"Craigslist Vehicles - Predictive Analysis"

BUAN/MKT 6337 - Spring 2020

Under the guidance of:

Mr Sourav Chatterjee

Group 10:

Ayush Singh (aks171430)

Chirag Hamirani (cph190000)

Harshavardhan Akiti (hxa180011)

RajaThushara Nama (rxn180011)

Tamanna Kawatra (txk190011)

Table of Contents

Executive Summary

The prices of new cars 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 price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase (Pal, Arora and Palakurthy, 2018). There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offer this service, their prediction method may not be the best. Besides, different models and systems may contribute to predicting power for a used car's actual market value. It is important to know their actual market value while both buying and selling. Therefore, to determine the appropriate value(price) we have used the craigslist's vehicle dataset. The dataset set contains car values from 1900s to 2021. This dataset was cleaned, pre-processed, and several predictive models were runned out of which polynomial regression was found to be the best one with a 69.98% R2 score.

Prospective Clients of Price Prediction Model

To be able to predict used cars market value can help both buyers and sellers.

Used car sellers (dealers): They are one of the biggest target groups that can be interested in the results of this study. If used car sellers better understand what makes a car desirable, what the important features are for a used car, then they may consider this knowledge and offer a better service.

Online pricing services: There are websites that offer an estimated value of a car. They may have a good prediction model. However, having a second model may help them to give a better prediction to their users. Therefore, the model developed in this study may help online web services that tells a used car's market value.

Individuals: There are lots of individuals who are interested in the used car market at some points in their life because they wanted to sell their car or buy a used car. In this process, it's a big corner to pay too much or sell less than its market value.

Model Building

We will be running following models in this project:

1. Multi - Logit Regression
    - Significance of Model & Variables
    - Prediction Model Equations
    - Business Inferences
2. Linear Regression
    - Multiple linear regression with all variables
    - Linear Regression with Lasso penalty
    - Linear Regression with Elastic net
    - Linear Regression with Forward selection
    - Linear Regression with Backward selection
    - Linear Regression with Stepwise selection
3. Polynomial regression
    - Stepwise Selection
4. Regression Trees
5. Logistic Regression : To predict the category of Used cars

Introduction About The Dataset

Craigslist is the world's largest collection of used vehicles for sale. The dataset is compiled from the same. It includes information like price, condition, odometer, and several other 22 columns. A brief description of the dataset is given below.

Total Number of Records - 539759
Total Number of Columns - 25
Total Missing Values - 2227259
Missing Value Percentage - 16.50 %

Cars Dataset

The CONTENTS Procedure

Data Set Name	WORK.CARS	**Observations**	539759
Member Type	DATA	**Variables**	25
Engine	V9	**Indexes**	0
Created	04/21/2020 20:22:54	**Observation Length**	6424
Last Modified	04/21/2020 20:22:54	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

The attributes include:
Vehicle features: price, year, manufacturer, condition, cylinders, fuel, odometer, title status, transmission, vin, drive, size, type, paint color, model
Location features: latitude, longitude, county, state, region
Other: URL, image url, description, region url, id

The variable price is considered as the target or response variable for our analysis. The main aim will be to determine how different features are affecting the price of the car.

Data Pre-Processing

Dropping Irrelevant Columns

Some variables in our dataset have no inherent value to our analysis. The first few variables we determined to have little value are “url”, “image_url”, and “region_url” variables. While they have some textual information in the hyperlink to the listing, we decided to remove these variables as they cannot easily be used in regression models. We also have many other variables related to the geographic location such as “county”, “state”, “region”, “lat”, and “long”. However, we don’t plan to use them since we are not focusing on geo-spatial analysis. Furthermore, we have dropped “vin”, “description” and “id” variables since they do not add value to our analysis. Lastly, we have dropped the “manufacturer”, “models” and “size” variable on the basis of 50+ classification categories and more than 50% missing values.

Removing Outliers

On running univariate, the price, year and odometer variables showed many outliers. Therefore, their values were restricted to 99% of the range values. The price was considered from range 500 to 100000, the odometer was considered till 300000 and finally, the year data was considered from 1960 till the year 2020.

Imputing Empty Records

Variables	Empty row count	% of Empty rows
Price	0	0
Year	0	0
Odometer	84365	17.58
Paint Color	152450	31.76
Type	131583	27.42
Drive	137394	28.63
Transmission	3334	0.69
Title Status	2216	0.46
Fuel	2710	0.56
Cylinder	190811	39.76
Condition	199791	41.63

General rule adopted for Imputation for this project:-

1. Delete rows if missing values are more than 10%
2. Impute with mean/median for continuous values
3. Impute with mode for classification variables

The reason for deleting values greater than 10% was decided to avoid any bias in the data. For example, if we were to impute 40% of the data for cylinders then 40% or 2/5th of the data would have been biased toward a particular value. Moreover, some studies point out that the limitations of mean imputation are almost absent if less than 10% of the data is missing and when the correlations between the variables are low (Raymond, 1986; Tsikriktsis, 2005)

Variables like condition, cylinder, and paint color were dropped. After dropping the empty records value changed as well. Thus, the remaining variables were imputed as shown below.

1. Odometer - 108995 - Mean
2. Fuel - gas - Mode
3. Transmission - automatic - Mode
4. Type - sedan - Mode
5. Drive - 4wd - Mode

Label Encoding - Ordinal Variables

The dataset contains two ordinal variables namely condition and cylinders

Obs	price	year	condition	cylinders	fuel	title_status	transmission	drive	type	paint_color	odometer
1	7995	2010	excellent	8 cylinders	gas	clean	automatic	4wd	truck	white	194050
2	4000	1995	excellent	8 cylinders	gas	clean	automatic	4wd	truck	grey	133000
3	16000	2011	excellent	6 cylinders	gas	salva	automatic	fwd	sedan	grey	85000
4	10950	2011	excellent	6 cylinders	gas	clean	automatic	fwd	sedan	red	43418
5	9400	2011	good	6 cylinders	gas	clean	automatic	4wd	SUV	blue	145000
6	4500	2012	excellent	6 cylinders	gas	clean	automatic	4wd	sedan	silver	155000
7	999	2016	excellent	4 cylinders	gas	clean	automatic	4wd	SUV	purple	71000
8	1495	2004	good	6 cylinders	gas	clean	automatic	4wd	SUV	red	196123
9	2800	2002	like new	6 cylinders	gas	clean	automatic	4wd	SUV	silver	190000
10	6795	2002	excellent	6 cylinders	gas	clean	automatic	4wd	SUV	black	108995

The variable condition and cylinders were converted into 6 and 8 different classes respectively

Obs	price	year	fuel	title_status	transmission	drive	type	paint_color	odometer	condition	cylinders
1	7995	2010	gas	clean	automatic	4wd	truck	white	194050	4	5
2	4000	1995	gas	clean	automatic	4wd	truck	grey	133000	4	5
3	16000	2011	gas	salva	automatic	fwd	sedan	grey	85000	4	4
4	10950	2011	gas	clean	automatic	fwd	sedan	red	43418	4	4
5	9400	2011	gas	clean	automatic	4wd	SUV	blue	145000	3	4
6	4500	2012	gas	clean	automatic	4wd	sedan	silver	155000	4	4
7	999	2016	gas	clean	automatic	4wd	SUV	purple	71000	4	2
8	1495	2004	gas	clean	automatic	4wd	SUV	red	196123	3	4
9	2800	2002	gas	clean	automatic	4wd	SUV	silver	190000	5	4
10	6795	2002	gas	clean	automatic	4wd	SUV	black	108995	4	4

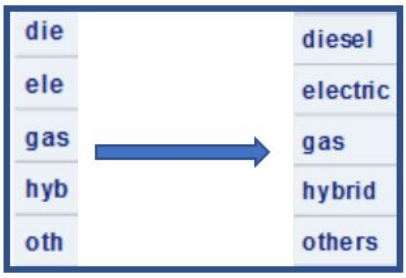
Transforming price and creating age Variable

On running the univariate procedure, the price was found to be not normal where a log transformation was performed and a new logPrice variable was created. Also, to make better analysis the year variable was subtracted with current year value 2020 to create a new age variable.

Obs	price	year	fuel	title_status	transmission	drive	type	paint_color	odometer	condition	cylinders	logPrice	age
1	7995	2010	gas	clean	automatic	4wd	truck	white	194050	4	5	8.99	10
2	4000	1995	gas	clean	automatic	4wd	truck	grey	133000	4	5	8.29	25
3	16000	2011	gas	salva	automatic	fwd	sedan	grey	85000	4	4	9.68	9
4	10950	2011	gas	clean	automatic	fwd	sedan	red	43418	4	4	9.30	9
5	9400	2011	gas	clean	automatic	4wd	SUV	blue	145000	3	4	9.15	9
6	4500	2012	gas	clean	automatic	4wd	sedan	silver	155000	4	4	8.41	8
7	999	2016	gas	clean	automatic	4wd	SUV	purple	71000	4	2	6.91	4
8	1495	2004	gas	clean	automatic	4wd	SUV	red	196123	3	4	7.31	16
9	2800	2002	gas	clean	automatic	4wd	SUV	silver	190000	5	4	7.94	18
10	6795	2002	gas	clean	automatic	4wd	SUV	black	108995	4	4	8.82	18

Formatting Fuel variable names

The dataset originally contained the fuel variable with length 3 so values of die, ele, hyb, oth were truncated. We formatted the dataset to change the length and format resulting in more legible values as posted below.



Exploratory Data Analysis

Data Overview

The dataset now has 189766 rows and 13 columns. It contains 5 numerical, 2 ordinal and 6 categorical variables.

Top 5 records

Obs	price	year	fuel	title_status	transmission	drive	type	paint_color	odometer	condition	cylinders	logPrice	age
1	7995	2010	gas	clean	automatic	4wd	truck	white	194050	4	5	8.99	10
2	4000	1995	gas	clean	automatic	4wd	truck	grey	133000	4	5	8.29	25
3	16000	2011	gas	salva	automatic	fwd	sedan	grey	85000	4	4	9.68	9
4	10950	2011	gas	clean	automatic	fwd	sedan	red	43418	4	4	9.30	9
5	9400	2011	gas	clean	automatic	4wd	SUV	blue	145000	3	4	9.15	9

Simple Statistics Results

The mean price of the car is $12159. On average, most cars in the dataset are from the year 2009. The average American seems to be driving their car for 108995 miles. The typical car has 5 to 6 cylinders.

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
price	189766	12159	9951	2307321217	500.00000	100000
year	189766	2009	7.97413	381169407	1960	2020
odometer	189766	108995	56417	2.06835E10	0	300000
condition	189766	3.65209	0.74382	693043	1.00000	6.00000
cylinders	189766	3.74619	1.23375	710899	1.00000	8.00000
logPrice	189766	9.08602	0.83973	1724217	6.21000	11.51000
age	189766	11.37144	7.97413	2157913	0	60.00000

Correlation Matrix

From the below matrix, we can observe that all the variables are significant at the 5% level.

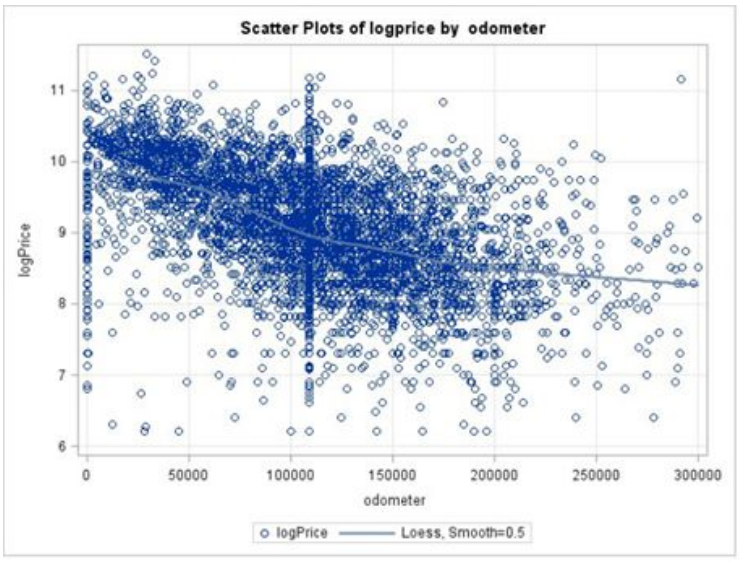
Some interesting insights,

1. Price has a negative relationship with the age of the car that is as the cars are getting older the price is decreasing.
2. Similarly, the odometer has a negative relationship, that is more the cars have been driven the lesser it's price will be.

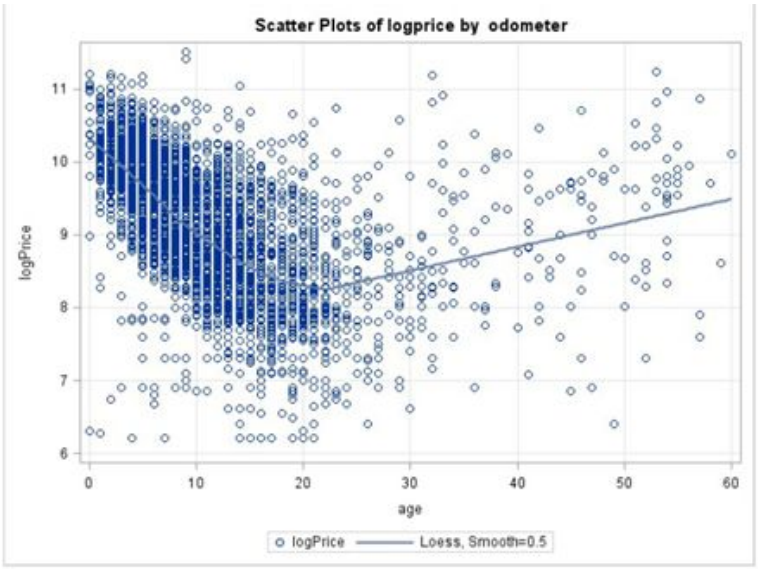
Pearson Correlation Coefficients, N = 189766
Prob > |r| under H0: Rho=0

	price	year	odometer	condition	cylinders	logPrice	age
price	1.00000	0.37962 <.0001	-0.48005 <.0001	0.21272 <.0001	0.28496 <.0001	0.88254 <.0001	-0.37962 <.0001
year	0.37962 <.0001	1.00000	-0.32122 <.0001	0.23463 <.0001	-0.17229 <.0001	0.44064 <.0001	-1.00000 <.0001
odometer	-0.48005 <.0001	-0.32122 <.0001	1.00000	-0.21720 <.0001	0.08818 <.0001	-0.48588 <.0001	0.32122 <.0001
condition	0.21272 <.0001	0.23463 <.0001	-0.21720 <.0001	1.00000	-0.08860 <.0001	0.28399 <.0001	-0.23463 <.0001
cylinders	0.28496 <.0001	-0.17229 <.0001	0.08818 <.0001	-0.08860 <.0001	1.00000	0.25338 <.0001	0.17229 <.0001
logPrice	0.88254 <.0001	0.44064 <.0001	-0.48588 <.0001	0.28399 <.0001	0.25338 <.0001	1.00000	-0.44064 <.0001
age	-0.37962 <.0001	-1.00000 <.0001	0.32122 <.0001	-0.23463 <.0001	0.17229 <.0001	-0.44064 <.0001	1.00000

Relationship between logPrice and Odometer The scatter plot between price and odometer reading shows a decreasing trend. They have negative correlation (-0.48). As the odometer reading increases, the price of used cars decreases.

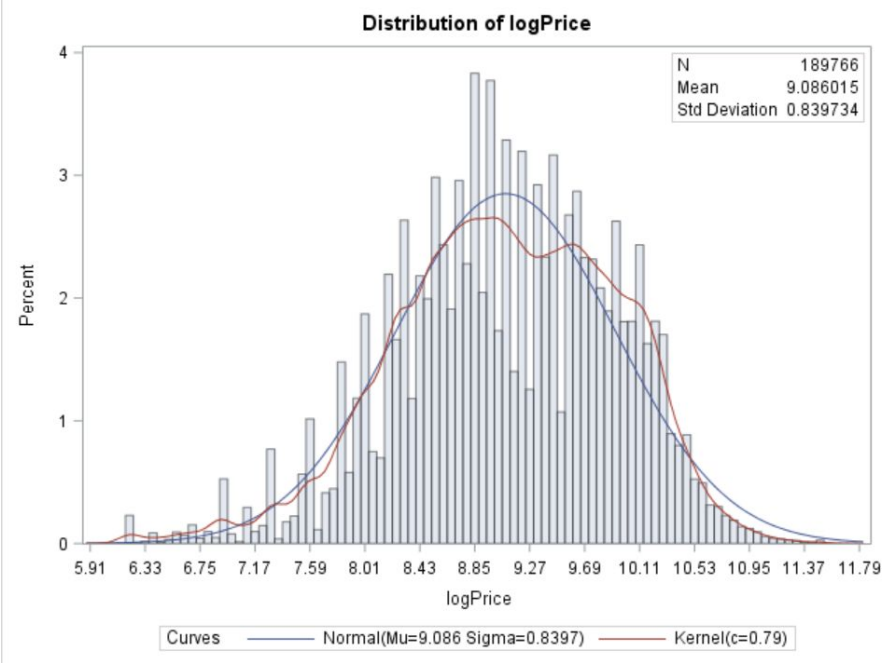


Relationship between logPrice and Age: The scatter plot between logprice and age shows that the price of the car first decreases with increase in the age of the car, but If the age of the car is greater than ~25 years, the price of used cars listed is increasing. It is possible if some Vintage cars are also listed. As vintage cars are priced higher, and their age is greater than 25 years this explains the trend. For example : Alfa Romeo Spider, Skylark etc. are few vintage cars listed in the data. The trends show that price and age are not linearly related and few higher degree polynomials to capture the trend.

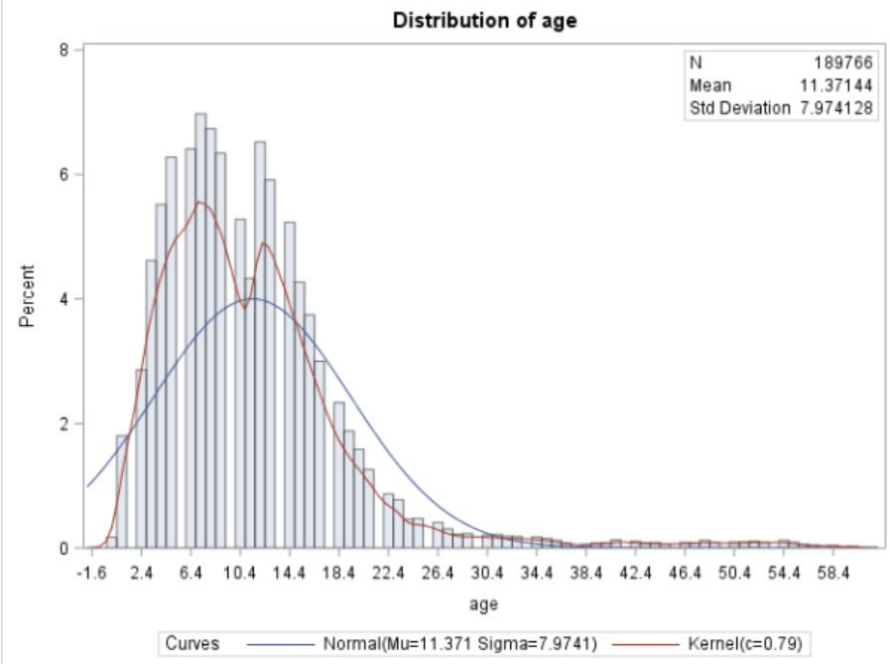


Distribution Analysis - Continuous Variables

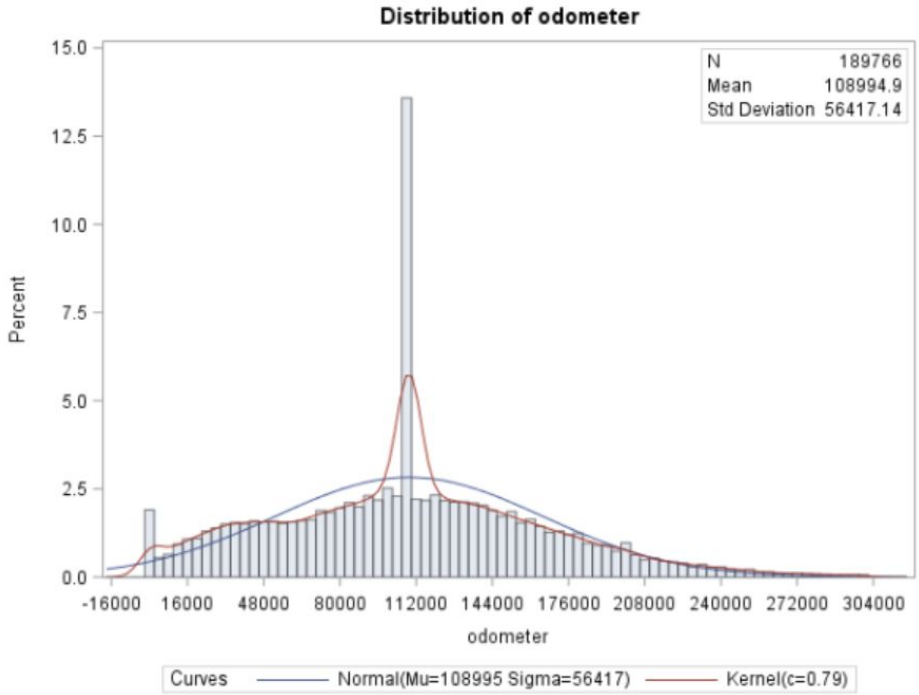
The logPrice variable appears to be normally distributed with a mean value of 9.08



The age variable seems to be skewed towards the right because of the large number of cars being listed around the year 2009



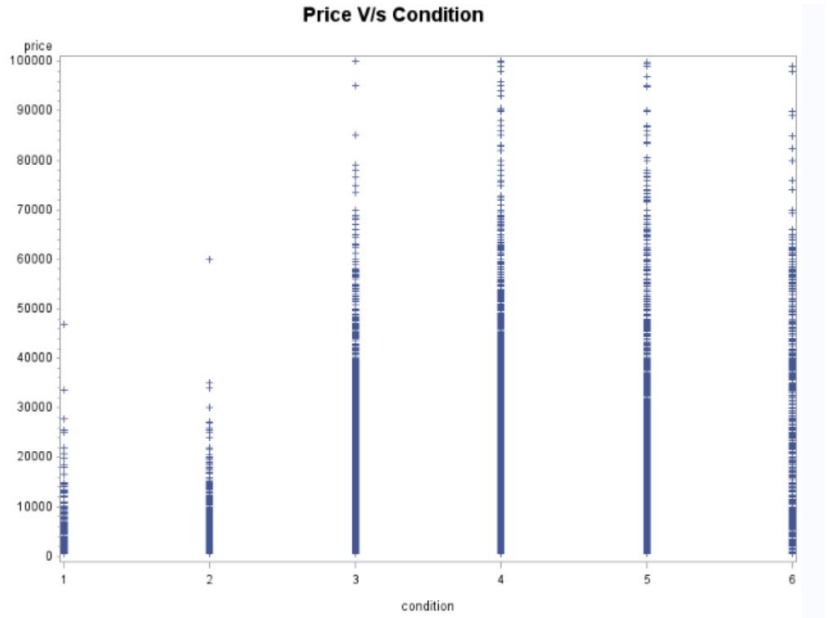
The odometer variable appears to be normally distributed. The peak at the center is due to the significantly high mean value and also the imputed value based on the mean imputation technique.



The below graph clearly suggests that the price of the car for poorly maintained cars is less as compared to well maintained or new ones.

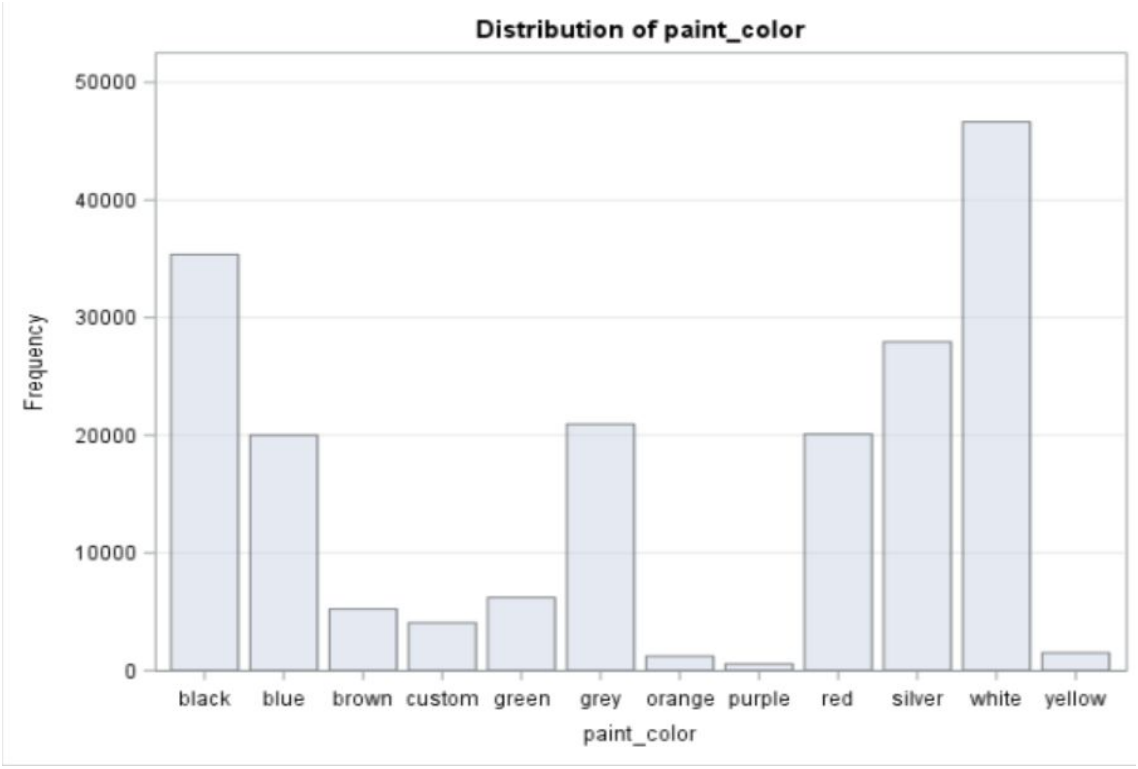
The encoding used for comparison is,

1. Salvage
2. Fair
3. Good
4. Excellent
5. Like new
6. New

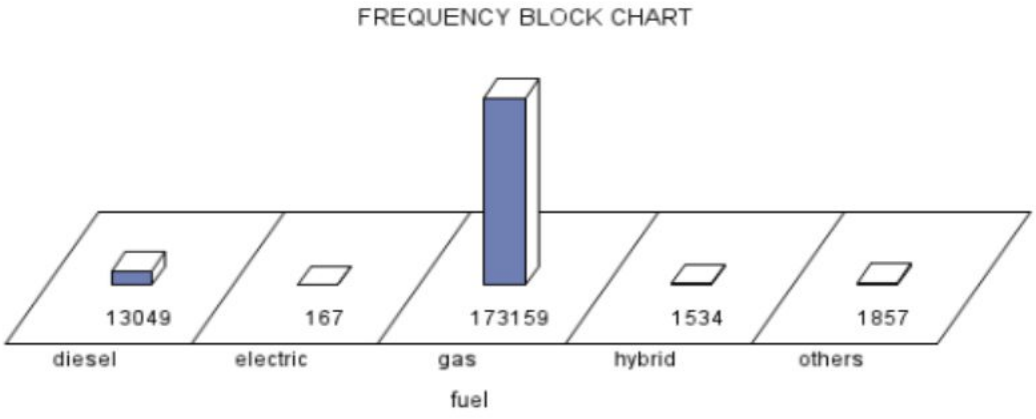


Distribution Analysis - Categorical Variables

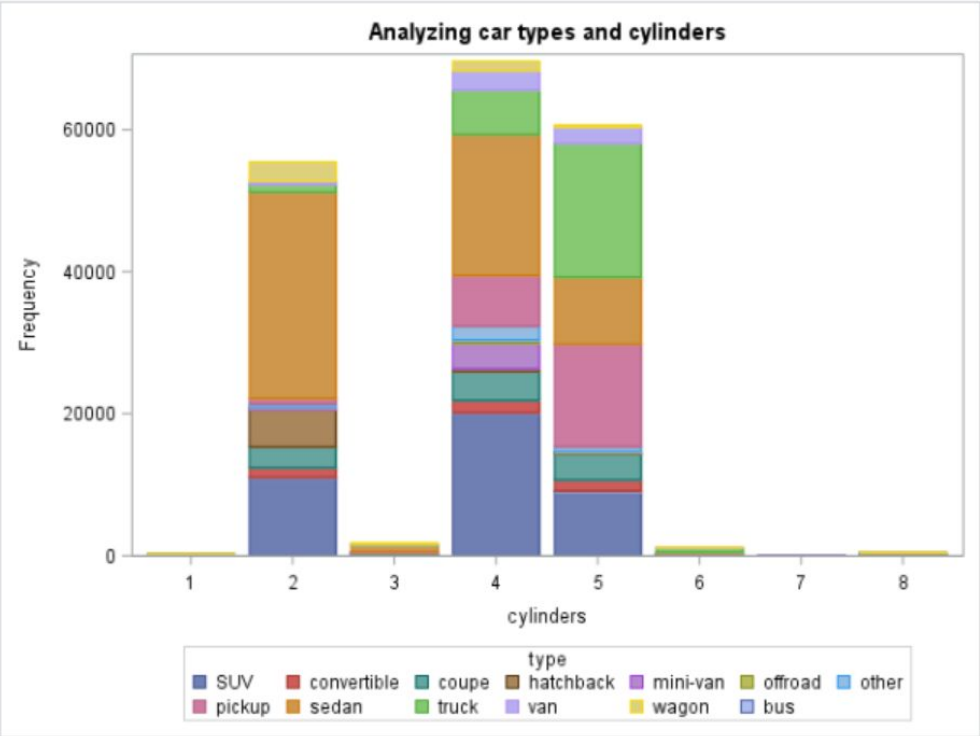
From the below graph, it appears that the white-colored cars are the most listed ones followed by black-colored cars



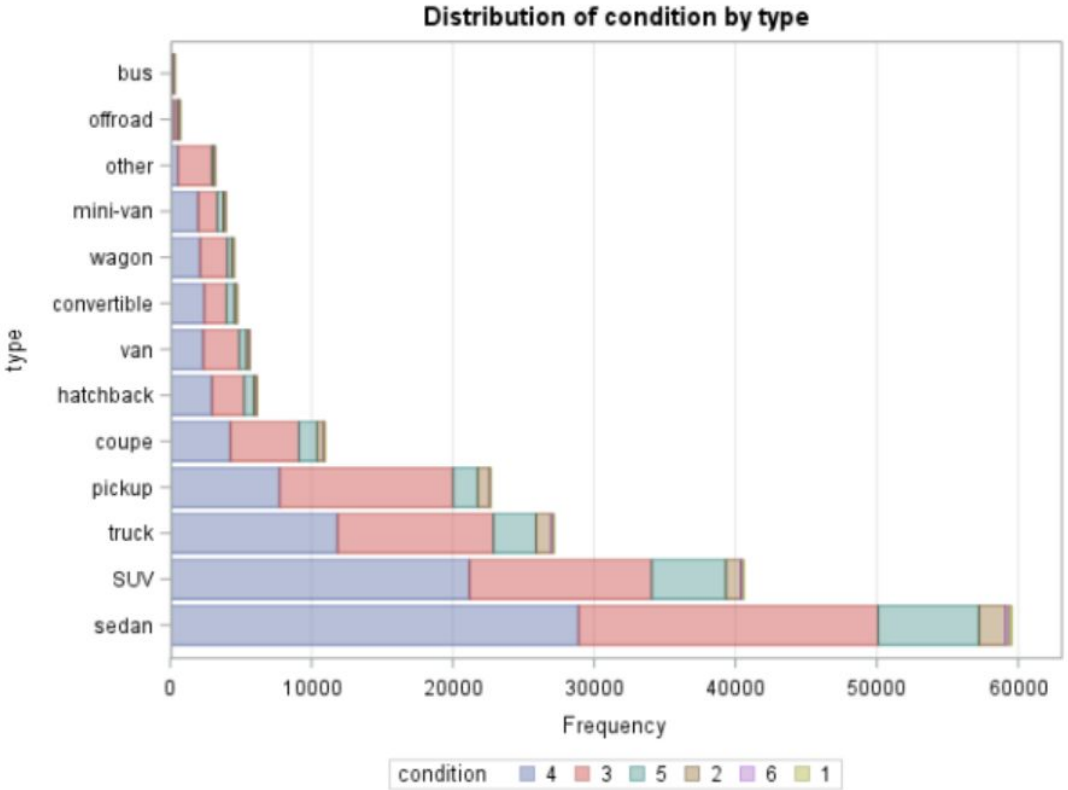
The frequency block chart for fuel suggests that the gas type cars are the most listed ones followed by diesel engines



From the below graph, it appears that the trucks have higher cylinder capacity than any other car type when it comes to 8 cylinders(encoded as 5)



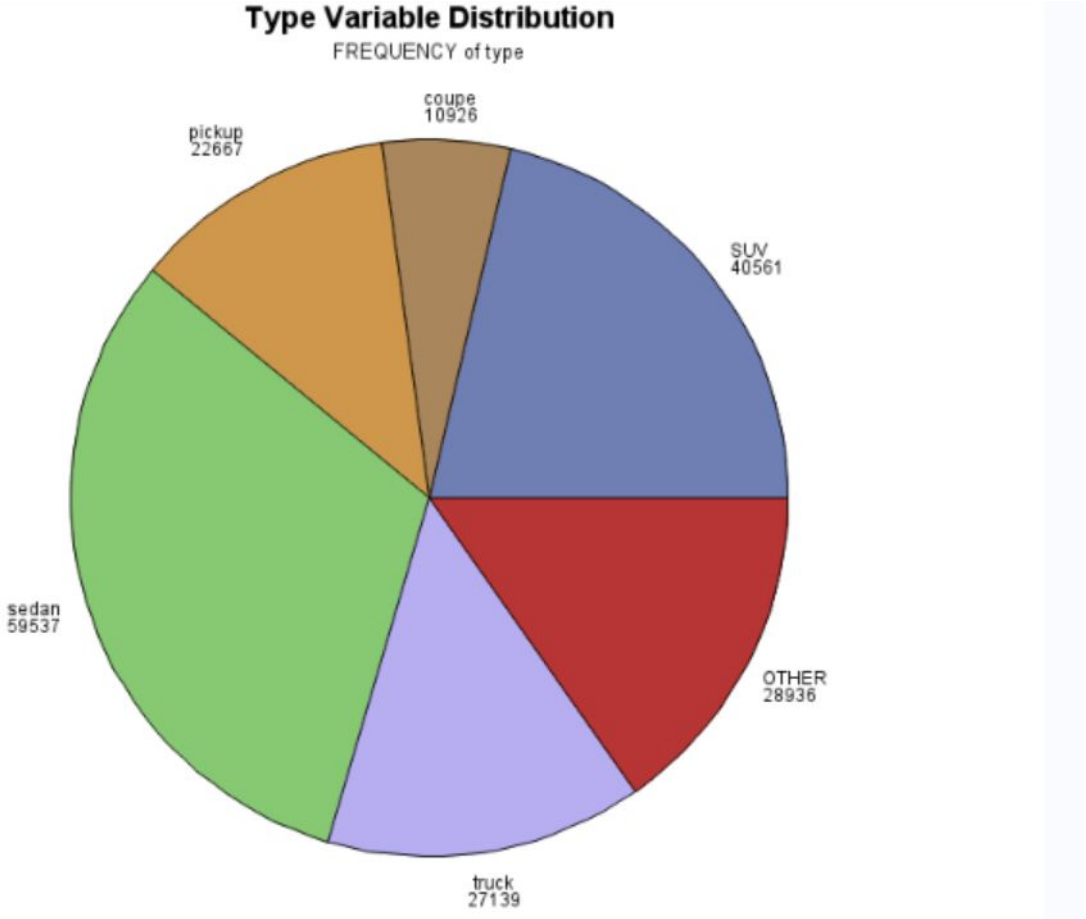
The dataset indicates that the majority of the SUV and sedan cars are in excellent condition



The below pie-chart lists the top 5 car types by listing namely,

1. Sedan
2. SUV
3. Trucks
4. Pickup
5. Coupe

This dataset has surprising results as most often Americans are associated with buying large vehicles like trucks and pickups however it's not the same case here.



Checking Impact of Categorical Variables on Price

Performing ANOVA for the variables to see if these variables have a significant impact on the dependent variable price.

As all the variables have a very high F statistic, we reject the null hypothesis and conclude that these variables have significant impact on the output variable.

Impact of Condition Variable

The ANOVA Procedure

Dependent Variable: price price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	988221259730	197644251946	1739.86	<.0001
Error	133699	1.518788E13	113597556.94		
Corrected Total	133704	1.6176101E13			

Impact of Cylinder Variable

The ANOVA Procedure

Dependent Variable: price price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	1.1726173E12	167516762660	1492.75	<.0001
Error	133697	1.5003484E13	112220047.47		
Corrected Total	133704	1.6176101E13			

Impact of year Variable

The ANOVA Procedure

Dependent Variable: price price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	61	5.3825738E12	88238914479	1092.55	<.0001
Error	133643	1.0793527E13	80763880.205		
Corrected Total	133704	1.6176101E13			

Impact of transmission Variable

The ANOVA Procedure

Dependent Variable: price price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	297814643329	148907321665	1253.86	<.0001
Error	133702	1.5878286E13	118758779.84		
Corrected Total	133704	1.6176101E13			

Impact of type Variable

The ANOVA Procedure

Dependent Variable: price price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	1.8709199E12	155909994036	1457.09	<.0001
Error	133692	1.4305181E13	107001025.47		
Corrected Total	133704	1.6176101E13			

Impact of drive Variable

The ANOVA Procedure

Dependent Variable: price price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.580579E12	790289523067	7239.43	<.0001
Error	133702	1.4595522E13	109164574.8		
Corrected Total	133704	1.6176101E13			

Impact of title_status Variable

The ANOVA Procedure

Dependent Variable: price price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	111837692731	22367538546	186.16	<.0001
Error	133699	1.6064263E13	120152456.88		
Corrected Total	133704	1.6176101E13			

Multinomial Logistic Regression

Description and Application

The data goes around cars, its attributes, and their resale value. Consider splitting the price into five groups which can roughly correspond to the purchase class of customers. The odds for cars being sold in a particular price group is greatly impacted by the car attributes.

Price Groups:

Price Group	Price Range	Record Count
1	<= 5,000	50620
2	5,001 – 10,000	55610
3	10,001 – 25,000	63740
4	25,001 – 40,000	16704
5	> 40,001	3092

To study the impact of car attributes on odds of sale price lying in a particular group, we will use multiple logistic regression. This will help us determine the significance of variables, their impact on odds in the collective equation. Furthermore, we will also look to narrow down effect due to certain categorical variables and if we can reason any anomalous behaviors.

Significance of Model & Variables

Amongst top important variables as per dataset are age of the car, number of cylinders, fuel type and odometer. While age and odometer are continuous variables, the number of cylinders is considered as ordinal (but continuous). The fuel type has five values: gas, diesel, electric, hybrid and others. We will convert these into additional five dummy variables. Since gas 91.25% records, when Logit is initially applied, it gives other fuel type values as insignificant (p-values not less than 0.05). Thus, we will run the regression with only gas fuel type.

While running the logit function with reference to group 5 (default), we get the following response. We get a very high AIC which implies a good model. Furthermore, the variables used show p-values less than 0.05, proving that variables selected are significant.

Logits modeled use Cgroup=5 as the reference category.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	516029.37	384209.56
SC	516069.98	384412.64
-2 Log L	516021.37	384169.56

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	131851.801	16	<.0001
Score	97755.9661	16	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
age	4	21885.6519	<.0001
odometer	4	29970.7335	<.0001
cylinders	4	27100.1802	<.0001
Fuel_gas	4	12900.7765	<.0001

Prediction Model Equations

The maximum likelihood table shows that each variable is significant, shown by p-value less than 0.05. These estimate values form the coefficient of equations below.

The equations below are deduced based on the table derived using multiple logit. For example, let us consider the odds ratio equation for probability of a point lying in price group 1 to probability of same point being in price group 5.

Analysis of Maximum Likelihood Estimates

Parameter	Cgroup	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	5.0285	0.1243	1636.8523	<.0001
Intercept	2	1	6.7931	0.1202	3194.9432	<.0001
Intercept	3	1	6.9626	0.1172	3531.8788	<.0001
Intercept	4	1	3.4781	0.1180	869.3351	<.0001
age	1	1	0.2459	0.00395	3871.7101	<.0001
age	2	1	0.1771	0.00391	2051.2935	<.0001
age	3	1	0.0583	0.00379	236.4040	<.0001
age	4	1	-0.0630	0.00407	240.0433	<.0001
odometer	1	1	3.3795	0.0336	10138.4909	<.0001
odometer	2	1	2.8021	0.0331	7150.0372	<.0001
odometer	3	1	1.9181	0.0323	3530.2105	<.0001
odometer	4	1	0.8380	0.0325	665.9721	<.0001
cylinders	1	1	-2.0728	0.0236	7741.5824	<.0001
cylinders	2	1	-1.8152	0.0233	6091.9122	<.0001
cylinders	3	1	-1.1261	0.0226	2474.7873	<.0001
cylinders	4	1	-0.2614	0.0222	138.8786	<.0001
Fuel_gas	1	1	5.6152	0.0586	9180.6497	<.0001
Fuel_gas	2	1	4.1806	0.0508	6775.9456	<.0001
Fuel_gas	3	1	2.6196	0.0449	3402.8913	<.0001
Fuel_gas	4	1	1.0108	0.0428	556.8136	<.0001

$$ln\left(\frac{P(Price\ Group=1)}{P(Price\ Group=5)}\right) = 5.03 + 0.25 * age + 3.38 * odometer - 2.07 * cylinders + 5.61 * fuel\_gas$$

$$ln\left(\frac{P(Price\ Group=2)}{P(Price\ Group=5)}\right) = 6.79 + 0.18 * age + 2.80 * odometer - 1.82 * cylinders + 4.18 * fuel\_gas$$

$$ln\left(\frac{P(Price\ Group=3)}{P(Price\ Group=5)}\right) = 6.96 + 0.06 * age + 1.92 * odometer - 1.13 * cylinders + 2.62 * fuel\_gas$$

$$ln\left(\frac{P(Price\ Group=4)}{P(Price\ Group=5)}\right) = 3.47 - 0.06 * age + 0.84 * odometer - 0.26 * cylinders + 1.01 * fuel\_gas$$

As per equation 1, odds ratio for price group 1 to price group 5

- Increases by 0.25 for increase by 1 year of age
- Increases by 3.38 for unit increase in odometer
- Decreases by 0.26 for single level increase in cylinders (2 to 3, or 10 to 12)
- Increases by 1.01 if car fuel type is gas

All the equations can be interpreted in similar fashion. These equations help predict odds ratio for any new point being in price groups 1 – 4 by price group 5. The value of odds ratio can be found by substituting variable values in above four equations.

Business Inferences

The multi-logit regression also gives point estimates for each variable and odds ratio of each price group with respect to group 5. With this we can observe cross price group impact for increase in particular variable.

Age

Point estimate is over 1 which shows higher odds for any car with higher age. This implies, with increase in age of a car, higher chances it will be sold in a lower price group.

For example, with increasing age, it will have 26.9% more probability to be sold in price group 1, as compared to price group 5.

Note that amongst group 4 and 5, the car with increasing age has more probability to be sold in price group 5. This behavior can be for rare vintage cars where increasing age contributes to price value.

Odds Ratio Estimates

Effect	Cgroup	Point Estimate	95% Wald Confidence Limits	
age	1	1.279	1.269	1.289
age	2	1.194	1.185	1.203
age	3	1.060	1.052	1.068
age	4	0.939	0.931	0.946
odometer	1	29.355	27.486	31.351
odometer	2	16.478	15.442	17.584
odometer	3	6.808	6.390	7.252
odometer	4	2.312	2.169	2.464
cylinders	1	0.126	0.120	0.132
cylinders	2	0.163	0.156	0.170
cylinders	3	0.324	0.310	0.339
cylinders	4	0.770	0.737	0.804
Fuel_gas	1	274.568	244.774	307.988
Fuel_gas	2	65.402	59.206	72.248
Fuel_gas	3	13.731	12.574	14.994
Fuel_gas	4	2.748	2.526	2.988

Odometer

Point estimates are way over 1. This shows the cars with high odometer readings are bound to be sold in lower price groups.

For Example, with increasing odometer readings, car will have about 28 times more chance to be sold in price group 1, 16 times more chances for group 2, 6 times more chances for group 3, and twice chance to be sold in group 4; all compared to chances of car being sold in group 5.

Cylinder

Point estimates are way less than 1. This implies the car with more cylinders has higher chances to be sold in higher groups. More accurate odds ratio increase can be observed by running logit in reference to price group 1.

Fuel

In the regression model we have considered only one dummy variable which tells if a car runs on gas fuel type or not. If it does, the car tends to be sold in lower price groups. The data is mainly around cars with fuel type as gas. So, the impact of other fuel types on the main model gets ignored.

Thus, independently we separated a dataset which contains all non-gas fuel types. This analysis is shown alongside. It helps us get a better perspective of price group dependency on fuel type by multi-logit.

Fuel type with value 'Others' and few of other comparisons highlighted in yellow are discarded as they have insignificant effect on price group odds. Overall, we can observe some patterns to develop insights in the rest of the fuel type cars.

We cannot observe any trend in odds for diesel or electric fuel cars, despite them having significant odds ratios. Hybrid cars however show clearer tendency to sold in lower price groups compared to group 5 (except group4)

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Fuel_diesel	4	154.4475	<.0001
Fuel_electric	4	148.5623	<.0001
Fuel_hybrid	4	912.1512	<.0001
Fuel_others	0	.	.

Odds Ratio Estimates

Effect	Cgroup	Point Estimate	95% Wald Confidence Limits	
Fuel_diesel	1	0.409	0.298	0.561
Fuel_diesel	2	0.763	0.573	1.016
Fuel_diesel	3	0.398	0.311	0.509
Fuel_diesel	4	0.309	0.241	0.398
Fuel_electric	1	0.438	0.233	0.824
Fuel_electric	2	0.733	0.437	1.229
Fuel_electric	3	0.127	0.077	0.209
Fuel_electric	4	0.041	0.020	0.082
Fuel_hybrid	1	32.772	15.292	70.237
Fuel_hybrid	2	41.410	19.563	87.654
Fuel_hybrid	3	4.272	2.040	8.946
Fuel_hybrid	4	0.573	0.259	1.268

Logistic regression

To run logistic regression we divided data into 5 buckets based on the price variable. We created a new column called Cgroup based on the price range. We used some of the variables as categorical variables such as drive, fuel, paint_color, title_status and transmission type.

Price range	Cgroup
<5k	1
5k to 10k	2
10k to 25k	3
25k to 40k	4
>40k	5

Response Profile

Ordered Value	Cgroup	Total Frequency
1	1	35535
2	2	39014
3	3	44453
4	4	11717
5	5	2118

Convergence

We got a satisfied model with AIC of 244215.78, when we tried with more number price groups we got higher AIC values. We got the best AIC of 182157.62 when we divided the data into only two groups(if the price is above 10K and below 10K).

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
4942.4762	114	<.0001

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	244215.78	149735.26
SC	244253.45	150130.77
-2 Log L	244207.78	149651.26

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	94556.5258	38	<.0001
Score	55244.9950	38	<.0001
Wald	50919.9176	38	<.0001

Stepwise Selection

We ran the model on stepwise selection, the Chi-Square has reduced to 574.8247 after adding all the variables.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	odometer		1	1	30803.8769		<.0001
2	type		12	2	24229.5117		<.0001
3	year		1	3	15820.5675		<.0001
4	drive		2	4	11230.3971		<.0001
5	fuel		4	5	9756.0258		<.0001
6	cylinders		1	6	5966.0707		<.0001
7	condition		1	7	3792.8860		<.0001
8	title_status		5	8	1230.1348		<.0001
9	transmission		2	9	1001.4129		<.0001
10	paint_color		11	10	574.8247		<.0001

Significance and Classification rate

We got a very good model based on the C-Statistic which is 0.878. From Somers D value we can say that there is good classification of data through all the Cgroups.

Association of Predicted Probabilities and Observed Responses

Percent Concordant	87.7	Somers' D	0.755
Percent Discordant	12.1	Gamma	0.757
Percent Tied	0.2	Tau-a	0.545
Pairs	6371498463	c	0.878

We divided the data into train and test in a 70:30 ratio, Error rate and R-square are almost similar for both train and test sets.

Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.TRAINING	132837	-121172	0.3683	242432.9	242433	242864	242864	0.590884	0.632637	.	0.498941
WORK.VALIDATION	56929	-51919.4	0.3689	103926.8	103926.8	104320.6	104320.6	0.592505	0.634211	.	0.499348

Prediction Model

The maximum likelihood table shows that most of the variables are significant except title_status(missi and parts), type (bus and other) and paint color with very low p values.

Analysis of Maximum Likelihood Estimates							
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	1	353.5	2.3938	21804.2168	<.0001	
Intercept	2	1	355.8	2.3967	22042.7340	<.0001	
Intercept	3	1	359.4	2.4013	22404.5957	<.0001	
Intercept	4	1	362.3	2.4036	22719.5314	<.0001	
year		1	-0.1764	0.00119	21954.0326	<.0001	-0.7747
fuel	diesel	1	-1.8279	0.0464	1549.2111	<.0001	-0.2762
fuel	electric	1	1.4173	0.1546	84.0246	<.0001	0.0800
fuel	gas	1	0.6259	0.0427	214.8275	<.0001	0.1130
fuel	hybrid	1	-0.2915	0.0643	20.5753	<.0001	-0.0215
odometer		1	0.000021	1.349E-7	23612.3246	<.0001	0.6437
title_status	clean	1	-0.5747	0.0865	44.1534	<.0001	-0.1038
title_status	lien	1	-1.1955	0.0957	155.9523	<.0001	-0.1097
title_status	missi	1	0.2860	0.2241	1.6291	0.2018	0.0202
title_status	parts	1	1.1134	0.3739	8.8658	0.0029	0.0767
title_status	rebui	1	0.1678	0.0895	3.5147	0.0608	0.0214
transmission	automatic	1	0.3647	0.0132	764.3894	<.0001	0.0958
transmission	manual	1	-0.3224	0.0184	306.5152	<.0001	-0.0617
drive	4wd	1	-0.5090	0.00900	3200.2490	<.0001	-0.2212
drive	fwd	1	0.6881	0.0109	4017.5921	<.0001	0.2701
type	SUV	1	0.2697	0.0202	178.7908	<.0001	0.0667
type	bus	1	0.00728	0.1415	0.0026	0.9590	0.000631
type	conve	1	-0.5111	0.0383	177.8171	<.0001	-0.0620
type	coupe	1	-0.1401	0.0284	24.3115	<.0001	-0.0219
type	hatch	1	0.7776	0.0352	488.1919	<.0001	0.1018
type	mini-	1	0.5540	0.0416	177.5386	<.0001	0.0642
type	offro	1	-1.3467	0.0908	219.7911	<.0001	-0.1214
type	offro	1	-1.3467	0.0908	219.7911	<.0001	-0.1214
type	other	1	-0.0141	0.0440	0.1029	0.7484	-0.00157
type	picku	1	-0.5105	0.0231	487.4106	<.0001	-0.1028
type	sedan	1	0.6845	0.0192	1273.5099	<.0001	0.1900
type	truck	1	-0.4379	0.0220	394.7449	<.0001	-0.0944
type	van	1	0.2191	0.0346	40.1365	<.0001	0.0279
paint_color	black	1	-0.1195	0.0181	43.4348	<.0001	-0.0265
paint_color	blue	1	0.1740	0.0211	68.1680	<.0001	0.0310
paint_color	brown	1	0.2622	0.0349	56.5012	<.0001	0.0270
paint_color	custom	1	-0.1546	0.0384	16.1955	<.0001	-0.0145
paint_color	green	1	0.2307	0.0333	47.9467	<.0001	0.0254
paint_color	grey	1	0.0726	0.0207	12.3358	0.0004	0.0131
paint_color	orange	1	-0.4307	0.0668	41.5379	<.0001	-0.0285
paint_color	purple	1	0.3612	0.0999	13.0748	0.0003	0.0208
paint_color	red	1	0.0465	0.0210	4.8955	0.0269	0.00827
paint_color	silver	1	0.1459	0.0191	58.1531	<.0001	0.0297
paint_color	white	1	-0.0812	0.0174	21.7376	<.0001	-0.0199
condition		1	-0.5297	0.00848	3902.3819	<.0001	-0.2172
cylinders		1	-0.4886	0.00631	5989.3819	<.0001	-0.3323

Confusion matrix

From the confusion matrix we got 63.17% train accuracy.

Frequency Percent Row Pct Col Pct	Table of F_Cgroup by I_Cgroup					
	I_Cgroup(Into: Cgroup)					
F_Cgroup(From: Cgroup)	1	2	3	4	5	Total
1	23000 17.31 64.72 68.51	10370 7.81 29.18 26.12	2085 1.57 5.87 4.05	76 0.06 0.21 1.01	4 0.00 0.01 0.69	35535 26.75
2	8699 6.55 22.30 25.91	21706 16.34 55.64 54.68	8550 6.44 21.92 16.61	57 0.04 0.15 0.76	2 0.00 0.01 0.35	39014 29.37
3	1467 1.10 3.30 4.37	7467 5.62 16.80 18.81	33948 25.56 76.37 65.95	1565 1.18 3.52 20.82	6 0.00 0.01 1.04	44453 33.46
4	273 0.21 2.33 0.81	114 0.09 0.97 0.29	6377 4.80 54.43 12.39	4821 3.63 41.15 64.13	132 0.10 1.13 22.84	11717 8.82
5	132 0.10 6.23 0.39	37 0.03 1.75 0.09	517 0.39 24.41 1.00	998 0.75 47.12 13.28	434 0.33 20.49 75.09	2118 1.59
Total	33571 25.27	39694 29.88	51477 38.75	7517 5.66	578 0.44	132837 100.00

For Test set we got an accuracy of 63.11%

Frequency Percent Row Pct Col Pct	Table of F_Cgroup by I_Cgroup					
	I_Cgroup(Into: Cgroup)					
F_Cgroup(From: Cgroup)	1	2	3	4	5	Total
1	9881 17.36 65.50 68.69	4287 7.53 28.42 25.54	886 1.56 5.87 3.98	29 0.05 0.19 0.88	2 0.00 0.01 0.90	15085 26.50
2	3725 6.54 22.45 25.90	9156 16.08 55.17 54.54	3689 6.48 22.23 16.58	26 0.05 0.16 0.79	0 0.00 0.00 0.00	16596 29.15
3	621 1.09 3.22 4.32	3297 5.79 17.09 19.64	14652 25.74 75.97 65.84	713 1.25 3.70 21.74	4 0.01 0.02 1.79	19287 33.88
4	94 0.17 1.88 0.65	36 0.06 0.72 0.21	2735 4.80 54.84 12.29	2072 3.64 41.55 63.19	50 0.09 1.00 22.42	4987 8.76
5	64 0.11 6.57 0.44	12 0.02 1.23 0.07	292 0.51 29.98 1.31	439 0.77 45.07 13.39	167 0.29 17.15 74.89	974 1.71
Total	14385 25.27	16788 29.49	22254 39.09	3279 5.76	223 0.39	56929 100.00

Linear Regression

In the linear regression model, we explain the linear relationship between a dependent variable and one or more explanatory variables. The usual method of estimating is Ordinary Least Squares (OLS). OLS minimizes the sum of the squared residuals. This method, along with a little calculus, leads to the closed form solution for the estimated parameters . Assuming that the error terms have finite variance and are uncorrelated with the regressors, this estimator is ***unbiased*** and ***consistent***. Further assuming that the variance is constant through the observations, the estimator is also ***efficient***.

Manipulating the dataset for Proc Reg procedure

As proc reg does not support categorical variables and partition statements, we have used following steps to modify the data:

- Scaling the data so as to avoid impacting one variable over the others. We have scaled odometer and age variables.
- Using proc GLMSELECT to create a design matrix: The design matrix stores each of the categories as individual variables and keeps one category as reference category in each category
- Using data step to partition the observations into training and validation set

After applying glm select and data partition, we get the data with 51 effects created in the model.

Categorical Variables	Reference Category
Condition	1
Cylinders	1
Paint_color	Yellow
Title Status	Salvage
Type	Wagon
Drive	rwd
Fuel	other
Transmission	other

Examples of Design matrix created for Ordinal and Nominal Variables

The design matrix of condition as an ordinal variable and Paint_color as a nominal variable has been shown below :

condition Level	Design Variables				
	1	2	3	4	5
1	0	0	0	0	0
2	1	0	0	0	0
3	1	1	0	0	0
4	1	1	1	0	0
5	1	1	1	1	0
6	1	1	1	1	1

paint_color Level	Design Variables										
	1	2	3	4	5	6	7	8	9	10	11
black	1	0	0	0	0	0	0	0	0	0	0
blue	0	1	0	0	0	0	0	0	0	0	0
brown	0	0	1	0	0	0	0	0	0	0	0
custom	0	0	0	1	0	0	0	0	0	0	0
green	0	0	0	0	1	0	0	0	0	0	0
grey	0	0	0	0	0	1	0	0	0	0	0
orange	0	0	0	0	0	0	1	0	0	0	0
purple	0	0	0	0	0	0	0	1	0	0	0
red	0	0	0	0	0	0	0	0	1	0	0
silver	0	0	0	0	0	0	0	0	0	1	0
white	0	0	0	0	0	0	0	0	0	0	1
yellow	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

Results of Linear Regression with Proc Reg:

The REG Procedure
Model: MODEL1
Dependent Variable: logPrice

Number of Observations Read	132732
Number of Observations Used	132732

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	56441	1128.81785	4014.49	<.0001
Error	132681	37308	0.28119		
Corrected Total	132731	93749			

Root MSE	0.53027	R-Square	0.6020
Dependent Mean	9.08498	Adj R-Sq	0.6019
Coeff Var	5.83677		

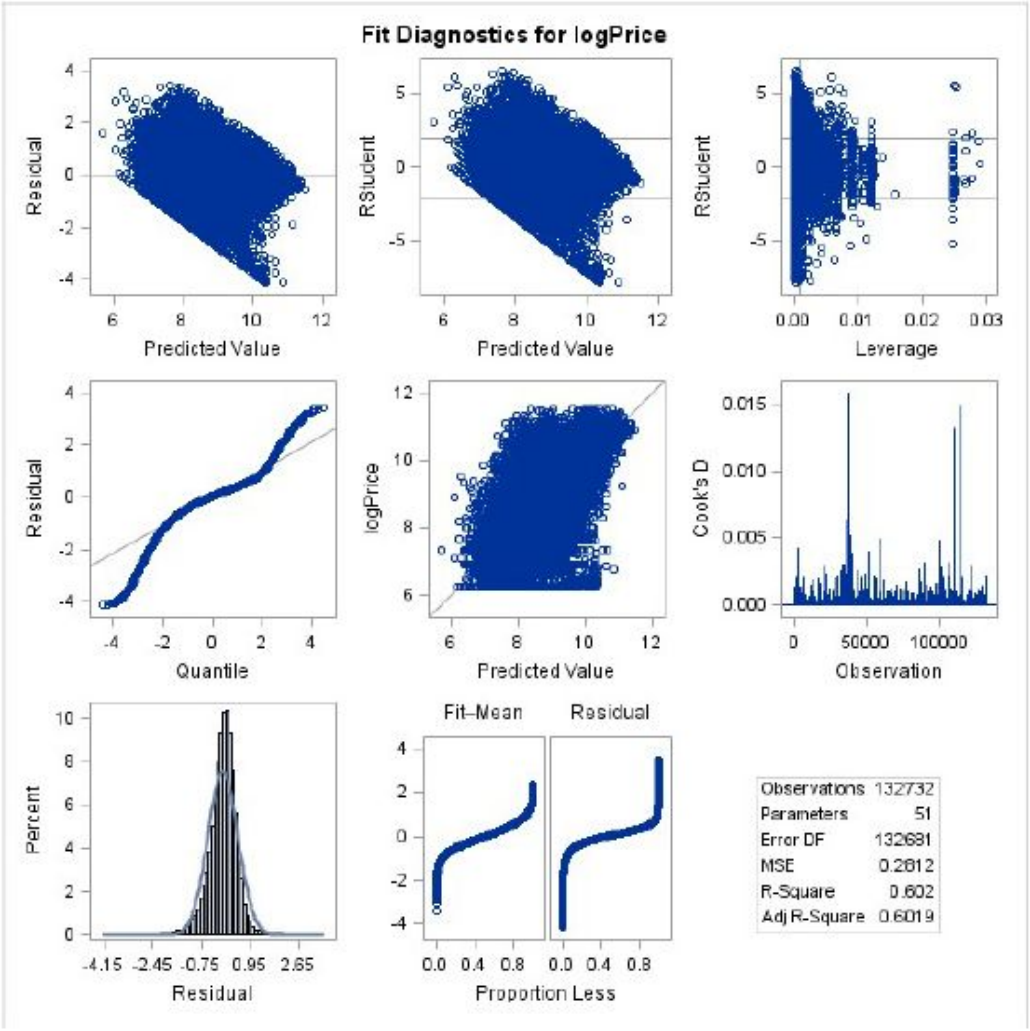
- As the F- statistic of the model is quite high, the model is statistically significant which states the price of used cars is dependent upon features like odometer, age of car, transmission, type, condition, fuel type and title status.
- The R2 of the model is 60.3% which states that 60.3% of the variation in the model is cumulatively explained by explanatory variables.
- Train RMSE of the model is 0.5302 Test RMSE: 0.5331

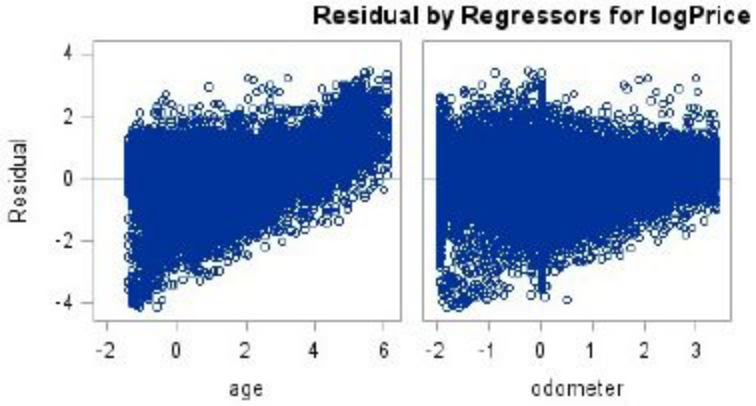
Model Equation

$$log(price) = 7.93 + (-0.276) * age + (-0.301) * odometer + 0.17 * condition2 + 0.701 condition3 + 0.184\, condition4 + 0.034 * condition5 + 0.166 * condition6 + 0.215 * cylinder2 + (-0.021) * cylinder3 + 0.184 * cylinder4 + \text{cylinder5} + 0.178 * \text{cylinder6} + 0.353 * \text{cylinder7} + -0.961 * \text{cylinder8} + 0.129 * \text{4wd} + 0.189 * \text{fwd} + 0340 * \text{fuel}_{\text{diesel}} + \cdots$$

Interpreting the Coefficients:

- With one- year increase in age of car the price of used cars drops by 27.6%
- If the odometer reading increase by one mile the price of cars drops by 30%
- Cars with condition “good” are priced higher by 70% than cars with condition salvage
- Cars with four- wheel drive (4wd) are priced higher by 13% than cars with rear wheel drive
- Cars with 4 cylinders (code value 2) are priced 21.5 % higher than cars with 3 cylinders (code value 1) . This intuitively also makes sense as the cars with higher cylinders have more value therefore they are priced higher.


Fit Diagnostics for logPrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
logPrice
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 132732
Parameters 51
Error DF 132681
MSE 0.2812
R-Square 0.602
Adj R-Square 0.6019


Residual by Regressors for logPrice
Residual
age
odometer

Parameter Estimates Table

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	7.93265	0.05174	153.33	<.0001	0
condition_2	condition 2	1	0.17178	0.03061	5.61	<.0001	1.10676
condition_3	condition 3	1	0.70197	0.00862	81.46	<.0001	1.18156
condition_4	condition 4	1	0.18470	0.00334	55.34	<.0001	1.28759
condition_5	condition 5	1	0.03441	0.00492	7.00	<.0001	1.16580
condition_6	condition 6	1	0.16690	0.02396	6.97	<.0001	1.03294
cylinders_2	cylinders 2	1	0.21551	0.03851	5.60	<.0001	1.00612
cylinders_3	cylinders 3	1	-0.02142	0.01495	-1.43	0.1518	21.84743
cylinders_4	cylinders 4	1	0.18208	0.01494	12.19	<.0001	22.24074
cylinders_5	cylinders 5	1	0.23563	0.00397	59.43	<.0001	1.64244
cylinders_6	cylinders 6	1	0.17873	0.01895	9.43	<.0001	1.60216
cylinders_7	cylinders 7	1	0.35308	0.06094	5.79	<.0001	5.93590
cylinders_8	cylinders 8	1	-0.96188	0.06578	-14.62	<.0001	5.63155
drive_4wd	drive 4wd	1	0.12935	0.00232	55.78	<.0001	1.58186
drive_fwd	drive fwd	1	-0.18860	0.00276	-68.25	<.0001	1.82930
fuel_die	fuel die	1	0.33955	0.01291	26.30	<.0001	5.92790
fuel_ele	fuel ele	1	0.06100	0.04429	1.38	0.1684	9.78342
fuel_gas	fuel gas	1	-0.25097	0.01217	-20.62	<.0001	7.50157
fuel_hyb	fuel hyb	1	-0.00345	0.01767	-0.20	0.8452	2.58263
paint_color_black	paint_color black	1	0.03583	0.00464	7.72	<.0001	1.65048
paint_color_blue	paint_color blue	1	-0.03523	0.00533	-6.61	<.0001	1.39046
paint_color_brown	paint_color brown	1	-0.08966	0.00875	-10.25	<.0001	1.26369
paint_color_custom	paint_color custom	1	0.04136	0.00975	4.24	<.0001	1.30785
paint_color_green	paint_color green	1	-0.09939	0.00813	-12.22	<.0001	1.25512
paint_color_grey	paint_color grey	1	-0.00293	0.00530	-0.55	0.5807	1.43336
paint_color_orange	paint_color orange	1	0.17622	0.01712	10.29	<.0001	1.98082
paint_color_purple	paint_color purple	1	-0.09321	0.02473	-3.77	0.0002	3.15713

paint_color_red	paint_color red	1	-0.01871	0.00532	-3.52	0.0004	1.39632
paint_color_silver	paint_color silver	1	-0.04044	0.00489	-8.27	<.0001	1.53369
paint_color_white	paint_color white	1	0.01968	0.00445	4.43	<.0001	1.84493
title_status_clean	title_status clean	1	0.15705	0.01605	9.79	<.0001	13.15888
title_status_lien	title_status lien	1	0.35693	0.01934	18.45	<.0001	4.94884
title_status_missi	title_status missi	1	0.03008	0.03853	0.78	0.4350	11.81401
title_status_parts	title_status parts	1	-0.55449	0.06960	-7.97	<.0001	36.54786
title_status_rebui	title_status rebui	1	-0.00245	0.01715	-0.14	0.8865	7.40646
transmission_automatic	transmission automatic	1	-0.13570	0.00339	-40.05	<.0001	1.23131
transmission_manual	transmission manual	1	0.03252	0.00466	6.98	<.0001	1.23329
type_SUV	type SUV	1	-0.07394	0.00517	-14.31	<.0001	2.52357
type_bus	type bus	1	0.00224	0.03518	0.06	0.9493	14.39339
type_conve	type conve	1	0.12678	0.00969	13.09	<.0001	2.13484
type_coupe	type coupe	1	0.05041	0.00708	7.12	<.0001	1.89728
type_hatch	type hatch	1	-0.19621	0.00892	-21.99	<.0001	2.09521
type_mini_	type mini-	1	-0.11589	0.01063	-10.90	<.0001	2.35066
type_offro	type offro	1	0.26979	0.02298	11.74	<.0001	6.67847
type_other	type other	1	0.11174	0.01131	9.88	<.0001	2.42911
type_picku	type picku	1	0.08318	0.00594	14.00	<.0001	2.23294
type_sedan	type sedan	1	-0.18916	0.00481	-39.31	<.0001	2.77276
type_truck	type truck	1	0.08024	0.00570	14.07	<.0001	2.34592
type_van	type van	1	-0.04203	0.00890	-4.72	<.0001	2.00199

Model Diagnostics

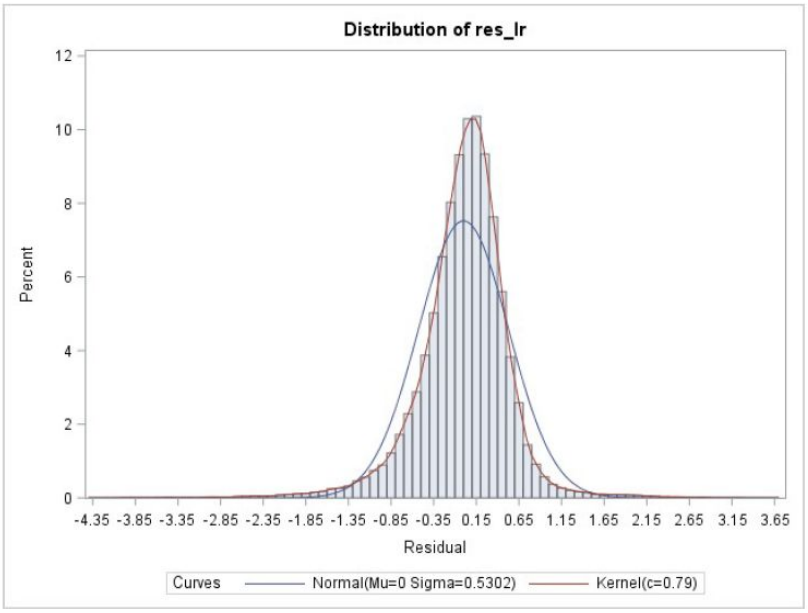
a. Checking multicollinearity

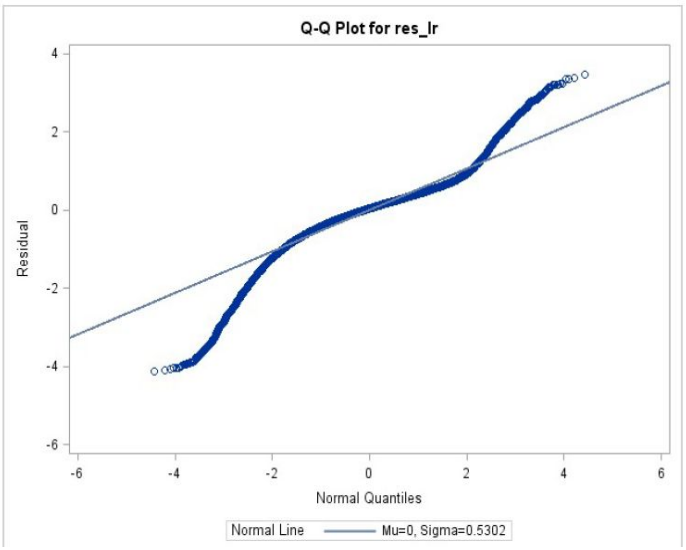
Analyzing the VIF in the above table, we see that few categories under cylinder and title status have VIF >10 representing multicollinearity in the model. This can happen if the proportion of cases in the reference category is small, the indicator variables will necessarily have high VIFs, even if the categorical variable is not associated with other variables in the regression model. This will only result in having p values of these indicator variables to be on the higher side. But the overall test that *all* indicators have coefficients of zero is unaffected by the high VIFs. And nothing else in the regression is affected. (ref: https://statisticalhorizons.com/multicollinearity)

b. Checking Model Assumptions

1. Test for Normality

- As p-value < 0.05 we reject the null hypothesis and conclude that the residuals are not normally distributed.
- As we have a large sample size and our goal is to estimate its coefficients and generate predictions in such a way as to minimize mean squared errors, we will go with these approximately normal residual plots.





Test for Normality

The UNIVARIATE Procedure
Fitted Normal Distribution for res_lr (Residual)

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	0.529502

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.07232	Pr > D	<0.010
Cramer-von Mises	W-Sq	246.50986	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1513.48547	Pr > A-Sq	<0.005

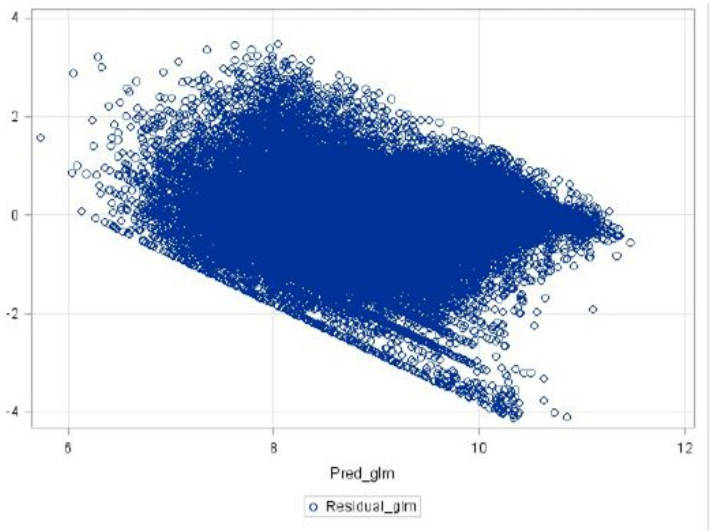
Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	-1.66494	-1.23181
5.0	-0.86547	-0.87095
10.0	-0.59181	-0.67858
25.0	-0.24488	-0.35714
50.0	0.04362	0.00000
75.0	0.29446	0.35714
90.0	0.52989	0.67858
95.0	0.69056	0.87095
99.0	1.34151	1.23181

2. Test for Heteroskedasticity

The plot between residual and fitted values shows a trend signifying heteroskedasticity. We performed a White Test to check for Heteroskedasticity. The null hypothesis for the test states the graph is heteroskedastic. As p-value is less than 0.001 we reject the null hypothesis and conclude that there is heteroskedastic in the model and variance of residual is not constant.

Implications of Heteroskedasticity

- **The least squares estimator is still a linear, unbiased and consistent estimator, but it is no longer BEST.** There is another estimator with a smaller variance: we can still avoid this implication as we have a large sample size so there is a probability of converging to true parameters.
- **The standard errors that are computed for the least squares estimator are incorrect** Confidence intervals and hypothesis tests that use these standard errors will be misleading.
- We have used **robust standard errors** to produce heteroskedastic - consistent standard errors.



Test for Homoskadasticity

The REG Procedure
Model: MODEL1
Dependent Variable: logPrice

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
1046	7552.20	<.0001

Solution for Heteroskedasticity

We have used the “acov” option of the proc reg feature to produce heteroskedastic consistent standard error. From the table we can see that Heteroskedastic standard errors are larger than OLS estimates. This states that Robust estimator are still not BEST but with a large enough sample size, this is not a problem as variance of the least squares estimator may still be sufficiently small to get precise estimates

Parameter Estimates

							Heteroscedasticity Consistent					
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	95% Confidence Limits		Heteroscedasticity Consistent 95% Confidence Limits	
Intercept	Intercept	1	7.90265	0.05174	153.33	<.0001	0.06990	113.49	<.0001	7.83125	8.03406	7.79566	8.06965
condition_2	condition 2	1	0.17178	0.03061	5.61	<.0001	0.04923	3.49	0.0005	0.11179	0.23178	0.07530	0.26827
condition_3	condition 3	1	0.70197	0.00862	81.46	<.0001	0.01146	61.24	<.0001	0.68508	0.71886	0.67950	0.72443
condition_4	condition 4	1	0.18470	0.00334	55.34	<.0001	0.00346	53.39	<.0001	0.17816	0.19124	0.17792	0.19148
condition_5	condition 5	1	0.03441	0.00492	7.00	<.0001	0.00521	6.60	<.0001	0.02477	0.04404	0.02419	0.04463
condition_6	condition 6	1	0.16690	0.02396	6.97	<.0001	0.03186	5.24	<.0001	0.11994	0.21386	0.10445	0.22935
cylinders_2	cylinders 2	1	0.21551	0.03851	5.60	<.0001	0.03897	5.53	<.0001	0.14002	0.29100	0.13913	0.29189
cylinders_3	cylinders 3	1	-0.02142	0.01495	-1.43	0.1518	0.01281	-1.67	0.0943	-0.05073	0.00788	-0.04652	0.00367
cylinders_4	cylinders 4	1	0.18208	0.01494	12.19	<.0001	0.01291	14.11	<.0001	0.15280	0.21136	0.15678	0.20738
cylinders_5	cylinders 5	1	0.23563	0.00397	59.43	<.0001	0.00417	56.45	<.0001	0.22786	0.24341	0.22745	0.24382
cylinders_6	cylinders 6	1	0.17873	0.01895	9.43	<.0001	0.01878	9.52	<.0001	0.14158	0.21588	0.14192	0.21553
cylinders_7	cylinders 7	1	0.35308	0.06094	5.79	<.0001	0.07638	4.62	<.0001	0.23364	0.47252	0.20337	0.50279
cylinders_8	cylinders 8	1	-0.96188	0.06578	-14.62	<.0001	0.09148	-10.51	<.0001	-1.09060	-0.83295	-1.14117	-0.78258
drive_4wd	drive 4wd	1	0.12935	0.00232	55.78	<.0001	0.00243	53.18	<.0001	0.12481	0.13390	0.12458	0.13412
drive_fwd	drive fwd	1	-0.18860	0.00276	-68.25	<.0001	0.00278	-67.77	<.0001	-0.19402	-0.18319	-0.19406	-0.18315
fuel_die	fuel die	1	0.33955	0.01291	26.30	<.0001	0.01755	19.35	<.0001	0.31424	0.36486	0.30516	0.37394
fuel_ele	fuel ele	1	0.06100	0.04429	1.38	0.1684	0.06587	0.93	0.3544	-0.02580	0.14781	-0.06811	0.19011
fuel_gas	fuel gas	1	-0.25097	0.01217	-20.62	<.0001	0.01710	-14.68	<.0001	-0.27482	-0.22712	-0.28448	-0.21745
fuel_hyb	fuel hyb	1	-0.00345	0.01767	-0.20	0.8452	0.01962	-0.18	0.8604	-0.03808	0.03118	-0.04190	0.03500
paint_color_black	paint_color black	1	0.03583	0.00464	7.72	<.0001	0.00505	7.09	<.0001	0.02673	0.04493	0.02593	0.04573
paint_color_blue	paint_color blue	1	-0.03523	0.00533	-6.61	<.0001	0.00589	-5.98	<.0001	-0.04568	-0.02478	-0.04679	-0.02368
paint_color_brown	paint_color brown	1	-0.08966	0.00875	-10.25	<.0001	0.00938	-9.56	<.0001	-0.10681	-0.07251	-0.10804	-0.07129
paint_color_custom	paint_color custom	1	0.04136	0.00975	4.24	<.0001	0.01142	3.62	0.0003	0.02224	0.06047	0.01897	0.06375
paint_color_green	paint_color green	1	-0.09939	0.00813	-12.22	<.0001	0.00983	-10.11	<.0001	-0.11532	-0.08345	-0.11865	-0.08013
paint_color_grey	paint_color grey	1	-0.00293	0.00530	-0.55	0.5807	0.00545	-0.54	0.5907	-0.01332	0.00746	-0.01360	0.00774
paint_color_orange	paint_color orange	1	0.17622	0.01712	10.29	<.0001	0.02096	8.41	<.0001	0.14266	0.20978	0.13515	0.21730
paint_color_purple	paint_color purple	1	-0.09321	0.02473	-3.77	0.0002	0.02945	-3.17	0.0015	-0.14168	-0.04474	-0.15093	-0.03550
paint_color_red	paint_color red	1	-0.01871	0.00532	-3.52	0.0004	0.00601	-3.11	0.0019	-0.02913	-0.00829	-0.03049	-0.00693
paint_color_silver	paint_color silver	1	-0.04044	0.00489	-8.27	<.0001	0.00518	-7.81	<.0001	-0.05003	-0.03085	-0.05058	-0.03030
paint_color_white	paint_color white	1	0.01968	0.00445	4.43	<.0001	0.00500	3.93	<.0001	0.01096	0.02839	0.00987	0.02948
title_status_clean	title_status clean	1	0.15705	0.01605	9.79	<.0001	0.03146	4.99	<.0001	0.12560	0.18851	0.09540	0.21870
title_status_lien	title_status lien	1	0.35693	0.01934	18.45	<.0001	0.03261	10.94	<.0001	0.31902	0.39485	0.29301	0.42085
title_status_missi	title_status missi	1	0.03008	0.03853	0.78	0.4350	0.07459	0.40	0.6868	-0.04543	0.10559	-0.11612	0.17628

title_status_parts	title_status parts	1	-0.55449	0.06960	-7.97	<.0001	0.14047	-3.95	<.0001	-0.69090	-0.41808	-0.82981	-0.27918
title_status_rebui	title_status rebui	1	-0.00245	0.01715	-0.14	0.8865	0.03187	-0.08	0.9387	-0.03606	0.03116	-0.06490	0.06001
transmission_automatic	transmission automatic	1	-0.13570	0.00339	-40.05	<.0001	0.00336	-40.35	<.0001	-0.14234	-0.12905	-0.14229	-0.12911
transmission_manual	transmission manual	1	0.03252	0.00466	6.98	<.0001	0.00520	6.26	<.0001	0.02339	0.04166	0.02234	0.04271
type_SUV	type SUV	1	-0.07394	0.00517	-14.31	<.0001	0.00626	-11.81	<.0001	-0.08407	-0.06381	-0.08621	-0.06167
type_bus	type bus	1	0.00224	0.03518	0.06	0.9493	0.05531	0.04	0.9678	-0.06671	0.07118	-0.10618	0.11065
type_conve	type conve	1	0.12678	0.00969	13.09	<.0001	0.01239	10.23	<.0001	0.10779	0.14577	0.10248	0.15107
type_coupe	type coupe	1	0.05041	0.00708	7.12	<.0001	0.00903	5.58	<.0001	0.03652	0.06429	0.03270	0.06811
type_hatch	type hatch	1	-0.19621	0.00892	-21.99	<.0001	0.00905	-21.67	<.0001	-0.21370	-0.17872	-0.21395	-0.17846
type_mini_	type mini-	1	-0.11589	0.01063	-10.90	<.0001	0.01127	-10.29	<.0001	-0.13672	-0.09505	-0.13797	-0.09381
type_offro	type offro	1	0.26979	0.02298	11.74	<.0001	0.02332	11.57	<.0001	0.22475	0.31484	0.22408	0.31551
type_other	type other	1	0.11174	0.01131	9.88	<.0001	0.01253	8.92	<.0001	0.08958	0.13390	0.08719	0.13629
type_picku	type picku	1	0.08318	0.00594	14.00	<.0001	0.00683	12.18	<.0001	0.07153	0.09482	0.06979	0.09656
type_sedan	type sedan	1	-0.18916	0.00481	-39.31	<.0001	0.00614	-30.82	<.0001	-0.19860	-0.17973	-0.20120	-0.17713
type_truck	type truck	1	0.08024	0.00570	14.07	<.0001	0.00670	11.98	<.0001	0.06906	0.09142	0.06711	0.09336
type_van	type van	1	-0.04203	0.00890	-4.72	<.0001	0.00926	-4.54	<.0001	-0.05947	-0.02459	-0.06017	-0.02388
age	age	1	-0.27540	0.00179	-154.21	<.0001	0.00357	-77.24	<.0001	-0.27890	-0.27190	-0.28239	-0.26841
odometer	odometer	1	-0.30059	0.00165	-182.02	<.0001	0.00204	-147.55	<.0001	-0.30383	-0.29735	-0.30458	-0.29660

3. Test for Independence

We have calculated Durbin Watson D statistics to test for autocorrelation. If the Durbin Watson D statistics value is close to 2, then there is no autocorrelation in the model and residuals are independent. As DW statistic ~1.85, we can still consider it satisfactory for modelling purposes as in practical problems errors are somewhat related.

Test for Autocorrelation

The REG Procedure
Model: MODEL1
Dependent Variable: logPrice

Durbin-Watson D	1.849
Number of Observations	113775
1st Order Autocorrelation	0.076

Linear Regression with GLMSelect

The optimal value is reached when all the effects are added in the model. The complete model specified in the MODEL statement is used to fit the model and no effect selection is done. You request this by specifying SELECTION=NONE in the MODEL statement. The model is chosen which has the lowest AICC, SBC value.
Train RMSE : 0.5269 Train MSE : 0.2775
Test RMSE : 0.5332 Test MSE : 0.2843

Dimensions	
Number of Effects	11
Number of Parameters	51

Regression Analysis with all variables using GLM select

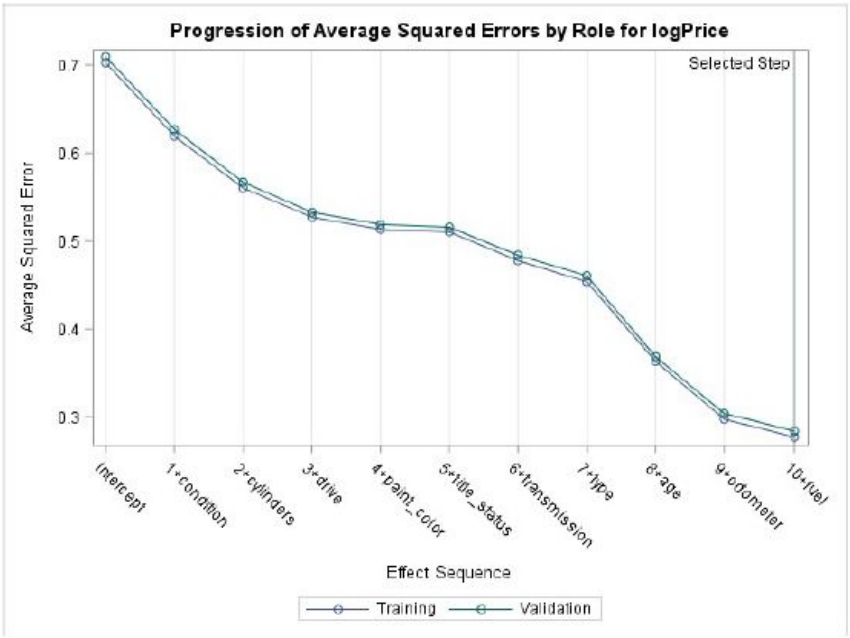
The GLMSELECT Procedure

Least Squares Summary

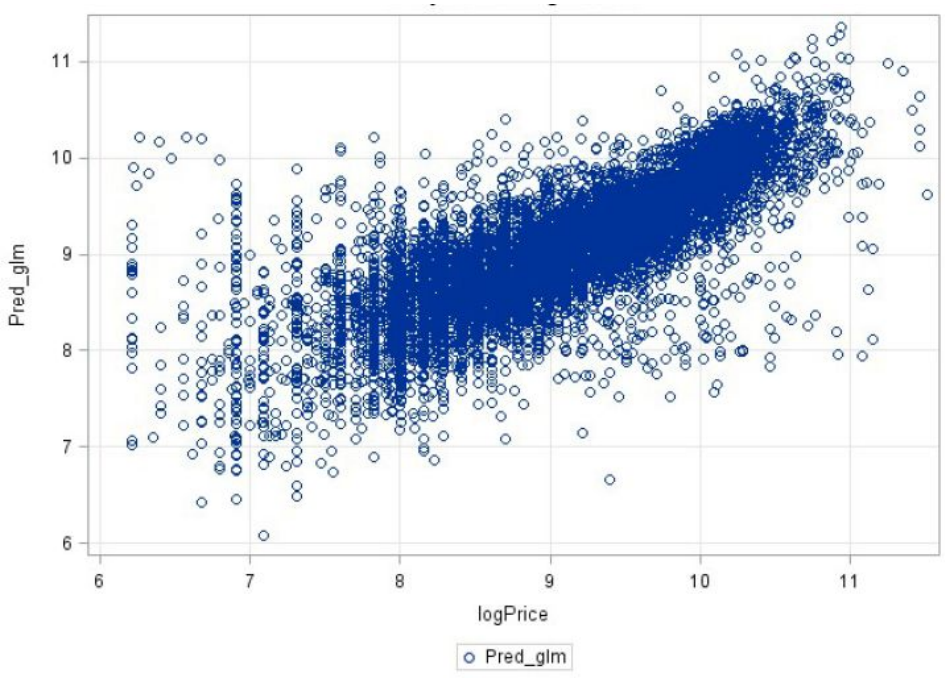
Step	Effect Entered	Number Effects In	Number Parms In	SBC	ASE	Validation ASE
0	Intercept	1	1	-40174.72	0.7025	0.7092
1	condition	2	6	-54481.22	0.6191	0.6270
2	cylinders	3	13	-65746.42	0.5604	0.5669
3	drive	4	15	-72676.20	0.5272	0.5327
4	paint_color	5	26	-75547.93	0.5134	0.5191
5	title_status	6	31	-76154.47	0.5105	0.5159
6	transmission	7	33	-83633.66	0.4779	0.4841
7	type	8	45	-89387.30	0.4538	0.4607
8	age	9	46	-114597.36	0.3636	0.3689
9	odometer	10	47	-137312.83	0.2977	0.3042
10	fuel	11	51	-145248.45*	0.2776	0.2843*

* Optimal Value of Criterion

.



The predicted v/s actual plot shows that there is a lot of scope of improvement as lot of points are away from the line.



Regularization

This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, **this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.**
We have implement Lasso and Elastic Net Linear Regression

Linear Regression with Lasso Penalty

For building this model we have used proc glm select with Lasso Penalty(l1 penalty). For choosing the best model we are implementing choose = cv criterion. The model which has the lowest cross validation scores is chosen. Using Lasso regression some of the parameters that do not help in decreasing the validation error are dropped.

Dimensions	
Number of Effects	11
Number of Effects after Splits	51
Number of Parameters	51

Regression Analysis with Lasso Penalty

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 45).

Effects:	Intercept condition_1 condition_2 condition_3 condition_4 condition_5 cylinders_1 cylinders_3 cylinders_4 cylinders_5 cylinders_6 cylinders_7 drive_4wd drive_fwd title_status_clean title_status_lien title_status_parts title_status_rebui paint_color_black paint_color_blue paint_color_brown paint_color_custom paint_color_green paint_color_grey paint_color_orange paint_color_purple paint_color_red paint_color_silver paint_color_white transmission_automatic transmission_manual type_SUV type_conve type_coupe type_hatch type_mini- type_offro type_other type_picku type_sedan type_truck type_van age odometer fuel_die fuel_gas

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	45	56312	1251.38602	4472.29	<.0001
Error	132601	37103	0.27981		
Corrected Total	132646	93415			

The model selection stopped at the 45th step. 46 parameters out of 51 were used. ***The model R2 remained the same***. Effects like fuel_type electric and fuel_type other, drive_fwd etc were removed from the system.The model which has the lowest Validation Average Squared Error (ASE) is chosen.

Linear Regression with ElasticNet

For building this model we have used proc glm select with ElasticNet. ElasticNet applies both l1 and l2 penalties. So few coefficients estimate to zero and few get penalized. For choosing the best model we are implementing choose = cv criterion. The model which has the lowest cross validation scores is chosen. Using ElasticNet regression some of the parameters that do not help in decreasing the validation error are dropped.

In ElasticNet, 47 parameters out of 51 have been used. ***The model R2 remained the same***. Effects like fuel_type electric and fuel_type other, drive_fwd etc were removed from the system ***However, the Test ASE got reduced which suggests that the model is generalizing better.***

Regression Analysis with Elastic Net

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 46).

ffects:	Intercept condition_1 condition_2 condition_3 condition_4 condition_5 cylinders_1 cylinders_2 cylinders_3 cylinders_4 cylinders_5 cylinders_6 cylinders_7 drive_4wd drive_fwd paint_color_black paint_color_blue paint_color_brown paint_color_custom paint_color_green paint_color_grey paint_color_orange paint_color_purple paint_color_red paint_color_silver paint_color_white title_status_clean title_status_lien title_status_parts title_status_rebui transmission_automatic transmission_manual type_SUV type_conve type_coupe type_hatch type_mini- type_offro type_other type_picku type_sedan type_truck type_van age odometer fuel_die fuel_gas

ote: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	46	56308	1224.08190	4374.13	<.0001
Error	132600	37108	0.27985		
Corrected Total	132646	93415			

**As ElasticNet had similar fit characteristics as Lasso we have omitted them from report*

Variable Selection in Linear Regression

In order to obtain a parsimonious model, various techniques to subset features of variables are deployed. We have used three subset selection methods namely *forward, backward and Stepwise selection. All the three models gave us the same result.*

For our analysis we have used the following select, stop and choose criteria:

Select: Based on the select criteria an effect is added in the model. For example in backward selection that effect is eliminated which yields lowest SBC value

Stop: Selection stops when further addition does not decrease/increase the selected criteria.

Choose: The model which has the lowest/ highest value based on criteria specified is chosen. For example, in the stepwise model we have chosen cv as the "Choose" criteria. So , the model which will give us lowest validation ASE will be chosen.

Model/ Criteria	Select	Stop	Choose
Forward	CV	CV	CV
Backward	SBC	SBC	CV
Stepwise	SBC	CV	CV

1. Forward Selection

In forward selection, we start with no predictors and then add predictors one by one. Each predictor added is the one (among all predictors) that helps to either increase or decrease the criterion that is specified in the selection criteria of the glm select module. If nothing is specified, a model is selected that tends to reduce SBC value.

- For our problem, we choose the Cross -Validation as the selection criteria. So predictors are added that tend to reduce the Cross -Validation error.
- Using the forward selection method, also all the effects are added as they tend to reduce the cross- validation score.
- But Using CV as the selection criterion, the test ASE got reduced, which suggests that the model generalizes better.

Regresion with forward Selection

The GLMSELECT Procedure

Forward Selection Summary

Step	Effect Entered	Number Effects In	Number Parms In	ASE	Validation ASE	CV PRESS
0	Intercept	1	1	0.7042	0.7073	93415.5414
1	odometer	2	2	0.5377	0.5410	71323.1013
2	type	3	14	0.4481	0.4494	59462.4137
3	age	4	15	0.3830	0.3862	50822.5375
4	cylinders	5	22	0.3505	0.3545	46515.8869
5	condition	6	27	0.3211	0.3249	42616.7482
6	fuel	7	31	0.2994	0.3026	39747.1603
7	drive	8	33	0.2864	0.2891	38022.8411
8	transmission	9	35	0.2827	0.2852	37536.7273
9	title_status	10	40	0.2809	0.2835	37308.6580
10	paint_color	11	51	0.2795	0.2820*	37132.1493*

* Optimal Value of Criterion

Selection stopped because all effects are in the final model.

2. Backward Selection

In **backward elimination**, we start with all predictors and then at each step, eliminate the least useful predictor (according to statistical significance). The algorithm stops when all the remaining predictors have significant contributions.

For our problem, the algorithm stopped at the first step as all the predictors had a significant impact.

Regression Analysis with Backward Selection

The GLMSELECT Procedure

Backward Selection Summary

Step	Effect Removed	Number Effects In	Number Parms In	SBC	ASE	Validation ASE	CV PRESS
0		11	51	-168476.70*	0.2795	0.2820*	37132.1493*

Selection stopped at a local minimum of the SBC criterion.

Stop Details

Candidate For	Effect	Candidate SBC		Compare SBC
Removal	paint_color	-167935.13	>	-168476.70

Regression Analysis with Backward Selection

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 0.

Effects:	Intercept condition cylinders drive paint_color title_status transmission type age odometer fuel

3. Stepwise Regression

Stepwise regression is like forward selection except that at each step, we consider dropping predictors that are not statistically significant, as in backward elimination.

- For our problem, stepwise regression also produces the same results, adding all the effects. Here model selection criteria is chosen at significance level, i.e. an effect stays in the model if its p-value is less than significance level. The model which gives lowest cross validation score is chosen
- *For stepwise selection* we have used an interaction term between age and condition to see the cumulative impact of condition and age of the car on price

The estimates show that there is only a slight increase in R2 of the model. Although the coefficient of age*conditions are statistically significant, they do not increase the model fit substantantially.

StepWise Regression with interaction of condition and age

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	SBC	ASE	Validation ASE	CV PRESS
0	Intercept		1	1	-46499.05	0.7042	0.7073	93415.5414
1	odometer		2	2	-82283.51	0.5377	0.5410	71323.1013
2	type		3	14	-106303.46	0.4481	0.4494	59462.4137
3	age		4	15	-127134.31	0.3830	0.3862	50822.5375
4	cylinders		5	22	-138819.09	0.3505	0.3545	46515.8869
5	condition		6	27	-150386.05	0.3211	0.3249	42616.7482
6	fuel		7	31	-159599.99	0.2994	0.3026	39747.1603
7	drive		8	33	-165468.44	0.2864	0.2891	38022.8411
8	transmission		9	35	-167155.23	0.2827	0.2852	37536.7273
9	age*condition		10	40	-168000.14	0.2808	0.2831	37291.2005
10	title_status		11	45	-168793.46	0.2790	0.2813	37060.0736
11	paint_color		12	56	-169322.18*	0.2776	0.2799*	36887.5871*

* Optimal Value of Criterion

StepWise Regression with interaction of condition and age

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 11.

Effects:	Intercept condition cylinders drive paint_color title_status transmission type odometer age*condition age fuel

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	55	56589	1028.88267	3704.39	<.0001
Error	132591	36827	0.27775		
Corrected Total	132646	93415			

Polynomial Regression

As the relationship between age and price is not linear. We have tried to fit a polynomial model with degree 2 of age with logPrice.

We have used proc glmselect" with polynomial effects to implement it.

The resultant model had the following characteristics:

- Adjusted R2 increased signifying age^2 is a relevant variable
- Test RMSE got reduced signifying the model fits the data well.
- Train RMSE also got reduced.

Polynomial Regression

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	ASE	Validation ASE	CV PRESS	F Value	Pr > F
0	Intercept		1	1	0.7042	0.7073	93415.5414	0.00	1.0000
1	odometer		2	2	0.5377	0.5410	71323.1013	41091.1	<.0001
2	type		3	14	0.4481	0.4494	59462.4137	2208.26	<.0001
3	age		4	15	0.3830	0.3862	50822.5375	22566.8	<.0001
4	age^2		5	16	0.3130	0.3163	41529.2586	29674.1	<.0001
5	cylinders		6	23	0.2756	0.2802	36573.9878	2570.68	<.0001
6	condition		7	28	0.2508	0.2550	33292.1979	2618.74	<.0001
7	fuel		8	32	0.2285	0.2318	30333.0579	3239.81	<.0001
8	drive		9	34	0.2170	0.2199	28811.9620	3504.53	<.0001
9	transmission		10	36	0.2141	0.2170	28429.5844	892.80	<.0001
10	title_status		11	41	0.2119	0.2151	28137.3415	282.56	<.0001
11	paint_color		12	52	0.2114	0.2147*	28086.0284*	25.62	<.0001

* Optimal Value of Criterion

Selection stopped because all effects are in the final model.

Age^2 variable entered at the 4th step signifying substantial drop in ASE value.

Polynomial Regression

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 11.

Effects:	Intercept condition cylinders drive paint_color title_status transmission type age age^2 odometer fuel

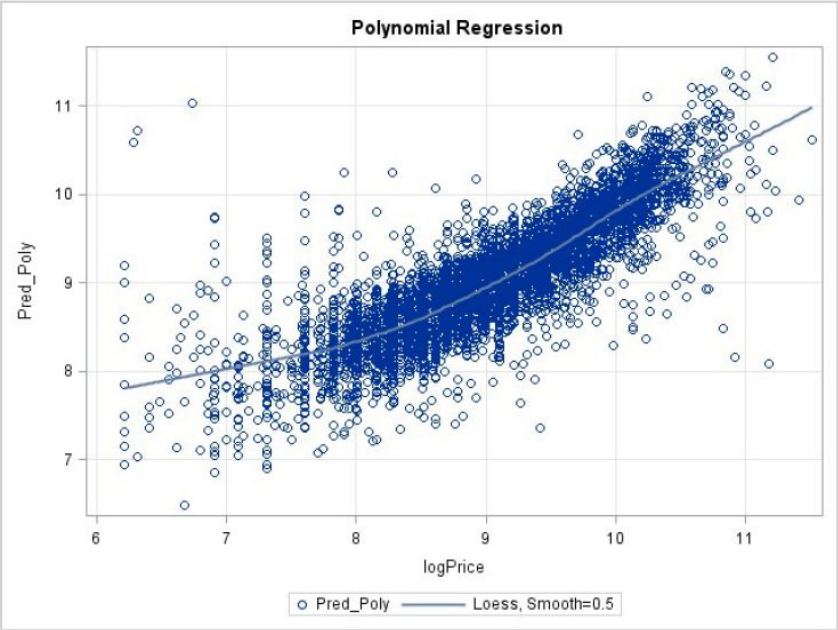
Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	51	65371	1281.78569	6060.37	<.0001
Error	132595	28044	0.21150		
Corrected Total	132646	93415			

Root MSE	0.45989
Dependent Mean	9.08870
R-Square	0.6998
Adj R-Sq	0.6997
AIC	-73368
AICC	-73368
SBC	-205508
ASE (Train)	0.21142
ASE (Validate)	0.21473
CV PRESS	28086

The fitted v/s actual plot of regression shows that the polynomial regression fit the data better than linear regression. There is still scope of improvement by using a weighted least square method in order to fit the data better. The model which has the lowest Validation Average Squared Error (ASE) is chosen.



Parameter Estimates

The parameter estimates show that age^2 is significant. Also addition of age^2 increased the r2 quite significantly. The resultant model now has R2~ 70%

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
type other	1	0.055246	0.009753	5.66	<.0001
type picku	1	0.035787	0.005196	6.89	<.0001
type sedan	1	-0.174678	0.004214	-41.45	<.0001
type truck	1	0.063560	0.004979	12.77	<.0001
type van	1	-0.070043	0.007822	-8.95	<.0001
age	1	-0.639576	0.002346	-272.62	<.0001
age^2	1	0.134544	0.000651	206.68	<.0001
odometer	1	-0.162862	0.001585	-102.78	<.0001
fuel dis	1	0.383419	0.011172	34.32	<.0001
fuel ele	1	0.023924	0.038338	0.62	0.5326
fuel gas	1	-0.222487	0.010502	-21.19	<.0001
fuel hyb	1	0.002110	0.015031	0.14	0.8884

Result of Polynomial Regression

Polynomial Regression

Obs	price	year	fuel	title_status	transmission	drive	type	paint_color	odometer	condition	cylinders	logPrice	age	_ROLE_	_CVINDEX_	Pred_Poly	Residual_poly
1	7995	2010	gas	clean	automatic	4wd	truck	white	1.5076104928	4	5	8.99	-0.171986414	VALIDATE	0	9.30745	-0.31745
2	4000	1995	gas	clean	automatic	4wd	truck	grey	0.4254926311	4	5	8.29	1.7090970272	TRAIN	2	8.65517	-0.36517
3	16000	2011	gas	salva	automatic	fwd	sedan	grey	-0.425312567	4	4	9.68	-0.297391977	VALIDATE	0	8.78806	0.89194
4	10950	2011	gas	clean	automatic	fwd	sedan	red	-1.162358021	4	4	9.3	-0.297391977	TRAIN	5	9.06375	0.23625
5	9400	2011	gas	clean	automatic	4wd	SUV	blue	0.6381939307	3	4	9.15	-0.297391977	TRAIN	3	9.00717	0.14283

**Remaining parameters in Appendix*

Regression Trees

The regression tree is created using RSS (variance split criteria) and cost-complexity pruning with a mx-depth=5.The total observations(189766) are divided into Training(132786) and Test set(56980) .

Number of Observations Read	189766
Number of Observations Used	189766
Number of Training Observations Used	132786
Number of Validation Observations Used	56980

Model Information	
Split Criterion Used	Variance
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	5
Maximum Tree Depth Achieved	5
Tree Depth	5
Number of Leaves Before Pruning	32
Number of Leaves After Pruning	31

Variable Importance

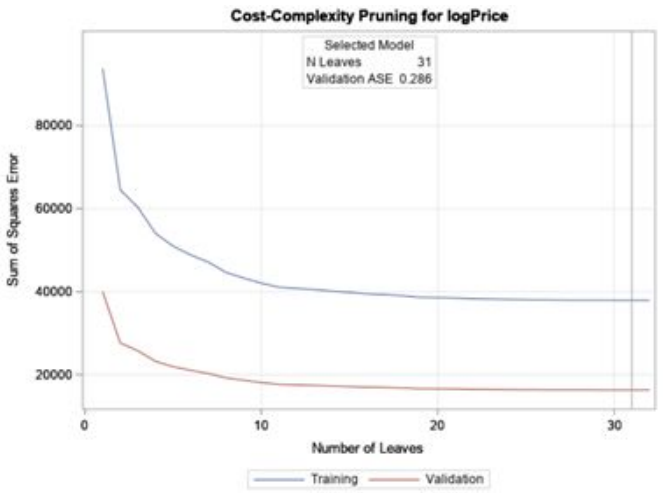
The variables used in the regression to predict price changes are: **Continuous**: year, Odometer **Categorical**: fuel, type, drive, title_status,paint_color,transmission**Nomina**l: condition, cylinders. From the variable Importance chart it can be inferred that the tree is growing using only 7 variables of the 10 used variables in the model. The variables paint_color and transmission are not causing any impact or change in the price. It can be inferred that one shouldn't be bothered about these 2 features for price comparisons when buying a used car and can consider the age of the car,no of cylinders, fuel type ,odometer reading, drive, type and condition of the car the most.

Variable Importance

	Training		Validation		Relative	
Variable	Relative	Importance	Relative	Importance	Ratio	Count
age	1.0000	171.2	1.0000	111.3	1.0000	1
cylinders	0.4664	79.8316	0.4567	50.8260	0.9792	3
odometer	0.4495	76.9287	0.4565	50.8052	1.0157	6
drive	0.4313	73.8134	0.4407	49.0466	1.0219	6
condition	0.3707	63.4497	0.3758	41.8186	1.0137	2
fuel	0.3272	56.0014	0.3187	35.4690	0.9741	5
type	0.2287	39.1501	0.2320	25.8179	1.0142	5
title_status	0.0698	11.9414	0.0679	7.5591	0.9736	2

Tree Pruning

The Regression tree is pruned on cost complexity criteria to 31 leaves with validation ASE(Average Squared error)=0.286.The model selects the optimal number of leaves in the tree where the ASE becomes constant and prevents the tree from overfitting.



Fit Statistics

The HPSPLIT Procedure

Fit Statistics for Selected Tree

	N Leaves	ASE	RSS
Training	31	0.2858	37947.7
Validation	31	0.2862	16309.4

The ASE for training data is 0.2858 which gives RMSE=0.534 and for testing ASE is 0.2862 which gives RMSE=0.536

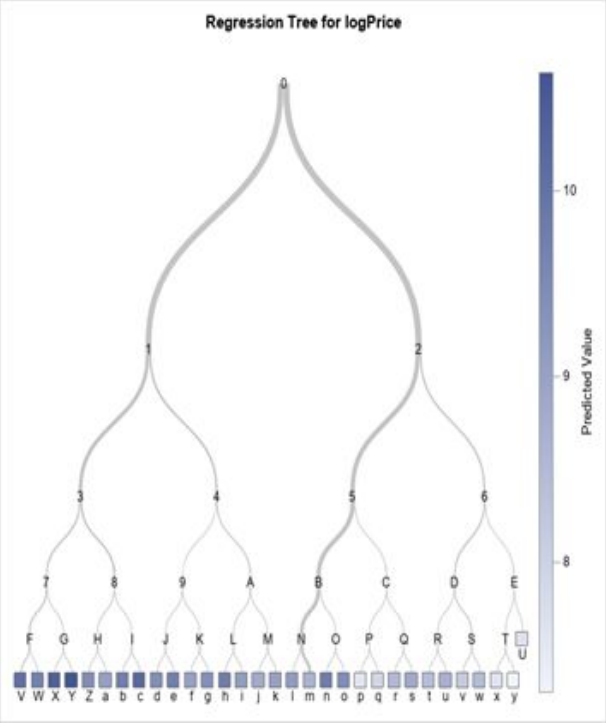
Train RMSE - 0.534

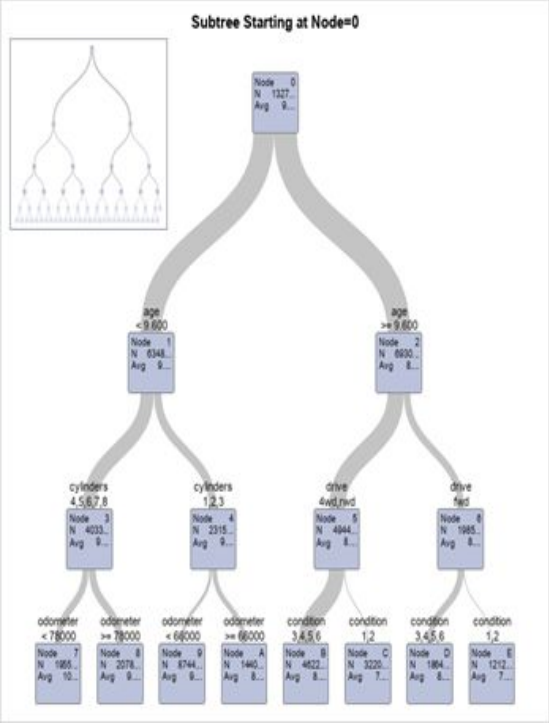
Test RMSE - 0.536

R-squared = 1-sse/sst=0.53

Tree Visualization

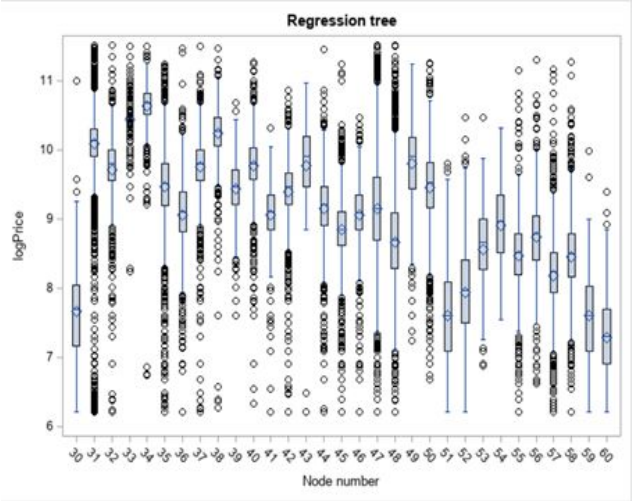
The model gives the option to view the whole selected tree after pruning with price prediction levels (The color bar on the side is indicating the price levels).Log transformation is done on price to eliminate skewness and to make it approximately normal. The model also gives the option to view a subtree/Zoomedtree.



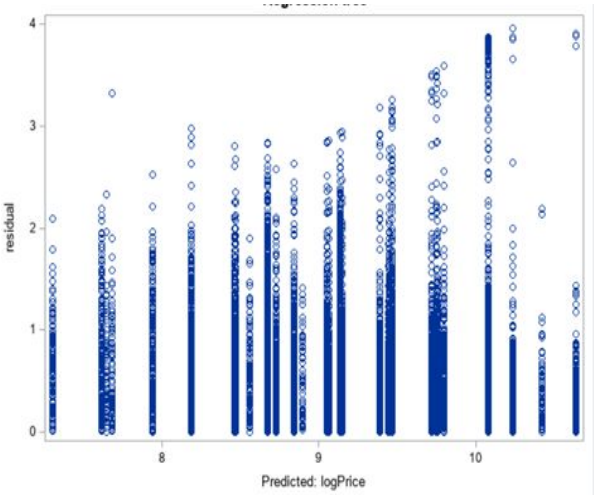


The price is higher for lower age cars, as the age increases the price decreases. The predicted prices of the used cars is higher for lower odometer readings and for the condition 3, 4, 5, 6(good, excellent ,like new ,new) and for cars with higher number of cylinders(4,5,6,8).

predicted price at final nodes



Residual vs predicted



Conclusion

Models/ Parameters	**R2**	**DOF**	**Train RMSE**	**Test RMSE**
Multiple Linear Regression ***(Proc Reg)***	60.2	50	0.5303	0.5332
Multiple Linear Regression ***(Proc GlmSelect no select)***	60.4	50	0.5270	0.5332
Multiple Linear Regression ***(Forward /Backward Selection)***	60.3	50	0.5288	0.5310
Stepwise Selection with Interaction Effects	60.58	55	0.5270	0.5290
Lasso Linear Regression	60.28	45	0.5289	0.5313
ElasticNet Regression	60.28	46	0.5290	0.5313
Polynomial Regression	69.98	51	0.4598	0.4634
Regression Trees	53		0.534	0.536

- Best Model is Polynomial Regression with Test RMSE 0.4634 and R2 = 69.98%
- We have similar results for regression analysis with proc reg and proc glmselect with no selection, forward selection and backward selection. This signifies that all the effects in the model are significant. Selecting Choose = cv for forward and backward resulted in lower test score for forward and backward selection.
- Using Regularization techniques few parameters have been eliminated. The R2 of the models remained the same which suggests that these parameters were not adding significant value.
- From logistic regression we got train accuracy of 63.17% and test accuracy of 63.11% . With 0.878 C-Statistic.

Sources

1. Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018, April). How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest. In Future of Information and Communication Conference (pp. 413–422). Springer, Cham.
2. https://statisticalhorizons.com/multicollinearity
3. https://blogs.sas.com/content/iml/2018/07/30/names-columns-design-matrix.html
4. https://support.sas.com/documentation/onlinedoc/stat/141/hpsplit.pdf

Appendix

Forward Selection Regression Results

Regresion with forward Selection

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 10.

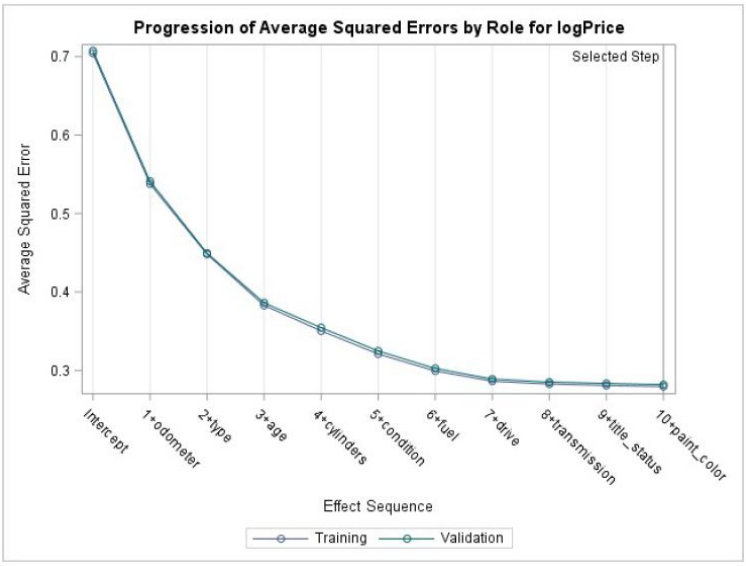
Effects:	Intercept condition cylinders drive paint_color title_status transmission type age odometer fuel

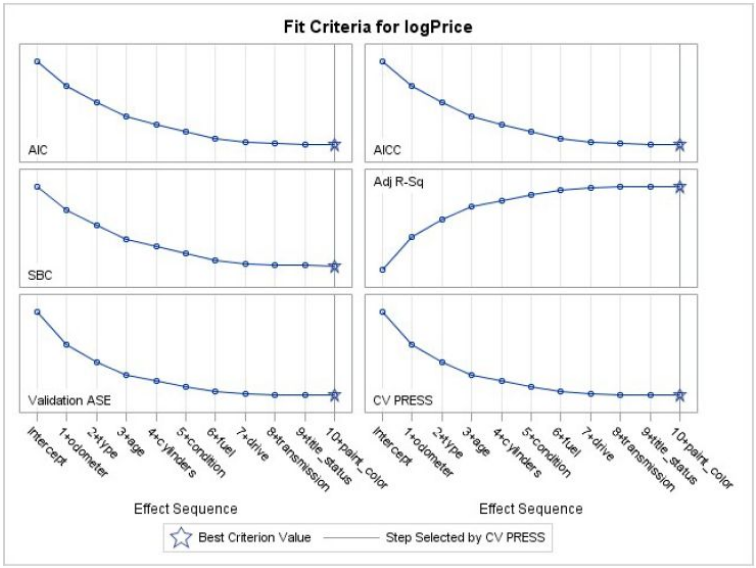
Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	56337	1126.73170	4029.27	<.0001
Error	132596	37079	0.27964		
Corrected Total	132646	93415			

Root MSE	0.52881
Dependent Mean	9.08870
R-Square	0.6031
Adj R-Sq	0.6029
AIC	-36327
AICC	-36327
SBC	-168477
ASE (Train)	0.27953
ASE (Validate)	0.28201
CV PRESS	37132

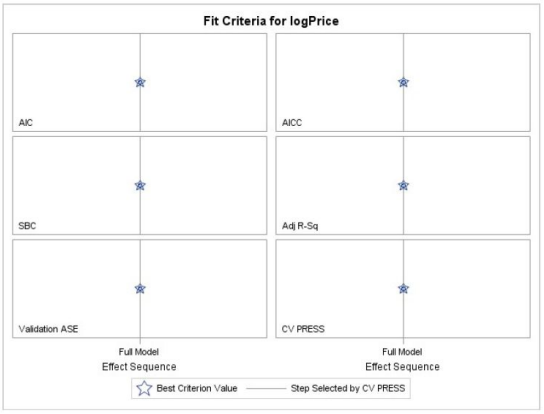


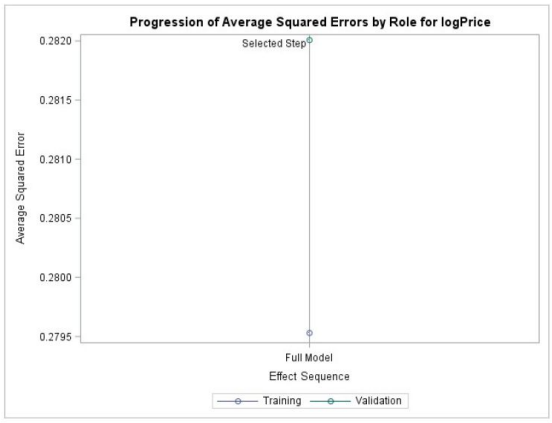


Backward Selection Regression Results

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	56337	1126.73170	4029.27	<.0001
Error	132596	37079	0.27964		
Corrected Total	132646	93415			

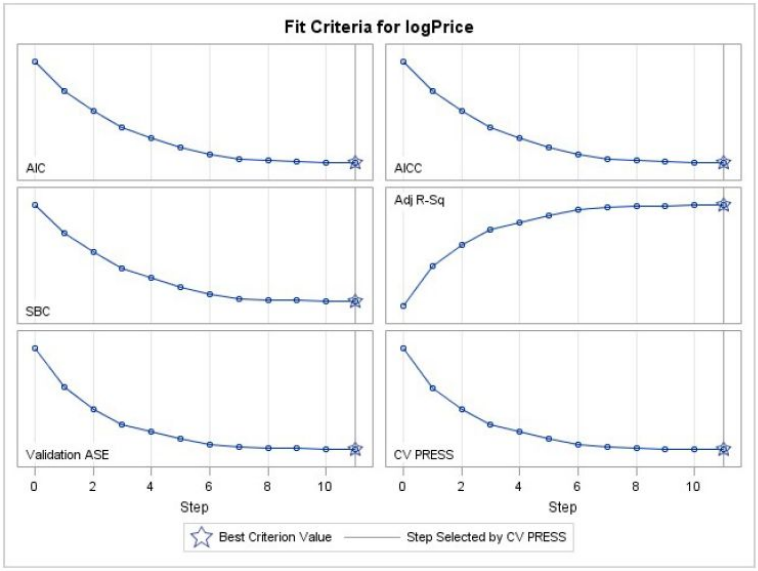
Root MSE	0.52881
Dependent Mean	9.08870
R-Square	0.6031
Adj R-Sq	0.6029
AIC	-36327
AICC	-36327
SBC	-168477
ASE (Train)	0.27953
ASE (Validate)	0.28201
CV PRESS	37132

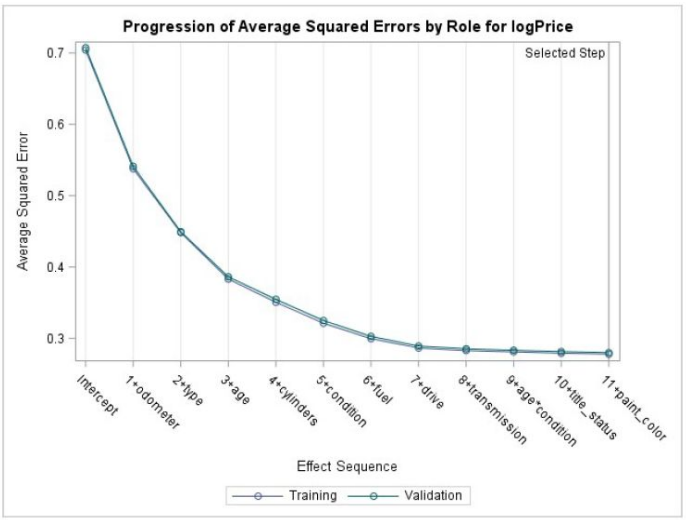




Stepwise Regression Results

Root MSE	0.52702
Dependent Mean	9.08870
R-Square	0.6058
Adj R-Sq	0.6056
AIC	-37222
AICC	-37222
SBC	-169322
ASE (Train)	0.27763
ASE (Validate)	0.27987
CV PRESS	36888





Lasso Regression Results

Root MSE	0.52897
Dependent Mean	9.08870
R-Square	0.6028
Adj R-Sq	0.6027
AIC	-36251
AICC	-36251
SBC	-168449
ASE (Train)	0.27971
ASE (Validate)	0.28225
CV PRESS	37125

The GLMSELECT Procedure

LASSO Selection Summary

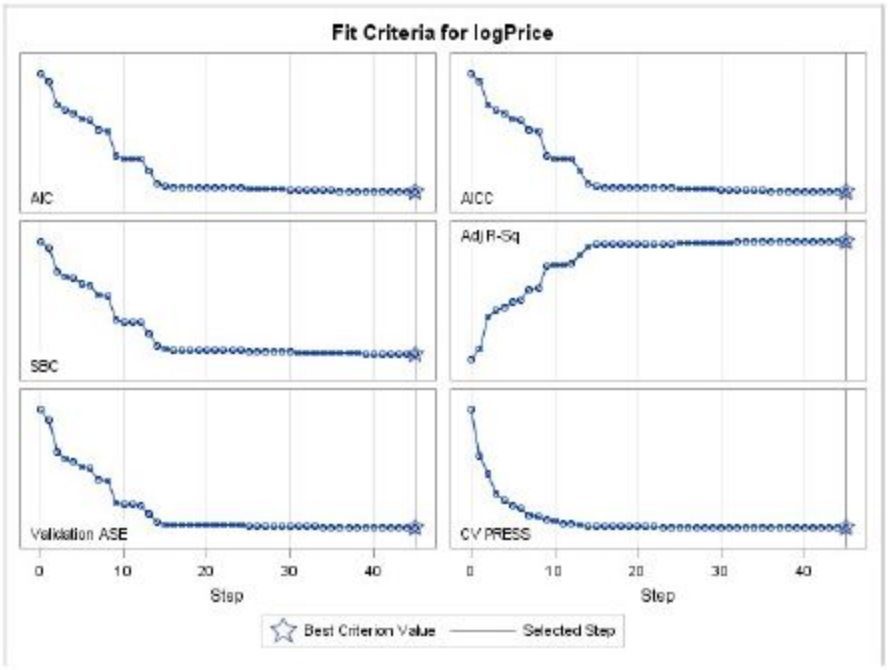
Step	Effect Entered	Effect Removed	Number Effects In	ASE	Validation ASE	CV PRESS
0	Intercept		1	0.7042	0.7073	93415.5414
1	odometer		2	0.6619	0.6649	71323.1013
2	age		3	0.5490	0.5522	62864.1198
3	cylinders_4		4	0.5271	0.5304	53220.1956
4	cylinders_3		5	0.5164	0.5197	50450.1200
5	condition_2		6	0.4954	0.4986	47663.1659
6	drive_fwd		7	0.4896	0.4928	46031.8111
7	fuel_gas		8	0.4550	0.4578	43401.3326
8	type_sedan		9	0.4477	0.4505	42329.4888
9	fuel_die		10	0.3690	0.3711	41070.4847
10	condition_3		11	0.3636	0.3657	40317.0050
11	drive_4wd		12	0.3630	0.3652	39176.2363
12	type_picku		13	0.3610	0.3631	38882.3639
13	type_truck		14	0.3307	0.3329	38685.9299
14	transmission_automatic		15	0.2977	0.3002	38072.8926
15	title_status_lien		16	0.2914	0.2939	37964.0442
16	paint_color_green		17	0.2901	0.2926	37906.9711
17	condition_4		18	0.2899	0.2925	37884.9849
18	paint_color_brown		19	0.2898	0.2924	37850.9993
19	title_status_clean		20	0.2897	0.2923	37826.7864
20	type_conve		21	0.2893	0.2919	37780.5493
21	title_status_rebui		22	0.2892	0.2918	37690.1573
22	type_hatch		23	0.2892	0.2918	37559.2221
23	cylinders_7		24	0.2890	0.2916	37506.5499
24	paint_color_silver		25	0.2878	0.2904	37479.5633
25	paint_color_black		26	0.2873	0.2899	37451.5670
26	type_coupe		27	0.2865	0.2892	37414.9726
27	paint_color_white		28	0.2855	0.2882	37391.8500
28	paint_color_blue		29	0.2854	0.2881	37387.1701
29	type_other		30	0.2852	0.2878	37374.8519
30	type_SUV		31	0.2843	0.2869	37309.9735
31	cylinders_1		32	0.2840	0.2866	37294.4204
32	cylinders_5		33	0.2830	0.2857	37253.0320
33	type_mini-		34	0.2830	0.2856	37229.6460
34	condition_5		35	0.2822	0.2849	37225.1442
35	type_offro		36	0.2819	0.2845	37189.4613
36	transmission_manual		37	0.2818	0.2845	37175.2929
37	condition_1		38	0.2816	0.2842	37173.0380
38	paint_color_orange		39	0.2809	0.2835	37158.2101
39	paint_color_purple		40	0.2805	0.2831	37154.5064
40	paint_color_red		41	0.2805	0.2831	37153.1664
41	paint_color_custom		42	0.2804	0.2830	37150.2006
42	type_van		43	0.2803	0.2829	37145.4280
43	cylinders_2		44	0.2803	0.2828	37145.1892
44	paint_color_grey		45	0.2800	0.2825	37144.4099
45	cylinders_6		46	0.2798	0.2824	37138.0267
46	title_status_parts		47	0.2797	0.2823*	37125.1107*

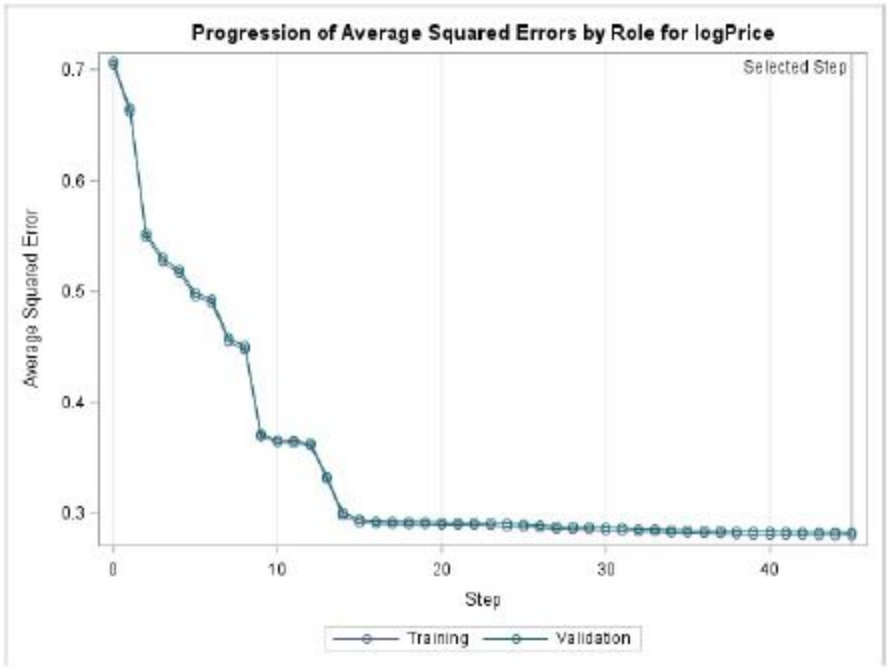
* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

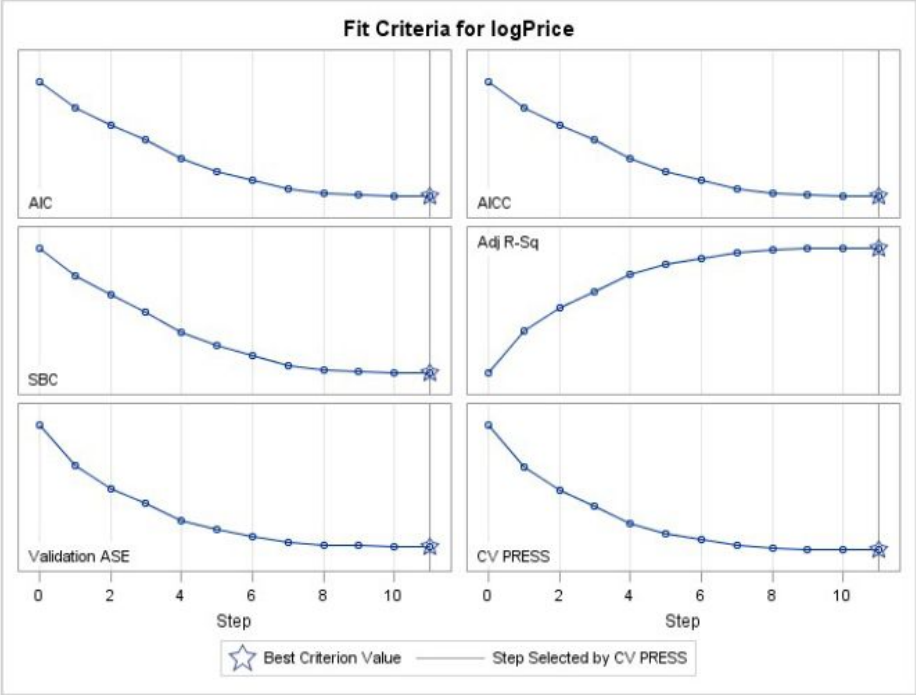
Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	fuel_hyb	37125.2968	>	37125.1107

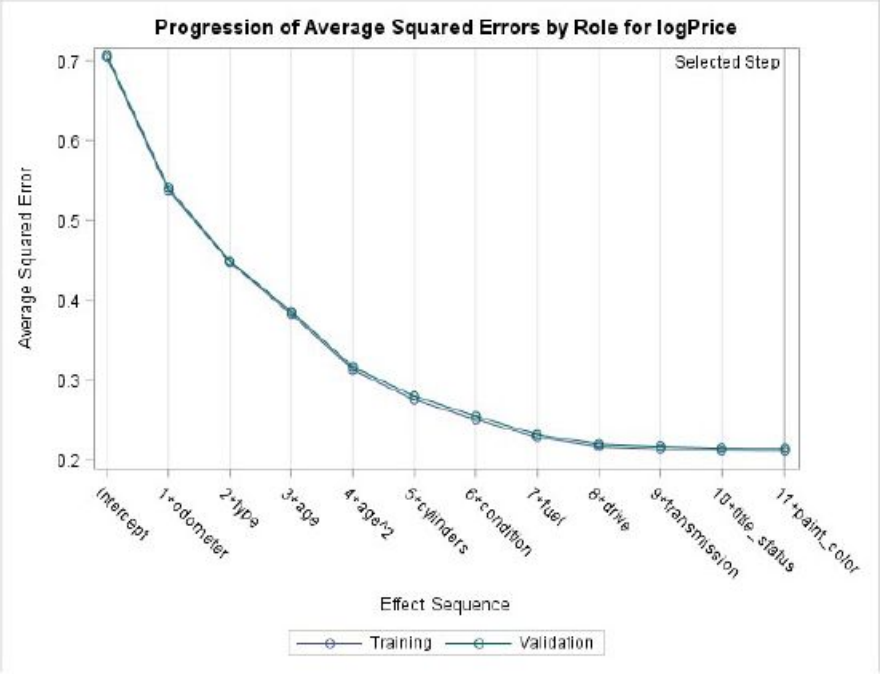

Fit Criteria for logPrice
AIC
AICC
SBC
Adj R-Sq
Validation ASE
CV PRESS
0
10
20
30
40
Step
Best Criterion Value
Selected Step


Progression of Average Squared Errors by Role for logPrice
Selected Step
Average Squared Error
0.7
0.6
0.5
0.4
0.3
0
10
20
30
40
Step
Training
Validation

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	8.084515
condition_1	1	0.078147
condition_2	1	0.696702
condition_3	1	0.184753
condition_4	1	0.038102
condition_5	1	0.100808
cylinders_1	1	0.250765
cylinders_3	1	0.159159
cylinders_4	1	0.238937
cylinders_5	1	0.182592
cylinders_6	1	0.130107
cylinders_7	1	-0.697311
drive_4wd	1	0.130677
drive_fwd	1	-0.186586
title_status_clean	1	0.044168
title_status_lien	1	0.239565
title_status_parts	1	-0.085622
title_status_rebui	1	-0.099219
paint_color_black	1	0.034019
paint_color_blue	1	-0.025579
paint_color_brown	1	-0.082579
paint_color_custom	1	0.027119
paint_color_green	1	-0.088967
paint_color_grey	1	-0.006716
paint_color_orange	1	0.113067
paint_color_purple	1	-0.065925
paint_color_red	1	-0.010623
paint_color_silver	1	-0.032234
paint_color_white	1	0.021692
transmission_automatic	1	-0.134089
transmission_manual	1	0.028555
type_SUV	1	-0.066397
type_conve	1	0.132196
type_coupe	1	0.052459
type_hatch	1	-0.177380
type_mini-	1	-0.099099
type_offro	1	0.221182
type_picku	1	0.083230
type_sedan	1	-0.187649
type_truck	1	0.079385
type_van	1	-0.028766
age	1	-0.275021
odometer	1	-0.300142
fuel_die	1	0.357974
fuel_gas	1	-0.232830

Polynomial Regression Results





Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.671138	0.043546	176.16	<.0001
condition 2	1	0.124914	0.026198	4.77	<.0001
condition 3	1	0.662026	0.007475	88.56	<.0001
condition 4	1	0.154560	0.002900	53.29	<.0001
condition 5	1	-0.013133	0.004263	-3.08	0.0021
condition 6	1	-0.021113	0.021048	-1.00	0.3158
cylinders 2	1	0.248502	0.031800	7.81	<.0001
cylinders 3	1	0.090651	0.013167	6.88	<.0001
cylinders 4	1	0.139230	0.013143	10.59	<.0001
cylinders 5	1	0.224716	0.003437	65.38	<.0001
cylinders 6	1	0.316179	0.016207	19.51	<.0001
cylinders 7	1	0.459833	0.053077	8.66	<.0001
cylinders 8	1	-1.212320	0.057863	-20.95	<.0001
drive 4wd	1	0.127822	0.002015	63.44	<.0001
drive fwd	1	-0.172658	0.002401	-71.92	<.0001
paint_color black	1	0.022938	0.004025	5.70	<.0001
paint_color blue	1	-0.020467	0.004623	-4.43	<.0001
paint_color brown	1	-0.047880	0.007567	-6.33	<.0001
paint_color custom	1	0.054858	0.008428	6.51	<.0001
paint_color green	1	-0.048237	0.007099	-6.80	<.0001
paint_color grey	1	-0.016695	0.004586	-3.64	0.0003
paint_color orange	1	0.054678	0.014613	3.74	0.0002
paint_color purple	1	-0.060295	0.021582	-2.79	0.0052
paint_color red	1	-0.005198	0.004603	-1.13	0.2587
paint_color silver	1	-0.008339	0.004241	-1.97	0.0493
paint_color white	1	-0.001965	0.003857	-0.51	0.6105
title_status clean	1	0.264058	0.013738	19.22	<.0001
title_status lien	1	0.386213	0.016604	23.26	<.0001
title_status missi	1	-0.370068	0.032711	-11.31	<.0001
title_status parts	1	-0.485212	0.059621	-8.14	<.0001
title_status rebui	1	0.085107	0.014693	5.79	<.0001
transmission automatic	1	-0.107726	0.002945	-36.58	<.0001
transmission manual	1	0.083101	0.004060	20.47	<.0001
type SUV	1	-0.068663	0.004511	-15.22	<.0001
type bus	1	0.004026	0.031181	0.13	0.8973
type conve	1	0.211709	0.008399	25.21	<.0001
type coupe	1	0.027764	0.006165	4.50	<.0001
type hatch	1	-0.165692	0.007759	-21.36	<.0001