Car Price Prediction Model Report

1. Introduction

This report walks through the development of a car price prediction model using linear regression. The goal was to estimate a car's selling price based on various attributes. We'll explore the dataset, dive into data preprocessing and analysis, build the model, and evaluate its performance.

2. Understanding the Data

The dataset (CarPricePrediction.csv) consists of 4340 records across 8 features, covering both car attributes and their respective selling prices.

Key Insights:

- No Missing Data: All entries were complete, which made initial data cleaning more straightforward.

- Data Types: Numerical: year, selling_price, km_driven; Categorical: name, fuel, seller_type, transmission, owner

Correlation Analysis:

- Newer cars (higher year) showed a moderate positive correlation (0.41) with price.

- Cars driven more (km_driven) had a weak negative correlation (-0.19) with price.

- Older cars typically had higher mileage, as seen from the inverse relation (-0.42) between year and km_driven.

Data Distributions:

- selling_price and km_driven were right-skewed, indicating that most cars are low-mileage and lower-priced, with a few high-end or high-mileage outliers.

- Most cars were manufactured in recent years.

Exploratory Visuals:

- Numerical features: Positive trend between year and price; Negative trend between km_driven and price

- Categorical features (via box plots):

  - Diesel cars typically sell for more than Petrol, CNG, or LPG cars.

- Dealer and Trustmark Dealer listings had higher prices than individual sellers.

- Automatic cars were generally more expensive than manual ones.

- First-owner vehicles fetched the highest resale values.

- Outliers: Clear price and mileage outliers were observed in the plots.

3. Data Preprocessing

To prepare the dataset for modeling, several key steps were taken:

- Dropped name column due to high cardinality.

- Used the Interquartile Range (IQR) method to remove outliers in selling_price and km_driven.

- Applied OneHotEncoding (drop='first') to avoid dummy variable trap.

- Used StandardScaler to normalize year and km_driven.

- Data split into 80% training and 20% testing.

Data Dimensions:

- X_train: (3472, 17), X_test: (868, 17)

- y_train and y_test for target variable

4. Building the Model

A Linear Regression model was chosen as a baseline due to its simplicity and interpretability. It was trained on the preprocessed training dataset.

5. Evaluating the Model

The model's performance was evaluated using common regression metrics:

Before Improvements vs After Improvements:

- MAE: 219,541.58 → 125,692.75

- MSE: 181.93B → 28.51B

- RMSE: 426,536.48 → 168,867.19

- R^2 Score: 0.40 → 0.54

After applying outlier removal and proper encoding, the model saw a significant boost in performance. Predictions became more accurate, and the model explained 54% of the variation in car prices.

6. Final Thoughts

The linear model performs reasonably well as a starting point. Further improvements using models like Random Forest or XGBoost may offer better accuracy by capturing complex patterns.