Final Project

Qin Xia, Wenhan Li, Yufeng Hu

2023-04-18

Abstract

We investigate the use of the diamonds dataset in R for building regression models to predict diamond prices. The initial analysis involves a simple linear regression model, which is then improved upon and used as the basis for various machine learning models, including random forests, decision trees, and supported vector machine for price prediction, in which price is changed to a factor variable. The performance of each model was then evaluated to see its accuracy. The results demonstrate the efficacy of machine learning models in predicting diamond prices. Overall, this study highlights the potential of machine learning for improving diamond price predictions, which could have significant implications for the diamond industry.

I. Introduction

As an iconic luxury good, diamonds have always been an important product in the jewelry market. Historically, diamond prices have been stable or slightly up, mainly because supply and demand are relatively balanced. However, the COVID-19 pandemic has had a dramatic impact on the global economy, leading to a decline of around 20% in diamond sales in 2020. But because the diamond supply chain is tightly controlled, prices have not fallen sharply. With the development of technology, the quality of synthetic diamonds is improving and the cost is decreasing, which could have an impact on natural diamond prices. In an article published on January 2, Kinisky pegged the global market for cultivated diamond jewelry at about $12 billion in 2022, up 38% from 2021, although natural diamond prices declined by about 20%. In addition, cultivated diamond jewelry sales accounted for more than 10 percent of the global total for the first time.

We are motivated to identify the best prediction model for diamond prices for two reasons. On one hand, knowing the trend of diamond prices and future trends is a good way for diamond investors to assess the value and potential earnings of diamonds. Consumers can also better grasp the diamond market situation and make reasonable choices of diamond quality and price. On the other hand, our team members are approaching the age of marriage, so it is very important for us to analyze and understand the prices of diamonds before when we potentially choose rings in the future. Just a bit of personal motivation. So in this study, we analyzed recent global diamond price data from the "diamonds" dataset, which includes data from GIA, EGL and other major diamond producers, and processed the data using multiple linear regression models to identify the best model for prediction. Our goal is to build a regression model to try to figure out the factors that affect the price of diamonds and determine the best predictive model for predicting diamond prices using selected factors. The rest of the paper breaks down into three sections: we will first discuss our data and methodology, then outline the results from our regressions, and lastly extrapolate our results and highlight this paper's shortcomings. An appendix is included at the end of the report, which holds figures and tables discussed but not shown in the report. Figures and tables in the appendix are labeled as "A#".

II. Methods

We started by cleaning the data and conduct exploratory data analysis. Then we construct a model using all available explanatory variables and we went from there to improve our predictive variables selection. We further utilize the residuals of the linear fit of the whole sample to determine necessary feature engineering.

The next step is to build the best prediction models using different methods. We divided our outcome variable, namely diamond prices, into 6 categories based off the distribution of prices chose five prediction methods:

Naïve Bayes, Decision Tree, Conditional Inference Tree, Random Forest, and Support Vector Machines to yield the best prediction model. Our decision over the five methods will be based on the performance of these models with confusion matrices for accuracy comparison. Since our outcome variable is a factor variable and therefore not continuous, RMSE analysis would not be applicable.

II.i Data Description The data diamonds come with package ggplot2 in R. The dataset contains information on the price and quality of approximately 54,000 diamonds. Each observation is made up of ten variables as shown in Table A1. We set the variable price as the outcome variable and the other nine as independent variables.

II.ii Data Cleaning Since the diamonds dataset is a built-in dataset in ggplot2, we directly load the diamond dataset using the data() function. In order to make the data clean and ready for analysis, we first identify if the dataset has any missing values. We do so by performing missing value processing using the mice package and see that there is no missing value in this dataset. We then check for repeated values in order to establish whether to duplicate the index and remove any duplicate rows. Lastly, we identify and remove initial outliers from the data.

II.iii Explorative Data Analysis We begin by examining the distribution of both diamond price and diamond weight. As we see from Table A2, the lowest diamond price is \$326, the highest is \$11,886, and the mean is \$3,160. In Figure A1, we can see from the histogram of diamond prices that the largest number of diamond prices are between \$0 and \$2000. Since the mean is greater than the median and the graph is positively skewed, there is potential positive outliers in the data. We can also see from the density curve (red) that number of diamond prices peaks at about \$1000, and faults occur at about \$2000, \$4000 and \$8000. By comparing the normal distribution curve (blue) with the density curve, we can see that when the price of diamond is less than \$8000, the two curves are very different; when the price of diamond is greater than \$8000, that difference appears smaller.

As we see from Table A3, the lowest diamond weight is 0.2 carats, the highest is 3.65 carats, and the mean is 0.7237 carats. In Figure A2, we can see from the histogram of diamond prices that the number of diamonds by weight is heavily concentrated on the lower end. Since the mean is greater than the median and the graph is positively skewed, there is potnetial positive outliers in the data. In addition, we notice that while the normal distribution curve only has one peak, the density curve has two, one at the lowest weight bracket and one just above one carat. Without this second peak, the distribution of weights appears to follow an exponential decay model. We believe that further analysis can be conducted here and elaborates on this in the Conclusion.

One interesting thing we noticed about the data was that when we analyzed the distribution of the cut variable, we saw that the price distribution when cut = Very Good is basically the same as that when cut = Premium (Figure A3). We used a series of tests to verify whether this variable has a significant impact on the price. The sample data is first divided into two groups: cut = Very Good for one group and cut = Premium for the other.

We first ran an Anderson-Darling test to see whether the prices of diamonds follow the normal distribution. Suppose:

H0: The price of diamond is subject to normal distribution.

H1: Diamond prices do not obey normal distribution.

As shown in Table A4, we find that $P < 0.001$, so we reject the null hypothesis and say that the distribution of diamond prices do not obey the normal distribution.

Since diamond prices do not obey normal distribution, we used the Wilcoxon rank sum test to see whether there is a significant difference between cut = Very Good price and cut = Premium price. Suppose:

H0: Very Good at the same price as Premium.

H1: Very Good is not the same price as Premium.

As shown in Table A5, we find that $P < 0.001$, so we reject the null hypothesis and say that the price of Very Good is significantly different from that of Premium. As such, we consider the cut variable when building the regression model.

III. Results

We begin by building a linear regression model using all nine explanatory variables. As shown by Table A6, our initial linear model showed an adjusted R-squared of 0.9179. We use this model as a basis and modify it to see if we can improve the fit.

As such, we proceed to create an influence plot and find that there are outliers in the data. We then proceed to identify the specific outliers and remove them from our data. After which we proceed to identify the variance inflation factor for each dependent variable in order to check for multicollinearity. As shown in Table A7, the variables x, y, and z have a statistic significantly over five, meaning that they potentially exhibit multicollinearity. We keep the carat variable because we believe that it is significant to the model since it has the highest correlation with the price variable (Table A8). Since multicollinearity exists in the model, variable screening is required. We remove the variables x, y, and z from the model and checks the variance inflation factor once more. The results in Figure A5 show that R squared can reach 92% after x,y, and z are deleted.

After the deletion of variables, the model was tested, and the results showed that the remaining variables were significant, and the adjusted R square rose to 0.9188 (Table A9). In addition, we see that multicollinearity no longer exists in the data (Table A10). As such, we proceed to analyze the residual of the updated model.

Figure 1. Residual plots of linear regression without x, y, and z.

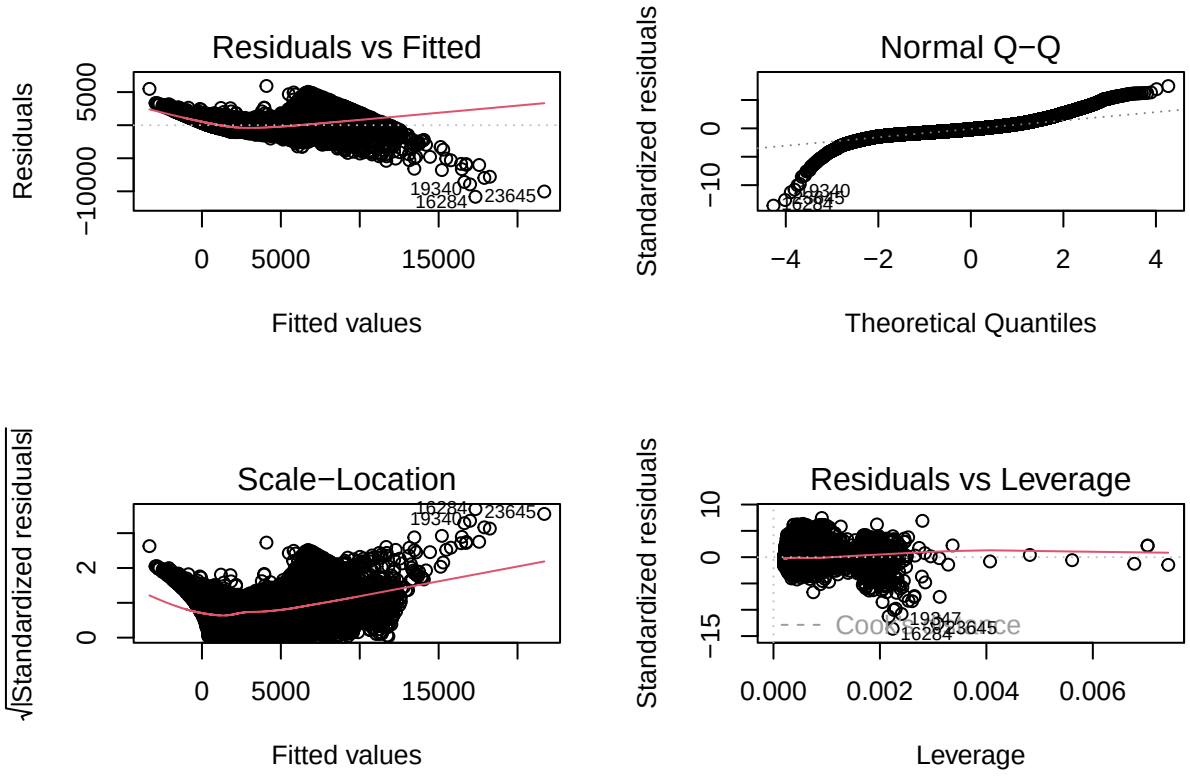
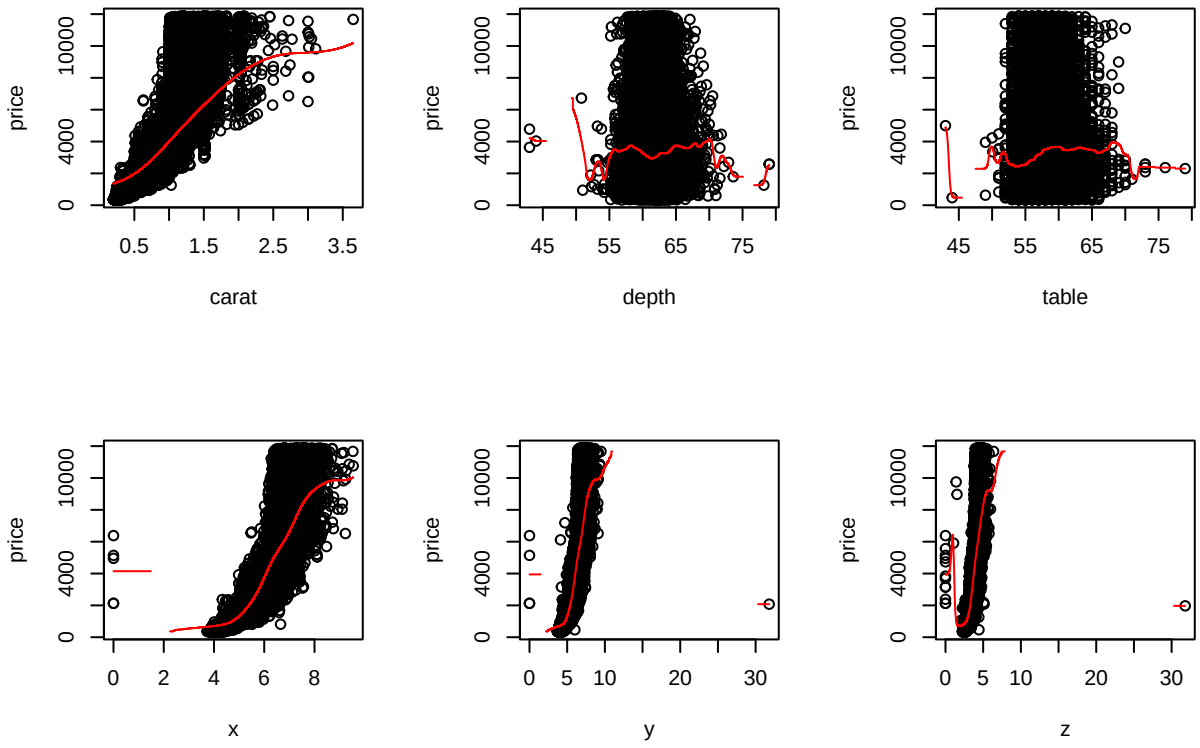


Figure 1 above shows various residual analysis plots of the model. Through the "residual vs fitted" graph, it is found that there is a curved relationship between the residual value and the fitting value, so the regression model does not meet the linear assumption. Through the "Normal Q-Q graph", we see that the points

on the graph arguably fall on the identified line, so the regression model satisfies normality. Through the "Scale-Location" graph, wee see that the graph shows non-horizontal trend, so the regression model does not satisfy homoscedasticity. Then, the correlation between variables is judged by the scatter plot and kernel density estimation curve of the variables and the price of diamonds. We then conduct variable transformation to improve the model effect.

Figure 2. Kernel smoothing graphs of independent variables against diamond price.




```
## null device
##           1
```

As shown in Figure 2 above, the explanatory variables depth and table may be related to a quadratic term. The regression model is then rebuilt and the terms depth^2 and table^2 are added onto the basis of the original model.

As we see from the results in Table A11, the adjusted R square of the model increases to 0.9202, which significantly improves the goodness of fit. Combining the regression coefficient of the model with the scatter plot, we find that the main factors affecting the price are diamond weight, cut quality, color, clarity and percentage of total depth. The more weight, the higher the price, which increases exponentially. The better the quality of the cut, the higher the price. The higher the color grade, the higher the price. The higher the clarity, the higher the price. The higher the percentage of total depth, the higher the price. The table variable, which is the width of the top of the diamond relative to the widest point, does not have a significant effect on the price. We select this regression as the one we will use for our predictive models.

We then identified the different price bracket factors that we will use for prediction. We use price bracket factors instead of nominal prices because it will give us a more practical prediction result. Nominal price predictions cannot practically identify prediction accuracy since and deviation from the original value is considered a false prediction. However, if we say that the prediction falls within the same price bracket as the

true value, then we can practically identify the prediction accuracy. Figure A6 shows that the price is divided into 6 grades: A(0-2000), B(2000-4000), C(4000-6000), D(6000-8000) , E(8000-10000), F(10000-12000).

We now fit the five previous mentioned prediction models. The results are shown below. All models use the same seed so the performance should be repeatable. In each of the confusion matrix results shown in below, "Actual" is the real value, "Predicted" is the predicted value, and the samples on the diagonal are those with accurate prediction.

Table 1. Naive Bayes prediction results.

```
##       Predicted
## Actual    A    B    C    D    E    F
##      A 4523  131    2    1    0    0
##      B  295 1639  102    1    0    0
##      C    0  266 1267  351   67   21
##      D    0   10  171  287  163   55
##      E    0    1   15  107  113   73
##      F    0    0    5   52  145  218
```

```
## [1] 0.7982343
```

The Naive Bayes prediction model shows an accuracy of 0.7977, which is the lowest in the models we selected.

Table 2. Decision tree prediction results.

```
##       Predicted
## Actual    A    B    C    D    E    F
##      A 4571   28    1    0    0    0
##      B  247 1851  119    0    0    0
##      C    0  165 1262  108    0    1
##      D    0    3  165  542  153   50
##      E    0    0   14  130  262  107
##      F    0    0    1   19   73  209
```

```
## [1] 0.862712
```

The Decision Tree prediction model shows an accuracy of 0.8583, which is the second lowest in the models we selected.

Table 3. Conditional inference tree prediction results.

```
##       Predicted
## Actual    A    B    C    D    E    F
##      A 4728   70    1    0    0    0
##      B   90 1808  105    0    0    0
##      C    0  169 1374  144    0    0
##      D    0    0   81  564   71    1
##      E    0    0    0   80  340   54
##      F    0    0    1   11   77  312
```

```
## [1] 0.9052673
```

The Conditional Inference Tree prediction model shows an accuracy of 0.8997, which is the third lowest in the models we selected.

Table 4. Random forest prediction results.

```
##       Predicted
## Actual    A    B    C    D    E    F
##      A 4741   78    1    0    0    0
##      B   77 1826   92    1    0    0
```

```
##      C    0  142 1377   96    0    0
##      D    0    1   90  629   72    3
##      E    0    0    2   64  353   54
##      F    0    0    0    9   63  310
```

```
## [1] 0.916179
```

The Random Forest prediction model shows an accuracy of 0.9139, which is the highest in the models we selected.

Table 5. Supported vector machine prediction results.

```
##       Predicted
## Actual    A    B    C    D    E    F
##      A 4725   77    1    0    0    0
##      B   93 1812  103    1    0    0
##      C    0  157 1343  104    1    1
##      D    0    1  115  619   84    4
##      E    0    0    0   69  345   87
##      F    0    0    0    6   58  275
```

```
## [1] 0.904573
```

The Supported Vector Machine prediction model shows an accuracy of 0.9010, which is the second highest in the models we selected.

IV. Conclusion

According to correlation analysis, carat has the largest correlation coefficient with the predictive variable price with a coefficient of 0.92. The main factors that affect price are diamond weight (carat), cut, quality, color, clarity, and percentage of total depth. We conducted model selection by starting with a linear model of all variables and altered the model to satisfy OLS assumptions, such as no multicollinearity and second order effects.

After we determined the regression, we used Naive Bayes, Decision tree, Conditional Inference Tree, Random Forest, and Support Vector Machines models to select the best prediction model. The data was divided between a training set and a test set to validate these machine learning models and determine which method gets the highest accuracy. The results show that the Random Forest model can obtain the highest accurate value of 91.39%.

As can be seen from the results, the test results that categorize variables, remove unimportant variables, etc., are better than the test results that are not removed. We can use this model to judge the price and trend of the current diamond market more accurately, so as to better choose suitable diamond products and price range and avoid being misled or cheated by price fluctuations. Investors in the diamond industry can also use this model to better understand the supply and demand situation of the diamond market, the causes and trends of price fluctuations, as well as the competitive landscape of the market. So as to better develop marketing strategy.

But there are some drawbacks to our study. For example, there are errors in the prediction results, which may be due to the insufficient correlation between independent variables and dependent variables, the small number of independent variables considered and the small sample size of data. We can add independent variables with higher correlation with dependent variables to improve the accuracy and prediction ability of the model. In the meantime, we can get more diamond data from reliable websites to solve this problem. Examples include Idex-Idex and GemKonnect. In addition, we notice in Figure A2 a spike in the number of diamond weights just above one carat, which can be explained by consumer preference against diamonds just below the one carat threshold. This can be potentially explored using a regression discontinuity model to identify consumer preference and producer preference for diamonds at this one carat threshold.

Appendix

Table A1 Variable description.

```
## # A tibble: 10 x 3
##    Variable Description                                        Values
##    <chr>    <chr>                                              <chr>
##  1 carat    Weight of the diamond in carats.                   0.2-5.01
##  2 cut      Cut quality rating.                                Fair, Good, Very ~
##  3 color    Diamond color rating.                              J (worst) to D (b~
##  4 clarity  Measurement of how clear the diamond is.           I1 (worst), SI2, ~
##  5 depth    Total depth percentage.                            43-79
##  6 table    Width of top of diamond relative to widest point. 43-95
##  7 price    Price in US dollars.                               $326-$18,823
##  8 x        Length in mm.                                      0-10.74
##  9 y        Width in mm.                                       0-58.9
## 10 z        Depth in mm.                                       0-31.8
```

Table A2. Summary of diamond prices.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     911    2155    3160    4670   11886
```

Table A3. Summary of diamond weights.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2000  0.3800  0.7000  0.7237  1.0100  3.6500
```

Table A4. Anderson-Darling test.

```
##
##  Anderson-Darling normality test
##
## data:  scale(diamonds1$price)
## A = 2441.2, p-value < 2.2e-16
```

Table A5. Wilcoxon rank sum and signed rank test.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  very_good$price and premium$price
## W = 65929791, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

Table A6. Linear regression with no second order variables.

```
##
## Call:
## lm(formula = price ~ ., data = diamonds1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11638.0   -450.8   -127.6    294.9   5928.4
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1412.562    300.572    4.700 2.61e-06 ***
## carat        8788.707     51.305  171.303  < 2e-16 ***
## cut.L         486.341     16.305   29.827  < 2e-16 ***
```

```
## cut.Q        -254.083     13.049  -19.472  < 2e-16 ***
## cut.C         120.726     11.272   10.710  < 2e-16 ***
## cut^4         -15.169      8.981   -1.689 0.091247 .
## color.L     -1432.238     12.768 -112.173  < 2e-16 ***
## color.Q      -497.342     11.627  -42.774  < 2e-16 ***
## color.C      -126.363     10.827  -11.672  < 2e-16 ***
## color^4        53.839      9.930    5.422 5.92e-08 ***
## color^5       -44.817      9.334   -4.801 1.58e-06 ***
## color^6       -28.250      8.441   -3.347 0.000818 ***
## clarity.L    3197.196     21.849  146.335  < 2e-16 ***
## clarity.Q   -1601.228     20.283  -78.943  < 2e-16 ***
## clarity.C     695.930     17.327   40.166  < 2e-16 ***
## clarity^4    -333.239     13.846  -24.067  < 2e-16 ***
## clarity^5     199.670     11.326   17.629  < 2e-16 ***
## clarity^6      26.923      9.865    2.729 0.006354 **
## clarity^7      80.117      8.740    9.167  < 2e-16 ***
## depth         -28.330      3.348   -8.461  < 2e-16 ***
## table         -16.467      2.120   -7.768 8.13e-15 ***
## x            -487.277     34.416  -14.159  < 2e-16 ***
## y              79.185     26.837    2.951 0.003172 **
## z             -30.861     25.135   -1.228 0.219537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 792.2 on 50378 degrees of freedom
## Multiple R-squared:  0.918,  Adjusted R-squared:  0.9179
## F-statistic: 2.451e+04 on 23 and 50378 DF,  p-value: < 2.2e-16
```

Table A7. Variance inflation factor for multicolinearity.

```
##         rstudent unadjusted p-value Bonferroni p
## 16284 -14.755321         3.6027e-49   1.8158e-44
## 23645 -14.449367         3.1456e-47   1.5855e-42
## 19340 -12.468609         1.2502e-35   6.3010e-31
## 19347 -12.002080         3.8453e-33   1.9381e-28
## 21863 -11.193124         4.7686e-29   2.4035e-24
## 22429 -11.039992         2.6409e-28   1.3311e-23
## 21759  -9.579899         1.0137e-21   5.1094e-17
## 19867  -9.381043         6.7939e-21   3.4243e-16
## 17197  -9.286977         1.6484e-20   8.3080e-16
## 23540  -8.708404         3.1733e-18   1.5994e-13
```

```
##              GVIF Df GVIF^(1/(2*Df))
## carat   31.135719  1        5.579939
## cut      1.997355  4        1.090327
## color    1.173062  6        1.013390
## clarity  1.393141  7        1.023966
## depth    1.842730  1        1.357472
## table    1.793577  1        1.339245
## x       94.150824  1        9.703135
## y       57.140043  1        7.559103
## z       20.165463  1        4.490597
```

Table A8. Correlation between numeric variables.

```
##            carat        depth      table       price           x           y
```

```
## carat 1.00000000  0.044617330  0.1843519 0.916963059  0.98129350  0.97342472
## depth 0.04461733  1.000000000 -0.2919922 0.003437801 -0.01720406 -0.02052194
## table 0.18435193 -0.291992154  1.0000000 0.128070684  0.19351455  0.18468097
## price 0.91696306  0.003437801  0.1280707 1.000000000  0.89583273  0.89147517
## x     0.98129350 -0.017204059  0.1935145 0.895832735  1.00000000  0.99074767
## y     0.97342472 -0.020521942  0.1846810 0.891475174  0.99074767  1.00000000
## z     0.95644038  0.114061269  0.1447365 0.869665712  0.96564848  0.95959697
##               z
## carat 0.9564404
## depth 0.1140613
## table 0.1447365
## price 0.8696657
## x     0.9656485
## y     0.9595970
## z     1.0000000
```

Table A9. Linear regression with no second order variables and no x, y, z.

```
##
## Call:
## lm(formula = price ~ . - x - y - z, data = diamonds1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10823.0   -476.7   -126.5    318.2   5923.0
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -1312.312    257.255   -5.101 3.39e-07 ***
## carat        7699.682     10.696  719.834  < 2e-16 ***
## cut.L         494.135     16.373   30.179  < 2e-16 ***
## cut.Q        -265.623     13.077  -20.312  < 2e-16 ***
## cut.C         127.348     11.255   11.315  < 2e-16 ***
## cut^4         -11.109      9.001   -1.234 0.217147
## color.L     -1408.806     12.783 -110.206  < 2e-16 ***
## color.Q      -481.246     11.659  -41.275  < 2e-16 ***
## color.C      -127.788     10.878  -11.748  < 2e-16 ***
## color^4        50.253      9.976    5.037 4.74e-07 ***
## color^5       -40.314      9.376   -4.300 1.71e-05 ***
## color^6       -27.442      8.481   -3.236 0.001214 **
## clarity.L    3204.771     21.939  146.076  < 2e-16 ***
## clarity.Q   -1548.145     20.222  -76.559  < 2e-16 ***
## clarity.C     657.669     17.312   37.990  < 2e-16 ***
## clarity^4    -318.391     13.895  -22.914  < 2e-16 ***
## clarity^5     188.729     11.367   16.603  < 2e-16 ***
## clarity^6      29.978      9.911    3.025 0.002491 **
## clarity^7      85.090      8.778    9.693  < 2e-16 ***
## depth         -11.189      2.911   -3.844 0.000121 ***
## table         -14.978      2.128   -7.038 1.97e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 796.1 on 50381 degrees of freedom
## Multiple R-squared:  0.9172, Adjusted R-squared:  0.9171
## F-statistic: 2.789e+04 on 20 and 50381 DF,  p-value: < 2.2e-16
```

Table A10. Variance inflation factor for multicolinearity after removing x, y, and z.

```
##             GVIF Df GVIF^(1/(2*Df))
## carat   1.340433  1        1.157771
## cut     1.937272  4        1.086173
## color   1.158982  6        1.012371
## clarity 1.338781  7        1.021059
## depth   1.379164  1        1.174378
## table   1.790162  1        1.337969
```

Table A11. Linear regression with all variables and second order variables.

```
##
## Call:
## lm(formula = price ~ . + I(depth^2) + I(table^2), data = diamonds1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11642.8   -449.5   -128.7    294.1   5856.2
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.395e+04  3.149e+03   -7.605 2.90e-14 ***
## carat        8.789e+03  5.128e+01  171.381  < 2e-16 ***
## cut.L        3.939e+02  1.977e+01   19.922  < 2e-16 ***
## cut.Q       -1.974e+02  1.481e+01  -13.327  < 2e-16 ***
## cut.C        1.011e+02  1.157e+01    8.732  < 2e-16 ***
## cut^4       -9.850e+00  9.005e+00   -1.094 0.274025
## color.L     -1.432e+03  1.276e+01 -112.162  < 2e-16 ***
## color.Q     -4.967e+02  1.162e+01  -42.749  < 2e-16 ***
## color.C     -1.264e+02  1.082e+01  -11.678  < 2e-16 ***
## color^4      5.341e+01  9.924e+00    5.383 7.38e-08 ***
## color^5     -4.436e+01  9.328e+00   -4.756 1.98e-06 ***
## color^6     -2.877e+01  8.435e+00   -3.411 0.000647 ***
## clarity.L    3.188e+03  2.187e+01  145.748  < 2e-16 ***
## clarity.Q   -1.592e+03  2.030e+01  -78.446  < 2e-16 ***
## clarity.C    6.906e+02  1.733e+01   39.843  < 2e-16 ***
## clarity^4   -3.291e+02  1.385e+01  -23.763  < 2e-16 ***
## clarity^5    1.975e+02  1.132e+01   17.443  < 2e-16 ***
## clarity^6    2.737e+01  9.859e+00    2.777 0.005497 **
## clarity^7    8.013e+01  8.734e+00    9.175  < 2e-16 ***
## depth        7.059e+02  9.259e+01    7.624 2.51e-14 ***
## table        8.474e+01  5.221e+01    1.623 0.104549
## x           -4.810e+02  3.442e+01  -13.977  < 2e-16 ***
## y            7.068e+01  2.684e+01    2.633 0.008467 **
## z           -2.697e+01  2.513e+01   -1.073 0.283119
## I(depth^2)  -5.999e+00  7.551e-01   -7.945 1.98e-15 ***
## I(table^2)  -8.700e-01  4.474e-01   -1.944 0.051851 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 791.7 on 50376 degrees of freedom
## Multiple R-squared:  0.9181, Adjusted R-squared:  0.918
## F-statistic: 2.258e+04 on 25 and 50376 DF,  p-value: < 2.2e-16
```

Figure A1. Distribution of diamond prices.

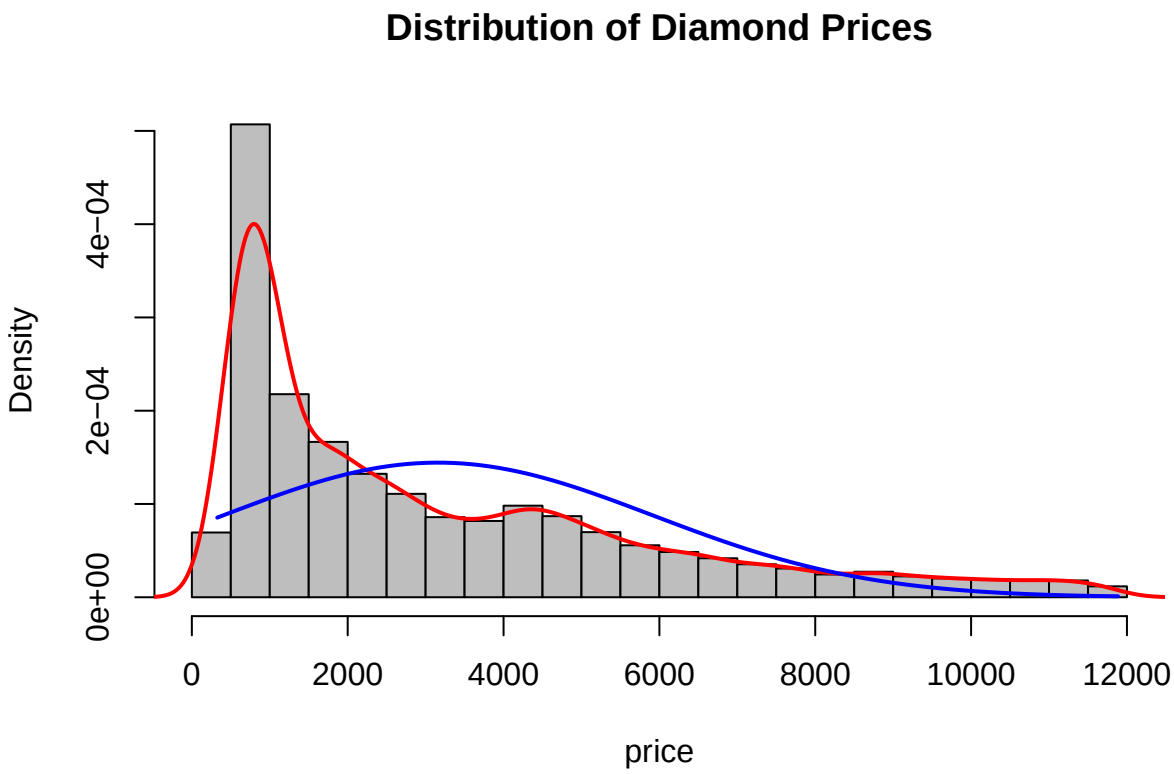


Figure A2. Distribution of diamond weights.

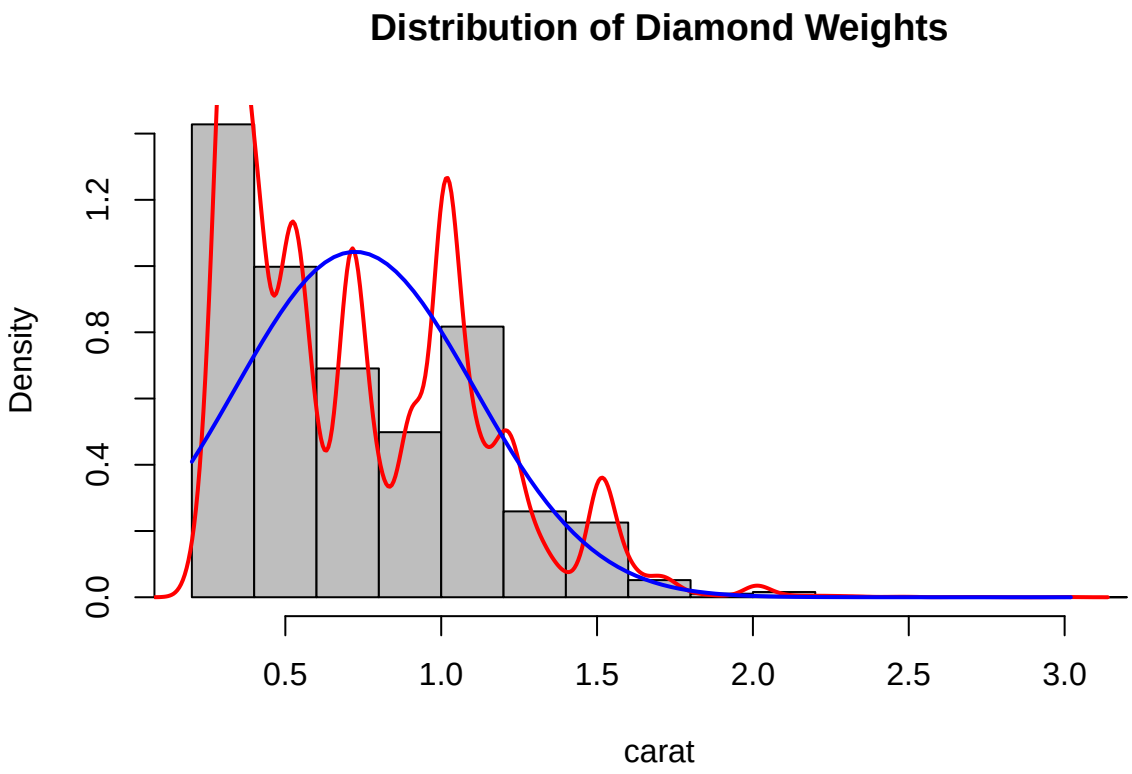


Figure A3. Distributions of diamond cuts.

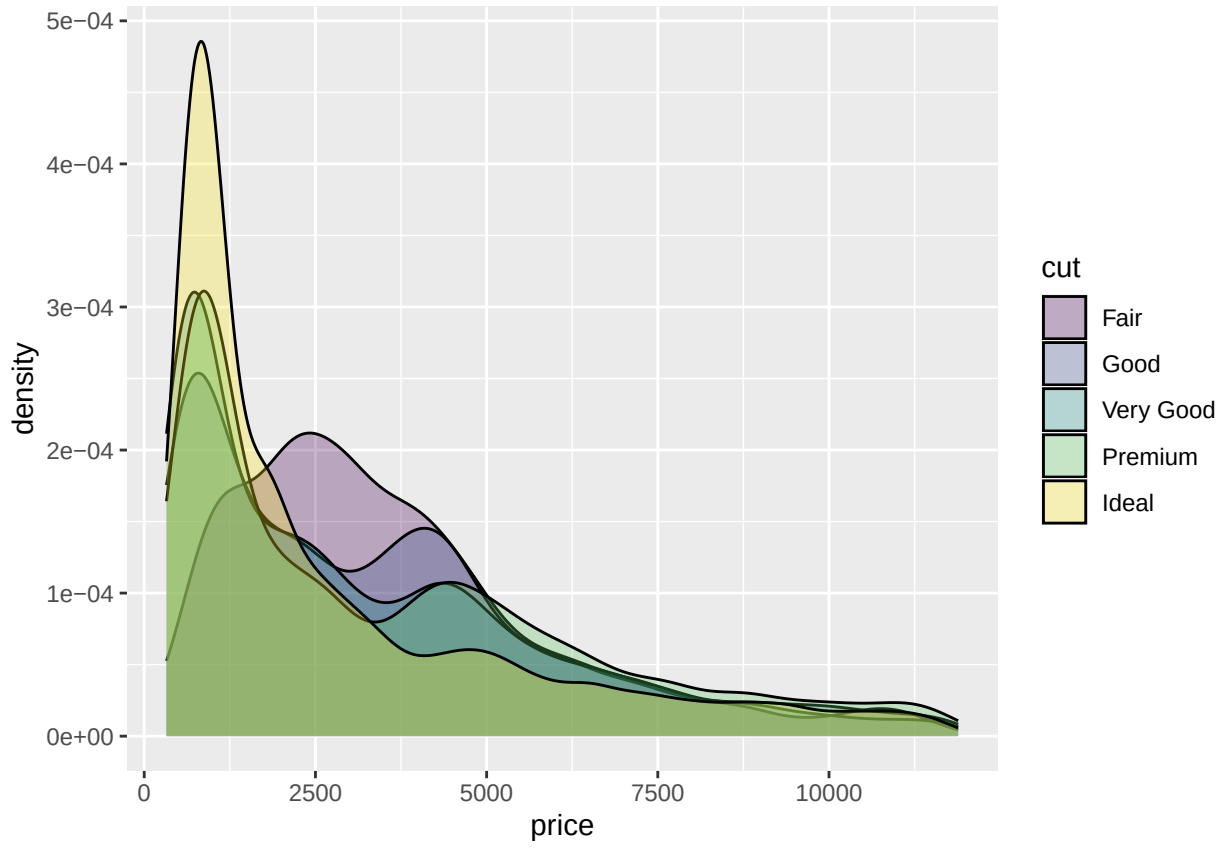


Figure A4. Variable screening for model selection.

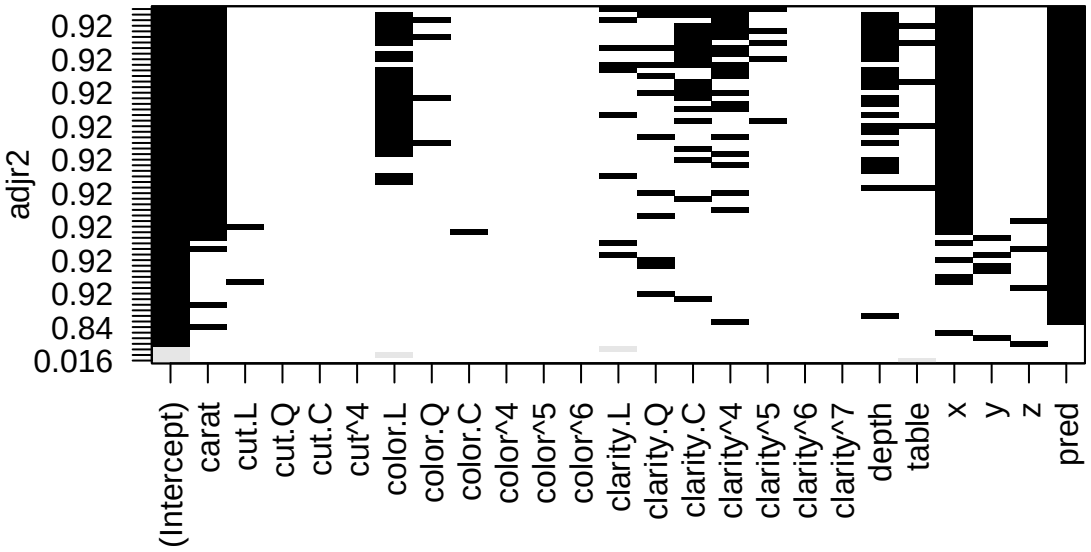
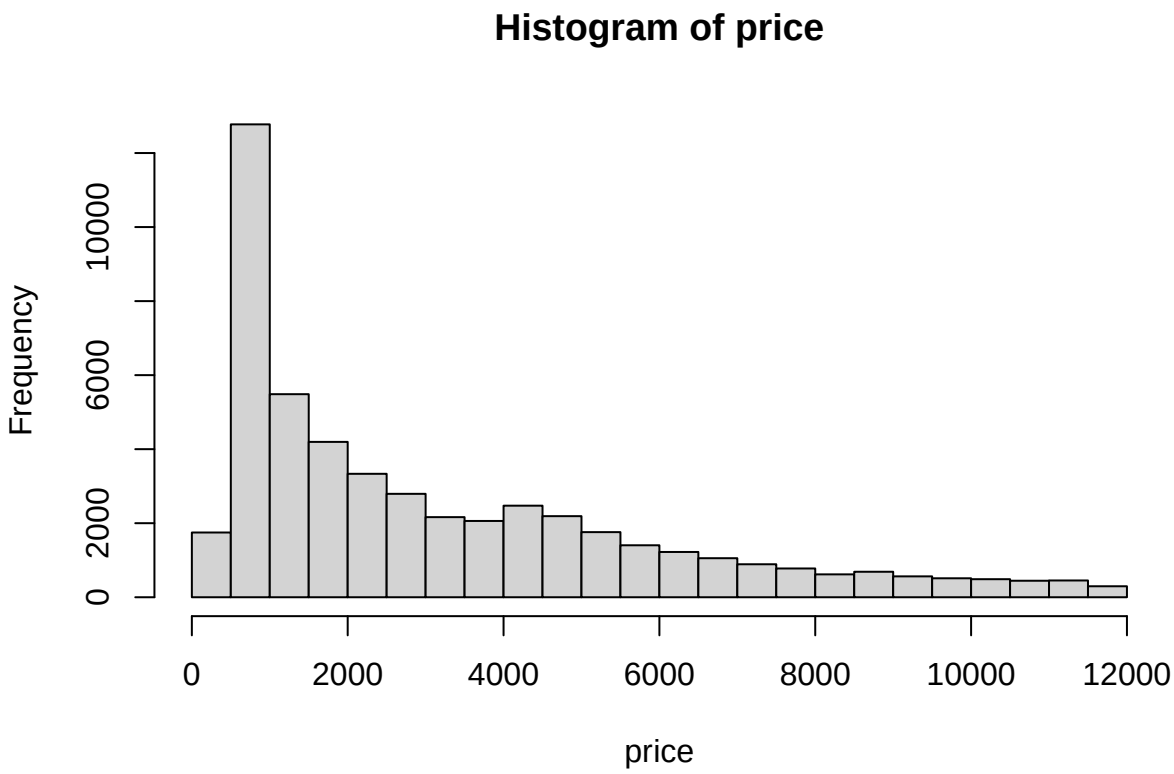


Figure A5. Histogram of diamond prices.


Histogram of price
Frequency
0
2000
6000
10000
0
2000
4000
6000
8000
10000
12000
price