Used Car Price Prediction by Using XGBoost

Tingyu Qian

The Pennsylvania State University, State College, Pennsylvania, 16802, United State

tjq5053@psu.edu

Abstract. This article demonstrates that by using methods such as Extreme Gradient Boosting (XGBoost), dummy variables, etc., the selling price can be accurately predicted according to the different conditions and variables of each used car. The used car dataset is divided into a training dataset and a test dataset according to the ratio of 83% and 17%. This article uses a total of three data processing methods to find the most accurate prediction method. The first is to remove the outliers of the training dataset and test dataset, and then directly use the xgboost prediction method for prediction. The second is to remove the outliers and remove the variable power that is most closely related to the price of the used car, and then use the xgboost prediction method to make predictions. The third method is to remove outliers and then normalize the training dataset and test dataset, finally using the xgboost prediction method to predict. The experimental results show that normalizing the dataset and then using XGBoost and dummy variables can be used to predict the selling price accurately and efficiently through the different usage conditions of each used car.

Keywords: Car Price; XGBoost; Variable Selection.

1. Introduction

More and more people are starting to buy used cars instead of new cars now because there are many advantages to buying a used car. First, People can save money by buying a used car. According to research, used cars cost nearly 50% less than new cars [1] Second, new cars depreciate quickly, while used cars are likely to increase in value through depreciation [1]. Therefore, many advantages of used cars will lead people to choose to buy used cars instead of new cars. Based on data from "New and used light vehicle sales in the United States from 2010 to 2021", people can find that used car sales are at least twice as high as new car sales every year from 2010 to 2021 [2].

Every used car will have a different price depending on how it is used. Different vehicle types, the mileage of the car, and the external conditions of the vehicle have all become factors for pricing used cars. Since each car will have different usage conditions, and not all factors can be expressed numerically, if people want to accurately predict the price of each used car, people need to use dummy variables and xgboost. In dummy variables, variables can use integers to represent nominal variables to facilitate our predictions. Xgboost, "stands for Extreme Gradient Boosting", is a "scalable distributed gradient-boosted decision tree (GBDT) machine learning library" [3]. There will be many decision trees in xgboost, and the decision trees are created in the sequential form [4]. Each variable will have a corresponding weight, which is then fed into the decision tree that predicts the outcome [4]. Variables that the decision tree predicts incorrectly will increase the weight of that variable, and then feed those variables into a second decision tree [4]. In the end, we will have efficient and accurate predictive models.

Buying a used car has become an essential thing for many families. For the choice of vehicle, it also becomes crucial to consider the situation of different vehicles and the price. The purpose of this article is to perform three different processing methods on the data of the training dataset and the test dataset and then use xgboost to make predictions respectively, and finally, find the most effective and accurate data processing method through the mean absolute error.

2. Methodology

2.1 Source of data

Dataset, from Kaggle [5], is used to predict used car price based on different usage conditions of each used car e.g. mileage, engine, and power. This dataset contains a total of 7253 groups of used cars with different conditions, from different years, from 1996 to 2019. Divide this data set into a training dataset and a test dataset according to the ratio of 83% and 17%. The training dataset, which is a subset to train a model, contains 6019 different used cars with 13 variables that are either numerical or categorical. The test dataset, which is a subset to test the trained model [6], contains 1233 different used cars with 12 variables that same as the training dataset except for the variable price.

2.1.1 Dependent variable

In the test dataset, the dependent variable y is price, and the price will be predicted by the 12 different usage conditions of each used car.

2.1.2 Independent variable

There are many possible factors that influence the price of used cars. For example, the name of the used car, location of the used car, when was this car bought, how many kilometers the car has been driven, how many people have bought the car, standard mileage of the car, engine of the car, the maximum power of the car, and how many seats does the car have. Among these is the transmission of the car (manual or automatic), and the fuel type of the car (CNG, Petrol, or diesel) are categorical variables that need to be converted to dummy variables in order to use to predict the price of used cars).

2.2 Data processing

Whether it is in the training set or the test set, there are very many prices for the first purchase of vehicles that are missing, so for these missing values, the best way to deal with it is to remove the variable of the price of the first purchase of the vehicle, because if people do not do any processing on missing values, it will have a great impact on the predicted price of our used car, resulting in the inaccurate used car price. Among the 13 variables, there are two categorical variables, which means that we need to use dummy variables to convert these two nominal variables into integers. This data set has two parts, one is the training dataset, and the other is the test dataset. The training dataset has 6018 different used cars, and the test dataset has 1233 different used cars.

2.3 Machine learning models

Using xgboost to predict the price of used cars is the best choice. xgboost is Extreme Gradient Boosting. Before understanding how xgboost works, people must first know what a decision tree is. A decision tree is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category [4]. A decision tree is complete when the split no longer adds value to the prediction [4]. Then people use boosting, which is a method of combining predictions from multiple models into a single model [7]. Boosting works by taking each predictor in order and modeling it based on the errors of its predecessors, with better performing predictors being given more weight, using the original data to fit the first model, and using the first model's residuals to fit a second model, and so on [7]. In the end, we will have a very accurate predictive model.

2.4 Evaluation metric

Upload the predicted price of the used car to kaggle, and kaggle will judge the accuracy rate based on the calculation error of the correct used car price.

3. Results and discussion

3.1 Data visualization

Table 1. Dataset variables description

Variable	Category	Description
Name	Categorical	Brand of the car
Location	Categorical	Location in which the car is being sold
Year	Categorical	Year of the car
Kilometers_Driven	Numeric	The total kilometers of the car has been driven by previous owner (s)
Fuel_Type	Categorical	Type of fuel used by car
Transmission	Categorical	Type of transmission used by car
Owner_Type	Categorical	Number of owner (s) who driven this car
Mileage	Numeric	Standard mileage offered by the car company
Engine	Numeric	Displacement volume of the engine
Power	Numeric	Maximum power of the engine
Seats	Numeric	Number of seats in the car
New_Price	Numeric	Price of a new car of the same brand
Price	Numeric	Price of a used car that will be predicted

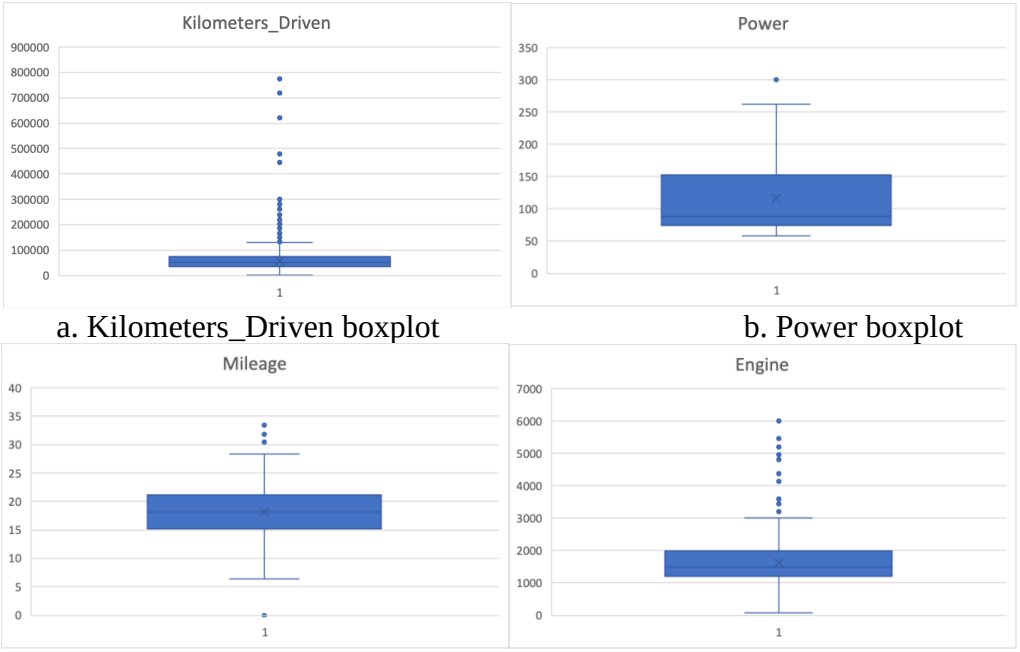
According to Table 1 above, people can find that there are 12 variables in this set of data, plus one variable that will be the predicted price of used cars. Among these 12 variables, 6 variables are categorical variables, and 6 variables are numerical variables.

Table 2. Dataset variables description

	Kilometers_Driven	Mileage	Engine	Power	Seats
Mean	58738.3803	18.13496094	1621.27645	113.2530497	5.27873515
Standard Deviation	91268.84321	4.58228913	601.3552326	53.87495737	0.80883955
Minimum	171	0	72	34.2	0
25% percentile	34000	15.17	1198	75	5
50% percentile	53000	18.15	1493	97.7	5
75% percentile	73000	21.1	1984	138.1	5
Maximum	6500000	33.54	5998	560	10

According to Table 2 above, we can find that the difference between the minimum value and the maximum value of variable kilometers_Driven is very large, which will have a great impact on the mean of variable kilometers_Driven because the mean is easily affected by outliers. There is also a relatively large gap between the minimum and maximum values of the variable engine and the variable power, so for data processing, people need to remove outliers first and then build a prediction model, otherwise, the prediction model will not accurately predict the used car price.

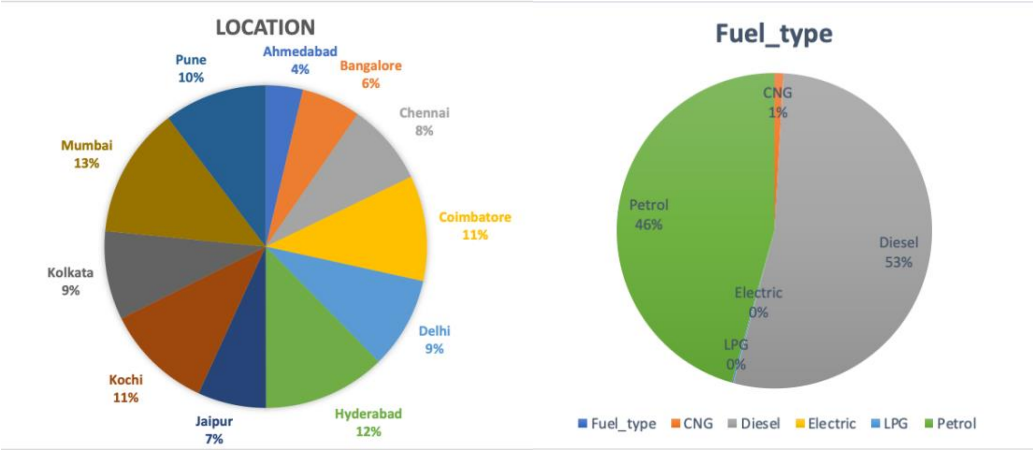
After analyzing the box plots of the four numerical variables from Fig. 1, people will find that each of the four numerical variables has at least one outlier. Because the maximum value of variable Kilometers_Driven is an outlier, the maximum value is removed to ensure the perfect presentation of the box plot for variable Kilometers_Driven. People can also find from the distribution of the four box plots above in Fig. 1 that the data of variable Kilometers_Driven is more widely distributed than the remaining 3 variables. The difference between the minimum value and the maximum value of the variable Kilometers_Driven is also very large, indicating that the range of the variable Kilometers_Driven is also very large. According to the calculation of the mean, the first quantile, the third quantile, the maximum value, etc. of the four numerical variables, people can have a rough understanding of the distribution of these numerical variables.



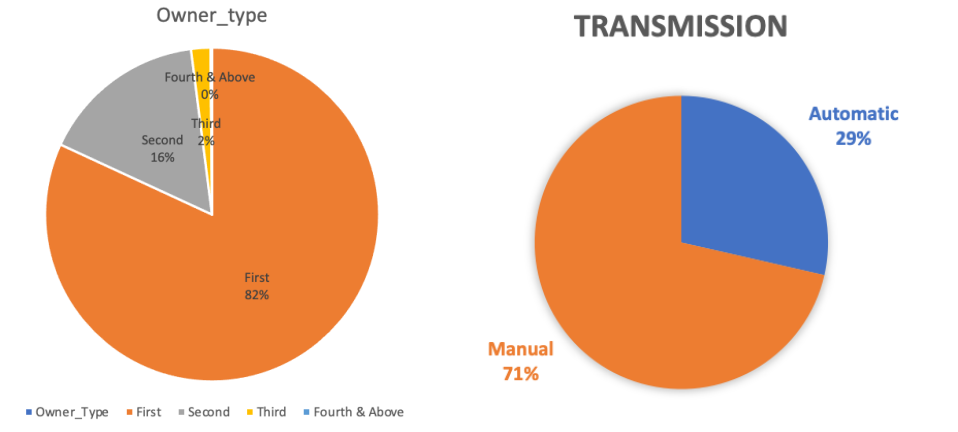
a. Kilometers_Driven boxplot b. Power boxplot

c. Mileage boxplot d. Engine boxplot

Fig. 1 Numeric variables boxplots (Photo credit: Original)



a. Location pie chart b. Fuel_type pie chart



c. Owner_type pie chart d.Transmission pie chart

Fig. 2 Categorical variables pie charts (Photo credit: Original)

After performing a pie charts analysis on the 4 categorical variables above in Fig. 2, people can clearly see the distribution of different categories in each categorical variable. By observing the above

4 categorical variables, people can find that the distribution of each category of variable location is very uniform, but the distribution of each category of variable fuel_type and variable owner_type is very uneven, and some categories almost account for 70 percent of the entire variable, but some categories can only account for 1 percentage of the entire variable.

3.2 Variable Selection

Table 3. Correlation between numeric variables

	Year	Kilometers_ Driven	Mileage	Engine	Power	Seats	Price
Year	1	-0.169369	0.285623	-0.068045	0.014531	0.007833	0.299475
Kilometers_ Driven	-0.169369	1	-0.060608	0.09303	0.03349	0.083072	-0.008249
Mileage	0.285623	-0.060608	1	-0.637258	-0.538844	-0.331576	-0.341652
Engine	-0.068045	0.09303	-0.637258	1	0.866301	0.401116	0.658047
Power	0.014531	0.03349	-0.538844	0.866301	1	0.10146	0.772843
Seats	0.007833	0.083072	-0.331576	0.401116	0.10146	1	0.055547
Price	0.299475	-0.008249	-0.341652	0.658047	0.7 72843	0.055547	1

Fig. 3 Coefficient matrix between numeric variables (Photo credit: Original)

Table 4. Correlation between numeric variables (except variable power)

	Year	Kilometers_Driven	Mileage	Engine	Seats	Price
Year	1	-0.172663	0.314417	-0.052538	0.012338	0.305800
Kilometers_Driven	-0.17266	1	-0.064991	0.090832	0.083101	-0.011263
Mileage	0.31442	-0.064991	1	-0.598638	-0.308226	-0.313877
Engine	-0.05254	0.090832	-0.598638	1	0.393435	0.659117
Seats	0.012338	0.083101	-0.308226	0.393435	1	0.052262
Price	0.305800	-0.011263	-0.313877	0.659117	0.052262	1

By looking at the coefficient matrix between different variables in Fig. 3, people can clearly know whether the relationship between different variables is positive or negative, which is very helpful for people to analyze between variables. When the correlation coefficient between two variables is closer to 1 or -1, it means that there is a very strong relationship between the two variables. When the correlation coefficient between two variables is close to 0, it means that the relationship between the two variables is very weak. When one variable becomes larger or smaller and the other variable also becomes larger or smaller, there is a positive relationship between the two variables [8]. When one variable is getting larger, but the other variable is getting smaller, it means that there is a negative relationship between the two variables [8]. After finishing the coefficient matrix chart with 7 numerical variables in figure 3, people can find that the correlation coefficient of variable power and

price is 0.77, compared with the remaining 5 variables and price correlation coefficient, 0.77 is the highest, so we removed variable power. After removing the power variable, a coefficient table was made again in Table 4.

3.3 Model training and evaluation

Table 5. Correlation between numeric variables after Normalization

	Year	Kilometers_Driven	Mileage	Engine	Power	Seats	Price
Year	1.000000	-0.173048	0.321565	-0.052197	0.014525	0.012333	0.305327
Kilometers_Driven	-0.173048	1.000000	-0.065253	0.091068	0.033503	0.083113	-0.011493
Mileage	0.321565	-0.065253	1.000000	-0.597699	-0.537729	-0.308226	-0.306593
Engine	-0.052197	0.091068	-0.597699	1.000000	0.866185	0.393337	0.658354
Power	0.014525	0.033503	-0.537729	0.866185	1.000000	0.101562	0.772566
Seats	0.012333	0.083113	-0.308226	0.393337	0.101562	1.000000	0.052225
Price	0.305327	-0.011493	-0.306593	0.658354	0.772566	0.052225	1.000000

Since the distribution of the data is very wide, the difference between the minimum value and the maximum value is very large, so normalization can make the difference between the data between 0 and 1, and it is also one of the methods that can make the prediction model more accurate. The formula for normalization is $zi = (xi - \min(x)) / (\max(x) - \min(x))$, where x is the ith value in the data and z is the normalized value of the ith value in the data [9]. Normalization is very important because normalization can put all data into one interval. If normalization is not used, variables measured at different scales will contribute differently to the analysis, which will eventually lead to bias [10]. After normalization, a coefficient table was made in Table 5. After observing the coefficients between the variables after data normalization (Table 5) and the coefficients between the original data variables (Table 3), we can find that the coefficients between the variables have changed slightly.

The mean absolute error of the original prediction is 1.464096, the mean absolute error of the prediction that deletes the variable power is 1.496584, and the mean absolute error of the prediction that uses the Normalization is 0.008922. Therefore, according to the mean absolute error of the different three methods, people can find that normalization is a really good way to make the prediction, and normalization can make the prediction model more accurate.

3.4 Limitation

The training dataset contains a total of 6019 sets of different used car data. Since under the new price variable (the price of a new car of the same brand), many new price data are missing in the training dataset, in order not to affect the accuracy of the prediction model, the best way is to delete the variable new price. However, when observing the data, people can also find that there are a total of 143 missing data under the power variable (maximum power of the engine). Since there are a total of 6019 sets of data, the best way is to delete 143 used car data corresponding to the 143 missing data under the power variable. After deleting 143 used car data, the accuracy of the prediction model will be increased. There are a total of 1234 sets of different used car data in the test dataset. Also because of a lot of absence of the variable new price, the best way is to delete the new price variable to increase the accuracy of the prediction. There are 32 missing data under the power variable, so deleting all 32 used car data corresponding to these 32 missing data under the power variable. If both the variables of new price and power have no missing values, the overall forecast can be made more accurate.

4. Conclusion

This paper studies 1 dataset but divides this dataset into a training dataset and a test dataset according to the ratio of 83% and 17%. The training dataset contains 13 variables and 6019 groups of used cars in different situations. The test dataset contains 12 variables and 1232 groups of used cars in different situations. The purpose of this article is to use the training dataset to make a suitable

prediction model to predict the price of used cars in 1232 different situations in the test dataset. Because there are missing values in the training set and test set, different treatments are made according to different situations of missing values, in order to make an accurate used car price prediction model. This paper also makes different visualizations for numerical variables and categorical variables. For numerical variables, people can clearly see the data distribution from the mean, first quantile, etc., and box plots. For categorical variables, people can clearly see the distribution of each category from the pie chart. This prediction is made because more and more people choose to buy used cars instead of new one because the price of used cars are cheaper, and the insurance premium is not so high. Therefore, making predictions for used cars in different situations can let people know the price of used cars. Having a good understanding makes it easier for people to decide whether to buy a used car.

References

[1] Rawhide Youth Services. "9 Advantages of Buying a Used Car Instead of New." Rawhide Youth Services, 21 Sept. 2015, https://www.rawhide.org/blog/car-tips/9-advantages-of-buying-a-used-car-instead-of-new/?gclid=Cj0KCQiAveebBhD_ARIsAFaAvrHL9hKasLoabnytqMICuCJcG_s43bBTVZvOEaaOkKy_vOzs2aMIoQ8aAp_sEALw_wcB.

[2] Carlier, Mathilde. "New and Used Light Vehicle Sales in the United States from 2010 to 2021." Statista, 22 July 2022, https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/.

[3] "What Is XGBoost?" NVIDIA Data Science Glossary, https://www.nvidia.com/en-us/glossary/data-science/xgboost/.

[4] "XGBoost." GeeksforGeeks, 11 July 2022, https://www.geeksforgeeks.org/xgboost/.

[5] Kasliwal, Avi. "Used Cars Price Prediction." Kaggle, 25 June 2019, https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction.

[6] "Training and Test Sets: Splitting Data Machine Learning | Google Developers." Google, Google, 18 July 2022, https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data.

[7] Mello, Arthur. "XGBoost: Theory and Practice - Towardsdatascience.com." Towards Data Science, 17 Aug. 2020, https://towardsdatascience.com/xgboost-theory-and-practice-fb8912930ad6.

[8] "Understanding Correlations and Correlation Matrix." Muthukrishnan, 7 May 2021, https://muthu.co/understanding-correlations-and-correlation-matrix/.

[9] Zach, zach. "How to Normalize Data between 0 and 1." Statology, 26 Apr. 2021, https://www.statology.org/normalize-data-between-0-and-1/.

[10] "How, When, and Why Should You Normalize / Standardize / Rescale Your Data?" Towards AI, 29 May 2020, https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff.