Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/105/2025.22745

The Research on Prediction of Used Car Prices Using Regression Model and LightGBM Model

David QIAN

Shenzhen Hong Kong Pui Kiu College Longhua Xin Yi School, Shenzhen, China
zhihaoqian924@gmail.com

Abstract: As the car industry has been developed rapidly, the used car market has also shown its potential due to the affordability and ability to make prediction on prices. This study aims to predict the prices of used car by using two methods: Linear regression model and LightGBM model and make comparison on the performance of two models. The dataset used are found in Kaggle which contain 5847 groups of data with 11 different variables in total affecting the price. In this paper, only 10 variables are chosen to process in the two models and evaluate the results. It has been found that LightGBM model are better than linear regression model with a higher efficiency, suitability and accuracy, with an R^2 value of 0.962 for the training set and 0.930 for the test set, compared to Linear Regression's R^2value of 0.700. Additionally, LightGBM demonstrates lower prediction errors (MAE: 1.103, MSE: 4.858, RMSE: 2.204) and better handling of large-scale data. To conclude, the LightGBM model has higher accuracy and is more suitable to predict the used car prices with higher efficiency especially when processing complex, large-scale data compared to linear regression model.

Keywords: Prediction, used car prices, machine learning.

1. Introduction

Under a society with high-speed changes and developments, the industrial market reform and elimination is rapid, it was believed that car industry has shown great potential especially for the markets of used cars, not only because of its low prices, various vehicle models and good car condition but also its availability of making prediction on the used car prices.

As more and more people are willing to own a car, the industry has caught fire and the prices of cars started to fluctuate include the used cars prices. Pattabiraman et al. expressed that the industry of cars has a booming development over the past ten years while 70 million of the cars have been manufactured in 2016, this has led to the growth of used cars markets [1]. However, Kriswantara et al. have stated that the global economy has been greatly affected during the period of pandemic and weakened consuming levels, which led to a good, alternative choice that was buying used cars for those people who wanted to own one [2].

Thus, the used cars market has started to shown its great potential. Samuruddhi et al. mentioned that commercial opportunity has been created because of the secondhand cars market, with affordable cost of buying the used cars. In addition, profit may be gained to buyers as they could sell it again after years of utilizing [3]. Unfortunately, although the used cars market seems to have many advantages, there are various factors affecting the prices which makes individuals or companies hard

DOI: 10.54254/2753-8818/105/2025.22745

to make relatively good decisions when buying so it is necessary to make estimation on the used car prices. Muti et al. have illustrated the purpose of making prediction and reason behind it. To ensure the most suitable prices without given a lower price than its exact value, the overestimated value of cars would result in a reduction of chance of sold it out and rise of selling time [4]. Pudaruth has proposed that the prices of used cars rely on numerous aspects however sometimes it was unable to seek for information about them, therefore he said that customers could only make decisions based on recent news about these factors [5].

As a result, it is of great importance to collect data by using different models or machine learning methods. AlShared indicated that demands of artificial system have existed in both buyers and sellers, however, it was troublesome to gather information because of the complex data, therefore, AIShared explained the necessity of managing the data and make any transformation in advance in order to fit directly to the model [6]. After the processing of data, models are available to be applied but choosing the right and appropriate machines is essential or else the process could carry forward slowly. Ahtesham expressed that obstacles have existed when determining what the cars worthen and compared to the prices sellers released since various factors like mileage, model, ages could affect the prices [7]. Apart from that, factors like engine power, fuel type and number of seats could all impact the prices. Chen et al. stated out that a precise estimation of used car prices can act as a catalyst to provide a healthy development of used car markets while it benefited both customers and sellers [8].

Different methods can be used for making prediction such as linear regression, random forest, neutral network and Xtreme boost algorithms. Chandak has stated that python is widely and commonly applied in the implementation of different machine learning models because of its large amounts of inbuilt methods in the form of packaged libraries [9]. Collar has demonstrated that the regression analysis is the foundational concept of Machine Learning as it could help with the processing of building relationship between different variables and output the most accurately-estimated combination of the variables [10].

In this paper, regression analysis and LightGBM model were chosen to make prediction of used cars prices while the dataset used to make analysis were found in Kaggle. This paper has also compared the efficiency, accuracy, suitability of these two models and it has been found that LightGBM was obviously better than linear regression model. Therefore, it is suggested that LightGBM model was more suitable.

2. Methods

2.1. Data source

The data utilized in this study was sourced from Kaggle which was updated by Sujay a year ago. The dataset's availability rating is 10.0 and it has collected the selling prices of used cars in different cities among India with various factors that affected the prices such as brand, location, year, kilometers driven, fuel type, transmission, owner type, mileage, engine, power and seats. In total, the datasets contained 5847 groups of figures.

2.2. Variable selection and description

The analysis in this paper has used all the 5847 sets of figures and 10 independent variables (location, year, kilometers driven, fuel type, transmission, owner type, mileage, engine, power, seats) out of 12 independent variables in total (Table 1). The dependent variable is the price. In order to use all the independent variables, numbers have given to some of the variables since the original variable names are not numeric variables, it is difficult to operate these into model analysis. Therefore, the paper has tried to create codes such as (-1,0,1,2) to replace the original words in the variables. For the variable

DOI: 10.54254/2753-8818/105/2025.22745

fuel type, '1' represents diesel oil while '-1' represents petrol oil. For the variable 'transmission','1' represents automatic while '-1' represents manual. For variable 'owner type','1' represents first, '0' represents second while '-1' represents third an others. Besides, number has also been coded to replace names of different cities in India for variable 'location'. In this paper, histogram of independent variables, the graph of correlation between X factors, table of linear regression analysis results, feature weight graph and table of LightGBM model results have been displayed.

Table 1: List of variables

variables	logogram	explanation
Location	X1	The place that the car has been sold
Year	X2	The year that the car has been manufactured
Kilometers driven(km)	X3	Travel distance of cars
Fuel type	X4	Two types: diesel oil and petrol oil
Transmission	X5	Two types: automatic and manual
Owner type	X6	The number of owners
Mileage(kmpl)	X7	The oil consumption rate per kilometers
Engine(cc)	X8	The capacity of engine
Power(bhp)	X9	The horsepower of cars
Seats	X10	The number of seats in cars
Price(lakhs)	Y	The price of used cars

2.3. Method introduction

This paper has applied both linear regression model and LightGBM model to analysis the relationship between X variables and Y factor. Linear regression is a model that apply statistic method to describe the linear relationship between dependent variable and independent variables by fitting a straight line or hyper plane. While this article has also applied machine learning method called LightGBM, it is a model that base on Decision Tree to process large-scale data and make prediction precisely with high efficiency. In this paper, liner regression model and LightGBM model has been applied respectively to analysis the history dataset and make prediction of the used car prices and it has been found that LightGBM model has more accurate prediction about the prices than the linear regression model and shown to more suitable in processing complex data.

3. Results and discussion

3.1. Descriptive analysis

This article has found the frequency and orthostates of data points of four of the variables among 10 which are year, mileage, power and engine. Below are the histograms of for variables. The histogram of year has reflected the distribution of year of the used cars and from Figure 1, it can be found that the distribution of the years of cars is relatively uniform, but cars that manufactured in recent years have accounted for a relatively higher percentage. Besides, the earliest year of manufacturing is 1998 while the latest year is 2019.

DOI: 10.54254/2753-8818/105/2025.22745

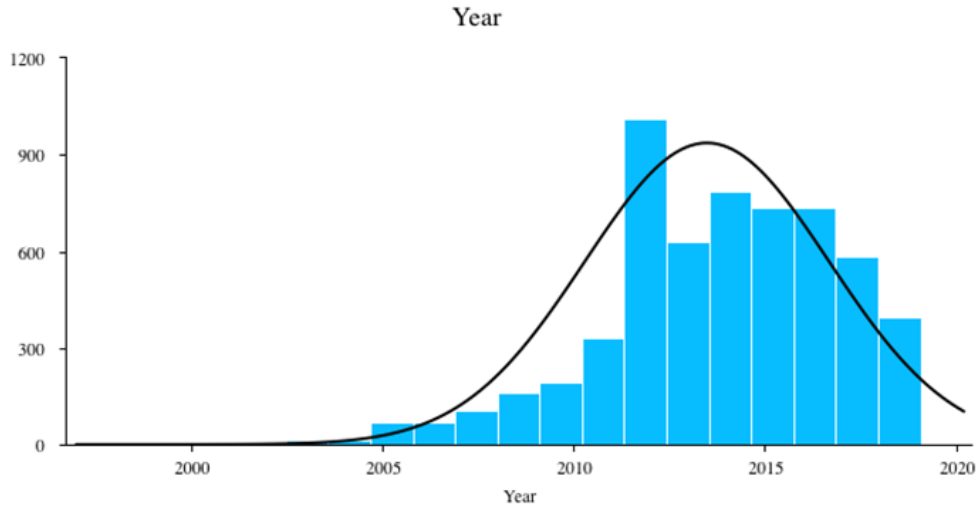

Figure 1: Histogram of year

The histogram of mileage is about the distribution of mileage of used cars. From Figure 2, it can be seen that the distribution of mileage is relatively concentrated and most of the cars have a mileage with middle level while cars with extremely high or low mileage account for only little percentage. The highest mileage found is 28.4 kmpl while the lowest mileage is 10.00 kmpl.

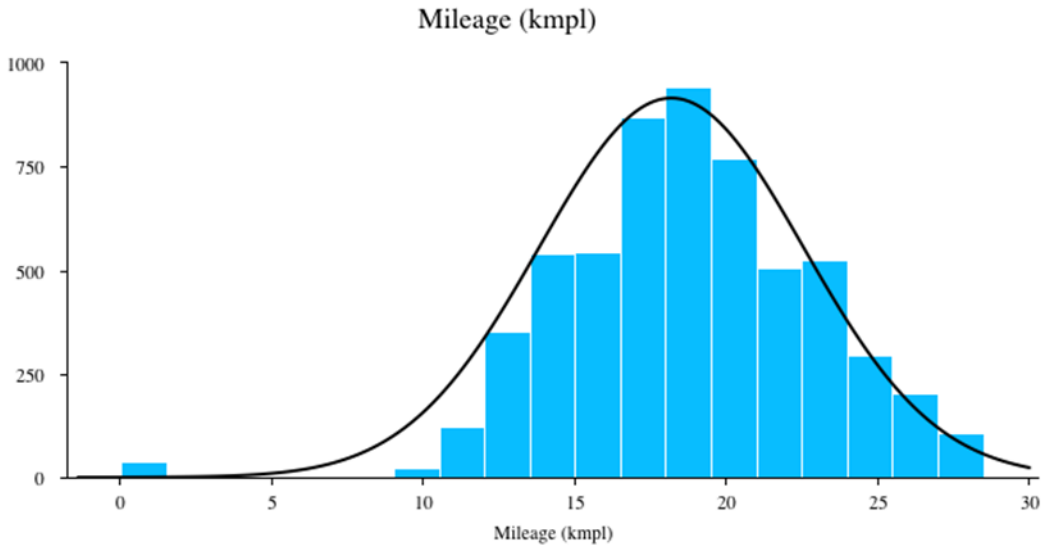

Figure 2: Histogram of mileage

The histogram of power is about the power of used cars and it has illustrated that cars with lower power have occupied most of the percentage and most of the cars have power with a middle level (Figure 3). Besides, cars with high power are fewer and the distribution of power of cars are relatively concentrated. The highest power of cars is 335.3bhp and the lowest power is 64bhp.

DOI: 10.54254/2753-8818/105/2025.22745

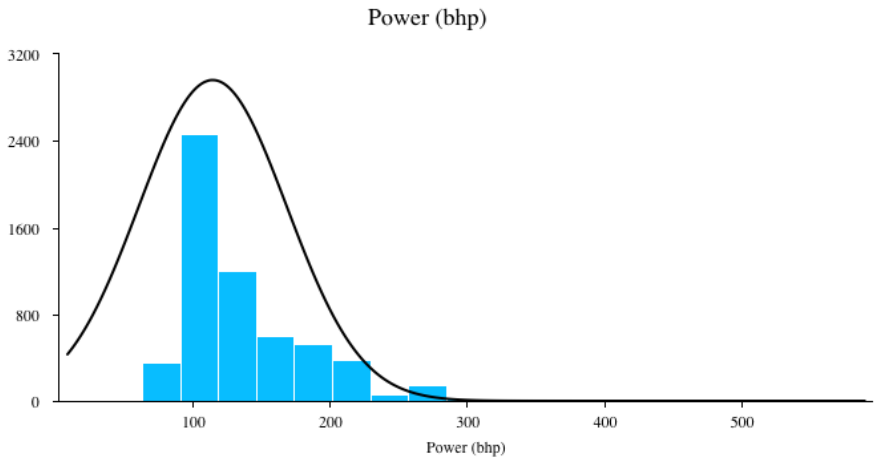

Figure 3: Histogram of power

The histogram of engine described the distribution of engine of cars. From Figure 4, it has been found that the engine of cars is widely distributed with a maximum value of 5998 cc and a minimum value of 624 cc. Therefore, it can be concluded that the difference of engine of cars is large as there are cars with a wide range of values of engine.

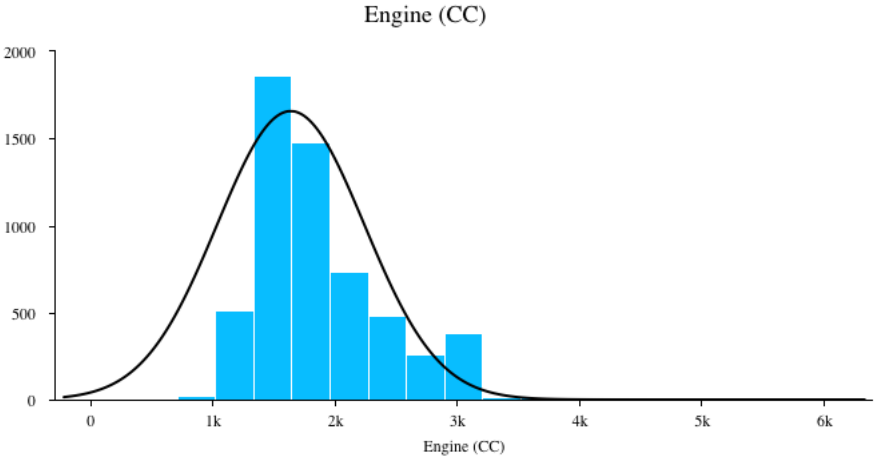

Figure 4: Histogram of engine

3.2. Correlation analysis results

From figure 5, it can be found that power has the highest relativity which means it has the greatest influence to price, while the second highest is the engine which also plays a dominant role on affecting the price. Apart from that, transmission, year, and fuel type all contributed to make influence to the used car prices. Besides, it can be concluded that mileage has an inverse proportion to price while location, seats, owner types and mileage did not have much impact on price.

DOI: 10.54254/2753-8818/105/2025.22745

Figure 5: The relativity of each variable to price

According to figure 6, there is a significant negative correlation between kilometers driven and location as well as owner type while a significant positive correlation between kilometers driven and fuel type, engine, seats and power. Moreover, it is found that the mileage has a dominant negative correlation between kilometers driven while year also shown the similar relativity. However, the correlation value will not show significance between kilometers driven and transmission.

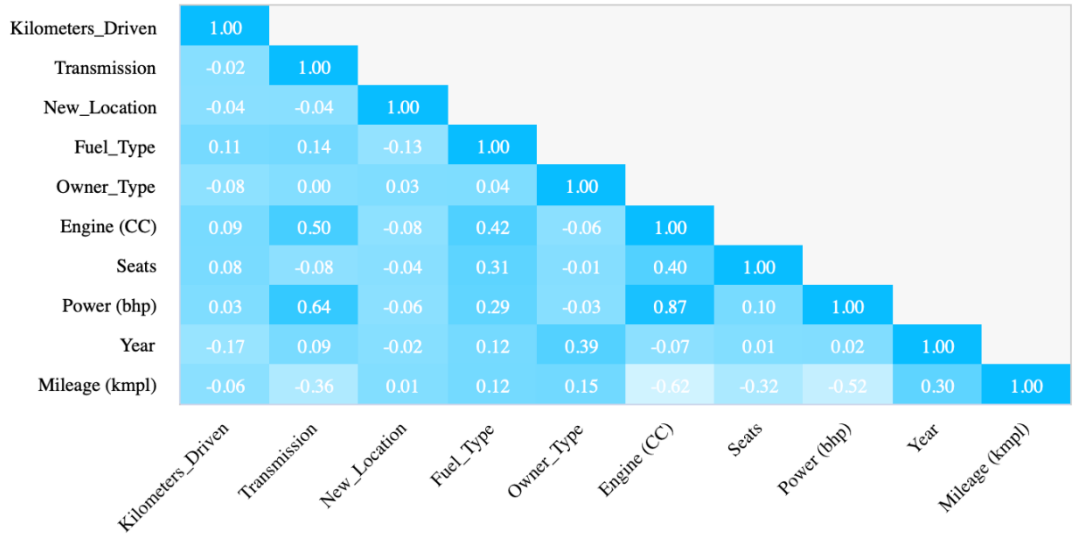

Figure 6: The correlation between X factors

3.3. Linear regression analysis

From the result of linear regression analysis of Table 2, it can be seen that the R^2 is 0.700 which means that the model can explain 70% of the price variation while the adjusted value of R2 is also 0.700, therefore, it can be concluded that the fit of the model is great. Apart from that, it can be found that there are several factors which has significant impact on used car prices: Power (bhp), Engine (CC), Transmission, Year, Fuel Type, Mileage (kmpl), Seats and Location while the kilometers

DOI: 10.54254/2753-8818/105/2025.22745

driven and owner type have a relatively unsignificant impact to the price. Moreover, the VIF value of engine and power are high which may lead to a certain multiple collinearities while other factors did not have such problems.

Table 2: The result of linear regression analysis

variable	Unstandardized Coefficients		Standardized Coefficients	t	p	Collinearity Diagnostics	
	B	Std. Error	Beta			VIF	Tolerance
Constant	-2095.121	60.523	-	-34.617	0.000**	-	-
Location	-0.211	0.028	-0.055	-7.631	0.000**	1.020	0.980
Kilometers Driven	0.000	0.000	0.009	1.160	0.246	1.053	0.950
Year	1.043	0.030	0.296	34.515	0.000**	1.429	0.700
Fuel Type	1.290	0.113	0.114	11.431	0.000**	1.937	0.516
Transmission	1.258	0.120	0.102	10.454	0.000**	1.838	0.544
Mileage (kmpl)	-0.236	0.031	-0.091	-7.615	0.000**	2.803	0.357
Owner Type	0.092	0.199	0.004	0.463	0.644	1.185	0.844
Engine (CC)	0.000	0.000	0.020	0.933	0.351	8.950	0.112
Power (bhp)	0.127	0.004	0.608	32.525	0.000**	6.815	0.147
Seats	-1.072	0.137	-0.077	-7.841	0.000**	1.858	0.538
R 2			0.700				
Adjusted R 2			0.700				
F			F (10,5836)=1363.756,p=0.000				
D-W value			0.763				

Note: dependent variable: price
* p<0.05 ** p<0.01

3.4. LightGBM model

From figure 7, it has been found that power holds the highest relative importance on affecting the used cars prices which accounts for 781.00% that result in a greatest impact of price. Besides, mileage also affect greatly to the price as it accounts for 625.00% while kilometers driven, year and location have similar significance of impact to prices. Other factors such as seats, transmission, fuel type and owner type have low importance.

DOI: 10.54254/2753-8818/105/2025.22745

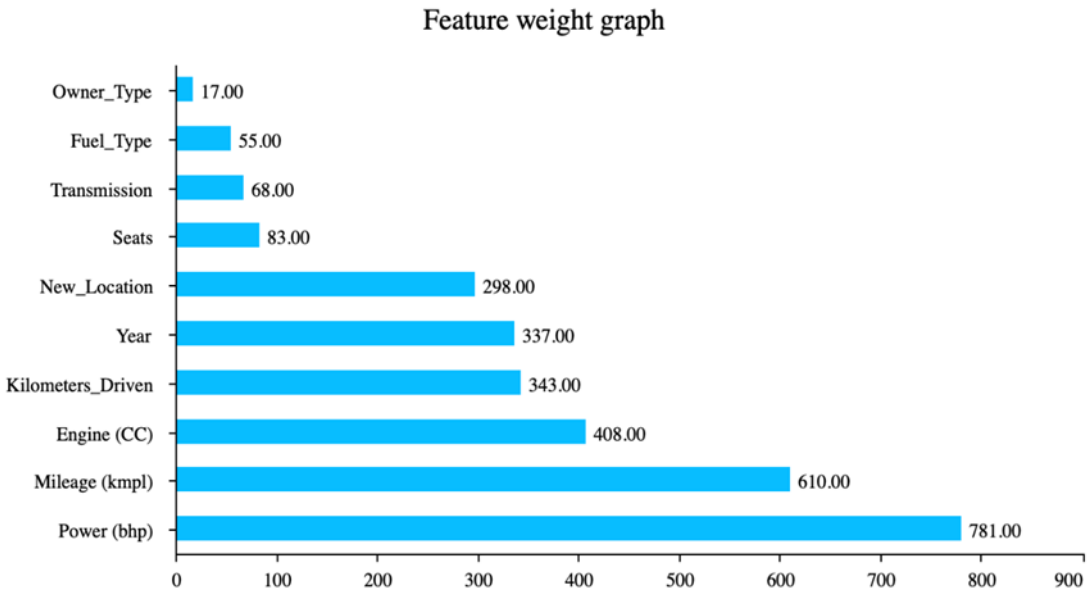

Figure 7: The relative importance of each variable

From the results (Table 3), it can be found that R^2 of the training set is 0.962 while the value of the test set is 0.930 which means the fit of the model is very high and although the performance of the test set is worse than that of training set, the model did not appear to have any overfitting problem. Moreover, it can be seen that the Mean Absolute Error (MAE) is 1.103, Mean Squared Error (MSE) is 4.858 and Root Mean Squared Error (RMSE) is 2.204 during the training of machine, therefore it can be concluded that the estimation error is low.

Table 3: The model evaluation results

Metric	Description	Training Set	Test Set
R-squared (R^2)	Goodness-of-fit measure, between 0 and 1, higher is better	0.962	0.930
Mean Absolute Error (MAE)	L1 loss, average difference between true and predicted values, closer to 0 is better	1.103	1.537
Mean Squared Error (MSE)	L2 loss, average of squared errors, closer to 0 is better	4.858	9.232
Root Mean Squared Error (RMSE)	Square root of MSE, average gap value	2.204	3.038
Median Absolute Deviation (MAD)	Absolute deviation from the median, robust to outliers, smaller is better	0.603	0.702
Mean Absolute Percentage Error (MAPE)	Average percentage error, robust to outliers, smaller is better	7.886	0.972
Explained Variance Score (EVS)	Measures the model's ability to explain data variance, between [0,1], higher is better	0.962	0.930
The mean square root logarithm error (MSLE)	When RMSE is the same, it punishes more for under-prediction	0.024	0.031

3.5. Compared results

In this paper, it has compared the two methods of understanding the factors affecting the used car prices and ways to make prediction, and it can be concluded that the LightGBM model is more

DOI: 10.54254/2753-8818/105/2025.22745

accurate in predicting the prices of used cars than the linear regression model. The specific comparison and analysis are shown below:

The value of R^2 of training set and test set for LightGBM model are 0.962, 0.930 respectively while the value of R^2 of training set and test set of linear regression model are both 0,700, it can be found that the fit of the LightGBM model is obviously higher than that of linear regression model.

The value of MAE, MSE, RMSE are 1.103, 1.537; 4.858, 9.232; 2.204, 3.038 for training set and test set, respectively by using LightGBM model. It can be seen that the error of prediction of the model was low which means the model has a high accuracy compared to liner regression model which has a D-W value of 0.76, it can be proved that the linear regression model may has a positive self-correlation in the residual as the model may miss some important variables or there may exist an incorrect form of function of model. Therefore, the LightGBM model has a better prediction of used car prices than linear regression model with a higher efficiency of processing large amount of data.

4. Conclusion

Through the research of the paper, it can be concluded that the prices of used car in India can be affected by various factors that are Kilometers Driven, Fuel Type, Year, Transmission, Owner Type, Mileage (kmpl), Engine (CC), Seats, Power (bhp) and Location while the power and mileage have the most significant impact on prices. The paper has compared the two methods applied in it which are LightGBM model and linear regression model and has found that the accuracy, suitability and efficiency of LightGBM model shown to be better than the regression model. And people can be inspired by the application of machine learning and prediction model in automotive industry and this paper can provide suitable suggestion for stakeholders in used car market in order to earn money by predicting the price. Additionally, other machine learning like deep learning or ensemble methods can enhance the accuracy of prediction. Overall, this paper has contributed to use machine learning method and explained its importance to tackle the real problems appeared in used car industry.

References

[1] *Venkatasubbu, P. and Mukkesh, G. (2019) Used cars price prediction using supe2rvised learning techniques. Int. J. Eng. Adv. Technol, 9, 13.*

[2] *Kriswantara, B. and Rifki, S. (2022) Machine learning used car price prediction with random forest regressor model. Journal of Information System, Informatics and Computing, 6, 40-49.*

[3] *Samruddhi, K. and Ashok-Kumar, R. (2020) Used car price prediction using k-nearest neighbor based model. Int. J. Innov. Res. Appl. Sci. Eng. (IJIRASE), 4, 686.*

[4] *Muti, S. and Kazım, Y. (2023) Using linear regression for used car price prediction. International Journal of Computational and Experimental Science and Engineering, 9, 11-16.*

[5] *Pudaruth, S. (2014) Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4, 753-764.*

[6] *AlShared, A. (2011) Used Cars Price Prediction and Valuation using Data Mining Techniques. International Journal of Computer Sciences, 12.*

[7] *Ahtesham, M. and Javairia, Z. (2022) Used car price prediction with pyspark. International conference on digital technologies and applications. Cham: Springer International Publishing.*

[8] *Chen, C.C., Hao, L.L. and Xu, C. (2017) Comparative analysis of used car price evaluation models. AIP Conference Proceeding, 1839.*

[9] *Chandak, A., et al. (2019) Car price prediction using machine learning." International Journal of Computer Sciences and Engineering, 7, 444-450.*

[10] *Collard, M. (2022) Price prediction for used cars: a comparison of machine learning regression models. Journal of Information System.*