

Machine Learning - Term 242

Assignment 1: Regression

1. Dataset

Geely Auto, a Chinese automobile manufacturer, aims to enter the U.S. market by establishing a local manufacturing unit to compete with American and European automakers. To ensure a successful market entry, Geely Auto has engaged an automobile consulting firm to analyze the key factors influencing car pricing in the American market, which may differ significantly from those in China.

The company seeks to answer the following questions:

- Which variables are most significant in predicting car prices?
- How well do these variables explain car pricing trends?

To facilitate this analysis, the consulting firm has compiled an extensive dataset covering various car types available in the American market. The dataset is publicly accessible on Kaggle and can be found at the following link: <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction/data>

2. Tasks

In this assignment, you are required to develop regression models using linear regression techniques covered in the course. Your objective is to predict car prices based on the provided dataset.

2.1 Model Development (30 points)

- Implement multiple regression models using linear regression methods covered in the course to estimate car prices.
- Apply appropriate feature selection techniques to improve model performance.

2.2 Model Evaluation (30 points)

- Conduct 10-fold nested cross-validation to assess model performance (20 points)
- Use the following performance metrics to evaluate the models (10 points):
 - Mean Absolute Error (MAE)
 - Root Mean Square Error (RMSE)

2.3 Analysis and Interpretation (24 points)

1. **Feature Importance** (3 points): Which features significantly impact car prices? Are all features equally important?
2. **Brand and Model** (3 points): How do a car's brand and model influence its price prediction?
3. **Horsepower** (3 points): Does higher horsepower always result in a higher price?
4. **Fuel Type and System** (3 points): How do fuel types and fuel systems affect car pricing?
5. **Engine Size** (3 points): Is engine size strongly correlated with car price?

6. **Wheelbase** (3 points): What impact does wheelbase have on car pricing trends?
7. **Risk Rating** (3 points): Does a higher risk rating (positive symboling) increase or decrease the predicted car price?
8. **Bore and Compression Ratio** (3 points): Are bore ratio and compression ratio statistically significant in determining car price?

3. Report and Submission

Deadline: March 15, 2025

Submission Platform: Blackboard

Your submission should include a single ZIP file, named with your student ID, containing the following:

1. **Source Code** (3 points): A clean, well-documented, and runnable codebase.
2. **Short Report** (10 points) (2-3 pages, PDF format):
 - **Experimental Setup** (5 points): Describe the dataset, preprocessing steps, and modeling approach.
 - **Comparison of Methods** (5 points): Compare different regression models and justify your choices.
 - **Analysis Answers** (as per the above questions in section 2.3).
3. **Timely Submission** (3 points): Assignments submitted after the deadline may incur penalties.

Ensure that your report is well-structured, concise, and professionally formatted.

Good luck!