DIAMOND PRICES

PROJECT ON PREDICTING THE DIAMOND PRICE FOR BID

PROJECT OVERVIEW

In this project we are going to help our company in predicting the prices for the diamonds. The Company wants to bid for the diamond but we don't have the right information for the correct price to bid. So, for that reason we are going to build a predictive model using ALTERYX to predict the price for the diamonds.

PROJECT DETAILS

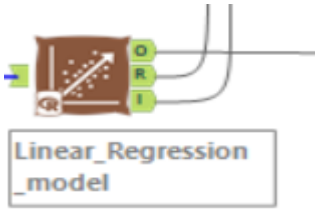
A diamond distributor has recently decided to exit the market and has put up a set of 3,000 diamonds up for auction. Seeing this as a great opportunity to expand its inventory, a jewelry company has shown interest in making a bid. To decide how much to bid, the company's analytics team used a large database of diamond prices to build **a linear regression model** to predict the price of a diamond based on its attributes. I, as the business analyst, am tasked to apply that model to make a **recommendation** for how much the company should bid for the **entire set of 3,000 diamonds.**

Data we are going to use for model **training is DIAMONDS.CSV** and for the **prediction NEW-DIAMONDS.CSV**

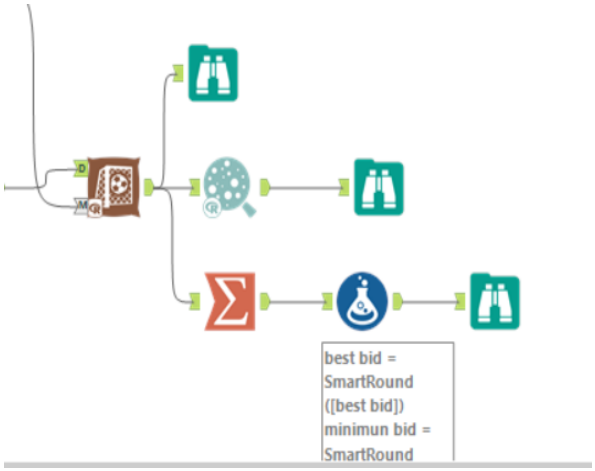
UNDERSTANDING MODeL AND QUESTIONS FROM IT

STEPS PERFORMED IN ALTERYX

- The very step was to load the file or the data set in which we are supposed to do all the necessary adjustment and RUN the linear regression model to get the desired results.
- We used an input data tool to get the data loaded.
- 

- After that to change the data type of variables we used a select tool to set data type from string to double so that we can perform our modeling
- 
- After that we used a linear regression tool to perform the regression on the data and in the configuration window we set the target variable PRICE and other variables as independent variables and connected the browse tool to the output end .
- 

- Then we loaded our new data set called new_diamond for predicting the price of diamonds and did the same thing with that connected the select tool in it to change the variable type to double.
- Then we used a score tool from a predictive tool set. This tool helped us in using the model for predicting the diamond prices to bid. And connected the ends of the score tool to summarize the tool to the total of prices of all the 3000 diamonds.

- 

- We also used a scatter plot to see the visual representation of the relationship between price and carat and at the end after summarize tool we used formula tool to round off the predicted prices of diamond.

Record	Field_1	carat	cut	cut_ord	color	clarity	clarity_ord	price
1	1	0.51	Premium	4	F	VS1	4	1749
2	2	2.25	Fair	1	G	I1	1	7069
3	3	0.7	Very Good	3	E	VS2	5	2757
4	4	0.47	Good	2	F	VS1	4	1243
5	5	0.3	Ideal	5	G	VVS1	7	789
6	6	0.33	Ideal	5	D	SI1	3	728
7	7	2.01	Very Good	3	G	SI1	3	18398
8	8	0.51	Ideal	5	F	VVS2	6	2203
9	9	1.7	Premium	4	D	SI1	3	15100
10	10	0.53	Premium	4	D	VS2	5	1857

In the first 10 rows of the data we got 7 important variables: carat, color,cut, cut_ord(this column represents the values of cut column in numeric where fair, good, very good, premium, ideal represent by numbers from 1-5 respectively.

Clarity and clarity_ord are also the same where : I1, SI2, SI1, VS1, VS2, VVS2, VVS1, and IF are represented by numbers in clarity_ord from 1-8 and at last price.

After, running linear regression model on the data we got our report

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5255.2	30.320	-173.33	< 2.2e-16 ***
carat	8363.4	13.565	616.55	< 2.2e-16 ***
cut_ord	160.4	5.513	29.09	< 2.2e-16 ***
clarity_ord	457.8	3.901	117.37	< 2.2e-16 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1348 on 49996 degrees of freedom
Multiple R-squared: 0.8862, Adjusted R-Squared: 0.8862
F-statistic: 129779 on 3 and 49996 degrees of freedom (DF), p-value < 2.2e-16

Which clearly says that all the 3 variables are significant to use and also the R-Squared is 0.8862 which is very nice and we can use this report to predict the further prices.

The equation we got from the report for calculating prices is

Prices = -5255.2 + 8363.4 * **Carat** + 160.4 * **cut_ord** + 457.8 * **clarity_ord**

According to the model, if a diamond is 1 carat heavier than another with the same cut, how much more should I expect to pay? Why?

To answer this question we will use the equation we got from the report

Prices = -5255.2 + 8363.4 * **Carat** + 160.4 * **cut_ord** + 457.8 * **clarity_ord**

As the other information regarding the diamond is not present we can assume that as default.

Price = -5255.2 + 8363.4 *2 + 160.4 + 457.8

Price = $12089.8

$**3726.4** you have to more (12089.8-3726.4)

If you were interested in a 1.5 carat diamond with a Very Good cut (represented by a 3 in the model) and a VS2 clarity rating (represented by a 5 in the model), how

much would the model predict you should pay for it?

In this question we got the value of other variables too and now we are going to fit these values in our equations to get prices

Prices = -5255.2 + 8363.4 * **Carat** + 160.4 * **cut_ord** + 457.8 * **clarity_ord**

Price = -5255.2 + 8363.4 * 1.5 + 160.4* 3 + 457.8 * 5

Price = **$10060.1**

What price do you recommend the jewelry company to bid? Please explain how you arrived at that number.

best bid	minimun bid	maximum bid
11750000	3800000	19750000

So I used a 95% confidence interval in the model to get the three values of prices and can be used as minimum , maximum and best fit.

The minimum price we can bid for the diamonds is $3800000

The maximum bid can go for $19750000

Best bid $11750000.

A FULL REPORT