Predicting Used Car Prices with Machine Learning

This project aims to address the challenges in the dynamic used car market by developing a machine learning model to accurately predict used car prices. The goal is to improve market efficiency, ensure fair transactions, and foster customer satisfaction for car dealerships, individual sellers, and buyers. Using a dataset from an online car marketplace with over 300,000 records, the study implements various models to estimate car prices based on features such as brand, model year, mileage, engine type, and accident history.

Menna Mahmoud EL-Bagoury

Youmna Wael Muhammed

Mariem Naeim

Problem Statement and Objectives

1

Market Complexity

The used car market faces challenges in determining fair prices due to numerous influencing factors.

2

Model Development

Develop a machine learning model to predict used car prices based on key features.

3

Factor Identification

Identify the most influential factors affecting used car prices.

4

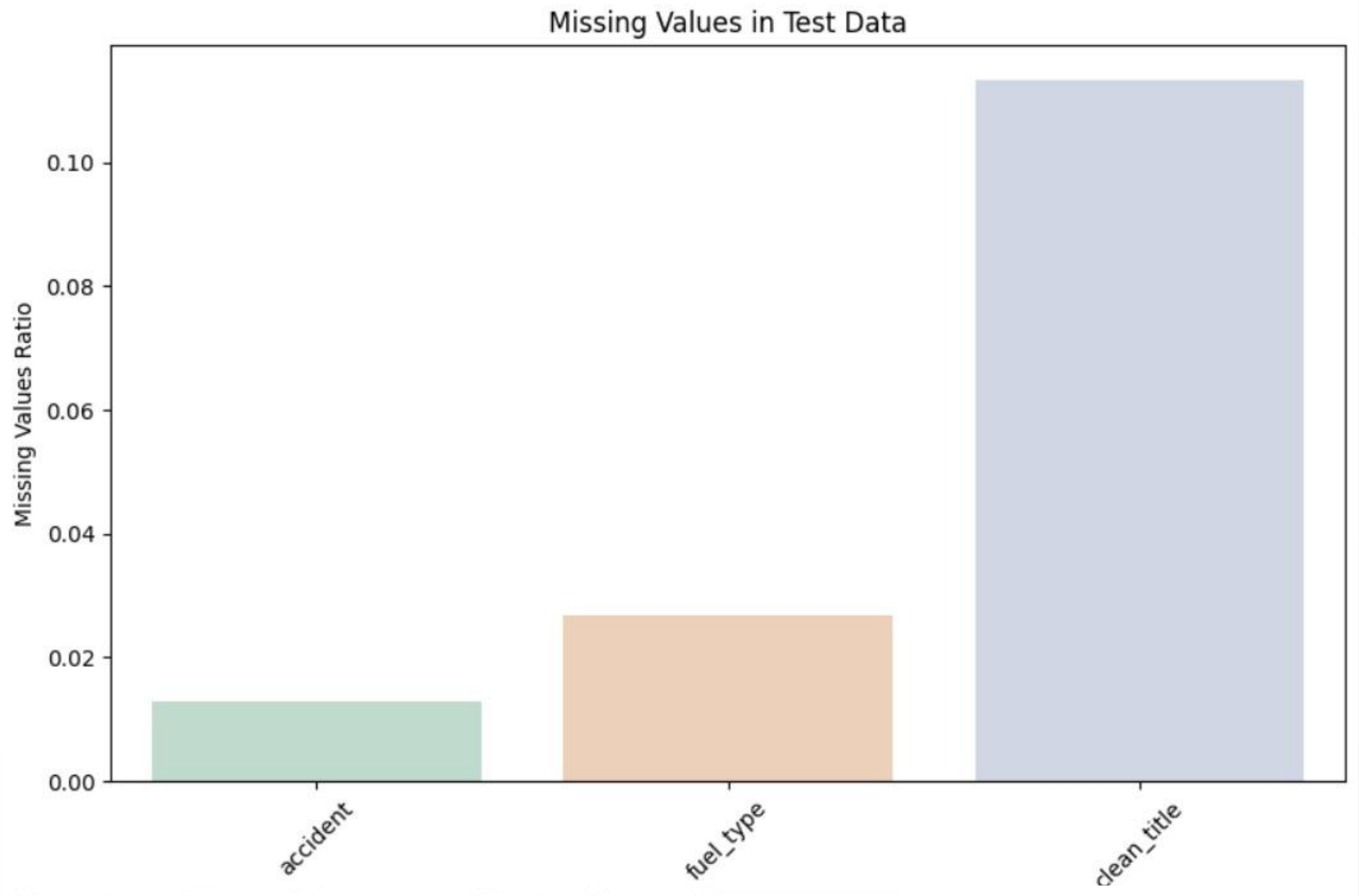
Improved Accuracy

Enhance prediction accuracy compared to traditional pricing models or manual strategies.

Dataset Overview

Training Dataset	**188,533 records**
Test Dataset	**125,690 records**
Total Columns	**13**

The dataset includes features such as id, brand, model, model year, mileage, fuel type, engine, transmission, exterior color, interior color, accident history, clean title status, and price. Several columns contained missing values that required handling during preprocessing.



Data Preprocessing

1

Missing Value Imputation

Handled missing values in fuel type, accident history, and clean title columns using iterative imputation techniques.

2

Feature Engineering

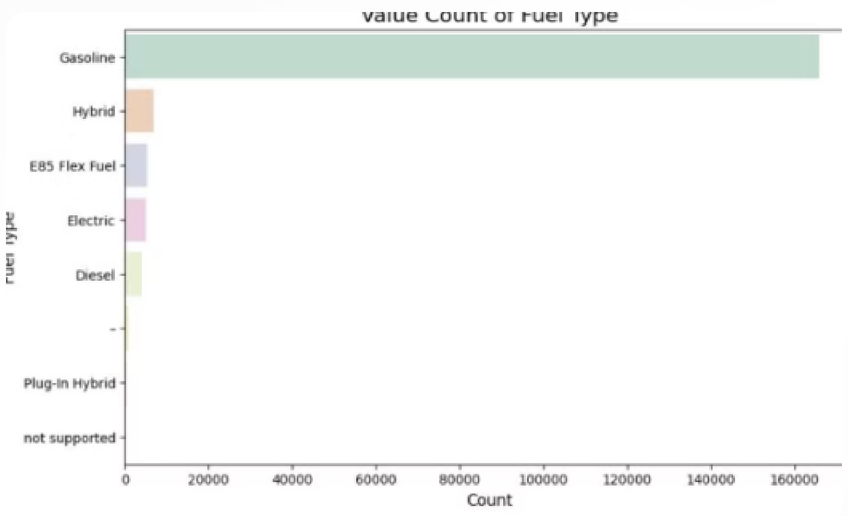
Created new features from the engine column, such as Horsepower and Displacement.

3

Encoding and Scaling

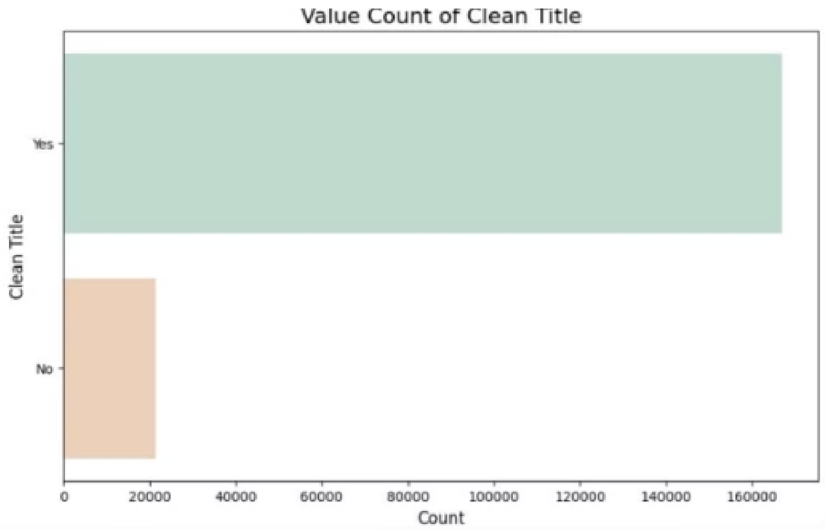
Applied one-hot encoding to categorical variables and standard scaling to continuous features.

Exploratory Data Analysis



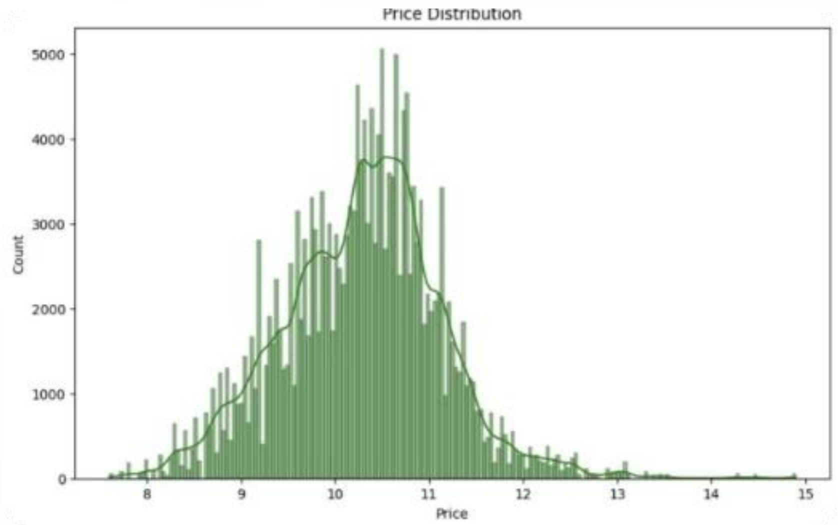
Fuel Type Distribution

Analysis of the distribution of fuel types among the cars in the dataset.



Clean Title Status

Visualization of the clean title status distribution in the dataset.



Price Distribution

Analysis of the price distribution across the used car dataset.

Table Before Data Preprocessing

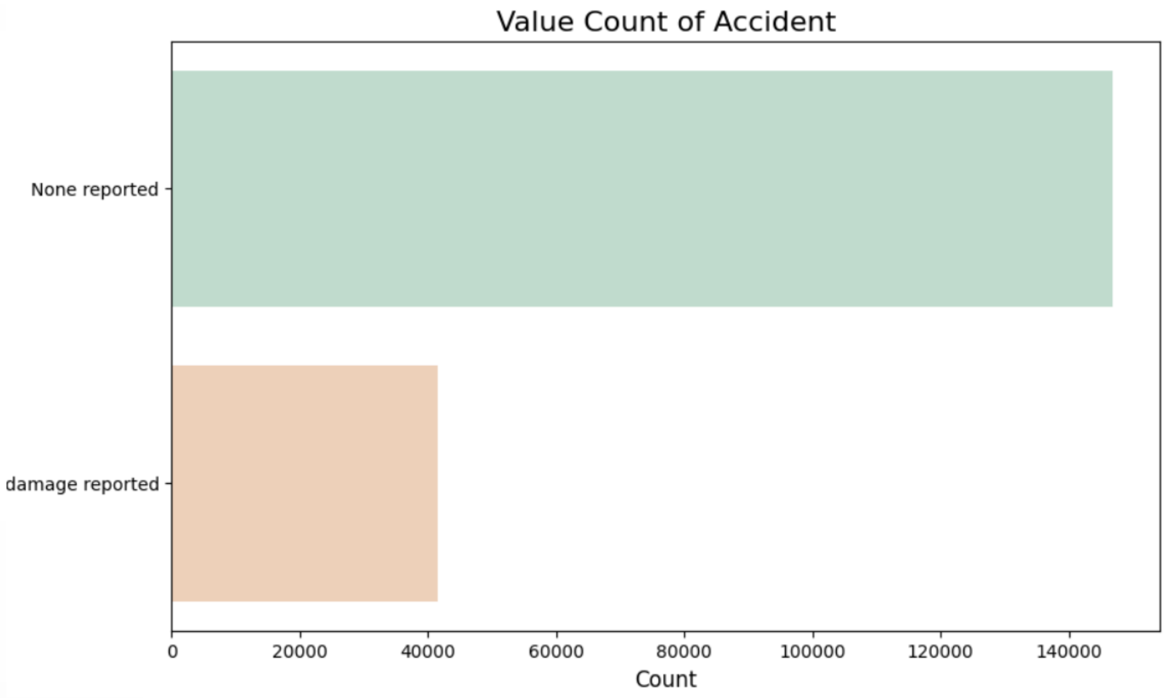
id	brand	model	model_year	milage	fuel_type	engine	transmission	ext_col	int_col	accident	clean_title	price
0	MINI	Cooper S Base	2007	213000	Gasoline	172.0HP 1.6L 4 Cylinder Engine Gasoline Fuel	A/T	Yellow	Gray	None reported	Yes	4200
1	Lincoln	LS V8	2002	143250	Gasoline	252.0HP 3.9L 8 Cylinder Engine Gasoline Fuel	A/T	Silver	Beige	At least 1 accident or damage reported	Yes	4999
2	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	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	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

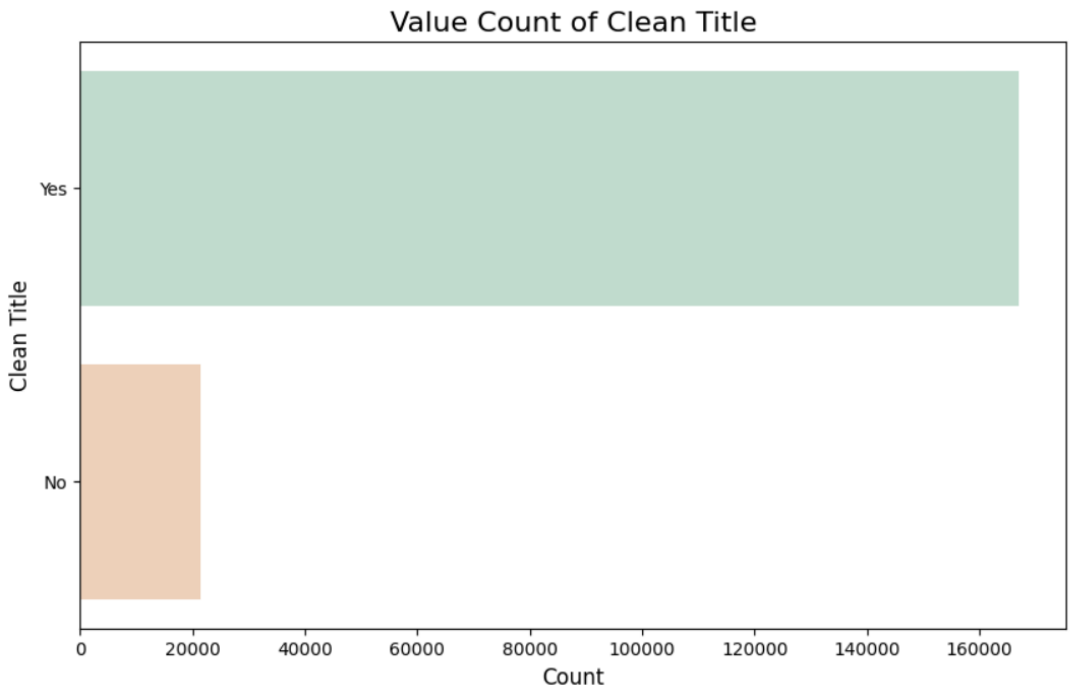
Dealing With Null Values

1- Null Values in “fuel_type”
is changed to “Electric”

2- Null Values in “clean”
is changed to “No”

3- Null Values in “accident”
is changed to “Non reported”





Data Preprocessing Summary

After dealing with null values:

1- Performing Feature Engineering using re library to deal with engine column and derive four columns:

(horsepower, displacement, engine type, cylinder count)

2- Handling missing values of theses four columns using:

a. Simple Imputation for Cylinder Column
b. Iterative Imputation for “displacement” and “cylinder type”

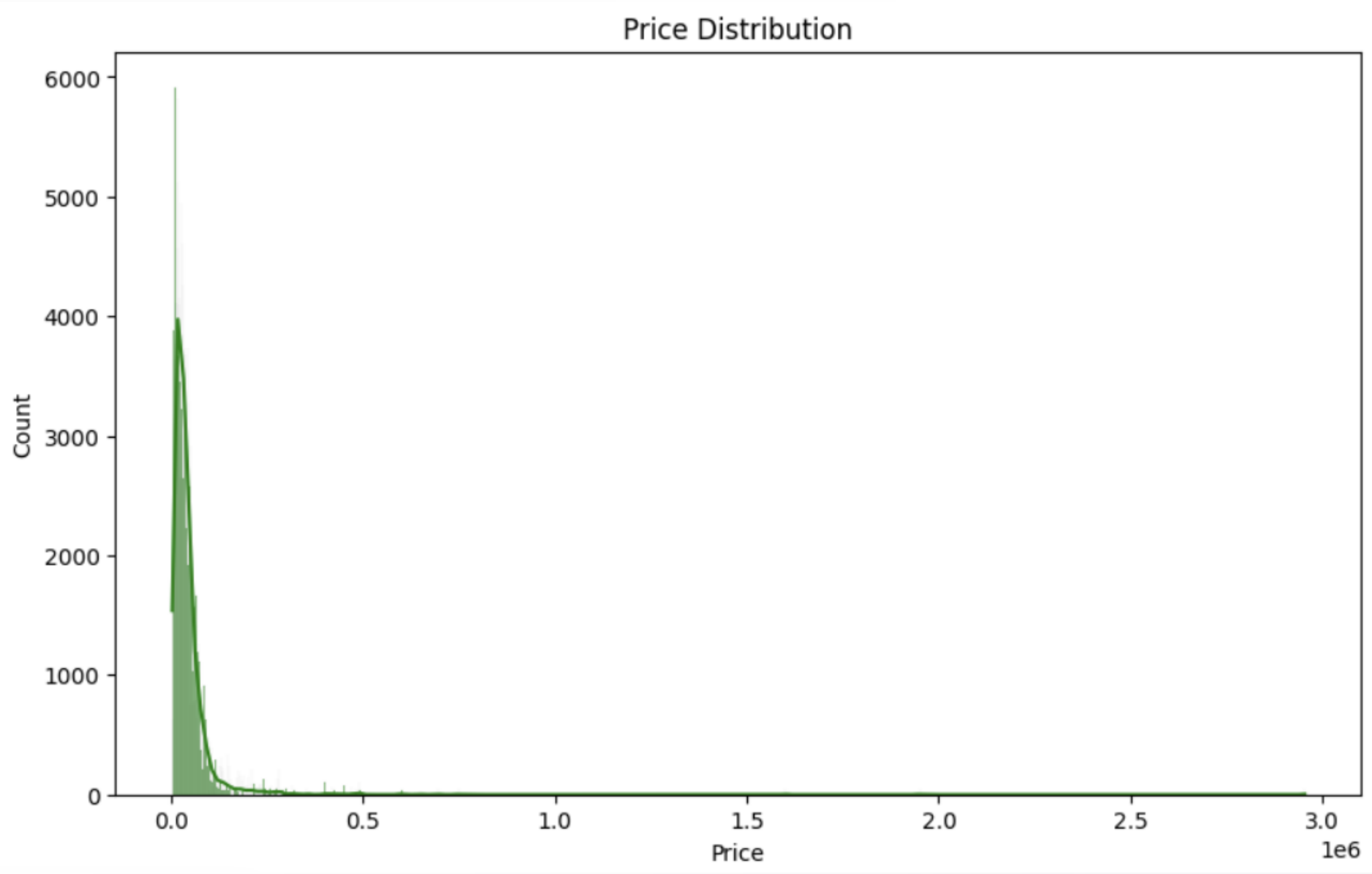
3- One Hot Encoding for the Categorical Columns:

`('brand', 'fuel_type’)`

4- Scaling for the Continous columns:

`('milage', 'Horsepower', 'Displacement', 'Cylinder Count', 'model_age’)`

Log Transformation For Price



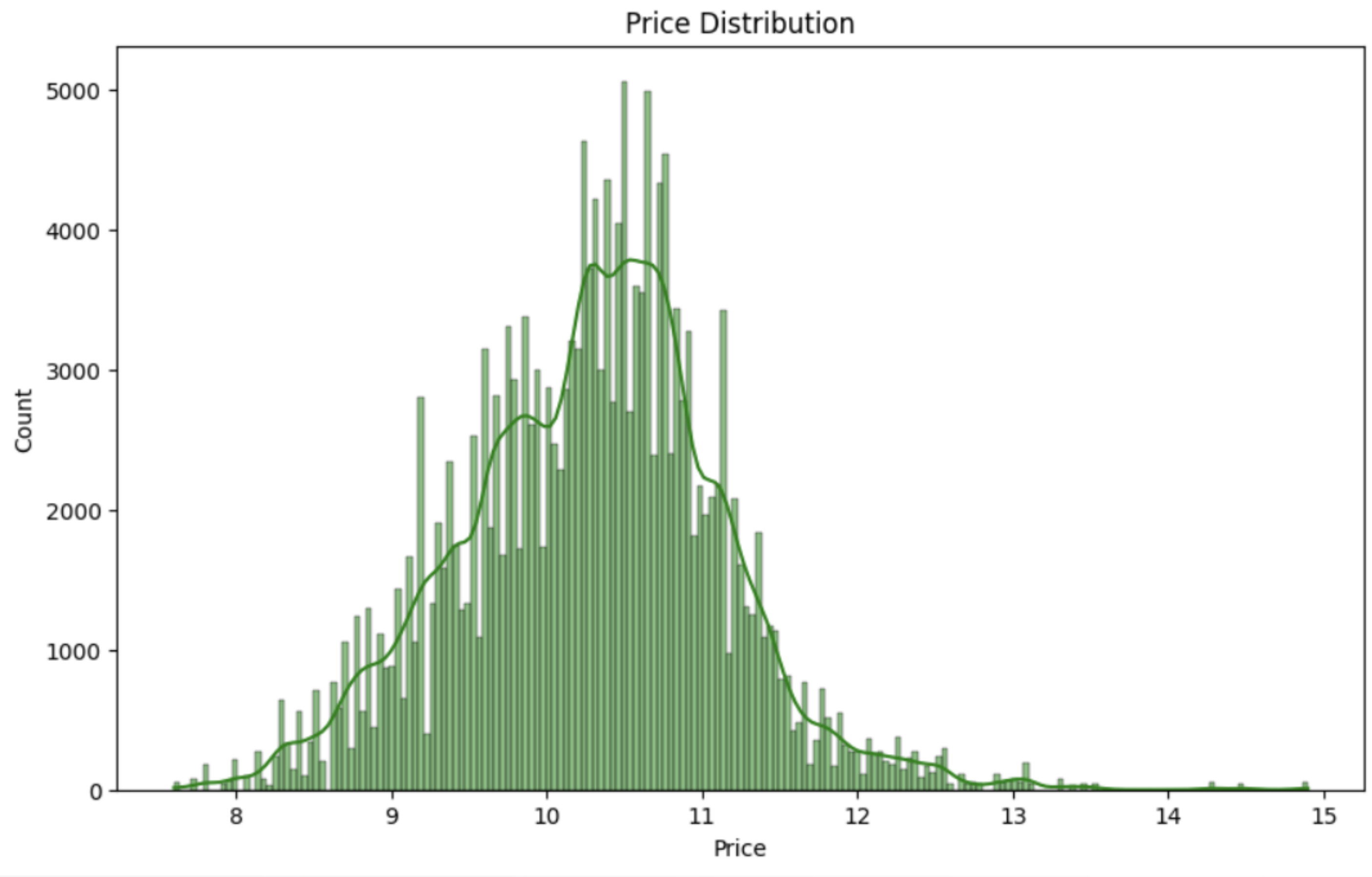
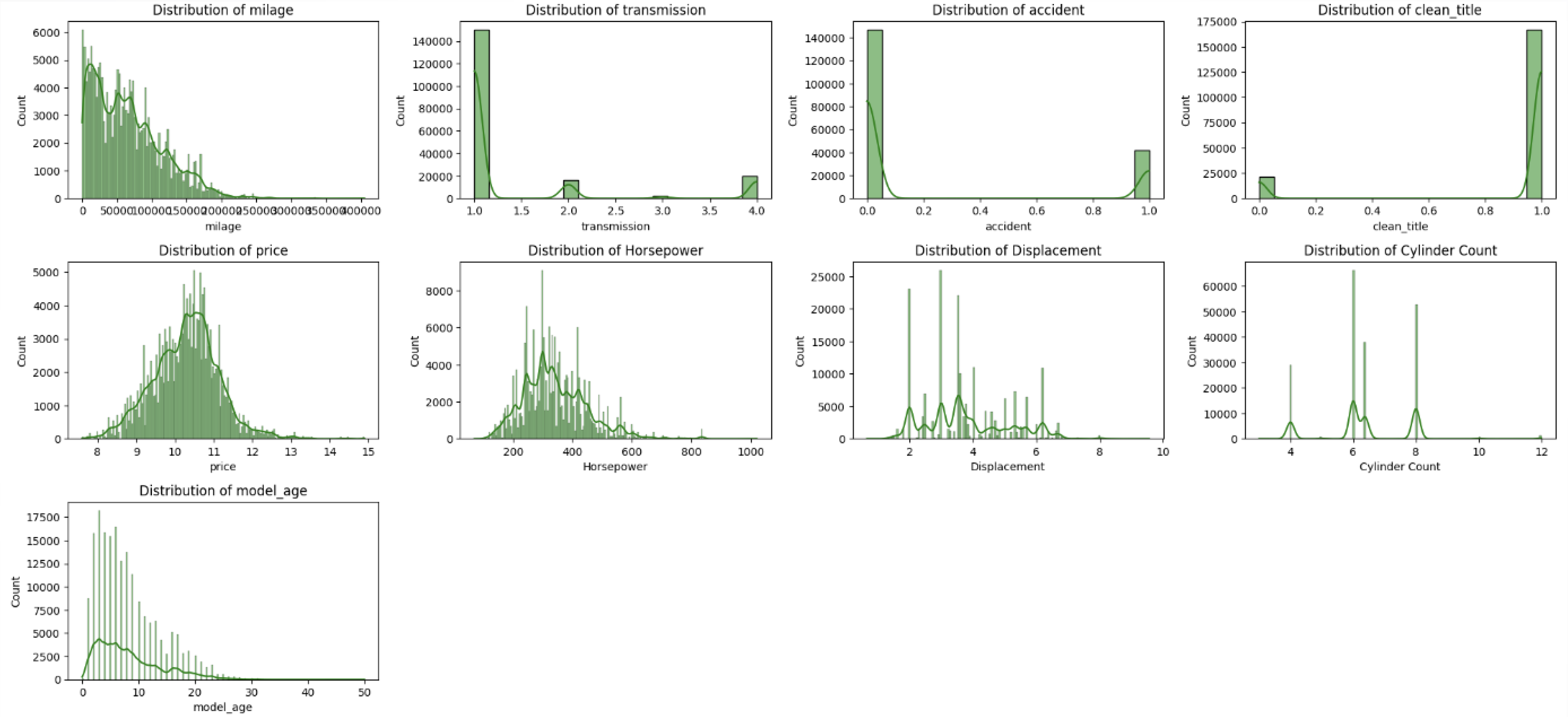


Table After Data Preprocessing

Note: Except For One Hot Encoding and Scaling

id	brand	milage	fuel_type	transmission	accident	clean_title	price	Horsepower	Displacement	Cylinder Count	model_age
0	MINI	213000	Gasoline	1	0	1	4200	172.000000	1.6	4.000000	17
1	Lincoln	143250	Gasoline	1	1	1	4999	252.000000	3.9	8.000000	22
2	Chevrolet	136731	E85 Flex Fuel	1	0	1	13900	320.000000	5.3	8.000000	22
3	Genesis	19500	Gasoline	4	0	1	45000	420.000000	5.0	8.000000	7
4	Mercedes-Benz	7388	Gasoline	1	0	1	97500	208.000000	2.0	4.000000	3
5	Audi	40950	Gasoline	1	0	1	29950	252.000000	2.0	4.000000	6
6	Audi	62200	Gasoline	1	0	1	28500	333.000000	3.0	6.000000	8
7	Chevrolet	102604	E85 Flex Fuel	1	0	1	12500	355.000000	5.3	8.000000	8
8	Ford	38352	Gasoline	1	0	1	62890	281.675334	2.7	6.374268	4
9	BMW	74850	Gasoline	4	0	1	4000	425.000000	3.0	6.000000	9

Visualizations Of column Values After Preprocessing



Note: Except For One Hot Encoding and Scaling

Model Development: Linear Regression

Model Performance

The Linear Regression model achieved an RMSE of 0.526 and an R-squared score of 0.61, indicating moderate performance in predicting used car prices.

Interpretation

This baseline model provides insights into linear relationships between features and car prices, serving as a benchmark for more complex models.

Model Development: Decision Tree

Hyperparameter Tuning

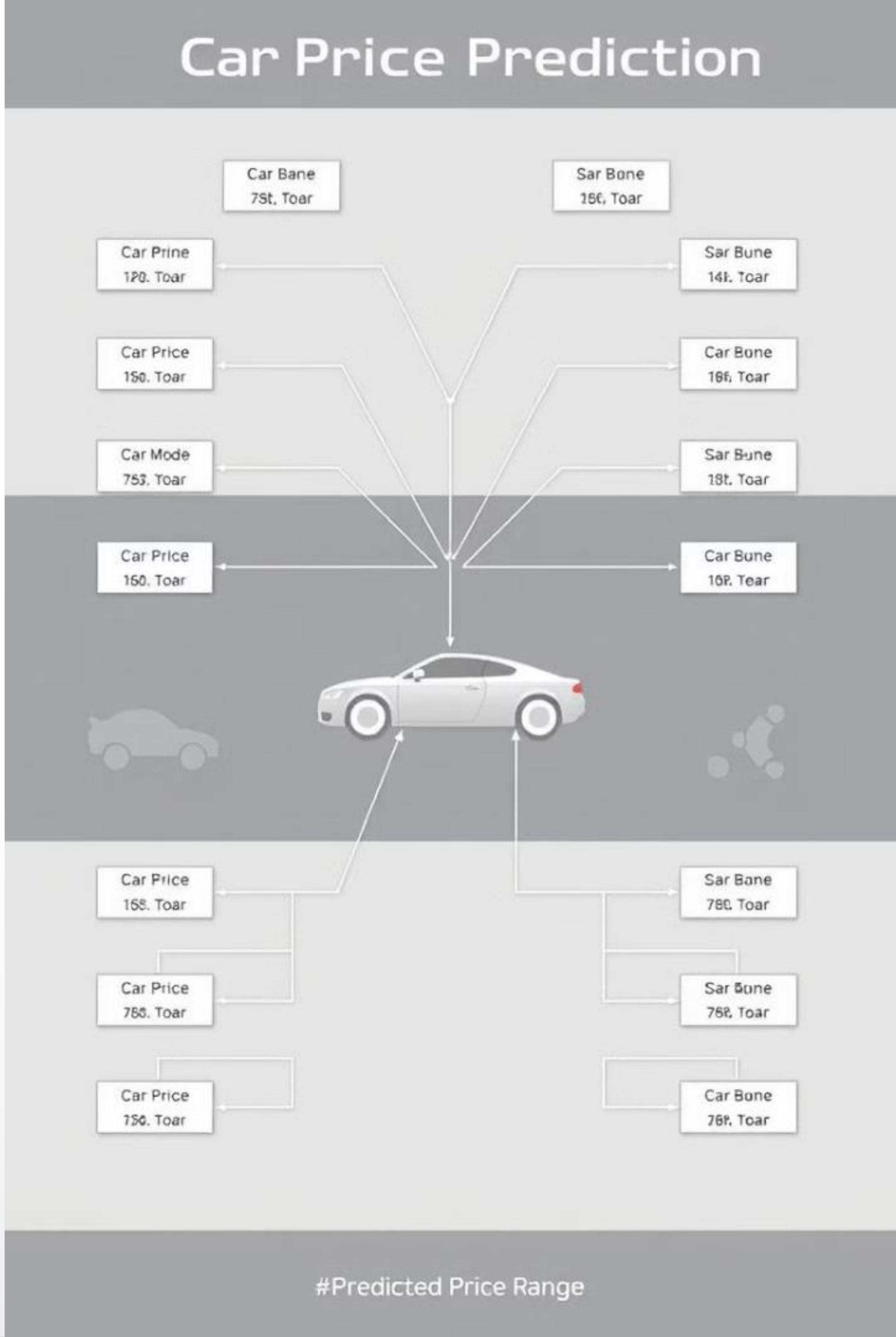
Optimized using RandomizedSearchCV with parameters: min_samples_split=5, min_samples_leaf=10, max_depth=10.

Performance Metrics

Achieved an RMSE of 0.51, MAE of 0.36, and an R-squared score of 0.63.

Model Insights

Demonstrated solid performance in predicting car prices, capturing non-linear relationships in the data.



Model Development: Random Forest

1

Hyperparameter Optimization

Tuned using RandomizedSearchCV: min_samples_split=5, min_samples_leaf=5, max_depth=10.

2

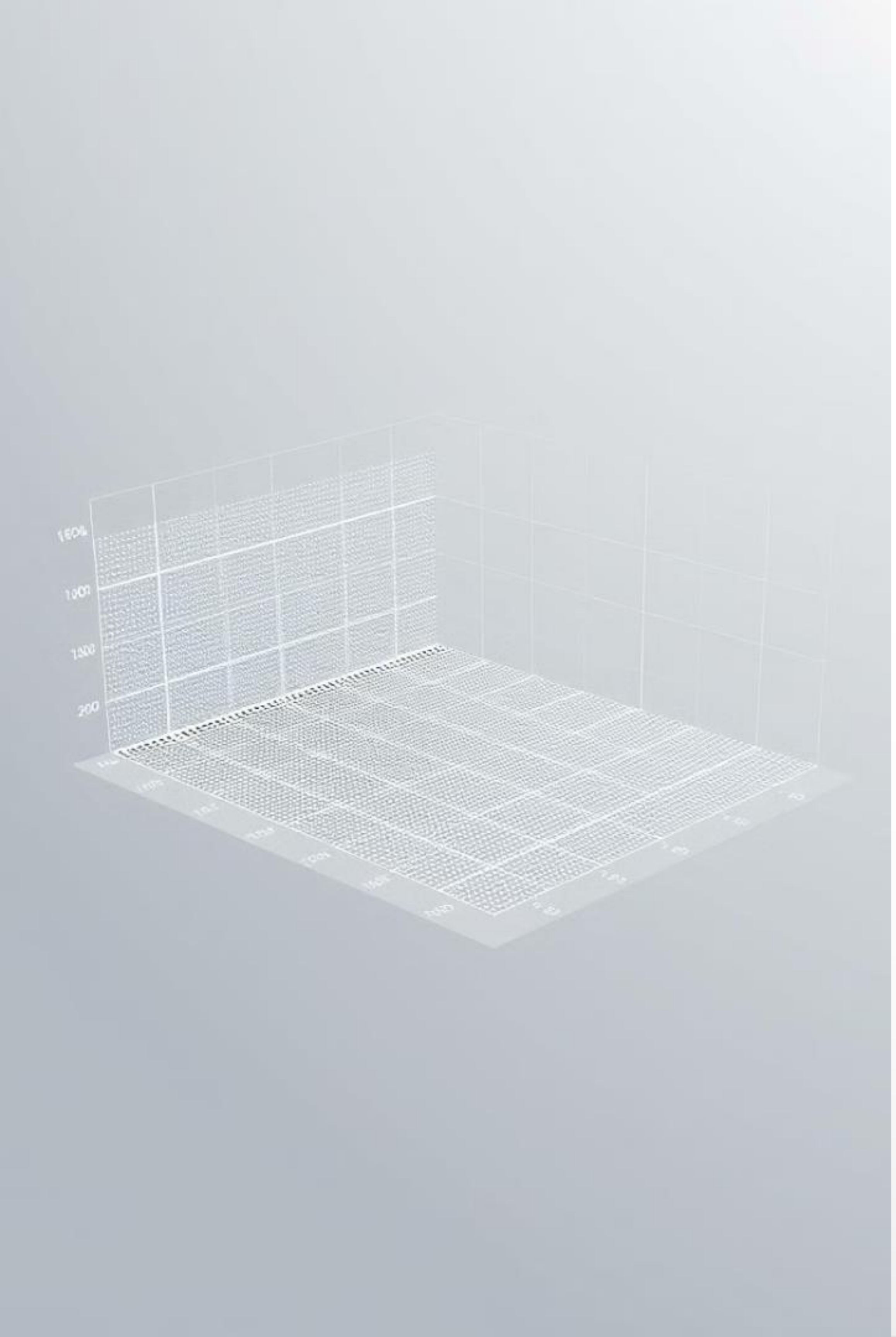
Superior Performance

Achieved RMSE of 69098.21, MAE of 0.35, and R-squared score of 0.65.

3

Best Model

Outperformed other models, demonstrating effectiveness in predicting used car prices.



Model Development: Support Vector Regressor (SVR)

1

Initial Implementation

SVR model was trained and optimized as part of the machine learning pipeline.

2

Performance Results

Yielded an RMSE of 71018.67, MAE of 0.42, and R-squared score of 0.53.

3

Comparative Analysis

Showed less predictive power compared to tree-based models in this specific use case.

Conclusion and Future Directions

Best Model

Random Forest outperformed other models in predicting used car prices.

Feature Engineering

Extracting the right features significantly impacted model performance.

Future Improvements

Explore model stacking, incorporate additional data sources, and further refine hyperparameters.

The project successfully developed a machine learning pipeline for predicting used car prices, with the Random Forest model showing the best performance. Future work could focus on expanding the dataset, experimenting with advanced models like XGBoost, and deploying the model in a real-time web application for practical use in the automotive market.

Questions Time

Thank you