Used Car Dataset: Prediction model report

Mark Movh

February 2022

Abstract

This regression analysis report looks into a used car dataset from 2018. It investigates how a prediction model may be created to train, test, analyse and predict the selling price of a car. Firstly, linear models are explored comparing the Original Price, and Kilometers Driven individually against Selling Price. The data was split on 70% training, and 30% testing. The first model with Original Price as the independent variable showed a Mean Square Error (MSE) of 2.51 and correlation of determination (r^2) of 78.03% on the final model. The second model with Kilometers Driven, showed much lower results, with the final model having an MSE of 5.92 and r^2 of 0.58%. Combining the variables presented a multilinear model, with Selling Price against both Original Price and Kilometers Driven. This model presented the most favourable results as the final model displayed an MSE of 1.809 and and r^2 of 89.14%.

Introduction

Owning and having access to a personal car can be extremely useful, whether it's needed for more official business such as work or school, to more leisure activities. However, how efficient cars are in terms of getting a person from point A to point B differs between models, and especially between other more specific mechanical aspects. Cars may either be bought brand-new, which isn't always necessarily the best choice. Used cars are an alternative, often being cheaper and more efficient price-wise. Considering a car can be fairly used or barely touched, it can be important to measure these differences and how they overall, impact the worth of the car. This report explores a real-world dataset which focuses on the statistics regarding prices in used cars.

The dataset explored is collected from CarDekho's website (Cardekho, 2018), which is centred on used card sales. Certain variables will be required to be picked from the dataset in order to determine how the selling price of a car may be measured. The dataset's contents range from descriptive (name of car, year), numerical (selling prices, original prices, distance driven) to technical factors (fuel type, seller type, transmission type, previous owners). Hence, there are various attributes that fall under the category of continuous and discrete. This statistical report investigates how the current selling price of a car, may be predicted and accurately determined through the use of such continuous variables. A regression model is to be created, to perform regression analysis.

Data Analysis

In this section the variables that are to be chosen are gone over, manipulated to gain further insight and why they were chosen at the end. For the regression analysis and model, there must be a dependent variable which is an output as a result of the independent variable. Meaning that the independent variable will be changing. In this statistical report, regardless of which independent variable is chosen, the aim is to predict the selling price of a car. Therefore, the independent variable (to be predicted) is the selling price. For the independent variable, the attributes needed to be narrowed down.

Only continuous values were required. This included the original price, and the kilometres driven. Considering the few pieces of data, it was further explored how these can be combined or manipulated in different ways to gain further insight into the relationship between them and the selling price. Firstly, the difference between all selling prices and original prices were collected, and mapped against selling price. This would be known as the change in price. Additionally, this change in price was further divided by the original price to see the price change rate. It was thought that there could be some existing pattern between the selling price and this change in price rate. A series of scatterplots was created for each of these variables being put up against the selling price. This way a relationship or correlation could already be visually confirmed, and could then be used for the regression model. See scatterplots below.

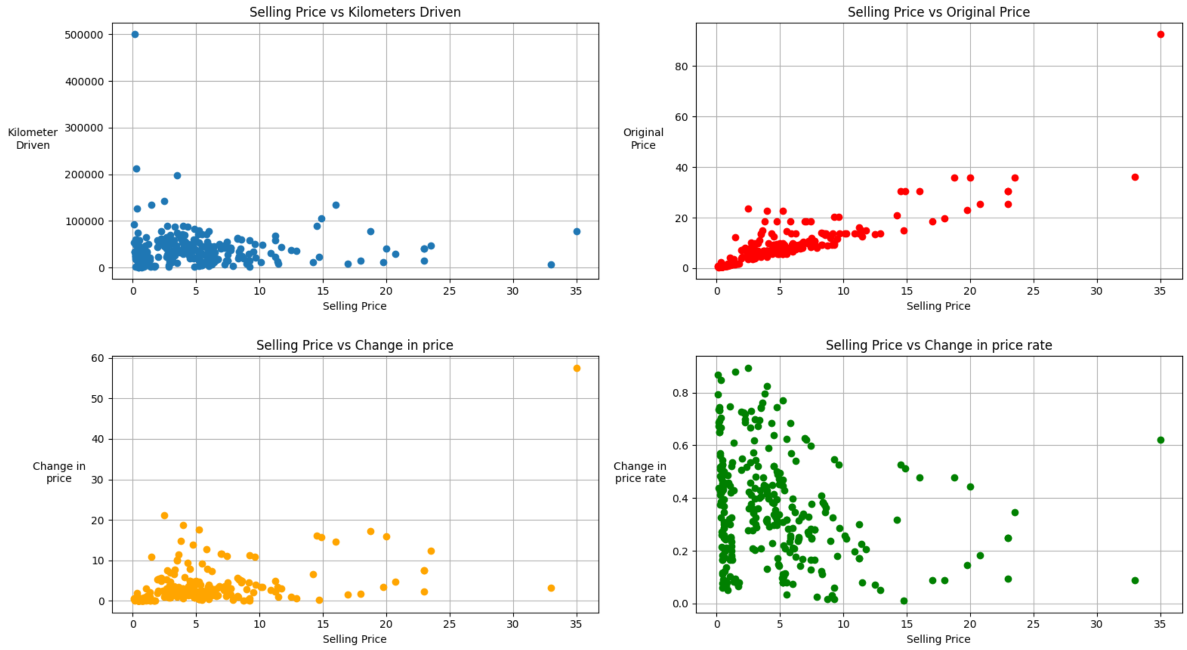

Figure 1: Scatterplots showing various possible independent variables against selling price of a used car

There are some visible outliers, however given the abundance of data these were considered not crucial to impacting the correlation. For example, in the case of "Selling Price vs Kilometers Driven", removing all values with Kilometers Driven over 200000 resulted in only 0.033 difference, which isn't that large of a difference. From this data it can be seen that there exists at least some relationship between the independent variables and the selling prices. Ultimately, it was decided to remove the additional information collected from the change in price, and the change in price rate, as it would be an unfair to use these to train a regression model because they already have some information on the selling price. The training of the dataset in turn would not be accurate. Taking this into account, two independent variables were chosen for the linear regression model. These were the kilometers driven (correlation of 0.146), and the original price (correlation of 0.908). Through this, it would be possible to examine how these two independent variables can influence the selling price, individually, and while combined.

Building the Regression Model

Linear/Non-Linear Models

As there were two variables selected, original price and kilometers as independent variables, these would be used to create different regression models. From these, it would be possible to then see how predictable the selling price of a used car is, and the models accuracy. In this report, the building of a model can vary, between linear and polynomial. In this case, the implication of linear and nonlinear imply that only one variable is used, but may still be subject to becoming polynomial. Transformations may be required to get more accurate results. The first step of this, is to select the appropriate degree for the final models.

The process to determining this consists of using a training model to see the accuracy of the relationship, and the accuracy of the line fitted with the graph. The methodology involves splitting the data into two categories; training and testing. With 30% of the data going to testing, and 70% to training the model. This split would be used to determine the accuracy of the final model. However, for the training model the training data was further split. Another set of testing and training data was created, with 20% going to testing, and 80% to training. See figure 2 for visual explanation. The reason for the higher choice of data in the training set T is because it was thought it would be more ideal to better train the training model, to select the right degree of k.

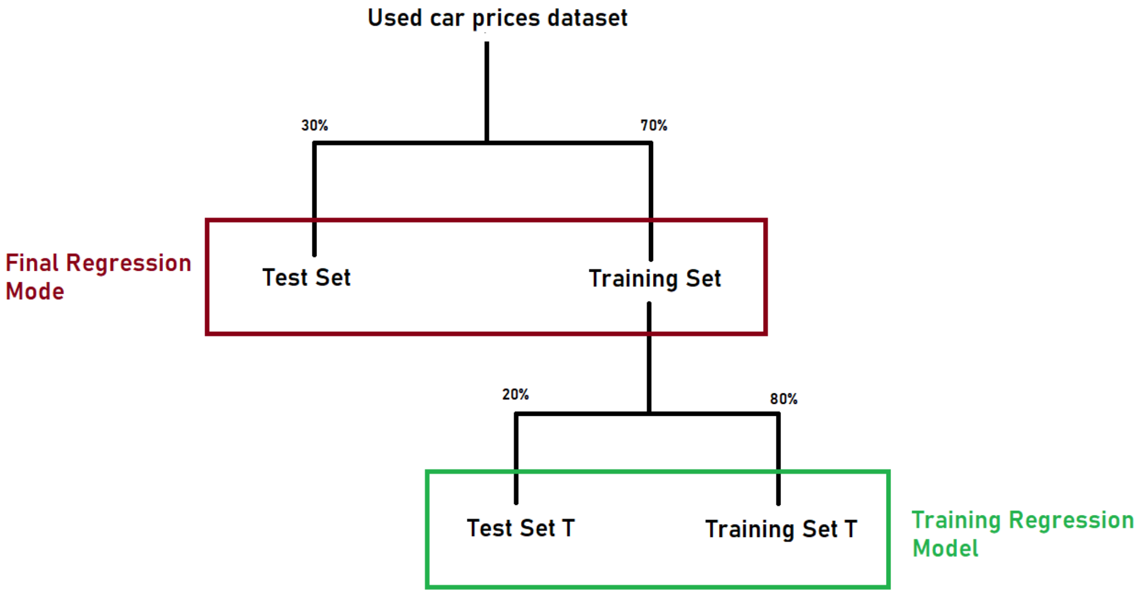

Figure 2: Tree Diagram showing split of dataset, and what data the two different models use

The degree of k in this case, would be to what degree the best fit line would need to be transformed to attain the most ideal results. In the methodology of this report, both the Mean

Square Error (MSE) and the Correlation of Determination (r^2) are looked at. The degree is synonymous with a variable's exponent; the higher the number, the more information is required and hence generating more complex models. The report hopes to avoid such scenarios, even if there are decreases in MSE or increase in r^2. Taking this into account, MSE and r^2 were gathered for all degrees up to 10. The first case where this was done was between the original price and the sales price. The results can be visible in figure 3.

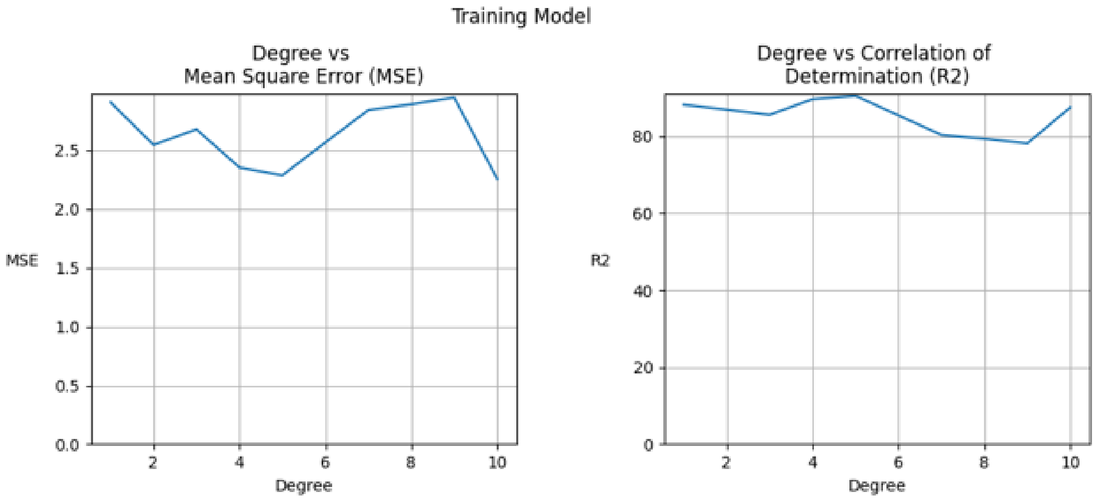

Figure 3: Mean Square Error and Correlation of Determination of the training model for Selling Price vs Original Price

What was visible from the results is that, as the complexity of the model increase, through the increase in degree, there is not much change in terms of r^2. Additionally, there was some decrease in the MSE, but otherwise not many big differences. Using the training model, it was decided that a degree of 2 would be used for the final model in this case, as it favoured the decrease in MSE from degree of 1, while avoiding complexity at degree 5. Which offered the best results, but not a large enough difference to choose a more complex model. This process was carried out for the case where kilometers driven were compared against selling price. Data is visible in the graph below.

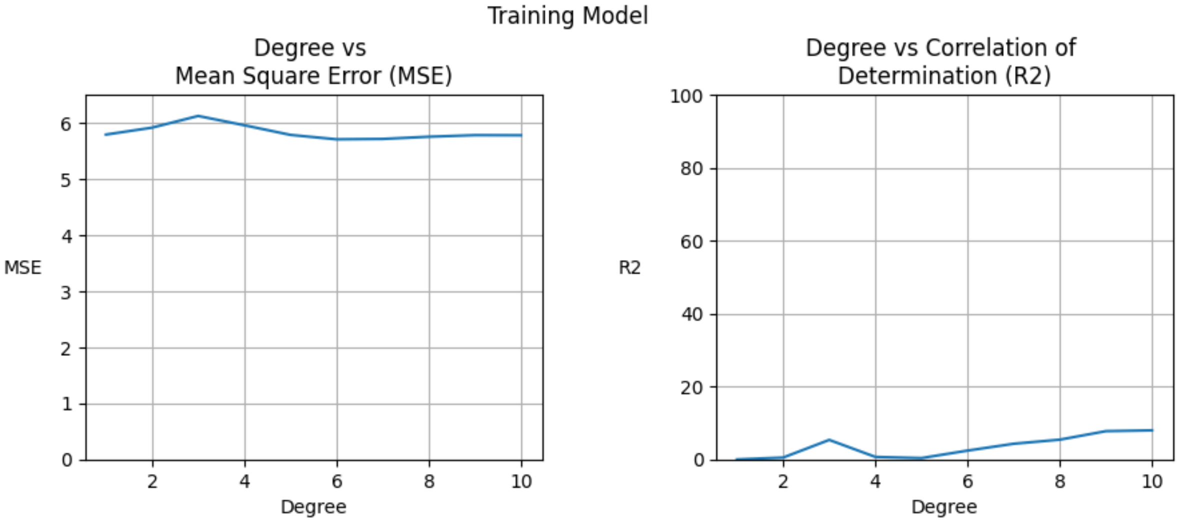

Figure 4: Mean Square Error and Correlation of Determination of the training model for Selling Price vs Kilometers Driven

Compared to the original price, the kilometers driven variable did not seem viable to be used to predict the price of a used car. The graphs show that the MSE is always at a high point, while the r^2 was always extremely low. The highest correlation of determination value was roughly 8%, this being at a degree of 10, which would be a very complex model that didn't show enough improvement to be selected. Considering the low differences between 2 and values above, the degree of 2 was selected for the final model again. Although with such results it can already be stated that there is sufficient evidence to support the idea that kilometers driven would not be an appropriate variable alone for regression analysis.

Taking these results into account, the final models would each use a degree of 2, as it was deemed appropriate for the best fit line. The same process would be carried out as before, however this time in the final model, the entire training set (70%) would be used, against the original testing set (30%). From this, new values of the MSE and r^2 can be found, and final conclusions may be drawn from the data. The process mentioned was carried out for both the original price and the kilometers driven as independent variables. The contents of the results, may be found in Table 1.

Variables			Training Model		Final Model	
Dependent Variable	Independent Variable	Degree of k	MSE	R^2	MSE	R^2
Selling Point	Original Price	2	2.55	86.86	2.510	78.03
	Kilometres driven	2	5.92	0.58	5.360	0.14

Table 1: Comparison of training models against final models in terms of MSE and r^2

Table 1 takes both independent variables separately, and notes down the results of the training models and the final models. Using the table, comparisons may be highlighted. Such as the change in the r^2 values. Interestingly, in each scenario the correlation of determination decreased, despite using the same degree of k. Perhaps in this scenario, it appears that the model may be overfitting and causing such negative impacts. This however, is not supported by the difference in MSE values. With the MSE, it is seen decreasing instead implying that the final model does contain some good fit. If such a linear regression analysis was to be continued, it could be a good idea to implement certain parameters that would limit how much the model can learn to avoid overfitting. Overall, it is still demonstrated that the Original Price variable shows favourable results, with the correlation of determination still being at 78%. The data found supports the premise that this independent variable may be used for regression analysis. Contrary, the r^2 and MSE values were too low and too high respectively, to deem it an accurate tool for the measurement of used car prices. However, what could potentially improve the model for original price, is to combine it with the kilometers driven, to create a multilinear model.

Multilinear Model

So far linear models had been attempted to go through regression analysis. This had some accuracy to predicting car sales prices, however it was deemed interesting to see how the two variables would work if they were combined. In this case, a multilinear model approach would be taken, taking into account both the original price, and the kilometers driven. The exact same approach was taken as for that of the linear regression models. A training model was first developed, on an 80%-20% (training-testing) split allowing for the appropriate degree of k to be selected. The figure below shows how the MSE and r^2 change over the course of degree of k. Same process described in Figure 2.

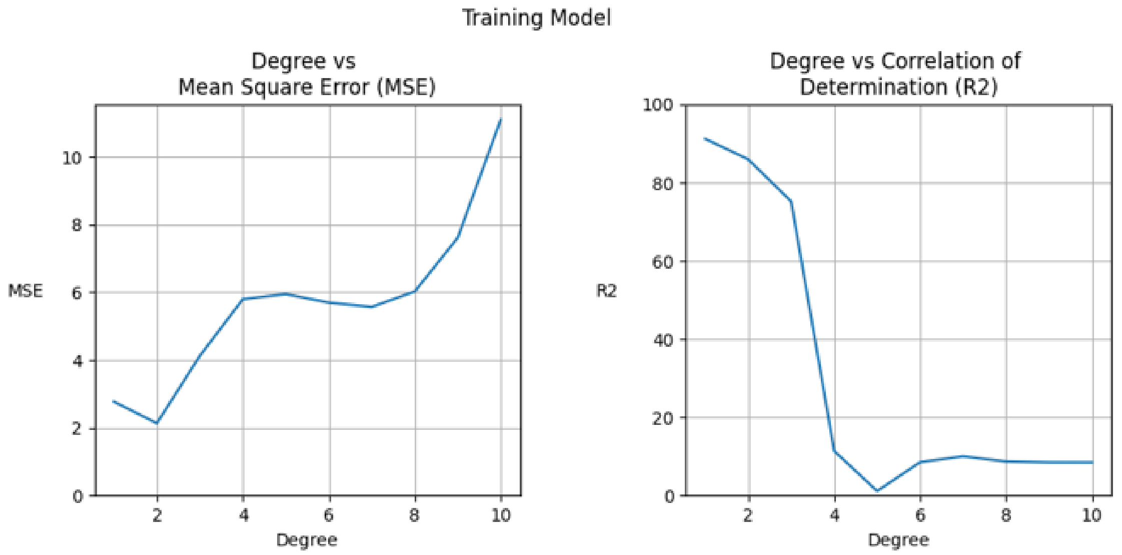

Figure 5: Mean Square Error and Correlation of Determination of the multilinear training model (Selling Price vs Original Price and Kilometers Driven)

Generally, what can be taken from the graphs is that the higher the degree of k is, the less wanted the results become. Essentially stating, that creating a more complex model creates less accurate results. At least this is what can be gathered from the training model. The degree of k was chosen once more to be set at 2. The reason being the lowest value in MSE (2.137), while also attaining a very high r^2 (86%). It met the requirements for a satisfactory final model. The final model was set to a degree of 2, with training on 70% of the total data, and using 30% for testing. Once more the results are recorded in another table; see table 2 below.

Variables			Training Model		Final Model	
Dependent Variable	Independent Variables	Degree of k	MSE	R2	MSE	R2
Selling Point	Original Price AND Kilometres driven	2	2.137	86	1.809	89.14

Table 2: Training Model compared against Final Model for Multilinear Regression

As was stated, the variables were combined, meaning that there is only one model, combining the two linear models and creating the multilinear regression model. In terms of regression analysis, it is possible to see that the training model's MSE and r^2 are either lower (2.137 down from 2.55) or extremely similar (86% against 86.86%) to that of the Original Price model. This alone shows the promising nature of this regression model. It was found that in the final model, the Mean Square Error was further decreased to 1.809, which is a decrease of 15.3%. This indicating that the best fit line of the final model fitted better (closer to the data points) than that of the training model. Further confirming the fact, r^2 also increased by 3% in the final model, leaving the final correlation of determination at 89.14%. In essence, it was shown that the combination of the two linear models lead to a more refined, and accurate regression model, which would be more suitable for regression analysis, and in the predicting of user car sales price.

Limitations

Such results may be open to limitations. Especially in terms of the data collection, and whether the values are entirely accurate themselves. Firstly, neither standardization or normalization was taken into account, which could have created the issues encountered. Several complications were stumbled upon throughout the process of creating the various models. The correlation of determination would often create unplausible values, as such confusion arose. As such, the methodology for calculating r^2 was changed, and lead to more ideal results. However, some values are still questionable. Such as the steep decline found in Figure 5. Such a tremendous decrease could have occurred because the data input was not rescaled into normal distributions. Standardization or normalization should have occurred to output more reliable results and more concrete predictions. Additionally, although outputs were mentioned, it would still be sufficed to remove noise to better prepare the model. Hence the conclusive remarks may not be entirely precise.

Conclusion

To conclude, this report presented a regression model which was then used for regression analysis in order to predict some variable. This variable being the selling price from the used car dataset. The process was to narrow down the useful data from the dataset, and omit any that weren't required. This left three continuous variables, which were used to first create two linear regression models. One for Original Price vs Selling Price, and another for Kilometers Driven vs Selling Price. Furthermore, training and final models were generated to first determine what degree of k was the right choice, and then the mean squared error and the correlation of determination were found. While the second didn't show any promising results, the Original Price model was deemed efficient enough to be able to predict correct results. To further inspect the prediction capabilities, and want for improvement, the two models were combined to create a multilinear regression model. The training model for this regression was shown to be as accurate as the Original Price model, while the final multilinear model had the highest correlation of determination value, and the lowest mean squared error. Taking into account the linear model consisting of just the Original Price as the independent variable, and the multilinear model taking into account both variables, it was clearly demonstrated that the selling price of a used car may be predicted accurately to roughly 90% through such regression analysis.

References

CarDekho. (2018). Car Price Dataset. Retrieved February 19, 2022, from https://www.cardekho.com/