Assignment 4

CS215

Abhineet Majety 23B0923 | *Mohana Evuri* 23B1017 | *Saksham Jain* 23B1074

Fall 2024

1 Parking Lot Problem

The data cleaning is done, removing the unnecessary null values wherever applicable.S

1.1 Forecasting total number of vehicles

We have used the **SARIMAX** model to forecast the total number of vehicles for the next week. The parameters are `order=(7, 0, 10)` and `seasonal_order=(1, 1, 2, 12)` We have the MAPE: 0.04 and the MASE: 0.50. The plots obtained are:

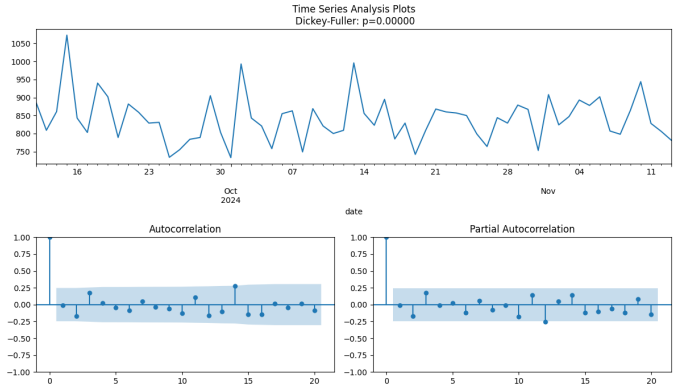

Figure 1: Time Series Analysis plots

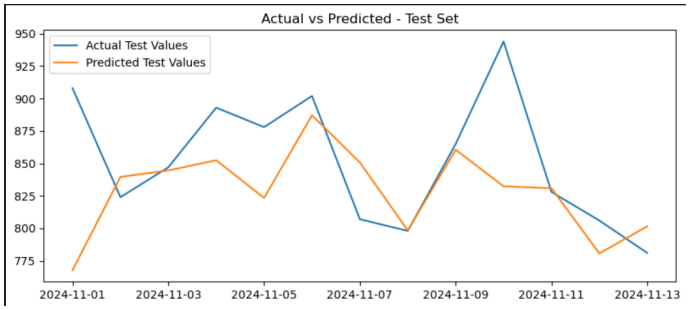

Figure 2: Test data and forecast

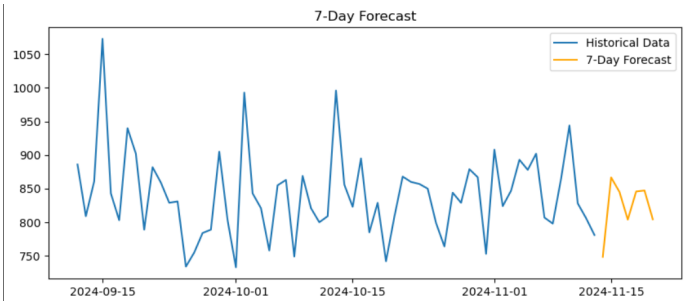

Figure 3: Forecasted plot

1.2 Forecasting average time spent

We have forecasted the average time spent using the **SARIMAX** model. The parameters `order=(3, 1, 3)` and `seasonal_order=(2, 1, 5, 7)` were used. We have obtained MAPE: 0.28 and MASE: 0.55. The plots obtained are here:

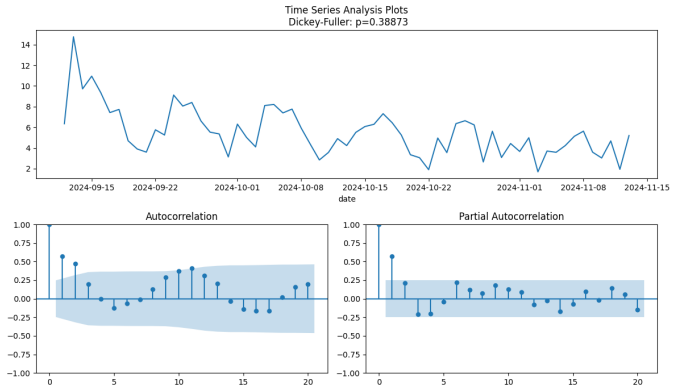

Figure 4: Time Series Analysis plots

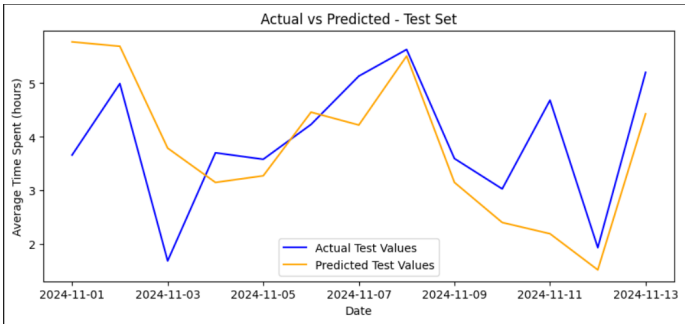

Figure 5: Test data and forecast

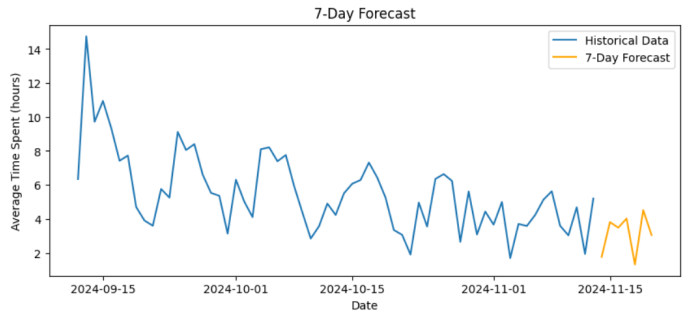

Figure 6: Forecasted plot

1.3 Smoothing strategies

We have used two strategies for smoothing of the data:

- Moving average smoothing
- Exponential smoothing

The code can be seen in `Q1c.ipynb`.

2 Forecasting on a Real World Dataset

2.1 Predicting `PASSENGERS CARRIED` from 2023 September to 2024 August

2.1.1 Using ARIMA model to forecast the data

We first cleaned the data by removing unnecessary whitespaces. We also renamed the months with an index to make it into an iterable format, following which we applied the **ARIMA** model to the data to predict future data. Note that we **included COVID-19** in the data as it is a prominent feature that shouldn't be ignored.

We have tried the ARIMA plots for various parameters until we settled for `ARIMA(order=(10, 1, 10))`. The **ACF** and the **PACF** plots are

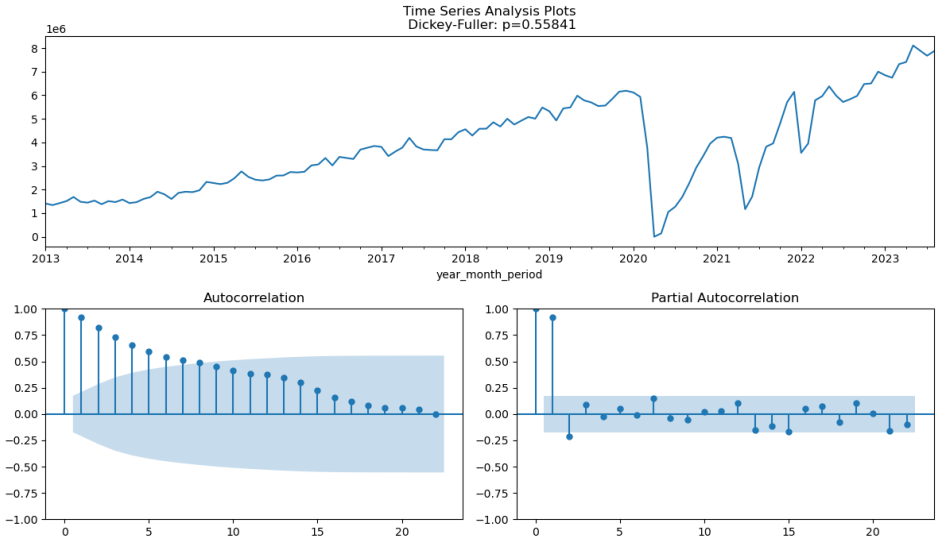

Figure 7: Time Series Analysis Plots

and the ARIMA model predictions are

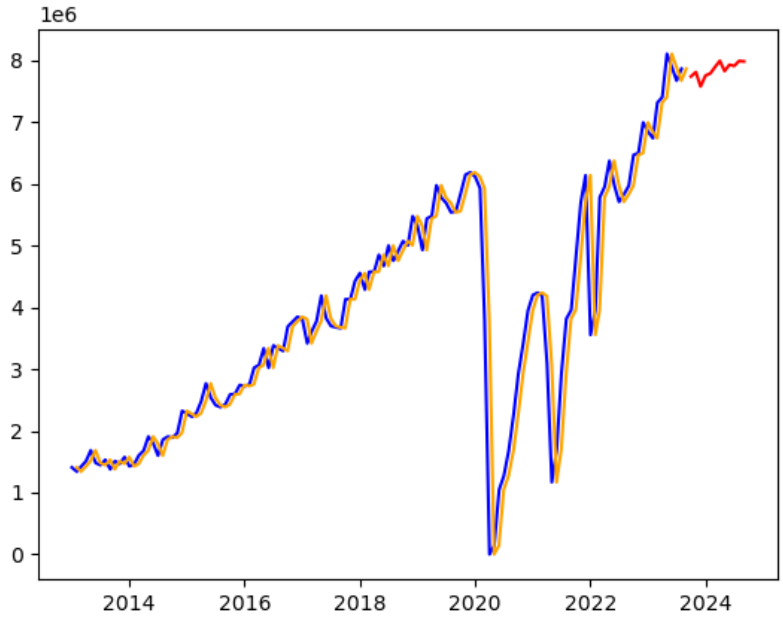

Figure 8: ARIMA applied (The red is the predicted part)

The `submission.csv` file in Kaggle gives a score of 2.84.

2.1.2 Using an LLM for prediction

Input prompt to LLM (Google AI Studio):

```
Given the following monthly airline passenger data, predict and display
the values for next 12 months (2023 SEP to 2024 AUG) for Passengers Carried,
you need not show any code or your though process, just the final predicted
values must be displayed, there's a season of 4 months: Airline A007, Year
2023, Month JAN, Passengers Carried: 6847384.0. Airline A007, Year 2023, Month
FEB, Passengers Carried: 6741948.0. Airline A007, Year 2023, Month MAR,
Passengers Carried: 7317288.0. Airline A007, Year 2023, Month APR, Passengers
Carried: 7406440.0. Airline A007, Year 2023, M...complete the prompt by
extracting the data from the attached CSV, after removing the covid null
values

-- The CSV file contents --
```

The output of the LLM (in Markdown):

```
* **2023 SEP:** 79,02,627
* **2023 OCT:** 79,58,513
* **2023 NOV:** 80,12,872
* **2023 DEC:** 80,65,695
* **2024 JAN:** 73,42,625
* **2024 FEB:** 72,37,189
* **2024 MAR:** 78,02,983
* **2024 APR:** 78,92,132
* **2024 MAY:** 85,92,379
* **2024 JUN:** 84,76,054
* **2024 JUL:** 82,57,578
* **2024 AUG:** 84,49,947
```

2.1.3 Using Prophet for model fitting

The Prophet model was used to make the data prediction. The output data graph:

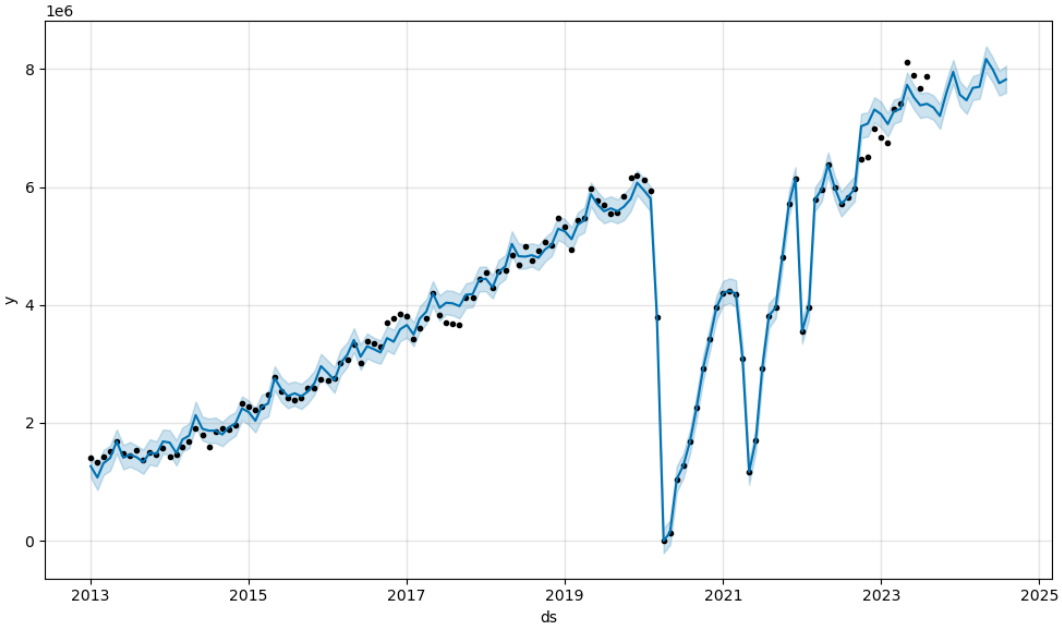

Figure 9: Model fitting using Prophet

The predicted values are:

```
YEAR_MONTH,PASSENGERS CARRIED
2023 SEP,7352169.4825807
2023 OCT,7210166.934052202
2023 NOV,7614914.074459398
2023 DEC,7957471.519386159
2024 JAN,7568591.618646984
2024 FEB,7475801.199378586
2024 MAR,7686314.721362665
2024 APR,7705897.56702898
2024 MAY,8178199.290168875
2024 JUN,7999134.73277676
2024 JUL,7766336.116382452
2024 AUG,7829486.003933291
```

2.2 Metrics to evaluate forecasts

The **fleet requirement** is constrained by the total number of passengers expected. Hence, there is a greater need for better forecasting during periods of high demand than during periods of low demand. The **human resources requirement** is constrained by the peak demand. Hence, the forecast for the peak demand must be accurate, i.e., the error must be more sensitive to the peaks.

Disadvantages of MAPE

- MAPE calculates the average percentage error across all periods and hence does not differentiate between high and low demand periods.
- When the actual values are small, the MAPE will show high sensitivity to the errors. In reality, the errors at low demand are not operationally significant.

Alternative evaluation metrics

- Weighted Absolute Percentage Error (WAPE) is a better alternative to MAPE. This is because it shows high sensitivity towards errors at high values. This addresses the requirement of better prediction at the peaks.
- Root Mean Squared Error (RMSE) is also a good metric because higher deviation during peaks impacts more than MAPE.

2.3 Test to study μ

To test if μ is different before and after COVID, we can use the **t-test** considering the two periods as two samples. In this test, the statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is used, where s is the sample standard deviation.