**Capstone 3: Diamond Price Prediction – Project Report**

Diamonds have been seen as a store of value and a luxury status symbol for generations, dating at least back to 400 B.C. when diamonds were first referenced in a Sanskrit manuscript. From a geologic perspective, these stones are considerable older: typically older than 1 billion years to approximately 3.5 billion years ago. This age carries academic value and contributes to the scarcity of the stones, but what really controls the value of diamonds are the various characteristics of the cut and polished final stone that is sold into wholesale or retail. The key characteristics of diamond quality are often referred to as the “Four C’s”: Carat, Cut, Colour, and Clarity.

Graphical user interface

Description automatically generated

Carat refers to the weight of a diamond expressed in carats, where 1 carat equals 0.2 grams. A carat is further divided into 100 points, thus a 0.5 ct diamond would be 50 points. Size is an important factor in determining diamond value, but two diamonds of equal weight can have very different values depending on their quality. Furthermore weight only indicates size and says nothing of proportion which can make two diamonds look very different and catch light differently (table, girdle, etc.; covered later). The “cut” of a diamond refers to not only the shape and style, but the proportions, symmetry, and finish of the diamond. It is the only property which is entirely dependent on human factors. Proportions and angles influence the internal reflection and dispersion of light leaving the diamond, which determines the “brightness and sparkle”. There is a trade-off in the cutting process: to maximize weight of the diamond or maximize the brilliance and beauty, which are considerations for the retail strategy and end-user. The “colour” of a diamond is evaluated with a visual assessment by comparing it to a set of master diamonds under laboratory conditions, ranging in value from “D” colourless to “Z” light yellow for gem-quality stones. Clarity refers to the number and size of inclusions or imperfections within the diamond which may not be visible to the naked eye.

Diagram

Description automatically generated with medium confidence

In addition to the “Four C’s” of diamond quality, two other key features are the “table” and “depth”. The “table” refers to the flat surface on the top of the diamond, and the “depth” refers to the length from the table to the bottom of the diamond, which is called the “culet”.

**Data Identification and Wrangling**

The primary data for this project is a dataset containing 53,940 diamond records with 10 different features. These features include: carat, cut, colour, clarity, table, depth, x, y, z, and price. This data contains a combination of numerical and categorical features. Numerical features include: carat, table, depth, x, y, z, and price, where table and depth are percentages. The remaining features are categorical and represent the remaining 3 “C’s” of the Four C’s of diamond quality: cut, colour, and clarity.

Table

Description automatically generated

The data cleaning process identified several features with records that needed to be rectified before continuing to explore the data. The issues were restricted to the ‘x’, ‘y’, and ‘z’ dimensional data, where there were zero values for each of the dimensional measurements that needed to be removed. Diamonds that were missing ‘x’ or ‘y’ values were also missing ‘z’ values, so the decision was made to drop all ‘z’ rows with a zero value (20 rows). Additionally, there were significantly and anomalously large values for ‘y’ and ‘z’ that dramatically skewed the distributions relative to the ‘x’ values; these instances were dropped (3 rows).

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Exploratory Data Analysis**

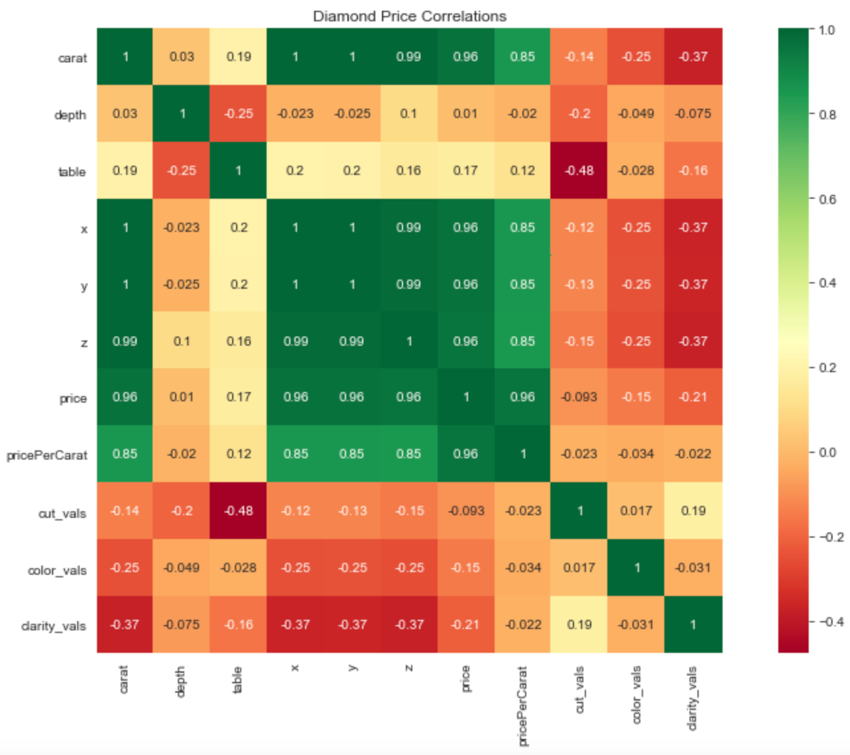
After wrangling and cleaning the data, it is imperative to understand the individual features within the datasets as well as the scale and relationship between features. During the Exploratory Data Analysis phase of the project, the features in the diamonds dataset were plotted and compared in various ways to best determine each features impact on price. The first step is to look at the commonly reported diamond characteristics.

Chart, bar chart

Description automatically generated

The diamonds in this dataset are mostly smaller than 3 carats and of higher quality cut values (‘Ideal’, ‘Premium’, or ‘Very Good’). Additionally, the clarity is skewed towards the middle of the quality distribution with most stones containing small or very small inclusions. There is a much more even distribution of color quality, apart from very few stones from the lower color quality categories.

The goal of this study is to predict diamond price based on the additional diamond features; thus, it is imperative that the correlation between features and price is well understood. Below is a correlation plot showing all the features in the dataset. From this plot, it is clear that the features with high correlation to price are predominantly describing the size of the diamond, whether it be carat weight or the dimensional data (x, y, z). An interesting observation from this correlation plot is that the remaining 3 C’s are uncorrelated or slightly negatively correlated with price.



The relationship between price and carat weight appears to be the strongest correlation and is worth exploring further. Plotted below is a chart of mean carat weight against diamond price with an error band representing the scatter of points around the mean. The red vertical dashed lines mark the 0.5 carat weight intervals. It is important to note that there is a large jump in price at these marked intervals and more price variability immediately before the marked interval. These price bands represent possible challenges in our price prediction as well as possible sources of price prediction error around these interval markers.

Chart, line chart

Description automatically generated

By grouping the diamond dataset into bins split by these carat weight markers, it is possible to see an even more important trend (image below). There is a clear trend in the skew of the price distribution above and below the 1.50-1.99 ct bin, which is rather evenly distributed around ~$10500. In diamonds smaller than 1.49 ct, the price skews lower with high-side outliers. In contrast, diamonds larger than 2.00 ct have a price skew higher with low-side outliers.

Chart, box and whisker chart

Description automatically generated Scatter chart

Description automatically generated

Given that there is a strong, positive correlation between price and carat weight, it is worth exploring the features in the dataset relative to the “Price per Carat” to explore the nature of the price outliers seen in the chart above. Shown in the plot below is the relationship between the 4C’s and the price per carat. Using the same carat bins as above, the top-left chart shows that the price per carat is generally the highest between 1.50 and 2.49 cts. The smaller stones have a much larger spread in the outliers than the larger stones, and the outliers are mostly to higher price per carat. Looking briefly at the other 3 plots, it is apparent that each of the other ‘C’ categories are characterized by outliers with higher price per carat. The smaller carat weight-higher price per carat diamonds could

Chart, box and whisker chart

Description automatically generated

Two other dimensional features in the dataset are the table and depth percentage values. The table represents the flat top of the diamond as a percentage of the x and y dimensions, and the depth represents the distance from the table to the culet (bottom point of the diamond) and is calculated by dividing the diamond’s total heigh by its total width. These features are indicative of both the size of a diamond as well as its cut quality, which together contribute to the brilliance of the diamond. In the plot below, there is a very narrow range for diamonds with a higher quality cut in both table and depth percentages, and much broader ranges for lower quality cut diamonds.

Chart, histogram

Description automatically generated

After removing the missing dimensional data in the data wrangling portion of the project, the ‘x’, ‘y’, and ‘z’ versus carat weight distributions looked quite similar. One hypothesis that can be tested from this observation is the symmetry of the diamonds in the dataset and the correlation to price or cut quality. Represented below is the symmetry value, calculated by dividing ‘x’ by ‘y’, of each diamond in the dataset versus price and colored by carat category. There are a number of conclusions that can be drawn from this plot. First, there is a tight distribution around 1.0 (+/- 0.05), with most of the outliers occurring in diamonds less than $10,000. This suggests that the majority of the diamonds in the dataset are round diamonds, and higher price diamonds have a much higher degree of symmetry. By pulling out the outliers, essentially any data points outside of the 0.9-1.1 symmetry range, there is no clear correlation between the outliers and lower cut quality due to the large number of higher quality cuts (bar chart below). This may suggest that these are a different shape of diamond and we are not specifically looking at round diamonds, but perhaps also oval or emerald shaped diamonds.

A picture containing text, sky

Description automatically generated Chart, bar chart

Description automatically generated

**Modeling: Preprocessing, Training, and Testing**

With a clear understanding of the variables and features within the dataset, the next step is to create suitable training and testing datasets from the existing dataset. The first step in preprocessing is to drop the columns created during the Exploratory Data Analysis, which were derived from the original features. Next, the categorical values were encoded as dummy columns, and the original categorical features were dropped (below). The target variable for the prediction is the ‘price’ column, so this was then separated from the main dataset, creating X and Y datasets. Finally, the dataset was split into training and test datasets using a 70/30 train/test split factor.

Text

Description automatically generated

Table

Description automatically generated

After separating the training and testing sets, the next step is determining which models to use and which model parameters to test. Since the target variable (‘price’) is numerical and due to the nature of the features in the dataset, ensemble regression models were chosen for their robust prediction accuracy. The three models used in this study are the AdaBoost Regressor (Adaptive Boosting Regressor, ABR), Random Forest Regressor (RFR), and GradientBoosting Regressor (GBR). Additionally, a second iteration of the Random Forest Regressor model was run using random search to determine the optimal range for grid search parameters. For each model, the workflow was consistent to the following: 1. Initiate a regressor, 2. Generate a dictionary of grid parameters to test, 3. Run a GridSearchCV with 5-fold cross-validation, 4. Fit the model to the X and Y training sets, 5. Run a 5-fold cross-validation using the best parameters from the grid search to determine model scores, 6. Predict Y using the X test set, 7. Generate standard error metrics for the model for the predicted values against the test set and compare.

Comparing the models used to predict the diamond price, the Gradient Boosting Regressor model performs best and the AdaBoost Regressor model performs worst, with 0.978 and 0.88 scores respectively. The mean absolute error is much lower for the GBR model, less than that of the ABR model.

Table

Description automatically generated

By plotting the predicted versus the actual values for the various models demonstrates clearly how well each model predicts the diamond price values. From the plot below, the RFR and GBR models appear similar in their distribution and differ from the more “stripped” ABR model. While these models appear different in graphical form, they all generally perform well in predicting diamond price.

Chart, scatter chart

Description automatically generated

The feature importance for the models represents how significant each feature is to the model generation. The figure below shows the feature importance for all 4 models. The most significant observation from these bar charts is that the two most important features in each model are the ‘y’ dimension and carat weight. From the exploratory data analysis phase of the project, both features had high correlation with price due to their description of diamond size, so this is neither surprising nor troubling. This will be explored futher in the scenario testing phase of the project.

Chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

**Scenario Testing and Business Impact**

To better understand the implications and business use-case of modeling the price of diamonds in the previous step, it is important to consider the scenarios in which price prediction would matter for a business, such as a diamond dealer or retailer. The profit of the business is directly tied to revenue from the value of the diamonds they sell minus costs associated from cutting/other equipment, operational expenditures, and raw stone costs. The chart below represents a hypothetical diamond retailer’s value driver tree, characterizing key value decisions and how they impact the overall business proposition. Since this dataset contains cut stones, it is worth considering scenarios in which a business might specify which diamonds they want to purchase or how they might predict the value of the diamonds they are set to receive to better understand future store revenue.

The first scenario assumes that either the granularity of the data a retail store might receive about diamonds won’t include the dimensional data or that the fidelity of that data might not be flawless due to the removal of rows earlier in the study on bad dimensional data. Thus, to test this scenario, the dimensional data, the ‘x’, ‘y’, ‘z’, and ‘symmetry\_vals’ columns, are removed from the training and test sets. Following the data removal, the GradientBoosting Regression model was run again using adjusted grid parameters to maximize the number of estimators. With these optimized parameters, the model scored quite well with an average score of 0.9807.

The second scenario assumes that the store might only have information about the “4C’s” of the diamonds they are set to receive and must determine how well they can predict the diamond values. In addition to the dimensional data dropped for the previous scenario, the ‘table’ and ‘depth’ columns were also dropped. After running the GBR model with adjusted grid parameters, the model scored only slightly lower than scenario 1 with a mean average score 0.9799.

The final scenario assumes that the store has confidence that the carat weight will suffice for all of the stones but requested additional information about the high-quality stones. In addition to the columns dropped for scenario 2, the lower quality cut clarity, and all of the color values have been dropped. Running the GBR model with the adjusted grid parameters, the model scored quite a bit lower than the other two but still quite well compared to the original AdaBoost Regression model, with an average score of 0.9066.

Table

Description automatically generated

Chart

Description automatically generated

The results of these three scenarios are compiled in the chart above and plots below for comparison. There is a degradation in model performance as additional information is pulled away from the training and test sets, but this is the cost-benefit balance a retail store must contend with. More information may allow for more predictable future revenue for a diamond retailer, but that information likely comes at a cost, and based on the model scores it may not all be necessary. It is clear from the bar charts of feature importance that with the absence of dimensional data, carat weight is by far the most important feature in each scenario model.

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

**Summary**

The data for the diamonds explored and modeled in this report represent a unique opportunity to understand primary controls on diamond pricing as well as offer strategic insight for businesses to influence operational value drivers. Through exploring the data, carat weight has a significant impact on diamond pricing. Diamond prices are more consistent within each 0.5 ct weight bin but have much higher variability near the upper limit of the 0.5 ct weight bin. Most of the diamonds have a high degree of symmetry, suggesting we are looking at mostly round diamonds. While there is no difference in the mean price per carat for the cut, clarity, and color features, the higher quality categories within those features have much higher price per carat values.

To model the price of the diamonds in the dataset, the categorical variables were encoded into dummy features which are suitable for ensemble method regression. Three separate models were chosen for preliminary modeling: AdaBoost Regression, Random Forest Regression, and GradientBoosting Regression. Each model utilized a grid of parameters to test for optimal performance and a 5-fold cross-validation prior to fitting the models. While all of the models performed quite well, the GradientBoosting Regression model performed the best with a mean score of 0.9786 and a mean absolute error of 289.16. From these models, there is a strong degree of confidence that diamond price can be predicted accurately from the dataset and features.

Several scenarios were tested as part of evaluating situational business impact for a hypothetical diamond retail store. The first test was dropping the most detailed information about the dimensions of the diamonds, followed by removing the ‘table’ and ‘depth’, and finally removing all but the most high-quality diamond categorical data but leaving the carat weight for all diamonds. All models ran with a GradientBoosting Regressor and scored quite high, suggesting that the highest value of information is the carat weight and the cost-benefit of the other information is less than optimal.