ACTL30008 Actuarial Analytics and Data I - 2025 Assignment 2

1 Data

This assignment aims at predicting the median housing price in a block group within a particular region. You are given some housing data of the target region.

The data set, which is named "Assignt2_data.csv", can be downloaded in Canvas. There are ten columns in the given data, such as the population, median income, median housing price, and so on for each block group in the given region. Each block group is a geographical unit that typically has a population of 600 to 3,000 people. There are 20,623 block groups in the given dataset that contains the following columns:

- id: the ID number of a given block group
- longitude: A measure of how far west a house is; a higher value is farther west
- latitude: A measure of how far north a house is; a higher value is farther north
- housingMedianAge: Median age of a house within a block; a lower number is a newer building
- aveRooms: Average number of rooms within a block per household, a group of people residing within a home unit
- aveBedrooms: Average number of bedrooms within a block per household

- population: Total number of people residing within a block
- medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
- medianHouseValue: Median house value for households within a block (measured in US Dollars)
- oceanProximity: Location of the house w.r.t ocean/sea

Based on the given data, you are required to complete the following tasks.

2 Tasks

2.1 Nonlinear model vs non-parametric model (17 marks)

In this part we aim to compare the prediction performances of GAM and KNN on the **medianHouseValue** variable using the given data. Ignore the censoring issue in the data in this section.

1. Discuss the pros and cons of GAM and KNN in the context of the current regression problem.
2. Train a GAM regression model and a KNN regression model (model improvements should be considered when feasible) and estimate their test MSE using the given data. Explain how you use the given data whenever needed.
3. Which model performs better and why? Justify your finding based on the bias-variance trade-off.

2.2 Classification models (19 marks)

Since the **medianHouseValue** variable was censored at 500,001, next we shall build some classifiers to predict the censoring for a given block group in the target region.

1. Generate a binary variable named **censoring** that takes value of 1 if medianHouseValue = 500,001; and 0 otherwise.

2. Suggest two classification methods that can be used to classify the **censoring** variable with reasoning.

3. Build your suggested classifiers and estimate their test error rates using the given data. Explain how you use the given data whenever needed.

4. Which classification method performs better here and why? The reasoning should be based on the bias-variance trade-off.

2.3 A hybrid approach (14 marks)

In this part we try to construct a procedure that combines the regression model and the classifier selected in 2.1 and 2.2, aiming to addressing the censoring issue in the given data.

1. Discuss the feasibility of this approach and design the main steps of it.

2. Using the given data to attempt the proposed procedure and estimate the test MSE of **medianHouseValue** predictions.

3. Does this approach predict **medianHouseValue** more accurately than the model given in 2.1.3? Justify your finding.

3 Instructions

- This assignment is due at 5pm on Sunday 1st June. Submit your files in Canvas under "Assignments".
- Generate an R markdown document covering tasks 2.1-2.3 and produce a pdf version for submission.
- Name your files by your group number. One submission is required for each group.
- This assignment counts for 15% in the total assessment of this subject.
- Mark deductions will be applied to late submissions. A 10% deduction in total marks may be applied to each day of delay.