*Available online at http://www.advancedscientificjournal.com*
*http://www.krishmapublication.com*
*IJMASRI, Vol. 1, issue 10, pp. 256 -262, Dec. -2021*
**https://doi.org/10.53633/ijmasri.2021.1.10.005**

**INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)**

**ISSN: 2582-9130** **IBI IMPACT FACTOR 1.5** **DOI: 10.53633/IJMASRI**

**RESEARCH ARTICLE**

# USED CAR PRICE PREDICTION

**[1]Vaibhav Gupta, [1]Sharma M.L and [1]Tripathi K.C**
*[1]Department of Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India*

## Abstract

Cars have become a necessity in this modern world. Every middle class family needs a vehicle or a mode of transport in order to move from one place to another. Not everyone is able to afford a new vehicle as they are costly and there's an added cost of taxes and various other expenses by both the provider/company of the car as well as the government. Moreover, not every customer is sure of spending a sum of their wealth on a certain car. The product might not meet their needs. The solution to this problem of having a car despite not being able to afford one is met by buying and selling second hand cars. It has become its own market now. There are already numerous companies and websites and app based services that serve as a mediator or a platform for the dealing of second hand or used cars and other vehicles. Establishment of such places is easy but there is another problem that still remains- How to price the used car appropriately at a price comfortable for both the seller and the buyer? Luckily, the Used Car Price Prediction systems exist and can be developed. Users might think that it's easy to determine the price of a used car, and whether there is even a need to have such a system. In truth, there are a lot of factors that are important in determining the price of a second hand vehicle. The quality of a vehicle deteriorates with age[1] of course but that is not all. Every single vehicle is different even when it is manufactured and sold as a new product and even more so when the same vehicle is used over time. Different people may use their vehicles more or less depending on their everyday activity, making kilometers driven as one of the important factors for the price prediction. It is obvious that a vehicle which is driven for 2000 kilometers in 1 year would be priced less than a vehicle which has been driven for only 500 kilometers in 2 years. This is just one of the factors that determine the price of a used car. In our Car Price Prediction System, we have used the Year of Manufacturing (used to determine the age of the vehicle by subtracting this from the

date of selling), the original maximum retail price of the vehicle (the price at which the vehicle was sold at from the manufacturing company/garage), the fuel type of the vehicle (Petrol, Diesel, CNG, Electric ; This affects the pricing severely as different fuel type engines have different prime performance periods and different rates of deterioration), Seller Type (Individual or Dealership), Transmission (Manual or Automatic), Number of past owners of the vehicle. Using all these factors[2], we are going to determine which model is best to determine a price for the used vehicle. For the Car Price Prediction System, Regression models[3]are used since these models give the results as a continuous curve instead of a categorized value as a result. Due to this, we can use the continuous curve to determine an accurate price for each and every scenario which won't be possible if the results obtained were in the form of a range. The final model of the system will implement the best suited algorithm and have a UI (User Interface) which make it possible for a user to be able to enter the values of these deciding factors and the system will predict the price for them.

**Keywords:** Car price prediction, machine learning, regression analysis, linear regression, correlation analysis

## Introduction

As mentioned above, the problem of predicting the price of a used car accurately is quite complex since there are a lot of affecting factors. The objective of this project is to determine a suitable and accurate price for a used vehicle based on the various features. In order to draw a conclusion, we implement several learning methods on our sample data set. The models that will be used for comparison are Linear Regression, Logistic Regression, Ridge Regression, Lasso Regression, Polynomial Regression and Bayesian Linear Regression[4]. We will train the models using each of the above techniques one by one and then observe the predictions and the accuracy of the trained model to figure out which regression model gives the best results from the Car Price Prediction System. Regression algorithms are the only techniques being tested since the other techniques are unnecessary. As we know, regression algorithms give the results of the structure of a continuous curve which is the ideal type of result for the Car Price Prediction System[5]. The other algorithms that give the final output in the form of a range are irrelevant since they do not contribute to the solution to the problem. Regression analysis is the predictive modelling technique which compares and establishes a relation between a dependent variable and the independent variable. The independent variables are the ones we mentioned earlier. Year of Manufacturing (used to determine the age of the vehicle by subtracting this from the date of selling), the original maximum retail price of the vehicle (the price at which the vehicle was sold at from the manufacturing company/garage), the fuel type of the vehicle (Petrol, Diesel, CNG, Electric ; This affects the pricing severely as different fuel type engines have different prime performance periods and different rates of deterioration), Seller Type (Individual or Dealership), Transmission (Manual or Automatic), Number of past owners of the vehicle. All of these are independent variables. The dependent variable is also the target of our problem which is the selling price. The graph that is obtained would be used to determine the value of X after the dependent factors Y has already affected the vehicle. Regression analysis is the technique that is used to solve regression based problems in ML. The regression technique is used primarily to calculate the strength of the predictor, forecast of the trend, time series, and in case of cause & effect relation.

## Requirements

### *Hardware Requirements:*

Windows 7,8,10,11
Mac OS X 10.11 or higher, 64-bit
Linux: RHEL 6/7, 64-bit
x86 64-bit CPU (Intel / AMD architecture)

4 GB RAM
Software Requirements:
Python 3.6 or above
PyCharm
Anaconda
PIP 2.7
Jupyter Notebook
Chrome
Notepad++
Vs Code/Sublime text editor

## Methodology

Every model in ML has two parts. First is the training phase. In this phase, the data is used to train the model. In other words, the data is used to calculate correlations between the different input factors and output factors. It is the sample data which is used as a reference for the testing phase. In the training phase, the data from our dataset is used to fit the model. Second is the testing phase. In this phase, we already have an existing trained model and all that's left is to give the trained model some sample input values and use these to gather new outputs. Since the correctness of the model depends on both training phase and testing phase as well as the training and testing data, the authenticity of the dataset is of utmost concern before fitting it into the model. The Car Price Prediction Algorithm uses the inputs given to determine an accurate output or selling price. It is necessary for the model to have the right algorithm used for this job. Hence we will be comparing the different regression algorithms and their performance in this experiment.

### *Objectives*

- To examine the data and determine the best factors for predicting the car price
- To compare various regression algorithms and to determine the best suited algorithm for the software.

5 GB free disk space

- To create a working model that predicts the Prices of the Used Cars.
- To make sure that the Car Price Prediction System has good accuracy.
- To implement a UI so as to take input values from the user to predict the selling price.

### *Procedure*

***First step:*** To collect a concrete and authentic dataset for the used cars and their features. The quality of the dataset is the cornerstone of the entire model and it's essential to obtain a good dataset first

***Second Step:*** Data Processing. Data Processing means converting the raw data available into data that is readable by the machine. It has various subparts. Data cleaning: Means removing the unwanted fields from the data and any other undesirable elements from the data. Data reduction: Reduction of data into simpler terms and taking only a part of the data for the first or second phase of the model. Data Transformation: Conversion of data into numerical forms that are easier for the machine to process and perform operations on.

***Third Step:*** Using the test regression algorithms to predict the final output

***Fourth Step:*** To observe and compare the output values using different algorithms and draw a conclusion for the best suitable algorithm for the Car Price Prediction System.

### *The Dataset*

The dataset used for the research is the used car data from a used car reselling website, Car Dekho.

```
df.head()
```

| | Car_Name | Year | Selling_Price | Present_Price | Kms_Driven | Fuel_Type | Seller_Type | Transmission | Owner |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ritz | 2014 | 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual | 0 |
| 1 | sx4 | 2013 | 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual | 0 |
| 2 | ciaz | 2017 | 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual | 0 |
| 3 | wagon r | 2011 | 2.85 | 4.15 | 5200 | Petrol | Dealer | Manual | 0 |
| 4 | swift | 2014 | 4.60 | 6.87 | 42450 | Diesel | Dealer | Manual | 0 |

**Fig:** Raw Dataset Head

As observed, the raw dataset has a few useless fields and some vague fields as well. So, we drop the car name field as it has no contribution whatsoever in predicting the price of the car. There are also fields like Fuel Type, Seller Type, and Transmission which do not have any mathematical values and have text values instead. But these values can't be eliminated like the car name since these do have an effect on the price of the car. In order to convert these fields into mathematical values for the ease of fitting these into our machine learning model, we use the one hot encoding technique.

```
final_dataset.head()
```

| | Selling_Price | Present_Price | Kms_Driven | Owner | No_Years | Fuel_Type_Diesel | Fuel_Type_Petrol | Seller_Type_Individual | Transmission_Manual |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 3.35 | 5.59 | 27000 | 0 | 7 | 0 | 1 | 0 | 1 |
| 1 | 4.75 | 9.54 | 43000 | 0 | 8 | 1 | 0 | 0 | 1 |
| 2 | 7.25 | 9.85 | 6900 | 0 | 4 | 0 | 1 | 0 | 1 |
| 3 | 2.85 | 4.15 | 5200 | 0 | 10 | 0 | 1 | 0 | 1 |
| 4 | 4.60 | 6.87 | 42450 | 0 | 7 | 1 | 0 | 0 | 1 |

**Fig:** Dataset after modifications

A correlation is drawn between all the variables in the table[6]. The correlations help us figure out which of the fields are highly correlated, that is, have a huge impact on the others fields when changed. A highly positive correlation between 2 fields say field A and field B means that change in field A has an impact on field B. And the positive relation here means that an increase in A would also result in increase in B and a decrease in A would give a decrease in B. Correlation between 2 variables doesn't necessarily have to be positive. A highly negative correlation is just as important for a good prediction. Let's say columns A and C and have highly negative correlation with each other. This means that an increase in the value in A would result in decrease in value in C and a decrease in value of A would result in increase in value of C.

In order to find the variables that would be useful for the prediction of car price, we would use the correlation between all the variables and filter out the variables that have a highly positive or negative correlation with the car price.

```
final_dataset.corr()
```

| | Selling_Price | Present_Price | Kms_Driven | Owner | No_Years | Fuel_Type_Diesel | Fuel_Type_Petrol | Seller_Type_Individual | Transmission_ |
|---|---|---|---|---|---|---|---|---|---|
| **Selling_Price** | 1.000000 | 0.878983 | 0.029187 | -0.088344 | -0.236141 | 0.552339 | -0.540571 | -0.550724 | -0. |
| **Present_Price** | 0.878983 | 1.000000 | 0.203647 | 0.008057 | 0.047584 | 0.473306 | -0.465244 | -0.512030 | -0. |
| **Kms_Driven** | 0.029187 | 0.203647 | 1.000000 | 0.089216 | 0.524342 | 0.172515 | -0.172874 | -0.101419 | -0. |
| **Owner** | -0.088344 | 0.008057 | 0.089216 | 1.000000 | 0.182104 | -0.053469 | 0.055687 | 0.124269 | -0. |
| **No_Years** | -0.236141 | 0.047584 | 0.524342 | 0.182104 | 1.000000 | -0.064315 | 0.059959 | 0.039896 | -0. |
| **Fuel_Type_Diesel** | 0.552339 | 0.473306 | 0.172515 | -0.053469 | -0.064315 | 1.000000 | -0.979648 | -0.350467 | -0. |
| **Fuel_Type_Petrol** | -0.540571 | -0.465244 | -0.172874 | 0.055687 | 0.059959 | -0.979648 | 1.000000 | 0.358321 | 0. |
| **Seller_Type_Individual** | -0.550724 | -0.512030 | -0.101419 | 0.124269 | 0.039896 | -0.350467 | 0.358321 | 1.000000 | 0. |
| **Transmission_Manual** | -0.367128 | -0.348715 | -0.162510 | -0.050316 | -0.000394 | -0.098643 | 0.091013 | 0.063240 | 1. |

**Fig:** Correlation table between the variables

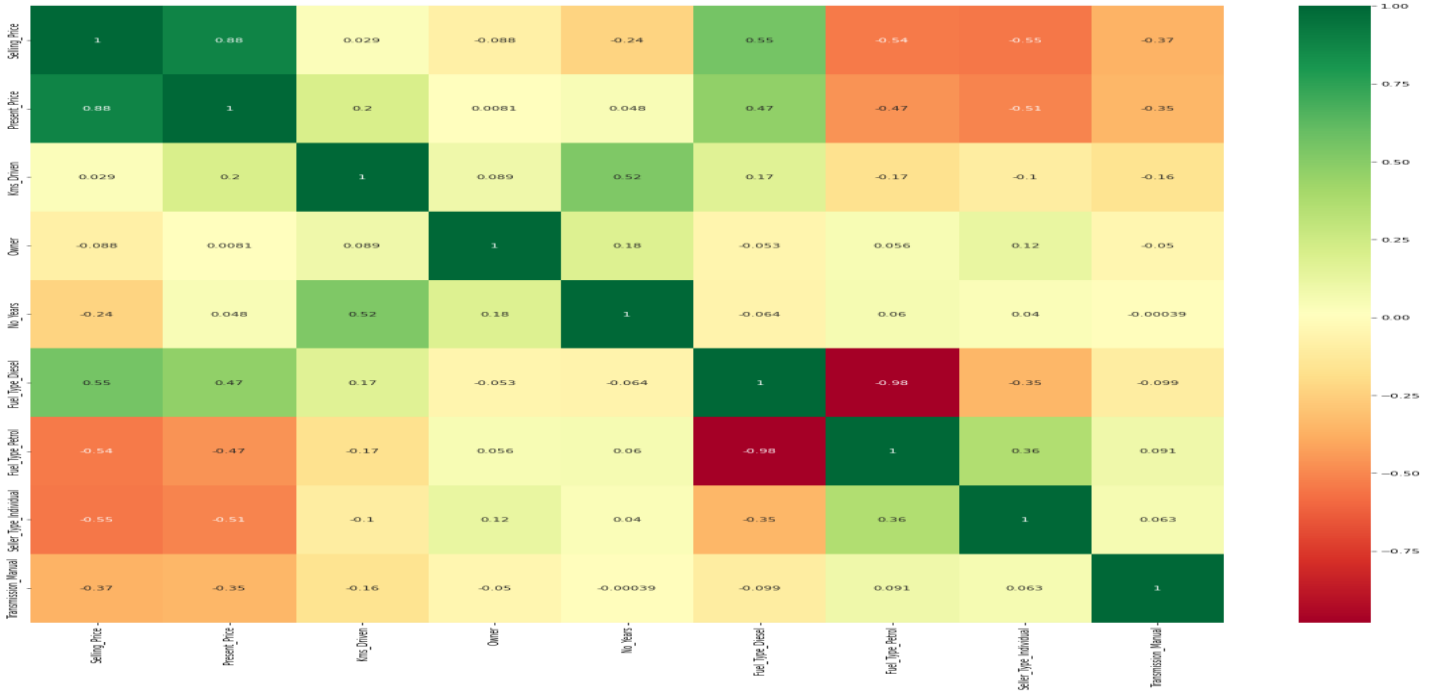



**Fig:** Correlation heat map between variables

Since correlation function determines the degree of relation between two variables, it will never have a value exceeding 1 and falling short of -1. A correlation value of 0 means that the variables do not have any impact on each other.

### *Linear Regression*

Linear regression method, as the name suggests uses a straight line to make the predictions. The plotting of this prediction line however, is not as simple.

The linear regression line has an equation of

$$Y = a + bX$$

Where X is the independent variable(s) and Y is the dependent variable that is to be predicted.

The linear regression line is usually the best fitting line going through a plot of Dependent variable values (Y axis) versus the independent variable values (x axis). The data used to determine this line is called the training data. It is

the data set which is used to train our data as implied.

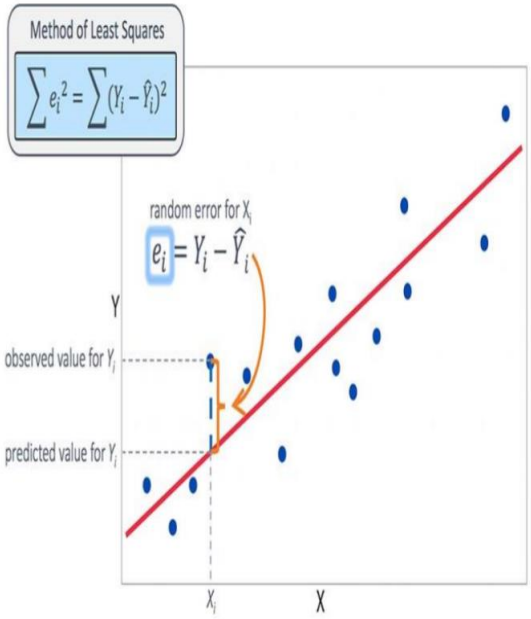



**Fig:** Linear Regression

In the method of least squares linear regression, we calculate the sum of squares of vertical distances of the training data points from the test line. The line which has the least sum of squares of vertical distances from the data points is the fitted line and is used for prediction. It is obvious that the accuracy of such a model can be increased by simply increasing the size of data.

Linear regression has an advantage when the requirement of the system is to get a continuous output line. And as the car price prediction system is of such nature, linear regression is the suitable type of regression for it.

Since there is more than one variable in the field X and only one independent variable Y,

The relation used for car price prediction is:

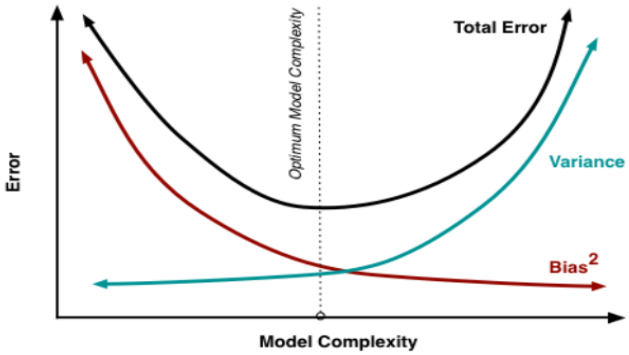
***y = a+ s1X1+s2X2+…… snXn***

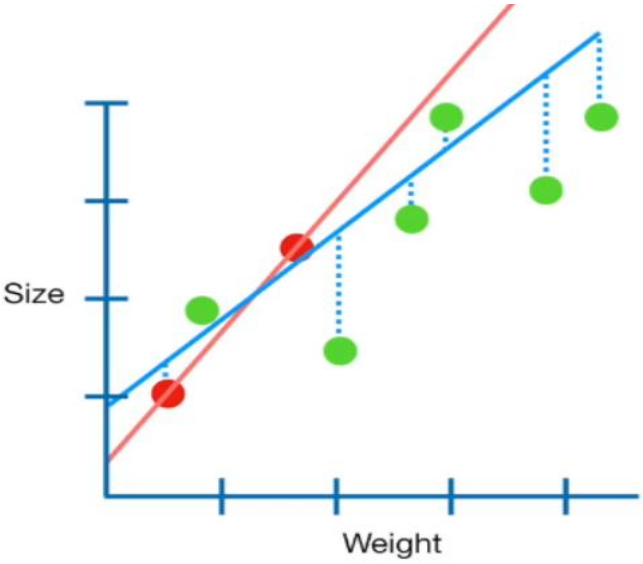
Here,

s1, s2, s3 …. Are the slopes of best fit lines of these factors a is the y intercept X1, X2, X3 are the independent variables y is the dependent variable which in this case would be our predicted price.

### *Ridge Regression*

Ridge regression is the regression technique useful when the data that is being analyzed and worked on is multicollinear in nature. Multicollinearity of data means when one predicted value from the numerous models is in a linear relation with the others. This is advantageous to increase the accuracy of the model. The ridge regression is a type of technique which relies on regularization of data. OLS estimate is the point where the data best fits the model. The ridge regression model works on modification of parameters to reduce the variance to a minimum. The tuning of the Lambda parameter is done in ridge regression.



In future, having better and concise data would result in the improvement of this proposed model. Adding more to the training set would improve the accuracy of the model by a lot. Due to the lack of data, the fuel types were limited to Petrol, Diesel, CNG in the current model. But in the future, having more data on electric vehicles would be beneficial for the car price prediction model.

**Conclusion**

The modern scenario has resulted in a need for the selling and buying of second hand vehicles. Because of this, the system is valuable and with society moving away from fossil fuels, it is inevitable that the old vehicles running on exhaustible resources such as Petrol, Diesel, and CNG would be discarded and the sales of electric vehicles would also increase. This would create a huge wave in the used car buying and selling industry making the used car price prediction system a must. In this research paper and in reference to the mentioned research papers, conclusions were drawn to make the car price prediction system accurate. A comparison was done between all the variables in the dataset and the variables that could be used in the prediction of car price were determined. The regression algorithms were compared. The proposed system will help in determining the prices of the used cars.

**Reference**


1. Listiani. M. 2009. "*Support Vector Regression Analysis for Price Prediction in a Car Leasing Application"*. Thesis (M.Sc). Hamburg University of Technology
2. Richardson. M. 2009."*Determinants of Used Car Resale Value"*. Thesis (BSc). The Colorado College
3. Gelman, A and Hill, J. 2006. *"Data Analysis Using Regression and Multilevel Hierarchical Models"*. Cambridge University Press, New York, USA
4. Quinlan, J. R. 1993. C4.5: "*Programs for Machine Learning"*. Morgan Kaufmann
5. Nitis Monburinon., Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, 2018. "*Prediction of Prices for Used Car by using Regression Models*" (ICBIR 2018)
6. Noor. K and Jan S. 2017. "*Vehicle Price Prediction System using Machine Learning Techniques* " (IJCA 2017)


*****