Jesse Toftum
James Webb

Dr. Grant Gustafson
Math 2270-4

PREDICTIVE ANALYTICS OF COMMERCIAL DIAMONDS

Introduction to Query

The market price of most commercial goods is rather formulaic. A metric of materials and labor with their respective frequency of production and rarity of skills needed, coupled with the successive profit margins determined by economic forces, forms the basis for the overall cost of a good. The calculations leading to this figure are relatively simple and surprisingly accurate when the variables of each phase of production are known.

However, determining the cost of a good by features of the final product alone is much more challenging. This is especially true in volatile market sectors where demand is highly artificial. An archetypal case illustrating this is the cosmetic diamond industry. Diamonds are popular in western jewelry due to a longstanding tradition of featuring the transparent white crystal in wedding or engagement rings. Cultural myths surrounding the extreme rarity of diamond have placed it at the highest tier of luxury gemstones, despite its fairly plentiful harvesting in the modern day.

This dynamic makes diamonds a particularly interesting case study in a top-down approach to cost analysis, as the factors are less straightforward than the utility of most materials. Note that the data of focus in this study is diamonds cut and refined for cosmetic use, not the market of diamond grit abrasives used in tooling and manufacturing. Here we will be primarily interested in answering two questions:

1) Which measurable qualities of a diamond are most strongly correlated to its market price?

2) With what accuracy can we create a cost-prediction model based on these qualities?

This will be accomplished through methods of machine-learning (ML) and least-squares regression.

Introduction to Methodology

While the full Python output is included, the semantics of its reasonably complicated source code will not be fully discussed, as it would distract from the core discussion of the groundwork mathematics. However, to serve as a readable context for the conceptual background to various parts of the process, several cells of code will be annotated here.

Qualitative Features

Below are the first 5 entries of our dataset, which had information on a total of 53,920 diamonds.

[6]:

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	5	6	4	61.5	55.0	326	3.95	3.98	2.43
2	0.21	4	6	5	59.8	61.0	326	3.89	3.84	2.31
3	0.23	2	6	7	56.9	65.0	327	4.05	4.07	2.31
4	0.29	4	2	6	62.4	58.0	334	4.20	4.23	2.63
5	0.31	2	1	4	63.3	58.0	335	4.34	4.35	2.75

Explanation of Variables [with Numerical Hot-Encoding Used]

- Carat: Mass (1 carat = 200 mg)
- Cut: Grade of cut, measured categorically as
  Fair [1] - Good [2] - Very Good [3] - Premium [4] - Ideal [5]
- Color: Grade of colorlessness, measured categorically as
  J [1] - I [2] - H [3] - G [4] - F [5] - E [6] - D [7]
- Clarity: Grade of blemish, measured categorically by inclusions as
  I_3 [1] - I_2 [2] - I_1 [3] (Included)
  SI_2 [4] - SI_1 [5] (Slightly Included)
  VS_2 [6] - VS_1 [7] (Very Slightly Included)
  VVS_2 [8] - VVS_1 [9] (Very Very Slightly Included)
  IF [10] - FL [11] (Flawless)
- Depth: Percentage ratio (x100) of total height by total width
- Table: Percentage ratio (x100) of top plane width by total width
- Price: Cost in USD
- X,Y,Z: Dimensions in mm of cubic extrapolation of diamond

The following data was also produced for each entry using the base variables.

	volume	density	carat_table	carat_depth	clarity_cut_color	carat_clarity_cut_color
1	38.202030	0.006021	12.65	14.145	120	27.60
2	34.505856	0.006086	12.81	12.558	120	25.20
3	38.076885	0.006040	14.95	13.087	84	19.32
4	46.724580	0.006207	16.82	18.096	48	13.92
5	51.917250	0.005971	17.98	19.623	8	

- Volume: X·Y·Z (mm3)
- Density: Carat/Volume (carat/mm3)

- Carat-Table: Carat·Table (carat)
- Carat-Depth: Carat·Depth (carat)
- Clarity-Cut-Color (CCC): Clarity·Cut·Color (unitless)
- Carat-Clarity-Cut-Color (CCCC): Carat·Clarity·Cut·Color (carat)

Here are some standard statistics for each variable across the entire dataset.

	carat	cut	color	clarity	depth
count	53920.000000	53920.000000	53920.000000	53920.000000	53920.000000
mean	0.797698	3.904228	4.405972	6.051502	61.749514
std	0.473795	1.116579	1.701272	1.647005	1.432331
min	0.200000	1.000000	1.000000	3.000000	43.000000
25%	0.400000	3.000000	3.000000	5.000000	61.000000
50%	0.700000	4.000000	4.000000	6.000000	61.800000
75%	1.040000	5.000000	6.000000	7.000000	62.500000
max	5.010000	5.000000	7.000000	10.000000	79.000000

	table	price	x	y	z
count	53920.000000	53920.000000	53920.000000	53920.000000	53920.000000
mean	57.456834	3930.993231	5.731627	5.734887	3.540046
std	2.234064	3987.280446	1.119423	1.140126	0.702530
min	43.000000	326.000000	3.730000	3.680000	1.070000
25%	56.000000	949.000000	4.710000	4.720000	2.910000
50%	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	59.000000	5323.250000	6.540000	6.540000	4.040000
max	95.000000	18823.000000	10.740000	58.900000	31.800000

	volume	density	carat_table	carat_depth	\
count	53920.000000	53920.000000	53920.000000	53920.000000	
mean	129.897567	0.006127	46.025483	49.276657	
std	78.219789	0.000178	27.744367	29.343722	
min	31.707984	0.000521	11.000000	11.800000	
25%	65.189759	0.006048	22.550000	24.520000	
50%	114.840180	0.006117	40.320000	43.594000	
75%	170.846415	0.006190	61.000000	64.872000	
max	3840.598060	0.022647	295.590000	328.155000	

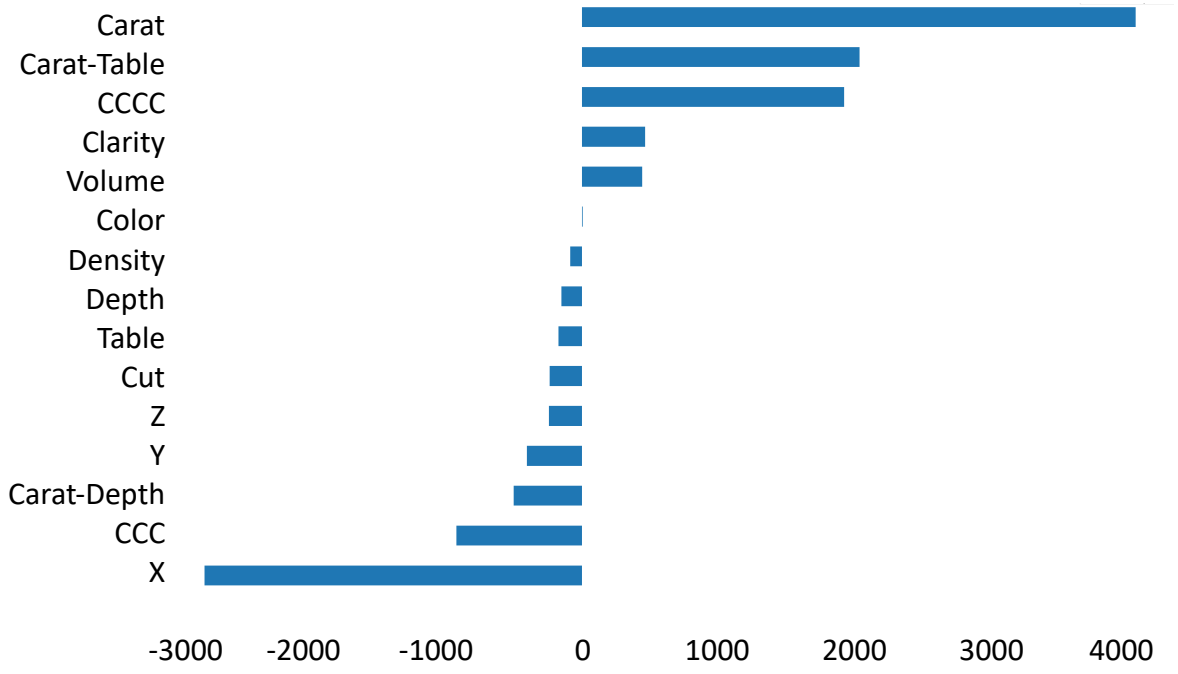
	clarity_cut_color	carat_clarity_cut_color
count	53920.000000	53920.000000
mean	105.551298	72.580075
std	61.382226	49.310398
min	3.000000	1.700000
25%	60.000000	37.260000
50%	96.000000	61.200000
75%	144.000000	94.400000
max	350.000000	409.500000

Feature Importance Through Lasso

1) Which measurable qualities of a diamond are most strongly correlated to its market price?

To answer our initial question, least absolute shrinkage and selection operator (lasso) was an obvious choice. Lasso is a method of analyzing regression through quantifying the predictive weight of certain features or variables in a least-squares regression model. It is also effective in other types of regression.

Here are the results of our lasso analysis, where price is the target of prediction constrained by the selected variable. Each figure is related through the lens of root-mean-square-error (RMSE).



	Feature Importance
x	-2759.110468
clarity_cut_color	-917.439737
carat_depth	-499.569926
y	-405.773400
z	-241.202641
cut	-236.358452
table	-174.657900
depth	-150.369281
density	-84.356368
color	6.781231
volume	440.914746
clarity	460.042407
carat_clarity_cut_color	1913.270318
carat_table	2027.520975
carat	4044.535181

Model Training Through Machine-Learning

2) With what accuracy can we create a cost-prediction model based on these qualities?

Our second question has a slightly more complex answer. We generated 20 linear regression models using ML with accompanying analysis standard to ML data manipulation. The computer science and mathematics behind this process is vast and far beyond the scope of explanation here, but at its core lies a very familiar process to linear algebra: support-vector machines (SVM). SVM uses techniques of vector algebra and normalized margins to perform linear regression.

The dictating figure chosen to understand each model's predictive power was RMSE, as it is a simple figure in USD that streamlines the process of comparing models. To broadly assess the validity of the generated models, we also created a dummy model. It returned an RSME of -3794.18, which is several times greater than anything seen below. This ensures us that the accuracy of the ML models below was not trivial.

Here are the RSME values for a simple average of all regressors, an average weighted by the individual accuracy of each regressor, and the best single regressor.

```
[48]: simple_average_regressors.score(test_features, test_target)

    'y_true: [1141  770  696 ...  571  778  999]'
    ('y_pred: [1309.5635402   409.33626366  522.7758654  ...  133.48408627  '
     '868.14109466\n'
     '  837.21288656]')
```

```
[48]: -726.3675103064977
```

```
[49]: weighted_average_regressors.score(test_features, test_target)

    'y_true: [1141  770  696 ...  571  778  999]'
    ('y_pred: [1332.10280425  511.93563139  596.47136552 ...  336.0316964   '
     '849.14351036\n'
     '  903.63981276]')
```

```
[49]: -625.2447757706211
```

```
[50]: best_single_regressor.score(test_features, test_target)

    'y_true: [1141  770  696 ...  571  778  999]'
    ('y_pred: [1397.54735918  674.858574    713.2889548  ...  597.28969038  '
     '806.14806765\n'
     ' 1025.94845563]')
```

```
[50]: -572.9178777392304
```

To help contextualize these figures, consider that the mean price of the dataset was $3,930.99 and the interquartile range between Q1 and Q3 (the “middle half”) was $4,374.25. This indicates an error margin with respect to RSME of 14.6-18.5% for the mean and 13.1-16.6% for the interquartile range. This could easily be considered a success. While the details won’t be discussed, a very popular form of ML regression called TPOT was also performed on the dataset to produce an RSME of -511.54. TPOT is an industry standard of sorts for ML analysis, and this finding further legitimizes the generated models.

Conclusion and Further Study

Methods of regression generation, analysis, and machine-learning through linear algebra and computer science are in constant development. A natural extension of our findings here would be an explicit detailing of the most accurate prediction models that could be used with multi-variable input data, as our initial focus has been on the regression processes and error analysis. While this would only be useful to a niche demographic of luxury diamond traders, the crystal is a staple in high-end asset diversification and these calculations extended might prove invaluable to investors. From an economics perspective, comparing these results to the supply-chain cost calculations as mentioned in the introduction would highlight important chokepoints and gateways in the entire process of diamond manufacture and selling. These applications only scratch the immutable surface of the elusive gemstone and the power of computational mathematics.