Summary of Findings

Key Insights from Exploratory Data Analysis (EDA)

The exploratory analysis provided valuable insights into the factors influencing car prices. It was observed that car prices vary significantly across the dataset, indicating the presence of both economy and luxury vehicles. Technical specifications such as engine size, horsepower, curb weight, and vehicle dimensions showed a strong positive relationship with price, suggesting that larger and more powerful vehicles tend to be more expensive.

Fuel efficiency variables, including city and highway mileage, exhibited a negative correlation with price, indicating that high-end vehicles often prioritize performance and comfort over fuel economy. The analysis also revealed the presence of outliers, particularly in price and engine-related features. These outliers represent premium vehicles and reflect realistic market behavior rather than data quality issues.

Categorical analysis further showed that brand, body style, fuel type, and engine type have a noticeable impact on pricing. Premium brands and luxury body styles consistently demonstrated higher average prices, emphasizing the role of brand value and design in the automobile market.

Features with the Most Impact on Price Prediction

Based on correlation analysis and model performance, the most influential features for predicting car prices include engine size, horsepower, curb weight, vehicle dimensions (length, width, and wheel base), car brand, body style, and fuel type. Among these, engine-related and weight-related features proved to be the strongest predictors, highlighting that mechanical capacity and vehicle build are key drivers of price. Brand-related features also played a significant role, reflecting consumer perception and market positioning.

Challenges Faced During Preprocessing and Modeling

Several challenges were encountered during the preprocessing and modeling stages. Missing values were present in both numerical and categorical features, requiring careful handling to avoid information loss. Additionally, some numerical attributes were stored as object data types, which had to be converted before analysis.

Outliers posed another challenge, particularly for regression models sensitive to extreme values. Encoding categorical features increased the dimensionality of the dataset, making feature selection more important. Distance-based models such as KNN required proper feature scaling, without which performance deteriorated significantly. Furthermore, Decision Tree Regression showed a tendency to overfit, which required hyperparameter tuning to improve generalization.

Additional Steps for a Larger Dataset

If a larger and more diverse dataset were available, several enhancements could be made. Advanced feature engineering techniques could be applied to capture complex relationships between variables. Regularization methods such as Ridge and Lasso regression could help control model complexity. Ensemble techniques like Random Forest and Gradient Boosting could be explored for improved predictive performance. Incorporating external factors such as market demand, resale value, and regional pricing trends would further strengthen the model. Additionally, more extensive cross-validation could be used to ensure robust and reliable results.

Conclusion

This study demonstrates that car prices are influenced by a combination of mechanical specifications, design characteristics, and brand value. Through systematic data exploration, preprocessing, and model evaluation, the KNN Regression model achieved the best predictive performance. The project highlights the importance of high-quality data preparation and appropriate model selection in building effective machine learning solutions.