LEARNING APPROACH FOR STUDY OF EMISSION STANDARDS ON USED CAR PRICES IN INDIA

A

M
A
C
H
I
N
E

Abstract

This study investigates the influence of emission standards on second-hand car prices in India using advanced machine learning techniques. Utilizing a comprehensive dataset from CarDekho, we performed two distinct analyses to explore this relationship. Initially, we excluded emission standards, employing various regression algorithms, with Random Forest and XGBoost achieving accuracies close to 94%. Upon introducing emission standards into the models, Random Forest's accuracy slightly improved to 94.25%, while XGBoost's accuracy decreased to 88.08%, highlighting different algorithmic responses to regulatory variables. These findings emphasize the critical role of emission standards in predictive modeling, offering valuable insights for policymakers and stakeholders in the automotive industry to make informed decisions that align with environmental objectives and market realities.

Keywords:

Predictive Modeling, Emission Standards, Automotive Economics, Regression Analysis, Algorithm Performance, Machine Learning, Random Forest, XGBoost, Environmental Regulations, Market Valuations

1. INTRODUCTION

The vibrant second-hand car market in India presents a unique economic fabric woven with consumer preferences, regulatory impacts, and environmental concerns. As the global automotive industry continues to evolve, marked by significant advancements in vehicular technology and stringent emission regulations, the Indian used car market has burgeoned. This growth is catalyzed by increased vehicle ownership turnover and consumer demand for cost-efficient transportation. Against this backdrop, our study embarks on an in-depth analysis to explore the influence of emission standards on second-hand car valuations, a subject that remains relatively unexplored in the Indian context. The Indian government's stringent Bharat Stage (BS) emission standards, aligned with European regulations, have significantly evolved over the years. These standards reflect a commitment to reducing vehicular pollution. The transition from BS-IV to BS-VI, skipping directly over BS-V, marked a substantial tightening of norms, reducing allowable emission limits and mandating advanced fuel and engine technologies. This regulatory trajectory not only affects vehicle manufacturers but also shapes the used car market dynamics, as older, less compliant vehicles may depreciate differently under the evolving regulatory landscape. Understanding the impact of these emission standards on second-hand car prices is crucial for various stakeholders, including policymakers, automotive companies, and consumers. This study aims to fill the gap in current literature by offering a focused analysis on the correlation between India's emission regulations and second-hand car pricing. By harnessing the predictive capabilities of advanced machine learning algorithms such as Random Forest and XGBoost, we unravel the intricacies of regulatory impact within the microeconomic framework of India's used car market. Our methodology involves a dual-phase approach. Initially, we establish a performance baseline by training our models without considering emission standards. This phase helps in understanding the primary factors influencing car prices. Subsequently, we integrate emission standards into these models, allowing us to quantitatively assess the impact of these regulations on model accuracy and pricing predictions. This approach provides granular insights into the direct economic repercussions of emission standards. The data for this study was sourced from a comprehensive dataset available on Kaggle, which includes detailed attributes for each vehicle such as make, model, year, mileage, and engine specifications. Extensive preprocessing was performed to clean and standardize the data. High cardinality issues were addressed by consolidating sparse categories, and rigorous data cleaning ensured the removal of erroneous and unrealistic data points. The dataset was then merged with manually compiled emission standards data, aligning each car with its corresponding emission regulation details. In the first phase of our analysis, various regression algorithms were applied to predict car prices without considering emission standards. Random Forest and XGBoost emerged as top performers, achieving accuracies close to 94%. In the second phase, emission standards were introduced into the models. Interestingly, while Random Forest's accuracy slightly improved to 94.25%, XGBoost's accuracy decreased to 88.08%. These results highlight the varying degrees of resilience that different algorithms exhibit when additional regulatory variables are introduced. The findings from this study underscore the sophistication required in predictive modeling for economic analyses and showcase the importance of integrating environmental regulations into economic models. By providing a comprehensive analysis of how emission standards influence second-hand car prices, our research offers valuable insights that can guide strategic planning and policy development. This approach ensures that environmental objectives are harmonized with market realities, promoting a more sustainable automotive market. In summary, our study illuminates the nuanced interplay between regulatory changes and economic outcomes in the Indian second-hand car market. The insights gained from this research are crucial for stakeholders aiming to navigate the complexities of market behaviors influenced by policy. This enriched understanding facilitates informed decision-making, aligning environmental goals with economic activities and contributing to sustainable development.

2. LITERATURE REVIEW

In preparing for our study on the impact of emission standards on used car prices in India using machine learning, we reviewed related work that, while focusing on price prediction, diverges in methodologies or specifics of interest, highlighting the uniqueness of our approach. For instance, a study by Kenneth Gillingham utilized machine learning techniques for predicting used car prices in the Croatian market. They employed supervised learning methods, demonstrating the potential of data mining for predictive accuracy but did not address emission standards, which are central to our research [1]. Another research effort by N. Pal, P. Arora, P. Kohli, D. Sundararaman, and S. S. Palakurthy highlighted the use of deep learning models to predict used car prices. They specifically emphasized the advantages of combining XGBoost and LightGBM for handling large feature sets, a methodology that complements our approach but is applied in a different context without considering regulatory impacts like emission standards [2]. Further, a study by E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric explored the application of machine learning in predicting vehicle prices in the context of the Internet of Things and smart manufacturing. They utilized models such as neural networks, decision trees, and support vector machines to optimize predictions, underscoring the potential of IoT data in refining economic models, but their work did not consider emission standards [3]. Additionally, S. Sinha, R. Azim, and S. Das integrated machine learning with vehicle telemetry data to enhance predictive accuracy in used car pricing. This research underscores the potential of IoT data in refining economic models but lacks the specific focus on emission standards that our study addresses [4]. A study by A. Fathalla, A. Salah, K. Li, K. Li, and P. Francesco explored the use of ensemble learning techniques to improve the robustness of price predictions in highly volatile markets. Their research illustrates the benefits of combining multiple predictive models, a strategy that aligns with our approach of using Random Forest and XGBoost but without the integration of emission standards into their models [5]. Chetna Longania, Sai Prasad Potharaju, and Sandhya Deore employed a mixed-model approach to investigate the effects of macroeconomic factors on used car prices. Their work provides a broader economic perspective that complements our emission-focused analysis by highlighting the importance of external economic variables in price prediction models [6]. Another relevant study by Aarone Steve J. Alexstan, Krishna M. Monesh, M. Poonkodi, and Vineet Raj used Random Forest and other machine learning algorithms to predict used car prices based on numerous features such as mileage, year of manufacturing, and engine size. However, their study did not consider emission standards [7]. A study by K.Samruddhi and Dr. R.Ashok Kumar employed the K-Nearest Neighbor (KNN) algorithm to predict used car prices. While their model demonstrated significant predictive accuracy, it did not incorporate emission standards into the analysis [8]. N.Sun, H. Bai, Y. Geng, and H. Shi developed a price evaluation model for second-hand cars using BP neural network theory, showcasing the potential of neural networks in price prediction. Their study, however, did not address the impact of emission regulations [9]. Research by S. Peerun, N. H. Chummun, and S. Pudaruth utilized artificial neural networks to predict the prices of second-hand cars, emphasizing the model's ability to handle complex datasets. Their work also lacked a focus on emission standards [10]. Lucija Bukvić, Jasmina Pašagić Škrinjar, Tomislav Fratrović, and Borna Abramović conducted a study using supervised machine learning for price prediction in the used car market. They identified the potential of various algorithms for predictive accuracy but did not incorporate emission standards [11]. Each of these studies contributes to the broader discourse on machine learning applications in automotive pricing but lacks the specific focus on regulatory impacts, particularly emission standards, which we address in our analysis. By integrating emission standards into our predictive models, we offer a more comprehensive understanding of their direct economic repercussions, enhancing the reliability and relevance of our findings for policymakers and stakeholders in the automotive industry. This examination not only underscores the relevance of our study but also enriches our understanding of varied methodologies and their applications across different markets and technological settings.

3. METHODOLOGY

This study employs a structured approach to examine the impact of emission standards on the pricing of second-hand cars in India, utilizing machine learning techniques to perform predictive modeling. The methodology is divided into several key components: data collection and preprocessing, model selection and development, and RESU.

3.1 DATA COLLECTION AND PREPROCESSING

Table. 1. Car Dataset Description

Column Name	Data Type	Description
loc	object	Location of the car
myear	int64	Manufacturing year of the car
bt	object	Body type of the car
tt	object	Transmission type of the car
ft	object	Fuel type of the car
km	object	Kilometres driven by the car
ip	int64	Insurance premium of the car
images	object	Number of images of the car
imgCount	int64	Count of images of the car
threesixty	bool	Whether 360 degree view of car is available or not
dvn	object	Dealer verification status of the car
oem	object	Original equipment manufacturer of the car
model	object	Model of the car
variantName	object	Name of the car variant
city_x	object	City of the car
pu	object	Price of car at the time of purchase
discountValue	int64	Discount offered on the car
utype	object	Type of the car seller
carType	object	Type of the car (hatchback, sedan, SUV, etc.)
top_features	object	List of top features of the car (e.g., ABS, airbags)
comfort_features	object	List of comfort features of the car (e.g., air conditioning)
interior_features	object	List of interior features of the car (e.g., leather seats)
exterior_features	object	List of exterior features of the car (e.g., alloy wheels)
safety_features	object	List of safety features of the car (e.g., ABS, airbags)

Color	object	Color of the car
Engine Type	object	Type of engine (very specific)
Max Power	object	Maximum power output of the engine in bhp
Max Torque	object	Maximum torque output of the engine in Nm
No of Cylinder	int64	Number of cylinders in the engine
Values per Cylinder	int64	Number of valves per cylinder in the engine
Value Configuration	object	Configuration of the valves in the engine (SOHC, DOHC, etc.)
BoreX Stroke	object	Bore and stroke dimensions of the engine
Turbo Charger	object	Indicates if the engine has a turbocharger
Super Charger	object	Indicates if the engine has a supercharger
Length	object	Length of the car in mm
Width	object	Width of the car in mm
Height	object	Height of the car in mm
Wheel Base	object	Wheel base of the car in mm
Front Tread	object	Front tread of the car in mm
Rear Tread	object	Rear tread of the car in mm
Kerb Weight	object	Kerb weight of the car in kg
Gross Weight	object	Gross weight of the car in kg
Gear Box	object	Gear box type of the car (4 speed, 5 speed, etc.)
Drive Type	object	Drive type of the car (RWD, FWD, etc.)
Seating Capacity	int64	How many people can sit in the car
Steering Type	object	The type of steering (power steering, manual steering, etc.)
Turning Radius	object	Turning radius of the car in meters
Front Brake Type	object	Front brake type of the car (disc, drum, etc.)
Rear Brake Type	object	Rear brake type of the car (disc, drum, etc.)
Top Speed	object	Top speed of the car in kmph
Acceleration	object	0-100 kmph acceleration time of the car in seconds
Tyre Type	object	Tyre type of the car (tubeless, tube, etc.)
No Door Numbers	int64	Number of doors in the car
Cargo Volumn	object	Amount of cargo space in the car in liters
model_type_new	object	Whether the car is a new car or a used car
state	object	State in which the car is located
owner_type	object	Owner type of the car (first, second, etc.)
Fuel Suppy System	object	Type of fuel supply system (Carburetor, Fuel Injection, etc.)
Compression Ratio	object	Compression ratio of the engine
Alloy Wheel Size	object	Size of the alloy wheels in inches
Ground Clearance Unladen	object	Ground clearance of the car when it is not loaded in mm

The data for this study as show in in Table 1 is sourced from a comprehensive dataset available on Kaggle, which in turn was scraped using an API. This dataset includes detailed attributes for each vehicle, such as make, model, year, mileage, engine specifications. To begin with the preprocessing, we have kept the columns shown in Table 1, it now has 62 columns, columns are hand-picked and will be further analyzed. Duplicated rows are deleted and the column names are lowered. significant efforts were made to address high cardinality issues in our dataset. High cardinality, where categorical features contain many unique values, can degrade model performance due to overfitting and increased computational complexity. To mitigate these issues, specific strategies were applied, particularly to features such as 'Value Configuration', 'Turbo Charger', 'Gear Box', 'Drive Type', 'Steering Type', 'Front Brake Type', 'Rear Brake Type', and 'Tyre Type'.One primary approach was the consolidation of sparse categories into more general ones, reducing the overall complexity of the data. For instance, in the 'Gear Box' column, various descriptions of gearbox types such as '5 speed', '6 speed', '5-speed', were standardized to fewer categories to maintain essential information while reducing granularity as shown in Fig 1 and Fig 2.

Fig.1. Value counts of each unique value in Drive-Type

```
Drive Type
fwd                                   27456
nan                                    4496
rwd                                    2248
awd                                    1082
2wd                                     648
4wd                                     570
2 wd                                    369
4x2                                     297
4x4                                     229
front wheel drive                       176
two wheel drive                          98
all wheel drive                          32
rear wheel drive with esp                29
two whhel drive                          26
permanent all-wheel drive quattro        21
rwd(with mtt)                            14
rear-wheel drive with esp                 7
4 wd                                      7
all-wheel drive with electronic traction  5
four whell drive                          2
3                                         1
```

Fig.2. Grouping of Values in Drive-Type Column

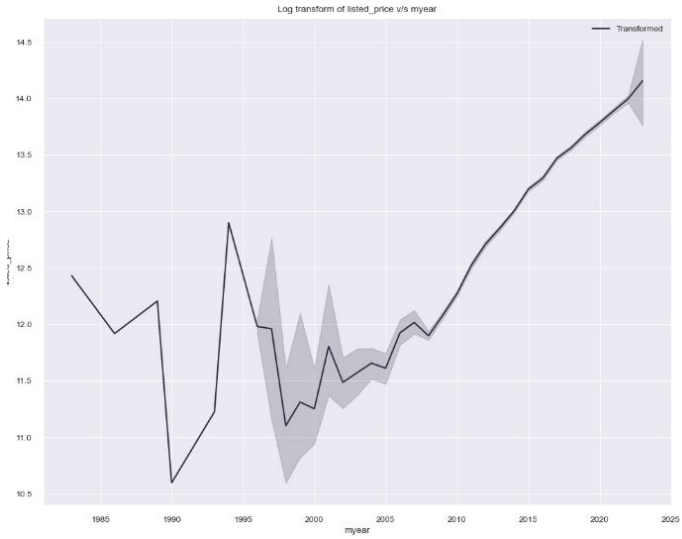
```
drive_type_mapping = {}
drive_type_mapping['fwd'] = ['fwd', 'front wheel drive']
drive_type_mapping['2wd'] = ['2wd', 'two wheel drive', '2 wd', 'two whhel drive']
drive_type_mapping['rwd'] = ['rwd', 'rear wheel drive with esp', 'rear-wheel drive with esp', 'rwd(with mtt)']
drive_type_mapping['awd'] = ['awd', 'all wheel drive', 'all-wheel drive with electronic traction', 'permanent all-wheel drive quattro']
drive_type_mapping['4wd'] = ['4wd', '4 wd', '4x4', 'four whell drive']
drive_type_mapping['nan'] = ['nan', '3']

mapping_dict = {v: k for k, lst in drive_type_mapping.items() for v in lst}
cars_df['Drive Type'] = cars_df['Drive Type'].replace(mapping_dict)
```

After addressing high cardinality, the focus shifted to rigorous data cleaning, for instance the columns of "'Max Power" had values like 70bhp@4000rpm, which were split into two columns "Max Power Delivered" and "Max Power At", similar operation was performed for Max Torque column. The columns which should be Boolean were converted so from string type, and all the columns which are needed to be converted to float or int were done so. After that, to increase readability, column names were changed for example 'tt' to 'transmission'. After basic cleaning we had four questions to answer before modelling, what factors have the biggest impact on the price of a vehicle? which features are important enough to keep in the model? how would we handle the missing values? how would

we handle the outliers?

There are only a few cars which are priced near Rs. 1 Crore, but one car has a price of 5.5 Crore. We can drop this row for better visualization and homogenization purposes. Next for the myear column which is the year of manufacture of the car, for the columns myeardrop the cars that are manufactured before 2005 which are 224 in number. We can see that the mean price of the cars manufactured before 2005 is very low compared to the rest of the cars. Even factoring in the exponential nature of the variable's relation with price, the pre-2005 data is very noisy as shown in Fig 3. Observations for the columns myearthe price increases exponentially with the year of manufacturing.

Fig.3. Log transform of listed_price vs myear



For the body column, everything seems to be fine here, except 7 of out of the 11 categories have cars which are less than 100 number. Observations for the column body Hatchbacks, sedans and SUVs make up for almost 95% of the cars in our dataset as shown in Fig 4

In the transmission column, everything is normal, we notice that there are a lot more manual cars than automatic cars, and automatic cars are more expensive. As one can see in Fig 5, the interquartile range for automatic cars is more than that of manual cars. For fuel column everything is as expected, we have 5 categories, petrol, diesel, cng, lpg and electric, with petrol being the most present. For the dvn column, which is nothing but the complete name of the car (oem, model and variant name), this column seems to be having very high cardinality, after performing an ANOVA test, we decided to drop the column.

Fig.4. Count of body types in our dataset

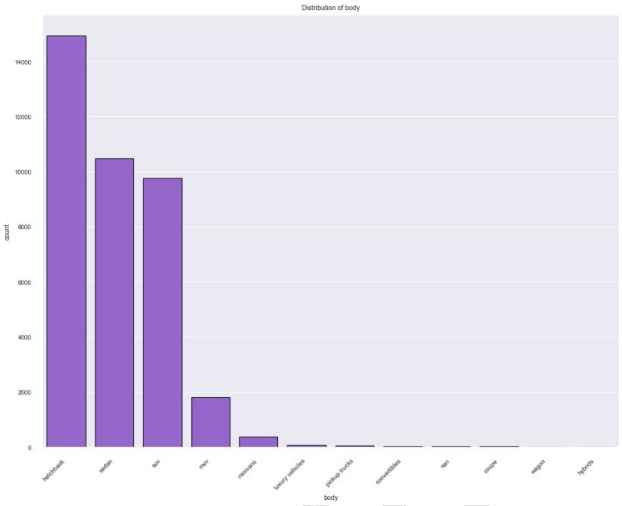
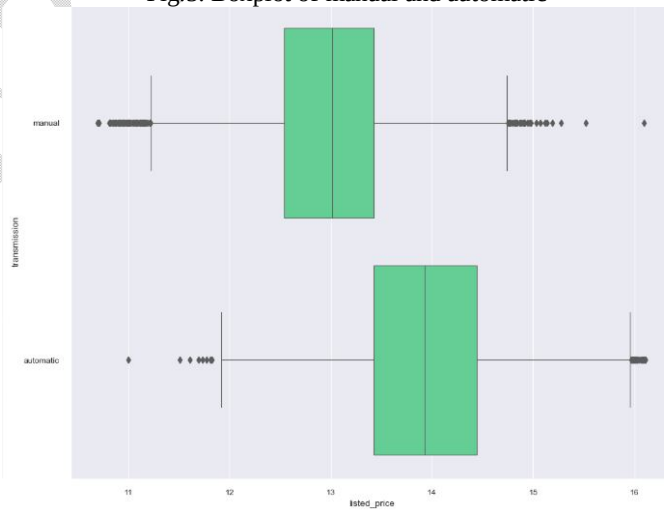


Fig.5. Boxplot of manual and automatic



For the oem column (manufacturer name column) we could drop some of the luxury car brands for better homogenization of the data, but we decided to keep them for generalization. If it adversely affects the model's performance, we could get rid of them later. For the “No. of cylinder column” we have an electric car claiming to have 16 cylinders, which is obviously wrong, we further inspected and dropped all such rows, for the Length, Width and Height columns we find the spearman correlation coefficient as shown in Fig 6, height is not very informative. drop the column length as it is highly correlated with width. If training a linear model, dropping one of them would be advisable, similarly, after studying the wheel base, wheel Base is VERY HIGHLY correlated with Length (coefficient of 0.91). We should drop one of them, but the choice is not entirely clear - Length has a lower influence on price but is less correlated with Width.The minimum value of the column Rear Tread is '15' which is not realistic and most probably were reported wrongly. The columns Rear Tread and Front Tread are VERY

HIGHLY correlated (0.92). So just drop the column Rear Tread. It will solve the collinearity of variables problem as well as help us avoid the wrongly reported values in the column. For the kerb weight and gross weight, we can drop gross weight since it is highly correlated with kerb weight as shown in Fig 7, and has 50% values missing

Fig.6. spearman correlation matrix of 'Length', 'Width', 'Height', TARGET

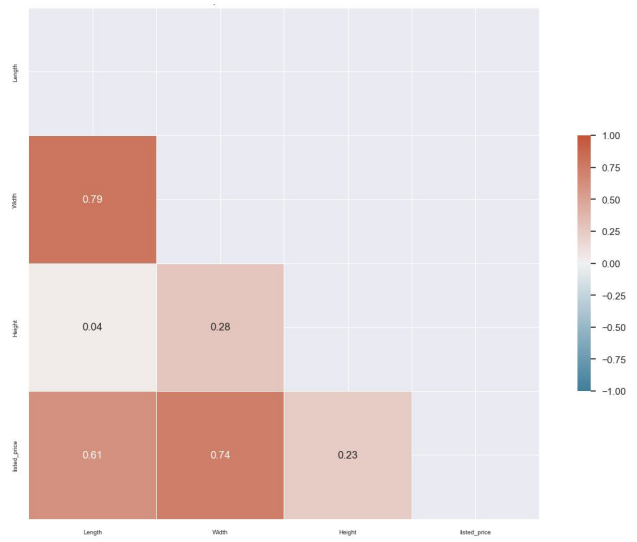
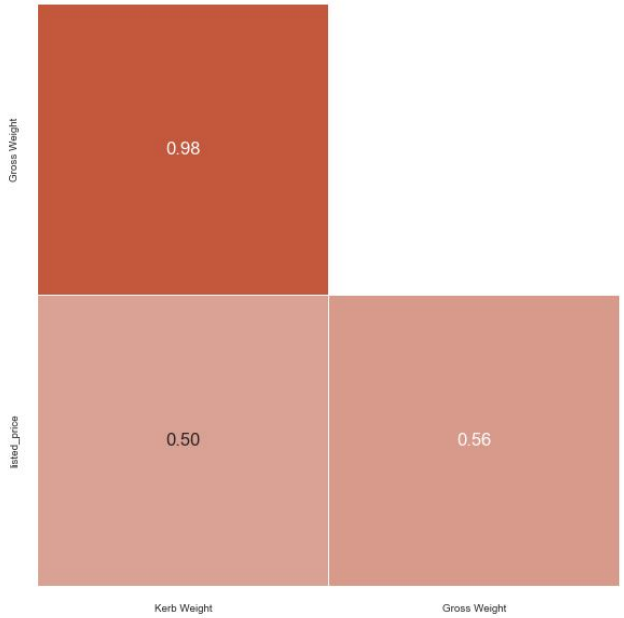


Fig.7.spearman correlation matrix of Gross Weight, Kerb weight and TARGET



For seats column we have dropped all rows with 0 seats, which is wrongly reported.We can see there are 2 cars which have a turning radius greater than 6000(m). Obviously, this data is wrongly reported, and we should nullify the data.

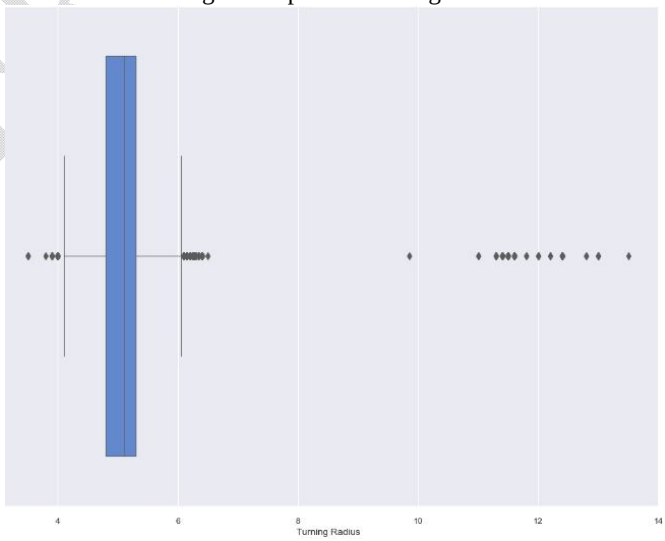
We musthandle the turning radius which are greater than, say 15m. One way would be to clip the values with a maximum and minimum range. But a better approach would be to make those values null straight awayas shown in Fig 8

For “Front Brake Type” and “Rear Brake Type” we can see from the chi2 test, that the degree of association is quite high between the 2 columns. It would be beneficial to drop one of them for interpretability, but here we decide to keep them since it will not harm the performance of the model as shown in Table 2, null hypothesis being there is no correlation between the two columns.

Table.2. chi2 test between the columns 'Front Brake Type' and 'Rear Brake Type'

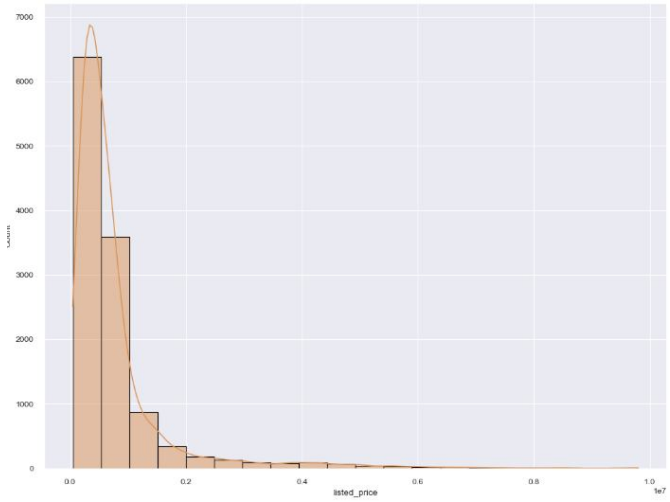
Chi2	p-value	dof
115615.93898398022	0.0	36

Fig.8. Boxplot of Turning Radius



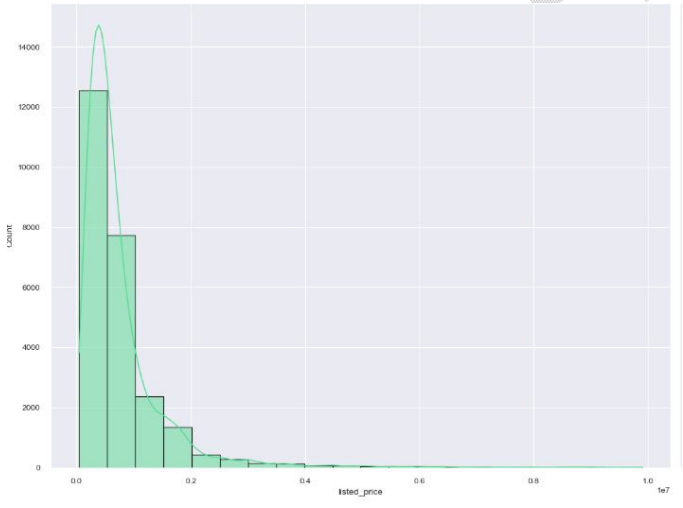
For doors we can see that there is not any major variation in the price of the car with the number of doors. But this is only becausemost of the cars have 4 or 5 doors, which is not a huge factor but for the rest, the price varies quite a bit. Therefore, it would not be wise to drop this column. However, as the above distributions suggest, the cars with 4 doors have a very similar distribution to those having 5 doors as shown in Fig 9 and Fig 10. We should just replace all the columns which have 5 as the Doors with 4. Since this is not a transformation only on the training set and can also affect predictions of new cars.

Fig.9. Distribution listed_price vs count of 4 doors



The “Ground Clearance Unladen”” column as well as the “Compression Ratio” column were dropped a they had more than 60% values missing, in “Alloy wheel size” column, rows with wheel size 7, which were 2 rows, were dropped as they were outliers. For the column “Max Torque At,”Gas cars do not output their peak torque at below 1000 rpm. Since we know this and the cars which say their output it below 1000 are not expensive cars is observed, it is most likely a mistake.Also, there is a car which has its peak torque delivered at 21000 rpm. This is also bad data and should be ignored in our EDA.columns Bore and Stroke Drop both the Bore and Stroke columns because over 60% and 90% data missing respectively. Similar operations were applied for all the other columns. After extensive operations we move onto creating our Regulations Data frame, for which data has been created manually.

Fig.10. Distribution listed_price vs count of 5 doors



The dataset representing emission standards over the years was meticulously compiled into a structured DataFrame, designed to encapsulate the evolution of vehicular emission norms in India. This DataFrame, labeled df_reg, includes several key columns that detail the progression of regulations aimed at controlling vehicular emissions.

'Norms': This column lists the various emission standards established over the years, starting from "1991Norms" through to "Bharat Stage-IV". Each entry represents a significant regulatory milestone that reflects tightening emission control measures.
'CO(g/km)': Carbon Monoxide (CO) emissions limits are specified for each set of norms, presented in ranges for some older norms and as precise values for more recent standards. These limits demonstrate the regulatory intent to reduce vehicular pollution over time.
'HC+ NOx(g/km)': Limits for Hydrocarbons plus Nitrogen Oxides (HC+NOx) are also specified, reflecting the combined regulatory caps on these pollutants. Initially listed as separate components or in less strict terms for earlier norms, the values become more stringent and are combined in more recent norms, indicating a shift towards more comprehensive emission controls.
'year': The years column chronologically aligns each norm with its implementation year, providing a timeline from 1991 to 2020. This temporal mapping is crucial for understanding the historical context and progression of emission standards in India.
In the preprocessing of emission standards data, the Carbon Monoxide (CO) and Hydrocarbons plus Nitrogen Oxides (HC+NOx) emissions values underwent standardization to ensure uniformity and precision. For CO emissions initially presented as ranges, an average of the two numbers was computed to obtain a single representative value. If the emission value was provided as a single number, it was directly converted into a float for consistency. Similarly, the HC+NOx values required cleaning to remove extraneous text and convert range values into single averages. This was achieved by stripping any parenthetical information and averaging numbers if presented as a range. This process of refining emission values ensures they are quantifiable and uniform across the dataset, allowing for more accurate analysis and modeling.

Following the cleaning of emission data, a crucial step involved integrating this information with the car dataset. Each car's year of manufacture was used to assign the most recent emission standard applicable at that time. If no prior standards were relevant, the entry was marked to reflect the absence of applicable norms. Finally, the car dataset and the cleaned emission standards data were merged based on the year of emission standard applicability, effectively aligning each car with its corresponding emission regulation details. This comprehensive approach not only improves the reliability of the analysis but also enhances the dataset's relevance for studying the impact of emission standards on car pricing.

Table.3. Our emission standards DataFrame post cleaning

Norms	**CO (g/km)**	**HC+ NOx (g/km)**	**Year**
1991 Norms	20.70	2.00	1991
1996 Norms	10.54	3.68	1996
1998 Norms	5.27	1.84	1998
India Stage 2000 Norms	2.72	0.97	2000
Bharat Stage-II	2.20	0.50	2005
Bharat Stage-III	2.30	0.35	2010
Bharat Stage-IV	1.00	0.18	2020

After processing and merging, the following columns have been dropped due to having more than 10% missing values
front_tread, top_speed, alloy_wheel_size, acceleration, kerb_weight, cargo_volume, turning_radius

3.2 MODEL SELECTION AND IMPLEMENTATION

We have divided our study into two parts, one is to predict with emission standards as features, and one is to predict the listed_pricewithout and then compare. First, we will use regression without taking emission standards.

3.2.1 WITHOUT EMISSIONS STANDARDS

We first encode our data. In the preprocessing of categorical data for machine learning, the choice of encoding technique is crucial for the effectiveness of the models. The use of LabelEncoder from sklearn.preprocessing for transforming categorical columns in our dataset is justified by its simplicity and computational efficiency, which is essential for handling large datasets swiftly. LabelEncoder is particularly suitable for algorithms requiring numeric input, as it encodes categories into a numerical format that algorithms can easily process. Moreover, it preserves any inherent ordinal relationships within the data, which can be beneficial for models where the order of categories impacts the outcome. Unlike one-hot encoding, LabelEncoder does not expand the feature space, thus avoiding an increase in dimensionality that could potentially lead to model complexity and overfitting issues. This method ensures a consistent and manageable transformation of categorical data, maintaining dataset integrity and aiding in the seamless integration of various features into predictive models. For numerical columns we used the StandardScaler from sklearn.preprocessing, which normalizes each feature to have zero mean and unit variance. This step is essential for models that rely on distance calculations or optimization algorithms, as it ensures all features contribute equally without any single feature dominating due to its scale. Additionally, scaling helps mitigate the impact of outliers by reducing the range within which the values vary, thereby preventing extreme values from disproportionately influencing the model's performance. This standardization of numerical data enhances the robustness and reliability of our machine learning models, ensuring that the outcomes are reflective of the underlying patterns in the data rather than artifacts of variable scales.

3.2.1.1 LINEAR REGRESSION

Linear Regression assumes a linear relationship as shown in Eq.(1) between the dependent variable (car price) and independent variables (car features). The model fits a line through the data points to minimize the sum of squared differences between the observed and predicted values.

Data Preprocessing

To preprocess the data, we:

- Encoded categorical variables using LabelEncoder, which converts categories into numeric values, preserving any ordinal relationships.
- Standardized numerical columns using StandardScaler to normalize each feature to have zero mean and unit variance. This step is essential for models that rely on distance calculations or optimization algorithms.

Feature Importance

Feature importance was assessed to understand the contribution of each feature to the predicted car prices. The coefficients of the Linear Regression model indicate the importance of each feature. A higher absolute value of a coefficient indicates a more significant impact on the prediction.

$$y = a + bx \qquad (1)$$

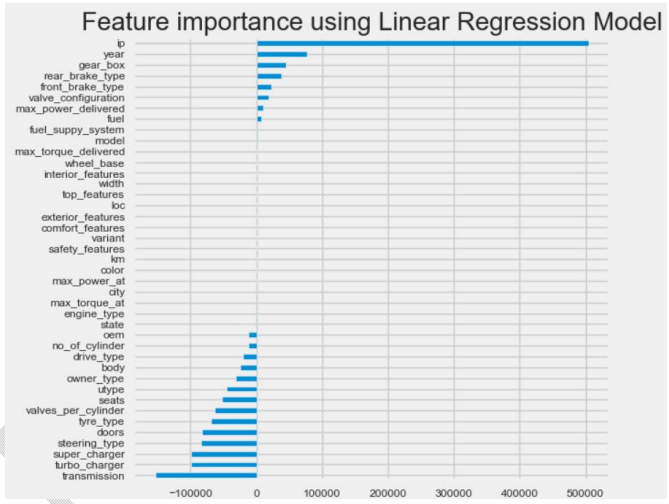
Where a and b are the intercept and slope respectively

$$b = \frac{(n\sum xy - (\sum x)(\sum y)}{(n\sum x\text{^}2 - (\sum x^2))}$$

$$a = \frac{\sum y - b(\sum x)}{n}$$

Since our main aim is to see the impact of features on our model, in Fig 11, one can see the importance of each feature in our predicted output.

Fig.11. Feature importance of Linear Regression



The feature importance graph produced by the Linear Regression model illustrates the impact of each feature on the predicted car prices. The x-axis represents the importance score, while the y-axis lists the features.

- **Positive Values**: Features with positive values indicate that an increase in these features will increase the car price.
- **Negative Values**: Features with negative values indicate that an increase in these features will decrease the car price.

In Fig 11, we see that features such as year, gear_box, and rear_brake_type have a significant positive impact on the car prices, while features like turbo_charger and transmission have a significant negative impact. This insight helps in understanding which factors most influence car prices, guiding both market strategies and policy decisions.

The scale of the graph is crucial in interpreting the impact magnitude. The x-axis values range from negative to positive, with the length of each bar representing the absolute value of the coefficient. A longer bar indicates a higher impact on the price, either positively or negatively. This scale helps visualize the relative importance of each feature, allowing for an intuitive understanding of their influence on car pricing.

This same methodology will be applicable throughout our study.

3.2.1.2RIDGE REGRESSION

In ridge regression, an additional term of “sum of squares of the coefficients” is added to the cost function in Eq.(2) Ridge regression essentially does is to try to minimize the sum of the error term along with sum of squares of coefficients which we try to determine. The

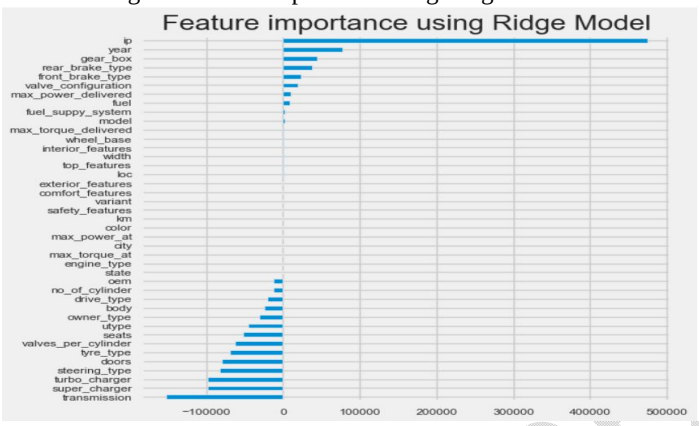
sum of the squares of the coefficients is called as 'regularization term' and it also has the regularization coefficient denoted by λ.

$$\beta ridge = ((X'X + \lambda Ip)^{-1})X'Y \qquad (2)$$

- βridge is the vector of coefficients estimated by the Ridge Regression.
- X represents the matrix of input features.
- X' is the transpose of X.
- Y is the vector of output/target values.
- λ is the regularization parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage.
- $I$is the identity matrix of appropriate size.
- The term λI ensures that the regression coefficients are shrunk towards zero to prevent overfitting

in Fig.12. one can see the importance of each feature in our predicted output

Fig.12. Feature importance using Ridge model



3.2.1.3 RANDOM FOREST

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

$$y = \frac{1}{B}\sum_{a=1}^{B}\tau_a(x) \qquad (3)$$

As shown in Eq.(3) B is the number of trees, and τ_a represents the prediction of the ath tree. Each of these models contributes uniquely to understanding the dynamics of used car pricing, leveraging different mathematical principles to minimize prediction error and optimize performance.

in Fig 13, one can see the importance of each feature in our predictied output., and in Fig 14 we can see the graph between actual and predicted

3.2.1.4XGBOOST

XGBoost is an advanced implementation of gradient boosting algorithm. This model uses a gradient boosting framework for making predictive models, which involves three main components: a loss function to be optimized, a weak learner to make predictions, and an additive model to add weak learners to minimize the loss function.

XGBoost involves an objective function that is a combination of a loss function and a regularization term as shown in Eq.(4). The loss function depends on the specific problem (e.g., regression, classification). For regression tasks like predicting car prices, it typically uses the squared error.

Fig.13. Random Forest Variable importance

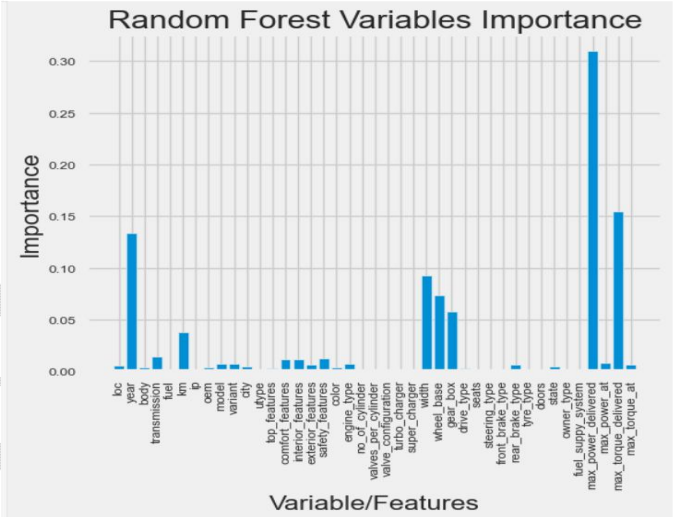
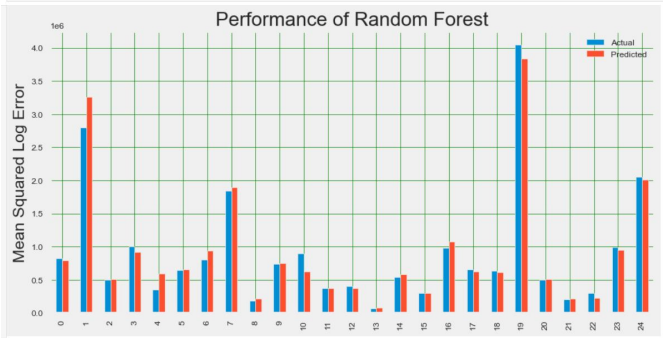


Fig.14.Performance of Random Forrest

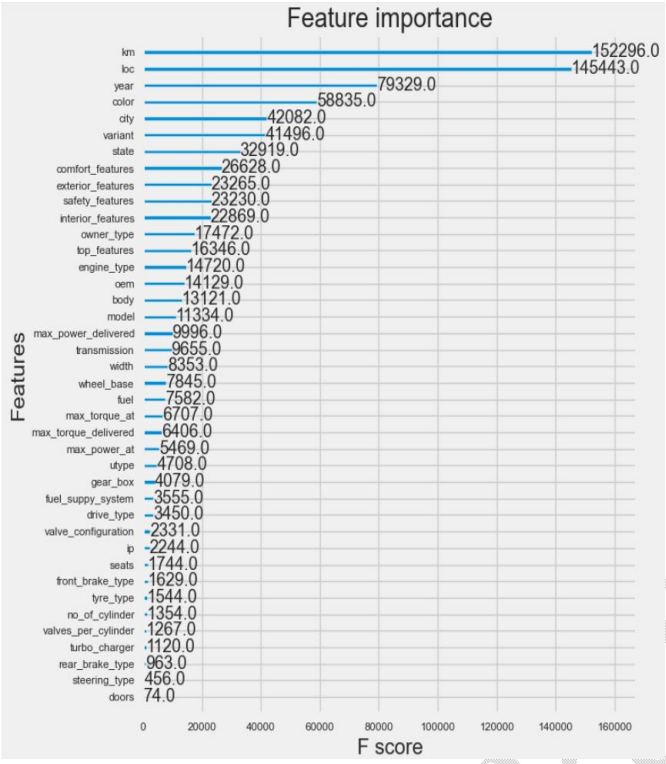


$$Obj = \sum_{i=1}^{n} l(yi, \hat{y}i) + \sum \Omega(fk) \qquad (4)$$

Here,l is the loss function that measures the difference between the predicted $\hat{y}i$ and actualyi values, and Ωrepresents the regularization term which helps in reducing overfitting by penalizing the complexity of the model.

Gradient and Hessian: XGBoost uses a second-order Taylor expansion of the loss function, involving both the gradient and the Hessian, which allows for more effective optimization.
Tree Ensemble Model: The model involves adding new trees that predict the residuals or errors of prior trees combined to make the final prediction more accurate.
In Fig 15, we can see the feature importancegraph

Fig.15.Feature importance in XGboost



In XGBoost, the feature importance graph uses the F-score to indicate the importance of features. The F-score is a metric that represents the number of times a feature is used to split the data across all trees in the model. Essentially, it measures how often a particular feature contributes to reducing the model's prediction error, with a higher F-score indicating greater importance.The F-score helps identify which features are most influential in making accurate predictions, providing insights into the model's decision-making process and highlighting the key factors affecting the target variable.

3.2.2 WITH EMISSIONS STANDARDS

Building upon the foundational models developed without considering emission standards, the next phase of our study introduces emission standards as a pivotal variable to assess their impact on second-hand car prices. This phase aims to explore whether the inclusion of regulatory compliance data, specifically the varying levels of emission standards over the years, adds significant predictive power or alters the dynamics captured by our models. By incorporating emission standards, we can more accurately reflect the influence of environmental regulations on market values, potentially uncovering nuanced relationships between regulatory compliance and vehicle pricing. This approach not only deepens our understanding of the automotive resale market but also aligns with global trends towards more environmentally conscious economic assessments. By integrating these standards, we expect to provide a more holistic view of the factors that influence used car prices, offering valuable insights for consumers, policymakers, and industry stakeholders about the economic impacts of emission regulations.
We will majorly be focussing on feature importance and accuracy scores.

3.2.2.1 LINEAR REGRESSION

Linear Regression was utilized to establish a baseline for understanding how linearly the variables, including emission standards, are associated with car prices. This model helps to quantify the direct impact of changes in emission standards on the pricing of used cars., we can see the feature importance graph in Fig 16

3.2.2.2 RIDGEREGRESSION

To account for potential multicollinearity between features, especially with the addition of emission standards, Ridge Regression was employed. This model extends Linear Regression by introducing an L2 regularization term, which helps to manage overfitting by penalizing large coefficients. We can see the feature importance in Fig 17

Fig.16. feature importance in Ridge regression

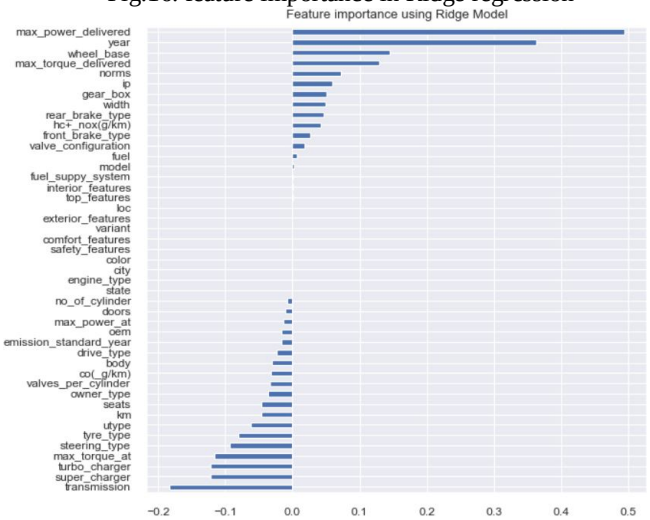
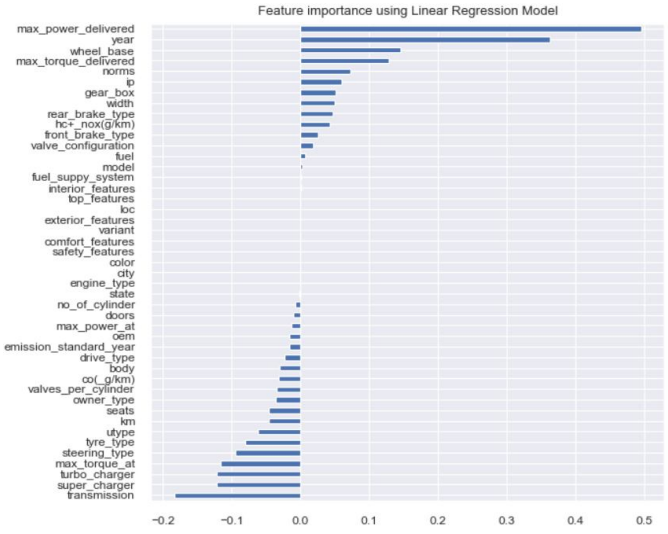


Fig.17. Feature importance in Linear Regression



3.2.2.3 RANDOM FOREST

As an ensemble method, Random Forest was used to handle the dataset's complexity and non-linear relationships more effectively than linear models. It constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. We can see the feature importance in Fig 18

3.2.2.4 XGBOOST

XGBoost was selected for its efficiency and performance in dealing with structured data. Known for its speed and performance, this gradient boosting framework builds sequential models that aim to correct the residuals of the previous models, thereby improving accuracy incrementally. We can see the feature importance in Fig 19

Fig.18. feature importance in random forest

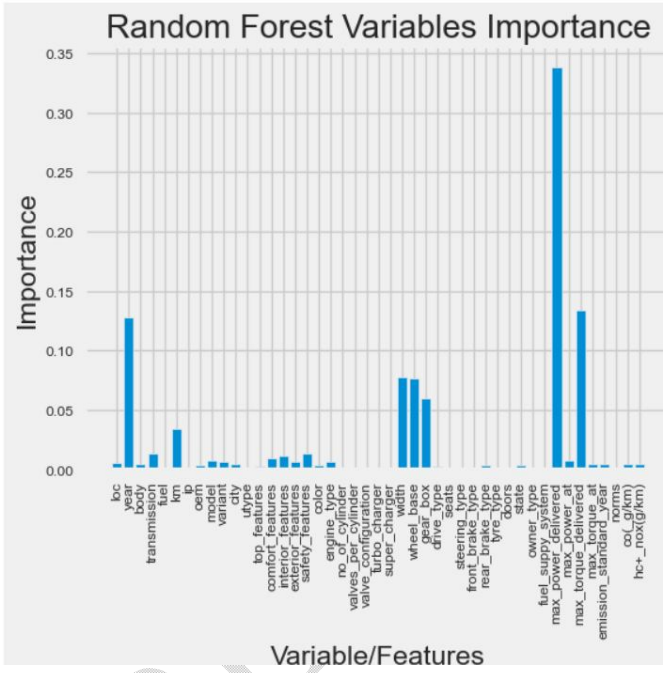
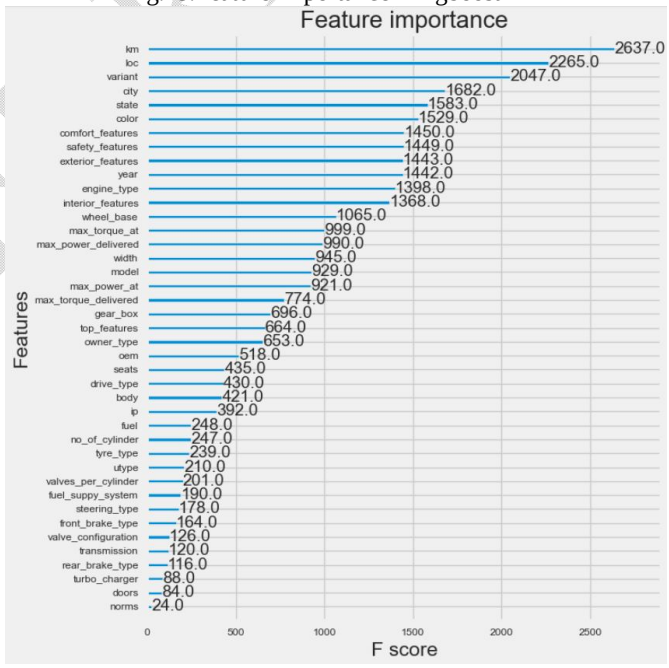


Fig.19. feature importance in xgboost



4. RESULTS AND FUTURE WORK

In table 4, we have the accuracy scores of each model in each case. In our study, we focused on using accuracy (R2 score) as the primary metric for evaluating the performance of our machine learning models. The choice of accuracy is primarily due to its intuitive interpretability and its ability to provide a straightforward measure of how well our models predict the actual prices of second-hand cars. Accuracy, represented as the R2 score, indicates the proportion of variance in the dependent variable (car prices) that is predictable from the independent variables (features).

While there are other metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which provide insights into the average magnitude of prediction errors, they do not offer the same level of interpretability as the R2 score. The R2 score, expressed as a percentage, directly communicates how well the model's predictions align with the observed data, making it easier for stakeholders to understand the model's effectiveness.

Table 4 Accuracy (R2 in % scores)

Accuracy (%)	Linear Regression	Ridge Regression	Random Forrest	XGBoost
With emissions	61.42	61.39	94.25	88.08
Without Emissions	70.15	70.16	93.98	93.31

The results from our predictive modelling clearly illustrate how the inclusion of emission standards influences model accuracy. Linear Regression and Ridge Regression models exhibited a noticeable decrease in accuracy when emission standards were considered, dropping from approximately 70% to just over 61%. This reduction highlights the models' limited capability to handle the additional complexity introduced by the emission standards. Conversely, ensemble methods, particularly Random Forest and XGBoost, managed the added complexity more effectively. Random Forest's accuracy improved slightly from 93.98% to 94.25% with the inclusion of emission standards, demonstrating its robustness and adaptability. Although XGBoost's accuracy decreased from 93.31% to 88.08%, it still maintained substantial predictive capability. An essential advantage of Random Forest and XGBoost is their ability to provide feature importance metrics. This feature is crucial as it helps identify which variables, including emission standards, significantly impact car prices. Even with the decrease in XGBoost's accuracy, the feature importance analysis revealed that emission standards play a meaningful role in pricing dynamics. This insight underscores the importance of including emission standards in predictive models, not only for regulatory compliance but also for enhancing the explanatory power of the models. Understanding the influence of each predictor is as vital as the prediction itself, especially in complex analytical tasks.

The results align with the current buying patterns in India, where emission standards are not a primary concern for most buyers. Indian consumers typically prioritize factors such as price, brand, and mileage over environmental regulations. This is reflected in the decreased accuracy of simpler models like Linear Regression and Ridge Regression, which struggle to account for the nuanced impact of emission standards. However, as government regulations become stricter, there will likely be a shift in consumer behaviour. The implementation of more rigorous emission standards could compel buyers to consider these factors more seriously in their purchasing decisions.

While this study provides valuable insights, several shortcomings need to be addressed in future research to enhance the robustness and applicability of the findings. The dataset used may have limitations in terms of the breadth and depth of features available. Future studies should aim to include more comprehensive datasets with additional relevant features to improve model accuracy and reliability. Although ensemble methods like Random Forest and XGBoost provide high accuracy, they are often seen as "black-box" models. Efforts should be made to improve the interpretability of these models to better understand the underlying relationships between features and outcomes. Future research should also consider incorporating more advanced features such as economic indicators, regional environmental policies, and consumer behaviour trends to provide a more holistic view of the market dynamics. Extending the analysis to other geographical regions with different regulatory frameworks and market conditions could validate the findings and provide more generalized insights. Conducting a temporal analysis to study how the impact of emission standards evolves over time could provide dynamic insights into the market adjustments to regulatory changes.

5. CONCLUSION

This study underscores the significant impact that emission standards have on the pricing of second-hand cars in India, a factor that has not been extensively explored in previous research. By employing a range of machine learning models, including Linear Regression, Ridge Regression, Random Forest, and XGBoost, we demonstrated that the inclusion of emission standards in predictive models affects their accuracy. Specifically, simpler models like Linear Regression and Ridge Regression saw a decrease in performance with the inclusion of emission standards, while robust ensemble methods like Random Forest and XGBoost maintained high accuracy and provided valuable insights through feature importance metrics. These findings are crucial, as they highlight the often-overlooked role of emission standards in the resale value of cars. The current Indian market primarily focuses on factors such as price, brand, and mileage, with less emphasis on environmental regulations. However, as the government implements stricter emission standards, these regulations will likely become more significant in consumer decision-making. Understanding this shift is essential for stakeholders, including policymakers, environmental regulators, automotive companies, and consumers, as it can guide more informed and sustainable decision-making. The study also points to several areas for future research. Enhancing data quality and availability, improving model interpretability, incorporating advanced features like economic indicators and consumer behaviour trends, and extending the analysis to other geographical regions are critical steps. Conducting temporal analyses to study the evolving impact of emission standards over time will provide dynamic insights into market adjustments to regulatory changes. In conclusion, the integration of emission standards into predictive models is not just about regulatory compliance; it enhances the explanatory power of the models and ensures that all significant factors influencing used car prices are considered. This comprehensive approach will lead to more accurate and holistic market assessments, driving smarter, more sustainable business strategies and policies that align with global environmental goals. By doing so, this research not only fills a critical gap in current literature but also provides practical implications for shaping the future of the second-hand car market in India.

Disclaimer (Artificial intelligence)

Option 1:
Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

REFERENCES

[1] Kenneth Gillingham, "Assessing Environmental Regulation in Automobile Markets," NBER report, 2022.

[2] N. Pal, P. Arora, P. Kohli, D. Sundararaman, S. S. Palakurthy, "How much is my car worth? A methodology for predicting used cars' prices using random forest," Future of Information and Communication Conference, 2018.

[3] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, J. Kevric, "Car Price Prediction using Machine Learning Techniques," 2021.

[4] S. Sinha, R. Azim, S. Das, "Linear Regression on Car Price Prediction," 2020.

[5] A. Fathalla, A. Salah, K. Li, K. Li, P. Francesco, "Deep end-to-end learning for price prediction of second-hand items," Knowl. Inf. Syst., 2020.

[6] Chetna Longania, Sai Prasad Potharaju, Sandhya Deore, "Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques," 2018.

[7] Aarone Steve J. Alexstan, Krishna M. Monesh, M. Poonkodi, Vineet Raj, "Used Car Price Prediction Using Machine Learning," 2020.

[8] K.Samruddhi, Dr. R.Ashok Kumar, "Used Car Price Prediction using K-Nearest Neighbor Based Model," 2020.

[9] N.Sun, H. Bai, Y. Geng, H. Shi, "Price evaluation model in second-hand car system based on BP neural network theory," 2018.

[10] S. Peerun, N. H. Chummun, S. Pudaruth, "Predicting the Price of Second-hand Cars using Artificial Neural Networks," The Second International Conference on Data Mining, Internet Computing, and Big Data, 2015.

[11] Lucija Bukvić, Jasmina Pašagić Škrinjar, Tomislav Fratrović, Borna Abramović, "Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning," 2020.