Used Car Price Evaluation using three Different Variants of Linear Regression

Zamar khan[1]
Virtual university,Lahore, Pakistan.

Abstract- This paper exhibits car price estimation framework with the application of three variants of linear regression models. Multivariable Linear Regression, Lasso Regression and Ridge Regression being two different variants of Linear Regression, by applying three different regression models we will select the one with highest accuracy rate for car price estimation. This research is presenting a framework in which price is estimated entity which is anticipated, and the value of car dependent on features like car's makerMD, modelBR, mileageML, manufactureMR yearDT, engineHP displacementDS, enginepowerEP, bodytypeBT, transmissionMT,combustiontype and pricePkr.

Keywords: Used car price prediction, price evaluation, Lasso Regression, ridge, lasso, Ridge Regression

1 Introduction

Car price estimation particularly is applicable if the car is not new, with increment in demand for used cars, there is a decrease of 8 in need of utilized car 2015, increasingly more car purchasers are discovering options of purchasing used cars. Individuals want to purchase cars through rent, which is a lawful agreement between purchaser and merchant. The dealer class incorporates direct merchant or outsider, business element or protection agency. Under rent contract, the purchasers pay ordinary portions of the thing bought for a pre-characterized timeframe. These rent portions are needy upon the assessed value of the car and accordingly, dealers are intrigued to think about reasonable evaluated value of the car. Estimated through investigations that computing reasonable assessed price of a trade-in car is significant just as trying. In this way, exact price estimation system is required for the trade-in cars. Estimation techniques of machine learning can be useful in such manner.Inductive and deductive are two different machine learning approaches which are being used in this research. When data is available and existing actualities are forming a new data than this is to be named as deductive approach while in inductive approach we make our assumptions on the basis of experiences of application of techniques.

We applied deductive approach of multiple linear regression, it makes new qualities dependent on existing qualities. A variable X is to be estimated and estimation is to be done on various features of car which are being donated by Y. Strategy: To begin with, we gather the information about trade-in Cars, distinguish significant highlights that mirror the price. After gathering all dependents and independents factors we made a relationship of Variable X which is dependent and Y variable which is independent. Estimator: This research discovers those variables, which are related with the car, is the best estimator of its price.

Price Estimation: This research has analyzed three different machine learning approaches i.e., [1] Linear Regression, Lasso Linear Model and Ridge Linear Model being two different variants of Linear Regression and after deducing result we come up with the most accurate machine learning algorithm for car price estimation.This paper is divided in such a way Section II talks about the works done by others that estimate used car price., Section III emphases on the methodology of our framework and Section IV gives the Data Analytical Statistics, Section V shows the result and the reaming part of this paper is for conclusion and future work.

2 Related work

A dataset from Kaggle has been used in this research. Dataset which is being used is licensed and it is having approximately 4.5 lacs of rows of data. The literature study gives scarcely any paper analysts have done comparative study of application three regression variants which we have done [2]. The Author depicts nonexclusive motor stage evaluating the value of an asset. It gives a price calculation matrix to do the price estimation. For the conclusion of used car price, this stage this stage figures out linear regression model. It can be offered by linear regression what highlights can be utilized for explicit kind of cars for such estimation. Random forest model has been applied for the price estimation model in this research.

Dataset from kaggle has been utilized by Zhang et al. the Author of this research [3]used dataset to perform price estimation of a trade-in car. The creator assesses the exhibition of a few arrangement strategies like Scaler vector machine ,decision tree and others to survey the presentation. Among every one of these models, random forest demonstrates finest role for their estimation task. The work utilizes five highlights (Company, HyPower, Miles Covered, TimeSold, CarLife) to play out the arrangement task after evacuation of unessential highlights and exceptions the dataset provided the accuracy of 83% from the test data. We likewise use Kaggle data-set to to do estimation of trade-in car prices. Be that as it may, the distinction lies in the consideration of not many progressively important highlights in estimation framework the value of the car, and car Type. These two highlights assume a significant job in estimateing the price of a trade-in car which is by all accounts given less significance in the paper. Furthermore, the scope of highlights year of enrollment, HyPower, the price is by all accounts limited [2] due to which test data gives minimal accuracy with respect to the expanded scope of the above highlights.

Researcher Kanwal Noor and Sadaqat Jan. [4] Utilized Minitab, they get the value being anticipated in the extra section "FIT". Aside from it, the leftover worth being the contrast among genuine and anticipated response. Assumed point gathered by examination is to assemble a framework which has the abilities of managing high intricacy and gives precise outcomes independent of the extent of the dataset. The information they used is accumulated from pakwheels dataset at the outset, 2000 records of trade-in cars were recorded. The gathered information included variable qualities for value, motor limit, shading, notice date, number of perspectives, mileage in kilometer, housepower, combination edges, transmission, kind of motor, Number city, adaptation, model, make and model year. [12] When the information assortment was finished, we prepared information utilizing numerous direct relapse method for value forecast. In this exploration, measurable programming Minitab was utilized in which we input the information and break down the outcomes through direct relapse application. We considered all traits first however sooner we pushed the estimator determination approach on our information, and get the most accurate factors skirted all other inconsequential factors as well.

The researcher [5] estimates the value of trade in cars in Mauritius approaching four variants of ML algorithms i.e., multiple linear regression, k-nearest neighbors and decision trees algorithm. Researcher utilized the reliable information gathered from every day ads at Mauritius. Utilizing the documented learning algorithms on the basis of furnished information an equivalent outcomes with not very great estimation precision. The distinction amongst the examination of task is that we rollout our evaluation on information from Kaggle, while theirs depends on the information gathered by day to day paper. Likewise, the creator utilizes basic and practically identical order algorithms that comply with our discoveries that utilizing a refined algorithm just like random forest can give great out comes with better statically analysis and accuracy result collection can give truly great outcomes.

Table 1 Car Price Estimation Researches

Objectives	Methodologies	Perdition Country	Limitations	Year	Ref
Used car price forecasting for buyer	Linear Regression	Germany	Old car were not showing accurate price as Linear regression is limited to linear relation only.	2017	[6]
Innovation in techniques	Neuro-fuzzy inference	Taiwan	Condition of car was very much dependent to price which was being ignore in neuro inference.	2009	[7]
Used car Price Estimation	Random forest	Kaggle (Different Countries)	Ensemble of price decision trees encounters interpretability which fails	2018	[1]

			to determine other variable significance.		
Greater Price Estimation Precision Achievement	Artificial neural network, support vector machine	Bosnia and Herzegovina	The cost estimation model is reduced to a specific value of the error on which vehicle training sample means where achieved.	2019	[8]
Price Estimation for used card with adjusted values.	Multivariable linear regression	Pakistan	High Causation in correlation between the variables like horsepower portrays the antithesis to the price.	2017	[4]

3 Methodology

This research is providing a framework for used cars price estimation which is based on three different Machine Learning Techniques, Linear Regression, Lasso Linear Model and Ridge Linear Model being two different variants of Linear Regression. Figure 1 represents the proposed Framework. It works in a rotation of four stages s1, s2, s3 and s4. Each stage compromises of application of different methods and techniques. In s1, data collected from dataset is preprocessed, at very initial stage our data which was to be investigated was very large it was more than 8 lacs in record, each record consisted of one dependent variable Price ‘x’ and many independent variables such as features of car as ‘y’. We took about 4.5 lac entries and after feature Selection we found the best possible price estimator.

Data is preprocessed in Jupiter notebook using most popular libraries of Python named as Pandas and scikit-learn[9]. Dataset was imported and following operations was performed for its preprocessing.

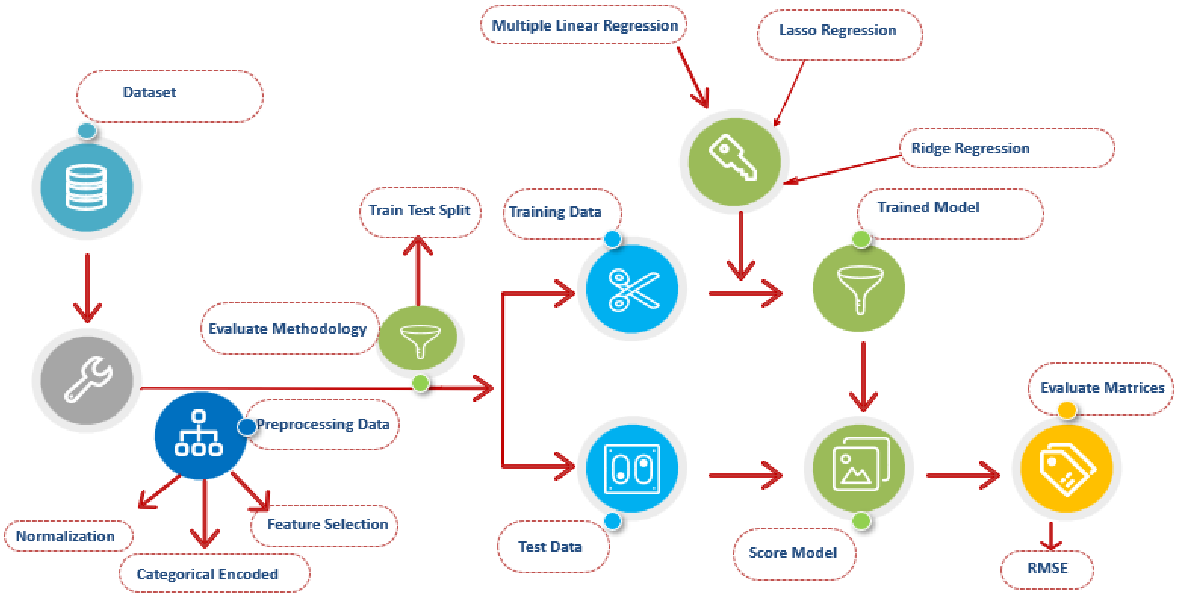

Figure 1. Framework Architecture

- Trimming data from 8 lacs to 4.5 lacs rows.
- Missing values were removed, and normalization was done.

- Feature Selection was done to find best price estimator.
- Categorical data Encoded for better results.

Table 1 shows the sample of data which we will be using as an input for the application of various regression techniques.

Sr	ID	maker_model	year	F_ID	fuel_type	price_eur
0	1	ford galaxy	2011	0	Diesel	10585
1	1	ford galaxy	2012	0	Diesel	11103
2	1	ford galaxy	1998	1	Gasoline	740
3	2	skoda octavia	2012	0	Diesel	8882
4	2	skoda octavia	2003	0	Diesel	4293

Table 1. Sample Input Data

[10]We have used Jupyter notebook and Pythons library scikit-learn for the implementation of regression algorithms

3.1 Multiple Linear Regression

An effective regression model is implemented to estimate the value of a car. Here car price is to be estimated and for this estimation we will be using estimators which are other attributes other than the price of car. We will be using a variable X for Input values i.e. Price, Engine Type and Age of Car and Y be the output i.e. Estimated Price, the equation below shows the linear regression correlation

$$Y=\beta_0+\beta_1 X \qquad (1)$$

Regression coefficients are represented by ß0, ß1, Y is the output which is to be estimated price and X shows the attributes which are to be fed as an Input. The equation given above shows the relation only when regression is linear with single input, if we have more than one input than our relation will be as following given in the equation no 2

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n \qquad (2)$$

In equation given above, X1,X2,X3,…Xn symbolize more than one input. For the number of particular inputs we can say n+1 regression coefficients are accumulated.

3.2 Lasso Regression

Decreasing the number of features, specification or properties of independent variable and estimate the result in such circumstance where independent variable are very few or likely to zero their lasso regression is very useful. In our framework we are left with three inputs only after feature selection and this mode of regression will best fit in. Lasso and its variations are principal to the field of compacted detecting. Under specific conditions, it can improve the specific arrangement of non-zero coefficients. Mathematical representation of this can be seen in equation given below.

$$\min_{\mathrm{w}} \frac{1}{2\mathrm{n}_{\mathrm{samples}}} ||\mathrm{X}_{\mathrm{w}} - \mathrm{y}||_2^2 + \alpha||\mathrm{w}||_1 \qquad (3)$$

The lasso measure in this way that understands the reduction of the least-square forfeit by the following expression below

$$\alpha||w||_1 \qquad (4)$$

We are left with a constant variable α with a ω i the l1-norm of the coefficient vector.

3.3 Ridge Regression

There was cardinality in our variable we have applied Ridge regression methodology for its removal and for better results. Ridge regression reports roughly few issues faced by the Ordinary Least Squares by striking a forfeit over the coefficient. The ridge coefficients reduce the accuracy of residual sum of squares that can be represented in mathematical equation below.

$$\min_{w} ||X_w - y||_2^2 + \alpha||w||_2^2 \qquad (5)$$

The complexity parameter $\alpha \geq 0$ handles the amount of shrinkage: the larger the value of α the larger the amount of shrinkage and thus the coefficients become more tough to form collinearity.

$$\mathrm{Cos}\,t(w) = RSS(w) = \sum_{i=1}^{N}\{y_i - \hat{y}_i\}^2 = \sum_{i=1}^{N}\left\{y_i - \sum_{j=0}^{M} w_j x_{i\,j}\right\}^2$$

You see that they are fundamentally the same as except RR introduces this new lambda which is a tuning parameter, increasing lambda leading to diminished variance with slight bias.The objective of Ridge regression is to produce most minimal MSE and it an accomplishes this by selecting the appropriate lambda. In general Ridge regression performs well when number of estimators are very huge (p > n), we took different experiments and after that analyzed the result using LS for this situation will create appraises that don't have exclusive solutions where Ridge regression will advance.

4 EXPLORATORY DATA ANALYSIS

When data is preprocessed, a visual exploration of data is done to analyze and collect intuitions for framework that would be accumulated for preprocessed data. We used graphical representation of data for explanatory data analysis and to discover the type of relation of dependent and independent variables.

Data Colums (total 5 columns):	
Id	453878 non-null
Maker model	453878 non-null
Manufacture year	453878 non-null
Fuel type	453878 non-null
Price eur	453878 non-null

Table 2 Number of rows of dataset

[11] We have used Python Library Panda for the statistical data analysis. Our dataset compromises of huge amount of data, we have 670 unique cars as a basic entity and 453878 rows of data varying on 670 unique entities. I would not be possible for us to plot graph of 670 unique entities so we will do sampling of data we take a fewer part of data for graphical representation and in tables we will analyzed total amount of data.

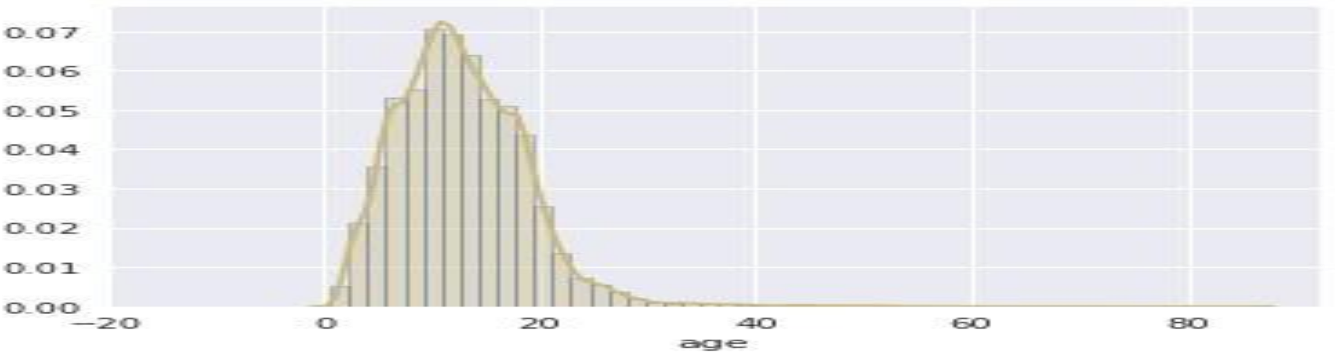

Figure 2 Car Age

Fig 2 represents the graphical view of age of car is calculated from its date of manufacturing till this current year (2020).

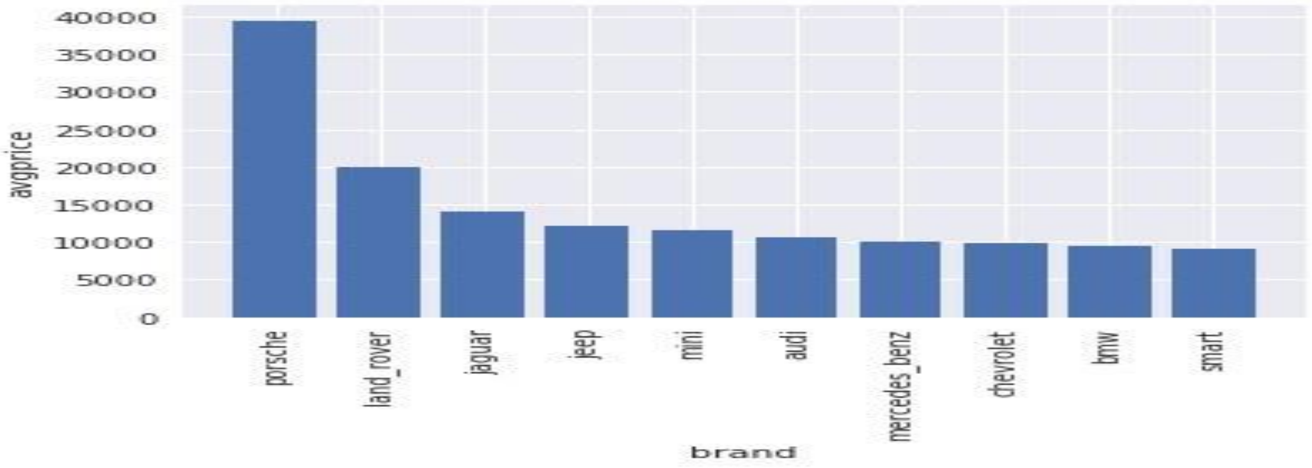

Figure 3 Brand wise Price Analysis

Fig 3 show the analysis of price which varies from brand to brand. Here price of 10 most expensive cars have been analyzed and represented in a bar chart.

5 EVALUATION

Python Sklearn model selection library has been used to test train evaluation in all of three regression variant we have split the matrices in two random subsets which are supposed to be called Test and Train Matrices. Test and train subset are made on Ratio of 80% train data and 20% test data.

Id	363102
Maker model	363102
Manufacture year	363102
Fuel type	363102
dtype	363102

Table 3 Train Data

Table 4 shows the Table of Number of Training Instances After Splitting the 80% of data randomly for train Matrix

Id	90776
Maker model	90776
Manufacture year	90776
Fuel type	90776
dtype	90776

Table 4 Test Data

Table 5 shows the Table of Number of Testing Instances After Splitting the 20% of data randomly for Test Matrix.

5.1 Evaluation Measures

We have used Root Mean Square Error (RMSE) scoring for evaluation matrices of all three regression algorithms of price estimation framework. Mathematical representation of RMSE is given equation below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\{y_i - \hat{y}_i\}^2} \cdot$$

RMSE is calculating the error deviation between the actual price and Estimated price. In our implementation our RMSE was Calculated between 0-1 which is considered as best results of evaluation matrices.

6 Results

[9] We have used Jupyter Notebook and Pythons Different Libraries in Sci-kit learn for the compilation of result. For the plotting of graphs, probability plotting graphs, Bar Charts etc.

6.1 Multivariable Regression

In fig 4 a graphical representation of over multivariable regression is exhibited if lasting tenets are distributed ordinarily or they are having greater amount of deviation. Probability Plot graph represented In Fig is having 1 root mean square value which can be represented as root mean square obtained. Linear Regression is applied and results are shown as normal probability graph. Percentage is shown in aloft slope line.

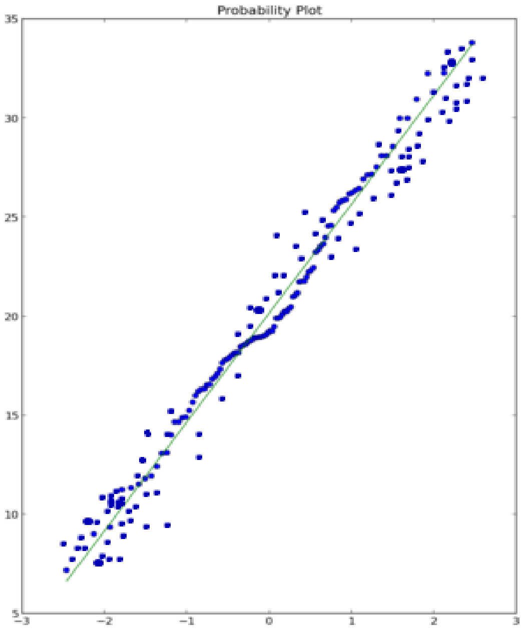

Figure 4 Probability Plot Graph

The In fig a straight line of red color can be seen this straight line is surrounded by the grouped blue dots the well-adjusted segmentation of points on line hence the notion is true and we can say that results are valid.

6.2 Lasso Regression

Analyzing Fig 5, the probability plot exhibits that Zero origin is playing a role in hosting other lines with the increment in lambda. Each line in the probability plot exhibits the amount of the variance for the framework, persistency of lambda proves to be the regularization parameter.

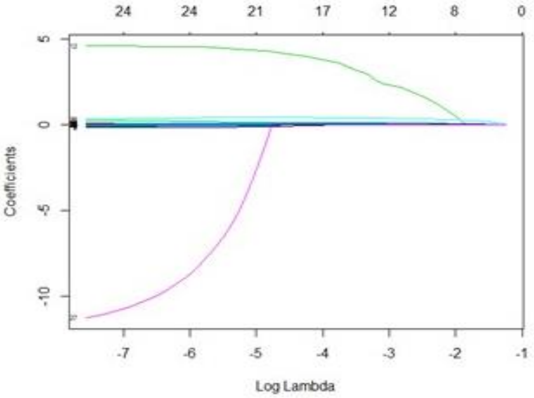

Figure 5

6.3 Ridge Regression

Based on Fig, the graphical representation of lambda is huge in capacity RMSE is high. The value of log lambda has approached to -2. The ideal rate of lambda for Ridge regression is 0.03200. Log lambda 1's graphical representation can be shown below in graph fig 6.

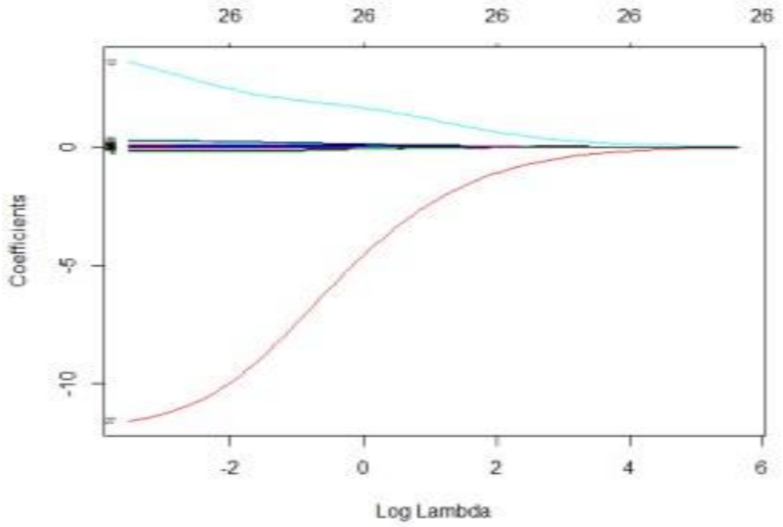

Figure 6

RMSE value is high, log of lambda value is likely to be -7, With such huge RMSE value is minimized and a straight line is drawn to show the results.

	Multivariable Linear Regression	Lasso Regression	Ridge Regression
RMSE	0.77	0.79	0.77

7 Conclusion

This framework estimates used car price doing various analytical operations and by treating it with different regression variants. An accuracy of 79% using Lasso regression model has been achieved. Among all regression variants lasso regression gives the most accurate result. The most relevant features used for this estimation are price, car brand, model, and its age, after feature selection.

8 Future Work

One could use this used car price estimation framework as reference, Once could use more advance ML models or methodologies i.e., fuzzy logic and genetic algorithms for car price estimation work aim to create a programmed, collaborative system that compromises of a source of used cars with their value.

9 References

[1] N.Pal, et al. “How Much Is My Car Worth? A Methodology for Predicting Used Cars’ Prices Using Random Forest”. in Future of Information and Communication Conference. 2018. Springer.

[2] R.Morabito, J. Kjällman, and M. Komu. “Hypervisors vs. lightweight virtualization: a performance comparison”. in 2018 IEEE International Conference on Cloud Engineering. 2018. IEEE.

[3] Xinyuan Zhang , Z.Z.a.C.Q., “Model of Predicting the Price Range of Used Car”. 2017.

[4] R.Noor, and S.J.I.J.o.C.A. Jan, “Vehicle price prediction system using machine learning techniques”. 2017. 167(9): p. 27-31.

[5] Pudaruth, S.J.I.J.I.C.T. “Predicting the price of used cars using machine learning techniques”. 2014. 4(7): p. 753-764.

[6] S.Lessmann, and S.J.I.J.o.F. Voß, Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. 2019. 33(4): p. 864-877.

[7] Wu, J.-D., C.-C. Hsu, and H.-C.J.E.S.w.A. Chen, “An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference”,. 2009. 36(4): p. 7809-7817.

[8] Gegic, E., et al., Car price prediction using machine learning techniques. 2019. 8(1): p. 113.

[9] developers, s.-l. Supervised learning. 2019 [cited 2020 27-01-2020]; Available from: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning.

[10] Dey, A. Data Preprocessing for Machine Learning. 2018 [cited 2020 27-01-2020]; Available from https://medium.com/datadriveninvestor/data-preprocessing-for-machine-learning-188e9eef1d2c.

[11] F.Nelli,Python data analytics: with pandas, numpy, and matplotlib. 2018: Apress.

[12] Muhammad Khan, & Asma Kanwal, & Sagheer Abbas, & Faheem Khan, & Taeg. Whangbo, (2021). Intelligent Model for Predicting the Quality of Services Violation. Computers, Materials and Continua. 71. 607-3619. 10.32604/cmc.2022.023480.