Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 2, February : 2024

ENHANCING PREDICTIVE MODELS FOR CAR PRICE ESTIMATION: A COMPREHENSIVE COMPARATIVE STUDY

Raavi Hemalatha, Karunasri Adina, Khaja Javeed Shaik, Sai Srikanta U, Jahnavi Satya Lakshmi S, D V Sasidhar S, Vishnu Institute of Technology

***Abstract*-- The automotive industry recognizes the profound influence of pricing on consumer behaviour and market dynamics. Accurate car price predictions not only ensure a competitive edge for manufacturers and dealers but also enhance transparency and trust for consumers in their purchase decisions. This research is devoted to the development of a sophisticated regression model to predict car prices [3]. Drawing on an extensive dataset encompassing diverse car attributes, the study employs an array of machine learning techniques, including linear regression, random forest, gradient boosting, and decision tree regression. The efficacy of each model is rigorously assessed through metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R- squared values. This comprehensive approach not only delivers actionable insights for stakeholders in the automotive industry [1] but also sets a benchmark for future research in predictive analytics within the sector.**

***Index Terms*- Car Price Prediction, Automotive Industry, Machine Learning, Regression Models, Model Evaluation, Linear Regression, Random Forest, Evaluation Metrics.**

I. INTRODUCTION

The automotive industry is an integral pillar of global economies [2], driving both technological advancements and consumer behaviour patterns. As the industry evolves, car pricing emerges as a crucial determinant that bridges manufacturer strategies with consumer decisions. In the contemporary automotive market, marked by rapid technological shifts and heightened competition, predicting car prices accurately has become more than just a business strategy it's pivotal for market sustenance and growth.

Understanding the myriad factors influencing car prices is essential for stakeholders ranging from manufacturers to consumers. A predictive model that encapsulates these factors offers a systematic approach to price determination, allowing manufacturers to align their offerings with market expectations, while consumers benefit from transparent and data-backed pricing. The Car Price Prediction dataset, central to this study, offers a rich tapestry of car attributes and their corresponding market prices. Such a dataset not only sheds light on prevailing market trends but also paves the way for the development of sophisticated predictive algorithms.

Harnessing this dataset, the objective is to delve deep into the intricacies of car pricing by leveraging a suite of regression algorithms. Regression models, as supervised learning techniques, are adept at predicting continuous outputs based on input features. Common regression techniques such as Linear Regression, Random Forest, Gradient Boosting, Lasso regression and Ridge regression, each come with their unique strengths and nuances. A comparative analysis of these algorithms, underpinned by rigorous evaluation metrics, offers a holistic view of their predictive capabilities.

In essence, this research revolves around a regression problem, aiming to predict car prices based on a plethora of features. A well-calibrated model holds the potential to revolutionize price determination in the automotive industry, fostering market efficiency, consumer trust, and informed decision-making.

II. EXPLORATORY DATA ANALYSIS

Understanding the data is an essential and crucial initial step in any machine-learning project [5], especially in the context of predicting used car prices. The deeper and more thorough our exploration of the data, the higher the quality of results we can anticipate. This depth of understanding is cultivated through rigorous data analysis. Through this process, we can extract vital information about the factors influencing car resale values, aiding us in making informed decisions throughout the research. The

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 2, February : 2024

dataset at hand presents a range of columns that detail specific car attributes [6]. Each of these attributes provides crucial insights into the potential resale value of a vehicle. Data cleaning, as always, remains a pivotal step. It empowers us to address missing values, eliminate duplicate entries, and filter out outliers to ensure the integrity of our dataset.

```
df.isnull().sum()

ID                  0
Levy                0
Manufacturer        0
Model               0
Prod. year          0
Category            0
Leather interior    0
Fuel type           0
Engine volume       0
Mileage             0
Cylinders           0
Gear box type       0
Drive wheels        0
Doors               0
Wheel               0
Color               0
Airbags             0
Price               0
dtype: int64
```

Figure 1: Null values in the dataset

In this work, pertaining to used car price prediction, one of the prominent findings was the absence of missing values in the dataset, as depicted in Figure 1. This is a commendable trait, especially given the breadth of attributes such as 'ID', 'Levy', 'Manufacturer', 'Model', and so forth [9]. The absence of missing values streamlined the preliminary stages of our research, allowing us to transition seamlessly to more advanced phases of data analysis without the need for imputation or other data rectification techniques. Various visualization techniques are used to represent the data visually [6].

```
# Histogram for numerical columns
numerical_columns = ['Prod. year', 'Cylinders', 'Airbags', 'Price']
for column in numerical_columns:
    plt.figure(figsize=(8, 6))
    sns.histplot(df[column], bins=20, kde=True)
    plt.xlabel(column)
    plt.ylabel('Frequency')
    plt.title(f'Histogram of {column}')
    plt.show()
```

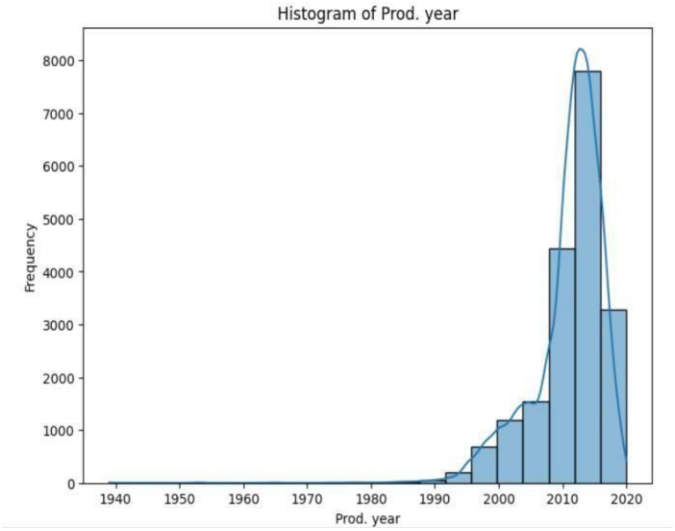

Figure 2: Histogram of Production Years

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 2, February : 2024
Visualizations such as histograms, scatter plots, or bar charts provide meaningful insights and make complex data more understandable and interpretable.

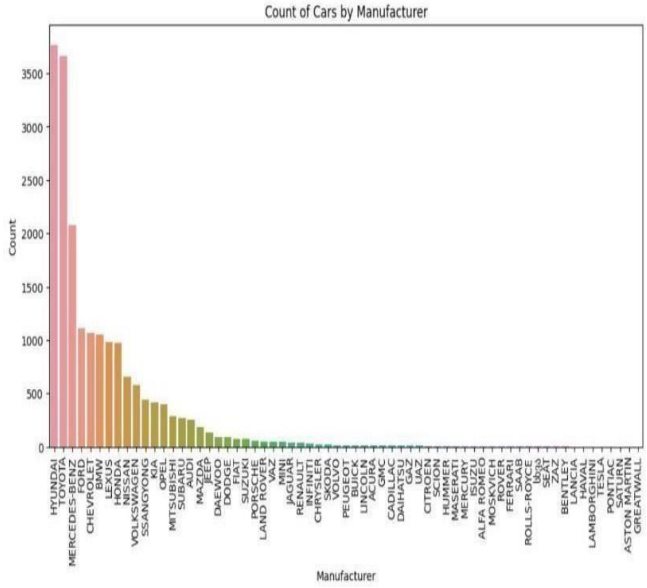

Figure 3: Count of cars by Manufacturer

III. METHODOLOGY

The dataset central to this research was procured from Kaggle, encompassing detailed attributes related to used cars. This data, instrumental in predicting used car prices, has been meticulously collated, detailing attributes such as 'ID', 'Levy', 'Manufacturer', 'Model', 'Prod. year', among others. To ensure the data's viability for regression analysis, a series of preprocessing steps were judiciously executed. These encompassed the transformation of categorical variables into numerical formats through techniques like one-hot encoding or label encoding.

Data visualization was employed to fathom the intricate nuances within the dataset. Visual tools, including histograms, bar plots, and heatmaps, shed light on the distribution of car prices in relation to variables like 'Manufacturer', 'Model', 'Fuel type', and so forth. These visualizations furnished a lucid understanding of potential relationships between the variables and the target attribute, 'Price'.

Transitioning to the predictive modelling phase, we leveraged a spectrum of regression algorithms to construct our price prediction models. This suite included Linear Regression, Lasso Regression, Decision Tree Regression [8], Random Forest
Regression, and Ridge Regression [4] as exemplified in Figure
4.

The 'Price' attribute was used as the target variable when each model was painstakingly trained on the dataset. After training, we evaluated each model's performance using important assessment criteria designed for regression tasks. These measurements included the Coefficient of Determination (R2) [7], Mean Squared Error (MSE), and Mean Absolute Error (MAE). We were able to determine which model was most accurate in forecasting used vehicle prices using these variables, and we chose the model with the best predictive ability.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 2, February : 2024

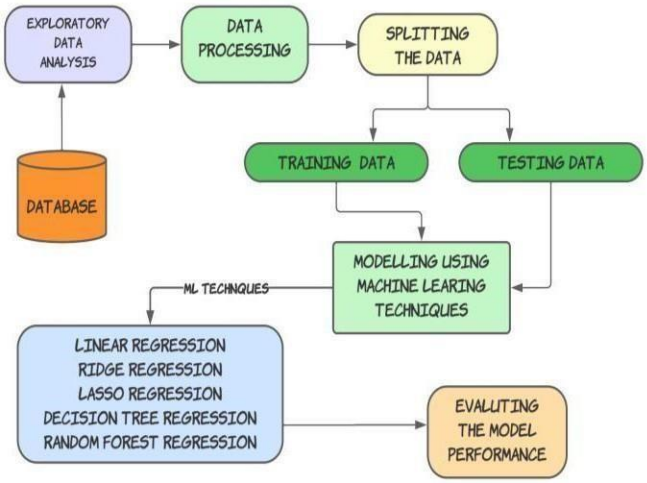

Figure 4: Workflow of Model

IV. MODELLING

Linear Regression:

Linear Regression is a foundational regression algorithm, primed to predict a continuous result variable depending on one or more predictor factors, such as car price. By assuming that the predictors and the result have a linear relationship, it models the relationship using a linear equation. In the context of predicting used car prices, the model assigns a weight to each feature, and by summing these weighted features and adding a bias term, it predicts the car's price.

Linear Regression establishes a connection between the target variable (Y) and one or multiple predictor variables (X) through a linear equation in the following form: **Y = β0 + β1X1 + β2X2 + ... + βnXn + ε Ridge Regression:**

An enhancement to the standard linear regression, Ridge

Regression introduces a regularization term. The added L2 regularization aims to prevent overfitting, especially in scenarios where multicollinearity exists among predictor variables. By controlling the magnitude of coefficients, Ridge Regression ensures the model retains its generalization capabilities across unseen data.

Ridge Regression extends Linear Regression by adding an L2 regularization term:

Y = β0 + β1X1 + β2X2 + ... + βnXn + λΣ(βi^2) + ε Decision Tree Regression:

Decision Tree Regression operates by segmenting the predictor space into distinct and non-overlapping regions. For a given predictor, the algorithm determines a value such that the resulting split minimizes the sum of squared differences of the target variable in the resultant regions. Each split is chosen to best segregate the car prices based on feature values. In our context, this could mean decisions based on attributes like 'Manufacturer', 'Mileage', or 'Fuel type'.

Random Forest Regression:

A 'forest' of decision trees are built during training via the ensemble learning technique known as Random Forest Regression. By employing bootstrapping and feature randomness when constructing each tree, it ensures model diversity. The final car price prediction is the averaged result from all individual trees. This ensemble approach aims to reduce the variance, increase the model's robustness, and prevent overfitting.

Lasso Regression:

Lasso Regression, short for "Least Absolute Shrinkage and Selection Operator," is a variant of linear regression designed for predicting continuous outcomes, such as used car prices. It incorporates L1 regularization, which not only prevents overfitting but also performs automatic feature selection, making it a valuable tool for simplifying models and identifying essential predictor variables.

Lasso Regression adds L1 regularization to Linear Regression:

Y = β0 + β1X1 + β2X2 + ... + βnXn + λΣ|βi| + ε

V. EVALUTION METRICS

Regression models utilize various evaluation metrics to assess their performance when applied to a dataset. These metrics offer valuable insights into different facets of the model's predictive
Based on the classification report, it's clear that Random Forest Regression surpasses the other regression models in performance. It attains the lowest values for MAE, MSE, and

VI. RESULTS

RMSE, underscoring its superior predictive accuracy. Furthermore, it boasts the highest R-squared (R^2) value of 0.720, which indicates a more optimal model fit to the dataset. accuracy. Some of the commonly employed regression evaluation techniques include:

(1) Mean Absolute Error (MAE)
(2) Mean Squared Error (MSE)
(3) Root Mean Squared Error (RMSE)
(4) Coefficient of Determination (R2)

We used a variety of regression measures to make sure our prediction models were accurate and reliable in projecting used automobile pricing. Each indicator offers a unique perspective from which the model's effectiveness can be evaluated, taking into account the complexities of continuous value prediction.

(1) Mean Absolute Error (MAE) MAE computes the average values of errors between the predicted car prices and actual car prices, without considering their direction. In the context of used car price prediction, MAE gives a direct insight into how much, on average, our predictions deviate from the actual prices in absolute terms. A lower MAE suggests better model performance

MAE = (1 / n) * Σ|Actual - Predicted|

(2) Mean Squared Error (MSE) MSE measures the average squared discrepancies between the prices of actual and projected cars. By squaring the errors, MSE amplifies the effect of larger errors, thus providing a more sensitive metric especially when larger deviations from actual prices are undesirable.

MSE = (1 / n) * Σ(Actual - Predicted)^2

(3) Root Mean Squared Error (RMSE) RMSE is the square root of Mean Square Error (MSE), and it is used to evaluate the performance of a predictive model. RMSE is particularly useful because it returns the error metric in the same units as the target variable, making it easier to interpret and understand the prediction accuracy

RMSE = √(MSE)

(4) R-squared (R^2) R-squared, also known as the coefficient of determination, quantifies the proportion of the variance in the target variable that can be explained by the predictor variables used in a statistical model [7]. It typically falls in the range of 0 to 1, where higher values signify a stronger fit of the model to the data, indicating that the predictor variables are more effective in explaining the variability in the target variable.

R^2 = 1 - (Σ(Actual - Predicted)^2) / (Σ(Actual - Mean)^2)

In conclusion, for this specific tourism review dataset, Random Forest Regression is the most accurate and effective regression technique for sentiment analysis and prediction.
The below table shows the performance of all the algorithms based on the classification report.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 2, February : 2024

Regression Technique	Mean Absolute Error (MAE) Value	Mean Squared Error (MSE) Value	Root Mean Squared Error (RMSE) Value	R-Squared (R^2) Value
Linear Regression	14456.94	1605202723.62	40064.98	-3.26
Ridge Regression	13607.89	1475559403.89	38413.01	-2.92
Lasso Regression	13796.72	1570769724.53	39632.94 1	-3.17
Decision Tree Regression	4939.62	151058327.79	12290.58	0.599
Random Forest Regression	4016.91	105403697.43	10266.63	0.720

Table 1: Performance of the Model

VII. FUTURE SCOPE

The realm of car price prediction is poised for significant growth and innovation in the years ahead. Leveraging increasingly extensive and diverse datasets, including real-time market data and customer sentiment analysis, promises more accurate predictions. Emerging technologies like deep learning and natural language processing hold potential for deeper insights from unstructured data sources, enriching pricing dynamics understanding.

Personalized pricing strategies, tailoring prices to individual customer profiles, could enhance customer satisfaction. Moreover, predictive models could be used to analyze market trends comprehensively, assisting manufacturers in datadriven decision-making for production, inventory, and marketing.

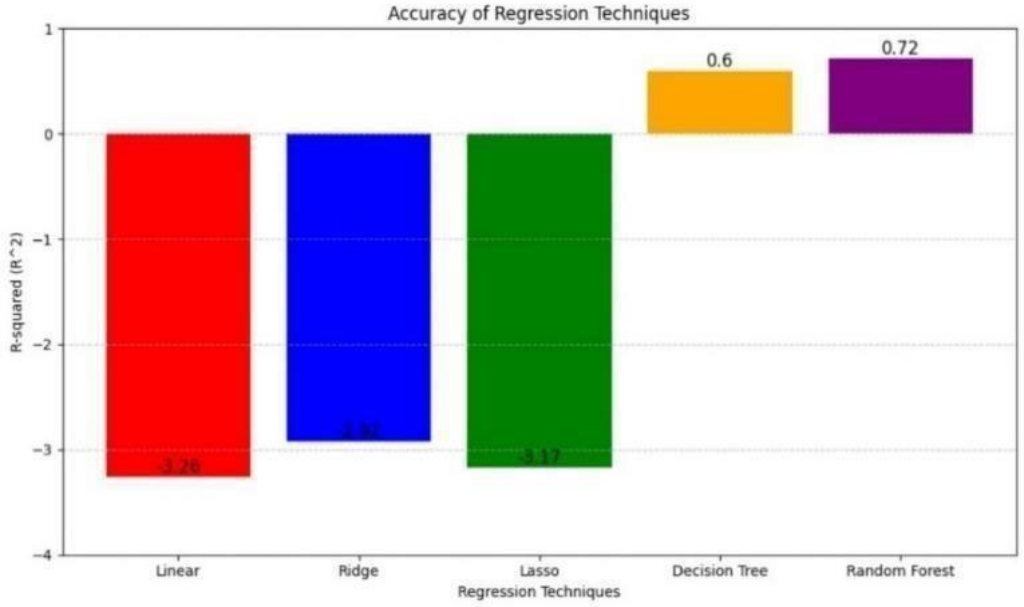

Figure 5: Accuracy of Regression Techniques

VIII. CONCLUSION

Among the regression techniques evaluated, Random Forest Regression demonstrates the highest accuracy in predicting the target variable. Its Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE) scores are the lowest, demonstrating that it makes predictions that are most accurate.

Hence, considering the unique characteristics of this dataset and the nature of the problem at hand, Random Forest Regression emerges as the optimal and precise choice for generating predictions.

REFERENCES

[1] Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed July 18, 2018.]

[2] Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).

[3] Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.

[4] Krupa, Jake and Minutti-Meza, Miguel, Regression and Machine Learning Methods to Predict Discrete Outcomes in Accounting Research (March 1, 2022). Journal of Financial Reporting, accepted, University of Miami Business School Research Paper No. 3801353, Available at SSRN: https://ssrn.com/abstract=3801353 or http://dx.doi.org/10.2139/ssrn.3801353

[5] L. Wilkinson. The Impact of Tukey's Exploratory Data Analysis, Chicago chapter of the American Statistical Association Spring Conference, May 5, 2000.

[6] L. Wilkinson, A. Anand, and R. Grossman. High dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. Visualization and Computer Graphics, IEEE

Transactions on, 12(6):1363-1372, 2006.Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27-31.

[7] Kvalseth, T.O., Cautionary Note about R2. The American Statistician, 1985. 39(4): p. 279-285 DOI:

https://doi.org/10.2307/2683704.

[8] Weka 3 - Data Mining with Open Source MachineLearning Software in Java. (n.d.), Retrieved from:https://www.cs.waikato.ac.nz/ml/weka/. [August 04, 2018].

[9] Used cars database. (n.d.) Retrieved from: https://www.kaggle.com/orgesleka/used-carsdatabase. [accessed: June 04, 2018].

[10] Silpa, N., Maheswara Rao VVR, M. Venkata Subbarao, M. Pradeep, Challa Ram Grandhi, and Adina Karunasri. "A Robust Team Building Recommendation System by Leveraging Personality Traits Through MBTI and Deep Learning Frameworks." In 2023 International Conference on IoT, Communication and Automation Technology (ICICAT), pp. 1-6. IEEE, 2023.

[11] Uppalapati PJ, Gontla BK, Gundu P, Hussain SM, Narasimharo K. A Machine Learning Approach to Identifying Phishing Websites: A Comparative Study of Classification Models and Ensemble Learning

Techniques. EAI Endorsed Scal Inf Syst [Internet]. 2023 Jun. 23 [cited 2023 Oct. 18];10(5). Available from: https://publications.eai.eu/index.php/sis/article/vie w/3300.

[12] Sunethra, B., Sreeya, C., Dhannushree, U., Nagaraj, P., Muthamil Sudar, K., Muneeswaran, V., A Systematic Parking

System Using bi-class Machine Learning

Techniques, 2022, June, 221, 226, 10.1109/ICSCDS53736.2022.9760903.

[13] Sudarshan E., Kumari D.A., Reddy Y.C.A.P., Balasundaram A., Mahender K., Machine learning based automatic vehicle alert system 2022, 10.1063/5.0081741,

https://www.scopus.com/inward/record.uri?eid=2-s2.0
85131855508&doi=10.1063%2f5.0081741&partnerID=
40&md5=4076587970d1817e8529729f517cd4f7.
[14] Manikanta Sirigineedi, M.Srikanth, Padma Bellapukonda, The Early Detection Of Alzheimer's Illness Using Machine Learning And Deep Learning
Algorithms, 2022, 10.47750/pnr.2022.13.S09.603, Journal of Pharmaceutical NegativeResults, https://www.pnrjournal.com/index.php/home/article/vie w/4470.