Machine Learning-Based Models for Accurate Car Prices Prediction

Chenguang Li

Department of FinTech, University College Dublin, Dublin, Ireland

li.chenguang@ucdconnect.ie

Abstract. The used car market is becoming more and more popular in various countries, but some car trading platforms have inaccuracies in predicting prices. So, it is necessary to select machine learning models with high accuracy to predict used car prices for both buyers and sellers. This study selected three factors that have the greatest impact on used car prices, which are: car name, car use year, and car use mileage. In the selection of dataset, this study chooses more than 2000 data from four brands of BMW, Volkswagen, Acura, and Tesla which are very popular in the market to predict. In the part of comparing the accuracy of machine learning models, this paper uses three machine learning models: linear regression, decision tree regressor and random forest regressor. Based on criteria such as root-mean-square error and R-variance, with K-Fold cross validation method to compare the advantages and disadvantages of every model. Through the analysis, all the test criteria consistently show that random forest regressor has the highest performance and can achieve an R-square value of 0.8562. After getting the suitable model, accurate price prediction will be realized in more car trading platforms to help buyers and sellers to understand the real price of used cars.

Keywords: Car Price Prediction, Machine Learning, Random Forest, Linear Regression, Decision Tree.

1. Introduction

If a middle-class family wants to buy a car, they have two choices in front of them: to buy a new car or to buy a used car. If the family only uses the car as an everyday means of transportation, there is little difference between buying a new and used car; And buying a new car is relatively expensive, and the value of a new car depreciates after purchase, so buying a used car at this time is a reliable option. The surge in new car prices caused by the pandemic has led to an increase in used car sales in Singapore, which are expected to increase at a high compound annual growth rate of 11.1% from 2020 to 2025 [1].

There is no doubt that machine learning models can handle the complex relationship between various features and car prices and secondly can also handle large datasets efficiently and with their advanced algorithms can achieve higher prediction accuracy, thus better bringing better decisions to people. The effectiveness of machine learning algorithms has been demonstrated in many domains including medical diagnosis, telecommunication and sentiment analysis etc. [2-5]. Therefore, it is very necessary to use machine learning model to predict the used car prices, which can provide the estimated value according to different situations of cars, greatly helping buyers and sellers to have a clear understanding of the price of used cars [6].

In previous research, some traditional methods, such as single variable linear regression and Multiple variable regression, were used in the prediction of automobile price. However, these traditional models have predictive limitations, for one thing, they can only deal with linear relationships between independent and dependent variables (i.e. car prices). In reality, many relationships are non-linear. Moreover, if there are outliers or extreme values in the data, it will have a significant impact on the prediction results, thus affecting the accuracy of the prediction.

With the further development of machine learning models in recent years, a variety of relatively advanced algorithms have emerged, such as artificial intelligence algorithms including random tree and neural network etc. For example, in the field of predicting stock market prices, the use of advanced models such as fuzzy logic and neural networks will be more efficient and accurate. Using

traditional methods to evaluate the return on investment of the stock market is a cumbersome, inefficient and expensive process [7]. And in the field of automobile price prediction, by applying three advanced machine learning techniques. Such as Support Vector Machine, Artificial Neural Network and Random Forest, the accuracy is as high as 87.38% [8]. Through the excellent results achieved in these fields, it can be found the powerful prediction ability of advanced machine learning methods.

This paper will use three regression machine learning models: Linear Regression, Decision Tree Regressor, and Random Forest Regressor to predict the used car prices. The used car dataset came from Kaggle, has 2,065 data from four popular brands including BMW, Tesla, Volkswagen and Acura. And the dataset includes three important variables that affect the price of a used car: the name of the car, the distance the car has been driven, and the number of years the car has been in use. Through the results of the three models, to judge which model is best for predicting the used car prices, and which factor has the greatest impact on the price of used cars, this is the research goal of this paper.

2. Methods

2.1. Dataset Preparation

The dataset used in this study for predicting car prices was web scraped from carvana.com/cars in Kaggle [9]. The original dataset has a total of 22,000 pieces of data containing 36 brands and different models of cars of each brand. In this study, the target variable is car prices. And there are three features that affect car prices, which are: the name of the car (The name column includes the make and model of the car), the year and the mileage of the car that has been used.

Since the collected dataset may not be in the most suitable format to be handed over to the machine learning model for manipulation. So, data preprocessing is necessary to ensure the efficiency and accuracy of the machine learning model. In the data Preprocessing, this study selected four popular car brands from the original dataset to form a new dataset: Volkswagen, Acura, BMW, and Tesla. A total of 2064 data are in the new dataset.

In the subsequent steps, some outliers were detected in the year column by introducing the tqdm method in tqdm.notebook [10]. After dropping these outliers, this study obtained the following trends by exploring Mean Car Prices and the name (brand) of the car, the year of use of the car, and the mileage.

The chart can be analysed and found: 1. Among the four car brands, Tesla has the highest mean price, followed by BMW, Acura and Volkswagen shown in Fig. 1. When all else is equal, the mean price of the used car tends to decrease as the mileage of the car increases, as shown in Fig. 2 (a) and (b). And shown in Fig. 2 (c) and (d), all else being equal, the mean price of a car tends to increase with the year of the car.

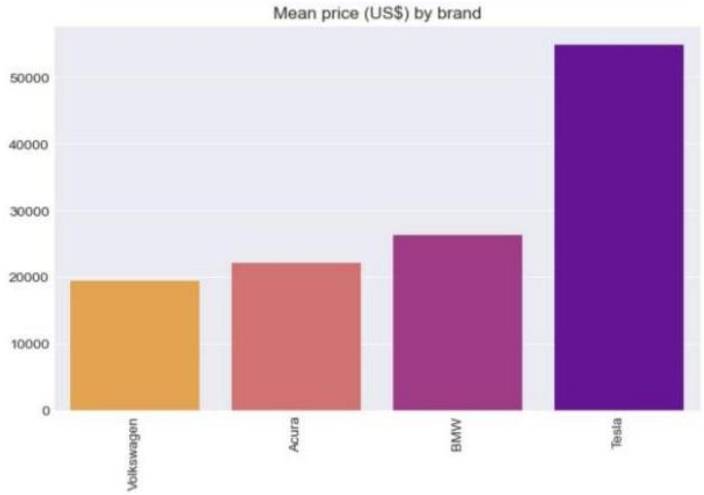

Figure 1. Mean price (US$) by brand (Photo/Picture credit: Original)

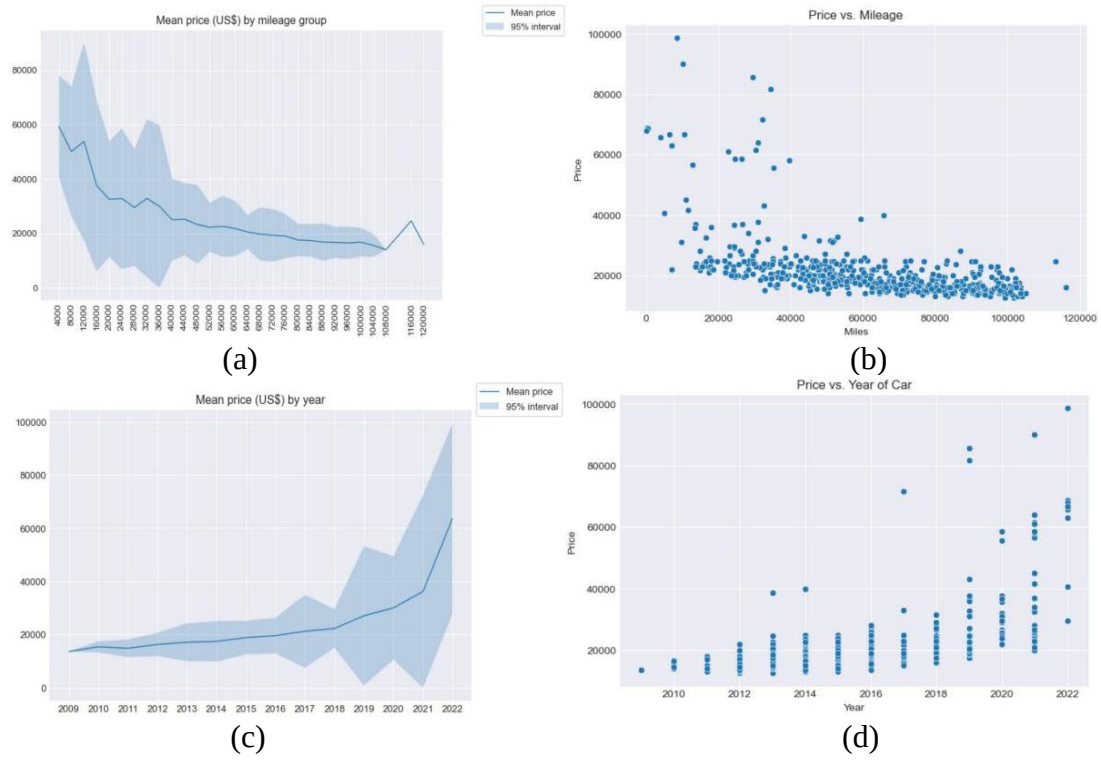

Figure 2. (a) Mean price (US$) by mileage group (b) Price (US$) by mileage (c) Mean price (US$) by year (d) Price (US$) by year of car. (Picture credit: Original)

There are many categorical values in this dataset, e.g., brand, model, etc. To enable machine learning models to better predict these data, this paper represents these types in numerical form. In the brand, model, and so on, there may be multiple different values. This paper uses LabelEncoder [11] in scikit-learn [12] for this task, label encoding compresses the data into integers. And moving to the part where the machine learning model predicts the price of the car, this study begins by dividing the dataset into a training set and a test set, with a Random state of 45, a test set size of 25% of the dataset, and a training set size of 75% of the dataset. Through drawing heat maps using Pearson's correlation coefficient, this study visualizes the correlation between these features shown in Fig. 3 [13].

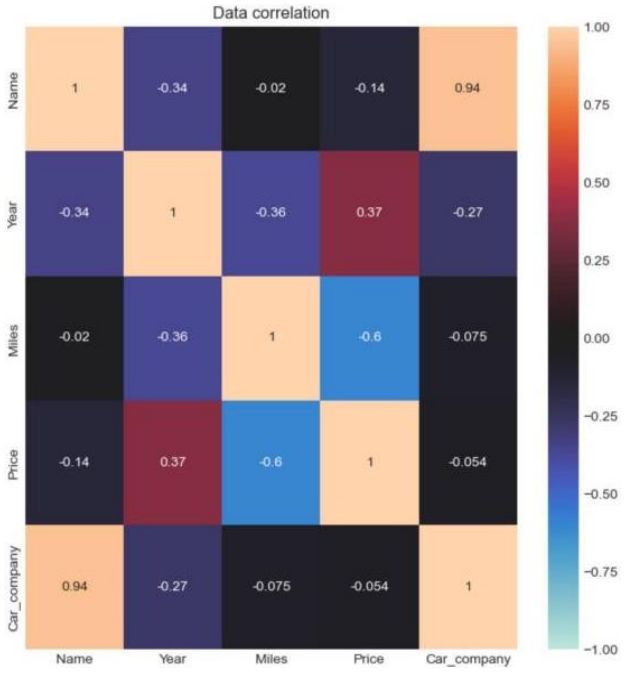

Figure 3. Heat map between every pair of features (Picture credit: Original)

2.2. Machine Learning Models

This paper imports three regression methods from sklearn, linear regression, decision tree regression, and random forest regression, to predict the training and test sets, and compares the performance of each regression model using four criteria: mean absolute error, mean squared error, root mean squared error, and R-Square. When using MSE, MAE, RMSE to compare model performance, this research uses K-Fold Cross-Validation [14] to calculate the different regression models values. Cross-validation makes effective use of limited data, and the evaluation results can be as close as possible to the performance of the model on the test set, which is conducive to the optimization of the model.

2.2.1 Linear Regression

Linear regression [15] is a common method used for regression analysis, and its basic idea is to find the best-fitting straight line (or hyperplane in higher dimensions) that describes the linear relationship between the dependent and independent variables. In the research, the independent variables are the name of the car, the year of use of the car and the mileage. The dependent variable is the price of the car.

2.2.2 Decision Tree Regressor

Decision trees [16] can capture nonlinear relationships and interactions between variables. The decision tree structure is employed in which the leaf nodes represent outcomes. Nodes other than leaf nodes are called decision nodes and they are segmented according to the binary question (yes/no). The objective of a decision tree is to construct a predictive model for a target variable by discerning straightforward decision rules from historical data.

2.2.3 Random Forest Regressor

Random Forest [17] is a composite learning approach that leverages numerous decision trees to generate classification or regression models. It comprises a multitude of decision trees operating collectively. Each tree forecasts the target category's value, and their predictions are amalgamated to yield more precise outcomes.

3. Results and Discussion

This study examines the expressiveness of the three regression methods by using K-Fold cross validation, and the Table 1 reveals that Mean Squared Error, Mean Absolute Error, and Root Mean Squared Error indicate that random forest regressor has the lowest error value. This paper also calculates the R-Square values of different machine learning methods, and they are Linear regression: 0.53020, Decision Tree Regressor: 0.79575, Random Forest Regressor: 0.85621. This also shows that Random Forest Regressor has achieved the best performance among the three methods.

Table 1. Performance of different models on the car dataset

Methods Models	Linear Regression	Decision Tree Regressor	Random Forest Regressor
Mean Squared Error	369268605.09	15436194.71	12732501.36
Mean Absolute Error	4278.39	2391.66	2032.25
Root Mean Squared Error	0.33206	0.29828	0.27132
R-Square	0.53020	0.79575	0.85621

Linear regression is the least expressive because it assumes a linear relationship between the dependent and independent variables of the dataset and fails to capture the nonlinear relationships present in the dataset [18]. For two machine learning methods that can fit nonlinearities. Firstly, since decision trees tend to overfit, this shows that they learn too well with training data and perform poorly with new, unseen data. Second, random forests contain multiple decision trees to make predictions

[19]. This will reduce the problem of overfitting decision trees and the effect of a single extreme value on the data, and generalize better to unseen data. Therefore, random forests will have the best performance.

This study also has some limitations and needs to be improved. 1. Only three machine learning methods are used in this study: decision tree regressor, linear regression and random forest regressor. Although some of these methods have good prediction performance, if more machine learning methods are compared, not only will the prediction results be more diversified and objective, but also better methods will emerge. 2. The automobile price dataset used in this study is derived from one source, which may not ensure that the predicted data are sufficiently diverse and adequate, and may not cover all automobile models and makes in the market. 3. Only three more important variable factors affecting the price of automobiles analyzed in this study were selected, which are: automobile names, age of automobile use, and mileage of automobile use. The other variable factors that affect the price of automobile are engine capacity, Oil for automobiles, hybrid and so on. 4.This study might be more objective if the types of cars were divided into gasoline, diesel, electric, and hybrid vehicles to study.

So, in future research, there will be four things that need to be done: 1. Use more expressive machine learning methods to analyze and compare, such as Support Vector Regression, Neural Networks, ElasticNet Regression, K-Nearest Neighbors (KNN) etc. 2. Search for datasets on automobile prices from multiple sources, ensuring that the datasets cover as much information as possible about the models and makes of automobiles available in the market. 3.The next study will expand the selection of factors affecting automobile prices by considering the effects of variables such as engine capacity, Oil for automobiles, hybrid etc. on the price of automobiles. 4. In the next study, the information in the data set will be divided into different vehicle types and analyzed individually to ensure that the results are more accurate and objective.

4. Conclusion

This study addresses the inaccuracy of car price prediction in used car platforms and proposes using of machine learning methods to increase the prediction accuracy of the platforms so as to provide valuable references for sellers and buyers. This paper uses three machine learning methods: Linear regression, Decision Tree Regressor, and Random Forest Regressor to predict car prices. By analyzing and comparing multiple methods, it is consistently concluded that Random Forest Regressor has the greatest performance among the three methods. The experimental results show that Random Forest Regressor is the most suitable of the three methods to provide car price prediction for used car platforms. In the future, this research will use more car categorization criteria and better machine learning methods to predict car prices in multiple ways in an effort to provide meaningful information.

References

[1] Singapore Used Car Market Outlook Report 2022-2025: Increasing Used Cars Demand Due to the Pandemic Contributes to Increase in Used Cars Sales During the Economic Crisis - ResearchAndMarkets.com. Businesswire, 2022, June 6. Available at: https://www.businesswire.com/news/home/20220606005609/en/.

[2] Qiu Y, Yang Y, Lin Z, Chen P, Luo Y, Huang W. Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. China Communications. 2020 Mar;17 (3): 46-57.

[3] Wu Y, Jin Z, Shi C, Liang P, Zhan T. Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis. arXiv preprint arXiv:2403.08217. 2024 Mar 13.

[4] Wang H, Zhou Y, Perez E, Roemer F. Jointly Learning Selection Matrices for Transmitters, Receivers and Fourier Coefficients in Multichannel Imaging. arXiv preprint arXiv:2402.19023. 2024 Feb 29.

[5] Li M, He J, Jiang G, Wang H. DDN-SLAM: Real-time Dense Dynamic Neural Implicit SLAM with Joint Semantic Encoding. arXiv preprint arXiv:2401.01545. 2024 Jan 3.

[6] Gajera P, Gondaliya A, Kavathiya J. Old car price prediction with machine learning. Int. Res. J. Mod. Eng. Technol. Sci, 2021, 3: 284-290.

[7] Miah MBA, Hossain MZ, Hossain MA, Islam MM. Price prediction of stock market using hybrid model of artificial intelligence. International Journal of Computer Applications, 2015, 111 (3).

[8] Gegic E, Isakovic B, Keco D, Masetic Z, Kevric J. Car price prediction using machine learning techniques. TEM Journal, 2019, 8 (1): 113.

[9] scikit-learn developers. sklearn.preprocessing.LabelEncoder — scikit-learn 0.22.1 documentation. 2019. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html.

[10] Costa-Luis C da. tqdm.notebook - tqdm documentation. Available at: https://tqdm.github.io/docs/notebook/.

[11] Kaggle Carvana - Predict Car Prices. Available at: https://www.kaggle.com/datasets/ravishah1/carvana-predict-car-prices.

[12] Scikit-learn. scikit-learn: machine learning in Python. 2019. Available at: https://scikit-learn.org/stable/.

[13] Benesty J, Chen J, Huang Y, Cohen I. Noise reduction in speech processing. Springer Science & Business Media, 2009, Vol. 2.

[14] Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The'K'in K-fold Cross Validation. ESANN, 2012, Vol. 102, pp. 441-446.

[15] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons, 2021.

[16] Quinlan JR. Induction of decision trees. Machine learning, 1986, 1: 81-106.

[17] Liaw A, Wiener M. Classification and regression by randomForest. R news, 2002, 2 (3): 18-22.

[18] Shanmugasundar G, Vanitha M, Čep R, Kumar V, Kalita K, Ramachandran M. A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. Processes, 2021, 9(11): 2015.

[19] Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest? In Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8, Springer Berlin Heidelberg, 2012, pp. 154-168.