Optimal Used Car Pricing and Recommendation System

A Machine Learning Approach

GROUP MEMBERS

Patricia Ngugi
Jeffrey Kanyi
Pamela Okinyo
Wafula Simiyu
Faith Mutisya
Brian Kariuki
Angela Kalelwa

Outline:

Introduction

Project Overview and Business

Research Framework

Research questions, objectives, and model selection.

Data & Modelling

Data sources, feature engineering, and model development.

Deployment

Deploying the system via a web app.

Conclusion

Overview of the system's anticipated benefits and impact.

Introduction

Introduction

The used car market is growing rapidly, but it often lacks transparency due to inconsistent pricing across regions, brands, and vehicle conditions. This project addresses these challenges by using machine learning to predict optimal prices and recommend similar vehicles based on user preferences. This approach aims to help users make informed choices quickly and with greater confidence.

Problem Statement

The used car market presents unique challenges, such as: Price Discrepancies, Variations in price by location, brand, age, and condition make it difficult for buyers and sellers to evaluate fair prices. Lack of Guidance: Buyers often lack tools to find comparable options, resulting in suboptimal purchase decisions.

OBJECTIVE:

To develop a machine-learning-powered system that predicts a car's market value based on its attributes, enhancing pricing accuracy and operational efficiency.

DOMAIN:

This project lies at the intersection of automotive resale, insurance services, and machine learning.

BUSINESS UNDERSTANDING:

Accurate car valuation is critical for insurance companies and car resellers. Insurance firms require precise appraisals to set premiums and calculate claim settlements, while resellers rely on competitive pricing for profitability. Current methods often lack reliability and consistency. The proposed system leverages historical car data, including attributes like make, model, mileage, and condition, to train machine learning models such as linear regression, decision trees, and gradient boosting. The solution offers real-time, user-friendly appraisals, improving decision-making, operational efficiency, and stakeholder trust.

TARGET AUDIENCE (Stakeholders)

The primary users of the system are:

- **Individual Sellers:** People who wish to sell their used cars and need an estimate of their vehicle's worth.
- **Car Buyers**: People looking to purchase used cars who need to know if the listed price is fair.
- **Car Dealerships:** Businesses that need a reliable tool for pricing to maintain profitability and customer satisfaction.
- **Financial Institutions and Insurance Companies:** Organizations that provide car loans or insurance policies and require accurate car valuations for underwriting and risk assessment.
- **Online Car Marketplaces:** Platforms that list used cars and could use the system to display fair price estimates for users.

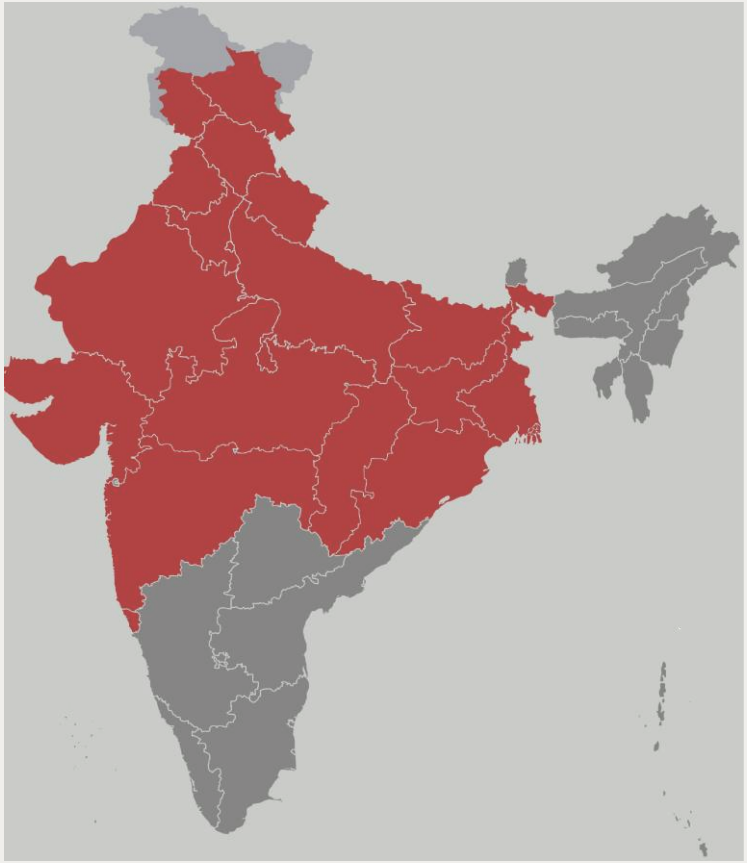
SOURCES OF DATA

- The data for this project was scraped from **CarDekho.com**, India's leading online automotive marketplace.
- CarDekho is renowned for providing detailed car specifications, expert reviews, comparisons, and price insights for a broad range of car models and brands.
- The platform features both new and used car listings, enabling users to explore options and make well-informed purchasing decisions.
- CarDekho has partnerships with numerous auto manufacturers, dealerships, and financial institutions, contributing to its extensive inventory and comprehensive car data.
- CarDekho also offers a rich, immersive experience with tools like the "Feel The Car" feature, enabling 360-degree interior and exterior views, along with live offers in various cities.

The data scraped includes used car listings from several major cities in India.

Locations covered:

- Ahmedabad
- Mumba
- Bangalore
- Gurgaon
- Jaipur
- Hyderabad
- Pune
- Kolkata
- Chennai



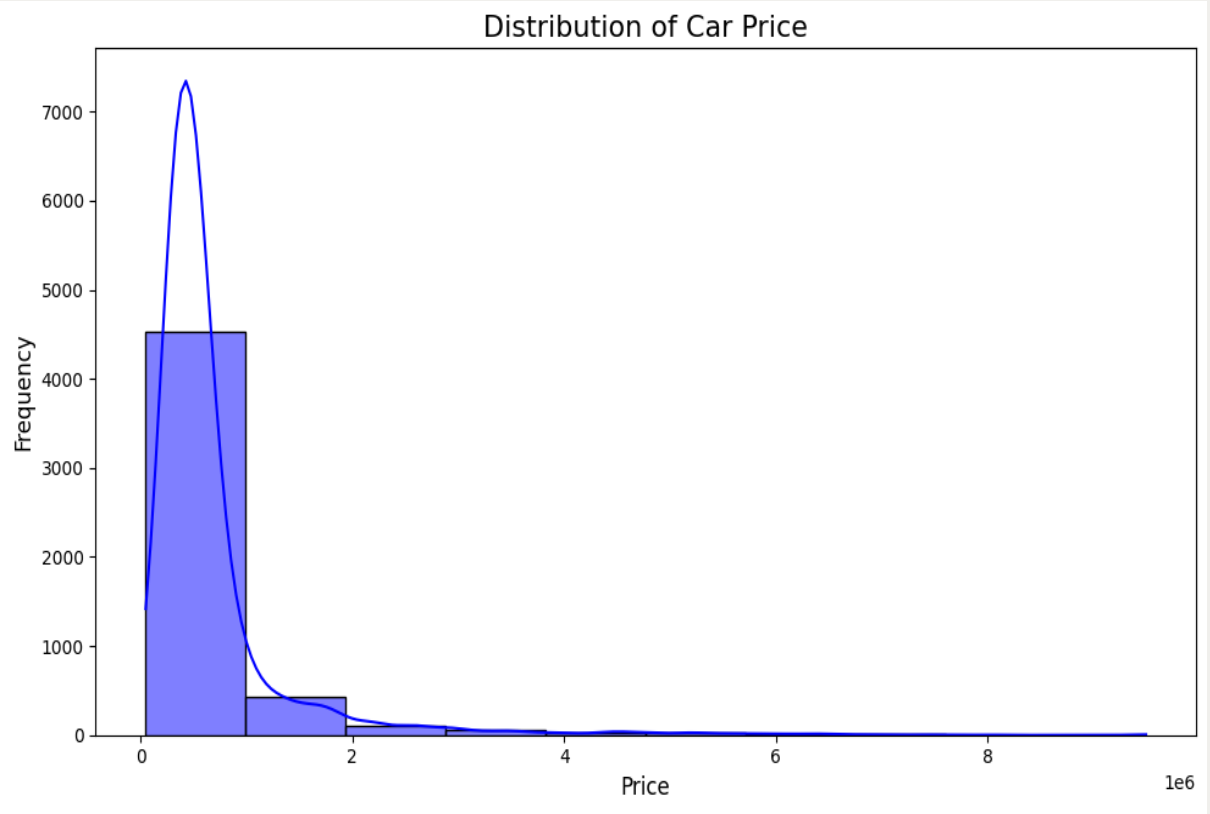
Dataset Characteristics
- Rows: 5197.
- Columns: 19.

Methodology

- **Understand Stakeholder Needs:** Identify business goals and key metrics.
- **Explore the Data:** Analyze and clean the data for patterns and insights.
- **Data Preprocessing:** Encode and prepare data for modeling.
- **Feature Selection:** Identify the most relevant features for prediction.
- **SelectKBest:** Selects the top k features with the highest scores.
- **Model Building & Training:** Create, train, and tune the models
- **Model Deployment & Monitoring:** Deploy the model through Streamlit and track its performance.

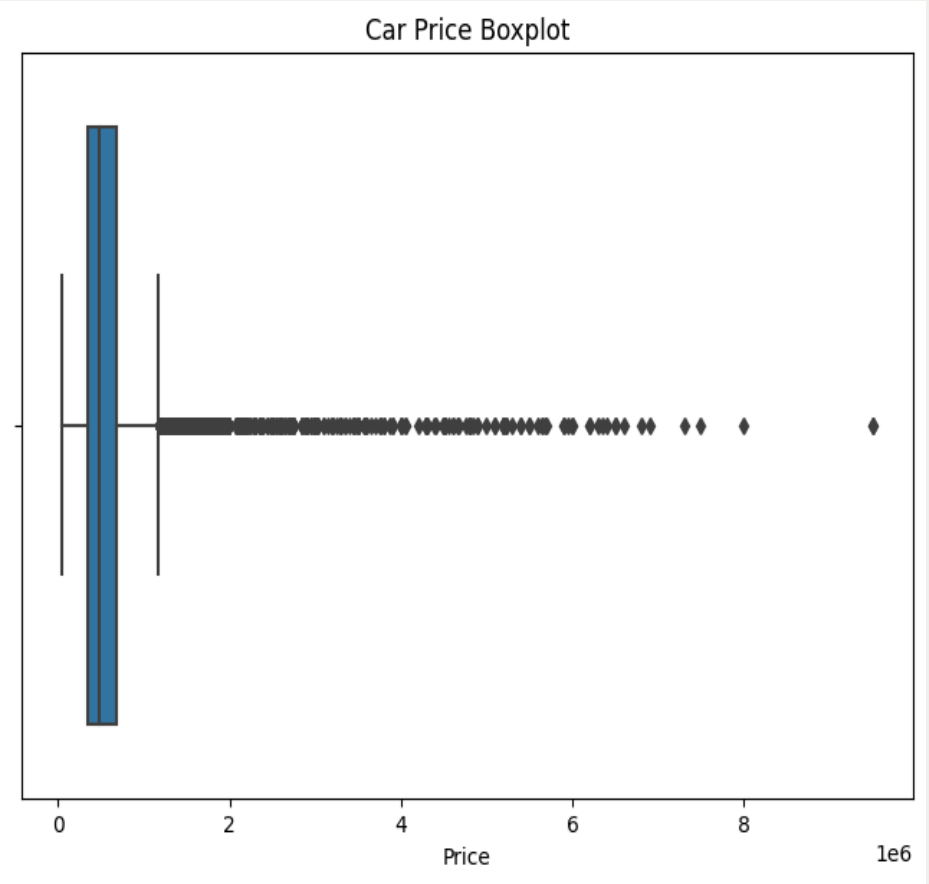
- **Handling Missing Values:**
  - Replaced NaN and infinite values with the median of respective columns to maintain data integrity.
  - Imputing columns with missing values with mode and median
  - Dropped irrelevant columns and rows
- **Outliers:**
  - Identified and treated outliers in key numerical features to prevent distortion in model predictions.
- **Standardization:**
  - Applied StandardScaler to standardize features, ensuring uniform scales for better model performance.
- **Final Dataset:**
  - Cleaned dataset with no missing or infinite values.
  - Ready for feature engineering and modeling.

Data Visualization



- Price distribution is right-skewed meaning most cars are priced lower with few cars at the higher end of the price spectrum.
- Peak at lower Price indicate low-priced cars.
- The long tail towards the higher end of the price range indicates presence of a few luxury, high-end cars.

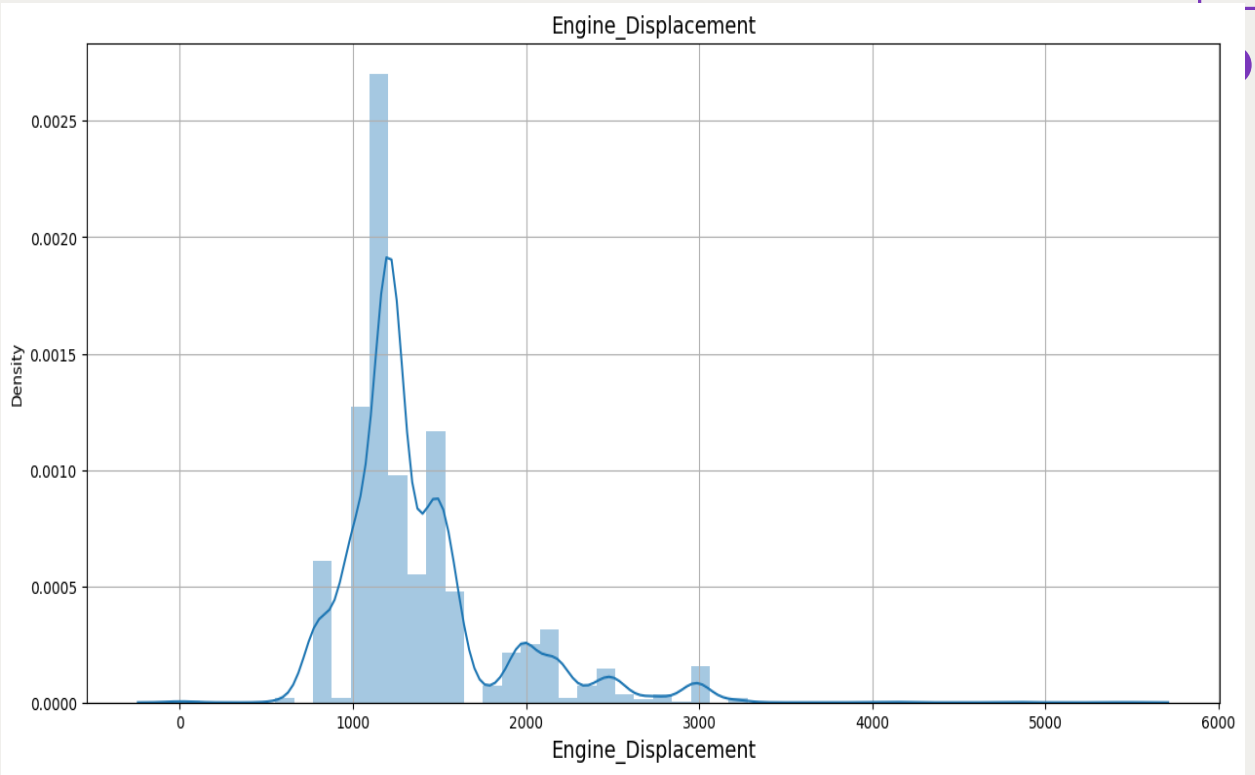
Car Price



- Most cars are priced lower since the majority of cars are clustered towards the lower price range..
- A few cars are very expensive since there are a few outliers with significantly higher prices.
- Price distribution is skewed meaning the data is unevenly distributed.

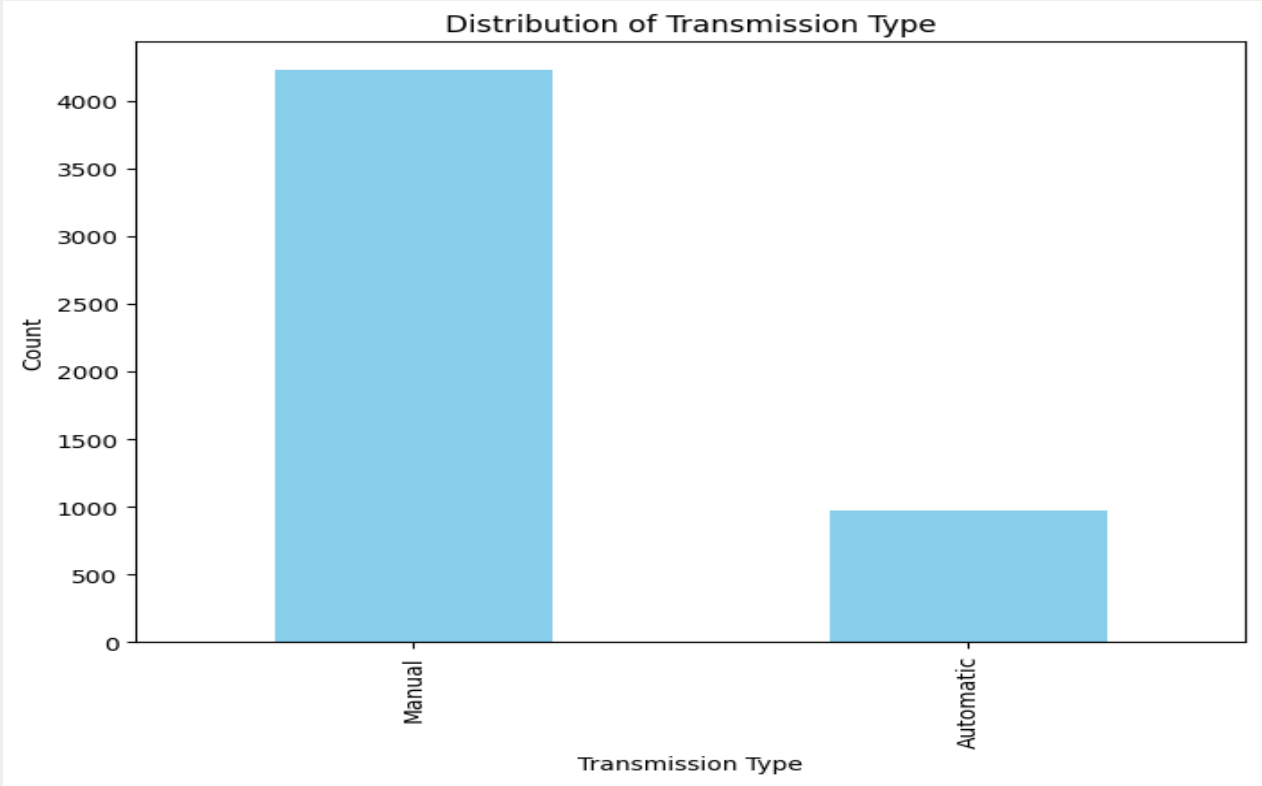
Engine Displacement

- Engine Displacement is having positive skewness
- Most of the cars having displacement in the range 1000-2000 CC.

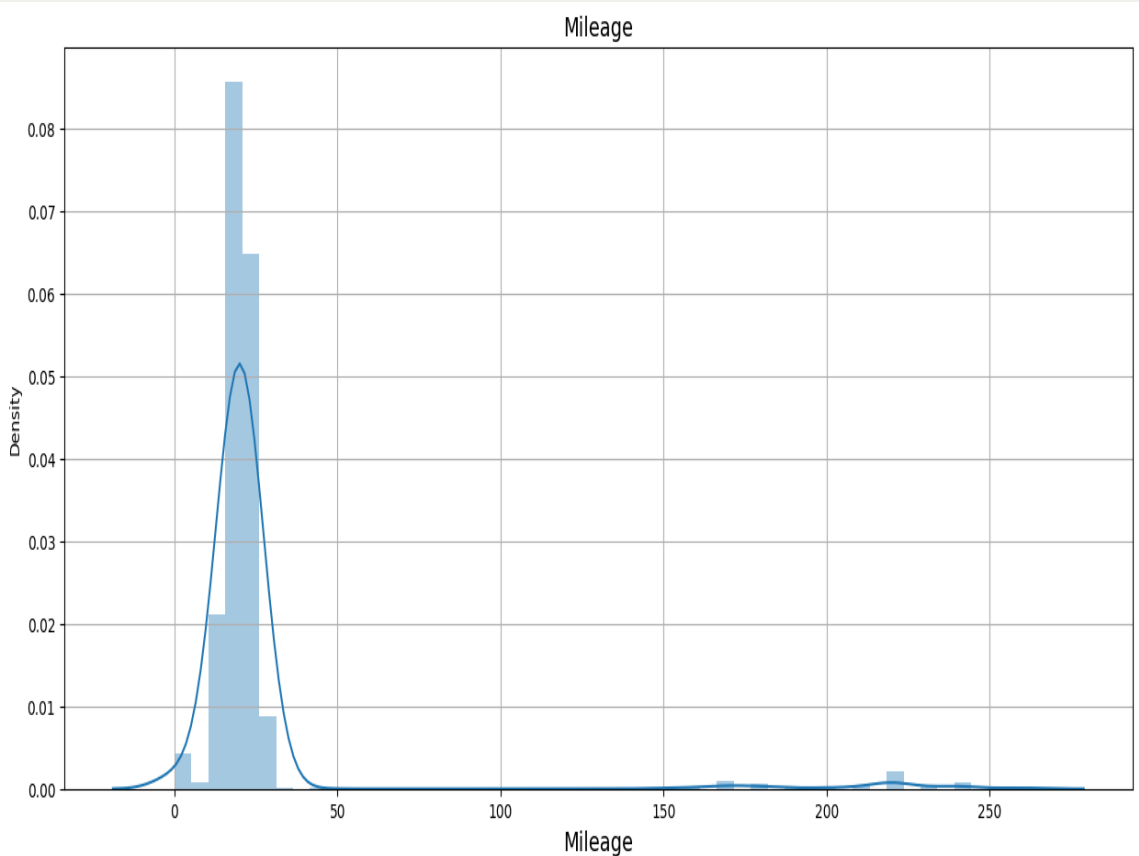


Transmission Type

- Maximum cars having Manual Transmission type, we can say a larger portion of the market prefers manual transmission cars, possibly due to factors like cost, fuel efficiency, or driving preference.
- We can say people are selling their Manual cars.

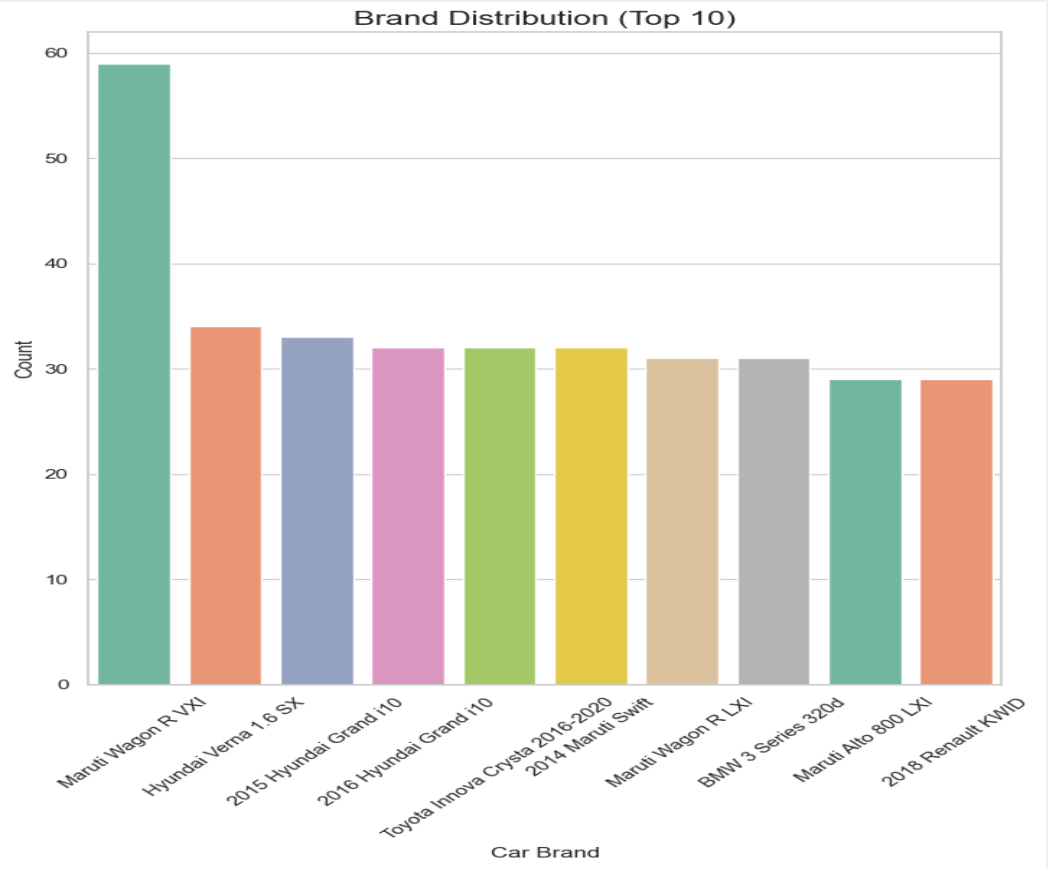


Mileage



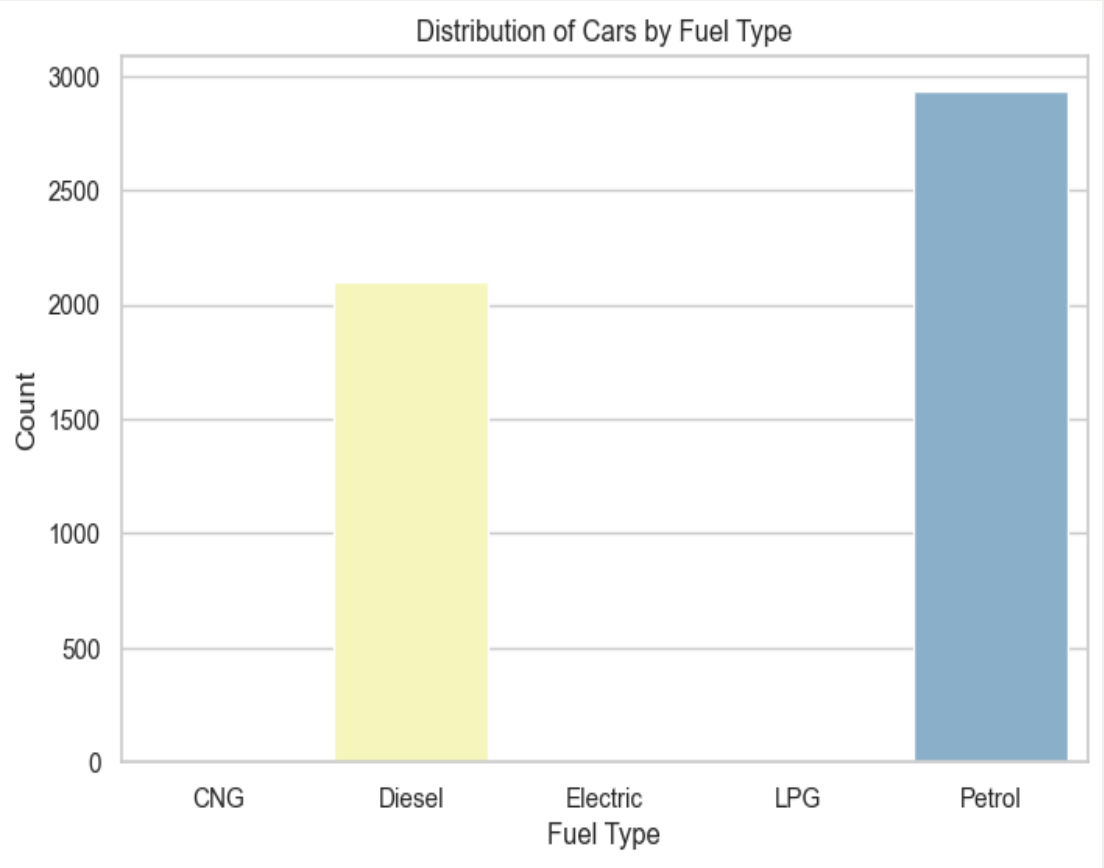
- Few cars having mileage more than 50 which not possible in old cars, these data are outliers.

Brand



- Dominance of Maruti and Hyundai brands indicate their popularity in the market.
- Inclusion of BMW 3 Series shows presence of premium car brands.
- Presence of newer models in the graph suggests preference of recent models.

Fuel



- The dominance of Diesel and Petrol suggests a preference for these fuel types in the market.

Feature Engineering & Selection

Missing and Infinite Values:

Handled by replacing NaN and infinite values with median, mean and mode.

Standardization:

Applied **StandardScaler** and **SelectKBest** to standardize and select top key features with the highest scores for our model.

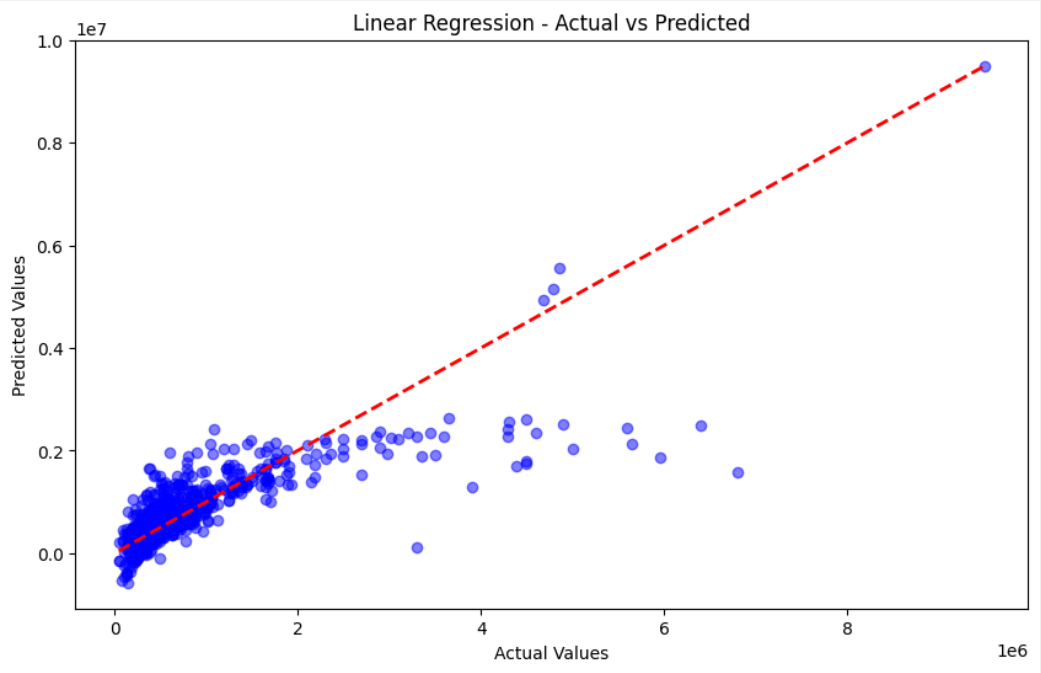
Feature Selection:

- **SelectKBest**: Selected 18 features using **f_regression** for statistical relevance.
- **Variance Inflation Factor (VIF)**: Checked for multicollinearity and ensured robustness in linear models.
- **PCA:** Reduced dimensionality by transforming features into principal components retaining as much variance as possible.

Model Performance - Linear Regression

Evaluation Results:

- **MSE:** 249,642,455,102.72
- **RMSE:** 499,462.17
- **R^2:** 0.641
- **MAE:** 262,972.37



Analysis:

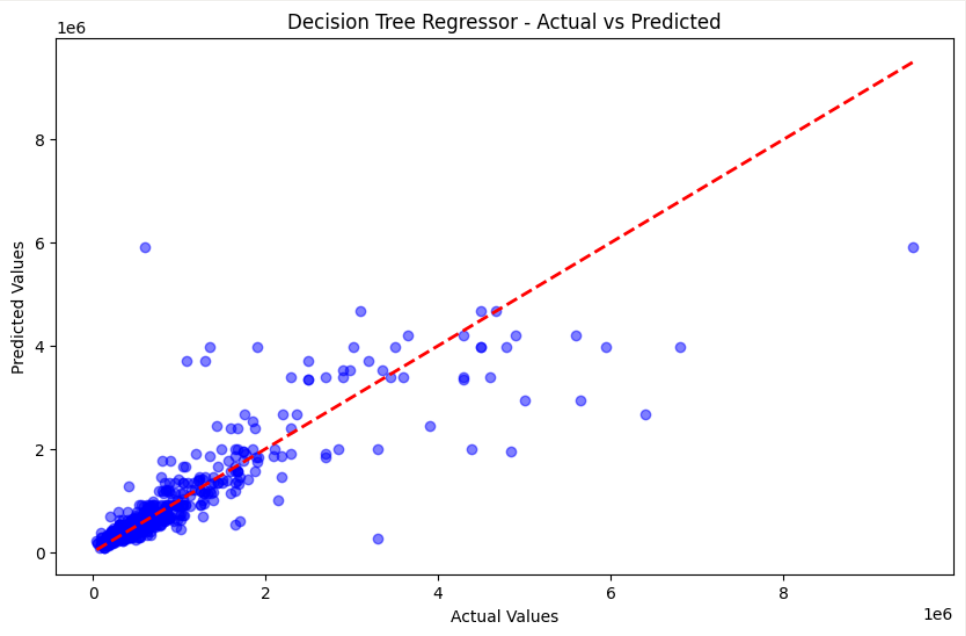
- R-squared is 0.641, which suggests the model explains about 64.1% of the variability in car prices is explained by the predictors. This is a decent result but shows room for improvement.

- The MSE and RMSE are quite large, indicating that the model's predictions have a significant deviation from the true values.

Model Performance - Decision Tree Regressor

Evaluation Results:

- **MSE**: 174,933,708,732
- **RMSE**: 418,250.77
- **R^2**: 0.748
- **MAE**: 160,572.33



Analysis:

- The R-squared value of 0.748 indicates that the model explains about 74.8% of the variance in the target variable, which is an improvement over Linear Regression.

- The RMSE and MSE are still large, but much lower than the Linear Regression's suggesting better prediction accuracy and smaller errors.

- The MAE is much lower than Linear Regression, indicating more accurate predictions on average.

Model Performance - Random Forest Regressor

Evaluation Results:

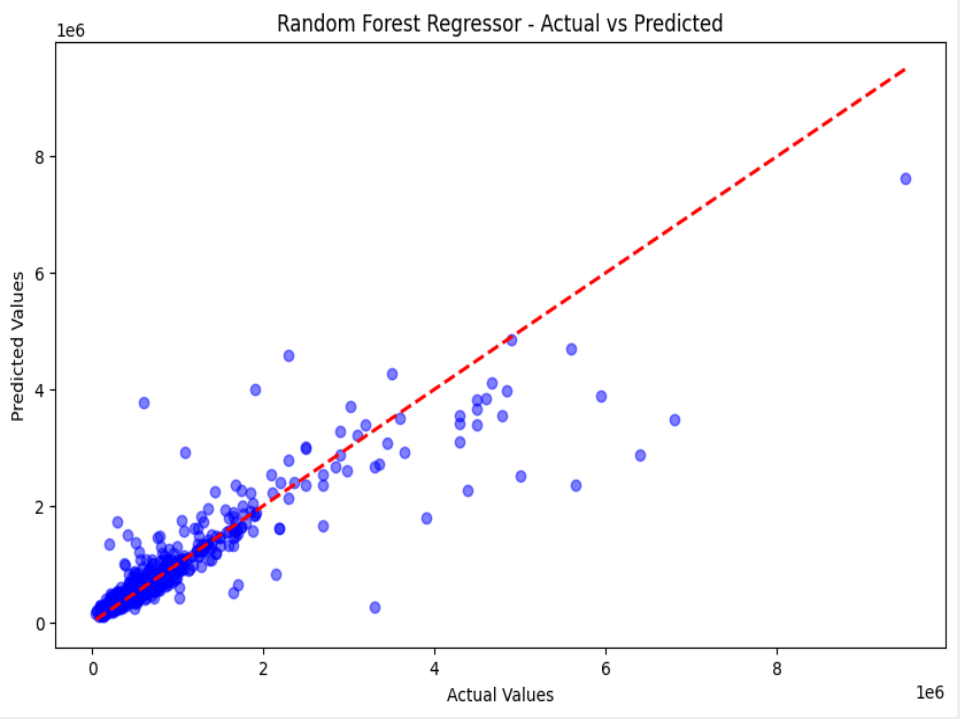
- **MSE:** 128,568,690,257
- **RMSE:** 358,564.76
- **R^2:** 0.8148
- **MAE:** 137,983.14

Analysis:

- The R-squared value of 0.8148 shows that the model explains about 81.5 % of the variance, which is a significant improvement over both Linear Regression and Decision Trees.

- The Random Forest Regressor shows even lower MSE and RMSE compared to both Linear Regression and Decision Tree models, suggesting that it has the smallest average prediction error.

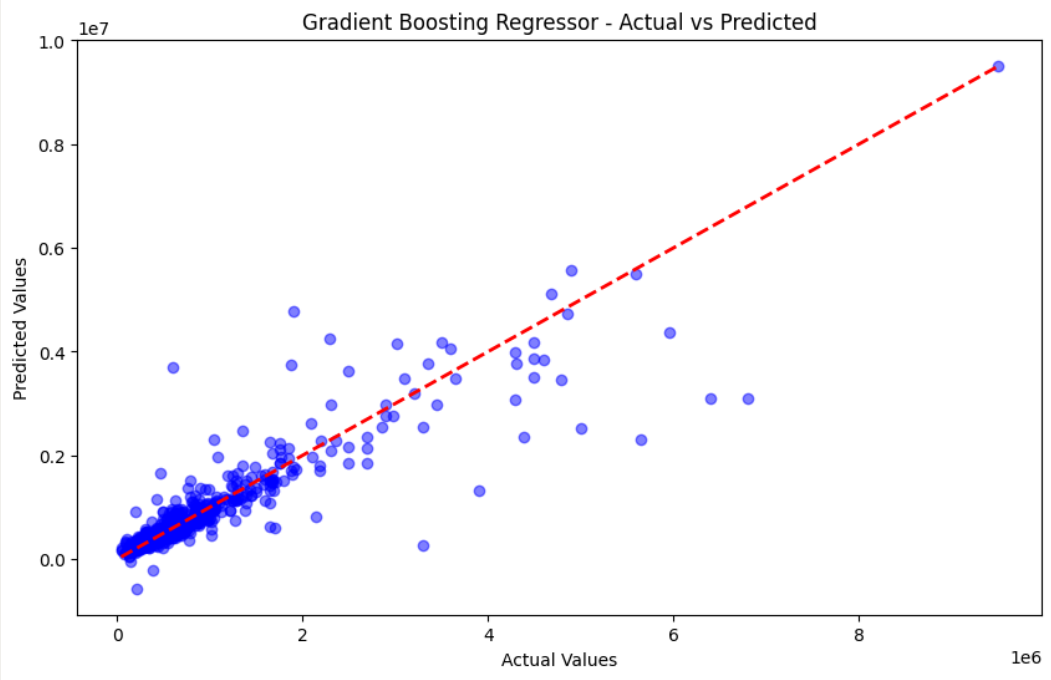
- The MAE is also lower than both previous models, suggesting that, on average, the Random Forest model makes smaller errors in its predictions.



Model Performance - Gradient Boosting Regressor

Evaluation Results:

- **MSE:** 128,268,327,433
- **RMSE:** 358,145.68
- **R^2:** 0.8153
- **MAE:** 140,197.49



Analysis:

- The R-squared value of 0.8153 indicates that the model explains about 81.5% of the variance in the target variable, the highest among all models.

- The Gradient Boosting Regressor has very similar MSE and RMSE values to the Random Forest model, indicating a similar level of prediction error.

Model Comparison & Conclusion:

- Gradient Boosting Regressor and Random Forest Regressor are best performing models.
- Decision Tree also shows strong performance but still slightly less effective than the Random Forest and Gradient Boosting models.
- Linear Regression has the highest MSE, RMSE, and MAE among all models, indicating its lower predictive performance compared to the others therefore the weakest model.

Final Model Selection

Best Model for Deployment: **Gradient Boosting Regressor (with SelectKBest)**.

Outperforms all other models in terms of RMSE, R^2, and MAE.

Provides a solid balance between accuracy and computational efficiency.

Model Selection Rationale:

Gradient Boosting with SelectKBest shows a strong fit for the data, with significant improvements over baseline models.

Deployment

- **Deployment Method:** Streamlit web app.
- **Model to Deploy:** Gradient Boosting Regressor (with SelectKBest for feature reduction).
- **Functionality:**
  - User inputs: Car features like mileage, age, fuel type, engine displacement, acceleration, transmission, brand.
  - Model outputs: Predicted car price.
  - Future enhancement: Real-time recommendations of similar cars based on user inputs.

Conclusion

Key Achievements

- Successfully developed a model that predicts various values of preferred vehicles and vehicle attributes.
- Successfully cleaned the data using various data cleaning methods to help evaluate performance of the model.
- Visualized various key attributes to help the end user to select their best recommended preference.
- Delivered insights on how car features influence the market price.
- Deployed the model using Streamlit for easy user access.
- End user is able to successfully identify a car market value based on its attributes

Recommendations:

- Since newer cars are more expensive, stakeholders should prioritize them in marketing and sales and offer discounts or trade-ins on older versions.
- They should market vehicles with larger engines, torque and max power.
- They should focus on selling low mileage cars or offer discounts on high mileage.
- Prioritize cars with fewer previous owners and fuel efficient.
- Cars with high accelerations can be marketed to specific target market.

THANK YOU!