

CAR PRICE INFLUENCE ANALYSIS: FEATURES

Ganesh Balaji R
Department of Computing
(Student – MSc Data Science)
Coimbatore Institute of Technology
(Affiliated to Anna University)
Coimbatore, India
71762132015@cit.edu.in

Ms Deivarani S
Department of Computing
(Assistant Professor)
Coimbatore Institute of Technology
(Affiliated to Anna University)
Coimbatore, India
deivarani@cit.edu.in

Kavin Chant P
Department of Computing
(Student – MSc Data Science)
Coimbatore Institute of Technology
(Affiliated to Anna University)
Coimbatore, India
71762132018@cit.edu.in

Dr Rajarajeshwari K
Department of Computing
(Assistant Professor)
Coimbatore Institute of Technology
(Affiliated to Anna University)
Coimbatore, India
rajarajeshwari@cit.edu.in

Abstract: In the ever-evolving Indian automobile market, understanding the driving factors behind car prices is crucial for manufacturers to refine product offerings and tailor marketing strategies effectively. This machine learning project aims to uncover the most influential features impacting car prices, assisting manufacturers in making informed decisions. The project leverages a comprehensive dataset containing over 1200 car models/variants, encompassing a wide range of features such as engine specifications, fuel efficiency, brand, model, and variant details. A comprehensive analysis was embarked upon to unveil the features that significantly influence car prices. The dataset was pre-processed, handling missing values and converting categorical variables into numerical representations for further analysis. The core of the project involves the development of predictive models for car prices, employing both Decision Trees and Random Forests regression algorithms. These models excel in capturing complex relationships between features and prices. Following the training process, the feature importance scores were extracted from each model. Visualizations, such as bar plots and heatmaps, aid in interpreting the results, highlighting the features with the highest importance scores. By systematically analyzing these insights, manufacturers gain a comprehensive understanding of the features that wield the greatest influence on car prices. Armed with this knowledge, they can prioritize improvements to specific attributes, adjust marketing campaigns to spotlight influential features, and better cater to evolving customer preferences. In conclusion, unveiling the feature importance enable the manufacturers to enhance their competitive edge by offering products that resonate with customers and maximizing their market presence. Through a combination of advanced machine learning techniques and insightful visualizations, this project contributes to a more informed and adaptive automotive sector.

Keywords: Indian automobile market, Machine learning, Feature importance, Predictive models, Marketing strategies

I. INTRODUCTION

A. Background

The automobile industry in India is known for its variety and constant changes making it an ever changing market for manufacturers. For car companies to succeed they need to

have an understanding of the factors that affect car prices. This paper explores the complexities of car pricing in India using machine learning techniques to uncover the features that have a significant influence on car prices in this market.

B. Objective

The main goal of this research is to discover the factors that have an impact, on the prices of cars. This will help car manufacturers make informed decisions. By analyzing a dataset containing than 1200 car models and variants along with, over 140 different features this study aims to provide valuable insights that can give manufacturers a competitive advantage and align better with customer preferences.

II. METHODOLOGY

A. Dataset Description

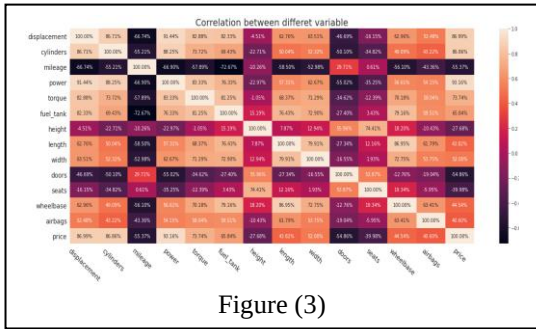
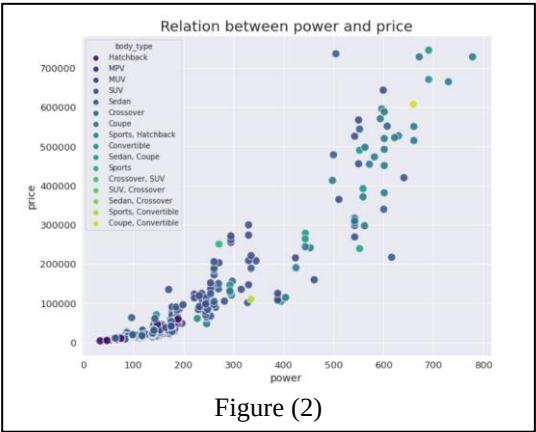
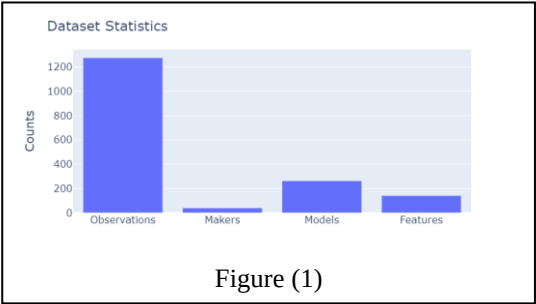
The dataset utilized in this study offers a comprehensive overview of the Indian automobile market, encompassing 1200 car models and variants across diverse brands and specifications. With 140 features detailing various aspects of the automobiles, the dataset provides insights into engine specifications, fuel efficiency, brand reputation, model specifics, and variant details. Features such as displacement, cylinders, drivetrain, emission norms, fuel type, kerb weight, gear count, ground clearance, fuel tank capacity, power, torque, and seating capacity, among others, are included. This extensive array of features enables a thorough analysis of the factors influencing car prices, thereby facilitating the development of predictive models and the extraction of feature importance scores for informed decision-making by manufacturers.

B. Exploratory Data Analysis

The Exploratory data analysis (EDA) is performed to understand the underlying patterns, relationships, anomalies, and trends in a dataset. EDA is essential because it informs the subsequent modeling process and reveals important characteristics that influence the price of the vehicle. The first step of EDA provided a concise overview of the structure of the data set. This is well represented in Figure (1), which

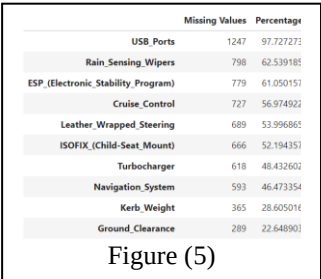
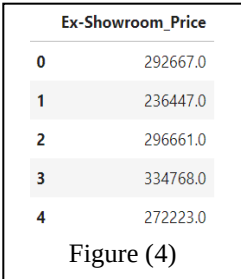
visualizes the overall observations, unique car manufacturers, number of models, and various features available for analysis. This basic overview paved the way for a deeper exploration of body type diversity and its distribution within the dataset, revealing common market trends and preferences. Additionally, the complex relationship between vehicle performance and price was investigated. Figure (2) shows a scatterplot showing the subtle interaction between performance and price and how these variables affect a car's market value.

This study was critical to understanding the dynamics of feature interaction and its impact on vehicle prices. Finally, a comprehensive heatmap of feature correlations was created, as shown in Figure (3). This heatmap becomes a valuable tool for visualizing the interdependencies between different vehicle features and identifying potential multicollinearity, thereby informing the feature selection process for subsequent modeling. The insights and visualizations gained during this phase served as the basis for subsequent data preprocessing, dimensionality reduction, and model development in this analysis.



C. Data Preprocessing

During data preprocessing, several essential transformations were applied to refine the dataset for model building. The 'Ex-Showroom_Price' column underwent conversion to numerical format, as illustrated in Figure (4).

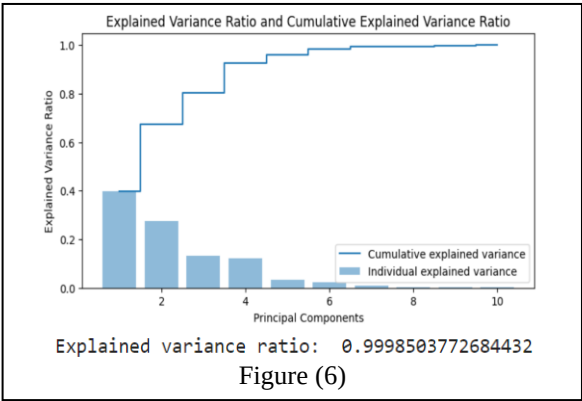


After a thorough examination of the dataset, the columns with missing values were identified and the columns with more than 70% missing data are dropped as depicted in Figure (5). For the remaining columns, missing values were addressed using median imputation for numerical columns and mode imputation for categorical ones. Subsequently, LabelEncoder() is applied to convert categorical columns into numerical representations, resulting in a uniformly formatted dataset. These preprocessing steps laid a robust foundation for further analysis and model development, aiming to uncover the significant features impacting car prices.

D. Dimensionality Reduction

In the context of reducing the dimensionality of the dataset while preserving its essential characteristics, Principal Component Analysis (PCA) is used. The data is first split into features X and target variable y, with column "Ex-Showroom_Price" serving as the target. Next, we applied PCA to the features with the number of principal components set to 10. This parameter is customizable and flexible based on the specific needs of your analysis.

When applying PCA, each principal component is formed as a linear combination of the original features, with the first principal component capturing the maximum variance in the dataset. Each subsequent component that was orthogonal to the previous component accounted for the largest possible residual variance.



The success of dimensionality reduction was quantified using the explained variance ratio, which represents the proportion of total variance in the dataset that is retained after applying PCA. This ratio is very important because it allows you to understand how much information is retained and lost during the conversion. The cumulative explained variance ratio shown in Figure (6) served as an indicator of the effectiveness of PCA in preserving important information of the dataset.

This detailed visualization of explained variance ratios provided a clear understanding of the trade-off between dimensionality reduction and information preservation. This allows you to make informed decisions about the number of principal components to use in subsequent analysis, ensuring that models based on transformed data are efficient and reflect the characteristics of the original dataset. Therefore, the application of PCA played an important role in improving the computational efficiency of the modeling process while maintaining data integrity and meaning.

E. Feature Scaling

In the pursuit of analysing the influence of various features on car prices, ensuring that different types of data are on a comparable scale is essential. This is known as Feature Scaling, a critical step in preparing the dataset for analysis.

Initially, the dataset is separated into features (X) and the target variable (y), which represents the car price to be analysed. Subsequently, the data is divided into training and testing sets, with 80% allocated for analysis and the remaining 20% reserved for validation. Upon dividing the data, Standard Scaling is employed to adjust the features, ensuring each has a mean value of 0 and a standard deviation of 1. This scaling is applied to both the training and testing sets, ensuring consistency and comparability across all features.

This preprocessing step is pivotal, laying a solid foundation for a thorough and reliable analysis of how various features influence car prices, and ensuring that the influence of each feature is assessed on a balanced and comparable scale.

F. Hyper Parameter Tuning

Hyperparameter tuning is employed to optimize the Decision Tree and Random Forest regression models in the analysis of car price influences. Unlike other parameters that the model learns from the data during training, hyperparameters aren't automatically optimized. Therefore, the right combination of hyperparameters can greatly influence a model's performance, making hyperparameter tuning a crucial step in the machine learning pipeline.

i. Decision Tree Hyperparameter Tuning

The Hyperparameter tuning for a Decision Tree regressor can be performed using Grid Search, a method that exhaustively tries all provided hyperparameter combinations to find the best model. Within a defined hyperparameter grid, parameters such as maximum depth, minimum

samples split, and others are varied. Utilizing GridSearchCV from scikit-learn, the model is trained using 5-fold cross-validation on PCA-transformed training data. The process identifies the optimal hyperparameters for the dataset, which are then stored in `best\_params\_dt` as in Figure (8). These optimal settings aim to enhance the model's predictive performance.

```
{
    'max_depth': 10,
    'min_samples_split': 2,
    'min_samples_leaf': 1,
    'max_features': 'auto'
}
```

Figure (7)

```
{
    'n_estimators': 100,
    'max_depth': None,
    'min_samples_split': 2,
    'min_samples_leaf': 1,
    'max_features': 'auto'
}
```

Figure (8)

ii. Random Forest Hyperparameter Tuning

The code demonstrates hyperparameter tuning for a RandomForestRegressor using RandomizedSearchCV. Unlike GridSearchCV, which tests all hyperparameter combinations, RandomizedSearchCV samples a set number, often providing faster results. Here a simplified hyperparameter grid is used with limited choices, making the random search process almost deterministic. After fitting the model to the data, the best parameters are extracted as in the Figure (8)

G. Model Development

Once the dataset underwent preliminary preprocessing, the spotlight turned to model development. Two primary algorithms, Decision Trees and Random Forests, were selected to predict car prices.

To ensure that these models were tailored for optimal performance, a search for the best hyperparameters was initiated. The `hyperopt` library was utilized for this endeavour, which aims to identify hyperparameters that minimize the negative mean squared error. The search space for hyperparameters was defined, encompassing various potential values for parameters such as `max\_depth`, `min\_samples\_split`, and `max\_features`.

Considering the high-dimensional nature of datasets, dimensionality reduction techniques can prove invaluable. For this study, Principal Component Analysis (PCA) was applied to the standardized training data, retaining the

majority of the variance while reducing the number of features.

Once the optimal hyperparameters were identified, Decision Tree and Random Forest models were constructed using these best-found parameters. They were then trained using the PCA-transformed training dataset.

H. Model Evaluation

i. Decision Tree Model Evaluation

After training, the decision tree model is evaluated using the root mean square error (RMSE). RMSE is a metric that measures the difference between observed and predicted values. For the test set, the RMSE value was approximately [76561.01] which indicates that the average prediction error for car price is of this magnitude. Cross-validation is used to evaluate the model’s robustness and overall applicability. For the Decision Tree, the cross-validation scores for the 5 folds of the model were approximately [201.33]. Moreover, an R-squared value of 89% was achieved, indicating that the Decision Tree model explained this percentage of the variance in the car prices.

ii. Random Forest Model Evaluation

The model was first tested on a test set, resulting in an RMSE of about [61561.01]. This indicates that the model’s predictions deviate approximately this much from the real car prices. Cross-validation was used to validate the random forest model. Cross validation scores for the random forest model were approximately [177.59] across 5 folds. Cross-validation further verified the random forest model’s reliability. The R-squared value for the Random Forest model was approximately 91%, signifying that this percentage of the variance in the car prices was explained by the model.

III. RESULTS

A. Feature Importance

i. Feature Importance in the Decision Tree Model

Feature importance provides insight into which features are most influential in making predictions. For the Decision Tree model, the significance of each feature was gauged, and the results were visualized in a bar plot. The y-axis of this plot lists the features while the x-axis denotes their respective importance scores. Features with higher scores have a greater impact on the model's predictions.

Upon examination of the Decision Tree's feature importance plot (Refer to Figure(9)), it becomes evident that Torque, Power, and Displacement emerge as the top three influential features. These

features play a pivotal role in determining the ex-showroom price of cars in the dataset.

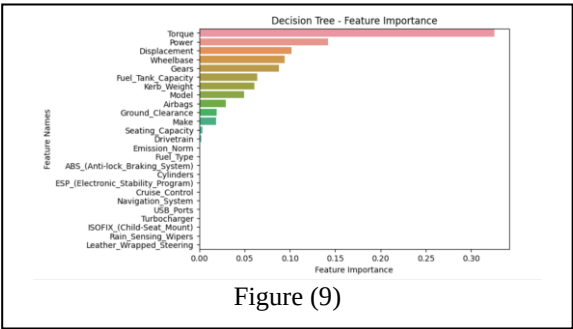


Figure (9)

ii. Feature Importance in the Random Forest

Similar to the Decision Tree, the Random Forest model's feature importance was assessed. Given the ensemble nature of Random Forests, this model calculates feature importance by aggregating the importance scores across all the trees in the forest.

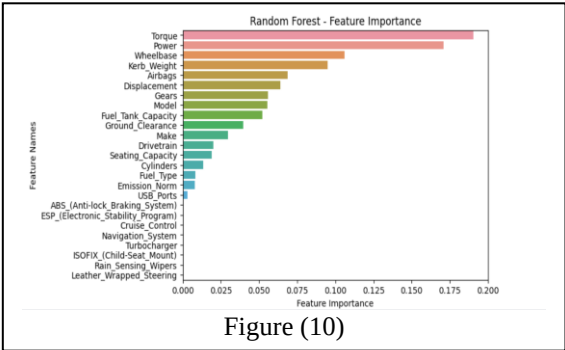


Figure (10)

From the Random Forest's feature importance plot (Refer to Figure (10)) : Torque, Power, and Wheelbase stand out as the most influential features. Their prominence signifies that they hold substantial predictive power in estimating the car prices within the dataset.

B. Model Comparison

To pinpoint which features are most pivotal in predicting car prices, the top 10 features, based on their importance scores were extracted for both the Decision Tree and Random Forest models. For the Decision Tree model, the leading features, in terms of their significance, are as follows:

Top 10 Important Features for Decision Tree:		
	Feature	Decision_Tree_Importance
12	Torque	0.325725
11	Power	0.142240
14	Wheelbase	0.094311
7	Kerb_Weight	0.060608
16	Airbags	0.029592
2	Displacement	0.102012
8	Gears	0.087945
1	Model	0.049510
10	Fuel_Tank_Capacity	0.063960
9	Ground_Clearance	0.019521

Similarly, for the Random Forest model, the most influential features are:

Top 10 Important Features for Random Forest:		
	Feature	Random_Forest_Importance
12	Torque	0.190499
11	Power	0.170893
14	Wheelbase	0.106114
7	Kerb_Weight	0.095059
16	Airbags	0.068903
2	Displacement	0.064095
8	Gears	0.055724
1	Model	0.055657
10	Fuel_Tank_Capacity	0.052148
9	Ground_Clearance	0.039618

Also a comparative analysis was conducted to identify features deemed important by both models. The intersection of the top features from the Decision Tree and Random Forest models revealed a set of common influential features. These features are: 'Wheelbase', 'Displacement', 'Model', 'Ground_Clearance', 'Airbags', 'Fuel_Tank_Capacity', 'Torque', 'Power', 'Gears', 'Kerb_Weight'. This overlap highlights the fact that both models agree on the importance of these specific characteristics in the forecasting process.

For a clearer understanding, the importance scores for the top features for each model were plotted side-by-side. The bar plots provide a comparative view of the importance of each feature within each model. The Decision Tree Feature Importance plot shows the importance of each top feature within the Decision Tree model, while the Random Forest Feature Importance plot, in salmon, shows the top features as seen in the Random Forest model. (Refer Figure (11))

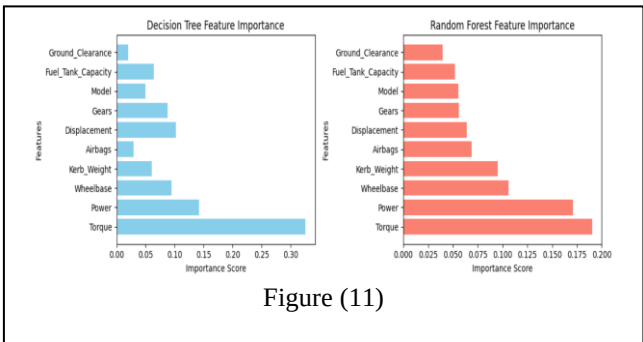


Figure (11)

IV. IMPLICATIONS

- Stakeholders:** These results can serve as a guide for stakeholders in the automobile industry, especially manufacturers and dealers. Understanding the key features that influence car prices allows them to make informed decisions regarding production, feature inclusion, and marketing strategies.
- For Consumers:** From a consumer perspective, recognizing these influential features can aid in making informed purchasing decisions. It provides clarity on what factors they are potentially paying a premium for and allows them to prioritize features based on their personal preferences and budget constraints.
- For Future Research:** This study sets a foundation for further exploration. Future research can delve into a deeper understanding of each top feature's impact,

possibly considering temporal or geographical variations. Moreover, incorporating more features or using other machine learning models might yield different insights, enriching our understanding of car price determinants.

- Model Enhancement:** While both models exhibited strong predictive capabilities, there's always room for improvement. Leveraging more advanced algorithms, incorporating additional data, or refining the feature engineering process could further enhance prediction accuracy.

V. CONCLUSION

Through a meticulous examination of a dataset comprising various car characteristics, the significance of specific features in determining car prices was elucidated using Decision Trees and Random Forest models. The models were fine-tuned for optimal performance using the "hyperopt" library. Subsequently, the data was scaled and dimensionally reduced via PCA, retaining five principal components for training.

The models' accuracy in predicting car prices is evident through their performance metrics. Specifically, the Decision Tree model achieved an R-squared value of 89%, while the Random Forest model, post-PCA, secured a score of 91%. These scores signify the models' robustness in predicting car prices, with lower values typically denoting superior performance.

A deeper dive into feature importance revealed attributes like "Power", "Torque", and "Displacement" as paramount. Such findings underscore that car prices are largely influenced by technical specifications and performance metrics. Furthermore, a side-by-side comparison of the top 10 influential features from both models revealed shared attributes, suggesting a consensus on their significance.

However, like all analytical endeavors, this one too has room for enhancement. Potential avenues for improvement include assimilating richer data, probing alternative machine learning paradigms, and delving deeper into feature interplays. While the models adeptly pinpointed critical features, corroborating these findings with real-world market trends would amplify their trustworthiness.

In essence, this analysis serves as a strategic compass for car manufacturers and marketers. By accentuating pivotal features like Wheelbase, Displacement, Ground_Clearance, Model, Airbags, Fuel_Tank_Capacity, Torque, Power, Gears, and Kerb_Weight, industry players can recalibrate their strategies to resonate with market demands and consumer inclinations.

REFERENCES

- [1] Ethem Alpaydin, "Introduction to Machine Learning", The MIT Press Cambridge, Massachusetts London, Fourth edition, 2020 (para 1,2,3,4,5)

- [2] Kevin Patrick Murphy ,”Probabilistic Machine Learning: An Introduction”, MIT Press, March 2022. (para -4, Kernel Machines, Ensemble Learning).
- [3] Machine Learning, Tom Mitchell, McGraw , 1997, 0-07-042807-7
- [4] Bishop, C. Pattern Recognition and Machine Learning. Berlin: Springer-Verlag, 2006.
- [5] Richard A.Johnson and Dean W.Wichern. Applied Multivariate Statistical Analysis,6 th Edition, Pearson Prentice Hall, 2007[Para 3 and Para 4]