Pricing Diamonds

BAN 525

Khayrayyah Haamid-Day

There are a variety of considerations that play a role in the pricing of diamonds though this was not truly standardized until the mid-twentieth century. The focus of this analysis will be on the five determinants of pricing diamonds which are carat size, clarity, color, depth, and cut of stone. The carat refers to the weight of the diamonds which are measured in metric carats, one carat is equal to 0.2 grams and the largest diamond found weighs in at over 3,000 carats. Clarity refers to the level of blemishes and imperfections present on the diamond, a diamond without these is rare and as such are expensive; there are 11 grades of clarity. For diamonds, the closest it is to colorless the more valuable it is, this is excluding colors like pink and blue. Diamonds tend to range from colorless to hints of yellow and brown. Depth of a diamond is a measurement of the diamond from top to bottom and is typically expressed as a percentage. The best depth is dependent on the type of cut the diamond is. The cut quality is sometimes considered the most important factor of a diamonds beauty and is determined by several different measurements. With this dataset, the two key factors of the cut are depth and table percentage.

This analysis will use “Price” as the response variable, and the predictor variables will be “Cut”, “Depth”, “Clarity”, “Color”, and “Carat Weight”. Preliminary analysis indicates there is a strong relationship between carat weight and price. In terms of cut, color, and clarity, there are diamond types that are rarer than others which can impact price. The statistical methods that will be used are the OLS method as a benchmark for comparison and two different Neural Network methods. The first NN method will consist of a one layer and three nodes with TanH activation function. The second NN method will be more complex by using two layer and three nodes for all three activation functions. Both models will also be constructed with random seed 123 replication.

The first method OLS, or Ordinary Least Squares, is a regression that works well as a benchmark model. This methods simplicity makes it easy to understand and highly interpretable. On the other hand, it is sensitive to outliers and tends to have overfit models. The other two models will be built with the Neural Network method. This method is very efficient as well as flexible as it allows for complex model building. A great aspect about this method is the number of inputs and layers that can be implemented. The downside to this method is that they are known as black boxes and can be incredibly difficult to interpret. They also rely on the training data which can lead to overfitting. The first NN will only have one layer and three nodes in TanH while the second will have two layers and three nodes in TanH, Linear, and Gaussian. Both models will also be built with a random seed of 123. A validation column will be used to build all three models with a 60/20/20 split into a training, validation, and test set as well as a random seed of 123.

Text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

The OLS method returned with very high RSquare consistent values in the training, validation, and test set but it also has a high RASE. The consistency among the three sets is a bit unexpected as usually there is a high RSquare in the training set that drastically drops in the test set. The first Neural Network model, which only used one layer and three nodes in TanH, had high RSquare values in all three sets with little variation among the sets and showed lower RASE than the benchmark model did. The final model is the second Neural Network model that was built with two layers and three nodes in all three activation functions. This model showed the highest RSquare values thus far for all three sets and had the lowest RASE compared to the previous two models. It is interesting to see what looks like a continuous model improvement from OLS to the final NN.

Graphical user interface, application

Description automatically generated

The model comparison allows for a clear understanding of each model in all three sets and shows very close values from RSquare to AAE. The OLS model maintained a 0.91 and the highest RASE and AAE throughout the sets while the NN models held the highest values and lowest RASE and AAE. Both NN models had extremely close test set values of 0.96 and 0.97 with the biggest difference being more evident in the RASE and AAE values. The Complex NN had the highest RSquare but the lowest RASE and AAE. Due to these values, the Complex NN is the best model in this analysis.

Text, table

Description automatically generated

Chart

Description automatically generated

As mentioned before Neural Networks are typically difficult to interpret due to the complexity of it so this analysis will focus on the model’s profiler to interpret the results. An important question to ask from this model is what predictor variable is the most important to the response variable, price. By far the most important variable is carat weight, which is logical as many expensive diamonds are of larger carat size. For example, the largest diamond weighs over 3,000 grams and has a price tag of over USD 400 million. The second most important variable is color wherein the less color is present, the more valuable it is because a colorless diamond is rare. The third most important variable is clarity which is referring to the presence of imperfections and blemishes that are present. Diamonds with poor clarity could appear dirty and have very visible imperfections which is not very desirable especially for an engagement ring.

Graphical user interface, chart

Description automatically generated

The prediction profiler can also be used to understand different possible scenarios, such as determining if buying a diamond ring that is on sale is a good value. If a diamond is on sale for $9,000, with a 1.5 carat, E color, VS1 clarity, 81 Depth, and Very Good cut, should it be bought? The profiler indicates that a diamond that meets all these criteria is a great purchase for $9,000 as the typical price is well over $10,000. It is predictions such as these that further prove a strong model. This prediction shows that the Complex Neural Network is a strong model that matches the factors that impact diamond prices in the jewelry industry.