Predicting Diamond Prices

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.

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, you will use a large database of diamond prices to build a model to predict the price of a diamond based on its attributes. Then you will use the results of that model to make a recommendation for how much the company should bid.

Step 1 – Understand the data: There are two datasets. *diamonds.csv* contains the data used to build the regression model. *new_diamonds_new.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. In project zero, these categories were represented by an ordinal variable, 1-5. You can decide to use the ordinal or categorical variable.
- *Clarity* represents the internal purity of the diamond, and falls into 8 categories: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF (in order from least to most pure). In project zero, these categories were represented by an ordinal variable, 1-8. You can decide to use the ordinal or categorical variable.
- *Color* represents the color of the diamond, and is rated D through J, with D being the most colorless (and valuable) and J being the most yellow.

Step 2 – Build the model: In project zero, the results were provided, but now you get to calculate them. A few things are different this time around.

- You have more potential predictor variables
- You now know how to use categorical variables, so no need to rely only on ordinal variables.

Go through the steps you've learned through the course to build the model and come up with a regression equation.

IMPORTANT: When using Alteryx, you do not need to manually create dummy variables before building the model. If you select a categorical variable, like cut or clarity, then Alteryx will automatically create the dummy variables and give you the correct regression output.

Step 3 – Calculate the predicted price for diamond: For each diamond, plug in the values for each of the variables into the equation. Then solve the equation to get the estimated, or predicted, diamond price.

Step 4 – 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 the that price, so your recommended bid price should represent that.