Case Study Report
Wyatt Rasmussen
DSC 550

Introduction:

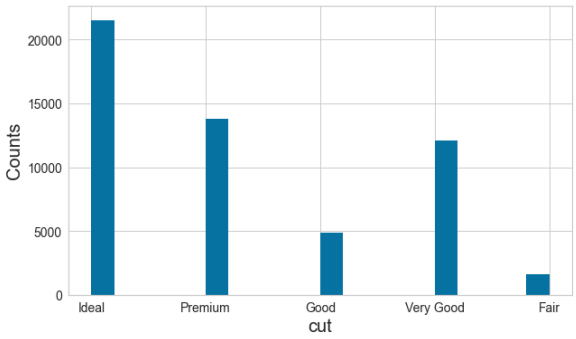
For my case study in DSC 550 I decided to research diamonds and located a data set full of information such as price, carat, cut, etc. about diamonds. I became interested in this topic due to recently getting engaged. The process of pricing diamonds has fascinated me since then. I found my data set on Kaggle.com (https://www.kaggle.com/shivam2503/diamonds). This data set was posted and free to use. The data set contains over 50,000 different diamonds with 10 variables and an index. Those variables are:

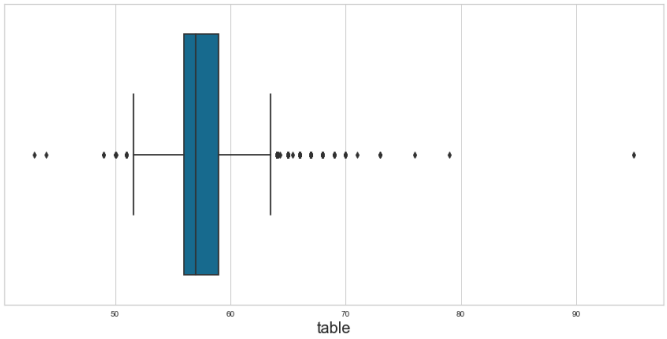
1. carat
    - unit of mass equal to 200 mg. Numerical data
2. cut
    - cut quality of the diamond. fair, good, very good, premium, ideal. Categorical data
3. color
    - quality of the cut. the worst being J to the best being D. Categorical data
4. clarity
    - how clear a diamond is. range of I1 being worst to IF being best. Categorical data
5. depth
    - depth percentage. height of diamond divided by its average girdle diameter. Numerical data
6. table
    - width of the top of the diamond relative to its widest point. Numerical data
7. price
    - how much is the diamond. Numerical data
8. length in mm
    - length of the diamond. variable 'x'. Numerical data
9. width in mm
    - width of the diamond. variable 'y'. Numerical data
10. depth in mm
    - depth of the diamond. variable 'z'. Numerical data

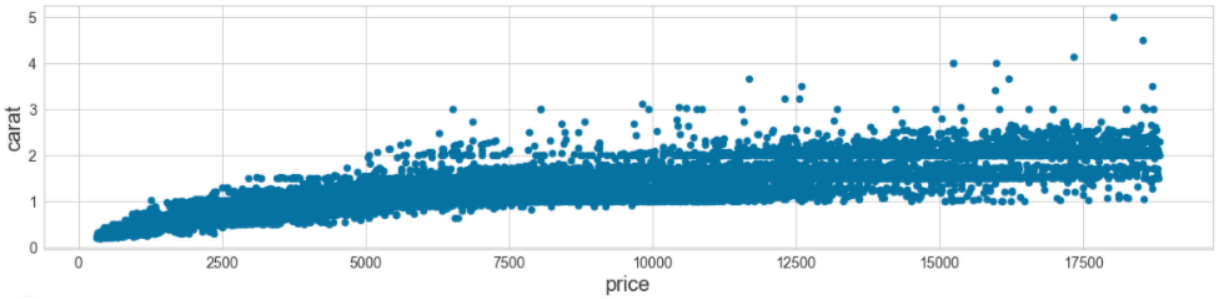
The key problem that I investigated in this project was to try to create a way to accurately price diamonds by using the features of the diamonds. Pricing diamonds is a very complex problem because each one is different and different variables affect the pricing greatly.

Summary of Findings

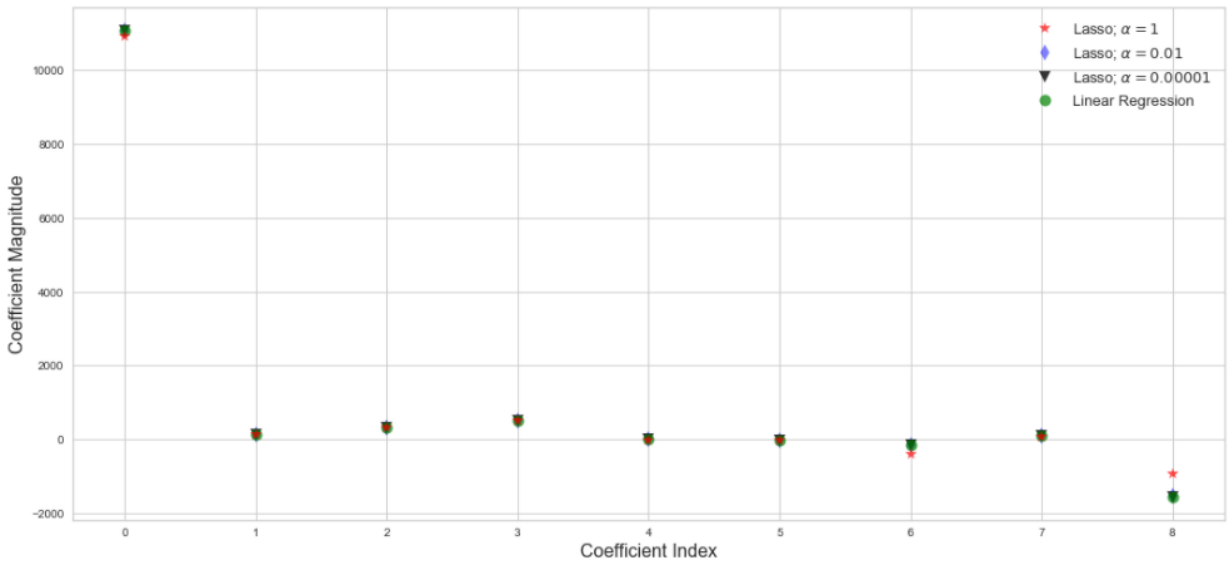
For this project I first started by creating an exploratory data analysis. I did this to examine the relationships between the data that I have. I created scatter plots, box plots, and bar graphs. This allowed me to begin to understand the data set that I was using.



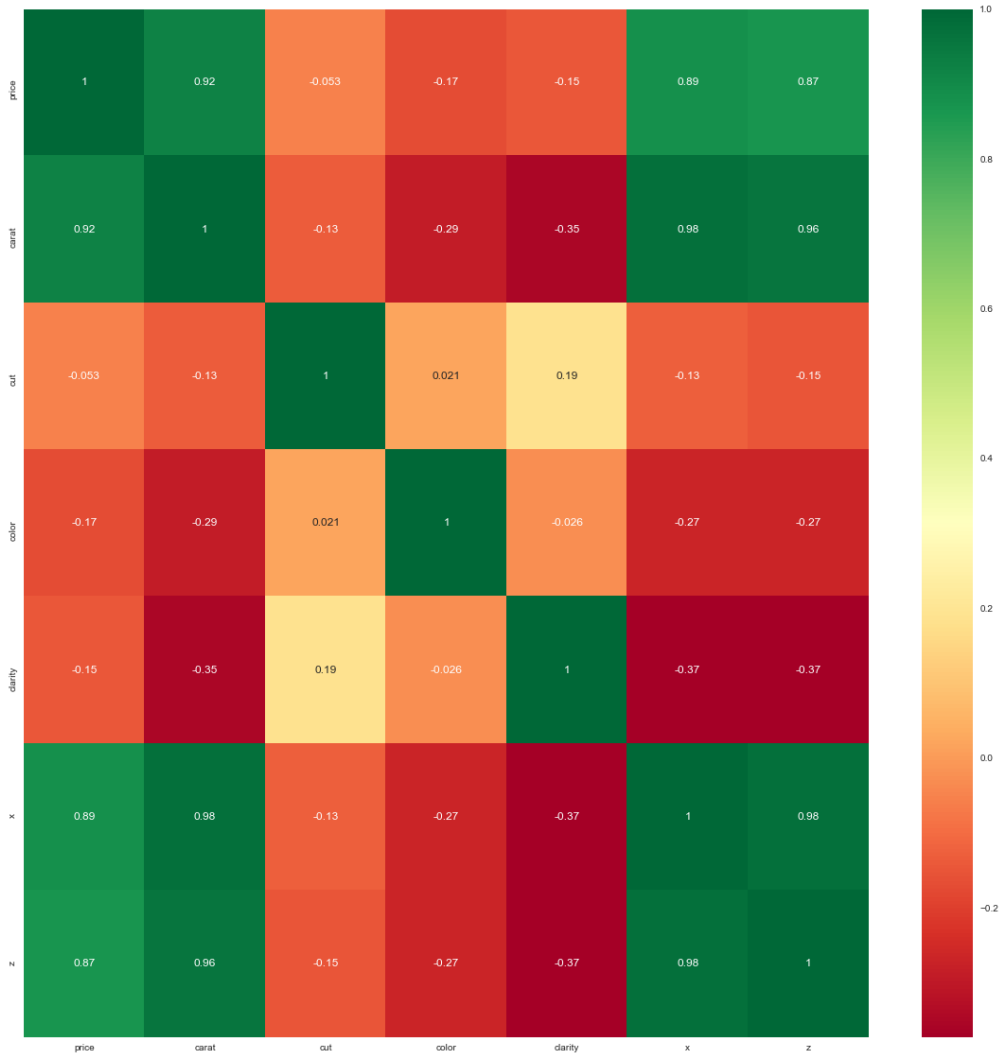




After investigating some of the relationships of the data I decided that the next logical step was to select the variables I wanted to use. Some would obviously have a little more impact on the target variable, or the price of a diamond. The technique I used to find the variables with the highest impact I created a Lasso Regression. Lasso regression is a way of regularizing the data to determine the best fit regression line. In this instance I used lasso regression to select the variables that seemed to have the highest impact on the price of a diamond. By examining the regression of the variables, I was able to eliminate the variables that were at zero and select the variables that were above zero



After using the lasso regression, I also wanted to look at the correlation between the variables I have selected and price. The best way for me to complete this process was using a correlation heat map to see which variables had correlation with each other.



After this process I was able to determine the variables that I wanted to create a model with. Those variables were **Carat**, **Cut**, **Color**, **Clarity**, **X** and **Z**.

Once the variables are selected the next step was to perform regressions with the data we are using. Regressions are typically used in models involving numbers. It fits a model to the data you have in order to predict the outcomes of future values. In this case we are using the variables we selected during feature selection to predict the price. Our models were pretty accurate at a scoring range between different regressions of 90.6% - 90.7%. This is a great score for our data.

```
 Ridge Regression Alpha = 0.01
Ridge Regresison Training Score:  0.9063765529268059
Ridge Regresison Test Score:  0.9078307387259955


 Ridge Regression Alpha = 20
Ridge Regresison Training Score:  0.9062222949090917
Ridge Regresison Test Score:  0.9075536429861389


 Linear Regression
Linear Regresison Training Score:  0.9063769814528755
Linear Regresison Test Score:  0.9078382378101414


 Lasso Regression
Lasso Regresison Training Score:  0.9063627559667576
Lasso Regresison Test Score:  0.9077798501553653


 Lasso Regression Alpha = 5
Lasso Regresison Training Score:  0.9060086563076835
Lasso Regresison Test Score:  0.9072741487986498
```

Using the data that we retrieved we can begin to understand what the key features to look for in more expensive diamonds. The feature selection gave us the features that are most important to a diamonds pricing. In order to find highly priced diamonds, someone should look to Carat, Cut, Color, Clarity, X and Z. This would enable retailers to find diamonds from retailers that are at a lower cost that what the model shows enabling them to get cheaper diamonds and sell them for a larger profit.

I would pitch this to a group of stakeholders to get buy-in that this model could help us find underpriced diamonds and to be able to create larger profit margin across the company. Since this model takes into account only the features that seem to be most important to the price of a diamond, we would be able to essentially exclude the other features that seem less important but may impact the price some. This would create a gap between our pricing and the pricing that salesmen are trying to sell a particular diamond for.

I think the biggest challenge for my model is that it only takes into account prices of the diamonds as they are sold to customers. This wouldn't include the retail markup that is almost impossible for me to calculate with the set of data I have. To accurately create a model for diamond shops I would need a dataset of non-marked up diamonds, to level out the retail markup.

Conclusion

Overall, this project introduced a lot of new themes and techniques that I didn't have very much previous experience doing. I think I created a model that accurately reflects the pricing of diamonds. Creating regressions of different types and getting accuracy scores of each type I can begin to see how the model is being fit and what could be done to improve upon an accuracy score of 90.6 – 90.7%.