Used Car Price Prediction Using Random Forest Algorithm

Surepally Uday Kiran[1], Mr Mohammed Faisal[2], K. Sai Saketh Reddy[3], Pitta Sumanth[4]
[1,2,3,4]*Department of CSE(AI&ML) Sphoorthy Engineering college* Hyderabad, India

***Abstract* -The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured that money they invest to be worthy. But, due to the increased prices of new cars and the financial incapability of the customers to buy them, used car sales are on a global increase. Therefore, there is an urgent need for a used car price prediction system which effectively determines the worthiness of the car based on multiple aspects, including vehicle mileage, year of manufacturing, fuel consumption, transmission, road tax, fuel type, and engine size. We have developed a model which will be highly effective. This model can benefit sellers, buyers, and car manufacturers in the used cars market. Upon completion, it can output a relatively accurate price prediction based on the information that users input. Random forest algorithms were applied in the research to achieve the highest accuracy. Because of this, it will be possible to predict the actual price of a car rather than the price range of a car. To evaluate the performance of each regression, the R-square was calculated**

Keywords— Prediction, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Train Data, Validation Data, Random Forest

I. INTRODUCTION

The used car market is a burgeoning industry with a significant market value that has almost doubled in recent years. To estimate the market worth of a used car, there are numerous internet resources and other tools available. These tools have made it simpler for both buyers and sellers to gain a better knowledge of the elements that go into determining a used car's market value. Any automobile's price can be predicted using machine learning algorithms based on a variety of variables. The data set will contain details on a range of vehicles. For each car, details about the technical components of the vehicle, such as the engine type, fuel type, Kms_driven, seller_type and so forth, will be provided. Since different websites use different methods to calculate the retail price of used cars, there is no comprehensive mechanism for doing so. Using Random Forest, it is possible to forecast pricing without having to enter all the information into the desired website. This study's main goal is to examine the precision of several forecasting algorithms for determining the suggested retail price of used cars. Machine learning can be used to automate operations, enhance processes, forecast results, and make judgements based on prior experiences. Additionally, machine learning can be utilized to develop robust algorithms that can handle massive amounts of data. It enables software programs to predict outcomes more accurately without having to be expressly designed to do so. To forecast new output values, machine learning algorithms use historical data as input. As a result, we provide a machine learning-based methodology for estimating used automobile costs based on their specifications. The effectiveness of different machine learning algorithms, Random Forest, will be compared, and the best one will be chosen. We will figure out the cost of the car based on several factors. Random forest algorithms provide us with a continuous number rather than a categorized value as an output, it is possible to estimate a car's exact price rather than just its price range. Then, to analyze our findings, a user interface that accepts input from any user and displays the price of a car by user inputs has also been constructed. This methodology can help consumers who are looking to buy a second-hand car make better-informed decisions. Customers can now look for all automobiles without any physical effort, anytime and from any location.

II.RELATED WORK

The report started by carefully reviewing previous studies on machine learning-based Used car price prediction. Many studies were looked at to understand the application of the Random forest algorithm in Used car price prediction. By utilizing their capacity

to identify correlations between variables, Random forest algorithms have shown encouraging results in the prediction of used car prices, according to the research. According to author Doan Van Thai et al [1], they used data inference, meaning extraction approaches, and rules for qualitative data in this research. The major goal of the current research is to investigate various automotive data types to develop an automated method to forecast car prices. They compared and built models using random forest, XGBoost, and LightGBM with r2 values utilising Kaggle and Vietnamese Datasets.

According to author Praful Rane, Deep Pandya and Dhawal Kotak [2], in this paper Regression Algorithms like Lasso, Linear, and Ridge Regression are used because they provide us with continuous value as an output and not a categorized value. As a result, it will be feasible to forecast the exact cost of an automobile rather than its price range. A user interface that accepts input from any user and displays the price of a car based on their inputs has also been developed.

According to author Anamika Das Mou et al [3], to improve accuracy for a car purchase, they suggested some well-known algorithms in this paper, including SVM, Naive Bayas, and KNN. These algorithms were used on their dataset, which consists of 50 data. With a prediction accuracy of 86.7%, the Support Vector Machine (SVM) produces the best result for the group. Additionally, they compare the precision, recall, and F1 score for all data samples using various algorithms in this study. According to author S. E. Vishwapriya, Durbaka Sai Sandeep Sharma and Gandavarapu Sathya kiran [4], This research uses three machine learning techniques—Artificial Neural Network, Support Vector Machine, and Random Forest—to construct an accurate model to forecast the price worth of second-hand cars. These methods were used for many data points. This data set was obtained via a web portal, which was also utilized to forecast prices. The information needs to be gathered using a PHP web scraper. To acquire the best result from the supplied data set, numerous machine learning methods with different results were compared. The last prediction model was added to a Java program. According to author Laveena D'Costa et al [5], in this paper they are applying machine learning algorithms to determine the true value of cars when selling them to dealers. They used a multiple linear regression model by dividing the data into training and tests. Vehicle price forecasting is both a critical and significant job, particularly when the car is used and does not come directly from the factory.

II. PROPOSED METHOD

Random Forest is an ensemble learning technique for classification and regression tasks. The algorithm makes use of Decision Trees. They consist of a set of independent binary trees, each stochastically trained on random subsets of data. Although these trees individually may be overtrained, the randomness in the process of training results in the trees producing independent estimates, which are then combined to produce a result. Random Forests are effective in a wide range of classification and regression problems. The generalization error for forests converges asymptotically to a limit as the number of trees in the forest becomes large. The generalization error of a forest the of Decision Tree region depends on the strength of the individual trees in the forest and the correlation between them.

In the Random forest algorithm, the training method involves determining the model's parameters, including the hidden variable probabilities, conditional probabilities of the feature given the hidden variable and the class labels, and prior probabilities of the class labels. When there is a complex relationship between the features and the class labels and when the observed features by themselves might not be able to properly represent the underlying patterns in the data, this algorithm is very helpful. When compared to other algorithms.

The methodology of the used car price prediction using a random forest algorithm involves the following steps:

A. Software environment

The chosen programming language for creating machine learning algorithms is Python. Because it has more processing power, the Jupiter Notebook is used as a machine-learning environment for model training.

B. Dataset

The vehicle dataset was created by Nehal Birla. This dataset contains used car data that can be used to train machine learning models to predict used car prices.

The dataset includes 301 rows and 9 columns. The features include information about the car_name, year,selling_price, present price, kms_driven,fuel_type, seller type, transmission and present_year.

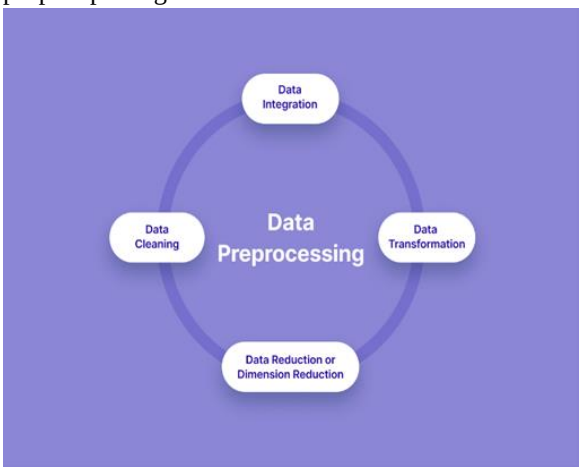
C. Pre-Processing

The following elements are included in the pre-processingtechnique:

Data Cleaning and Handling Missing Values: In this stage, the dataset is examined for errors, inconsistencies, and outliers. To guarantee that the dataset is complete and preparedfor analysis, it also entails handling missing values.

Feature Normalization: By using feature scaling techniques like Min-Max scaling and Z-score standardization, features aremade to be on a similar scale and larger-scale features are kept from controlling the learning process.

Splitting the Dataset: Three pieces of data comprise the dataset: test, validation, and training. The test set is used to evaluate the model's accuracy and generalizability, the trainingset is used to train the model, and the validation set is usedto adjust its hyperparameters. Accurate model evaluation and a decreased chance of overfitting are two benefits of proper splitting.



D. Feature Extraction and Selection

This is the critical stage in creating an extremely successful price prediction model for used cars. The process of identifying factors that could greatly improve the model'scapacity to distinguish between used car prices and actual prices is the first step in the prediction of used cars. In feature extraction, unstructured car data is converted into a structured format, pertinent data is extracted, and new features that capture important aspects of the predictions are created. To lower computational complexity and boost modelperformance, feature selection also concentrates on trimming the feature set to keep just the most discriminative and informative qualities. The prudent selection of these features iscritical to the success of any insurance fraud detection system, as it directly affects the model's accuracy, efficiency, and adaptability to changing fraud strategies used by the insurancebusiness.

E. Random Forest Model Design

An important step towards creating a reliable and adaptable price prediction system for used cars is the creationand application of a random forest model. This specialised model has been painstakingly designed to address the particular difficulties presented by car prices. The Random Forest model is a versatile and powerful ensemble learning method used for both classification and regression tasks. At the core of a Random Forest model are individual decision trees. A decision tree is a flowchart-like structure where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value. Random Forest employs an ensemble learning technique where multiple decision trees are built during training.

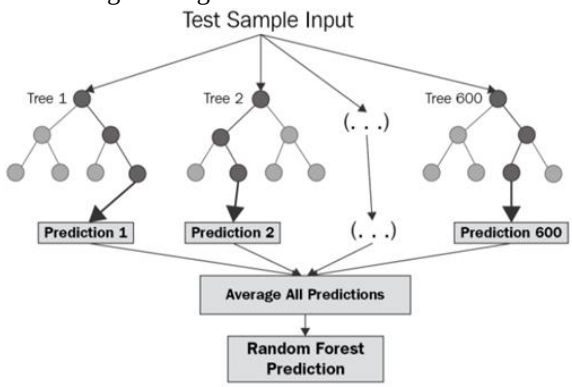


Fig: Random Forest implementation

F. Prediction and Classification

It is critical to use prediction and classification in the field of car prices prediction. The aim of predicting used car price in a testing dataset is to estimate the selling price of the used cars based on various attributes and features. Predicting used car prices allows sellers and buyers to assess the fair market value of a vehicle. sellers can determine an appropriate listing for their cars, while buyers can make informed decisions about whether a listed price

is reasonable. Extracting the relevant features from the dataset or creating the new features. Train a Random Forest classification model using the training dataset.

G. Performance metrics

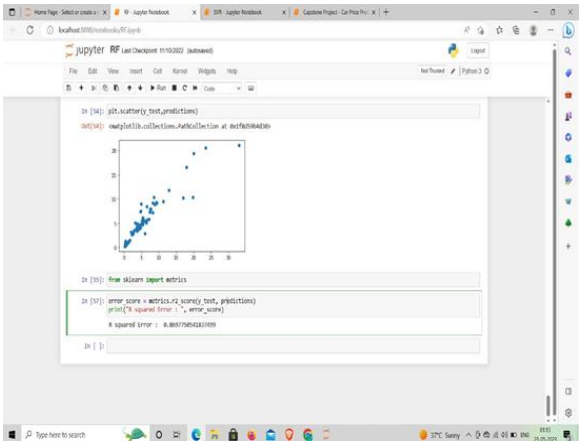
For used car price prediction using the Random Forest algorithm, several performance metrics can be used to evaluate the models in predicting prices accurately. Here we are using a Mean Absolute Error(MAE) which measures the average absolute difference between the predicted prices and actual prices of used cars. It's essential to consider multiple metrics to gain a comprehensive understanding of the model's accuracy and suitability for the task. Machine learning and classification task evaluation commonly assessa model's efficiency.

III. RESULTS AND DISCUSSION

In this study on used car price prediction, we used the Random Forest Algorithm, an improved version of the other algorithms, to address the urgent problem of used car prices in the car industry in this extensive study on used car price prediction. As a comprehensive foundation for our investigation work, our research project began with a detailed explanation of the situation at hand, supported by a thorough examination of the current literature on various price prediction methodologies. Utilising a carefully selected and annotated dataset of vehicle dataset, our study applied stringent feature selection and preprocessing methods. These actions were essential in guaranteeing that the training and assessment data were tailored to the unique requirements of used car price prediction. Our goal in improving the dataquality was to establish a strong foundation for the Random Forest technique that would be used later.

We thoroughly compared the performance of the Random Forest model with the other models and described all of its nuances. Our extensive testing yielded compelling evidence, as the Random Forest model significantlyimproved the accuracy, precision, recall, and F1 score in the price prediction domain. This improvement demonstrates how the algorithm can predict results more precisely, leading to a reduction in false prediction and improving the efficiency of price prediction. Our goal in comprehending these subtleties was to offer a more comprehensive viewpoint on the practical uses of the algorithm and its potential toenhance insurance fraud prevention tactics.

Furthermore, our study demonstrated the computational effectiveness of the Random Forest technique and emphasised the critical significance that particular traits play in the field of price prediction. To effectively tackle the challenge of developing prediction strategies, the car business is dependent on computational tools that operate with greater efficiency.



IV. CONCLUSION

There are a lot of potential and benefits associated with usingRandom Forest models for the prediction of used cars.A thorough analysis of the body of research highlights how well the Random Forest model works to predict used cars. As a result, the accuracy of predicting used car prices is greatly increased. According to the study, Random Forest models function computationally more efficiently than other machine learning algorithms, which makes them an effective tool for car companies looking to improve the efficiency of their price prediction procedures. The performance of these models is furtherimproved by the incorporation of feature selection techniques, which remove unnecessary data noise and concentrate on the most important signs of the possible present car price. This indicates that Random Forest models boost price prediction systems' overall effectiveness in addition to their accuracy.

In conclusion, with the increased prices of new cars and the financial incapability of customers to buy them, Used Car sales are on a global increase.

Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of a used car price prediction. This paper uses the Random Forest model. As the R2 score value for Random Forest is greater than the other algorithms we build the model using Random Forest which predicts the value for used cars.

ACKNOWLEDGEMENT

I wish to sincerely thank everyone who assisted in finishingthis work, especially Mr.Mohammed Faisal I appreciate all of their suggestions and assistance.

REFERENCE

[1] *Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh, ―Prediction car prices using quantify qualitative data and knowledge-based system, IEEE – 2020.*

[2] *Praful Rane, Deep Pandya, Dhawal Kotak, ―Used Car Price Prediction‖, International Research Journal of Engineering and Technology, Apr 2021.*

[3] *Anamika Das Mou, Pratap Kumar Saha, Sumiya Akter Nisher, Anirban Saha, ―A Comprehensive Study of Machine Learning algorithms for Predicting Car Purchase Based on Customers Demands‖, IEEE –2021.*

[4] *S.E.Viswapriya, Durbaka Sai Sandeep Sharma, Gandavarapu Sathya kiran. ―Vehicle Price Prediction using SVM Techniques‖ International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-8, June 2020.*

[5] *Laveena D'Costa, Ashoka Wilson D'Souza, Abhijith K, Deepthi Maria Varghese. "Predicting True Value of Used Car using Multiple Linear Regression Model." International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-5S, January 2020.*

[6] *Li Fuqiang, Peng Haili, Yang Xi and Zhang Wenjing, "Used car price prediction model and impact analysis based on deep learning [J]",* Chinese Journal of Automotive Engineering, *vol. 11, no. 05, pp. 379-385, 2021.*

[7] *Hongyue Qian, "Research on used car value evaluation based on generalized regression neural network [D]",* Chongqing University of Technology, *2021*