predicting_diamond_prices_Report

November 13, 2018

1 Data Act Report

1.0.1 by Travis Gillespie

1.1 Table of Contents

Introduction
The purpose of this project is to get familiar with Udacity's project submission processes, plus understand some of the terms that will be used throughout the program. The visuals included in this report represent data gathered, assessed, cleaned, and graphed in the *predicting_diamond_prices.ipynb* file. This section analyzes the visualizations created during the data wrangling process.

First, I had to get familiar with both datasets. Then, predict the prices of diamonds using the new dataset. I was able to calculate the predicted prices using the given formula below for the regression model. I used python to build the predicted prices and bids columns, created two scatter plots, then wrote the dataframe out to a new csv file.

$$-5269 + (8413 \times \ Carat\ ) + (158.1 \times \ Cut\ ) + (454 \times \ Clarity\ )$$

Understand the Model, and Visualizing the Data
Insight One: Price Difference

- According to the linear model provided, if a diamond is 1 carat heavier than another with the same cut and clarity, how much more would the retail price of the heavier diamond be? Why?

    - $predictedPriceHeavyDiamond - \ predictedPriceLightDiamond = \$8413$

– The original calculation can be found in the file *predicting_diamond_prices.ipynb*. Also, a
mockup of these calculations is below.

```python
# Declare variables for rows in df that meet the following criteria

# carat = 1
# cut = "Premium"
# clarity = "SI1"
predictedPriceLightDiamond = df_predicted.loc[(df_predicted['carat'] == 1) &
                             (df_predicted['cut'] == 'Premium') &
                             (df_predicted['clarity'] == 'SI1')].iloc[0]['price']
predictedPriceLightDiamond




# carat = 2
# cut = "Premium"
# clarity = "SI1"
predictedPriceHeavyDiamond =  df_predicted.loc[(df_predicted['carat'] == 2) &
                                (df_predicted['cut'] == 'Premium') &
                                (df_predicted['clarity'] == 'SI1')].iloc[0]['price']
predictedPriceHeavyDiamond

# predictedPriceLightDiamond = 5138.4
# predictedPriceHeavyDiamond = 13551.4
# predictedPriceHeavyDiamond - predictedPriceLightDiamond = 8413
```

Insight Two: Visualization

Insight Three: Visualization

- 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), what retail price would
the model predict for the diamond?

  – $-5269 + (8413 \times \ 1.5\ ) + (158.1 \times \ 3\ ) + (454 \times \ 5\ )$
  – $= \$10,094.80$

- What strikes you about this comparison? After seeing this plot, do you feel confident in the
model's ability to predict prices?

  – This graph displays a positive correlation between carat and price. There are fewer
observations with diamonds that are 2 carats and greater, so it is difficult to determine
how close the predicted values would be to the line of best fit. I found it interesting that
some prices are below zero. Does this mean the seller would pay the buyer money for
taking the diamond off their hands? Obviously not, this would be a foolish business
model.

- There are probably some modifications that need to be made to the formula for this regression model. Another simple solution to cleaning up the dataset by removing rows that contain a negative number in the price column. As shown earler in the dataframe labaeled *df_predictedNegativePrices*, there are 291 rows that contain a negative price that could potentially be dropped from the dataset, or better yet, use these items to improve the algorithm.
- Regardless of the future modifications, I feel confident in the initial formula. The relationship between carat and predicted price is considered a strong positive linear relationship of appoximately 0.98. Since, bid price is 70% of predicted price, the corrilation coefficient between carat and bid is identical (i.e. 0.98).

Recommendation

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

    - As expained in the intro section, I was able to use the linear regression model to calculate the predicted prices.
    - I then calculated what 70% of the predicted prices sum value.
    - If the goal is to bid at 70% of the sum total price, my recommendation is to place a bid at $8,213,465.93 .

Resources

1. LaTex Math Symbols