Used Car Price Prediction and Feature Importance Analysis Using XGBoost and SHAP

Tianran Li

School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215000, China

Tianran.Li23@student.xjtlu.edu.cn

Abstract. The used car market has rapidly expanded; therefore, the importance of the accurate and transparent pricing of a used car has increased. Traditional approaches to pricing usually rely on linear regression to infer a clear-cut mathematical relationship between vehicle attributes and pricing. This paper presents an optimized machine learning approach to predict the used car price using an Extreme Gradient Boosting (XGBoost) model. Above all, it always beats the baseline regression models. Using a dataset of 7253 records from India, it compares Linear regression, Ridge, Lasso, Random Forest, and XGBoost models. The findings from experiments reveal that XGBoost offers the highest predictive performance with a score of 0.87, and also possesses the lowest root mean squared error, compared to 0.80 for Linear Regression. To make our model more interpretable, it uses Shapley Additive Explanations (SHAP) to analyze how power, year, km driven, transmission, and engine influence price. The understanding and improvement of online pricing transparency will pave the way for consumers, dealerships, and online players to take effective action.

Keywords: Machine learning regression used to evaluate used car pricing.

1. Introduction

The used car market is influenced by the age of a vehicle, kilometres driven, engine specification, and demand in the market [1, 2]. The used car market is not as simple as it looks because of many hidden factors. An accurate prediction of price is necessary for the buyer, seller, and dealership. Traditional statistical methodologies for model fitting, such as linear regression (LinR), cannot capture the complex non-linear relationships typically prevalent in this domain [3-5].

This project is a continuation of a Kaggle project that used LinR [3]. This study uses Extreme Gradient Boosting (XGBoost) to study which attributes affect used-car prices the most [6-8]. The primary aim is to make better predictions and offer interpretable insights. The present paper describes an optimized XGBoost model achieving a 0.87 R^2 score. Further, it applies Shapley Additive Explanation to discover auto price drivers per cent [9]. The result ensures a certain enhancement in the pricing decision-making by consumers, dealers, and online platforms.

2. Related Work

The prediction of used-car prices is gaining popularity in the field of data science and has become a common topic in various data science competitions and open-source projects [3, 6, 7]. The Kaggle platform provides assorted datasets and benchmark models for various researchers dealing with the problem [1]. Many studies use LinR as the baseline method since it generally uncovers base-level associations between the vehicle features and pricing, yet may fail to capture other complicated relationships due to potential nonlinearities [3-5].

This project aims to further predict the price of used cars based on already existing research done on Kaggle. It makes use of Ridge, Lasso, and some other advanced machine learning models, such as Random Forest (RF), XGBoost, etc., to improve upon the baseline LinR performance, as simple regression models are found to be limited in their effectiveness [6, 8, 10]. In addition to predicting vehicle price accuracy, this study emphasizes interpretability by looking at the most influential features [2].

3. Methodology

3.1. Dataset Description

The information used in this study originates from the Kaggle public site, comprising 7253 records of Indian used cars [10]. Each record is a transaction and will contain any technical characteristics of the vehicle and transaction details. The fundamental identifiers include serial number, model name, sales place, vehicle-specific features such as the year of manufacture, mileage, fuel type, transmission type, ownership history, odometer reading, engine displacement, power, seating capacity, etc [2, 7]. The dataset consists of New Price (showroom price, often missing for older vehicles) and the second-hand price, and the column Price (which is the actual transaction price) is the target variable. The dataset consists of various brands, regions, and conditions of the vehicles, making it ideal for second-hand vehicle price prediction modelling

3.2. Data Preprocessing and Feature Engineering

The dataset used featured various data types, numerous missing values, and formatting inconsistencies that required heavy preprocessing. The first thing the team did was run through some steps to clean the data. Here, they ensure that variables like Engine and Power, which were in the string format like (1498 CC or 103.5 bhp), were converted to a number format. Also, they ensured that the Mileage, which was also in string format with varying levels of units (kmpl and km/kg), was standardized. The irrelevant identifier S.No. was removed. The missing values were treated using imputation techniques. Continuous variables like Engine, Power, and Mileage were imputed with their respective median values, while categorical variables like Seats were imputed with the mode. To ensure the validity of the model, records that did not have the Price target were removed.

Feature engineering aimed to enhance the dataset quality. A new feature-car age was created by subtracting the manufacturing year of a car from 2021, which reflects the depreciation of the car. The New Price variable was cleaned by removing any textual units like “Lakhs” and converting it to a numeric. However, a significant number remained missing. For categorical variables, One-Hot Encoding was used for Fuel Type, Transmission, Owner Type, and Location [3]. The Name variable was dropped because it had very high cardinality. A high cardinality variable leads to sparsity in the feature space.

3.3. Model Selection

Both LinR models and tree-based ensemble models are used to ensure the comparability of results [1, 3, 10]. The linear models were LinR, ridge regression, as well as lasso regression. Due to their simplicity and interpretability, linear models are highly regarded, but do not capture non-linearity a lot.

The LinR expects that a linear relationship exists between the target variable and the predictors. The parameters of the linear model are estimated by least squares. But if the characteristics are highly correlated, this model is susceptible to multicollinearity and overfitting. The problem of overfitting can be addressed using ridge regression, which modifies the cost function. It helps improve the generalization ability of the model. On the other hand, Lasso regression performs both regularization and variable selection simultaneously. It does this by adding an L1 regularization term that sets some coefficients to 0.

In contrast, tree-based models such as RF and XGBoost are more advanced methods that can easily model nonlinear models, as well as intricate interactions between variables. Random forest is an application of ensemble learning [4, 10]. It utilizes bootstrap sampling and takes the mean of the predicted values of each of them. A large number of decision trees are created using a random sampling of features. The mean value is taken. This reduces the dispersion. Similarly, overfitting is not a problem here. XGBoost (Extreme Gradient Boosting) takes things a step further as it builds the trees sequentially in a way that each new tree learns from the residuals of the previous tree. It is very

powerful for tabular data because it uses advanced techniques such as gradient-based optimization, normalization, and contraction [1, 6, 7].

These models have proven to be effective predictors on structured data [1, 6, 8]. For example, used car price predictions, where variable interrelationships are highly nonlinear.

3.4. Evaluation Metrics

Three commonly used regression metrics were used to evaluate the performance of the model. Mean Absolute Error (MAE) was selected as a measure of the average error made in the prediction. It provides a good indication of accuracy. Root mean squared error (RMSE) was included, which penalizes bigger errors more severely and hence gives insight into the model's sensitivity to outliers. To conclude, the determination coefficient (R^2) was calculated to show the percentage of variation of the target variable explained by the model. To make the results more robust and generalizable, all the models were subjected to a five-fold cross-validation. The aim was to ensure that performance comparison was not dependent on a train-test split.

4. Result

4.1. A comparison of RF and XGBoost

Through the testing of regression models, it was determined that RF and XGBoost were the two best-performing methods by far, both significantly outperforming LinR methods [1, 10]. The outcome indicators for different models are consolidated in Table 1.

Table 1. The comparison results of model performance

Model	MAE	RMSE	R2
LinR	3.47	6.57	0.80
Ridge	3.22	6.40	0.81
Laaso	4.70	8.47	0.67
RF	2.44	5.84	0.84
XGBoost	2.16	5.28	0.87

The comparison results show that XGBoost is slightly better than RF on all measurement metrics, proving that its overall fitting capability is stronger [1, 7, 8]. Moreover, it is much stronger at variance capturing. XGBoost is a predictive analysis model that relies on an iterative learning process and has features that make it a powerful and effective model.

The predicted value has a close linkage to the actual one. Fig. 1: Compared Actual and Expected Values.

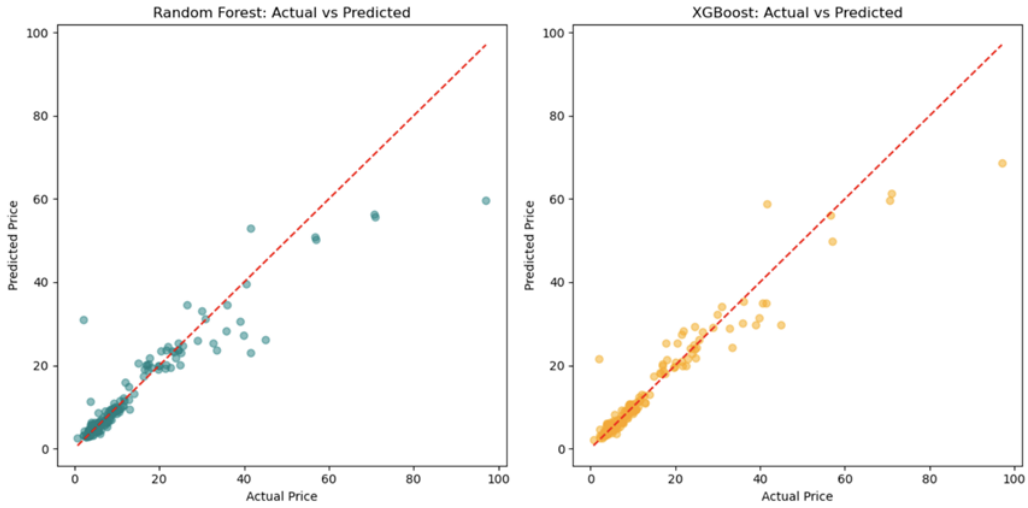

Fig. 1 Predicted vs. actual values (Photo/Picture credit: Original).

The predictions made by XGBoost cluster closer to the diagonal line than those made by RF, indicating that, overall, the bias and variance of predictions are less than those of RF.

Thus, based on these experiments and evidence from previous studies, the final model selected for deeper interpretability analysis based on SHAP is XGBoost.

4.2. SHAP Analysis for Feature Interpretability.

XGBoost is the highest predictive model, but tree-based ensemble models are often criticized for a lack of interpretability. To overcome this, the authors used SHAP (Shapley Additive Explanations), a model-agnostic interpretability scheme based on cooperative game theory [9].

The SHAP framework assigns individual feature contributions for prediction [9]. It indicates how much a feature adds to or subtracts from the prediction of the car price. Two types of SHAP visualizations were employed.

Fig. 2 shows what inputs contribute most on average across all predictions.

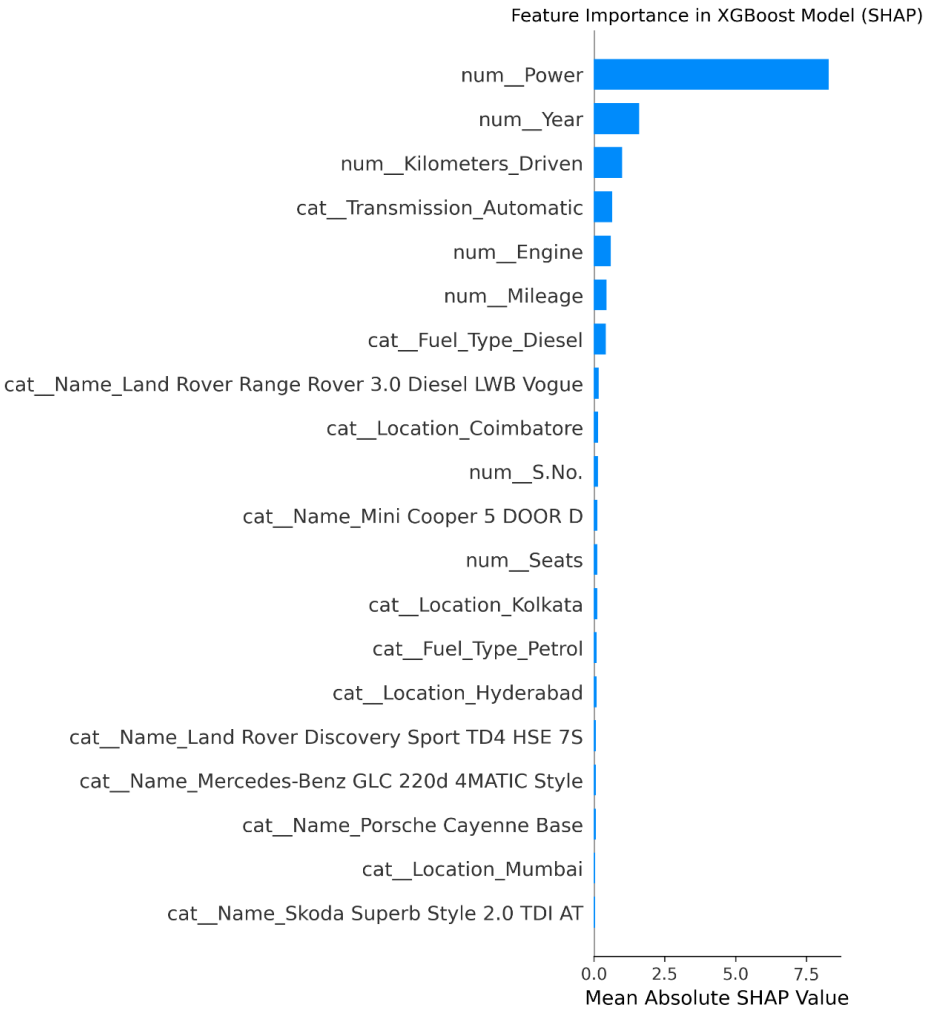

Fig. 2 Feature importance (Photo/Picture credit: Original).

The analysis found that the Power, Year, Kilometers Driven, transmission, and Engine were the most important.

As it can see in Fig. 3, the Summary Plot enables the visualization of the size and direction of the impact of their features.

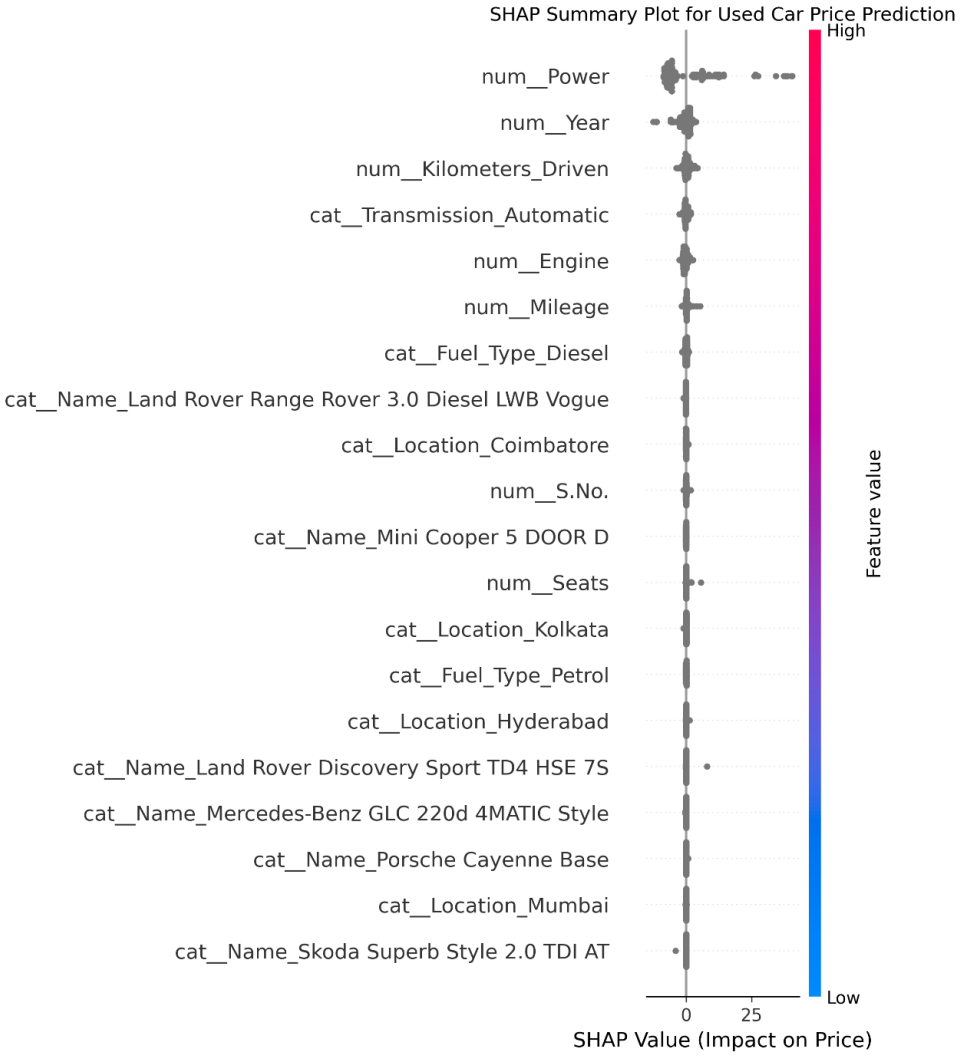

Fig. 3 SHAP beeswarm plot (Photo/Picture credit: Original).

In analysing the data, it can be seen that more horsepower and a larger engine displacement correlate with a price increase. This is clearly because of such a feature on the vehicle, enhancing the latter's performance greatly. Furthermore, a newer model year, which corresponds to a lesser age of the vehicle, generally is more expensive, owing to common sense. On the contrary, the kilometers driven hurt the price, but are one of the yardsticks. In other words, the more you drive, the less the price will be, but not with maximum impact. The more the damage, the greater will be the impact on the price. A more important point is the type of transmission. For instance, an automatic will cost more than a manual [7, 8].

This analysis of interpretability confirms what it knows regarding a used car price; it also provides quantitative evidence. This study contributes both practically and theoretically to the valuation of used cars by integrating predictive accuracy (using XGBoost) and interpretability (using SHAP).

5. Conclusion

This writing demonstrates that XGBoost outperforms other techniques while predicting the prices in a used car dataset, with R^2 matching 0.87. Moreover, the study applies SHAP to reveal that power and year are two important features. Gaining insights into relevant markets enables market transparency and decision-making, and interpretable AI.

Using the dataset from Kaggle, which contained 7,253 transactions, this study constructed an optimised machine learning framework to predict the used car prices. LinR was the baseline model, and tree-based ensemble models performed significantly better than LinR. The performance

comparison showed that the XGBoost model had the highest R-squared value of 0.87. Moreover, it also had the minimum MAE and RMSE, reflecting its robustness in studying feature interactions. The SHAP analysis revealed what affects the price of used cars. Power, Year, Kilometers Driven, transmission, and Engine.

In conclusion, XGBoost and SHAP can obtain an interesting balance of associated with their interpretation and prediction capabilities. Future projects may include advanced deep learning techniques that lead to better analysis and predictions, further enlarging the scope of this framework.

References

[1] Bergmann S, Feuerriegel S. Machine learning for predicting used car resale prices using granular vehicle equipment information. Expert Systems with Applications, 2025, 263: 125640.

[2] Huang J. Price prediction and analysis of price influencing factors for second-hand car sales in AutoTrader based on XGBoost algorithm. World Scientific Research Journal, 2025, 11(9).

[3] Darade Y. Predict used car prices using linear regression. Kaggle, 2021.

[4] Li C. Machine learning-based models for accurate car prices prediction. Highlights in Business, Economics and Management, 2024, 40.

[5] Pal N, Arora P, Sundararaman D, Kohli P, Palakurthy S. How much is my car worth? A methodology for predicting used car prices using RF. Future of Information and Communications Conference (FICC), 2017.

[6] Qian T. Used car price prediction by using XGBoost. BCP Business & Management, 2023, 44: FIBA 2023.

[7] Naveen Reddy S, Kumar S. A comparative analysis of XGBoost model and AdaBoost regressor for prediction of used car price. Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT), 2023: 441–446.

[8] Guo S, Zhang B. Revolutionizing the used car market: Predicting prices with XGBoost. Proceedings of the 4th International Conference on Signal Processing and Machine Learning, 2024, 48.

[9] Wang H, Liang Q, Hancock J T, Khoshgoftaar T M. Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. Journal of Big Data, 2024, 11: 44.

[10] Fatima S, Hussain A, Amir S B, Ahmed S H, Aslam S M H. XGBoost and RF algorithms: An in-depth analysis. Pakistan Journal of Scientific Research, 2023, 3(1): 26–31.