CS 215
Assignment 4

Due Date: 26^{th} October, 2024

Instructions

1. You should type out a report containing all the answers to the written problems in Word (with the equation editor) or using LaTeX, or write it neatly on paper and scan it. In either case, prepare a single pdf file.
2. The report should contain names and roll numbers of all group members on the first page as a header.
3. Once the pdf and code are ready, `zip` them into one folder with the name `N` = `A3-FirstStudentRollNumber-SecondStudentRollNumber-ThirdStudentRollNumber.zip` Please name the zip file `N.zip` You may use the command `zip -r N.zip N` from the directory containing your submission directory `N` to do this. If the assignment was done in a group of size not equal to 3, the name of the folder and zip file will be `A3` followed by the roll numbers of each member; each item separated by hyphens. Any letters in the roll number must be capitalized. The submissions for this assignment **MUST** follow the given directory structure and assignments not adhering to this risk not being graded.
4. Upload the file on moodle BEFORE 11:59 pm on the due date. No assignments will be accepted thereafter.
5. Note that only one student per group should upload their work on moodle, though all group members will receive grades.
6. If using any LLM based tool to generate code make sure to understand and verify it, as some parts of the assignment might be auto graded and errors while running the code will lead to zero in that question. Make sure your code does not call obscure libraries beyond common ones like (numpy, pandas, sklearn, sktime, statsmodels, pytorch, gluonTS, prophet, u8darts or others shared in the collab link in class slides)
7. For all questions: plots, answers and evaluation metrics go into the report. The code part for the that question goes into Q{ques no}{part no}.ipynb e.g. Q1a.ipynb
8. Please preserve a copy of all your work until the end of the semester.

Question 1. Parking Lot Problem. Download the parkingLot.csv file from kaggle link. This file has over 9 weeks data (mock not actual) which comes from two camera feeds each recording from either entrance or exit of the parking lot of a mall. The OCR on the camera sensor records number plates of exiting or entering vehicles and uploads it to a central database(in burst of every 20 min) in the following format.
camera_id (one of 001(at entry) or 002(at exit)),
vehicle_no (string eg "MH24E0367"),

timestamp (pd.timestamp),
Since the data is uploaded async by both feeds it won't be in a strict chronological order. There are some other systemic issues too: each camera receives scheduled update/maintenance once a week when it just updates null for every vehicle it sees for 20 mins. The mall remains closed from 12 am to 5 am everyday. Demand increases slightly on weekends/holidays. Sometimes a sensor might replace one of the characters with another in the number plate. Based on these assumptions. Clean the dataset/ along with any additional data engineering needed for the following forecasting tasks, You are only allowed to use either ARIMA/ETS and linear regression. Split into train/test set. Report both the MASE (Mean absolute scaled error) and MAPE(Mean absolute percentage error) scores for both the tasks. [**7 + 7 + 6**]

- **Part a.** Forecast total number of vehicles entering the parking per day, for next 7 days.
- **Part b.** Forecast avg time spent in the mall by a vehicle entering on a particular day, for the next 7 days.
- **Part c.** For each of the above two parts experiment with at least 2 other outlier smoothing or missing value imputation strategies as preprocessing steps before being fed into the model. Check this link. Report the scores just like before.

Question 2. Forecasting on a Real World Dataset.

Enter the kaggle competition with your roll-no as name for more details. Max number of submissions are limited to 4 per day/team. This dataset provides monthly operational metrics for a major Indian airline from 2013. It includes information on the number of departures, flight hours, distance flown, passenger traffic, available seat kilometers, freight carried, and mail carried. You can use any popular time series library for this task. (Sktime, u8darts, GluonTS, etc..) Do the necessary Feature Engineering/Data Cleanup required to create time-series in your choice of library. [**12 + 4 + 4**]

1. You need to predict 'PASSENGERS CARRIED' from 2023 Sep to 2024 Aug.
   - Part(a): Only Models discussed in class or its variant are allowed for submission for the competition leaderboard. Try to beat the benchmark submission.(Naive Drift)
   - Part(b): Use an LLM to generate predictions see section 3 of LLMTIME for some ideas on tokenisation of input, handling null values. Report only your best working prompting strategy and their evaluations.[You are not expected to train/finetune an LLM nor use multimodal LLM i.e you input pdf or screenshot, your input to the LLM should be `PROMPT` + string generated from timeseries dataframe ideally by a script]
   - Part(c): Train a Global Model say prophet or any other. These models are fit once on related time series and used later for prediction, instead of fit being called for every time series separately. You can use all of the other columns present in the dataset to create related Time-series. Again Only report evaluations by this method in the report.

In Report

2. The leadership at this company wants this demand forecasts for planning next quarter requirements for fleet management and to hire human resources for the next quarter. Can you argue why MAPE [Mean Absolute percentage error] may not be a good metric to evaluate the forecasts. Suggest a metric that might be better for this case.[Hint: Fleet requirement are usually constrained by the total no of passengers expected, while human resources requirement will be usually constrained by peak demand expected]
3. For this part assume that the ΔY represents the first differenced series for the above and it is weakly stationary and can be modelled as $\Delta Y = \mu + \mathcal{N}(0, \sigma)$ with σ known and μ is an unknown constant. What kind of test will you perform to test if μ was different precovid (data before DEC 2019) and after (data after JAN 2022). No need to test, just report the test that you'll use.