PREDICTING SELLING PRICE OF USED CAR

STAT 614 Final Project Report

By
Praveen Chandrasekaran

DECEMBER 12, 2021

ROCHESTER INSTITUTE OF TECHNOLOGY
Rochester, New York

Table of Contents

pc2846@rit.edu

STAT 614 Project Report: Predicting Selling Price of Used Car

Praveen Chandrasekaran <pc2846@rit.edu>

1.0 Introduction

The used cars market has grown rapidly in recent years and expected to grow more over the next five year in the United States of America. The increasing demand for used cars has encouraged the industry to collect enormous amounts of data. This project uses such a dataset for analysing, understanding and interpreting relationships between variables and the how much value is being added by each variable in building a tool that can predict the selling price of a used vehicle. Statistical tests performed on this data showed that Linear Regression is the best suited model for prediction.

2.0 Objective

miles covered by the used car, make, model, year of manufacture, engine specification, etc. can influence the actual worth of the used car. From the perspective of the seller, it has always been a dilemma to price a used car appropriately. Based on the selected dataset, the main objective of this project is to use an appropriate statistical model to predict the selling price of a used car based on 4 regressors.

3.0 Dataset

For this project, the original dataset is sourced from a user profile on Github. This dataset includes 6 variables and 4,006 observations, but this was reduced to 5 variables for reducing the complexity. Out of 5 variables, 4 are predictors and they are brand, model year, engine capacity, miles driven. The response variable is the selling price. A random sample of 22.5% was taken from the original dataset which resulted in 901 observations in total. After which, outliers and missing data were removed the get the final dataset of 647 data points. All input variables are continuous type except for two, one is categorical nominal and another one is quantitative discrete.

4.0 Data Pre-processing

Histograms and box plots were plotted in order to get a better understanding of the data. Observation showed that the dataset had many outliers due to the price sensitivity of used cars. Typically, cars of latest models and that have low mileage sell for higher price, but there were many observations that didn't conform to this. Vehicle damage history and engine conditions affect this price to a larger extent. Since we do not have any data regarding the history of the vehicle, outliers were removed by setting a lower and upper limit using the IQR. Outliers are shown in Fig. 1 (Left)

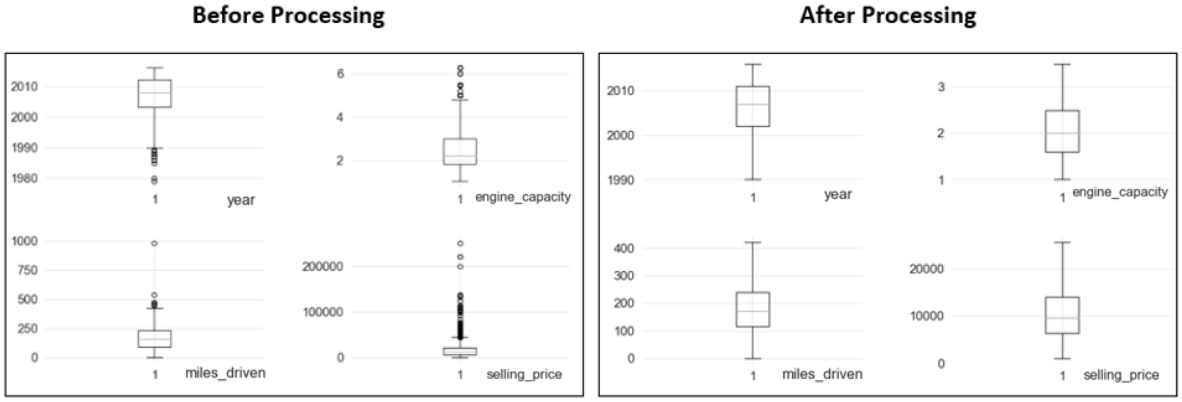

Fig 1: (Left) Box plot for raw data. (Right) Box plot after removing outliers using IQR

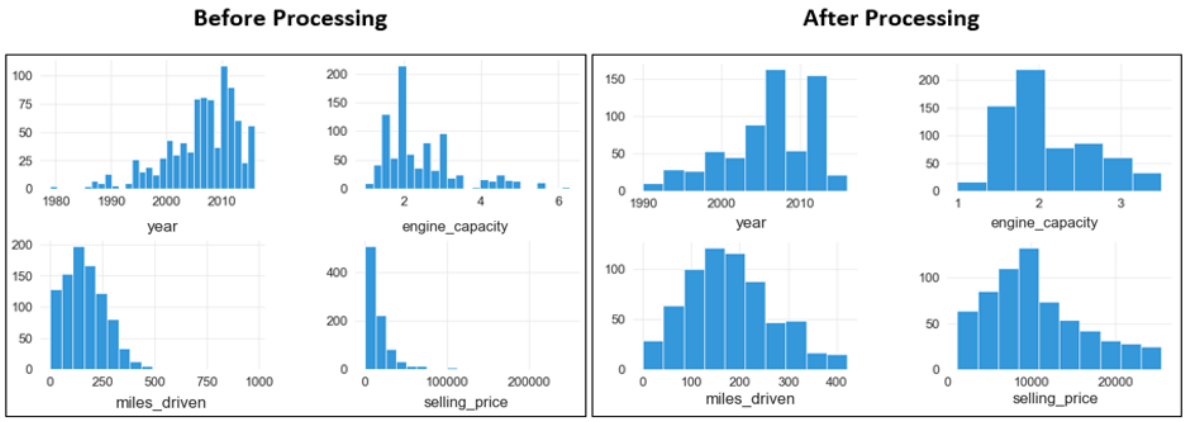

Fig 2: (Left) Histogram for raw data. (Right) Histogram after removing outliers using IQR

Interpretation: After removing outliers, high skewness observed in the Fig 2 (Left) histograms were reduced and resulted in Fig 2. (Right) histograms. However, Year, miles driven, engine capacity and selling price were skewed left, right, right, left respectively. Removal of outliers helps in verifying Regression assumptions better.

5.0 Analyses of Linearity

In order to analyse the degree to which the numeric features are linearly related to price, scatter plot and regression line were plotted. There seemed to be a fair degree of linearity for these 3 features vs price. The visuals provoked a thought that there could be relationships between the input and response variables respectively. The basic regression assumptions were verified as a next step.

Scatter Plots with Regression Line

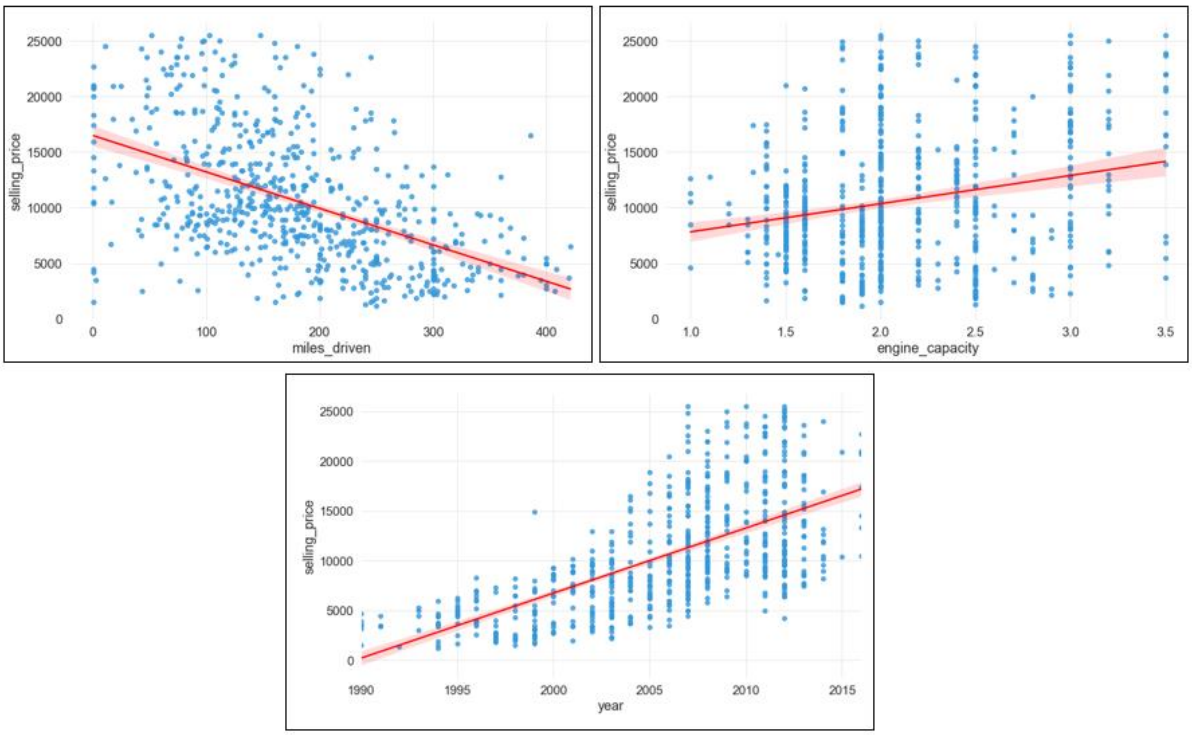

Fig 3: (Top left) Miles driven vs Price. (Top right) Engine capacity vs Price. (Bottom) Year vs Price

Interpretation from Fig 3: Considerable slopes are observed in Fig 3: (Top left) Miles driven vs Price and Fig 3: (Bottom) Year vs Price; this shows that there is linear relationship. Slope of engine capacity vs Price is subtle but it is sensible to fit a regression line, however this observation has to be confirmed statistically by conducting more tests.

6.0 Verification of Assumptions

The assumptions of Linear Regression were verified by checking the residual by predicted plot and residual normal quantile plot, considering all 4 features as input variables.

Residual by Predicted Plot **Normal Quantile Plot**

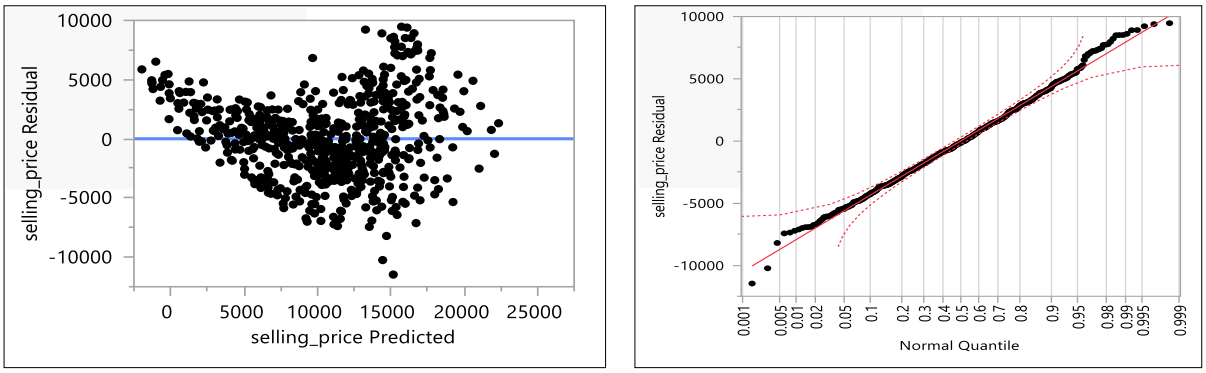

Fig 4: (Left) Selling Price Residual vs Predicted. (Right) Normal Quantile plot of Selling Price Residual

Interpretation from Fig 4: In Fig 4: (Left) the residuals are scattered but not perfectly random, if at all there is a pattern observed it is not very serious. The assumption of constant or homogeneous variance does not appear to be violated. In Fig 4: (Right) the residuals form a straight line in the Normal Quantile Plot. The residuals do not appear to violate the assumption that they are normally distributed. This allows us to fit the model and perform hypothesis testing as the next step.

7.0 Hypothesis Test

- H_0: $\beta 1$ (vehicle_brand [Audi]) = $\beta 2$ (vehicle_brand [BMW]) = $\beta 3$ (vehicle_brand [Mercedes-Benz]) = $\beta 4$ (vehicle_brand [Mitsubishi]) = $\beta 5$ (vehicle_brand [Renault]) = $\beta 7$ (year) = $\beta 8$ (engine_capacity]) = $\beta 9$ (miles_driven)
- H_1: $\beta j \neq 0$ for at least one j
- Level of significance (α) is taken as 0.05
- F_0 = 138.0793 (F Ratio value taken from the Table 1)
- P-value = 2* ($F_{0.05}$, 636 > 138.0793) = < 0.0001* (P-value taken from Table 1)
- Since 0.0001* is too small and lesser than the level of significance (α) 0.05, therefore, we reject H_0.
- Decision for hypothesis: Null hypothesis rejected
- Conclusion for hypothesis: At 5 % level of significance, we have sufficient evidence to conclude that there is a linear relationship between the input variables (vehicle brand, year, engine capacity, miles driven) and the output variable (price). The model fits the data well.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	1.4385e+10	1.5983e+9	138.0793
Error	636	7362011106	11575489	**Prob > F**
C. Total	645	2.1747e+10		<.0001*

Table 1: F_0 and P-value are used for hypothesis test

Interpretation from Table 1: The hypothesis test concludes that we reject null hypothesis using the value from the Table 1. The large F_0 value shows that the model explains the variation of price that is observed. P-value of <.0001* is small enough to indicate convincing significance. This implies that the model as a whole does a good job of fitting with the dataset. P-values are probability of getting an even more extreme statistic given the true value being tested is at hypothesized value, usually at 0.

8.0 Implications about all β values using t-test

Parameter estimates table is constructed to interpret the impact that each variable has on the selling price. The P-values and for each variable is observed and compared with the level of significance 0.05 that we have considered for out test. The parameter estimate table is shown as Table 2.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1373988	68880.16	-19.95	<.0001*
vehicle_brand [Audi]	1200.8166	407.9403	2.94	0.0034*
vehicle_brand[BMW]	1105.7824	392.0228	2.82	0.0049*
vehicle_brand[Mercedes-Benz]	1522.3082	337.9404	4.50	<.0001*
vehicle_brand [Mitsubishi]	-1238.434	401.2572	-3.09	0.0021*
vehicle_brand [Renault]	-2797.88	361.7431	-7.73	<.0001*
vehicle_brand [Toyota]	412.98334	393.5631	1.05	0.2944
year	687.40845	34.22743	20.08	<.0001*
engine_capacity	3993.8874	302.325	13.21	<.0001*
miles_driven	-14.46845	2.200866	-6.57	<.0001*

Table 2: Parameter estimates table shows significance for each variable

Interpretation from Table 2: $\beta 1$ (vehicle_brand [Audi]), $\beta 2$ (vehicle_brand [BMW]), $\beta 3$ (vehicle_brand [Mercedes-Benz]), $\beta 4$ (vehicle_brand [Mitsubishi]), $\beta 5$ (vehicle_brand [Renault]), $\beta 7$ (year), $\beta 8$ (engine_capacity]), $\beta 9$ (miles_driven):- Based on t-test, we can claim that all brands (except for Toyota), year, engine capacity and miles driven contributes significantly to the linear regression model, given the other regressors are included in the model.

$\beta 6$ (vehicle_brand [Toyota]):- Based on the t-test, we can claim that brand Toyota does not contribute significantly to the model, given the other regressors are included in the model. (Toyota has P-value > 0.05 and does not significantly impact the price of the car)

9.0 Conclusion

A multiple regression model was performed to investigate whether the mentioned input variables could significantly predict the selling price of used car. The final Model as a whole is significant and this is concluded from the small P-value from the ANOVA table, Table 1. Moreover, each input variable included is significant and this is concluded from the P-values for each variable from the Parameter Estimate Table, Table 2, except for the brand Toyota. The estimate values in the Table 2, helps in answering which brand increases the sales price of a car. For instance, the interpretation of $\beta 3$ (vehicle_brand [Mercedes-Benz]), is that for every Mercedes-Benz vehicle, the expected selling price increases by 1522.3 units. Conversely, for $\beta 5$ (vehicle_brand [Renault]), for every Renault vehicle, the expected selling price decreases by 2797.88 units. The estimate of the $\beta 10$, year variable helps us interpret that as every year newer the car is, it's expected selling price increases by 687.4 units. Which means newer cars are more expensive than the old ones. Estimate of $\beta 8$, engine capacity allows us to interpret that increase in engine volume by 1 litre, increases the sales price by 3993.88 units, which means that engine specification is one of the important factors that people see and increase the price of a car. Another estimate for the 9th variable $\beta 9$, miles driven is that for every mile that is covered by the car, the selling price of the car is decreased by 14.46 units. This concludes that as the car is used a lot by driving, its value in resale market decreases. This concludes that there is a causal relationship between the input and output variables. JMP uses Effect Coding to read the 7 level categorical brand names in the dataset. During the model fitting, brand the brand names are encoded as -1s, 0s and 1s. Due to this the Volkswagen brand will not be seen in the Table 2. The base model assumes that the car is Volkswagen and considers it as a benchmark. All other brand dummies give a comparison with it.

Summary of Fit

Title	Values
RSquare	0.661471
RSquare Adjusted	0.65668
Root Mean Square Error	3402.277

Table 3: Summary of fit showing R^2, adjusted R^2 and RMSE values for validation

Interpretation from Table 3: Since adjusted R^2 accounts for penalty based on number of additional terms and since this model used many input variables, adjusted R^2 value is being considered. Adjusted R^2 = 65.66% of the variability in the selling price of used car can be explained by all the input variables. 65.66% shows that this model has a considerable explanatory power if not having a very high value like >80%. RMSE value is little high, this is because of the noise in the data, this could also be due to low number of observations considered in dataset.

10.0 Recommendation

The limitations that were faced during the project was to use a dataset that has to be lesser than 1000 observations. Statistical tests showed that the multiple linear regression model is significant, however, the adjusted R^2 value is observed to be 65.66%, which is considered to be low if this project is to be used as a prediction tool. The root mean squared value is also bigger and this is likely due to the smaller dataset used. This dataset size limit brings down the accuracy of the predictive model. There is a very high possibility that the adjusted R^2 value obtained will be higher if model is fitted on the actual bigger version of dataset consisting of 4000+ data points. It is recommended to use the actual data to improve the score or R^2 value.

11.0 Software used

1. Python 3
2. JMP Pro.
3. Microsoft Excel

12.0 Links to resources

1. Final dataset
2. Python pre-processing
3. Python data analysis
4. JMP report

pc2846@rit.edu

Appendix – Figures and Tables

pc2846@rit.edu