Daniel Sanchez

BAN 525

Dr. Cetin Ciner

Module 5: Assignment 1 - Pricing Diamonds

June 13th, 2021

Introduction

Marilyn Monroe famously quipped, “Diamonds are a girl’s best friend.” James Bond reminds us that “Diamonds are forever.” Pop singer Rihanna encourages listeners to “shine bright like a diamond.” Diamonds have established their prominence in modern society as a premium luxury. But what exactly makes a diamond so desirable that it justifies an often-exorbitant price? Advertisements routinely highlight the carat weight of featured diamonds. How significant is that one feature, though, and do any other features contribute to the price of a diamond?

In order to better understand the determinants of diamond pricing, data analysis may provide some key insights. The data set in this analysis comes from Adiamor, a major online diamond retailer. Variables in this data set include Price, Carat Weight, Clarity, Color, Depth, Cut, Table, and Report. This array of variables is a mixture of nominal and continuous features, representing observations from 2,690 individual diamonds. We will select Price as the response variable, with all other variables serving as candidate predictors. Table and Report will be excluded.

The analysis in this report has been performed in JMP Pro 16, using the following methods:

1.) Ordinary Least Squares (OLS)

2.) Neural Net, Simple (using one layer, three nodes with a TanH activation function)

3.) Neural Net, Complex (using two layers, three nodes with three activation functions)

Analysis and Model Comparison

Ordinary Least Squares(OLS): This method will be used as a benchmark for the analysis of the continuous variable, Price. OLS fits a model by minimizing the squared differences between observed and predicted values. Since all of the model effects are included in the final model, overfitting can become an issue.

Neural Net - One-Layer, Three Nodes, One Activation Function (TanH): This method is helpful for modeling very complex non-linear relationships. Intended to mimic how the human brain processes information, this method sends inputs through nodes, where a transformation function is applied. In this case, the zero-centered, S-shaped hyperbolic tangent(TanH) function allows for a non-linear relationship with outliers and approximates a linear regression for mid-range values.

Neural Net - Two-Layers, Three Nodes, Three Activation Functions (TanH, Linear, Gaussian): This more complex model accounts for TanH, in addition to Linear and Gaussian functions. The Linear regression function will be applied to variables without being transformed. The Gaussian function will apply a bell-shaped, normal distribution to the predictor variables.

Since the Diamonds data set in this analysis is cross-sectional (as opposed to being time-series), the cross-validation method for this analysis involves randomization. The data set has been split into three parts. The *Make Validation Column* facility in JMP Pro 16 was selected, applying a random seed of 123. Implementing a 60-20-20 split of the 2690 total observations, 1614 rows were used to train the models, 538 rows were tied to validation, and 538 rows were set aside for testing.

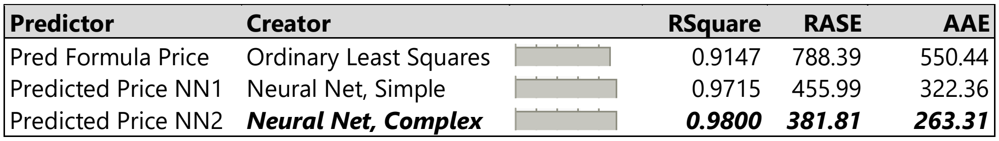
Although this data set is not time-series in nature, it is worth noting that overall diamond pricing may be subject to inflationary pressures and other market forces of their time. A predictive model of cross-sectional diamond pricing would likely need to be monitored, re-evaluated, and updated as new data emerges.

Aiming to determine the predictors of diamond pricing, the Price variable was selected in JMP Pro 16 as the Y variable. All of the other variables in the data set (apart from Table and

Report) were included as model effects. The Validation Column was implemented, and the analysis was run with all three methods. As predictions from each model were generated, they were saved and stored in new columns in the data set. The figure below shows the comparison of how the three models performed with the test set.

Model Comparison

Measures of Fit for Price - Test Data



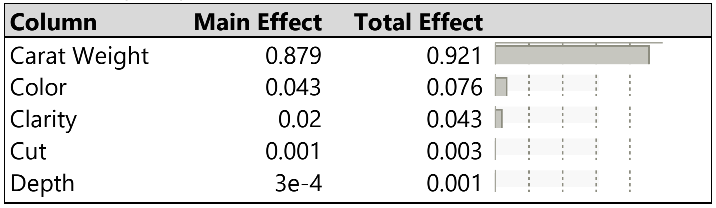
Predictor	Creator	RSquare	RASE	AAE
Pred Formula Price	Ordinary Least Squares	0.9147	788.39	550.44
Predicted Price NN1	Neural Net, Simple	0.9715	455.99	322.36
Predicted Price NN2	***Neural Net, Complex***	***0.9800***	***381.81***	***263.31***

When the predictions of each model were applied to the new and unbiased test data, the Complex Neural Net model outperformed the others. Its RSquare value of 0.9800 was quite a bit higher than OLS (0.9147), and marginally higher than the Simple Neural Net model (0.9715). Compared to the other models, Complex Neural Net also exhibited the lowest RASE and AAE values (381.81 and 263.31, respectively). This Complex Neural Net model will be more closely examined in the next section of the analysis.

Interpretation

Variable importance analysis of the Complex Neural Net model shows that Carat Weight is by-far the most predictive variable of Price, with a total effect of 92.1%. This is very similar to its main effect (87.9%). The only other features that appear to have any bearing on Price are Color and Clarity. However, their total effects are small (7.6% for Color, 4.3% for Clarity). The total effects of Cut and Depth are negligible (less than 0.4%).

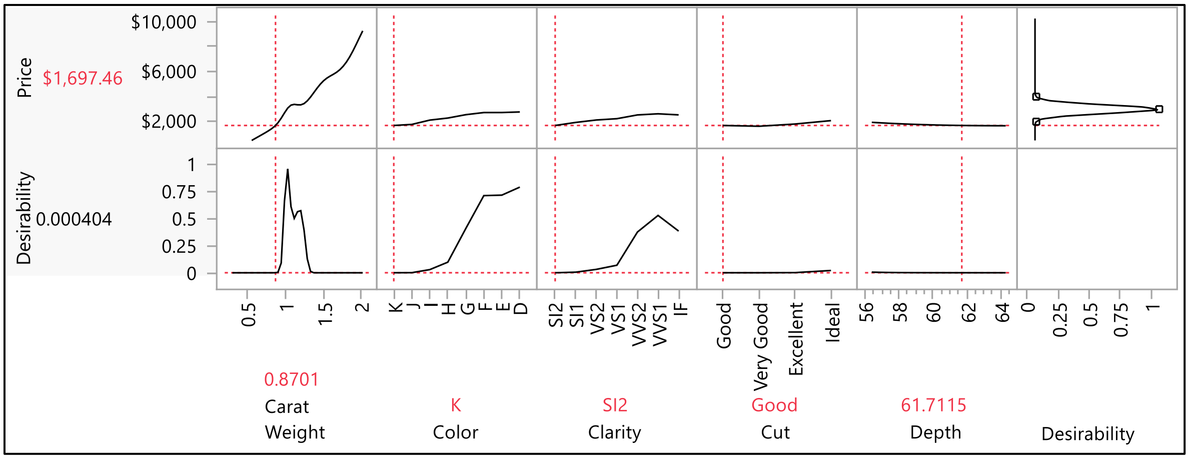
Variable Importance: Independent Uniform Inputs - Complex Neural Net
Summary Report



Column	Main Effect	Total Effect
Carat Weight	0.879	0.921
Color	0.043	0.076
Clarity	0.02	0.043
Cut	0.001	0.003
Depth	3e-4	0.001

The Prediction Profiler displayed below helps to visualize the relationship of the candidate variables to Price and each other. The steep slope of Carat Weight illustrates its effect on Price. The modest inclines of Color and Clarity show a more subtle positive correlation with the response variable. Manipulating values in the Profiler also reveals that Carat Weight, Color and Clarity have an impact on the other variables in the data set. The relatively flat lines of the remaining features (Cut and Depth) show their minimal effects on Price. Adjusting these unimportant variables does not appear to have much influence on any of the other variables, either.

Prediction Profiler - Complex Neural Net



This analysis confirms that carat weight alone is highly predictive of diamond pricing. The lesser features of color and clarity are only marginally influential. The cut and depth of a diamond do not predict its price.

There may be features outside of this data set that contribute to a diamond's price. For example, in a secondary market, the history or lineage of the diamond's ownership could weigh heavily on its desirability. The analysis in this report, however, is useful for the pricing of generalized sets of retail diamonds. While inflationary pressures and market demand may also impact overall prices, it is likely that carat weight will remain the leading predictive feature.