DA 204o: Data Science in Practice
Course Project

Predictive Modelling for Used Car Pricing
(Project 16)

- Manikanda Sakthi, manikandasa1@iisc.ac.in
- Anfaal Obaid Waafy, anfaalwaafy@iisc.ac.in
- Vimalraj K, VimalrajK@iisc.ac.in
- Abhilasha Kawle, abhilashak@iisc.ac.in

Image source: Internet

Problem Statement

- Background of the problem
  - Buyers and sellers challenge → Determine fair market value of used-car
  - Reason → Lack an accurate, data-driven estimate
    - Manual Inspection and estimation – inconsistent and inaccurate
    - Different platforms show wide price variance.
- Why is it important?
  - Customers → Transparency, Trust
  - Business (Marketplace, Dealerships)
    - Automated, reliable price prediction system for informed decisions
    - Increase sales conversions by reducing negotiation gaps
- Objectives of the project
  - Build a data-driven model to predict market price of used car based on their features
  - Provide model interpretability and a deployable pipeline.
- How can Data Science solve the problem?
  - Leverage historical listings and feature engineering to learn price drivers

Data Science Workflow

1. Problem Definition
2. Data collection and preparation
3. Data Exploration
4. Data Modeling
5. Reflection or Inference Phase
6. Communicating & Visualizing Results

Predictive Modelling for Used Car Pricing

- **Objective in Business Terms**
  - Market Value estimation for Dealerships accounting Depreciation and Quality Control
  - **Increase Sales Conversions** – Enhance transparency → customer trust,

- **How Will Your Solution Be Used?**
  - Car marketplaces, dealerships, and resale platforms → **Reliably estimate** the *fair market value* of a used car
  - Sellers → list a competitive price; Buyers → avoid overpaying

- **Current Solutions/Workarounds:** Manual Inspection & Expert Appraisal, Rule-Based or Heuristic Pricing, Market Comparison Tools

- **How Should You Frame This Problem?** Offline **supervised learning** problem, predicting car re-sell price on historical data

- **How Should Performance Be Measured?** Use **MSE, R2** to measure accuracy of prediction models

- **Is the Performance Measure Aligned with the Business Objective?** Yes, as it aids financial planning and asset management

- **What Would Be the Minimum Performance Needed?** Aim for at least prediction of 90%

- **What Are Comparable Problems?** Prediction models from **house re-sell price prediction** can be adapted.

- **List the Assumptions and verification of them:**
  - Price can be predicted based on training on historical data - small-scale tests to confirm predictions
  - Available data is sufficient and reliable - Run preliminary analyses for outliers behavior

Data Science Workflow

1. Problem Definition
2. Data collection and preparation
3. Data Exploration
4. Data Modeling
5. Reflection or Inference Phase
6. Communicating & Visualizing Results

Step 2a: Data Understanding

- Vehicle Dataset from CarDekho (Kaggle) — vehicle features and selling prices.
- Link: https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho

- The dataset from Car Dekho contains information about **8,128 used cars** sold in the past.
- Each row represents a car listing with the following attributes:
- **Car Identification & Basic Info:**
  - Full car brand and model , Manufacturing year , Original Selling Price, Re-sale price
- **Car Usage History:**
  - Kms Driven , Ownership history, Seller type (Individual, Dealer, Trustmark Dealer)
- **Car Performance Specifications:**
  - Fuel type (Diesel, Petrol, LPG, CNG) , Transmission type (Manual, Automatic)
  - Mileage, Engine Volume, Power, Torque seats: Number of seats

Step 2a: Data Understanding

- **Car Name Structure**: Contains both brand and model information combined (e.g., "Maruti Swift Dzire VDI").
- **Mixed-Format Columns**: Performance attributes (mileage, engine, max_power, torque) contain numeric values + units. Eg : '190Nm@ 2000rpm – Split

	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

Step 2b: Data Preparation

- **Missing Values**: The dataset contains NULL values in car performance attributes:
  - Mileage, engine, seats: 221 missing values
  - max_power: 215 missing values
  - torque: 222 missing values
  - Missing value filled using KNN imputation

```
# 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
```

- **Feature Engineering** on split Numeric and units
  - Parsed Engine (CC) Max Power (bhp).
  - Parsed and standardize mileage → km/kg to kmpl using fuel density ratios
  - Extract and standardize torque value → kgm to Nm
  - separate RPM from torque and standardize → Range to average
  - Add Vehicle Age = (Current Year - Model Year).
  - Add Km_Per_Year = (Total Kms / Age).
  - Extracted 'Make' and 'Model' from car names.

Based on Domain Knowledge

Step 2a: Feature Engineering

	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

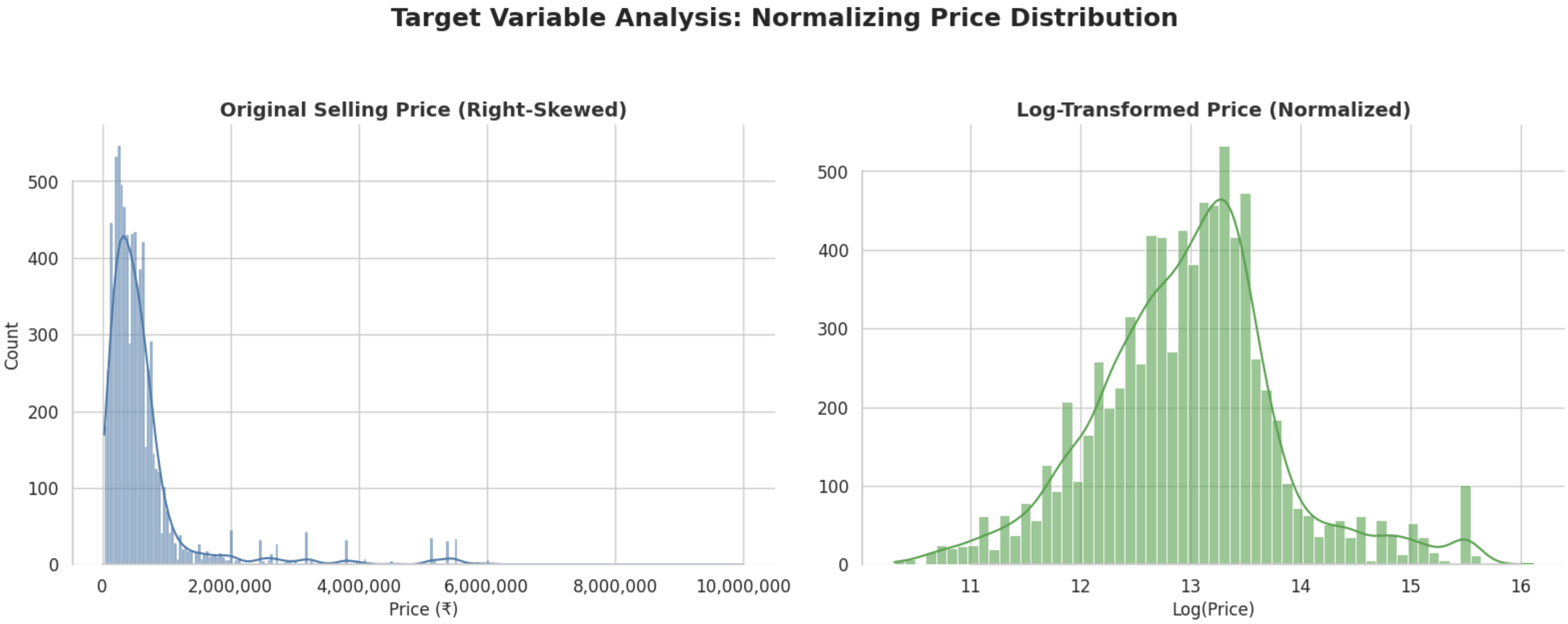
```
# Fuel density ratios (kg/L) used for conversion
fuel_density = {
    'Petrol': 0.74,
    'Diesel': 0.832,
    'LPG': 0.51,
    'CNG': 0.615
}
return row['mileage_value'] / density
```

Data Science Workflow

1. Problem Definition
2. Data collection and preparation
3. Data Exploration
4. Data Modeling
5. Reflection or Inference Phase
6. Communicating & Visualizing Results

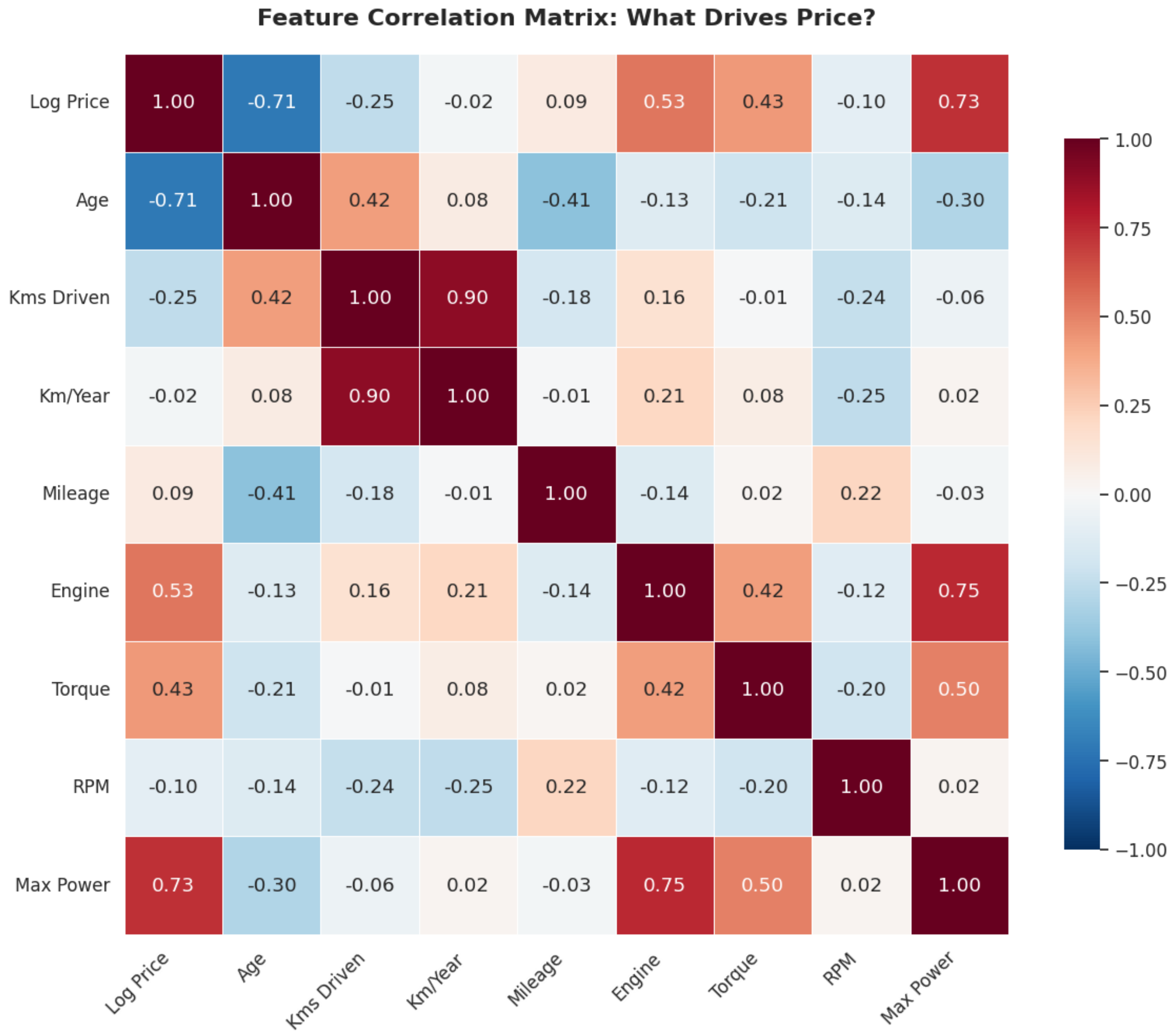
Step 3a: Data Exploration - Target Variable Analysis

- **The Skewness Problem**: The raw *selling_price* distribution was highly right-skewed, meaning a small number of expensive luxury cars were distorting the average.
- **The Solution**: We applied a Log Transformation *(log(1+x))* to the target variable to reduce skewness.
- **The Result**: As shown in the green chart, the data now follows a near-normal distribution, which satisfies the assumptions required for regression modeling.



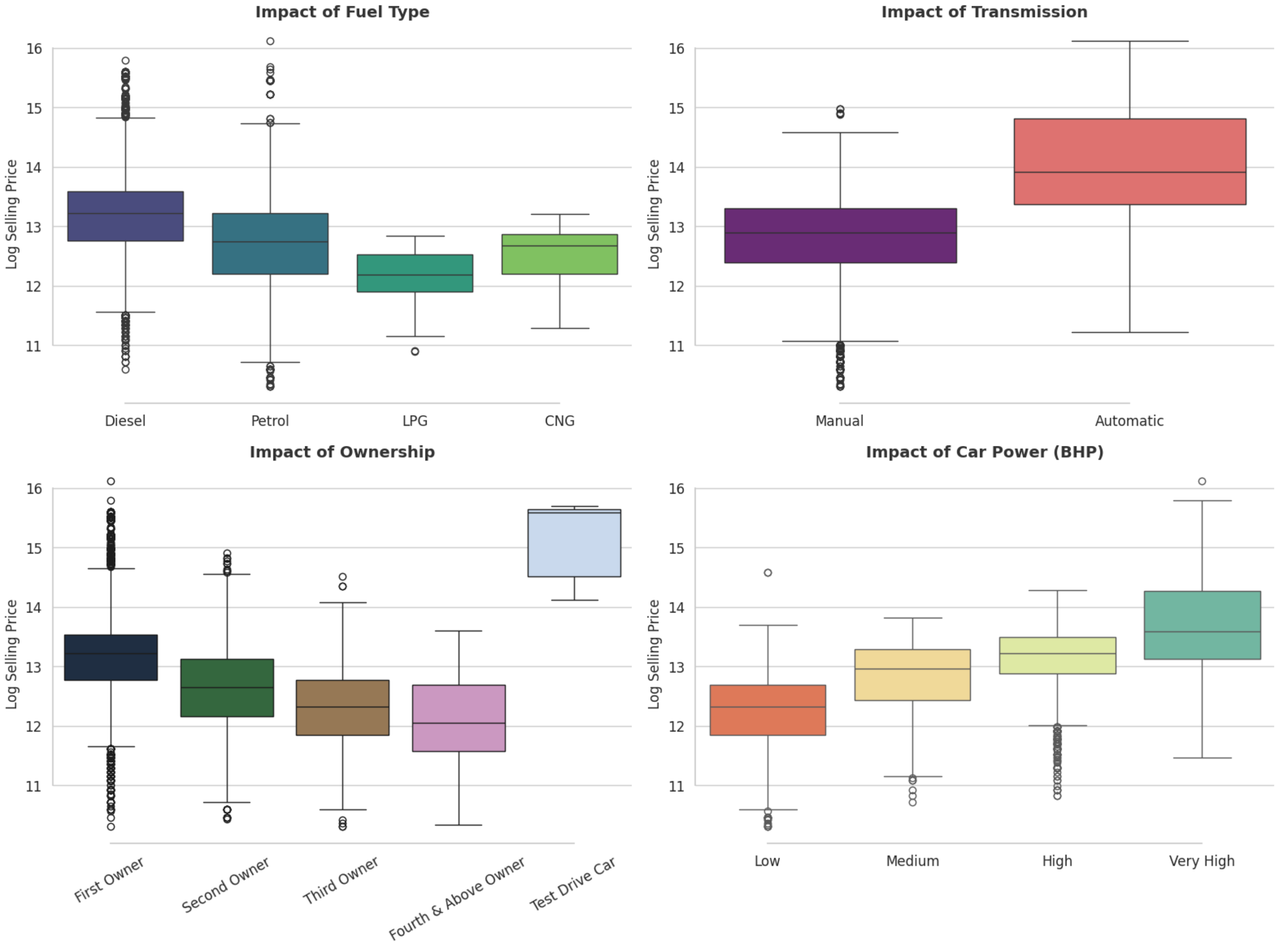
Step 3b: Feature Correlation Matrix (Heatmap)

- **Key Insight**: Identified the strongest numerical drivers of price.
- **Positive Correlation**: Max Power (0.73) is the single biggest predictor of higher value.
- **Negative Correlation**: Vehicle Age (-0.71) is the primary depreciation factor.
- **Moderate Impact**:
  - Engine Volume (0.53) and Torque (0.43) – Imply car performance impact pricing positively
  - Kilometers Driven (-0.25) negatively affects price but is less significant than the car's age.



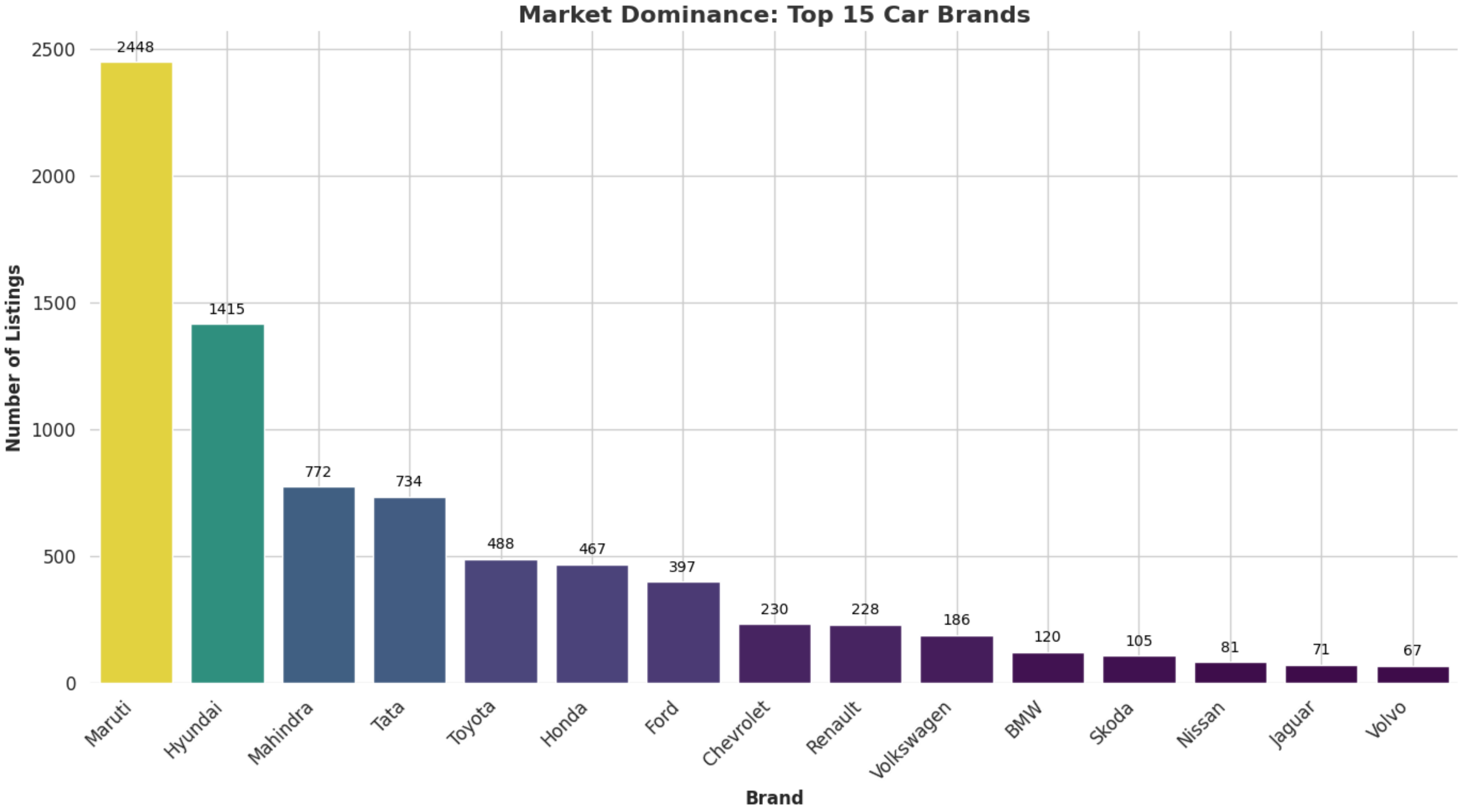
Step 3c: Categorical Feature Impact (Box Plots)

- **Fuel**: Diesel cars retain higher median value compared to Petrol or CNG.
- **Transmission**: Automatic vehicles command a significant price premium over Manuals.
- **Ownership**: Valuation drops sharply after the First Owner; subsequent owners see accelerated depreciation.
- **Power Group**: Higher power bands ("High", "Very High") consistently correlate with higher median prices.



Step 3d: Market Composition (Top Brands)

- **Dominant Players**: The dataset is heavily skewed towards mass-market leaders: Maruti, Hyundai, and Mahindra.
- **Implication**: The model is highly robust for common Indian family cars due to the high volume of data.
- **Outliers**: Luxury brands (Mercedes, BMW) are present but represent a smaller fraction, treated as high-value outliers in earlier analysis.



Data Science Workflow

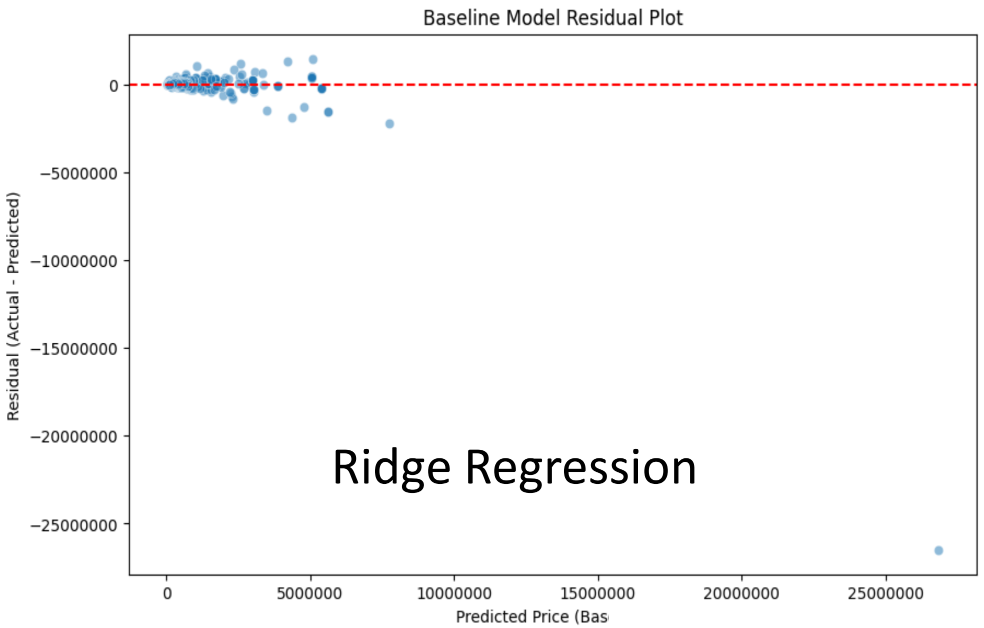
1. Problem Definition
2. Data collection and preparation
3. Data Exploration
4. **Data Modeling**
5. Reflection or Inference Phase
6. Communicating & Visualizing Results

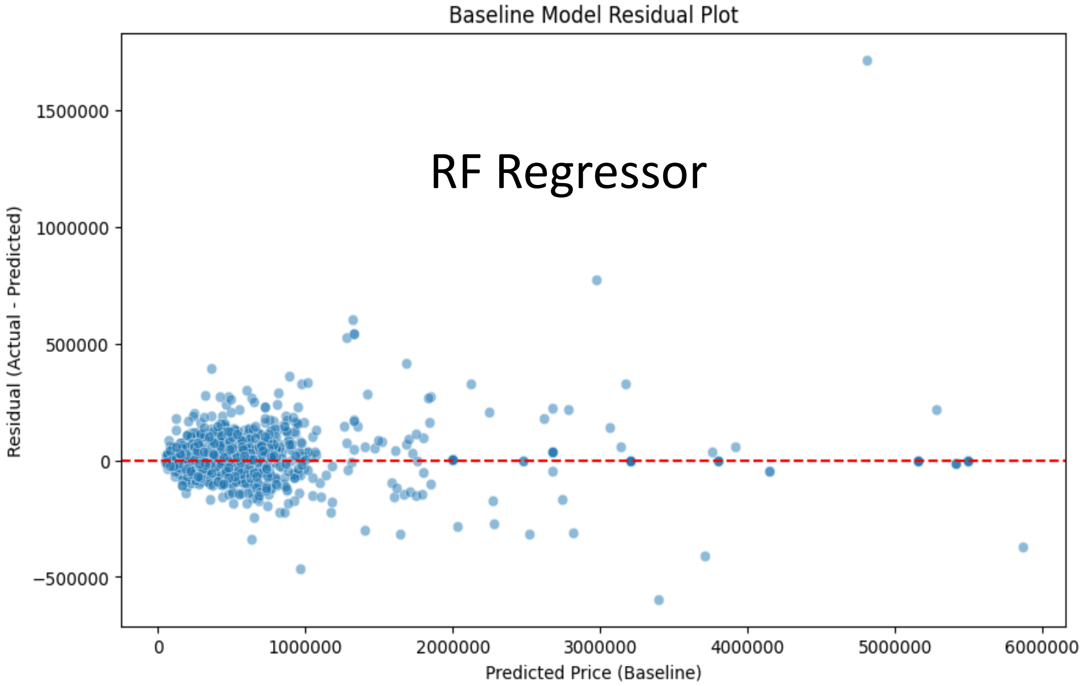
Model Development

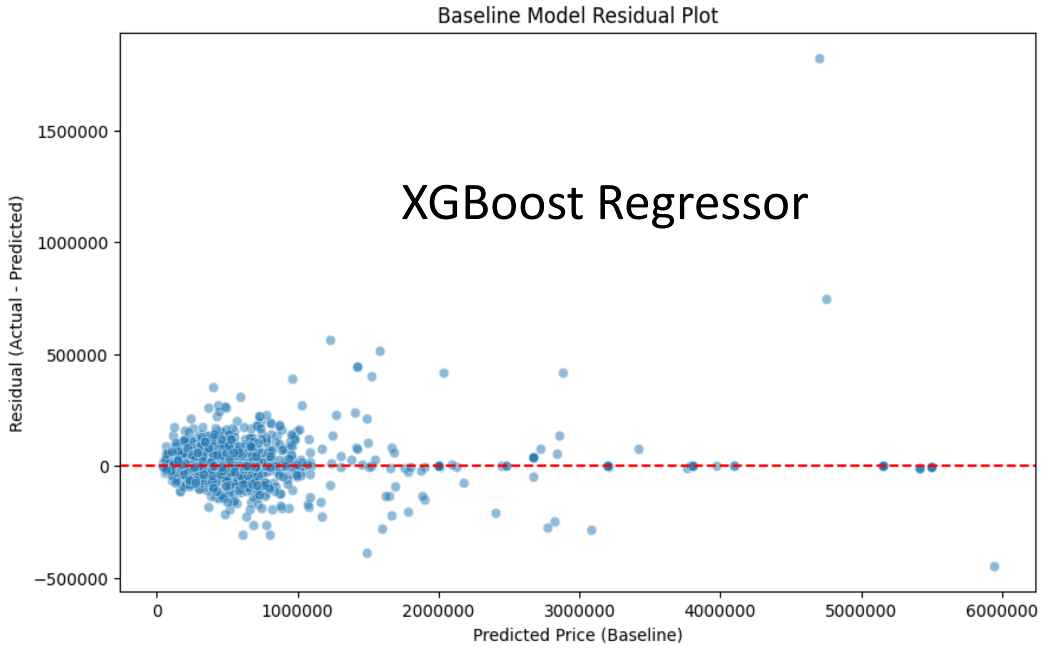
- Perform stratified train-test split (85% train, 15% test)

Models	Parameters	Train Accuaracy	Validation Accuracy
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	0.984	0.949

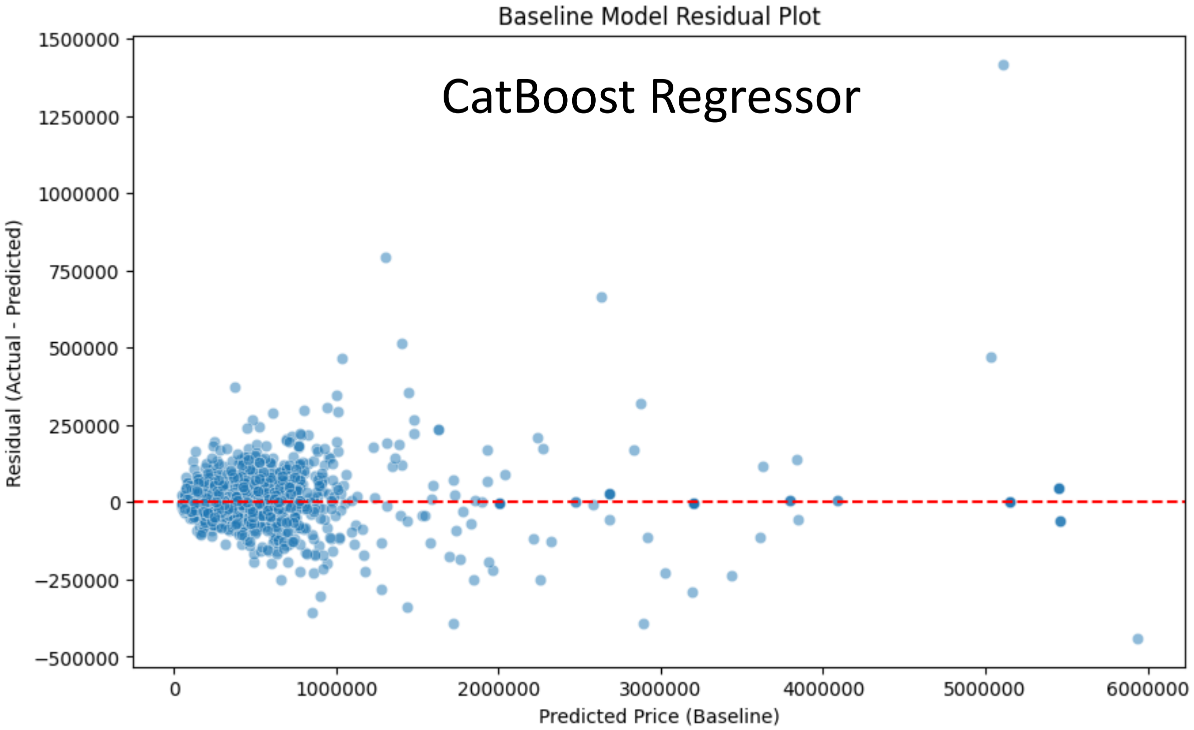
Model Development

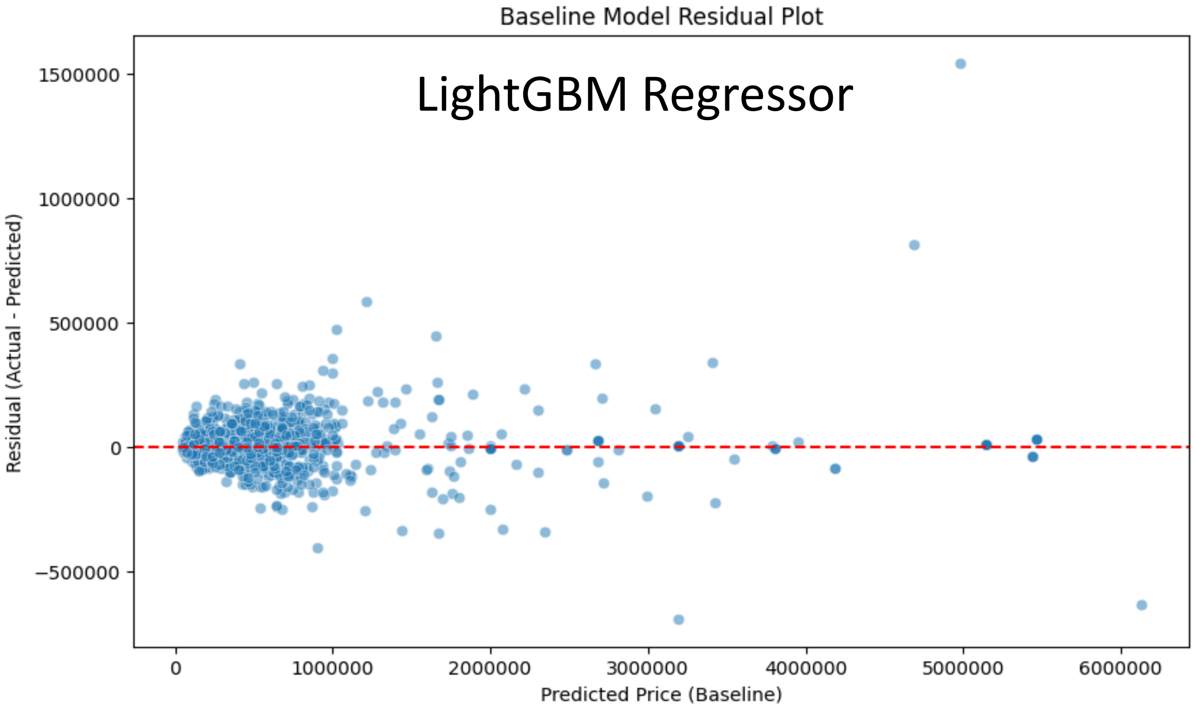






Model Development



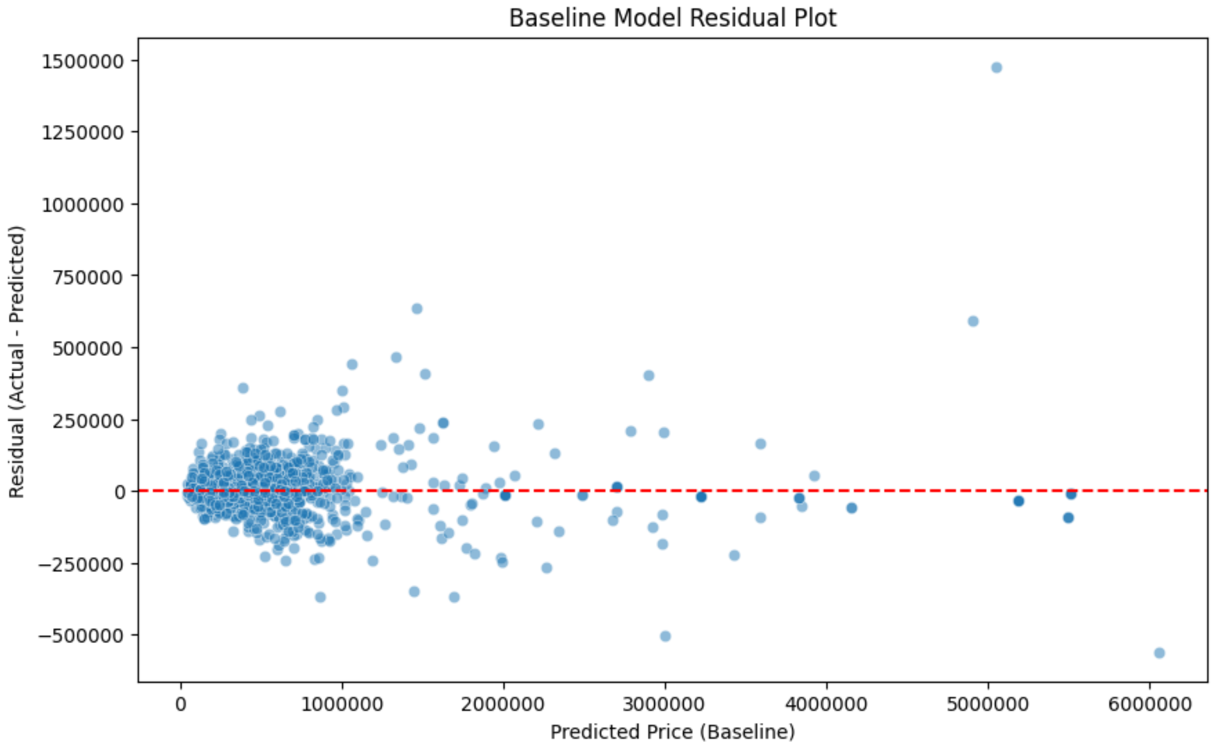


Stacking Ensemble

- **Base Models**
  CatBoost Regressor
  LightGBM Regressor
  XGBoost Regressor

- **Meta Model**
  Linear Ridge Regression

- **Accuracy**
  Training - 0.985
  Validation – 0.951



- K-fold validation with k=5, Out-of-Fold Cross Validation didn't improve the accuracy.
- Polynomial feature engineering, Target encoding (for make and model) also didn't improve accuracy

Data Science Workflow

1. Problem Definition
2. Data collection and preparation
3. Data Exploration
4. Data Modeling
5. Reflection or Inference Phase
6. Communicating & Visualizing Results

Interpretability & Uncertainty

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

- **Uncertainty Quantification**
  - Quantile regression used to generate prediction intervals.
  - Communicate results as ranges (e.g., ₹*X*–₹*Y*) rather than single point estimates

- **Business Value of Interpretability**
  - Stakeholders can see *why* a car is priced a certain way.
  - Confidence intervals build trust by showing realistic bounds.
  - Helps identify missing human-like signals (brand reputation, depreciation curves).
    - New brand car price can now be estimated using the performance parameters

Data Science Workflow

1. Problem Definition
2. Data collection and preparation
3. Data Exploration
4. Data Modeling
5. Reflection or Inference Phase
6. Communicating & Visualizing Results

Model Evaluation & Results

- **Evaluation Approach**
  - Train/test split (85/15) with cross-validation.
  - Metrics tracked: R^2, RMSE, MAE, prediction interval coverage.
  - Honest target encoding to avoid leakage.
- **Performance Outcomes**
  - Best Model: LightGBM with target encoding.
  - Train R^2: 0.984
  - Test R^2: 0.950
  - RMSE: ≈ ₹99, 649
  - MAE: ≈ ₹57,448
- Conclusion: Gradient Boosting reduced error by ~30%.
- Residual analysis shows predictions are unbiased.

GUI to predict the price

🚗 Used Car Price Predictor

Enter your car specifications below to get an estimated resale price range. The model uses advanced machine learning techniques with quantile regression to provide a realistic price range (lower, median, and upper estimates) with an R^2 score of 0.950.

Required Information

Car Name *
Full name including brand and model
Hyundai i20 Sportz

Manufacturing Year *
Year the car was manufactured
2016

Kilometers Driven *
Total kilometers the car has been driven
80000

Fuel Type *
Type of fuel
Petrol

Transmission *
Type of transmission
Manual

Owner *
Number of previous owners
Second Owner

Optional Information

Mileage (Optional)
Fuel efficiency with unit (kmpl for Petrol/Diesel, km/kg for LPG/CNG)
18.5 kmpl

Engine (Optional)
Engine displacement in CC
1197 CC

Max Power (Optional)
Maximum power output in bhp
83 bhp

Torque (Optional)
Torque specification with RPM
115Nm@ 4000rpm

Number of Seats (Optional)
Number of seats in the car
5

🚀 **Predict Price**

💰 **Estimated Resale Price Range**

Lower Estimate (5th percentile): ₹403,907 **Median Estimate (50th percentile):** ₹530,328 **Upper Estimate (95th percentile):** ₹619,037

Recommended Price Range: ₹403,907 - ₹619,037

Data Science Canvas

Project: Predictive Modelling for Used Car Pricing

Team: Manikanda Sakthi, Anfaal Obaid Waafy, Vimalraj K, Abhilasha Kawle

Problem Statement

Business Case & Value Added
Price Optimization for Dealerships accounting Depreciation and Quality Control

Value Add : Better financial planning and asset management

Data Landscape
we need data about
- The vehicle's attributes (make, model, year, fuel type, transmission, engine size),
- Its usage/condition (kilometers driven, number of owners)
- Pricing information (original price and selling price).
- Additional seller and location details help capture market variations and improve prediction accuracy.

Model Selection
- For used-car price prediction, regression models are most suitable to predict a continuous value.
- Linear Regression for baseline performance, then move to more powerful algorithms such as Random Forest, XGBoost, or Gradient Boosting, which handle nonlinear relationships and mixed data type
- Tree-based models generally perform best because they capture complex feature interactions without heavy preprocessing.

Model Requirements
- Complete ML pipeline
- Key preprocessing
- Feature engineering
- Various Models
- Evaluation

Software & Libraries
- Python 3.10+
- Core: pandas/numpy (data), sklearn (pipeline/impute/scale/models)
- ML: lightgbm (primary), category_encoders (target encoding), (interpretability), catboost

Skills
- Python programming & data manipulation (pandas/numpy)
- Feature engineering: parsing mixed formats, unit normalization, derived features
- ML preprocessing: imputation, scaling, encoding, pipelines
- Model development: regression (LightGBM/XGBoost), ensembles, CV, hyperparameter tuning
- Evaluation & interpretability: metrics calculation

Execution & Evaluation

Model Evaluation

Performance metrics: R^2, RMSE, MAE -> check train vs test consistency.
Residuals: look for bias or heteroscedasticity.
Feature drift: Correlation of price to car age, fuel type, etc.
Data quality: missing values, parsing errors, outliers.

How to interpret :
High R^2 but similar across train/test = good fit.

RMSE/MAE must be judged relative to average car price.
Residual plots reveal systematic under/overprediction.

Drift or poor coverage = recalibration needed.

Data Storytelling Target group Requirements:

Clarity: Simple, easy-to-read outputs.

Context: Metrics explained in business terms (errors in ₹).

Trust: Show validation steps and uncertainty ranges.

Actionability: Highlight key drivers (age, km, fuel).

Visuals: Simple GUI based dashboard

Effective Communication
1.Translate metrics into real-world meaning.
2. Use simple GUI based dashboard
3. Tell a clear story: problem -> solution -> impact.
4. Present predictions as ranges (₹X–₹Y) for confidence.

Data Collection & Preparation

Data Selection & Cleansing
- Companies like car dekho have collected this data historically
- Clean the data by handling missing values, correcting inconsistent entries, removing duplicates,

Data Collection
The data is already available from various sources. We have selected dataset from well established on-line re-sell car dealer – Car Dekho

Data Integration
Ingestion: Data is loaded from a static, consolidated CSV repository for batch processing.
Homogenization: Standardized units (bhp, CC, kmpl) across different manufacturers to ensure comparable numerical inputs.

Explorative Data Analysis
Target Distribution: selling_price was highly right-skewed. Applied Log-Transformation [log(1+x)] to normalize the distribution for regression.
Correlations: Vehicle_Age showed the strongest negative correlation (-0.71) with price.
Categorical Insights:
Transmission: Automatic cars command a significant price premium over Manual.
Fuel: Diesel cars retain higher resale value compared to Petrol/CNG.
Ownership: Price depreciation significantly accelerates after the "First Owner."

Future Work

- Data Expansion and Enrichment
- Real World Integration
- Business Extensions

Thank You