Assignment 4

Sagnik Nandi (23B0905)

Jainendrajeet (23B1008)

Sumedh S S (23B1079)

October 28, 2024

1 Parking Lot Problem

(a) Forecast total number of vehicles entering the parking per day

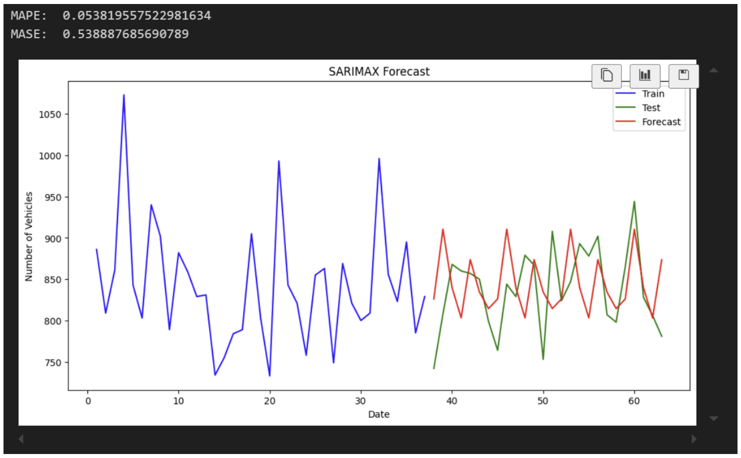

Figure 1: Train data(blue), validation data(green) and Predictions generated by **Seasonal ARIMA** model (red)

- **MAPE** : 0.053

- **MASE** : 0.538

Predictions for next week (days indexed from 0):

Day	Entry count
63	828
64	823
65	820
66	882
67	844
68	825
69	860

Table 1: Forecast of Number of Cars Entered

(b) Forecast avg time spent in the mall by a vehicle entering on a particular day, for the next 7 days

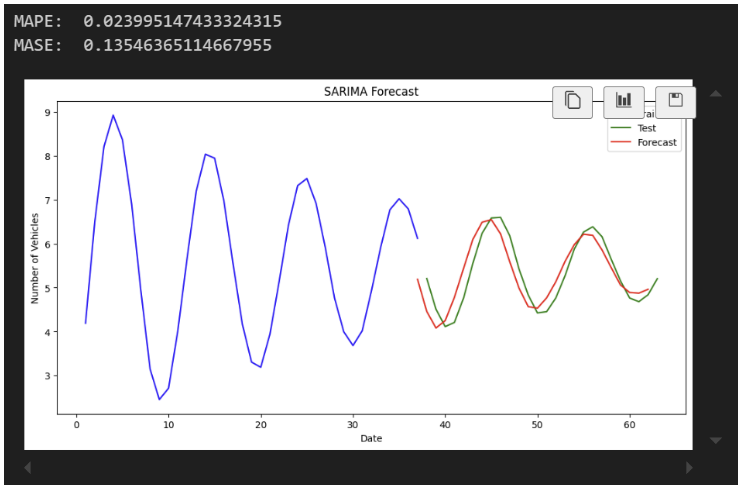

Figure 2: Train data(blue), validation data(green) and Predictions generated by **Seasonal ARIMA** model (red)

- **MAPE** : 0.023

- **MASE** : 0.135

Predictions for next week (days indexed from 0):

Day	Average time (hours)
63	5.796953
64	6.231002
65	6.445560
66	6.381449
67	6.073602
68	5.669879
69	5.283271

Table 2: Forecast of Average Time Spent

(c) Experiment with at least 2 other outlier smoothing or missing value imputation strategies

1.0.1 1a continued

Strategy1-Remove outliers using Z score calculation:

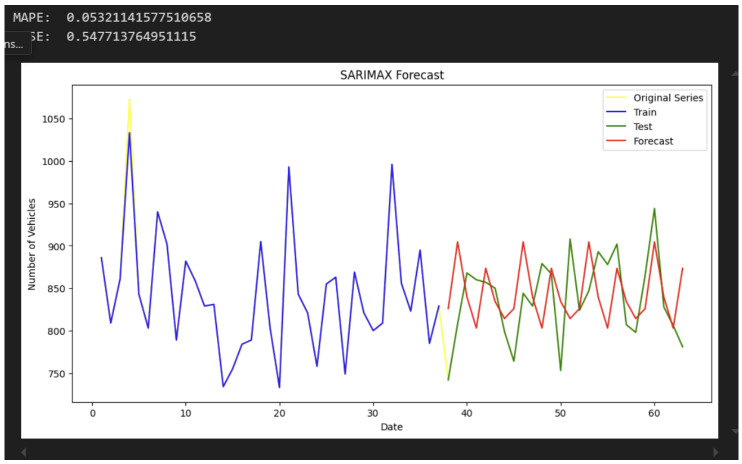

Figure 3: Train data(blue), validation data(green) and Predictions generated by **Seasonal ARIMA** model (red)

- **MAPE** : 0.053
- **MASE** : 0.567

Strategy2-Use rolling window average to smoothen the outliers:

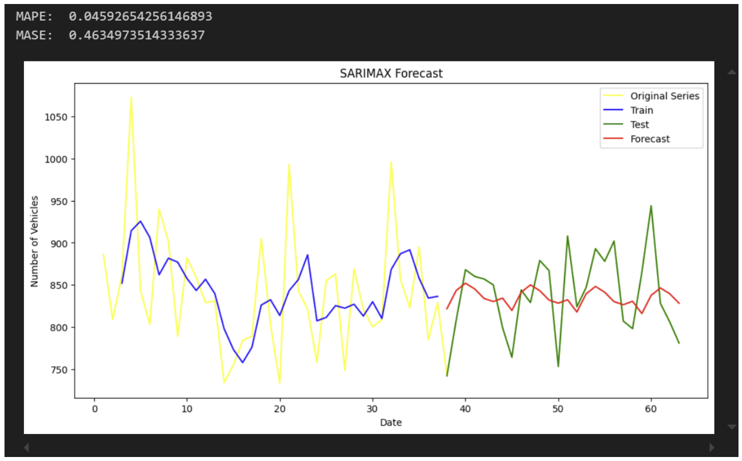

Figure 4: Train data(blue), validation data(green) and Predictions generated by **Seasonal ARIMA** model (red)

- **MAPE** : 0.046

- **MASE** : 0.482

1.0.2 1b continued

Strategy1-Match camera records which could not be matched exactly but can be done so with one character mismatch:

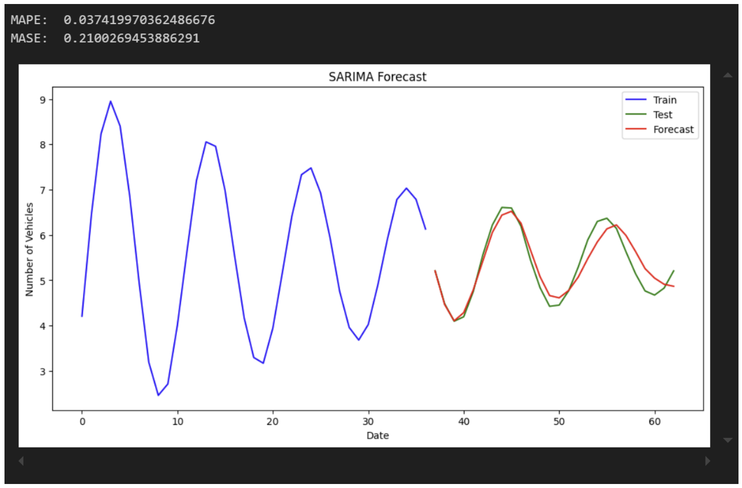

Figure 5: Train data(blue), validation data(green) and Predictions generated by **Seasonal ARIMA** model (red)

- **MAPE** : 0.037
- **MASE** : 0.210

Strategy2-Match the remaining entry and exit items with the null values:

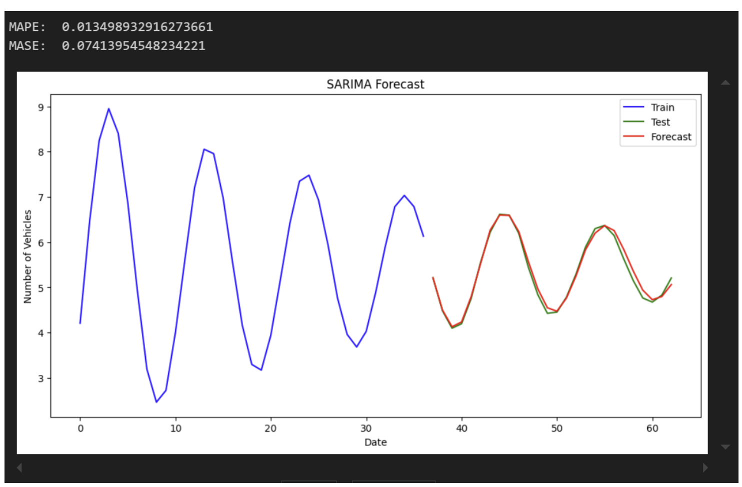

Figure 6: Train data(blue), validation data(green) and Predictions generated by **Seasonal ARIMA** model (red)

- **MAPE** : 0.013
- **MASE** : 0.074

2 Forecasting on a Real World Dataset

2.1 Prediction

(a) Using methods discussed in class

The plot and MAPE obtained by different apporaches are shown below.

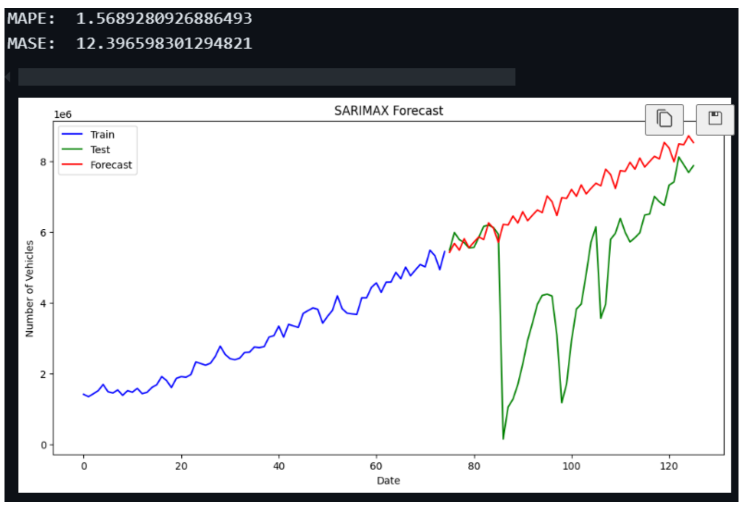

Figure 7: Train data(blue), validation data(green) and Predictions generated by **SARIMAX (1,1,1)x(1,1,1,12)** model (red)

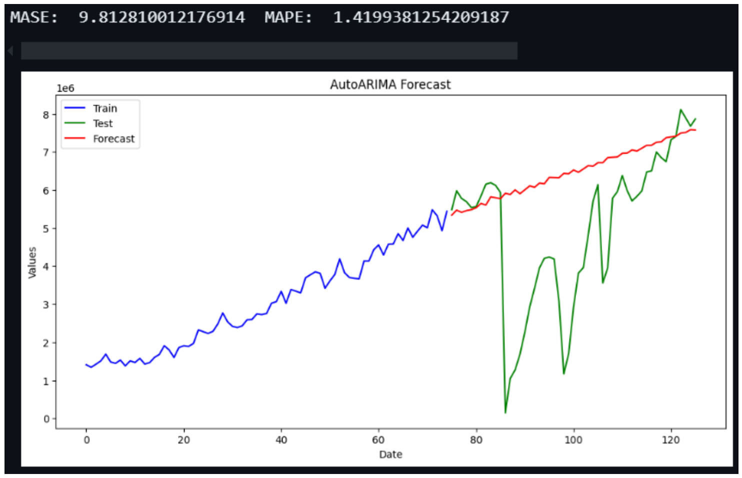

Figure 8: Train data(blue), validation data(green) and Predictions generated by **AutoARIMA (seasonality=true)** model (red)

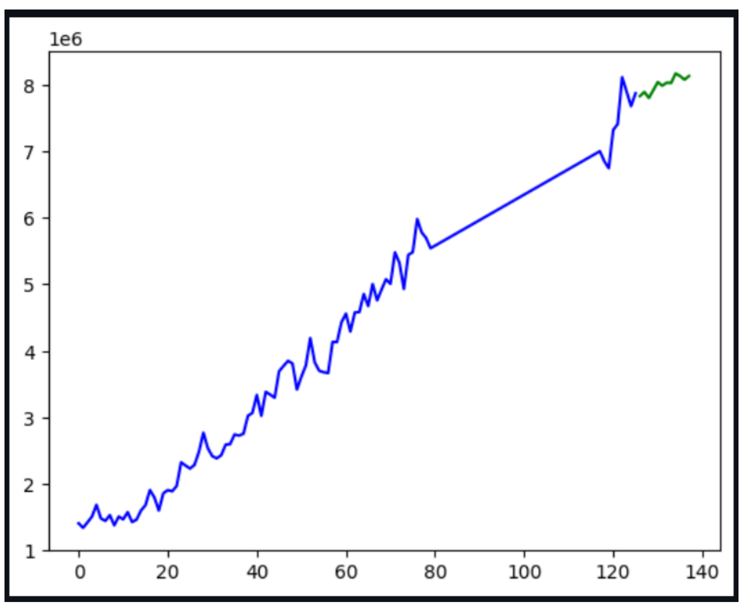

Figure 9: Data(blue) and Predictions generated by **SARIMAX (0,1,1)x(1,0,2,6)** model where Covid period was removed and interpolated (green)

(b) Using LLM

Given prompt:
Given the dataset (passengers carried by an airline over the years 2013 Jan-2023 Aug) below, forecast next 12 values of the time series.
(y-tokens as generated by python code)
Return the answer in JSON format, containing two keys: 'time-idx' and 'forecast-col', and list of values assigned to them. Return only the forecasts, not the Python code.
Tokenization:
Each number is converted to a string of digits separated by spaces and then included in the prompt to GPT model.
Prediction:
(7879448.797360663, 7820390.150911682, 7936876.144944265, 7918937.200867041, 8018894.378393498, 7923624.878678311, 7792785.668688422, 7883954.739400877, 7950816.722030679, 8206739.740450005, 7912619.781882194, 7866005.377980492
)

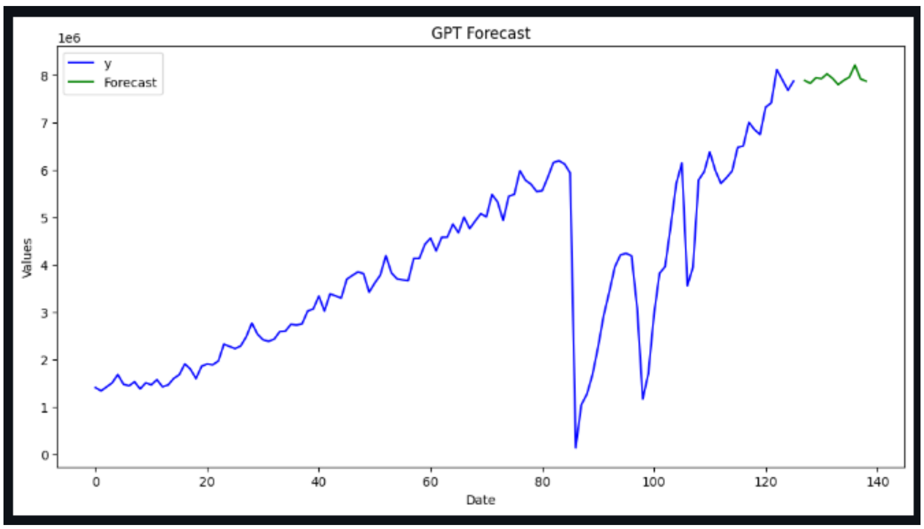

Figure 10: Data(blue) and Predictions generated by GPT (green)

(c) Using Global Model

Predictions (after training on whole data and considering covid period as an anomaly):
(A regressor based on covid period was added to the prophet model.)

SN	Date	Passengers Carried
126	2023-08-31	7.240524e+06
127	2023-09-30	7.916970e+06
128	2023-10-31	8.283024e+06
129	2023-11-30	8.747879e+06
130	2023-12-31	8.143012e+06
131	2024-01-31	7.394197e+06
132	2024-02-29	7.938122e+06
133	2024-03-31	7.767246e+06
134	2024-04-30	7.472996e+06
135	2024-05-31	7.396513e+06
136	2024-06-30	7.540277e+06
137	2024-07-31	7.786391e+06

Table 3: Predicted number of passengers by prophet

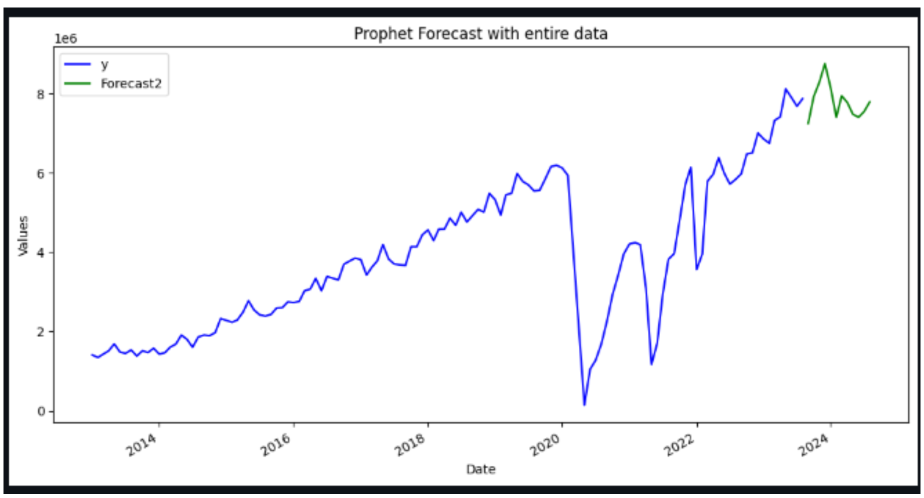

Figure 11: Data(blue) and Predictions generated by Prophet (green)

2.2

Why MAPE may not be a good metric
1. MAPE treats all percentage errors equally, whether they occur during high or low demand periods. However, operational planning for fleet and staffing focuses more on peak demand rather than average demand.
2. A small percentage error during high-demand months can have a bigger operational impact than a large error during low-demand months, but MAPE doesn't account for this.
Alternative Metric – Root Mean Squared Error (RMSE)

RMSE squares the errors, so it gives more weight to large deviations, which are more critical for peak demand planning.

2.3

To test if the mean (μ) of the differenced series (ΔY) is different pre-COVID (before Dec 2019) and post-COVID (after Jan 2022), the appropriate statistical test would be a **two-sample z-test for means**.

Reason for Choosing the Z-Test:

1. The question specifies that σ (the standard deviation of the series) is known, which is a condition for using the z-test.
2. The objective is to check if the mean of μ has changed across two independent periods. And z-test is suitable for this.