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: R², 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 R² 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."