Assignment 2

- Total Score: **20 points**
- Number of case studies: **2**
- **Number of submissions allowed on Canvas: 1 (Zip and submit your answers in a PDF together with your R file; once you are completely sure)**
- **Submission:**
  - Answer the questions given with each case in **a single document** and save it as a PDF <your-full-name>.pdf.
  - Save your R script as <your-full-name>.R. Submit your R script with the above PDF of answers to the in the Canvas folder - **Assignment 2 (submission)**
  - **Zip the above 2 files for submission as** <your-full-name>.zip
  - No email submissions allowed.
  - Due date: Refer canvas.
  - No submissions allowed after grace period of 1 week from due date (refer canvas)
- Follow the instructions as under for each case, datasets, and related analysis.
- Remember to provide brief answers to questions where explanations are required.
- Use seed(123) for your R coding.
- Provide a well-documented R code file.
  - You must document your code to help understand your work and avoid any penalty.
  - Undocumented codes will attract score penalty of up to 5 marks. Please follow best practices from the class sessions.

Case-1: Boston Housing Prices

Dataset: Boston Housing Dataset

URL:
https://raw.githubusercontent.com/selva86/datasets/master/BostonHousing.csv

Business Context: You work as a senior data analyst for "Boston Property Advisors," a real estate investment firm specializing in residential properties in the Greater Boston area. The company has been commissioned by institutional investors to develop a predictive model for housing prices that can:

- Estimate property values for potential acquisitions before conducting expensive physical appraisals.
- Identify undervalued properties by comparing model predictions with current market prices

- Understand key value drivers to guide renovation and development decisions
- Create investment strategies based on neighborhood characteristics that most strongly influence home values
- Provide data-driven recommendations to clients about which areas offer the best investment potential
- Support portfolio management by predicting ROI for different property types and locations

Your analysis will directly impact investment decisions worth millions of dollars, so accuracy, interpretability, and business insights are critical. The model must be robust enough for real-world deployment and transparent enough for client presentations.
You will need to translate statistical findings into actionable investment strategies that non-technical stakeholders can understand and implement.

Instructions

- Download and analyze the Boston Housing dataset (URL link above).
- Perform comprehensive EDA including correlation analysis and distribution checks.
- Build MLR model to predict the TARGET '**medv'** using random train/test split (70:30).
- Check model assumptions (linearity, normality, homoscedasticity, multicollinearity)
- Calculate variable importance and provide business insights.
- Graders will also look at the quality of the model you have developed.

QUESTIONS (10 marks)
[Answer these in your Answer document]

1. After performing EDA on the Boston Housing dataset, which variable shows the strongest negative correlation with the target variable 'medv' (median home value)? (1 marks)
2. Share your final fine-tuned trained model summary below and provide interpretation of the summary in bullet points including variable coefficients. (1 + 2 marks)
3. What is the difference between your train and test accuracy. What does it tell you? (1 +1 marks)
4. If you retain 'rm' in your final model, what is the coefficient for 'rm' (average rooms per dwelling). What is the correct business interpretation? (1 marks)
5. Identify the top 2 predictors of the medv price based on variable importance analysis of your final model. (1 marks)
6. What is the R-squared of your final model? Provide a business interpretation. (1 + 1 marks)

Case-2: Heart Disease Risk Prediction

Dataset: Heart Disease Dataset (Cleveland)

URL: https://raw.githubusercontent.com/rashida048/Datasets/master/Heart.csv

Target Variable: AHD (Angiographic Heart Disease: **Yes/No**)

Business Context : You are the lead data scientist at "CardioPredict Analytics," a healthcare technology company that partners with hospitals and clinics to improve early detection of cardiovascular disease. The organization has been contracted by a major healthcare network to develop a clinical decision support system that can:

- Predict heart disease risk using non-invasive clinical parameters available during routine checkups
- Assist physicians in screening decisions by flagging high-risk patients who need immediate cardiac evaluation
- Optimize resource allocation by identifying which patients should be prioritized for expensive diagnostic tests (angiography)
- Reduce healthcare costs by preventing unnecessary procedures while ensuring no high-risk patients are missed
- Support early intervention strategies that could prevent heart attacks and save lives
- Provide interpretable risk factors that doctors can communicate to patients for lifestyle modification counseling
- Integrate with electronic health records to provide real-time risk assessments during patient visits

This is a binary classification problem where the false negatives (missing diseased patients) could be life-threatening, while false positives create patient anxiety and unnecessary medical costs.

Remember that in healthcare applications:

- False Negatives (missing sick patients) are typically more costly than False Positives
- Sensitivity is often prioritized over Specificity for screening
- Model interpretability is crucial for physician adoption

Instructions

- Download and analyze the Heart Disease dataset from the provided URL
- Perform clinical EDA focusing on class distribution, predictor relationships, and medical interpretation.
- Build and validate logistic regression model to predict the TARGET '**AHD'** using 80:20 stratified train/test split.
- Evaluate model performance using confusion matrix, ROC analysis, sensitivity, specificity, etc. (threshold 0.5).
- Calculate and interpret odds ratios with clinical significance for healthcare decision-making.
- Graders will also look at the quality of the model you have developed.

QUESTIONS (10 marks)
[Answer these in your Answer document]

1. After examining the target variable 'AHD' (heart disease), what is the exact class distribution (percentages for Yes/No)? Is this a balanced or imbalanced dataset? (1 mark)
2. Share your final fine-tuned logistic regression model summary below and provide interpretation of the summary in bullet points including significant predictors and their odds ratios. (1 + 2 marks)
3. Is your final model generalizable? (2 marks)
4. If you retain 'MaxHR' in your final model, what is the odds ratio for 'MaxHR' (maximum heart rate)? What is the correct clinical interpretation? (1 mark)
5. Identify the top 2 predictors of heart disease risk based on variable importance analysis of your final model. (1 mark)
6. What is the AUC (Area Under Curve) of your final model? Provide a clinical interpretation of this performance metric. (1 + 1 marks)

Data Dictionaries

Variable	Description	Type	Range/Values
crim	Per capita crime rate by town	Continuous	0.006 - 88.976
zn	Proportion of residential land zoned for lots over 25,000 sq.ft.	Continuous	0 - 100 (%)
indus	Proportion of non-retail business acres per town	Continuous	0.46 - 27.74 (%)
chas	Charles River dummy variable (1 if tract bounds river; 0 otherwise)	Binary	0, 1
nox	Nitric oxides concentration (parts per 10 million)	Continuous	0.385 - 0.871
rm	Average number of rooms per dwelling	Continuous	3.561 - 8.780
age	Proportion of owner-occupied units built prior to 1940	Continuous	2.9 - 100 (%)
dis	Weighted distances to five Boston employment centres	Continuous	1.130 - 12.127
rad	Index of accessibility to radial highways	Discrete	1 - 24

tax	Full-value property-tax rate per $10,000	Continuous	187 - 711
ptratio	Pupil-teacher ratio by town	Continuous	12.6 - 22.0
black	$1000(Bk - 0.63)^2$ where Bk is the proportion of black population by town	Continuous	0.32 - 396.9
lstat	% lower status of the population	Continuous	1.73 - 37.97 (%)
medv	Median value of owner-occupied homes in $1000s (TARGET)	Continuous	5.0 - 50.0

Variable	Description	Type	Range/Values
Age	Age of the patient in years	Continuous	29 - 77 years
Sex	Gender (1 = male; 0 = female)	Binary	0, 1
ChestPain	Chest pain type	Categorical	typical, atypical, nonanginal, asymptomatic
RestBP	Resting blood pressure (mm Hg)	Continuous	94 - 200
Chol	Serum cholesterol in mg/dl	Continuous	126 - 564
Fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)	Binary	0, 1
RestECG	Resting electrocardiographic results	Categorical	normal, stt, lv hypertrophy
MaxHR	Maximum heart rate achieved	Continuous	71 - 202
ExAng	Exercise induced angina (1 = yes; 0 = no)	Binary	0, 1
Oldpeak	ST depression induced by exercise relative to rest	Continuous	0.0 - 6.2
Slope	The slope of the peak exercise ST segment	Categorical	upsloping, flat, downsloping
Ca	Number of major vessels (0-3) colored by fluoroscopy	Discrete	0, 1, 2, 3

Thal	Thallium stress test result	Categorical	normal, fixed, reversible
AHD	Angiographic Heart Disease (TARGET)	Binary	No, Yes