Diamonds

Stanislav Fatkhutdinov

University of Maryland Global Campus

Assignment 3

DATA 610 Section 9040

Decision Management Systems

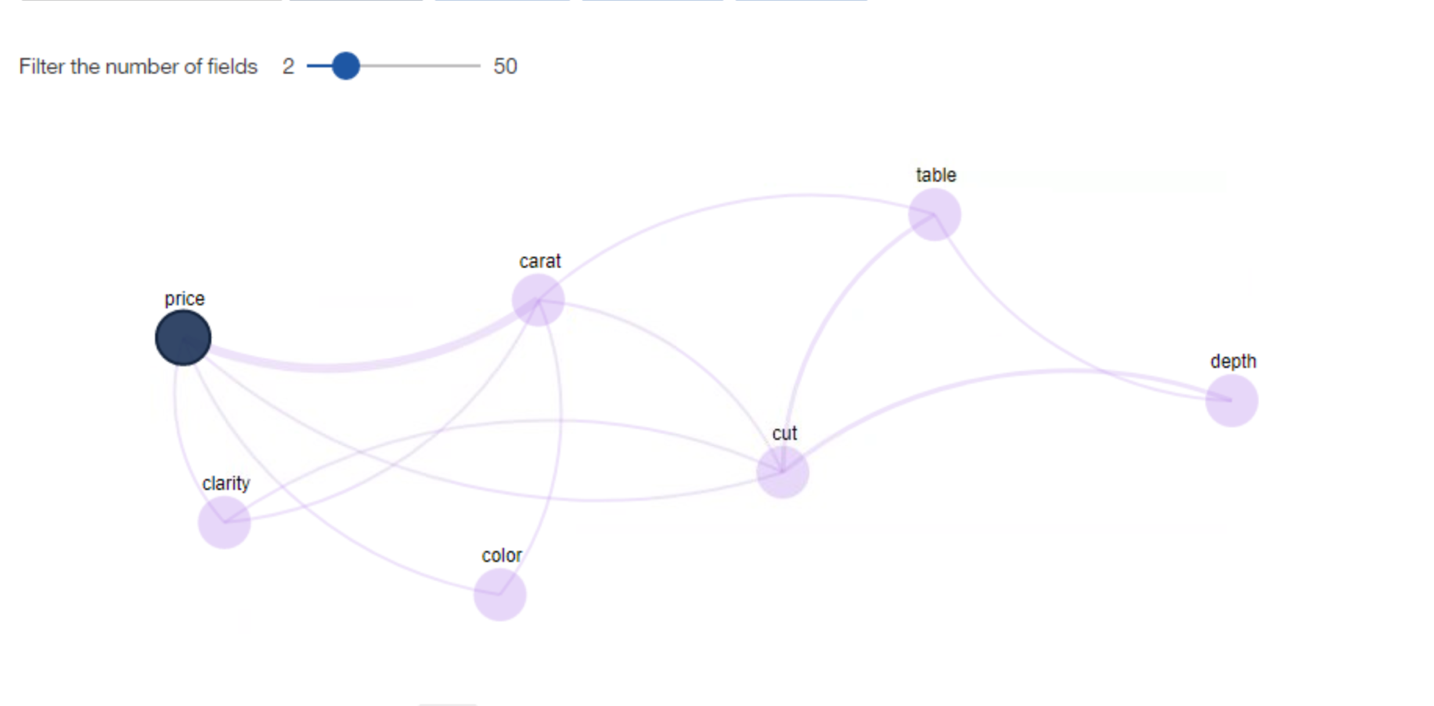
Vahe Heboyan

Diamonds

**Introduction:**

The dataset describes the features of 54,000 diamonds. The set consists of several features: price in US dollars; carat weight of the diamond; cut quality; diamond color and clarity; x, y, z dimensions, and the derived from the depth and another derived property – table. The price ranges from $326 to $18,823 per diamond. The carat weight fluctuates between 0.2 and 5.01 carats. The cut quality is divided between Fair, Good, Very Good, Premium, and Ideal. Diamonds are also separated by color, where J is the worst, and D is the best color. The clarity feature is measured by how clear the diamond is; it ranges from I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF, where 'I1' is the worst clarity and 'IF' is the best possible clarity for a diamond. Other features of the diamonds in the dataset are x, y, and z dimensions representing length, width, and depth. Depth and Table are calculated properties of the diamonds describing the total depth percentage and the width of the top of a diamond relative to the widest point, respectively (Magrawal, 2017). The following formula calculates depth:

Figure 1 below represents the snapshot of the data described above. I've decided to hide-from-user several columns in the dataset. The number column is redundant since the Cognos Analytics automatically creates Row Id; plus, the number column is not relevant to the diamonds' features. I also hid the x, y, z column since their derived property of depth is the one I'm going to use for my predictions. For this assignment, I will use the Diamond dataset to create a decision tree to predict the diamonds' price and carat based on the features.



Figure

**Predictive Models, Results, and Discussion:**

My first model was designed to predict a diamond's price based on its features. Using Cognos Analytics (CA), I've created the decision tree classifier using the driver analysis visualization tool. The following Figure 2 provides the initial decision tree that separates the diamond data based on a particular diamond's features. The CA software calculated (IBM Cognos Analytics) that the carat weight is the most crucial factor affecting the diamond's price; it is five times as important as any other field in the dataset.

A picture containing graphical user interface

Description automatically generated

Figure

The program initially divided the diamonds into four categories, diamonds more massive than 1.13 carats, diamonds between 0.9 and 1.13 carats, between 0.53 and 0.9, and diamonds lighter than 0.53 carats. The most expensive diamonds were more massive than 1.13 carats. Conversely, the least expensive diamonds were the lightest ones. The data demonstrated that almost 40% of the diamonds in the dataset weigh less than 0.53 carats. Still, most of those diamonds are homogeneous and do not have further separating characteristics affecting the price. This observation is more evident when looking at the tree sunburst diagram below in Figure 3.

Chart, sunburst chart

Description automatically generated

Figure

Based on the CA tree diagram classification, the most expensive diamonds that are more massive than 1.13 carats, have clarity of VS1 through IF and are E, F, D, G colors. In the Diamond dataset, those diamonds accounted for only 4% of all diamonds. The second most expensive color combinations were J, H, and I.

The second model was directed to determine the carat weight of a particular diamond based on the dataset's characteristics. The carat weight was mostly dependent on the diamond's price, which confirms the first model's findings. The price was four times as important to predict the diamond's carat weight than any other feature present in the dataset. The tree diagram classifying the carat weight of the diamonds is presented in Figure 4 below.

A picture containing graphical user interface

Description automatically generated

Figure

As seen in Figure 4 above, the CA divided diamonds into five price categories, the cheapest diamonds, below $828, weighted the least, and the most expensive, naturally, weighted the most (IBM Cognos Analytics). However, there is a slight deviation when determining the weight of the diamond. The color mattered only for 12% of the dataset, where is clarity played a more prominent role in deciding the carat weight of a diamond. Figure 5 demonstrates that the most massive diamonds with clarity SI2 and I1, which are a little strange since those clarities are considered the lowest when measuring the diamond clarity.

Chart, sunburst chart

Description automatically generated

Figure

I've decided to look closer into the relationship between the carat and price for each clarity to see if the tree diagram was correct in predicting that the most expensive and the most massive diamonds were SI2 clarity. Figure 6 confirmed my suspicions and also demonstrated that clarity I1 and SI2 are the most massive diamonds.

Chart, histogram

Description automatically generated

Figure

It was interesting that the cut did not play a significant role during the decision tree model creation. Even though the premium cut was more expensive, the rest on average, second only by the fair cut diamonds, didn't play a significant role during the initial model creation to determine the diamonds' price and carat weight. However, in Figure 7, it is clear that premium cut diamonds weighing the most have the highest price, followed closely by lighter high-priced fair cut diamonds.

Chart, bar chart

Description automatically generated

Figure

**Organizational Application:**

I think the decision tree is suitable for developing an initial insight when dealing with large datasets. My organization can benefit from implementing this model to sort relevant data for our projects quickly. I can't discuss any details of those projects, but I can extrapolate to a closely related example. My HR department can use this quick tool when determining the initial salary offer for a new applicant based on their skills and experiences. The applicant's data can be aggregated from their resume and organized against the backdrop of contracting costs to determine how much money will be the most optimal for the individual and the company.

**Conclusion:**

The decision trees are a great way to organize and analyze extensive data; however, it lacks precision when necessary to see granularity. As seen in my example, the diamond cut wasn't determined to be a relevant feature affecting the price or weight of the diamonds; however, a closer look reveals that premium cuts are more expensive.

# References

IBM Cognos Analytics. (n.d.). *Explorations User Guide.* Retrieved September 2020, from www.ibm.com: https://www.ibm.com/support/knowledgecenter/SSEP7J\_11.1.0/com.ibm.swg.ba.cognos.ca\_explorations.doc/ca\_explorations.pdf?view=kc

Magrawal, S. (2017, May 24). *Diamonds.* Retrieved from www.kaggle.com: https://www.kaggle.com/shivam2503/diamonds/metadata