INDR 450/550 HOMEWORK 4, Due Date: May 23, 2022

- Please work in groups of two or three or (individually) and submit one file for each group with all names.
- Please perform all computations in python.
- In addition to the python notebook, **submit a short typed summary report** that includes the results (error tables, prediction intervals etc.) of all exercises. Also add a general assessment of the methods (which method is the best, which should be avoided etc.). **The report is part of the overall grade.**
- The data file includes the monthly sales of Audi vehicles in Turkey as well as the values of some corresponding predictors from January 2006 for 108 consecutive months.

Exercises

1. (35 points) Regression based methods
   (a) Split to data into a training set (months 5 to 84) and a test set (months 85 to 108). Fit a least squares regression to training data with the following predictors:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 t + \beta_2 t^2 + \\ &\quad + \beta_3 x_{1t} + \beta_3 x_{2t} + ... + \beta_{14} x_{11t} \\ &\quad + \beta_{15} d_{1t} + \beta_{16} d_{2t} + \beta_{17} d_{3t} + \epsilon_t \end{aligned}$$

   where:
   x_{it} is an indicator (dummy) for month i ($i = 1, 2, , , 11$). Note that we only need to use 11 of the 12 monthly indicators in the regression and we skip month 12 d_{it} is a difference at lag i, $d_{it} = y_{t-i} - y_{t-i-1}$.
   (b) (5 points) Comment on the significance of the predictors and R^2 value. Compute the MSE and RMSE of the fit on the training data.

(c) (5 points) Compute the MSE and RMSE of the fit on the test data.

(d) (25 points) Now fit a lasso regression to shrink the full model. Experiment with different penalty parameters (referred to as 'gamma' in sklearn library) and compute the test RMSE for a few cases. Report the most significant predictors based on the lasso optimization.

2. (50 points) Tree based methods

(a) (10 points) Fit a regression tree using all predictors . Experiment with a few different values for the depth of the tree and report the train and test MSE and RMSE.

(b) (10 points) Fit a bagged regression tree using bagging and using all predictors. Report the train and test MSE and RMSE.

(c) (15 points) Fit a random forest. Experiment with different numbers of features and report the train and test MSE and RMSE.

(d) Plot the importance of the predictors for the best random forest.

(e) (15 points) Fit a boosted tree. Experiment with different depths and learning rates and report the train and test MSE and RMSE.

(f) Plot the importance of the predictors for the best boosted tree.

3. (15 points) Interpreting the results.

(a) (10 points) Complete the below results summary table. For full regression, you can note the statistically significant predictors at 5%. For lasso you can list the predictors that remain after shrinkage and for random forest and boosted tree you can list the predictors based on the Importance measure order. List also the specifications (maximum depth, maximum number of features etc.) whenever they apply.

(b) (5 points) Using the summary table, please explain which three or four predictors are the most important ones to predict the sales of Audis. Which of the methods would you recommend for predictipon for this data?

Table 1: Comparison Table

Method	**Train RMSE**	**Test RMSE**	**Predictors**	**Spec.**
Full Regression				
Lasso				
Regression Tree			-	
Bagged Tree			-	
Random Forest				
Boosted Tree				