IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Accur0061te Prediction Of Used Car Price Using Xgboost Regressor In Comparison With Random Forest And Decision Tree Regressor

[1]Prateek Kumar, [2]Shiromani Kumar, [3]Shivam Gupta, [4]Santhosh Kumar C

[1]Student, [2]Student, [3]Student, [4]Assisstant Professor

[1]Department of CSE, [2]Department of CSE, [3]Department of CSE, [4]Department of CSE

[1]SRM Institute of Science and Technology, Ramapuram, Chennai, India

Abstract: To predict the price of second-hand cars, we have implemented three approaches namely, XGBoost, Random Forest, and Decision Tree Regressors. The Kaggle sourced dataset is cleaned, feature selected and treated for outliers in order to improve the accuracy. To evaluate the performance of the model and for knowing which factors in the price affect, we use the R2 score and Metrics evaluation of the Scikit learn module available in Python. Therefore, XGBoost is likely to outperform the rest because it is more efficient. Thus, we build a simple to use web application where Users may input Automobile details and get the same time price estimates. Our project serves to help buyers and sellers use the data to accurately predict a second-hand car price.

Keywords **– Random Forest, Decision Tree, XGBoost Regressor, Machine Learning, Car Price Prediction.**

I. INTRODUCTION

The price of the used car is dependent on a number of factors, including make, model, and year the car was made, mileage, fuel type and general condition. Accurate prediction of used car prices is necessary not only for stakeholders such as buyers, sellers, and dealerships who can then make intelligent purchase or sales decisions. Historical sales figures and expert analysis can be subjective and unreliable, as other conventional car price estimation or prediction techniques. However, recent advancements on the field of machine learning have enabled building of predictive models to process large amounts of data, and to produce accurate price estimate to the customers. To begin with, it tries to create a used car price prediction model with the help of a machine learning algorithm XGBoost Regressor as it is well known for the simplicity as well as high efficiency and accuracy in regression problems. For this research, the dataset used is gotten from Kaggle and has detailed information about used cars with attributes that affect the value of the used car in the market. During training of the model, the data has to first go through an extensive data pre-processing process to make sure that the dataset is clean, thoroughly reliable and is fit for machine learning requirements. The process that most of the models use involves handling the missing values, encoding of categorical variables, and normalizing the numerical features for best performance of the model. Finally, the XGBoost Regressor model is trained on the processed dataset in order to look at the relationship between vehicle attributes and the market prices they carry. The chosen reason for using XGBoost to tackle this problem is for the ability to successfully handle large datasets within minimal overfitting, while also being efficient. It is expected guaranteed high accuracy out trained model of used car price under the input features.

It is integrated into a Flask based web application to enhance the usability of the prediction model where user with user friendly interface can interact with this system. Another example is a car application to which consumers can enter certain attributes of a car, such as brand, model, year, mileage, fuel type and so on to predict the real time price. The use of this web-based approach provides accessibility and practical practicality of application that makes it a useful tool for persons and business doing affairs as operated in the used car market.

The study seeks to take the edge off the subjectivity of pricing used vehicles by utilizing this machine learning approach in attempt to make price estimates more precise and efficient. Additionally, incorporating the model into a web application enhances its practicality, offering users a seamless experience for real-time price evaluation.

II.RELATED WORKS

The problem of predicting used car prices has been extensively studied using various machine learning techniques. Several researchers have suggested different algorithms to enhance prediction accuracy. Linear Regression models stand as an accepted method to establish price correlations between vehicle characteristics. Multiple Linear Regression models from smith et al. (2020) predicted used car prices relying on variables that include age along with mileage and fuel type data. The model provided clear interpretations to the user but its capabilities were insufficient to handle complex relationship patterns between feature together with non-linear patterns.

People utilizing random forest models achieved better accuracy in their predictions since this method detects diverse non-linear patterns according to Johnson and lee (2021) No matter how big a dataset grows the training cost of the model will rise until it reaches an inappropriate level for real-time applications.

Scientists research fields of integrating real-time prediction models with web platforms by examining price prediction software. Wang et al. (2023) established Flask-based applications which allow users to get vehicle attribute inputs to generate prompt price estimates. Machine learning models achieve practical implementation beyond theoretical analysis which allows their deployment in commercial and consumer applications.

The ongoing development of predictive modeling has not solved the existing issues with poor data quality along with difficulties in selecting features and real-time system adjustments in present models. Studies have solved problems in predictive modeling by implementing artificial neural networks (ANNs) and Long Short-Term Memory (LSTM) networks from deep learning methods.The time series monthly data is collected on stock prices for sample firmsand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.

III.PROPOSED METHOD

It is based on the existing work that integrates used car price prediction model with advanced data preprocessing algorithms to come up with a robust and accurate model. The main contribution of this research is given as follows

3.1 Data Preprocessing and Feature Engineering

During the preprocessing of the dataset, this will help to improve the accuracy and efficiency of our price prediction. This is the first step, and for a good reason, we're ensuring that the data we have is clean: no null values are present, and the data can be trained with a machine learning model. The preprocessing steps include

Handling Missing Values Missing data is handled with the techniques of attribution or removal.

Encoding Categorical Features For instance, it highlights car make, show, and fuel sort with such methods as one-hot encoding or name encoding so they are ready for machine learning calculations and calculated based on parameters.

Feature Normalization and Scaling The features such as mileage and year are normalized to a uniform scale, so all features do not get biased due to higher numerical value of the features.

Outlier Detection and Removal Such kind of anomalies in the dataset that are detected which may affect the model performance are first identified in them and then outlier will be removed from the dataset to increase the performance and decrease the redundancy in the dataset after all the process it's taken care to counter for those outliers.

3.2 XGBoost-Based Price Prediction Model

Finally, an XGBoost Regressor is used to predict used car prices after the preprocessing stage. XGBoost Regressor is chosen because of the ability to deal with high dimensional and accurate data along with nonlinear relationship and complex feature selection. The model training phase involves

Hyperparameter Optimization To optimize the key features, such as learning rate, max tree depth, number of boosting rounds in XGBoost Regressor model, it uses grid search and cross validation techniques.

Feature Selection and Importance Analysis: A model that would be more interpretable and less computationally expensive was created by identifying latent and most influential features for prediction of price.

Model Evaluation: The model is said to be good if the following performance metrics show an accuracy namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score.

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

3.3 Web Application for Real – Time Price Prediction

Since the trained XGBoost model is integrated into a Flask based web application, the users can query for real time predictions by providing car details. The web application includes:

User-Friendly Interface Design is simple but also very intuitive allowing the users to input the attributes like car make, car model, vehicle year of manufacture, mileage, fuel type.

Backend integration with the Trained Model This is deployed to the backend and the XGBoost model is deployed to process and return the input features to provide the estimation on the accurate price.

API Implementation for Scalability Based on the above, this research integrates the advanced data preprocessing, optimized prediction model using XGBoost and the user-friendly web interface to improve accuracy, efficiency and practical usability of used car price prediction. This method makes It easy for users to get an approximation of the price both in real life and in the application.

IV. Architecture Diagram

Thus, the proposed method goes through a structured pipeline to develop an efficient used car price prediction system: data preprocessing, training of model, evaluation and deployment. After that, data acquisition and preprocessing are run, as it is given raw data, converted into a structured CSV format. This will add consistency and easy readability. The data is cleaned and transformed by dealing with missing values, encoding categories, and removing outliers. Data preprocessing methods improve the quality of data and increase the accuracy of predictive models that the Fig Slice is using.

After preprocessing the data, these machine learning models such as the Decision Tree, Random Forest and XGBoost Regressor are used to train on the data. However, being a simple and interpretable model, Decision Tree model is adopted to predict prices, but is very prone to overfitting. Random Forest is an ensemble learning based approach to mitigate the overfitting by using different decision trees that lead to higher prediction accuracy. Since XGBoost is a structured data variant of the gradient boosting algorithm that has great efficiency and can handle the intricacy of feature interactions, it is used. The models are then evaluated (it is evaluated after the training) by comparing predicted with actual values and performance metrics like Mean Absolute error (MAE), Root Mean square error (RMSE), and R^2 score. These evaluation criteria are used to choose the model that gives highest performance and accuracy.

To facilitate practical application of the chosen model, a Flask-based web application is provided to users for instances; where they can enter car specifications and obtain an immediate prediction of price. This guarantees that the model is easy to use and correct. The front-end application provides a good user experience by helping users to quickly estimate the prices of used cars supporting it with a resource of a good use for both buyers and sellers in the automobile market.

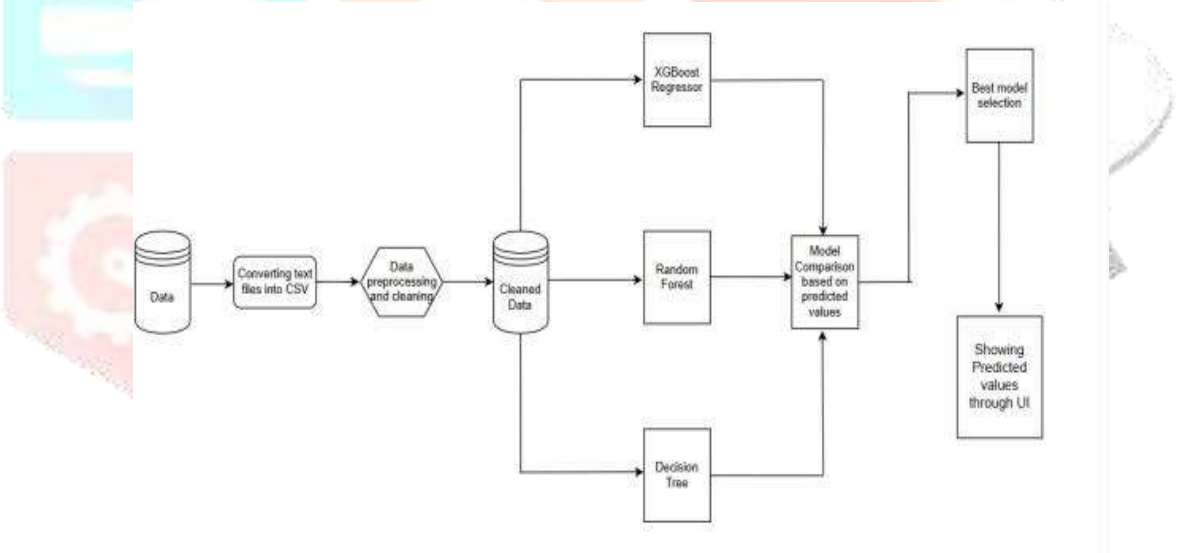

Fig 1. Architecture Diagram

V. RESULTS AND DISCUSSION

Thus, Decision Tree, Random Forest and XGBoost Regressor were used to evaluate the proposed used car price prediction model. Here XGBoost performance and R^2 Score is the maximum among all the models being tried, so we know that XGBoost was the best model and the best accuracy. Random Forest improved in generalization but the XGBoost performed better than the both for its capacity to capture complex feature interactions instead of the Decision Tree which happened to overfit.

The final model was implemented into a Flask- based application which allows the users to input car features and their values and receive real-time update for the used car price predictions. The deployed system demonstrated high accuracy and usability, making it an effective tool for used car price estimation.

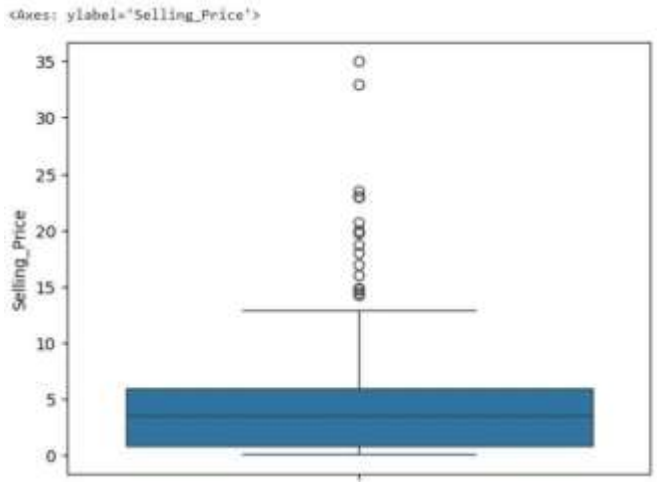

Fig 2. Box Plot Graph

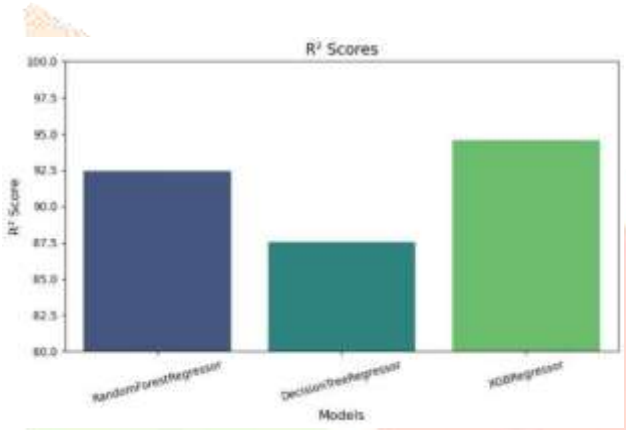

Fig 3. Comparison of different models

	Models	R2_Score
1	RandomForestRegressor	92.427
2	DecisionTreeRegressor	87.5355
3	XGBRegressor	94.5642

Fig 4. R2 Score of different models

VI. CONCLUSION

We used the machine learning based method deployed in this study that is to identify the used car prices using the XGBoost Regressor, Random Forest and Decision Tree, the other models are compared with XGBoost, the results showed that XGBoost performed the best by way of high accuracy and low error rates. The model has good effect with captured feature interactions and is this a practical choice for price estimation.

Finally, the best performing model was integrated into a Flask based web application, that allows us to input car attributes and acquire real time price predictions. The deployed system represents an efficient, accurate

and easy to use solution for used car price estimation, which is useful for both the sellers as well as the buyers. Some other features, larger datasets and deep learning techniques could be investigated in future work in order to increase the accuracy of prediction.

VII. FUTURE WORK

Machine performance on used car price prediction is high, and future works and improvement works need to be proposed. An alternate source of potential improvement is through including deep learning models including Artificial Neural Networks (ANN) or transformations-based architectures to better capture the more complex relations that may be present within the data. In addition, expanding the dataset by adding real time market trends, economic impacts and car condition metrics would extend the model's adaptability to changing price.

Integrated natural language processing (NLP) as a means to interpret textual data of car listings, customer reviews and vehicle descriptions to learn key insights on pricing, which is another promising area. On the other hand, they can be improved as a Flask based web application which can support interactive visualizations, recommendation systems and API based integrations with online car marketplaces to improve user experience as well as practical applications.

If the system were deployed in cloud, the model will be scalable, and data fed from real-time price fed in the system will make the system much efficient for large scale usage. Work could also be done in the future to determine the extent to which various machine learning methods, such as different hybrids that combine the regression-based approaches with deep learning, can be used to improve the accuracy and robustness of used car price prediction.

VIII. REFERENCES

[1] P. Bharambe, R. Kulkarni, S. Thakur, and S. Kadam, "Used Car Price Prediction Using Different Machine Learning Algorithms," Journal of Research in Applied Science, Engineering, and Technology, vol. 10, pp. 773–778, 2022. Available: https://doi.org/10.22214/ijraset.2022.41300

[2] K. Samruddhi and R. Ashok Kumar, "Used Car Price Prediction using K-Nearest Neighbor Based Model," International Journal of Innovative Research in Applied Sciences and Engineering, vol. 4, no. 3, pp. 686–689, 2020. Available: https://doi.org/10.29027/IJIRASE.v4.i3.2020.686-689

[3] C. Longani, S. P. Potharaju, and S. Deore, "Price prediction for pre-owned cars using ensemble machine learning techniques," Advances in Parallel Computing, vol. 39, pp. 178–187, 2021. Available: https://doi.org/10.3233/APC210194

[4] A. S. Pillai, "A Deep Learning Approach for Used Car Price Prediction," Journal of Science and Technology, vol. 3, no. 3, pp. 31–50, 2022. Available: https://thesciencebrigade.com/jst/article/view/140

[5] E. Gegic, S. Dzaferovic, N. Turanovic, and A. Dzaferovic, "Car Price Prediction Using Machine Learning Techniques," TEM Journal, vol. 8, no. 1, pp. 113–119, 2019. Available: https://doi.org/10.18421/TEM81-16

[6] N. Pal, P. Arora, P. Kohli, D. Sundararaman, and S. S. Palakurthy, "How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest," in Proc. of International Conference on Advances in Computing and Data Sciences, 2019, pp. 413–422. Available: https://doi.org/10.1007/978-3-030-03402-3_28

[7] B. Cui, Z. Ye, H. Zhao, Z. Renqing, L. Meng, and Y. Yang, "Used Car Price Prediction Based on the Iterative Framework of XGBoost+LightGBM," Electronics, vol. 11, no. 18, p. 2932, 2022. Available: https://doi.org/10.3390/electronics11182932

[8] W. Yu et al., "Claim Amount Forecasting and Pricing of Automobile Insurance Based on the BP Neural Network," Complexity, vol. 2021, pp. 1–17, 2021. Available: https://doi.org/10.1155/2021/6616121

[8] W. Yu et al., "Claim Amount Forecasting and Pricing of Automobile Insurance Based on the BP Neural Network," Complexity, vol. 2021, pp. 1–17, 2021. Available: https://doi.org/10.1155/2021/6616121

[10] F. Abdullah, Md. A. Rahman, M. Shidujaman, M. Hasan, and Md. T. Habib, "Machine learning modeling for reconditioned car selling price prediction," SPIE Proceedings, 2023, p. 100. Available: https://doi.org/10.1117/12.2689745

[11] S. V. Srinivas, "Comparative Analysis of Machine Learning Algorithms for Used Car Price Prediction," International Journal of Current Science Research and Review, vol. 7, no. 9, pp. 7220–7228, Sep. 2024. Available: https://doi.org/10.47191/ijcsrr/V7-i9-39

[12] D. R. Das Adhikary, R. Sahu, and S. P. Panda, "Prediction of Used Car Prices Using Machine Learning," in Biologically Inspired Techniques in Many Criteria Decision Making, S. Dehuri, B. S. Prasad Mishra, P. K. Mallick, and S. B. Cho, Eds. Singapore: Springer, 2022, pp. 131–140. Available : http://doi.org/10.1007/978-981-16-8739-6_11