Predicting Auto MPG Using Linear Regression and Naive Bayes

Lab Assignment Report

Name : Vimal Kumar Verma

Roll N0: M24MAC015

Course Name: Machine Learning

Department of Computer Science Indian Institute of Technolog Jodhpur

Abstract

This report details the analysis of the Auto MPG dataset to predict the fuel efficiency of cars using Linear Regression and Naive Bayes classification. The goal was to build a hybrid model that combines the strengths of both methods to enhance prediction accuracy. The results from the three models — Linear Regression, Naive Bayes, and the Hybrid Model — are compared and analyzed.

1 Introduction

1.1 Objective

The objective of this assignment is to predict the miles per gallon (MPG) of cars using a combination of Linear Regression and Naive Bayes classification models. The goal is to explore both regression and classification techniques to build a hybrid model.

1.2 Dataset

The Auto MPG dataset used in this is taken from Kaggle. The dataset contains information on various aspects of cars such as displacement, horsepower, weight, and acceleration, which were used as features to predict MPG.

2 Data Preparation

2.1 Loading and Cleaning the Data

The dataset was loaded and preprocessed to handle missing values and extract relevant numerical features such as displacement, horsepower, weight, and acceleration.There were no missing values but there are some inconsistencies('?') in horsepower column that are handled with median of the same column.

2.2 Feature Selection

Numerical features: Displacement, Horsepower, Weight, Acceleration. Categorical features : origin Target variable: Miles per Gallon (MPG).

2.3 Train-Test Split

The dataset was split into training and testing subsets, ensuring that the models could be trained on one portion of the data and evaluated on unseen data to assess their performance.
1. Training Dataset=70%
2. Testing Dataset= 30

3 Modeling and Implementation

3.1 Linear Regression Model

Linear Regression is a supervised learning method used to predict the target variable, MPG, by analyzing the relationship between it and key numerical features (displacement, horsepower,weight, and acceleration). The model assumes a linear relationship between these variables.

- **Model Training:** A Linear Regression model was trained using the numerical features.

- **Results:** The model was evaluated using Mean Squared Error (MSE) and R-squared score.

- **Analysis and Results:** After training, the model generated predictions, which were assessed MSE and R-Squared. The results suggest that while Linear Regression offers a reasonable estimate, it doesn't perfectly capture the complexity of fuel efficiency, as it relies solely on numerical features.-

  Mean Squared Error on Train : 14.264772335505414
  R-squared Error on Train: 74.5%

  Mean Squared Error on Test : 18.474297233505414
  R-squared Error on Test: 69%

3.2 Naive Bayes Model with Categorical Features

Naive Bayes is a probabilistic classification method based on Bayes' Theorem.That assumes that no two features are dependent on each other.

- **Feature Engineering:** Numerical features were converted into categorical features

- **Column: displacement**
    - **Low:** ≤ 104.25
    - **Medium:** 104.25 - 262.00
    - **High:** ≥ 262.00
- **Column: horsepower**
    - **Low:** ≤ 76.00
    - **Medium:** 76.00 - 125.00
    - **High:** ≥ 125.00
- **Column: weight**
    - **Low:** ≤ 2223.75
    - **Medium:** 2223.75 - 3608.00
    - **High:** ≥ 3608.00

The target variable, MPG, was categorized into three groups:

- **Low MPG:** Less than 20 MPG.
- **Medium MPG:** Between 20 and 30 MPG.
- **High MPG:** Greater than 30 MPG.

- **Model Training:** A Naive Bayes classifier was trained to predict the MPG category.The model was built from scratch and trained using these categorical features.
- **Evaluation:** The classifier was evaluated using accuracy metrics.
- **Evaluation Metric and Analysis:**After training, the model generated predictions, which were assessed by Accuracy.The model demonstrated a reasonable level of accuracy in classifying vehicle fuel efficiency at an accuracy on 71% on unseen data.
    - **Accuracy on Training dataset:** 75%
    - **Accuracy on Testing dataset:** 71.6%

3.3 Hybrid Model: Combining Naive Bayes and Linear Regression

Categorical predictions from the Naive Bayes model, which classify MPG into categories such as low, medium, or high, were converted into numerical values.These values were then included as an extra feature in the Linear Regression model. This approach allowed the Linear Regression model to utilize both continuous numerical data and the categorical classifications provided by Naive Bayes.

- **Model Training:** The HybridModel class is designed to leverage a linear regression model for predictive analysis. It includes methods for training the model, making predictions, and evaluating the model's performance.
- **Evaluation:** The hybrid model was evaluated using MSE and R-squared score.

  Mean Squared Error on Test : 14.72381420241375
  R-squared Error on Test: 74%
- **Comparison:** The performance of the hybrid model is compared with the original Linear Regression model-
  - **MSE :** : The Hybrid model has a lower MSE on the test set compared to the original Linear Regression model, indicating potentially better prediction accuracy on new data.
  - **R-Squared**: The Hybrid model has a higher R-Squred score on the test set, suggesting it fits the test data better and explains more variance.
  - Hence combining the numerical features and output from Naive Bayes increases the correct prediction of MPG

4 Conclusion

In this assignment, we evaluated three different models for predicting car MPG using the Auto MPG dataset. The Linear Regression model provided a solid starting point with its ability to predict MPG from numerical features but struggled with complex data relationships. The Naive Bayes classifier, effective with categorical data, did not perform as well on its own. The hybrid model, which combined Linear Regression with Naive Bayes, proved to be the most effective approach. It demonstrated how combining different models can harness the strengths of each to improve overall performance.

5 Appendix

Please find below the code notebook used for implementing Assignment Colab Notebook link.