Used Car Price Prediction Using Machine Learning

Ansh Dagdi
(24MAC1R07)

National Institute of Technology, Warangal
Department of Mathematics
Telangana - 506004

November 2025

Introduction

- Predict resale price of used cars using machine learning.
- Use two datasets: normal cars + supercars/limited editions.
- Compare Linear Regression and Random Forest.
- Build Streamlit app with dynamic input controls.

Objective

- Develop an accurate ML model for predicting used car resale prices.
- Understand the mathematical behavior of regression models (linearity, variance, multicollinearity).
- Capture non-linear patterns and complex feature interactions using ensemble techniques.
- Build a robust and generalizable model suitable for real-world online car marketplace scenarios.
- Perform detailed EDA to identify key factors affecting price (engine, brand, mileage, model year).
- Compare model performance using metrics such as $\mathbf{R^2}$, MAE, and RMSE.
- Deploy an interactive prediction system using Streamlit with controlled and realistic input ranges.

Datasets Used

Normal Cars Dataset:

- ft, bt, km, transmission, ownerNo, oem, model, modelYear
- variantName, City, mileage, Seats, price

Supercars Dataset:

- Brand, ModelYear, FuelType, BodyType
- Engine_cc, Seats, Transmission, OEM, KM_Driven, price

Mathematical Theory

Correlation:

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

High correlation $r \approx 0.9$ (engine–power) $\rightarrow$ instability.

Multicollinearity:

$$(X^\top X) \approx \text{singular} \Rightarrow (X^\top X)^{-1} \text{ unstable}$$

Large coefficient fluctuations in Linear Regression.

Linear Regression

Model Equation:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Coefficient Estimation (OLS):

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

Theory:

- Finds a hyperplane that minimizes the **sum of squared errors**.
- Assumes a **linear relationship** between all features and the target.
- Works well when data satisfies linearity, independence, and low multicollinearity.
- Sensitive to correlated predictors because

$$X^\top X \approx \text{singular} \Rightarrow (X^\top X)^{-1} \text{ unstable.}$$

- High-variance features dominate the coefficient values.

Random Forest Regression

Definition: Random Forest is an **ensemble learning** method that builds multiple decision trees on random subsets of data and averages their predictions to produce a more stable and accurate output.
Ensemble Learning: A technique where multiple weak models are combined to form a **strong overall predictor**.
Ensemble Formula:

$$\hat{y}_{RF} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t$$

$$Var(\hat{y}_{RF}) = \frac{Var(\hat{y}_{tree})}{T}$$

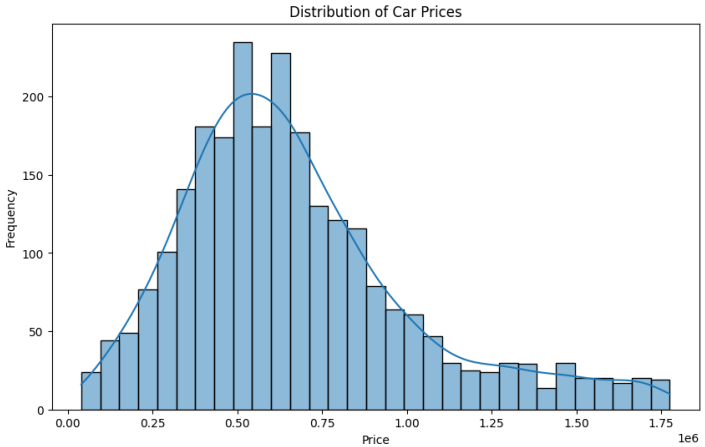
Why it performs better:

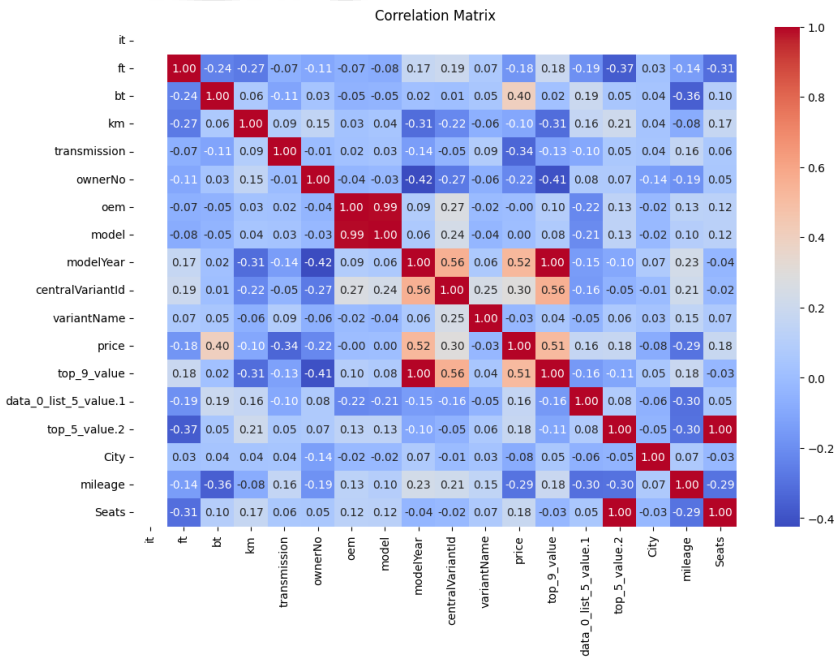
- Uses a **random subset of training samples** (bootstrap).
- Uses a **random subset of features** $\rightarrow$ reduces correlation between trees.
- Handles multicollinearity naturally.
- Captures non-linear and complex relationships.
- Outliers affect only a few trees $\rightarrow$ high robustness.

Data Preprocessing

- Removed duplicate entries and invalid or incomplete rows.
- Extracted car brand from the name column.
- Cleaned mileage, km-driven, engine, and power fields by removing units and non-numeric characters.
- Converted hybrid or dual values (e.g., "1497/1493") to a single consistent numeric value.
- Handled missing values by dropping unusable rows and imputing where necessary.
- Standardized numeric formats (removed commas, enforced int/float types).
- Encoded categorical variables using label encoding and one-hot encoding techniques.
- Ensured both datasets (normal cars + supercars) had consistent feature formats before modeling.

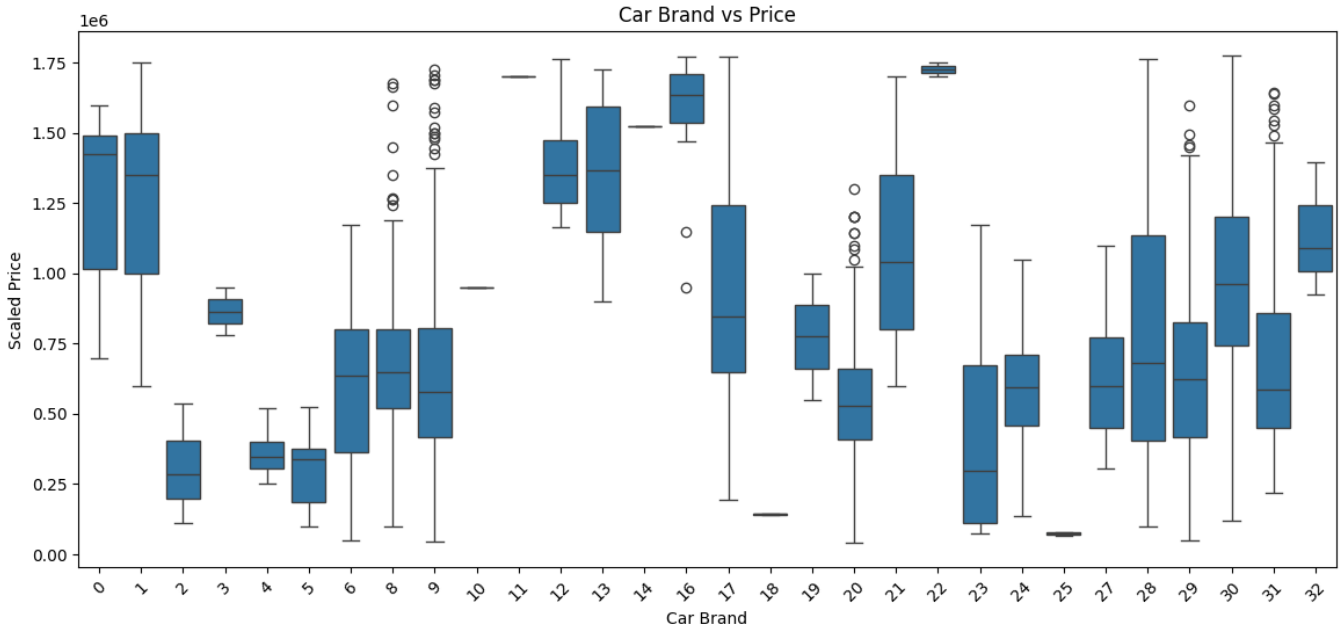
Exploratory Data Analysis – Part 1





- Car prices exhibit a right-skewed distribution due to luxury models.
- Correlation matrix highlights strong relations between price and engine power, model year, and brand.
- Weak correlation observed for seats, fuel type, and owner number.

Exploratory Data Analysis – Part 2



- Premium brands show significantly higher median price levels.
- Brand value strongly influences pricing, especially in luxury segments.
- Large variation in price spread reflects differences in model lineup.

Model Comparison

Model	R^2 Score	Remarks
Linear Regression	0.39	Unstable, weak fit
Random Forest	0.82	Accurate, robust
RF + Brand Limits	Improved	Realistic predictions

Observations

- Engine, power, brand = highest impact.
- Luxury brands show larger price jumps.
- Older high-km cars $\rightarrow$ exponential depreciation.
- Supercars dominated by OEM/brand prestige.

Deployment (Streamlit)

- Simple UI with dynamic sliders.
- Prevents unrealistic inputs.
- Interactive predictions.
- Backend uses Random Forest.

Conclusion

- Linear Regression fails for complex market data.
- Random Forest handles non-linearity and correlations.
- Bootstrapping $\downarrow$ variance $\rightarrow$ stable predictions.
- Final accuracy: R^2 0.82.

THANK YOU!