Alec Engel
June 13, 2021

Introduction

The question at hand here is to determine what if any diamond variables play a key role in predicting the price of a diamond. We want to run several cross-validation models to see what variables can be strongly correlated with our price response variable. These cross-validation methods will allow us to split our dataset into training, validation and test portions. The split allows for better predictions of future outcomes with reduction in potential overfitting and removing unnecessary model noise amongst highly correlated independent variables. We will next take a look at our variables using a benchmark model of ordinary least squares and then two neural network models.

Predictor variables consist of diamond factors that may cause increase or decrease of its monetary value. The most important variables in this data set are the response variable of price; along with predictor variables of carat weight, color, clarity, cut and depth. As we will elaborate later, the carat weight will have highly unique information as it pertains to the correlation with diamond price value.

In order to narrow down our predictor variables, we ran several statistical models and judged our outcomes. Statistical methods include cross-validation tactics of ordinary least squares (OLS) as well as default and more complex neural networks (NN). The latter method of neural networks produces extremely powerful and telling models that create a solid narrative when dealing with nonlinear data. Neural networks use hidden layer nodes to help model very complex variable relations between a model's inputs and outputs. Transformation is then applied and used to provide stronger model predictions. With neural networks, it is also important that we include a validation portion of our data set as one drawback with neural networks because they can overfit if this validation is not present.

Analysis and Model Comparison

For the diamond price response variable, we first used OLS which involves the inclusion of all variables. While using all variables, there are often noisy and unnecessary data variables that are accounted for. Our neural network models performed much better while taking several different transformation approaches. As a default, we first used one layer of transformation

with TangentH and then used a more complex version of neural networks using two layers with additional transformations of Linear and Gaussian. TangentH transformation function uses a similar “S” shape as a logarithmic function while Linear is similar to a linear regression model and Gaussian transformation function is similar to a normal (bell shaped) distribution. To get the clearest picture and better predictive abilities, we take the most stock in our complex version of neural network that includes two layers and three nodes for all three transformation functions. Although extremely complex, this advanced model is our best bet for accurately predicting the price value of a diamond.

We have created a cross-validation column which splits our data into training (60%), validation (20%) and testing (20%) sets to allow our model to gain predictive attributes and score our R squared values and error rates for how we believe future data will look. As previously mentioned, this is an important step and validation must be included in neural network models to tell the model when to stop running and prevent overfitting. After cross-validation, we created predictor formulas with each of the above three methods described. Each method was then added as a dataset column so that we could easily perform model comparisons.

Model comparison results can be seen below. First with the training sets, we see all three models with high performance. All three are above 0.91 R squared with relatively low route average standard error rates. These however are our training sub sets of data and we want to focus more towards our validation and testing metrics. Moving to validation and test notations, we see consistency with our training data but the winner with a slight edge in R squared value is our more complex model of neural networks that explains nearly 98% on our unbiased test data (R squared of 0.978). Although I will dig into each model to investigate strongest profile variables, I want to choose the strongest complex neural network model here as my predominant model.

Model Comparison

Predictors

Measures of Fit for Price

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Training	Pred Formula Price OLS	Fit Least Squares		0.9160	687.27	497.06	1614
Training	Predicted Price NN Default	Neural		0.9713	401.79	287.30	1614
Training	Predicted Price NN Complex	Neural		0.9829	310.18	221.44	1614
Validation	Pred Formula Price OLS	Fit Least Squares		0.9185	643.29	484.61	538
Validation	Predicted Price NN Default	Neural		0.9676	405.50	300.44	538
Validation	Predicted Price NN Complex	Neural		0.9764	346.06	250.53	538
Test	Pred Formula Price OLS	Fit Least Squares		0.9147	788.39	550.44	538
Test	Predicted Price NN Default	Neural		0.9741	434.19	301.12	538
Test	Predicted Price NN Complex	Neural		0.9779	401.03	275.87	538

Interpretation

Per below, we can see that three prediction profiler variables were leaned on to maximize our overall model performance on test data. These top three parameter estimates in our complex neural network model are carat weight, color and diamond clarity. We can see that a diamond's cut and its depth don't play much if any role in determining a diamond's price. However, there is obviously heavy weight on a diamond's worth solely due to its carat weight. Per our model, a diamond's carat weight will have a total effect (including non-linear variable interactions) of over 88% of its price. So, when determining a diamond's worth, the first thing that needs to be deciphered is its weight. This makes a lot of sense as most jewelry stores measure prices of diamonds on weight scales and it is an easy measurement to account for.

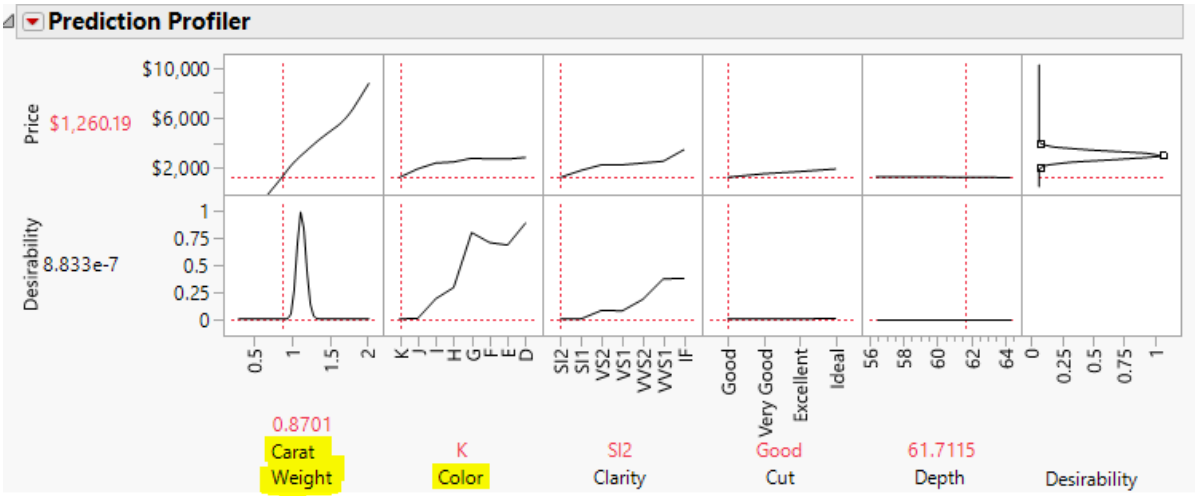
Variable Importance: Independent Uniform Inpu

Summary Report

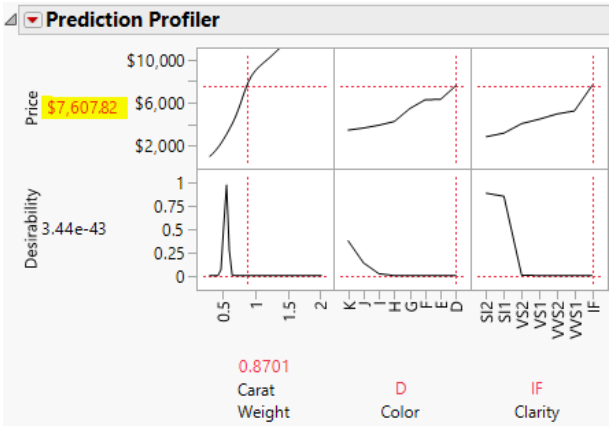
Column	Main Effect	Total Effect	.2 .4 .6 .8
Carat Weight	0.847	0.882	
Color	0.057	0.089	
Clarity	0.036	0.061	
Cut	0.002	0.005	
Depth	4e-4	0.001	

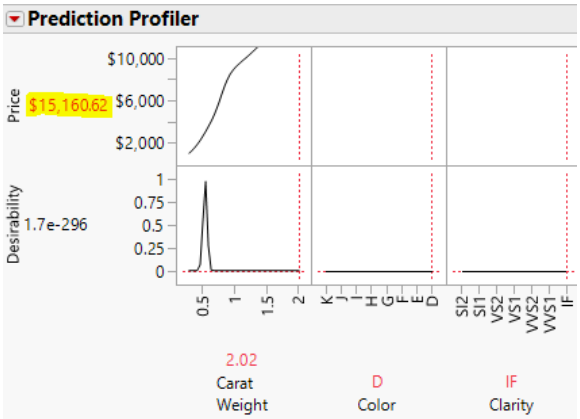
Carat weight and color will be our most important parameter estimates as they show both high estimate numbers as well as low standard errors. For carat weight, we see that as weight of the diamond increases, the diamond's price looks to simultaneously rise in nearly a linear 45-degree angle. There looks to be no leveling off here, so as stated previously, the

bigger the diamond the more valuable it will be. A similar notion to a lesser extent can be used with a diamond's color in relationship to its value. A diamond which is colorless will represent "D" while the higher alphabetically you go increases color (up to "K" in this scenario). So, we can decipher here that as a diamond becomes more colorless, its value increases as in the color of "D" being most valuable and the color "K" holding very little value. Both of these positive relationships with diamond prices help us to accurately predict its worth.



When predicting with our profilers and performing some manipulation, we can back up our theory of important variables being carat weight, color and clarity. I have highlighted our most valuable color of "D" and our most valuable clarity of "IF" with our average carat weight in this category of 0.87 to get a price of $7,607. But, if we increase carat weight to 2, our price doubles to $15,160.





Carat weight is clearly our strongest contributor and variable for predicting diamond prices. Each variable showed somewhat positive relationships and increased ability to accurately predict positive prices of diamonds. Some consideration can be given to a diamond's color and clarity with slightly over 9% predictability of its price. But again, carat weight is our strongest predictor variable as its relationships explains over 88% of variations in the price of a diamond.