أكاديمية مسك
MISK ACADEMY

UDACITY

TECH LAB

Predictive Analytics for Business

Project#1 Predicting Diamond Prices

Name: Marwan Saeed Alsharabbi

Date: 29-11-2019

2019

INTRODUCTION

Predicting Diamond Prices

This project is designed for three main reasons:

- To give you a feel for what you'll be doing throughout the Nanodegree Program
- To introduce you to Udacity's project submission and review process
- To make sure you feel comfortable with the basics before you begin. If it feels too easy, don't worry. We have some great stuff in store for you.

Project Overview

A jewelry company wants to put in a bid to purchase a large set of diamonds, but is unsure how much it should bid. In this project, you will use the results from a predictive model to make a recommendation on how much the jewelry company should bid for the diamonds.

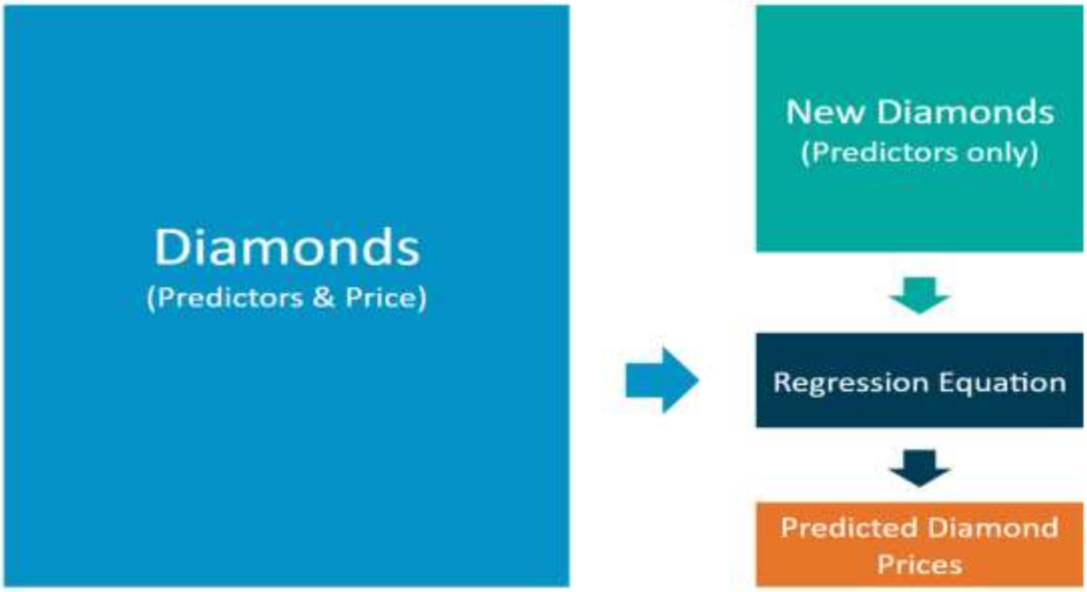
US Number System

All numbers that will be presented in this Nanodegree program will be based on the US numbering system where 5,269 is "five thousand two hundred sixty nine" and 158.1 is "one hundred fifty eight point one" where 1 is a decimal number. This is **very** important so please take note of this.

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. You, as the business analysts, are tasked to apply that model to make a recommendation for how much the company should bid for the entire set of 3,000 diamonds.

The following diagram represents the analysis at a high level. Since the model is already built, your analysis will focus on the right side of the diagram.



Diamonds
(Predictors & Price)

New Diamonds
(Predictors only)

Regression Equation

Predicted Diamond Prices

The linear regression model provides an equation that you can use to predict diamond prices for the set of 3,000 diamonds. The equation is below:

Price = -5,269 + 8,413 x **Carat** + 158.1 x **Cut** + 454 x **Clarity**

Let us understand what a linear regression procedure is Linear regression attempts to fit a linear relationship between a variable of interest, and a set of variables that may be related to the variable of interest.

I will introduce regression analysis, and have an overview of regression across four elements.

Overview of Regression

- ❖ Modeling *Developing a regression model.*
- ❖ Estimation *Using software to estimate the model.*
- ❖ Inference *Interpreting the estimated regression model .*
- ❖ Prediction *Making predictions about the variable of interest*

The first steps for a legitimate analysis Predicting Diamond Prices

Step 1 – Understand the data: There are two datasets.

- diamonds.csv contains the data used to build the regression model.
- new_diamonds.csv contains the data for the diamonds the company would like to purchase.

Both datasets contain carat, cut, and clarity data for each diamond. Only the diamonds.csv dataset has prices. You'll be predicting prices for the new_diamonds.csv dataset.

- Carat represents the weight of the diamond, and is a numerical variable.
- Cut represents the quality of the cut of the diamond, and falls into 5 categories: fair, good, very good, ideal, and premium. Each of these categories are represented by a number, 1-5, in the Cut_Ord variable.
- Clarity represents the internal purity of the diamond, and falls into 8 categories: I1, SI2, SI1, VS1, VS2, VVS2, VVS1, and IF. Each of these categories are represented by a number, 1-8, in the Clarity_Ord variable.
- Note: Transforming category variables to ordinal variables like this is not always appropriate, but we've done it here for simplicity.

Step 2 – Calculate the predicted price for diamond:

For each diamond, plug in the values for each of the variables into the linear model (equation). Then solve the equation to get the estimated, or predicted, diamond price. We suggest using a spreadsheet tool like Excel, Numbers, or Google Sheets. You could also do it in Alteryx and/or Tableau if you already have a license. If you don't have a license yet, you'll receive one after your free trial

Step 3 – Make a recommendation:

Now that you have the predicted price for each diamond, it’s time to calculate the bid price for the whole set. Note: The diamond price that the model predicts represents the final retail price the consumer will pay. The company generally purchases diamonds from distributors at 70% of that price, so your recommended bid price should represent that.

I will not give more details on how to create the model because it already exists will carry out the required steps for the project

Step 1: Understanding the Model

Answer the following questions:

1. 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?

   The formula created by the linear regression model is: Price = -5,269 + 8,413 x Carat + 158.1 x Cut + 454 x Clarity. The coefficient for Carat is 8,413, therefore if cut and clarity are the same in the two diamonds, and the first element is a constant, one carat heavier will increase the price in $8,413. That is, for each additional carat, holding the rest of properties equal, the price will increase by the amount of the Carat coefficient.

   The estimated value of this coefficient is a negative 5,269

   The generic interpretation of the coefficient is for every diamond is increase 1 carat, the Price variable increases. All other variables remaining at the same level. So in this particular case, the interpretation translates to when the diamond is 1 carat heavier increases, which is 8,413, then the Price, will increase by the amount of the Carat coefficient the opposite is true. All other variables remaining in the same level. Also the cut_ord and clarity_ord has an effect for the diamond price coefficient.

2. 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?

   The formula is: Price = -5,269 + 8,413 x Carat + 158.1 x Cut + 454 x Clarity. So, entering the value of the corresponding coefficients in the formula we obtain the price for the specified diamond. The predicted price is

   Price = -5,269 + 8,413 x 1.5 + 158.1 x 3 + 454 x 5 = $10,094.80.

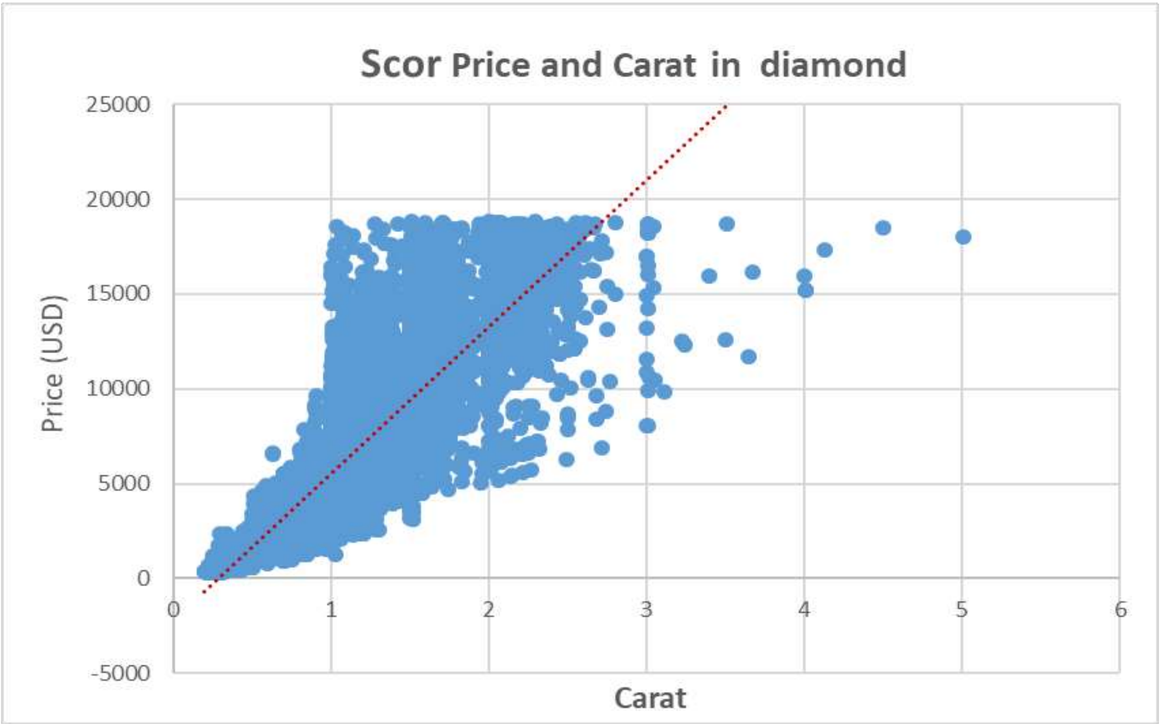
   Price for the specified the diamond in this Model = **$10,094.80**

Step 2: Visualize the Data

Make sure to plot and include the visualizations in this report. For example, you can create graphs in Excel and copy and paste the graphs into this Word document.

1. Plot 1 - Plot the data for the diamonds in the database, with carat on the x-axis and price on the y-axis.
2. Plot 2 - Plot the data for the diamonds for which you are predicting prices with carat on the x-axis and predicted price on the y-axis.
   - **Note**: You can also plot both sets of data on the same chart in different colors.
3. What strikes you about this comparison? After seeing this plot, do you feel confident in the model’s ability to predict prices?

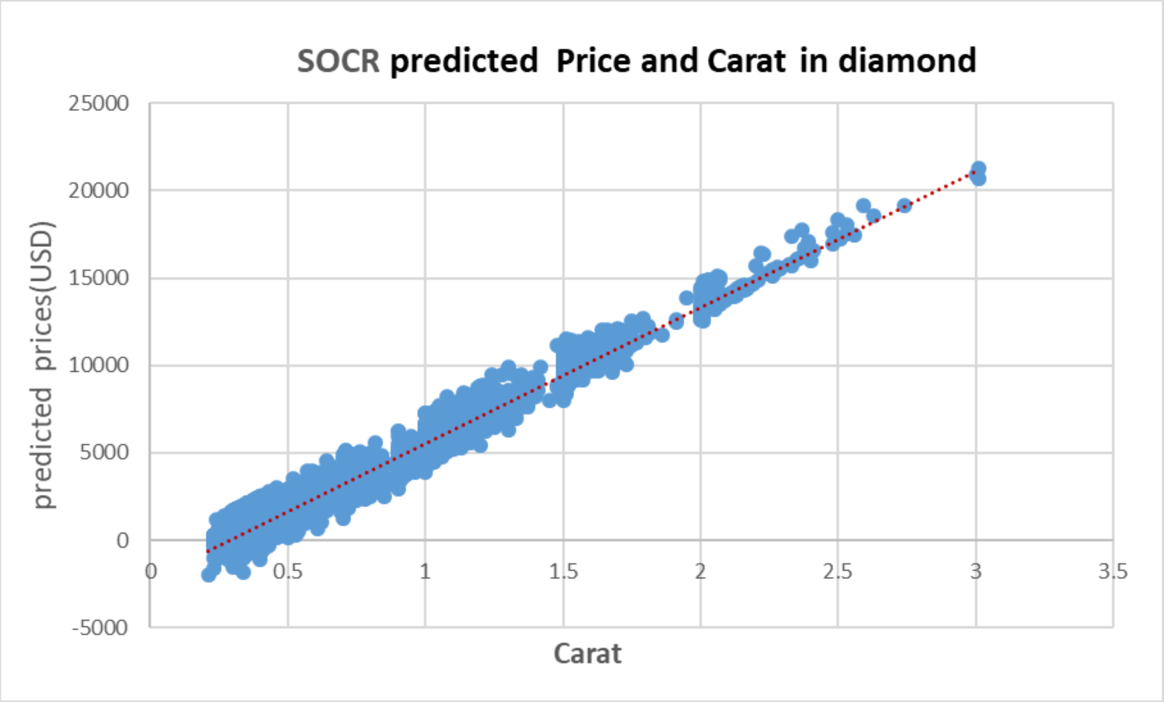
Plot 1 - Plot the data for the diamonds in the database, with carat on the x-axis and price on the y-axis.



Plot 1

Plot 1: Scatter plot of the price for 50,000 diamonds of the diamonds.csv the data used to build the regression model, not only carats are taking into account, but cut, clarity and a negative constant. So, we are not considering the necessary factors to account for all calculating measures of Center, measures of Spread and predict the prices. Moreover, even if we considered all these factors in the formula, in my opinion other characteristics influence the price of a diamond I see the Correlation high positive.

Plot 2 - Plot the data for the diamonds for which you are predicting prices with carat on the x-axis and predicted price on the y-axis.



Plot 2

Plot 2: Scatter plot of the predicting price for 50,000 diamonds and carat
I used the new_diamonds.csv the data used to calculate predicting prices for the formula in regression model shows that the predicted prices in **plot2** in each interval of number of carats are much more compact than the known prices of the already sold diamonds in **plot 1** The model gave the best fit line but sometimes it underestimated the price of the diamonds and even gave negative prices for those diamonds with less desirable characteristics. But, although the model less predicted the value of many diamonds, the prediction of the total price to pay for the 3,000 diamonds seems to be accurate (good average. Correlation is Positive when the values increase together

Step 3: Make a Recommendation

Answer the following questions:

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

Based on the predictive model's findings, and taking into account the jewelry company Generally purchases diamonds from distributors at 70% of the retail price, I used the formula from the linear regression model (based on the characteristics and prices of the previous diamond sales), applied it to the 3,000 diamonds and then I added up all those predicted prices. Finally, I multiplied that amount ($11733522.76) by 0.70 to get the final predicted bid of $ 8213465.932 bid for the new set of 3,000 diamonds is ~ $ 8213465.932

I wish success to all.

Marwan Saeed Alsharabbi