Car Price Prediction Using Machine Learning

Ketan Agrahari[1], Ayush Chaubey[2], Mamoor Khan[3], Manas Srivastava[4]
[1,2,3,4]*Department of Computer Science and Engineering, Raj Kumar Goel Institute of Technology, AKTU*

Abstract **- The demand for used cars has increased significantly in the past decade and it is prognosticated that with Covid-19 outbreak this requirement will augment considerably. Hence to enhance the reliability, with the expansion of the used car market, a model that can forecast the current market price of a used automobile on the basis of a variety of criteria. This analysis can be used to study the trends in the industry, offer better insight into the market, and aid the community in its smooth workflow. The aim of this research paper is to predict the car price as per the data set (previous consumer data like engine capacity, distance traveled, year of manufacture, etc.). The result of these algorithms will be analyzed and based on the efficiency and accuracy of these algorithms, the best one of them can be used for the said purpose.**

Index Terms **- Machine Learning, Linear Regression, Lasso Regression, Correlation.**

I.INTRODUCTION

The used automobile market is a growing business with a market value that has nearly doubled itself in previous years. The rise of online websites and other tools like it have made it easier for both buyers and sellers to get a better understanding of the factors that determine the market value of a used car. Based on a set of factors, Machine Learning algorithms may be used to forecast the price of any automobile.

The data set will include information on a variety of automobiles. There will be information regarding the vehicle's technical elements, such as the engine type, fuel type, the kilometers per liter, and more, for each car.

There is no universal mechanism for establishing the retail price of used automobiles because different websites employ different methods to create it. By using statistical models to anticipate pricing, it is possible to obtain a preliminary price estimate without having to enter all of the details into the desired website. The main purpose of this study is to compare the accuracy of two different prediction models for estimating a used car's retail price.

As a result, we offer a Machine Learning-based methodology for predicting the prices of secondhand cars based on their characteristics. The cost is calculated using the amount of characteristics. Then, to illustrate our findings, we construct a responsive website that includes all of the countless used car listings. Our efforts culminated in this deployed service, which integrates data, machine learning, and features. This methodology can assist consumers looking to purchase a used car in making more informed judgments. Customers can now look for all automobiles in a region without physical efforts, anytime and from any location.

In this research, we used linear regression and lasso regression to develop a price model for used automobiles in a comparative research. Data was gathered from Kaggle for each algorithm. The main goal of this study is to discover the best predictive model for estimating the price of a used car.

II. LITERATURE REVIEW

With the recent arrival of internet portals, buyers and sellers may obtain an appropriate status of the factors that ascertain the market price of a used automobile. Lasso Regression, Multiple Regression, and Regression Trees are examples of machine learning algorithms. We will try to develop a statistical model that can forecast the value of a pre-owned automobile based on prior customer details and different parameters of the vehicle. [2] This paper aims to compare the efficiency of different models' predictions to find the appropriate one.

On the subject of used automobile price prediction, several previous studies have been conducted.

To anticipate the value of pre-owned automobiles in Mauritius, Pudaruth employed naive Bayes, k-nearest neighbours, multiple linear regression, and decision trees. However, because there were fewer cars

observed, their results were not good for prediction. In his article, Pudaruth concluded that decision trees and naive Bayes are ineffective for continuous-valued variables.[4]

To anticipate the price of a vehicle, Noor and Jan employed multiple linear regression. They used a variable selection methodology to determine the variables that had the highest influence and then eliminated the remainder. Only a few variables are included in the data, which were utilised to create the linear regression model. With an R-square of 98 percent, the outcome was outstanding. [4]

Peerun et al. conducted study to assess the neural network's performance in predicting used automobile prices. However, especially on higher-priced cars, the estimated value is not very close to the real price. In forecasting the price of a used car, they found that support vector machine regression outperformed neural networks and linear regression by a little margin. [4]

To accurately anticipate the price of a car, many different approaches have been used in the digital world, ranging from machine learning approaches like multiple linear regression, k-nearest neighbor, and naive bayes to random forest and decision tree to the SAS enterprise miner. In [7], [8], [9], [10] and [11] all of these solutions took into account distinct sets of attributes when making predictions based on the historical data used to train the model.

We attempted to construct a web application where a user may verify the effective market price of their automobiles using a model for prediction based on the factors that have the greatest impact on vehicle prices.

III. METHODOLOGY

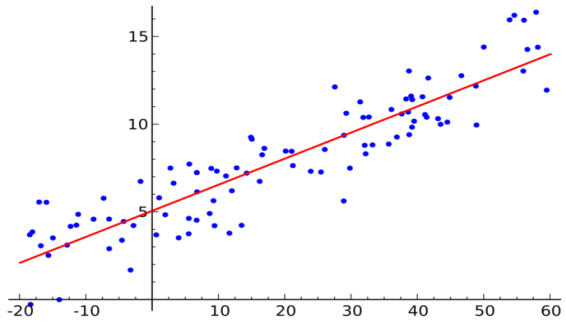
The main goal of this method is to give users an accurate estimate of how much has to be paid for the given vehicle. The model may give the customer a record of possibilities for various automobiles based on the details of the automobile the customer wants. The system assists in providing the customer with sufficient data to help him to reach a conclusion.

The used automobile market is expanding at an exponential rate, and vehicle vendors may profit from this by offering incorrect prices to capitalise on the demand. As a result, a system that can predict the price of a car based on its parameters while also taking into consideration the costs of competing vehicles is necessary. Our system fills in the gaps by providing buyers and sellers with an estimate of the car's value based on the best algorithm available for price prediction.

A. LINEAR REGRESSION

Regression is a technique for predicting a dependent factor using independent factors.

The approach is typically employed for predicting and calculating correlations between independent and dependent factors. The regression model establishes the relationship between independent factors and dependent factors i.e. linearly or exponentially.



Linear regression is a sort of regression analysis in which there is only one independent variable and the independent(x) and dependent(y) variables can be bound in a linear relationship. In the graph above, the red line is called the best fit straight line. We aim to plot a line that best predicts the data-points based on the given data-points. The linear equation provided below can be used to represent the line.

y = a0 + a1 * x #Linear_Equation

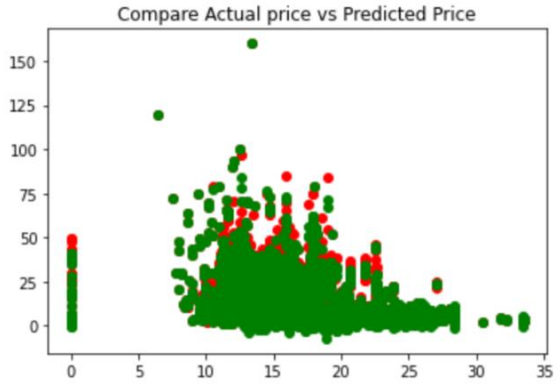
B. COST FUNCTION

The cost function is used to determine the best feasible values for a0 and a1, which can be used to obtain the most feasible fit line for the points plotted against data. Since we focus on obtaining the best values for a0 and a1, we employ this into a minimization issue through which we aim to minimise the disparity between the expected(anticipated) and the actual(truth) values.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

To minimise, we use the function mentioned above. The error difference is measured by the difference between expected and ground truth values. We square the error difference, add the data-points together, then divide the total by the number of data-points. This gives you the average squared error across all of your data-points. As a result, the Mean Squared Error (MSE) function is another name for this cost function. Now, we'll use the MSE function to adjust the values of a0 and a1 until the MSE-value reaches minima.



C. LASSO REGRESSION

Least Absolute Shrinkage and Selection Operator (LASSO) is similar to ridge regression, but it includes an absolute term as a punishment function to reduce error. The L1 form of regularisation is also known as Lasso regression. By penalising the L1 norm (manhattan distance) of the coefficient vector, Lasso regression helps to tackle overfitting by causing some of the coefficients to shrink to zero as the value of lambda grows. L1 regularisation aids in the optimal feature selection in this way. This sort of regression is best for models with a lot of multicollinearities or for automating parts of the model selection process like variable selection and parameter removal.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

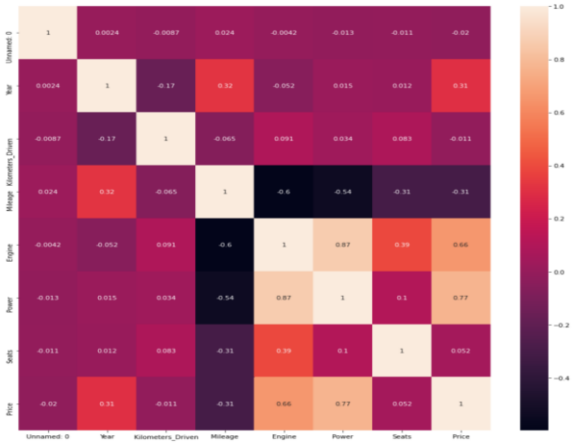
D. DATA SET

The collected data has the following fields in table II.

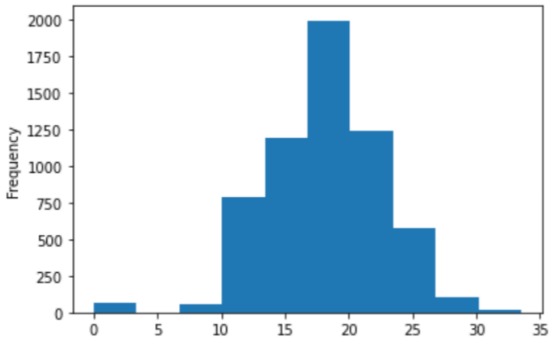
Fields	Types	Description
Name	Text	Name of Car Model
Location	Text	Car Registration City
Year	Date	Year of Manufacture
Kilometer	Number	Kilometers driven
Fuel_Type	Dropdown	E.g. Diesel, Petrol, etc.
Transmission	Dropdown	E.g. Automatic, Manual
Owner_Type	Dropdown	E.g. First, Second
Mileage	Number	Fuel efficiency
Engine	Number	Engine capacity/volume
Power	Number	Power produced
Seats	Number	Seating capacity
Price	Number	Selling price

E. ATTRIBUTE CORRELATION MATRIX

Correlation tells the relationship between the mean of two attributes. Although correlation can apply to any statistical relationship, it is most commonly used to describe the degree to which two variables are linearly related. In the data set used, attribute correlation is depicted in the diagram below.



F. DATA FREQUENCY GRAPH



IV. CONCLUSION

With the rise in auto ownership, the used automobile market is ripe for growth. The healthy development of the used car market requires an accurate used car pricing evaluation.

Since the developed system can be real-time and user friendly in terms of its handling, it is an overall unique proposal idea that is simple to implement and gives

overall customer satisfaction, proving to be a profitable business idea.

The authors of this study compared Linear Regression to Lasso Regression. The data for this study was gathered from Kaggle and then analysed using the Python programming language.

REFERENCES

[1] Doan Van Thai, "Prediction car prices using quantify qualitative data and knowledge-based system."

[2] Pattabiraman Venkatasubbu, "Used Cars Price Prediction using Supervised Learning Techniques."

[3] Nitis Monburinon, "Prediction of Prices for Used Car by Using Regression Models"

[4] https://towardsdatascience.com/used-car-price-prediction-using-machine-learninge3be02d977b2

[5] https://www.semanticscholar.org/paper/vehicle-Price-Prediction-System-using-Machine-Noor-Jan/fc87ead6754b188b1b8629db77badf361fd24a22

[6] https://www.docsity.com/en/research-project-proposal-online-car-rental-system/5232831/

[7] Comparative Analysis of Used Car Price Evaluation Models, Tongji University, Shanghai 200000, China.

[8] Nitis Monburinon, "Prediction of Prices for Used Car by Using Regression Models", 5th International Conference on Business and Industrial Research, (ICBIR), Bangkok, Thailand, 2018

[9] Jaideep A Muley, "Prediction of Used Cars' Prices by Using SAS EM", Oklahoma State University

[10] Nabarun Pal, "A methodology for predicting used cars prices using Random Forest", Future of Information and Communications Conference, 2018

[11] Kuiper, Shonda, "Introduction to Multiple Regression: How Much Is Your Car Worth?" - Journal Of Statistics Education, 2008.