Prediction of the price of used cars based on machine learning algorithms

Yian Zhu

School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China

zyazya0529@shu.edu.cn

Abstract. The used car market is really huge, being caused by a variety of reasons, so it matters to forecast the transaction price of used cars. The study can help people make a better choice when they are willing to purchase a second-hand car. Machine learning is the most effective tool for predicting things at present, and that's why machine learning models are chosen to realize the study. Three machine learning models are involved in the study, SVM, XGBoost and neutral networks. The predictive effect on used car prices is evaluated with accuracy and precision. In the end of the study, the results of the three models are compared with each other, and then a model with a best effect of prediction will be selected. The result of the comparison shows that SVM performs best in the study of used car transaction price forecast. The models used in the study are comparatively shallow, so perhaps some more advanced models will be added to the project in the future.

Keywords: Machine learning, neural network, XGBoost, SVM, price of used cars

1. Introduction

Nowadays, the whole society is stepping into the 5G era. The 5G technology supports many application scenarios, expanding from mobile Internet to mobile Web of things expansion [1]. Meanwhile, the government will support to build up high-speed, mobile and safe next-generation information infrastructure. Driverless technology in the 5G era is becoming more advanced and electric cars are widely available, which lead to the result that a great number of cars flowing in the market have to be disposed of (be scrapped or be sold as used cars). So, this project is chosen, using the knowledge of machine learning to predict the transaction prices in the used car market, so as to grasp the second-hand car market situation more effectively. The prediction can help people who has a will to buy a second-hand car for reference.

The reason for choosing the machine learning model is that it's really hard to make prediction, and the relationship between the variables used for prediction and the predicted variables is difficult to be found [2, 3]. However, some machine learning models can solve this problem in a very simple way [4].

This paper uses three prediction models, namely XGBoost [5], support vector machine (SVM) [6] and neural network [7] to estimate the transaction prices of second-hand cars, and then compares the prediction effect.

More detail information will be introduced deeply in the following parts. The section of Method introduces the data pre-processing and the design ideas of the three models; the section of Result makes a comparison between the performances of the three models.

2. Method

To study the prediction of used car prices, a dataset from the Aliyun is chosen, which contains 150000 training data and 50000 testing data.

The first part is data pre-processing, which is divided into three sections, data cleaning, data dimension reduction and features selection. In the process of data cleaning, 1) the columns independent of the target value are deleted, such as “Sale ID” and “name”, 2) outliers are handled by deleting the training set specific data, 3) missing values are handled by using classified characteristics to fill the mode and using continuous characteristics to fill the average. As for the data dimension reduction, two types of methods are introduced. The former one is PCA, a linear dimension reduction method. The latter one is a nonlinear dimension reduction method called ISOMAP. In the process of features selection, as the non-explanatory ability of neural networks, a lot of features that is unknown will be generated, so only the outliers and the missing values have to be dealt with.

2.1. *Dataset*

In this study, a second-hand car price prediction dataset from Aliyun [8] is leveraged. The Aliyun is a data set website commonly used in China.

The features of the data used are shown in the figure below These features are the Sale ID (the unique code of the used car), name, reg Data (Date of car registration), model (car type code), brand, body type (limousine:0, sedan car:1, truck:2, bus:3, open car:4, coupe:5, BPV:6, agitating lorry:7), fuel type (gasoline:0, diesel:1, LPG:2, natural gas:3, hybrid power:4, others:5, electric power:6), gearbox (manual gear:0, automatic catch:1), power (engine power), kilometre (the number of kilometres the car has travelled), not Repaired Damage (yes:0, no:1), region Code, seller(individual:0, non-individual:1), offer Type(quotation submitted:0, quotation requested:1), create Date(when the car goes on sale), price(the target of prediction), vSeries features(15 anonymous features of v0-14). Of the above features, several features have been desensitized. They are respectively name, model, brand and region Code. This can effectively protect the user's privacy, so as they can use the prediction system at ease, without the worry about privacy

The size of the data set for training is 150000, and the size of the data set for testing is 50000. The data set is sufficiently large to be universal, so that the experiment does not overfit.

Figure 1, 2 and 3 show the learned distribution of the dataset, using three different distribution models. They are respectively Johnson Distribution, Normal Distribution and Log Normal Distribution.

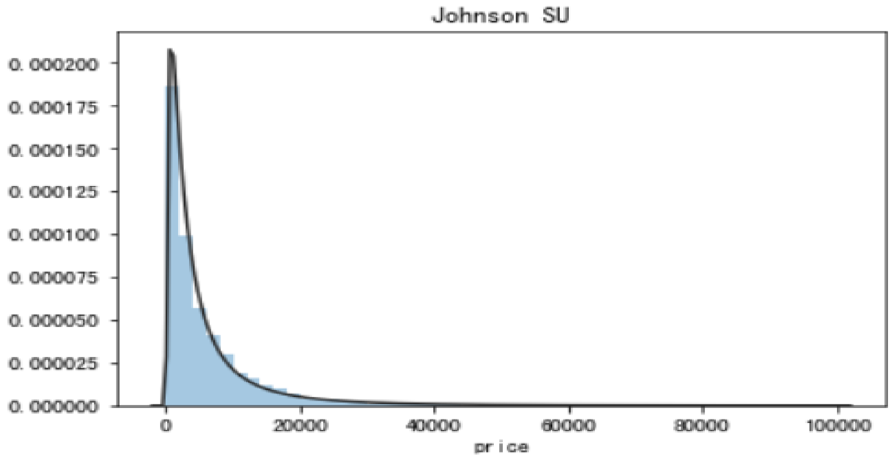

Figure 1. Johnson distribution of the price in dataset.

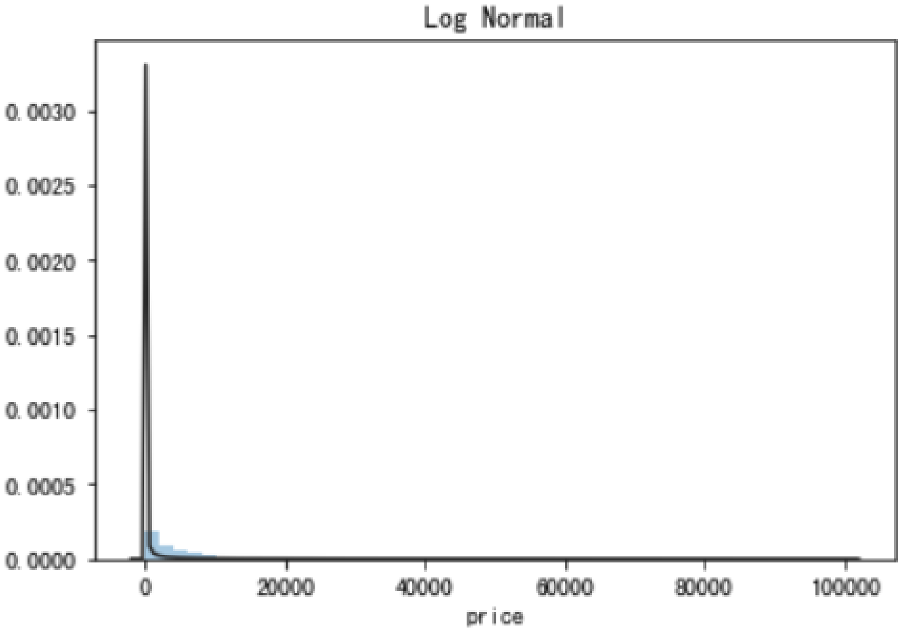

Figure 2. Normal distribution of the price in dataset.

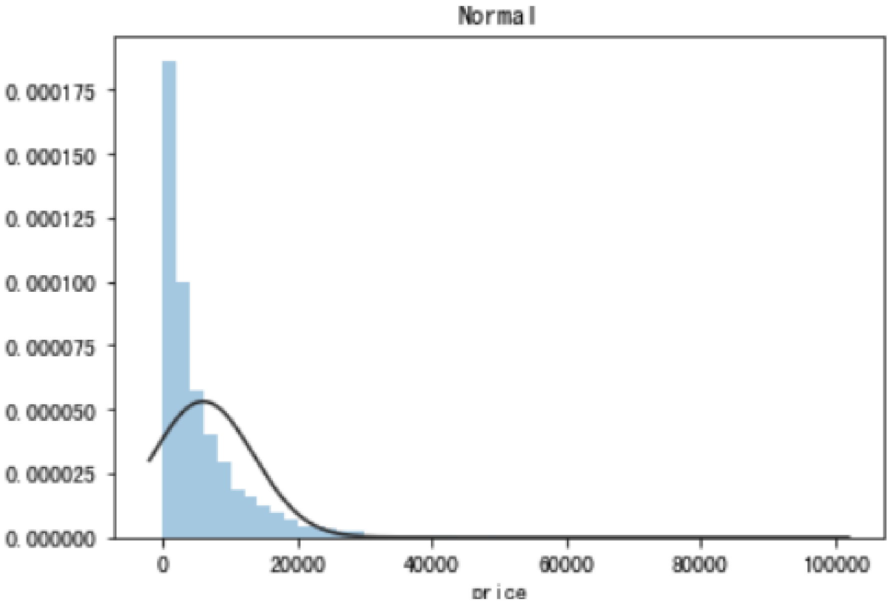

Figure 3. Log normal distribution of the price in dataset.

It can be seen that the data doesn't follow a normal distribution, therefore, the data has to be processed first. The best fit is Johnson Distribution.

2.2. Data preprocessing

The process of data preprocessing includes three parts, data cleaning, data dimension reduction and features selection.

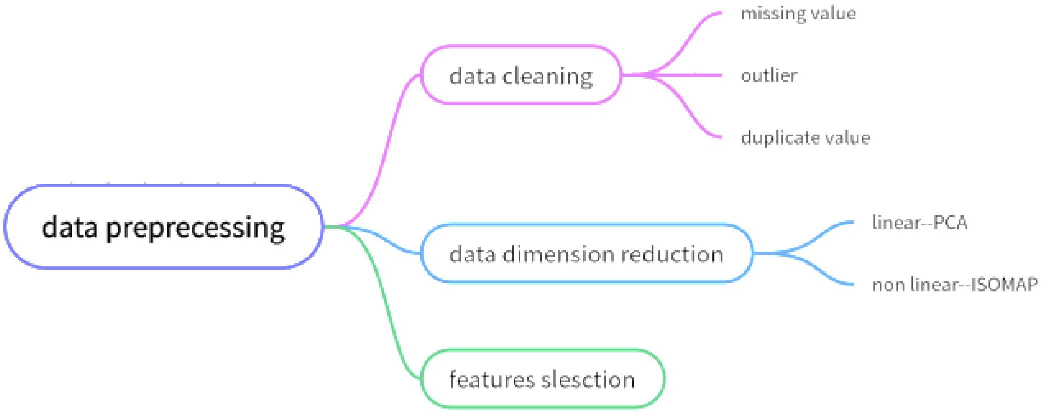

Figure 4. Flow chart of the data pre-processing.

2.2.1. Data cleaning. The part of data cleaning can also be divided into three sections, missing value processing, outlier processing and duplicate value processing. 1) Solve the long tail distribution by changing the target value ‘price’. 2) Delete columns that are not related to the target value, such as ‘SaleID’, ‘name’. The length of ‘name’ may become a new feature. 3) Outlier processing: delete specific data that owned by the training set. For example, the value of ‘seller’=2 will be deleted. 4)Missing value processing: fill the classification features with mode and fill the concatenated features with mean. 5) After some other treatments are added to the data set, the columns that don't change in value can be deleted. 6) Outlier processing: delete the value that does not conform to the requirements of the data set. For example, the value of ‘power’ is required to be in the interval between 0 and 600, so if the value of ‘power’ is above 600, it should be truncated to 600. Values of ‘notRepairedDamage’ that are not numerical should be replaced by np.nan, and then the model will deal with them by itself.

2.2.2. Data dimension reduction. Two different ways of data dimension reduction are used in this study, principal component analysis (PCA) [9] and ISOMAP [10], and they are linear and nonlinear respectively. PCA is a basic linear dimension reduction method. It acquires eigenvectors of the covariance matrix between features through svd. The first k eigenvectors with the largest eigenvalues are selected as the main components. ISOMAP applies to data embedding in floating examples. The fluid is actually a two-dimensional distribution plane. In three dimensions, the distance between points on a fluid cannot be calculated using the traditional Euclidean space. Instead, the geodesic distance represents the actual distance between these two points. Performing like this can better fit the fluid data. But in this study, both these two ways of data dimension reduction don’t perform well. After searching some researches, the reason why PCA and ISOMAP both make a poor performance is that the observed value in this study contain category label. In this kind of situation, supervised dimension reduction methods should be concerned, such as PLS, Linear Discriminant Analysis and Neighbourhood Component Analysis. The two methods mentioned above are both unsupervised dimension reduction methods, and these methods cannot get the information of the category of the observation value, meanwhile supervised dimension reduction methods can directly group data points of the same label together through category information. So, when reducing the dimension of the data by PCA and ISOMAP, some important information is lost in the process of data dimension reduction. The data set that has been processed is not adopted.

2.2.3. Feature selection. The part of feature engineering can be divided into five steps. 1) time and region feature: a bunch of new features like the year, the month, the day can be got from features like ‘regDate’ and ‘createDate’. And then subtraction can be made between them to get some new features like age and days of use. 2) classification features: the continuous features that can be classified are divided into buckets, except ‘kilometer’, which has already been divided into buckets. The classification features ‘brand’, ‘model’, ‘kilometer’, ‘bodyType’ and ‘fuelType’ are used to cross their features with ‘price’, ‘days’ and ‘power’. Through the feature crossing, total, variance, maximum, minimum, mean, mode, kurtosis and so on can be acquired. A great number of new features can be found in this step. LightBGM (be used with XGBoost) can help select the features directly. After selecting one group at a time, these features are retained('model_power_sum','model_power_std', 'model_power_median', 'model_power_max', 'brand_price_max', 'brand_price_median', 'brand_price_sum', 'brand_price_std', 'model_days_sum','model_days_std', 'model_days_median', 'model_days_max', 'model_amount','model_price_max', 'model_price_median','model_price_min', 'model_price_sum', 'model_price_std', 'model_price_mean'). 3) continuous features: Although there are a lot of anonymous features, it doesn’t matter the analysis, and the importance of outputs of these features are very high in lgb, so they are all retained. 4）Supplementary feature engineering： It mainly deals with the very important features of the output. In phase 1, by multiplying the 14 anonymous features, 14 * 14 features can be got. Using sklearn's automatic feature selection, the final selection contains 'new3*3', 'new12*14', 'new2*14','new14*14', ‘new3 * 3’, ' new12 * 14’, ' new2* 14’, ' new14* 14'. In phase 2, 14 * 14 features were obtained by adding 14 anonymous features, and then all these features are put into lgb to train. A great number of new features are generated, so a few redundant features have to be deleted. The method used is to delete high correlation variables (eliminate the correlation > 0.95). In phase 3 and 4, there is no obvious effects, so these two phases can be negligible. 5) feature engineering of neural network: All of the above features are engineered for the tree model. Because of the uninterpretability of Neural Network, it can generate a large number of unknown features, so in the data pre-processing, dealing with outliers and missing values is enough

2.3. Models

In this project, three models are chosen, XGBoost, SVM and neural network.

XGBoost is one of the most famous machine learning methods in boosting model. It has the advantage of efficiency, flexibility and protability. It works under Gradient Boosting framework. It could solve many machine learning problems efficiently and accurately, using parallel tree lift.

Support Vector Machine (SVM) is a conventional supervised linear classifier. It aims at learning a decision maximum-margin hyperplane for separating data. It is a robust classifier, which is one of the common kernel learning methods. By using kernel techniques, it can be adapted for nonlinear scenarios.

Neural network is widely used in deep era. It is composed of many of hierarchically arranged neurons. Empowered by massively parallel processing and good ability of self-organizing and self-learning, it is equipped with excellent learning capacity.

A five-layer neural network is introduced in this project. 1) The training model uses a small batchsize of 512. Although there may be a small deviation in the direction of decline, it has a large return on the convergence speed and can be converged within 2000 generations. 2) Neutral networks don’t have to worry too much about feature engineering. 3) Adjust the regularization coefficient and use regularization to prevent overfitting. 4) Adjust the learning rate, analyze the errors in the training process, and choose the time when the learning rate drops to adjust. 5) Use cross validation and ten-fold cross to reduce overfitting. 6) Adam is chosen as the optimizer for gradient descent. It is an optimizer with good comprehensive capability at present, with advantages such as high computation efficiency and less memory demand

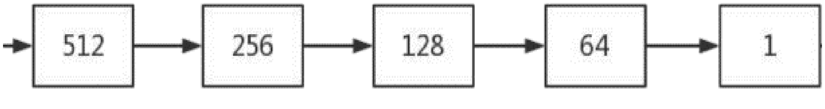

Figure 5. The number of nodes of the fully connection layer in the neural network.

3. Result

In this part, the predictive effect of three models on used car prices are compared according to their R^2 score. The R2 score is calculated as 1 minus the residual sum of squares over the sum of squares of deviations.

As can be seen from Table 1, the R^2 score of XGB is the highest, while which of SVM is the lowest. Xgb algorithm is the best processing result, followed by neural network, and the worst is SVM algorithm.

Table 1. The R^2 score results of the three models.

	R^2 score
XGBoost	0.9823
SVM	0.2258
Neural Network	0.5692

4. Discussion

In the study, it is found that SVM, XGBoost and Neural Network can all fulfil the task of predicting transaction prices of used cars. XGBoost performs best in the prediction task, meanwhile SVM performs worst. The present study showed that SVM and neural network could get good prediction effects, and they are both mature models in the field of prediction. So, after comparing with the present studies, the factors that lead to the poor effect in this study are the wrong way to reduce dimension and the number of the layers is not enough. However, XGBoost is a newly-developing method of prediction. It requires a second derivative, so the target function doesn't have a MAE. That's also why accuracy is chosen as the evaluation criteria instead of MAE. Xgboost's approximation algorithm may give better data results, and the approximation algorithm may be more average. Because greedy algorithm is easy to implement, so choose greedy algorithm. In the study, the accuracy of XGBoost has already reached 90%, which means that it gets a good prediction effect. So some traditional methods can be taken place by some new methods.

5. Conclusion

Because of the development of technology and the changes taken place in people's lifestyle, the market of second-hand cars is being expanded step by step, so the prediction of the price is important for both sellers and purchaser. In this work, three models are leveraged to predict the price of used cars. They are SVM, XGBoost and neural network. The chosen dataset is from Aliyun, where 150000 data are chosen for training and the rest 50000 for testing. The R2 score is used to measure the effectiveness of the three models. Results demonstrate that the XGBoost perform the best, which could achieve the R2 score of about 0.98. The neural network is at the second place, which obtains the R2 score at about 0.57 and the SVM performs the worst. In the future, this work could not only benefit the price prediction of used cars, but can be added to many different fields, such as stock prediction can give people reference to buy which stock.

References

[1] Skubic, B., Bottari, G., Rostami, A., Cavaliere, F., & Öhlén, P. (2015). Rethinking optical transport to pave the way for 5G and the networked society. Journal of lightwave technology, 33(5), 1084-1091.

[2] Hristova, Y. (2019). The second-hand goods market: Trends and challenges. Izvestia journal of the union of scientists-varna. Economic Sciences Series, 8(3), 62-71.

[3] Sun, N., Bai, H., Geng, Y., & Shi, H. (2017). Price evaluation model in second-hand car system based on BP neural network theory. In 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 431-436.

[4] Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. Big Data Research, 2(3), 87-93.

[5] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794.

[6] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.

[7] Bishop, C. M. (1994). Neural networks and their applications. Review of scientific instruments, 65(6), 1803-1832.

[8] Aliyun. (2020). Second hand car price prediction. URL: https://tianchi.aliyun.com/competition/entrance/231784/information

[9] Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.

[10] Balasubramanian, M., & Schwartz, E. L. (2002). The isomap algorithm and topological stability. Science, 295(5552), 7-7.