CS215
Assignment4

Kaushal Vijayvergiya 23B0901
Kamyak Channa 23B0949
Tanishq Mandhane 23B0986

October 2024

1 Parking Lot Problem

(A) SARIMA Forecast for Next 7 Days

Day	Forecasted Vehicles
1	858
2	856
3	888
4	905
5	877
6	858
7	896

Table 1: SARIMA Forecast of Vehicles Entering

Summary

Mean Absolute Scaled Error (MASE): 0.7518
Mean Absolute Percentage Error (MAPE): 5.76%

(B) ARIMA Forecast on Average Time Spent for Next 7 Days

Day	Forecasted Time (mins)
1	340
2	368
3	381
4	375
5	351
6	320
7	291

Table 2: ARIMA Forecast on Average Time Spent

Summary

Mean Absolute Scaled Error (MASE): 0.0837
Mean Absolute Percentage Error (MAPE): 1.32%

(C) SARIMA Forecasts Using STL and Rolling for Next 7 Days

Date	Forecasted Vehicles (STL)
2024-11-14	855
2024-11-15	875
2024-11-16	869
2024-11-17	877
2024-11-18	882
2024-11-19	877
2024-11-20	886

Table 3: SARIMA Forecast of Vehicles Entering Using STL

STL Forecast Summary

Mean Absolute Scaled Error (MASE): 0.9996
Mean Absolute Percentage Error (MAPE): 7.84%

Date	Forecasted Vehicles (Rolling)
2024-11-14	862
2024-11-15	854
2024-11-16	858
2024-11-17	876
2024-11-18	864
2024-11-19	855
2024-11-20	877

Table 4: SARIMA Forecast of Vehicles Entering Using Rolling

Rolling Forecast Summary

Mean Absolute Scaled Error (MASE): 0.9340
Mean Absolute Percentage Error (MAPE): 7.04%

Day	Forecasted Time (Rolling, mins)
1	359.58
2	391.66
3	400.22
4	382.29
5	344.48
6	300.58
7	266.48

Table 5: ARIMA Forecast on Average Time Spent Using Rolling

Rolling Forecast Summary

Mean Absolute Scaled Error (MASE): 0.3827
Mean Absolute Percentage Error (MAPE): 6.72%

Day	Forecasted Time (STL, mins)
1	339.20
2	367.93
3	380.83
4	373.39
5	348.41
6	314.98
7	285.16

Table 6: ARIMA Forecast on Average Time Spent Using STL

STL Forecast Summary

Mean Absolute Scaled Error (MASE): 0.2250
Mean Absolute Percentage Error (MAPE): 3.64%

2 Forecasting on a Real World Dataset

2.1.1

We were able to reduce the error to roughly 1.50. The ETS model for predictions.

2.1.2

The prompt we gave was "Below is a tokenised version of time series data. Predict the next 12 values of the time series data." This was then followed by the tokenized data which was generated using a python script.

2.1.3

Using prophet model, we were able to reduce the error to roughly 2.54. We used the month and year for predictions and additionally we included a variable which took care of the effects of COVID-19 on the number of passengers.

MAPE calculates the absolute percentage error, making it extremely sensitive to small actual values. If certain months have lower passenger counts or other metrics, even a small forecast error can yield a large percentage error.

This could lead to misleading insights, causing the model to over-prioritize accuracy during low-demand periods, which aren't as critical to planning as peak periods. It prioritizes all the time steps equally, but for an airline it is more important to improve the accuracy during the high demand periods because they need

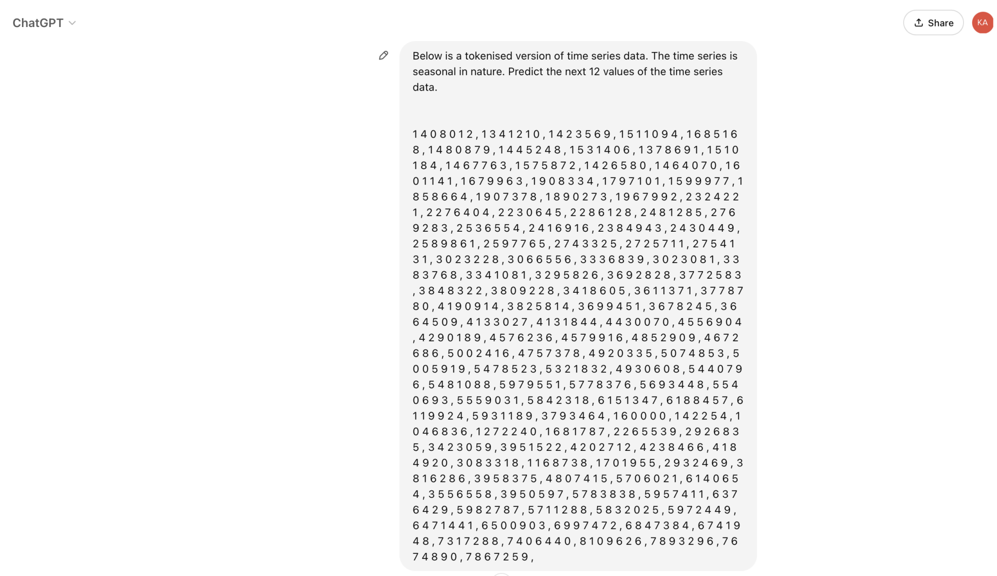


Figure 1: Prompt

to hire to avoid under-staffing and inadequate fleet allocation during such times.

Also MAPE does not care whether the forecasted demand is more or less than the actual demand, but that is extremely important for fleet requirements. Under-forecasting can result in passenger capacity shortages and staff strain, whereas over-forecasting leads to under-utilized resources.

A better error estimate would be the **Weighted Mean Squared Error(RMSE)**. It is given by

$$\sqrt{\frac{\sum_{i=1}^{n} w_i (a_i - p_i)^2}{\sum_{i=1}^{n} w_i}}$$

where a_i represents the actual value and p_i represents the predicted value and w_i represents the weight given to each observation.

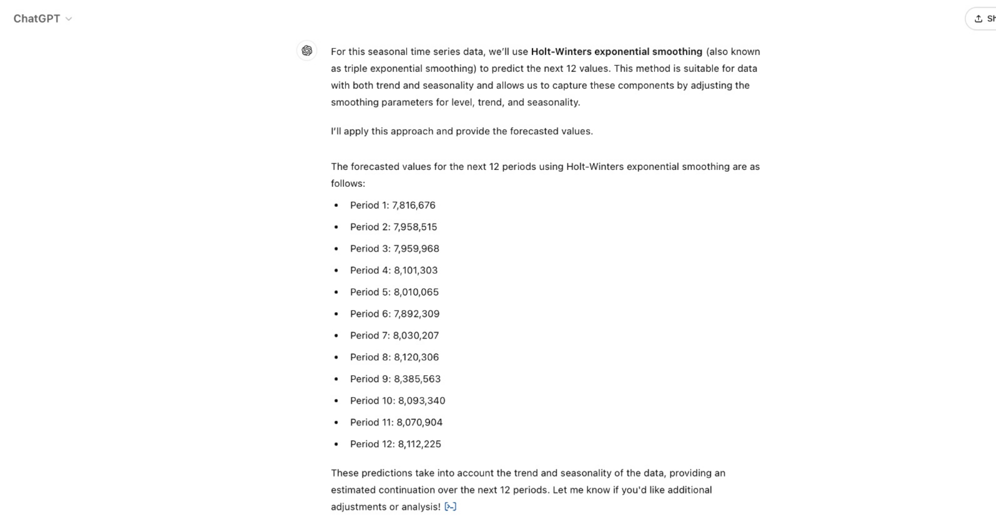
WMSE is more suitable because it allows assigning higher weights to periods with higher expected demand. This is helpful for airlines because large forecast errors during high-demand periods are more drastic than during low-demand times and so they should be assigned higher weights.

To test if the means before and after covid are same or not, we can perform a two sample t-test. Since the variances of the two samples are the same and the samples are independent (it follows from the fact that the series is stationary), the requirements of the t-test are fulfilled.

In hypothesis testing terms, it represents the probability of obtaining the observed data (or more extreme values) under the assumption that the null hypothesis is true.

We form the null hypothesis (H_0) that the means of the two samples are the same. We now calculate the test statistic

$$t = \frac{(\bar{\mu}_{\text{pre}} - \bar{\mu}_{\text{post}})}{s\sqrt{\frac{1}{n_{\text{pre}}} + \frac{1}{n_{\text{post}}}}}$$



ChatGPT

For this seasonal time series data, we'll use **Holt-Winters exponential smoothing** (also known as triple exponential smoothing) to predict the next 12 values. This method is suitable for data with both trend and seasonality and allows us to capture these components by adjusting the smoothing parameters for level, trend, and seasonality.

I'll apply this approach and provide the forecasted values.

The forecasted values for the next 12 periods using Holt-Winters exponential smoothing are as follows:

- Period 1: 7,816,676
- Period 2: 7,958,515
- Period 3: 7,959,968
- Period 4: 8,101,303
- Period 5: 8,010,065
- Period 6: 7,892,309
- Period 7: 8,030,207
- Period 8: 8,120,306
- Period 9: 8,385,563
- Period 10: 8,093,340
- Period 11: 8,070,904
- Period 12: 8,112,225

These predictions take into account the trend and seasonality of the data, providing an estimated continuation over the next 12 periods. Let me know if you'd like additional adjustments or analysis!

Figure 2: Result

where s is given by

$$\sqrt{\frac{(n_{\text{pre}}-1)s_{\text{pre}}^2+(n_{\text{post}}-1)s_{\text{post}}^2}{n_{\text{pre}}+n_{\text{post}}-2}}$$

where s_{pre} and s_{post} and the variances of the data before and after COVID respectively. This is now compared to the critical value based on the degrees of freedom $(n_{pre} + n_{post} - 1)$ and the set significance level.

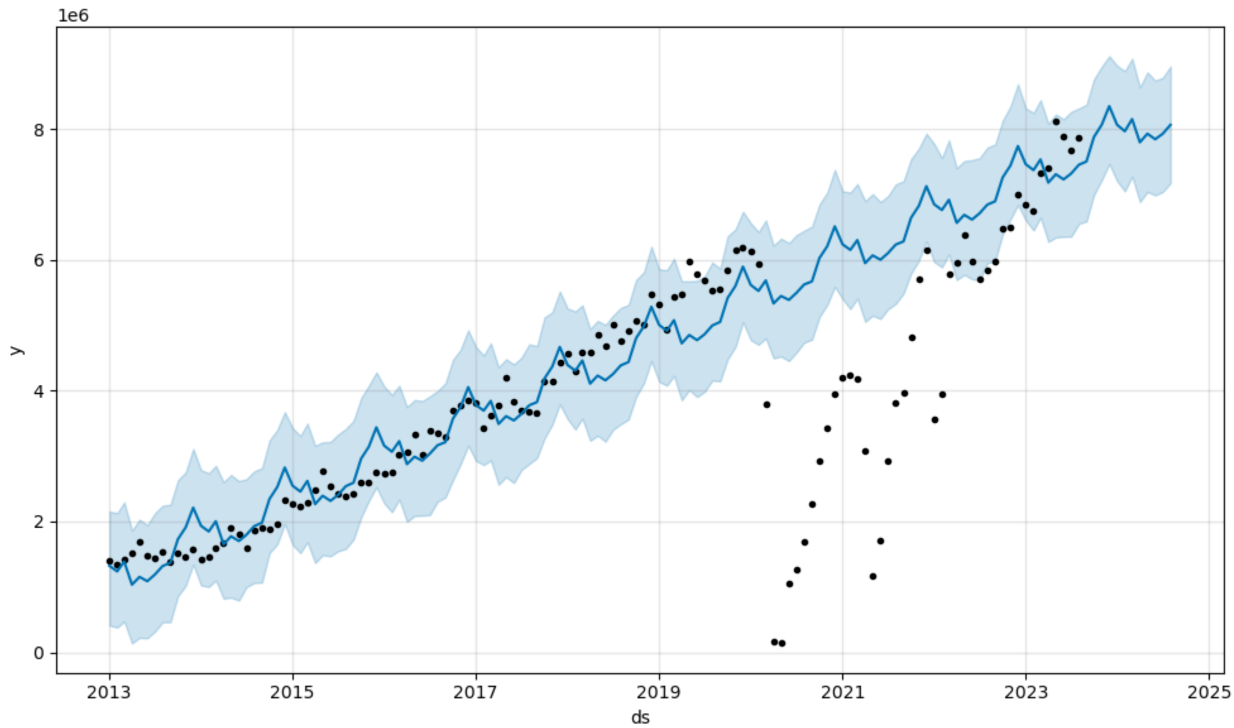


Figure 3: Plot