Journal of Theoretical and Applied Information Technology
15th May 2025. Vol.103. No.9
© Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SMART PRICING SOLUTIONS FOR USED CARS USING DECISION TREE AND RANDOM FOREST MODEL

PRATAP KUMAR CHAMUPATY[1], UDITA J. MONANI[2], BISWAJIT DAS[3], PRASANT KUMAR PATTNAIK[4]

[1,3]Kalinga School of Management (KSOM), Kalinga Institute of Industrial Technology - DU, India
[2,4]School of Computer Engineering, Kalinga Institute of Industrial Technology - DU, India

ABSTRACT

The used car market is vast and is affected by many parameters, making it is essential to predict prices of used cars. This work recommends customers in making informed decisions while planning to purchase used car. Presently, Machine Learning emerged as the most effective methods for prediction of prices of used cars considering correlated factors such as mileage and vehicle age. This work applied two machine learning models includes Decision Tree and Random Forests on large used car dataset so collected from Kaggle. The prediction accuracy and precision of these models thoroughly evaluated. The comparison of results output of two models is made to find out the effectiveness and reliability for analyzing large database. The study concludes that Random forest model emerged as the best model in comparison to Decision Tree model in prediction accuracy.

Keywords: *Machine learning; random forest; decision tree; used/pre owned cars; regression; predictions*

1. INTRODUCTION

All new cars from a particular brand, model and product year are offered at a uniform retail price, excluding any additional features. This pricing is set by the manufacturer. The buyers are confidently purchase a new car without uncertainty, as models are available with same retail price, ensuring reliability with latest technology and guaranteed quality by the manufacturers. There is a rapid growth of the used car market in India due to expanding size of middle class, characterized by increased disposable income, increased demand for personal transportation, easy availability of financing options and shorter new car replacement cycles. As per the data collected from Daswel Auto Statistics, the sales of new cars to used cars presented in Figure-1. The figure shown in FY2025 is a projection one basing on current sales growth of used car and new car.

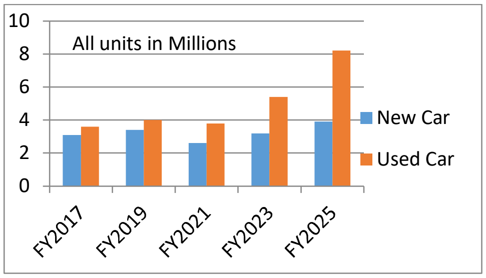


Figure-1. New Car sales to Used Car Sales in India (in millions)

The used car market is experiencing a significant growth, with its value nearly twice in recent years. The adoption of new technologies and online portals such as CarDekho, CarTrade, CarWale, OLX, Spinny, Droom, Car24, QuikCars as well as manufacturer used car sales outlets such as Maruti True Value, [1] Hyundai Promise, Mahindra First-Choice and Honda Auto Terrace, has facilitated access to vital information for both buyers and sellers regarding the trends and factors that affect used car valuations in market. Among these portals, the capability to accurately estimate the price of a used car stands out as a crucial feature. Due to huge price fluctuation in the market and intense competition among sellers, it necessitate for a reliable tool that can swiftly and accurately predict used car prices [2].

The used car market is growing rapidly, nearly doubling its value in recent years. Online platforms like Carsome and OLX have helped buyers and sellers stay informed about the trends that affect used car prices ("Used Cars Price Prediction Using Supervised Learning Techniques,") [2]. Machine Learning algorithms can predict a car's retail value based on specific features [3]. As business and activities grow, cars have become essential for many people. New vehicles come with advanced features, which leads to higher prices. Because of this, many individuals opt for used cars that are still in good condition. One effective method for predicting car prices is through Machine Learning techniques [4].

Different websites use unique algorithms to set prices for used cars, so there is no standard method for pricing. By developing statistical models to predict these prices, you can estimate a car's value without entering information on a specific website. It is wise able to apply different prediction models to estimate the retail price of a used car and have to evaluate their level of accuracy [5].Data mining includes several classification algorithms, such as K-Nearest Neighbor (KNN), Decision Tree, Random Forest, Support Vector Machines (SVM), Recurrent Neural Network (RNN), and Conventional Neural Network (CNN). This study focused on the Random Forest and Decision Tree methods. Both of these methods are used to classify datasets effectively. Random Forest enhances accuracy by randomly selecting child nodes for each parent node. The results from each tree are combined, and the most frequent classification is chosen. The Decision Tree is a widely used method for classification in supervised learning. It follows a top-down, step-by-step approach. A decision tree consists of a root node, several decision nodes, and leaf nodes. The root node indicates the most significant attribute of the data set that helps in making the best prediction [6].

The scope of this work includes how machine learning can analyze the selling price of used cars. Currently promoting the use of environmentally friendly vehicles by offering discounts impacts the sales of used cars. We are here employs machine learning techniques, specifically random forest and decision tree methods, to improve the accuracy of the analysis.

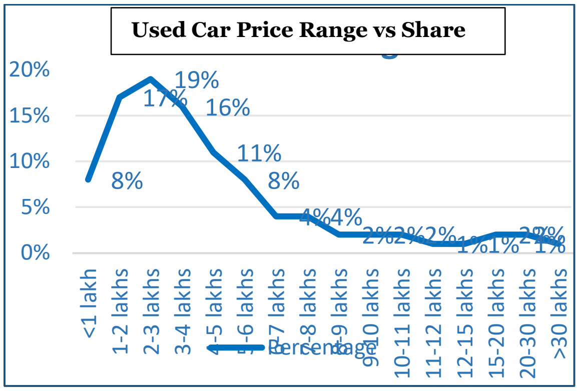


Figure-2. Used Car Price Analysis (Source: das Wale Auto Statistics on Used Car vs New Car)

On analyzing the dataset, it is evident that the used car Price Analysis graph (Fig-2) that more than 70% of used car sales come from price range below Rs.5 lakhs and about 90% of the used sales happens at a price bracket below Rs.10 lakhs.

One of the main assumptions in this study is that the dataset applied to predict used car prices is representative of actual market conditions, including varied factors like brand, model, production year, mileage, fuel type, transmission, and condition. Nevertheless, some limitations apply. The model does not consider external economic forces such as inflation, fluctuations in fuel prices, government decisions, and local demand-supply patterns, which have a profound effect on the prices of cars. The influence of non-standardized adjustments, after-sales accessories, and localized bargaining patterns is also not directly captured in the dataset, which may create price discrepancies. The research is based on the assumption that information collected from websites like CarDekho, OLX, and manufacturer-authorised outlets is genuine business transactions and unpolluted by price manipulation. Yet another shortcoming is that machine learning algorithms, especially Decision Tree and Random Forest, are based on past data patterns, which fail to keep pace with abrupt changes in the market or new trends such as increasing use of electric vehicles (EVs) and environment-conscious price inflation. Additionally, the research fails to investigate ensemble learning using deep learning models or real-time integration of data, which would boost predictive performance in subsequent studies. In spite of these shortcomings, the research gives a firm basis for automatic price estimation and identifies the contribution of machine learning to explaining and predicting the used car market dynamics in India.

2. RESEARCH METHOD

To make this study reproducible, a systematic research method protocol was adopted, which included both qualitative and quantitative approaches. The prediction of used car prices based on Decision Tree and Random Forest models was the objective of the study, and the whole process was systematically segregated into four major phases: dataset collection and comprehension, data preprocessing and feature engineering, model training and validation, and comparative performance analysis. Every phase was intended to increase the precision and consistency of the predictions and facilitate replication and verification of the results. The first phase, dataset acquisition and understanding, consisted of obtaining the dataset from publicly accessible data repositories like Kaggle.

The dataset contained essential features like car make, model, manufacturing year, mileage, fuel type, transmission, seating capacity, and overall condition. A preliminary exploratory data analysis (EDA) was performed to evaluate the completeness, distribution, and correlation of features, as well as detect possible outliers and inconsistencies in the data. Statistical methods and visualization tools were employed to identify the effect of each feature on price prediction. The second phase, data preprocessing and feature engineering, was designed to pre-process the dataset for training. Missing values were managed through proper imputation methods, including mean/mode substitution for numerical and categorical variables or K-Nearest Neighbors (KNN) imputation when required.

Data was preprocessed by eliminating duplicates and unnecessary variables that did not play a crucial role in predicting price. Feature scaling and encoding methods, including Min-Max normalization and one-hot encoding, were used to maintain uniformity throughout the dataset. Moreover, feature selection techniques such as correlation analysis and Principal Component Analysis (PCA) were applied to remove duplicate attributes and dimensionality reduction. The third phase, model training and validation, consisted of applying Decision Tree and Random Forest algorithms with Python's Scikit-Learn library. The dataset was divided into training (80%) and test (20%) subsets to test model generalization. The Decision Tree model was trained with different hyperparameters like maximum depth, minimum samples per leaf, and criterion choice (Gini impurity vs. entropy).

For the Random Forest model, a collection of decision trees was trained using bootstrap sampling, with the number of estimators, maximum depth, and feature subsets varied to maximize performance. The models were also tested based on the important performance measures of accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. The fourth stage, comparative performance analysis, was to check how effective each model was at predicting used car prices in each of the categories of prices (low, medium, and high).

By adopting this systematic research approach, the research is made repeatable and enables other researchers to verify and extend the results. Future research can extend this framework by using real-time pricing information, other machine learning algorithms, and domain-specific variables influencing used car valuation.

2.1. Machine Learning

Machine Learning (ML) is a part of Artificial Intelligence (AI) that enables systems to make decisions based on data. By using a set of training data, the ML model is refined to create an effective predictive model. This model can then analyze new data in the future [7]. Machine learning is used for classifying data. It can identify patterns in data, with or without prior training. Humans can program computers to help machines learn from the information they receive or recognize patterns on their own [8].

2.1.1. Random Forest

Random Forest (RF) is an algorithm that uses a recursive binary splitting method to build trees for classification and regression purposes [9, 10]. This algorithm offers several benefits, including low error rates, effective classification performance, and the ability to manage large datasets efficiently. It also serves as a strong tool for estimating missing data. Random Forest creates multiple independent trees by randomly selecting subsets from the training sample and input variables for each node [11].

The Random Forest (RF) method enhances accuracy by generating child nodes randomly for each node [12]. This technique builds a decision tree made up of root nodes, internal nodes, and leaf nodes, selecting attributes and data randomly according to certain rules [13]. The root node is the topmost point in the decision tree. Internal nodes are where branches split, each having at least two outputs and one input. The leaf node, or terminal node, is the final point with a single input and no output. To start the decision tree, the entropy value is calculated to assess the impurity level of the attribute and to obtain information value. The formula for calculating entropy is shown in equation 1, while equation 2 is used for calculating information gain [14].

$$Entropy\ (Y) = -\sum_{i} p\ (c|Y) log_{2}\, p(c|Y)\ (1)$$

Where Y is the set of cases and p (c|Y) is the proportion of Y values to class c:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$Entropy\,(Y)\;= \\ -\sum Y_{v}Y_{\alpha}Entopy\,(Y_{v})\,)_{vtValues\,(\alpha)}\,(2)$$

Where Values(a) is all possible values in case set a. Yv is a subclass of Y with class v which is related to class a. Yes are all values that correspond to a

2.1.2. Decision tree

A Decision Tree is a popular algorithm for making choices. It organizes problem-solving criteria or nodes into a tree-like structure. Essentially, it serves as a predictive model using a hierarchy [15]. Each tree consists of branches that represent attributes. These branches guide users to the next step until reaching an endpoint, known as a leaf. The data used in a Decision Tree is typically formatted in a table, containing various attributes and records [16].

3. RESULTS AND ANALYSIS

The paper uses the research methodology based on four stages, which can be seen as follows:

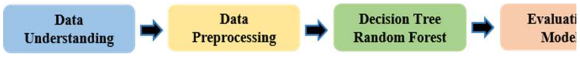
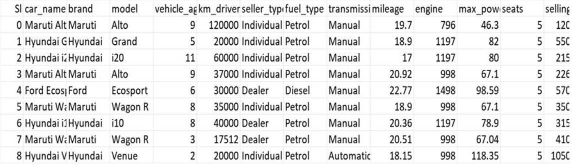

Data Understanding → Data Preprocessing → Decision Tree Random Forest → Evaluati Mode

Figure-3. Flow Diagram Of The Proposed Research Methodology

3.1. Data Understanding

This stage is the data collection stage. Here, we analyze the data to understand what we will use. We identify issues by examining the details in the data and searching for interesting insights. The data for this study comes from a Kaggle dataset. We used secondary data collection methods from kaggle.com, which includes an unnamed serial number, Car_name, brand, model, vehicle_age, km_driver, seller_type, fuel_type, transmission, mileage, engine, max_power, seats, selling_price as shown in Figure-3.



Sl	car_name	brand	model	vehicle_ag	km_driver	seller_typ	fuel_type	transmissi	mileage	engine	max_pow	seats	selling
0	Maruti Alt	Maruti	Alto	9	120000	Individual	Petrol	Manual	19.7	796	46.3	5	120
1	Hyundai G	Hyundai	Grand	5	20000	Individual	Petrol	Manual	18.9	1197	82	5	550
2	Hyundai i2	Hyundai	i20	11	60000	Individual	Petrol	Manual	17	1197	80	5	215
3	Maruti Alt	Maruti	Alto	9	37000	Individual	Petrol	Manual	20.92	998	67.1	5	226
4	Ford Ecos	Ford	Ecosport	6	30000	Dealer	Diesel	Manual	22.77	1498	98.59	5	570
5	Maruti Wa	Maruti	Wagon R	8	35000	Individual	Petrol	Manual	18.9	998	67.1	5	350
6	Hyundai i1	Hyundai	i10	8	40000	Dealer	Petrol	Manual	20.36	1197	78.9	5	315
7	Maruti Wa	Maruti	Wagon R	3	17512	Dealer	Petrol	Manual	20.51	998	67.04	5	410
8	Hyundai V	Hyundai	Venue	2	20000	Individual	Petrol	Automatic	18.15	998	118.35	5	1050

Figure-4. Sample Dataset

Here we presents data in Figure-4. from a Kaggle datasets that focuses on predicting car prices. This datasets identifies which factors are important for estimating those prices and how effectively each factor reflects a car's value. It include fifteen attributes and contains 15417 rows of data. After processing the total data set and removing outliers, the final Decision Tree Visualization presented in Figure-5.

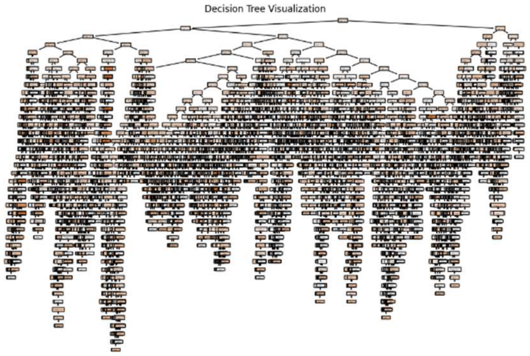


Figure-5. Decision Tree Visualization

On comprehensive analysis of data set, the percentage of branded cars as per the sales records is presented in Figure-6. Out of those, the most preferred sales of used car for top manufacturers are as follows:

1. Maruti - 32.39%
2. Hyundai- 19.35%
3. Honda- 9.64%
4. Mahindra- 6.56%
5. Toyota- 5.15%
6. Ford- 5.13%

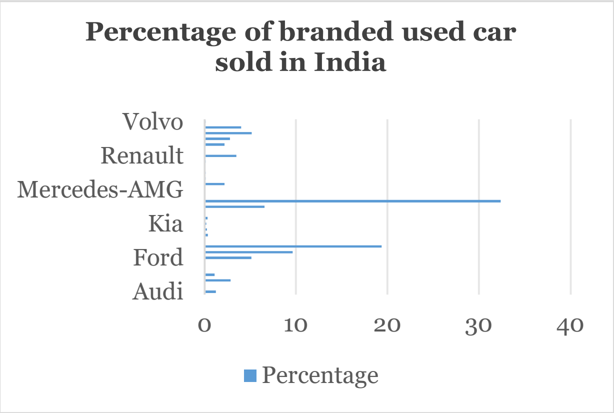


Figure-6. Percentage Of Branded Used Cars Sold In India (From Kaggle Dataset)

3.2. Data Preprocessing

Data preprocessing techniques are essential for getting raw data ready for analysis or machine learning. This step ensures that the data is clean, consistent, and ready for use. This research utilizes several preprocessing techniques, including data cleansing, aggregation, and checking for missing values.

On analysis, the graphical representation between selling prices and KM travelled for used cars is reflected in Figure-7.

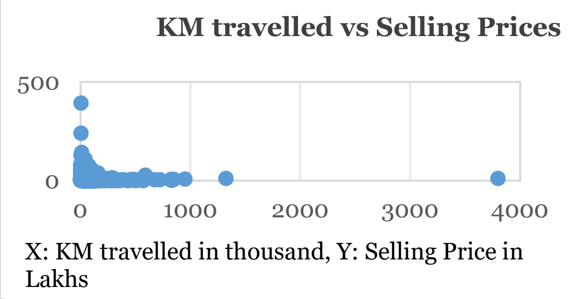


Figure-7. Relationship Between The Price Of Vehicle And Kilometer Travelled

3.3. Decision tree algorithm

In Figure-8, we are trying to demonstrate how to use the decision tree method with RapidMiner tools. We start the analysis by importing the car price datasets using read CSV. Next, the dataset is divided basing on the available variables. After filtering the data, it is linked to the decision tree. Then we applied the model and connects the performance to generate accurate results for each datasets.

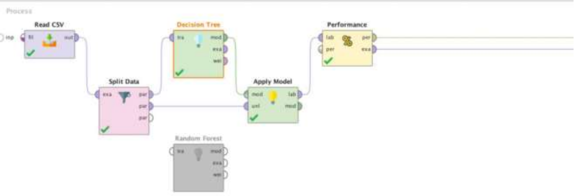


Figure-8. Process of Implementing the Decision Tree algorithm

The decision tree model's performance in predicting the prices and conditions of used cars based on user preference is judged through error curve. It shows decision model's error rate pn validation of chosen dataset. It decreases initially as the decision tree accepts meaningful on training data set and increases after a certain level of complexity tends to over fit to the training datasets which results in a decline in its ability to generalize. Absolute error assesses the prediction accuracy of used car prices. The optimal point on the error curve is reached when the validation error is at its lowest, achieving a balance between complexity and accuracy. The error curve is reflected in Figure-9.

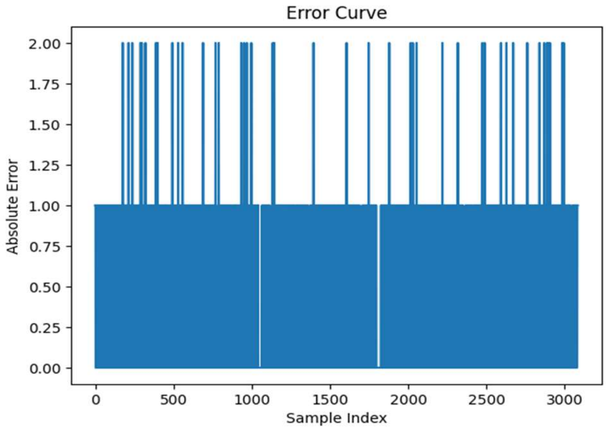


Figure-9. Error Curve On Sample Index In Decision Tree Algorithm

In purchase of a used car, each buyer has its own choice and importance of various features such as age, mileage, brand, fuel type, transmission, condition, ownership history, locations. The used car features such as vehicle age and distance covered are significantly contribute for price predictions. The feature importance for price predictions of used car is presented in Figure-10.

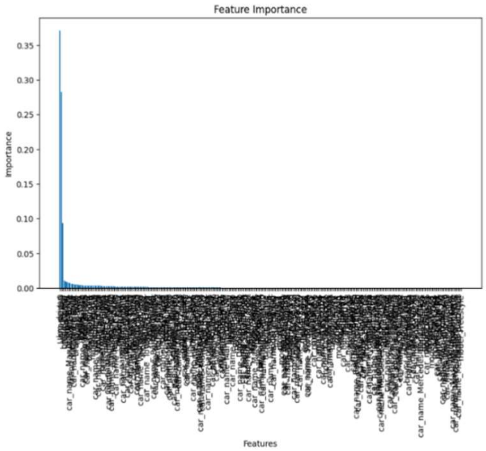


Figure-10.Feature Importance On Decision Tree Algorithm

The Actual Prices vs Predicted price analysis by using decision tree model gives an idea how well the model performs in predicting used car prices. Plotting the prices in graphs visually reflect model's performance. When, both actual and predicted prices plot on ideal line where Y=X, denotes the model is able to predict prices more accurately. The output presented in Figure-11.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

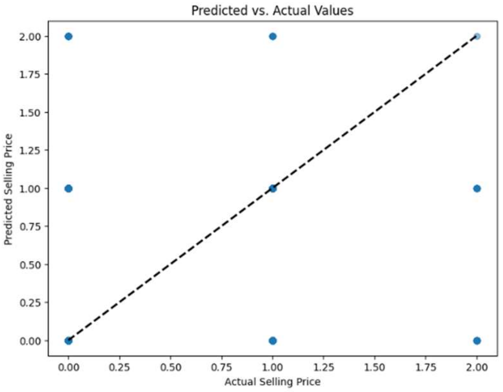


Figure-11. Actual Selling Price Vs Predicted Price On Decision Tree Algorithm

On running the datasets in decision tree model, we obtained the residual plot, which helps us in finding the systematic errors (underpricing or over pricing) trends and scattered around zero. This reflected unbiased prediction in price of used cars and suggest that decision tree so chosen is performing well, which depicted in Figure- 12.

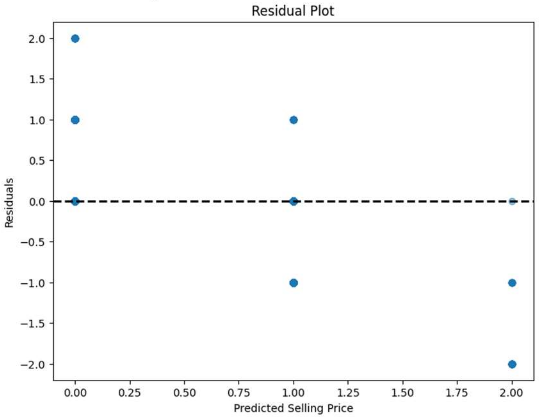


Figure-12. Residual Plot (Residuals Vs Predicted Selling Price) On Decision Tree Algorithm

The Receiver Operating Characteristic (ROC) curve in decision tree model is generally used for classification of data set on some sort of characteristic. This curve reflected the balance between true or positive rate of sensitivity with the false positive rate at various thresh hold values. Since our scope of research work is for prediction of used car prices, which involves regression analysis, so ROC curve does not directly applied being of categorical one. However, we have applied the area under ROC concept to this research problem by converting regression problem into a classification problem. Here we have classified the target variables i.e. prices into low prices and high prices categories. The nature and performance of decision tree model in running data sets are judged basically from its area under the ROC curves. It provides us a single scalar value by testing data sets in decision tree model. As per the criteria, area under curve (AUC) value ranges from 0 to 1. After running the model, AUC so obtained from three data sets are between ranges 0.52 to 0.71, which implied the Poor Performance of datasets analysis by decision tree models in predicting prices of used cars. The analysis reflected in graphical form in Figure-13.

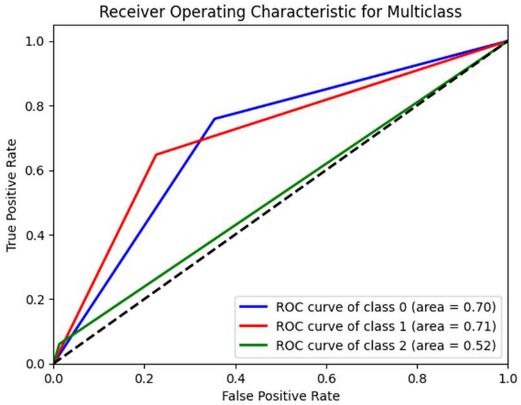


Figure-13. ROC (Receiver Operating Characteristics) On Decision Tree Algorithm

3.4. Random forest algorithm

We are taking ensemble methods to address the limitations associated with individual prediction models. This ensemble consists of multiple models that are trained on subsets of data derived from the same dataset. The primary aim is to achieve greater accuracy compared to a single prediction model. The improved accuracy is attributed to the involvement of several models in the decision-making process with logic that a forecasting model that integrated multiple models enhanced accuracy of predictions [17].

The Random Forests model algorithm is constructed by recognizing the diversity that is crucial for enhancing the model's predictive capabilities. A higher level of diversity within the dataset ensures that each decision tree classifier received a varied mixture of data, thereby facilitating effective training of the base classifiers. The performance of individual tree significantly influences the overall accuracy of the Random Forest model. To enhance the accuracy of this model, we trained each base learner on a diverse subset of data. Here the Random Forest algorithm is trained on various subsets of features instead of solely selecting the most features from the dataset. This introduces randomness and contributed in achieving improved accuracy.

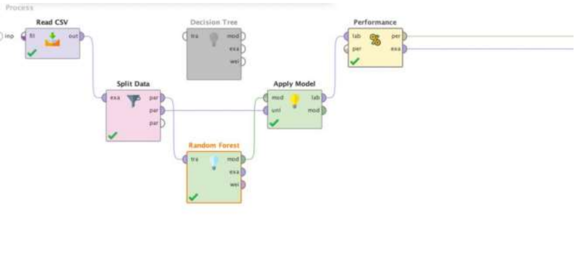


Figure-14. Random Forest Implementation Process

While applying Random Forest model for predicting car prices of used cars, the error curve so obtained by running data sets shows how error changes with different model adjustments, such as number of trees or their depth. This error curve analysis helpful in evaluating the random forest model's effectiveness and fine-tuning its hyper parameters for a better performance. It decreases as the number of trees increases. The fitness of random forest model to data set is judged from its training error and test error levels. After running the model, the test errors stabilizes at a low level signifies that this as an optimal model for generalization and prediction of prices as minimal over-fitting or under-fitting has been made. The Error Curve of Random Forest model analysis is reflected in Figure-15.

The feature importance in a Random Forest model reflects shows how much each factor affects the model's ability in making predictions of prices of used cars.

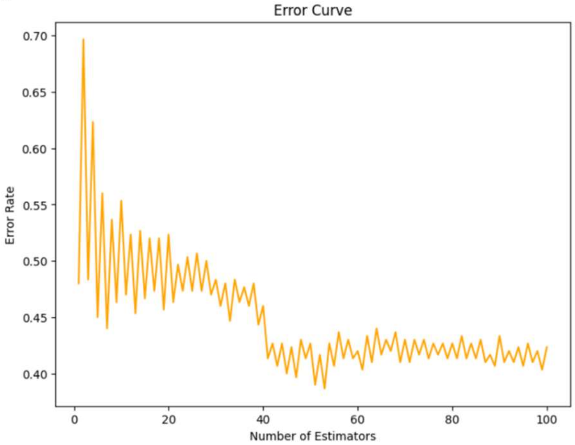


Figure-15. Error Curve On Sample Index On Random Forest Algorithm

When doing predictions, we understand the importance of features to earmarks its importance and significance in price predictions. This feature importance study reveals the importance of each feature on uncertainty in proper predictions of prices. The analysis is presented in Figure-16 which evaluates the random forest models performance or accuracy declines when feature's values are shuffled randomly. A significant declines in graph shows that the feature is important for model in predicting the prices.

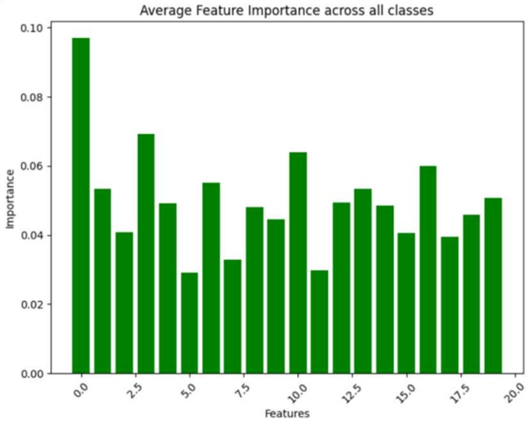


Figure-16. Feature Importance On Random Forest Algorithm

In order to evaluate the performance of Random Forest model in predicting prices of used cars, we herewith comparing the actual selling price with the predicted prices generated through this model by using the viable datasets. This comparison reflected with a scatter plot (actual sales price vs. predicted prices) in Figure-17, wherein points are align along a 45 degree line.

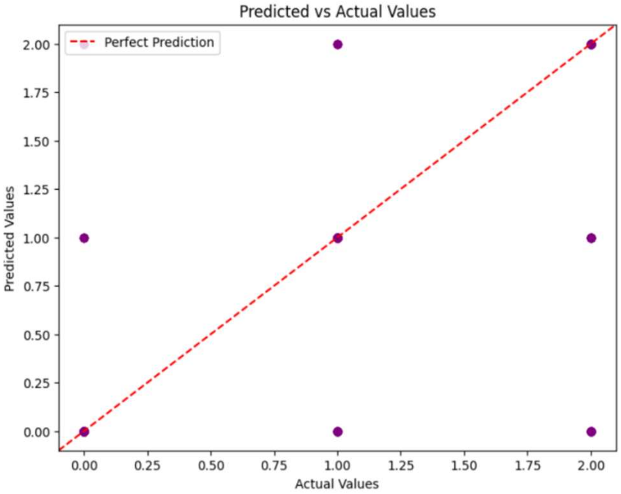


Figure-17. Actual Selling Price Vs Predicted Price On Random Forest Algorithm

We run the data sets in random forest model to evaluate ow well the model is capable enough in predicting prices of used cars. Residuals means difference between the actual prices to the predicted prices. If it is positive residual means actual prices is more than the predicted prices, then it is evaluated as under predicted. Whenever, the actual prices is less than the predicted prices, it returns negative residual value, which denotes it is over predicted. For a well fitted random forest model, the residuals should be randomly distributed around the zero line on the horizontal axis in Figure-18 and the distributions of residuals are constant across all levels of the actual vs predicted prices.

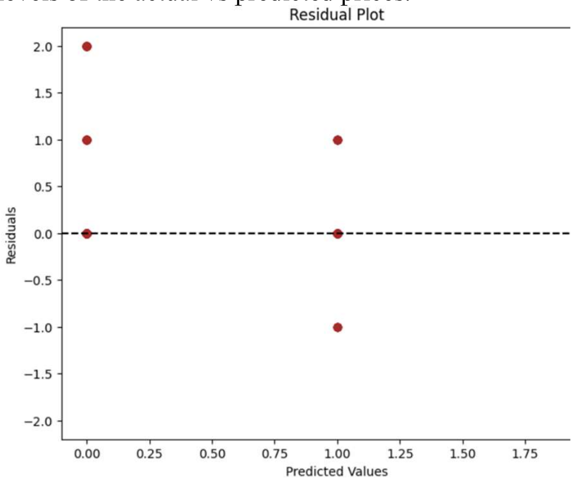


Figure-18. Residual Plot (Residuals Vs Predicted Selling Price) On Decision Tree Algorithm

Generally Receiver Operating Characteristic(ROC) in random forest model used for classification problems but not in regression analysis of predicting the prices of used cars. However, we are doing analysis under ROC by converting regression problems into a classification problems such as low price, medium price and high price target levels in available data sets basing on sales prices. The ROC for multi-class categories (low,medium and high), we compute ROC curve for each class against all others and calculate micro-average area under curve (AUC) under ROC. After running the model, AUC so obtained for Low price ROC is 0.90, Medium Price ROC is 0.91 where as High Price ROC is 0.84. In all three categories, ROC for Low Price, Medium Price and High Prices are close to 1 reflects the random forest model's excellent performance in predicting prices of used cars. The graphical output presented in Figure-19.

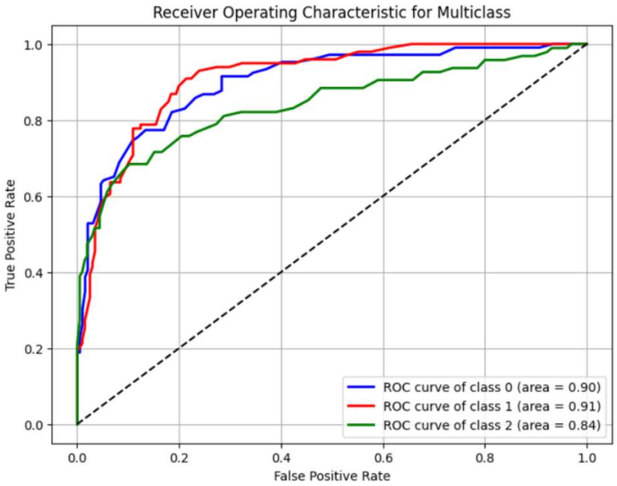


Figure-19. ROC (Receiver Operating Characteristics On Random Forest Algorithm

4. DISCUSSION AND COMPARISON

Comparison with similar published work emphasizes strengths and weaknesses in using Decision Tree and Random Forest models for predicting the price of used cars. The strengths of our research agree with existing work, showing that Random Forest dramatically outperforms Decision Tree because of its ensemble learning feature, resulting in greater AUC values (0.90, 0.91, and 0.84). This validates results in other research that Random Forest efficiently minimizes overfitting and enhances predictive performance in intricate datasets by combining several decision trees and averaging their predictions. This method assists in reducing the possibility of overfitting and improving model generalization, hence rendering the model more dependable in forecasting used car prices for various categories. Moreover, the Decision Tree model, even with its lower AUC scores (0.70, 0.71, and 0.52), is still interpretable and computationally efficient, and hence ideal for rapid, rule-based price estimation. This result is in line with earlier research that points out Decision Tree's capability to generate easily interpretable decision rules, which are beneficial in applications where transparency and interpretability are needed. But the restrictions of the Decision Tree model are its vulnerability to overfitting and instability in price prediction for high-value vehicles, as evidenced in its low AUC score of 0.52 for the category with high prices.

This is consistent with existing research identifying the Decision Tree's propensity to create excessively complex structures when confronted with diverse price distributions, resulting in poor generalization on new data. Moreover, whereas Random Forest stabilizes predictions, it is still computationally intensive and requires a lot of processing power and memory, which might be less than optimal for real-time pricing implementations, particularly in large-scale deployment. Another shortcoming seen in both models is their failure to capture external economic influences like inflation, regional market conditions, seasonal demand, and fuel price volatility, which play a very significant role in used car valuations. While feature engineering can aid in the incorporation of some of these factors, old machine learning algorithms such as Decision Tree and Random Forest cannot dynamically adjust for a sudden shift in the market, a drawback highlighted in prior research as well. One noteworthy fact highlighted in our study is that both the models perform poorly for high-price category predictions, an issue also evidenced in prior work. This is due to the volatility in luxury car pricing, where subjective aspects like brand image, special features, and limited series play a role in volatile price changes.

The challenge in forecasting high-end used car prices indicates that including other features like brand loyalty, resale value patterns, and professional appraisals might enhance model accuracy. In addition, research indicates that the inclusion of hybrid models, deep learning methods, and sophisticated optimization techniques such as Particle Swarm Optimization (PSO) may boost predictive performance, especially for upscale vehicles. Artificial Neural Networks (ANNs) and Transformer-based architectures are some examples of deep learning models that have demonstrated the capability to learn intricate nonlinear connections and thus might be the optimal solutions to future enhancements of used car price forecasting. Additionally, explainability methods like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) may be utilized to gain a better insight into the most important factors driving price predictions, enhancing user trust and model transparency.

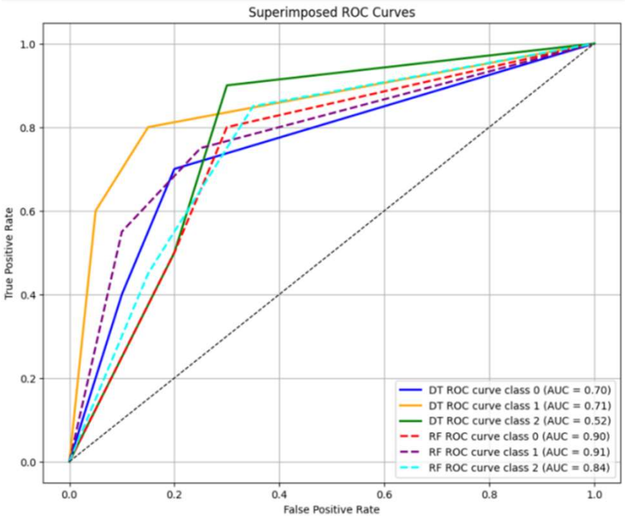


Figure-20. Superimposed AUC of ROC for DT& RF Algorithm

5. CONCLUSION

The paper of predicting prices for used cars has not been thoroughly explored. In most research works, efforts were made to predict the prices. There is still a gap what inspired to do the research, as many sellers price of used cars based on the brands. Only a limited number of studies have focused on predicting prices for used cars. The novelty of work includes a method that combines exploratory data analysis with features from both current and past data to forecast trends in the used cars market. This research paper is trying to compare two supervised machine learning models i.e. Decision Tree and Random Forest models for predicting used car prices. The paper is able to provide significant findings, assesses the performance of each model in running the datasets extracted from Kaggle website and explores the implications for further research. The datasets included important features such as car make, model, and year of manufacturer, mileage, transmission, condition, fuel type and seats which are crucial in predicting used car prices.

With the application of the Decision Tree model, we are able to generate results that have easy to understand with reasonable accuracy in predicting prices of used cars. It is observed that it over fits the training data in most of cases thus results in complicated splits. The AUC-ROC score suggest that the model had average effectiveness and struggled to predict prices on new or untrained data set. The Random Forest model, uses an ensemble method, provided a better predictive accuracy and reliability as compared to the Decision Tree model in predicting prices of used cars. It minimized over-fitting by averaging predictions from multiple trees. The AUC-ROC score suggest that the Random Forest model had excellent efficiency in predicting used car prices as it closed to accuracy. The Random Forest model emerged as a stronger and more dependable model for predicting prices of used car markets in comparison to Decision Tree model.

6. FUTURE SCOPE

The following can be obtained in future in used car price prediction by using optimization methods, deep learning models, better datasets, and enhanced explainability. Particle Swarm Optimization (PSO) can be used to tune hyperparameters for Decision Tree and Random Forest models, optimizing parameter values like tree depth and feature selection, reducing overfitting, and increasing predictive accuracy. Second, advanced deep learning methods, such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks, could be investigated in order to fit complex relationships among data and thereby give more accurate price estimates. Transformer-based frameworks, which have found widespread usage in natural language processing, can also be adapted to sequential and contextual pricing trends. Extending the dataset beyond Kaggle using live data from dealership reports, government vehicle registration databases, and internet car marketplaces would increase model strength and responsiveness to market changes. External variables like economic health, fuel prices, and seasonal demand fluctuations further enhance predictive accuracy. Additionally, incorporating explainability methods like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) will improve transparency to enable users to understand and rely on model predictions. Future work should also entail deploying these models into cloud platforms, mobile apps, or APIs to facilitate real-time price estimation and dynamic trend analysis. The integration of edge computing and federated learning would further enhance efficiency by enabling localized price forecasting without compromising data privacy. Through these developments, future research can create highly accurate, scalable, and interpretable predictive models, transforming the used car market by offering sellers and buyers data-driven, real-time insights for informed decision-making.

REFERENCES

[1] Ardiansyah GT, Hasibuan MS, Santosa S, Heikal J. Mapping the Wuling vehicle market with K-Means Clustering: An effective digital marketing strategy. JurnalFokusManajemenBisnis. 2024 Sep 3;14(2):136-50.

[2] Kriswantara B, Sadikin R. Machine learning used car price prediction with random forest regressor model. JISICOM (Journal of Information System, Informatics and Computing). 2022 Jun 2;6(1):40-9.

[3] Chandak S, Chandak A. Impact of IT and Sustainability on Supply Chain Performance: A DEMATEL Analysis of the Indian Automobile Industry. IUP Journal of Supply Chain Management. 2023 Jun 1;20(2):5-22.

[4] Amalia D, Diana N. Pengaruh BOPO, CAR, dan FDR TerhadapProfitabilitas Bank BukopinSyariahPeriode 2013-2020. JurnalIlmiahEkonomi Islam. 2022 Apr 11;8(1):1095-102.

[5] Gajera H, Pulugurtha SS, Gore N, Ghasemi A, Duvvuri S, Kodupuganti SR. Modeling and Estimating the Effect of a Mix of Varying Levels of Automated Vehicles on Operational Performance of Urban Freeways. Transportation Research Record. 2024:03611981241287195.

[6] Dutta A, Rathore AP. Estimating Ergonomic Compatibility of Cars: A Fuzzy Approach. Procedia Computer Science. 2020 Jan 1;167:506-15.

[7] Putra RP, Yuvenda D, Setiyo M, Andrizal A, Martias M. Body city car design of two passengers capacity: a numerical simulation study. Automotive Experiences. 2022 Apr 18;5(2):163-72.

[8] Sabri M, Danapalasingam KA, Rahmat MF. A review on hybrid electric vehicles architecture and energy management strategies. Renewable and Sustainable Energy Reviews. 2016 Jan 1;53:1433-42.

[9] Putra PH, Azanuddin A, Purba B, Dalimunthe YA. Random forest and decision tree algorithms for car price prediction. JurnalMatematika Dan IlmuPengetahuanAlamLLDikti Wilayah 1 (JUMPA). 2024;4(1):81-9.

[10] Wardana TR, Oetomo W, Hartatik N. ANALISIS AKTIVITAS PASAR TANJUNG ANYAR TERHADAP KINERJA LALU LINTAS DI JALAN RESIDEN PAMUJI KOTA MOJOKERTO DENGAN METODE PKJI. Journal of Scientech Research and Development. 2024 Jun 9;6(1):563-70.

[11] Yang H, Wang H. Signaling control of the constitutive androstane receptor (CAR). Protein & cell. 2014 Feb;5(2):113-23.

[12] Singh PK, Sarkar P. A framework based on fuzzy Delphi and DEMATEL for sustainable product development: A case of Indian automotive industry. Journal of Cleaner Production. 2020 Feb 10;246:118991.

[13] Dutta A, Rathore AP. Estimating Ergonomic Compatibility of Cars: A Fuzzy Approach. Procedia Computer Science. 2020 Jan 1;167:506-15.

[14] Putra PH, Azanuddin A, Purba B, Dalimunthe YA. Random forest and decision tree algorithms for car price prediction. JurnalMatematika Dan IlmuPengetahuanAlamLLDikti Wilayah 1 (JUMPA). 2024;4(1):81-9.

[15] Amjad R, Sulong Ghazali ST. An intelligent approach to image denoising.

[16] Maruthi R, Sankarasubramanian DK. Multifocus Image based on the information level in the region of the images. JATIT. 2005:2005-7.

[17] Yasmina Am, Bentaleb Y, Sharippudin Sn, Mahadi N, Zakaria Wn, Sardjono W, Perdana Wg, Hadi Vh, Mutiara Ab, Refianti R, Hastuti K. Latent Modeling For Predicting Multidimensional Data. Journal of Theoretical and Applied Information Technology. 2024 Jan;102(1).