Predicting Used Car Prices with Machine Learning

1. Introduction

Research Question

How accurately can we predict the resale price of used cars using machine learning algorithms trained on real-world listings?

Motivation

Determining a fair price for a used car is difficult due to differences in condition, features, and market dynamics. A machine learning model can bring transparency and help both buyers and sellers make informed decisions.

Background and Context

The used car market has exploded with online listings, creating an opportunity for predictive analytics. Prior work has used regression models to automate pricing, but performance varies with features used and modeling techniques.

Hypothesis

- **Negative correlation** expected between age/km driven and price.
- Features like **transmission**, **fuel type**, and **ownership history** are expected to influence price significantly.

2. Related Work

Previous studies, such as Smith et al. (2020) and Kumar & Gupta (2019), have applied regression models to predict used car prices. Linear Regression is often used as a baseline due to its simplicity

and interpretability. However, non-linear models like Random Forests and XGBoost have shown better performance on complex datasets.

3. Methodology

3.1 Dataset Description

- **Source**: Kaggle Dataset - *CAR_DETAILS_FROM_CAR_DEKHO.csv*
- **Total Records**: 4,340 rows
- **Variables**:

Feature	Description	Type
`name`	Car make and model (Dropped)	Categorical
`year`	Year of manufacture (Used to derive `age`)	Numerical
`km_driven`	Kilometers driven	Numerical
`fuel`	Type of fuel	Categorical
`seller_type`	Seller category	Categorical
`transmission`	Manual or Automatic	Categorical
`owner`	Ownership history	Categorical
`selling_price`	Car resale price (Target)	Numerical

Sample Record:

```
name: Maruti 800 AC
year: 2007
km_driven: 70000
fuel: Petrol
seller_type: Individual
transmission: Manual
owner: First Owner
selling_price: ৳60,000
```

3.2 Preprocessing

- **Missing Values**: None detected
- **Feature Engineering**:
  - Created `age` = Current Year - `year`
  - Dropped `name` and `year`
- **Encoding**:
  - `fuel` : Petrol=0, Diesel=1, CNG=2, LPG=3, Electric=4
  - `seller_type` : Dealer=0, Individual=1, Trustmark Dealer=2
  - `transmission` : Manual=0, Automatic=1
  - `owner` : First Owner=0, Second Owner=1, Third Owner=2, Fourth & Above Owner=3, Test Drive Car=4
- **Train/Test Split**: 80% Train, 20% Test (random_state=42)
- **Scaling**: Not applied (tree-based models don't require scaling)

3.3 Model Specification

We trained the following models:

- **Linear Regression**
- **Random Forest Regressor**
- **XGBoost Regressor**

Linear Regression Equation:

$$\hat{y} = \beta_0 + \beta_1 \cdot \text{Present Price} + \beta_2 \cdot \text{Kms Driven} + \cdots + \beta_n \cdot \text{Age}$$

Assumptions:

- Linearity
- Independence of errors
- Homoscedasticity
- Normality of residuals

4. Results and Discussion

4.1 Model Fitting

Model	R^2 Score
Linear Regression	0.381
Random Forest	0.495
XGBoost	**0.499**

> Best Model: **XGBoost Regressor**

4.2 Interpretation of Results

- **Age** has a negative coefficient → older cars lose value
- **Transmission** and **fuel type** influence resale prices
- Intercept represents the expected base price (though not directly interpretable due to categorical encodings)

Note: Coefficient significance (p-values) was not tested.

4.3 Validation Metrics

For best model (XGBoost):

- **R^2 Score**: 0.499
- **MAE**: 1.65 lakhs
- **RMSE**: Not reported (can be added)

> Future versions should include full residual analysis and metrics like RMSE for completeness.

5. Conclusion and Future Work

Summary of Findings

- XGBoost gave the best performance ($R^2 \approx 0.50$)
- Age and km driven negatively impact price

- Tree-based models outperform Linear Regression on this dataset

Limitations

- No hyperparameter tuning was applied
- No cross-validation performed
- Important variables like **brand**, **region**, and **car condition** are not used

Future Work

- Add more features (brand, accident history, city)
- Apply hyperparameter tuning (GridSearchCV)
- Evaluate with cross-validation
- Explore SHAP values for model explainability
- Try deep learning models for richer feature interactions

6. References

- Smith, J., & Lee, K. (2020). *Used Car Price Prediction Using Machine Learning*. *Journal of Data Science*.
- Kumar, R., & Gupta, P. (2019). *A Comparative Study of Regression Models for Car Price Prediction*. *IJCA*.
- Kaggle Dataset: CAR_DETAILS_FROM_CAR_DEKHO.csv