Capstone Team 4

Cade culver, justin merryman, Karla Murphy, and Patricia Berry

diamond pricing

Executive Summary

Add this after the write-up has been done

Table of Contents

Diamond Introduction 3

Data Information 4

Tableau 5

Machine Learning 6

Conclusions 10

Limitations 10

Future Work Recommendations 10

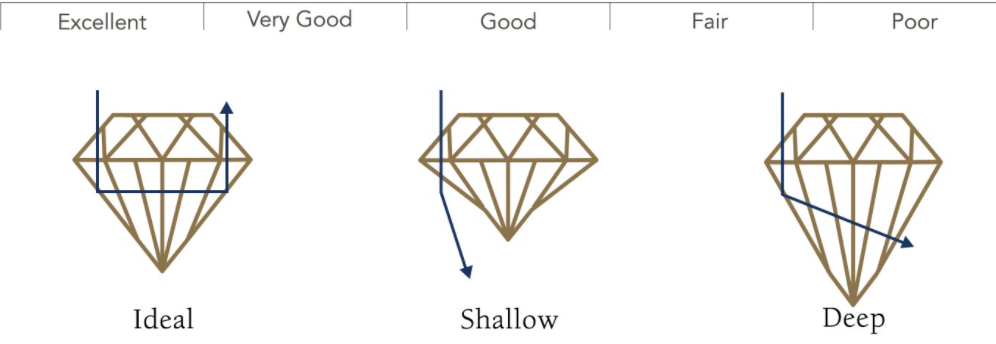
Resources 11

About Us 12

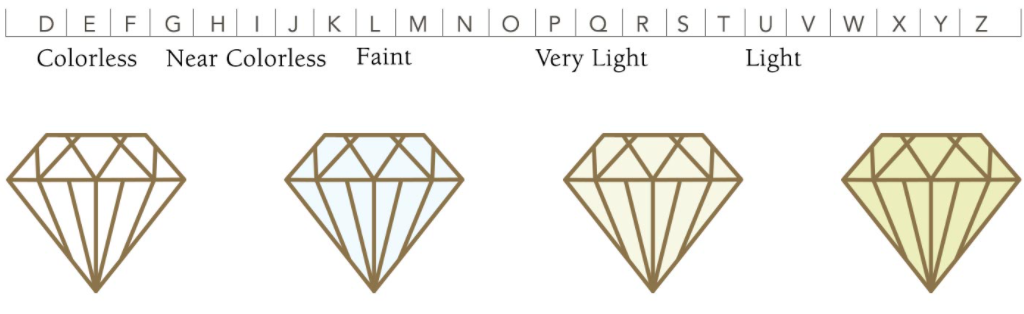
Diamond Introduction

Diamonds have been defined as a girl’s best friend, but we wanted to know more about the diamond trade. Our team invested pricing and the anatomy of a diamond to determine whether we could predict pricing and help the user to buy the best choice for their budget and preference. To choose the perfect diamond, you must follow the 4C’s that determine the price of the diamond.

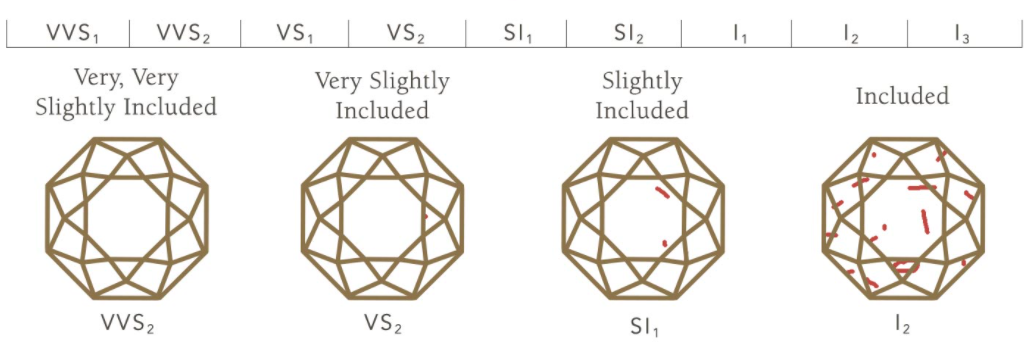
Cut – The most important of the 4Cs is cut. Cut drives how much a diamond sparkles.



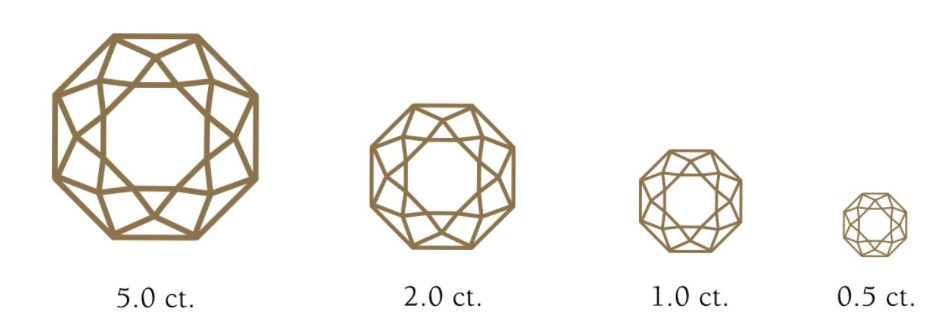
Color – Color is the second most important feature in diamond selection. The less color that a diamond has the higher the grade of the diamond.



Clarity – clarity defines the degree that a diamond is flawed. Clarity grade is based on the size, number, position of, and color of the diamond’s flaws.

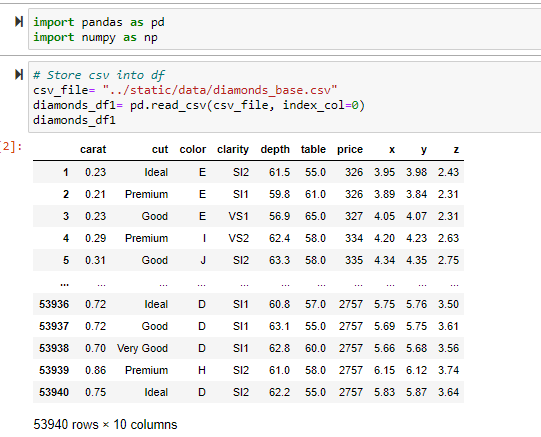


Carat Weight – carat weight refers to a diamond’s total weight not to size.

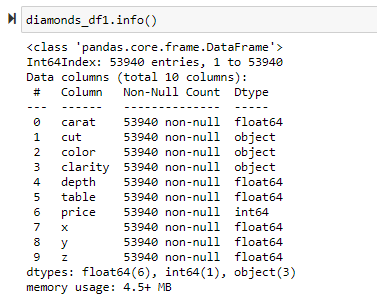


Data Information

The data was selected from Kaggle and was in a csv format. The data set was medium in size and we were able to use it as is without taking a random sample. The data selected was very clean and only one column was deleted as it mirrored the index column.



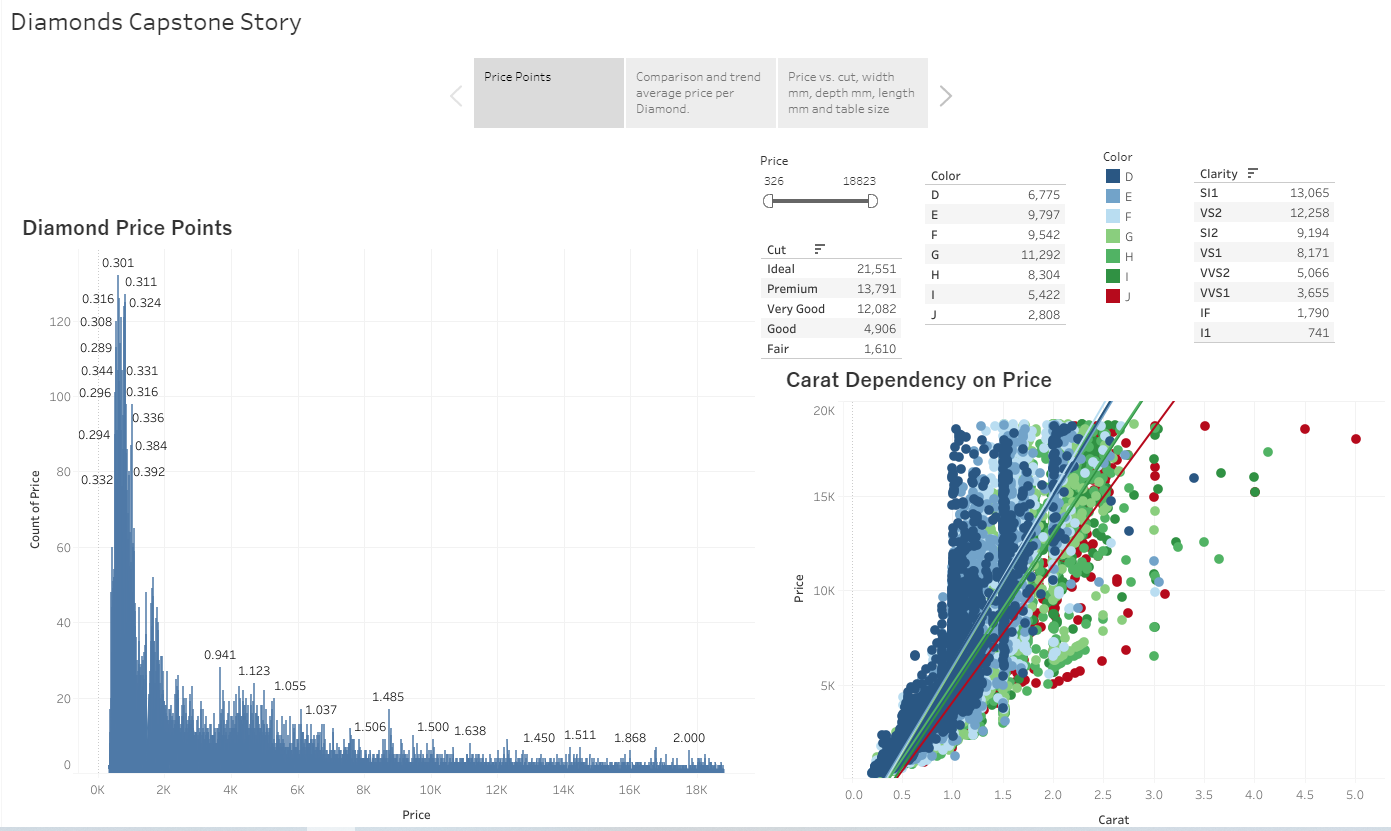
We also made sure that the data did not have any null values.



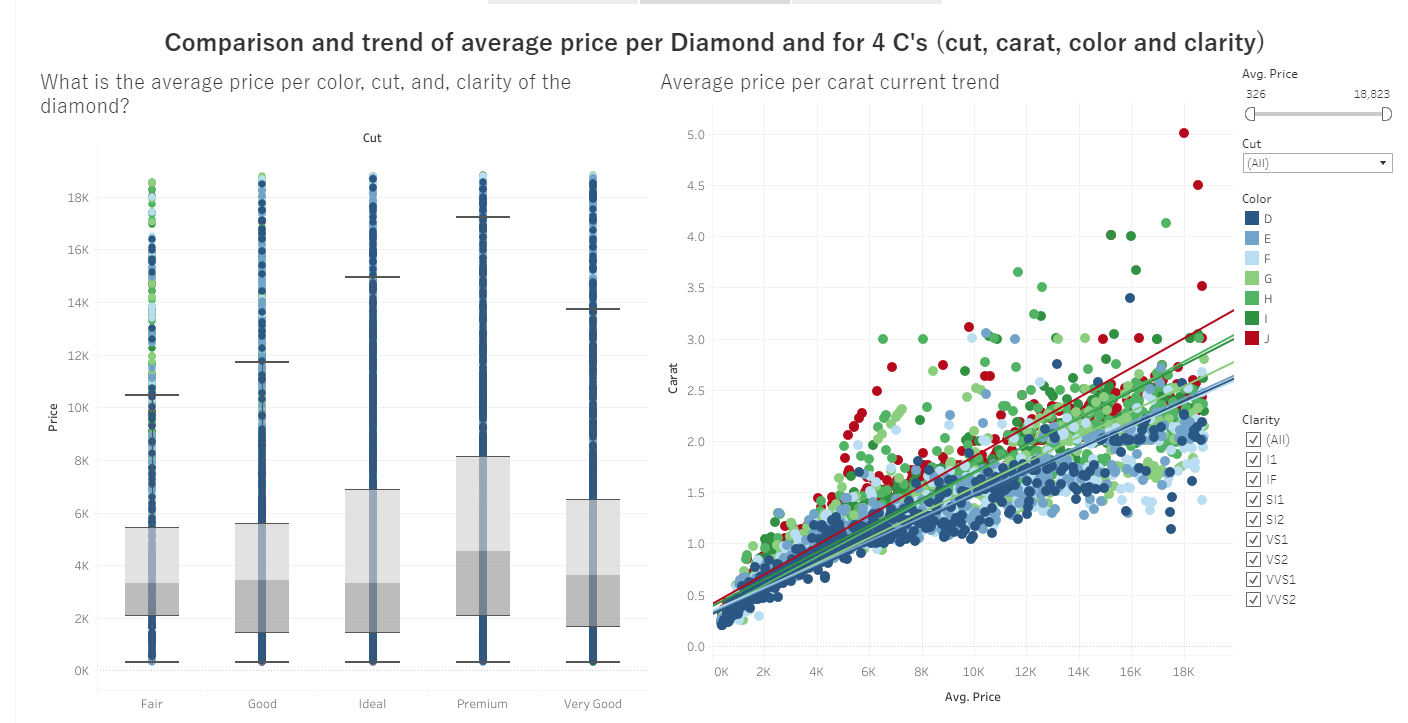
Tableau

Three dashboards were created to visualize the different features of diamonds. We chose the green-blue diverging color scheme for our charts due to the cool appearance and that the color scheme would fit with the chosen bootswatch template we chose for the html.

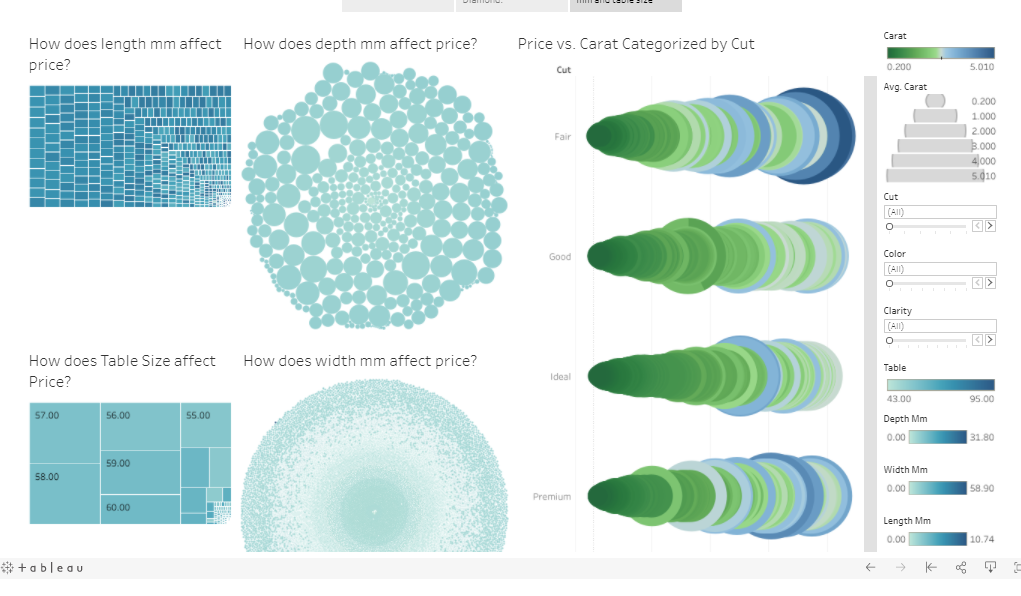
The first dashboard is broken up into two segments. The chart on the left identifies the number times a certain price point occurred. The maximum price in this case was over $18 K. Most of the diamonds sold are below $2 K. The second chart highlights the carat dependency on price. Filters at the top allow the user to choose cut, color and clarity.



The second dashboard covers the comparison and trend of average price per Diamond and the 4C’s (cut, carat, color, and clarity. There are filters to switch between the 4C’s to understand the trend between the 4 different features. The first chart looks at how cut affects price. The second chart looks at the average price per carat in a scatter chart and identifies the trends.

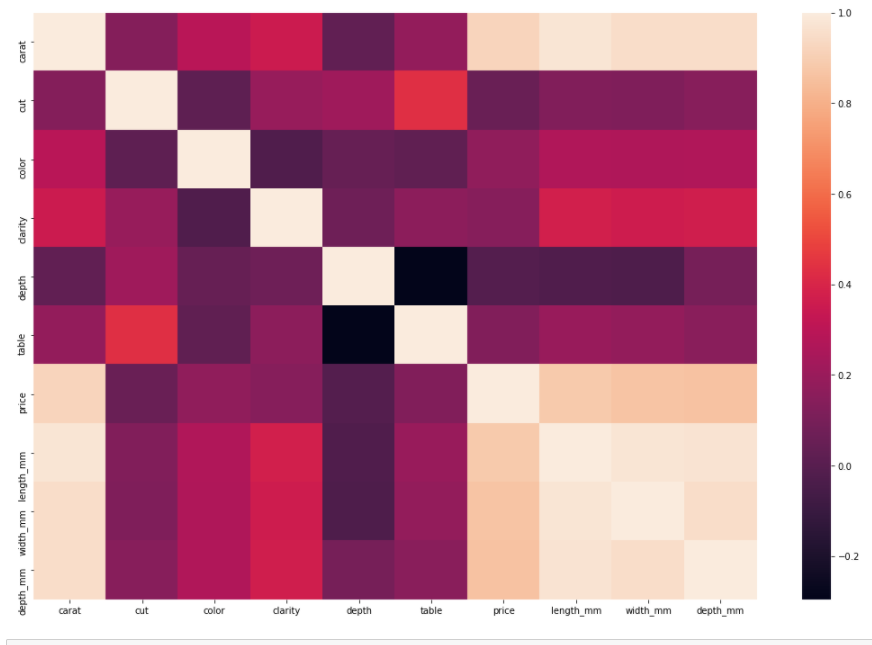


The final dashboard dives into the remaining data points contained within the data set and how price affects those measures. We look at length mm, width mm, depth mm, carat weight and table size. Filters allow the user to switch between cut, color and clarity. The length mm chart is a square chart that starts at the largest impact and dives down to the lower point. The table size chart is square chart that starts at the largest impact and dives down to the lowest point. The depth mm chart is a circle scatter chart where the price identifies the size of the bubble. The width mm chart is a circle chart that plots the layers of price driven by the width mm of the diamond. The final chart is a circle bar chart that highlights how carat weight determines the diamond price.



Machine Learning

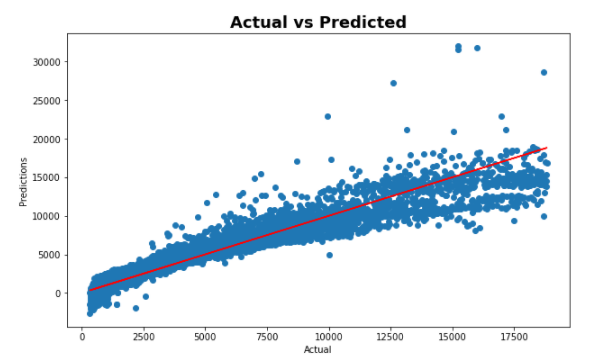
The machine learning module was created to give an end user the opportunity to identify the price of a diamond based on their choices between carat weight, color, clarity and cut. The data was put through several tests to identify the best model to use for the predictive application. The first step was to identify correlations within the data.



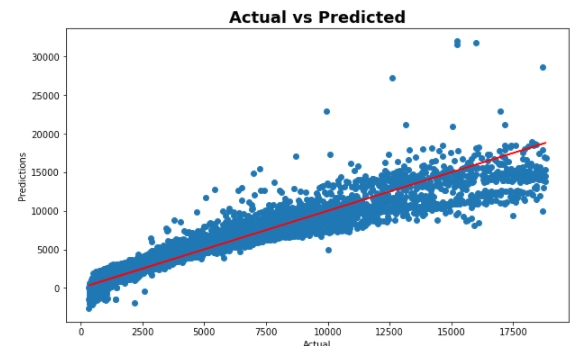
The next step was to look at the different models.

Linear Regression Models

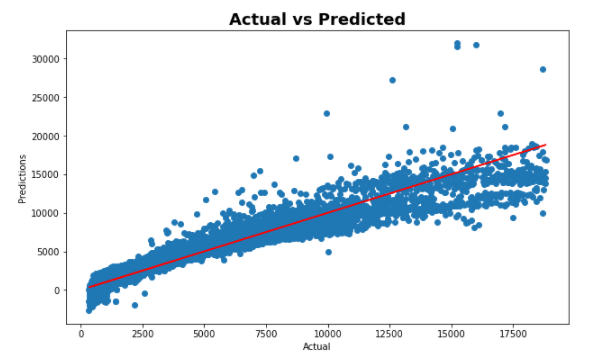
LinearRegression - the in sample R2 score was 0.9092517263784572 and the in sample RMSE was 1201.9520820408259. The out sample R2 score was 0.9076446121106826 and the out sample RMSE was 1211.675332479086.



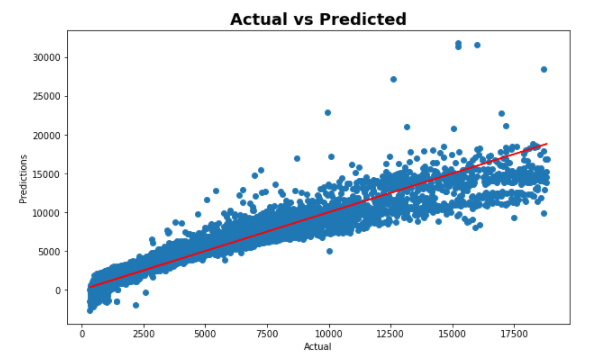
The KNeighborsRegression model - the in sample R2 score was 0.9092517263784572 and the in sample RMSE was 1201.9520820408259. The out sample R2 score was 0.9076446121106826 and the out sample RMSE was 1211.675332479086.



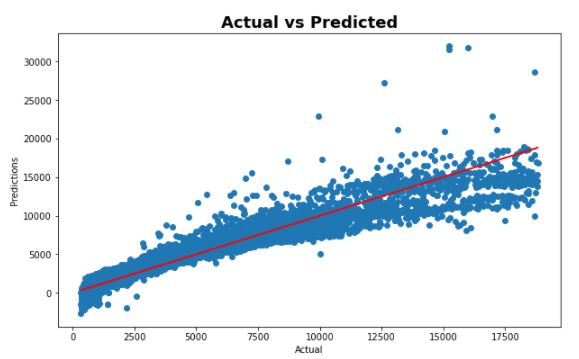
The SVR model - the in sample R2 score was 0.9092517263784572 and the in sample RMSE was 1201.9520820408259. The out sample R2 score was 0.9076446121106826 and the out sample RMSE was 1211.675332479086.



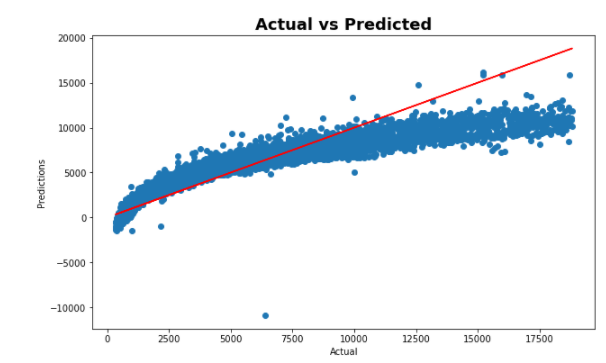
The Lasso model - the in sample R2 score was 0.9092361303192542 and the in sample RMSE was 1202.0553617492924. The out sample R2 score was 0.9076434629479834 and the out sample RMSE was 1211.6828707927464.



The Ridge model - the in sample R2 score was 0.9092513195555165 and the in sample RMSE was 1201.954776203001. The out sample R2 score was 0.9076494175687261 and the out sample RMSE was 1211.6438089759688.

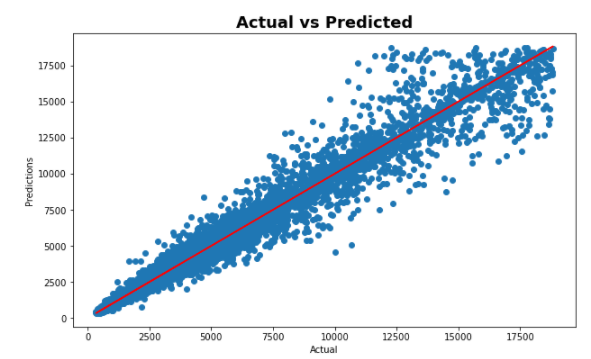


The ElasticNet model - the in sample R2 score was 0.8104628540520513 and the in sample RMSE was 1737.061267132269. The out sample R2 score was 0.8197579525330729 and the out sample RMSE was 1692.7125115007573.

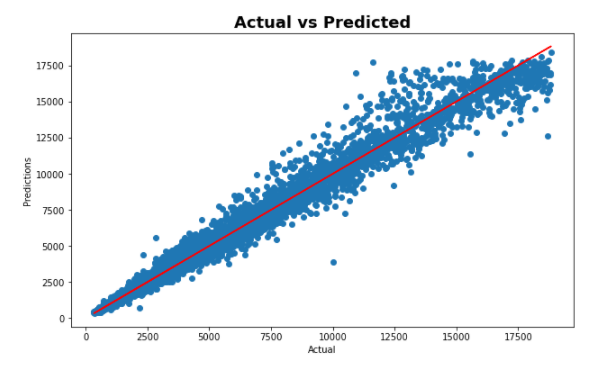


Trees Models

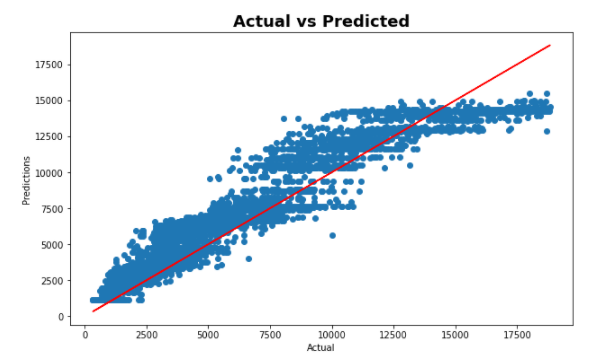
The DecisionTreeRegressor model - the in sample R2 score was 0.9999948355708338 and the in sample RMSE was 9.067323921015339. The out sample R2 score was 0.9666252386506127 and the out sample RMSE was 728.3904324945589.



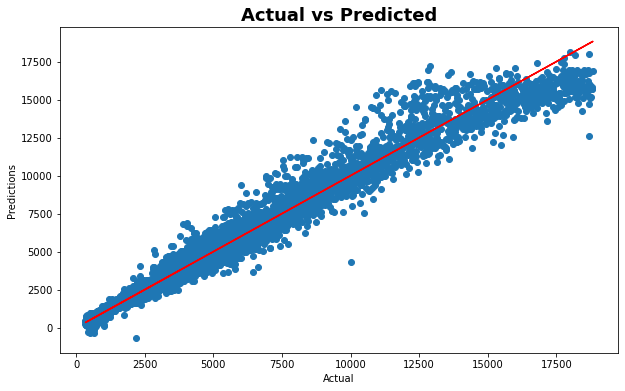
The RandomForestRegressor – the in sample R2 score was 0.997422047746516 and the in sample RMSE was 202.58412171224964. The out sample R2 score was 0.981658235117793 and the out sample RMSE was 539.9791992422503.



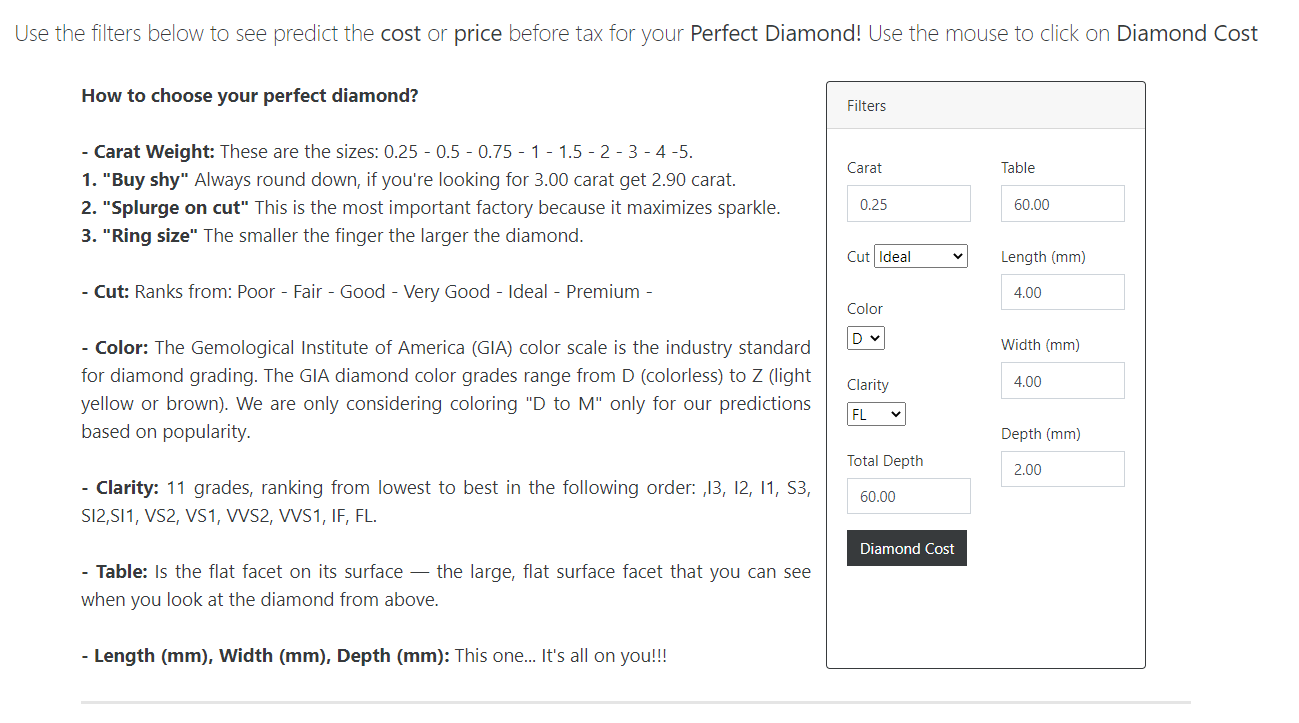
The AdaBoostRegressor – the in sample R2 score was 0.9150623620214242 and the in sample RMSE was 1162.8349006414346. The out sample R2 score was 0.9131433607546641 and the out sample RMSE was 1175.0508439806226.



The GradientBoostRegressor – the in sample R2 score was 0.9773703273701563 and the in sample RMSE was 600.2153281248416. The out sample R2 score was 0.9766551379985087 and the out sample RMSE was 609.1872479312117.



The model we chose for the project was the LinearRegression because it has the lowest MSE and highest R2 score and does not overfit as bad as the Decision Tree. The predictor:



Conclusions

Limitations/Bias

Resources

Diamonds: <https://www.kaggle.com/shivam2503/diamonds>

Diamonds In-Depth Analysis: <https://www.kaggle.com/fuzzywizard/diamonds-in-depth-analysis>

Prediction Diamond Prices: <https://www.kaggle.com/himahima/prediction-diamond-prices>

Regression graph examples: <https://datascienceplus.com/wp-content/uploads/2016/02/Screen-Shot-2016-02-15-at-6.39.53-PM.png>

Regression graph examples: <https://cme195.github.io/assets/lectures/Lecture4_Visualizing_Data_files/figure-revealjs/unnamed-chunk-41-1.png>

Tableau example: <https://public.tableau.com/profile/zitong.yang#!/vizhome/DiamondAttributes-PriceClarityandColor/Story1>

Data cleaning inspiration: <https://analyticsindiamag.com/tutorial-get-started-with-exploratory-data-analysis-and-data-preprocessing/>

EDA example: <https://www.excelr.com/exploratory-data-analysis-in-data-science>

Diamond research: <https://www.lumeradiamonds.com/diamond-education/index>

Diamond research: <https://www.bluenile.com/education/diamonds?gclid=da6dfee664eb16716dca9e624e867e33&msclkid=da6dfee664eb16716dca9e624e867e33&click_id=173535549&utm_source=bing&utm_medium=text&utm_campaign=Diamonds+%3A+Education+Prospect_Exact_Desktop&utm_content=Diamonds%3AGrading%3AChartutm_term%3Ddiamond+grading+chart>

Diamond research: <https://www.ihatestevensinger.com/engagement-rings/101-diamonds-women/?lang=en_US>

About Us



Cade Culver:



Justin Merryman:



Karla Murphy:



Patricia Berry: