www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Advanced Machine Learning Techniques for Accurate Used Car Price Prediction: A Data-Driven Approach

Mr K.Sreenivasa Reddy[1]
Asst. Professor
Department of CSD
TKR College of Engineering & Technology
ksrinivasreddy@tkrcet.com[1]

V.Vaishnavi[2]
B.Tech(Scholar)
Department of CSD
TKR College of Engineering & Technology
vaishnavireddyvelpoor@gmail.com[2]

M.Fani Goud[3]
B.Tech(Scholar)
Department of CSD
TKR College of Engineering & Technology
mfanigoud@gmail.com[3]

P.Abhiram[4]
B.Tech(Scholar)
Department of CSD
TKR College of Engineering & Technology
abhiram2709@gmail.com[4]

K.Manigoud[5]
B.Tech(Scholar)
Department of CSD
TKR College of Engineering & Technology
manigoudkollu@gmail.com[5]

ABSTRACT

The manufacturer sets the price of a new car in the industry, with the government incurring some additional expenditures in the form of taxes. Customers purchasing a new car may thus be confident that their investment will be worthwhile. However, due to rising new car prices and buyers' financial inability to purchase them, used car sales are increasing globally. A USED CAR PRICE PREDICTION SYSTEM that efficiently assesses the car's worthiness using a range of factors is required. The current system involves dealers deciding prices at random, leaving buyers unaware of the car's actual worth. Sellers, too, lack knowledge of appropriate pricing. To address this issue, a highly effective model is proposed. Regression algorithms are employed, producing continuous values rather than classified values. This allows for precise price estimation rather than generalized price ranges. A user interface has also been created to allow users to input data and view predictions.

Keywords: *Used Car Price Prediction, Regression Algorithms, Machine Learning, Linear Regression, Ridge and Lasso Regression, Bayesian Ridge Regression, Decision Tree, Random Forest, XGBoost, Gradient Boosting.*

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

1.INTRODUCTION

Determining if the quoted price of a used car is fair is a difficult process owing to the numerous elements that influence a used vehicle's market pricing. The goal of this research is to create machine learning models that can properly anticipate the price of a used car based on its features so that buyers can make informed choices.

We create and analyze numerous learning algorithms using a dataset that includes the selling prices of various brands and models. We will compare and choose the best machine learning algorithms such as Linear Regression, Lasso Regression, Ridge Regression, Bayesian Ridge Regression, Decision Tree Regression, Random Forest Regression, XG Boost Regression, and Gradient Boosting Regression. The price of the car will be determined by a number of factors. Regression algorithms are used because they produce a continuous value rather than a categorized value, allowing us to predict the actual price of a car rather than the price range of a car. A user interface has also been created that takes input from any user and shows the price of a car based on the inputs

2.RELATED WORK

Numerous studies have explored machine learning techniques for predicting used car prices by analyzing various datasets and algorithms. Pudaruth (2014) employed methods such as multiple linear regression, k-nearest neighbors, and decision trees to predict prices, though small sample sizes impacted accuracy. Kuiper (2008) demonstrated a multivariate regression model, highlighting variable selection as a key factor for improving predictions.

Pal et al. (2019) utilized Random Forest regression on Kaggle datasets, achieving high accuracy by focusing on features like price, kilometers driven, and vehicle type. Gegic et al. (2019) incorporated neural networks and support vector machines to predict car prices in Bosnia and Herzegovina, achieving an accuracy of 87.38%. Listiani (2009) demonstrated the superiority of Support Vector Machines over traditional regression methods when working with high-dimensional datasets. Additionally, Dholiya et al. (2019) developed a web-based system using multiple linear regression to estimate prices dynamically, catering to user preferences.

These studies provide a foundation for the effective use of regression and ensemble learning techniques while highlighting the importance of preprocessing and feature engineering in improving predictive accuracy.

However, the need for models tailored to regional markets, such as India, remains underexplored, motivating this research.

3.METHODOLOGY

Figure 1: Workflow of Study

3.1 Data Gathering

The source of the data is the web portal Kaggle.com, where vehicle data sets are

Index in Cosmos
MAY 2025, Volume 15, ISSUE 2
UGC Approved Journal

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

provided by Cardekho for the sale and purchase of cars. The dataset contained the following features: car name, year, selling price, present or current price, kilometres driven, fuel type: diesel, petrol, or CNG (compressed natural gas), seller type: dealer or individual, transmission: automatic or manual, owner (number of previous owners).

3.2 Create Environment

An environment is created using the Anaconda prompt. This environment would separate our project area from the other default environment (base) or other previously created environments. All the packages, libraries, and modules that we need can be manually installed in the environment created in this way, making it an advantageous step. In such an environment, we can make changes according to our needs.

3.3 Data Reading

The first step is to import and read the csv file for the research. The dataset is extensively examined in terms of null values, shape, columns, numerical and categorical features, dataset columns, unique values of each feature, data information, and so on.

3.4 Data Pre-processing

Some of the data features were renamed for clarity (Present Price = Initial Price, Owner = Previous Owners), and some features that were not important for analysis were removed. In exploratory data analysis, we use statistical graphics and other visualisation techniques to describe the important aspects of data. Top Selling Vehicles, Year vs. Number of Available Vehicles, Selling Price vs. Initial Price, Vehicle Fuel Type, Transmission Type, Seller Type, Age, Selling Price v/s Age, Selling Price v/s Seller Type, Selling Price v/s Transmission, Selling Price v/s Fuel Type, Selling Price v/s Previous Owners, Initial Price vs Selling Price, Selling Price v/s Kilometers Driven, pairplot, heatmaps, and other visualisations are used to gain a better understanding of data. Following EDA, One Hot Encoding approach is used to deal with the dataset's categorical features.

After that, the dataset's correlation characteristics are generated and International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 09 Issue: 12 | Dec 2022 www.irjet.net p-ISSN: 2395-0072 © 2022, IRJET | Impact Factor value: 7.529 | ISO 9001:2008 Certified Journal | Page 357 thoroughly analysed by visualising several plots. Then the features allocation of data is where the dependent feature (Selling Price) and independent features (Initial Price, Kilometers Driven, Previous Owners, Age, and so on) are then allocated for further processing.

3.5 Train-Test Split

Once the dependent and independent features have been assigned, we proceed with the splitting of the dataset into training and testing data. We use 80% of the data to train our model and 20% to test it.

3.6 Model Building

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Following the Train-Test split, data modeling is complete, and the process of building the model begins. The model is defined, along with a few parameters, for future implementation. After the model is built, various algorithms are used to create the final results. After building the model, the following algorithms are used for predictive analysis.

4.IMPLEMENTATION DETAILS

Adding a new feature Age, which determines the number of years the vehicle has been used, is stored in the final dataset, and the year attribute is dropped.

4.1 Exploratory Data Analysis

In this stage, we summarize the major characteristics of data using statistical graphics and other visualization tools. Various graphs and charts are plotted to gain a better understanding of the dataset and the relationships between its features.

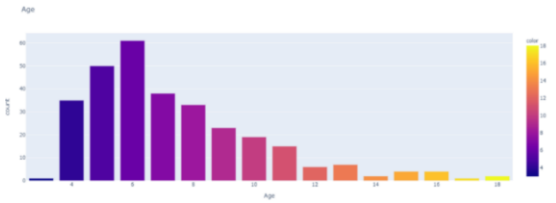


Figure 2: Count w.r.t Age

Vehicle count in relation to vehicle age: The following bar graph depicts the number of vehicles of a certain age.

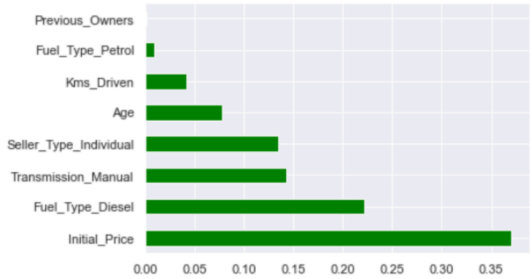
Figure 3: Selling price v/s Age

Comparison of each vehicle's selling price vs. age: The chart below depicts the selling price and age of a certain car. And it is easy to conclude that the selling price is high for a car of a young age.

4.2 One Hot Encoding

The one hot coding approach is used to deal with the categorical variables in the dataset. It generates a sparse matrix or a dense array based on the parameters while creating a binary column for each category or parameter.

Fuel Type, Seller Type, and Transmission were the three categorical variables in our dataset. Following one hot encoding, these variables are given a binary representation, so that for a car with a Fuel Type of Diesel, the value of Fuel_Type_Diesel is a binary 1 and the value of Fuel_Type_Petrol is a



binary 0. The same procedure is applied for the remaining category variables.

Figure 4: Feature Importance

Feature Importance of dataset: The feature importance technique provides a score to features in a feature set based on their usefulness in predicting the target variable. Initial Price is the most relevant feature in the provided dataset, while Previous Owners is the least important.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

4.3 Model Building

After the train-test split of the dataset, modeling is complete, and the process of building the model begins. For final implementation, the model is created with a few parameters, such as the algorithm, x train, y train, x test, and y test. After the completion of the model, various algorithms are used to generate the final results.

4.4 Developing a Web Application:

A web application is then made using HTML, CSS, and JavaScript in the frontend and using the Flask Framework of Python in the backend. This web application allows any user to enter parameters and calculate the estimated selling price of a used car. To view the results, the user must enter values for variables such as year, initial price (in lakhs), kilometers driven, and previous owners, as well as select options for parameters such as fuel type, transmission type, and seller type.

Figure 5: Web Application

5. LITERATURE SURVEY

The prediction of used car prices has garnered significant interest in the field of machine learning, with researchers exploring various algorithms and methodologies. Pudaruth (2014) utilized multiple linear regression, k-nearest neighbors, and decision trees for predicting car prices in Mauritius. While results showed comparable accuracy among these methods, limitations like small sample sizes and challenges in handling numeric values were noted, particularly for decision tree and Naive Bayes models. Kuiper (2008) introduced multivariate regression for predicting prices of General Motors vehicles, emphasizing the importance of variable selection in improving prediction models and demonstrating that publicly available data can yield reliable forecasts.

Further advancements were seen in Pal et al. (2019), where Random Forest regression was applied to Kaggle datasets. By focusing on features like price, kilometers driven, and vehicle type, the model achieved high accuracy rates of 83.62% for test data. Similarly, Gegic et al. (2019) integrated neural networks, support vector machines (SVM), and Random Forest algorithms to forecast car prices in Bosnia and Herzegovina, achieving 87.38% accuracy through a combination of techniques and web scraping for data collection.

Richardson (2009) explored regression models to analyze how hybrid vehicles retain their value better than conventional cars, driven by environmental concerns and fuel efficiency. This study also highlighted the role of additional variables like age, mileage, and fuel efficiency (MPG) in influencing car prices. Listiani (2009) demonstrated the

effectiveness of SVM for handling high-dimensional datasets, outperforming linear regression by avoiding issues like overfitting and underfitting, especially for large-scale data.

More recent works have leveraged ensemble learning techniques. A study on XGBoost by Dholiya et al. (2019) demonstrated its robustness in predicting car prices by integrating historical data with advanced feature engineering. Meanwhile, research by Huang et al. (2023) incorporated feature selection methods like Recursive Feature Elimination (RFE) alongside algorithms like LightGBM and CatBoost, achieving high R^2 scores and demonstrating the impact of model fusion.

These studies collectively highlight the evolution of methodologies for used car price prediction. From traditional regression techniques to ensemble learning and neural networks, researchers have consistently improved accuracy and scalability. However, challenges such as adapting models to regional markets, incorporating real-time data, and handling diverse datasets remain open for future exploration. These works establish a strong foundation for developing robust, adaptable, and practical machine learning systems for the automotive industry.

6. CONCLUSION

The problem of predicting used car prices accurately is multifaceted, involving several challenges such as data availability, feature variability, and market dynamics. This study addresses these challenges by employing various regression algorithms to build a robust prediction model. The results demonstrate that machine learning techniques, especially ensemble methods like Gradient Boosting and Decision Tree Regression, excel in capturing the relationships between features such as age, mileage, fuel type, and transmission, thereby delivering precise price estimates. The use of one-hot encoding for categorical features and correlation analysis during preprocessing enhanced the model's ability to identify key contributors to price prediction, such as initial price and kilometers driven.

While linear regression provided a baseline understanding of the relationships, its performance was limited by its inability to handle non-linear patterns in the dataset. Ensemble methods like Random Forest and Gradient Boosting addressed this limitation by leveraging multiple decision trees, which improved accuracy and reduced errors. The Decision Tree Regression model, in particular, stood out as the best performer, achieving an R^2 score of 0.9544, the highest among all tested algorithms. This highlights the importance of non-linear models in capturing the complexities of used car price prediction.

One critical aspect of this study was the effective handling of categorical variables such as fuel type, seller type, and transmission type. By transforming these variables into binary features using

one-hot encoding, the model was able to evaluate their individual contributions without introducing noise or bias. Additionally, exploratory data analysis revealed valuable insights, such as the strong inverse relationship between a car's age and its selling price, as well as the direct proportionality between initial price and selling price. These insights are not only useful for model training but also provide actionable information for dealerships and sellers.

Another significant contribution of this work is the development of a user-friendly web application that allows users to input car details and receive real-time price predictions. By integrating the machine learning model into a Flask-based backend, the system ensures scalability and ease of deployment. This practical application bridges the gap between theoretical research and real-world usability, providing an invaluable tool for buyers and sellers in the used car market.

Despite the success of the proposed model, there are areas for improvement. For example, incorporating real-time data from online marketplaces and user reviews could enhance the model's adaptability to dynamic market conditions. Additionally, leveraging advanced techniques like deep learning and neural networks may further refine accuracy, particularly for high-dimensional datasets. Future research should also explore regional adaptations of the model to account for localized market trends and buyer preferences, ensuring its broader applicability in diverse markets like India. Overall, this study demonstrates the effectiveness of machine learning in addressing complex pricing challenges while laying the groundwork for future advancements in the field.

REFERENCES

[1] A latent factor-based bayesian neural networks model in cloud platform for used car price prediction, august 2024

[2] Seeking in ride-on-demand service: a reinforcement learning model with dynamic price prediction, september 2024

[3] House price prediction: a multi-source data fusion perspective, september 2024

[4] Flight price prediction web-based platform: leveraging generative ai for real-time airfare forecasting, april 2024

[5] A novel used vehicles price prediction model based on denoising autoencoder with convolution operation, march 2023

[6] A deep learning-based cryptocurrency price prediction model that uses on-chain data, june 2022

[7] Day-ahead electricity price forecasting based on hybrid regression model, october 2022

[8] Trip pricing scheme for electric vehicle sharing network with demand prediction, november 2022

[9] Predicting the regional adoption of electric vehicle (ev) with comprehensive models, august 2020

[10]Predicting available parking slots on critical and regular services by

exploiting a range of open data, august 2018

[11] "Mood As Information: 20 Years Later," Psychological Inquiry, Vol. 14, No. 3, Pp. 296–303, 2003, Doi: 10.1207/S15327965PLI1403&4_20.

[12] "Happy And Mindless, But Sad And Smart? The Impact Of Affective States On Analytic Reasoning," In Emotion And Social Judgments, J. P. Forgas, Ed., New York, NY, USA: Garland Science, 2020, Pp. 55–71.

[13] "A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios," IEEE Trans. Affective Comput., vol. 14, no. 1, pp. 747–760, Jan./Mar. 2023, doi: 10.1109/TAFFC.2021.3063387.

[14] "A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles," 2020, arXiv:2005.08626.

[15] "Face detection techniques: A review," Artif. Intell. Rev., vol. 52, no. 2, pp. 927–948, 2019, doi: 10.1007/s10462-018-9650-2.

[16] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "Yolo-face: A real-time face detector," Vis. Comput., vol. 37, no. 4, pp. 805–813, 2021, doi: 10.1007/s00371-020-01831-7.

[17] T. Cootes, E. Baldock, and J. Graham, "An introduction to active shape models," Image Process. Anal., vol. 328, pp. 223–248, 2000.

[18] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, Jun. 2001, doi: 10.1109/34.927467.

[19] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit., 2010, pp. 1078–1085, doi: 10.1109/CVPR.2010.5540094.

[20] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," Multimedia Tools Appl., vol. 76, no. 6, pp. 7803–7821, 2017, doi: 10.1007/s11042-016-3418-y.

[21]Attention-Emotion-Enhanced Convolutional LSTM for Sentiment Analysis – IEEE Journals & Magazines.

[22] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 915–928, Jun. 2007, doi: 10.1109/TPAMI.2007.1110.

[23] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "LBP with six intersection points: Reducing redundant information in LBP-TOP for micro expression recognition," in Proc. 12th Asian Conf. Comput. Vision (ACCV), Cham, Switzerland: Springer-Verlag, 2015, pp. 525–537.