Homework 3

You can submit in groups of 2.
Due 2/28, 1pm.

All assignments need to be submitted via github classroom:
https://classroom.github.com/g/H0sNqUJD
and via gradescope.

The goal of this homework is to provide a realistic setting for a machine learning task. Therefore instructions will not specify the exact steps to carry out. Instead, it is part of the assignment to identify promising features, models and preprocessing methods and apply them as appropriate.

The overall goal is to predict the fuel efficiency of car models based on historical data collected by the department of energy that can be found at
https://www.fueleconomy.gov/feg/download.shtml

The measure of fuel efficiency that you should predict is Combined Unrounded adjusted Fuel Economy("Comb Unrd Adj FE - Conventional Fuel").
Do not use any of the other measurement results, only features of the car provided by the manufacturer, to predict the efficiency. The main performance metrics is R^2 and homework grades will depend on your test-set score.
Document your process as appropriate, in particular in how and when you used the test set.
The main goal is to predict the 2018 data from the 2015-2017 data. You should also evaluate your final model for Task 1 and Task 3 making an i.i.d. assumption, i.e. splitting data at random.
For Tasks 1-3, document which features you used and why.

Task 1 Linear Models

Measure performance of a linear model using the provided features with appropriate preprocessing.

Task 2 Feature Engineering

Build non-linear features or derived features from the provided column. Try to improve the performance of a linear model with these.

Task 3 Any models

Use any regression model we discussed (trees, forests, gradient boosting, SVM) to improve your result. You can (and probably should) change your preprocessing and feature engineering to be suitable for the model. You are not required to try all of these models.

Task 4 Feature Selections

Identify features that are important for your best model. Which features are most influential, and which features could be removed without decrease in performance? Does removing irrelevant features make your model better?