Assignment 4

CS215: Data Analysis and Interpretation

Shaik Awez Mehtab, 23B1080
Vaibhav Singh, 23B1068
Satyam Sinoliya, 23B0958

Solutions

SOLUTION 1

Part (a)

SOLUTION 2

Forecasting on a Real World Dataset

2.1

Part (a)

Our data has

- Trend
- Seasonality (3 months)
- Varying variance

To remove varying variance I took log of the data, then took difference with lag 3 to remove seasonality.

Training the model

Since there's a clear seasonality, I added a `SARIMA` (seasonal `ARIMA`) model. I'm using the parameters which can by easily deduced using `ACF` and `PACF` plots of the preprocessed data. Also we evaluate errors (`MASE` and `MAPE`).

Using and Interpreting our Model

Finally, we solve the question! Predicting the number of `PASSENGERS CARRIED` from 2023 September to 2024 August. Since we added some preprocessing to our data, we need to reverse it. Here are the things we did

- Took log of the data
- Took difference with lag 3 (season)

So I first reverse the difference by adding data at 3 values before. Then I reverse the log by taking exp of the data.

Part (b)

This is the prompt given:

```
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.
```

The final evaluations by the LLM GPT4o are:

Year & Month	Passengers Carried
23 SEP	7789024
23 OCT	8122317
23 NOV	8225536
23 DEC	8504783
24 JAN	7739815
24 FEB	7631294
24 MAR	8124572
24 APR	8218501
24 MAY	8802926
24 JUN	8550792
24 JUL	8289312
24 AUG	8494387

Part (c)

After applying the prophet model, the final predictions are:

Year & Month	Passengers Carried
23 SEP	1.316906e+06
23 OCT	2.067208e+06
23 NOV	2.391833e+06
23 DEC	2.002217e+06
24 JAN	2.112962e+06
24 FEB	2.170063e+06
24 MAR	2.191516e+06
24 APR	2.404104e+06
24 MAY	2.605715e+06
24 JUN	2.192019e+06
24 JUL	2.120529e+06
24 AUG	2.404104e+06

2.2

In forecasting demand for fleet management and human resource planning, Mean Absolute Percentage Error (MAPE) may not be the best metric for evaluation. This is because

- **Sensitivity to Low Passenger Volumes:** MAPE can inflate errors in months or seasons with low passenger volumes because it calculates percentage error. For fleet management, which focuses on total passenger volume over a quarter, these small-volume periods can skew MAPE disproportionately and lead to misleading conclusions about capacity requirements.
- **Human Resource Needs Based on Peak Demand:** For staffing, peak demand periods are often more critical than average levels. MAPE does not account for these peak demands, which are essential to ensuring adequate staffing during high-demand times.

The metric that could be better for this case is **Mean Absolute Scaled Error (MASE)** due to the following reasons:

- MASE is scale-invariant, making it useful across high- and low-demand periods without the issues MAPE has in low-volume situations.
- By scaling against average demand from a baseline period, MASE provides a clearer picture of prediction error across fluctuating demands, making it helpful for capturing both total and peak periods—key for both fleet and human resources needs.

2.3

Given that $\Delta Y =$ (first-differenced series) is weakly stationary and can be represented as:

$$\Delta Y = \mu + \mathcal{N}(0, \sigma) \tag{1.1}$$

where σ is known and μ is an unknown constant, we are tasked with testing if μ differs between the pre-COVID (before December 2019) and post-COVID (after January 2022) periods.
We can use a **Two-Sample t-test** for comparing means of μ across the two periods. Given σ is known and assuming normal distribution of ΔY, the two-sample t-test will allow us to test if the mean demand (represented by μ) significantly changed between the pre- and post-COVID periods.