Data Challenge

Please solve the data challenge collaboratively in your group. Submission deadline is March 2, 2025. Please submit your answers in this PDF and a CSV file containing your predictions.

Wholesale Manager

You manage a wholesale store. The data file `juice.csv` contains orange juice sales (`sales`) and prices (`price`) of different grocery stores that you deliver. Your product range contains three different orange juice brands: `Tropicana`, `Minute Maid`, and `Dominicks`. Some stores advertise/feature specific orange juice brands, which is indicated by the dummy variable `feat`. The data contains also the store ID (`id`).

You deliver new grocery stores. The new stores sent you the file `new_grocery.csv`, which contains the planned prices and advertisements for the different brands. Your job as wholesale manager is to predict the sales of the new grocery stores and deliver the right amount of orange juice.

1. Load the data files. Describe your data cleaning procedure. Are there missing values? Do you transform some variables?

2. How do you partition the data?

3. Asses the performance of Lasso, Ridge, Tree, and Forest estimators in the test sample. Which estimation procedure has the best prediction power? How did you select the tuning parameters? How did you transform covariates?

4. Predict the sales of the new grocery using the estimation procedure with the best prediction power. Save these predictions together with the store ID. Submit the sales predictions and the store ID in one csv-file.