Introduction

In today's data-driven automotive market, accurately estimating used car prices is essential for both buyers and sellers. The goal of this project is to develop a machine learning pipeline capable of predicting the resale value of used cars based on their technical, categorical, and derived features. We built a full-fledged Streamlit web application integrated with the trained ML model to provide real-time pricing predictions.

Project Workflow Overview

1. **Data Ingestion**
    - Scraped structured & nested data from Excel files (new_car_detail, new_car_specs, etc.)
    - Parsed JSON-like strings using ast.literal_eval() and flattened using custom logic
    - Consolidated data from **6 major cities** into a single DataFrame
2. **Data Cleaning & Preprocessing**
    - Removed missing/null values and non-uniform units (₹, Lakh, kmpl, etc.)
    - Extracted structured attributes (e.g., Color, Engine Type, Top Speed) from nested JSONs
    - Feature engineered variables like car_age = 2025 - model_year
3. **Exploratory Data Analysis (EDA)**
    - Visualized missing data, outliers, correlation heatmap, and feature distributions
    - Identified major influencing factors on car prices
    - Removed outliers using IQR (Interquartile Range) and optional capping for feature robustness
4. **Feature Engineering & Encoding**
    - Separated categorical & numerical columns
    - Used **One-Hot Encoding** for high-cardinality fields (brand, model, variant, etc.)
    - Final feature matrix had **785 columns** after encoding, optimized from initial 3300+

Model Development

Models Evaluated:

Model	R^2 Score	MAE (₹)	RMSE (₹)	CV R^2
Linear Regression	0.890	167,205	331,377	0.832
Ridge Regression	0.891	166,745	330,668	0.832

Model	R^2 Score	MAE (₹)	RMSE (₹)	CV R^2
Lasso Regression	0.890	167,311	331,691	0.827
Gradient Boosting	0.909	161,701	301,405	0.887
Random Forest	0.937	111,030	250,938	0.899
XGBoost	**0.944**	**98,500**	**234,800**	**0.912**

Final Pipeline:

- StandardScaler (for numeric feature scaling)
- XGBRegressor (best overall performance)
- Pipeline saved as XGBoost_best_car_price_model.pkl

Streamlit Web Application

Highlights:

- Fully dynamic dropdowns based on selected brand → model → variant
- Filters: fuel type, body type, gear box, drive type, city
- Numerical sliders for engine specs: mileage, power, acceleration, top speed
- One-click prediction and formatted result output
- Expandable panel to view input summary

Example Input UI:

Below is a sample input used to demonstrate the working of the Streamlit application:

Feature	Sample Input
Brand	Toyota
Model	Toyota Camry
Variant	Hybrid
Fuel Type	Petrol
Body Type	Sedan
Transmission	Automatic
Gear Box	6 Speed
Drive Type	FWD

Kilometers Driven	70,000
Model Year	2014
Engine Displacement	1949.5 cc
Mileage	19.16 km/l
Max Power	158.2 bhp
Torque	213 Nm
Top Speed	170 km/h
Acceleration	12.9 sec
Color	White
City	Kolkata
Car Age	11 years (Derived)

→ **Estimated Price: ₹ 10,75,870**

Model Evaluation Summary

- **MAE under ₹ 1 Lakh**: Considered excellent for used car price range
- **Feature Importance**: Brand, engine_displacement, model_year, and kms_driven most influential
- **Challenges**: Close predictions for different car types (e.g., sedan vs hatchback) suggest opportunity to enrich with insurance history, accident record, and customer sentiment

---

Deployment & Integration

- Model serialized using pickle & joblib
- All expected input columns stored as model_features_columns.pkl
- Streamlit interface runs locally or deployable on platforms like **Streamlit Cloud**, **Heroku**, or **AWS**

Conclusion

- Successfully built an **end-to-end ML system** with >94% R^2 score
- Deployed as a **real-time web application** using Streamlit
- Reusable modular ML pipeline with clean input handling

Future Scope

- Integrate live price listings from **CarDekho API** or **web scraping**
- Include **image analysis** of vehicle condition (deep learning)
- Push to production using **Docker**, **CI/CD**, and **cloud hosting**