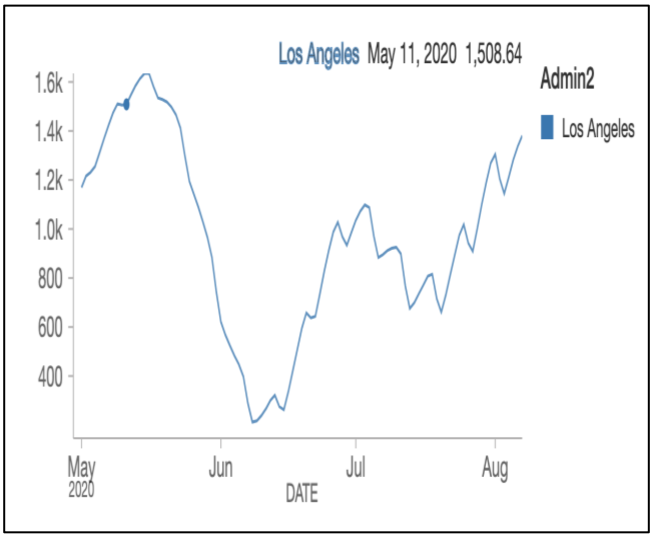


Fig 2: Time Series Forcast for deaths in Los Angeles Dated May 16, 2020

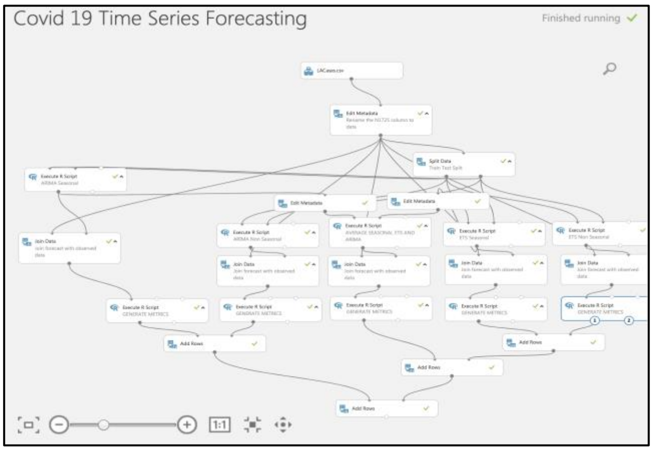


Fig 3: COVID 19 Time Series Forecasting in Azure ML.

4.2 ARIMA

Using the ARIMA model, you can forecast a time series using the series past values as we tried forecasting the number of cases and deaths. In this project, we build an optimal ARIMA model and extend it to Seasonal ARIMA and Non-Seasonal ARIMA. ARIMA, short for 'Auto-Regressive Integrated Moving Average' is a class of models that 'explains' a given time series based on its own pastues, that is, its lags and the lagged forecast errors, so that equation can be used to forecast future values. In this project, we have implemented ARIMA (AutoAuto-Regressiveegrated Moving Average) Seasonal & Nonseasonal, ETS using an R programming language.

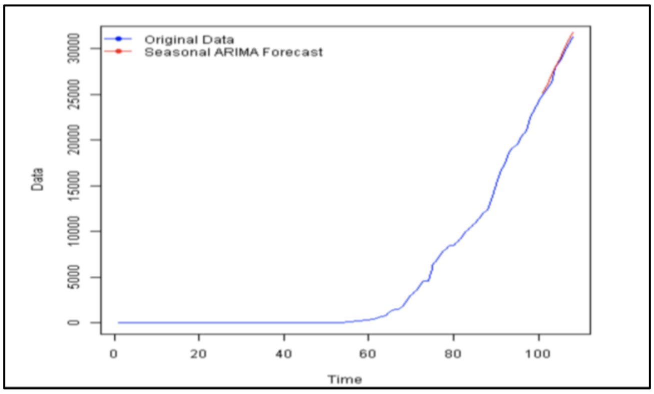


Fig 4: Seasonal ARIMA Forecast

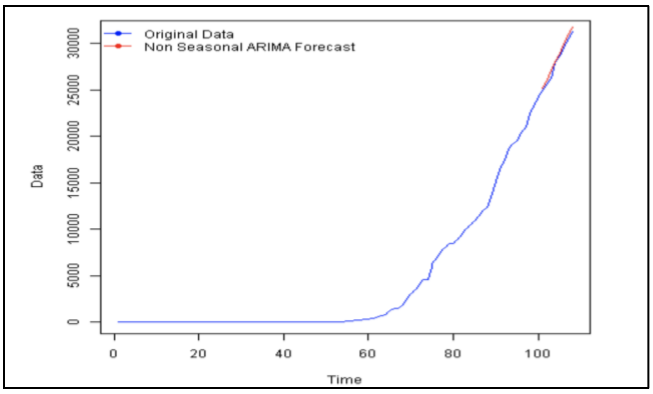


Fig 5: Non-Seasonal ARIMA Forecast

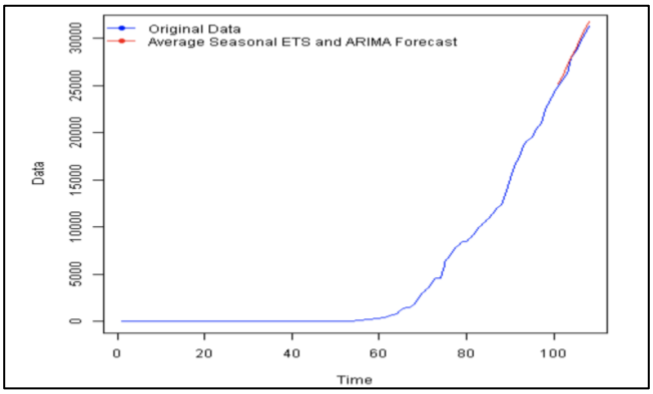


Fig 6: Average Seasonal ETS and ARIMA Forecast

4.3 Linear Regression

Linear Regression is a machine learning algorithm built on supervised learning. It implements a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

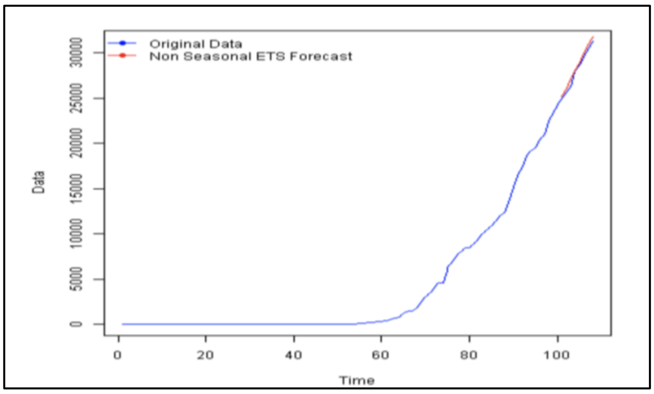


Fig 7: Non-Seasonal ETS Forecast

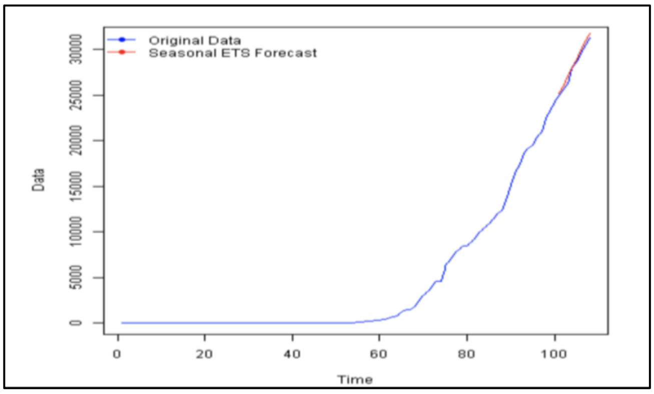
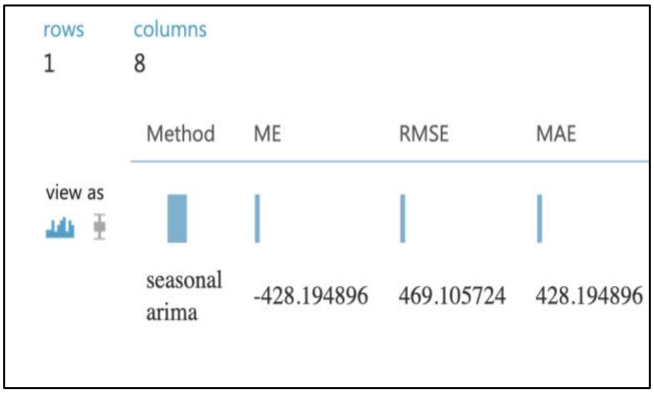


Fig 8: Seasonal ETS Forecast

4.4 Facebook Prophet

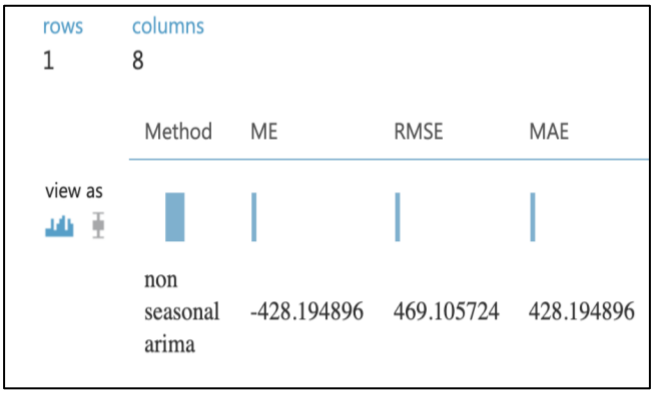
The prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. The prophet is open source software released by Facebook's Core Data Science team. It is fast, accurate, Fully automatic, and includesthe feature of tunaforecastsasts. We have used python facebook prophet in Databricks for prediction of the number of deaths and cases in Los Angels and New York which are two megacities in United States where Corona disease is spreading exponentially.



rows	columns
1	8

Method	ME	RMSE	MAE
seasonal arima	-428.194896	469.105724	428.194896

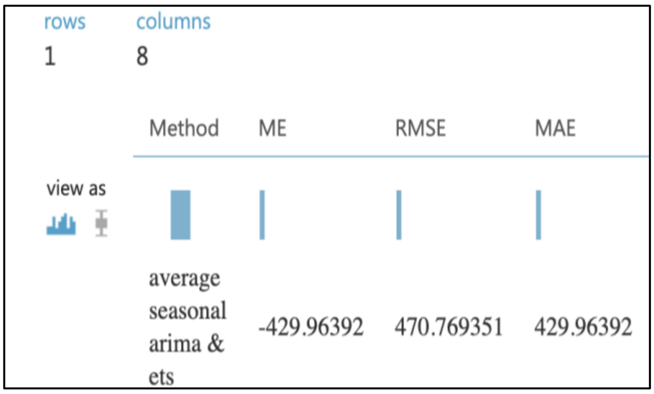
Fig 9: Seasonal ARIMA Root Mean Square Error



rows	columns
1	8

Method	ME	RMSE	MAE
non seasonal arima	-428.194896	469.105724	428.194896

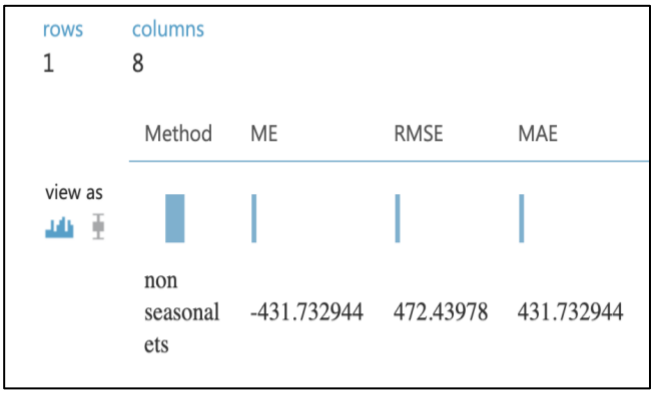
Fig 10: Non Seasonal ARIMA Root Mean Square Error



rows	columns
1	8

Method	ME	RMSE	MAE
average seasonal arima & ets	-429.96392	470.769351	429.96392

Fig 11: Average Seasonal ARIMA & ETS Root Mean Square Error



rows	columns
1	8

Method	ME	RMSE	MAE
non seasonal ets	-431.732944	472.43978	431.732944

Fig 12: Non Seasonal ETS Root Mean Square Error

5. CONCLUSION

Azure ML and Spark ML are powerful platforms for machine learning. Below is the summary of our experiment:

Azure ML	**RMSE**
Arima Seasonal	469.105724
Arima Non-Seasonal	469.105724
Average Seasonal ETS and Arima	470.769351
ETS Seasonal	471.43979
ETS Non Seasonal	472.43978
Spark	**AUC**
Logistic Regression Model	0.916030534351145
Decision Tree Regression Model	0.9416058394160584
Oracle BDCE	**AUC**
Logistic Regression Model	0.9596774193548387
Decision Tree Regression Model	0.965753424658

Table 2. Summary/Comparison Table

Following conclusion can be drawn from our project:

- Weather Factor is not prominent in prediction of deaths & casesdue to COVID.
- In Azure ML, ARIMA seasonal and non seasonal forecast is better compared to other algorithms.
- In comparison with WORLDOMETER, we have decent similar results for predicting number of cases and deaths.

6. Links

- **Dataset:**
  - **https://github.com/datasets/covid-19**
  - **https://www.accuweather.com**
- **GitHub:**
  - **https://github.com/shradha5410/5560**

References

- S. Jain, “Top 10 Benefits of Online Shopping (and 10 Disadvantages),” ToughNickel, 2018. [Online]. Available: https://toughnickel.com/frugal-living/Online-shopping-sites-benefits.
- M. Woolf, “Playing with 80 Million Amazon Product Review Ratings Using Apache Spark,” minimaxir, 02-Jan-2017. [Online]. Available: https://minimaxir.com/2017/01/amazon-spark/.
- Github Link: https://github.com/monika2403/mmishra2/tree/master/CIS%205560
- Dataset Link: https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_US_v1_00.tsv.gz