QBUS2820 Assignment 2: Queensland public transport

Overview

The assignment consists of forecasting the number of passengers passing through several public transport stations across Queensland.

You will create and validate a methodology that considers the forecasts models seen in class, compute point, probabilistic forecasts and estimate the errors of the model.

Context and Data

Each time series measures the number of passengers that finished their trip at a given destination station, every month (a monthly time series). This way, we can get an estimate of the behavior of Queensland's public transport system through the years, the impact of COVID19, etc.

At this level of resolution, forecasts are interesting for getting a sense of 'what is going on' in the public transport system. We could identify potential areas that are on the rise (or declining) for real state investing (not financial advice), which parts of the public system would need to be downsized or upsized, hiring staff, etc. Probabilistic forecasts might give us a more complete picture and help us identify potential extreme events such as identifying potential stations that are at risk of exceeding their current maximum operating capacity.

Realisticaly, much more information is available, for example, not only destination but pairs of (origin, destination) stations would identify specific trips (we do not include those in this assignment). The dataset might include bus, train, even ferry trips.

The dataset comes in a csv format, with columns:

- `month`: Date in year-month format.
- `destination_stop:` number with the 'id' of the station. If you are interested in recovering the actual name of the stations, the following table can be consulted: https://www.data.qld.gov.au/dataset/go-card-transaction-data/resource/9ca69d04-0629-45f3-811a-11f1fc6b2ee9
- `tot`: The number of passengers counted for that month.

From this data, you would create several time series, one for each station, and then forecast them.

The original source can be found on: https://www.data.qld.gov.au/dataset/go-card-transaction-data

The forecast problem

The 'forecasts' to be computed and reported:

- Point forecasts: You will forecast the number of passengers for the next 12 months.
- Probabilistic forecasts: For each time series, the 95% quantile across the 12 forecasted months, kind of the extreme 'load'

- Performance: Estimate of the mean absolute error of the predictions for each time series.
- Aggregate: Using the individual time series in the dataset as a 'sample' from the population of all stations in the Queensland public transport, forecast the size of the total transport system, comparing the sum of the trips across the last 12 months in the dataset (Jun 2021-July 2022) to the sum of the forecasts, an report it in percentage change (forecasted_year / last_observed_year)

It is important that you create a complete methodology that compares several of the models covered in the unit (naïve, seasonal naïve, exponential smoothing family, ARIMA family) and identifies the best one using a proper validation mechanism (a version temporal crossvalidation). The methodology should be applied to each time series. Because of the number of time series in the dataset, you are not required to manually apply the methodology to each time series, you can use a few time series to design the methodology and then apply it automatically to all time series in the set.

What you need to submit

- A **notebook (.ipynb file)** that runs the methodology, creates the forecasts and documents the decisions and results along the way.
    - Divide it into sections and document clearly what you are doing using markdown cells before the code of each section. You can use one (or more) cells for the methodological discussion, this should be separated from the cells that clarify technical (programming parts). Failure to explain/Incompete sections of the notebook might lead to strong penalties for those sections (this is, do not just have 'code').
    - The filename of the notebooks should be 'STUDENTID_ ASG2_nbook_QBUS2820.ipynb'
    - No pdf document is needed, make sure that the notebook is as clear as possible, you can find many examples online that interweave code and analysis.
- The notebook must have a **Results** section at the end where you will report the forecast items (the four points in the forecast problem section of this document), using plots for the point and probabilistic forecast and text output for the performance and aggregate change. **You will also report on which model was used to point forecast each series**, to get a rough idea of the dynamics of the series (is it seasonal? Does it have a trend?, etc.)

Marking

Percent of the total grade that is dedicated to each part of the assignment.

- Visual aspect of the notebook (10%). Proper sectioning, plots, text and code cells structure, etc.
- Methodology: Point forecast (40%), Probabilistic(25%), Estimate the error (15%), Aggregate (10%).
- Other errors might penalize the final grade, e.g. Notebook does not run, the format of submission is not correct, etc.