Perspectives on Business Analytics
Prof. Dr. Gruetzemacher
Barton School of Business

WICHITA STATE UNIVERSITY

LINEAR REGRESSION MODEL ON PREWONED FORD CARS

Based on Data Set from Kaggle

Table of content

List of figure

1.INTRODUCTION

Shopping for a used car is like going on a treasure hunt. There are amazing deals out there, and with the emergence of the Internet as a car shopping tool, you have every chance of finding a good deal on a car that meets your needs and fits your budget.

There are plenty of incentives to buy used instead of new:

It will save you money on car insurance, registration, taxes and depreciation, which is the loss in a car's value due to wear and tear over time. It also makes sense because cars have never been more reliable. It's not unusual for some vehicles to be trouble-free for well over 100,000 miles.

To buy a good used car here are some ground rules that we all need to follow:

- ✓ Set your budget
- ✓ Choose the right car
- ✓ Check for alternative cars
- ✓ Price the car with the grading of Year, Model, Mileage and Vehicle condition.
- ✓ Check whether the car is clean title.
- ✓ Test drive the car
- ✓ Inspect the car
- ✓ Negotiate the best price & Close the deal

So, we have taken a data set of ford used cars with different car models which will be repurchased by the customers in the future. Our goal is to develop a Linear Regression Model in order to see, how the other variables affect the price of the pre-owned ford cars in order to predict the future purchase. Therefore, the data analysis report represents the process of developing this model. To make statements about the accuracy and generalizability of the model, the data set, and its variables as well as the process of cleaning the data and the method used are described in Chapter 2. Chapter 3 shows the results, Chapter 4 discusses the results and their limitations and, finally, Chapter 5 summarizes the conclusions.

2.METHOD

The following section briefly describes the data set used to develop the model, the process of pre-processing the data, and the methodology used to build and identify the most accurate model.

Data set:

Our data set consist of 1000 entries with one dependent variable and 7 independent variables. The Year ranges from 2013 to 2020 (More than 5 years' data). Where the Categorical variables are Model, Transmission and fuel type, later we have created dummy variables for each variable in which we got 18 models, 3 transmission and 3 fuel types. Hence the total dummy variables here are 24.

1	Model	Year	Price	Transmissior	Miles Travelled	Fuel Type	MPG	Purchased Car
2	Fiesta	2017	12000	Automatic	15944	Petrol	57.7	0
3	Focus	2018	14000	Manual	9083	Petrol	57.7	0
4	Focus	2017	13000	Manual	12456	Petrol	57.7	0
5	Fiesta	2019	17500	Manual	10460	Petrol	40.3	1
6	Fiesta	2019	16500	Automatic	1482	Petrol	48.7	1
7	Fiesta	2015	10500	Manual	35432	Petrol	47.9	0
8	Puma	2019	22500	Manual	2029	Petrol	50.4	1
9	Fiesta	2017	9000	Manual	13054	Petrol	54.3	1
10	Kuga	2019	25500	Automatic	6894	Diesel	42.2	0

Figure 1 Data set from kaggle.

Data pre-processing:

We have run the data in tableau to check whether the data have any errors or missing values but we found nothing, we believe that the data is clean and suitable for our analysis.

Methodology:

In order to develop the flexible regression model, we have taken the data from Kaggle site, for buying and selling used cars. The price is the main factor people consider,

what are the features the used cars will have with reasonable pricing? So, we considered the price as the Dependent variable and all other variables as independent variables. Here

we used a linear regression model in SPSS to know the correlation between the price and all other variables.

The Linear regression analysis is used to predict the value of the Price variable based on other variables' values.

Y= a + b1X1 + b2X2 + b3X3 + b4X4 …….. + bnXn

Several dummy variables also need to be created to capture the dataset's categorical variables properly. Here in the data set, the categorical variables are Model, Fuel Type, Transmission.

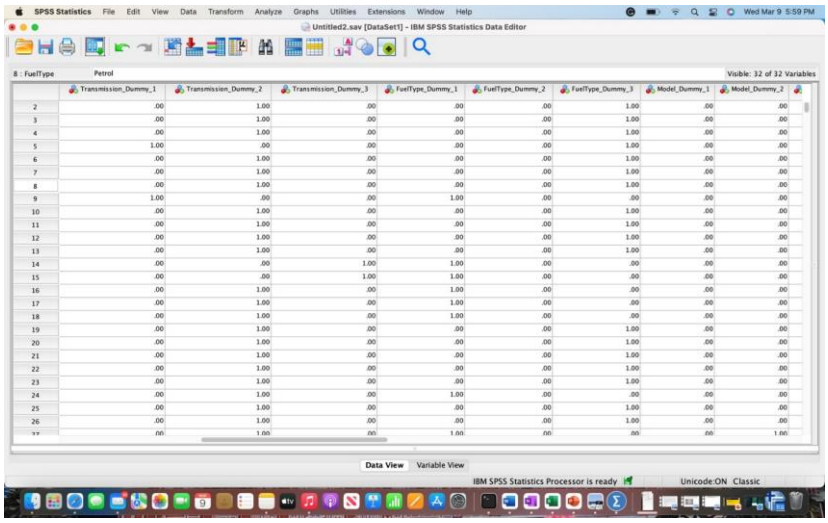

Figure 2 for Model, Transmission and fuel type Dummy variables

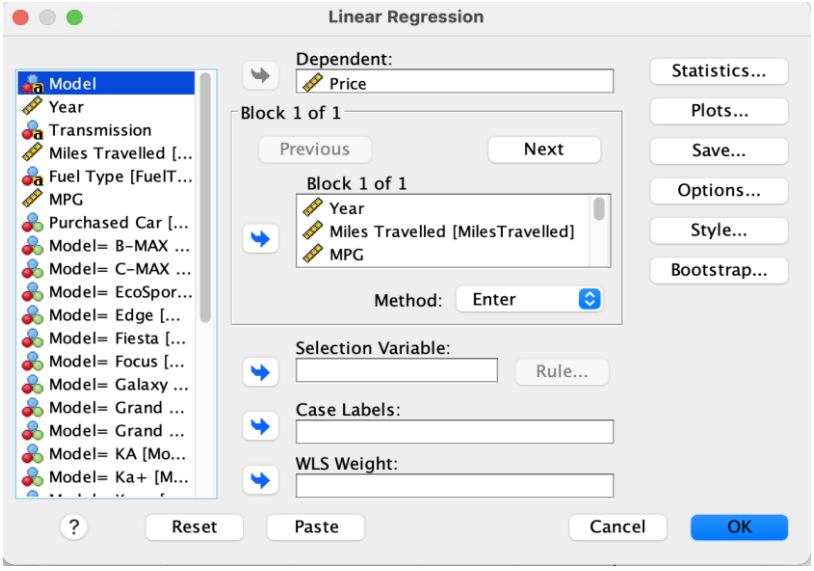

Figure 3 Linear Regression SPSS.

3.RESULTS

In results from SPSS output we see that the R square is 86.1%, Adjacent R square value is 85.7%. Here is the figure 4 given below.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.928[a]	.861	.857	1738.370

a. Predictors: (Constant), PurchasedCar=1.0, MPG, Model= KA, Model= Grand Tourneo Connect, Model= Tourneo Connect, Model= Puma, FuelType=Hybrid, Model= S-MAX, Model= Galaxy, Model= Grand C-MAX, Model= C-MAX, Model= Ka+, Model= Edge, Model= Tourneo Custom, Model= EcoSport, Transmission=Semi-Auto, Miles Travelled, Model= Mustang, Model= Mondeo, Model= Focus, Model= Kuga, Year, Transmission=Manual, FuelType=Petrol, Model= Fiesta

Figure 4 Model 1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.819E+10	25	727424217	240.715	.000[b]
	Residual	2.943E+9	974	3021930.89		
	Total	2.113E+10	999			

a. Dependent Variable: Price

b. Predictors: (Constant), FuelType=Petrol, Purchased Car, Model= Focus, Model= Puma, FuelType=Hybrid, Model= KA, Model= Tourneo Connect, Model= Mustang, Model= Galaxy, Model= Tourneo Custom, Model= Grand Tourneo Connect, Transmission=Semi-Auto, Model= Edge, Model= S-MAX, Model= Grand C-MAX, Model= Ka+, Model= C-MAX, Model= EcoSport, Year, Model= Mondeo, MPG, Miles Travelled, Transmission=Manual, Model= Kuga, Model= Fiesta

Figure 4 Model 2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-2413232.6	95593.553		-25.245	<.001
	Year	1207.818	47.385	.434	25.490	<.001
	MPG	-139.628	9.796	-.296	-14.254	<.001
	Miles Travelled	-.061	.005	-.218	-12.909	<.001
	Purchased Car	-21.912	111.529	-.002	-.196	.844
	Model= C-MAX	-14.161	506.458	-.001	-.028	.978
	Model= EcoSport	-158.911	474.869	-.009	-.335	.738
	Model= Edge	7150.782	865.923	.120	8.258	<.001
	Model= Fiesta	398.453	439.127	.041	.907	.364
	Model= Focus	2568.734	438.633	.251	5.856	<.001
	Model= Galaxy	6077.172	1100.809	.072	5.521	<.001
	Model= Grand C-MAX	-223.368	623.877	-.006	-.358	.720
	Model= Grand Tourneo Connect	1783.681	853.272	.030	2.090	.037
	Model= KA	-1056.125	885.305	-.016	-1.193	.233
	Model= Ka+	-3993.034	539.524	-.146	-7.401	<.001
	Model= Kuga	2578.521	473.118	.204	5.450	<.001
	Model= Mondeo	2698.287	665.146	.066	4.057	<.001
	Model= Mustang	16285.005	841.345	.295	19.356	<.001
	Model= Puma	6276.110	1102.315	.075	5.694	<.001
	Model= S-MAX	6862.880	745.087	.141	9.211	<.001
	Model= Tourneo Connect	816.104	892.788	.013	.914	.361
	Model= Tourneo Custom	4412.055	1035.168	.061	4.262	<.001
	Transmission=Manual	-816.463	295.784	-.055	-2.760	.006
	Transmission=Semi-Auto	-863.221	365.531	-.045	-2.362	.018
	FuelType=Hybrid	1994.050	1840.647	.014	1.083	.279
	FuelType=Petrol	-2600.072	224.007	-.272	-11.607	<.001

a. Dependent Variable: Price

Figure 4 Model 3

Here we observe that 15 variables are significant i.e. Constant, Year, MPG, Miles travelled, Model Edge, Focus, Galaxy, Ka+, Kuga, Mondeo, Mustang, Puma, S-Max, Tourneo Custom, Petrol.

Y = -2413232.6 + 1207.81(Year) – 0.061(Miles travelled) – 21.912(Purchased car) – 14.161(C-Max) - 158.911(Eco sport) + 7150.782 (Edge) + 398.453(fiesta) + 2568.734(Focus) + 6077.172(galaxy) - 223.368(grand C-MAX) + 1783.681(grand tourneo connect) - 1056.125(KA) - 3993.034(Ka+) + 2578.521(Kuga) + 2698.287(Mondeo) + 16285.0(puma) + 6282.8(s-max) + 816.104(tourneo connect) + 44112.0(tourneo connect) - 816.463(Manual) - 863.221(semi-auto) + 1994.050(hybrid) - 2600.072(petrol)

4.DISCUSSION

The Model R square is higher which is 86.1%, we believe it is the good sign for this model but we still need some more external factors or variables so that the data can be more reliable

The external factors could be adding up to the independent variables such as Title type, Rating or car score, Frequency of data and many more could be added to get the proper pricing value for the car.

Some of the variables were negatively correlated with the constant which show the negative trend and some of the other variables are positively correlated which shows the positive trend with the constant.

For better understanding about the data, we ran the data in Bar graphs SPSS by choosing specific variables. Here is the figure 5 Bar Graphs

- X axis is MPG
- Y axis is Mean price

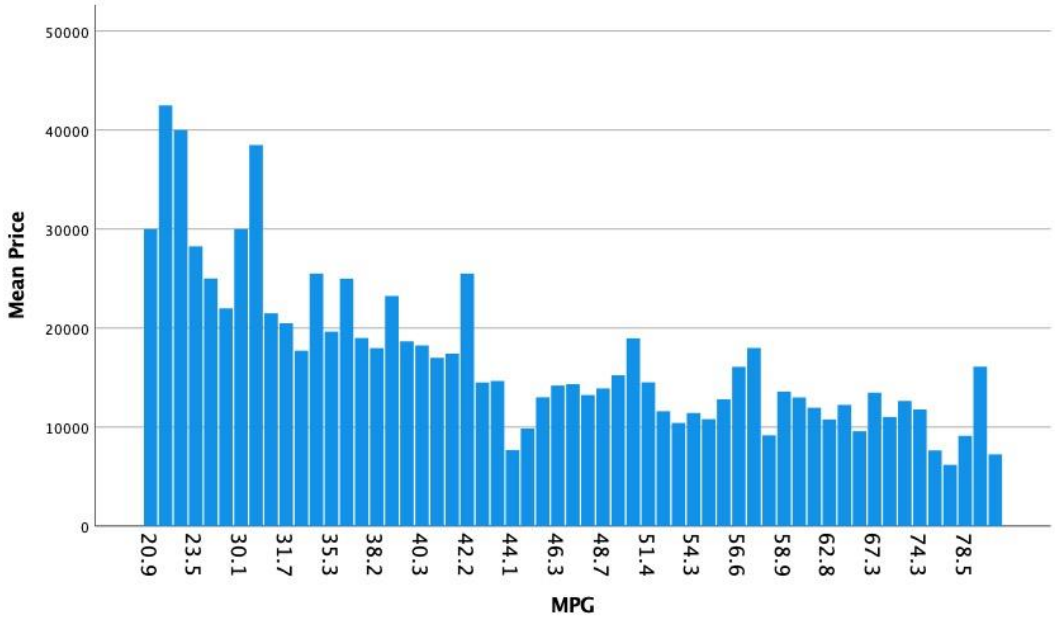

Figure 5 Bar graph 1

We observe that with the decrease of the Price, MPG is increasing, both the variables are correlated.

- X axis is MPG
- Y axis is Mean price

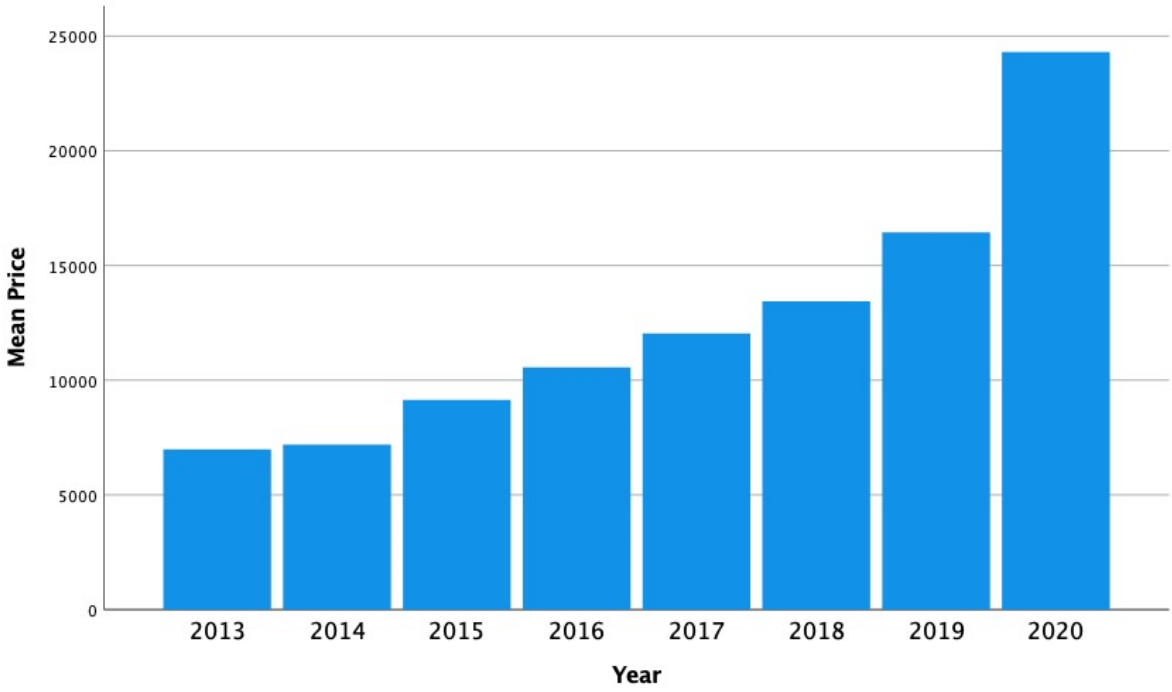

Figure 5 Bar graph 2

Similarly, with the Increase of the year's model, the price also increases

5.PRESCRIPTIVE ANALYSIS

Prescriptive analytics factors information about possible situations or scenarios, available resources, past performance, and current performance, and suggests a course of action or strategy. It can be used to make decisions on any time horizon, from immediate to long term.

Based on our understanding, how the other variables effecting the price. We observe LP predicts that the model Edge, Mustang and S- Max have been purchased often when compare to other models. So, we could analyze in the future that most likely there can be more purchases for these types of Models

Therefore, to increase the sales with better price range ford should consider the variables such as MPG, Transmission, Fuel type.

6.CONCLUSION

This project sets out to determine that running a linear regression model with 1000 entries of data with one dependent variable and 7 independent variables. Year ranges from 2013 to 2020 (More than 5 years' data). Where the Categorical variables are Model, Transmission and fuel type.

Despite the limitation discussed in previous chapter with the variables to be added for more reliable outcome, Our R square value is 86.1% and Adjacent R square value is 85.7%. The independent variables correlate with the dependent variable which is a good sign for both the buyers as well as the sellers.

Few models such as; Mustang, Puma, Kuga, S-Max, Galaxy, Focus, Edge are highly significant from figure 4 model 3 Which can be considered as most preferred vehicles in the market

As we already discussed in chapter 4, fig 5 bar graphs that as the price is decreasing the MPG is increasing, with the increase of the year models similarly the price of the vehicle also increasing.

Therefore, we conclude and predict that the sellers have the true car price value for the buyers in the Market.

REFERENCE

Text Book: Business analytics data analysis and decision making by S. Christian Albright Wayne L. Winston.

https://www.investopedia.com/terms/p/prescriptive-analytics.asp

https://www.nerdwallet.com/article/loans/auto-loans/buy-used-car

https://www.kaggle.com/datasets