Fachbereich AIHI
Universität Salzburg
Ferndinand Ferber, BSc., Univ.-Prof. Dr. Wolfgang Trutschnig

Salzburg, June 2024
SS 23/24

Exercise Sheet 11 @'Applied AI Using R'

This week we complete the hackathon from last week - the one of you whose forecasts is the best (in terms of mean absolute error MAE) on 5 ATMs scores an extra 5 points and qualifies for a (paid) IDA-Lab internship (if desired).
Furthermore we will work with missing data and outliers.

Exercise 45 (ATM forecasting - hackathon)**.**
Download new ATM data via `load(url("http://trutschnig.net/ATM5.RData"))`. The data contains daily withdrawn amounts at five (new) ATMs in the period 2007-01-01 till 2009-02-01. Your main task is to forecast the daily withdrawn amounts at each of the 5 ATMs for each of the 7 days from 2009-02-02 to 2009-02-08. Proceed as follows:

1. Use ggplot2 to produce a quick plot of the 5 timeseries and save them in a pdf.
2. Train you top model from last week (or any other model you consider good) for each of the new ATMs and calculate the forecasts. Notice that you need to forecast 1 day ahead, 2 days ahead, ..., 7 days ahead.
3. Add the forecasts to the plots produced before and save them in a pdf.
4. Collect the forecasts in a dataframe containing the date in the first, the ATM number in the second, the forecast for the day in the third column, your name in the last column and export the dataframe.
5. Bring the dataframe to the next class.

Exercise 46.
In this exercise you are asked to impute missing data in a dataset related to the red and white wine variants of the Portuguese "Vinho Verde" grape. For a detailed description of the data see (Cortez et al. 2009) as linked below.

1. Download the dataset from `https://ufile.io/cnay2lk5` and parse it into `R`.
2. Install the `{misty}` package and answer the following questions about the `misty::na.test()` function:
    1. What is the null hypothesis of the test?
    2. Run the hypothesis test on the wine data. What is the p-value?
    3. How do you interpret the result of the test?
3. Create the plots from slides 13 to 15. What do you observe? Do you think the missingness is MCAR, MAR, or MNAR?

4. Propose a way to handle this type of missingness (CCA/imputation/...?) and do it in `R`.

[1] Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." *Decision Support Systems* 47 (4): 547–53. `https://doi.org/10.1016/j.dss.2009.05.016`.

Exercise 47.
In this exercise we perform an outlier detection on a dataset regarding credit card fraud. The feature `time` contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature `Amount` is the transaction amount and the feature `Class` is the response variable (1 = fraud and 0 = normal). All other features are the result of a PCA-transformation of the original features due to privacy issues. If solving the tasks takes too much computation time, use a subset of the data.

1. Download the dataset from `https://ufile.io/bo6osyla` and parse it into `R`.
2. Remove the `Class` column.
3. Use reconstruction-based outlier detection. Does this method identify the fraudulent transactions as outliers? Bonus if you try out other encoder/decoders.
4. Use model-based outlier detection. Does this method identify the fraudulent transactions as outliers? Bonus if you try different columns as a target for the models.
5. Use density-based outlier detection. Does this method identify the fraudulent transactions as outliers? Bonus if you try different values for k.