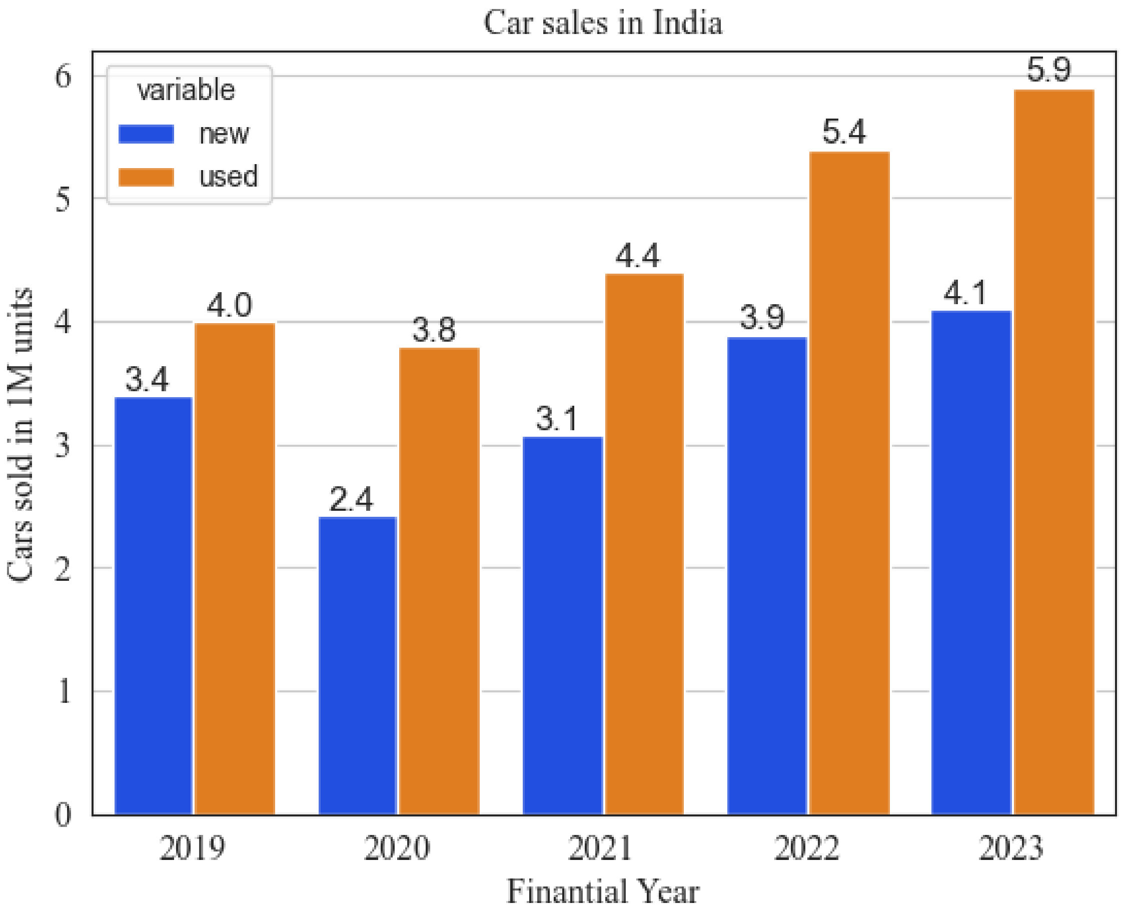
Cars4U

USED CARS PRICE PREDICTION

Grzegorz Finke / Grzegorz.Finke@gmail.com

PROBLEM DEFINITION

- Notable shift from new car purchases to pre-owned vehicles.
- Surge in the popularity of used cars is particularly significant, **8.2 million** by FY2025.
- Organized segment is expected to expand to **45%** by FY2025.



Ref. [1, 2, 3, 4]

PROBLEMS TO SOLVE

COMPETITION

Business potential
has been recognized.

Major players
(Audi, BMW, Mercedes-Benz, Porsche)

Online automotive marketplaces
(**OLX Autos** or **CarTrade**).

Ref. [1, 3, 5]

METHODOLOGY

Pricing used cars is
particularly complex

Numerous factors influencing price

Changing trends
and customers' preferences

Unexpected situations

SOLID TOOLS

that help gain an
advantage in the market

KEY QUESTIONS

1. How to build a predictive model with high accuracy?

2. Which features influence the price?

3. How to improve price predictions?

SOLUTION APPROACH

Identify data limitations

Make dataset uniform

Categorize cars by price segment

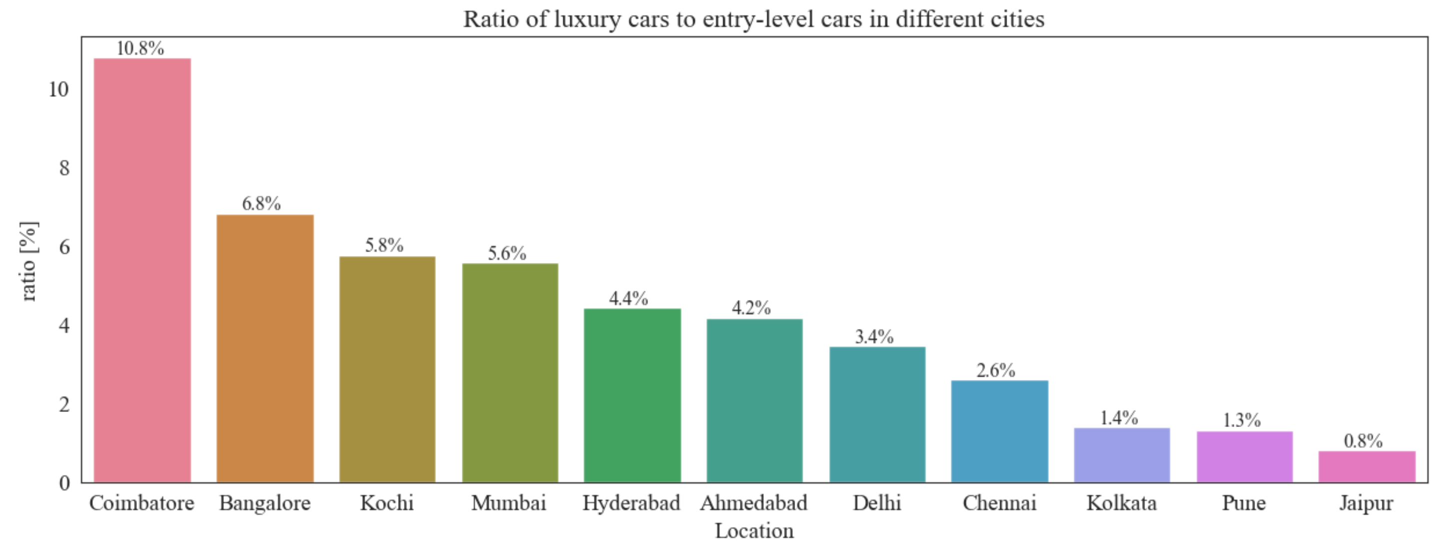
Improve dataset

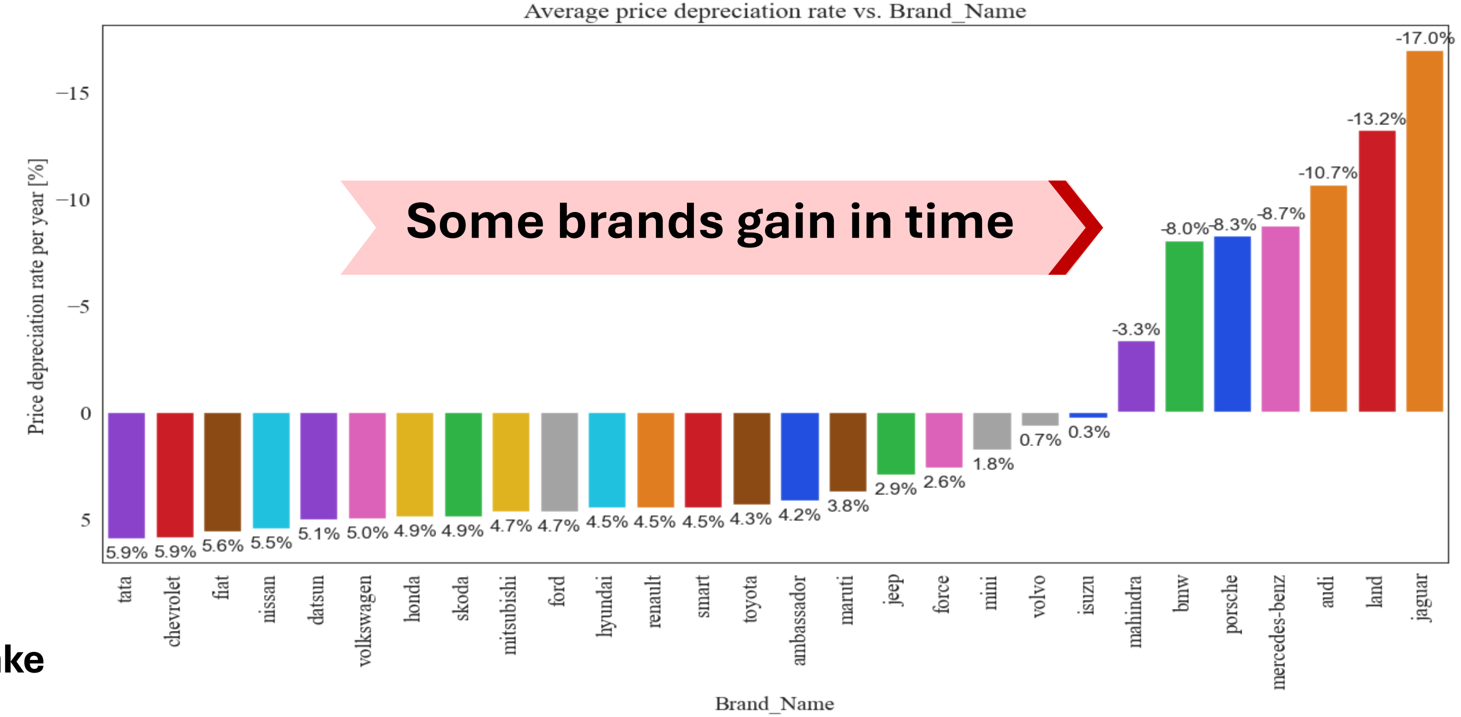
Increase diversity by adding more features relevant to predicted value

Find interesting relationships within the dataset

SOLUTION APPROACH

Find interesting relationships within the dataset.





SOLUTION APPROACH

CHOOSE THE BEST MODEL

METRICS

RMSE

(Root Mean Square Error)

MAE

(Mean Absolute Error)

MAPE

(Mean Absolute Percentage Error)

R^2

(R-squared)

PROPOSED MODEL

Start with SIMPLE solutions

Helps to identify weak spots faster

Explore more COMPLICATED approaches

Remember to balance complexity and computational resources.

GRADIENT BOOSTING REGRESSION

Combines the strengths of multiple weak learners

New models are sequentially trained to correct the errors made by previous models.

Provides insights into feature importance, helping in understanding the underlying data.

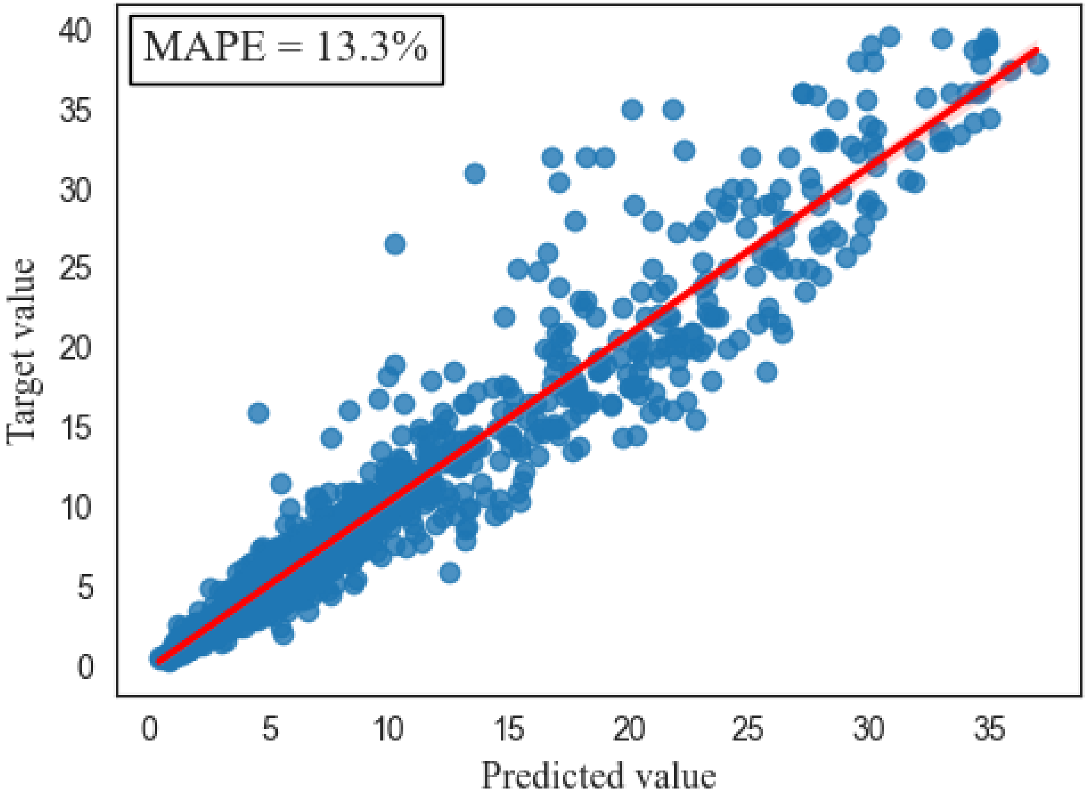
FINAL SOLUTION

Residual Distribution

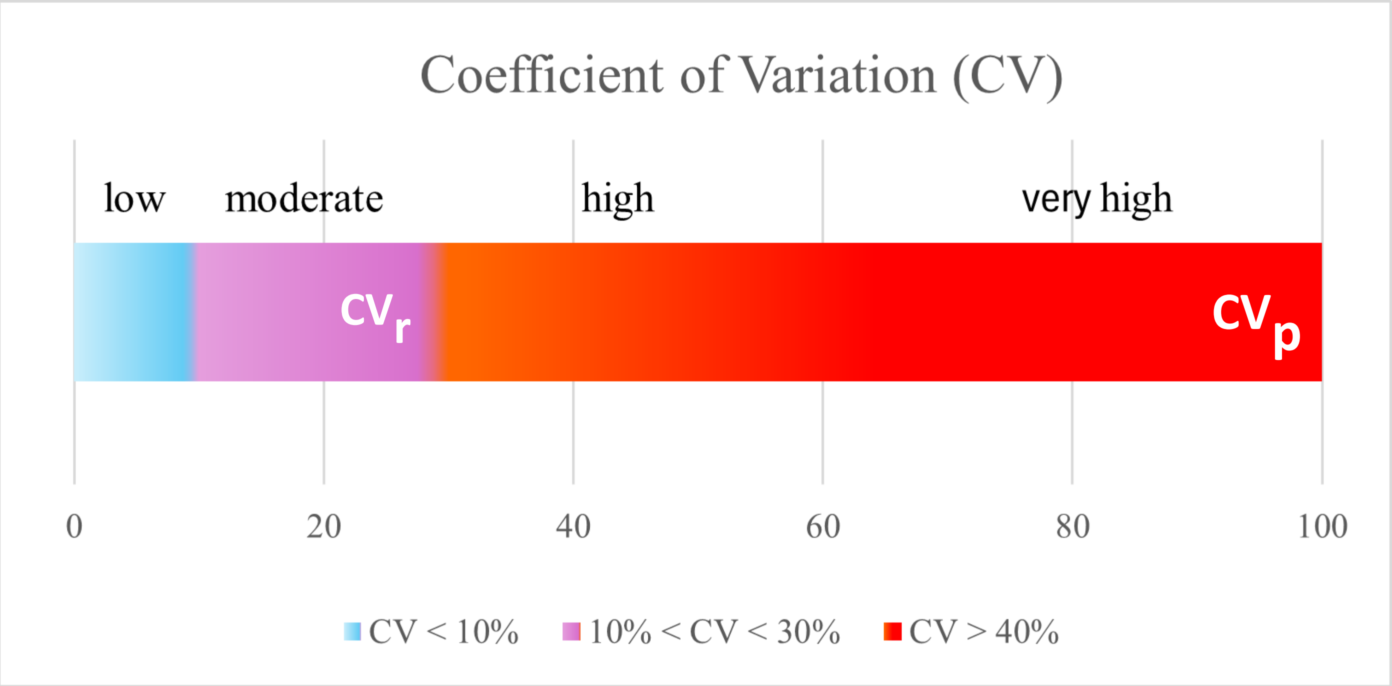
As car prices increase, the distribution of residuals becomes sparser. This behavior supports the data segmentation approach and suggests, that further segmentation of entry-level and mid-tier cars could enhance model performance, which is recommended for business planning.

MAPE = 13.3%

Used cars prices in 100.000 INR



FINAL SOLUTION



CV_p - Coefficient of Price Variation
CV_r - Coefficient of Residual Variation

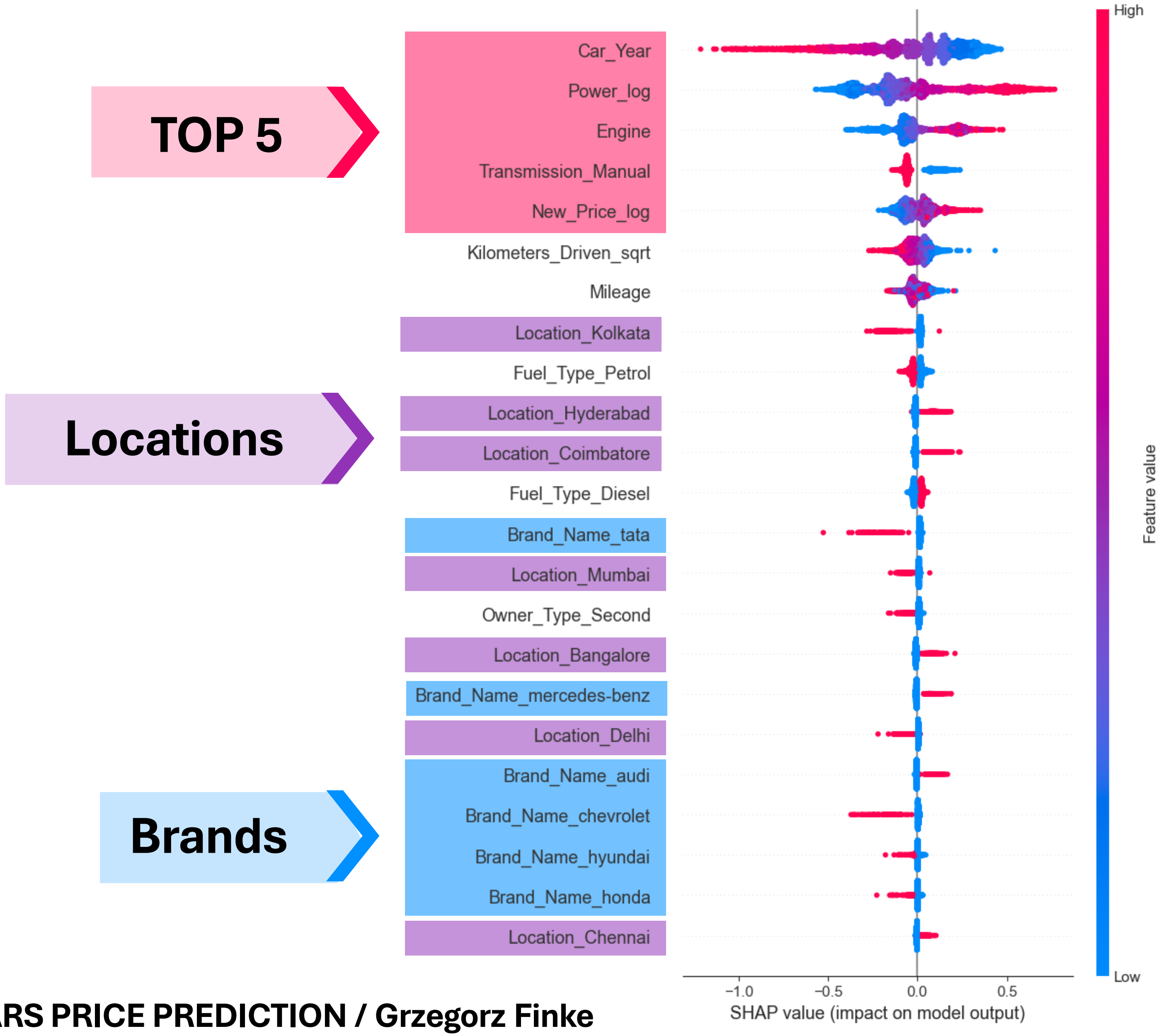
Residual Values

$CV_p = 91\%$
$CV_r = 25\%$

The model has significantly reduced the variability relative to the target variable. This indicates that while the model performs well, the high variance in the target variable may contribute to the residual variance.

FINAL SOLUTION

SHAP* vs. feature value



***SHAP -** Shapley Additive Explanations

BUSINESS SOLUTIONS AND IMPLEMENTATIONS

Focus on High-Impact Features

- Use the most important features to create targeted marketing campaigns that highlight the top features.
- Older cars pricing should be more competitive. Offering warranties or certified pre-owned programs can help mitigate concerns about the car age.

Tailor Marketing Strategies

- Use the model's insights to guide procurement.
- Regularly update the inventory based on market trends and model predictions to ensure alignment with customer preferences in specific locations.

Price Adjustment Mechanism

- Develop and integrate dynamic pricing tools that adjust prices based on the top features' current market trends.

BUSINESS SOLUTIONS AND IMPLEMENTATIONS

Enhance Customer Experience

- Educate customers about the importance of key features in determining a car's value.
- Educate customers on the value of pre-owned cars, emphasizing benefits like cost savings and environmental impact.

Consider Market Segmentation and Consumer Preferences

- Tailor marketing campaigns to target market segments based on car age, brand, or car price.
- Recognize customers' preferences to offer personalized deals and promotions.

Future Trends Monitoring

- Stay ahead of market trends.
- Monitor changes in customer preferences for electric vehicles.
- Monitor emerging segments concerning technological innovations.

RISKS AND CHALLANGES

1 There is no such thing as a perfect model; even the most sophisticated **models have limitations** and require ongoing refinement to adapt to new data and changing market conditions.

2 **Data quality is critical** to model effectiveness; inaccuracies, biases, or incomplete datasets can lead to unreliable predictions, making robust data validation and preprocessing essential.

3 **High competition from powerful players** can make market entry difficult, requiring a focus on innovation, strategic positioning, and a deep understanding of market dynamics to differentiate and succeed.

4 **Customer sentiment and trends** are inherently variable, influenced by changing economic conditions, technological advancements, and social dynamics, which force constant monitoring of the market and regular data updates.

SUMMARY

Price is variable and depends on multiple factors; incorporating more features improves prediction accuracy.

Data quality is crucial for model effectiveness; proper data cleaning and preprocessing should be integral parts of the data pipeline.

Integrate **up-to-date data** to reflect the latest trends.

Look for relationships between features to better understand the market and develop effective business strategies.

Ensure the **dataset aligns with customers' expectations** for optimal model accuracy.

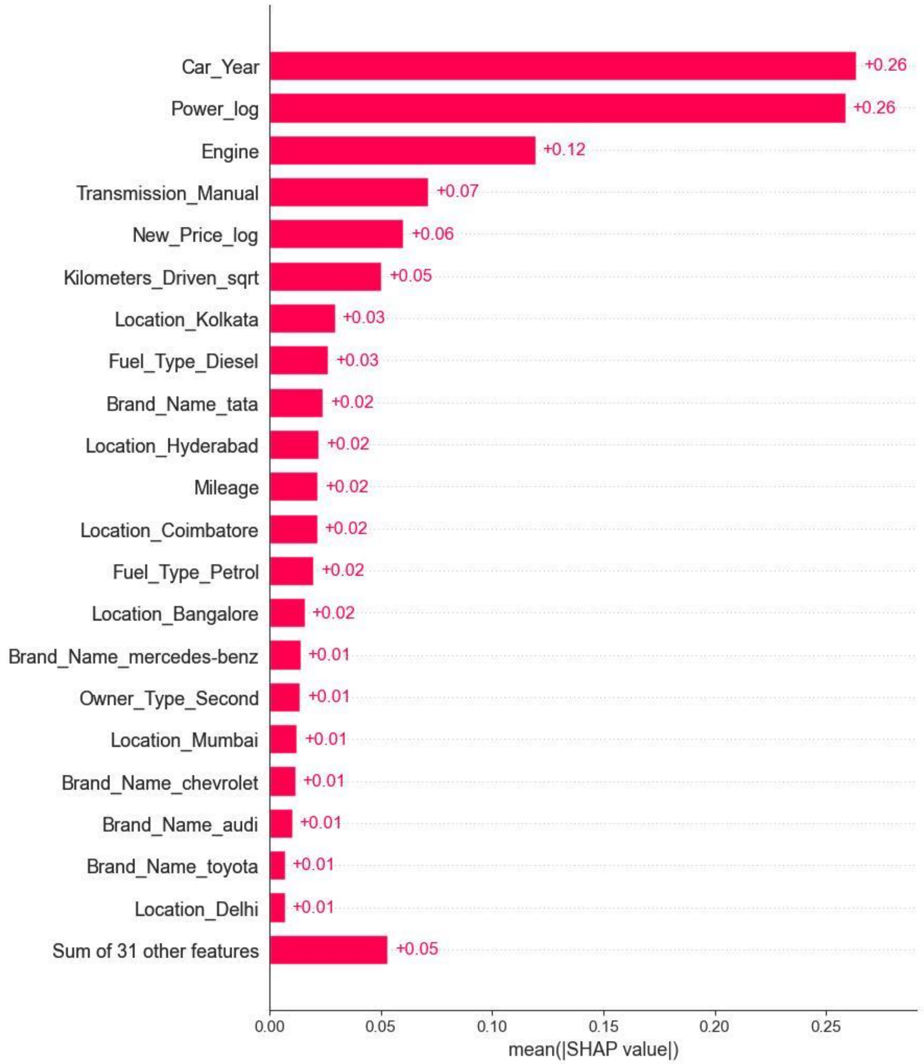
The developed model predicts used car prices with a **MAPE of 13.3%**.

BIBLIOGRAPHY

1. Autocar India
2. HT Auto
3. Mordor Intelligence
4. LeadSquared
5. 6Wresearch

APPENDIX

Gradient Boosting Regression Model



Mean absolute SHAP values represent the average impact of each feature on the model's predictions (Price_log)

These values provide insights into the relative importance of each feature in the model.

Gradient Boosting Regression Model

Model performance

Metric	Train	Test
Root Mean Squared Error (RMSE)	1.32	1.99
Mean Absolute Error (MAE)	0.77	1.05
Mean Absolute Percentage Error (MAPE)	10.37%	13.32%
R-squared (R^2)	0.96	0.93

Tuned parameters

Model Parameters	Value
max_depth	5
max_features	0.5
min_samples_leaf	3
n_estimators	120

XGBoost Model

Model performance

Metric	Train	Test
Root Mean Squared Error (RMSE)	1.32	1.99
Mean Absolute Error (MAE)	0.77	1.05
Mean Absolute Percentage Error (MAPE)	10.37%	13.32%
R-squared (R^2)	0.96	0.93

Tuned parameters

Model Parameters	Value
max_depth	7
colsample_bytree	0.5
n_estimators	120
reg_alpha	0.1
reg_lambda	0.1

XGBoost vs. GBR

Performance comparison of the models

Metric	XGB	GBR
Root Mean Squared Error (RMSE)	1.91	1.99
Mean Absolute Error (MAE)	0.96	1.05
Mean Absolute Percentage Error (MAPE)	12.1	13.32
R-squared (R^2)	0.94	0.93

- XGBoost model was also evaluated and showed an improvement in accuracy compared to the GradientBoostingRegressor.
- XGBoost as more sophisticated model offers several advantages over the GBR algorithm e.g., includes regularization to prevent overfitting at the cost of higher computational demands, however. Therefore, a balance must be struck between the complexity of the model and the available computational resources. Prioritize simpler, less resource-intensive improvements before moving to more complex techniques.