Problem Set 2: Feature and Target Engineering

Course responsible: Ana Alina Tudoran, anat@econ.au.dk
Tutorial responsible: Camille Pedersen, cpe@econ.au.dk
Administrator: Gitte Isager, gi@econ.au.dk
Version 1: 14.09. 2023

Note: Before proceeding with this tutorial, it is recommended the student is familiarized with Hands on Machine Learning, Chapter 3 (aprox. 1 h)

Learning objectives

The goal of this tutorial is to revisit the ideas covered in the lecture on Feature Engineering and apply them in the R environment.

Data

The dataset includes features such as make, model, year, engine, and other properties of different cars. The data was originally used to predict the price of the car. Acknowledgments to Edmunds and Twitter for providing the data. Variables labels and description:

- "Make": the brand of the car (factor)
- "Year": the year the car was produced (factor)
- "Engine.Fuel.Type": the fuel type of the car (factor)
- "Engine.HP": the engine horsepower of the car (integer)
- "Engine.Cylinders": the engine capacity in cylinders of the car (integer)
- "Transmission.Type": the car transmission type of the car (factor)
- "Driven_Wheels": the type of driven wheels of the car (factor)
- "Number.of.Doors": the number of doors of the car (integer)
- "Market.Category": the market category of the car (factor)
- "Vehicle.Size": the car size (factor)
- "Vehicle.Style": the car style (factor)
- "highway.MPG": highway miles per gallon of the car (integer)
- "city.mpg": city miles per gallon of the car (integer)
- "Popularity": popularity of the car (integer)
- "MSRP": Manufacturer Suggested Retail Price. MSRP is the price that a product's manufacturer recommends it be sold for at the point of sale. Any retail product can have an MSRP, but the term is frequently used with automobiles. Currency is unknown (integer)

Problem 1

1. Upload the data in RStudio and familiarize yourself with the variables and their meaning.
2. Check the variable type (e.g., factor, integer, numeric, etc.) and adapt it to the variable described in the text.
3. Delete the variable "Model" from the dataset and reflect on the reason for deletion
4. Set a seed and randomly partition the data into training and test set (70%/30%)
5. Evaluate MSRP distribution and check various transformations to normalize it. Considering the target engineering methods discussed in the course, decide what transformation will be most appropriate.
6. Evaluate the missing data. Decide which method of treating missing data will be applied.
7. Evaluate the features with zero and near-zero variance. Decide which variables will be eliminated.
8. Display the distributions of the numeric features. Decide what type of pre-processing to be implemented later.
9. Display the distributions of factor features. Decide what type of pre-processing to be implemented later.
10. Based on the data exploration above, proceed by creating a *blueprint* that prepares the data for predicting MSRP, using the ***recipes*** package. More specifically,
    - Set up a recipe including all the steps desired for data pre-processing. Reflect on the order of these steps and how they will influence the final dataset.
    - Prepare and bake the training and testing data. Finally, you should have two datasets (baked_train and baked_test) ready for analysis.
    - Display the size of the new datasets. Conclude how they changed compared to the original datasets.
11. Reflect on the possibility of developing the blueprint within each resample iteration. If time allows, implement this approach in the ***caret*** library when training a knn regression model to predict MSRP (see an example in the lecture)