

# Pricing Diamonds

## Introduction

“A Diamond is a girl’s best friend” is a cliché that’s been said many times in the Diamond Industry. The online diamond manufacturer Adiamor takes their product to another level by offering Gemological Institute of America (GIA) certified

diamonds by creating their own all-natural engagement, wedding rings and customized jewelry.<sup>1</sup>

“Before GIA established the 4 C’s” in 1931 by Robert M.

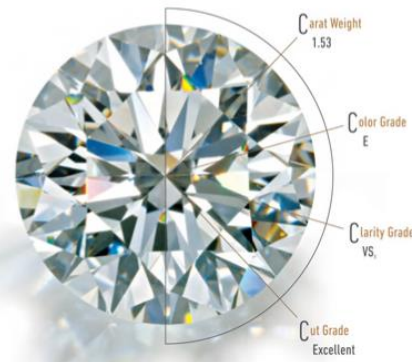
Shipley, there were no established common guidelines

to judge the quality of the diamonds, themselves<sup>2</sup>. For our data set from Adiamor, we are looking at the 4 C’s: Carat weight, Color Grade, Clarity Grade, and Cut Grade in addition to a Depth characteristic in order to predict the price of the diamond itself.

Our given data set has eight variables and 2, 690 undated observations or rows. There are four continuous variables of **Carat Weight** [*Weight of diamond listed to the nearest hundredth of a carat*], **Depth** [*(for round diamonds) and/or width (for fancy shapes) relative to diameter*], **Table** [*Table facet (for round diamonds) and/or width (for fancy shapes) relative to diameter*], and Diamond **Price**<sup>3</sup>. The data set also has 4 nominal or categorical descriptor variables with **Color** [*Color grade of the diamond*], **Clarity** [*Clarity grade determined under 10x magnification*], **Cut** [*Cut grade of the diamond*] and **Report** [*GIA report*]. For the analysis of determining which variables predict the Price, we will eliminate the Table and Report variables.

Following the 4C’s as shown in the graphic above, it would be reasonable to expect Depth to not be a large factor in predicting the price or value of a diamond. It would also be reasonable that the order of the variables in the GIA 4 C’s would predict which of these variables are most important: Carat Weight. Likewise, Cut Grade would be expected to be the least of the 4 C’s as it is listed last.

UNDERSTANDING THE 4 C’s OF DIAMOND QUALITY



<sup>1</sup> <https://www.adiamor.com/About-Adiamor>

<sup>2</sup> <https://www.gia.edu/gia-news-research-purchase-diamond-engagement-ring>

<sup>3</sup> <https://www.adiamor.com/Education/Diamond-Certification>

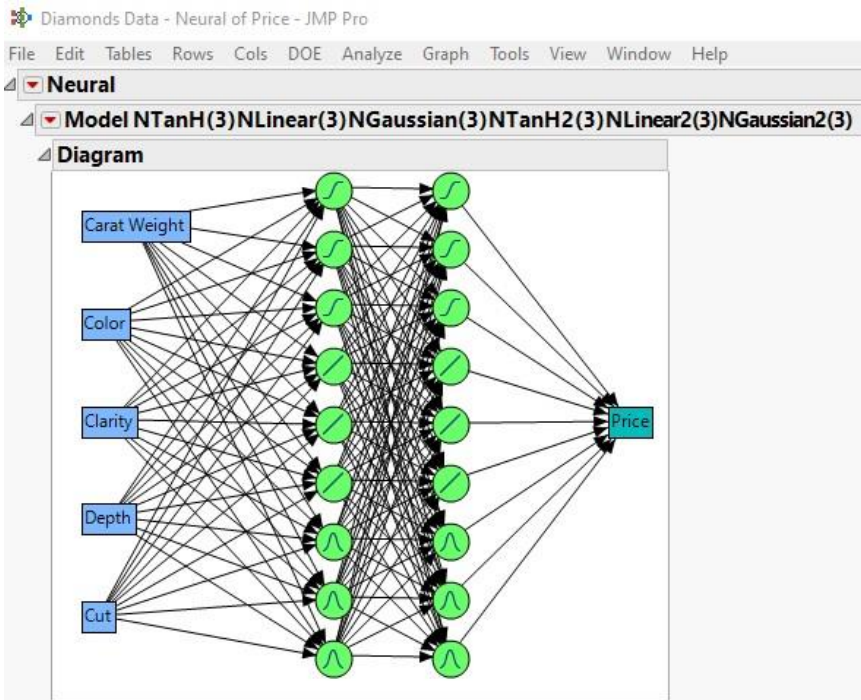
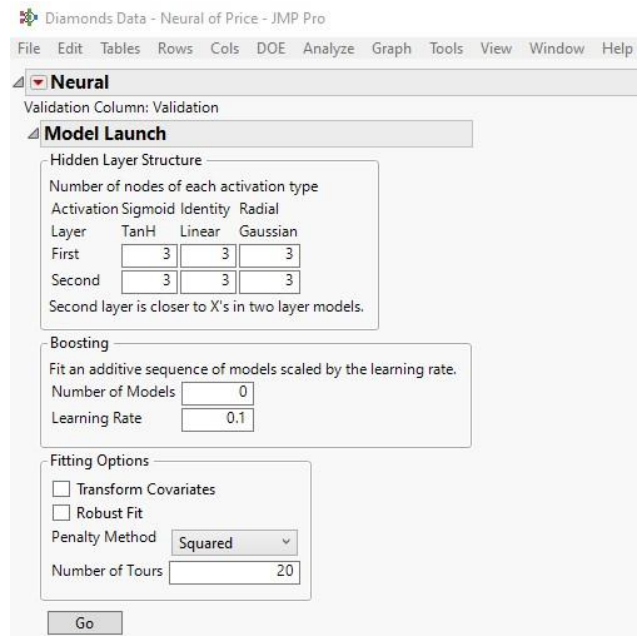
# Pricing Diamonds

## Analysis and Model Comparison

The analysis will start with the Ordinary or Least Squares Model, which is the benchmark, the oldest of the strategies, to compare with. Then two additional models will be used, both of them will be a neural network model which is a simplified model based on the way neurons operate in layers within the human nervous system. According to IBM, there are typically three layers in a neural network model: 1) an input layer, 2) one or more hidden layers and 3) an output layer<sup>4</sup>. Similar to the penalizing methods with Lasso and Elastic Net, there are built in penalties which make estimation models perform more accurately. Additionally, there is a Random Seed of 123 in the models so that each time the model is run, the data is split in the same ways and is reproducible. The model fitting option for number or Tours is set to 20 so that the process is repeated 20 times to ensure a better prediction.

The difference between the two neural network models will be in the number of nodes of each

activation type that decides whether a neuron will be activated or not. The first model will be calculated using the default Neural Network settings which include a first layer with three nodes in the TanH activation function but no other layers or activation types. The second model will be the most complex neural network model with both a first and second layer with three nodes in the TanH (*the hyperbolic tangent function which is supposed to account for linear and non-linear functions*), Linear

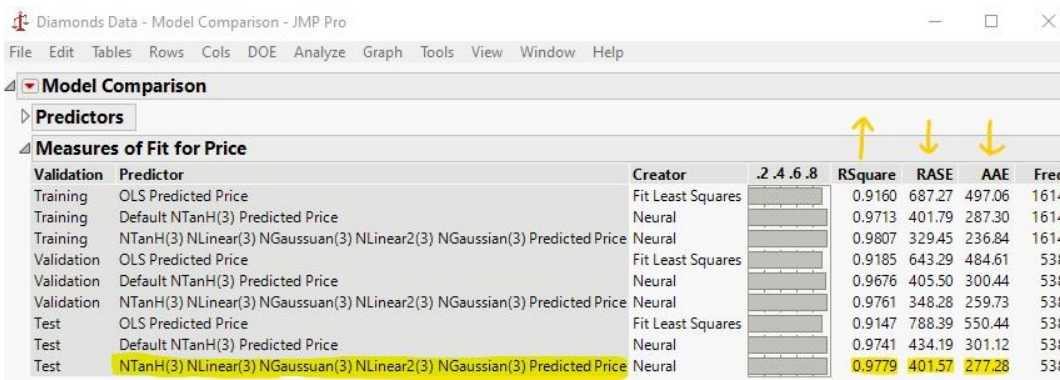


<sup>4</sup> <https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=networks-neural-model>

# Pricing Diamonds

(which is similar to the linear regression model and the linear combination of predictor variables is not transformed) and Gaussian (a bell-shaped function, which is similar to the normal distribution density function) activation function types. Both neural network models will use the Squared Penalty Squared method and 20 for the number of tours, as mentioned above.

For the data set, a Validation column will be created to allow cross-validation to occur within the modeling training phase at the same time estimation of the parameters of variables occurs. The



The screenshot shows the JMP Pro Model Comparison window. The 'Predictors' section is expanded, showing 'Measures of Fit for Price'. The table below lists various models and their performance metrics. The last row, 'NTanH(3) NLinear(3) NGaussian(3) NLinear2(3) NGaussian(3) Predicted Price', is highlighted in yellow. Three yellow arrows point to the 'RSquare', 'RASE', and 'AAE' columns for this model.

Validation	Predictor	Creator	.2	.4	.6	.8	RSquare	RASE	AAE	Freq
Training	OLS Predicted Price	Fit Least Squares					0.9160	687.27	497.06	1614
Training	Default NTanH(3) Predicted Price	Neural					0.9713	401.79	287.30	1614
Training	NTanH(3) NLinear(3) NGaussian(3) NLinear2(3) NGaussian(3) Predicted Price	Neural					0.9807	329.45	236.84	1614
Validation	OLS Predicted Price	Fit Least Squares					0.9185	643.29	484.61	538
Validation	Default NTanH(3) Predicted Price	Neural					0.9676	405.50	300.44	538
Validation	NTanH(3) NLinear(3) NGaussian(3) NLinear2(3) NGaussian(3) Predicted Price	Neural					0.9761	348.28	259.73	538
Test	OLS Predicted Price	Fit Least Squares					0.9147	788.39	550.44	538
Test	Default NTanH(3) Predicted Price	Neural					0.9741	434.19	301.12	538
Test	NTanH(3) NLinear(3) NGaussian(3) NLinear2(3) NGaussian(3) Predicted Price	Neural					0.9779	401.57	277.28	538

training phase will contain 1,614 observations or 60%. The Validation data subset which works to determine when to stop when it has found that the model is no longer improving, has 538

observations or rows. Lastly, the Testing data subset is used to determine how good the final model is once the Training phase is complete.

## The Best-Chosen Model

In comparing the three models (the Ordinary or Least Squares Method Model, the Default Neural Network Model, and the Most Complex Neural Network Model), the last model clearly outperforms the other two. The Ordinary or Least Squares Model had the highest  $R^2$  value, and both the Root Average<sup>2</sup> Error (RASE) and the Mean/Average Absolute Error (RMSE) were substantially higher than both the Neural Network models. The two Neural Network models were comparable, but the Most Complex Neural Network model stood out with a 0.38 higher  $R^2$  value, a 32.63 lower RASE value and a 23.84 lower RMSE value. It took longer to computer those 20 tours of data modeling, but the time was worth it.

# Pricing Diamonds

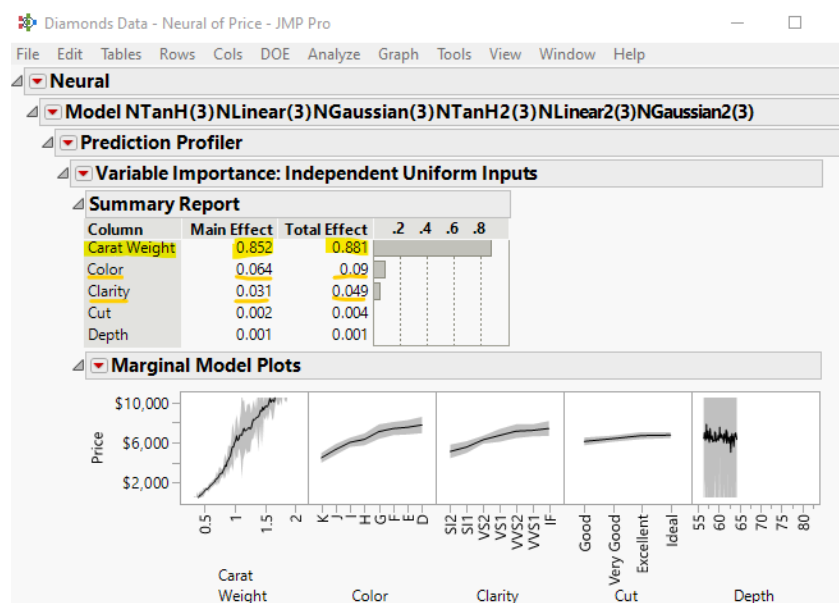
With the most complex neural network model with both a first and second layer with three

nodes in the TanH, Linear and Gaussian activation function types, the data can then be best analyzed using JMP's Prediction Profiler Tools. From this we can determine that the Carat Weight has a direct relationship with the Price of the Diamond; when one goes up, the other does too. The cut with its cut grades of K, J, I, H, G, F, E & D, has a distance 2<sup>nd</sup> place in its effect on Diamond prices, but it too is a direct relationship. Clarity with its S12, S11, VS2, VS1, VVS2, VVS1, & IF Clarity grades, also has a low direct

relationship with Diamond Prices. Depth has a value range of 56-80 but has an inverse relationship with Diamond Prices; when Depth goes higher, the price goes lower. Cut has a scale of four categories and it has a mixed effect on Diamond price with "Very Good" having a negative effect but "Good", "Excellent", and "Ideal" having a positive effect.

## Interpretation

Next, the order of the importance of variables is examined using JMP's Variable Importance: Independent Uniform Imports tool from its Prediction Profiler which clearly sets Carat Weight as the largest factor of Diamond Price. Since there wasn't much difference between the Main Effect and the Total Effect, so there isn't much variable interaction but it has a whopping 88% of the effect on Diamond Price, all on its own.



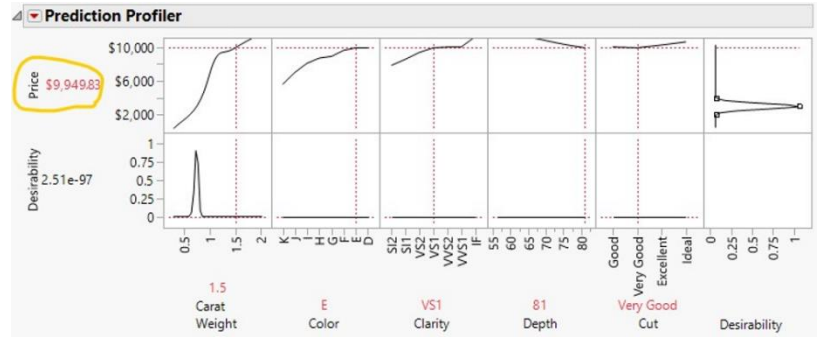


# Pricing Diamonds

Color and Clarity both show some variable interactions, but if you look at the Main Effect on Price, Color has just 6.4% and Clarity has just 3.1%. As expected the Cut Grade is the least important of the 4 C's and Depth just barely registers an effect on Diamond Price.

The model was then used to formulate a prediction of Diamond Price with a Carat

Weight of 1.5, a Color Grade of E, a Clarity Grade of VS1, a Depth Measurement of 81 and a Cut grade of Very Good. Using JMP's Prediction Profiler, the data was manipulated to reflect the test case and the estimated predicted price of such a diamond was found to be \$9,949.83.



The prices in this undated data set must be accepted to be a sample of the goods in the snapshot of time from which they were taken from. The diamond might never change but the demand and the technology making diamonds will undoubtedly change. For instance, a family heirloom of the Hagen family is a 5 stone, princess cut diamond ring engraved with "To my daughter December 25, 1865." For insurance purposes, it was appraised by a professional jeweler whose results basically said that it was priceless as an antique. However, the value had to be assigned to the item as if it had to be reconstructed as a replacement, which is tens of thousands less than it would actually cost to recreate. The cut of the diamond is hand-cut, and technology today cannot reproduce the cut, unless you have someone from a museum like the Smithsonian do it as a piece of art. So, the data set reflects the demand and the technology in its own snapshot of time just like the family heirloom represents the added personal value.