PREDICT CAR PRICES

A Machine Learning Approach to Car Price Prediction

Presented by: Utkarsh Jaiswal (22scse1040378)

utkarshj19@gmail.com

Guide Name- Ms. Kalyani Singh

ID- BCA3039

Introduction

- Problem Statement:

Traditional Car Pricing: Many traditional methods fail to account for the complex interaction of variables like engine size, vehicle dimensions, fuel efficiency, etc. Need for Accurate Pricing: Proper car price prediction is crucial for manufacturers, dealerships, and consumers to make informed decisions.

Objective:

Develop a predictive model for car prices using a comprehensive dataset. Predict car prices based on features such as engine size, horsepower, fuel efficiency, and more.

Dataset Overview

- Dataset Details:

Total Number of Features: 10Sample

Features : Engine Size (Numeric)Horsepower (Numeric)Year of Manufacture (Categorical)Car Make (Categorical)Fuel Efficiency (Numeric - MPG)

Dataset Source: Google Drive CSV file Link: https://drive.google.com/file/d/1JT-pWgT4iIThbC8DqGzSJ2Wv0wfpO0WW/view?usp=drivesdk

Data Preprocessing

- Steps Performed:

Missing Data Handling: Replaced missing values with the average value for features like horsepower and fuel efficiency. Categorical Encoding: Converted categorical features (e.g., car make, model) into numerical values using one-hot encoding. Data Normalization: Applied log transformation to the 'Price' variable to reduce skewness and improve model performance. Why Preprocessing is Important: Ensures that data is clean and machine-learning models can interpret it effectively.

Feature Engineering

- Engineered Features:

Car Age: 2023 minus the year of manufacture.Fuel Efficiency Group: Grouped cars based on their MPG (High, Medium, Low).Power-to-Weight Ratio: Created a new feature that represents the vehicle’s performance by dividing horsepower by the car's weight.Purpose: Improve prediction accuracy by adding more meaningful features to the model.

Machine Learning Model

- Model Selection:

Ridge Regression was chosen to handle multicollinearity and to prevent overfitting.

Training Process:

Dataset Split:

Training Set: 60%

Validation Set: 20%

Test Set: 20%

Why Ridge Regression:

Penalizes large coefficients, handles numeric and categorical features, and helps generalize well on unseen data.

Model Performance

- Evaluation Metric:

RMSE (Root Mean Squared Error)

Training RMSE: 0.45

Validation RMSE: 0.55

Graph: Show a line graph comparing Predicted vs Actual Prices for the validation set.

code

- https://colab.research.google.com/drive/1o62rSUC_E6VA1TGc5vYEuKp1sNLrqhAC?usp=sharing

Multiple Model Evaluation

- Models Used:
- - Ridge Regression (original)
- - Linear Regression
- - Random Forest Regressor
- Data Split: Training: 60% , Validation: 20% , Test: 20%
- Evaluation Metric: Root Mean Squared Error (RMSE)
- Why Multiple Models:
- - To compare performance and choose the most accurate model.

Actual vs Predicted Price (Validation Set)

- Visualization:
- - Scatter plots comparing actual vs predicted prices.
- - Diagonal line shows perfect prediction.
- Insight:
- - Random Forest shows better fit with less scatter.
- - Linear and Ridge Regression models show more variance.

Model Accuracy (RMSE Comparison)

- RMSE Results:
- - Ridge Regression: ~0.51
- - Linear Regression: ~0.51
- - Random Forest: ~0.13
- Insight:
- - Lower RMSE means better performance.
- - Random Forest performed best for this dataset.

Results & Insights

- Key Findings:

Engine Size and Car Age have a significant impact on price predictions. The model performs well with an RMSE of 0.55, indicating accurate predictions.

Applications:

For Dealerships: Optimizes pricing strategies based on vehicle specifications.

For Consumers: Provides insights on market prices for various car models and types.

Challenges

- Challenges Faced:

Data Availability: Some features had missing values, which required careful handling.

Feature Selection: Balancing between including enough features for accuracy and avoiding overfitting.

Interpreting Results: Ensuring that the model's output is understandable for non-technical stakeholders.

Conclusion

- Summary:

Developed an accurate predictive model for car prices using a robust dataset. The model can inform decisions for car manufacturers, dealerships, and consumers. Final Thoughts: Future improvements and additional data can make the model even more accurate and useful for various sectors of the automotive industry.

Thankyou