Ministry of Higher Education and Research
Djilali Bounaama University of Khemis Miliana

Faculty of Materials Science and Computer Science
Department of Artificial Intelligence

Report about Predicting Used Car Prices with Linear Regression

Master 1 Artificial Intelligence

Presented by:

Douadjia Abdelkarim

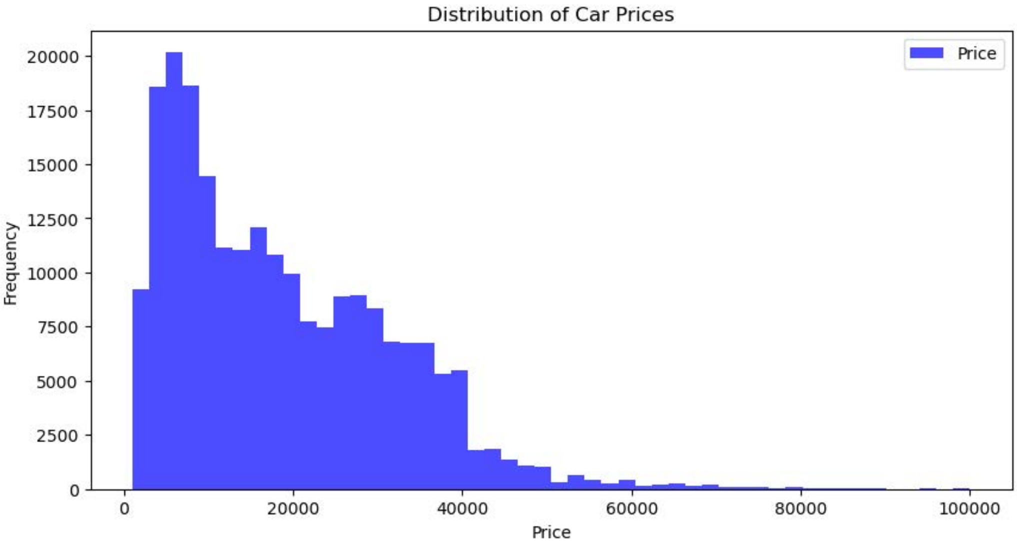
Report on Predicting Used Car Prices with Linear Regression

1. Introduction

The purpose of this project is to predict the price of used cars based on various features such as mileage, year, model, brand, and fuel efficiency. A publicly available dataset, such as Kaggle's "Used Cars" or UCI's "Car Evaluation" dataset, is utilized for this purpose. The project follows a structured approach that includes data exploration, feature engineering, model training, evaluation, and analysis of feature importance.

2. Data Exploration and Pre-processing

- ✓ The dataset was loaded and explored to understand its structure and identify missing values and outliers.
- ✓ The dataset contains 426,880 entries and 26 features, including price, year, manufacturer, model, odometer, fuel type, and transmission.
- ✓ Data cleaning steps were performed, including handling missing values and outliers.
- ✓ Distribution plots were created to visualize the spread of the target variable (car prices) and key features.



Most cars are priced under $20,000, with a long tail of luxury vehicles.

3. Feature Engineering

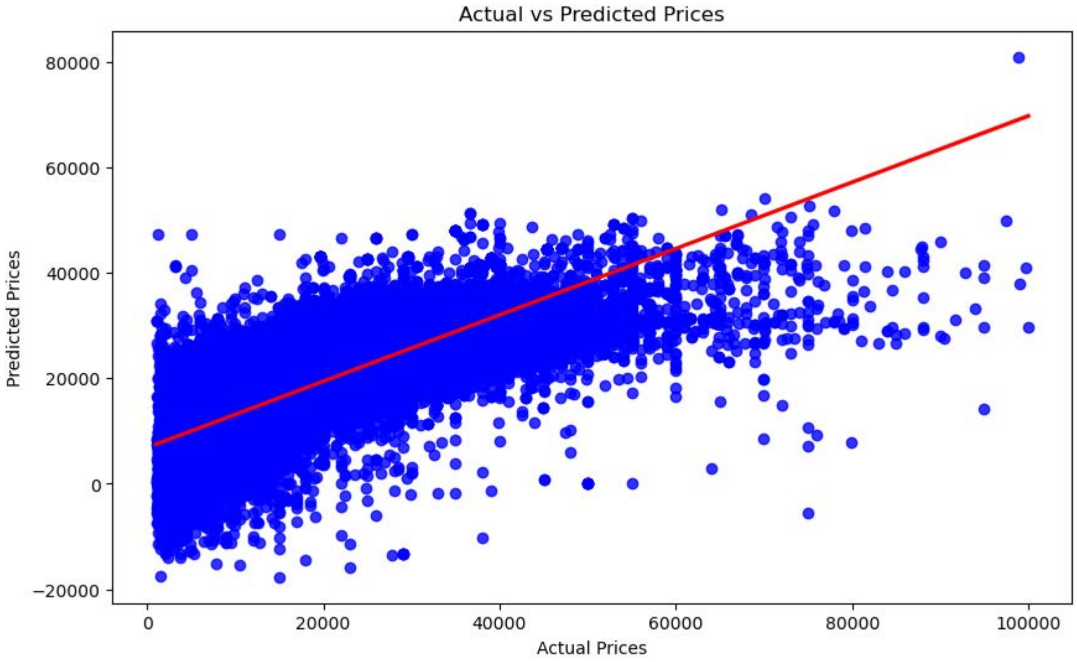
- ✓ Relevant information was extracted from features, such as computing the age of the car from its manufacturing year.
- ✓ Categorical variables were encoded using label encoding.
- ✓ Numerical features were normalized or scaled to improve model performance.

4. Training a Linear Regression Model

- ✓ The dataset was split into training and testing sets to evaluate the model's generalizability.
- ✓ A linear regression model was trained using scikit-learn and a custom implementation developed in class.

5. Model Evaluation

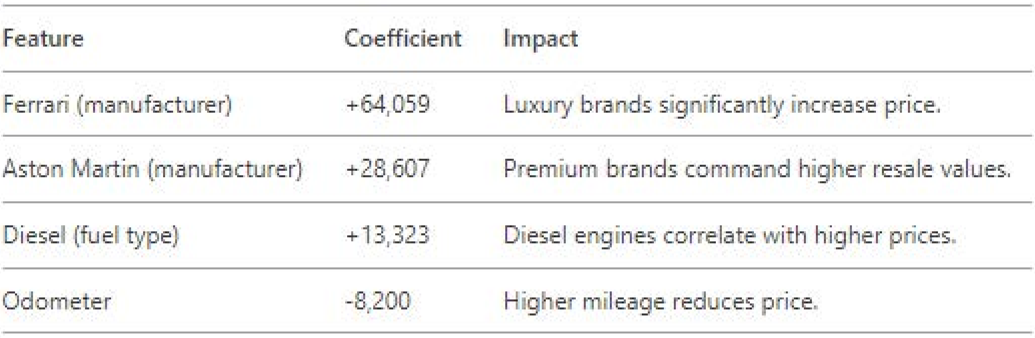
- ✓ The model was evaluated using Mean Squared Error (MSE) of 63,529,711.82 and an R-squared score of 0.63.
- ✓ Scatter plots were generated to visualize the predicted prices against actual prices.



Comment the plot:
The model performs well for lower-priced cars (<$50k) but struggles with luxury vehicles.
Scatter plots comparing predicted vs. actual prices: The spread in the scatter plot suggests that while the model captures the general trend, there are deviations that could be minimized with further tuning.

6. Feature Importance Analysis

- ✓ The coefficients of the linear regression model were analyzed to identify the most influential features:
  - Manufacturer (Ferrari, Aston-Martin, Tesla) had the highest positive impact on price.
  - Fuel type (Diesel) also showed a significant impact.
  - Odometer readings had a moderate impact on price.
- ✓ These findings suggest that luxury brands and fuel efficiency play a key role in pricing.
- ✓ Buyers should consider brand reputation and mileage, while sellers can leverage premium brands for higher pricing.



Feature	Coefficient	Impact
Ferrari (manufacturer)	+64,059	Luxury brands significantly increase price.
Aston Martin (manufacturer)	+28,607	Premium brands command higher resale values.
Diesel (fuel type)	+13,323	Diesel engines correlate with higher prices.
Odometer	-8,200	Higher mileage reduces price.

Implications:

- **Buyers:** Prioritize low-mileage vehicles from reputable brands.
- **Sellers:** Highlight brand value and fuel efficiency for premium pricing.

7. Predictions for New Unseen Cars (Bonus)

- ✓ The trained model was used to predict prices for unseen car data.
- ✓ Example Unseen Car Data Used for Prediction:
  - Manufacturer: Toyota
  - Model: Corolla
  - Year: 2018
  - Odometer: 40,000 miles
  - Fuel Type: Gasoline
  - Transmission: Automatic
- ✓ Predicted Price: $26,067.73
- ✓ The model can be applied in real-world scenarios such as automated pricing systems for used car marketplaces.
- ✓ Limitations include the model's reliance on available features and potential overfitting to dataset biases.

8. Conclusion

This project provided a comprehensive hands-on experience in applying linear regression to a real-world problem. The analysis highlighted key factors influencing used car prices, and the trained model demonstrated its predictive capability. Future improvements could include using advanced regression techniques and incorporating additional features to enhance prediction accuracy. Additionally, addressing dataset biases and integrating real-time pricing data could further improve model performance