International Journal of Engineering Research and Sustainable Technologies
Volume 2, No.2, June 2024, P 20-25
ISSN: 2584-1394 (Online version)

AUTOMATED PRICING PREDICTIONS FOR PRE-OWNED VEHICLES USING RANDOM FOREST

Jayashri Kethini Umapathi
Bank of America,Charlotte,North Carolina,USA
*** Corresponding author email address**: jayashri.kethiniumapathi@bofa.com

https://doi.org/000000/000000/

Abstract

The pricing of new vehicles in the automotive industry is determined by manufacturers, augmented by additional costs imposed by the government in the form of taxes. Consequently, customers purchasing a new car can be assured that their investment is justified. However, due to escalating prices of new vehicles and the financial constraints faced by many consumers, the market for used cars is experiencing significant global growth. This underscores the pressing necessity for an effective used car price prediction system that can accurately assess the vehicle's value based on various factors, including mileage, year of manufacture, fuel consumption, transmission type, road tax, fuel type, and engine size. We have developed a highly effective model designed to serve the needs of sellers, buyers, and manufacturers within the used car market. Upon completion, this model will deliver relatively precise price predictions based on the information provided by users. The Random Forest algorithm was employed in this research to maximize accuracy, enabling the prediction of an actual vehicle price rather than merely a price range. To evaluate the performance of each regression model, the R-squared metric was calculated.

Keywords: Prediction, MAE, MSE, RMSE, Train Data, Validation Data, Random Forest

1. Introduction

The pre-owned automobile market represents a rapidly expanding sector with a substantial market valuation that has nearly doubled in recent years. To ascertain the market value of a used vehicle, a plethora of online resources and analytical tools are readily accessible. These innovations have facilitated a more nuanced understanding for both buyers and sellers regarding the myriad factors that influence a used car's market valuation. Machine learning algorithms can effectively forecast the price of any automobile by analyzing a diverse array of variables. The dataset will encompass intricate details about various vehicles, including specifications pertaining to technical components such as engine type, fuel type, kilometers driven, and seller type, among others. However, due to the disparate methodologies employed by different platforms to assess the retail price of used cars, a unified mechanism for valuation remains elusive. Utilizing the Random Forest algorithm, one can harness these variables to enhance predictive accuracy and derive more reliable market valuations.

It is feasible to forecast pricing without the necessity of inputting comprehensive information into the designated website. The primary objective of this study is to evaluate the accuracy of various forecasting algorithms in determining the suggested retail price of pre-owned vehicles. Machine learning can automate processes, refine operational efficiencies, predict outcomes, and facilitate decision-making based on historical data. Moreover, machine learning enables the creation of robust algorithms capable of processing vast datasets, allowing software applications to predict results with heightened precision without being explicitly programmed for each scenario.

To forecast new output values, machine learning algorithms [7] leverage historical data as input. Consequently, we propose a machine learning-based approach for estimating the costs of used automobiles based on their specifications. This study will compare the effectiveness of various machine learning algorithms, with particular [8] emphasis on the Random Forest algorithm, to identify the most effective model. We will assess vehicle pricing based on an array of factors. The Random Forest algorithm, in particular, provides a continuous numerical output rather than merely a categorical value, thereby facilitating precise estimations of a vehicle's exact price rather than a mere price range.

Additionally, we have developed a user interface that allows users to input parameters and receive a calculated price for a vehicle based on their specifications. This methodology equips consumers in the pre-owned car market with the tools to make more informed purchasing decisions. Buyers can now explore all available vehicles with minimal physical effort, anytime and from any location.

2. Related Work

The report started off by carefully reviewing previous studies on machine learning-based Used car priceprediction. Many studies were looked at in order to understand the application of the Random Forest algorithm in Used car price prediction. By utilizing their capacity to identify correlations between variables, Random Forestalgorithm have shown encouraging results in the prediction of used car price, according to the research. [1] According to authors Doan Van Thai et al., this research employs data inference, meaning extraction methodologies, and [6] qualitative data rules. [6] The primary objective of their study is to explore various automotive data types with the intent of developing an automated system for forecasting vehicle prices. The authors compared and constructed models utilizing Random Forest, XGBoost, and LightGBM [7], evaluating their performance through R^2 values derived from both Kaggle and Vietnamese datasets.

[2] This study employs regression algorithmssuch as Lasso, Linear, and Ridge Regression, which are adept at yielding continuous numerical outputs rather than categorical classifications. Consequently, this approach enables the precise forecasting of an automobile's exact cost rather than merely estimating a price range. Additionally, a user interface has been developed that allows any user to input relevant data and receive an accurate car price based on their specifications. In a related work [3], several prominent algorithms, including Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbours (KNN), have been proposed to enhance accuracy in car purchasing decisions. Their dataset comprises 50 samples, and the SVM algorithm achieved the highest predictionaccuracy at 86.7%. The study further analyzes precision, recall, and F1 score across all algorithms to evaluateperformance metrics comprehensively.

Similarly, [4] utilized three machine learning methodologies—Artificial Neural Networks, Support Vector Machines, and Random Forests—to construct a robust model for forecasting the value of pre-owned vehicles. This analysis was based on a substantial dataset obtained through a web portal, which required data collection via a PHP web scraper. To deriveoptimal results from the dataset, various machine learning techniques were compared, culminating in the integration of the final predictive model into a Java application.

Moreover, [5] explored the application of machine learning algorithms to ascertain the true value of vehicles when selling to dealers.They employed a multiple linear regression model, partitioning the data into training and testing subsets. Accuratevehicle price prediction is a crucial and significant task, especially for pre-owned cars that do not originate directlyfrom the manufacturer.

3. System Methods

Random Forest is an advanced ensemble learning methodology utilized for both classification and regression tasks. This algorithm leverages Decision Trees, which are constituted of multiple independent binary trees, each trained stochastically on random subsets of the dataset. While these individual trees may exhibit tendencies toward overfitting, the inherent randomness during the training process facilitates the generation of independent predictions from each tree. These predictions are subsequently aggregated to yield a comprehensive result.

The efficacy of Random Forests has been well-documented across a diverse array of classification and regression challenges. Notably, the generalization error of the ensemble converges asymptotically to a specific limit as the quantity of trees in the forest increases. This generalization error is influenced by the robustness of the individual trees as well as the degree of correlation among them.

In 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.

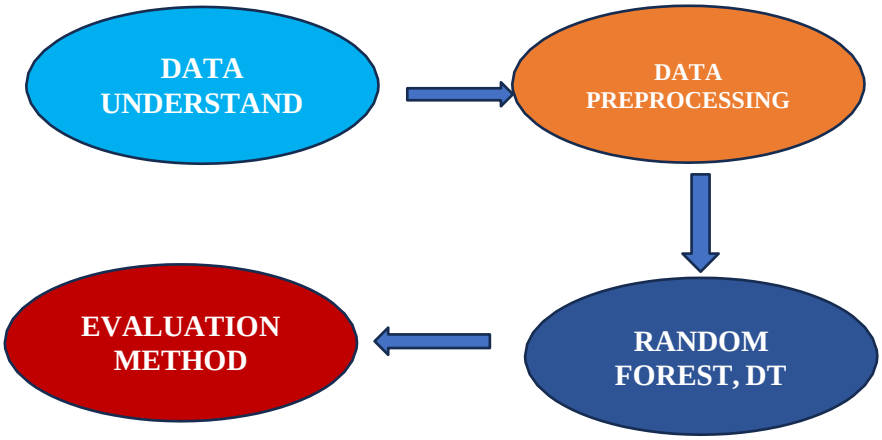

Fig 1. Proposed System

Figure 1 illustrates the system proposed for the methodology of the used car price prediction using random forest algorithm involves the following steps:

3.1 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.

3.2 Dataset

The vehicle dataset, created by Nehal Birla. This dataset contains used car data that [9] can be used to train machine learning models to predict used car price [9]. 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.,

3.3 Pre-Processing

The following elements are included in the pre-processing technique through Figure 2, Data Cleaning and Handling Missing Values: In this stage, the dataset is examined for errors, inconsistencies, and outliers. In order to guarantee that the dataset is complete and prepared for analysis, it also entails handling missing values. Feature Normalization: By using feature scaling techniques like Min-Max scaling and Z-score standardization, features are made to be on a similarscale 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 accuracyand generalizability, the training set is used to train the model, and the validation set is used to adjust its hyper parameters. Accurate model evaluation and a decreased chance of overfitting are two benefits of proper splitting.

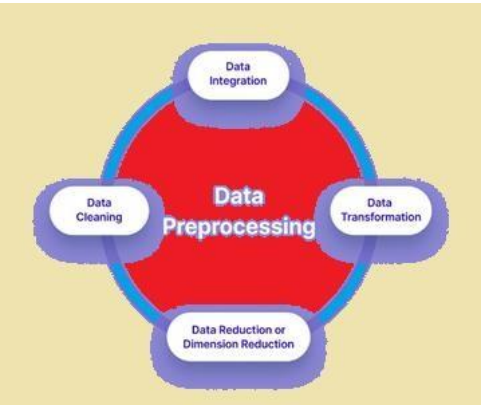

Fig 2. Pre Processing

3.4 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's capacity to distinguish used car price and actual price is the first step in the prediction of used car. 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. In order to lower computational complexity and boost model performance, feature selection also concentrates on trimming the feature set to keep just the most discriminative and informative qualities. The prudent selection of these features is critical 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 insurance business.

3.5 Random Forest Model Design

A pivotal advancement in the development of a robust and adaptable price prediction system for used vehicles lies in the creation and implementation of a Random Forest model. This specialized framework has been meticulously engineered to tackle the unique challenges associated with automotive pricing. The Random Forest model represents a highly versatile and potent ensemble learning technique, adept at handling both classification and regression tasks.

At its core, the Random Forest model comprises individual decision trees. Each decision tree functions as a flowchart-like structure, wherein each internal node signifies a test on a specific attribute, each branch denotesthe outcome of that test, and each leaf node embodies either a class label or a numerical value. The RandomForest leverages an ensemble learning approach, constructing multiple decision trees during the training phase, thereby enhancing predictive accuracy and resilience against overfitting. Figure 3 illustrates the implementation of Random Forest Algorithm.

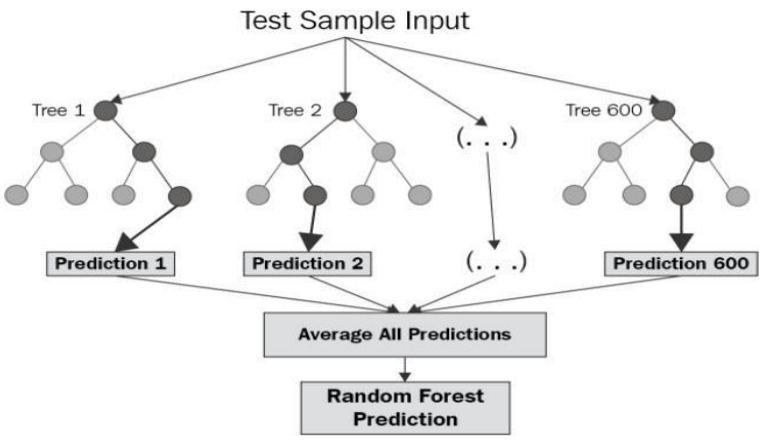

Fig 3. : Implementation of Random Forest Algorithm

3.6 Prediction and Classification

It is critical to use prediction and classification in the field of car price 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 form the dataset or creating the new features. Train a Random Forest classification model using the training dataset.

3.7 Performance metrics

For used car price prediction using the Random Forest algorithm, several performance metrics can be used to evaluate the model's in predicting prices accurately. Here we are using a Mean Absolute Error (MAE) it 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 the model's accuracy and suitability for the task. Machine learning and classification task evaluation commonly assess a model's efficiency.

4. Results and Discussion

In this study on used car price prediction, we used Random Forest Algorithm, an improved version of the other algorithms, to address the urgent problem of used car price 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 detailed explanation of the situation at hand, supported by a thorough examination of the current literature on various price prediction methodologies. Utilizing 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 thetraining and assessment data were tailored to the unique requirements of used car price prediction. Our goal in improving the data quality was to establish a strong foundation for the Random Forest technique that would be usedlater.

We thoroughly compared the performance of the Random Forest Model Figure 4, with the other models and described all of its nuances. Our extensive testing yielded compelling evidence, as Random Forest model significantly improved 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 efficiencyof price prediction. Our goal in comprehending these subtleties was to offer a more comprehensive viewpoint on thepractical uses of the algorithm and its potential to enhance insurance fraud prevention tactics. Furthermore, our studydemonstrated the computational effectiveness of the Random Forest technique and emphasized the critical significance that particular traits play in the field of price prediction. In order to effectively tackle the challenge of developing prediction strategies, the cars business is dependent on computational tools that operate with greater efficiency.

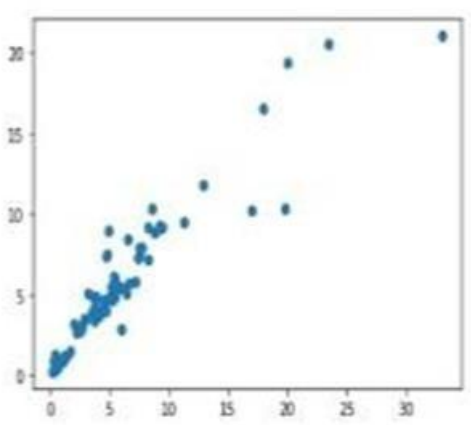

Fig 4. RFM Method

5. Conclusion

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

References

1. Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh, Predictioncar prices using quantify qualitative data and knowledge-based system‖, IEEE – 2020.
2. Praful Rane, Deep Pandya, Dhawal Kotak, ―Used Car Price Prediction‖, International ResearchJournal of Engineering and Technology, Apr 2021.

3. Anamika Das Mou, Protap Kumar Saha, Sumiya Akter Nisher, Anirban Saha, ―A ComprehensiveStudy 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 RecentTechnology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue- 5S, January 2020.
6. Challa Lakshmi Lasya, S. Pooja, S. Jeyashree, C. Ambhika, G. Eswari. "Forecasting PreOwnedCar Prices Using Machine Learning" , 2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), 2023
7. www.ijfmr.com
8. Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain,Computing and Security - Volume 2" , CRC Press, 2023
9. Prasenjit Chatterjee, Morteza Yazdani, Francisco Fernández-Navarro, Javier Pérez Rodríguez."Machine Learning Algorithms and Applications in Engineering" , CRC Press, 2023