International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 4, July Year 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Unveiling Price Patterns in the Automotive Market Through Regression Techniques

Abarna D*, Shenbaha S**
*(Department of Mathematics, Dr. N. G. P. Arts and Science College, Coimbatore Email:abarnadhandapani13@gmail.com)
**(Department of Computer Science with Data Analytics, Dr. N. G. P. Arts and Science College, Coimbatore Email: sshenbaha@gmail.com)

Abstract
In recent times, there have been studies on the use of regression model algorithms for prediction in many areas such as stock market; finance; health care, etc. The proposed case study gives the regression model for the prediction of price of cars. The factors contributing to the car price prediction is of six factors. The prediction of a car price has become a high interest research area, as it ultimately transforms the future world. Many studies have been proposed for the prediction of car price. Thus, we develop a linear regression model that can estimate car prices based on various features such as car name, selling price and kilometers driven etc. The model's predictions are validated against a dataset so as to predict the value of car approximately. Numerous features are examined for reliable and accurate predictions. Statistical methods like linear and multiple regression algorithms, decision tree algorithm, random forest techniques and their performance is compared to predict the accuracy of that particular model used. The method is very efficient and has higher accuracy in polynomial regression model.

Keywords: Car Price Prediction, Polynomial Regression, Python Programming.

I INTRODUCTION
In today's busy working world, business plays a very important role in molding the global economy and innovation. From very small startups to multinational companies, businesses of all fields are rapidly evolving and leading to cut above the rest. The automobile industry is a vital part of the global economy, with millions of vehicles being bought and sold every year. In most developed are influenced by a complex array of factors, including vehicle characteristics, market trends, and economic conditions. To address this challenge, this study proposes a car price prediction model using a regression approach. By analyzing a comprehensive dataset of used car countries, around 70 percent of the vehicles manufactured are automobiles. In this paper, predicting car prices is crucial for both buyers and sellers, as it enables informed decision-making and maximizes the investment. So purpose is to explain the price of cars where the basic statistical method regression is used, and which helped to estimate predictions. However, car prices

prices and attributes, our model aims to identify the key factors driving car prices and provide accurate predictions.
Based on the statistical and mathematical analysis, there are various types of regression models. Here we refer to multiple regression modelsin which

there is more than one independent variable. The steps of the model are discussed in the methodology section with subsequent analysis.

II LITERATURE REVIEW

Huseyn Mammadov proposed a model for predicting car prices in USA using linear regression model. This paper was published in the year 2021 and graphically represented the model using heatmap correlation. The data used for the prediction were collected from the web portal fred.stlouisfed.org using web scraper, written in Python/Jupyter programming language. The dataset includes the following characteristics: car ID, wheelbase, car length, car width, car height, horsepower, etc. It produced result of 85% predictability.

Sümeyra MUTİ[1] and Kazım YILDIZ[2] gave a model for car price prediction using linear regression. This study has data set that includes the features and price information of vehicles in Turkey as the year 2020. As a result, when selected 1/3 of the data set as the test data, it was observed that the R2 score for the prediction success of model was 73%.

Another author named AravindSasidharanPillai, presented a deep learning approach to car price prediction in the year 2022, where he used dataset of 30 popular US car brands. Experimental results showed that an ANN model with a mean absolute percentage error of 11percent and an R2 value of 0.96 outperforms the random forest model with a MAPE of 14 percent and an R2 value of 0.94.

In this case study, we will proceed with polynomial regression model carried outand visualized using Python programming language anticipating with minimum error percentage.

III METHODOLOGY

This study schematizes the hypotheses and presents them in Fig 1. The objective of the model is to predict the car price based on the attributes like car name, year, kilometers driven, fuel type, seller type, transmission, owner, mileage, engine , maximum power and number of seats of the vehicle. These factors contribute to the overall selling price of the vehicle.

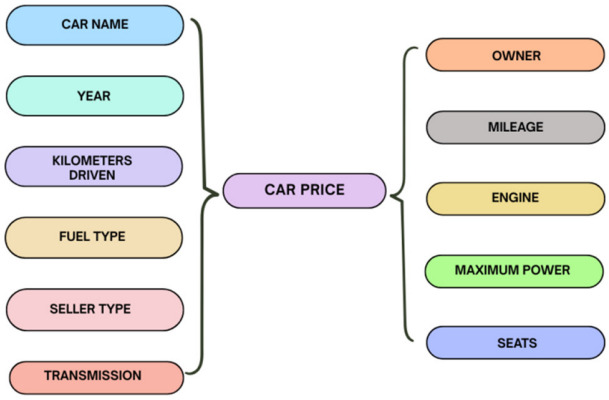

Figure 1:Attributes of car price prediction

3.1 Preparation of the dataset

The objective of the case study is to predict the selling price of a car based on factors such as car name, year, kilometers driven, fuel type, seller type, transmission, owner, mileage, engine , maximum power and number of seats. The model is based on the parameters, and then is evaluated to predict the present price of particular model of a car. The prediction of future values of car price involves few steps of data manipulation to obtain the cumulative trend so as to match the original dataset of the past. Let's have a look at the dataset and analyze which model would fit in.

name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	ileage(km/ltr/k	engine	max_power	seats
Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4	1248	74	5
Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14	1498	103.52	5
Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7	1497	78	5
Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23	1396	90	5
Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1	1298	88.2	5
Hyundai Xcent 1.2 VTVT E Plus	2017	440000	45000	Petrol	Individual	Manual	First Owner	20.14	1197	81.86	5
Maruti Wagon R LXI DUO BSIII	2007	96000	175000	LPG	Individual	Manual	First Owner	17.3	1061	57.5	5
Maruti 800 DX BSII	2001	45000	5000	Petrol	Individual	Manual	Second Owner	16.1	796	37	4
Toyota Etios VXD	2011	350000	90000	Diesel	Individual	Manual	First Owner	23.59	1364	67.1	5
Ford Figo Diesel Celebration Edition	2013	200000	169000	Diesel	Individual	Manual	First Owner	20	1399	68.1	5
Renault Duster 110PS Diesel RxL	2014	500000	68000	Diesel	Individual	Manual	Second Owner	19.01	1461	108.45	5
Maruti Zen LX	2005	92000	100000	Petrol	Individual	Manual	Second Owner	17.3	993	60	5
Maruti Swift Dzire VDi	2009	280000	140000	Diesel	Individual	Manual	Second Owner	19.3	1248	73.9	5

Figure 2: Dataset display

3.2 Data collection

The dataset used for this case study consists of some attributes taken from online sources. It contains for about 8000 data in all variables. This model can also be carried out in natural conditions like a survey that can be taken among the people who is willing to buy or sell cars. The visualization of the model can be done in MATLAB or EXCEL as graph, scatter diagram, etc.

Initially, some of the data were taken and analyzed which factor would majorly affect the car price prediction model. Let us name the independent variables taken in this model from X_1 to X_{11} and dependent variable as Y. Each independent variable contributes to the overall price and any change in those variables (X_1 to X_{11}) might affect the selling price of the car.

3.3 Data preprocessing

This section will check dataset for missing values and check the data types of different features like numeric, string or float etc. The accuracy of car price prediction models depends more on the quality of the data used to train them.

- Data cleaning: Ensures the data is consistent, accurate and stable.
- Missing values: Cells with no values of the attribute.
- Chaotic formatting: Cells with uneven entries of the variables such as ‘Ertiga’, ‘ERTIGA’ or ‘ertiga’.
- Removing outliers: Data may have unusual patterns with very high or low values among them.
- Erroneous data: Mistakes or errors may contain like numeric values stored as text.
- Duplicate entries: Human errors may lead to duplicate records in data sometimes.
- Data transformation: Converting the categorical values into numeric values for encoding.
- Data integration: Combining the data taken from various sources.
- Training and testing: Avoid over fitting and ensure the model generalizes well to unobserved data.
- Data splitting: Splitting the data for training and testing for model evaluation.

The process carried out in this paper is illustrated in Fig.3.

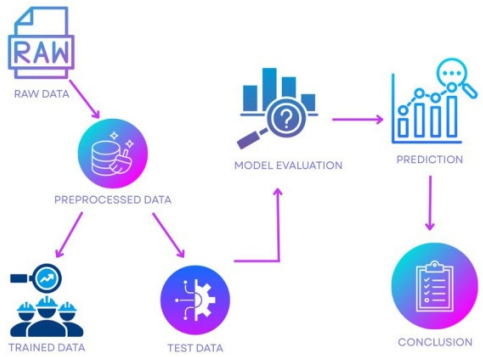

Figure 3: Process

The dataset undergoes preprocessing procedures and are collected to put in the suitable regression models. Preprocessed dataset is shown in Fig.4.

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage(km/ltr/kg)	engine	max_power	seats
0	1307	2014	450000	145500	1	1	1	0	23.400000	1248.000000	74.00	5.000000
1	1607	2014	370000	120000	1	1	1	2	21.140000	1498.000000	103.52	5.000000
2	385	2006	158000	140000	3	1	1	4	17.700000	1497.000000	78.00	5.000000
3	781	2010	225000	127000	1	1	1	0	23.000000	1396.000000	90.00	5.000000
4	1349	2007	130000	120000	3	1	1	0	16.100000	1298.000000	88.20	5.000000
5	693	2017	440000	45000	3	1	1	0	20.140000	1197.000000	81.86	5.000000
6	1397	2007	96000	175000	2	1	1	0	17.300000	1061.000000	57.50	5.000000
7	1064	2001	45000	5000	3	1	1	2	16.100000	796.000000	37.00	4.000000
8	1908	2011	350000	90000	1	1	1	0	23.590000	1364.000000	67.10	5.000000
9	286	2013	200000	169000	1	1	1	0	20.000000	1399.000000	68.10	5.000000
10	1521	2014	500000	68000	1	1	1	2	19.010000	1461.000000	108.45	5.000000

Figure 4: Preprocessed data

3.4 Regression model

One of the most common methods used in analyzing a relationship between two or more variables is regression analysis. We examine the relationship between two or more independent variables and create a model that can be used for predictive purposes. Our aim is to create a model and study the relationship between variables when one dependent and several independent variables are present. We denote by Y the random variable to be predicted, also called the *dependent* variable (or response variable) and x_i the *independent* (or predictor) variables used to model (or predict) Y. The process of finding a mathematical equation that best fits the noisy data is known as *regression analysis*. The technique of regression is one of the most popular statistical tools to study the dependence of one variable with respect to another. There are different forms of regression: *simple, linear, nonlinear, multiple*, and others. The primary use of regression model is prediction. When using a model to predict Y for a particular set of values $x_1,\ldots,x_k$, one may want to know how large the error of prediction might be. In this case study, we have used polynomial regression since the dataset is more suitable for this model.

Polynomial regression model

Polynomial regression is a type of regression analysis where the relationship between the predictor variable(s) and the response variable is modeled for prediction.

The general form of a polynomial regression model is:

$$Y=\beta_0 + \beta_1 X^2 + \ldots.+ \beta_n X^n + \varepsilon$$

Where Y is the dependent (or response) variable, X is the independent (or predictor) variable, β_0, β_1,...., β_n are regression coefficients, and the ε is the error term.

For model validation, the mean absolute error (MAE), the mean squared error (MSE) and the root mean squared error (RMSE) were used to evaluate the performance of regression model with our dataset containing various data points using the following formulas:

1. The mean absolute error, or MAE, is a measure of the errors that occur when a model is evaluated. It can be calculated by comparing the actual values with the sample size.
2. The mean square error, or MSE, is a statistical measure that shows the difference between the predicted and observed values of a model. When a model has no errors, theMSE is zero. On the other hand, if a model has high error value, it can be calculated using the formula.
3. The RMSE is a measure that shows the extent to which the residuals are spread outacross the data points in the regression line. It is also used to measure the concentration ofthe data in the optimal fit.
4. The R^2 value is the measure of proportion of variance in the dependent variable that is predictable from the independent variable(s) in a regression model. It ranges from 0 to 1.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Figure 5: Regression Metrics

Let us name the variables of the polynomial regression as follows:

- Independent variable: Car name (X_1), Year (X_2), Kilometers driven (X_3), Fuel type (X_4), Seller type(X_5), Transmission (X_6), Owner (X_7), Mileage (X_8), Engine (X_9), Maximum power (X_{10}) and Seats (X_{11}).
- Dependent variable: Selling price (Y).

Among the independent variables, attributes like car name, fuel type, seller type, transmission and owner have string values. So conversion of string values to float values helps in obtaining more accuracy.

Here we used polynomial regression model and we obtain the R^2 value as 0.889, which represent a good model. The prediction of the selling price of cars involves few steps of data manipulation to obtain the cumulative trend so as to match the actual dataset.The regression equation gives the predicted values of cars which help in obtaining a more suitable model. Below are some predicted values of price of cars from the model evaluated.

Actual_selling_price	Predicted_selling_price
450000	390513.6905
370000	458528.2576
158000	130771.7298
225000	334752.1681
130000	205985.9691
440000	513768.2367
96000	304892.1186
45000	-105615.4573
350000	365291.1485
200000	284967.5071
500000	570384.9085
92000	152736.0876
280000	207628.4225
200000	227101.9131
180000	187939.9358

Figure 6: Actual vs. Predicted Selling price of Cars

3.5 Model Visualization

Visualization helps identify patterns, outliers and relationships in the data. It evaluates the model performance and identifies areas for improvement. Some common ways of visualization includes histograms, bar charts, pie charts, heat maps, scatter plots, etc. By choosing the right visualization method, we can effectively communicate insights and help understand the data. Here the regression fitting of the data with predicted values of the evaluated data is visualized as bar chart using python code.The bar chart representing the predicted values of present price of cars is given below.

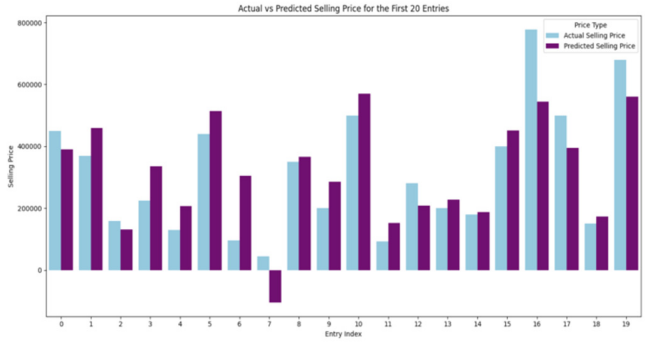

Figure 7: Bar chart denoting the selling price of cars

Another way of visualization is done as line graph which is as follows:

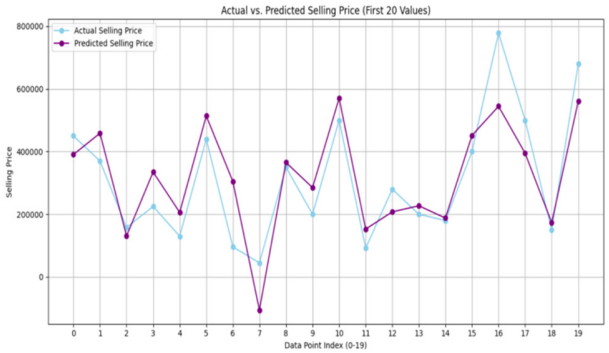

Figure 8: Line graph of variation in predicted vs. actual prices

Both the bar chart and line graph gives a very clear view of the predicted model. Since the visualization is done only for the first 20 entries of the dataset, it seems that the accuracy of the model might be low. The model has accuracy of 89% which is fairly accurate and reasonably good, based on the large dataset of over 8000 entries, despite the visualization only showing a small subset of 20 entries.

IV CONCLUSION

According to the statistical analysis, the proposed methodology predicts the selling price of cars using regression model. With the comprehensive dataset of over 8000 entries helps in accurately predicting the actual selling price of cars. The model's performance evaluation metrics indicate a good fit, suggesting that it can be a useful tool in the industry. The visualization seems to have some variations to the original data but have only fewer deviations when comparing to the overall dataset. Apart from the polynomial regression, we can even choose other regression models according to the variables involved and the accuracy obtained in that model. More appropriate datasets can be used to obtain better results. It's more likely that you'll obtain a good model if the data undergoes preprocessing procedures. This model can not only be used in the automotive industry but also be applied in various sectors to predict outcomes, or estimate approximate profit values.

References

[1] Sümeyra MUTİ1, Kazım YILDIZ2, (2023). Using Linear Regression For Used Car Price Prediction.DOI: 10.22399/ijcesen.1070505.

[2] Debanjan Parbat a ,MonishaChakraborty b (2020). A python based support vector regression model for prediction of COVID19 cases in India.

[3] HuseynMammadov Carlo Bo University of Urbino, Italy, Car Price Prediction in the USA using Linear Regression (2021).

[4] Muhammad Abdullah, Car Price Prediction Project using Random Forest Model, 2025.

[5] Sukhmandeep Singh Brar, Car Price Prediciton Dataset, 2024.

[6] Debanjan Parbat [a], Monisha Chakraborty [b,] A python based support vector regression model for prediction of COVID19 cases in India, 2020.

[7] Gelman, A., Hill, J. (2006). Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, New York, USA.

[8] Eric H. Au, et al. Prediction Modeling—Part 1: Regression Modeling (2020).

[9] Navdeep Tangri et al. A Predictive Model for Chronic Kidney Disease Progression (2011).

[10] Zhenya Qi, Yudong Feng, et al. Enhancing Hydropower Predictions using XGBoost and SVR Models (2025).

