Big Data Analytics - Systems and Methods

Portfolioprüfung Part 2: Implementation

Prof. Dr. Benjamin Buchwitz

Task Description

The exercise for the implementation is to generate a forecast for a collection of time series or a single time series. The dataset is either your own proposed time series dataset or the common dataset below.

The specific task is to choose and recommend a model or forecasting process that fits to your data *in the best possible way* and evidently generates suitable forecasts for a practically relevant horizon. Try all applicable forecasting methods you know including simple benchmarks. Do not forget to discuss the forecasting methods, examine their residuals and forecasting errors and present evidence that outlines why one model should be favored over another. The following bullet points aide in guiding you through the analysis:

- Your work should start with an introduction to the topic and answer the following questions:
  - Why are the topic and data set - you work with - of interest? Why is it important to make forecasts in this domain?
  - What is the exact forecasting problem and what distinguishes a good from a bad forecast (evaluation metrics)?
  - What is the aim of your work?
  - What is already known (from literature or practice)? From this and your forecasting goals infer a hypothesis. What do you assume and expect concerning the models behavior and their forecasts?
  - Give a *very brief* overview of the steps you will take in the following presentation/paper/poster.
- Describe the dataset including relevant variables and time series' you will examine:
  - Which variables are of interest? Describe at least the y and t variable.
  - Visualize the time series. What is the range of values? What is the granularity? Are the observations equidistant? How many observations exist? Are there any missing values in the time series (how many)? Are there any further peculiarities?
  - How is the time series characterized? Trend? Seasonal patterns? Cycles?
- Then, try different forecasting methods and examine the results. Do not forget to split your data set in a training and test set before evaluating the forecasts. The chosen split needs to be justified by proper argumentation (especially if you are not using time series cross validation).
  - In order to select an *appropriate* forecasting method start with the simple forecasting methods: Is one of the simple forecasting methods already appropriate? Why or why not? What is the benefit of more complex methods and how do you use them to your advantage? Provide empirical and theoretical arguments. Plot all values of interest to your argumentation. Always examine the results and consider a plethora of evaluation measures (eg. RMSE, MAE, MAPE and MASE).
- In the end of your presentation/paper/poster, give a conclusion.

Remarks

Depending on your specific dataset not all of the bullets above may be fully appropriate. So it is up to you which points are the most relevant to your data and support your argumentation in the best way possible. Generally there are two possible paths.

- **Single Time Series:** And in depth analysis and forecasting a single time series is usually done when the forecast are of *higher value*. This means your work should go into all the details of the series and in fine tuning of the models. This is especially true if the series of interest is (relatively) short. So, when dealing with a single series the analysis and the corresponding work can feature every little aspect in great detail.
- **Collection of Time Series:** When working with a collection of time series each time series in the collection and therefore each forecast is of *lower value*. It is also unlikely that one model fits all time series as they usually exhibit some differences that can only be leveraged by specialized models. So when dealing with a collection of time series the emphasis is on the process of model selection rather than on the peculiarities of a single model. Results and evaluation measures are usually highly aggregated, plots contain multiple series/forecasts/measures and the whole analysis is supported by adequate (representable) examples.

General Dataset

If you have not proposed a dataset or your proposed dataset is not suitable for the domain of the course, you must use the following dataset:

German Petrol Prices https://dev.azure.com/tankerkoenig/_git/tankerkoenig-data

To generate a subset of time series please follow these steps:

1) Download the dataset
2) Choose a selection criteria for a subset of petrol stations (eg. all stations in a specific region or a specific brand)
3) Convert the event based observations into a time series and choose a suitable granularity (eg. hourly, daily, ... the resulting series should exhibit substantial movement)
4) Aggregate the generate time series into a small subset by choosing an aggreation level (eg. type of petrol) and an aggregation measure (eg. arithmetic mean)
5) Use the result small collection for the case study

Formal Criteria

Each case study has to be solved **individually** and **autonomously**. Your results must be submitted via the link in the Moodle course and must follow the specified format. The submitted **zip**-file (*#######.zip*) must contain at least the following:

- The **raw** data including a comprehensive description of the dataset and the variables used for the case study including all code and functions that were used to generate the time serie(s) for the actual forecasting task.
- The complete code for all analyses including descriptive statistics, model selection and estimation as well as derived evaluation results.
- Proper documentation for the complete code and all steps that guide the reader through the flow of your analyses.
- The poster in the final **pdf**-format as well as its editable source files.
- A signed declaration of authorship (Ehrenwörtliche Erklärung).