Zackery Bradley

BAN 525

20220220

Model 5 Assignment 1

On behalf of this week’s assignment, we have been tasked with examining the pricing of diamonds. Traditionally, diamonds are normally priced by utilizing the “4 C’s” strategy, which includes the cut, clarity, color, and carat weight of the diamond. In our dataset, we will add an additional criterion by also examining the diamonds depth. A diamonds cut is measured of how well-proportioned the dimension of the diamond is, and how these positions create the sparkle and brilliance of the cut. Color is widely considered a significantly important characteristic of a diamond or lack thereof. Diamonds that have less color are considered a higher grade and are thus more expensive. The clarity of a diamond is measured by the purity of rarity of the stone. The clarity is measured by a graded scale that ranges from imperfect (|1, |2, |3) to flawless (FL). The weight of a diamond is standard, as the bigger the diamond, the more expensive the stone. The weight of a diamond is typically what drives the price point. Lastly, the depth of a diamond is a measurement of how well the stone refracts light. Essentially, the depth of a diamond is how well the diamond sparkles. Diamond depth can be somewhat tricky to evaluate, as diamonds that are too shallow will have light simply pass through them while diamonds that are too deep will have poor sparkle and make the stone appear smaller. It is believed that by following conventional wisdom, we can make a hypothesis that the prices of the diamond will be largely affected by the overall weight of the diamond, followed by the color and clarity. Throughout our analysis, we will be using various methods to ensure we get the best model for the data in question. These methods will include our standard method, Ordinary Logistic Regression (OLS), in addition to using Neural Network (NN) and Complex Neural Network (CNN) methodology. OLS is typically used as a method for estimating the parameters that are unknown in the model. This method helps us to find the relationships between independent and dependent variables. NN models are considered very flexible algorithms that are able to model extraordinarily complex relations between model inputs in outputs. Each NN consists of an unput layer, in addition to one or more hidden layers followed by an output layer. The hidden layers are made up of one or more nodes, in which a linear combination of the input variables is transformed. In JMP Pro, there are three types of functions that can be utilized in a NN model consisting of TanH, Linear, and Gaussian. TanH, a hyperbolic tangent function, is similar to the shape of a logistic function. Linear is similar to a linear regression model. Gaussian is a bell-shaped function, which is similar to the normal distribution density function.

Analysis and Model Comparison

Regarding our predominant methods as previously stated, OLS can be described as the “workhorse” method and is a standard operating tool in statistics analysis. On the contrary, OLS tends to produce estimates with a great deal of variance regarding big data. In all, OLS typically generates poor forecasts. Our second method, NN also comes with its own list of advantages and shortcomings. As previously mentioned, NN typically has good predictive ability and is capable of capturing complex relationships. However, NN has no variable selection mechanism, so utilizing JMP’s variable importance is vital when identifying the most important variables in the data. Procedurally, we will be employing a cross-validation analysis. Cross-validation techniques are built on the concept of leaving out a part of the data out of the estimation process as a buffer. As the predictions of the model stop improving as the data holdout process occurs, model growth then stops, and estimates are obtained. This creates a ‘split’ in the data. These splits are then broken down into three parts in the terms of our model. Those parts include training, validation, and testing splits respectfully. Training data is the portion that is used to estimate our model. Validation data is not used in estimate directly, but instead works in the background to determine the optimal point in which the model stops. Lastly, the test data is never actually used in the model estimation. Test data is used to represent “new observations” and assists in providing and unbiased analysis of the predictive ability of the model. With the validation method only being utilized indirectly, we choose our model that displays the best performance on the test data. Upon examination of our data, we can see that while all methods were similar in analysis, one method stood out in comparison to the others. Our Complex Neural Network method (CNN) proved to be the best fit for our data, as CNN calculated an RSquare of 0.9777942 in addition to a RASE of 402.16345. Our simple NN method performed convincingly worse in calculation, measuring an RSquare of 0.9699057 followed by a RASE of 468.17819. When comparing these measurements from the visuals below, we can see that CNN is the appropriate choice of methodology as far as the performance of this particular dataset is concerned.

**Model NTanH(3)**

**Test**

**Price**

| **Measures** | **Value** |
| --- | --- |
| Rsquare | 0.9699057 |
| RASE | 468.17819 |
| Mean Abs Dev | 311.76717 |
| -LogLikelihood | 4071.4697 |
| SSE | 117924661 |
| Sum Freq | 538 |
|  |  |

**Model NtanH(3)NtanH2(3)Nlinear2(3)Ngaussian2(3)**

**Test**

**Price**

| **Measures** | **Value** |
| --- | --- |
| Rsquare | 0.9777942 |
| RASE | 402.16345 |
| Mean Abs Dev | 262.01699 |
| -LogLikelihood | 3989.6989 |
| SSE | 87013667 |
| Sum Freq | 538 |

After thoroughly sifting through our data, our chosen model displayed which variables were of most importance. Shown in the visual below, we can see that the variables responsible for the biggest impact of diamond pricing. Our initial hypothesis was proven to be correct, as carat weight had a substantial impact on the pricing of diamonds with a total effect of 0.91. Color also seemed to be somewhat significant, it measured to have a total effect of 0.84, Our third most important variable was clarity, as its total effect came out to 0.048.

**Predicted Price CNN**

| **Column** | **Main Effect** | **Total Effect** |  |
| --- | --- | --- | --- |
| Carat Weight | 0.869 | 0.91 |  |
| Color | 0.052 | 0.084 |  |
| Clarity | 0.025 | 0.048 |  |
| Cut | 0.001 | 0.003 |  |
| Depth | 5e-5 | 1e-4 |  |

We further explored out data by analyzing the prediction profilers below. We experimented by placing our most significant variable, carat weight, on the lowest scale and then again on the highest scale to examine the relationship between the other variables evaluated in our dataset. From our experiment, we can see that all of the variables measured seem to have a positive correlation in to each other, with some variables being detrimentally more impacted than others as earlier indicated. By examining our minimum and maximum parameters on the carat weight in regard to price, we can see that the impact is of great significance.

Prediction Profiler (Carat Weight at minimum)

A picture containing chart

Description automatically generated

Prediction Profiler (Carat Weight at maximum)

A picture containing chart

Description automatically generated

In closing of our analysis, we can say with certain confidence that Complex Neural Network (CNN) performs adequately as opposed a simple Neural Network (NN) in the discussion of this particular dataset. This can be seen on our overall analysis of the other performance methods CNN has proven to be superior statistically speaking. Given the competition in the jewelry industry, it is vital that companies have access to the best modeling analysis methods available in order to stay relevant. By seeing the significance of each variable associated with the pricing of diamonds, companies are able to make better economic decisions on each diamond purchase, potentially saving the companies millions in lost revenue.