**MIS-5560/ MIS-4560**

**Group 5**

**Prepared For: Professor Yazan Roumani**

**Prepared By:**

**Aishwarya Sangal**

**Deepa Palariya**

**Savannah Schwenk**

**Vasudha Gulati**

**Gemstone Price Prediction**

# Introduction

The prediction of the price of a Gemstone for a company is crucial in order to figure out the higher profitable stones and lower profitable stones so as to have a better profit share. Not only is this important for the company to distinguish between the most and least profitable Gemstones, it will also help the company with the marketing aspect and retain their customer base. Consequently, we are going to investigate the variables that are most important in correctly indicating the price that can help identify the most and least profitable Gemstones.

# Data Collection

The dataset “**Gemstone Price Prediction**” from Kaggle was used to explore this topic. Since our outcome variable is numeric, we will only use three models: Multiple Regression Model, CART Model and Neural Network. We will evaluate the model performance, determine which model performs better and will identify the most important variables. The dataset contains the prices and attributes defining the price of a Gemstone of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). This large dataset will help us come up with accuracy in terms of building and analyzing the models to predict price. We will not be using the “ID” variable since it is a unique identifier and holds no significance in our analysis. An explanation of all the variables is as follows:

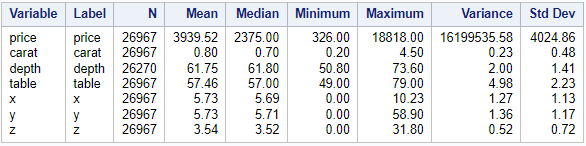
|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| CARAT | Carat weight of the cubic zirconia |
| CUT | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal |
| COLOR | Color of the cubic Zirconia with D being the best and J the worst |
| CLARITY | Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3 |
| DEPTH | The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter |
| TABLE | The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter |
| PRICE | The Price of the cubic zirconia |
| X | Length of the cubic zirconia in mm |
| Y | Width of the cubic zirconia in mm |
| Z | Height of the cubic zirconia in mm |

# Data Exploration and Visualization

Initially, we explored each of the 9 variables using a variety of graphs and statistical summaries to better understand the data. Our target variable is the price in the dataset which is a numeric outcome. We have to understand each variable in detail in order to understand its significance while predicting price.

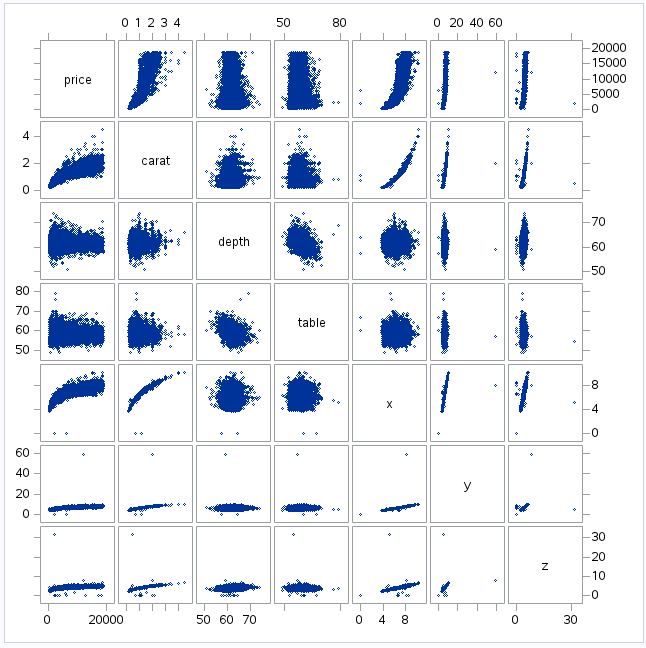
# Numerical Summary -

The below summary shows the statistics of all the numerical variables that are being considered to predict the price. This shows the distribution of these values using minimum, maximum, variance and standard deviation.

****

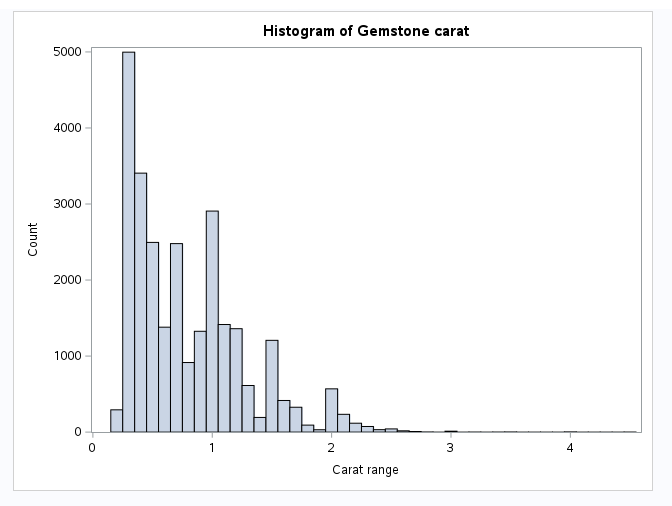
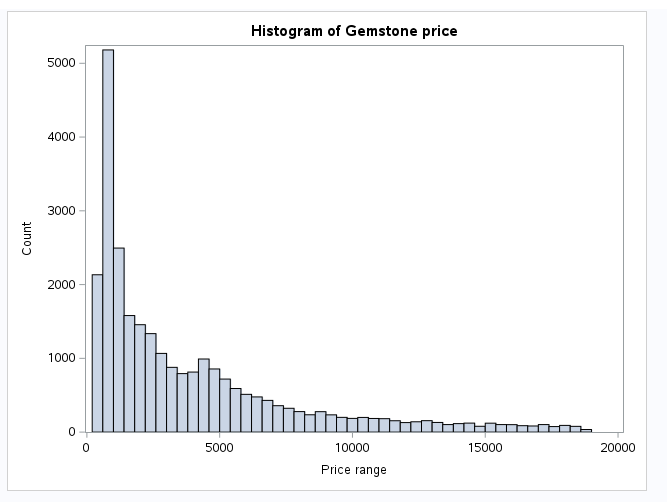
# Scatter Plot -

The below plot depicts that the price is having high positive correlation with carat and x (length of gemstone) variables i.e. when value of carat and x increases the value of price also increases which makes them important variables to consider.

****

# Histogram -

As we have seen above that price and carat are highly correlated, so we will look into the distribution of these variables. Price is right skewed, widely distributed and the mean values lie somewhere between 150 - 250. As for carat it is also right skewed, widely distributed and the mean values lie somewhere between 0.25 - 0.35.

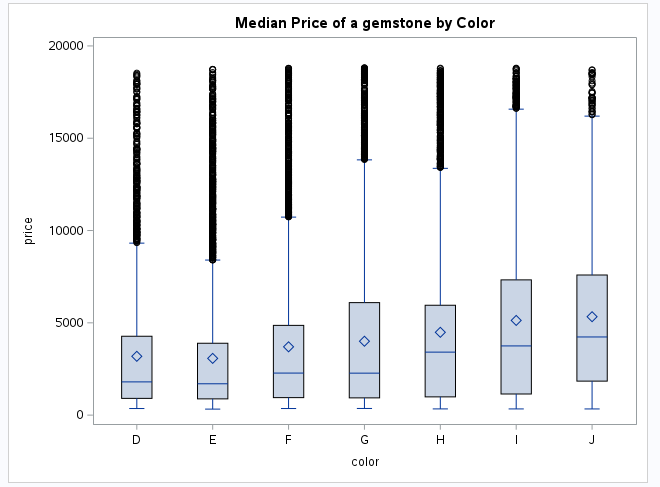
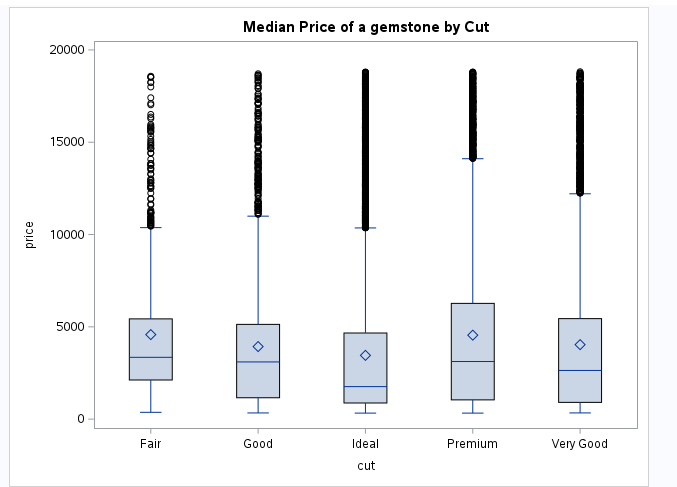
****

# Boxplot -

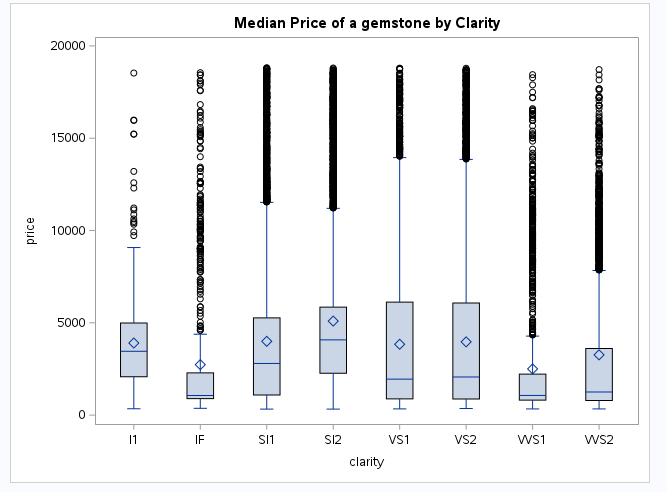
There are 3 categorical variables in the dataset whose distribution with respect to price can be seen as below. The 3 variables are cut, color, and clarity.

Cut is generally having mean more than median for all the values. The minimum of Fair is higher than any other cut. The interquartile range is maximum for Premium. There are many outliers for the cut but maximum is in Ideal and minimum for Fair.

Color has mean more than median for all the values. The minimum of all colors are similar but maximum value varies for all the colors.The interquartile range is maximum for I. There are many outliers for the color but maximum is in E and minimum for J.

****

Clarity has mean more than median for all the values except IF (Internally flawless) and VVS1 (Very very slightly included) where mean is more than median and interquartile range which shows that there are many values in lower range but some higher values of price pulling up the mean. The interquartile range is maximum for VS1 and VS2. There are many outliers for the clarity but maximum is in VVS1 and minimum for F1.

****

# Data Analysis

Our analysis will be based on three types of models: Multiple Linear Regression Model, CART Model, and Neural Network for our outcome variable “PRICE” which is a numeric variable. The predictors included all variables in the dataset. In all three models created, a 70/30 split was used on the data and a seed of 12345. We will be evaluating the model performance and determine which model performs better.

**Scenario 1**:

**Multiple Linear Regression**

A multiple regression model as the outcome variable is numerical. The data was split into 70% training and 30% validation, and forward, backward, and stepwise selection were used to ensure that the best performed model was concluded as the best multiple regression model.

Stepwise: Validation ASE is 1245219.

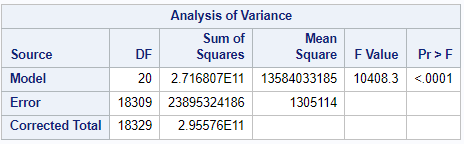
Forward: Validation ASE is 1245219.

Backward: Validation ASE is 1244600.

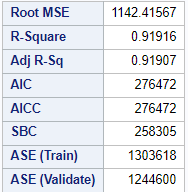
The multiple regression model using stepwise and forward selection performed exactly the same with all variables being left in the final model, and had an AUC of 276475. These two multiple regression models performed well but the multiple regression model using backward selection was the best model because it has lowest ASE values although it has also selected all 9 variables in the final model. It has an AUC of 276472.

**Evaluating the Model:**

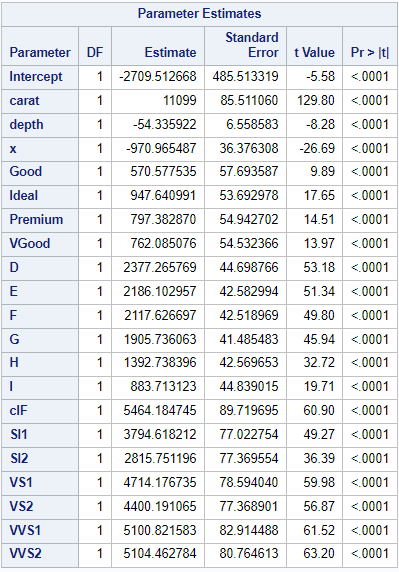
1. **F-test**: Clearly, p-value < 0.0001 which is less than alpha (0.05) which justifies that the model is statistically significant.

****

1. **Adjusted-R-square**: Adj R-Sq value is 0.91907 which means 92% of the variation in Y(PRICE) is explained by the independent variables. Adj R-Sq indicates a good Model Fit.

****

1. **Slope P-value**: The p-value for all the variables is less than alpha which indicates these variables are the significant predictors of gemstone price.

****

**Model Performance:**

Selected Model: **Backward**

Training ASE is **1303618**.

Validation ASE is **1244600**.

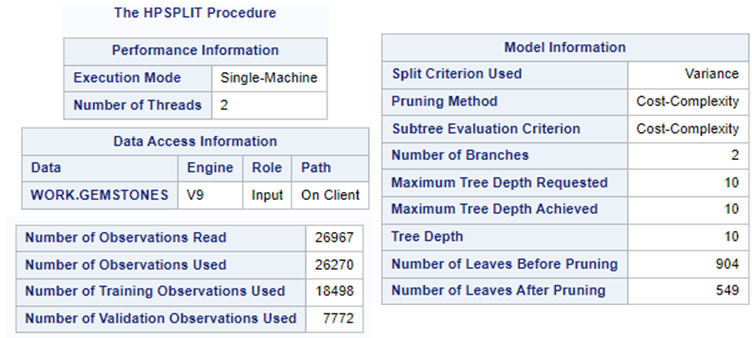
There are no signs of overfitting as the validation set did better than the training set.

Overall, the multiple regression model performed very well on the data.

**Scenario 2**:

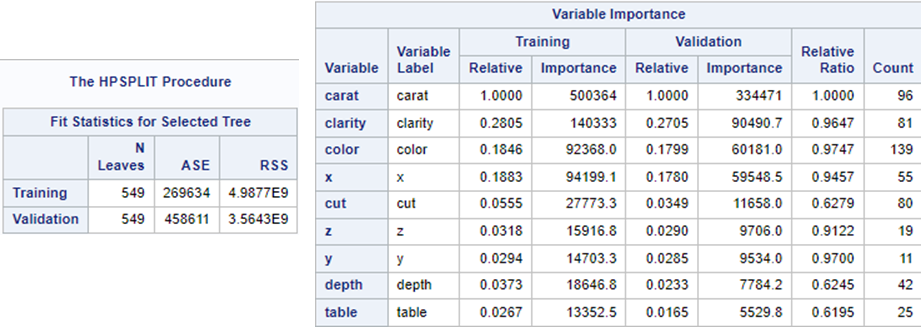
**CART - Classification and Regression Tree**

The CART Model was the next model to be tested. The model was built using RSS since our outcome was numeric. The data was partitioned into 70% training set and 30% validation set**.**

****

The above model information describes the number of observations used to build the CART model**.** Also, there were 904 leaves before pruning and 549 after pruning.

**Evaluating Model’s Performance**

****

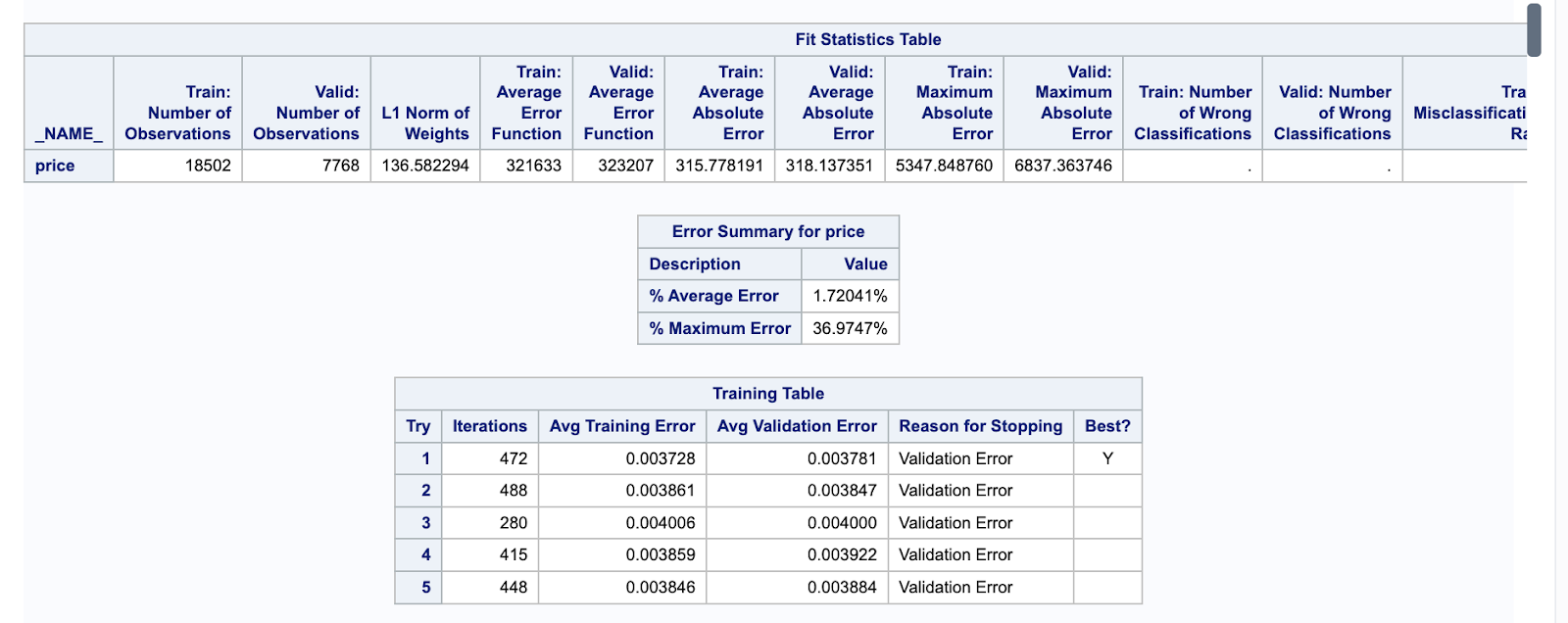
In general, the lower the value of ASE the better model fit and accuracy. According to the results above, the ASE of the validation set for CART is much higher than ASE of the training set of CART which indicates overfitting. Therefore, The CART model did not perform well.

Coming to the variable importance, Carat, Clarity and Color are the top three important variables based on the relative importance of the validation set.

**Scenario 3**:

**Neural Network**

**Model 1:**

For 10 hidden variables the average absolute error on the training and validation set are very similar. This shows that there is no overfitting with the data. The solution was found on the 1st try.****

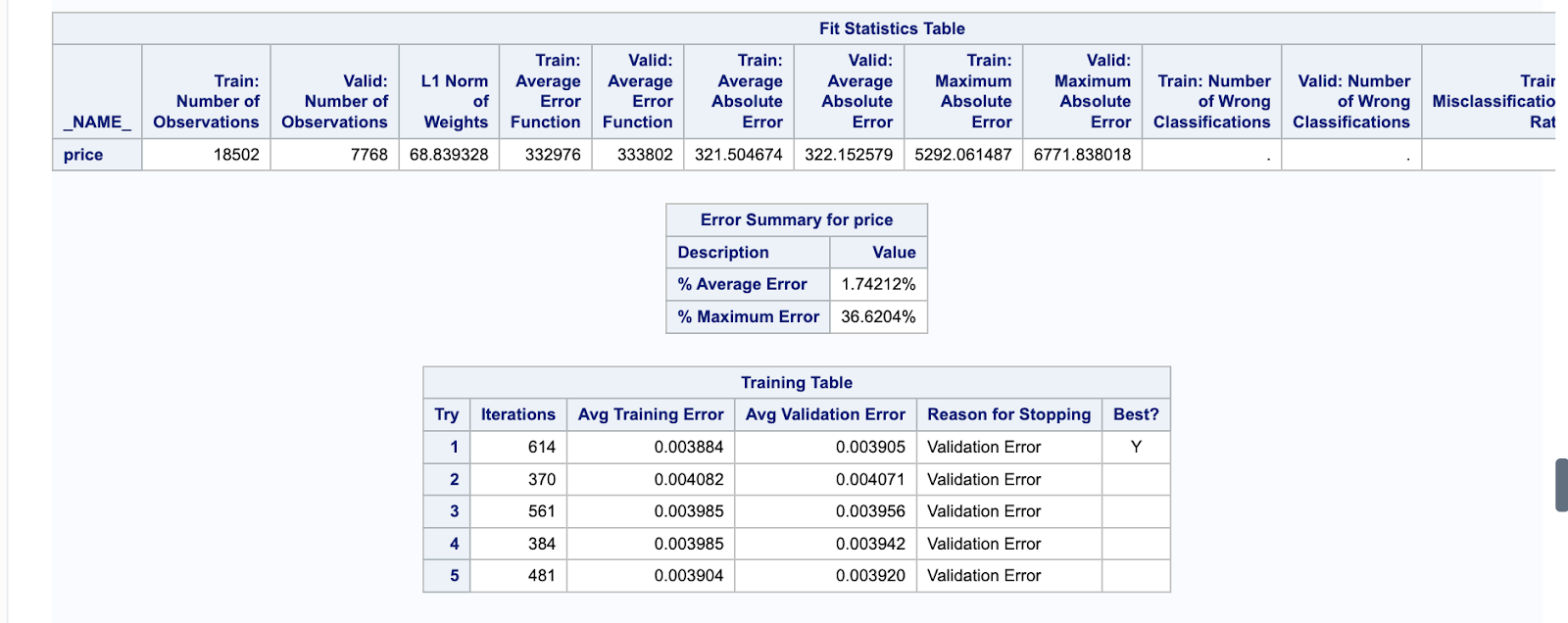
**Model2:**

Using 20 hidden variables, both the training and validation errors went down and the solution was found on the 5th try now.



**Model3:**

Using 5 hidden variables compared to 10 hidden variables, the errors went up on both sets but are still close to one another and there was a solution on the first try.



**Model4:**

Using 7 hidden variables compared to 7 hidden variables, the training set is now higher than the validation set, but the solution was still found on the first try.

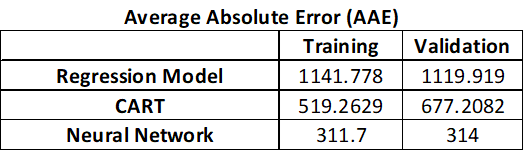


**Model Performance:**

Model 2 performed the best out of all the models we have tried. Using 20 hidden variables, both the training and validation errors went down and the solution was found on the 5th try.

# Summarizing Models

The 3 models are summarized below with their AAE which is calculated from square root of ASE for Regression and CART model. Results clearly show that the neural network performs the best with minimum error.



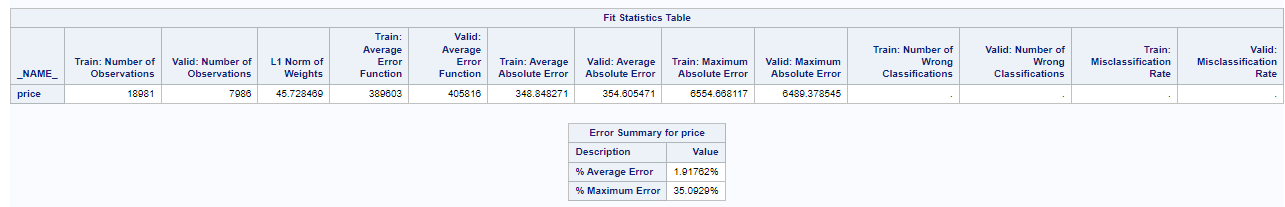
**Model Performance:**

As a result, we considered the CART model to choose the 3 most important variables and re-built the neural network with 3 important variables recognized from the CART model i.e. carat, color, and clarity.

**Rebuilt Neural Network:**

The important variables are identified from the CART model i.e. carat, color, clarity. The Neural Network is rebuilt to achieve a parsimonious model.

The final model is built on 3 important variables and 2 hidden layers with neurons 3 & 10 respectively. The model shows a good fit and good performance and no signs of overfitting.



**Average Absolute Error (AAE)**

|  | **Training** | **Validation** |
| --- | --- | --- |
| **Neural Network** | **311.7** | **314** |
| **Neural Network (considering important variables)** | **348.8** | **364.6** |

# Conclusion:

As per our results, the three most important variables are Carat, Color and Clarity which has been validated by our correlation matrix and as per the article published in Ganoksin, “The valuation of a gemstone is derived from the "FOUR C's": carat, color, clarity, and cut. Understanding all four of these is vital background to the buyer”. Additionally, the Price and Carat are highly correlated, hence Carat is identified as one of the important variables to predict the price of gemstone. Also, according to our analysis (Box plot analysis discussed above) the gemstones with SI2 and SI1 clarity are more profitable and the popularity of colored gemstones have increased for their right price and uniqueness where color J is the most profitable.

A few shortcomings for our project were there were limited dependent variables in the dataset and the data is 2 years old. The gemstone prices are heavily influenced by market conditions, so data should be continuously updated for more accurate and relevant predictions.

Despite these shortcomings, we were successfully able to determine the top 3 variables affecting the price of gemstones which has been justified by “The 4 C’s of “Gemstones Prediction” article published in Ganoksin.

# References

1. Kaggle, “Gemstone Price Prediction”, <https://www.kaggle.com/datasets/colearninglounge/gemstone-price-prediction?select=cubic_zirconia.csv>.
2. The 4 C’s of “Gemstones Prediction”, <https://www.ganoksin.com/article/4-cs-gemstone-valuation/>
3. Gemstone market, <https://www.futuremarketinsights.com/reports/gemstones-market>

# Appendix:

**Multiple Linear Regression:**

proc import out=gemstone datafile="/home/u62188546/sasuser.v94/DataScience Project/cubic\_zirconia.xlsx"

dbms=xlsx replace;

sheet="cubic\_zirconia";

run;

proc freq data=gemstone;

table cut color clarity;

run;

data gemstone1;

set gemstone;

if cut="Good" then Good=1; else Good=0;

if cut="Ideal" then Ideal=1; else Ideal=0;

if cut="Premium" then Premium=1; else Premium=0;

if cut="Very Good" then VGood=1; else VGood=0;

if color="D" then D=1; else D=0;

if color="E" then E=1; else E=0;

if color="F" then F=1; else F=0;

if color="G" then G=1; else G=0;

if color="H" then H=1; else H=0;

if color="I" then I=1; else I=0;

if clarity="IF" then cIF=1; else cIF=0;

if clarity="SI1" then SI1=1; else SI1=0;

if clarity="SI2" then SI2=1; else SI2=0;

if clarity="VS1" then VS1=1; else VS1=0;

if clarity="VS2" then VS2=1; else VS2=0;

if clarity="VVS1" then VVS1=1; else VVS1=0;

if clarity="VVS2" then VVS2=1; else VVS2=0;

run;

proc freq data=gemstone1;

table cut color clarity Good Ideal Premium VGood D E F G H I cIF SI1 SI2 VS1 VS2 VVS1 VVS2;

run;

proc hpreg data=gemstone1 seed=12345;

model price = carat depth table x y z Good Ideal Premium VGood D E F G H I cIF SI1 SI2 VS1 VS2

VVS1 VVS2;

partition fraction(validate=0.3);

run;

proc hpreg data=gemstone1 seed=12345;

model price=carat depth table x y z Good Ideal Premium VGood D E F G H I cIF SI1 SI2 VS1 VS2

VVS1 VVS2;

selection method=stepwise(choose=validate);

partition fraction(validate=0.3);

run;

proc hpreg data=gemstone1 seed=12345;

model price=carat depth table x y z Good Ideal Premium VGood D E F G H I cIF SI1 SI2 VS1 VS2

VVS1 VVS2;

selection method=forward(choose=validate);

partition fraction(validate=0.3);

run;

proc hpreg data=gemstone1 seed=12345;

model price=carat depth table x y z Good Ideal Premium VGood D E F G H I cIF SI1 SI2 VS1 VS2

VVS1 VVS2;

selection method=backward(choose=validate);

partition fraction(validate=0.3);

run;

**CART**

proc import out=GemStones datafile= "/home/u62277456/sasuser.v94/cubic\_zirconia.xlsx"

dbms=xlsx replace; sheet = "Data";

run;

proc hpsplit data=GemStones nodes=detail;

class cut color clarity;

model price=cut color clarity carat depth table x y z;

grow rss;

prune cc;

partition fraction(validate=0.3 seed=12345);

run;

**Histogram and side by side boxplots:**

proc import out=gemstone datafile="/home/u61150156/sasuser.v94/Data Science/cubic\_zirconia.csv"

dbms=csv replace;

run;

proc sgplot data=gemstone;

vbox price/category=cut;

title "Median Price of a gemstone by Cut";

run;

proc sgplot data=gemstone;

vbox price/category=color;

title "Median Price of a gemstone by Color";

run;

proc sgplot data=gemstone;

vbox price/category=clarity;

title "Median Price of a gemstone by Clarity";

run;

proc sgplot data=gemstone;

histogram price/scale=count;

title "Histogram of Gemstone price";

xaxis label="Price range";

yaxis label="Count";

run;

proc sgplot data=gemstone;

histogram carat/scale=count;

title "Histogram of Gemstone carat";

xaxis label="Carat range";

yaxis label="Count";

run;

**/\*Neural Network\*/**

proc import out=GEM datafile="/home/u61165957/sasuser.v94/cubic\_zirconia.xlsx"

dbms=xlsx replace;

sheet="cubic\_zirconia";

run;

proc hpneural data=GEM;

partition fraction(validate=0.3 seed=12345);

target price/level=int;

input cut color clarity/level=nom;

input carat depth table x y z/level=int;

hidden 10;

train maxiter=1000 numtries=5;

run;

proc hpneural data=GEM;

partition fraction(validate=0.3 seed=12345);

target price/level=int;

input cut color clarity/level=nom;

input carat depth table x y z/level=int;

hidden 20;

train maxiter=1000 numtries=5;

run;

proc hpneural data=GEM;

partition fraction(validate=0.3 seed=12345);

target price/level=int;

input cut color clarity/level=nom;

input carat depth table x y z/level=int;

hidden 5;

train maxiter=1000 numtries=5;

run;

proc hpneural data=GEM;

partition fraction(validate=0.3 seed=12345);

target price/level=int;

input cut color clarity/level=nom;

input carat depth table x y z/level=int;

hidden 7;

train maxiter=1000 numtries=5;

run;

**/\*Final neural network built on important variables\*/**

proc hpneural data=GEM;

partition fraction(validate=0.3 seed=12345);

target price/level=int;

input color clarity/level=nom;

input carat /level=int;

hidden 3;

hidden 10;

train maxiter=100 numtries=5;

run;