Predictive Modelling for Used Car Pricing

(Team 16)

Abhilasha Kawle - abhilashak@iisc.ac.in

Manikanda S Subramaniam
manikandasa1@iisc.ac.in

Anfaal O Waafy -anfaalwaafy@iisc.ac.in

Vimalraj K - vimalrajk@iisc.ac.in

1. Motivation and problem statement

Car buyers and seller face challenges of determining fair market value of used cars due to lack of accurate, data driven price estimation. Traditional price estimation involves Manual Inspection & Expert Appraisal, Rule-Based or Heuristic Pricing, and Market Comparison Tools. Because of human errors there is wide range of price variation in different platforms.

Objective in Business Terms: To increase sales conversion by building customer trust . This can be done by enhancing transparency using a data driven model to predict market price based on car features that account for Depreciation and Quality Control.

1.1 Approach

It's a supervised learning problem, predicting car re-sell prices on historical data, and is handled offline. Use MSE, R2 to measure accuracy of prediction models. We have used the “Data science Workflow” Methodology

2. Data set and data preparation

2.1 Dataset Overview

Our analysis utilizes the “Vehicle Dataset from CarDekho”, which contains information on 8,128 used cars sold in India. The raw dataset consists of following:

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm

2.2. Feature Engineering

- Missing value filled using KNN imputation
- **Car Name Structure**: Contains both brand and model information combined à Separated
- **Mixed-Format Columns**: Performance attributes (mileage, engine, max_power, torque) contain numeric values + units. Eg : '190Nm@ 2000rpm – Values and units were split
- **Unit normalization** : Domain Knowledge like fuel density was used for converting kg/l to kmpl for milage, kgm to Nm for torque and RPM was average instead of range.

	name	make	model
0	Maruti Swift Dzire VDI	Maruti	Swift
1	Skoda Rapid 1.5 TDI Ambition	Skoda	Rapid
2	Honda City 2017-2020 EXi	Honda	City
3	Hyundai i20 Sportz Diesel	Hyundai	i20
4	Maruti Swift VXI BSIII	Maruti	Swift

	max_power	max_power_value	max_power_unit
0	74 bhp	74.00	bhp
1	103.52 bhp	103.52	bhp
2	78 bhp	78.00	bhp
3	90 bhp	90.00	bhp
4	88.2 bhp	88.20	bhp

	torque	torque_value	torque_unit	torque_nm	rpm_avg
0	190Nm@ 2000rpm	190.0	nm	190.000000	2000.0
1	250Nm@ 1500-2500rpm	250.0	nm	250.000000	2000.0
2	12.7@ 2,700(kgm@ rpm)	12.7	kgm	124.544455	2700.0
3	22.4 kgm at 1750-2750rpm	22.4	kgm	219.668960	2250.0
4	11.5@ 4,500(kgm@ rpm)	11.5	kgm	112.776475	4500.0

	mileage	mileage_value	mileage_kmpl
0	23.4 kmpl	23.40	23.400000
1	21.14 kmpl	21.14	21.140000
2	17.7 kmpl	17.70	17.700000
3	23.0 kmpl	23.00	23.000000
4	16.1 kmpl	16.10	16.100000
5	20.14 kmpl	20.14	20.140000
6	17.3 km/kg	17.30	33.921569
7	16.1 kmpl	16.10	16.100000

```
# Fuel density ratios (kg/L) used for conversion
fuel_density = {
    'Petrol': 0.74,
    'Diesel': 0.832,
    'LPG': 0.51,
    'CNG': 0.615
}
return row['mileage_value'] / density
```

3. Exploratory Data Analysis

Comprehensive Exploratory Data Analysis was conducted to identify the underlying structure of the data, detect outliers, and determine the primary drivers of car resale value.

3.1 Descriptive Statistics

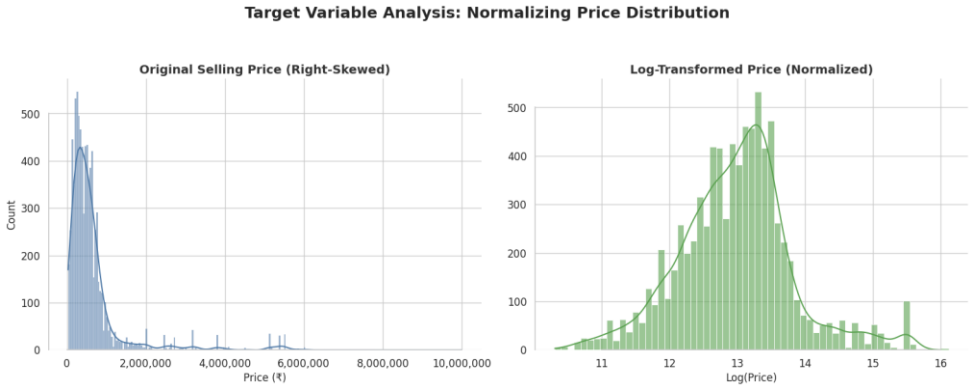
After initial preprocessing and cleaning, the dataset revealed the following statistics:

1. **Price Range**: The selling prices ranged significantly, from a minimum of ₹29,999 to a maximum of ₹10,000,000, with an average resale price of approximately ₹638,271.
2. **Vehicle Age**: The dataset covers a wide range of vehicle ages, with a mean vehicle age of approximately 11 years.

Usage: The average distance driven was approximately 69,819 km, indicating a mix of lightly used and heavily driven vehicles.

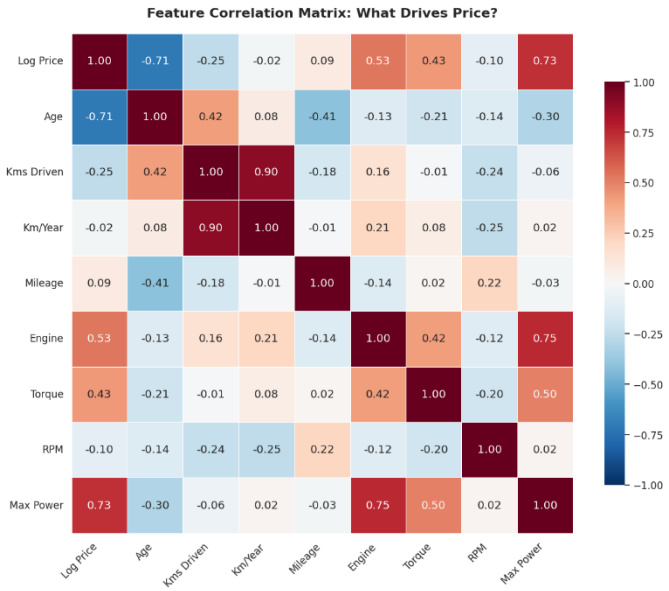
3.2 Target Variable Analysis: Normalization

The primary target variable, selling_price, exhibited a highly right-skewed distribution in its raw form. This skewness is typical in pricing data, where a small number of high-value luxury vehicles distort the distribution range. To address this and satisfy the normality assumptions required for regression modeling, we applied a Log Transformation to the target variable which successfully normalized the price distribution, reducing the influence of outliers and creating a bell-curve shape suitable for linear and tree-based models.



3.2 Numerical Feature CorrelationsA correlation matrix was generated to quantify the linear relationships between numerical features and the log-transformed selling price. The key findings were:

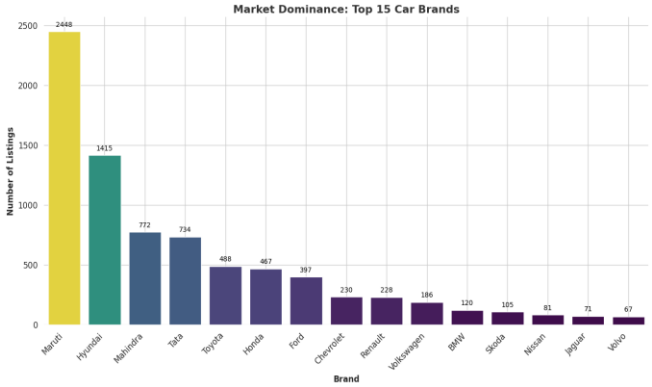
1. **Max Power (+0.73)**: Strongest positive correlation with price. Higher performance (bhp) is the single most significant predictor of a higher resale value.
2. **Vehicle Age (-0.71)**: Strong negative correlation. Older cars depreciate price significantly
3. **Moderate Impact:**
    a. **Engine Volume (0.53) and Torque (0.43)** – Imply car performance impact pricing positively
    b. **Kilometers Driven (-0.25)** negatively affects price but is less significant than the car's age.

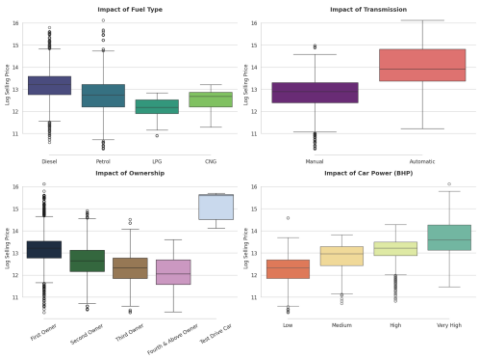


3.3 Categorical Feature Analysis

We analyzed categorical variables to understand how market segments influence pricing:

1. **Transmission**: There is a clear price premium for Automatic vehicles compared to Manual transmission models.
2. **Fuel Type**: Diesel vehicles consistently retain a higher median resale value compared to Petrol or CNG variants.
3. **Ownership History**: A significant drop in value was observed as the number of owners increased. "First Owner" cars command the highest prices, with depreciation accelerating for "Second" and "Third" owners.
4. **Brand Dominance**: The dataset is heavily dominated by mass-market manufacturers such as Maruti, Hyundai, and Mahindra, which influences the baseline pricing model.





4. Data Model Development

Evaluation Approach :

- Train/test split (85/15) with stratified cross-validation.
- Metrics tracked: R^2, RMSE, MAE, prediction interval coverage.
- Honest target encoding to avoid leakage
- K-fold validation (k=5), OOF CV didn't improve accuracy

Models	Parameters	Train R2	Validation R2
Ridge Regression	L2 regularization	0.933	0.905
Random Forest Regressor	Estimators = 50 Max_depth =12	0.981	0.944
XGBoost Regressor	Estimators = 500 Learning_rate = 0.05 Max_depth =12	0.998	0.945
CatBoost Regressor	Iterations=800, Learning_rate=0.05, Depth=8	0.980	0.949
LightGBM Regressor	Estimators = 800 Learning_rate = 0.03 num_leaves = 31	0.984	0.950
LightGBM + Target Encoding	Learning_rate = 0.05 num_leaves = 31	0.984	0.949

- Stacking model, Target encoding also didn't help with accuracy improvement

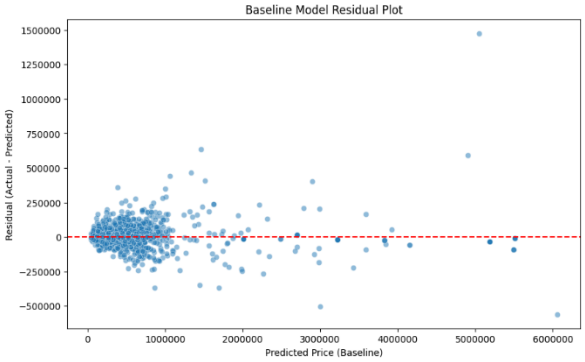
5.Key results and insights

.5.1 Performance Outcomes

- Best Model: LightGBM with target encoding.
- Train R^2: 0.984, Test R^2: 0.950
- RMSE: ≈ ₹99, 649, MAE: ≈ ₹57,448

5.2 Conclusion:

- We were able to predict price with 95% accuracy
- To account for uncertainty we included range estimation using quatile regression using LightGBM model
- Using car performance features, any new car model price can be predicted



6. Limitations

- Underprediction for luxury/high-end cars.
- Overprediction for rare fuel types (LPG/CNG).
- Residual plots reveal larger errors for expensive cars.

7. Contributions by each team member

Abhilasha Kawle:

- Data preprocessing, schema standardization, Car Nameparsing, Feature engineering, Initial model analysis usingXGBoost/LightGBM

Anfaal Obaid Waafy:

- Exploratory Data Analysis, visualizations, descriptive statistics.

Manikanda Sakthi:

- Baseline linear regression models, evaluation protocol, Stacking ensemble

Vimalraj:

- Advanced models, hyperparameter tuning, prediction intervals, deployment demo.