Master’s Degree in Data Science – A.A. 2023/2024
Department of Computer Science
University of Verona

Projects for Statistical Models for Data Science – Second Module

Statistical Models for Data Science

Ilaria Boscolo Galazzo

Preliminary Information

- **Final Examination:** Theory (Oral exam) + Project (Python, code + comments) on real scenarios. A discussion of the chosen project will be done, the same day of the theory part or in a separate session. During the exam, you will briefly illustrate what you have done, the pipeline you have devised and implemented, discuss and comments the main findings you have achieved.

- In the next few slides, a list of mini-projects is provided. As you will see, the general steps to be applied are highly similar in all cases, only the data type and the research question are different. As we have seen during the different lessons, it is essential to first have a good understanding of our time series, their patterns and characteristics, and then attempt to build any appropriate model and produce sensible forecasts.

Project 1 – Weather Forecasting

Brief description of the dataset

This dataset provides training data on weather from 1st January 2013 to 31st December 2016. An additional testing set is available, enclosing the period 1st January 2017 to 31st March 2017, which corresponds to the forecast interval. There are 4 parameters describing weather, that are *meantemp, humidity, wind_speed,* and *meanpressure*.

Project steps

The main aim of this project is to perform *weather forecasting* for the period January-March 2017.

Following what we have seen during the lessons, a comprehensive pipeline should be devised, including:

1. Loading, converting and cleaning of the data;
2. Exploring the dataset with descriptive statistics and frequency analysis. Use appropriate graphs to visualise the data at hand;
3. Describing the time series patterns (visually and numerically);
4. Focusing on *meantemp* (temperature as dependent variable), build the most appropriate model(s) to forecast the data for the specified period. In particular, use as test data those contained in the corresponding csv file. Evaluate the model performance using different metrics. Visualise the results with appropriate graphs.

For the different steps, comment on the main results and any relevant observation/finding you have noticed.

Project 2 – Wind Power Generation forecasting

Brief description of the dataset

This dataset contains daily measurements of temperature, wind capacity and wind generation (i.e., wind production) from 2017 to 2019 in Germany.

Project steps

The main aim of this project is to forecast the *wind power generation* for the last month of 2019 (December). Following what we have seen during the lessons, a comprehensive pipeline should be devised, including:

1. Loading, converting and cleaning of the data;
2. Exploring the dataset with descriptive statistics and frequency analysis. Use appropriate graphs to visualise the data at hand;
3. Describe the time series patterns (visually and numerically);
4. Choosing the appropriate model(s) to forecast the required data for the specified period. Evaluate the model performance using different metrics. Visualise the results with appropriate graphs.

For the different steps, comment on the main results and any relevant observation/finding you have noticed.

Project 3 – Property Price Forecasting

Brief description of the dataset

This dataset provides information about the mean home prices across a number of US states and housing types collected in the period from 30th April 1996 to 31st December 2017. The data are available for 44 states and different house types, in particular 1/2/3/4/5+ bedrooms and single family residence.

Project steps

The main aim of this project is to perform *house price forecasting* for the last two years (2016/2017), whose data should be used as testing set.

Following what we have seen during the lessons, a comprehensive pipeline should be devised, including:

1. Loading, converting and cleaning of the data;
2. Exploring the dataset with descriptive statistics and frequency analysis. Use appropriate graphs to visualise the data at hand;
3. Focusing on different US states and housing types (choose the ones you prefer, more than one), describing the time series patterns (visually and numerically);
4. Choosing the appropriate model(s) to forecast the data for the specified period for each selected state/house type. Evaluate the model performance using different metrics. Visualise the results with appropriate graphs.

For the different steps, comment on the main results and any relevant observation/finding you have noticed.

Project 4 – Pharma Sales Forecasting

Brief description of the dataset

This dataset contains information on the quantity of pharmaceutical drugs sold in 6 years (period: 2014-September 2019, daily data). In particular, drugs were classified into eight different categories following the Anatomical Therapeutic Chemical (ATC) Classification System, ranging from anti-inflammatory and antirheumatic products to drugs for obstructive airway disease and antihistamines for systemic use.

Project steps

The main aim of this project is to forecast *the quantity of drugs sold for each of the eight categories* in the May-September 2019 period (5 months). Following what we have seen during the lessons, a comprehensive pipeline should be devised, including:

1. Loading, converting and cleaning of the data;
2. Moving from daily to weekly frequency, calculate for each category the total drugs sold per week. This will become the variable to forecast, for each of the eight categories;
3. Exploring the dataset with descriptive statistics and frequency analysis. Use appropriate graphs for the visualization;
4. Describing the time series patterns (visually and numerically) of the different variables;
5. Choosing the appropriate model(s) to forecast the total drugs sold per week for the specified period, for each category. Evaluate the model performance using different metrics. Visualise the results with appropriate graphs.

For the different steps, comment on the main results and any relevant observation/finding you have noticed.

Project 5 – GDP Forecasting

Brief description of the dataset

This dataset contains estimates of total GDP (gross domestic product) and its components for several countries over the period 1970-2020. This can inform on how these measures have changed over time, and allows to investigate the usefulness of GDP as a measure of wellbeing. These information are available for 220 countries and 17 different indicators were derived.

Project steps

The main aim of this project is to forecast the *total GDP* for the last 10 years (2010-2020), for five different states of your choice. Following what we have seen during the lessons, a comprehensive pipeline should be devised, including:

1. Loading, converting and cleaning of the data;
2. Focusing on a subset of variables besides GDP (such as exports, imports, manufactoring, gross capital), explore the selected dataset with descriptive statistics and frequency analysis. Use appropriate graphs to visualise the data at hand,
3. Describing the time series patterns (visually and numerically) of the selected variables;
4. Choosing the appropriate model(s) to forecast the total GDP data for the specified period. Evaluate the model performance using different metrics. Visualise the results with appropriate graphs.

For the different steps, comment on the main results and any relevant observation/finding you have noticed.

Project 6 – Predictive Maintenance

Brief description of the dataset

This dataset contains different information related to predictive maintenance, the main one being telemetry time series data collected from sensors and representing the hourly average of voltage, rotation, pressure, vibration collected from 100 machines (2015). Information about failure history and machine features are also available.

Project steps

The main aim of this project is to build a model for *predictive maintenance*. Following what we have seen during the lessons, a comprehensive pipeline should be devised, including:

1. Loading, converting and cleaning of the data;
2. Focusing on a few machines, explore the selected dataset with descriptive statistics and frequency analysis. Use appropriate graphs to visualise the data at hand,
3. Describing the time series patterns (visually and numerically) of the selected variables;
4. For each machine you have selected, verify when any failure occurred during 2015. Then, filter the telemetry data and select a time window encompassing 10 days before and after the failure occured to observe any abnormalities. Use this as training set to then define the most appropriate variable and model(s) to predict a future failure according to the failure records (window of +/-10 days as for training). Evaluate the model performance using different metrics. Visualise the results with appropriate graphs.

For the different steps, comment on the main results and any relevant observation/finding you have noticed.

Project 7 – Item Sales Forecasting

Brief description of the dataset

This dataset contains 5-year store-item sales data (period: 1st January 2013-31st December 2017). These data have been collected for 50 different products at 10 different stores.

Project steps

The main aim of this project is to forecast the *total amount of items sold in each store* in the last two months of 2017 (November/December). Following what we have seen during the lessons, a comprehensive pipeline should be devised, including:

1. Loading, converting and cleaning of the data;
2. Exploring the dataset with descriptive statistics and frequency analysis. Use appropriate graphs to visualise the data at hand;
3. Considering the number of total items sold in each store, describe the time series patterns (visually and numerically);
4. Choosing the appropriate model(s) to forecast the number of total items sold in each of the 10 stores for the specified period. Evaluate the model performance using different metrics. Visualise the results with appropriate graphs.

For the different steps, comment on the main results and any relevant observation/finding you have noticed.

Extra - Build your own project!

In this last scenario, you are free to choose one of the following general topics (or even define yourself the main focus), find the most appropriate dataset and perform all the necessary steps to solve the problem at hand.
Some suggested topics are listed below, but the possibilities are not limited to this list:

1. Energy consumption forecasting on individual household in North Italy;
2. Forecasting the stock price of a given company stock using different temporal spans;
3. Petrol price/consumption/production forecasting using data from the last N years;
4. Tourism forecasting in a given country using different time frequencies (e.g., monthly or quarterly data);
5. Web traffic time series forecasting for different web pages and articles;
6. Forecasting COVID-19 new cases/death in different countries.