Assignment 3: Diamonds Price Estimation

MEF - BDA 503

Nov 21, 2017

Your assignment consists of finding the price of a diamond given its properties. You will use the `diamonds` data set in ggplot2 package (which is inside tidyverse). You need to do your exploratory analysis well and come up with a predictive model. Your performance depends on the difference between the actual price of the diamond and the predicted price by the model. Use the `price` column as the response variable and other columns (except `diamond_id`) as predictors.

You are recommended to use CART but welcome to use any advanced method you like. Add your exploratory analysis to form a basis of your model and include references (with links) if you are inspired from similar analysis. Use the following code (and random seed) to form your train and test data. Remember, you should train your model on the train data and your real performance depends on the test data.

```
set.seed(503)
library(tidyverse)
diamonds_test <- diamonds %>% mutate(diamond_id = row_number()) %>%
    group_by(cut, color, clarity) %>% sample_frac(0.2) %>% ungroup()

diamonds_train <- anti_join(diamonds %>% mutate(diamond_id = row_number()),
    diamonds_test, by = "diamond_id")

diamonds_train
```

```
## # A tibble: 43,143 x 11
##    carat       cut color clarity depth table price     x     y     z
##    <dbl>     <ord> <ord>   <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  0.23     Ideal     E     SI2  61.5    55   326  3.95  3.98  2.43
##  2  0.21   Premium     E     SI1  59.8    61   326  3.89  3.84  2.31
##  3  0.23      Good     E     VS1  56.9    65   327  4.05  4.07  2.31
##  4  0.29   Premium     I     VS2  62.4    58   334  4.20  4.23  2.63
##  5  0.24 Very Good     J    VVS2  62.8    57   336  3.94  3.96  2.48
##  6  0.24 Very Good     I    VVS1  62.3    57   336  3.95  3.98  2.47
##  7  0.26 Very Good     H     SI1  61.9    55   337  4.07  4.11  2.53
##  8  0.22      Fair     E     VS2  65.1    61   337  3.87  3.78  2.49
##  9  0.23 Very Good     H     VS1  59.4    61   338  4.00  4.05  2.39
## 10  0.30      Good     J     SI1  64.0    55   339  4.25  4.28  2.73
## # ... with 43,133 more rows, and 1 more variables: diamond_id <int>
```

```
diamonds_test
```

```
## # A tibble: 10,797 x 11
##    carat   cut color clarity depth table price     x     y     z
##    <dbl> <ord> <ord>   <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  3.40  Fair     D      I1  66.8    52 15964  9.42  9.34  6.27
##  2  0.90  Fair     D     SI2  64.7    59  3205  6.09  5.99  3.91
##  3  0.95  Fair     D     SI2  64.4    60  3384  6.06  6.02  3.89
##  4  1.00  Fair     D     SI2  65.2    56  3634  6.27  6.21  4.07
##  5  0.70  Fair     D     SI2  58.1    60  2358  5.79  5.82  3.37
##  6  1.04  Fair     D     SI2  64.9    56  4398  6.39  6.34  4.13
##  7  0.70  Fair     D     SI2  65.6    55  2167  5.59  5.50  3.64
##  8  1.03  Fair     D     SI2  66.4    56  3743  6.31  6.19  4.15
```

```
##  9  1.10  Fair     D     SI2  64.6    54  4725  6.56  6.49  4.22
## 10  2.01  Fair     D     SI2  59.4    66 15627  8.20  8.17  4.86
## # ... with 10,787 more rows, and 1 more variables: diamond_id <int>
```