Final Project Report

Title: *Auto-Insight: Predicting Ownership, Usage, Price and Insights from Indian Car Listings*

Authors: Binayak Chakraborty, Supratim Dey, Rohit Agarwal, Rituparno Chatterjee

Emails: binayakc@iisc.ac.in , supratimdey@iisc.ac.in , rohitagarwal@iisc.ac.in , rituparnoc@iisc.ac.in

Description of the dataset(s) and data preparation

- Source: Kaggle dataset (140K rows, 12 fields, ~11 MB CSV).

- Dataset: Kaggle (https://www.kaggle.com/datasets/milapgohil/car-dataset)

- Features: Brand, Model name, Year, Price, KM Driven, Fuel Type, Transmission, Owner, Car Age

- **Preprocessing:**

- Removed duplicates and invalid entries.
- Standardized units and formats.
- Handled missing values using median/mode imputation.
- Outlier detection and removal (e.g., extreme mileage or unrealistic prices).
- Converted categorical variables (fuel type, transmission, brand) into numeric encodings.

- **Feature Engineering:**

- Derived features such as car age, mileage per year, and "Car Points" (a composite score reflecting usage and depreciation).
- Brand category mapping for grouping similar models.

Methodology

- Exploratory Data Analysis (EDA)

- Distribution analysis of numerical fields (price, km driven, year).
- Correlation studies between price and features such as mileage, age, and brand.
- Visualization: scatterplots, boxplots, heatmaps to detect trends and anomalies.

- Feature Engineering

- Created derived variables (car age, mileage per year).
- Encoded categorical features.
- Designed "Car Points" metric to capture depreciation trends.

- Models Used

- Baseline: Linear Regression (interpretability).
- Tree-based models: Decision Tree, Random Forest.
- Boosting models: XGBoost, LightGBM, AdaBoost.
- Ensemble: Weighted ensemble combining multiple regressors.

- Evaluation Strategy

- Metrics: MAE, RMSE, R^2.
- Baseline comparison against mean/median predictor.
- Cross-validation to ensure stability across car segments.

Key Results

- **Initial results:** XGBoost achieved very high R^2 (0.998) but with suspiciously low error metrics, suggesting possible data leakage or overfitting.
- **Refined evaluation:** After hyperparameter tuning and feature re-evaluation, models achieved more realistic performance:
    - Random Forest, XGBoost, LightGBM, and AdaBoost all converged to $R^2 \approx 0.29$ with MAE ≈ 288K INR.
    - Linear Regression performed slightly worse ($R^2 \approx 0.26$).
    - Decision Tree and Weighted Ensemble underperformed significantly.

Insights:

- Car age, mileage, and brand are the strongest predictors of resale price.
- Transmission and fuel type have secondary but notable influence.
- Feature importance analysis highlights mileage per year and brand category as key drivers.

Limitations:

- Dataset bias: Kaggle dataset may not fully represent the Indian market.
- Model instability: High variance across folds and potential overfitting.
- Feature limitations: “Car Points” metric did not initially align with actual pricing trends.

Future Enhancements:

- Integration with real-time APIs (e.g., Cars24).
- Incorporation of neural networks for non-linear feature interactions.
- Allow user-uploaded datasets for broader applicability.

- Improved feature engineering (regional effects, demand trends).

Contributions by Each Team Member

- **Binayak Chakraborty:** Model training, validation & presentation.
- **Supratim Dey:** Model training, validation & Streamlit dashboard development.
- **Rohit Agarwal:** Data cleaning, exploratory data analysis & report.
- **Rituparno Chatterjee:** Data cleaning & exploratory data analysis.

Motivation and problem statement

The Indian used-car market is one of the fastest-growing automotive segments, yet it remains highly unstructured. Buyers often struggle to assess whether a listed price is fair, while sellers face uncertainty in setting competitive valuations. Traditional pricing methods rely heavily on subjective judgment, leading to inefficiencies, mistrust, and missed opportunities.

Our project, **Auto-Insight**, is motivated by the need to bring transparency and data-driven decision-making into this space. By analyzing large-scale car listing data, we aim to uncover the true factors that influence resale value—such as brand, mileage, year of manufacture, age, fuel type, and ownership history. This not only helps buyers avoid overpaying but also enables sellers to benchmark their vehicles against market trends.

The problem we address is twofold:

- **For buyers:** Lack of reliable guidance on whether a car is overpriced or undervalued.
- **For sellers:** Difficulty in setting fair prices that attract buyers while maximizing returns.

By applying exploratory data analysis (EDA) and machine learning (ML), Auto-Insight seeks to predict car prices accurately and generate actionable insights into ownership and usage patterns. Ultimately, the project contributes to standardizing valuations in the Indian used-car ecosystem, fostering trust and efficiency for all stakeholders.