Management Summary (Aberle, Kleinhenz)

Introduction

The management summary is about the third assignment of the course Data Analytics. Participants of the following assignment are Esther Kleinhenz and Thomas Aberle. The Kick-Off happened at the 23th of January and was finished with finalising the management summary at the first of February.
The assignment was done to practice procedures by using classifiers and proof that training a dataset can value different informations in an appropriate way.

The following short overview shows how the assignment was compiled:

1. First steps to understand the dataset and join it to one common dataset
2. Create some features to get to know the needed procedure
3. Split data into test and training set
4. Segmenting the data into single and non-single households and completing the five-fold cross validation
5. Calculating the average of accuracy
6. Creating a guess and a biased guess
7. Putting all the result into one barplot

Course of action

The features were chosen by some suggestions found in the slides fo the lecture (Feature Selection). Features like:
night to noon consumption,
highest consumption,
mean consumption,
sum of the consumption,
dinnertime consumption,
lowest consumption,
lunchtime consumption,
morning consumption,
afternoon consumption,
evening consumption,
weekday consumption,
weekend consumption,
dinner to lunch ratio,
min to max ratio and
morning to evening ratio.

Training and test set were put into relation, as recommended in the lecture (80% to 20%) and were randomly ordered by “setting a seed“.

To train a model, glm.fit was used to keep it as simple as possible and all the features were put into the regression model.

Calculating the accuracy of the guessed values showed an appropriate probability, so the five-fold cross validation was performed. In-between the different lists and matrixes were printed and determined as reasonable.

The accuracies were put into a vector to calculate the average accuracy and the standard deviation to compute the occurring error. A barplot was generated to visualise the current status.

Biased and random guess are realised with a counter and loops which are iterating through the “single“ row and taking samples. After compiling these guesses the average accuracy can be calculated and the output was issued and rated as appropriate.

The interpretations are added to the code as comments.