🚙 Used Car Price Prediction — Detailed Data Analysis & Modelling Report

1. Introduction

The used car market has grown significantly in recent years, with a wide variety of factors influencing how vehicles are priced. Buyers and sellers often face difficulty estimating the fair value of a vehicle due to the complex combination of mechanical, physical, and market-driven attributes.
This project aims to:

1. Understand how different car features influence price.
2. Identify the strongest predictors of price.
3. Perform exploratory data analysis (EDA) to uncover patterns and trends.
4. Build and evaluate machine learning models capable of predicting car prices accurately.
5. Provide actionable insights for sellers, buyers, and businesses.

The dataset includes thousands of car listings across multiple brands, body types, and conditions. Through analysis and modeling, the project reveals valuable pricing patterns and offers a production-ready prediction workflow.

2. Dataset Overview

The dataset contains the following columns:

Make, Model, Price, Year, Kilometer, Fuel Type, Transmission, Location, Color, Owner, Seller Type, Engine, Max Power, Max Torque, Drivetrain, Length, Width, Height, Seating Capacity, Fuel Tank Capacity

2.1. Structure and Data Types

- **Categorical Features:**
  Make, Model, Fuel Type, Transmission, Location, Color, Owner, Seller Type, Drivetrain
- **Numerical Features:**
  Price, Year, Kilometer, Engine, Max Power, Max Torque, Length, Width, Height, Seating Capacity, Fuel Tank Capacity

2.2. Dataset Characteristics

- Contains entries from a wide range of automobiles, from affordable hatchbacks to high-end luxury vehicles.
- Cars differ significantly in year, power, usage, and mechanical specifications.

- Some columns require transformation (e.g., Power/Engine units) for proper numeric interpretation.

3. Data Cleaning & Preparation

To ensure high-quality analysis and modelling, the dataset was thoroughly cleaned.

3.1. Handling Missing Values

- Missing **Price** → rows removed (critical target variable).
- Missing values in **Fuel Type, Transmission, Color, Owner, Seller Type** were replaced with "Unknown".
- Missing numeric values (if any) were cautiously imputed based on median of similar groups.

3.2. Standardizing Numerical Columns

- Extracted numeric values from strings (e.g., "100 bhp" → 100).
- Converted all dimensions to integers (Length, Width, Height).
- Converted Engine CC, Max Torque, Max Power to numeric.

3.3. Categorical Cleaning

- Unified inconsistent labels:
  Example: "manual", "Manual", "MANUAL" → **Manual**
- Standardized Make and Model names where needed.

3.4. Outlier Removal

Cars below ₹50,000 or above ₹40,00,000 were removed to stabilize model performance.

3.5. Final Clean Dataset

After cleaning, the dataset became more structured, consistent, and ready for EDA and modeling.

4. Exploratory Data Analysis (EDA) — Detailed Insights

EDA plays a crucial role in understanding patterns. Below are key insights derived from the dataset and visual graphs.

4.1. Price Distribution

- Price is **highly right-skewed**, meaning most cars are low- to mid-priced, with a small number of expensive cars pushing the distribution right.
- A **log transformation** helps linearize this skewed distribution and improves model performance.

4.2. Correlation Heatmap

- Revealed strong positive correlations:
    - **Fuel Tank Capacity** (~0.59)
    - **Width, Length**
    - **Engine**
    - **Max Power**
- Negative correlations:
    - **Kilometer Driven**
    - **Transmission** (after encoding)
- Multicollinearity:
    - Engine ↔ Max Power
    - Length ↔ Width ↔ Fuel Tank Capacity

This multicollinearity affects linear models but not tree models.

4.3. Scatter Plot Insights

Price vs Max Power

- Clear upward trend: higher power → higher price.
- Luxury cars form visible high-power clusters.

Price vs Engine

- Larger engines generally have higher price.
- Some high-engine entries denote SUVs and premium cars.

Price vs Fuel Tank Capacity

- Vehicles with large tanks (SUVs, MUVs) tend to be more expensive.
- Correlates with body size and performance.

4.4. Categorical Feature Insights

Fuel Type

- Diesel vehicles tend to be more expensive than Petrol in many cases.
- CNG/LPG vehicles are at the lower end of pricing.

Transmission

- Automatic cars consistently show higher prices than Manual cars.

Owner

- First-owner cars command the highest prices.
- Each additional owner reduces price progressively.

4.5. Make & Model Analysis

- Popular brands (Maruti Suzuki, Hyundai) dominate listing count but also show wide price variations.
- Luxury brands (BMW, Mercedes, Audi) appear less frequently but occupy the higher price range.

5. Feature Importance Analysis

Using Random Forest, feature importance rankings were calculated:

Top Predictors

1. **Max Power**
2. **Engine CC**
3. **Fuel Tank Capacity**
4. **Year**
5. **Width / Length / Height**
6. **Fuel Type**
7. **Transmission**
8. **Kilometer Driven**

Least Important

- Color
- Seller Type

- Fine-grained Location details

These findings guided feature selection for modeling.

---

6. Model Building

Three model strategies were tested:

6.1. Linear Regression

Performed poorly due to:

- Strong multicollinearity
- Non-linear relationships
- Skewed price distribution
- Categorical complexity

Produced unstable and sometimes negative predictions.

6.2. Random Forest Regressor (Best Model)

After:

- Outlier removal
- Log(Price) transformation
- Feature engineering
- Parameter tuning

Random Forest showed:

- High accuracy
- Robust performance
- No negative predictions
- Handles categorical variables well (after encoding)

Final Parameters Used

- n_estimators: 300–500
- max_depth: 20–30
- random_state: 42
- min_samples_leaf: 1

6.3. XGBoost / CatBoost (Optional Upgrades)

These models provide even higher accuracy and handle categorical variables more efficiently.

---

7. Model Performance Evaluation

After improvements, final model metrics were:

R^2 Score: 91.40%, showing strong predictive ability.

Mean Absolute Error (MAE): ₹249,917.63

Root Mean Squared Error (RMSE): ₹152,502.15

This is a significant improvement from initial performance:

- MAE was ~₹4.7 lakhs
- RMSE was ~₹17.8 lakhs
- R^2 ~72%

---

8. Business-Level Insights

Price Drivers

- Mechanical power and engine capacity are the primary determinants of used car pricing.
- Year and Kilometer driven (age & use) strongly affect depreciation.
- Brand alone is *not* a strong predictor—model + engine + power combination matters more.

Market Insights

- SUVs and high-engine cars dominate the upper pricing tiers.
- Petrol economy cars dominate the lower price range.
- Automatic transmissions are increasingly valued in higher-end segments.

---

9. Recommendations

For Model Deployment

- Use Random Forest or XGBoost with log-transformed target.
- Keep dimensional and power-based features in the model.
- Retrain model every 6–12 months as market prices change.

For Data Improvement

- Add features like service history, accident records, insurance validity, and variant/trim.
- Improve accuracy by including real market depreciation curves.

---

10. Conclusion

The analysis provides a thorough understanding of car pricing behavior in the used car market. Mechanical specifications, power, dimensions, year, fuel type, transmission, and mileage are the strongest price determinants. Using a carefully cleaned dataset, engineered features, and a tuned RandomForest model, we can reliably predict car prices with high accuracy.

The final model is ready for integration into a prediction system such as:

- A web app
- A mobile app
- A dealership valuation tool
- A consumer-facing price estimator

This project establishes a strong foundation for building a data-driven used car valuation system.