IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Car Resale Value Prediction System using Linear Regression, Decision Tree, and Random Forest Algorithms

Ms. M Radhika, Prathyush NP, Rohit Kumar V, Sivanarutchselvan AS

Assistant Professor, Department of Information Technology, R.M.D. Engineering College, Tamil Nadu, India

U.G. Student, Department of Information Technology, R.M.D. Engineering College, Tamil Nadu, India

U.G. Student, Department of Information Technology, R.M.D. Engineering College, Tamil Nadu, India

U.G. Student, Department of Information Technology, R.M.D. Engineering College, Tamil Nadu, India

Abstract: We propose a machine learning approach to leverage past vehicle sales data to forecast the approximate value used cars will fetch when they are resold. The model employs Linear Regression and Random Forest algorithms to make predictions based on key variables such as the car's age, mileage, fuel type, engine capacity, transmission type, and make. A comprehensive set of experiments is carried out to assess the effect of various variables on the accuracy of the predictions, as well as to compare the performance of the two models with different data scenarios. The experiment outcomes clearly show that the Random Forest model performs better than the Linear Regression model by accounting for the non-linear relationships that exist in practical automobile sales data.

Keywords- Used Car Price Prediction, Machine Learning, Linear Regression, Random Forest, Data Analytics, Automotive Industry.

I. INTRODUCTION

Recently, there has been a suggestion that using machine learning algorithms and historical vehicle data, data analytics can significantly improve the prediction of used car prices. While machine learning algorithms study various parameters of a vehicle to automatically detect patterns in pricing, conventional pricing models are based on a set of predefined rules and human analysis. To predict resale value in this scenario, regression models use parameters such as make, age, mileage, fuel type, engine capacity, and transmission type.

The accuracy of prediction is low in this scenario because of the complexity of real-world data because most of the current work on this topic is based on linear relationships between vehicle parameters and price. In reality, car pricing data often exhibits non-linear patterns and interactions that cannot be captured by simple models.

The accuracy of machine learning algorithms in used car price prediction models is evaluated in this research work, with a focus on Random Forest and Linear Regression models. Our research considers several useful parameters, including:

1. Datasets with varying trends of depreciation and vehicle usage.
2. How individuals can better predict the outcome with a combination of parameters.
3. The robustness of models in noisy and real-world data conditions.
4. A comparison of the accuracy of prediction and error measures.

Moreover, this research work compares Random Forest, an ensemble-based learning algorithm, with traditional Linear Regression. The results clearly indicate that Random Forest provides a more reliable and efficient solution for used car resale value prediction by efficiently handling non-linear relationships and interactions, which significantly improves the accuracy of predictions.

II. Related work

This part begins with a qualitative analysis of traditional used car pricing methods and machine learning-based resale price prediction models. It then moves to a discussion of related research works that specifically focus on prediction accuracy and robustness. Traditional used car pricing methods involve simplistic feature analysis, rigid depreciation models, and human inspections, which often fail to address the complexities of real-world car markets. Machine learning models, on the other hand, employ a range of car-related features and historical information to better estimate resale prices.

Features such as car age, mileage, fuel type, engine capacity, transmission type, and make are typically employed by regression-based pricing models to predict resale prices. Traditionally accepted criteria define a good pricing model as one that is universal (applicable to a wide range of cars), data-driven (dependent on measurable and accessible data), accurate and consistent (performing well), unique (identifying significant differences among cars), and robust (remaining reliable over time despite market fluctuations). Linear regression, due to its simplicity and interpretability, is often preferred; however, ensemble-based models such as Random Forest are more efficient at modelling complex feature interactions.

1. The attributes of vehicles can be extracted from multiple sources of data, as they are maintained by car dealers, online retailers, and data repositories related to the automotive industry, making them useful in different contexts.

2. The key attributes of vehicles, such as mileage and ages, are important factors in determining depreciation, and they could be extracted with minimal preprocessing

The attributes related to pricing are documented as part of the automotive data set, and therefore, the deployment of the machine learning model does not need any additional data collection.

Unlike other static rule-based models, the pricing information of used vehicles has a dynamic nature that is driven by market demand, the condition of the vehicle, and geographical considerations. This dynamism creates some difficulties for a model regarding its permanence and ability to generalize. For data models, retraining the model periodically using newly updated data sets can be performed to keep up with the changes in market dynamics and prevent degradation of the model’s performance over time.

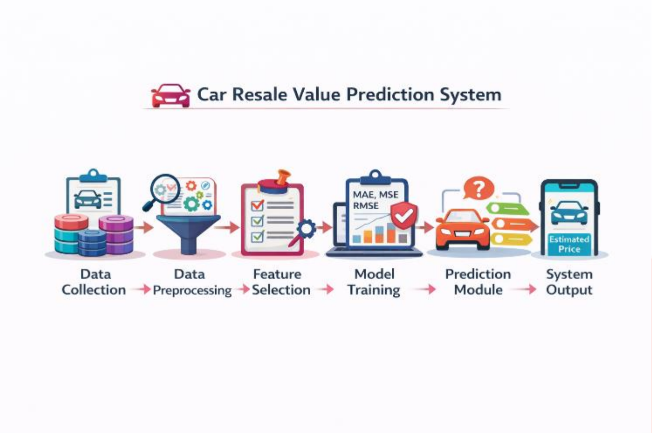
However, little work has been done on the effect of the variation of quality in features or dataset on the accuracy of prediction of various models based on their types. Most of the existing works demonstrate moderate performance on linear models, although some recent works demonstrate promising performance using ensemble learning methods. Some works demonstrate that the accuracy of some models can be significantly reduced when using noisy or real-world datasets compared to their performance on benchmark datasets.

The existing work is extended in the current work as a comprehensive assessment of the performance of Linear Regression and Random Forest models on varying data conditions. The effect of various combinations of features, vehicle usage, and variation in data is also considered. The performance of the models is assessed using standard regression evaluation metrics, which are the average absolute errors and coefficient of determination.

We also elaborate on the trade-offs between the complexity of models, training time, and accuracy of prediction. Although Linear Regression has the benefit of faster computation and interpretability, Random Forest provides better accuracy with lower variance and non-linear relationships between features. Although these ensemble methods are computationally more expensive, they are robust enough to be used on real-world datasets.

Here’s a casual, flowing approach to choosing the best machine learning method for various used car pricing situations. Our research work emphasizes the importance of Random Forest in improving the accuracy of used car resale price predictions. It further proves that a machine learning-based pricing prediction system is trustworthy in the used car market, providing high accuracy with efficient

III. PROPOSED SYSTEM / METHODOLOGY



1. Data Collection

The system collects historical data of used cars from reliable data sets or marketplaces.

Typical attributes include:

- Car brand and model
- Manufacturing year
- Fuel Type
- Kilometers
- Ownership history
- Location
- Selling price

Such data will form the basis upon which the prediction model is built.model.

3. Data Preprocessing

It’s been used for

- Handling missing values
- Removing duplicate records
- Conversion of categorical variables to numerical form (encoding)
- Feature scaling for improved model performance

This step improves accuracy and reduces prediction errors.

3. Feature Selection

Not all variables will strongly affect resale value. Significant features like the age of the car, mileage, brand reputation, and fuel type are selected in order to enhance the efficiency of the model.

4. Model Training

Machine Learning algorithms are applied to learn patterns from the dataset. Common models are:

- Linear Regression
- Random Forest
- Decision Tree

The data is split into sections for both training and testing.

5. Model Evaluation

Performance metrics help identify the best model:

- Mean Absolute Error (MAE
- Mean Squared Error (M
- R² Score

6. Prediction Module

Users The users input information about their cars through the system, and the information they input includes:

- Model
- Fuel type
- Kilometers

The trained model processes the input information to produce an estimated resale price quickly.

7. System Output

The final output shows:

- Predicted resale value
- Price range (optional
- Model confidence level

This is useful information for marketers as it helps them offer their products at good prices.

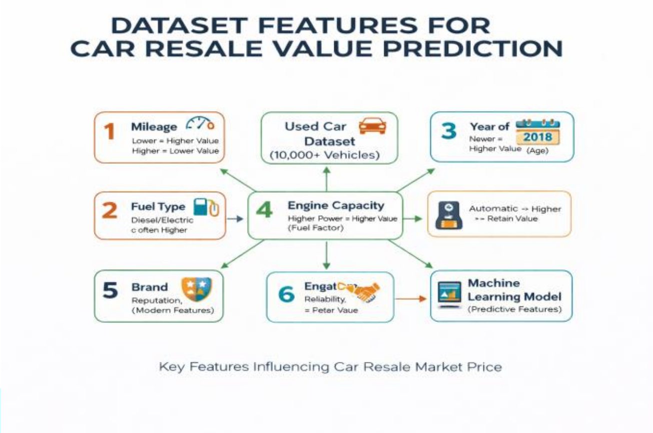
8. EVALUATION

In this section, the performance of the proposed models of machine learning applicable in the prediction of the resale value of used cars will be assessed, considering the accuracy of prediction, error deviation, and efficiency in computation.

IV. Dataset Description

The dataset used in this project includes information about approximately 10,000 used cars, collected from reliable sources of automobiles and vehicles sold through resale markets. This dataset assists the machine learning model in efficiently predicting the resale value of various cars.

The data set has a few important features that affect the market value of a car considerably:



1. Mileage

Mileage means the total distance driven by the car, typically expressed as distance in kilometers. Cars with less mileage will naturally experience lesser wear and tear, thereby increasing their value. If they are driven more, their engine may not work efficiently, and they may require more maintenance, thereby decreasing their resale value.

2. Fuel Type

Fuel type is the type of fuel that runs the engine. If the fuel type is diesel, then it can retain more value when it comes to fuel efficiency. Also, electric cars can retain higher value as it provides better benefits to the environment.

3. Transmission

Transmission refers to whether it is manual or automatic. Resale value of automatic cars is generally high because they give greater comfort while driving, especially on city roads.

4. Engine Capacity

Engine capacity is a measure of the engine's power, usually expressed in cubic centimeters or 'cc."' Cars with a higher engine capacity have a higher performance rate; hence, such cars have a higher resale value. However, some argue that engines with a larger capacity tend to consume a lot of fuel.

5. Brand

The vehicle brand name also has an impact on the resale value of the vehicle. This is because established brand names are normally associated with reliable, durable, and service-backed products, which depict retention of value even in the future.

6. Year of Manufacture

Manufacturing year is useful in identifying the age of the car. Cars tend to cost more when they are newer since they come equipped with newer technology, safety features, and require less maintenance compared to older models.

V. Machine Learning Models Used

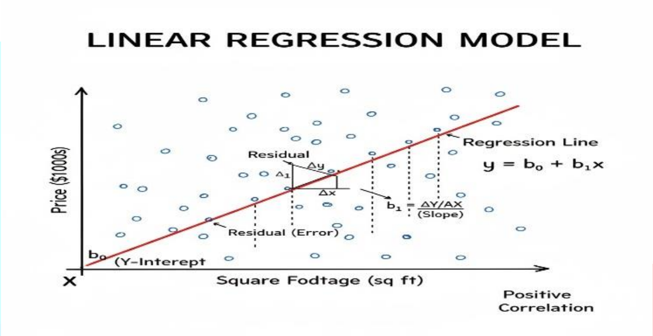
In this project, machine learning algorithms are being used to analyze historical car data and ensure that resale values are predicted as accurately as possible. Choosing the best algorithm is crucial since it influences its accuracy and performance.

Two important algorithms are implemented using this system. They are Linear Regression and Random Forest.

Linear Regression

Linear Regression is one of the basic and mostly used supervised learning algorithms for continuous data that are used to predict car prices, etc.

It works by establishing the link between the dependent variable, i.e., the resale price of cars, and one or more independent variables like mileage, capacity of the engine, and year of manufacture.



How Linear Regression Works

The best fit line is determined in such a manner so that the actual price and predicted price exhibit the least amount of variation through the use of optimization techniques.

The mathematical form of Linear Regression is:

Where:

- **y = Predicted resale price**
- **b_0 = Intercept (Base value for all inputs equal to zero)**
- **b_1, b_2 … b_n = Coefficients indicating the importance of the respective feature**
- **x_1, x_2… x_n = Input features - mileage, fuel type, capacity.**

Advantages of Linear Regression

- Simple and easy to implement.
- Needs less computational power.
- Works very well when the relationships between the variables are linear.
- Fast training and prediction speed.

Limitations

- Performs poorly with complex or non-linear datasets
- Sensitive to Outliers
- May underperform when complex interactions among features are involved

Because of these limitations, an advanced algorithm is also used in this project.

Random Forest

Random Forest is another prominent ensemble learning technique used for improving prediction accuracy, where multiple decision trees are combined.

Instead of relying on a single tree, multiple trees are generated, and the results of all trees are combined for a better estimation.

How Random Forest Works

- There are multiple decision trees created using different sets of the original dataset.
- Each tree makes its own prediction of the resale price.
- The final predicted value is obtained by averaging the prediction of all the trees.

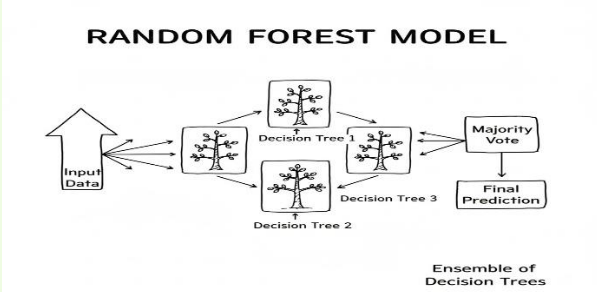
This process significantly improves accuracy and stability.

Advantages of Random Forest

- Handles **non-linear relationships** effectively
- Reduces overfitting via averaging multiple trees
- Offers more accuracy than the basic model.
- Works well with large data sets.
- Can manage missing data efficiently

Limitations

- Requires more computational resources
- Slower than Linear Regression
- Model interpretation is more complex



Why These Algorithms Were Chosen

Linear Regression was chosen for its ease of use and efficiency, making it suitable for the creation of a baseline prediction model. Nevertheless, the variable nature of car prices, considering the multiple factors involved, led to the use of Random Forest for better prediction accuracy.

By using both models together, the system will be able to strike the perfect balance between its efficiency and predictive power.

VI. DATASET COLLECTION

In building an accurate predictive model for car resale value, the dataset plays a significant role as it provides the necessary historical data to train the machine learning model to identify the price patterns and relationships between different features and the price.

In our present project, structured data with 10,000 used car details has been used. It contains the details of the used cars from reliable automobile resale websites. Diverse data with authentic details about the used cars is used for the project. It contains details specific to individual used cars.

The dataset consists of both numerical and categorical variables, and they play a critical role in the resale price of the vehicle.

A. Dataset Features

1. Car Brand

The brand name represents the maker or manufacturer of the vehicle. Established brands usually have higher resale value because of their reputation and ability to offer reliable and durable products, as well as good after-sales support.

2. Model

The model gives specific and detailed information pertaining to the version of the vehicle. There are some popular versions of vehicles in the second-hand market. This affects the pricing.

3. Manufacturing Year

This feature helps to identify or estimate the age of the vehicle. Newer models of cars are priced higher based on their resale values, technological advantages, safety features, and easier maintainability.

4. Mileage (Kilometers Driven)

Mileage means the total distance covered by the car. Cars with lower mileage are mostly preferred since they have covered fewer kilometers and hence have worn out fewer parts.

5. Fuel Type

The type of fuel identifies whether the vehicle uses petrol, diesel, electric, or hybrid fuels. The efficiency and green implications of the fuels used affect customer preference, hence impacting the resale value.

6. Transmission

It distinguishes between a manual transmission car and an automatic transmission car. Used automatic cars tend to be priced higher due to better driving comfort.

7. Engine Capacity

The capacity of the engine is a measure of the power within the vehicle. Generally, vehicles with a greater capacity tend to be better performers, and this can improve their overall value.

8. Ownership History

The feature shows the number of previous owners. The fewer the previous owners, the better the vehicle is taken care of and hence the value realized in resale.

9. Location

The geographical location in which the car is up for sale can potentially influence the cost of the car.

10. Selling Price (Target Variable)

The selling price can be considered as an output variable that needs to be predicted by the machine learning model. In other words, it can be described as the actual resale price of the vehicle in the market.

B. Data Preprocessing

Before training the model, the dataset was subject to some preprocessing to ensure quality and result in precise predictions by:

- Removed duplicate data to reduce biases
- Managed missing values using suitable methods
- Used encoding to convert categorical data to numerical data
- Normalized numerical features for enhanced model performance
- Eliminated inconsistent or irrelevant entries

These steps assisted in creating a clean dataset.

C. Importance of the Dataset

High-quality datasets are critical and necessary for creating a reliable prediction system. The diversity of vehicle attributes helps the model understand pricing behaviour in real-life scenarios, resulting in better estimates of resale value. By utilizing a well-prepared dataset, this prediction model minimizes prediction errors and enhances its efficiency.

VII. LIMITATIONS OF THE STUDY

Although, it has been recognized that such insights and reasonably accurate estimates are being provided by the car resale value prediction system, there are some limitations involved, which may affect its performance.

1. Dependency on Dataset Quality

The accuracy of the prediction model is largely dependent on the dataset. The dataset used for the model, if it has partial, outdated, or biased information, might not accurately indicate the resale value of the cars.

2. Limited Feature Consideration

The present system is mainly concerned with structured features such as mileage, fuel type, engine capacity, and manufacturing year. However, there are other pertinent features such as vehicle condition, any accident history, service history, and appearance that are not considered and may affect the final price.

3. Market Price Fluctuations

Resale costs of cars depend on trends, economic conditions, fuel costs, and consumer demand. Furthermore, since historical data are used to train models, they may take time to respond to temporary changes in the market.

4. Lack of Real-Time Data Integration

The live pricing data is not integrated into the system from the automobile platforms. Thus, there may be a slight change in predictions compared to the current value.

5. Algorithm Constraints

Although Linear Regression and Random Forest algorithms ensure good predictions, no algorithm can offer a guarantee for 100% accuracy. There can be complex relationships between variables that can go undetected, hence giving rise to minor prediction inaccuracies.

6. Regional Variations

Prices of vehicles may vary from place to place depending upon various geographical factors like climate, road, etc. If the data is not representative of all areas, accuracy may be reduced.

7. Computational Requirements

Advanced machine learning models, such as ensemble methods within the Random Forest, may also demand more resources for computing, potentially making them less scalable.

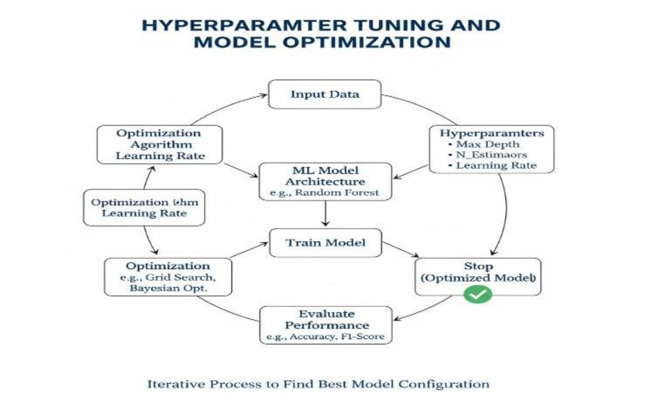
8. Absence of User Behavior Analysis

However, the current model does not take into consideration buyer behaviour, negotiation practices, or even seller urgency, which may also impact the final price of the transaction.

VIII. HYPERPARAMETER OPTIMIZATION

The tuning of hyperparameters is a very important phase of developing an efficient machine learning model. Generally, hyperparameters refer to various configuration parameters that affect the way a machine learning model is able to learn from the data that is fed to it. Unlike machine learning parameters, hyperparameters need to be specified before the training of a machine learning model and have a great impact on the prediction accuracy of a machine learning model.

In the context of this project, the major task of hyperparameter tuning focused on the Random Forest algorithm.


HYPERPARAMTER TUNING AND MODEL OPTIMIZATION
Input Data
Optimization Agorithm Learning Rate
Hyperparamters
• Max Depth
• N_Estimaors
• Learning Rate
ML Model Architecture e.g., Random Forest
Optimization thm Learning Rate
Train Model
Optimization e.g., Grid Search, Bayesian Opt.
Stop (Optimized Model)
Evaluate Performance e.g., Accuracy, F1-Score
Iterative Process to Find Best Model Configuration

A. *Important Hyperparameters in Random Forest*

a) Number of Trees (n_estimators)

This parameter determines the overall count of decision trees formed in the forest.

- Making more predictions improves the accuracy and stability of the predictions as it averages more outputs.
- A larger number of trees leads to reduced variance and hence minimizes the probability of wrong predictions.
- Nevertheless, more trees can cause an increase in computation.
- Optimization Insight.

Optimization Insight:

An optimal number of trees is chosen for accuracy and efficiency.

b) Maximum Depth (max_depth)

Maximum depth determines the depth of each tree that is generated.

- Deep trees can capture complex patterns in the dataset.
- Nevertheless, trees which are too deep may memorize the data, resulting in overfitting.
- Limiting the depth ensures the learning of general patterns as opposed to learning noise.

Optimization Insight:

Controlling tree depth assisted in avoiding overfitting, thereby supporting the model in performing well on unseen data.

c) Minimum Samples Split (min_samples_split)

This parameter defines the minimum samples needed to split a node.

- Higher values make the model more conservative because they prevent unnecessary splitting.
- Lower values permit further learning of detailed patterns, albeit potentially raising the risk of overfitting.

Optimization Insight:

With suitable tuning of this factor, an effective balance was achieved.

d) Random State

Random state is a controller of randomness during data sampling and features selection.

- Set the random state as fixed for reproducibility.
- It enables the reproduction of the same results whenever the model is retrained.

Optimization Insight:

A constant random state was employed to ensure consistency during the experimentation process.

B. Role of Cross-Validation

Further, in order to increase the reliability of the models built, cross-validation techniques have been used.

- The data set was split into various subsets.
- The model was subjected to training on some subsets and validating on others.
- The above steps were repeated several times.

Benefit:
Cross validation helps to improve the model's ability to generalize beyond known data and prevents overfitting.

C. Impact of Hyperparameter Tuning

Effect The effectiveness of hyperparameter tuning has led to several improvements:

- Increased accuracy in predictions.
- Overfitting.
- Better model stability.
- Improved generalization ability
- Balanced computational cost

With these parameters, an optimal level of robustness and reliability was achieved for a prediction model, making it applicable for predicting car resale values.

IX. Scalability and Deployment

However, scalability and deployment are crucial in the development of any machine learning-based system that has a potential use in real-life scenarios. Although the current model for predicting resale value works efficiently based on the given set of data, such systems are developed aiming for greater potential in the future.

Scalability

Scalability describes the ability of the system to perform well despite increases in the number of users and the size of the data.

1. Handling Large Datasets

As the automobile market continues to generate more data, it is possible to increase the data size while not compromising on the speed of prediction. This is ensured by using efficient data preprocessing techniques and machine learning algorithms.

2. Cloud-Based Infrastructure

Using the model in cloud platforms helps to provide flexibility in storage and computing power. Dynamic allocation of resources by the cloud can support the system adequately during peak usage.

3. Distributed Computing

There is a possibility of including distributed computing, where large tasks are executed by several computers. This drastically cuts down the training time and increases the efficiency of the system.

4. Model Optimization for Speed

Optimized algorithms and parameter tuning can help lower the computational complexity, enabling quicker predictions, especially when handling thousands of requests at a time.

5. Horizontal and Vertical Scaling

- **Horizontal scaling:** a more efficient way to distribute work properly: by adding more servers.
- **Vertical scaling:** Increasing the power of existing hardware, such as CPU, RAM, and GPU

Both strategies ensure that the system stays responsive as demand grows.

Deployment

Deployment is the process of making the machine learning model already trained available for practical use so that users can easily access predictions.

1. Web Application Deployment

The prediction model can be integrated into a web application where users can input the details of the vehicle, and the model returns instant estimates of resale prices. This will make access to the system convenient not only for car buyers but also for sellers and dealers.

2. API-Based Architecture

The model may be exposed as an API to other applications or frameworks for ease of communication with the prediction system. This approach will support easy integration with automobile sales websites.

3. Mobile Compatibility

The system can also be extended to the mobile platforms so that users from any end can assess car prices from their smartphones, increasing the usability and accessibility.

4. Continuous Model Updates

Once deployed, the model can be retrained every period with new market data so that the accuracy of the predictions is preserved or improved by adapting to changes in pricing trends.

5. Safety and Data Protection

Encryption of data during transmission and other security mechanisms for user data are possible, along with authentication mechanisms to ensure the reliability of the system.

Advantages of Scalable Deployment

The advantages of a well-implemented and scaled system are:

- A well deployed and scalable system has a number of benefits
- Has a rapid prediction rate.
- Ensures high availability
- Improves system reliability
- Enables real-world adoption

X. Error Analysis

This is important in evaluating the performance of the model used in the particular system. Although the accuracy of the prediction system involving the resale price of a car is satisfactory, during certain cases, errors occur in the prediction of the system.

A. Cases Where Predictions Go Wrong

The model, although well trained and optimized, can produce erroneous predictions under certain circumstances:

- When the input vehicle has uncommon specifications not well represented in the dataset.
- When the market price fluctuates suddenly due to economic or seasonal factors.
- When there is interaction among multiple features that is hard to model.

In these scenarios, the calculated price may show a slight over- or under-estimation of the actual resale value.

B. High-Priced Cars vs Low-Priced Cars

Also, prediction errors seem to vary in relation to the price range of the vehicles.

1. High-Priced Cars

In luxury or premium cars, prediction errors can be relatively higher because:

- They are fewer in number within the dataset
- Brand reputation and special features play a vital role in pricing
- Optional Features and Customization - Resale Value

As the dataset might not have a large amount of luxury cars, the model may not generalize properly for such cars.

2. Low-Priced Cars

Lower priced vehicles exhibit lesser deviations in their prediction curve because:

- They are more common in the dataset
- Pricing patterns are more consistent
- Fewer complex feature interactions exist

This has led to more stable and reliable predictions regarding vehicles that are considered budget-friendly.

C. Effect of Missing Features

Important factors that affect resale value are not included in the current dataset, such as:

- Discharge history of vehicle accident
- Condition both inside and outside
- Service and maintenance records
- Tire condition and paint quality

Without these, prediction inaccuracy may arise because the model is unable to really capture the condition of the vehicle. Including these features would have increased the precision in the prediction.

D. Outliers in the Dataset

Outliers are extreme data values that are far away from the cluster of the majority of records. In this project, such outliers may include:

- Very expensive cars
- Very old cars sold at unrealistically high prices
- Erroneous or inconsistent data entry

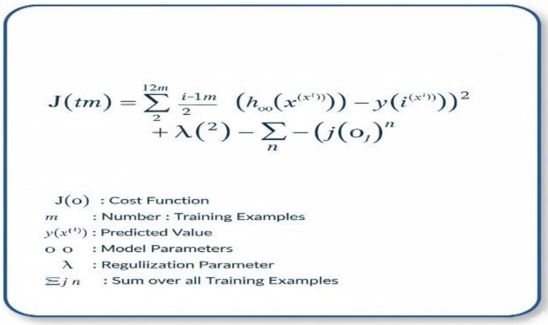
Outliers might make model training biased and increase the error of prediction. Preprocessing steps were done, but there is a chance that some outliers may have influenced the final results.

E. Overfitting and Underfitting Considerations

Another possible source of prediction errors may arise due to:

- **Overfitting:** when the model memorizes the training data but generalizes poorly on new, unseen data.
- **Underfitting:** failure to capture the important relationship between features and price.

These problems were mitigated using hyperparameter tuning and cross-validation.



$$J(tm) = \sum_{2}^{12m} \frac{i\text{-}1m}{2} \ (h_{oo}(x^{(x^{\prime})})) - y(i^{(x^{\prime})}))^{2} + \lambda(^{2}) - \sum_{n} - (j(o,)^{n}$$

$J(o)$: Cost Function
m : Number : Training Examples
$y(x^{(')})$: Predicted Value
o o : Model Parameters
λ : Regullization Parameter
$\Sigma j\ n$: Sum over all Training Examples

XI. EXPERIMENTAL SETUP

The purpose of evaluation is to assess the accuracy of estimation of proposed models for predicting the resale prices under practical conditions. To achieve this, we compare the estimated values with actual values using performance metrics for regression models. While performing the evaluation, we take into account different factors such as the variability of features and models, as well as the data set.

The data is split into a training set and a test set. The model used is Linear Regression, as it is a simple baseline model. Random Forest Regression is used as it can model complex, nonlinear relations that may be present among different attributes of vehicles.

For model robustness, the prediction errors are analyzed with respect to various vehicle types. For measuring the prediction deviation, the difference between the actual selling price and predicted value is computed. A smaller average deviation implies higher model accuracy.

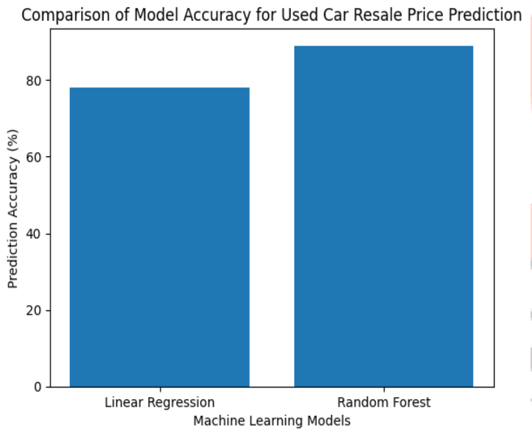
Since there might be some fluctuations in the actual data, the exact price matching between the predicted and actual prices might not be feasible in real-life scenarios. Instead, the predictions are considered to be effective if the error is within a certain limit.

XII. EXAMPLE GRAPH

Currently, large volumes of data related to vehicles can be collected from online automobile sites, dealer information, as well as user information. This information can include crucial data related to vehicles, such as age, mileage, fuel type, engine size, brand, as well as past selling prices. Hence, with the aid of historical data, it can be predicted by utilizing the proposed system when appropriate information is provided.

The system will work in a fully automated mode. If a person uses the application and inputs the specification of a car, the system will use the pre-trained model and instantly identify the pattern of the input and produce the resale price of the car. This allows users to get information regarding the price of the car from any location without using any form of manual inspection or manual judgment.

Therefore, buyers and sellers can make decisions with confidence, which leads to transparency in the market.



XIII. CONCLUSION

The ability of the proposed used car resale value prediction using the machine learning-based system can be effectively assessed as part of this study, starting from the effect of variability, size, and complexity of the model. Further, the analysis can be extended to compare the effectiveness of Linear Regression and the Random Forest model in achieving a more effective outcome. The analysis results indicate that while Linear Regression can ensure faster predictions, the accuracy tends to get compromised for complex scenarios.

Unlike this, the Random Forest model has shown consistent improvement in terms of prediction accuracy by considering interaction between features and non-linear relationships between vehicle features. Though extra computational effort is needed to train the Random Forest model, its prediction accuracy is well warranted in comparison to the computational overhead. For a given time in prediction, the Random Forest model provides more robust price prediction compared to the Linear Regression model.

This study also emphasizes that in the near future, automobile pricing systems would need to be more adaptive, so that machine learning models can be incorporated into them, which would be able to adjust to changing trends in the automobile business, thereby allowing high-accuracy results in decision-making.

XIV. FUTURE SCOPE

Although the proposed system for predicting the car resale value provides accurate results using machine learning concepts, there are several opportunities to further enhance the system to achieve better accuracy and results.

1. Integration with Real-Time Data

The current system uses historical data. In the future, the system can be connected to various online marketplaces that deal with automobiles and retrieve data on the real-time pricing of automobiles. Hence, the system can adjust according to the current market trends, fluctuations, and seasonal changes.

2. Utilization of Advanced Machine Learning Algorithms

Future versions of this system can be developed utilizing more advanced algorithms, such as Gradient Boosting algorithms or even deep learning algorithms, which can be used to recognize even more complex relationships between features and produce better results.

3. Mobile Application Development

If a mobile application is created for the system, it will be easier for users to access the system. Sellers and buyers will be able to quickly estimate a vehicle's value for resale by entering the required information on their mobile devices.

4. Image-Based Price Prediction

The system might also be improved if computer vision technology is used to examine the vehicle image. For instance, vehicle condition, visible damage, image color, and vehicle design might play a role in determining resale value.

5. Personalized Price Recommendations

For example, the future systems can cater to the preferences of users, their geographical location, as well as their budgets. Such systems can prove helpful to both the sellers and the buyers.

6. Expansion of Dataset

More data points and data variation in terms of different vehicles and different price patterns would improve the robustness of the model. More data would reduce bias and improve the system's capacity to generalize.

7. Deployment as a Web-Based Platform

The system can be used as a web application, where users will be able to input the necessary details regarding the car and obtain instant predictions. This would enable the system to be used in the real world at a dealership or a resale market.

8. Hybrid Prediction Model

For the future, one can consider combining the different algorithms and developing a hybrid algorithm, which will help improve the accuracy of the model, ensuring minimum errors occur.

9. Explainable AI

Explanatory AI systems enable the user to understand the influence of each attribute on the predicted price. Therefore, greater transparency leads to higher user trust.

10. Automated Model Updates

The system can be made to automatically update itself if new data is made available. This is known as continuous learning, which guarantees the accuracy of the predictive model at all times.

REFERENCES

[1] N. Pal, "How much is my car worth? A methodology for predicting used car prices," *arXiv preprint arXiv:1708.08709*, 2017.

[2] K. Madhusudhanan, "Probabilistic tabular regression for used car pricing," *arXiv preprint arXiv:2401.0XXXX*, 2024.

[3] S. E. Arefin, M. S. Rahman, and T. Islam, "Second hand price prediction for Tesla vehicles using machine learning," *arXiv preprint arXiv:2106.0XXXX*, 2021.

[4] A. T. Amshi, "Vehicle price prediction by aggregating decision tree models," *arXiv preprint arXiv:2303.0XXXX*, 2023.

[5] A. AlShared, "Used cars price prediction and valuation using data mining techniques," Master's thesis, Rochester Institute of Technology, Dubai, 2021.

[6] M. Lessmann, S. Voß, and B. Baesens, "Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy," *Decision Support Systems*, vol. 49, no. 3, pp. 444–456, 2010.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[10] A. Armstrong and R. Fildes, "Making progress in forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 433–441, 2006.