

Used Car Price Analysis and Prediction Using R

Project Overview

This project aims to analyze and predict the prices of used cars using a dataset containing various attributes such as brand, model, year, mileage, transmission type, fuel type, and ownership history. The objective is to develop a robust model that can classify cars into price categories and predict their exact price.

Technologies Used

- **Programming Language:** R
- **Libraries:** dplyr, cluster, caret, randomForest, e1071, nnet, rpart, Metrics, recipes

Dataset Overview

The dataset consists of multiple attributes:

- **Brand & Model:** Identifies the make and model of the car.
- **Year & Age:** Represents the manufacturing year and derived age.
- **Mileage (kmDriven):** Distance the car has been driven.
- **Transmission:** Manual or Automatic.
- **Ownership History:** First, second, or more owners.
- **Fuel Type:** Petrol, Diesel, or other types.
- **Price (AskPrice):** The price listed for the used car.

Data Preprocessing

1. **Handling Missing Values:** Checked for and treated missing or inconsistent data.
2. **Feature Engineering:** Derived new features such as age from the manufacturing year.
3. **Categorical Encoding:** Converted categorical variables into factors for model compatibility.
4. **Data Normalization:** Scaled numerical variables for better model performance.

Exploratory Data Analysis (EDA)

- Summary statistics of key variables.

- Visualization of price distributions based on fuel type, transmission, and ownership.
- Correlation analysis between features and price.

Machine Learning Models

Regression Models (For Price Prediction)

- **Linear Regression:** Baseline model to estimate car price.
- **Random Forest Regression:** Used for capturing non-linear relationships.
- **Neural Network Regression:** Applied for complex feature interactions.

Classification Models (For Price Category Prediction)

- **Decision Trees:** Simple model to classify cars into price ranges.
- **Random Forest Classifier:** Enhanced model for better accuracy.
- **Support Vector Machine (SVM):** Applied for better classification of price categories.

Model Evaluation

- **Regression Metrics:** RMSE, MAE, and R-squared.
- **Classification Metrics:** Accuracy

Results and Findings

- Random Forest performed best in both regression and classification tasks.
- Higher car age and mileage negatively impact price.
- Automatic transmission cars tend to have higher prices.
- Ownership history influences resale value significantly.

Conclusion

This project successfully builds a predictive model for used car prices using R. The findings can help buyers and sellers make data-driven pricing decisions. Future improvements can include additional features such as location-based pricing and advanced deep learning models.