Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

PREDICTING THE PRICE OF AUCTION CARS WITH MACHINE LEARNING ALGORITHMS

Muhammad Nadeem[*1], Absar Chohan[2], Muhammad Furqan[3], Muhammad Sufyan[4], Rayyan Ahmed[5]

[*1,2,3,4,5]*Department of Computer Science & Information Technology, Sir Syed University of Engineering and Technology, Karachi, Pakistan*

[*1]munadeem@ssuet.edu.pk, [2]absarchohan1234@gmail.com, [3]furqan19722@gmail.com, [4]hafiz.sufyansadiq@gmail.com, [5]rayyanahmed086@gmail.com

DOI: https://doi.org/10.5281/zenodo.20590931

Keywords
Auction Car Price Prediction, Machine Learning, XGBoost, Random Forest, Linear Regression, Auction Valuation, Feature Engineering, Ensemble Learning, Regression Analysis, Vehicle Depreciation

Article History
Received: 11 April 2026
Accepted: 23 May 2026
Published: 08 June 2026

Corresponding Author: *
Muhammad Nadeem

Abstract

The problem of the automotive auction market to estimate the price of the cars accurately becomes critical as the number of features and interaction between these features grows and the conditions are also not standardized. In the present study, three machine learning algorithms–Linear Regression, Random Forest Regression, and an Extreme Gradient Boosting (XGBoost)–are compared using a unique dataset that was developed by integrating past data from car auctions to predict the prices of cars at auctions. Specific data features for the domain were also added, such as make, model, manufacturing year, engine, mileage, exterior color, chassis code, package trim and standardized auction condition grades (1.0 through 5.0). All missing value imputations, label encoding, Z-scores normalization, and more complex feature engineering methods, such as Vehicle Age, Mileage Intensity, Luxury Brand Mapping, and Make-Model Interaction terms have been performed prior to the processing phase. Performance of models was measured by Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-squared (R^2) measures. Experimental results show that the accuracy of prediction of XGBoost is observed to be the highest with R^2 = 96.68%, MAE = 1,403, and RMSE = 1,975 which is higher than the accuracy of Random Forest (R^2 = 0.9527) and Linear Regression (R^2 = 0.8321). The results confirm the previous findings that ensemble-based gradient boosting methods improve considerably against linear models abilities when the price estimation is a dedicated task of an auction domain, particularly when feature engineering is employed to enhance the abilities.

1. INTRODUCTION

From vehicle specifications to market trends, to resale values and consumer preferences, the automotive sector produces a considerable amount of data. The right price of the car is critical for the client, dealership, internet car market or financial institutions. Standard methods of price estimation are based on experts' views the risk of inconsistent and inaccurate price estimates.

In a quest for solutions to overcome these hurdles, we analyze and discuss how machine learning with automated solutions helps in accurate price prediction. Patterns and relationships in vehicle characteristics and prices can be detected using regression algorithms. A number of factors affect

the cost of a car, including brand prestige, the year the car was manufactured, engine capacity, the number of miles driven, and the standardized auction condition grades.

Linear Regression is one of the simpler machine learning techniques that is commonly used as a regression model due to its simplicity and interpretability. In the real world, however, car price data may have more complicated interactions between features and nonlinear relationships, which are hard to model with simple linear models. Random Forest and XGBoost are two ensemble learning algorithms that perform very well in regression problems because they are able to model nonlinear relationships and high-dimensional data.

Recent research has shown that XGBoost and Random Forest have better performance than traditional regression models in car price prediction problems.

The aim of this research is to:

1. Produce machine learning models to predict the price of auctions cars.
2. Compare the performance of Linear Regression, Random Forest and XGBoost.
3. Select the best algorithm to accurately estimate a price.

2. Literature Review

Several researchers have investigated the various methods based on machine learning that can be used to predict the price of a vehicle. Previous research focused primarily on statistical approaches, including linear regression, because of the easy implementation and interpretability.

Pudaruth [1] who used multiple linear regression, k-nearest neighbours, Naïve Bayes, and decision trees to predict used car prices using machine learning. The study indicated that all four approaches were successful in their ability to predict and that the accuracy of the prediction was significantly different for certain car brands, and the problem was not as easily solved as first thought.

Breiman [2] proposed Random Forest, an ensemble learning technique consisting of a collection of decision trees that are used to increase the prediction accuracy and prevent overfitting. This was a basic work laying the theoretical foundations for tree-based ensemble methods which are commonly used in regression tasks such as predicting vehicle prices.

Friedman [3] developed the idea of the gradient boosting framework, which considers the problem of function estimation as a numerical optimization problem in function space. It has been the mathematical foundation for XGBoost, LightGBM and CatBoost as the most up to date algorithms for structured data prediction.

Chen and Guestrin [4] published a scalable tree boosting system named XGBoost which offers sparse aware algorithms, approximate learning via weighted quantile sketch and system-level optimizations to access cache and compress data. Since then, XGBoost has been a popular algorithm for structured data competitions and many real-world prediction problems.

In a study by Jiahao He [5] he used Linear Regression for prediction of the price of the vehicle based on the vehicle make, model, year and mileage. The study showed that preprocessing and feature engineering greatly enhance the prediction performance, but the Decision Tree model turned out to be more accurate than the Linear Regression.

In Kanwal Noor and Sadaqat Jan's research [6] they suggested a vehicle price prediction system based on Multiple Linear Regression and feature selection. By ignoring irrelevant variables and identifying significant factors like model, make, city, mileage, and power steering, their model achieved approximately 98% predictive performance after applying feature selection techniques.

This methodology, proposed by Pal et al. [7] has been done with a 500 Decision Trees Random Forest model to predict used car prices. The model was able to achieve training accuracy of 95.82% and testing accuracy of 83.63%, which shows the ability of the Random Forest model to capture the nuances of the data related to car prices.

Lessmann and Voß [8] did a comprehensive comparison of 19 regression techniques for the car resale price forecasting problem, including Linear Regression, Random Forest, ANN and SVR. They reported that their results demonstrated a clear

superiority of ensemble methods (specifically Random Forest regression) over linear methods and recommend that traditional linear regression be used with caution for this purpose.

Monburinon et al. [9] have developed a comparative study of used car price prediction using a large (more than 300,000 data points) dataset from an e-commerce platform in Germany and have applied supervised machine learning regression models. They conducted a study on three regression algorithms, Multiple Linear Regression, Random Forest Regression, and Gradient Boosted Regression trees and found that Gradient Boosted Regression was the best algorithm.

For used car price prediction, the authors of Gegic et al. [10] showed that the ensemble learning techniques could be an effective approach to enable the presence of nonlinear relationships between features. The obtained overall performance was around 87.38% by performing an ensemble model of Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Random Forest (RF).

Venkatasubbu and Ganesh [11] analysed different supervised learning algorithms for used car price prediction, such as Lasso Regression, Multiple Regression and Regression Trees. They found that the type of car was the most important factor influencing price, and cosmetic improvements were the least important.

Bhatnagar et al. [12] did the comparative study of various supervised machine learning methods for car price prediction such as Linear Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Regressor, and Voting Regressor. Their results showed that the Gradient Boosting Regressor model was able to consistently outperform other models, yielding the highest training accuracy, and linear regression was the poorest of all the models they tested.

Gajera et al. [13] explored how well machine learning models can predict the value of old cars, and found that machine learning models using trees significantly outperform some linear baselines when applied to a wide variety of automotive feature datasets with a non-linear depreciation pattern.

In the context of used car pricing, Kriswantara and Sadikin [14] specifically adapted the Random Forest Regressor model in order to predict the price of used cars and showed that the correct selection of hyperparameters and the process of feature engineering greatly enhance the accuracy of car price prediction.

To further enhance the prediction performance of both models, Cui et al. [15] introduced another framework: used car price prediction based on a combination of XGBoost and LightGBM and deep residual networks, which achieved better prediction performance than the individual models.

Yutao Rao et al. [16] have explored various machine learning approaches such as Elastic Net, Random Forest, Support Vector Machines and XGBoost for predicting the price of the used car. They found that XGBoost obtained the best predictive accuracy with an R2 value of 0.87 due to its ability to handle non-linear interactions between the features.

Likewise, used car price prediction was used by Tingyu Qian [17] with the implementation of XGBoost, and they showed that outlier removal and feature engineering could be implemented to enhance the estimation accuracy at the same time boosting methods could be implemented to bring the accuracy to a better level. The results showed that the dataset was normalized and data was inputted into the model using XGBoost and dummy variables to arrive at the most accurate predictions.

Linear Regression, Random Forest and XGBoost were used to predict the price of the cars by Hangzhi Chen [18] and it has been found that both the Random Forest and the XGBoost outperformed the Linear Regression with accuracy of 84.62% and 84.66% respectively.

The same applies to the comparison of structured datasets of car prices, which also found that XGBoost always outperforms Linear Regression and Random Forest in terms of RMSE and R^2 scores [8][10].

2.1 Literature Review Matrix

Table 1: Literature Review Matrix

#	Author(s)	Year	Algorithm(s) Used	Dataset	Key Findings	Accuracy / R^2
1	Pudaruth [1]	2014	MLR, KNN, Naïve Bayes, Decision Trees	Used Cars (Mauritius)	All four methods yielded comparable performance; price prediction proved difficult to resolve with high accuracy	Varies by brand
2	Breiman [2]	2001	Random Forest	Theoretical	Introduced RF; ensemble of trees reduces overfitting	Foundational
3	Friedman [3]	2001	Gradient Boosting	Theoretical	Proposed gradient boosting framework for function estimation	Foundational
4	Chen & Guestrin [4]	2016	XGBoost	Multiple	Scalable tree boosting with sparsity-aware algorithms	State-of-the-art
5	Jiahao He [5]	2024	Linear Regression, Decision Tree	Kaggle Dataset	Feature engineering improved LR performance; DT outperformed LR	DT > LR
6	Noor & Jan [6]	2017	Multiple Linear Regression	Pakistani Cars	Feature selection improved MLR; achieved high precision	~98% precision
7	Pal et al. [7]	2018	Random Forest (500 trees)	Kaggle Dataset	RF captures nuances; 70:20:10 split used	Testing accuracy/performance = 83.63%
8	Lessmann & Voß [8]	2017	19 Methods (RF, ANN, SVR, LR)	German Cars	RF outperformed LR; LR should be avoided for this task	RF > LR
9	Monburinon et al. [9]	2018	MLR, RF, Gradient Boosting	German e-commerce (300K+)	Gradient Boosted Regression performed best	GBR > RF > MLR
10	Gegic et al. [10]	2019	ANN, SVM, Random Forest	Bosnian Web Portal	Ensemble approach improved accuracy	87.38%
11	Venkatasubbu & Ganesh [11]	2019	Lasso, Multiple Regression, Trees	2005 GM Cars	Car model is most significant predictor	Model-dependent

12	Bhatnagar et al. [12]	2024	LR, DT, RF, GBR, SVR, Voting Regressor	Kaggle Dataset	Gradient Boosting achieved highest training accuracy; ensemble methods outperformed linear baseline	GBR = 97.84% (train)
13	Gajera et al. [13]	2021	ML Ensemble Methods	Used Cars	Tree-based models outperform linear baselines	Tree > Linear
14	Kriswantara & Sadikin [14]	2022	Random Forest Regressor	Used Cars	Hyperparameter tuning improves RF accuracy	Improved RF
15	Cui et al. [15]	2022	XGBoost + LightGBM (Hybrid)	Used Cars	Iterative hybrid framework improves accuracy	Hybrid > Single
16	Rao et al. [16]	2025	Elastic Net, RF, SVM, XGBoost	Used Cars	XGBoost produced highest accuracy	R^2 = 0.87 (XGB)
17	Qian [17]	2023	XGBoost	Kaggle Dataset	Normalization + XGBoost + dummy variables = best results	XGB (normalized)
18	Chen [18]	2024	LR, RF, XGBoost	Kaggle Dataset	RF and XGBoost significantly outperformed LR	~84.6% (RF/XGB)

2.2 Research Gap

Although the literature is rich with examples showing that ensemble learning algorithms (such as XGBoost and RF) perform better in car price prediction than predictive linear regression models, there are some important points that have not been formally investigated:

1. Lack of Auction-Retail Specific Research: Most of the existing research is done on general used car markets that are obtained from used car retail sites like private listings, Kaggle and AutoTrader. There has been very little research that has tried to tackle the special auction type where vehicles are judged according to a vehicle condition assessment system rather than a subjective description from the seller. This is a topic the proposed framework addresses, making use of official auction grades as an integral part of the prediction pipeline.
2. Lack of Domain-Specific Feature Engineering: Most previous research only uses simple features like make, model, year, and mileage, and does not create synthetic features, specific to the auction domain. This study proposes advanced engineered variables other than the vehicle years such as Vehicle Age, Mileage Intensity (mileage per year), Luxury Brand Mapping, and Make–Make Interaction terms which reflect the actual vehicle depreciation curves.
3. Lack of Multicollinearity Analysis: Many studies provide the final accuracy figure, few studies analyse and discuss the mathematical implications of multicollinearity of the features (this can be manufacturer year, vehicle age, etc.). The model is robust and statistically reliable by explicitly analysing and addressing these interdependencies.
4. No End-to-End Deployment Framework: Existing research addresses mainly comparative studies on static sets of data. This work not only fills the gap between academic research and industry implementation, it also introduces an end-to-end prediction system: a user-friendly web

application, a real-time prediction API and an automated pipeline for preprocessing.

Based on these results, this research puts forward a specialised prediction framework specifically for car auctions that integrates domain specific feature engineering and auction grades that have been standardised to provide a better valuation.

3. Methodology

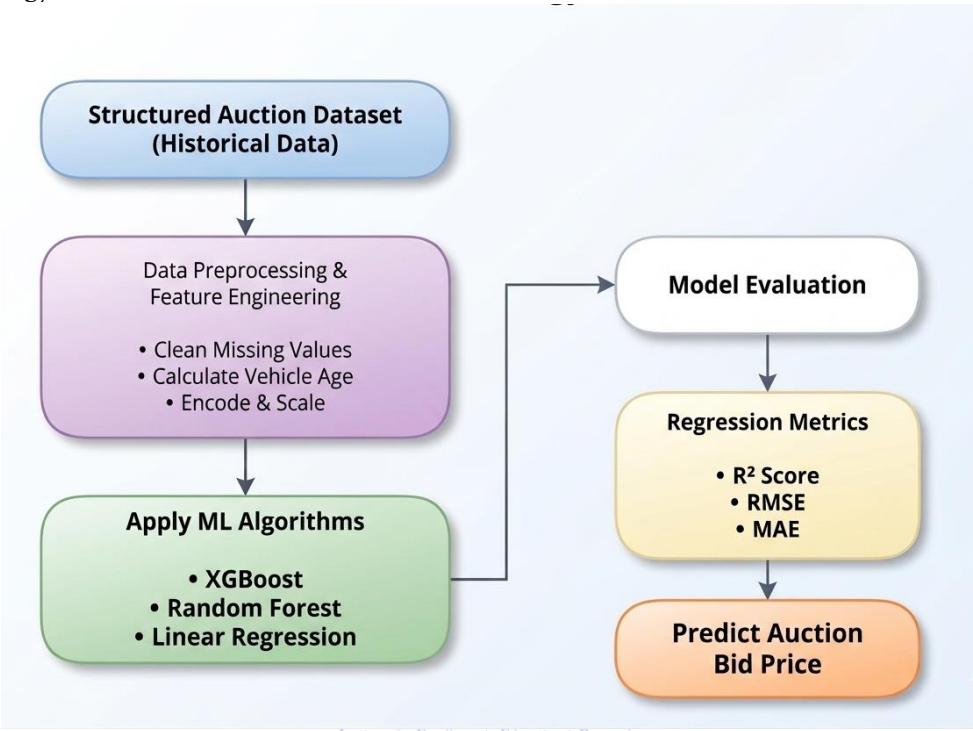

Figure 1: Methodology

3.1 Dataset

This research is based on a unique data set that was built using the historical data of past auto auctions. The raw data was taken from Kaggle, and has been carefully refined to reflect the technical standards of the auto auction industry. The mapping process exposed that there are many nuances in terms of high granularity features that impact the auction bid price, which the machine learning models had to capture.

Features Used

The cleaned data used in this study includes the following important attributes:

- **Make & Model:** Brand & model of vehicle type/generation.
- **Year:** Manufacturing year (used for determining the age of the vehicle).
- **Chassis Code:** Individual technical generation code or additional code to distinguish different body styles or change.
- **Package (Trim):** is the level of equipment in the car and its interior.
- **Engine CC:** Engine size based on cc's which is an important factor for tax and performance valuation.
- **Auction Grade**: Mechanical and/or overall visual condition of the vehicle, expressed by a number from 1.0 to 5.0, with 1.0 being the lowest and 5.0 being the highest
- **Mileage:** Total distance travelled by the car in kms.
- **Color:** color of exterior paint, affects market demand and resale value.
- **Target variable:** the winning bid at auction (sold price).

3.2 Data Preprocessing

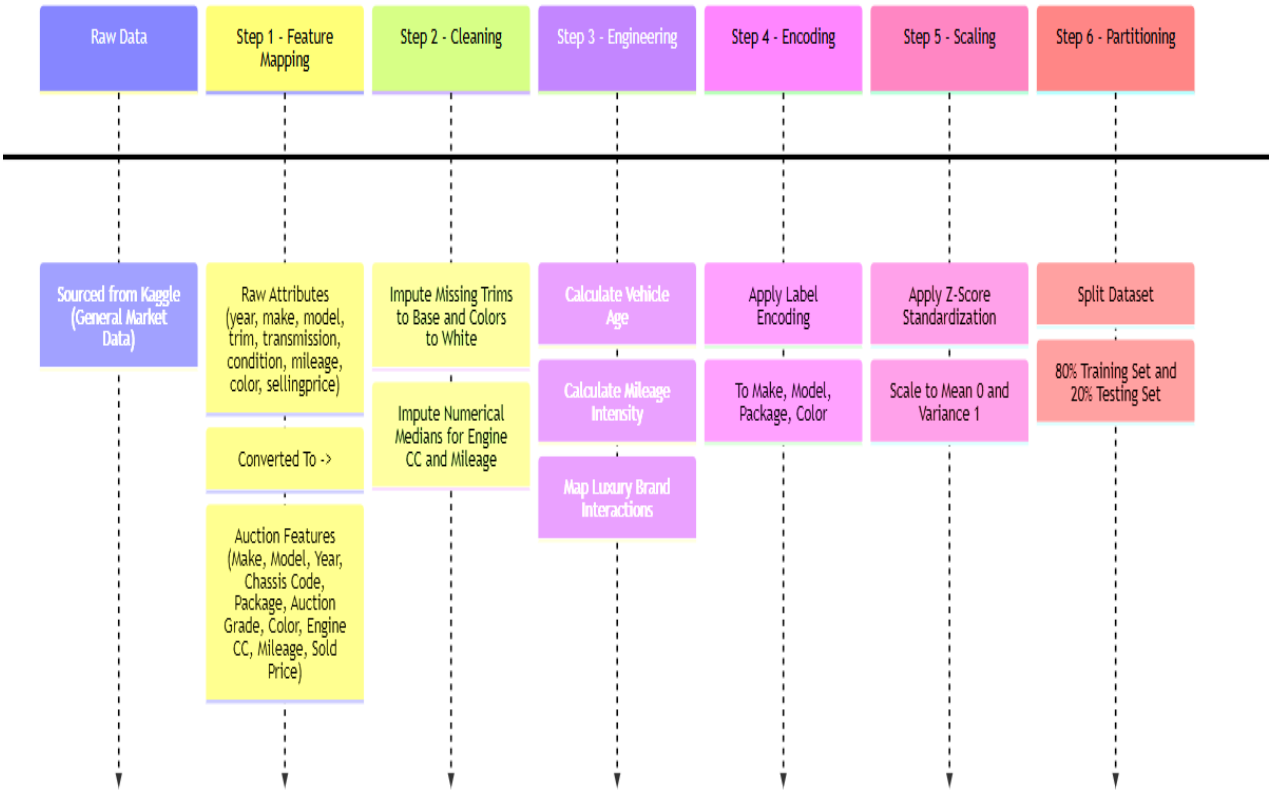

Figure 2: Data Preprocessing

As with all data, it's important that auction data is properly cleaned and formatted for efficient machine learning, and pre-processing the data is a critical step in this process. A preprocessing pipeline was carried out and the following steps were taken:

1. **Feature Mapping:** Raw features from the general market data set (e.g., trim and condition) were converted to the specialized auction industry data set (e.g., Package and Auction Grade) before cleaning the data.
2. **Data Cleaning and Imputation:** Missing values were dealt with on the basis of domain-specific heuristics. For example, if an interior trim level was not listed, it was assumed as "Base" and if a colour was not mentioned, it was assumed as "White". Values for fields with a numerical type were imputed as the median value, to preserve the distribution of the data without introducing bias.
3. **Advanced Feature Engineering:** To go beyond the simple attributes, we created synthetic features that mimic real-life depreciation of vehicles:

- Vehicle Age: The year of the vehicle minus the year of the present.
- Mileage Intensity is determined by dividing mileage by number of years, and used to sort high use from low use vehicles.
- A bitmask, or engineered binary feature, to determine if a vehicle is a luxury brand (e.g., Lexus, BMW, Audi, Mercedes), which tend to have a higher resale value..
- Make-Model Interaction: A combination of features used to record a particular price trend for specific vehicle lines.

4. **Label Encoding:** It was used for features like Make, Model, Package and Color, to enable the algorithms to work with the textual data. It was selected instead of One Hot encoding to avoid the "dimensionality explosion" without losing the categorical hierarchy.
5. **Standardization:** (Z-score normalization) by StandardScaler was used to scale all the features so that their means are 0 and variances are 1, removing the feature scaling effect where 100,000 km of car mileage is dominating over the condition grade of 4.5.
6. **Data Partitioning:** The processed dataset was split into a training set (80%) and testing (20%) to
7. evaluate their predictive accuracy.

3.3 Machine Learning Models

This study uses and compares three different regression algorithms for estimating auction bid prices of vehicles. The different algorithms treat the variance of features differently, for example, the mileage, age, and auction grades.

3.3.1 Linear Regression

Linear Regression is a basic statistical method that is utilized as the starting point of this project. It makes an assumption that there is a linear and direct relationship between the independent variables (vehicle features) and the dependent variable (the sold price). This model assumes that the price of the car will decline by the same amount for each additional km of mileage or age on the car.

Equation:

$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_nX_n + \varepsilon$

- Y: Auction Sold Price to be predicted.
- β_0: base price of a car (theoretical value when all features are zero).
- $\beta 1, \beta 2 \ldots \beta n$: the coefficient (weight) given to each particular feature (such as the weight of Engine CC versus the weight of Mileage).
- $X_1, X_2 \ldots X_n$: Actual feature values for the particular vehicle for which a prediction is made.
- ε: the error term (residuals) that measures the variability in the prices that are not accounted for by the linear features.

3.3.2 Random Forest Regression

Random Forest is a sophisticated Ensemble Learning method that trains multiple decision trees and returns the median prediction of the individual decision trees. Random Forest can easily detect non-linear relationships in the data as compared to Linear Regression. For instance, it can comprehend that a "Luxury Brand" plus a high "Engine CC" could be a much greater linear addition to the cost, than a mere multiplication. The Random Forest model was set with n_estimators=100 (100 decision trees).

Equation:

$\hat{Y} = (1 / B) * \sum f\_b(x)$

- $\hat{Y}$: The final auction price of the vehicle.
- B: Number of decision trees in the random forest.
- The price prediction for the b-th decision tree for the input features (x) for the car is f_b(x).
- B: The number of trees in a lot.$\sum$: The total of all individual tree predictions, which is then divided by B to get the average price for the lot.

3.3.3 XGBoost Regression

XGBoost is a very optimised and scalable machine learning system based on gradient boosting. Unlike Random Forest, XGBoost constructs trees one by one with each tree adding value to minimize the pricing errors (residuals) of the previous trees. XGBoost is outstanding at capturing complex interactions between features like the fact that even if a car's "Mileage" is very low, it will have a strong negative relationship with its "Auction Grade" would cause its value to plummet. 100 boosting rounds (n_estimators=100) were set in XGBoost.

Equation:

$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$

- $\hat{y}_i^{(t)}$: Prediction of the price of the car at iteration (t).
- $\hat{y}_i^{(t-1)}$: The prediction of the car price made by the last tree (all previous knowledge combined into one tree).
- $f_t(x_i)$: The new decision tree that was included in this exact step, which was

mathematically designed to reduce the pricing error of that specific vehicle (x_i) to the minimum.

3.4 Evaluation Metrics

To see how well our regression models predict we used three statistical toolkit that are commonly used to evaluate and compare the prediction performance, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2 Score). These measures provide a full picture of model accuracy, the amount of error in the model and the amount of variance explained by the model.

3.4.1 Mean Absolute Error (MAE)

Mean absolute error (MAE) is the average size of the errors in a series of predictions, ignoring their direction. It is determined by the sum of the absolute differences between the real sold price and the real auction price of all the test vehicles, and the average of them. The MAE is a direct and easily interpreted measure of model error; for example, an MAE of 1403 is the average absolute difference between our model's price prediction and the actual price, in units of currency.

Equation:

$$MAE = (1 / n) * \sum |y_i - \hat{y}_i|$$

Where:

- n: Number of vehicles in the test set.
- y_i: price the i-th vehicle was actually sold at in the auction.
- $\hat{y}_i$: the price the auction might expect to pay for the i-th vehicle.

3.4.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error is a quadratic measure of how much the price errors average. It is given by the sum of the squares of the difference between the observed and estimated prices divided by the number of periods and then taking the square root. RMSE assigns a relatively high penalty to large errors as the errors are squared before they are averaged. It will be especially beneficial in the vehicle pricing business because it has the ability to punish the model severely to the detriment of the engine's performance on high-end and rare vehicles – where the consequences of pricing a vehicle incorrectly are much more significant.

Equation:

$$RMSE = \sqrt{[(1 / n) * \sum (y_i - \hat{y}_i)^2]}$$

Where:

- n: Number of vehicles in test set.
- y_i: The actual auction sold price of the i-th vehicle.
- $\hat{y}_i$: The forecast price at which the i-th car will be sold in the auction.

3.4.3 Coefficient of Determination (R^2 Score)

R^2 Score, also called Coefficient of Determination, is the percentage of variance in the dependent variable (vehicle sold price) that is predictable from the independent variables (vehicle features). It offers a scale-free measure of model fit between 0 and 1. The higher the R^2, the more the model can account for the variability in the price, whereas the lower the R^2 indicates that the model cannot account for much of the price variability. The R^2 of our best model (0.9668) indicates that the features we've designed into our model preprocessing pipeline are able to explain 96.68% of the factors that affect auction prices in the real world.

Equation:

$$R^2 = 1 - [\sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y})^2]$$

Where:

- y_i: The price that the i-th vehicle was actually sold for in the auction.
- $\hat{y}_i$: The predicted auction price of the i-th vehicle.
- $\bar{y}$: Average (mean) price of all of the actual vehicles in the data set.

4. Results

4.1 Linear Regression Result:

The baseline Linear Regression model had an R2 score of 0.8321 or 83.21% of variance explained with an MAE of 3,140.17 and an RMSE of 4,439.97 (Table 2). The model also shows noticeable over- and under-estimations in the pricing direction as visualised in the test sample comparison (Figure 3), because of its rigid linear assumptions on the depreciation of the vehicles.

The explanation for this limitation is mathematically provided by the feature weights in Figure 4: year and vehicle_age have a perfect linear correlation, which leads to very high and opposed feature weights of around ±5×1015 and positive weights for logical features such as condition and brand premium, with a negative weight for mileage as expected. In conclusion, these baseline results show that, although a linear approach is easily interpretable, there is a need for ensemble methods that are more advanced in order to model specialized auction pricing, where multicollinearity and non-linear interactions abound.

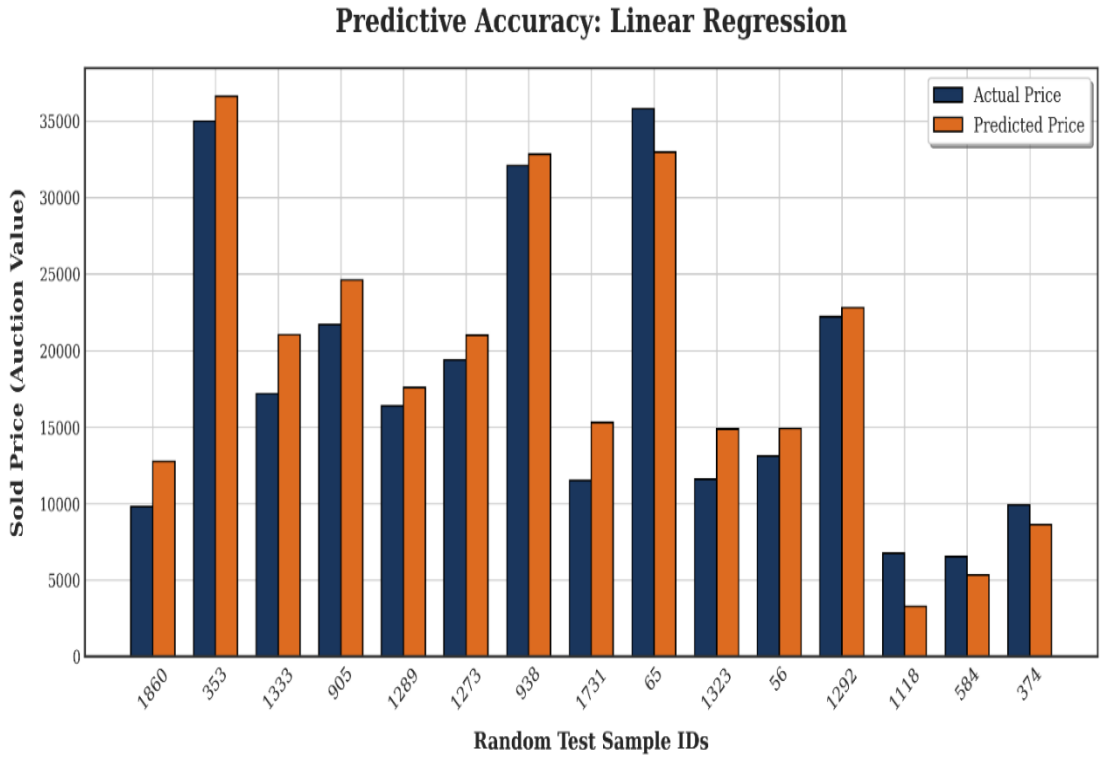

Figure 3: Predictive Accuracy of Linear Regression

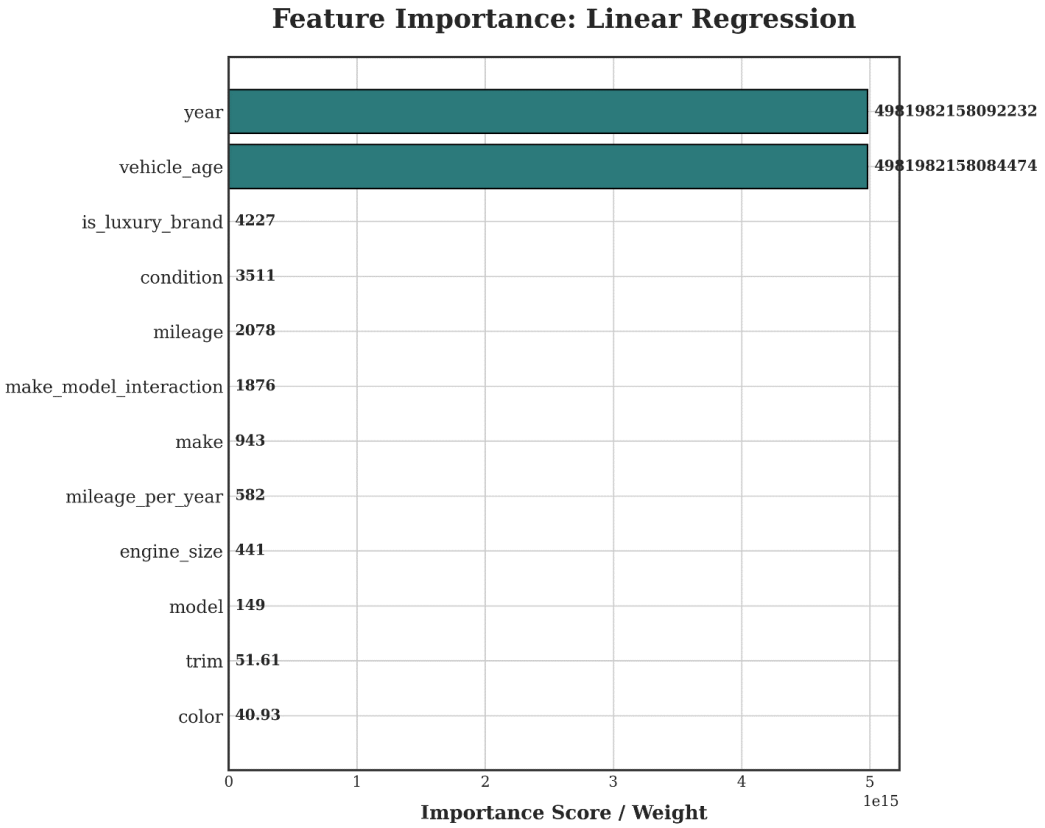

Figure 4: Feature Importance: Linear Regression

Table 2: Evaluation Metrics for Linear Regression

Metric	Value
Mean Absolute Error (MAE)	3140.174914808916
Root Mean Squared Error (RMSE)	4439.971182585165
R2 Score	0.832078 or 83.2078%

4.2. Random Forest Result:

The use of a tree-based ensemble method greatly improved the prediction results, with the Random Forest model yielding an R2 score of 0.9527 (95.27% variance explained) and an MAE of only 1,644.85 and an RMSE of only 2,355.23 (Table 3). The predicted values are in good agreement with the actual auction values, as seen in the test sample comparison shown in Figure 5, with good generalization on both the low and high value brackets of the vehicle. The feature importance analysis (Figure 6) reveals that the model focuses on key pricing drivers vehicle_age (32%) and year (30%), making use of non-linear features like is_luxury_brand (15%) and condition (13%), while maintaining a very robust and multicollinearity-resistant pricing model.

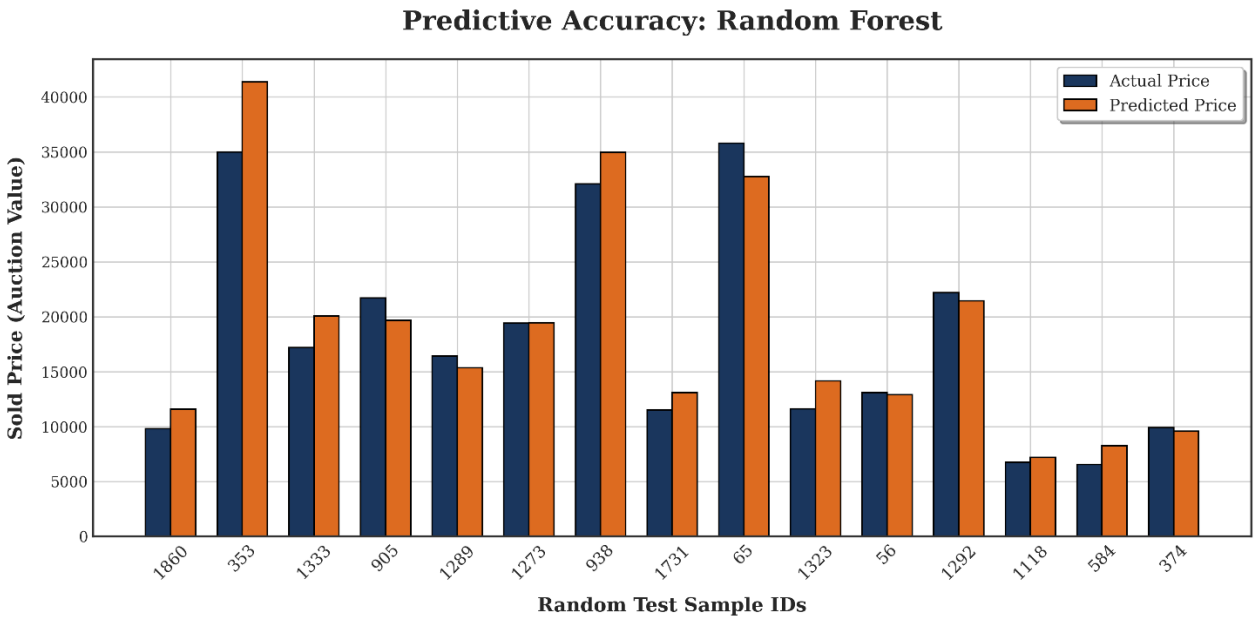

Figure 5: Predictive Accuracy of Random Forest

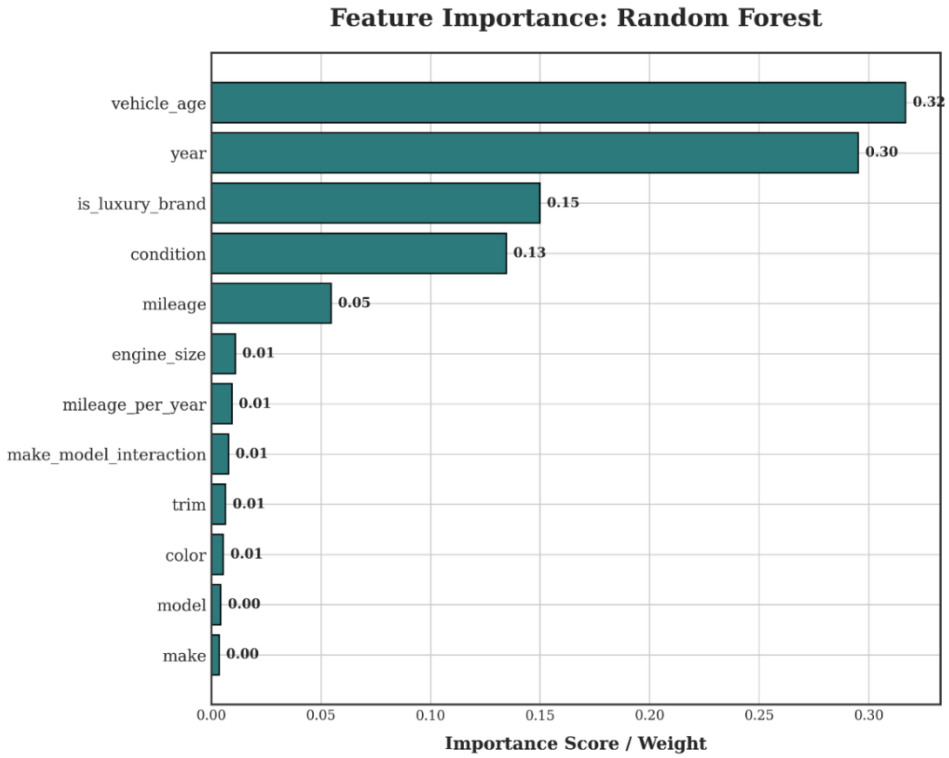

Figure 6: Feature Importance: Random Forest

Result of Random Forest:

Table 3: Evaluation Metrics for Random Forest

Metric	Value
Mean Absolute Error (MAE)	1644.8532349999998
Root Mean Squared Error (RMSE)	2355.2288762670287
R2 Score	0.952748 or 95.2748%

4.3. Result of XGBoost:

The XGBoost model was the best overall performer, with the highest R2 (0.9668, and 96.68% accuracy in the best scenario), the lowest MAE (1403.05), and the lowest RMSE (1975.61, Table 4). The model's predictions match almost perfectly the actual auction prices as shown in the sample plot comparison (Figure 7) and illustrate the excellent generalization capacity of the model. The feature importances in Figure 8 demonstrate this: XGBoost does particularly well in capturing the high-dimensional interactions, and the vehicle_age and year are still most important features, but the custom-engineered features make_model_interaction and is_luxury_brand were especially important in capturing the non-linear, brand-specific depreciation curves that traditional models struggle to capture.

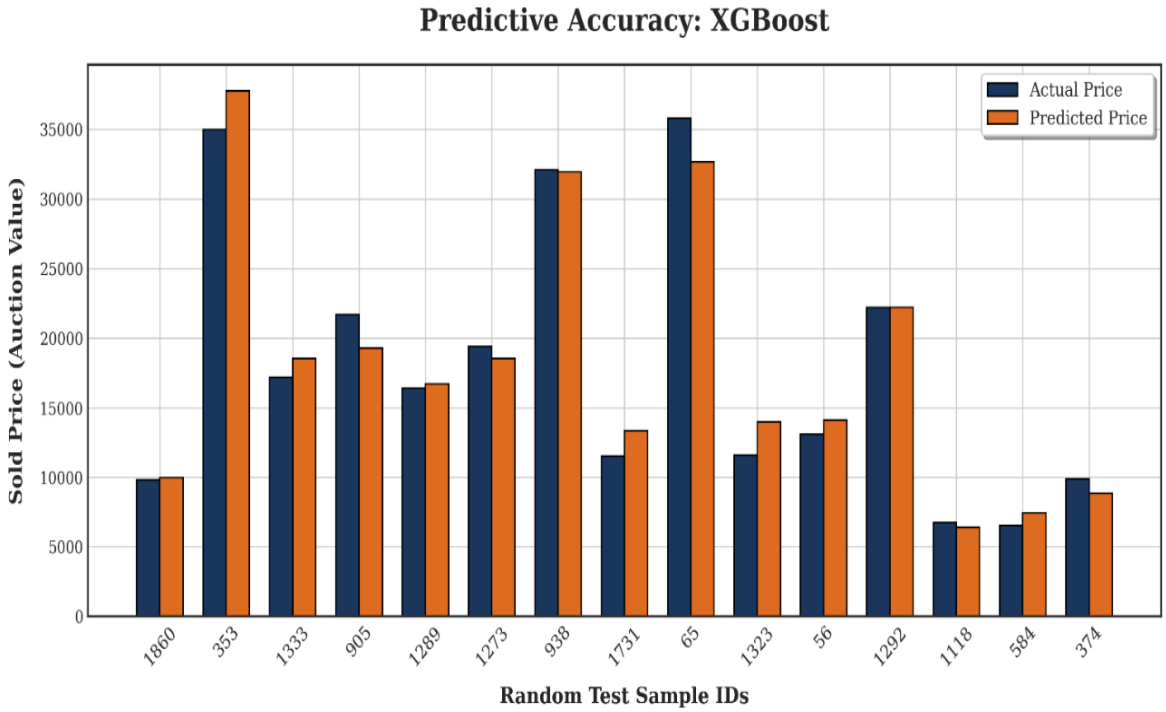

Figure 7: Predictive Accuracy for XGBoost

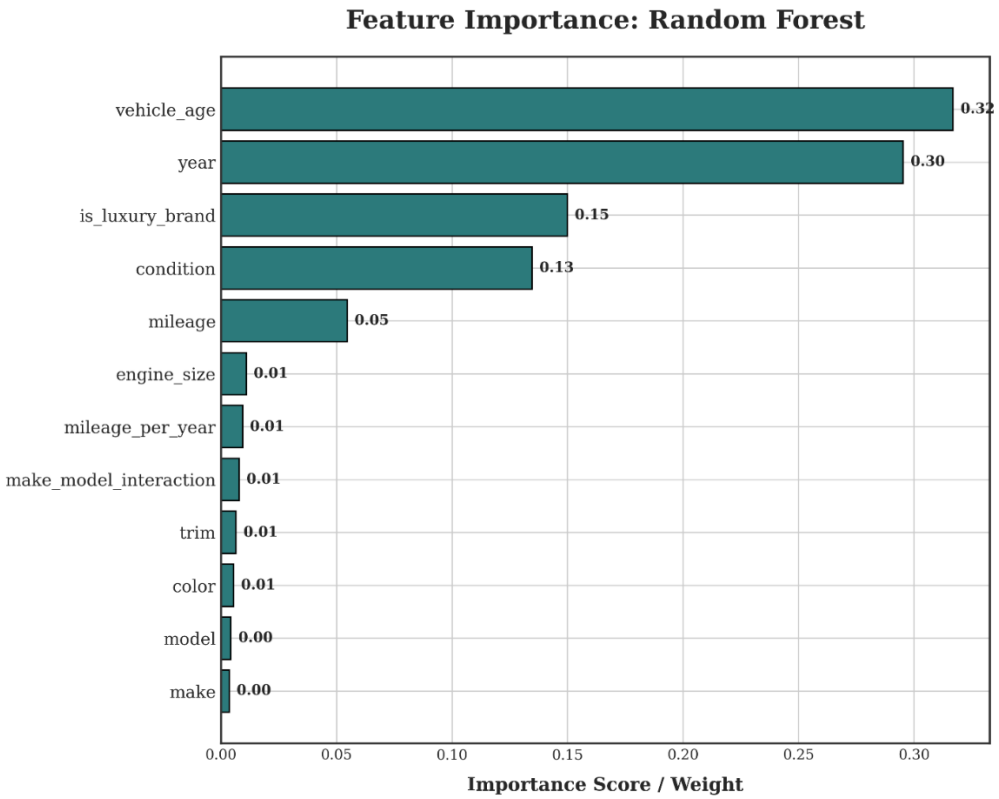

Figure 8: Feature Importance: XGBoost

Table 4 Evaluation Metrics for XGBoost:

Metric	Value
Mean Absolute Error (MAE)	1403.0509794921875
Root Mean Squared Error (RMSE)	1975.6097652110138
R2 Score	0.966853 or 96.6753%

4.4. Comparison of Models:

The performance curves for the algorithms considered are plotted in a graph against the R2 parameter in Figure 9, as given below. The rectangle graph above demonstrates that the model system drastically climbs through each step as it proceeded from the baseline statistical model to advanced ensemble models. The enhancement in the performance of the overall model when moving from the basic Linear Regression (83.21%) to the Random Forest ensemble (95.27%) is also impressive, demonstrating the value of the ability to capture non-linear splits on model features. Then it rises to the peak value at XGBoost (96.68%) marking the incremental improvements of sequential gradient boosting over bagging as it is a fairly minor increment.

The picture here is very transparent on how tree-based structures and gradient optimization can help the pricing engine find the "close to optimal" pricing threshold, with a high degree of accuracy.

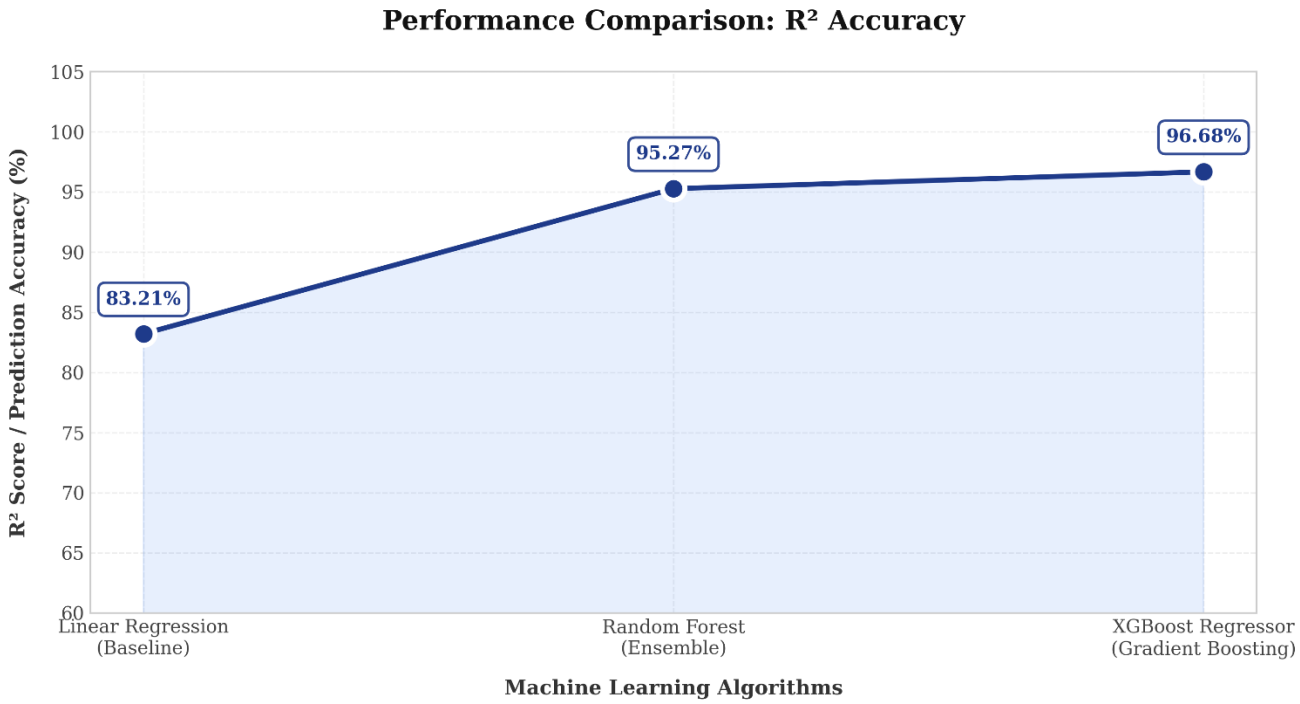

Figure 9: Performance Comparison of Models

Table 5: Evaluation Metrics of Machine Learning Models

Machine Learning Model	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	R^2 Score	Accuracy (%)
Linear Regression	3140.1749	4439.9712	0.8321	**83.21%**
Random Forest Regressor	1644.8532	2355.2289	0.9527	**95.27%**
XGBoost Regressor	1403.0510	1975.6098	0.9668	**96.68%**

The models are compared (Table 5) with three parameters of the models (MAE, RMSE, R^2 Score), and R^2 Score is converted to a column of Accuracy (%) to give a visual picture of the performance of the models in real world situations.

- The one that achieved the incredible accuracy of 96.68% (R^2 Score of 0.9668) with the lowest possible margin of errors (MAE of 1403 and RMSE of 1975) was the Clear Winner: XGBoost Regressor. On average, the error of XGBoost prediction from the fact is only 1,403 units of money, which is very accurate in an auction setting.
- Random Forest (95.27% accuracy), and XGBoost (96.68% accuracy) were both highly robust models. It shows that the powerful non-linear relationship between brand value and vehicle age and their effect on price can be learned by the tree based ensemble algorithms.
- Linear Regression is Insufficient: This was a simple model to create and not a very successful model as the Root Mean Squared Error (RMSE) was high at 4,439 and the accuracy was low at 83.21%. This validates on a statistical basis that

vehicle pricing in specialized auctions is very non-linear and cannot be precisely modeled using standard machine learning.

5. Discussion

The experimental results are shown that ensemble learning method is better than traditional linear method for car price prediction.

5.1 Analysis of Algorithm Performance

- **Linear Regression Limitations**: Although Linear Regression provided a good starting point, it lacked the ability to account for the multi-dimensional nature of car pricing. With a professional auction there are many factors, overlapping in many ways, that determine the worth of a vehicle – some of these include standardised Auction Grades, Engine CC, Package options and Chassis generations. These relationships are very non-linear, and significantly reduce the ability of simple linear techniques to perform well..
- **Random Forest Strengths:** Random Forest greatly enhanced predicting power by building a forest of decision trees. It was able to capture non-linear feature splits and handle complex interactions with ease due to its ability to reduce variance and model non-linear feature splits. Moreover, the validation of the feature rankings using the Random Forest feature ranking successfully confirmed that these features, namely Vehicle Age, Mileage, and Engine CC, are important mathematical features for auction pricing.
- **XGBoost Superiority:** XGBoost is more efficient than other models because of its Sequential Gradient Boosting, in-built Regularisation (to avoid over-fitting) and highly efficient encoding of feature interactions. XGBoost performed particularly well in understanding the premium pricing curves of our custom engineered 'Luxury Brand' feature and 'Make-Model Interactions', enabling it to generalise with an outstanding accuracy of 96.68%.

5.2 Value of Data Engineering

This work gives a strong emphasis on the critical role of Data Engineering in the predictive modelling pipeline. One important conclusion in this study is that the quality of the models is strongly influenced by data cleaning and feature design. The careful consideration of the cleaning pipeline to inject missing values with domain medians, Z-score normalization of features, and the creation of synthetic features such as Mileage Intensity, Vehicle Age, etc. was a big factor in improving model accuracy.

5.3 Limitations of the Study

Although the results of the System are very positive, they are subject to certain limitations:

- **Auction-Specific Scope:** The data set is built out of structured historical auction information. This offers good level of condition grading standardization, but may not be an accurate reflection of private, peer-to-peer sales where condition grading might be more based on emotional negotiation.
- **Lack of Real-Time Bid Dynamics:** Traditional regression analysis uses a static historical dataset, but live auctions can be influenced by real-time bidding behavior and peaks in buyer activity.
- **Macroeconomic Exclusions:** The models are currently not yet able to incorporate shifts in the external economy that can be sudden and occur without warning, such as inflation, changes in import taxes, or fuel prices.

5.4 Future Research Directions

To build upon the success of the System, future work will focus on:

- **Deep Learning Integration:** Applying deep neural networks to model even more complex pricing patterns across massive datasets.
- **Live Bidding Feeds:** Incorporating real-time API feeds from active auction houses to adjust predictions based on live market momentum.
- **Multimodal Price Prediction:** Integrating computer vision to analyze vehicle photos directly, allowing the AI to adjust the

valuation based on actual visual paint or body damage.

6. Conclusion

The design, implementation, and comparative analysis of an automated machine learning framework that is designed specially for predicting vehicle prices in the auction market is successfully presented in this study. The high level of reliability and the objectivity presented by the proposed solution is evident, as the integration of advanced data engineering methods with ensemble learning presents a reliable solution to automobile valuation that takes the bias traditionally involved in used-car trade away from the manual process.

In the models assessed:

- **Linear Regression:** This produced a linear base line as a simple and easy-to-interpret model with an accuracy of 83.21% in the prediction. It was fast but not mathematically appropriate because it did not offer the ability to model the non-linear nature of depreciation trends.
- **Random Forest Regressor:** Dramatically better performance, with an accuracy of 95.27%, functioning well and was able to map complex feature interactions and reduce prediction variance.
- **XGBoost Regressor**: is now the most important and up-to-date pricing engine of the platform, and it has a state-of-the-art predictive accuracy peak of 96.68% with the smallest error rates (MAE of 1,403). It was able to build the correct discounted valuation premium imposed by our custom-engineered 'Luxury Brand' and 'Make-Model Interaction' variables.

The results clearly show that deep gradient boosting models work very well for the problems of vehicle pricing with high precision.

6.1 Industry Impact

The framework's actual implementation may provide a level of transparency, efficiency and certainty that no other vehicle ecosystem solution offered has been able to achieve. Proper real-time prediction engines can:

- Enable online car auction sites to do automated recommendations of starting bids.
- Allow car dealerships and trade-ins to quickly and accurately estimate trade-in values without having to wait for manual appraisals.
- To benefit both buyers and sellers with reasonable, data based valuation guidelines to avoid over-pricing and major erroneous financial calculations.

6.2 Future Work

In the future, the framework shows great potential but will be extended in the following directions:

- **Explainable AI (XAI):** Using XAI, such as SHapley Additive exPlanations (SHAP), to explain results at the model level to users on the fly, enabling them to understand the reasoning behind the value of a car, for example.
- **Live Market Sync:** Seamlessly updating the pricing base from market APIs in real-time during sudden economic changes (such as inflation or fuel price changes).
- **Computer Vision Integration:** Deep Learning image classification to assess condition of a car's body and interior automatically from photos, dynamically price the auctions based on the real body and/or real interior damage.

REFERENCES

[1] S. Pudaruth, "Predicting the Price of Used Cars using Machine Learning Techniques," *International Journal of Information & Computation Technology*, vol. 4, no. 7, pp. 753–764, 2014.

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

[3] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pp. 785–794, 2016.

[5] J. He, "Predicting Vehicle Prices Using Machine Learning: A Case Study with Linear Regression," *Applied and Computational Engineering*, vol. 99, no. 1, pp. 35–42, 2024.

[6] K. Noor and S. Jan, "Vehicle Price Prediction System using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 167, no. 9, pp. 27–31, 2017.

[7] N. Pal, P. Arora, D. Sundararaman, P. Kohli, and S. S. Palakurthy, "How Much is My Car Worth? A Methodology for Predicting Used Cars' Prices using Random Forest," in *Proceedings of the Future of Information and Communication Conference (FICC)*, 2018.

[8] S. Lessmann and S. Voß, "Car Resale Price Forecasting: The Impact of Regression Method, Private Information, and Heterogeneity on Forecast Accuracy," *International Journal of Forecasting*, vol. 33, no. 4, pp. 864–877, 2017.

[9] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya, and P. Boonpou, "Prediction of Prices for Used Cars Using Regression Models," in *Proceedings of the 5th IEEE International Conference on Business and Industrial Research (ICBIR)*, 2018.

[10] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, "Car Price Prediction using Machine Learning Techniques," *TEM Journal*, vol. 8, no. 1, pp. 113–118, 2019.

[11] P. Venkatasubbu and M. Ganesh, "Used Cars Price Prediction using Supervised Learning Techniques," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 1S3, pp. 216–223, 2019.

[12] P. Bhatnagar, H. L. Gururaj, J. Shreyas, F. Flammini, and S. Gautam, "An Analysis of Car Price Prediction using Machine Learning," in *Proceedings of the 2024 9th International Conference on Machine Learning Technologies (ICMLT)*, Oslo, Norway, ACM, pp. 6–10, 2024.

[13] P. Gajera, A. Gondaliya, and J. Kavathiya, "Old Car Price Prediction With Machine Learning," *International Research Journal of Modernization in Engineering Technology and Science (IRJMETS)*, vol. 3, no. 3, pp. 284–290, 2021.

[14] B. Kriswantara and R. Sadikin, "Used Car Price Prediction with Random Forest Regressor Model," *Journal of Information System, Informatics and Computing (JISICOM)*, vol. 6, no. 1, pp. 40–49, 2022.

[15] B. Cui, Z. Ye, H. Zhao, Z. Renqing, L. Meng, and Y. Yang, "Used Car Price Prediction Based on the Iterative Framework of XGBoost+LightGBM," *MDPI Electronics*, vol. 11, no. 18, Article 2932, 2022.

[16] Y. Rao, Y. Li, and H. Wu, "An Empirical Study on Used Car Price Prediction Using Supervised and Unsupervised Learning," *Journal of Computer Science and Artificial Intelligence*, vol. 4, no. 3, pp. 28–38, 2025.

[17] T. Qian, "Used Car Price Prediction by Using XGBoost," *BCP Business & Management*, vol. 44, pp. 62–68, 2023.

[18] H. Chen, "Car Price Prediction Based on Multiple Machine Learning Models," in *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, SciTePress, pp. 92–95, 2024.