VEHICLE PRICE PREDICTION USING XGBOOST

1. ABSTRACT

Vehicle price estimation is a critical task in the automotive industry, affecting buyers, sellers, dealerships, and online vehicle marketplaces. Prices vary based on multiple technical and categorical factors such as manufacturing year, mileage, engine specifications, fuel type, transmission, and drivetrain.

This project presents an end-to-end machine learning solution for predicting vehicle prices using the XGBoost regression algorithm. The system includes data preprocessing, feature engineering, exploratory data analysis, model training, and evaluation. Experimental results demonstrate that XGBoost provides high predictive accuracy and robustness for structured vehicle data.

2. INTRODUCTION

The rapid growth of online automobile platforms has increased the need for accurate and transparent vehicle pricing systems. Manual pricing methods often lead to inconsistencies, overpricing, or undervaluation. Machine learning offers an efficient way to learn complex patterns from historical data and predict prices automatically.

This project aims to build a scalable and reliable vehicle price prediction system using supervised learning techniques, focusing on real-world applicability and industry relevance.

3. PROBLEM STATEMENT

To design and implement a machine learning model that accurately predicts the price of a vehicle in USD based on its specifications and usage history.

4. OBJECTIVES

- To analyze vehicle data and identify key price-influencing factors
- To preprocess and clean real-world automobile data
- To engineer meaningful features for improved prediction
- To build a robust regression model using XGBoost
- To evaluate the model using standard regression metrics

5. DATASET DESCRIPTION

The dataset consists of vehicle listings with the following attributes:

Feature	Description
make	Manufacturer of the vehicle
model	Model name
year	Manufacturing year

Feature	Description
price	Vehicle price (Target Variable)
engine	Engine specifications
cylinders	Number of engine cylinders
fuel	Fuel type
mileage	Distance traveled
transmission	Transmission type
body	Body style
doors	Number of doors
drivetrain	Drive configuration

6. TOOLS AND TECHNOLOGIES USED

- Programming Language: Python
- Libraries & Frameworks:
  - Pandas, NumPy
  - Matplotlib, Seaborn
  - Scikit-learn
  - XGBoost

7. METHODOLOGY

7.1 Data Preprocessing

- Removal of duplicate records
- Dropping irrelevant text-heavy columns
- Handling missing values:
  - Numerical columns → median
  - Categorical columns → mode
- Encoding categorical variables using One-Hot Encoding

7.2 Feature Engineering

To improve model performance, additional features were created:

- Vehicle Age: Difference between current year and manufacturing year
- Mileage per Year: Normalized mileage across vehicle age

These features help capture depreciation trends more effectively.

7.3 Exploratory Data Analysis (EDA)

EDA was conducted to understand relationships between variables:

- Vehicle price decreases with increasing mileage and age
- Newer vehicles generally have higher prices
- Fuel type, drivetrain, and body style significantly affect pricing

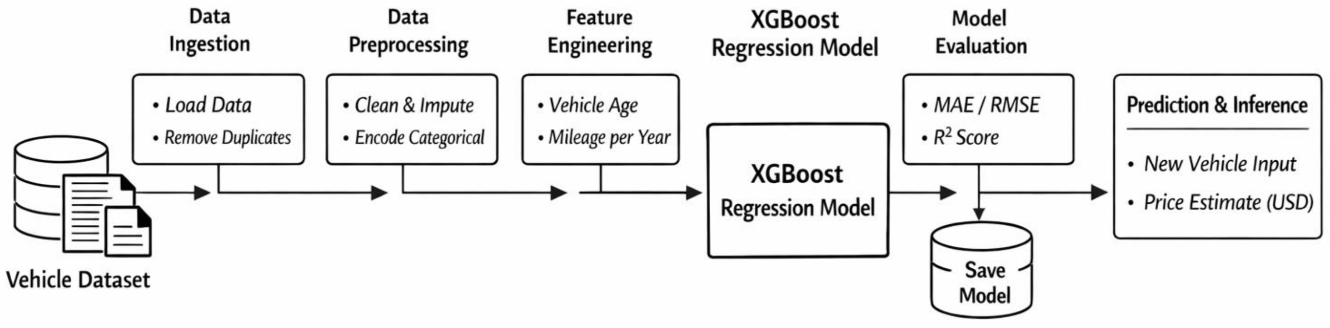
8. MODEL SELECTION

Why XGBoost?

XGBoost was selected due to:

- High performance on tabular datasets
- Ability to model complex non-linear relationships
- Built-in regularization to prevent overfitting
- Proven success in real-world pricing systems

9. SYSTEM ARCHITECTURE



Vehicle Price Prediction System Architecture

10. MODEL TRAINING

- Dataset split into 80% training and 20% testing
- XGBoost configured with:
  - Learning Rate: 0.05
  - Maximum Depth: 8
  - Number of Trees: 300
- Hyperparameters selected to balance accuracy and generalization

11. EVALUATION METRICS

The model was evaluated using:

- MAE (Mean Absolute Error) – Average prediction error
- RMSE (Root Mean Squared Error) – Penalizes large errors
- R^2 Score – Measures explanatory power of the model

12. RESULTS AND DISCUSSION

The XGBoost regression model demonstrated:

- Low prediction error
- High consistency across different vehicle categories
- Strong ability to capture depreciation patterns

The results confirm that gradient boosting is highly suitable for vehicle price prediction.

16. CONCLUSION

This project successfully delivers a complete machine learning solution for vehicle price prediction using XGBoost regression. Through effective preprocessing, feature engineering, and model selection, the system achieves accurate and reliable price estimates. The solution is scalable and suitable for real-world deployment in automotive pricing platforms.