Used Car Price Prediction and Market Analysis for the Indian Market

Business Problem

For online marketplaces such as Cars24, accurate pricing of used cars is essential to remain profitable, turn over inventory quickly, and build trust with buyers and sellers. Pricing cars above the market price, referred to as the as-is price, could lead to vehicles sitting in inventory for long periods, while pricing cars below the market price may result in revenue losses. This project addresses this business issue by creating a data-driven pricing model to estimate a used car's market price (known as the as-is price) based on specific attributes.

Background/History

Over the last ten years, India has witnessed a substantial increase in the used car market due to aspirational consumption, increasing prices of new cars, and the emergence of certified pre-owned organized players like Cars24. Online marketplaces bring transparency and convenience to an otherwise unregulated and fragmented market. It is essential to understand the vehicle depreciation and value drivers to be competitive in this space.

Datasets

The primary dataset used in this project is 'cars24data.csv'. The dataset claims to provide scraped data from Cars24.com, an online marketplace for used cars in India. 'cars24data.csv' contains comprehensive information on pre-owned car listings. It includes 1,445 records of used car listings from the Cars24 platform.

Data Dictionary:

- **Model Name:** The make and model of the car (e.g., "2017 Maruti Swift VXI").
- **Price:** The listing price of the car in Indian Rupees (INR). (Target Variable)
- **Manufacturing Year:** The year the car was manufactured.
- **Engine Capacity:** The engine capacity in cubic centimeters (CC).

- **Spare key:** Whether a spare key is available ('Yes' or 'No').
- **Transmission:** The transmission type ('Manual' or 'Automatic').
- **KM driven:** The total kilometers driven.
- **Ownership:** The number of previous owners (1, 2, or 3).
- **Fuel Type:** The car's fuel type ('Petrol', 'Diesel', 'CNG').
- **Imperfections:** The number of documented imperfections.
- **Repainted Parts:** Repainted parts count.

Kaggle Link: https://www.kaggle.com/datasets/amanrajput16/used-car-price-data-from-cars24

Data Preparation:

1. A new feature, `Car_Age`, was created by subtracting the `Manufacturing_year` from the current year (2025).
2. Categorical – We transformed text-based features (`Spare key`, `Transmission`, and `Fuel type`) into a numerical representation using one-hot encoding.
3. Due to its high cardinality, we removed the `Model Name` feature from this first analysis.
4. We divided the dataset into a training set (80%) and a test set (20%) to train the model and evaluate the results without bias.

Methods

We used a supervised machine learning method to forecast the price of secondhand cars.

1. **Baseline Model (Linear Regression):** We first created a baseline level of performance using a Linear Regression model. The model tries to find a simple linear relationship between the car's features and the price.
2. **More Advanced Model (Random Forest):** We trained a more complicated ensemble model, a Random Forest Regressor. This model comprises several individual decision trees and aggregates their predictions to yield a more accurate and robust predicted price for a given car.

3. **Evaluation:** We evaluated the models on the previously unseen test data using two metrics:
    a. **Mean Absolute Error (MAE):** This is the average Absolute Error one gets between the modified price predicted by the model and the car's actual price.
    b. **R-squared (R^2):** The car's features predict a proportion of the variance in the price.

Analysis

The Random Forest model performed significantly better than the Linear Regression baseline.

Random Forest Performance:

- **R-squared (R^2):** 0.83 (This model can explain 83% of the variance in car prices).
- **Mean Absolute Error (MAE):** ₹54,435** (The model typically mispredicts the price by around ₹54,435).

The main takeaway from our analysis was identifying the most influential factors in determining the price of used cars. By interpreting the Random Forest model indirectly, we can rank the feature's importance:

1. **Engine Capacity (63% importance):** This was the highest-ranking feature. Cars with higher engine capacities have substantially higher values.

2. **Car Age (23% importance):** This was the second most important feature, as expected, the older the car is, the cheaper its value is.

3. **Kilometers Driven (5.5% importance):** The value decreases when a car drives more kilometers.

4. **Imperfections & Repainted sections:** Although the car's physical condition is not more important than the three previously listed factors, it still has an essential, measurable effect on the price.

Conclusion

We conducted the analysis using several assumptions:

- The data scraped from Cars24 is an accurate and unbiased representation of the listings for the used car market.

- The features in the dataset will be enough to predict the price of the car.
- The relationships between the features and price observed in the historical data will hold for future price predictions.

Assumptions

This project successfully created a machine learning model that can model and forecast the price of used cars with acceptable accuracy (R-squared = 0.83). The intrinsic features of a vehicle, particularly its engine size and age, primarily determine its resale price. The Random Forest model is a valuable method to aid in interpreting the complex relationships existing in the used car market. The project has produced findings useful for understanding pricing in online car markets.

Limitations

The research features the following constraints:

- **Feature Scope:** The model excludes the vehicle's 'Model Name' or brand. The vehicle's brand influences pricing, and while including this feature would have enhanced model accuracy, it requires more complex feature engineering approaches (e.g., target encoding, embedding).
- **Sample Size:** The dataset consists of 1,445 observations, which is quite limited for training a complex machine learning model. A larger dataset would further enrich the model's strength and accuracy.
- **Market Factors:** The model does not factor in external market forces such as geographical location (city/state), seasonality, real-time supply and demand, or macroeconomic conditions, which could impact the pricing of cars.
- **Static Model:** The model relies on a historical dataset and remains static; however, because the used car market is dynamic, we must periodically retrain the model to maintain its accuracy.

Challenges

Several issues arose throughout the course of the project:

- **Feature Engineering concerning Model Name:** The 'Model Name' column contained valuable information about the make and model of the car. The challenge, however, was high cardinality. We could not leverage the `Model Name` without significant feature engineering. In this case, we could use sophisticated techniques like target encoding and brand or model-level features.
- **Data Sparsity:** Although the dataset was clean and tidy, the relatively small number of observations (1,445) reduced our statistical power and limited the complexity of the models we could build without overfitting.
- **Model Interpretability versus Accuracy:** The Random Forest model was limited in interpretability compared to the simple Linear Regression model, which was both desirable and appropriate here. We should note that explaining why a Random Forest model predicts a specific instance can be complicated.

Future Uses/Additional Applications

The work carried out in this project could be expanded into several exciting directions:

- **Dynamic Pricing API:** We could operationalize the model as a REST API. We would enable other programs (including the Cars24 website or in-house pricing tools) to obtain real-time price estimates for new listings.
- **Customer-Facing Price Estimator:** A simplified model version could be operationalized as a web application to enable potential sellers to receive an immediate estimate of their car value, a sure way to increase user engagement and leads.
- **Residual Value Forecasting:** The model could also be modified to forecast the future residual value of cars in the marketplace, which could be helpful for leasing companies, fleet managers, and those providing financial products.
- **Market Trend Analysis:** Retraining the model regularly, monitoring the feature importance for metrics within the model, and finding trends in the used vehicle market are feasible.

Recommendations

From the analysis, we have crafted the following recommendations for a company like Cars24:

- **Implementation of Data-Driven Pricing:** Use the Random Forest Model as an auxiliary option to support pricing specialists. We could create a more standardized pricing process, reduce the manual effort of making the pricing quote, and help reduce potential pricing errors due to subjective pricing judgments.
- **Displaying Relevant Features:** In both your marketing and your listing pages, give prominence to the pricing drivers identified by this analysis: engine size, lower age, and fewer kilometers driven. We could assist in communicating and justifying the vehicle's price to the buyer.
- **Opportunistic Inventory Procurement:** We can also use the model to identify markets with the potential to procure undervalued cars. Conversely, the model can identify vehicles that may be comparatively overpriced and present challenges in selling the listing.
- **Invest time into the 'Model Name' as a Feature Engineering exercise:** Investing the appropriate resources to properly engineer the `Model Name` feature would be worthwhile. Having the brand and model name captured is the single quantitative feature engineering artifact that will significantly impact its performance, ultimately leading to more precise recommendations and market signaling.

Implementation Plan

To incorporate this price prediction model into a business process at a firm such as Cars24, we suggest a stepwise approach to implementation:

1. **Phase 1: Offline Validation & Refinement (1-2 Months):**
    - Back-testing: Rigorously run the model back on data it has not seen, including data from other times and other parts of the country (if we have those).

- Specialist Review: Show the predictions to a pricing specialist and hear their comments. They may be able to help find any systematic errors in the model or the logical flaws.
- Feature Engineering: After hearing from the specialists and running the model back, plan for feature engineering, especially for the 'Model Name' and a location (or not).

2. **Phase 2: Internal Tool Development (2-3 Months):**
    - API Development: Once we have the trained model, package it as a secure, internal-facing REST API.
    - Dashboard Development: Build a simple web application, or dashboard, that pricing specialists can use to put in the details of the car they are looking at and receive the model price, and key drivers of the prediction. We keep a "human in the loop" and use it as a better decision-making approach.
3. **Phase 3: A/B Testing (3-6 months):**
    - Controlled Experimental Design: Use model recommendation pricing for a subset of new listings and maintain pricing like normal for the control group.
    - Metrics for Impact: Track the metrics for the two groups (e.g., time-to-sell, final sale price compared to listing price - is it within an acceptable range, and customer satisfaction levels, etc.). We will provide quantitative evidence of the model's impact on the business.
4. **Phase 4: Full Integration & Monitoring (Ongoing):**
    - Full Integration into Workflow: If the A/B test in phase 3 went well, use the recommendation to price new listings and fully integrate the model pricing directly into the existing pricing and inventory management system(s).
    - Monitoring Use & Impact: Set up use and monitoring systems, and make any adjustments based on feedback.

- Retraining Schedule: Create a schedule for retraining the model on new data to adapt to the market changes.

Ethical Assessment

Before the implementation of this project, a comprehensive ethical evaluation is necessary. The fundamental ethical concerns highlighted in this assessment are algorithmic bias, transparency, and data privacy.

- Algorithmic Bias: Our investigation found that features such as `Engine capacity` and `Car_Age` are strong price predictors. While these are reasonable features, ensuring that the model does not unintentionally create or exacerbate unjust bias is essential. For instance, if the training data underrepresented certain car types, the model may unjustly punish certain car types.
    - Mitigation: We must consistently audit the model for bias across car types, brands, and other sensitive features. Increasing the size and ensuring the representativeness of the data is a critical means of mitigating this risk.
- Transparency and Explainability: The Random Forest model is more of a "black box" than a simple linear model. The lack of transparency can be problematic for both the internal user (pricing specialists) and the external user (customer), who are likely interested in the rationale of the price assigned.
    - Mitigation: We utilized feature importance to provide a general rationale; however, for an individual prediction, methods such as SHAP (Shapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) should be considered for implementation into the internal tool as a means of explaining why we assigned a particular price to a specific car.
- Data Privacy: The information in the dataset was obtained from a public dataset on Kaggle and does not appear to contain any Personally Identifiable Information (PII).

- Mitigation: We will perform a comprehensive data privacy review before introducing any organizational proprietary data into the model, ensuring that all customer and transaction data has been anonymized and processed in a manner consistent with privacy laws.

10 Questions an Audience Would Ask

1. How accurate is your price prediction model in real-world scenarios?
2. You mentioned the model doesn't include the car's brand or model. Wouldn't that be one of the most important factors?
3. How do you plan to keep the model current with changing market trends?
4. Could this model be used to predict the price of other types of vehicles, like motorcycles or trucks?
5. What was the most surprising finding from your analysis?
6. How do you account for regional differences in car prices across India?
7. You mentioned ethical concerns like algorithmic bias. What specific steps will you take to ensure your model is fair?
8. How long would it take for a company like Cars24 to implement your model?
9. What are the limitations of using a dataset from a single platform like Cars24?
10. How could a private seller use this model to their advantage?

References

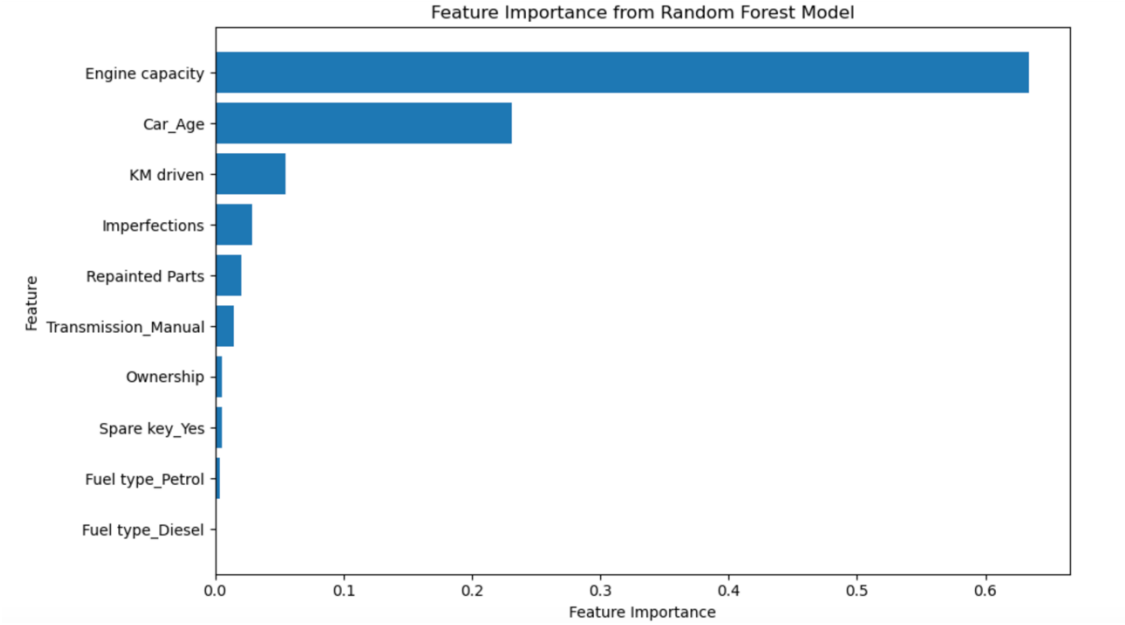
Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

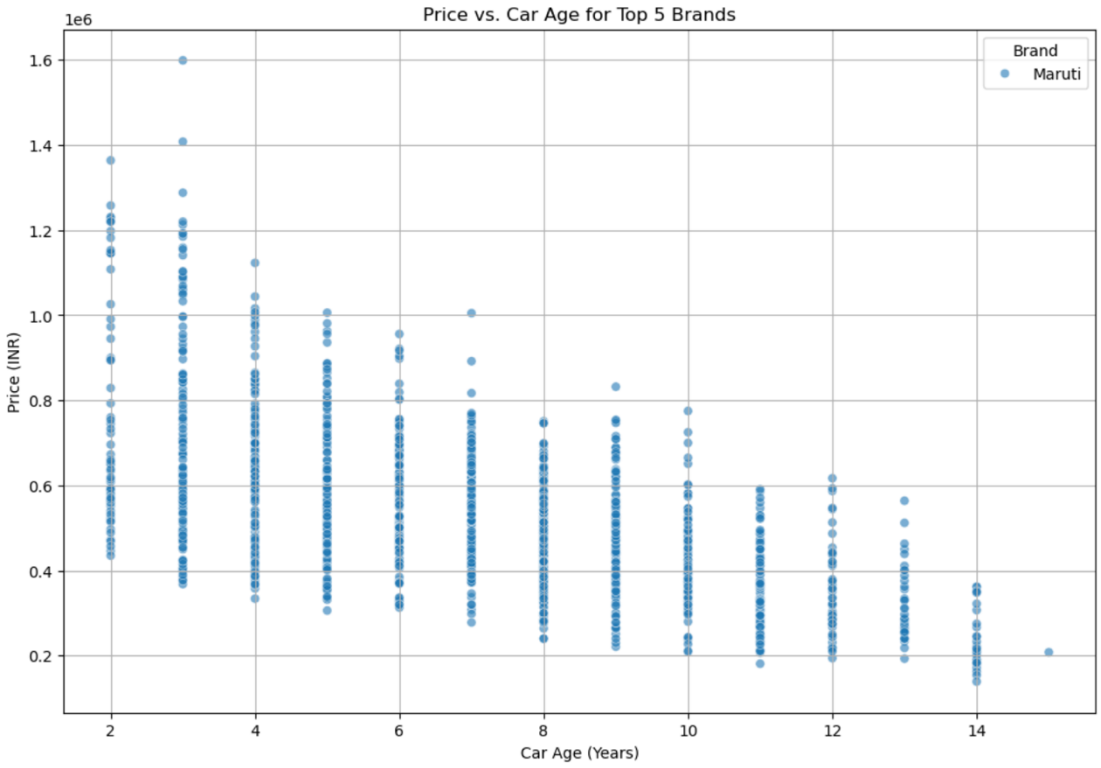
GeeksforGeeks. (2025). *Analyzing Selling Price of used Cars using Python*. Retrieved from https://www.geeksforgeeks.org/python/analyzing-selling-price-of-used-cars-using-python/

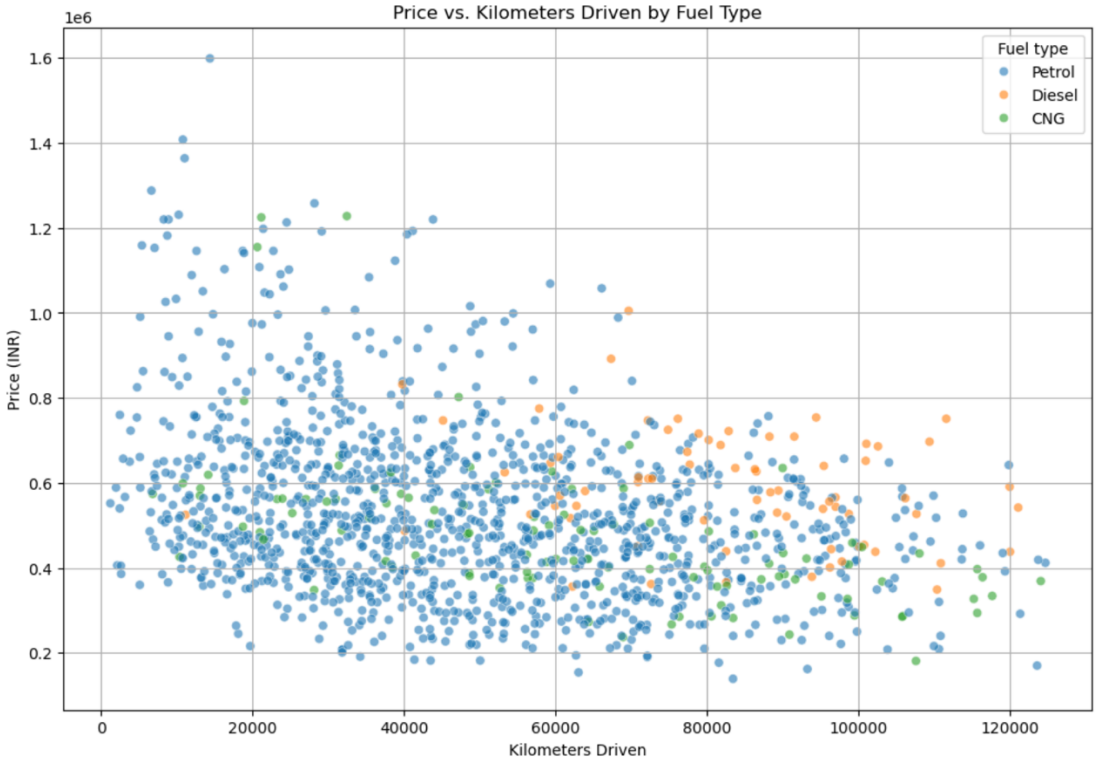
IMARC Group. (2023). *India Used Car Market: Industry Trends, Share, Size, Growth, Opportunity and Forecast 2025-2033*. Retrieved from https://www.imarcgroup.com/india-used-car-market

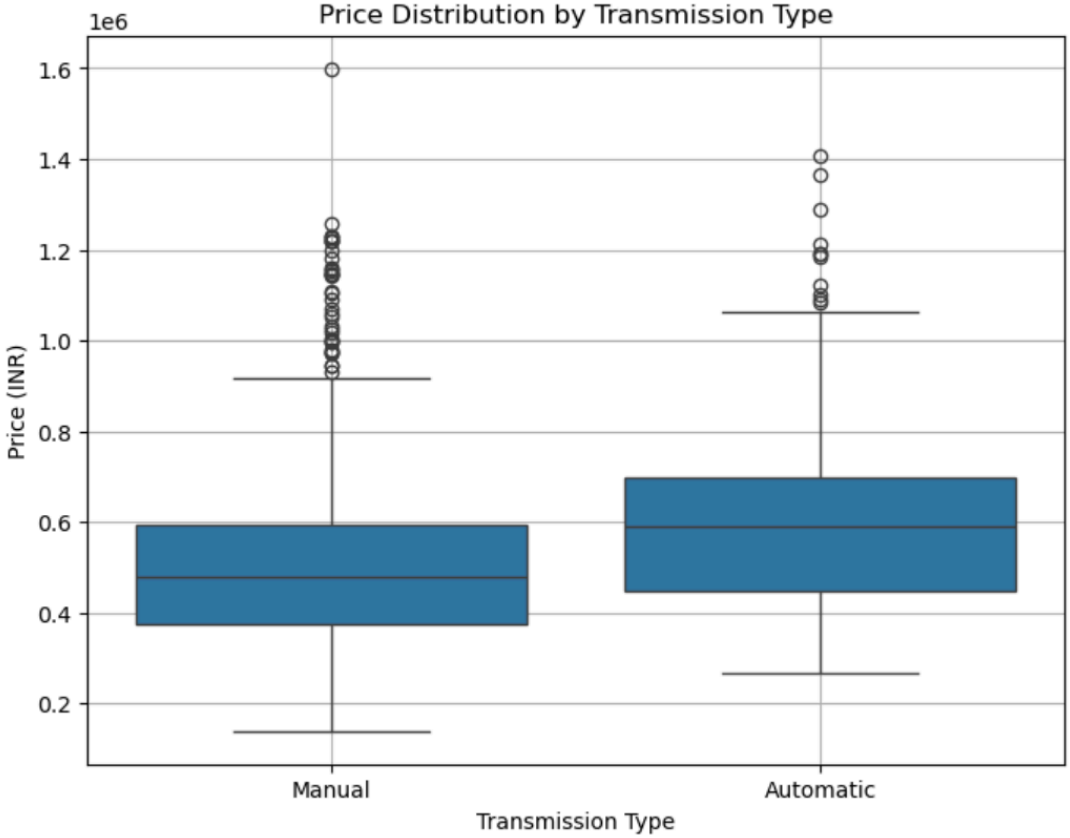
Appendix

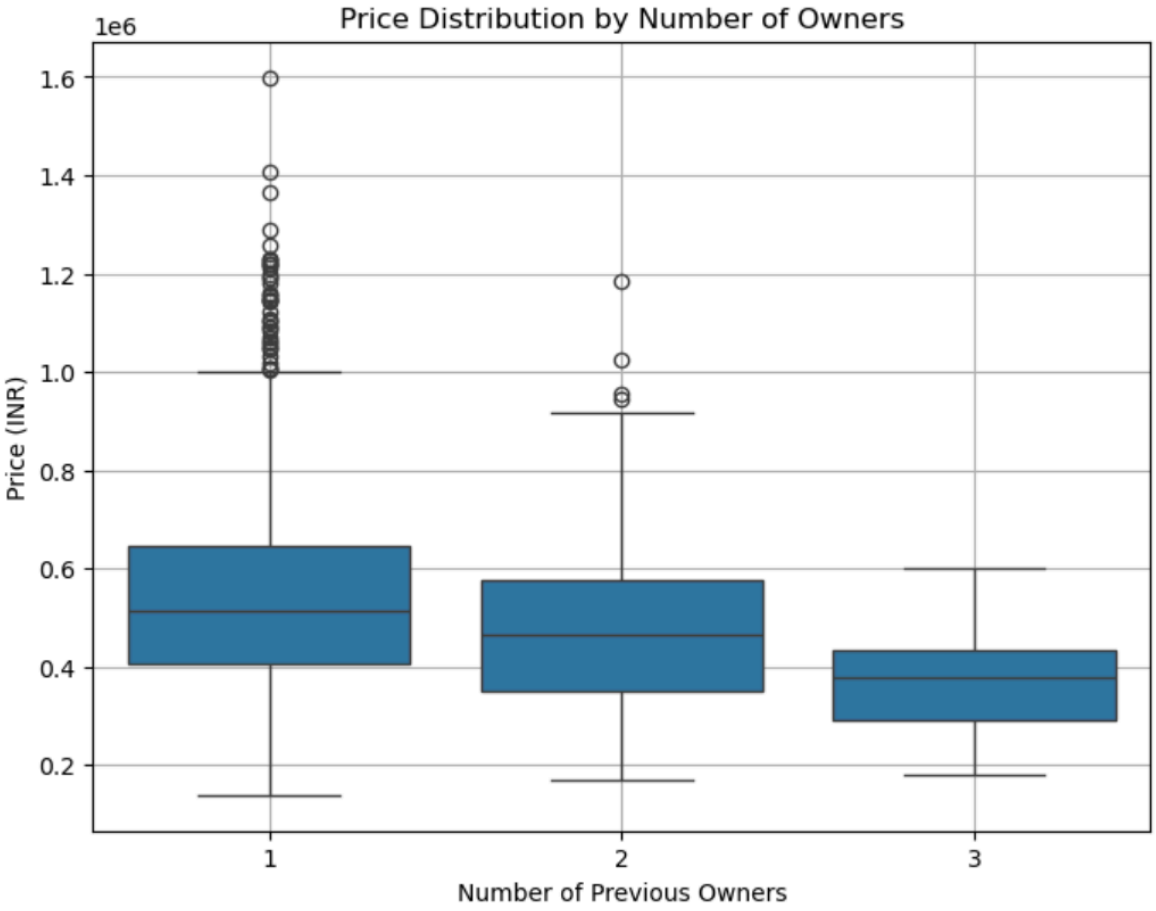
The plots below explain the model-building process and show the results of the models we built.






Price vs. Kilometers Driven by Fuel Type
1e6
Fuel type
Petrol
Diesel
CNG
Price (INR)
Kilometers Driven
0
20000
40000
60000
80000
100000
120000
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6


Price Distribution by Transmission Type
1e6
Price (INR)
Manual
Automatic
Transmission Type
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6


Price Distribution by Number of Owners
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Price (INR)
1
2
3
Number of Previous Owners