Used Cars Price Prediction

Foundations for Data Science

Capstone Project

15.08.2024

Contents / Agenda

Executive Summary

This project outlines Cars4U's initiative to develop a robust pricing model for the used car market. By utilizing advanced machine learning techniques, the company aims to optimize pricing strategies and enhance market positioning.

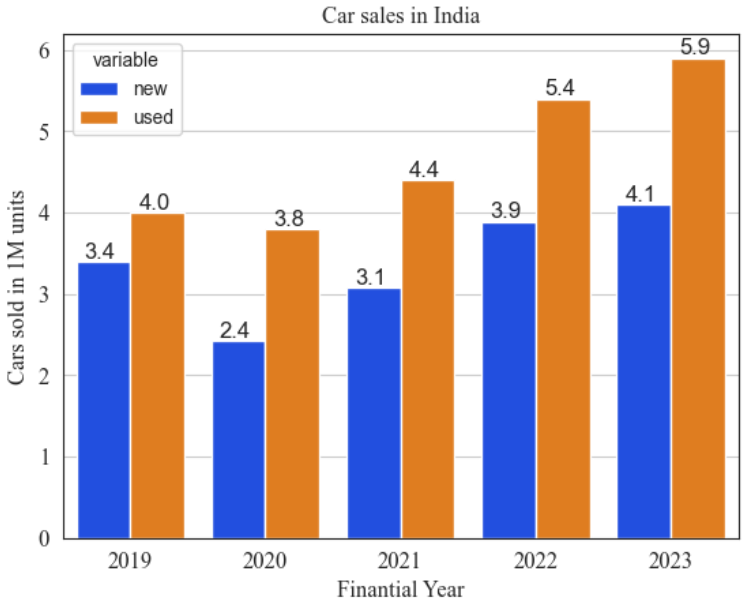
The GradientBoostingRegressor model, developed using data up to 2019, serves as a solid foundation and has the flexibility to incorporate current market trends. The model was built with data limited to entry-level and mid-tier price ranges as they show better market consistency than luxury cars. While the proposed model demonstrates good performance, considering the high variation in prices in the used car market, there is still room for improvement. Future enhancements should include:

1. Integrating up-to-date data.
2. Adding more predictive features.
3. Ensuring high-quality new data.
4. Segmenting cars into different price ranges to create more tailored marketing and sales strategies.

To maximize the model's potential and profitability, it is advisable to address all these aspects.

Problem summary

Recent trends in the Indian car market indicate a notable shift from new car purchases to pre-owned vehicles. Although both segments are growing, the surge in the popularity of used cars is particularly significant, with sales projected to reach **8.2 million** units by FY2025 [1, 2].



This business potential has been recognized by major players such as **Suzuki, Mahindra, Volkswagen, Audi, BMW,** and **Porsche** [1, 3, 5] as well as by online automotive marketplaces like **OLX Autos** or **CarTrade,** which number is increasing rapidly across India.

Currently, the majority of transactions in the used car market are informal, with only 17% to 25% of the market being organized. However, this organized segment is expected to expand to **45%** by FY2025 [1, 3, 4], presenting substantial opportunities for new businesses, especially that technological advancements and digitalization has already made buying and selling used cars more accessible and trustworthy [2, 3, 4].

High competition in this area requires solid tools that help gain an advantage in the market.

A critical challenge for new entrants, though, is determining the optimal pricing strategy to ensure profitability. Pricing used cars is particularly complex due to the numerous factors influencing their value. Therefore, the key objective of this project is to **build a predictive model for used car prices** in India to optimize pricing strategies and enhance business profitability.

The regression model developed in this study aims to provide insights into:

- **Feature Identification:** Recognize the features that most significantly impact the final price.
- **Model Accuracy:** Build a predictive model with high accuracy in car price estimation.

The analysis and model predictions discussed will assist the company in:

- Developing effective business strategies,
- Managing inventory efficiently,
- Supporting strategic decision-making,
- Optimizing revenue.

By addressing these areas, the company can create a competitive edge in the market.

Solution Design

The objective of this research was to create a trustworthy model that predicts used car prices with high accuracy and repeatability. To achieve this, it was essential to understand the relationships between model components, their impact on the target value, and estimate their importance for model development. The research was composed of two parts:

- Analysis of features relationships,
- Price prediction model development.

Additionally, I separated the luxury/sports cars from the entry-level/mid-tier cars. To distinguish these two groups, I used a market price threshold of 4,000,000 INR (~$47,500) [6, 7, 8]. The decision to exclude the high-price car segment was based on two primary reasons:

- **Impact of Outliers**: Luxury and sports cars introduced strong outliers that negatively influenced the model's accuracy. These outliers made the results less reproducible across

different training and testing splits, reducing the model's reliability. By excluding these outliers, the model provided more consistent and trustworthy results.

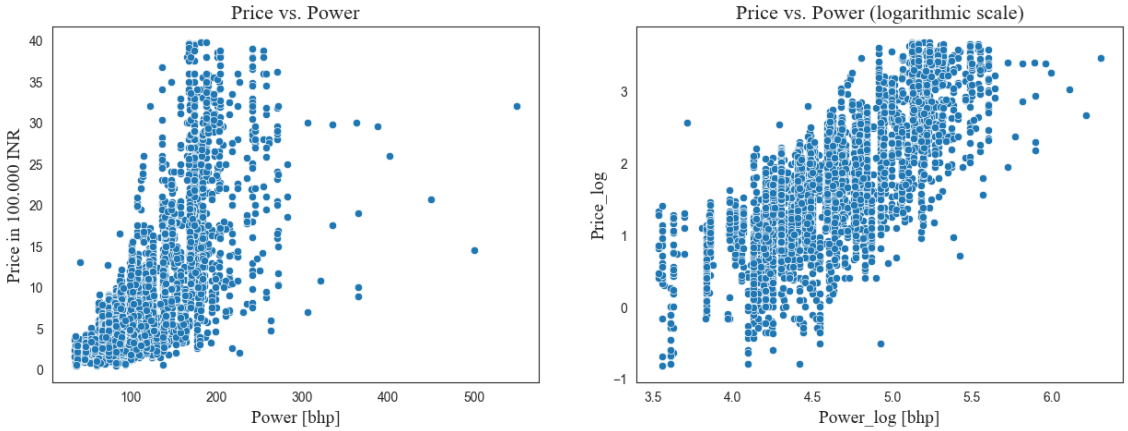
- **Distinct Market Dynamics**: The market for luxury and sports cars differs significantly from that of entry-level cars. Factors such as pricing range, depreciation rates, durability, target demographics, and customer expectations vary widely between these segments. A model that includes both would be very general or would necessitate handling these differences resulting in its increased complexity.

By creating separate datasets for luxury/sports cars and entry-level/mid-tier cars, I ensured that the datasets were more consistent, leading to more reliable analysis and predictions.

Analysis of Features Relationships

To better understand the influence of various features on car prices and their interrelationships, I created additional features, including car age, depreciation rate, average mileage per year, or extracted brand name to have a more general view. Below are some interesting relationships observed:

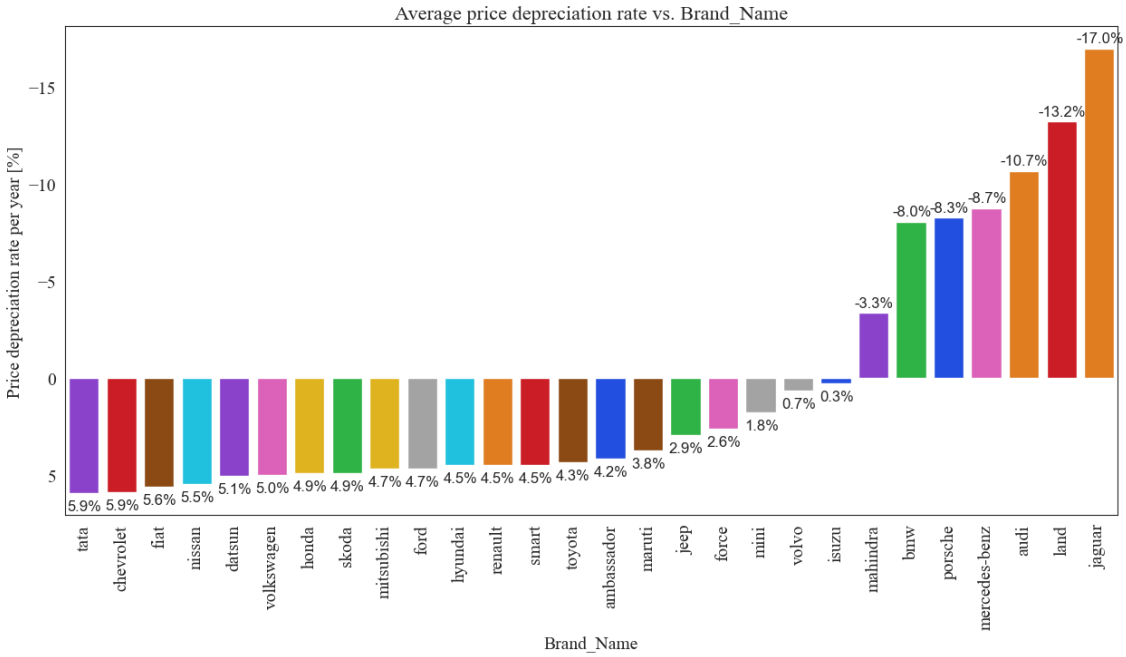
Power vs. Price (correlation 0.79)



Observations:

- The power of a car is highly correlated with its price, indicating that more powerful cars tend to be more expensive. This feature is crucial for model predictions and business targeting.
- Despite excluding luxury and sport cars from the dataset, some outliers remain, which could negatively impact error metrics. This could be due to very simple grouping methodology used. However, it also shows how important it is to maintain a consistent and uniform dataset.

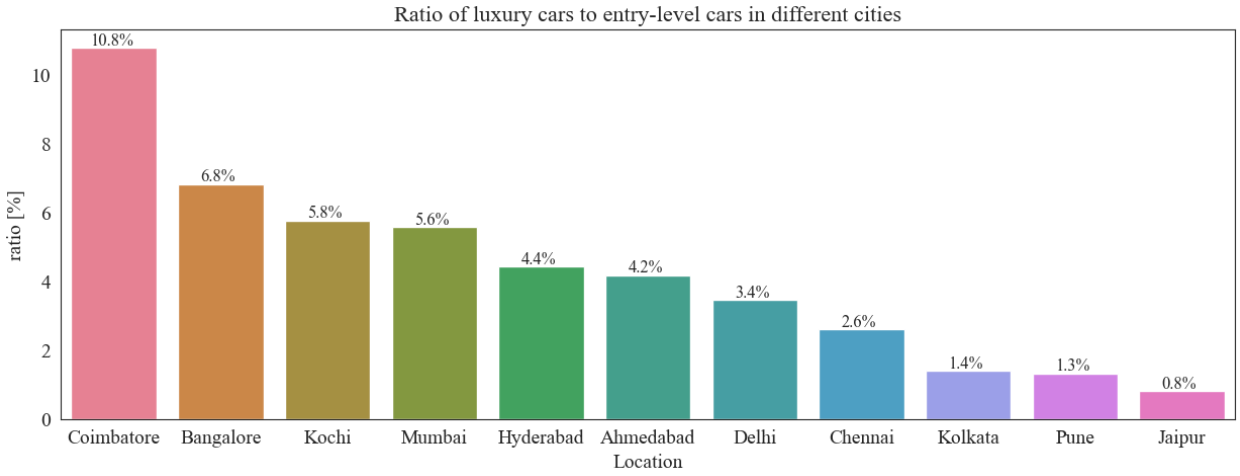
Price Depreciation Rate vs. Brand Name



Observations:

- Most car brands show a price depreciation over time, which is expected. However, certain brands exhibit value growth, particularly some European 'big players' and popular Indian cars like Mahindra.
- These unusual relationships might indicate data errors or extreme cases, especially when the dataset contains a small number of such cars.
- Understanding these depreciation patterns can aid in managing inventory efficiently and support strategic decision-making.

Luxury/sport Cars to Entry-level/mid-tier Cars Ratio



Observations:

- Comparing the ratio of luxury/sport cars to entry-level/mid-tier cars can provide a quick insight into market targeting and potential business localizations. This should be, of course, followed by deeper analysis for comprehensive understanding.

The insights gained from analyzing feature relationships are valuable for developing effective business strategies. Such observations underscore the importance of thorough feature analysis in refining predictive models and crafting strategic business initiatives.

Price Prediction Model Development

Research Phase Overview

During the research phase, multiple models were explored to predict used car prices, including both linear and non-linear methods. The evaluation process involved the use of full and partial datasets, as well as hyperparameter tuning to optimize models performance. After extensive evaluation, the final proposed model is a **GradientBoostingRegressor** (GBR). This model was fine-tuned to achieve an optimal balance between accuracy and reliability. Even though the model development focused on a partial dataset that includes entry-level and mid-tier cars, it can be easily implemented on the high-tier segment as well.

Model Development and Dataset

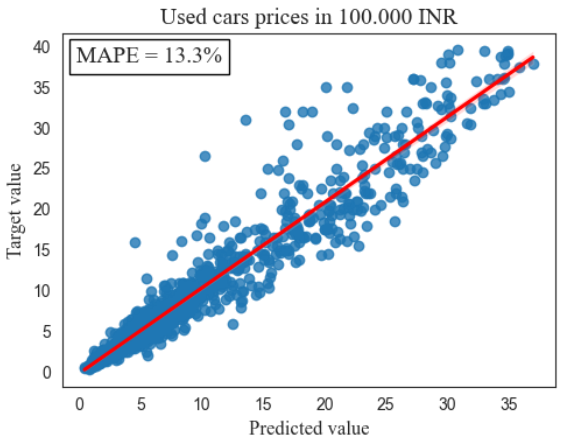
The model development focused on a partial dataset encompassing entry-level and mid-tier cars. However, the model can be easily adapted to the high-tier car segment. Given the high variability in used car prices (coefficient of variation, CV = 91%), the chosen GBR model demonstrated good accuracy on the test data, with the following performance metrics:

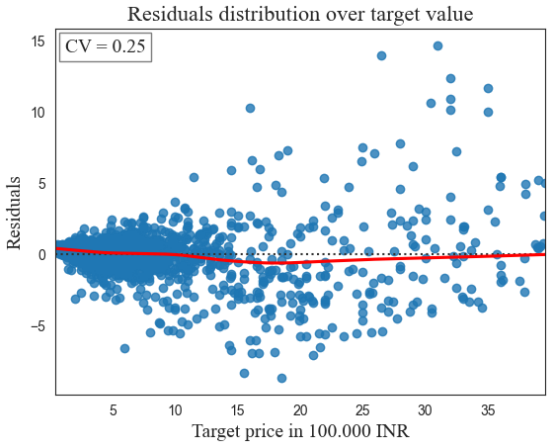
Metric	Value
Root Mean Squared Error (RMSE)	1.99
Mean Absolute Error (MAE)	1.05
R-squared (R^2)	0.93
Mean Absolute Percentage Error (MAPE)	13.3%

These metrics were calculated on the original scale after applying the exponential function to the predictions.

Model Performance and Residual Analysis

The visual relationship between the predicted values and the actual target values indicates the model's accuracy.





The key observations from the residual analysis are:

- **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 and is recommended for business planning.
- **Residual Values:** Residual values tend to be higher for higher-priced cars. The coefficient of residual variation (CV) is 25%, which is moderate. Considering the high CV for the target price (91%), 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.

Analysis and Key Insights

The approach to analysis mirrors the approach to solution design, encompassing both the examination of feature relationships and the direct evaluation of the model's results. This dual approach provides a comprehensive understanding of the factors influencing car prices and the model's predictive capabilities.

Wider Perspective from Feature Analysis

From the broader perspective provided by feature analysis, the key insights emphasize the importance of dataset uniformization. Different car segments cater to varying customer expectations, and for optimal model accuracy, the analysis should focus on specific price tiers. This segmentation ensures that the model accounts for the unique characteristics and demands of each market segment.

Price Diversification and Coefficient Variation

The prices of new cars exhibit significant diversification, which is reflected in a high coefficient of variation. This variability must be carefully managed to improve model precision. The presence of latent features, which are challenging to detect straightforwardly, such as customer sentiment, also impacts car prices. Understanding these hidden factors requires ongoing market trend analysis, continuous data updates, and cross-analysis of various features specific to the predefined market.

Key Insights from SHAP Analysis

The SHAP (SHapley Additive exPlanations) value plot provides detailed insights into the impact of different features on the model's output. Here are the key observations from the SHAP analysis:

1. **Most Influential Features:**

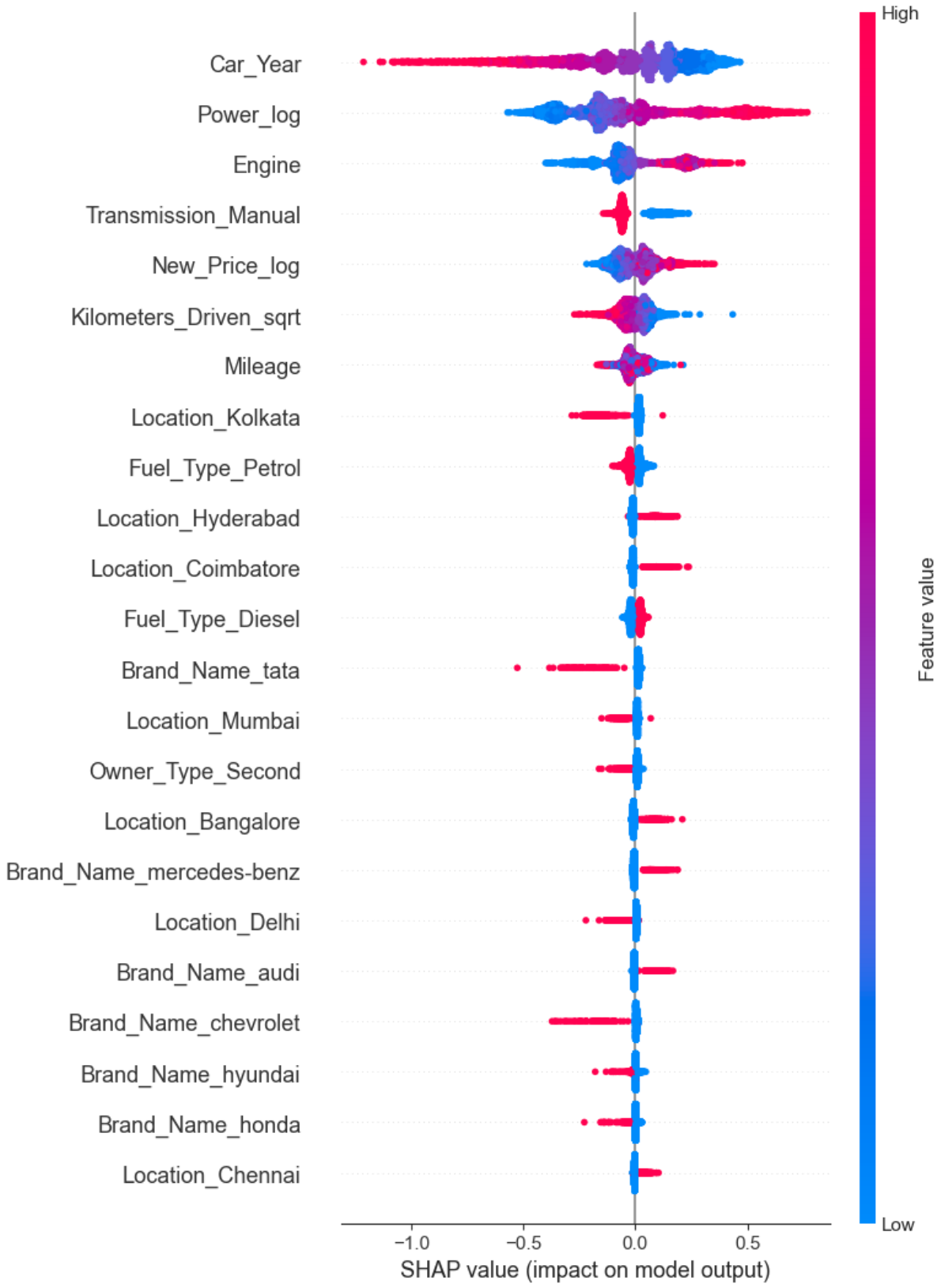
- **Car_Year:** Older cars have a significantly stronger negative impact on the final price compared to newer cars.
- **Power_log:** Higher engine power has a more positive impact on the final car price than lower engine power has on price decrease. This trend is also seen, though to a lesser extent, in the Engine feature. Such a relationship underscores the importance of performance in pricing.
- **New_Price_log:** More expensive cars tend to keep their value over the years more than cheaper cars.

2. **Hidden preferences:**

- **Mileage:** The impact of mileage [km/l] on the final price is relatively symmetrical, and does not provide clear advantage or disadvantage, based on this value alone. This shows that customers who decide on more fuel consumable vehicles value other car features more.
- **Brand_Name:** Cars from Tata and Chevrolet have a negative impact on prices, while European brands like Audi and Mercedes have a positive impact. This might show clients preferences or might be correlated with the current trends.

3. **Locations**

- Cities like Mumbai, Kolkata, and Delhi are associated with lower car prices, making them cheaper markets.
- In contrast, cities like Coimbatore, Hyderabad, and Bangalore tend to have higher car prices, indicating more expensive markets.


Car_Year
Power_log
Engine
Transmission_Manual
New_Price_log
Kilometers_Driven_sqrt
Mileage
Location_Kolkata
Fuel_Type_Petrol
Location_Hyderabad
Location_Coimbatore
Fuel_Type_Diesel
Brand_Name_tata
Location_Mumbai
Owner_Type_Second
Location_Bangalore
Brand_Name_mercedes-benz
Location_Delhi
Brand_Name_audi
Brand_Name_chevrolet
Brand_Name_hyundai
Brand_Name_honda
Location_Chennai
−1.0
−0.5
0.0
0.5
SHAP value (impact on model output)
High
Feature value
Low

Recommendations for Further analysis

Potential for Model Improvement

The current model for predicting used car prices performs well and serves as a reliable tool for market orientation. However, there are opportunities for further refinement and improvement. To enhance the model's accuracy and reliability, the following recommendations are proposed:

1. **Incorporate up-to-date data:** The model was developed using data up to 2019. Given the rapid changes in market trends and customer preferences, relying on outdated data may limit the model's effectiveness. Therefore, it is recommended to:
   - **Constant monitor and update data:** Regularly update the dataset with current market information to reflect the latest trends and consumer behavior. This will help ensure the model remains relevant and accurate in predicting current market conditions.
   - **Retrain model periodically**: Implement a process for periodically retraining the model with the latest data. This practice will help in maintaining the model's predictive accuracy and relevance over time.
2. **Capturing more variability in target prices:** The used car market is characterized by high price variability, and addressing this can further improve the model's performance. Several strategies can be employed:
   - **Incorporating Additional Features:** Including more predictive features could help capture more variability in the target prices.
     - **Interior equipment**, Features like leather seats, air conditioning, and electric windows are significant as they impact the car's perceived value.
     - **Additional electronic devices** Car audio systems, GPS, and HUD are increasingly important to buyers and can significantly influence pricing.
     - **Advanced Driver Assistance Systems (ADAS)**: Safety and convenience features like ABS, ESP, cruise control, and parking sensors are highly valued and should be included in the model.
   - **Ensuring High-Quality Data:** High-quality data is critical for improving model accuracy. Focus on acquiring data that is free from missing values, outliers, and inconsistencies. Proper data cleaning and preprocessing steps should be established as part of the data pipeline.
   - **Further Segmentation:** Segmenting cars into different price tiers can help address outliers and reduce variability within each segment, leading to a more robust and accurate model. This approach will also support the creation of more tailored marketing and sales strategies.
   - **Exploring Advanced Modeling Techniques:** Considering exploration of advanced modeling techniques such as Support Vector Regression (SVR), Neural Networks (Deep Learning), or Extreme Gradient Boosting (XGBoost, see Appendix 2) may be

beneficial and offer improved accuracy by capturing complex relationships in the data. While these techniques can enhance accuracy, they are computationally intensive. 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.

3. **Implementing cutting-edge technology:** Leveraging AI for dynamic pricing adjustments in real-time offers a significant opportunity to optimize pricing strategies and increase profitability in the used car market. By continuously adapting to market conditions and customer behavior, businesses can maintain a competitive edge and maximize revenue opportunities. However, careful consideration of data quality, resource requirements, customer perception, and regulatory compliance is essential for successful implementation.

Recommendations for Business and Implementation

Market Segmentation and Consumer Preferences:

- **Market Segmentation:** The analysis suggests clear segmentation in the market based on car age, power, and brand, which can guide pricing strategies and inventory decisions.
- **Consumer Preferences:** Preferences for newer, more powerful cars and premium brands like Audi and Mercedes are evident, which can inform marketing and sales strategies.
- **Implementation Recommendations:** Tailor marketing campaigns to target these segments effectively, offering personalized deals and promotions.

Focus on High-Impact Features:

- **Power and Car Year:** These features have the highest importance. Marketing and pricing strategies should emphasize these features if their values are assessed as valuable.
- **Engine Size and Initial Price:** Given their significant importance, these features can be considered when setting prices or during customer consultations to justify pricing.
- **Implementation Recommendations:** Use the most important features to create targeted marketing campaigns that highlight the top features. For example, a campaign could showcase cars with high power and recent manufacturing years, emphasizing their reliability and performance. For 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:

- **Regional Pricing Strategies:** Regional differences in car prices suggest the need for location-specific pricing strategies. Understanding these variations can help in optimizing

sales and inventory across different cities. However, this should be preceded by market research as the high impact on prices may not always mean profitable margins.

- **Brand-Specific Campaigns:** Focus on popular brands like Audi or Mercedes-Benz. Custom campaigns for these brands can attract brand-loyal customers. On the other hand, do not forget about customers with a lower economic status who most likely will look for cheaper vehicles.
- **Implementation Recommendations:** Use the model's insights to guide procurement. Depending on a target group prioritize acquiring cars by their appropriate features. Regularly update the inventory based on market trends and model predictions to ensure alignment with customer preferences in specific locations.

Price Adjustment Mechanism:

- **Data update:** Monitor and adjust prices based on real-time data, ensuring prices are aligned with market demand and supply conditions.
- **Implementation Recommendations:** Develop and integrate dynamic pricing tools that adjust prices based on the top features' current market trends (e.g., brand, location, fuel type, transmission). This ensures competitive pricing and maximizes sales.

Enhance Customer Experience:

- **Transparent Information:** Provide detailed information on the top features that influence the car's price.
- **Ownership History:** Cars with more than one previous owner (Second, Third, Fourth & Above) have a negative impact on prices. Emphasizing cars with fewer previous owners in marketing materials can help attract customers looking for better value.
- **Implementation Recommendations:** Educate customers about the importance of key features in determining a car's value. Provide resources and tools that help customers understand how features like power, year, and engine affect pricing. Educate customers on the value of pre-owned cars, emphasizing benefits like cost savings and environmental impact. Highlighting the rigorous inspection and certification processes can build trust and encourage purchases.

Future Trends Monitoring:

- Stay ahead of market trends by monitoring changes in customer preferences for electric vehicles and other emerging segments concerning technological innovations. Gradually increase the focus on these areas to stay relevant and competitive.

Bibliography:

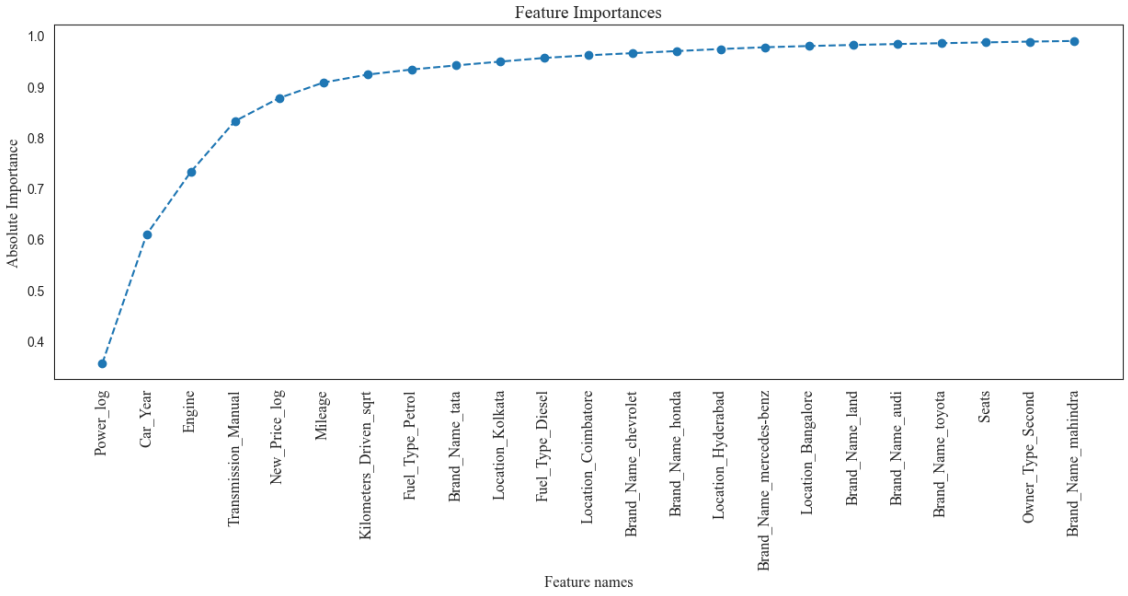
1. Autocar India
2. HT Auto
3. Mordor Intelligence
4. LeadSquared
5. 6Wresearch
6. CarWale
7. CarDekho
8. ZigWheels

Appendix

Appendix 1: Feature Importance Analysis (GBR)

Feature Importance Analysis

Additional analysis was conducted to determine the importance of features for predicting the target variable. To enhance clarity, the feature importance chart was limited to the top 23 features out of 51. These 23 features collectively account for 99% of the overall importance, indicating that the model could potentially be simplified by approximately half without a significant loss in performance.



Comparison of Full and Trimmed Datasets

The performance metrics for the full dataset and the trimmed dataset (limited to the top 23 features) are as follows:

Metric	Full dataset	Trimmed dataset
Root Mean Squared Error (RMSE)	1.99	2.01
Mean Absolute Error (MAE)	1.05	1.06
R-squared (R^2)	0.929	0.928
Mean Absolute Percentage Error (MAPE)	13.31%	13.55%

Decision on Feature Selection:

Despite the ability to reduce the number of features significantly, I made the decision not to limit the range of variables. The primary reasons for this decision are:

- **Negligible Reduction in Overfitting:** The reduction in overfitting achieved by trimming the features was minimal.

	Metric	Train	Test
Full dataset	R-squared (R^2)	0.97	0.93
Trimmed dataset	R-squared (R^2)	0.96	0.93

- **Impact on Error Metrics:** There was a slight increase in error metrics when using the trimmed dataset. Although the differences are small, the preference was to maintain the broader feature set to avoid any potential negative impact on model accuracy.

	Metric	Test
Full dataset	Mean Absolute Percentage Error (MAPE)	13.32
Trimmed dataset	Mean Absolute Percentage Error (MAPE)	13.55

Appendix 2: XGBoost vs. GBR

XGBoost model was also evaluated and showed an improvement in accuracy compared to the GradientBoostingRegressor.

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 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.

Appendix 3: GBR & XGBoost parameters

Gradient Boosting Regressor

- Model performance

Data	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 parameter	Value
max_depth	5
max_features	0.5
min_samples_leaf	3
n_estimators	120

XGBoost

- Model performance

Data	Train	Test
Root Mean Squared Error (RMSE)	0.77	1.91
Mean Absolute Error (MAE)	0.47	0.96
Mean Absolute Percentage Error (MAPE)	6.45%	12.1%
R-squared (R^2)	0.99	0.94

- Tuned parameters

Model parameter	Value
max_depth	7
colsample_bytree	0.5
n_estimators	120
reg_alpha	01
reg_lambda	0.1

Appendix 4: Gradient Boosting Regressor mean absolute SHAP values (Price_log)