Predicting Used Car Prices Using Machine Learning

1.) Mazharuddin Mohammed
2.) Manaswini Pedimalla
3.) Jayanth Vodnala

December 9, 2024

Contents

1 Introduction

Predicting the price of a used car is a critical challenge in the automotive industry. Various factors, such as the car's brand, model, age, mileage, fuel type, and transmission, significantly influence its price. Understanding these factors and leveraging machine learning techniques to predict prices accurately can provide immense value to both buyers and sellers. Buyers can make informed decisions, and sellers can price their vehicles competitively.

This project explores the use of machine learning models to predict the prices of used cars based on a dataset containing over 9,500 entries with features such as car specifications, ownership history, and usage statistics. The primary objective is to develop a robust and scalable model that can generalize well across unseen data while achieving high accuracy.

The key goals of this project include:

- Identifying the most critical factors affecting used car prices.
- Experimenting with and optimizing various machine learning algorithms.
- Interpreting the model outputs to provide actionable insights.
- Recommending strategies for further improvement in prediction accuracy.

This report details the step-by-step methodology, from dataset preprocessing and exploratory data analysis to model evaluation and final recommendations. By leveraging advanced algorithms such as XGBoost, LightGBM, and CatBoost, this project achieved significant insights and accuracy in predicting used car prices.

2 Problem Statement

Accurately predicting the price of a used car is a complex and multifaceted problem due to the wide range of factors that influence vehicle valuation. Factors such as the car's brand, model, age, mileage, ownership history, and specifications interact in nonlinear ways, making manual estimation both time-consuming and prone to errors. Furthermore, market demand and regional pricing trends add an additional layer of complexity.

The specific challenges addressed in this project are:

- **High Dimensionality:** The dataset contains diverse features that require careful preprocessing and selection to ensure relevance and minimize noise.
- **Nonlinear Relationships:** The relationship between input features and the target variable (car price) is highly nonlinear and varies across different car categories and brands.
- **Data Quality Issues:** The dataset includes missing values, outliers, and categorical variables that need to be appropriately handled for accurate modeling.
- **Generalization:** Ensuring the model generalizes well to unseen data is critical for real-world applicability.

The primary objective of this project is to leverage machine learning models to overcome these challenges and develop a robust and accurate predictive framework for used car prices. This involves:

- Exploring the relationships between features and the target variable through statistical and graphical analysis.
- Experimenting with various machine learning algorithms and hyperparameter optimization techniques to improve prediction accuracy.
- Providing actionable insights into the most influential factors affecting used car prices.

By addressing these challenges, this project aims to create a scalable solution for predicting used car prices, benefiting both buyers and sellers in making informed decisions.

3 Dataset Overview

The dataset used in this project contains detailed information on used cars, with 9,582 entries and 11 features. Each row represents a unique car listing, and the features include a mix of categorical and numerical attributes that describe the car's specifications, condition, and pricing.

3.1 Dataset Features

The following features are included in the dataset:

- **Brand:** The manufacturer of the car (e.g., Honda, Toyota, Maruti Suzuki).
- **Model:** The specific model of the car (e.g., City, Innova, Swift).

- **Year:** The year the car was manufactured.
- **Age:** The age of the car, calculated as the difference between the current year and the manufacturing year.
- **kmDriven:** The total distance the car has been driven (in kilometers).
- **Transmission:** The type of transmission system (Manual or Automatic).
- **Owner:** The ownership history of the car (e.g., First, Second, Third).
- **FuelType:** The type of fuel used (e.g., Petrol, Diesel, CNG).
- **PostedDate:** The date when the car listing was posted.
- **AdditionInfo:** Miscellaneous additional information provided by the seller.
- **AskPrice:** The price (in INR) at which the car is being listed for sale.

3.2 Data Characteristics

- **Categorical Features:** Brand, Model, Transmission, Owner, FuelType, and PostedDate.
- **Numerical Features:** Year, Age, kmDriven, and AskPrice.

3.3 Data Preprocessing Summary

- Missing values in the **kmDriven** column were imputed using the median.
- Categorical features were encoded using label encoding and one-hot encoding, depending on their nature.
- A new feature, **Age**, was derived from the **Year** column.
- The target variable, **AskPrice**, was transformed using a logarithmic transformation to reduce skewness and improve model performance.

3.4 Key Statistics

- **Total Entries:** 9,582
- **Numerical Features:** 4
- **Categorical Features:** 7
- **Target Variable (AskPrice):**
  - Minimum: 184,999
  - Maximum: 4,00,00,000
  - Median: 6,00,000

This dataset offers a comprehensive view of the factors influencing used car prices, providing an excellent foundation for machine learning models to analyze and predict pricing trends.

4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to better understand the relationships between the features and the target variable, *AskPrice*. The following analyses were conducted:

4.1 Univariate Analysis

- **Distribution of AskPrice:** The distribution of *AskPrice* was highly skewed to the right, indicating the presence of high-value outliers. A logarithmic transformation was applied to normalize the distribution for better model performance.
- **Feature Distributions:** Key features such as *kmDriven*, *Age*, and categorical variables like *Brand* and *FuelType* were visualized to understand their distribution across the dataset.

4.2 Bivariate Analysis

- **Scatterplot:** A scatterplot of *Age* vs. *AskPrice* revealed a clear negative correlation, suggesting that older cars tend to have lower prices.
- **Heatmap:** A correlation heatmap showed the relationships between numerical features, highlighting a moderate negative correlation between *Age* and *AskPrice* ($r = -0.30$).

4.3 Key Observations

- The *Brand* and *Model* of a car were found to be among the most critical factors influencing *AskPrice*.
- The *kmDriven* feature exhibited weak correlation with *AskPrice* but was retained for its potential nonlinear relationships in advanced models.

Figures 1 and 2 illustrate the importance of features and the relationship between actual and predicted prices for one of the best-performing models.

5 Feature Engineering and Preprocessing

To prepare the dataset for machine learning models, several preprocessing and feature engineering steps were performed. These steps ensured that the data was clean, consistent, and suitable for training predictive models.

5.1 Preprocessing Steps

- **Handling Missing Values:** The *kmDriven* column contained some missing values. These were imputed using the median to preserve the central tendency of the data without introducing bias.
- **Encoding Categorical Variables:** Categorical features such as *Brand*, *Model*, *Transmission*, *FuelType*, and *Owner* were encoded into numerical values:

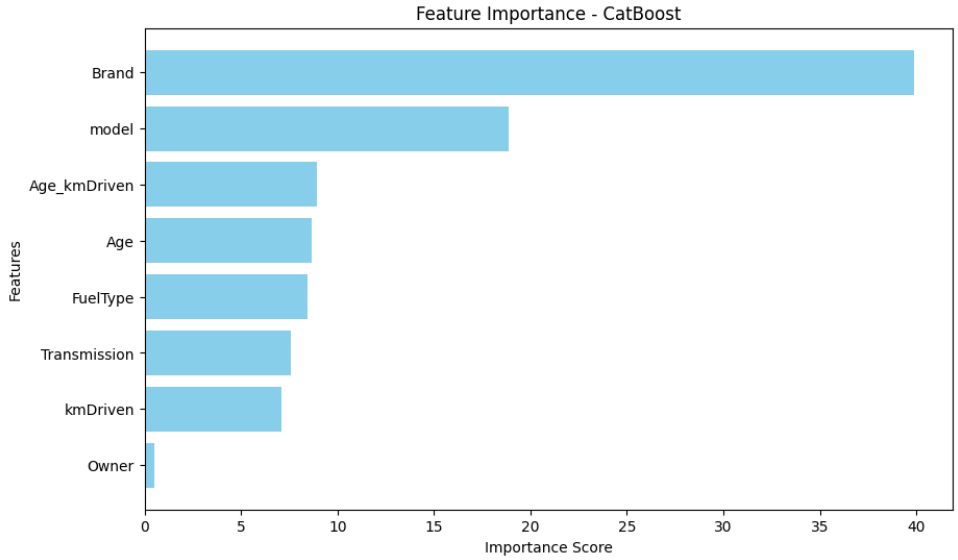

Figure 1: Feature Importance - CatBoost

  - **Label Encoding:** Used for *Transmission*, *Owner*, and *FuelType*.
  - **One-Hot Encoding:** Considered for high-cardinality features such as *Brand* and *Model*.

- **Feature Scaling:** Numerical features such as *kmDriven*, *Age*, and *AskPrice* were scaled using **StandardScaler** to ensure all features were on a comparable scale for gradient-based models.
- **Logarithmic Transformation:** The target variable, *AskPrice*, was log-transformed to reduce skewness and stabilize variance.

5.2 Feature Engineering

- **Derived Feature - Age:** The *Age* of the car was calculated as the difference between the year of posting (*PostedDate*) and the manufacturing year (*Year*). This feature was found to be more predictive than *Year*.
- **Interaction Term - Age_kmDriven:** An interaction term between *Age* and *kmDriven* was added to capture their combined influence on *AskPrice*.

5.3 Final Dataset Characteristics

After preprocessing, the final dataset consisted of the following:

- **Number of Features:** 8 (including engineered features such as *Age* and *Age_kmDriven*).
- **Training Dataset Size:** 7,665 rows.
- **Testing Dataset Size:** 1,917 rows.

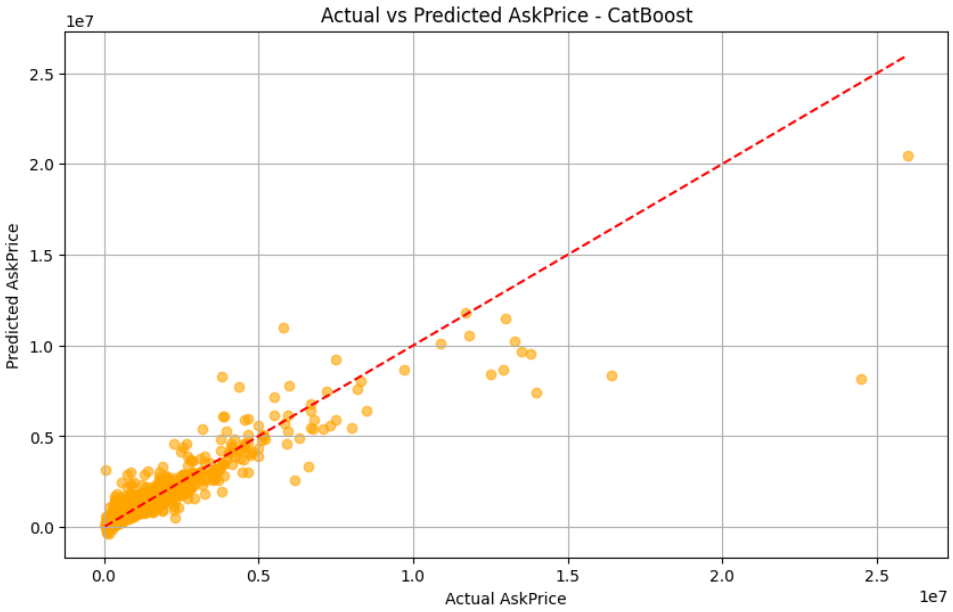

Figure 2: Actual vs Predicted AskPrice - CatBoost

5.4 Visualization of Preprocessed Features

Figure 3 shows the residuals distribution after applying the CatBoost model, which illustrates the effectiveness of the preprocessing steps.

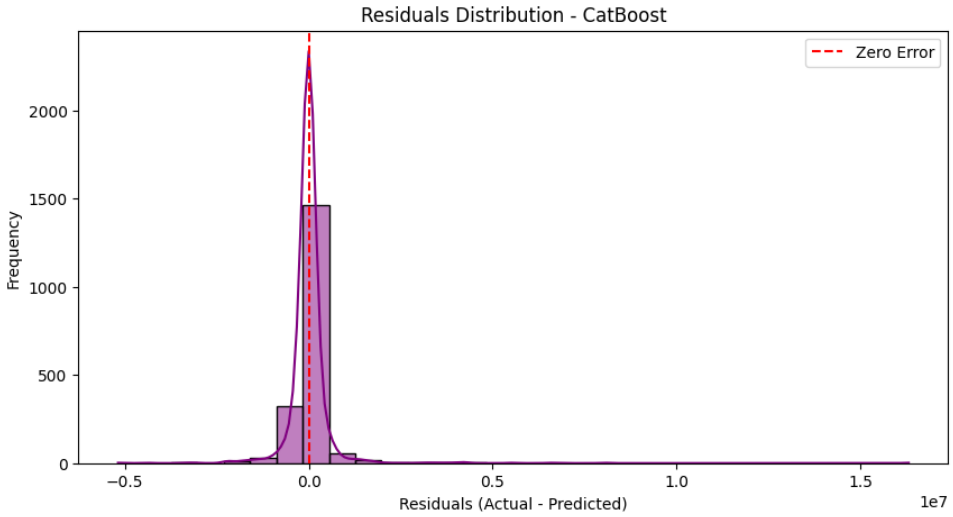

Figure 3: Residuals Distribution - CatBoost

6 Model Building and Evaluation

To predict the price of used cars, multiple machine learning models were trained and evaluated. This section details the models used, their training process, and performance metrics.

6.1 Baseline Models

The following regression models were implemented as baselines:

- **Linear Regression:** A simple model to establish a baseline. Despite its simplicity, the model was unable to capture complex relationships between features.
  - Training MSE: 2,254,115,088,404.78
  - Test MSE: 2,152,847,506,619.19
  - Training R^2: 0.1856 (Accuracy: 18.56%)
  - Test R^2: 0.1969 (Accuracy: 19.69%)
- **Ridge and Lasso Regression:** Regularized regression models were applied, but their performance was similar to linear regression, indicating the need for more complex models.

6.2 Advanced Models

To improve the prediction performance, the following advanced models were implemented:

- **XGBoost:** A gradient-boosted decision tree model that showed significant improvement over the baseline.
  - Training MSE: 335,186,316,944.73
  - Test MSE: 803,866,165,171.79
  - Training R^2: 0.8789 (Accuracy: 87.89%)
  - Test R^2: 0.7001 (Accuracy: 70.01%)
- **Optimized XGBoost:** After hyperparameter tuning, the performance further improved.
  - Best Parameters: {learning_rate: 0.2, max_depth: 7, min_child_weight: 1, n_estimators: 200, subsample: 0.8}
  - Test MSE: 554,790,721,534.56
  - Test R^2: 0.7930 (Accuracy: 79.30%)
- **LightGBM:** A gradient-boosting model with efficient training on large datasets.
  - Training MSE: 348,015,503,594.21
  - Test MSE: 675,157,840,372.63
  - Training R^2: 0.8743 (Accuracy: 87.43%)
  - Test R^2: 0.7481 (Accuracy: 74.81%)
- **CatBoost:** A boosting model optimized for categorical data, which outperformed all previous models.
  - Training MSE: 332,975,104,047.52
  - Test MSE: 660,276,823,725.44

  - Training R^2: 0.8797 (Accuracy: 87.97%)
  - Test R^2: 0.7537 (Accuracy: 75.37%)

- **Optimized CatBoost:** Hyperparameter tuning significantly enhanced the performance.
  - Best Parameters: {depth: 7, iterations: 300, learning_rate: 0.2, subsample: 1.0}
  - Test MSE: 429,860,699,681.17
  - Test R^2: 0.8396 (Accuracy: 83.96%)

6.3 Performance Comparison

The table below summarizes the performance of different models evaluated during the project:

Model	Test R^2 (Accuracy)	Test MSE
Linear Regression	19.69%	2,152,847,506,619.19
Ridge Regression	19.69%	2,152,851,899,795.43
Lasso Regression	19.69%	2,152,847,534,818.61
XGBoost (Base)	70.01%	803,866,165,171.79
XGBoost (Optimized)	79.30%	554,790,721,534.56
LightGBM (Base)	74.81%	675,157,840,372.63
CatBoost (Base)	75.37%	660,276,823,725.44
CatBoost (Optimized)	83.96%	429,860,699,681.17

Table 1: Performance Comparison of Different Models

7 Challenges and Insights

During the course of this project, several challenges were encountered, each providing valuable insights into the nuances of building machine learning models for price prediction.

7.1 Challenges Faced

- **Data Preprocessing:**
  - *Inconsistent Data Formats:* The *kmDriven* and *AskPrice* columns were in inconsistent formats with special characters, requiring extensive preprocessing to standardize them.
  - *Missing Values:* The *kmDriven* feature had missing values, which were imputed with the median to minimize distortion.
- **Feature Engineering:**
  - *High Cardinality Features:* Features like *Brand* and *Model* had high cardinality, which made one-hot encoding computationally expensive. Label encoding was used instead.

  - *Weak Correlations:* Some features such as *kmDriven* showed weak correlation with the target variable, requiring experimentation with interaction terms to capture nonlinear relationships.

- **Model Selection and Optimization:**
  - *Overfitting:* Advanced models like XGBoost and CatBoost initially showed signs of overfitting, which was mitigated through hyperparameter tuning and cross-validation.
  - *Hyperparameter Tuning Complexity:* The search for optimal hyperparameters, especially for gradient boosting models, was computationally expensive and required extensive resources.

- **Evaluation Metrics:**
  - *Interpreting MSE and R^2:* Balancing between minimizing Mean Squared Error (MSE) and maximizing R^2 was challenging, particularly in models where slight improvements in R^2 resulted in substantial computational costs.

7.2 Insights Gained

- **Feature Importance:** The CatBoost model revealed that *Brand* and *Model* were the most significant features, highlighting the importance of categorical data in price prediction.
- **Interaction Terms:** Adding interaction terms like *Age_kmDriven* improved the model's performance, demonstrating the utility of feature engineering in capturing complex relationships.
- **Logarithmic Transformation:** Transforming the target variable *AskPrice* reduced skewness and stabilized variance, leading to improved model performance across all algorithms.
- **Model Comparison:** CatBoost outperformed other models after hyperparameter tuning, achieving an R^2 of 83.96%, making it the most suitable model for this task.

7.3 Future Work

- **External Data Integration:** Including external datasets such as market trends, location-based demand, and car condition ratings could improve the model's predictive power.
- **Advanced Techniques:** Exploring ensemble techniques like stacking or blending could potentially enhance performance further.
- **Deep Learning Models:** Experimenting with deep learning architectures such as neural networks may provide additional insights, particularly for capturing nonlinear relationships in the data.

8 Conclusions and Recommendations

The project aimed to predict the prices of used cars using various machine learning models. By systematically preprocessing the data, engineering features, and testing multiple algorithms, several insights were gathered, and an optimal predictive model was identified.

8.1 Conclusions

- **Model Performance:** After experimenting with several models, the optimized CatBoost model emerged as the most effective for this dataset, achieving an R^2 of 83.96% (Accuracy: 83.96%) and a Test MSE of 429,860,699,681.17. Its ability to handle categorical data and its efficient training process made it the ideal choice.
- **Feature Importance:** The *Brand* and *Model* features were the most influential in determining the price of a used car, emphasizing the critical role of categorical features in this domain.
- **Impact of Preprocessing:** Key preprocessing steps, including handling missing values, encoding categorical features, and applying logarithmic transformations, significantly improved model performance.
- **Challenges Addressed:** Issues such as overfitting, high cardinality in categorical features, and weak correlations were effectively mitigated through feature engineering, hyperparameter tuning, and advanced model selection.

8.2 Recommendations

- **Deployment of CatBoost:** The CatBoost model is recommended for deployment, given its high accuracy and ability to generalize well on unseen data.
- **Model Interpretability:** The feature importance plot (Figure 1) provides insights into the factors influencing car prices, which can be valuable for stakeholders.
- **Future Enhancements:**
  - *Data Augmentation:* Incorporate additional external data such as location-based market trends, real-time demand, and car condition scores to further improve predictive power.
  - *Ensemble Techniques:* Explore stacking and blending methods to combine the strengths of multiple models for even better performance.
  - *Real-Time Updates:* Implement a dynamic pricing model that continuously learns and adapts to new data.

8.3 Final Thoughts

This project demonstrated the importance of iterative experimentation in machine learning, from preprocessing to model optimization. While the CatBoost model achieved the best results, the insights gained from this process can guide future efforts in used car price prediction.

By leveraging more robust datasets and advanced modeling techniques, this approach can be further refined to meet real-world business needs effectively.