# CS215 Data Analysis and Interpretation

## Assignment 4

**Aditi Singh (23B1053)**

**Abhilasha Sharma Suman (23B1011)**

**Navya Garg (23B0982)**

Indian Institute of Technology, Bombay

October 2024

# Contents

# 1 Parking Lot Problem

## 1.1 Part a

In this part we had to forecast the total number of vehicles entered for the next 7 days. We started with data cleaning processes in which we did the following steps

- Removed any columns which had null entries
- Removes any duplicate rows in the data
- Created a merged data frame and added vehicles based on vehicle number and removed those datapoints where either entry or exit was not provided.
- Removed all the data points whose exit and entry times were in between 12 to 5

Following these steps, we got our merged data frame in which each vehicle had an entry and an exit time. Using these we were able to calculate the number of vehicles entering the mall each day yielding us the required dataset. Following this we tried fitting several models in the time series Which included ARIMA and Exponential Smoothing. We can see that the Exponential Smoothing method performs better than Auto ARIMA in this case. Following this I even tried using Prophet and Sarima to predict the values. However, their performance was not as expected. So I obtained the **MAPE 5.43%** and **MASE 0.8027** using Exponential Smoothing

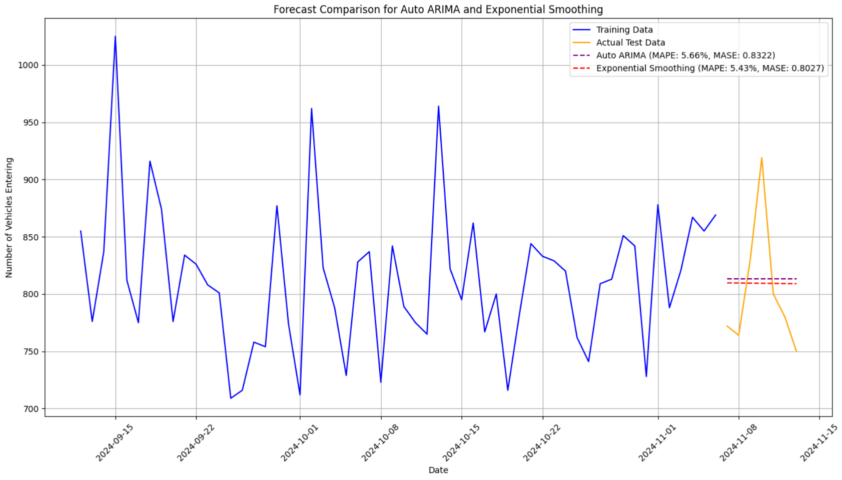

Figure 1: Data prediction comparison with ARIMA and Exponential Smoothing

## 1.2 Part b

Similar to the above case I used the filtered data and calculated the difference between the exit time and entry time to calculate the average time spent. Following this I fitted the AUTO Arima model in this and obtained a perfect fit with **MAPE 1.55%** and **MASE 0.2417**. Here Exponential Smoothing was not accurate and Auto ARIMA was way better.

## 1.3 Part c

Here I used two important strategies. For the Computation of missing values, I use **Backward Filling**. This method simply fills missing values by copying the most recent previous or next valid value in the series. With the help of this, I obtained the obtained values for missing days in a

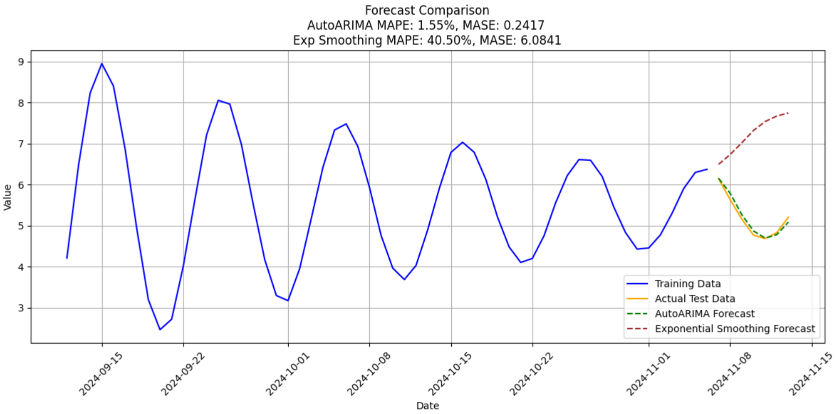

Figure 2: Average Time spent

row and was successfully able to calculate the values of missing values. The graphs and analysis is provided in `Q1c.ipynb`.

For Outlier Smoothing I applied the method of **Rolling Mean Smoothing**. This method replaces each value with the mean of its neighbourhood, smoothing extreme outliers. Again the analysis is present in the Q1c.ipynb file. One of the graphs showing the effect of Rolling mean smoothing is attached here With this our prediction for No. of vehicles had an **MAPE:- 1.09**

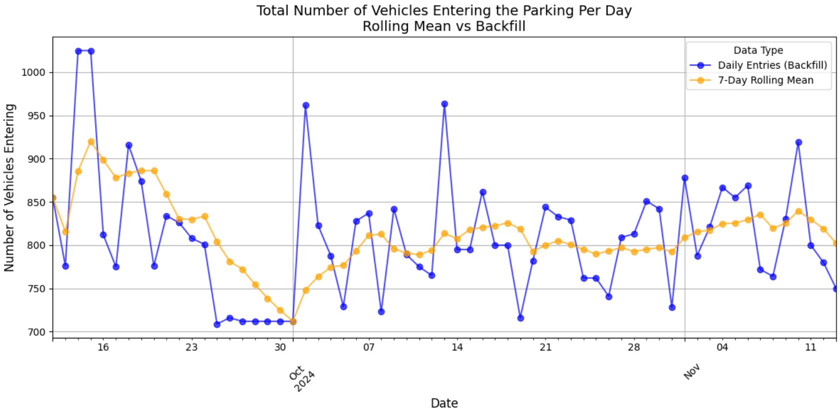

Figure 3: Rolling Mean Smoothing in no. of vehicles

and **MASE :- 0.6952**. For the prediction of Average time spect we had an **MAPE:- 1.51** and **MASE:-0.5887**.

I then used the method of z-score calculation to remove the outliers. The Z-score, also known as the standard score, is a statistical measurement that describes how many standard deviations a data point is from the mean of a dataset. Typically, a Z-score above 3 or below -3 is considered an outlier. This means the data point is more than 3 standard deviations from the mean, which is relatively rare in a normal distribution. I have used this principle and removed the outliers. The relevant graphs are attached in the `Q1c.ipynb`. So these were a few of the methods which we used in this part.

# 2 Forecasting on a Real World Dataset

## 2.1 Predicting the number of Passengers Carried

1. **Part1: Predicting the number of Passengers Carried using ARIMA Model**
   Since there is an anomalous depression in the total number of passengers carried in the years 2020, 2021, 2022 and 2023, we can not use this data to predict future values as we can see that the values have started returning to the original increasing trend roughly from 2023 onwards and so we dropped the data from these years.
   Now considering only the data from 2013 to 2020, we used an **ARIMA** model with

   $$p = 10 \qquad d = 2 \qquad q = 20$$

   and trained it over the data. Then using this model we forecasted the number of passengers carried for the next four years, including the required timeframe, "Sep 2023 to Aug 2024" and this brought us an **MAPE Score** of **2.20** on Kaggle

2. **Part2: Using an LLM to generate predictions**

   To provide a refined dataset to the LLM used (ChatGPT), firstly I cleaned the data by converting all numeric values to floats, creating a year_month column. Then for the data from the Covid period, since it was significantly out of line with the rest of the data, I removed the data from January 2020 to December 2021. This was the basis of the first few prompts I gave. The rest of the prompts were basically to somewhat fine-tune the predictions made by the LLM.

   To fine-tune the data I first tried a few different parameters in the ARIMA model but later switched to the SARIMA model to better capture the seasonality associated with the air traffic data. After adjusting the parameters of the SARIMA model, I got my final submission which got a score of 3.87 on the Kaggle contest.

3. **Training a Global Model to make predictions**

   In this part, we used the pre defined **ARIMA** function from the statsmodels.tsa.arima.model library. As we did in the other cases, we cleaned the data by removing the data from the Covid years (2020 and 2021) from the training data and the data is indexed with the dates. Then we used ARIMA model to predict the values for the next 12 months (Sept 2023 to Aug 2024) and visualised the data and the predictions using a plot to show the forecasted values plotted along with their uncertainty intervals to visualize expected trends.

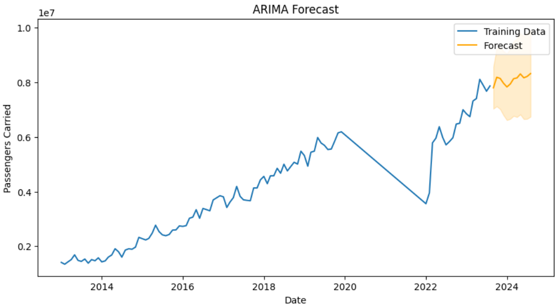

Figure 4: Visualising Global Model Prediction

This approach ensures that the model leverages all available data and related time series, producing comprehensive forecasts for future passenger counts.

## 2.2 Finding a fair metric to evaluate the forecast

Mean Absolute Percentage Error (MAPE) is not a good metric to judge our forecast of the passengers to be carried by the airlines in the future due to the following reasons:

1. **Dependence on the actual value of predicted parameter**: Due to percentages being involved, we are required to divide the error by the actual value. This means for a smaller actual value we could get a more significant error than we would for a larger actual value even for the same magnitude error.
2. **Cannot differentiate between Overestimated and Underestimated Values**: A 5% overestimate and a 5% underestimate are both the same as MAPE but they could have a huge impact on the business of an airline company. An underestimate could lead to lesser resources which could lead to customer disgruntlement and negatively impact the perception of the services provided by the company. On the other hand, overestimating could lead to wastage of resources and would potentially drain resources from a place where they are actually required (poor allocation of resources).
3. **Fluctuation in Demand**: MAPE cannot adequately account for seasonal fluctuations in demand for airline travel which would, again, lead to our forecast being evaluated incorrectly.

While considering a business like this, we need to keep in mind that we have to decide upon the allocation of two types of resources: Fleet Resources (aircraft and all equipment required by them) and Human Resources (crew, pilots, ground staff). Depending on which resource we have a more pressing need for, we can provide different weights to them (significance of the need for said resource) and then evaluate our predictions keeping these in mind. This can be accomplished by using **Weighted Mean Absolute Percentage Error** (WAPE):

$$\text{WMAPE} = \frac{W_{\text{fleet}} \cdot \sum \left| \frac{Y_{\text{fleet}}^{\text{actual}} - Y_{\text{fleet}}^{\text{forecast}}}{Y_{\text{fleet}}^{\text{actual}}} \right| + W_{\text{HR}} \cdot \sum \left| \frac{Y_{\text{HR}}^{\text{actual}} - Y_{\text{HR}}^{\text{forecast}}}{Y_{\text{HR}}^{\text{actual}}} \right|}{W_{\text{fleet}} + W_{\text{HR}}}$$

1. We can create the weight criteria looking at the impact the passenger traffic has had on both resources in the company's past.
2. Implementing the weighted metrics, we can regularly evaluate the optimality of the weights assigned and alter them based on changing business priorities, historical performance, and any significant shifts in passenger demand patterns. Using feedback loops and ongoing analysis, resource allocation can continuously be made optimal.

Thus this can allow for a more nuanced evaluation of the predictions and allow to gather further insights into the company's operations.

## 2.3 Working with Pre-Covid and Post-Covid data

$\Delta Y$ is the first differenced time series and is given to be weakly stationary with a constant mean $\mu$ and a normal distribution $\mathcal{N}(0, \sigma)$. Now we wish to see whether $\mu$ was different in pre and post covid times. To accomplish this we use the **two sample t test** and compare both the samples. As we do this, we have assumed that The data in each period is drawn from a normally distributed population with a known variance $\sigma$. We also assume that the samples are independent being separated by the COVID period.

- To conduct the test, we follow this equation

$$t = \frac{\bar{X}_{preCovid} - \bar{X}_{postCovid}}{\sqrt{\frac{\sigma^2}{n_{preCovid}} + \frac{\sigma^2}{n_{postCovid}}}}$$

We then compare the value of t-statistic against a critical value from the t distribution to check if the difference between the means is statistically significant.