Predict Diamond Price by Its Properties

Anqi Wu, Ruiqian Tang, Runru Shen, Xinran Chen

Carnegie Mellon University

Introduction & Background

Diamonds, symbols of faithfulness and luxury, often vary significantly in price despite being similar in size. This variation is primarily attributed to the famous "Four Cs" – cut, color, clarity, and carat weight. However, other factors also play a crucial role in determining a diamond's value. In this project, we aim to closely examine the characteristics of diamonds to understand how these attributes influence their market value.

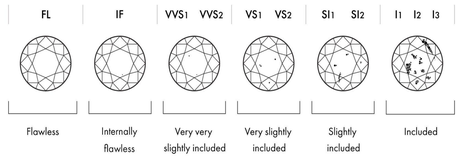


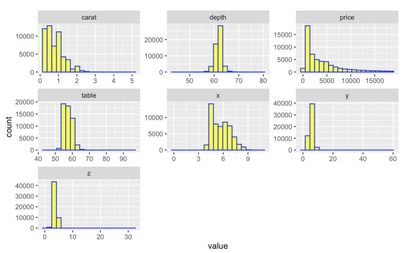
Figure 1

Our objective is to develop a model that can accurately predict the price of a diamond based on its properties.

Data Processing

The dataset includes 53,940 entries, each containing of 10 attributes: carat, cut, color, clarity, x, y, z, table, depth, along with the target variable, price. On the right side is a table detailing our variables along with their respective descriptions.

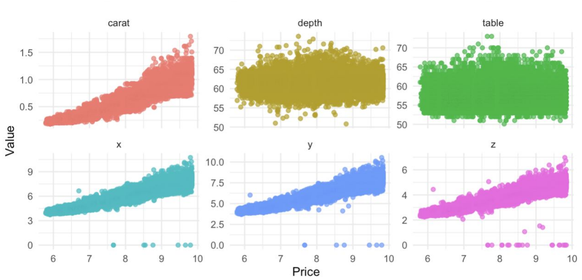
Target Variable:	Descriptions:
price	◇ The price of the diamond
Numerical Predictors:	**Descriptions:**
carat	◇ diamond weight (1 carat – 200 milligrams)
x	◇ length of diamond (millimeters)
y	◇ width of diamond (millimeters)
z	◇ depth/height of diamond (millimeters)
table	◇ width of top part of diamond relative to widest point (percentage)
depth	◇ depth of top part of diamond from the widest point, relative to total depth (percentage)
Categorical Predictors:	**Descriptions:**
cut	◇ graded quality (Fair, Good, Very Good, Premium, Ideal)
color	◇ graded color (J is worst, to D, which is best)
clarity	◇ graded measurement of clarity (I1, SI1, SI2, VS1, VS2, VVS1, VVS2, IF, in that order from worst to best)



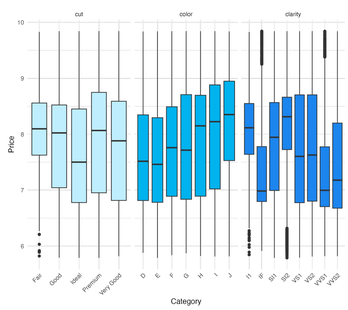
Histograms showing the distribution of each variable

By observing the histograms for each quantitative variable, we found that the distributions of carat and price were highly right-skewed, and variable y, z, depth and table have outliers. Thus, we applied a log transformation to carat and price to normalize their distributions. Then, we filtered out values in the depth and table that were either below 50 or above 75, values in y that were greater than 20, and values in z that is greater than 10. This process removed 18 data points from our dataset.

The scatter plots below demonstrate that depth and table remain relatively constant as the price increases, suggesting that these two properties do not significantly influence the price. Conversely, carat, x, y, and z exhibit a more pronounced positive correlation with price.



Faceted Scatter Plots of Price vs. Quantitative Variables



Boxplots of Price vs. Categorical Variables

It is interesting to note that diamonds categorized in the lower color grades tend to exhibit higher median prices. We later found this counterintuitive trend appears to be linked to the higher values in the y and z dimensions.

Methods

We then employed a range of regression models including a Linear Regression (L1 Regularization) as our baseline, along with Lasso Regression, Random Forest, XGBoost, and K-Nearest Neighbor. The performance of these models was evaluated based on Mean Square Error (MSE) and R-squared metrics.

Analysis and Results

Preparing Data:

Before evaluating different predictive models, we implemented a split of our dataset, allocating 70% for training and the remaining 30% for testing. For each selected model, we followed a two-step process: first, training the model using the training dataset, and then applying the model to predict outcomes based on the test set's predictive variables.

Metric Used:

We measured the model's performance by calculating the Mean Square Error (MSE) between the actual prices in the test set and the predicted prices derived from the model. Additionally, we computed the R-squared value to assess the proportion of variance in the dependent variable that is predictable from the independent variables.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Mean · Error · Squared

Figure 2

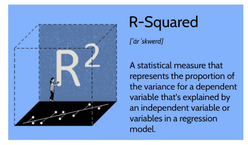
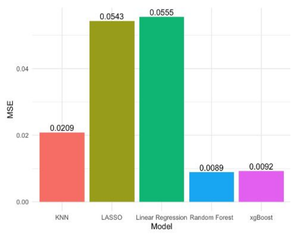
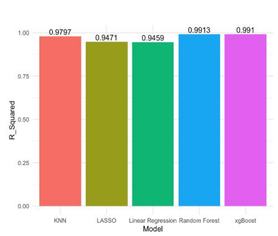


Figure 3

Model Performance Comparison:



MSE Comparison of Models



R_Squared Comparison of Models

In our analysis, both Random Forest and XGBoost models significantly outperform other models in terms of performance. Random Forest model slightly edges out with a lower Mean Square Error (MSE) and a marginally higher R-squared value than the XGBoost model. The strength of the Random Forest lies in its construction of multiple decision trees during training, ultimately outputting the mean prediction of individual trees for regression tasks, thereby offering robust and reliable predictions.

Best Model Parameter:

We attained our optimal model performance with the Random Forest algorithm by using its default parameters.

	ntree	mtry	nodesize	classwt	Samplesize
Description	Number of trees to grow	Number of variables randomly sampled as candidates at each split	Minimum size of terminal nodes	Prior probabilities of the classes.	Size of the sample to draw from the training dataset.
Value	500	3	5	NULL	37745

Conclusion

Employing Random Forest, our research has unveiled key insights into diamond price prediction. We achieved the highest R-squared score and the lowest MSE among the models considered. Notably, diamond width (y), height (z), length (x) and carat emerged as the pivotal predictors in this analysis. Our findings hold promise for future studies, which may involve the discovery of additional influential features and a comprehensive grid search for hyperparameter fine-tuning.

Reference

1. Figure 1 retrieved from an online source : https://www.ritani.com/blogs/education/f-color-diamonds-the-ultimate-buying-guide
2. Figure 2 retrieved from suboptimal: https://suboptimal.wiki/explanation/mse/
3. Figure 3 retrieved form investopedia: https://www.investopedia.com/terms/r/r-squared.asp