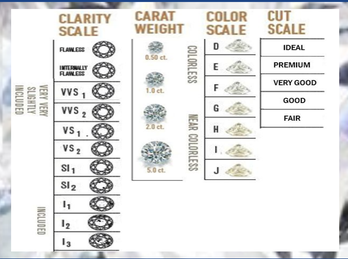
Predicting a Diamond's Price from Its Properties

By: Deanna Badger, Dom Casalnuovo, Darren Cheng, Wendy Flores-Brito, Josh Rasco
36-600: Overview of Statistical Learning and Modeling

Background & Introduction

Diamonds are the most popular gemstone in the world with a market value of $68 billion as of 2020.[1] The value of a diamond is determined by its properties, such as its cut, size, color, and clarity. **In this project, we use these properties to learn a regression model to predict the price of a diamond.**



Methods

- We took a subset of 20,000 data points from the full dataset to expedite statistical analysis.
- We split the subset of data into test and training sets with 70% used for model training and 30% used for model testing.
- The variables `price`, `carot`, `x`, `y`, and `z` exhibit multicollinearity; this does not impact predictions.
- We tested the following models using mean-squared error (MSE) to choose the best: Linear Regression, Decision Tree, Random Forest, eXtreme Gradient Boosting (XGB), and K-nearest Neighbors (KNN).
- Linear regression with variable selection was performed to determine important predictor variables.

Data Pre-Processing

We analyze data of 53,940 diamonds from Tiffany & Co.[2] The processed data contains nine predictor variables of diamond properties and the diamond `price` in U.S. dollars. The predictor variables are summarized in **Table 1**.

Predictor Variable Name	Variable Description
`carat`	Carat weight of diamond
`cut`	Ideal is best, Fair is worst
`color`	D is best, J is worst
`clarity`	FL is best, I3 is worst
`depth`	Total depth % = 2*z/(x+y)
`table`	Width of diamond top relative to widest point
`x`	Length, mm
`y`	Width, mm
`z`	Depth, mm

Table 1. Summary of predictor variables from dataset.

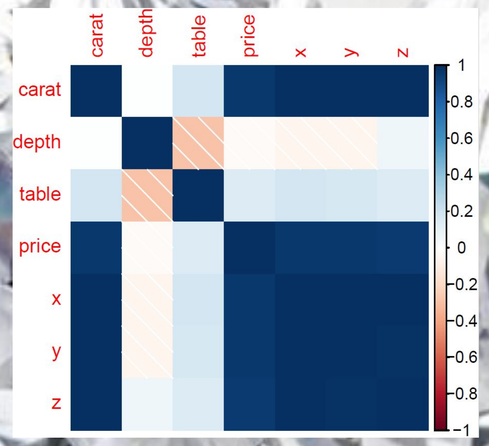


Figure 1. Correlation plot for numerical variables after transformation showing that `price`, `carot`, `x`, `y`, and `z` are highly correlated.

Exploratory Data Analysis. After removing 39 outliers from the dataset, we log-transform the predictor variable `carat` and the response variable `price` to reduce skewness. We visualize the distributions of the predictor variables that are factors (`cut`, `clarity`, and `color`) by making histograms. Based on the proportions of each secondary factor being relatively constant across each of the primary factor variables, we conclude that there is little relationship between the factor variables. We find that there is a strong linear correlation between a diamond's `price` and its dimensions (`x`, `y`, and `z`) as well as `carat` weight (**Figure 1**).

Analysis & Results

A summary of the MSEs for the different models can be found in **Table 2**. The linear regression and random forest models had the lowest MSEs. **Figure 2** displays parity plots for the random forest and linear regression models. These plots show good agreement between the observed and the model predicted prices for our test sets.

Performing variable selection on the linear regression model did not improve the MSE of the original linear regression model, but it did achieve the same MSE with less variables; `carat`, `cut`, `color`, `clarity`, `depth`, and `x` were retained, while `table`, `y`, and `z` were removed. A variable importance plot from the random forest model, shown in **Figure 3**, indicates that `clarity`, `color` and `cut` are the most important variables, respectively.

Note that KNN does not accept factor variables; therefore, `cut`, `color`, and `clarity` were not included in this model. As the random forest model shows these are the most important variables, this limitation likely limits the accuracy of the KNN model.

Model	Test Set MSE
Linear Regression	0.003
Decision Tree	0.017
Random Forest	0.002
eXtreme Gradient Boosting	0.01
K-Nearest Neighbors	0.004

Table 2. Summary of mean-squared error (MSE) results from models.

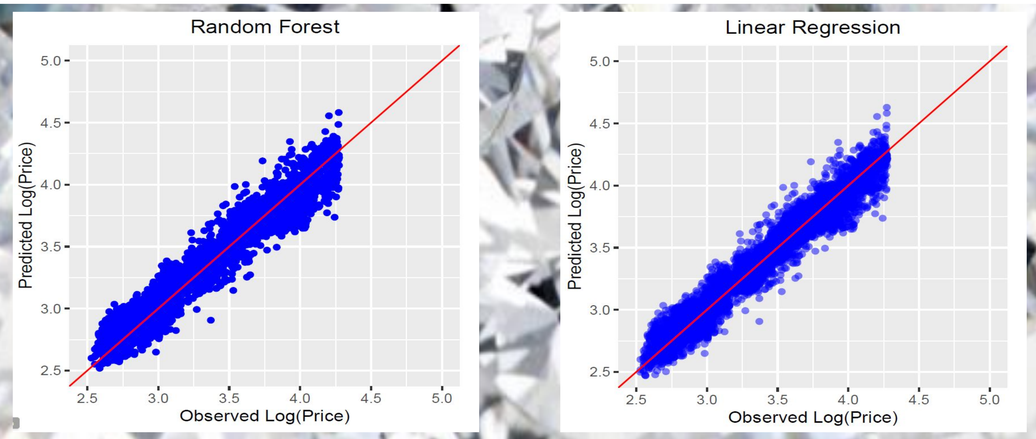


Figure 2. The predicted log(price) vs. observed log(price) for the two best models: Random Forest (left) and Linear Regression (right).

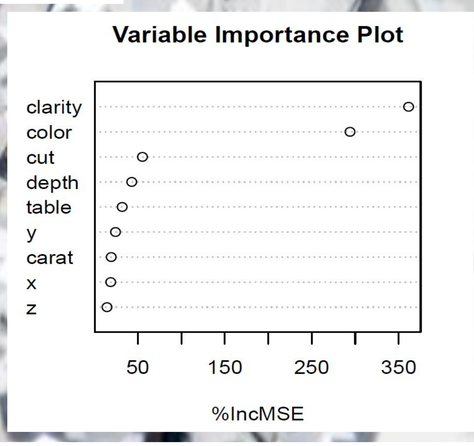


Figure 3. Variable importance plot for random forest showing that the factor predictor variables `clarity`, `color`, and `cut` are the most important.

Conclusions

Overall, the models indicate that there is a linear relationship between the predictor variables and the response variable `price`. Random forest and linear regression models were the best at predicting diamond price, with MSEs of 0.002 and 0.003, respectively. The random forest model indicates that `clarity` and `color`, followed by `cut`, are the most important variables in predicting diamond price.

References

1. M. Garland. (2021). Market value of diamond jewelry worldwide from 2010 to 2020 (in billion U.S. dollars). https://www.statista.com/statistics/585267/diamond-jewelry-market-value-worldwide/
2. S. Agrawal. (2016). Diamonds. https://www.kaggle.com/shivam2503/diamonds