ENOVA Data Scientist

Diamond Assignment

Michael Golaski

Variable Creation and Imputation

**Measurements**: I separated the Measurements column into 3 separate columns for length, width, and depth2 so that I would be able to work with actual continuous variables rather than a weird categorical variable

**Depth Imputed and Depth Imputed Flag**: I found the equation for Depth % online which was:

(Depth2/Width)\*100

I created a new column for Depth Imputed if the actual % wasn’t missing I used that, but for missing values I used the equation to fill in the blanks. I then created a column next to it as a flag to tell which values were imputed (1) and which were the original (0).

**Color\_Adj**: Based on research from multiple online sites I saw that diamonds are graded on color and there were messy values in the dataframe so I decided to bin the Color grade together as following:

|  |  |
| --- | --- |
| Grades | Binned Group |
| D, E, and F | Colorless |
| G, H, I, and J | NearColorless |
| K, L, and M | FaintColor |
| N-Z | ColoredDiamond |

Image 1

**Clarity\_Adj**: Just like Color\_Adj based on online research I saw that diamonds are graded on clarity and there were messy values in the dataframe so I decided to bin the Clarity grades together as well:

|  |  |
| --- | --- |
| Grades | Binned Group |
| FL ,IF | Interally Flawless |
| VVS1, VVS2 | Very Very Slightly Included |
| VS1, VS2 | Very Slightly Included |
| S1, S2 | Slightly Included |
| I1, I2, I3 | Included |
| Anything else | Other |

Image 2

**Imputed Medians**: For Table I imputed the medians for the missing values and created a missing flag for this column as well

**Imputed Missing**: For Symmetry, Cut, and Polish I imputed ‘Missing’ as a value for the missing values as there wasn’t a way I could find to impute a more intuitive value for these columns.

**Fixing Values**: For Symmetry and Shape some of the values were capitalized and spelled incorrectly so I fixed those messy values to help create more defined groups.

Short Answer Problems

Problem 1:

|  |  |  |
| --- | --- | --- |
| Vendor | Median Price | Average Price |
| 1 | $1,795 | $2,363 |
| 2 | $11,385 | $27,123 |
| 3 | $8,592 | $10,278 |
| 4 | $12,675 | $15,726 |

Image 3

As Image 3 shows there was a huge difference between Median and Average Prices for Vendor 1 vs the rest of the Vendors. After seeing the stark difference, I ran an Anova test to see if there was a difference in means between the groups. I found an extremely low p-value from the test which leads to the conclusion that there is a statistically significant difference between the groups Average Price. I also looked at other categorical variables for Vendors. I created a variable early in the analysis for Clarity which grouped the clarity grades into more general categories. The worst type of clarity is ‘Included’ and I saw that Vendor 1 sold 91/148 or 61.5% of all ‘Included’ diamonds which leads me to believe that they were trying to sell cheaper diamonds. They also sold the lowest amount of ‘Internally Flawless’ which is one of the best categories.

Another Category I compared with was Color\_Adj which was created to bin diamonds together based on how much color the diamond had. After a little bit of research, I saw that the less color means a higher price Vendor 1 also sold the most diamonds that contained color which was 351/885 or 39.7%.

Finally, I looked at the median carat value grouped by vendor and saw that Vendor 1 typically sold diamonds around .53 carats while vendors 2,3, and 4

Problem 2:

Carat and Retail Price shared a strong positive relationship I first looked at the correlation between the two variables and saw that it was 0.7413. Afterwards I looked at the median price by Cart and saw that there was a steady increase in price until about 4 carats to 5 carats and saw the Retail Price almost double. As shown in the image below the blue line is a fitted linear model that shows as the Carats increase so does the Price. And after 2.5 Carats the price starts to spread out more. This may be because the typical engagement ring is around 1-2 Carats in the United States which means the market for diamonds could shift for larger diamonds.

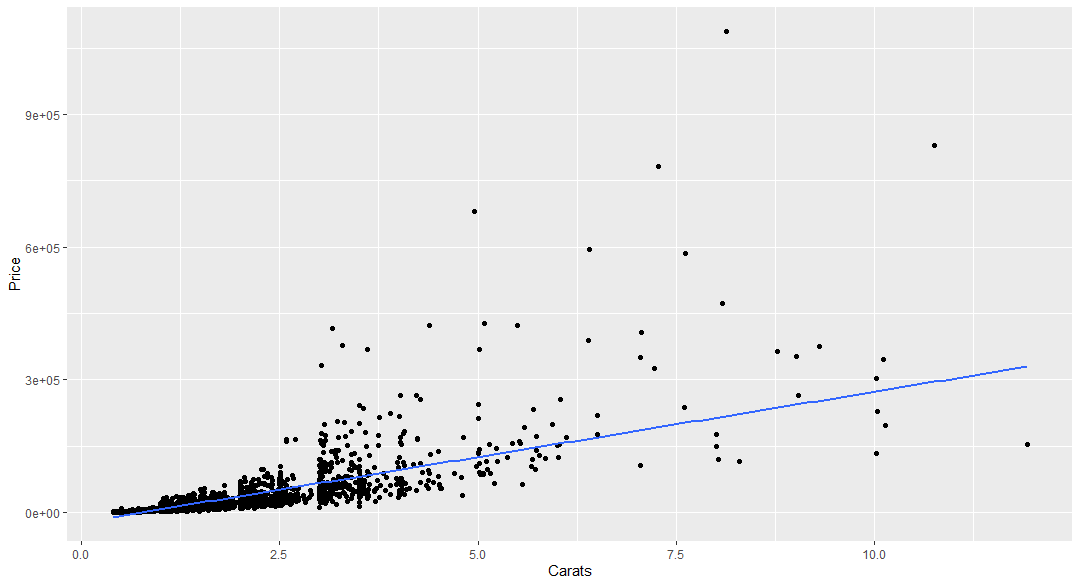


Image 4

Process Explanation:

As I started the analysis, I wanted to use more powerful models to increase predictive power and worry less about the relationship and interpretation of the variables. I settled on two models and started by predicting Price of the diamonds. The two models I used were a random forest model and an xgboost model. I split the dataframe into train and test sets to help validate my model further.

I developed the models in similar methods. I started by creating a complex model with all variables in the train data set. I then tuned the models for the best parameters with and was checking the models for Variable Importance to see which variables impacted the models the most. After the first model was created and tuned, I used a random variable from a normal distribution to help with Variable selection. By creating a random variable and using it in the model I could see if it was a more powerful predictor than other variables from the train set. If the random variable I had more predictive power and the variable had low importance I would remove it from the final model. Both models were very similar. Symmetry being the only difference between the two.

My final Random Forest Model ended up with the following predictors for Price:

Carats + Clarity\_Adj + Color\_Adj + Length + Width + Depth2 + Symmetry

My final XGBoost Model ended up with the following predictors for Price:

Carats + Clarity\_Adj + Color\_Adj + Length + Width +Depth2

After I created my final models I needed to compare the two I looked at the Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 value. I chose to look at the R2 value as the deciding factor for which model to use.

Random Forest R2: .868

XGBoost R2: .639

Random Forest blew the XGBoost away in terms of R2 so I decided to move forward with Random Forest as my final model to predict Price.

After my analysis between the two models there was more time left so I developed another Random Forest Model to predict the LogPrice. It was extremely accurate at predicting LogPrice and had an R2 value of .985, but when I exponentiated the predicted Log Price and compared it with the actual price I saw that the original random forest model was more powerful still so I decided that anymore side investigation would cost too much time so I moved forward with the original random forest model to predict Price.

After I made my predictions on the Price of the diamonds, I realized it would be extremely difficult to go one by one to determine whether I would purchase the diamonds. So, I went back to developed another Random Forest Model to predict Retail Price. I then took the difference of the two predictions and tried to see where the biggest gain could possibly exist. I was hopeful this would help the selection process I although I know that strictly relying on models can be a bad thing so I looked at the median price between different groups (Carats vs Color\_Adj) and (Carat vs Clarity\_Adj) were some of the examples. Unfortunately Known\_Diamond\_Conflict was very empty in the offers data set so when I used the model it would only predict where the column was populated. I decided to keep it in as the first model because it was an extremely important predictor variable in the model and then I would use another model without it to help the rest of the predictions.

So my final two Random Forest Model ended up with the following predictors for Retail Price were:

Carats + Clarity\_Adj + Color\_Adj + Length + Width + Depth2 + Symmetry + Known\_Conflict\_Diamond + Regions

Carats + Clarity\_Adj + Color\_Adj + Length + Width + Depth2 + Symmetry + Regions

If there was a large gain, but the Median Price for those groups was completely different from my prediction I stayed away from selecting that diamond. I put the code that I used for this at the bottom of my r-script.