Guide to Assignment 2: Temperature Prediction

FEM11149 - Introduction to Data Science

Instructor: dr. A. Tetereva
TAs: L. de Wit and M. Praum

October, 2025

Introduction

This is a guide for the second group assignment of the course Introduction to Data Science in the Data Science and Marketing Analytics Master program at the Erasmus School of Economics.Not all steps here have to appear in your final report. However, they are useful for you to learn and use in your life as data scientist whenever dealing with a dataset.

You should **not** use this file structure for your final report. Instead, you should organize your answers in the format of a business report, following the instructions from Canvas. This means that your team should **not** organize the document as 'question' followed by 'answer' ! Instead, make a report and include the answers from the above questions in your text.

The grading for this group assignment follows the grading policy specified on Canvas (https://canvas.eur.nl/courses/50901/pages/grading-policy-for-group-assignments-and-presentations). Write after each name the percentage of the grade points that should be awarded to the person. If you have four team members who worked equally, write (25%) after each name. Check Canvas for more details and another example.

This document is structured as follows:

1. Motivation for the assignment
2. The task

Note: The dataset is different for each group. Hence, different groups will have different results.

Motivation for the assignment

You are part of the data science team of a company in the energy sector. The demand for energy is highly dependent on the weather, and to better anticipate rises in the demand for energy and avoid blackouts, your team is given the task to predict extreme weather conditions one day ahead. The starting point for this project is some general weather data that can easily be accessed from CERRA. The data is pre-processed and provided to you in the file **a2_data_group_x.csv**. The data contains 17 weather variables related mostly to temperature, wind, and precipitation, all with daily observation frequency. Each group is given 12 years of data for one of five major European cities. In particular, the variable **Location ID** indicates whether you observe data from Paris (0), Rome (1), Amsterdam (2)[1], London (3), or Madrid (4).

[1]The most beautiful out of the five.

The task

As mentioned, our interest lies in extreme weather conditions, and for this assignment, we focus on the maximum temperature observed during the day.[2] Since we are going to be making one day-ahead predictions, the first step of the assignment will be a pre-processing step.

1. Construct your matrix of independent variables by considering all observations of all variables, except those in the last row, and construct your vector of the dependent variable by considering all observations of the maximum temperature, except the first one in the sample.
2. As an initial exploratory analysis, plot all independent variables against the dependent variable and compute the correlations. Discuss your findings. Note that you will most likely have too many figures to include in the main text, so some of the figures can go in the appendix.
3. Before we can actually start doing the interesting work, there is first one more essential step. Make a train-test split of your data. Note that you are using time series data!
4. Generate principal components on your training data. Only consider the independent variables here.
5. Someone in your team asks you why you are using PCA on this kind of data. We are interested in extreme weather conditions, but is it not true that PCA does not work particularly well for data with extremes? Discuss this concern.
6. Determine the number of principal components that you will use to proceed with your analysis. You should have learned about various ways of doing this, so please report the results from at least two methods. Do they agree on the number of principal components to select? Can you explain why?
7. Construct (a) biplot(s) and compute the loadings. Can you interpret the principal components? Discuss why it is perhaps not trivial to interpret some of the principal components.
8. Decide which variables are best explained by the principal components. In particular:
    - **Odd-numbered groups** investigate what is the total proportion of variance explained by the selected number of principal components. Furthermore, you should construct bootstrap confidence intervals for the total variance explained by the selected number of principal components. Also, which variables are best explained by the selected number of principal components?
    - **Even-numbered groups** investigate what is the total proportion of variance explained by the first principal component. Furthermore, you should construct bootstrap confidence interval for the total variance explained by the first principal component. Also, which variables are best explained by the first principal component?
9. Fit the principal component regression on the training data. The easiest way to do this is to consider the pcr function of the pls package.
10. Also fit a benchmark multiple linear regression model on the training data.
11. Obtain predictions for the testing data using both models and compare their performances using an appropriate metric. What do you find? Did you expect this?
12. In order to assess the sensitivity of the principal component regression to the selected number of principal components, perform training and prediction of two more principal component regression models, but with one fewer, and one more principal component than you used before. Discuss your results.
13. You want to critically assess your prediction results. Hence, from the testing data, group the observations of the dependent variable into ten roughly equal groups depending on how high the maximum temperature was on that day. That is, the first group contains the 10% of observations in the testing data where the observed maximum temperatures were the lowest. The second group contains the 10% of observations where the observed maximum temperatures were slightly higher. This continues until the tenth group, which contains the 10% of largest observed maximum temperatures in the testing data. Plot the metric that you used before for the different methods and the different groups. See the example below:[3]
14. Write a conclusion on your overall findings in this assignment. You are free to discuss what you find most interesting, but you can think about answering (some of) the following questions:

[2] Note that all procedures could be generalized to the minimum temperature observed during the day as well.
[3] Do not forget to also discuss the figure that you obtain, of course.

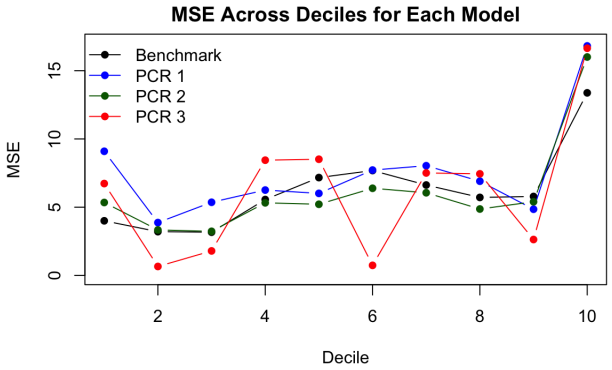

Figure 1: **DO NOT COMPARE YOUR RESULTS TO THIS FIGURE** because it merely plots randomly simulated noise. It is just to clarify what kind of figure you should make here.

- Do you think PCA is a useful tool for reducing the dimensionality in this particular dataset?
- Is the predictive performance of your methods very dependent on how high the maximum temperature was on a given day? Why do you think this is? And if so, what could you potentially do to reduce this problem?
- How do you think the penalized regression methods from the first group assignment would have performed here?