Prediction of Car prices based on Used Car Prices

Akhil Reddy Kommareddy
Department of Computer Science
Eastern Kentucky University
Richmond, Kentucky
Akhilreddy_kommar@mymail.eku.edu

Sai Bharghavi Mupparaju
Department of Computer Science
Eastern Kentucky University
Richmond, Kentucky
Saibharghavi_mupp@mymail.eku.edu

Abstract*: It has been a while since cars were produced in large numbers. As a result, there has been a significant increase in the used car market, which has its own industry. Assigning a correct price to a car can be challenging when buying or selling it. To determine the value of a car, both buyers and sellers must be aware of better trends and factors. To predict the price of cars using the features of the cars in the dataset, we will use Machine Learning algorithms like Linear Regression, Multiple Linear Regression, Lasso Regression and Ridge Regression. The best model for this dataset will also be determined by comparing the prediction performance of these models.

Keywords—Linear Regression, Multiple Linear Regression, Lasso Regression, Ridge Regression.

I. Introduction

Used car sales have almost doubled in value in the past few years, making it one of the fastest-growing industries. Due to the emergence of online portals, both the buyer and seller are more aware of the trends and patterns determining the used car's value. By using machine learning algorithms, one can predict a car's retail value based on certain characteristics of used cars data.

Our inspiration for this project originated from a current issue. We wanted to purchase a brand-new automobile a few months back. We are thinking about a lot of different aspects and features at that point. Typically, a car has several attributes, like the make, model, MPG, fuel tank capacity, etc. As a result, estimating the price based on the characteristics we needed became quite challenging for us. We were aware that we wanted a reasonably priced vehicle with good MPG and horsepower. Alternatively, one of our friends is trying to sell his car but is unsure of how much to ask. Therefore, we came up with the concept to create this project that will predict the price of car. The main objective of this project is to use the prediction models to predict the retail price of a car and compare the performance of those models.

II. Literature review

Due to the recent development of online marketplaces, buyers and sellers now have access to accurate information about the variables influencing the market value of cars. Examples of machine learning algorithms include linear regression, multiple linear regression, ridge regression and lasso regression. Based on previous customer information and various vehicle parameters, we will attempt to create a statistical model that can predict the value of a car. [1] This paper aims to compare the efficiency of different models' predictions to find the appropriate one.

Previous research has been done on the topic of used car price prediction. Naive Bayes, k-nearest neighbors, multiple linear regression, and decision trees were used by Pudaruth to predict the value of used cars in Mauritius. However, since fewer cars were seen, their results weren't very predictive. In his article, Pudaruth concluded that decision trees and naive Bayes are inefficient for variables with continuous values. [2]

Noor and Jan used multiple linear regression to predict the cost of a vehicle. They employed a variable selection process to identify the factors that had the greatest influence before excluding the others. The data, which were used to build the linear regression model, contain only a small number of variables. The result was excellent, with an R-square of 98%. [2]

A study was undertaken by Peerun et al. to evaluate how well the neural network performed in forecasting used car prices. The estimated value is not very close to the true price, especially for more expensive cars. They discovered that, when predicting the cost of a used car, support vector machine regression performed somewhat better. [2]

In the digital age, a variety of methods have been employed to precisely predict the cost of an automobile, ranging from machine learning techniques like multiple linear regression, k-nearest neighbor, and naive bayes to random forest and decision trees to the SAS enterprise miner. All these solutions [3], and [4] used different sets of attributes to make predictions based on the historical data used to train the model. We tried to build a web application that would allow users to check the actual market value of their cars using a prediction model based on the variables that have the biggest effects on car prices. [5]

III. DATASET

The dataset used for the prediction models was collected from internet sources.

https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data

Price: Retail price of used cars

Symbolling: Rating corresponds to the degree to which the auto is riskier than its price indicates.

Normalized losses: the relative average loss payment per insured vehicle year.

Make: The manufacturer of the car

Fuel type: Represents the type of fuel the vehicle consumes.

Aspiration: 'breathing' engines, defines those that take in air under normal means at normal atmospheric pressures.

num-of-doors: The number of doors in the car.

body-style: Most cars are divided into 2-box and 3-box body styles, with up to four "pillars**."** The pillars refer to posts or supports around the vehicle's windows.

drive-wheels: It is the wheel and tire assembly that pushes or pulls the **vehicle** down the road.

engine location: The engine location in the car either front or rear.

Wheelbase: Ddistance between the center of the front wheels and the center of the rear wheels.

Length: represent the length of car.

Width: represent the width of car.

Height: represents the vertical distance between the ground and highest point of car.

curb-weight: Weight of the vehicle includes a full tank of fuel and all standard equipment.

engine-type: Category of the engine types like petrol or diesel.

Num of cylinders: It's a chamber where fuel is combusted and power is generated.

engine -size: The size of an engine is determined by the amount of space (volume) there is in an engine's cylinders.

Fuel system: The fuel system in a vehicle consists of a few components that help transfer fuel from the tank to the engine for combustion.

Bore: engine's bore is the diameter of each cylinder.

Stroke: stroke is the distance within the cylinder the piston travels

compression-ratio: Ratio of the volume of the cylinder and its head space.

Horsepower: It is the metric used to indicate the power produced by a car's engine.

peak-rpm: peak revolutions per minute which represents how fast any machine is operating at a given time.

city-mpg: refers to driving with occasional stopping and braking.

Highway-mpg: The average a car will get while driving on an open stretch of road without stopping or starting, typically at a higher speed.

Data preprocessing on the above is as follows. In the initial stage of preprocessing, unknown data, such as '?' are removed and replaced with NAN. Check for null values throughout the data and count how many there are in each column as the unknown items have now been replaced with NAN in the second step. There are 41 null values for "normalized losses." Replace the four null values in "stroke" with mean. "bore" has 4 nulls values, insert mean therein. "horsepower" has 2 null values replace them with mean. Replace the two null values in "peak-rpm" with mean. Replace the two missing values in "num-of-doors" with "four" as most sedans (84%) have four doors. Four doors being the most common, it is most likely to happen drop the whole row. "price": 4 missing data, simply delete the whole row Reason: the price is what we want to predict. Any data entry without price data cannot be used for prediction; therefore, any row now without price data is not useful to us.

In the next step, some of the data which has to be in numerical form is in the object form. So, convert these types of data into a numerical form such as int and float. Scale the data's length, width, and height simultaneously to create a variance that is between 0 and 1. To do this, divide each parameter by the highest value in the relevant column. The data will then be converted to decimal points. In the next steps convert some of the useful categorical data into numerical data by using PD dummies and then we will drop the categorical data which is not required for our purpose. After this just reformat the column positions and pick the data into a new CSV file.

IV. MAIN APPROACH

The major objective of this approach is to provide users with a precise estimate of the price that must be paid for the specified car. Vendors of used cars may benefit from the exponential growth of the market by setting erroneous pricing to take advantage of the demand. It is necessary to determine the relationship between the price and the dataset's independent variables to move forward. For that, factors like engine size, horsepower, and highway mpg are considered. Through random sampling, the data is divided into training (80% - 160 records) and testing (20% - 41 records) data sets.

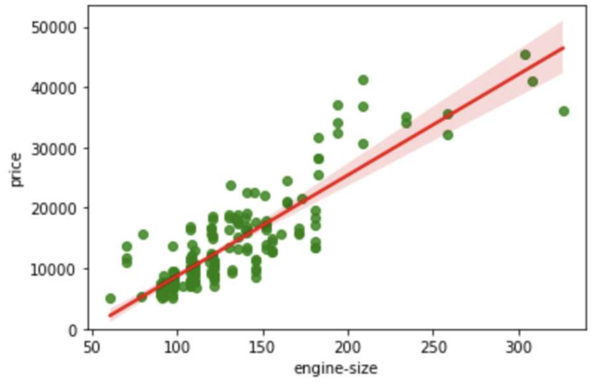

Fig 1. Engine-size VS Price

As the engine size increases in the graph above, the price also rises, demonstrating a positive direct association between these two factors. According to this finding, engine size is a reliable predictor of pricing. Similarly, it is evident from the graph below that as horsepower rises, so do prices. This allows us to also think of horsepower as a reliable price indicator.

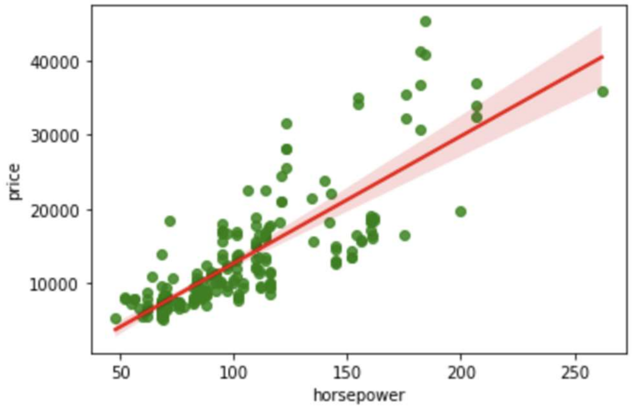

Fig 2. horsepower VS Price

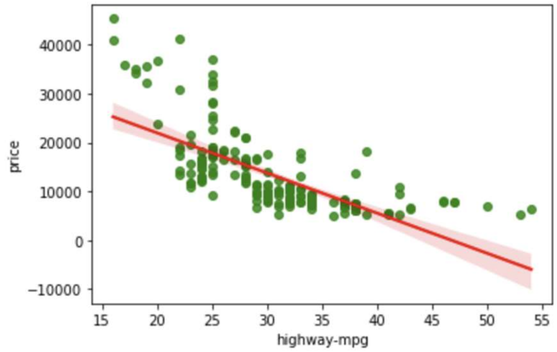

Fig 3. Highway-mpg VS Price

The graph above demonstrates that the highway-mpg and price are negatively correlated. Here, the price has an inverse relationship with the highway-mpg variable. Therefore, the highway-mpg is a reliable indicator of price.

A. *Linear Regression*

The regression model determines the linear or exponential relationship between independent and dependent variables. The line can be represented by the linear equation given below.

y = a0 + a1 * x

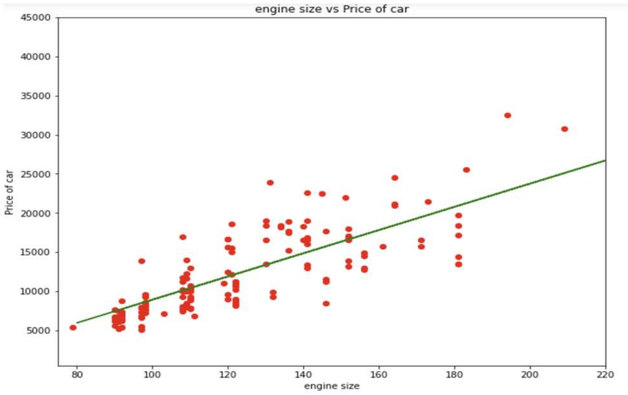

Fig 4. Engine-size VS Price Prediction

B. *Multiple Linear Regression*

Like linear regression, multiple regression attempts to predict a value based on two or more factors, but with more than one independent value. Consider the following as the independent variables for the model in the first step: horsepower, wheelbase, and engine size. The regression model's price predictions are considered.

The full training dataset will be used as the input for the regression in the next stage, and the result of the anticipated price is considered.

C. *Ridge Regression*

When there are several highly correlated variables, ridge regression is utilized. By penalizing the variable coefficients, it helps avoid overfitting. By incorporating a penalty term to the error function that limits the size of the coefficients, ridge regression lessens overfitting. To avoid the overfitting and model complexity from the linear regression model, the ridge regression is considered.

$$L_{hridge}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j-1}^{m} w_j\hat{\beta}_j^2.$$

The training dataset is used to train the model and the test dataset are considered to get the relatively accurate predicted prices

D. Lasso Regression

In contrast to ridge regression, least absolute shrinkage and selection operator (LASSO) incorporates an absolute term as a punishment function. Regression of this type works well for models with several multicollinearities or for automating steps in the model selection process, such as parameter removal and variable selection.

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

For this regression, the model is fitted with the training dataset of used cars. Later the predicted price results are considered from the trained model.

V. Evaluation metric

The metrics used for evaluating this project or R-squared value and the mean squared value.

R-Squared Value: The *coefficient of determination* is a measure that provides information about the goodness of fit of a model. In the context of regression, it is a statistical measure of how well the regression line approximates the actual data. It is therefore important when a statistical model is used either to predict future outcomes or in the testing of hypotheses [6].

$$\begin{aligned} R^2 &= 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}, \\ &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}. \end{aligned}$$

Mean Squared Value: Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD). [7]

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

VI. Results & analysis

In the linear regression model, the Engine size versus price has yielded an R-squared value of 0.724 and a Mean squared value of 33696986 whereas the linear regression with the horsepower versus prize has yielded an R-squared value of 0.623 and a Mean squared value of 46089111. with multiple linear regression now when 4 parameters are passed the R-squared value is 0.745 and the Mean squared value is 31085980 with the multilinear regression. When entire columns of the training data are passed into the model we got the R-squared as 0.803 and the Mean squared value of 31085980. Similarly, the ridge regression R-squared value of 0.802 and the Mean squared value of 24220389 were shown. At the same time in lasso regression with the whole data set, we got the R-squared value of 0.802, and the Mean squared value of 24176093.

In the linear regression when the single parameter is passed then the R-squared value is low this is happening because the price value is not dependent upon the single variable of the data. So to get the best price value we need to consider different parameters at the same time and coming to the multilinear regression models every model has given approximately near the same R squared values but the mean squared value of lasso regression is low so we can say that lasso regression has performed well out of all the models.

VII. Future work

In this project, the prediction of the prices of cars is based on the used car data. For now, only numerical data is considered but in the future, categorical data can also be considered to have the classification analysis on the data. The categorical values of the data set are like make model style etc.. by considering these factors one can provide a better & accurate prediction of price, and with help of classification algorithms, this can be achieved easily. In the future, these models will help to develop a complete end-to-end application that will help stakeholders in the car market. The buyers and sellers in the car market can interact with that application and can provide the required features of the car. The application will provide the predicted price of the car based on the inputs from the stakeholder. These machine learning algorithms can provide a better design and patterns to determine the prices of cars and one can also easily add a graphical user interface to have a better experience while developing this application and predicting the accurate information of this show in car prices of the cars.

VIII. References

[1] M. G. Pattabiraman Venkatasubbu. [Online]. Available: https://www.ijeat.org/wp-content/uploads/papers/v9i1s3/A10421291S319.pdf.

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

[3] L. H. a. C. X. Chuancan Chena. [Online]. Available: https://aip.scitation.org/doi/10.1063/1.4982530.

[4] N. Monburinon, "Prediction of Prices for Used Car by Using Regression Models".

[5] K. Agrahari, "Car Price Prediction Using Machine Learning," [Online]. Available: https://ijirt.org/master/publishedpaper/IJIRT151705_PAPER.pdf.

[6] https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html

[7] M. S. Value. [Online]. Available: https://statisticsbyjim.com/regression/mean-squared-error-mse/.

Team contribution

Since it is a team project, we decided to split our responsibility towards it. The below mentioned tasks are our contribution for this project. Both of us involved in giving the presentation for the project because that is the best way to showcase our individual efforts in the project.

Akhil Reddy Kommareddy:

Dataset collection from the internet source.

Worked on the data preprocessing of the dataset.

Analyzed the dataset with two regression models (Multiple and Ridge) and observed the predicted results.

Involved in dataset section and the results section of the final report.

Sai Bharghavi Mupparaju:

Presented the original idea for the project.

Worked on the exploratory analysis of the variables for linear regression model.

Worked on analysis of Lasso Regression model.

Involved in the literature survey and main approach section of the final report.