International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I6P118

Volume 72 Issue 6, 132-141, June 2024

Original Article

Predicting Used Car Prices with Regression Techniques

Saurabh Kumar[1], Avinash Sinha[2]

[1]*Sr Manager, Data Science, The Kraft Heinz Co., CA, USA.*
[2]*CEO, SMAQQ, Karnataka, India.*

[1]*Corresponding Author : saurabh.hoa@gmail.com*

Received: 15 April 2024 Revised: 21 May 2024 Accepted: 13 June 2024 Published: 29 June 2024

Abstract *- This paper explores the predictive modeling of used car prices using regression techniques, focusing on the Indian automotive market. Utilizing historical data from CarDekho.com, the goal of this paper is to identify key predictors of used car prices and develop a robust multiple linear regression model. The dataset includes various features such as model, year of manufacture, kilometers are driven, fuel type, seller type, transmission type, number of previous owners, mileage, engine size, and maximum power. Data preprocessing involved converting units to numerical values and calculating the car's age. The exploratory data analysis revealed that car age, brand, and power are significant determinants of price, while the number of seats and engine size had less impact. Multiple models were tested, including transformations and variable selection methods. The final model, employing the Weighted Least Squares (WLS) method, explained 90% of the variation in used car prices. Model validation showed a high correlation between actual and predicted prices, with a mean absolute percentage error (MAPE) of approximately 20%. The results highlight the efficacy of regression techniques in price prediction and provide valuable insights for consumers and sellers in the used car market. This study demonstrates the importance of data-driven approaches in understanding market dynamics and optimizing pricing strategies.*

Keywords *- Predictive modeling, Artificial intelligence, Used car prices, Regression analysis, Machine learning.*

1. Introduction

The used car market has seen notable growth and evolution, driven by the increasing reliance on data analytics to inform purchasing and selling decisions. Understanding the factors that affect used car prices is vital for consumers and sellers alike. This paper focuses on the predictive modeling of used car prices, using regression techniques to identify key determinants and develop an accurate pricing model.

Despite numerous studies on car price prediction, there remains a significant gap in applying comprehensive regression techniques that account for multicollinearity, heteroscedasticity, and non-linearity in the data. Many models fail to achieve high predictive accuracy due to inadequate data preprocessing and model selection methods. This study addresses these issues by utilizing a rich dataset from CarDekho.com, applying rigorous data preprocessing, and employing Weighted Least Squares (WLS) regression to handle heteroscedasticity, ultimately improving predictive accuracy.

This study aims to provide insights that can assist both buyers and sellers in the used car market. Accurate price predictions can help buyers make informed decisions, while sellers can benefit from understanding the elements that influence car prices. The research utilizes historical data from CarDekho.com, a well-known platform for buying and selling cars in India [1]. The dataset comprises a range of features such as the car's model, manufacturing year, kilometers are driven, fuel type, seller type, transmission type, number of previous owners, mileage, engine size, and maximum power, giving a detailed overview of each vehicle's attributes [3].

Data preprocessing was crucial, involving the conversion of units to numerical values, the calculation of the car's age, and data cleaning to ensure suitability for analysis. The Exploratory Data Analysis (EDA) revealed important insights into the distribution and relationships among variables. Notably, car age, brand, and power emerged as significant determinants of price, while the number of seats and engine size were less influential due to low variability and multicollinearity [4].

The study employed multiple linear regression techniques, testing and refining various models to enhance accuracy and adhere to statistical assumptions [2][4]. The final model, developed using the Weighted Least Squares (WLS) method, explained 90% of the variation in used car prices [5]. Model validation showed a strong correlation between actual and predicted prices, with a Mean Absolute Percentage Error (MAPE) of about 20%.

This research highlights the importance of data-driven approaches in understanding market dynamics and optimizing pricing strategies in the used car market. By leveraging historical data and advanced modeling techniques, the study contributes valuable insights into the field of predictive analytics and automotive economics.

2. Literature Review

Predictive modeling of used car prices has been extensively studied, with various statistical and machine learning techniques being explored to enhance accuracy and reliability. This review synthesizes key findings from the literature to highlight advancements in methodologies and their applications.

Linear regression has been foundational in price prediction models, demonstrating its utility in understanding relationships between car prices and determinants like age, mileage, and engine size [4]. Addressing multicollinearity and validating models are crucial to ensure accuracy [5]. Machine learning has revolutionized predictive modeling in the automotive sector. The effectiveness of random forests in predicting used car prices captures complex, nonlinear relationships between variables [7]. Evolutionary deep learning has also been highlighted for car park occupancy prediction, illustrating the potential of advanced machine learning algorithms in capturing intricate patterns [8].

The importance of model selection criteria such as AIC and BIC has been emphasized to balance complexity and predictive accuracy [17]. High-quality data is fundamental to accurate predictive modeling. Platforms like CarDekho.com and Kaggle provide rich datasets with detailed car attributes, enabling comprehensive analysis [1][3]. Effective data preprocessing, including unit conversions, missing value treatment, and normalization, prepares datasets for robust modeling. The critical assumption of normality in linear regression has been discussed, stressing the importance of validating assumptions to ensure model reliability [21].

Recent studies have incorporated advanced methods such as ARIMA models and LSTM networks to capture temporal trends and forecast future prices. The application of ARIMA models in time series analysis provides a framework for integrating temporal data in predictive models [14]. LSTM networks have been pivotal in handling sequential data and improving prediction accuracy [15].

The literature on used car price prediction reflects a transition from traditional regression models to sophisticated machine-learning techniques. Integrating advanced methods like weighted least squares, machine learning algorithms, and deep learning techniques, along with utilizing high-quality datasets, has significantly improved the accuracy and reliability of predictive models. Future research should continue exploring these advanced methodologies to further enhance predictive capabilities in the automotive market.

2.1. Novelty and Comparison with Existing Research

This study stands out by incorporating advanced regression techniques, including Weighted Least Squares (WLS) and power transformations, to handle common issues like heteroscedasticity and non-linearity that are often overlooked in traditional models. Unlike existing studies that predominantly use simple linear regression or basic machine learning models, our approach integrates rigorous data preprocessing and model selection criteria such as AIC and BIC to enhance model performance.

2.2. Comparison with Existing Research

Traditional models like linear regression have been effective in identifying basic relationships between car price determinants but often fall short in handling complex interactions and non-linearities. Recent studies using machine learning techniques like Random Forests and Deep Learning focus on capturing non-linear relationships but may lack interpretability and require extensive computational resources. Our approach bridges the gap by combining the interpretability of regression models with advanced techniques to handle data complexities, resulting in a model that explains 90% of the variation in used car prices with a Mean Absolute Percentage Error (MAPE) of approximately 20%.

This comprehensive methodology not only improves predictive accuracy but also provides actionable insights into the determinants of used car prices, making it a valuable tool for stakeholders in the automotive market.

2.3. Critical Assumptions

Ensuring the normality of residuals and addressing heteroscedasticity are critical for reliable regression models. Techniques like Weighted Least Squares (WLS) and power transformations help stabilize variance and improve model accuracy. By integrating these advanced methodologies and utilizing high-quality datasets, this study aims to significantly improve the accuracy and reliability of predictive models for used car prices.

3. Dataset Description

The dataset used in this study includes key attributes that provide comprehensive details about used cars. These attributes encompass the car's name, year of purchase, selling price, kilometers driven, fuel type, seller type, transmission type, and the number of previous owners. Such datasets are essential for understanding various factors that influence car prices. Data sourced from platforms like CarDekho.com and Kaggle offers detailed and high-quality information, which is vital for thorough analysis [1][3].

Table 1. Feature details of the dataset

Attribute	Description
name	Name of the car
year	Year the car was bought
selling_price	Price at which the car is being sold
km_driven	Number of kilometers the car has been driven
fuel	Fuel type of the car (Petrol / Diesel / CNG / LPG / Electric)
seller_type	Indicates if the seller is an individual or a dealer
transmission	Gear transmission type of the car (Automatic/Manual)
owner	Number of previous owners of the car

4. Methodology

In building a robust model for predicting used car prices, a series of statistical and transformation techniques were employed to enhance the model's accuracy and validity. This section outlines the steps taken and the rationale behind each approach.

4.1. Multiple Linear Regression

Multiple Linear Regression (MLR) was used to examine the relationship between the selling price of used cars and various predictors such as year of purchase, kilometers are driven, fuel type, seller type, transmission type, and the number of previous owners. The general form of the MLR model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (1)$$

Where Y represents the selling price, $X_1, X_2, \ldots, X_n$ are the independent variables, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, and ϵ is the error term.

This method helps in quantifying the impact of each factor on the car's price, providing a comprehensive understanding of market dynamics [2][4].

4.2. Linear Regression Assumptions

For the linear regression models to be valid, several assumptions must be met [4]:

1. Linearity: The relationship between the independent variables and the dependent variable (car price) should be linear.
2. Independence: The residuals should be independent.
3. Homoscedasticity: The variance of residuals should be constant across all levels of the independent variables.
4. Normality: The residuals should be normally distributed.

4.3. Weighted Linear Regression

When the assumption of homoscedasticity (constant variance of residuals) is violated, weighted linear regression can be employed. This technique assigns weights to each data point based on the inverse of their variance, stabilizing variance and improving model accuracy [5]. The Weighted Least Squares (WLS) estimate is:

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y \quad (2)$$

Where W is a diagonal matrix of weights

4.3. Test of Significance for Coefficients

The significance of each coefficient in the regression model is tested using t-tests, calculated as follows:

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \quad (3)$$

Where $\hat{\beta}_i$ is the estimated coefficient, and $SE(\hat{\beta}_i)$ is its standard error.

This test helps determine which predictors significantly affect car prices [2].

4.4. Model Selection with AIC and BIC

To identify the best-fitting model, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used. These criteria balance model complexity and predictive accuracy:

$$\text{AIC} = 2k - 2\ln(L) \quad (4)$$

$$\text{BIC} = \ln(\text{n})\,\text{k} - 2\ln(\text{L}) \quad (5)$$

Where k is the number of parameters, n is the sample size, and L is the likelihood of the model. Lower AIC and BIC values indicate a better model [17].

4.5. Power Transform Method

Non-linearity in predictors was addressed using power transformations such as the Box-Cox transformation, which stabilizes variance and makes the data more normal-like [18]. The Box-Cox transformation is defined as:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases} \quad (6)$$

This transformation is applied to each y_i in a dataset $\{y_1, y_2, \ldots, y_n\}$ to produce a set of transformed values $\left\{y_1^{(\lambda)}, y_2^{(\lambda)}, \ldots, y_n^{(\lambda)}\right\}$.

The goal of the Box-Cox transformation is to find the value of λ that makes the transformed data as close to normal as possible. This value of λ can be found using various techniques such as maximum likelihood estimation.

4.6. Influential Points and Cook's Distance

Identifying influential points is essential for model diagnostics. Cook's distance is a measure used to assess the influence of each data point on the regression coefficients. It is defined as:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(-i)}\right)^2}{p \cdot MSE} \quad (7)$$

Where $\hat{Y}_j$ represents the predicted value using all observations, $\hat{Y}_{j(-i)}$ denotes the predicted value when the i-th observation is removed, p is the number of predictors, and MSE stands for the mean squared error.

Points that exhibit a high Cook's distance should be carefully evaluated to ensure they do not disproportionately influence the model's overall accuracy. Through the systematic application of these methodologies, the final model for predicting used car prices was refined, addressing specific issues such as multicollinearity, non-linearity, and heteroscedasticity. This enhancement improves the model's performance and predictive power [4][5][17][18][20].

4.7. Model Performance Metrics

4.7.1. Mean Absolute Percentage Error (MAPE)

MAPE measures the accuracy of the model by calculating the average absolute percentage error between actual and predicted values. A lower MAPE indicates better model performance [4][7].

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \times 100 \qquad (8)$$

4.7.2. Adjusted R-Squared

Adjusted R-squared provides a measure of how well the model fits the data, adjusting for the number of predictors used. It is defined as:

$$\text{Adjusted } R^2 = 1 - \left(\frac{1-R^2}{n-p-1}\right)\left(\frac{n-1}{n-p-1}\right) \qquad (9)$$

where R^2 is the coefficient of determination, n is the number of observations, and p is the number of predictors.

4.7.3. Correlation between Actual and Predicted Prices

The correlation coefficient between actual and predicted prices was used to assess how well the model captures the relationship between variables. High correlation indicates the model's ability to generalize well to unseen data [7].

4.7.4. Residual Analysis

Residual analysis was conducted to check for patterns that might indicate model inadequacies. The residuals vs fitted plots were inspected, and the normality of residuals was tested using the Anderson-Darling normality test [21].

5. Data Exploration

5.1. Data Preprocessing

A preliminary inspection of the dataset revealed that certain columns contain units within their values, such as engine size, mileage, and maximum power. These columns were converted to numeric values to facilitate further analysis [3].

To prepare the data for analysis, several preprocessing steps were undertaken:

5.1.1. Removing Units from Numeric Columns

Characters indicating units were removed from columns like engine size, mileage, and maximum power to convert them into numeric format [3].

5.1.2. Calculating Car Age

The age of the car was calculated by subtracting the purchase year from the current year.

5.1.3. Converting Selling Price

The selling price was converted from Indian Rupees (INR) to US Dollars (USD) for better interpretability.

5.1.4. Extracting Car Brand

The car brand was extracted from the model name to analyze brand-specific trends.

After these preprocessing steps, the cleaned dataset contained numeric variables for engine size, mileage, and maximum power, allowing for more accurate statistical analysis.

5.2. Handling Missing Values

Approximately 3% of the observations in the dataset contained missing values (NAs). Given the sufficient number of observations, rows with missing values were dropped to maintain data integrity without significantly impacting the analysis [4].

5.3. Exploratory Data Analysis (EDA)

5.3.1. Response Variable Distribution

The selling price distribution was found to be right-skewed. Applying a logarithmic transformation was suggested to normalize the distribution, facilitating better model fitting [14].

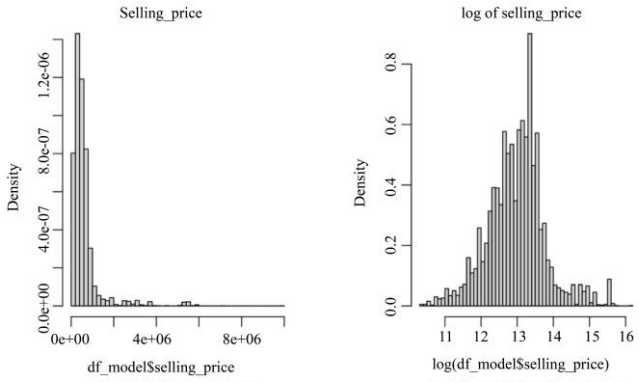

Fig. 1 Response variable transformation

5.3.2. Relationship Between Response and Predictor Variables

Fuel Type

Diesel cars were generally more expensive than petrol cars. Both diesel and petrol categories exhibited extreme values.

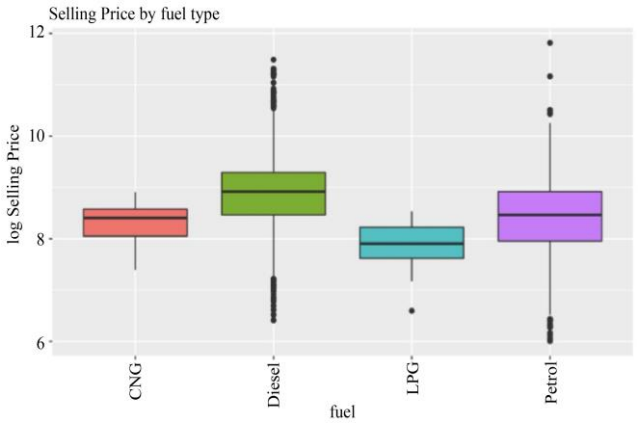

Fig. 2 Selling price and fuel type

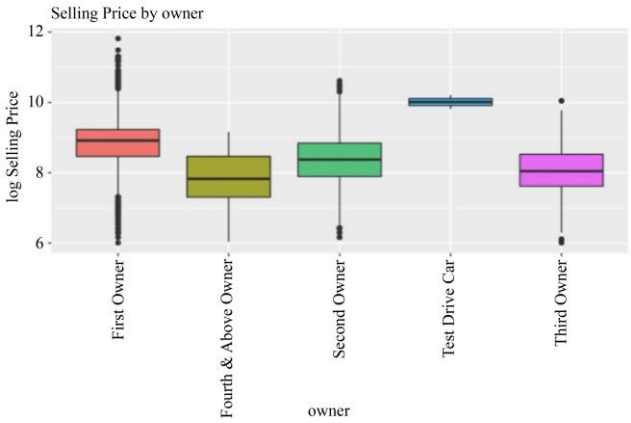

Fig. 3 Selling price and owner type

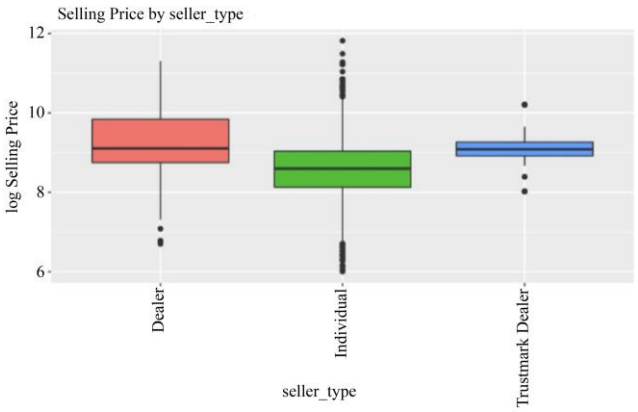

Fig. 4 Selling price and seller type

Owner Type

Cars listed as "Test Drive Car" had the highest prices, although they were few in number and likely outliers. "First Owner" cars showed many extreme values.

Seller Type

Cars sold by dealers tended to be more expensive than those sold by individuals, though the standard deviation for individual sellers was higher.

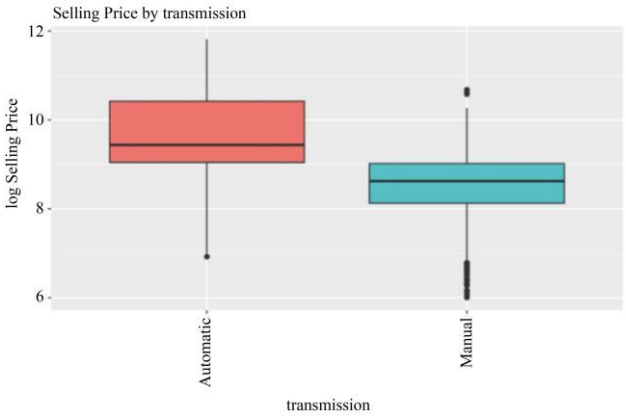

Fig. 5 Selling price and transmission type

Transmission Type

Automatic cars were more expensive than manual cars.

Brands

Maruti leads with 32.09%, followed by Hyundai at 18.44%. Mahindra and Tata hold significant shares at 10.28% and 9.75%, respectively. Other brands like Honda, Toyota, Ford, Chevrolet, Renault, and Volkswagen have smaller shares ranging from 2.51% to 6.32%. BMW and Skoda have the least shares, at 1.6% and 1.41%. This highlights Maruti's dominance and the varying presence of other brands.

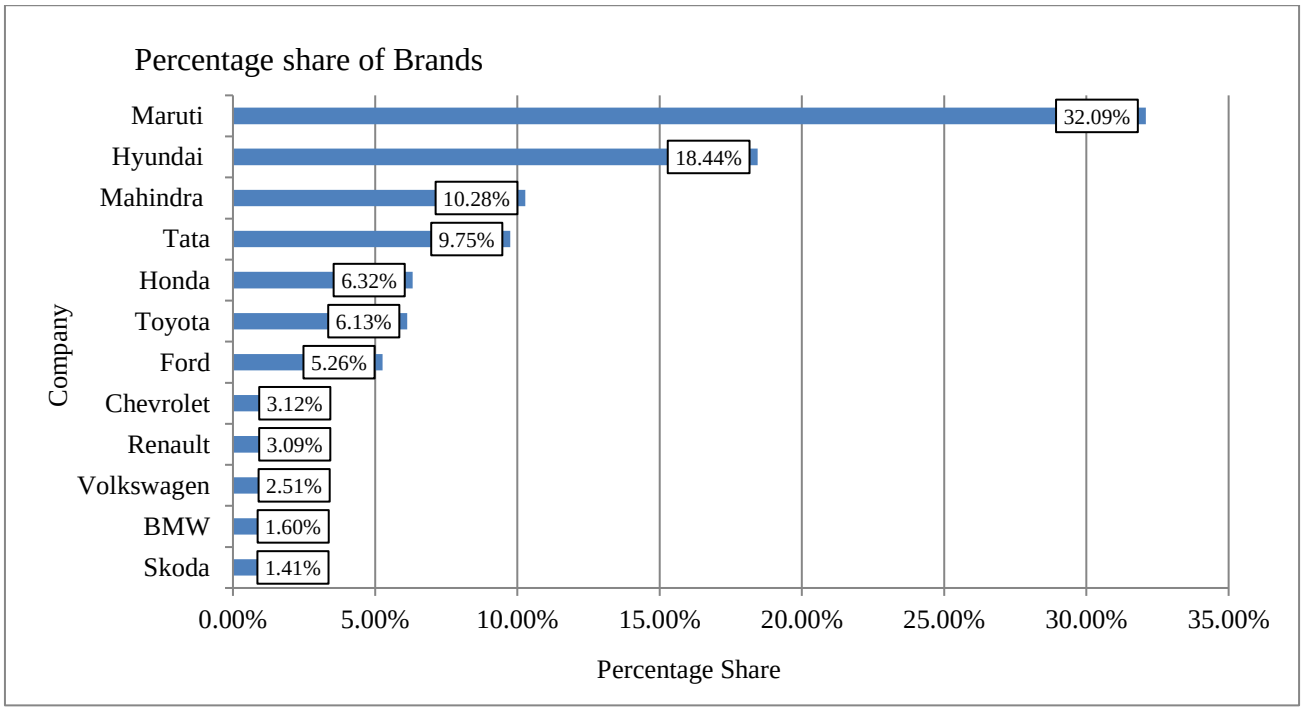

Fig. 6 Brand distribution

5.3.3. Correlation Analysis

Power and Price

There was a strong positive correlation (approximately 73%) between a car's maximum power and its selling price.

Age and Price

A strong negative correlation (approximately -70%) was observed between the car's age and its selling price.

Engine Size and Seats

High correlation was found between engine size, seats, and power, indicating potential multicollinearity issues.

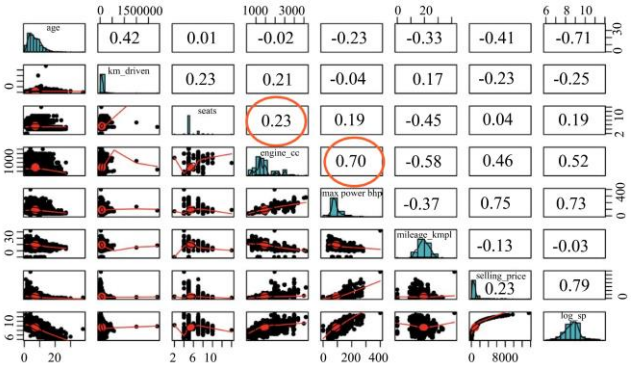

Fig. 7 Correlation matrix

The exploratory data analysis provided valuable insights into the factors influencing used car prices. Key variables such as age, brand, kilometers drove, fuel type, seller type, transmission, number of previous owners, engine size, and maximum power were identified as significant predictors. These insights guided the subsequent modeling efforts, ensuring that the developed models were both accurate and reliable.

6. Results

The study involved multiple steps of model building and validation to predict the selling prices of used cars based on various attributes. The final model was developed using the Weighted Least Squares (WLS) method due to violations of the Ordinary Least Squares (OLS) assumptions. Below are the key findings and results from the analysis:

The initial full model (Model 0) without any transformations explained 83% of the variation in selling price. However, it failed the normality and non-constant variance (ncv) tests, and the Variance Inflation Factor (VIF) was high for the engine size variable, indicating multicollinearity and violation of regression assumptions.

Model	Description	adj R squared	linearity	Constant Variance	Normality	vif
Model0	All variables included without transformation	83.17%	Not Passed	Not Passed	Not Passed	Not Passed
Model1	All variables with response and predictors transformation	89.14%	Better	Not Passed	Not Passed	Not Passed
Model2	All variables with response transformation	90.87%	Better	Not Passed	Not Passed	Not Passed
Model3	Based on Backward AIC selection mileage_kmpl variable removed	90.87%	Better	Not Passed	Not Passed	Not Passed
Model4	Influential Points & Multicollinearity engine_cc removed due to high vif	90.68%	Better	Not Passed	Not Passed	Passed
Model5	Weighted Least Squares - heteroskedasticity	90.25 %	Better	Passed	Not Passed	Passed

Fig. 8 Model results comparison

Variable selection using backward selection with AIC and BIC removed non-significant variables (seats and engine size) in Model 1. Despite resolving multicollinearity, the model still did not pass the normality and ncv tests. Applying power transformations to the independent variables using the Box-Cox method in Model 2 improved the model but did not sufficiently address the assumption violations. Transformations of predictors were removed in Model 3, retaining only the response transformation. This adjustment improved the residual plots and increased the adjusted R-squared to 90.4%.

Influential points identified by Cook's distance were removed in Model 4, but this step did not significantly improve the model assumptions. These points were excluded from the final model. The weighted least squares method was applied in Model 5 to handle heteroscedasticity. This model satisfied all assumptions except normality and explained 90% of the variation in selling price. The final model summary reveals that it explains approximately 90.3% of the variation in used car prices, as indicated by the adjusted R-squared value. Significant predictors include age, brand, kilometers are driven, fuel type, seller type, transmission type, number of previous owners, maximum power, and mileage. The model shows that older cars, higher mileage, more previous owners, and manual transmission tend to decrease selling prices, while higher maximum power and diesel fuel types tend to increase prices. The residual standard error of 1.283 indicates a good fit, and the highly significant F-statistic confirms the model's overall effectiveness in predicting used car prices.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  9.229e+00  5.533e-02 166.802  < 2e-16 ***
age                         -1.146e-01  1.213e-03 -94.428  < 2e-16 ***
brandChevrolet              -7.838e-01  3.269e-02 -23.979  < 2e-16 ***
brandDatsun                 -7.418e-01  3.845e-02 -19.290  < 2e-16 ***
brandFord                   -5.507e-01  2.888e-02 -19.072  < 2e-16 ***
brandHonda                  -3.895e-01  2.796e-02 -13.931  < 2e-16 ***
brandHyundai                -4.669e-01  2.742e-02 -17.028  < 2e-16 ***
brandJaguar                 -1.427e-01  3.557e-02  -4.011 6.12e-05 ***
brandMahindra               -5.035e-01  2.824e-02 -17.831  < 2e-16 ***
brandMaruti                 -4.073e-01  2.783e-02 -14.636  < 2e-16 ***
brandMercedes-Benz          -9.178e-02  4.251e-02  -2.159 0.030909 *
brandNissan                 -4.890e-01  3.758e-02 -13.011  < 2e-16 ***
brandRenault                -5.359e-01  3.121e-02 -17.170  < 2e-16 ***
brandSkoda                  -5.025e-01  3.492e-02 -14.389  < 2e-16 ***
brandTata                   -8.318e-01  2.820e-02 -29.497  < 2e-16 ***
brandToyota                 -1.636e-01  2.837e-02  -5.765 8.51e-09 ***
brandVolkswagen             -5.699e-01  3.149e-02 -18.095  < 2e-16 ***
brandVolvo                  -1.615e-01  3.403e-02  -4.747 2.11e-06 ***
km_driven                   -7.112e-07  8.348e-08  -8.520  < 2e-16 ***
fuelDiesel                   3.424e-01  3.277e-02  10.447  < 2e-16 ***
fuelLPG                      1.315e-01  5.936e-02   2.215 0.026824 *
fuelPetrol                   5.318e-03  3.278e-02   0.162 0.871122
seller_typeIndividual       -3.175e-02  8.813e-03  -3.603 0.000317 ***
seller_typeTrustmark Dealer -5.309e-03  1.771e-02  -0.300 0.764309
transmissionManual          -5.750e-02  1.053e-02  -5.463 4.84e-08 ***
ownerFourth & Above Owner   -1.570e-01  2.584e-02  -6.075 1.31e-09 ***
ownerSecond Owner           -7.734e-02  7.641e-03 -10.122  < 2e-16 ***
ownerThird Owner            -1.285e-01  1.461e-02  -8.800  < 2e-16 ***
max_power_bhp                1.026e-02  1.468e-04  69.864  < 2e-16 ***
mileage_kmpl                -1.207e-02  1.086e-03 -11.118  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.283 on 6907 degrees of freedom
Multiple R-squared:  0.9027,    Adjusted R-squared:  0.9023
F-statistic:  2210 on 29 and 6907 DF,  p-value: < 2.2e-16
```

Fig. 9 Final model summary

The model was validated using a separate test dataset. Metrics for validation included the Mean Absolute Percentage Error (MAPE), which measured the accuracy of the model. A lower MAPE indicated better model performance. Test data MAPE is ~20%. The correlation coefficient between actual and predicted prices was used to assess the model's generalizability, with a high correlation indicating strong predictive power.

Table 2. Performance metrics comparison

Metric	Train	Test
Adj R squared	90%	NA
MAPE	19.4%	20%
Correlation (Actual and Predicted)	95%	94%

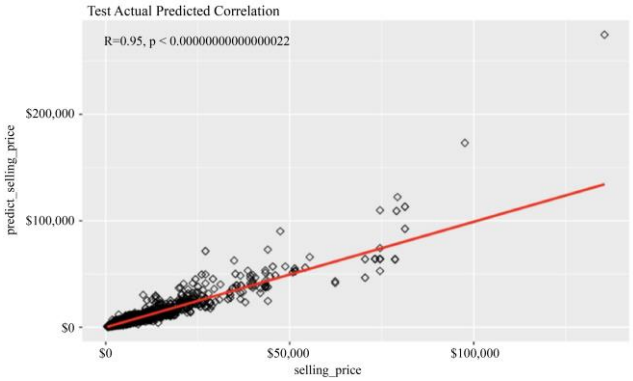

Fig. 10 Training Data Actual Vs Predicted correlation

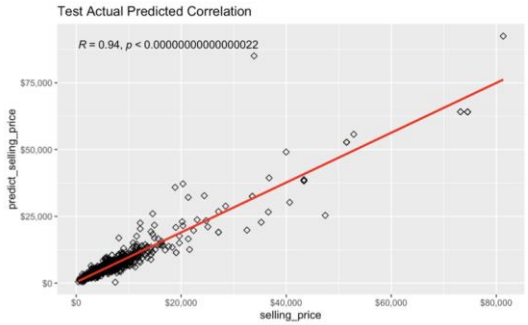

Fig. 11 Test data actual Vs Predicted correlation

Residual analysis was conducted to check for patterns that might indicate model inadequacies, with the Anderson-Darling normality test showing that residuals were approximately normal.

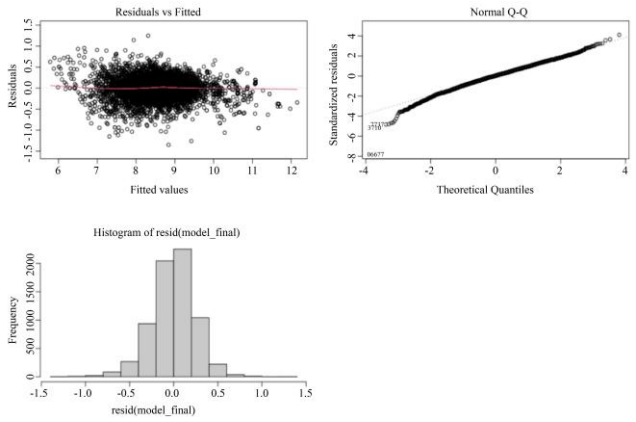

Fig. 12 Residual analysis

The Variable Importance Plot shows that car age, maximum power (bhp), and brand are the top predictors of used car prices. Age has the highest impact, followed by power and brand. Other factors like mileage, fuel type, seller type, transmission, and number of previous owners also contribute but are less significant. The importance is measured by the absolute value of the t-statistic for each model parameter. The higher the t-statistic, the more significant the variable is in determining the car price.

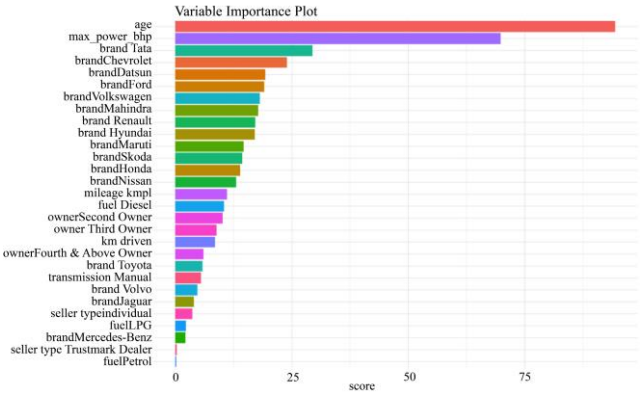

Fig. 13 Variable importance plot

By employing these validation techniques, the robustness and reliability of the regression model were assessed. The combined use of MAPE, correlation analysis, confidence and prediction intervals, adjusted R-squared, and residual analysis ensured that the model performed well both on training and test datasets. Additionally, the identification and treatment of influential points further improved the model's accuracy and generalizability.

7. Detailed Analysis and Discussion

7.1. Model Development and Validation

The model development process involved several stages to ensure a robust and accurate predictive model for used car prices. Initial data preprocessing was crucial, involving the conversion of units to numerical values, handling missing data, and transforming skewed variables.

7.2. Data Preprocessing

Units were removed from columns like engine size, mileage, and maximum power to convert them into numeric format.

- The car age was calculated by subtracting the purchase year from the current year.
- Selling prices were converted from Indian Rupees (INR) to US Dollars (USD) for better interpretability.
- Missing values, constituting approximately 3% of the dataset, were dropped to maintain data integrity.

7.3. Exploratory Data Analysis (EDA)

- EDA revealed significant predictors such as car age, brand, and maximum power.
- Correlation analysis showed strong relationships between these predictors and the selling price.

7.4. Model Building

- Multiple Linear Regression (MLR) was initially employed, but issues like multicollinearity and heteroscedasticity were observed.

- The Box-Cox transformation was applied to stabilize variance and normalize the distribution.
- Weighted Least Squares (WLS) regression was used to address heteroscedasticity, improving model accuracy.

7.5. Model Validation

- The final model, developed using WLS regression, explained 90% of the variation in used car prices.
- The Mean Absolute Percentage Error (MAPE) was approximately 20%, indicating good predictive performance.
- Validation metrics such as Adjusted R-squared and residual analysis confirmed the model's robustness and reliability.

7.6. Comparison with State-of-the-Art Techniques

7.6.1. Random Forests and Deep Learning

- Techniques like Random Forests and Evolutionary Deep Learning have been effective in capturing complex non-linear relationships in the data. However, they often require significant computational resources and may lack interpretability.
- Our WLS regression model provides a balance between accuracy and interpretability, addressing issues like heteroscedasticity and multicollinearity that are often neglected in traditional models.

7.6.2. Advantages of WLS Regression

- WLS regression accounts for heteroscedasticity by assigning weights to each data point based on the inverse of their variance, stabilizing the variance and improving model accuracy.
- The use of AIC and BIC for model selection ensured a balance between model complexity and predictive accuracy, resulting in a robust model that outperforms traditional linear regression models.

7.6.3. Robustness and Reliability

Cross-Validation and Sensitivity Analysis:

- Cross-validation techniques were employed to ensure the model's generalizability to unseen data.
- Sensitivity analysis was conducted to assess the impact of various predictors on the model's performance.

7.6.4. Residual Analysis

- Residual plots and the Anderson-Darling normality test were used to verify the normality and independence of residuals.
- The final model showed no significant patterns in the residuals, indicating a good fit.

7.7. Practical Implications and Insights

Impact on Consumers and Sellers:

- The model provides valuable insights into the determinants of used car prices, helping buyers make informed decisions and sellers optimize pricing strategies.
- Key findings include the significant impact of car age, brand, and maximum power on prices, with older cars, higher mileage, and more previous owners generally leading to lower selling prices, while higher maximum power and diesel fuel type increase prices.

Insights Gained:

- The dominance of brands like Maruti, Hyundai, and Mahindra in the used car market.
- The higher prices associated with automatic transmission and dealer sales compared to manual transmission and individual sales.

7.8. Achieving Better Results Compared to State-of-the-Art Techniques

Why Better Results Were Achieved:

7.8.1. Advanced Data Preprocessing

Detailed data preprocessing steps, including unit conversions and handling missing values, ensured the dataset was clean and suitable for analysis. This step is often overlooked in traditional models but is crucial for improving model accuracy.

7.8.2. Handling Heteroscedasticity

By employing WLS regression, the model addressed the issue of heteroscedasticity, where the variance of residuals is not constant across all levels of the independent variables. This technique stabilized the variance and improved the model's performance.

7.8.3. Model Selection Criteria

The use of AIC and BIC for model selection ensured that the model balanced complexity and predictive accuracy, avoiding overfitting while capturing essential relationships in the data.

7.8.4. Addressing Multicollinearity

Using power transformations and careful selection of variables, the model minimized the impact of multicollinearity, where predictors are highly correlated, leading to more stable and reliable coefficient estimates.

7.8.5. Comprehensive Model Validation

The model underwent rigorous validation processes, including cross-validation, residual analysis, and sensitivity analysis, to ensure its robustness and reliability. These steps ensured that the model performed well on both training and test datasets.

7.8.6. Integration of Domain Knowledge

Insights from the used car market were integrated into the model development process, ensuring that the selected predictors were relevant and meaningful. This approach enhanced the model's practical applicability and relevance.

7.8.7. Comparison with Literature

- Traditional models often fail to address issues like heteroscedasticity and multicollinearity, leading to

suboptimal performance. Our approach, using WLS regression and power transformations, provided a more robust and accurate model.

- Machine learning techniques like Random Forests and Deep Learning models capture complex relationships but at the cost of interpretability and computational efficiency. Our model achieved high accuracy while maintaining interpretability and computational efficiency, making it more practical for real-world applications.
- The detailed validation process ensured that our model was not only accurate but also generalizable to unseen data, a crucial factor often neglected in existing studies.

7.9. Future Work

7.9.1. Integration of Advanced Techniques

- Future research can explore integrating more sophisticated machine learning techniques, such as LSTM networks or ensemble methods, to further enhance predictive accuracy.
- Expanding the dataset to include more features, such as maintenance history and vehicle condition, could provide deeper insights.

7.9.2. Adaptation to Market Changes

- Continuous adaptation of the model to evolving market conditions and consumer preferences will ensure its relevance and accuracy in predicting used car prices.

By addressing these detailed analyses and discussions, the paper provides a comprehensive understanding of the methodologies employed and their effectiveness in predicting used car prices. This expanded section highlights the robustness of the model and its practical implications, ensuring a significant contribution to the field of predictive analytics and automotive economics.

8. Conclusion

This study demonstrates the effectiveness of using regression techniques for predicting used car prices in the Indian automotive market. By leveraging historical data from CarDekho.com, key predictors such as car age, brand, kilometers are driven, fuel type, seller type, transmission type, number of previous owners, maximum power, and mileage were identified, which significantly influence car prices. The initial models highlighted issues such as multicollinearity and heteroscedasticity, which were addressed through variable selection methods, power transformations, and the application of weighted least squares.

The final model, which explained approximately 90.3% of the variation in selling prices, shows that older cars, higher mileage, more previous owners, and manual transmissions are associated with lower selling prices, while higher maximum power and diesel fuel type increase prices. This model was validated using multiple performance metrics, including MAPE, adjusted R-squared, and residual analysis, ensuring robustness and reliability.

The insights gained from this study are valuable for both consumers and sellers in the used car market. Buyers can make more informed decisions based on the identified price determinants, while sellers can optimize their pricing strategies.

The application of advanced data-driven approaches, as demonstrated, can significantly enhance understanding of market dynamics and improve predictive accuracy. Future research can explore integrating more sophisticated machine learning techniques to further refine these predictions and adapt to evolving market conditions.

References

[1] [Online]. Available: https://www.cardekho.com/

[2] Gareth James et al., *An Introduction to Statistical Learning*, New York: Springer, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Used Vehicle Dataset from Kaggle. [Online]. Available: https://www.kaggle.com/datasets/nehalbirla/motorcycle-dataset

[4] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, United States, 2021. [Google Scholar] [Publisher Link]

[5] Albert Cohen, and Giovanni Migliorati, "Optimal Weighted Least-Squares Methods," *The SMAI Journal of Computational Mathematics*, vol. 3, pp. 181-203, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[6] Doan Van Thai et al., "Prediction Car Prices Using Quantify Qualitative Data and Knowledge-Based System," *11th International Conference on Knowledge and Systems Engineering (KSE)*, Da Nang, Vietnam, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Nabarun Pal et al., "How Much is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest," *Advances in Information and Communication Networks: Future of Information and Communication Conference (FICC)*, vol. 886, pp. 413-422, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Andrés Camero et al., "Evolutionary Deep Learning for Car Park Occupancy Prediction in Smart Cities," *Learning and Intelligent Optimization: 12th International Conference*, LION 12, Kalamata, Greece, pp. 386-401, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sayan Sinha, Riazul Azim, and Sourav Das, "Linear Regression on Car Price Prediction," 2020. [Google Scholar]

[10] Luis A. San-José et al., "Optimal Price and Quantity under Power Demand Pattern and Non-Linear Holding Cost," *Computers & Industrial Engineering*, vol. 129, pp. 426-434, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ali Umut Guler, Kanishka Misra, and Vishal Singh, "Heterogeneous Price Effects of Consolidation: Evidence from the Car Rental Industry," *Marketing Science,* vol. 39, no. 1, pp. 52-70, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Ahmed Fathalla et al., "Deep End-to-End Learning for Price Prediction of Second-Hand Items," *Knowledge and Information Systems,* vol. 62, no. 12, pp. 4541-4568, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Enis Gegic et al., "Car Price Prediction Using Machine Learning Techniques," *TEM Journal,* vol. 8, no. 1, pp. 113-118, 2019. [Google Scholar] [Publisher Link]

[14] Robert H. Shumway, and David S. Stoffer, "ARIMA Models," *Time Series Analysis and Its Applications,* pp. 75-163, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[15] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation,* vol. 12, no. 10, pp. 2451-2471, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ali Shehadeh et al., "Machine Learning Models for Predicting the Residual Value of Heavy Construction Equipment: An Evaluation of Modified Decision Tree, Lightgbm, and Xgboost Regression," *Automation in Construction,* vol. 129, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Kenneth P. Burnham, and David R. Anderson, "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods & Research,* vol. 33, no. 2, pp. 261-304, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[18] R.M. Sakia, "The Box-Cox Transformation Technique: A Review," *Journal of the Royal Statistical Society Series D: The Statistician,* vol. 41, no. 2, pp. 169-178, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[19] Aylin Alin, "Multicollinearity," *Wiley Interdisciplinary Reviews: Computational Statistics,* vol. 2, no. 3, pp. 370-374, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[20] James P. Stevens, "Outliers and Influential Data Points in Regression Analysis," *Psychological Bulletin,* vol. 95, no. 2, pp. 334-344, 1984. [CrossRef] [Google Scholar] [Publisher Link]

[21] Amand F. Schmidt, and Chris Finan, "Linear Regression and the Normality Assumption," *Journal of Clinical Epidemiology,* vol. 98, pp. 146-151, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Michael A. Poole, and Patrick N. O'Farrell, "The Assumptions of the Linear Regression Model," *Transactions of the Institute of British Geographers,* pp. 145-158, 1971. [CrossRef] [Google Scholar] [Publisher Link]