Abstract:

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But, due to the increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and its value in the present day scenario. In fact, the seller also has no idea about the car's existing value or the price he should be selling the car at. To overcome this problem we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value. Because of which it will be possible to predict the actual price of a car rather than the price range of a car.

1. Introduction

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing simple linear regression models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models . We will use learning algorithms like Linear Regression, Depending on various parameters we will determine the price of the car. Regression algorithms are used because they provide us with continuous value as an output and not a categorized value because of which it will be possible to predict the actual price of a car rather than the price range of a car.

2. Method

For this project, we used the dataset on used cars from all over India available on Kaggle [1]. The data set comprises 301 examples with different units like car name, year, transmission, fuel type, kilometers driven, selling price, present price, owner and selling type.

The data for this dataset is incorporated by using SRS(Simple Random Sampling). As this was a dataset which was a combination of some other used cars datasets available in kaggle. Please refer to figure 2.1 for better understanding of SRS.

There are two primary phases in the Model: 1. Training phase: The model is trained by using the data in the data set and fits a model (line) based on the algorithm chosen accordingly. 2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked accordingly on the basis of the results incurred from the study.

Upon the pre study we have notified some of the potential variables which were impacting the results and also important for this study. They were defined clearly in table 1.

Variables	Description
Kms_driven	The no of kilometers driven are shown by this integer
Selling Price	The selling price of a used car in lakhs is shown by this integer.
Transmission	It shows us the type of the car whether it is manual or automatic.
Year	This integer shows the year of the car manufactured on.
Car name	This variable measures the model of the car like its brand.

Table . 1

2.1 Model for the test

Linear Regression attempts to model the relationship between two variables by fitting a linear equation to observed data. The other is considered to be a dependent variable. We are going to find the relationship between **Kilometers driven** and **Selling Price** and see how a unit change in kilometers driven is changing the selling price of the car. Refer to image 2.1.1 for SLR model. We used R programming language throughout the study

The equation for the Simple Linear Regression is

$y = \beta0 + \beta1x + \varepsilon$

where y = Selling Price, x = kilometers driven

β0 = intercept coefficient,

β1 = Kilometers driven coefficient

ε is the standard error and is i.i.d.

$\varepsilon i \sim N(0, \sigma2)$.

3. Results

The preprocessing of data went through some summary statistics of the dataset for our response variable, predictor of interest and some of our confounders which were clearly shown in the below table 2.

Variables	Mean(sd) or n
Kms_driven	4.66 lakhs (5.08)
Selling_price	36947.21km (38886.88)
Fuel Type	C = 2, D= 60, P = 239
Transmission	A = 35 M = 258

Table 2

As part of checking for the model fit we plotted the relationship between our predictor of interest and response variable which was shown in the figure 3.1. The data was scattered good enough to perform the model hence continued the model after preprocessing though there were some outliers which were answered in our next step of analysis.

As part of the study the main step which we needed to perform was the condition check(LINE), and upon testing the preprocessed data it was clear that the data had passed all the conditions for which the visualization was shown in the figure 3.2-3.4 with well labeled.

3.2 Model Computation

The simple linear regression tries to find the best line to predict selling price on the basis of youtube advertising budget.

The linear model equation can be written as follow: selling price = b0 + b1 * kms driven.The R function lm() can be used to determine the beta coefficients of the linear model:

Null Hypothesis H0: $\beta 0 = 0$
Alternate Hypothesis H1: $\beta 0 < 0$

As we were interested in knowing the negative relation (as kilometers increase the price decreases) we defined the model accordingly. The results of the test are shown in the table 3.From the output of our model it is noted the estimated regression line equation can be written as follow: $\widehat{selling\ Price}$=4.52+0.000003815*kms driven.

t-stat	p-value	CI-low	CI-High
0.505	0.614	-0.0000114	.87e-05

Table 3: For X

The values were changed after removing the extreme outlier of kilometers driven(Removed all the values which were above 1 lakh). The results were shown in the below table 4.

t-stat	p-value	CI-low	CI-High
2.11	0.0358	1.92e-06	5.586e-05

Table 4 : For X

So from the above results we were 95% confident that the estimated value of y will be in between our CI low and CI high for both the cases.

4. Discussions

So based on the results from our model if we need to estimate the selling price at 14590 kms we can get it by substituting in line equation

$\widehat{selling\ Price}$=4.52+0.000003815*14590.

From above we can clearly say that the car will be sold at 4.58 Lakhs. But we want to find the negative relationship and the model shows us the positive relationship between x and y.

So we don't have enough evidence to portray that the model is accurate. And it is noted for every unit change of kms driven there is 0.000003815 change in selling price of used car.After extreme outlier : for every unit change of kms driven there is .0002889 change in selling price.

4.1 Limitations and Future scope

From the deep dive observation of the model summary we note some significant limitations for the model. The potentials which we observed were the confounders and they are the year of car's name (the brand), year of manufacturing and fuel type. low correlation ($-0.2 < x < 0.2$) probably suggests that much of variation of the outcome variable (y) is not explained by the predictor (x)

So for fixing them and to get the desired and accurate result we want to continue the study further. So to fix it we want to build an MLR model and split (train and test) model for accurate results. And also if needed we

are interested to build a deep learning model and work on it for the best fit model.

Conclusion: The model has worked on a good note but it is not yielding us the desired outcome which we were looking for, but we are confident that we can fix it by continuing the study.

5. References

1.https://www.kaggle.com/code/vijayaadithyanvg/car-price-prediction-used-cars/data

2. Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, "Prediction of Prices for Used Car by using Regression Models" (ICBIR 2018)

3.http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/#formula-and-basics

6. Appendix

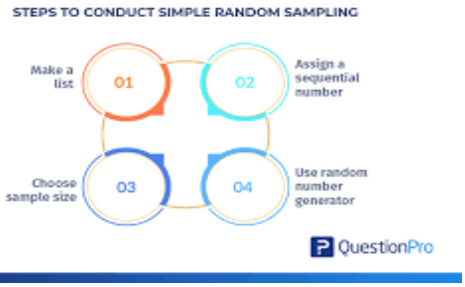

Figure: 2.1

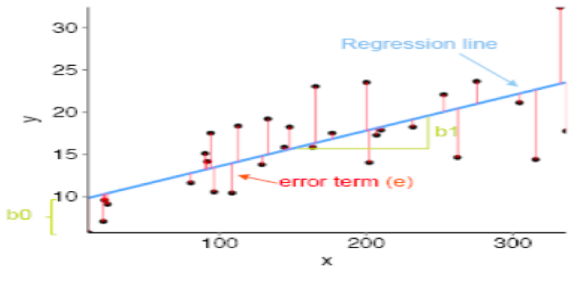

Figure: 2.1.1

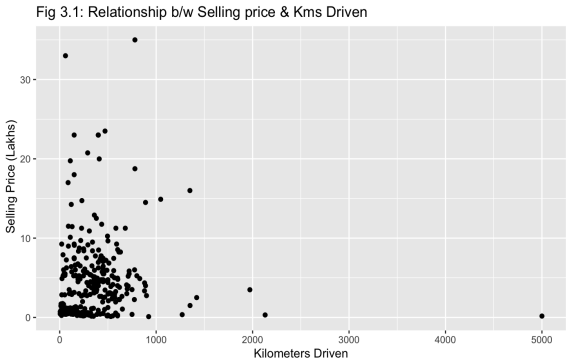

Figure 3.1

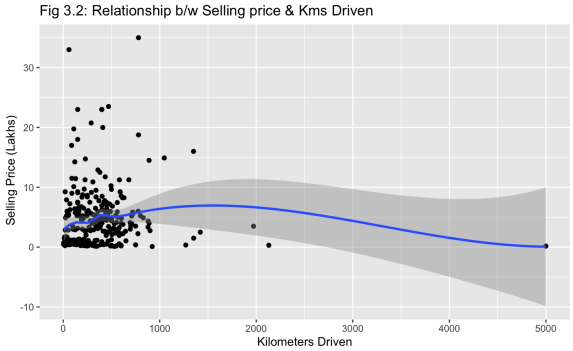

Fig : 3.2: Linearity

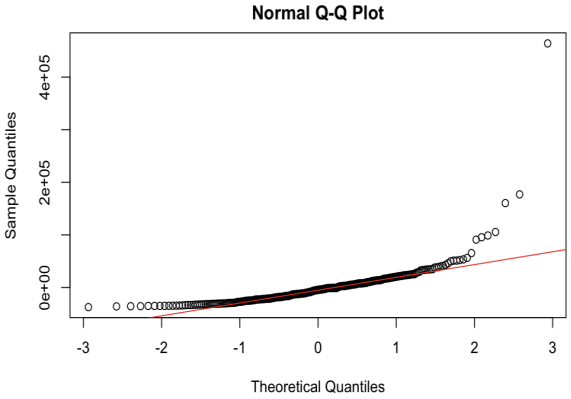

Fig : 3.3 Normality

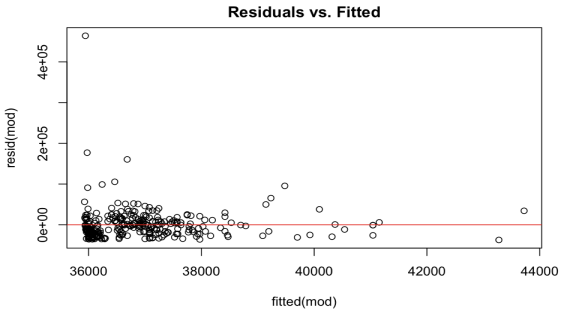

Fig : 3.4 Equal Variance