Machine Learning
Assignment 2 - Supervised Learning (Regression)

Vikas Thammanna Gowda

Due: 04/09/2025

Instructions

Collaboration Policy

This is an individual assignment. You may discuss concepts, problem formulations, and approaches to solving the problems, but you must write your own code and explanations.

- You are **not allowed** to share solutions, source code, or exact approaches.
- Any external sources (books, online resources, discussions, etc.) you refer to must be cited in your write-up.
- You do not need to cite course lecture notes, textbooks, or materials provided as part of the course.

Assignment Structure

Your submission consists of two parts:

Coding Component (Submit as `<your_name>_PA02.ipynb`)

- Use Markdown cells in Jupyter Notebook to add each question before solving it.
- Write clean, readable, and well-commented code.
- Define functions for repetitive tasks instead of redundant code.
- Ensure all visualizations are clear, properly labeled, and provide meaningful insights.
- Use at least two different types of visualizations for each data exploration question (e.g., histogram and box plot).

Report Write-up (Submit as `<your_name>_PA02.pdf`)

- Add each question to your write-up before answering them.
- The report should mirror the coding component and provide interpretations of results.
- Use Times New Roman, size 14 for questions, size 12 for answers.
- Ensure the document is justified and structured.
- Include properly labeled figures and tables, centered with captions.
- All the plots must be complete, be of the same size, and be centered with a figure number and a figure name.
- Clearly explain decisions regarding missing data handling, feature selection, scaling, and outlier removal.

Dataset Description

You will use the dataset `Clean_Used_Car_Sales_PA02.csv` for this assignment. This dataset contains cleaned and preprocessed used car sales data for supervised learning tasks. You will build regression models on this data.

Assignment Questions: Coding vs. Write-Up

Each question involves both a coding component (implementation) and a write-up component (interpretation).

1. Simple Linear Regression & Random Forest Regression

Coding:

- Use 20% of the data as your test set.
- For each numerical feature in your dataset, apply:
  - Simple Linear Regression
  - Random Forest Regression

  to predict the car's price.
- Create a dataframe (or a table) containing the Testing MSE and R-squared for each model's predictions on every numerical feature.

Write-up:

- Present the Testing MSE and R-squared values in a table.
- Provide the equations of Simple Linear Regression in the form:

$$y = mx + c$$

- Comment on your findings regarding which features appear to be better predictors of price.

2. Multiple Linear Regression & Regularization

Coding:

- Use 20% of the data as your test set.
- From the table in Question 1, choose the best three numerical features.
- Apply:
  - Multiple Linear Regression
  - Lasso Regression
  - Ridge Regression
  - Elastic Net

  to predict the car's price.

Write-up:

- Provide the final equation of the Multiple Linear Regression model (with the chosen three features).

- Create a table showing Training MSE, Testing MSE, Bias, Variance, and R-squared for each of the four models.
- Discuss:
  - How do the Bias and Variance compare across these models?
  - Which model performs best, and why?

3. Regression on All Features

Coding:

- Use 20% of the data as your test set.
- Apply:
  - Multiple Linear Regression
  - Lasso Regression
  - Ridge Regression
  - Elastic Net

  using **all features** to predict car's price.

Write-up:

- Provide the final equation of the Multiple Linear Regression model.
- Create a table showing Training MSE, Testing MSE, Bias, Variance, and R-squared for each of the four models.
- Discuss:
  - How do the Bias and Variance compare across these models?
  - Which model performs best, and why?
  - Do you see any improvement in the results compared to the results in Question 2? **Explain**.

4. Principal Component Analysis (PCA)

Coding:

- Use 20% of the data as your test set.
- Apply PCA on all features before performing Multiple Linear Regression.

Write-up:

- Determine how many PCA components are required to achieve at least 25%, 50%, 75%, and 90% of the R-squared value obtained in 3.
- Discuss how dimensionality reduction is affecting performance and interpretability.

Grading Rubric (100 Points)

Category	Criteria	Points
Code Quality	Code is well-structured, commented, and readable.	10
Tables Quality	Clear, labeled, and informative.	10
Regression models	Correct application of regression models and PCA.	50
Discussion of results	Quality of explanations, rationale, and justifications.	30
Total	**Final Score**	**100**

Table 1: Grading Rubric

Final Submission Checklist

- Jupyter Notebook (`.ipynb`)
- PDF Write-up (`.pdf`)
- All tables included
- Formatted report with proper justifications
- Sources cited where applicable