Christopher Dillard 10/1/2023

Neural Networks to Predict Diamond Prices

I will be using two Neural Network models to identify the determinants of diamond prices. My candidate determinants are as follows: carat size, clarity, color, depth, and cut of stone. The dataset used in this assignment comes from the online retailer, Adiamor.

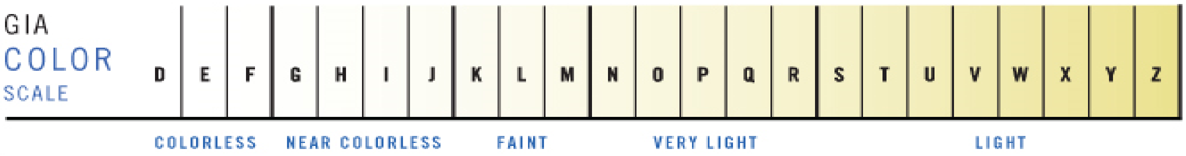
Adiamor references the 4 C's of diamonds: Carat size, Clarity, Color, and Cut, which were my predictor variables along with the depth of the gemstones, a continuous variable measuring height from the cutlet to the table.

Carat Size is a continuous variable which measures the weight of the gemstone, where each carat is equivalent to 200 milligrams (GIA, 2023).

Despite Carat Size being such an important variable, two diamonds with the same weight can have vastly different appearances depending on their clarity, color, and cut. The factor of rarity plays a big role in the valuation of diamonds, considering how roughly 1 million rough diamonds must be mined until one can be cut into a 1.00 carat diamond (Adiamor, 2023).

Clarity is a categorical variable with 7 levels in the dataset. Inclusions and blemishes are the two main types of characteristics that affect the clarity of a diamond. While blemishes are surface level defects, inclusions are inside of the gem. The problem with these marks is in how they affect how light disperses within the gem, otherwise messing with it's brilliance.

Color is another categorical variable with 8 levels, alphabetically ranging from D-K in the dataset where D is the highest color grade. Adiamor bases their colors on the GIA scale. According to their educational glossary on their website, D is considered to be the closest to a true "colorless" diamond, while each letter above D has slightly more color to it due to increasing levels of impurities present in the gemstone, until reaching Z.



Infographic retrieved from Adiamor (https://www.adiamor.com/Education/Diamond-Color)

Cut is a categorical variable with 5 levels: F (Fair), G (Good), VG (Very Good), EX (Excellent), and AF (Affinity Ideal). Diamonds achieve their sparkly brilliance by cutting the gem in such a way that light enters the gem and creates the most dispersion, thus resulting in a shimmery shine. Two factors contribute to the grading of a diamond's cut: Polish and Symmetry. Polish refers to the diamond's surface-level defects, including nicks, polish lines, and abrasions. Symmetry, on the other hand, references the diamond's shape and how well it's facets align. So, in accordance with the GIA grading scale, both Polish and Symmetry are considered when classifying a diamond's cut.

My baseline model in this assignment was the Ordinary Least Squares (OLS). OLS models provide the benefit of being relatively easy to interpret yet also have the downside in how they assume linear relationships between predictor and response variables, despite there being much more complexity in real-world data. The inability to capture nonlinear relationships set these models behind more robust tools like Neural Networks which I explore in this assignment.

Neural Networks (NNs) derive their name from neurons, given that they were designed to mimic the way that humans learn (Han, Su-Hyun, et al., 2018). Just like humans learn through pattern recognition from random events, so do these model in the way new data is introduced into the model across its layers and nodes. Neural Networks are an older statistical modelling method than other robust tools that I had previously explored, such as Decision Tree models, Random Forest, and Elastic Net and Lasso for penalized regression. So, this assignment will explore how powerful Neural Networks are and why they are still popular despite their age.

I will be setting the random seed for my NNs to 123 for the sake of controlling replicability. That being said, the inherent random nature of NNs makes it so that even if the initial weights are the same, the model will continue to randomly split the data into different layers and nodes which make their result vary every time the model is run.

Neural Networks are an older statistical modelling method than other robust tools that I had previously explored, such as Decision Tree models, Random Forest, and Elastic Net and Lasso for penalized regression. So, this assignment will explore how powerful Neural Networks are and why they are still popular despite their age.

Since NNs are less replicable than previous statistical models, I will be analyzing the results with small margins assumed in the performance on the testing set, since the initial weights are chosen at random. For the same reason, my interpretation of variable importance in the model will consider the potential for slight differences. NN are prone to overfitting, especially under two conditions where there is not a large enough dataset or when the model is too complex. Overly complex models with multiple layers and activation functions can lead to overfitting. EXPLAIN HOW

The cross-validation column has a 60-20-20 split for training, validation, and testing splits. This information is stored in the “Validation” column, which will come into play when running the model comparisons to test for their accuracy on the unbiased, testing split. For replicability purposes, the random seed was set to 123.

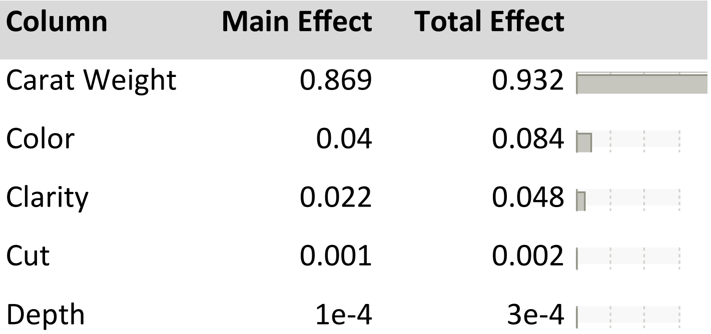
Model Comparison Table:

Validation	Predictor	Creator	RSquare	RASE	AAE	Freq
Training	Pred Formula Price OLS	Fit Least Squares	0.9160	687.27	497.06	1614
Training	Predicted Price NN 1	Neural	0.9666	433.44	311.58	1614
Training	Predicted Price NN2 Complex	Neural	0.9781	350.89	250.83	1614
Validation	Pred Formula Price OLS	Fit Least Squares	0.9185	643.29	484.61	538
Validation	Predicted Price NN 1	Neural	0.9608	446.19	328.04	538
Validation	Predicted Price NN2 Complex	Neural	0.9705	387.11	281.36	538
Test	Pred Formula Price OLS	Fit Least Squares	0.9147	788.39	550.44	538
Test	Predicted Price NN 1	Neural	0.9692	473.33	331.04	538
Test	Predicted Price NN2 Complex	Neural	0.9684	480.02	324.01	538

As seen in the table, the best model was the first Neural Network, which was not the most complex one. Although the model performed only slightly better, it shows that complexity in model does not necessarily lead to more accurate prediction models.

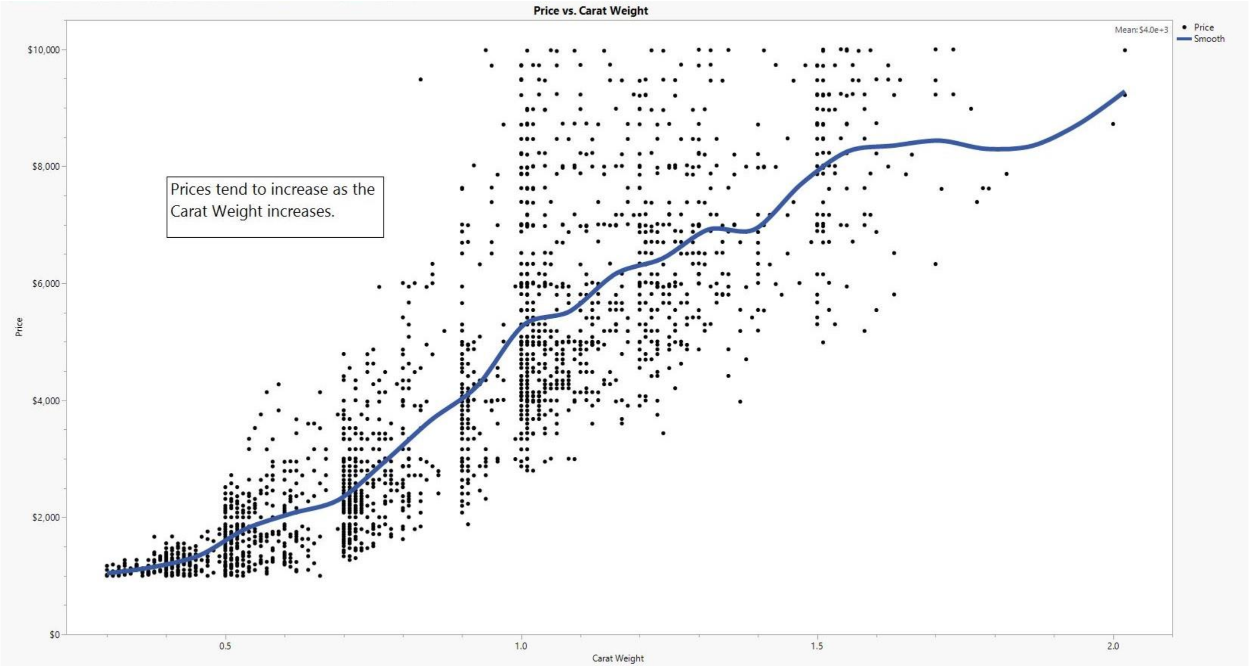
As discussed previously, overfitting is an unfortunate problem that can be associated with Neural Networks, which is likely the primary factor as to why the more complex NN model performed worse on the Testing split, despite having the best R Square value on the Validation set.

So, in order to fully see the scope of the model's performance and extract some key takeaways, I reopened the model's script to assess its performance on the testing set. The extract model shows the variables terms of their importance, based on their independent uniform inputs (subject to slightly vary even with the same random seed 123):



Column	Main Effect	Total Effect
Carat Weight	0.869	0.932
Color	0.04	0.084
Clarity	0.022	0.048
Cut	0.001	0.002
Depth	1e-4	3e-4

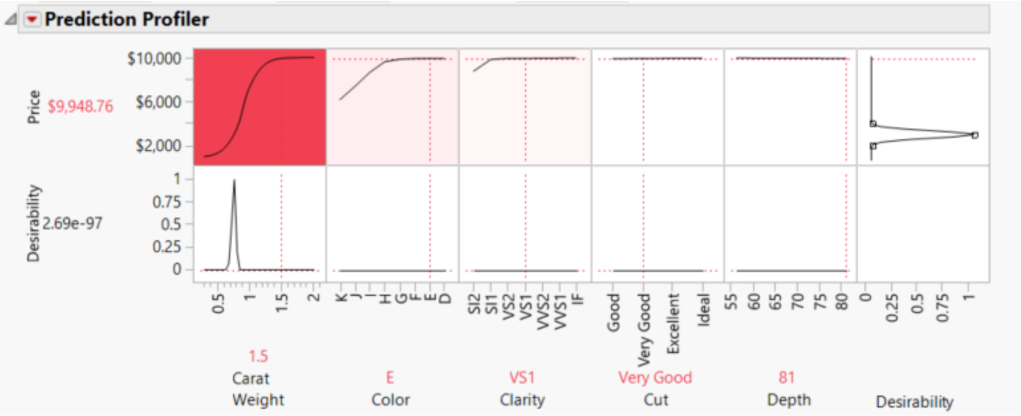
As shown, Carat Weight had a roughly 93% explanative power in this model in predicting diamond prices. Now, looking back at the dataset, I wanted to visualize the relationship between price and carat weight in a scatterplot with a smooth line with a lambda value of 0.05:



There appears to be a strong, positive relationship between a diamond's carat weight and it's price. That being said, the relationship is also not perfectly linear and there is a wide range of dispersion around the smooth line, particularly for diamonds with carat weights between 1.0 and 1.5. In these case, we could assume that secondary characteristics that the model deemed to be significant like color and clarity can account for the differences.

The model created in this assignment can help us determine whether a diamond is reasonably priced on Adiamor's site. Using the Prediction Profiler tool from the best Neural Network model, I will determine whether a diamond with the following characteristics could reasonably be priced at $9,000:

Carat: 1.5, Color: E, Clarity: VS1, Depth: 81, Cut: Very Good.



Price $9,948.76

Desirability 2.69e-97

Carat Weight 1.5

Color E

Clarity VS1

Cut Very Good

Depth 81

If I were to find a diamond on Adiamor's site with the specified characteristics for $9000, I could be convinced to make the purchase given that my model predicted that such a diamond should cost $9,948.76. My model explains roughly 97% of the variance in diamond prices in the testing set, so I can be very confident that the predictions are reliable, especially with such a large margin between the sales price and what the model predicted. That being said, I could also look into other diamonds with slightly better characteristics and rerun the model. So, I would idealy keep the same carat size, since any increase in size would presumably affect the cost much more than improvements in other characteristics like color, clarity, cut, and depth.

References:

“Adiamor - Diamond Education Glossary.” *Adiamor Diamonds and Fine Jewelry*, www.adiamor.com/Glossary/Depth. Accessed 28 Sept. 2023.

Han, Su-Hyun, et al. “Artificial Neural Network: Understanding the Basic Concepts without Mathematics.” *Dementia and Neurocognitive Disorders*, U.S. National Library of Medicine, Sept. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6428006/.

“Learn What Carat Means and What Diamond Carat Measures: 4Cs of Diamond Quality by Gia.” *GIA 4Cs*, 9 Dec. 2019, 4cs.gia.edu/en-us/diamond-carat-weight/.