**Project Report: Car Dheko - Used Car Price Prediction**

Name: Dinesh V

Date: October 20, 2025

Project: End-to-End Used Car Price Prediction Application

**1. Executive Summary**

This report details the development of an end-to-end machine learning solution for **Car Dheko**, designed to predict the price of used cars. The primary business objective was to enhance customer experience and streamline sales by providing an accurate, data-driven pricing tool.

A machine learning pipeline was successfully built, beginning with data collection and extensive preprocessing. An **XGBoost Regressor** model was trained and selected for its high performance, achieving an **R-squared ($R^2$) value of over 0.90**.

The final model was deployed as an interactive web application using **Streamlit**. This application allows users (customers or sales staff) to input a car's features and receive an instant, accurate price estimate, along with data visualizations to provide market context.

**2. Introduction & Problem Statement**

**2.1. Business Problem**

The used car market is characterized by high price volatility and information asymmetry. Customers often lack confidence in pricing, and sales representatives rely on manual heuristics, leading to inconsistent and suboptimal price setting.

Car Dheko required a tool to standardize this process, leveraging historical sales data to provide accurate and transparent price estimations.

**2.2. Project Objectives**

1. **Analyze** historical car sales data to identify key factors influencing price.
2. **Develop** a high-accuracy machine learning regression model to predict used car prices.
3. **Deploy** the model as a user-friendly, interactive web application for easy access by non-technical users.

**3. Data Methodology**

**3.1. Data Sourcing**

The project utilized a consolidated dataset (structure\_cars\_final.csv) containing historical data on used car listings.

* **Key Features:** oem (make), model, modelYear, kms (kilometers driven), fueltype, transmission, bodytype, ownerNo, and price (target variable).

**3.2. Data Preprocessing**

The raw data was noisy and required significant cleaning to be suitable for modeling. This process was performed in the datapreprocessing.ipynb notebook.

1. **Handling Missing Values:** Imputed or removed null values where appropriate.
2. **Type Conversion:** Converted critical columns like price and kms from object/string types to numerical integers, removing non-numeric characters (e.g., "kms", "₹", ",").
3. **Outlier Removal:** Identified and removed extreme outliers using the **Interquartile Range (IQR)** method. This step was crucial for preventing the model from being skewed by unrealistic data points.
4. **Final Dataset:** The cleaned and prepared data was saved as cars\_eda.csv to serve as the master dataset for all subsequent analysis and modeling.

**3.3. Exploratory Data Analysis (EDA)**

EDA was performed to uncover insights and validate hypotheses about the data. These insights were later integrated into the Streamlit application's dashboard.

* **Price vs. Kilometers:** A strong negative correlation was observed; as kilometers driven increase, the car's price generally decreases.
* **Price vs. Body Type:** SUVs and Sedans were found to have a higher average price point compared to Hatchbacks.
* **Price Distribution:** The target variable (price) was right-skewed, which was a key consideration during modeling.

**4. Modeling & Evaluation**

This phase (modelling.ipynb) focused on preparing features and selecting the best-performing model.

**4.1. Feature Engineering**

1. **Categorical Features:**
   * bodytype was encoded using LabelEncoder.
   * All other nominal categorical features (e.g., oem, fueltype) were transformed using OneHotEncoder to create binary columns.
2. **Numerical Features:**
   * All numerical features (e.g., kms, modelYear, engine\_cc) were standardized using StandardScaler to bring them to a common scale, improving model convergence and performance.

**4.2. Model Selection & Training**

Several regression algorithms were trained and evaluated (e.g., Linear Regression, Decision Tree, Random Forest). The **XGBoost Regressor** was selected as the final model.

* **Reasoning:** XGBoost (Extreme Gradient Boosting) is a powerful algorithm that consistently delivers high performance. It excels at capturing complex, non-linear relationships between features, which is characteristic of pricing data.

**4.3. Performance Evaluation**

The model was evaluated on a held-out test set (20% of the data) to simulate real-world performance.

* **R-squared ($R^2$):** The final model achieved an **$R^2$ score greater than 0.90**. This indicates that over 90% of the variance in used car prices is explainable by the features included in the model, demonstrating a very high level of accuracy.
* **Root Mean Squared Error (RMSE):** The model also achieved a low RMSE, signifying that its price predictions are, on average, very close to the actual sale prices.

**5. Deployment**

The trained model and all preprocessing objects (scaler, encoders) were saved as .pkl files for production use.

* **Application:** A customer-facing web application was built using **Streamlit** (import pickle.py).
* **User Interface:** The app provides a simple and intuitive form where a user can select a car's make, model, year, fuel type, and other features.
* **Functionality:**
  1. User inputs car details.
  2. The app loads the saved preprocessing pipelines to transform the user's input.
  3. The saved best\_xgb\_model.pkl model predicts the price.
  4. The application instantly displays the **Estimated Price in Lakhs (₹)**.
  5. The dashboard also includes interactive charts from the EDA (e.g., "Price by Body Type," "Distance vs. Price") to give the user valuable context about the market.

**6. Conclusion & Future Work**

**6.1. Conclusion**

This project successfully delivered an end-to-end, data-driven solution for used car price prediction. The final Streamlit application provides an accurate, reliable, and user-friendly tool that solves the core business problem, empowering both customers and sales staff with transparent pricing.

**6.2. Future Work**

* **Data Expansion:** Integrate data from more cities and include more features (e.g., accident history, color, features) to improve accuracy.
* **Automated Retraining:** Create a pipeline to automatically retrain the model on new sales data to prevent model drift.
* **Feature Enhancement:** Add a feature to the app that shows listings of "similar cars" to the one being evaluated.