A Comparative Study of Linear Regression and Random Forest Models for Predicting Used Car Prices

Chiyu Zhou

The University of Sydney, Sydney, NSW 2006, Australia
czho0682@uni.sydney.edu.au

Abstract. This study deeply analyzed the problem of used car price prediction based on machine learning methods. By constructing two models, linear regression and random forest, and comparing their prediction performance, the essential influence of model structure on price prediction accuracy and generalization ability was explored. The study used public data sets for strict data preprocessing and feature engineering. The results showed that the random forest model was significantly better than linear regression in prediction accuracy, which was particularly prominent in the scatter plot of actual and predicted prices. At the same time, through the feature importance analysis of random forests, it was found that the number of engine cylinders and fuel type have a key impact on vehicle pricing, which further confirmed the ability of random forests to effectively capture nonlinear features. Although there is a certain skewness in the residual distribution of random forests, it is suggested that advanced models such as gradient boosting trees and external data can be further introduced in the future to improve prediction accuracy and robustness.

Keywords: Used Car Price Prediction, Linear Regression, Random Forest Regression

1 Introduction

As the global automobile market gradually develops towards refinement and digitalization, used car price prediction has become an important research topic of common concern in academia and industry. As one of the key factors affecting transaction decisions, accurate prediction of automobile prices can not only help consumers and dealers effectively reduce the risks brought by information asymmetry in the transaction process, but also help improve the overall efficiency of market transactions and further promote the healthy development of the industry.

However, the process of determining automobile prices is extremely complex, which includes nonlinear interactions between many technical parameters and market factors, making it difficult for traditional statistical prediction methods to effectively deal with it. Machine learning technology has gradually shown significant application advantages due to its powerful data analysis capabilities and ability to capture nonlinear features.

A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_18

Recent studies have shown that models based on machine learning algorithms perform well in vehicle price prediction problems, especially in capturing complex nonlinear relationships and interactions between high-dimensional features. Sharma and Sharma used a linear regression algorithm combined with one-hot encoding for prediction and achieved a high prediction accuracy [1]. This result indirectly reflects the importance of data preprocessing and feature transformation for improving the performance of linear models. But their study also illustrates the naïveté of linear regression models on high-dimensional data with complex structures. In contrast, Bharambe et al. also pointed out that the prediction performance of the Lasso regression model, by in-depth comparison with multiple regression algorithms, is much better than that of traditional linear regression, which means introducing a regularization mechanism can help the generalization ability of the model with high-dimensional sparse data structure [2]. However, more advanced nonlinear ensemble learning models, such as random forests, have shown more outstanding performance due to their low assumptions on data distribution and their ability to automatically capture nonlinear relationships [3]. Gajera and Gondaliya further confirmed in their study that instance-based nonlinear methods such as K-nearest neighbor (KNN) can achieve lower prediction errors and higher R^2, further verifying the application potential and practical advantages of nonlinear models in automobile price prediction tasks [4].

Accordingly, this study chose two traditional models in the field of machine learning, including LR and RF, and took advantage of public automobile transactions data. By means of the data preprocessing, feature engineering, and the severe model training and evaluation, the practical performance and the suitable environment of various models in the used car price prediction task were profoundly studied.

2 Dataset and Preprocessing

The data used in this study comes from the public vehicle price prediction dataset provided by the Kaggle platform. The dataset records the transaction information of 1,002 cars of different brands, models, and configurations, involving 17 different dimensions of features such as vehicle manufacturing year, fuel type, number of cylinders, and drive mode, and uses the vehicle listing price as the target variable of the study [5].

To ensure the high quality and validity of the model input data, the study strictly implemented the data cleaning procedure in the preprocessing stage, in which the records with missing target variable prices were deleted to avoid interfering with model training. Considering the multicollinearity problems and feature redundancy that may occur in model training, some high-cardinality or unstructured text information in the original data set was excluded, such as engine model, interior, and exterior color etc. In addition, in response to the problem of missing data, this study eliminated missing variables and used median filling to reduce the potential impact of outliers on data distribution [6]. In the end, the original data set retained 979 high-quality valid records.

On this basis, this study conducted deep feature engineering operations on categorical features to increase the adaptability of categorical variables to regression modeling. By performing one-hot encoding on variables such as vehicle brand model, transmission, and drive mode, the effective conversion of categorical features to numerical features is achieved [7]. This eliminates the false distance and sequence assumptions that may arise in model construction, allowing the feature matrix to maintain a reasonable structure after the dimension is expanded to more than 40, enhancing the ability of subsequent regression models to capture complex nonlinear relationships and generalization performance.

3 Methodology

Based on the supervised learning paradigm, this study selected two classic models, LR and RF, for modeling and analysis. The LR model is a classic prediction method based on the least squares method. It makes a strict linear assumption about the relationship structure in the data. Although this assumption can make the model have high explanatory transparency and computational efficiency, it may also oversimplify the nonlinear feature interaction effects and complex patterns that are widely present in real data, thereby limiting the generalization performance of the model in actual situations [5].

In stark contrast to linear models, random forest regression models are based on decision trees. Through random sampling and feature subspace splitting, the prediction results of a large number of weak learners are integrated to effectively capture the complex nonlinear relationships and high-order interactions between features [7]. The random forest algorithm is especially suitable for use in this work, because in the process of predicting the price of used cars, there exist numerous complicated relationships, multidimensional mutual influence among features between which are difficult to premodel [1]. Random forests are good in the way that they automatically capture and learning these patterns into the models, without a lot of pre-assumed, and they are tailored for the complex, data-intensive application context used in car price prediction [3].

To strictly evaluate the predictive ability of the model, this paper selects two traditional statistics, RMSE and determination coefficient (R^2) as evaluation criteria. RMSE evaluates the absolute prediction performance of a model by computing the average error between the pv and the rv. The lower the error, the more accurate the model. R^2 examines the model's ability to explain the data fluctuation trend from a relative perspective. The closer the value is to 1, the stronger the model's explanatory power for the target variable [8].

4 Experiment and Training Details

This study follows the standardized operation process under the supervised learning paradigm, uses Python as the modeling basis, and combines scikit-learn for construction and evaluation [8]. The training data and test data are randomly divided

into 80% and 20% to ensure the generalization ability of the model on unseen data, while avoiding the potential risks of overfitting and information leakage. The parameter training process of the linear regression model directly implements the analytical solution based on the training set, and uses the least squares method as the objective function to achieve the optimal parameter estimation, while the random forest model determines the best hyperparameter combination through a five-fold cross-validation strategy [7].

5 Experiment Result

The experimental results show that the model structure has a significant impact on the prediction performance in the used car price prediction task. From the prediction scatter plot of the linear regression model (Fig. 1), its point cloud is roughly distributed around the diagonal reference line, reflecting that the model has a certain linear fitting ability for most price samples. However, in the high price range, the scatter points deviate significantly from the ideal prediction line, and the predicted values are systematically underestimated. This is consistent with the model's medium-level performance results of RMSE of approximately US$11,102 and R^2 of only 0.674. This phenomenon shows that the model failed to capture the nonlinear jump trend of high-end vehicle prices, resulting in a significant increase in the degree of prediction distortion. The model still has difficulty in handling the complex nonlinear characteristics and multidimensional interactions implied in the price distribution [1].

Fig. 1. Linear Regression - Actual vs. Predicted Car Prices (Photo/Picture credit: Original).

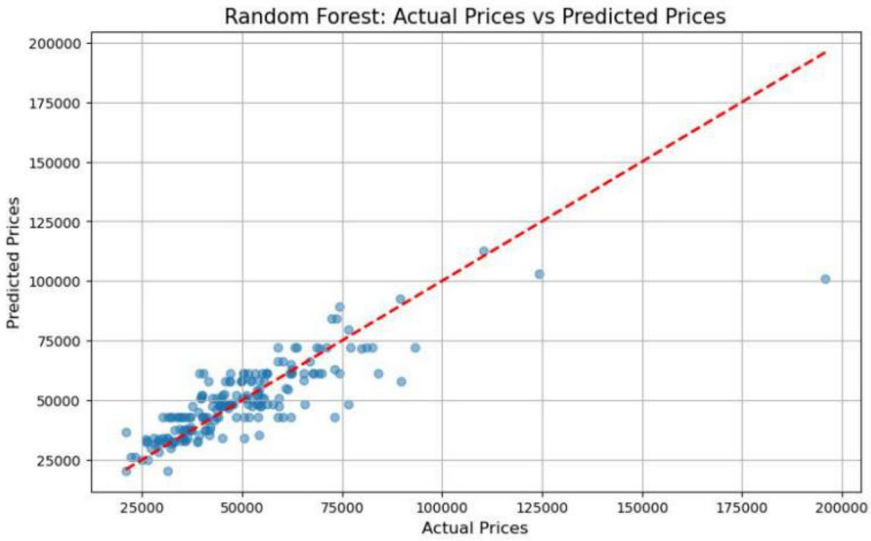

Fig. 2. Random Forest - Actual vs. Predicted Car Prices (Photo/Picture credit: Original).

In comparison, the random forest model shows a superior prediction ability. As can be seen from the prediction scatter plot in Fig. 2, the prediction values of the random forest model are more closely clustered around the actual values, with its RMSE reaching $11,032 and R^2 increasing to 0.678, both of which are better than the corresponding indicators of LR. This reflects the advantages and practical application value of the random forest model in predicting used car prices. At the same time, the feature importance ranking of the prediction results of the two variables of engine cylinder number (cylinders) and fuel type (fuel_Gasoline) presented in Fig. 3 is also consistent with the empirical observations of the real market (Fig. 4). The research of Pal et al. shows that engine performance and fuel economy often occupy a core position in the market valuation process, directly affecting the market demand and valuation level of vehicles [3].

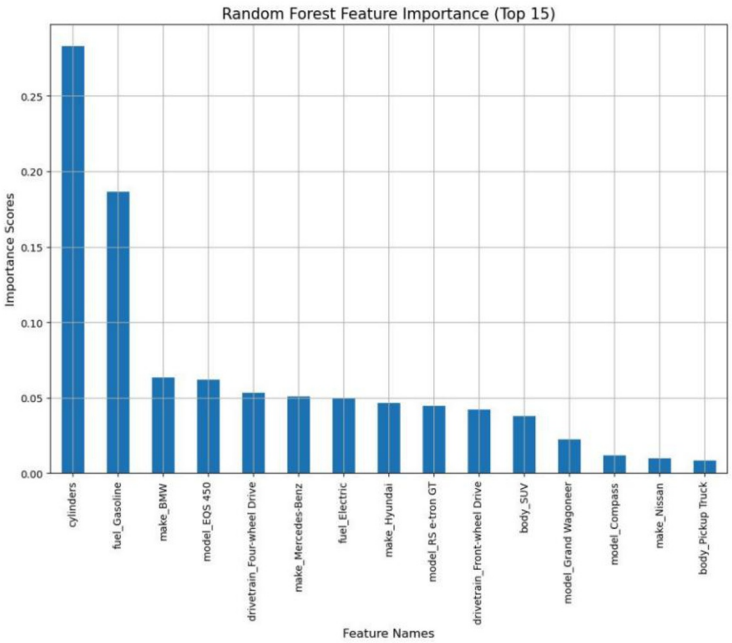

Fig. 3. Random Forest Model - Top 15 Feature Importance Scores (Photo/Picture credit: Original).

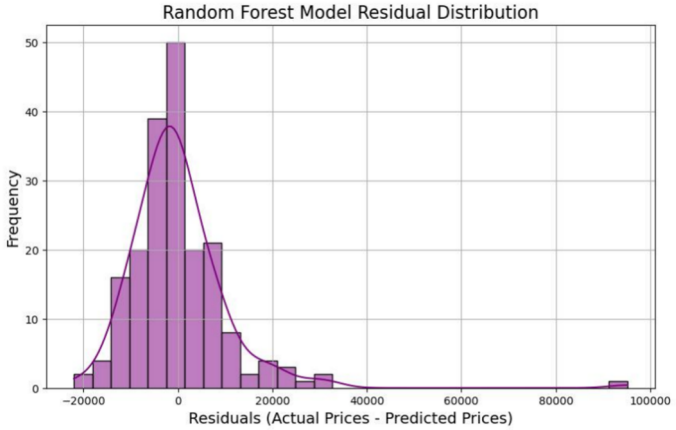

Fig. 4. Residual Distribution of Random Forest Model Predictions

6 Discussion

The difference in prediction performance revealed by the experimental results is essentially due to the difference in the degree of adaptation between different model architectures and complex data features. The linear regression model maps the input

variables and the prediction target with a deterministic linear function, which is destined to have structural deficiencies in capturing nonlinear and high-order feature interactions. Especially when it comes to multi-dimensional and highly heterogeneous vehicle characteristics such as brand, model, transmission type and drive mode, this single model structure is more likely to expose the amplification effect of prediction errors, which is manifested in the obvious deviation of the prediction of the high price range in Fig. 1 from the actual value. The price of high-end vehicles in the data is affected by more complex nonlinear interaction factors, making it difficult for the linear model to be effectively described by simple linear coefficients, resulting in the structural accumulation and significant deviation of prediction errors [5].

In stark contrast to the LR model, the RF recursively partitions the feature space and performs non-parametric local fitting through a large number of decision trees, thus possessing natural nonlinear fitting capabilities and sensitivity to high-order interactive features [9]. The random forest model does not rely on specific assumptions about the data distribution, but rather uses the advantages of model integration and the hierarchical structure of decision trees to flexibly capture and characterize the complex and local structural characteristics within the used car price data. The structural superiority of this model is directly reflected in the prediction scatter plot in Fig. 2, where the scatter points are significantly closer to the diagonal ideal prediction line, especially in the medium and high price ranges, showing the stability and consistency of the prediction [10]. However, although the random forest is superior in terms of prediction accuracy, its inherent black box structure limits the model's ability to intuitively explain and interpret the mechanism of variable action, especially in scenarios where the decision-making level requires high model transparency and interpretability. This limitation may significantly weaken the practical application value and recognition of the model. At the same time, the significant increase in the computational complexity and training time of the random forest, as well as the complexity of the parameter optimization process, has also increased the cost and difficulty of model development and maintenance to a certain extent.

7 Conclusion

Through in-depth comparison and analysis of two machine learning models, LR and RF, in the task of used car price prediction, the study clearly reveals the performance differences and deep-seated reasons of the model structure when processing complex data features. The LR model is limited by its inherent linear mapping assumption. Although the overall prediction accuracy is acceptable, there is a significant error deviation in the prediction of high-end vehicle prices. This is particularly evident in the scatter plot, which is manifested as a significant structural underestimation between the actual and predicted prices. In contrast, the RF model relies on the non-parametric decision tree integration mechanism to successfully capture the complex and nonlinear interactive relationship between multi-dimensional variables, thereby showing a more stable and accurate prediction ability. As shown in the scatter

distribution, the prediction results are closer to the ideal prediction line, especially in the mid-to-high price range. In addition, the feature importance analysis of the random forest model further verifies the key role of the number of engine cylinders and fuel type in vehicle pricing. However, the inherent black box characteristics of the random forest model limit its advantages in model interpretability, and the higher computational complexity and training costs also need to be treated with caution. Therefore, future research should try to integrate advanced models such as gradient boosting trees (XGBoost) or deep learning algorithms, while introducing external macroeconomic factors and multimodal data such as vehicle images and user reviews, to find a more robust and optimized balance between accuracy, model transparency, and generalization performance.

References

1. Sharma, D., Sharma, A.: Car price prediction using machine learning algorithms. International Journal of Computer Applications 175(22), 24–27 (2020)
2. Bharambe, V.A., Gite, S., Shinde, S.: Used car price prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT) 10(4), 91–95 (2021)
3. Pal, N., Arora, P., Sundararaman, D., Kohli, P., Palakurthy, S.S.: How much is my car worth? A methodology for predicting used cars prices using random forest. In: Abraham, A., Cherukuri, A., Melin, P., Gandhi, N. (eds.) FICC 2018. AISC, vol. 887, pp. 413–422. Springer, Cham (2019)
4. Gajera, A., Gondaliya, N.: Old car price prediction with machine learning. International Research Journal of Engineering and Technology (IRJET) 6(4), 3639–3642 (2019)
5. Shaban, A., Kareem, O.S.: Comparative evaluation of linear regression, tree-based regression, and neural network models for structured car price prediction. Polaris Global Journal of Scholarly Research and Trends 4(1), 1–11 (2025)
6. Zhang, Y., Zheng, Y.: A machine learning approach to used car price prediction. Procedia Computer Science 183, 758–765 (2021)
7. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, É.: Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
9. Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub, 2nd edn. (2022)
10. Chen, T., & Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), 785–794 (2016)