Cars4u: Car Price Prediction

Comprehensive Analysis and Modelling

Author: Dmitry Luchkin, Data Analyst
Date: 3 August 2024

Agenda

- Project Overview
- Methodology
- Comprehensive Analysis
- Model Building
- Conclusions

Project Overview

Context

The **Car4u** project aims to develop a model to predict used car prices using historical data.

Goals

- Explore and prepare the Cars4u dataset
- Estimate car prices based on features like location, brand, and technical specs.

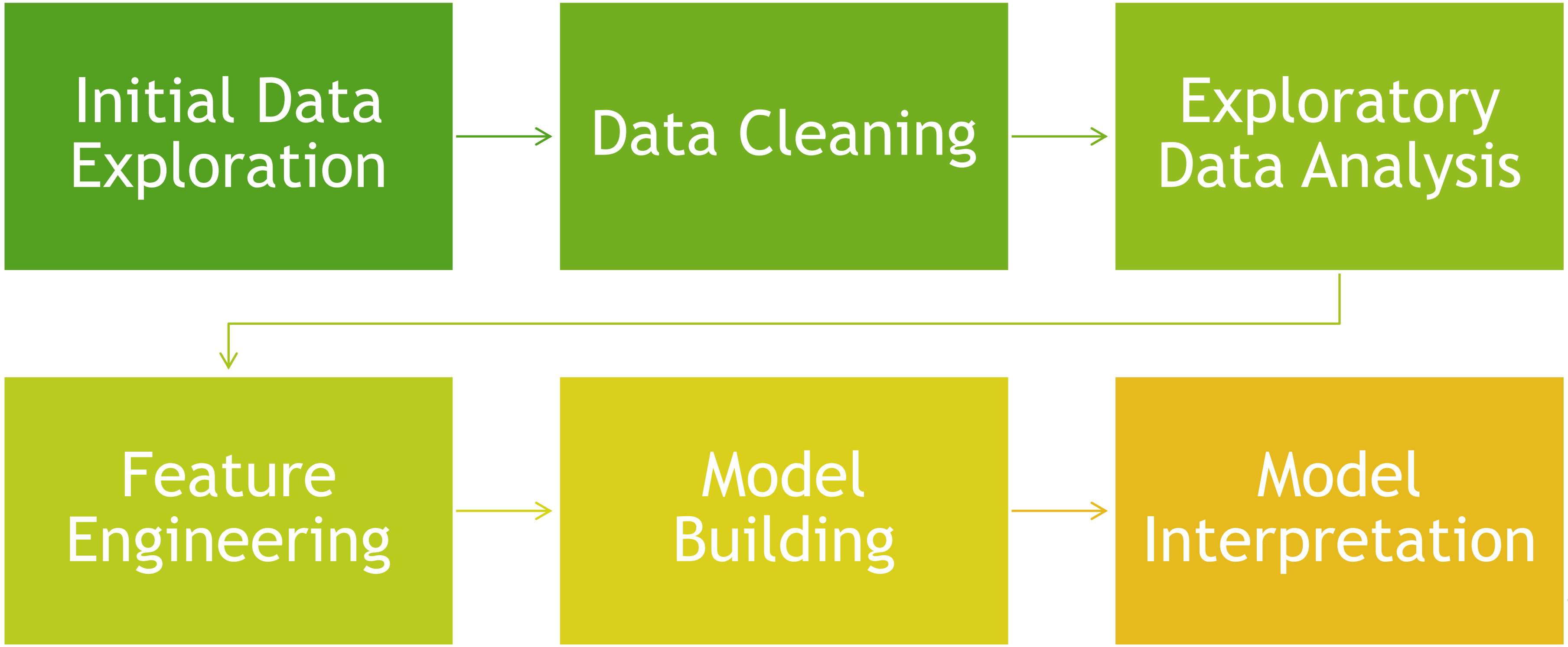
Key Findings

- The model predicts prices with high accuracy.
- Significant factors influencing price include:

car segment, location, engine size, number of seats, car age, mileage, kilometers driven, transmission type, and new car price.

Methodology

Processing Steps



Initial Data Exploration

Data Cleaning

Exploratory Data Analysis

Feature Engineering

Model Building

Model Interpretation

Comprehensive Analysis

Data Overview

The source of the dataset is Kaggle*

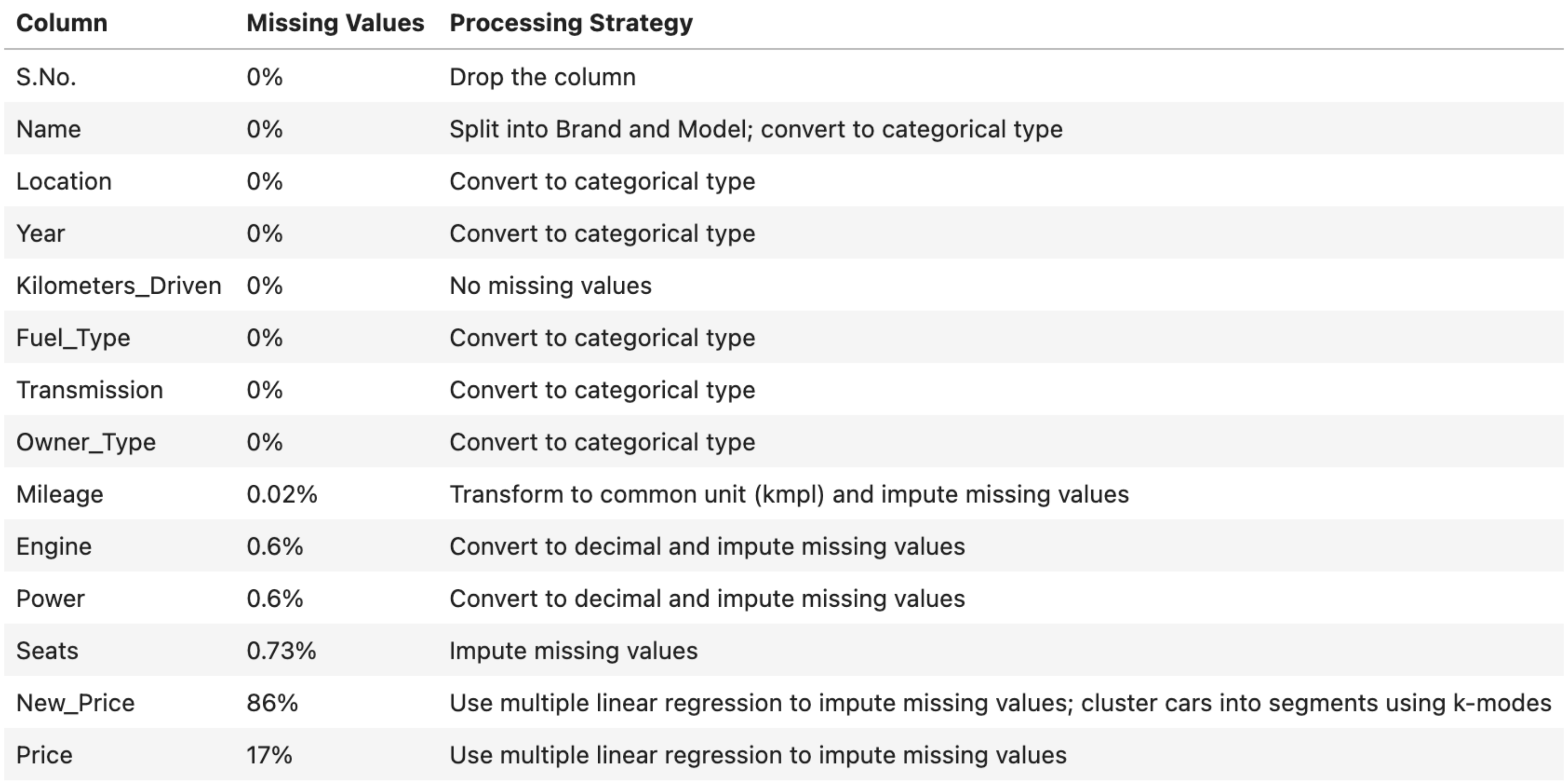
It includes 7253 rows and 14 columns.

No duplicate rows were found in the dataset.

* https://www.kaggle.com/datasets/sukhmanibedi/cars4u

Data Processing

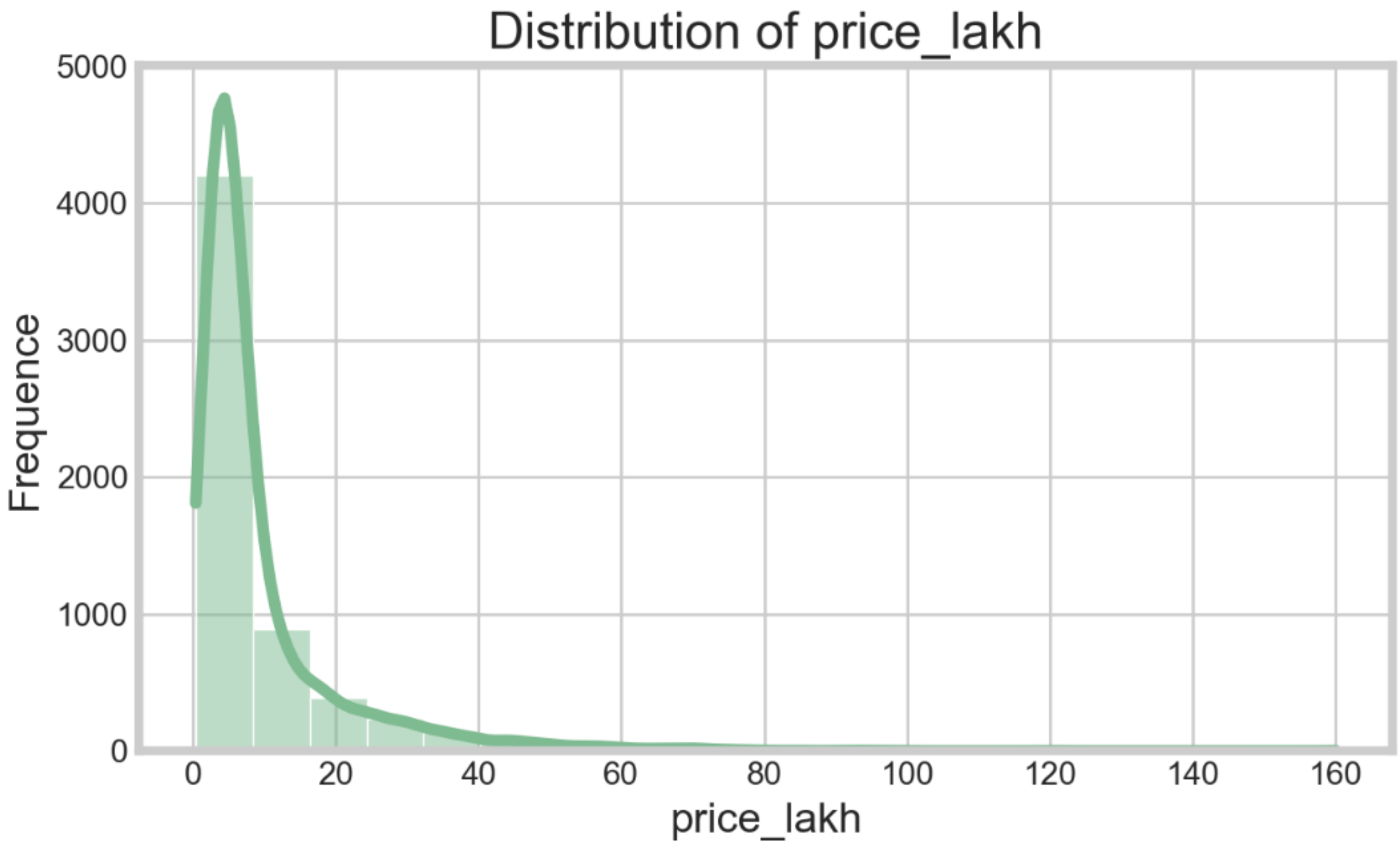
The initial data overview revealed missing values in the Mileage, Engine, Power, Seats, New_Price, and Price columns.



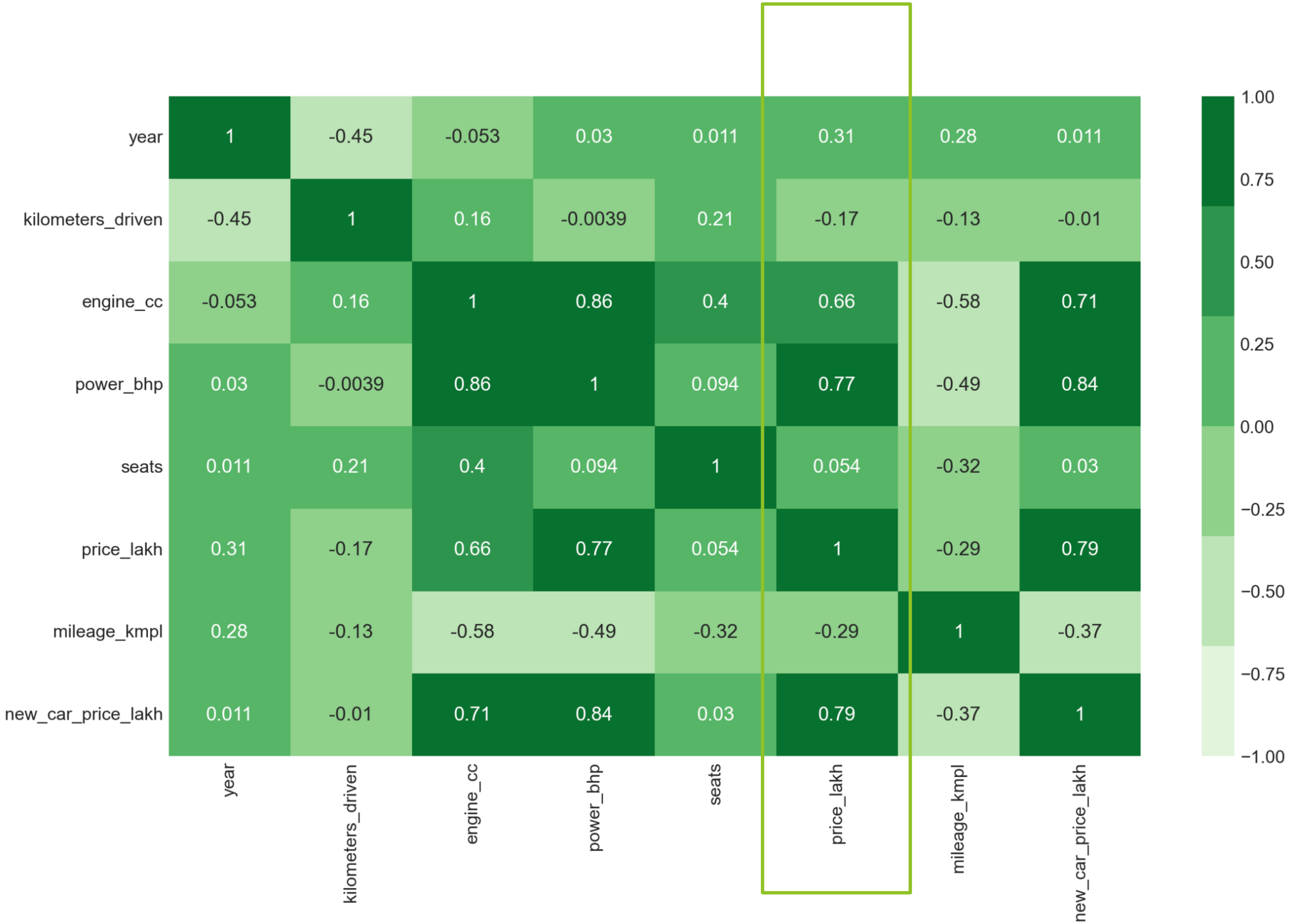
Column	Missing Values	Processing Strategy
S.No.	0%	Drop the column
Name	0%	Split into Brand and Model; convert to categorical type
Location	0%	Convert to categorical type
Year	0%	Convert to categorical type
Kilometers_Driven	0%	No missing values
Fuel_Type	0%	Convert to categorical type
Transmission	0%	Convert to categorical type
Owner_Type	0%	Convert to categorical type
Mileage	0.02%	Transform to common unit (kmpl) and impute missing values
Engine	0.6%	Convert to decimal and impute missing values
Power	0.6%	Convert to decimal and impute missing values
Seats	0.73%	Impute missing values
New_Price	86%	Use multiple linear regression to impute missing values; cluster cars into segments using k-modes
Price	17%	Use multiple linear regression to impute missing values

Car Price Distribution

- The distribution of car prices is right-skewed (skewness=**3.34**).
- The car price values need to be log transformed.



Car Price Correlation



	year	kilometers_driven	engine_cc	power_bhp	seats	price_lakh	mileage_kmpl	new_car_price_lakh
year	1	-0.45	-0.053	0.03	0.011	0.31	0.28	0.011
kilometers_driven	-0.45	1	0.16	-0.0039	0.21	-0.17	-0.13	-0.01
engine_cc	-0.053	0.16	1	0.86	0.4	0.66	-0.58	0.71
power_bhp	0.03	-0.0039	0.86	1	0.094	0.77	-0.49	0.84
seats	0.011	0.21	0.4	0.094	1	0.054	-0.32	0.03
price_lakh	0.31	-0.17	0.66	0.77	0.054	1	-0.29	0.79
mileage_kmpl	0.28	-0.13	-0.58	-0.49	-0.32	-0.29	1	-0.37
new_car_price_lakh	0.011	-0.01	0.71	0.84	0.03	0.79	-0.37	1

Model Building

Features Engineering

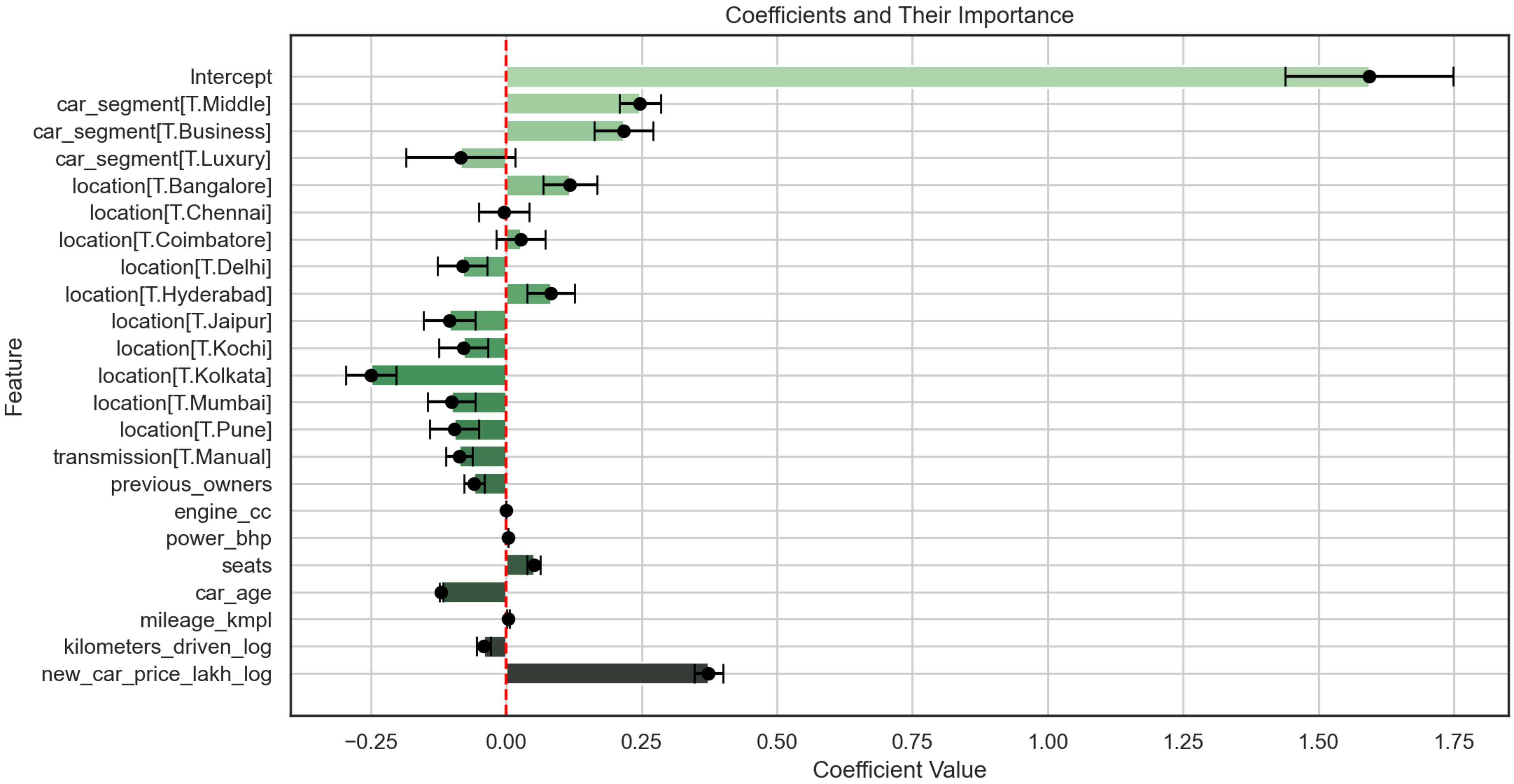
- From `owner_type` to `previous_owner`
  The category was encoded to a number of previous owners.
- From `year` to `car_age`
  The year of manufacturing was transformed to a car age.
- From `model` and `brand` to `car_segment`
  To avoid the multicollinearity cars were clustered by a new car price.

Build the Model

Hold-out validation 80/20

Metric	Training	Testing
Sum of Squared Errors (SSE)	340.1914	85.4336
Mean Absolute Error (MAE)	1.2140	1.2153
Mean Squared Error (MSE)	1.0733	1.0735
Root Mean Squared Error (RMSE)	1.0360	1.0361
Symmetric Mean Absolute Percentage Error (SMAPE)	15.9%	16.8%
R^2	0.9075	0.9061
Adjusted R^2	0.9071	0.9044

Model Interpretation



~90% of variable of used car price explained by the model

~1.21 Lakh is the average error

Significant predictors:

previous onwers, engine CC, power bhp, number of seats, car age, mileage km/l, kilometers driven, new car price, location, car segment

No multicollinearity

Conclusions

- Understanding which factors most influence car prices helps Car4u set competitive pricing and optimize inventory.
- The model assumes linear relationships and may not capture all complexities of the pricing dynamics.
- The model accurately predicts car prices based on car segment, location, and other features.
- **Future Work:** Explore additional features, non-linear models, and external factors to further improve predictions.
- **Recommendations:** Implement the model to guide pricing decisions and continually update the model with new data.

Thank You