Machine Learning Assessment

Bibek Sapkota

Student ID: 23189618

Word Count: 2697

Total Page Count: 15

February 2025

Contents

List of Figures

Abstract

This project presents a machine learning approach to predicting car prices based on key vehicle attributes such as mileage, brand, horsepower, and fuel type. The dataset underwent extensive preprocessing, including feature engineering and missing value imputation. Various regression models were evaluated, with XGBoost emerging as the best-performing model based on RMSE. Feature importance analysis using SHAP highlighted mileage and brand as the most influential factors in price determination.

1 Introduction

The automotive market is highly dynamic, with vehicle prices influenced by various factors such as brand, model, age, mileage, fuel type, and accident history. Accurately estimating car prices is essential for buyers, sellers, and dealerships to make informed decisions. The workflow includes data loading, exploratory data analysis (EDA), feature engineering, model training with multiple algorithms, evaluation using regression metrics, interpretability via analysis, and deployment through an API. This report summarizes the key methodologies, findings, and implementation steps of the project.

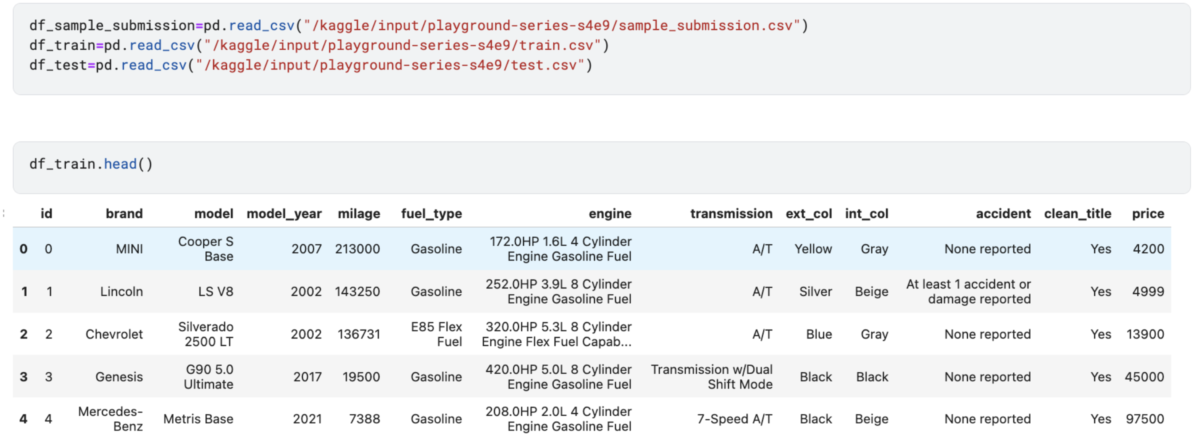
2 Problem Definition

Objective: The used car market is highly dynamic, with prices influenced by multiple factors such as brand, model, year, mileage, fuel type, and accident history. Traditional pricing methods often rely on subjective assessments, leading to inconsistencies and inefficiencies. This project aims to build a machine learning model for accurate price prediction, improving transparency and aiding better decisions for buyers, sellers, and dealerships.

3 Dataset Selection and Analysis

3.1 Data Loading

he dataset was loaded from a CSV file and its integrity was verified by inspecting the first few rows. A snippet of the code used is provided below:



```
df_sample_submission=pd.read_csv("/kaggle/input/playground-series-s4e9/sample_submission.csv")
df_train=pd.read_csv("/kaggle/input/playground-series-s4e9/train.csv")
df_test=pd.read_csv("/kaggle/input/playground-series-s4e9/test.csv")
```

```
df_train.head()
```

	id	brand	model	model_year	milage	fuel_type	engine	transmission	ext_col	int_col	accident	clean_title	price
0	0	MINI	Cooper S Base	2007	213000	Gasoline	172.0HP 1.6L 4 Cylinder Engine Gasoline Fuel	A/T	Yellow	Gray	None reported	Yes	4200
1	1	Lincoln	LS V8	2002	143250	Gasoline	252.0HP 3.9L 8 Cylinder Engine Gasoline Fuel	A/T	Silver	Beige	At least 1 accident or damage reported	Yes	4999
2	2	Chevrolet	Silverado 2500 LT	2002	136731	E85 Flex Fuel	320.0HP 5.3L 8 Cylinder Engine Flex Fuel Capab...	A/T	Blue	Gray	None reported	Yes	13900
3	3	Genesis	G90 5.0 Ultimate	2017	19500	Gasoline	420.0HP 5.0L 8 Cylinder Engine Gasoline Fuel	Transmission w/Dual Shift Mode	Black	Black	None reported	Yes	45000
4	4	Mercedes-Benz	Metris Base	2021	7388	Gasoline	208.0HP 2.0L 4 Cylinder Engine Gasoline Fuel	7-Speed A/T	Black	Beige	None reported	Yes	97500

Figure 1: Loading and Displaying data

3.2 Dataset Overview

The dataset comprises **188,533** instances and **13** attributes, including categorical and numerical features. The target variable is `price`, making this a regression problem. The dataset includes the following columns:

- `id`: Unique identifier for each record.
- `brand`: Car manufacturer or brand name.
- `model`: Specific model of the car.

- `model year`: Manufacturing year of the car.
- `milage`: Total distance traveled by the vehicle (in miles).
- `fuel type`: Type of fuel used (e.g., Gasoline, Diesel, Electric).
- `engine`: Engine details including power and fuel consumption.
- `transmission`: Type of transmission system (e.g., Manual, Automatic).
- `ext col`: Exterior color of the car.
- `int col`: Interior color of the car.
- `accident`: History of reported accidents.
- `clean title`: Indicates whether the car has a clean title with no major damage.
- `price`: The target variable representing the selling price of the car.

3.3 Missing Values

Several features contain missing values, which need to be handled appropriately to ensure model reliability. The percentage of missing values for key columns in the datasets is as follows:

- `fuel type`: 2.69%
- `accident`: 1.30%
- `clean title`: 11.36%

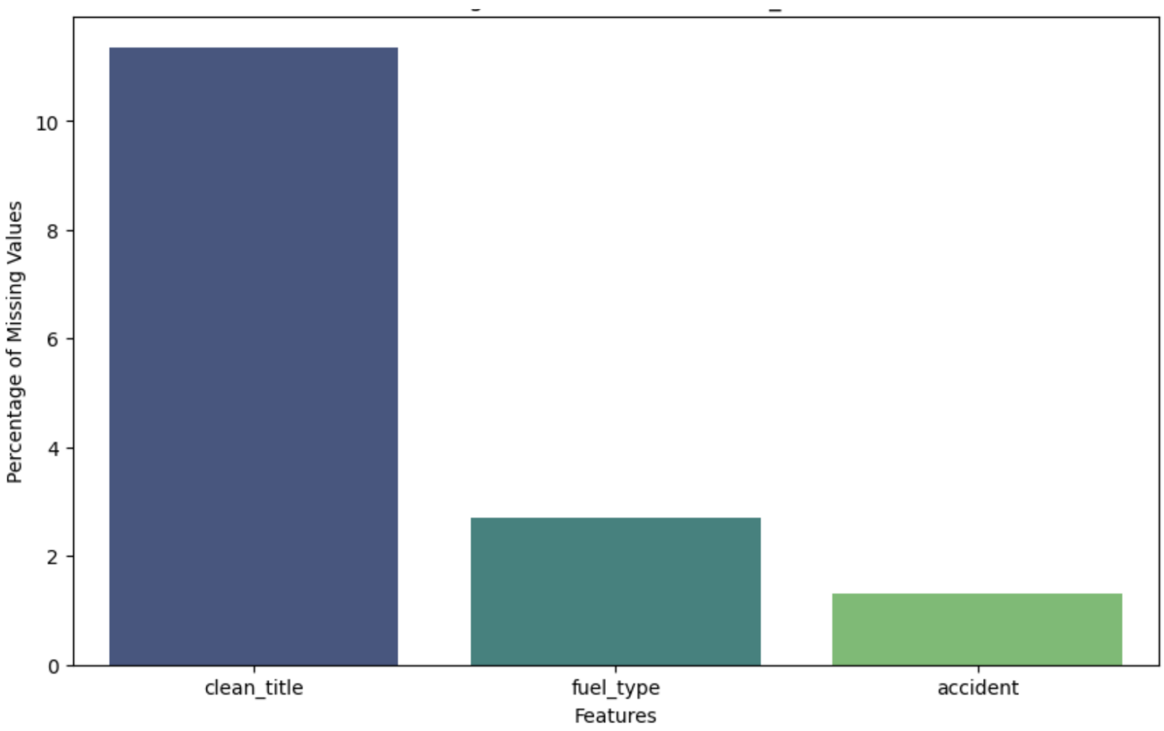


Figure 2: Plot of missing values

The missing data distribution was visualized using a heatmap to highlight patterns and correlations among missing values. Additionally, a correlation matrix was employed to assess dependencies between missing values across features. These visualizations provided key insights that guided the choice of imputation methods, ensuring minimal bias while preserving data integrity.

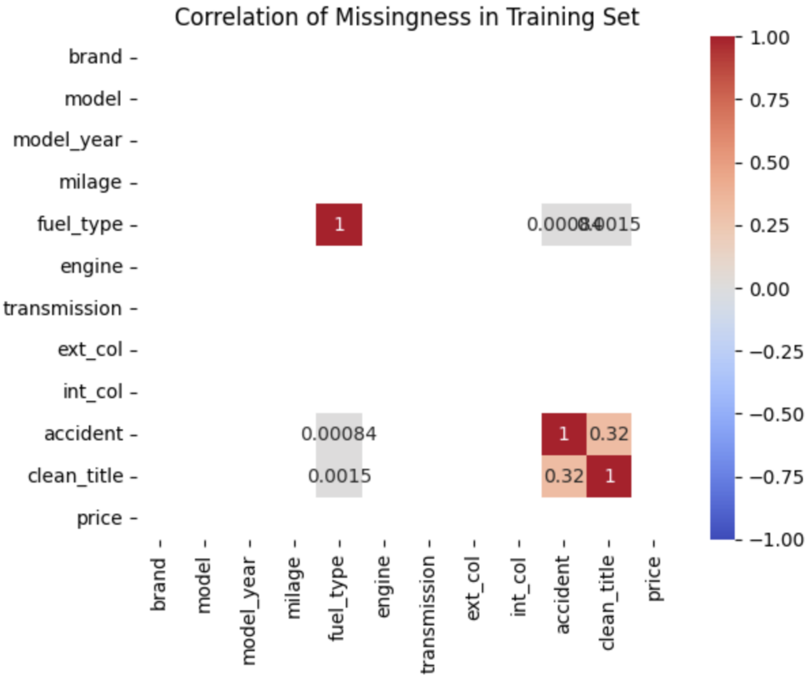


Figure 3: correlation of missing data

To handle these missing values, categorical features were imputed using the most logical default values. Specifically:

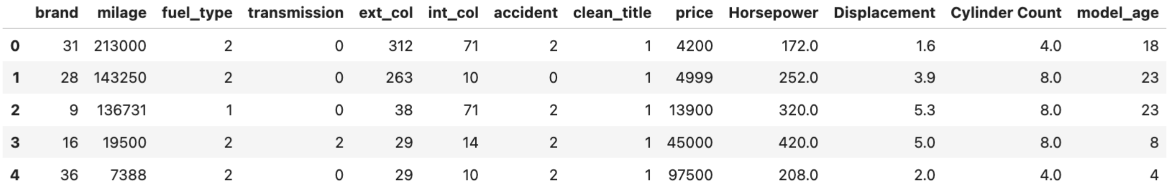
- Missing values in `fuel type` were replaced with `'Unknown'`.
- Missing values in `accident` were replaced with `'No'` to indicate no reported accidents.
- Missing values in `clean title` were also replaced with `'No'`, signifying that the vehicle does not have a clean title.

3.4 Data Preprocessing and Feature Engineering

- Extracted numerical features from the `engine` column, including horsepower, displacement, engine type, cylinder count, and fuel type.
- Created a new feature `model age` by computing the difference between the current year and the manufacturing year.
- Removed redundant features such as `engine`, `model`, `model year`, `Engine Type`, and `Fuel Type` after extracting relevant information.
- Handled missing values:
  - Used default values for categorical features (e.g., `'Unknown'` for fuel type, `'No'` for accident and clean title).
  - Applied **Iterative Imputation** for numerical features like `Horsepower` and `Displacement`.

  - Used **Simple Imputer** (mean strategy) for `Cylinder Count`.

- Standardized the `transmission` feature by mapping various transmission types to predefined categories (e.g., Automatic, Manual, Variator, Triptronic, Other).
- Encoded categorical variables using **Label Encoding** for features such as `brand`, `fuel type`, `ext col`, `int col`, `accident`, `clean title`, and `transmission`.
- Saved label encoders for future use, ensuring consistency in encoding between training and test data.
- Applied feature scaling and normalization where necessary.



	brand	milage	fuel_type	transmission	ext_col	int_col	accident	clean_title	price	Horsepower	Displacement	Cylinder Count	model_age
0	31	213000	2	0	312	71	2	1	4200	172.0	1.6	4.0	18
1	28	143250	2	0	263	10	0	1	4999	252.0	3.9	8.0	23
2	9	136731	1	0	38	71	2	1	13900	320.0	5.3	8.0	23
3	16	19500	2	2	29	14	2	1	45000	420.0	5.0	8.0	8
4	36	7388	2	0	29	10	2	1	97500	208.0	2.0	4.0	4

Figure 4: plotting dataset after encoding

3.5 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the structure and characteristics of the dataset. Various visualizations and statistical techniques were employed to uncover patterns, distributions, and relationships among features.

3.5.1 Word Cloud of Car Models

To visualize the most common car models in the dataset, a word cloud was generated. The figure below illustrates the most frequent model names:

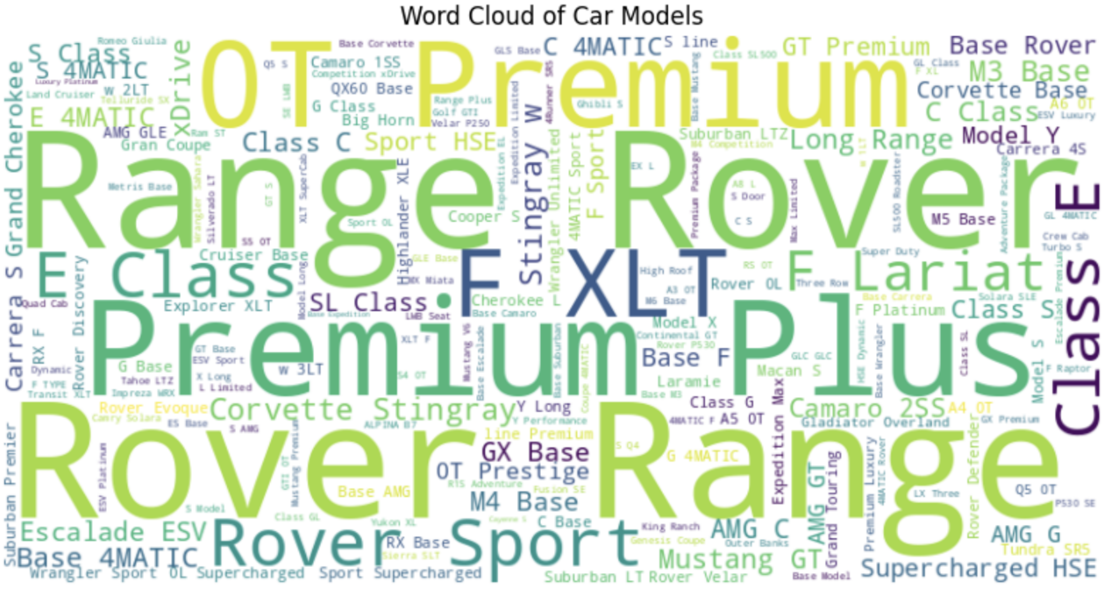


Figure 5: World cloud of car models

Insights: The word cloud reveals that some car models, such as Range Rover" and Premium," appear more frequently in the dataset. This suggests that luxury and high-end models are well-represented. The presence of premium models implies that the dataset may contain a substantial number of high-value vehicles, which could influence the pricing patterns.

3.5.2 Average Price by Car Brand

A bar chart was plotted to analyze the average price of vehicles based on brand. The results indicate that luxury brands such as Rolls-Royce, Ferrari, and Lamborghini have significantly higher prices:

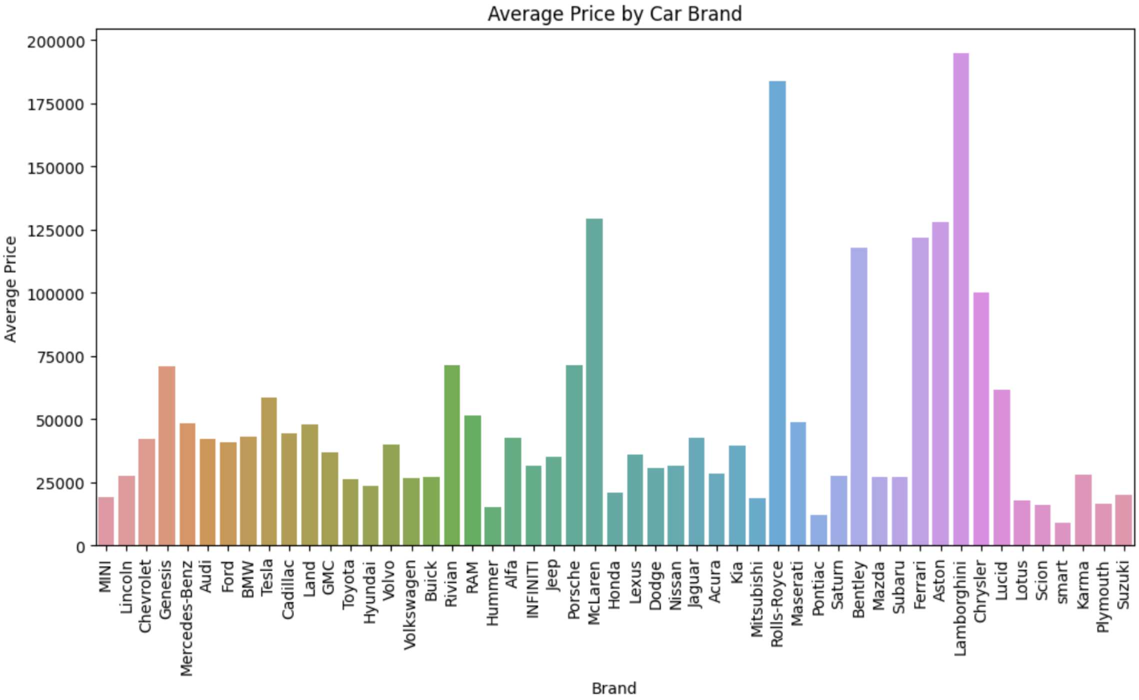


Figure 6: Avergae price by Car brand

Insights: The chart shows a significant price variation among brands, with high-end brands commanding much higher prices. This observation is expected, as luxury brands produce premium vehicles with advanced features, high-performance engines, and superior craftsmanship. Additionally, some mid-range brands have considerable price variation, possibly due to model variations and trim levels.

3.5.3 Pairplot of Numerical Features

A pairplot was generated to visualize relationships between numerical variables. Strong correlations were observed between model age and mileage:

Insights: The pairplot provides a detailed view of interactions among numerical features. Several key observations include:

- Model Year and Mileage: There is a clear trend where older cars tend to have higher mileage. This relationship aligns with expectations, as older vehicles have been driven longer.
- Mileage and Price: A strong negative correlation is observed between mileage and price, reinforcing the idea that higher mileage leads to lower car valuation.

- Price Distribution: The price variable exhibits a right-skewed distribution, with a few exceptionally high-priced vehicles. This suggests that luxury cars or rare models significantly influence the dataset.

- Clusters of Data Points: Some distinct groupings can be seen, indicating that certain vehicle types, possibly specific brands or models, have unique distributions for mileage and price.

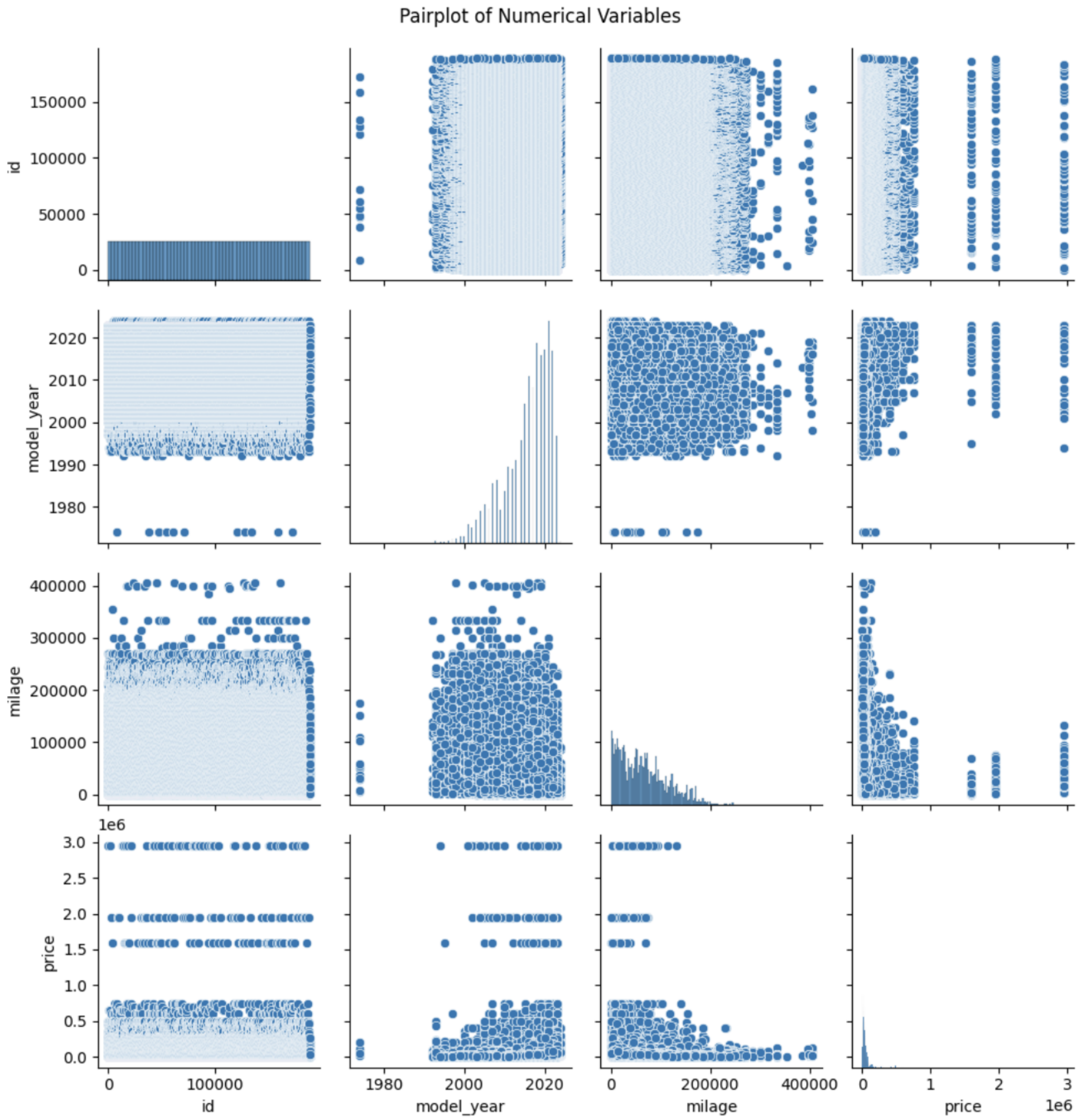


Figure 7: Pairplot of Numeric Values

This analysis highlights the importance of considering feature interactions when building predictive models. The strong associations between model year, mileage, and price suggest that these variables are crucial for car price estimation.

3.5.4 Correlation Heatmap

A heatmap was generated to visualize correlations among numerical features. The strongest correlation was observed between horsepower and displacement, as well as a negative correlation between mileage and price:

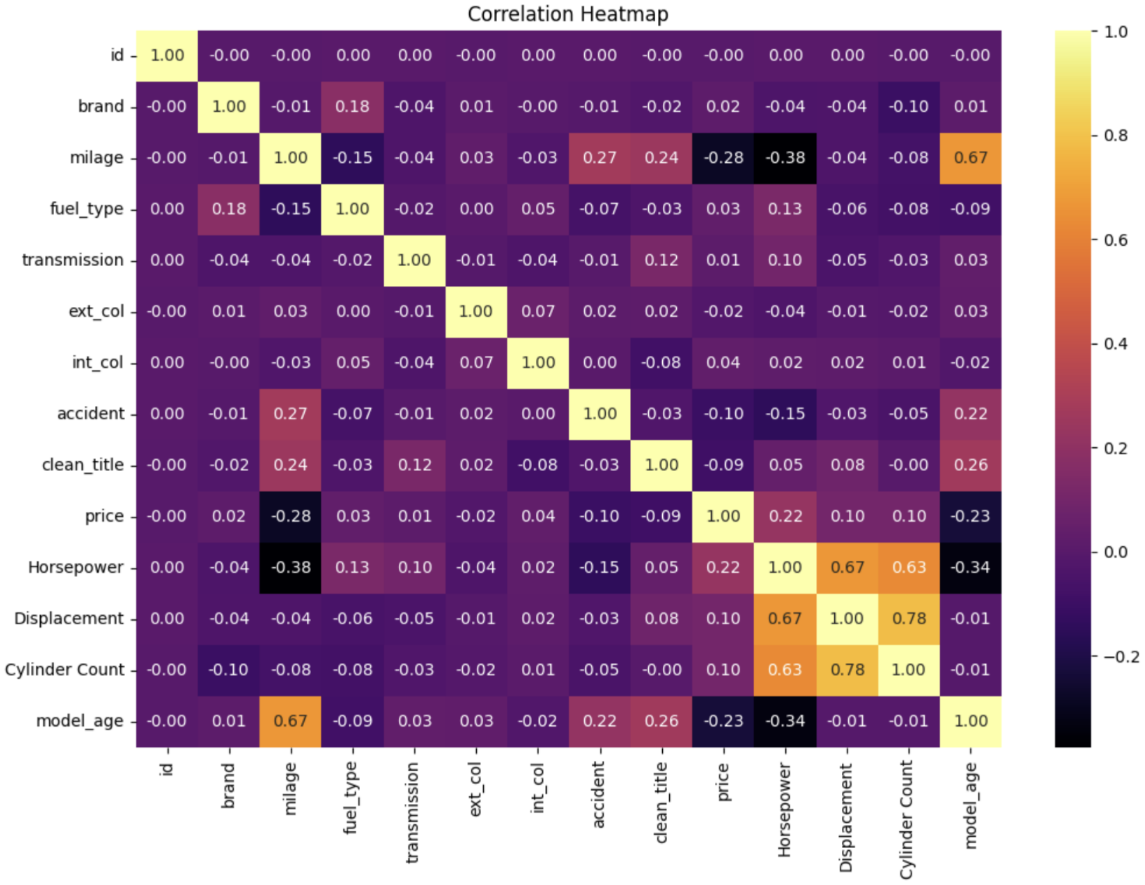


Figure 8: Correlation Heatmap

Insights: The negative correlation between mileage and price reinforces findings from the scatter plot. Also, horsepower and displacement are highly correlated, meaning one may be redundant in the model. The correlation between model age and mileage also confirms that older vehicles tend to have higher mileage, which directly impacts their pricing

4 Model Development and Evaluation

To evaluate different models for car price prediction, a 5-fold cross-validation strategy was implemented. This approach ensured that the model's performance was validated on multiple subsets of the data, reducing the risk of overfitting and improving generalizability. The models were trained and evaluated on different folds of the dataset, and the overall performance was assessed using the Out-of-Fold (OOF) RMSE scores.

A preprocessing pipeline was constructed to streamline the data preparation process. The pipeline included:

- Standardizing numerical features using **StandardScaler** to ensure uniformity across different scales.
- Transforming categorical features using one-hot encoding and label encoding where necessary.
- Implementing a **ColumnTransformer** to apply transformations separately to numerical and categorical features.
- Ensuring that missing values were handled appropriately before model training.

The final pipeline was integrated into an **XGBoost** regression model, which demonstrated the best predictive accuracy among tested models. GPU acceleration was enabled using the `tree_method=gpu_hist` setting to enhance training efficiency. Hyperparameters were fine-tuned using **Optuna**, optimizing key parameters such as the learning rate, depth, and regularization constraints.

The results of different models tested using cross-validation are summarized below:

- **Linear Regression** - Overall OOF RMSE: 75,049.7913
- **Decision Tree** - Overall OOF RMSE: 74,017.4018
- **Random Forest** - Overall OOF RMSE: 73,140.0250
- **AdaBoost** - Overall OOF RMSE: 73,952.9107
- **CatBoost** - Overall OOF RMSE: 73,119.5593
- **XGBoost** - Overall OOF RMSE: 72,832.0142 (Best Model)
- **LightGBM** - Overall OOF RMSE: 72,904.8041
- **Ridge Regression** - Overall OOF RMSE: 74,748.1006
- **Lasso Regression** - Overall OOF RMSE: 75,063.8828
- **MLP Regression** - Overall OOF RMSE: 73,970.3535

Among these models, **XGBoost** demonstrated the lowest RMSE, making it the most suitable model for car price prediction. The trained model was saved for deployment, ensuring that users could input vehicle specifications and receive price estimates efficiently.

5 Model Interpretation

5.1 Feature Importance Analysis

Feature importance analysis was conducted using a Random Forest model to identify the most influential features affecting car price prediction. The results indicate that mileage, model age, and exterior color play crucial roles in determining vehicle prices:

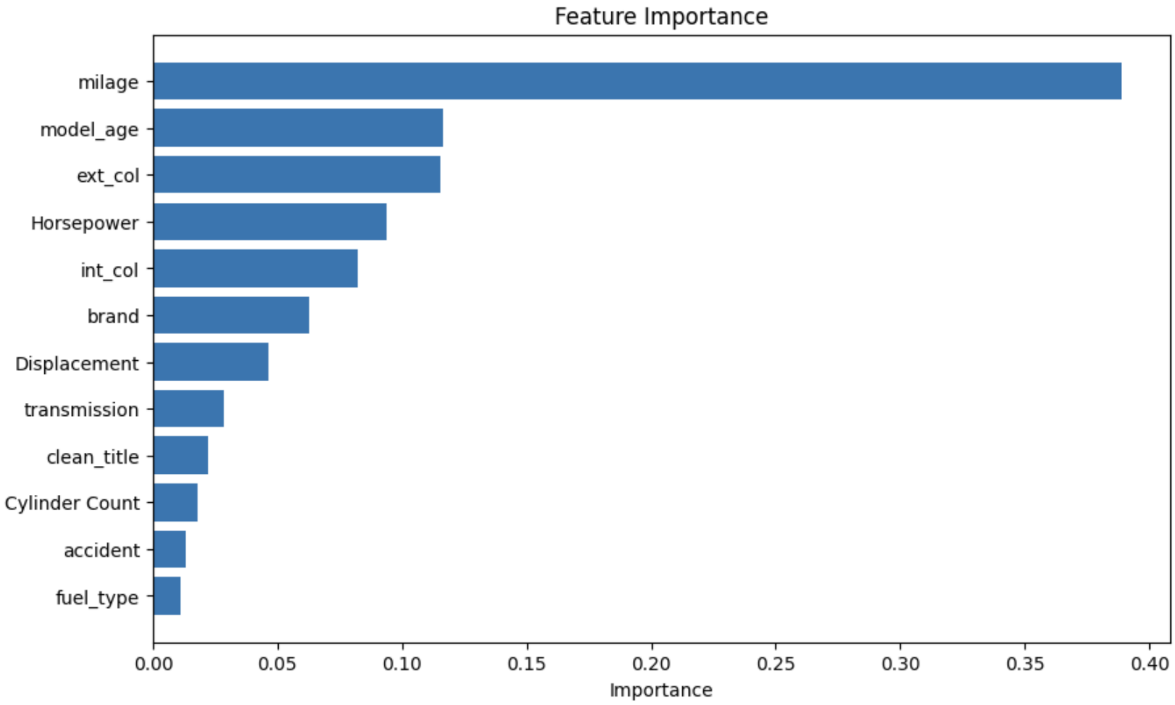


Figure 9: Feacture Importance using random forest

Insights: Mileage is the most critical factor in determining a car's price, followed by model age. Exterior and interior color also contribute, which may be linked to brand perception and desirability. Surprisingly, some non-obvious features like transmission type and accident history have lower importance than expected, suggesting that these factors might be less significant when compared to mileage and model year.

5.2 Model Interpretation with SHAP Analysis

To interpret the impact of each feature on the model's predictions, SHAP (SHapley Additive exPlanations) values were calculated for the XGBoost model. The SHAP summary plot below provides an overview of how different features contribute to price predictions:

Insights:

- Mileage: As observed earlier, mileage is the most influential factor in price prediction. Higher mileage negatively impacts price, which aligns with market trends where vehicles with higher mileage typically have lower resale value.

- Brand: Brand also plays a significant role in price determination. Luxury brands contribute positively to higher prices, whereas budget brands influence lower prices.

- Horsepower and Engine Displacement: Vehicles with higher horsepower and larger engine displacement tend to have higher prices, indicating that performance-based attributes are crucial in pricing.

- Exterior and Interior Color: The impact of color suggests that aesthetic preferences and market demand influence car valuation. Some colors may be more desirable than others, affecting the resale price.
- Transmission and Accident History: Transmission type shows a lower impact compared to other factors, indicating that buyers may prioritize mileage and brand over transmission type. Similarly, accident history plays a role but is not the primary determinant of price.

```
Computing SHAP values...
```

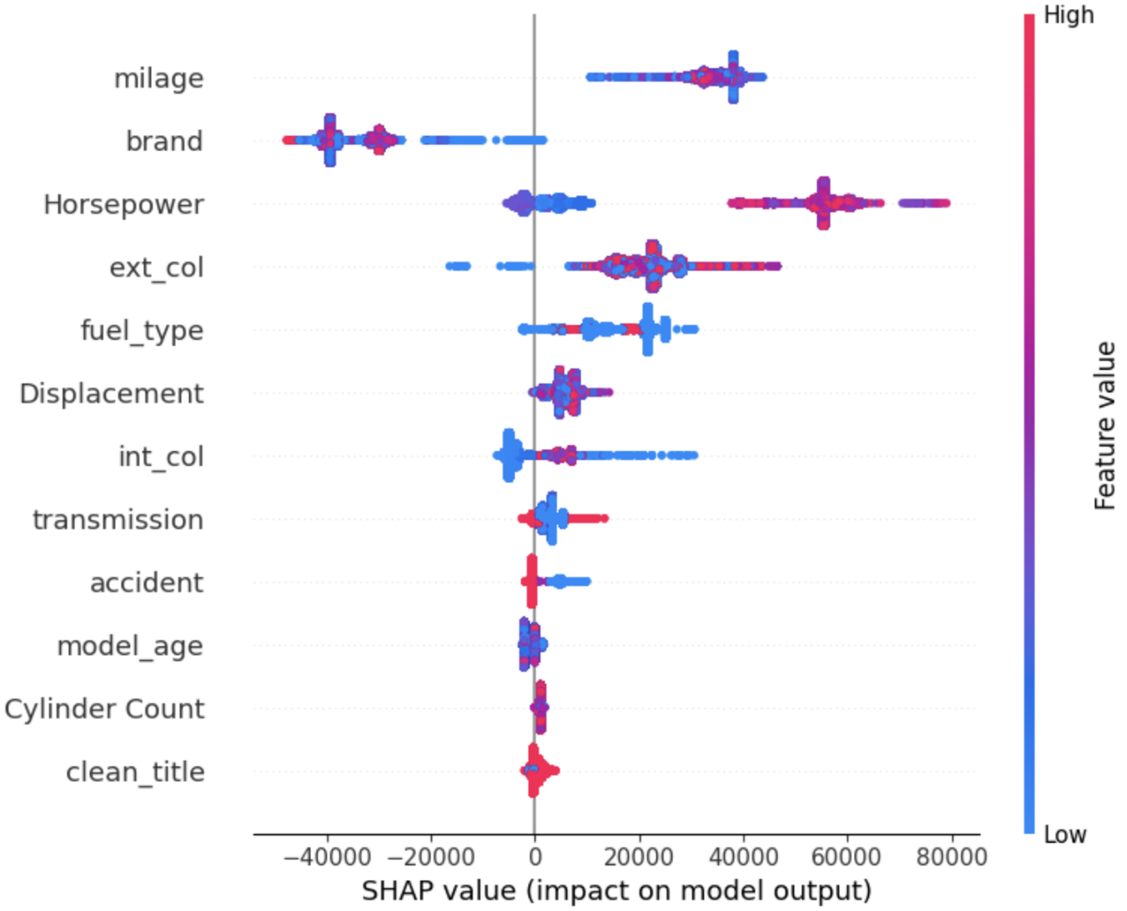


Figure 10: Shape impact plot

The average impact of each feature on the model output magnitude is visualized in the SHAP bar plot below:

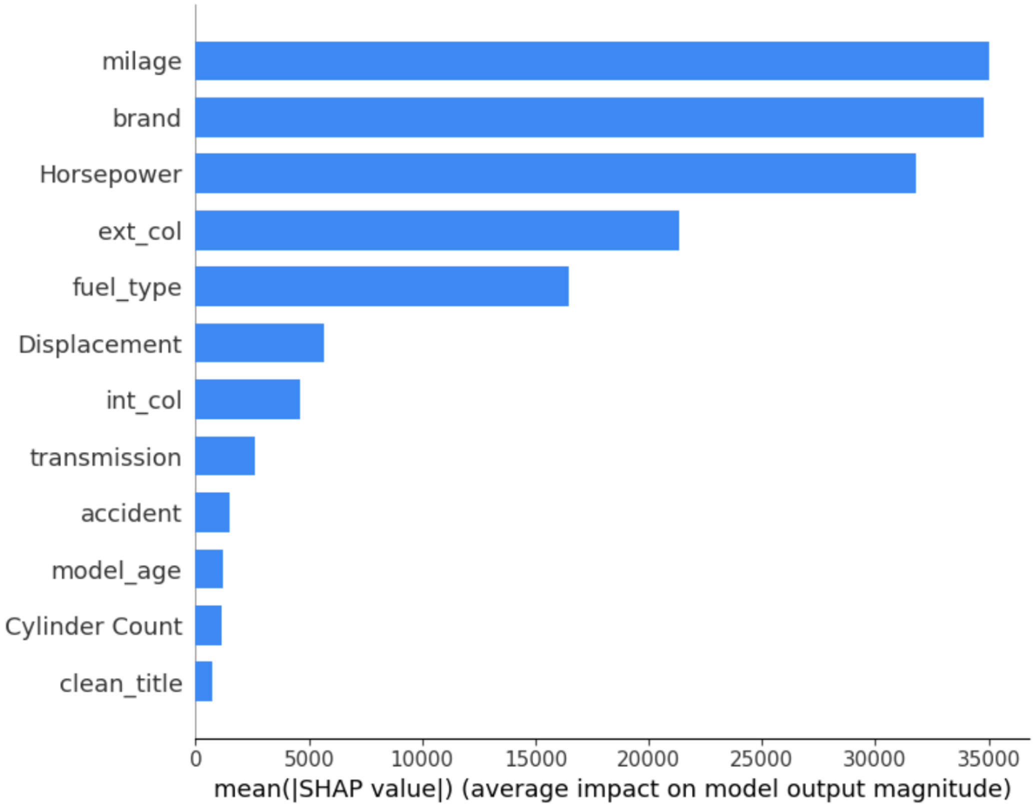


Figure 11: shape bar plot

Further Insights: This visualization confirms that mileage, brand, and horsepower are the top contributors to price prediction. While categorical variables such as transmission and accident history do influence price, their impact is relatively lower compared to numerical factors such as engine displacement and mileage.

Understanding feature importance through SHAP analysis allows for a more interpretable model. It ensures that predictions are aligned with real-world pricing trends and provides confidence in the model's decision-making proces

6 Conclusion

This project successfully implemented a machine learning-based car price prediction system using real-world vehicle data. The dataset was extensively analyzed, and various preprocessing techniques, including feature engineering and missing value imputation, were applied to enhance model performance.

Multiple regression models were trained and evaluated, with XGBoost emerging as the most effective model due to its lower RMSE and superior predictive accuracy. Feature importance analysis using SHAP revealed that mileage, brand, and horsepower were the most influential factors affecting car prices.

Key findings from the study include:

- Higher mileage negatively impacts vehicle prices, confirming its role as a critical depreciation factor.
- Luxury brands retain higher resale values, significantly influencing the prediction model.
- Horsepower and engine displacement contribute positively to price, highlighting the demand for high-performance vehicles.
- Exterior and interior color preferences show a minor but noticeable effect on vehicle valuation.
- Transmission type and accident history have relatively lower influence on pricing decisions.

The results demonstrate that machine learning can effectively predict vehicle prices with high accuracy. The model's interpretability through SHAP values ensures that its predictions align with industry insights, making it a valuable tool for car dealerships and buyers.

Future improvements:

- Incorporating additional features such as market demand trends and economic indicators to enhance the model's robustness.
- Refining the model by exploring deep learning approaches or ensemble techniques for better generalization.
- Extending the model's functionality to predict price depreciation over time based on usage patterns.

The deployment of this model through an API provides a practical and accessible solution for price estimation. Overall, this study highlights the potential of machine learning in automating and optimizing price predictions in the automotive industry.