Assignment Two: Linear Regression for prediction

Total Marks: 15 points

# The problem

You are required to predict the median price of houses in Boston by using linear regression and write a report on the process. The data set is Boston\_Housing.csv on the avenue to learn. It has 506 records with 14 variables. The goal is to build a linear regression model to predict the median property value. Use the Jupyter notebook from week 4 (updated) posted on the avenue to learn. The objective of this assignment is to demonstrate that:

* You know how to create a Jupyter notebook from scratch.
* You have an adequate understanding of concepts related to predictive model building using linear regression.
* You know what predictive linear regression’s application in business is.

# The data dictionary

Here is the description of variables in Boston Housing Dataset

|  |  |
| --- | --- |
| Name | Description |
| CRIM | Crime rate |
| ZN | Percentage of residential land zoned for lots over 25,000 ft2 |
| INDUS | Percentage of land occupied by nonretail business |
| CHAS | Does tract bound Charles River (= Yes *if tract bounds river*, = No *otherwise*) |
| NOX | Nitric oxide concentration (parts per 10 million) |
| RM | The average number of rooms per dwelling |
| AGE | Percentage of owner-occupied units built prior to 1940 |
| DIS | Weighted distances to five Boston employment centers |
| RAD | Index of accessibility to radial highways |
| TAX | Full-value property tax rate per $10,000 |
| PTRATIO | Pupil-to-teacher ratio by town |
| LSTAT | Percentage of the lower status of the population |
| MEDV | Median value of owner-occupied homes in $10,000s |
| CAT.MEDV | Is the median value of owner-occupied homes in tract above $30,000 (CAT.MEDV = 1) or not (CAT.MEDV = 0) |

# Tasks

## **25% for:** Data visualization for initial variable investigation and selection.

* **10% for:** What are the distributions for the dependent and independent variables?
  + Create histograms (use different colors for different variables) and boxplots.
  + Run a summary statistic for variables (predictors and outcome) and write a few lines about the distribution, outliers, … for each graph.
* **10% for:** How do variables relate to each other?
  + Create scatterplots / a scatterplot matrix.
  + Use the categorical variable for the colors in your scatterplot matrix and report what it shows.
* **5% for:** What are three or four most important characteristics of the neighbourhood for predicting the median value of houses in Boston?
  + Look for patterns in the scatterplots.

## **10% for:** Data preparation and partition for training, validation, and testing.

* **5% for:** Convert the categorical variable to dummy variable(s).
  + How many dummy variables do you need for the categorical variable?
  + What is the difference between choosing *p* dummy variables and *p-1* dummy variables for a categorical variable with *p* categories?
* **5% for:** Create training (70%) and validation (30%) partitions.
  + Why do we need training and validation/test partitions?
  + If you need to have both validation and test partitions in addition to the training partition, what do you need to do (How should you change the code you used for training/validation to create training/validation/test)?

## **30% for:** Run a linear regression with all available variables.

* **15% for:** What is the mathematical formula of the regression model obtained? What is the R2 of your model on the training partition? What is the error of your model (Regression statistics) for the validation partition?  
  **5% for:** You should avoid using one of the variables (features). Which feature is that? And why should you avoid using it?
* **5% for:** How do you calculate the predicted median value for housing and the prediction error (residual) for each record?
* **5% for:** Show the error distribution for training and validation? What are the differences between them?

## **30% for:** Automated variable selection.

* **10% for:** Using exhaustive, what is the adjusted R2 and AIC for the best model with five features?
* **10% for:** Using forward selection, what is the best set of variables you will use? What are the criteria used for selection?
* **10% for:** Using Statmodels, fit the “best model” with the “best set of variables” you found by forward selection.

## **5% for:** Feedback

Report how much time you have spent on this assignment, what problem you have faced and what you have learned.

## **5% bonus for:** Dummy variables

* Run the linear regression with all available variables twice:
  + First, with *p-1* dummy variable(s) for the categorical variable
  + Second, with *p* dummy variable(s) for the categorical variable

What is the difference between the two models? Why do you have different intercepts and coefficients for dummy variables?