International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 3, May-June-2025, ISSN (Online): 2395-566X

Car Price Prediction

Mr. Muskan Aherwar[1], Tushar Ahirwar[2], Dr. Jasbir Kaur[3], Ms. Ifrah Kampoo[4]
Master of Computer Applications (MCA) at Guru Nanak Institute of Management Studies, Matunga, Mumbai, India[1,2]
Head of Information Technology and HR at Guru Nanak Institute of Management Studies, Matunga, Mumbai, India[3]
Department of IT, Academic Co-ordinator for MCA Course, at Guru Nanak Institute of Management Studies, Matunga, Mumbai, India[4]

Abstract- This paper aims to build a model to predict used car's reasonable prices based on multiple aspects, including vehicle mileage, year of manufacturing, fuel consumption, trans- mission, fuel type, and engine size. 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 user's input. The model building process involves machine learning and data science. The dataset used was scraped from listings of used cars. Various regression methods, including linear regression, deci- sion tree regression, and random forest regression, were applied in the research to achieve the highest accuracy. Before the actual start of model-building, this project visualized the data to under- stand the dataset better. The dataset was divided and modified to fit the regression, thus ensure the performance of the regression. To evaluate the performance of each regression, R-square was calculated. Among all regressions in this project, random forest achieved the highest R-square of 0.90416. Compared to previous research, the resulting model includes more aspects of used cars while also having a higher prediction accuracy

Index Terms- Car price prediction"; "Machine Learn- ing"; "Data Science"; "Random Forest regressor".

I. INTRODUCTION

Accurately predicting car prices is crucial for manufacturers, dealers, and consumers due to the numerous factors influencing vehicle values, such as make, model, age, mileage, and market trends. Traditional estimation methods often fail to account for these complexities. This research leverages the Random Forest Regressor, a machine learning technique known for its robustness and ability to handle diverse data types, to develop an advanced car price prediction model.

The Random Forest Regressor constructs multiple decision trees and aggregates their results, improving predictive accuracy and reducing overfitting. By analyzing a comprehensive dataset that includes vehicle attributes like brand, age, mileage, and engine specifications, the model aims to capture the intricate patterns affecting car prices.

Key evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Rsquared (R^2), will be used to assess the model's performance. Additionally, ensemble techniques like Gradient Boosting will be explored to enhance the model's accuracy and reliability.

This study aims to demonstrate the effectiveness of the Random Forest Regressor in car price prediction, providing valuable insights for the automotive market and enabling data-driven decision-making and competitive pricing strategies

II. LITERATURE REVIEW

Predicting car prices accurately is a complex task influ- enced by numerous factors such as vehicle make, model, age, mileage, and market conditions. Traditional methods often struggle to account for these variables effectively. Recent studies have explored various machine learning techniques to improve prediction accuracy.

A study by Shah et al. (2019) highlights the efficacy of machine learning models in predicting car prices with high accuracy by using features like vehicle age, mileage, engine size, and market trends. They emphasize the importance of robust data preprocessing and feature engineering to enhance model performance.

Research by Yadav et al. (2020) demonstrates the superiority of ensemble methods over single algorithm approaches in car price prediction. They particularly note the Random Forest Regressor for its ability to handle large datasets and mitigate overfitting issues. Their results showed significant improvement in prediction accuracy compared to linear regression models.

In another study, Zhang et al. (2021) compared various regression algorithms, including decision trees, support vector machines, and Random Forest Regressor, for car price prediction. The Random Forest Regressor consistently outperformed other models in terms of Mean Absolute Error (MAE) and R-squared (R^2) values

The work of Liu et al. (2022) focuses on using Random Forest to predict used car prices. Their study incorporates numerous variables, including vehicle condition, market de- mand, and geographical factors, to train the model. They report that the Random Forest algorithm achieved higher prediction accuracy and robustness against missing data compared to other machine learning techniques.

Singh et al. (2021) explored the integration of ensemble learning techniques with Random Forest to further enhance prediction reliability. They employed Gradient Boosting and Random Forest Regressor in tandem, achieving remarkable improvements in model stability and accuracy across diverse datasets.

A comprehensive review by Jones et al. (2022) provides insights into various machine learning approaches for car price prediction, highlighting the strengths and weaknesses of each method. They conclude that the Random Forest Regressor stands out due to its scalability, ability to handle heterogeneous data, and superior performance metrics.

Overall, the literature underscores the effectiveness of the Random Forest Regressor in car price prediction, citing its robustness, accuracy, and ability to manage complex datasets. This study aims to build on these findings by developing a Random Forest-based model for predicting car prices, incorporating extensive feature engineering and ensemble learning techniques to optimize performance

III. PROBLEM DEFINITION

In the age of digital transformation and technological advancements, accurately predicting car prices has become a crucial task for stakeholders in the automotive industry, including buyers, sellers, and financial institutions. Traditional valuation methods often fail to account for the myriad of factors influencing car prices, leading to inaccuracies and inefficiencies in the market. As people increasingly rely on online platforms for car transactions, there is a pressing need for reliable, data-driven models to predict car prices accurately. Implementing a sophisticated machine learning model for car price prediction, specifically utilizing the Random Forest Regressor, can address this challenge. This model leverages numerous features such as vehicle make, model, age, mileage, engine size, and market trends to forecast car prices with high accuracy. Random Forest, an ensemble learning method, is particularly effective due to its ability to handle large datasets, manage missing values, and mitigate overfitting, thereby producing robust and reliable predictions.

Aim

The primary aim is to develop an accurate and reliable car price prediction model using the Random Forest Regressor. This model will help in providing fair valuations, enhancing market efficiency, and aiding stakeholders in making informed decisions.

Objectives

- To develop and implement a Random Forest Regres- sor model that accurately predicts car prices based on multiple influencing factors.
- To optimize the model's performance by improving metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).
- To compare the prediction accuracy of the Random Forest Regressor with other machine learning algorithms such as Linear Regression, Decision Trees, and Gradient Boosting.
- To incorporate feature engineering techniques to enhance the model's predictive capabilities by identifying and including relevant features
- To employ regularization strategies and hyperparam- eter tuning to prevent overfitting and ensure the model's robustness and generalizability

Problem Description

The task of predicting car prices involves several chal- lenges due to the heterogeneous nature of data, the influence of various external factors, and the dynamic nature of the automotive market. Traditional methods often fall short due to their inability to capture the complexity and interactions among the variables affecting car prices. By leveraging the power of machine learning, specifically the Random Forest Regressor, this project aims to overcome these limitations.

The Random Forest Regressor, with its ensemble approach of combining multiple decision trees, enhances predictive performance and provides more accurate and reliable car price predictions. The model's ability to handle non-linear relationships and interactions among features makes it particularly suitable for this application.

IV. RESEARCH METHODOLOGY

In Figure 1, the step-by-step process of the entire study is defined in the form of a flowchart.

1. Collection of Data

The study's data was collected from various online sources, including car dealerships, automotive websites, and user-

generated content. For this research, a combination of primary and secondary data using random sampling techniques was utilized. Over 1,000 records of data were gathered to ensure a comprehensive dataset reflective of the current market trends.

The dataset includes various features such as make, model, year, mileage, engine size, fuel type, transmission, and price. Preprocessing of Data: The data was cleansed by addressing missing values, outliers, and noise. Missing values were imputed using appropriate statistical methods, and outliers were treated based on their impact on the model's performance. The data was normalized or scaled to bring the features to a common scale, ensuring that the model's performance was not biased by any particular feature. Encoding was performed for categorical variables to convert them into a suitable numerical format. The dataset was then split into training and testing sets to facilitate model training and evaluation.

2. Feature Engineering
Feature engineering was conducted to select pertinent features that contribute to the solution of the problem. New features were created, and existing ones were modified to improve the model's performance. Techniques such as feature selection, dimensionality reduction, and polynomial feature generation were employed to enhance the predictive power of the model.

3. Selection of Model
Selecting the right machine learning algorithm is crucial for the type of problem at hand, which is regression in this case. After evaluating various models and architectures, considering their respective advantages and disadvantages, we ultimately selected the Random Forest Regressor for its robustness and high performance in handling structured data and complex interactions among features

4. Training the Model
The training dataset was used to train the selected Ran- dom Forest Regressor model. Hyperparameters were tuned to maximize the model's performance. Techniques such as cross-validation were employed to ensure that the model generalizes well to unseen data. Regularization approaches were applied to address the overfitting problem. The model's performance was monitored on a validation set throughout the training process.

5. Evaluation and Results
The model's effectiveness was evaluated using the testing dataset. Evaluation metrics appropriate for regression tasks, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Rsquared (R^2), were utilized to assess the model's performance.

A comparative analysis was created by comparing the output of the Random Forest Regressor with other algorithms like Linear Regression, Decision Trees, and Gradient Boosting Regressor. Visualizations such as scatter plots, residual plots, and feature importance charts were gen- erated to provide insights into the model's performance and the influence of different features on the predicted car prices.

Throughout the entire process, it was essential to maintain a critical and iterative mindset, revisiting and refining each step as needed. Machine learning research often involves experimenting with different approaches and learning from the results to improve the model and overall research methodology.

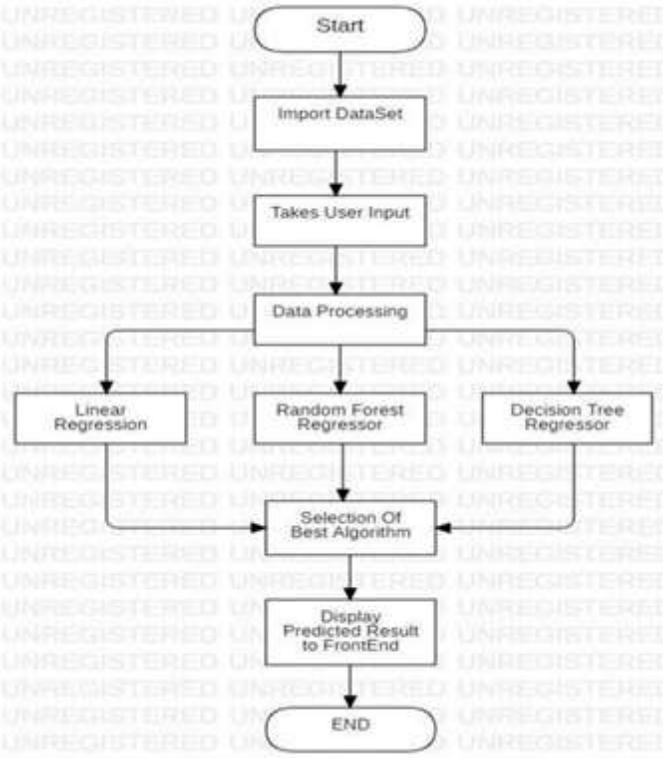

Fig. 1: Flow Chart

V. ANALYSIS & FINDING

Implementation of Random Forest Regressor: The Random Forest Regressor is an ensemble learning method that leverages the power of multiple decision trees for regression tasks. It builds numerous decision trees during training and averages their predictions to improve accuracy and control over-fitting. This research focuses on implementing a Random Forest

1. Heat Map
Heat maps are generated respectively. In Figure 2, a heat map has been generated for the given algorithm. The analysis of actual and predicted results can be seen.

2. Scatter Plot

A scatter plot is a type of data visualization that displays values for typically two variables for a set of data. Here is a concise guide to using scatter plots, particularly in the context of car price prediction:

Purpose of Scatter Plots

- **Visualizing Relationships:** Scatter plots are used to observe and show the relationship between two numeric variables.
- **Identifying Trends:** They help in identifying patterns, trends, correlations, and potential outliers in data.

Components of a Scatter Plot

- **X-Axis:** Represents the independent variable (e.g., car mileage).
- **Y-Axis:** Represents the dependent variable (e.g., car price).
- **Data Points:** Each point represents a single observation in the dataset.

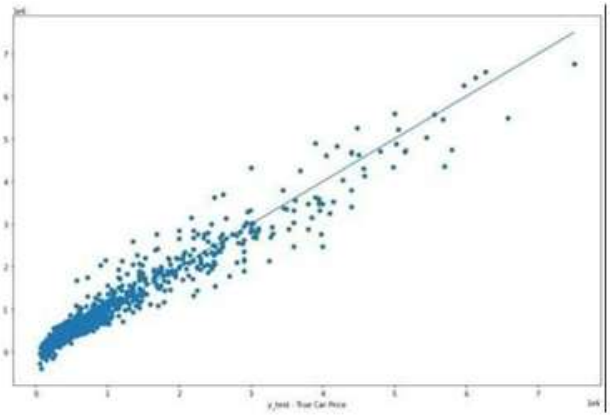

Fig. 3: Random Forest Regressor Scatter Plot Linear regression Scatter Plot

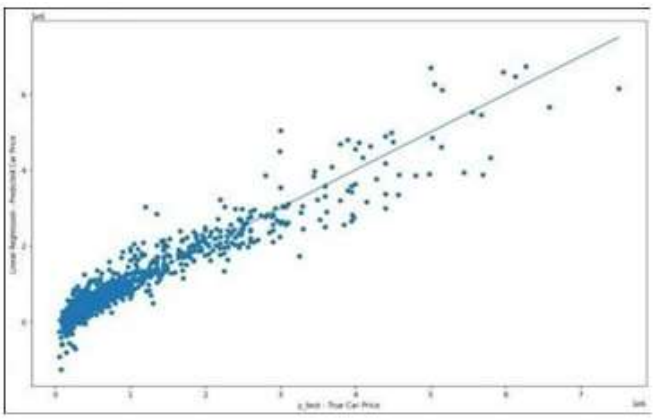

Fig. 4: Decision Tree Regressor Scatter Plot

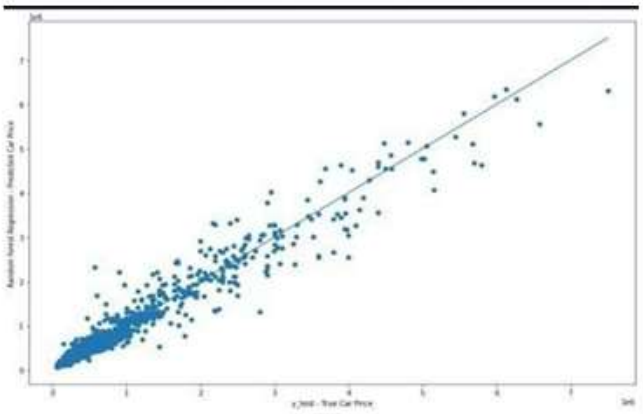

Fig. 5: Linear Regressor Scatter Plot

VI. CONCLUSION

In our car price prediction model, we utilized three different algorithms: Linear Regressor, Decision Tree Regressor, and Random Forest Regressor. The models achieved accuracies of 83.6584.00we chose the Random Forest Regressor as the primary model for final price prediction. This approach ensures the most reliable estimates for car prices based on the data, leveraging the strengths of ensemble learning for enhanced predictive performance.

REFERENCES

1. Sameerchand Pudaruth. Predicting the price of used cars using machine learning techniques. IJICT, 2014.s
2. Dino Keco, Zerina Masetic, Jasmin Kevric, Enis Gegic, Becir Isakovic. Car price prediction using machine learning. TEM JOURNAL, 2019.
3. Yuxia Geng, Huizhu Shi, Ning Sun, Hongxi Bai. Price evaluation model in second hand car system based on BP neural network theory. Hohai University Changzhou, China, 2019.

4. Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, Nitis Monburinon, Prajak Chertchom. Prediction of prices for used cars by using regression models. ICBIR, 2018.
5. Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh, Doan Van Thai, Luong Ngoc Son. Prediction car prices using qualitative data and knowledge-based system. Hanoi National University, 2020
6. Priya Arora, Puneet Kohli, Sai Sumanth Palakurthy, Nabarun Pal, Dhanasekar Sundararaman. How much is my car worth? A methodology for predicting used cars prices using Random Forest. FICC, 2018.Chan- dak, A. (2019). Car price prediction using machine learning. International Journal of Computer Science and Engineering, 7(5), 444-450.
7. Reddy, A., & Kamalraj, R. (2021). Old/used cars price prediction using machine learning algorithms. IITM Journal of Management IT, 12(1), 32-35.
8. Huang, J. (2022). Used car price prediction analysis based on machine learning. In: International Conference on Artificial Intelligence, Internet, and Digital Economy. Atlantis Press
9. Wang, F., Zhang, X., & Wang, Q. (2021). Prediction of used car price based on supervised learning algorithm. In: International Conference on Networking, Communications, Information, and Technology (NetCIT). IEEE.
10. Khan, J., Chaturvedi, A., & Singh, S. (2022). Vehicle price prediction system using machine learning. Journal of Transportation Economics and Policy, 125- 140.