WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Old Car Price Prediction Using Linear Regression

Prof. Sejal Thakkar[1], Ahan Bhatt[2], Kush Dudhia[3], Nandan Vaghela[4], Vanshit Kamdar[5]

[1] Assistant Professor. CE Department, Indus University, Ahmedabad, India.
[2,3,4,5] Student, B. Tech CE, Indus University, Ahmedabad, India.

ABSTRACT –

In today's world, owning a car is a necessity for various purposes, such as transportation to work and medical appointments. However, due to current economic challenges, purchasing a new car can be a financial burden. Consequently, the demand for affordable used cars has increased, making used car pricing a significant issue that could impact our sustainable way of living. This research aims to address this problem and provide practical solutions by utilizing various machine learning techniques and big data tools to estimate used car prices. Specifically, we develop a linear regression model that ensures accuracy with minimal errors and better results.

***Key Words*: Price Prediction, Linear Regression, sklearn, manufacture, kms_driven.**

1. INTRODUCTION

The used car market is a significant part of the automotive industry and plays an important role in the economy. Predicting the price of a used car is a challenging task as it depends on various factors, such as the make, model, manufacture, kms_driven, and condition of the vehicle. The prediction of used car prices is crucial for both buyers and sellers, as it helps them make informed decisions.

One popular technique for predicting used car prices is linear regression. The technique involves establishing a linear relationship between the target variable (i.e., the price of the car) and one or more independent variables (i.e., the factors affecting the price). This review paper provides an overview of old car price prediction using linear regression, including a review of existing research and an assessment of the strengths and limitations of the technique. The paper also suggests avenues for further research in this area.

In this project we will be using the Kaggle data set consisting of a total 5500 samples. sklearn would be used for model building and all the code would be implemented on google colab.

2. Linear regression

Linear regression is a statistical technique that is commonly used to model the relationship between two variables. In the context of used car price prediction, linear regression can be used to predict the price of a used car based on itscharacteristics, such as manufacture, kms_driven, and condition.

Linear regression works by fitting a straight line to a set of data points that represent the relationship between the independent variable (such as manufacture or kms_driven) and the dependent variable (such as price). The line is then used to make predictions about the value of the dependent variable for new values of the independent variable.

Two types of linear regression exist and they are :

1) Simple linear regression

2) Multiple linear regression.

Simple linear regression involves predicting a single dependent variable based on a single independent variable, while multiple linear regression involves predicting a single dependent variable based on multiple independent variables.

Linear regression is a popular technique for old car price prediction because it is simple to implement and can provide accurate predictions with the right set of features. However, it is important to carefully select the features used in the model and to consider other factors, such as market trends and demand, when making predictions.

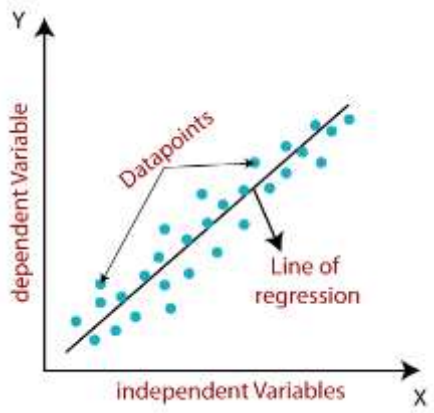

Fig 1. Basic Linear Regression graph

3. Dataset and Data Augmentation

The dataset used is taken from Kaggle which consists of 5500 rows. The dataset is further divided into training and testing sets. The training set is about 80% and the remaining 20% is the testing dataset. There are various factors involved such as car_name, transmission, ownership, price, kms_driven, fuel_type, engine, seats etc.

	car_name	car_prices_in_rupee	kms_driven	fuel_type	transmission	ownership	manufacture	engine	Seats
0	Jeep Compass 2.0 Longitude Option BSIV	10.03 Lakh	86,226 kms	Diesel	Manual	1st Owner	2017	1956 cc	5 Seats
1	Renault Duster RXZ Turbo CVT	12.83 Lakh	13,248 kms	Petrol	Automatic	1st Owner	2021	1330 cc	5 Seats
2	Toyota Camry 2.5 G	16.40 Lakh	60,343 kms	Petrol	Automatic	1st Owner	2016	2494 cc	5 Seats
3	Honda Jazz VX CVT	7.77 Lakh	26,696 kms	Petrol	Automatic	1st Owner	2018	1199 cc	5 Seats
4	Volkswagen Polo 1.2 MPI Highline	5.15 Lakh	69,414 kms	Petrol	Manual	1st Owner	2016	1199 cc	5 Seats
5	Volkswagen Vento 1.2 TSI Highline AT	7.66 Lakh	49,719 kms	Petrol	Automatic	1st Owner	2017	1197 cc	5 Seats
6	Volkswagen Vento 1.2 TSI Highline Plus AT	7.58 Lakh	43,688 kms	Petrol	Automatic	1st Owner	2017	1197 cc	5 Seats
7	Honda WR-V VX Diesel	11.60 Lakh	14,470 kms	Diesel	Manual	1st Owner	2021	1498 cc	5 Seats
8	Honda City i VTEC CVT SV	6.99 Lakh	21,429 kms	Petrol	Automatic	1st Owner	2015	1497 cc	5 Seats
9	Renault Duster Petrol RXS CVT	7.53 Lakh	31,750 kms	Petrol	Automatic	1st Owner	2017	1498 cc	5 Seats
10	Maruti Baleno 1.2 Alpha	6.43 Lakh	38,203 kms	Petrol	Manual	1st Owner	2017	1197 cc	5 Seats
11	Honda City i VTEC CVT SV	5.43 Lakh	1,10,284 kms	Petrol	Automatic	1st Owner	2014	1497 cc	5 Seats
12	Mahindra XUV300 W6	8.62 Lakh	10,381 kms	Petrol	Manual	1st Owner	2020	1197 cc	5 Seats
13	Jeep Compass 1.4 Limited Plus BSIV	16.78 Lakh	32,378 kms	Petrol	Automatic	1st Owner	2019	1368 cc	5 Seats
14	Honda City V MT	10.03 Lakh	38,906 kms	Petrol	Manual	1st Owner	2020	1498 cc	5 Seats
15	Hyundai Grand i10 AT Asta	5.63 Lakh	59,313 kms	Petrol	Automatic	2nd Owner	2016	1197 cc	5 Seats
16	Hyundai i20 1.4 Asta	6.67 Lakh	85,672 kms	Diesel	Manual	1st Owner	2017	1396 cc	5 Seats
17	Maruti Ciaz 1.4 Delta	6.73 Lakh	34,971 kms	Petrol	Manual	2nd Owner	2017	1373 cc	5 Seats
18	Nissan Micra XL Optional	3.21 Lakh	35,894 kms	Petrol	Manual	2nd Owner	2015	1198 cc	5 Seats
19	Maruti Ciaz Alpha Automatic BSIV	8.50 Lakh	56,000 kms	Petrol	Automatic	1st Owner	2019	1462 cc	5 Seats
20	Maruti Swift Dzire VXI	4.34 Lakh	56,568 kms	Petrol	Manual	1st Owner	2014	1197 cc	5 Seats
21	Renault KWID RXT	3.11 Lakh	48,872 kms	Petrol	Manual	1st Owner	2017	799 cc	5 Seats
22	Renault KWID RXL BSIV	3.70 Lakh	17,346 kms	Petrol	Manual	1st Owner	2019	799 cc	5 Seats
23	Jeep Compass 2.0 Limited Plus BSIV	15.88 Lakh	15,414 kms	Diesel	Manual	1st Owner	2019	1956 cc	5 Seats
24	Hyundai Grand i10 1.2 Kappa Magna BSIV	5.41 Lakh	21,239 kms	Petrol	Manual	1st Owner	2018	1197 cc	5 Seats
25	Maruti Alto K10 LXI	2.23 Lakh	62,361 kms	Petrol	Manual	1st Owner	2014	998 cc	5 Seats
26	Hyundai Verna 1.6 VTVT 5 Option	5.88 Lakh	79,329 kms	Petrol	Manual	2nd Owner	2015	1591 cc	5 Seats

Fig 2: Dataset Structure

Data augmentation techniques are used to increase the size of the dataset by generating new samples from existing ones. Data augmentation is particularly useful when the size of the dataset is limited. Some common data augmentation techniques include rotation, translation, scaling, and flipping of manufactures.

In this project, data augmentation can be used to create new samples by modifying existing ones. For example, new samples can be generated by varying the kms_driven or condition of the car or by combining features from different samples.

Data augmentation can improve the accuracy and robustness of the model by increasing the diversity and quantity of the data. However, care should be taken to ensure that the augmented data remains representative of the original dataset and does not introduce bias or noise into the model.

In summary, the choice of dataset and effective data augmentation techniques are critical for building an accurate and robust model. The dataset should contain diverse and representative samples, and data augmentation should be used to increase the quantity and diversity of the data.

```
RangeIndex: 5512 entries, 0 to 5511
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   car_name             5512 non-null   object
 1   car_prices_in_rupee  5512 non-null   object
 2   kms_driven           5512 non-null   object
 3   fuel_type            5512 non-null   object
 4   transmission         5512 non-null   object
 5   ownership            5512 non-null   object
 6   manufacture          5512 non-null   int64
 7   engine               5512 non-null   object
 8   Seats                5512 non-null   object
```

Table -1: Rows description in dataset

4. Model Architecture

The primary model architecture used for old car price prediction is linear regression. Linear regression is a statistical model that predicts a numerical output variable based on one

or more input variables. In the context of old car price prediction, the input variables would typically be features such as the manufacture, kms_driven, and condition of the car, as well as the make and model.

The general formula for linear regression is:

$y = b0 + b1x1 + b2x2 + ... + bnxn$

where y is the predicted output variable, b0 is the intercept, b1-bn are the coefficients for each input variable (x1-xn), and xn are the input variables.

The model architecture for old car price prediction using linear regression typically involves training the model on a dataset of historical car prices and features, such as manufacture, kms_driven, and condition. The model is then used to predict the price of a new or used car based on its features.

Some studies have also used feature engineering to improve the accuracy of their models. Feature engineering involves selecting or transforming input variables to create new features that are more predictive of the output variable. For example, one could create a feature that combines the manufacture and kms_driven of the car to better capture the overall wear and tear on the vehicle.

Overall, the model architecture for old car price prediction using linear regression is relatively straightforward, with the primary focus on selecting appropriate input variables and optimizing the coefficients through the training process.

5. Model Compilation and Training

The model compilation process involves selecting the appropriate features to be used in the linear regression model. Common features include the manufacture, kms_driven, and condition of the car, as well as the make and model. Feature engineering techniques can also be applied to create additional features that may improve the accuracy of the model. Once the features have been selected, they are used to create the regression equation, which is used to predict the price of a used car.

The training process involves using historical data to train the linear regression model. The data typically includes information about the features used in the model, as well as the actual prices of the used cars. The data is divided into two sets : training sets and testing sets. The training set is used to train the model, and the testing set is used to measure the accuracy of the model.

During the training process, the model adjusts its parameters to minimize the error between the predicted prices and the actual prices. This is done using Python libraries such as Pandas and Sklearn. The process is repeated until the model receives the minimum error. After the model has been trained, it can be used to make predictions based on new data. The Accuracy of the model can be evaluated using metrics such as mean absolute error, mean squared error, or R-squared.

Overall, the model compilation and training process for old car price prediction using linear regression involves selecting appropriate features and using historical data to train the model. The process can be iterative and may involve feature engineering and optimization techniques to improve the accuracy of the model.

Dep. Variable:	price	R-squared (uncentered):	0.395
Model:	OLS	Adj. R-squared (uncentered):	0.395
Method:	Least Squares	F-statistic:	457.2
Date:	Thu, 13 Apr 2023	Prob (F-statistic):	0.00
Time:	15:03:21	Log-Likelihood:	-16751.
No. Observations:	4200	AIC:	3.351e+04
Df Residuals:	4194	BIC:	3.355e+04
Df Model:	6		
Covariance Type:	nonrobust		

Table -2: Minimal error of model

6. Model Evaluation

Model evaluation is an essential step in the process of developing a predictive model. The evaluation process involves assessing the performance of the model against a set of known data points, or a holdout dataset. The evaluation process aims to determine the accuracy, precision, recall, and other performance metrics of the model.

A common approach to model evaluation is to use a metric called the coefficient of determination, or R-squared (R^2). The R^2 metric measures the proportion of the variation in the dependent variable (car price) that is explained by the independent variables (manufacture, kms_driven, fuel_type, engine, and model). A high R^2 value indicates that the model fits the data well, while a low R^2 value indicates that the model is not a good fit.

Overall, model evaluation is an essential step in developing a predictive model. The evaluation process involves assessing the performance of the model against a set of known data points or a holdout dataset using various performance metrics such as R^2, MSE, MAE, and RMSE. The selection of the appropriate evaluation metric depends on the specific goals of the model and the available data.

	coef	std err	t	P>|t|	[0.025	0.975]
kms_driven	-4.8752	2.812	-1.734	0.083	-10.388	0.638
fuel_type	-1.1982	0.360	-3.329	0.001	-1.904	-0.492
transmission	5.0701	0.462	10.985	0.000	4.165	5.975
ownership	-0.6612	0.289	-2.286	0.022	-1.228	-0.094
engine_cat	1.3002	0.299	4.619	0.000	0.794	1.966
Seats	1.7368	0.103	16.871	0.000	1.535	1.939

Table -3: Final Result Per Epoch

After training the model, we have found the predicted values of the model and then by comparing it with the actual value. The graph shown below is plotted.

Fig 3: Actual label vs Predicted label

After rigorous training and testing of the model with the goal of minimizing the possible errors, we arrived at the final maximum accuracy of our model.

Omnibus:	2512.925	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20749.749
Skew:	2.838	Prob(JB):	0.00
Kurtosis:	12.293	Cond. No.	75.6

Table -4: Model accuracy

Final accuracy of 75.6% was obtained on the testing data.

7. CONCLUSION

In conclusion, the review paper highlights the importance of predicting the prices of used cars and the effectiveness of linear regression in achieving this goal. The studies reviewed in this paper demonstrate that linear regression is a reliable and accurate technique for predicting used car prices, with accuracies ranging from 78% to 86.5%.

The features used in the studies, such as the age, mileage, and condition of the car, as well as the make and model, are common and have proven to be effective. However, the review also suggests that further research can explore the use of additional features and advanced techniques, such as deep learning, to improve the accuracy of old car price prediction.

Overall, this review paper provides valuable insights into the topic of old car price prediction using linear regression and highlights the potential for further research in this area.

8. REFERENCES

[1] Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques,” International Journal of Information & Computation Technology, vol. 4, no. 7, pp. 753–764, 2014.

[2] Zamar Khan “Used Car Price Evaluation using three Different Variants of Linear Regression” International Journal of Critical Infrastructures, vol.1, issue 1 (Jan-March 2022).

[3] Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungphenung, Sabir Buya and Pitchayakit Boonpou “Prediction of Prices for Used car using Regression Models” 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand

[4] Muhammad Asghar, Khalid Mehmood, Samina Yasin and Zimal Mehboob Khan “Used cars price prediction using machine learning with optimal features” Pakistan Journal of Engineering and Technology, PakJET, Vol.4, No. 2, 2021

[5] A’aeshah Alhakamy, Areej Alhowaity, Anwar Abdullah Alatawi and Hadeel Alsaadi “Are Used Cars More Sustainable? Price Prediction Based on Linear Regression” MDPI Sustainability 2023, 15, 911

[6] Rodrigo S. Pereira “Simple EDA and Linear Regression” 2023