# Used Car Price Prediction - Final Report

Alan Tom Akkara

09/27/2025

## Problem Statement

The used car market is vast, dynamic, and influenced by a wide range of factors such as vehicle age, mileage, fuel efficiency, drivetrain, transmission type, and even exterior or interior color. Buyers often struggle to determine whether a listed price is fair, while sellers and dealerships face the challenge of setting competitive prices that maximize profit without discouraging potential buyers.

Traditional pricing approaches rely heavily on manual appraisal, heuristics, or outdated reference guides, which often fail to capture real-time market dynamics. This leads to inefficiencies such as overpriced vehicles remaining unsold, underpriced vehicles selling too quickly without maximizing revenue, and general mistrust between buyers and sellers.

The objective of this project is to build a predictive model that can accurately estimate the price of used cars based on their features. By leveraging data-driven machine learning techniques, the model aims to reduce uncertainty in car valuation, support dealerships in pricing strategy, and empower buyers with transparent, fair market estimates.

## Overview of Dataset

The dataset used in this project consists of approximately 750,000 used car listings, containing a mix of numerical and categorical features that describe each vehicle, its history, and seller-related information. The target variable for prediction is the car price.

During the initial review, the dataset was found to have missing values in several columns and some inconsistencies were also present in a few columns. Through data wrangling, these issues were resolved and the dataset was cleaned and organized into a structured format.

## Data Wrangling

During the data wrangling stage, the dataset was first inspected for missing values and inconsistencies. Rows with incomplete or irrelevant information were either removed or corrected to ensure data quality. Duplicate entries were also identified and dropped to avoid bias in the analysis.

Next, categorical variables such as fuel type, transmission, and brand were converted into numerical representations to make them compatible with machine learning models. Outliers, particularly in features like mileage and price, were examined and handled appropriately to reduce their impact on predictions.

Finally, the dataset was standardized by aligning units and ensuring consistency across features. After completing the data wrangling process, the dataset was refined to 690,370 rows × 17 columns.

## Exploratory Data Analysis

The exploratory data analysis focused on understanding the relationships between features and the target variable, price, and identifying patterns that could inform model development. Distributions of numerical features such as car age, mileage, and MPG were examined, revealing trends like older cars and higher mileage generally being associated with lower prices.

Categorical variables, including transmission type, drivetrain, and colors, were analyzed to determine their impact on price. Visualizations such as boxplots for top categories and scatterplots for key numerical features were used to identify patterns and potential outliers. A correlation

heatmap of numeric features highlighted which variables had the strongest linear relationships with price, providing insights for feature selection.

The EDA stage helped in identifying key trends, potential data issues, and the most influential features, laying the groundwork for preprocessing and model training.

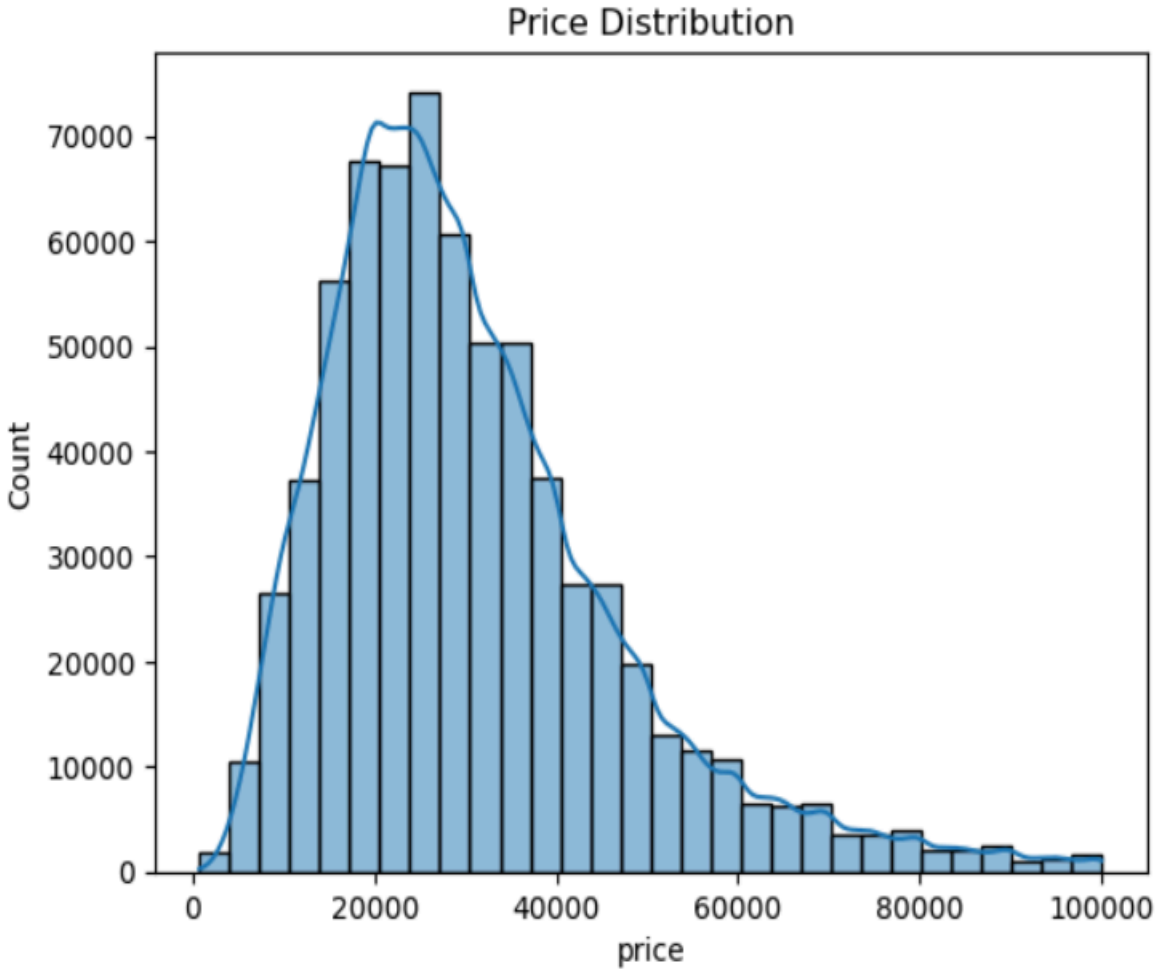


Figure 1: Distribution of Car Prices.

This histogram illustrates the distribution of used car prices across the dataset. Most vehicles are clustered around the mid-price range, indicating that the majority of cars are reasonably priced for typical buyers. There are fewer cars at the very low or very high ends, representing outliers or luxury/specialty vehicles. By visualizing the spread and frequency of prices, we can better understand the overall

market trends and identify extreme values that may need special handling during modeling. This figure is essential for recognizing the natural variation in car prices and setting realistic expectations for predictive performance.

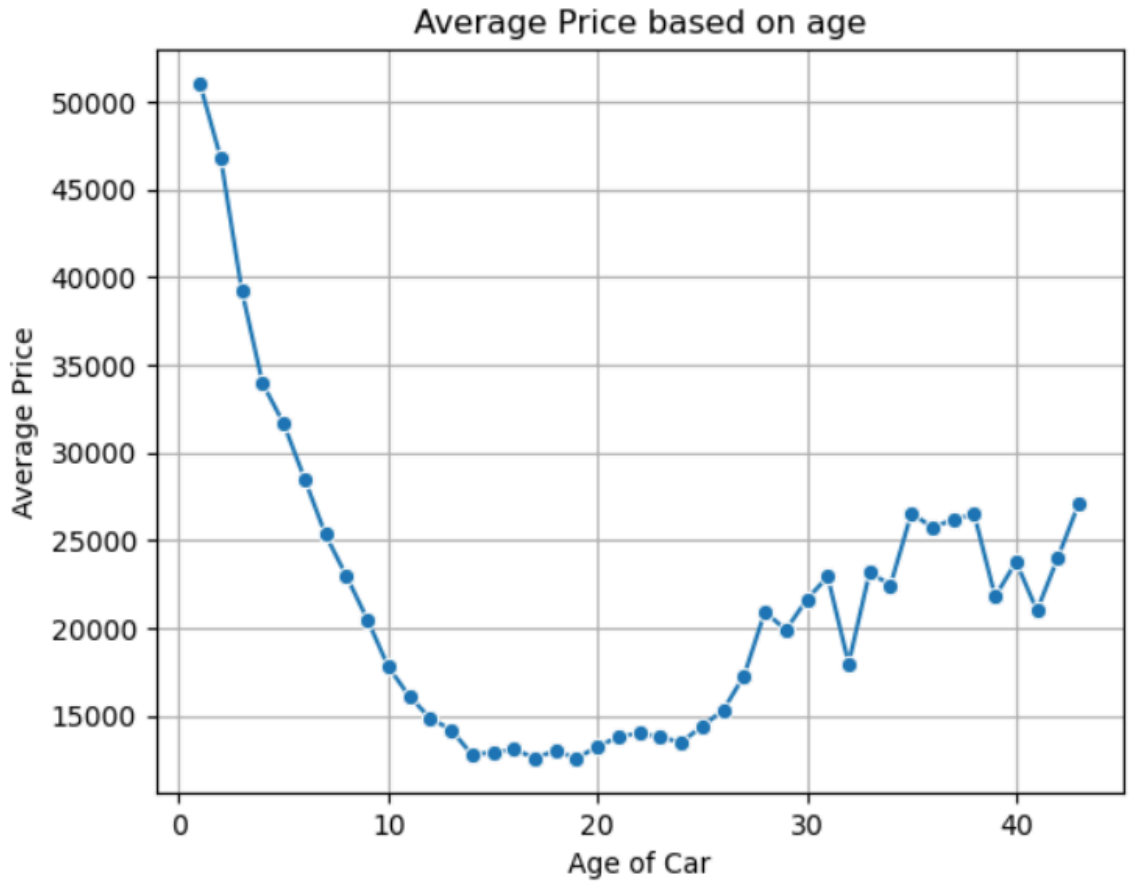


Figure 2: Relationship between car age and price.

This line plot shows the mean price of cars at each age. As expected, the average price decreases steadily as car age increases, reflecting depreciation over time. Interestingly, after a certain age, the trend slightly reverses, with some older cars showing higher average prices. This may be due to classic, vintage, or well-maintained vehicles that retain value despite their age. The plot confirms that car age is a strong predictor of price while also highlighting exceptions that may require additional consideration in modeling.

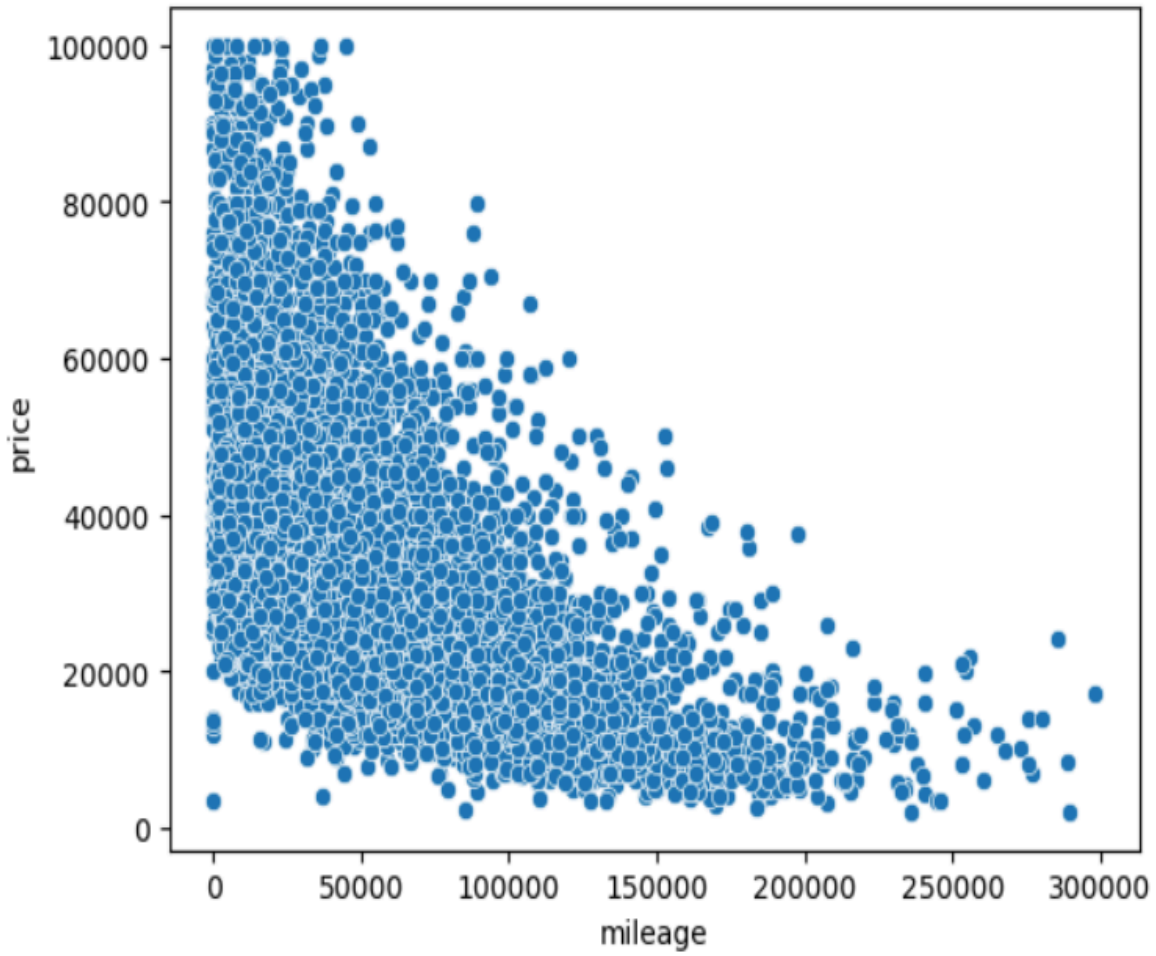


Figure 3: Relationship between car mileage and price.

This plot illustrates the relationship between a car's mileage and its listed price. As expected, there is a clear negative trend: cars with higher mileage generally have lower prices. The plot also shows some variability in prices at similar mileage levels, which can be attributed to differences in features such as brand, age, condition, and fuel type. By aggregating or sampling the data, the trend becomes more visible without the clutter of individual points, making it easier to interpret. This visualization confirms that mileage is a strong predictor of price and should be included in the predictive model.

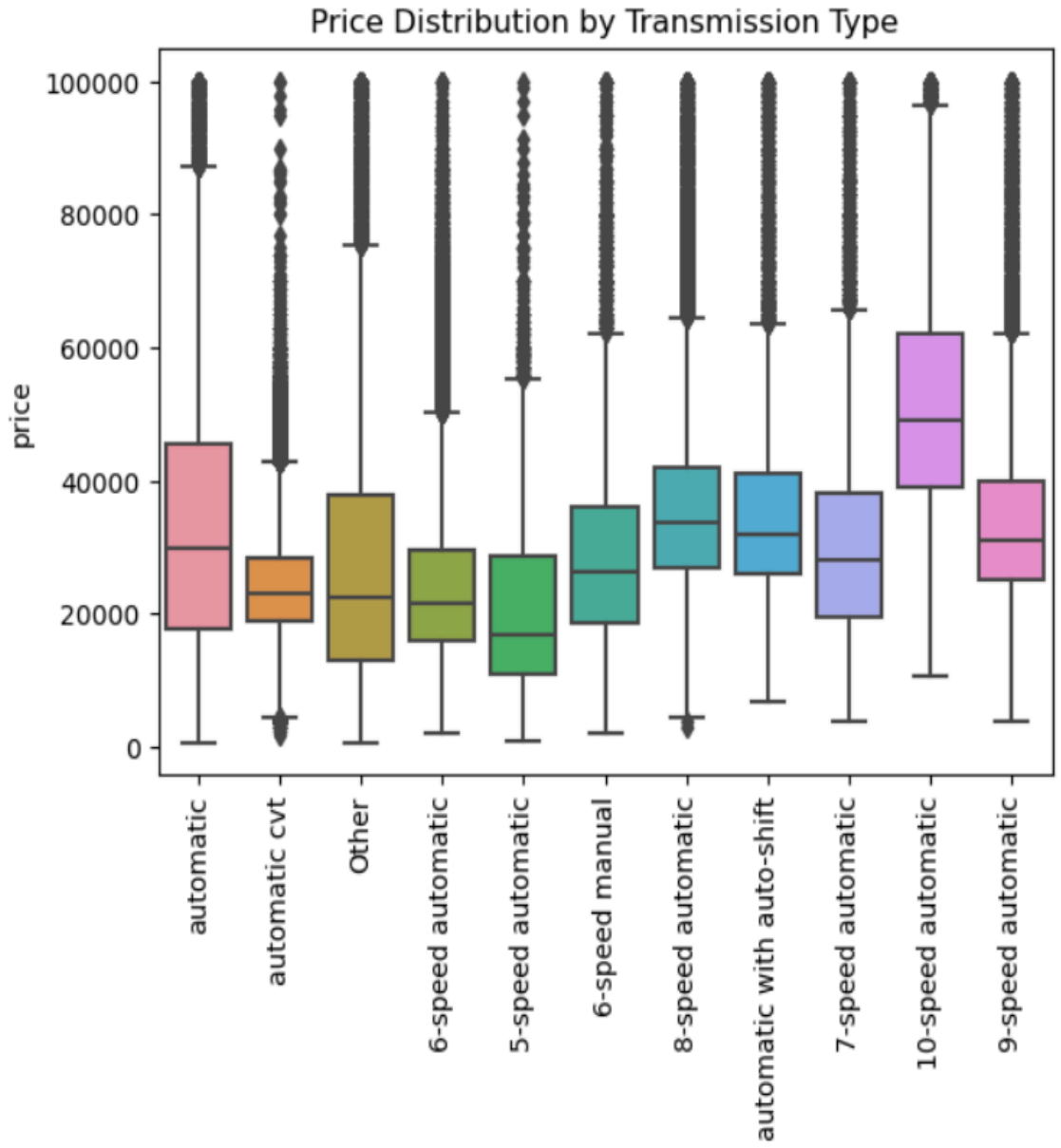


Figure 4: Comparison of price distributions across the most common transmission types.

The boxplot compares price distributions for the top 10 transmission types, grouping all less common types into an “Other” category. Automatic transmissions tend to show higher median prices, while manual and less common transmissions have lower medians. The plot also shows the range and spread of prices within each transmission type, highlighting variability caused by other factors such as car age, model, and mileage. This figure demonstrates how a categorical feature like transmission can influence pricing and justifies the need to encode such variables for machine learning.

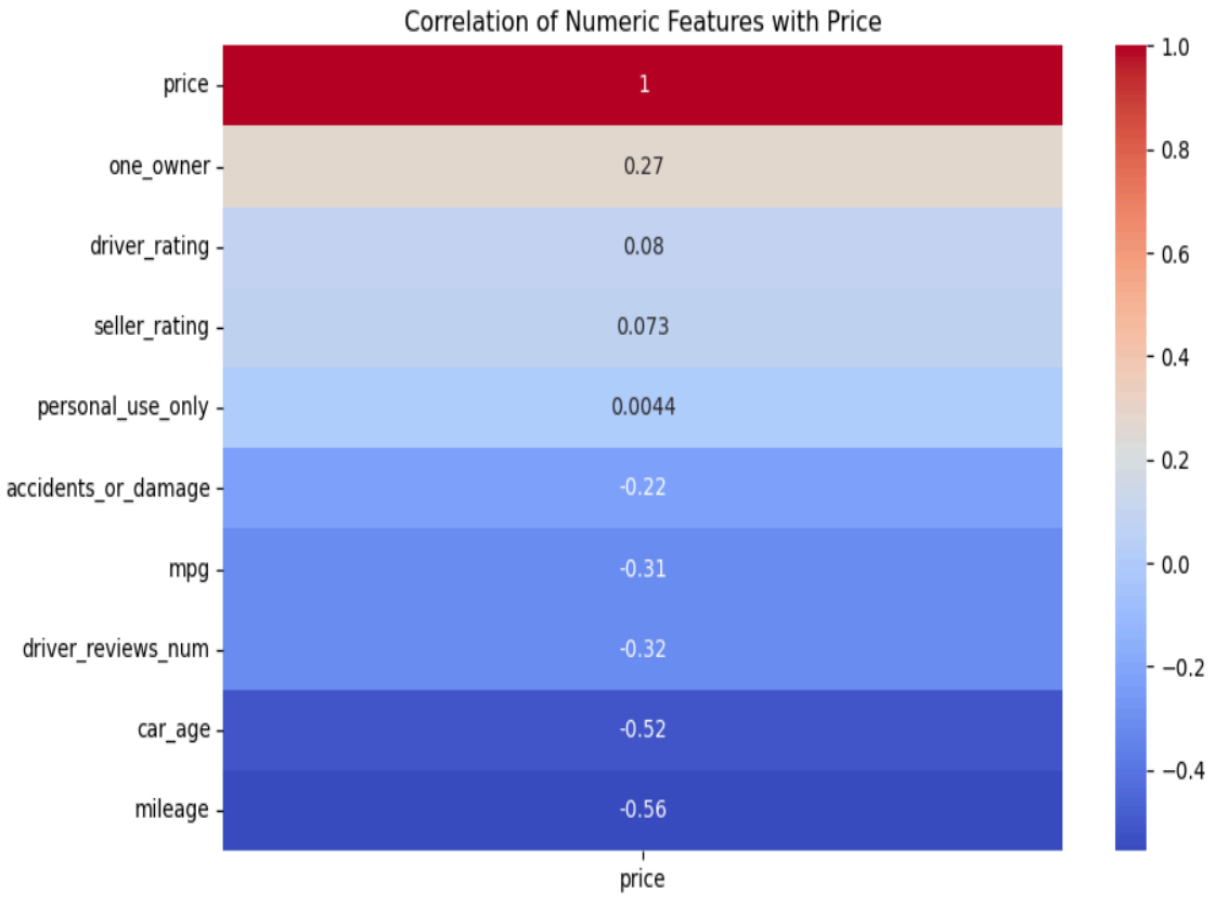


Figure 5: Heatmap of correlations between numeric features and price.

The heatmap visualizes the strength and direction of linear relationships between numeric features and the target variable, price. Features such as car age and mileage show strong negative correlations, indicating that higher values in these features tend to correspond to lower prices. Positive correlations may appear for features like MPG or certain ratings. By highlighting which numeric features are most strongly associated with price, this figure provides guidance for feature selection and helps detect multicollinearity, ultimately improving model performance and interpretability.

## Model Selection

To identify the best-performing model, multiple algorithms were trained and evaluated on the dataset, including Linear Regression, Gradient Boosting, Random Forest, XGBoost, and LightGBM. Each model was

compared based on its ability to minimize prediction error and maximize explanatory power, using performance metrics such as **Root Mean Squared Error (RMSE), Mean Absolute Error (MAE),** and $R^2$ score.

- Linear Regression served as the baseline, achieving an $R^2$ of ~0.71, but it struggled to capture the nonlinear relationships in the data.

- Gradient Boosting showed improvement with an $R^2$ of ~0.79, but still fell short compared to more advanced ensemble methods.

- Random Forest and XGBoost both demonstrated strong performance, achieving $R^2$ scores above 0.92 and 0.93 respectively, though Random Forest exhibited signs of overfitting.

- LightGBM delivered the best overall results, with a test $R^2$ of 0.94 and RMSE of ~3857, outperforming all other models. Additionally, cross-validation confirmed its stability and generalizability.

Given its superior accuracy, scalability, and interpretability (through feature importance scores), LightGBM was selected as the final model for predicting used car prices.

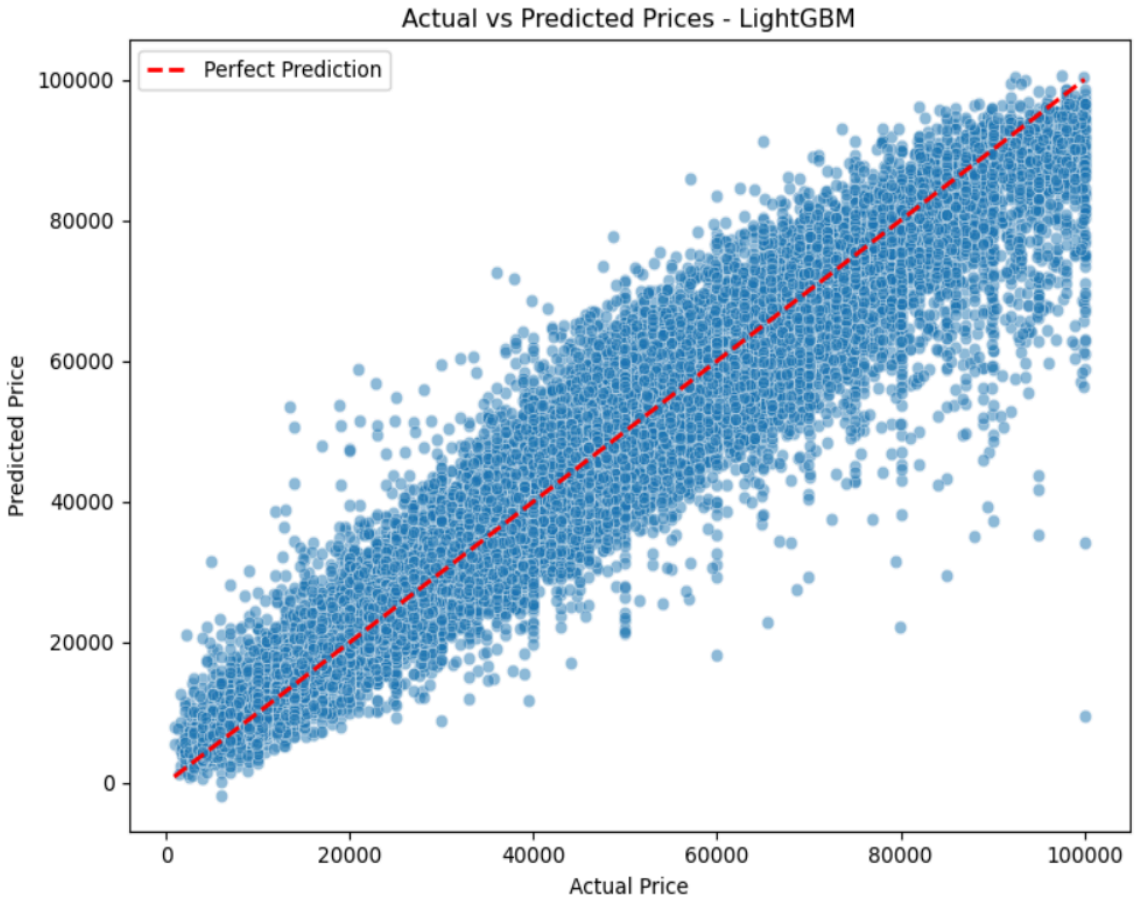


Figure 6: Scatter plot of actual vs predicted car prices using LightGBM.

This plot compares the actual car prices with the prices predicted by the LightGBM model. Most of the points align closely along the diagonal line, indicating that the model is highly accurate in capturing real market values. While a few outliers exist where predictions deviate significantly, the overall trend confirms the model's strong predictive power.

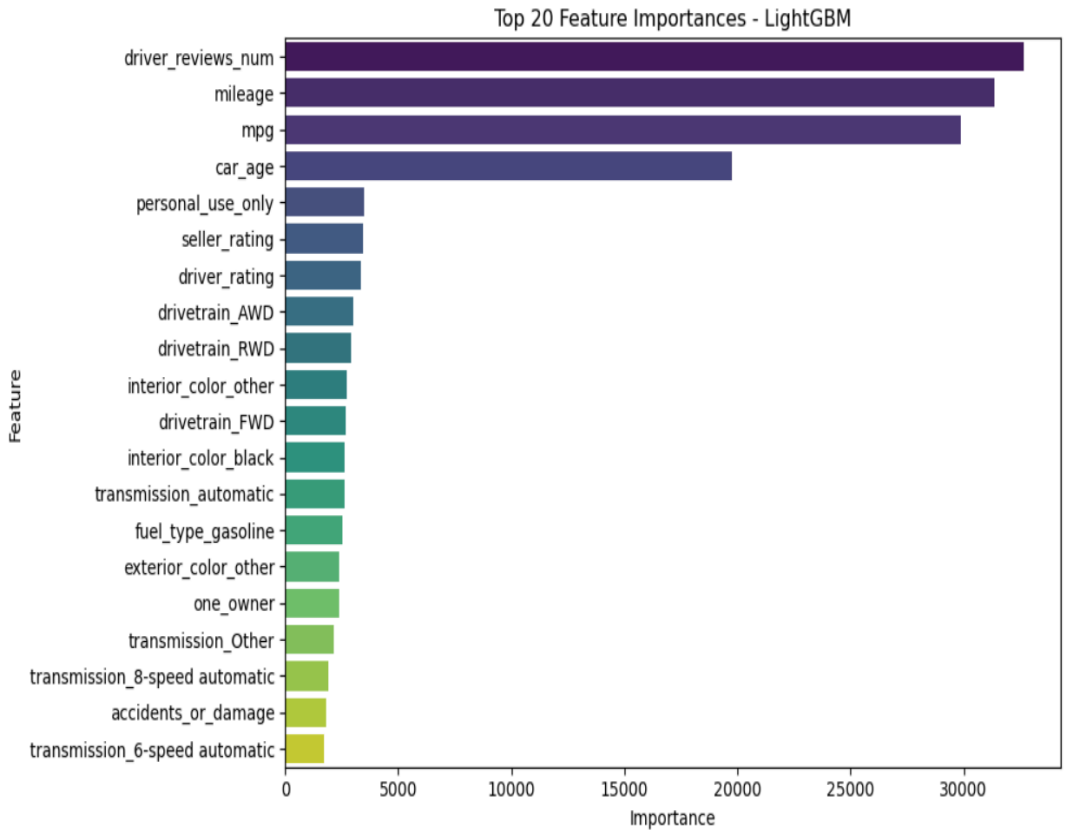


Figure 7: Top features influencing car price predictions in the LightGBM model

The feature importance chart highlights the top variables driving price predictions in the LightGBM model. Car age and mileage dominate as the strongest predictors, followed by transmission type, drivetrain, and exterior color. This breakdown provides valuable insights into which factors matter most in determining used car prices, making the model not only accurate but also interpretable.

| Error_Category | Total Cars | Mean Actual Price | Error ≤ 10% (%) |
|---|---|---|---|
| Manufacturer | | | |
| Lexus | 4174 | 33729.742214 | 76.329660 |
| Mazda | 2922 | 23050.818960 | 73.271732 |
| Honda | 6650 | 24608.190376 | 73.037594 |
| Volvo | 1847 | 36109.319437 | 72.441798 |
| Subaru | 4420 | 25089.121493 | 72.036199 |
| GMC | 5612 | 39606.717391 | 70.990734 |
| Buick | 2600 | 23267.728846 | 70.269231 |
| Cadillac | 2968 | 34666.745620 | 70.249326 |
| BMW | 6498 | 35830.715605 | 69.744537 |
| Audi | 3325 | 35053.821353 | 69.172932 |
| INFINITI | 2222 | 30304.543654 | 69.171917 |
| Hyundai | 4270 | 21582.560890 | 68.969555 |
| Toyota | 11207 | 30096.844561 | 68.707058 |
| Mercedes-Benz | 6593 | 39151.451236 | 68.451388 |
| Tesla | 1109 | 47512.549143 | 68.349865 |
| Lincoln | 2028 | 34974.718442 | 68.343195 |
| Kia | 6525 | 22401.283525 | 68.275862 |
| Mitsubishi | 1080 | 19424.600000 | 66.851852 |
| Jeep | 7890 | 31088.353232 | 66.387833 |
| Land Rover | 2178 | 46990.070707 | 66.115702 |
| Volkswagen | 4491 | 23316.080828 | 65.531062 |
| Ford | 14425 | 32572.158960 | 64.811092 |
| Jaguar | 1918 | 29287.994786 | 64.546403 |
| Porsche | 1445 | 54237.115571 | 64.013841 |
| Nissan | 8821 | 23019.266070 | 63.394173 |
| Chevrolet | 10409 | 30565.075512 | 63.022384 |
| Dodge | 4483 | 26820.315414 | 62.703547 |
| Chrysler | 2218 | 23231.490532 | 62.353472 |
| RAM | 3746 | 43493.901495 | 59.770422 |

The table above summarizes model performance by **car manufacturer**, showing the **total number of cars**, **average actual price**, and the **percentage of predictions within 10% of the actual value**.
The analysis reveals that **Lexus, Mazda, Honda, and Volvo** achieved the highest accuracy, with over **72%** of predictions falling within 10% of actual prices.
These brands tend to have more consistent pricing patterns, which helps the model perform better.

Meanwhile, **Chevrolet, Dodge, Chrysler, and RAM** displayed lower accuracy (around **60–63%**), likely due to greater variability in models and trims.
This analysis helps identify brands where the model excels and where brand-specific refinements could improve prediction accuracy.

# Future Research

While the current project demonstrates strong predictive performance, several areas remain open for further exploration. First, additional features such as service history, accident severity, or regional economic factors could be incorporated to improve accuracy and capture more of the real-world drivers of car prices. Second, advanced natural language processing (NLP) techniques could be applied to seller descriptions or customer reviews, which often contain valuable signals about vehicle condition and desirability.

Another avenue for research is exploring deep learning models, such as neural networks, which may capture more complex, non-linear relationships in the data compared to traditional machine learning algorithms. Finally, expanding the dataset to include international markets or integrating external data sources, such as fuel prices, could help build a more robust and generalizable model.