Building a Diamond Price Prediction Model

Xan Lamoreux

Professor Dr. Shanker

Fundamentals of Machine Learning (MIS-64060-002)

5 December 2021

**Purpose/Objectives:**

The objectives of this assignment are threefold:

1. To use real-world data
2. To use the appropriate machine learning technique for the business problem, and
3. To present the solution to top-level management.

**Directions:**

1. You may use any source to get your data as long as the data is real-world data. You can either collect data, or download from competitions like Kaggle. Make sure to clearly mention in your report, both the source of data, and how it satisfies the condition for being real-world.
2. Apply whatever machine learning techniques you think are appropriate for the problem. Clearly specify what the central objectives are for this study.
3. Write a paper that describes the objectives, your approach, the results, and your conclusions. It is not necessary to describe the step-by-step account of what you did. Remember that this report is for top-level management. Similarly, also create a presentation of no more than 3 slides (excluding the initial slide) that summarizes the problem, the approach, and the conclusions.

**Background and source of the data:**

The data I have chosen was collected from a Kaggle competition; Predicting diamond prices, Diamonds | datamad 1021-rev <https://www.kaggle.com/c/diamonds-datamad1021-rev/data> (Kaggle, n.d.). The goal of this competition is the prediction of the price of diamonds based on their characteristics (weight, color, quality of cut, etc.). Kaggle has also determined that the evaluation metric chosen for this competition is the RMSE (Root Mean Squared Error).

**Features of the data provided:**

* id: only for test & sample submission files, id for prediction sample identification
* price: price in USD
* carat: weight of the diamond
* cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
* color: diamond colour
* clarity: a measurement of how clear the diamond is
* x: length in mm
* y: width in mm
* z: depth in mm
* depth: total depth percentage = z / mean(x, y) = 2 \* z / (x + y) (43--79)
* table: width of top of diamond relative to widest point (43--95)

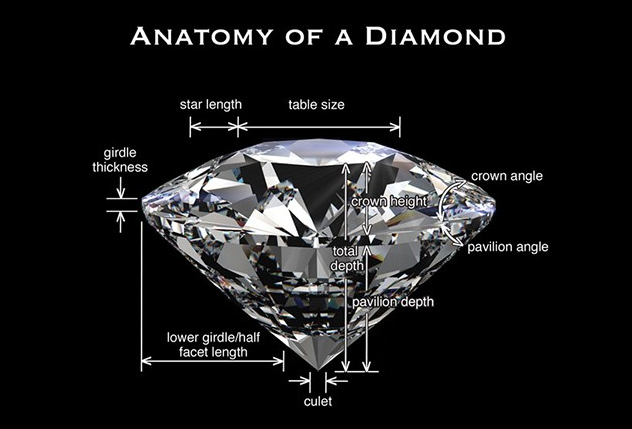
**How the data satisfies the condition for being real-world:**

Replacing a manual process for calculating the price of diamonds to a Machine Learning process increases speed and accuracy saving on labor costs and reducing errors and improving inventory valuation. “Diamonds have a set international market price, unlike other gems. In this respect, diamond is somewhat like gold. However, diamond pricing is much more complicated than gold. While every ounce of gold is the same, diamonds must be graded to determine their value. Even though the 4 Cs is the basic standard for grading, a myriad [of] other factors influence diamond value” (Shang, 2021). The 4 Cs that are the base of grading a diamond are **c**olor, **c**ut, **c**larity, and **c**arat weight. However, there are a number of other properties which can influence the value of a diamond because “gemstones are graded by the rarity of their features” (Clark, 2021). The data provided by the Kaggle competition is in-line with the current prices of diamonds as published by the International Diamond Exchange which can be found at <http://www.idexonline.com/> (IDEX, n.d.). It is very rudimentary for pricing a diamond, but will provide consumers a general idea of what they can expect to pay for a diamond by entering an expected weight range, cut, and clarity.

The data and general competition goal are relevant to me as I work for Signet Jewelers, parent company to many nationally recognized jewelry retail brands. In my previous role, I was responsible for planning and forecasting many different metrics (including productivity) and I would often use historical data to forecast an assumption which would be used to predict what would happen in the future. Though this approach was often close enough for setting budget spend, having the ability to predict with greater accuracy would have assisted me in allocating more dollars to needed supplies and human capital resources, resulting in greater efficiency of operations and more productivity.

**Problem:**

Diamond and Gemologists use a manual process for calculating the value of a diamond. Although each measurement of a stone must be viewed and tracked manually for accuracy, the calculation process is also manual through hand-written calculation sheets where the worker refers back to a reference sheet for the value of each measurement. That leaves room for human errors as workers are unable to read their own handwriting, or they enter values in the wrong section of the calculation sheet thus incorrectly valuing the diamond they are measuring. My goal to solve this problem will be to provide a database that can calculate the value of a diamond for them.



Calculation sheet for each stone contains:

* Shape
* Weight
* Clarity
* Cut
* Color
* Impurities
* Any physical occlusions which would lower the value of the stone
* All of the physical measurements listed in the image on the left

**Approach:**

Using a data file which includes many combinations of the measurable variables for diamond grading along with the current value of a diamond that has those characteristics would allow me to determine which characteristics are most relevant to predicting diamond pricing for any stones that I do not have listed in this data set. This would be useful in building an end-user database that workers can reference easily. Workers can quickly enter the measurements of a stone and the database would calculate the price of the stone for them. The computer algorithm would replace the manual process of measuring each variable, writing down the value of each variable, and then using a manual calculation sheet to price the diamond. This promotes accuracy, improves productivity, and reduces human error.

**Potential Future Improvements by Investing:**

Looking to the future of diamond grading, we could further improve accuracy by investing in high-definition laser measurement devices. These devices would automatically capture every facet of the diamond in seconds. The device would replace the human process for entering the measurements into the database, so where one person could grade and enter each measurement into the database at approximately 25 pieces per hour (pph), the laser’s automatic input could increase the pph to 1,200.

**Results:**

The overarching goal of putting this Machine Learning algorithm into practice is to predict the price of a diamond based on its characteristics (weight, color, quality of cut, etc.). The program results, as I have currently written it, has a 91.37% accuracy rate. This level of accuracy will improve inventory valuation resulting and more accurate bookkeeping. This is far more accurate than hand-written calculations. The long-term potential for the future growth of this project could reduce spend, increase productivity, eliminate errors, and confidently value diamond product with little to no errors.

References

Clark, D. (2021). A Consumer’s Guide to Gem Grading. Retrieved December 5, 2021, from International Gem Society: https://www.gemsociety.org/article/a-consumers-guide-to-gem-grading/

IDEX. (n.d.). International Diamond Exchange. Retrieved from http://www.idexonline.com/

Kaggle. (n.d.). Predicting diamonds prices. Retrieved December 1, 2021, from Kaggle: https://www.kaggle.com/c/diamonds-datamad1021-rev/data

Shang, P. (2021). The Fifth C: What Determines Diamond Cost? Retrieved December 5, 2021, from International Gem Society: https://www.gemsociety.org/article/what-determines-diamond-cost/