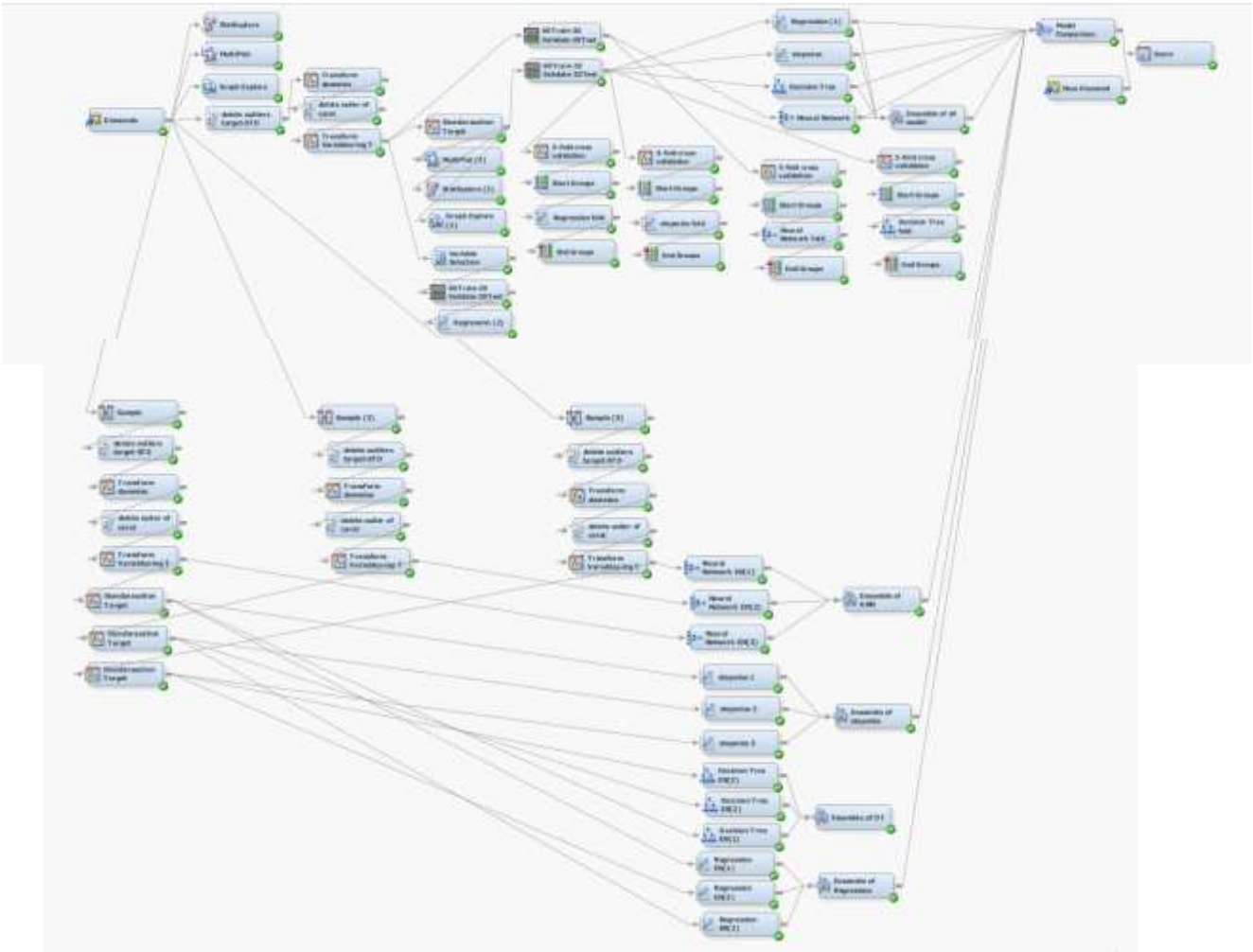
Predictive Analytics

Table of content



Executive summary

A jewelry company wants to put in a bid to purchase a large set of diamonds but is unsure how much it should bid. In this project, results from a predictive model to make recommendation on how much the jewelry company should bid for the diamonds.

In this project, data of the diamonds and new diamonds have on common columns, in order to predict the new diamonds, price we need to use diamonds therefore we have to use common features. In this project Univariate analysis and Bivariate analysis has been done to find the missing value outliers and treat them and discover correction or feature selection. As a result, price and carat have been cleaned by deleting outliers and log transformation in order to meet the linear regression assumptions and other models used in this project such as stepwise regression, ANN and Decision Tree.

After applying the models, the cross validation has been deployed to validate the result of the initiated model, then we ensemble the models both with itself or other models. The result of comparison shows that the ensemble of the ANN model is the best choice for the prediction. And we use them to predict the price of the new diamonds.

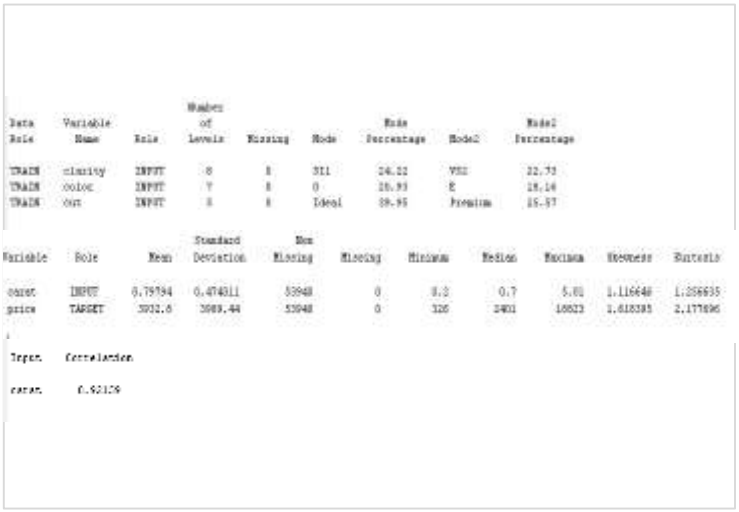
Column1	carat	cut	cut_ord	color	clarity	clarity_ord	log price	price
1246	0.21	Premium	4	E	SI2	2	5.85677626	349.5953
366	0.23	Very Good	3	E	VVS2	6	6.27386261	530.5226
654	0.23	Very Good	3	F	VVS1	7	6.29780409	543.3774
755	0.23	Very Good	3	F	VS1	4	6.12729505	458.1951
1125	0.23	Good	2	E	VVS2	6	6.27473466	530.9855
1202	0.23	Good	2	G	VVS1	7	6.24363412	514.7257
1228	0.23	Very Good	3	E	VVS2	6	6.27386261	530.5226
1254	0.23	Very Good	3	E	VVS1	7	6.34046512	567.06
1295	0.23	Very Good	3	D	VS2	5	6.19311045	489.3659
1746	0.23	Very Good	3	H	VVS1	7	6.14411712	465.9681
1909	0.23	Good	2	E	VVS1	7	6.32131203	556.3024
2513	0.23	Very Good	3	E	VVS1	7	6.34046512	567.06
2530	0.23	Good	2	E	VVS2	6	6.27473466	530.9855
2706	0.23	Ideal	5	E	SI2	2	5.871258	354.6949
327	0.24	Ideal	5	E	VVS2	6	6.39928053	601.4122
667	0.24	Ideal	5	G	VS1	4	6.19351818	489.5655
670	0.24	Very Good	3	E	VVS1	7	6.38565208	593.2715
906	0.24	Very Good	3	E	VVS2	6	6.31764705	554.2673
927	0.24	Very Good	3	E	VS1	4	6.20292342	494.1917
1081	0.24	Good	2	F	VVS1	7	6.32100351	556.1308
1280	0.24	Good	2	F	VS1	4	6.16780286	477.1366
1749	0.24	Very Good	3	J	VVS2	6	5.89503682	363.2302
1882	0.24	Ideal	5	F	IF	8	6.52887397	684.6269
2158	0.24	Very Good	3	F	VVS1	7	6.34224842	568.0721
65	0.25	Very Good	3	E	VVS2	6	6.36175402	579.2615

Explanatory Data Analysis
univariate & Bivariate Analysis

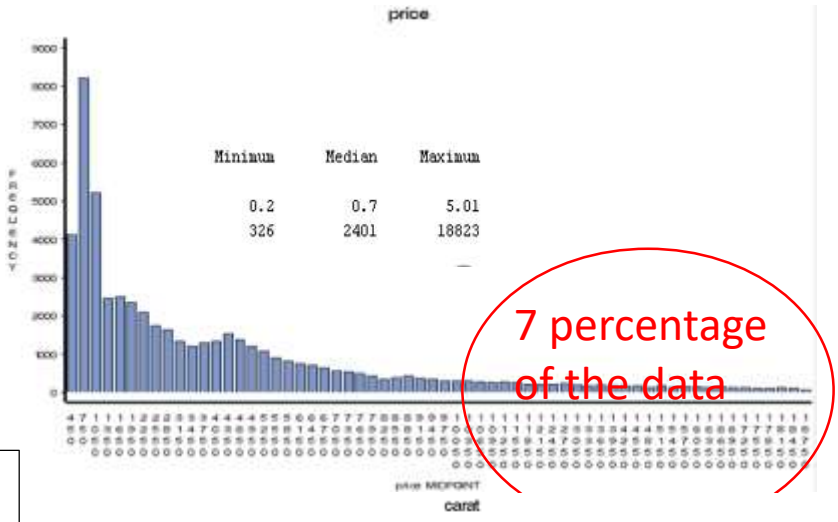
The statistic values show there are two type of variables, character variable and numeric variables.

The character variable should be turn to be dummies or numeric variable in order to be used in the regression, ANN.

The numeric data are skewed, and we need to fix them, using outlier treatments and transformation to have less bias on the model.

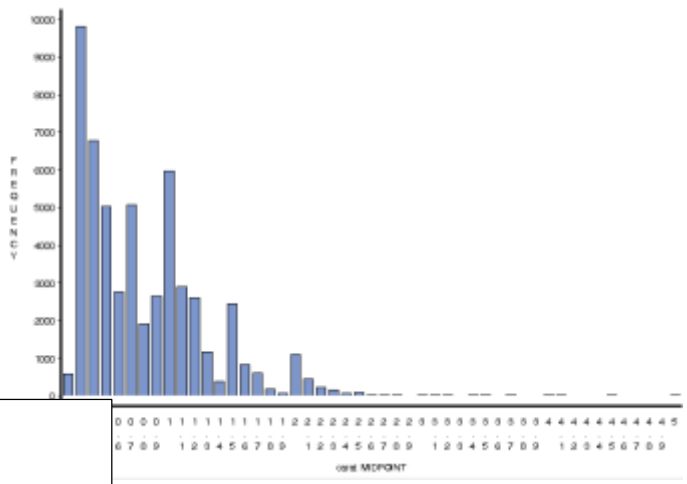


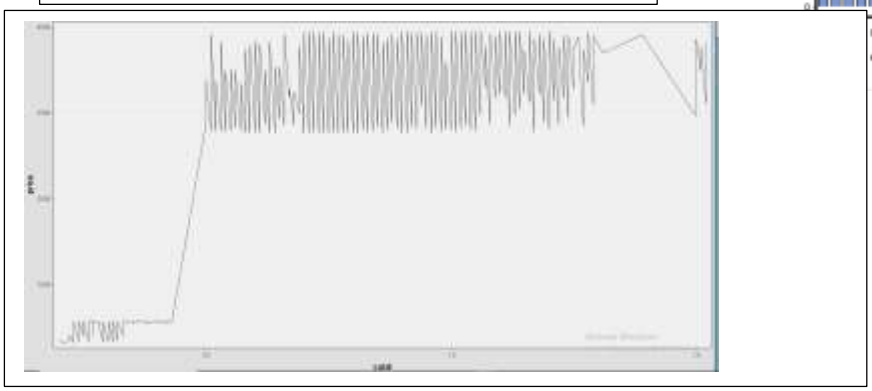
Price is skewed data, after various transformations and outlier treatments the best possible way to handle this part id using STD filter and Log transformation.



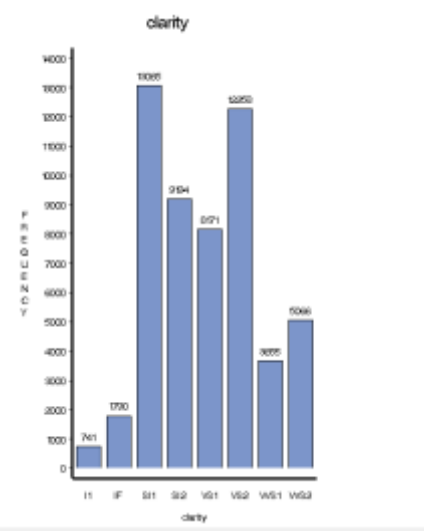
Carat is skewed data, after deep investigation, and looking at the carat respect to price we can see there is a fluctuation and we cannot do Binnig to squeeze skewedness. the best possible way to handle this part id using STD filter.

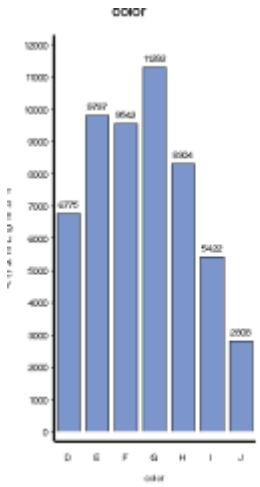
The total outliers is 663 point and half of them will be delete by deleting outlier of the price. So deleting almost 300 data points wont effect the model badly.

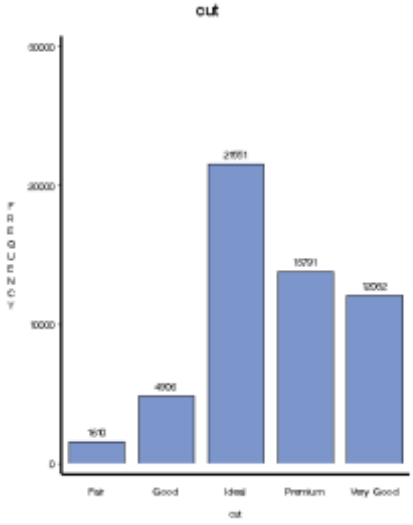




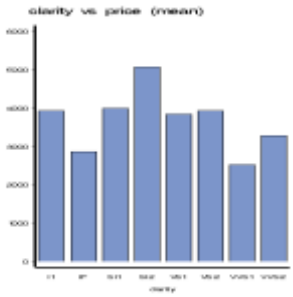
Univariate analysis of the character variables shows the frequencies of each class of the variables varies and there some class with less frequencies that we need to keep in mind for selecting sampling method

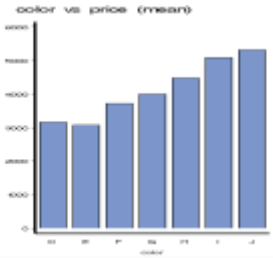


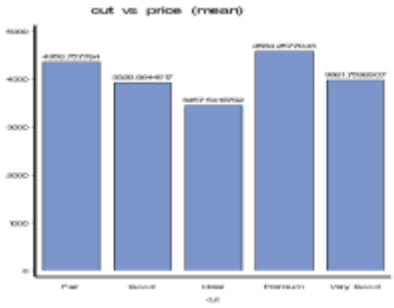




The bivariate analysis shows there are a correlation between carat and price. different value of others variable have slightly change in the price

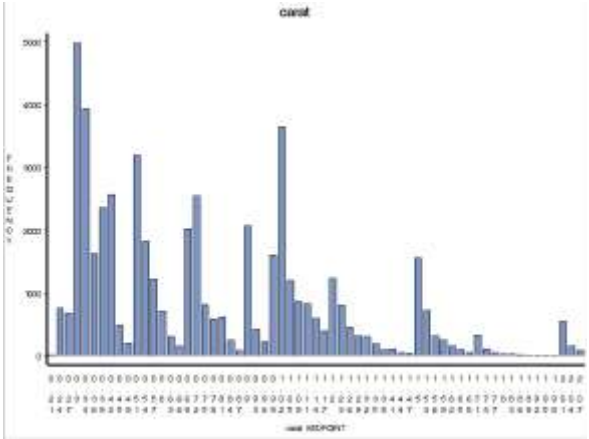


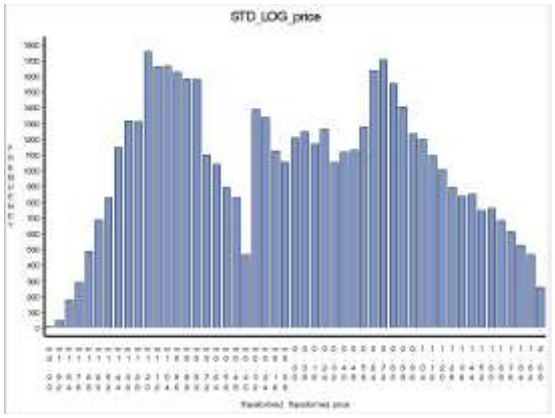




Changes after cleaning outliers and transformation, standardize the target and creating dummies we have these statistic figures,

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum
carat	INPUT	0.756986	0.416509	52281	0	0.2	0.7	2.08
STD_LOG_price	TARGET	4.77E-11	1	52281	0	-1.99495	0.016703	2.004229



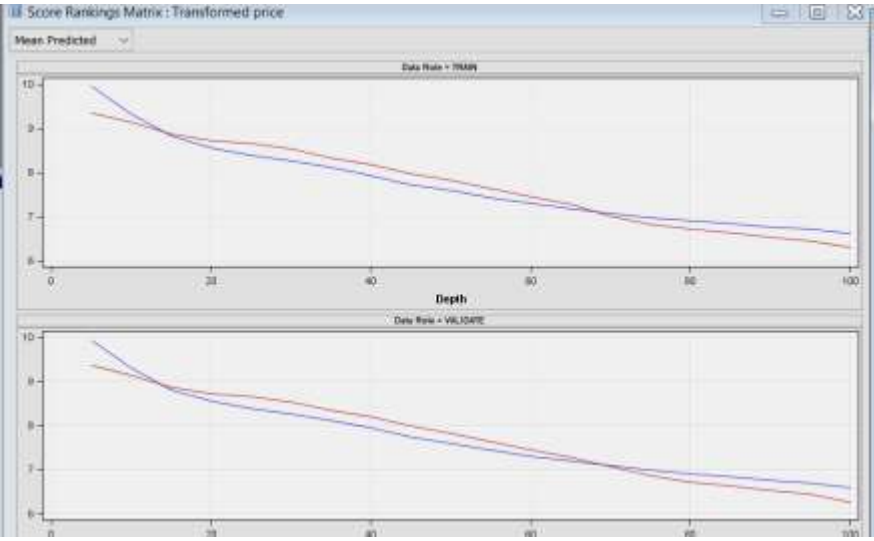


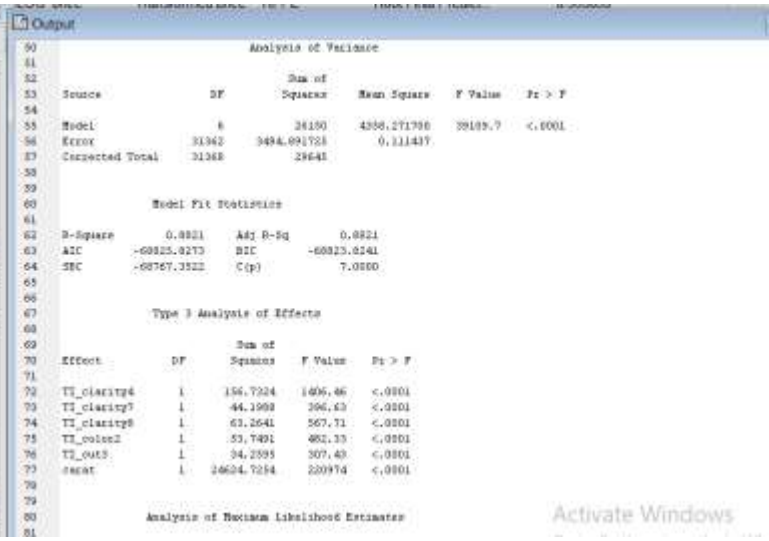
Now everything is ready for applying the models.

Linear regression models (regression and stepwise)

Assumptions of regression: normality and equality of variance in the error term, linearity of the ID, non-heteroscedasticity

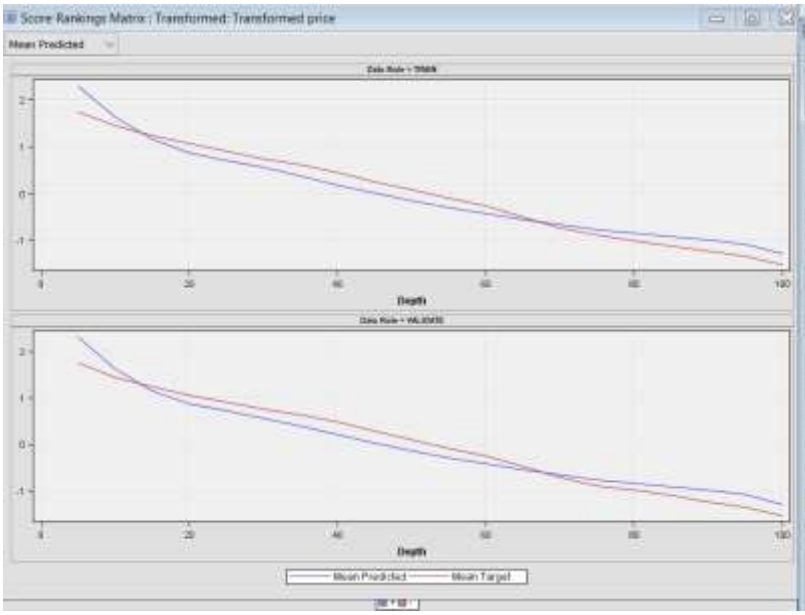
Linear regression with feature selection Is showing 88 percentage of the variation is data is captures and f value is high enough the ASG is .11 which low but it can be lower.

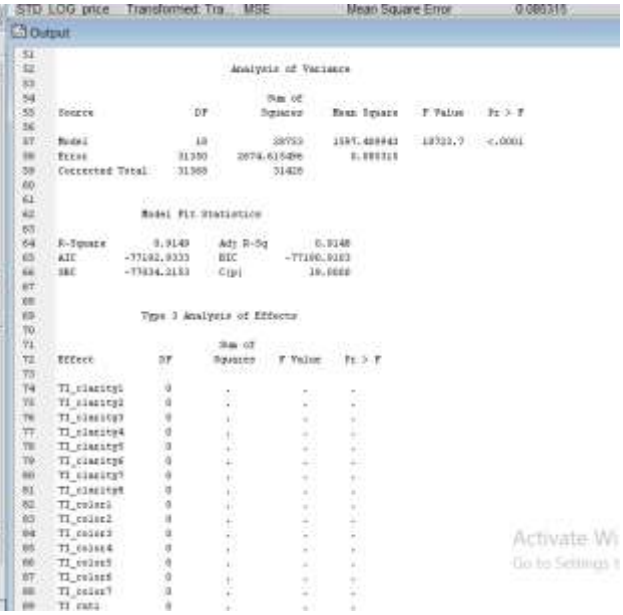




Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	-68825.83	.	.
ASE	Average Squared Error	0.11	0.11	0.11
AVERR	Average Error Function	0.11	0.11	0.11
DFE	Degrees of Freedom for Error	31362.00		

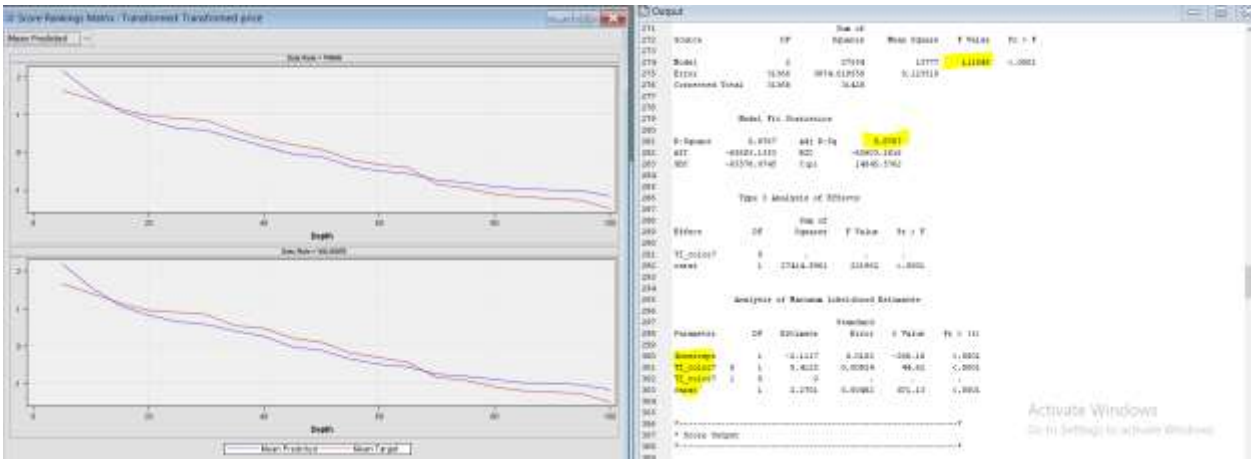
Regression on the Whole model without feature selection :





This is showing the 91 percentage for the Adjusted R2 and F value is 18723.7 there are two features that needs to be remove from the model those might effect on the f value. Using stepwise we can see the f value significantly increase as the crap variables will be removed. There ASE here is 0.09 less than previous model.

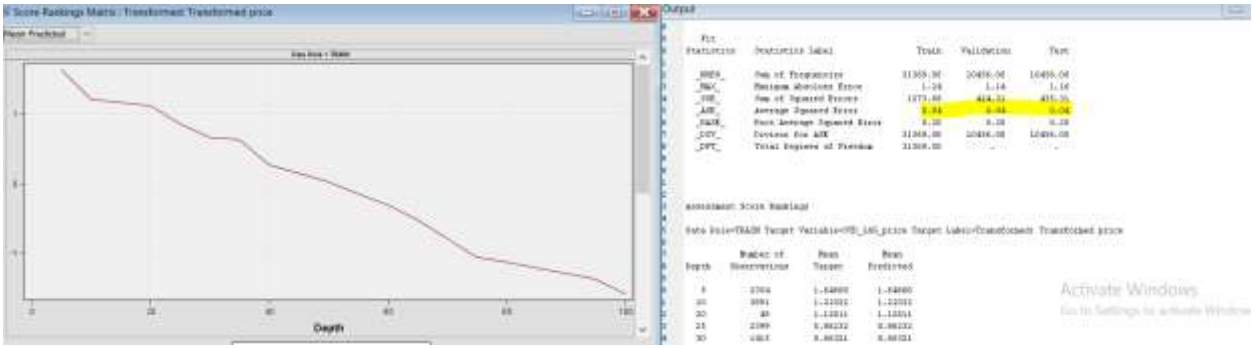
Stepwise Regression:



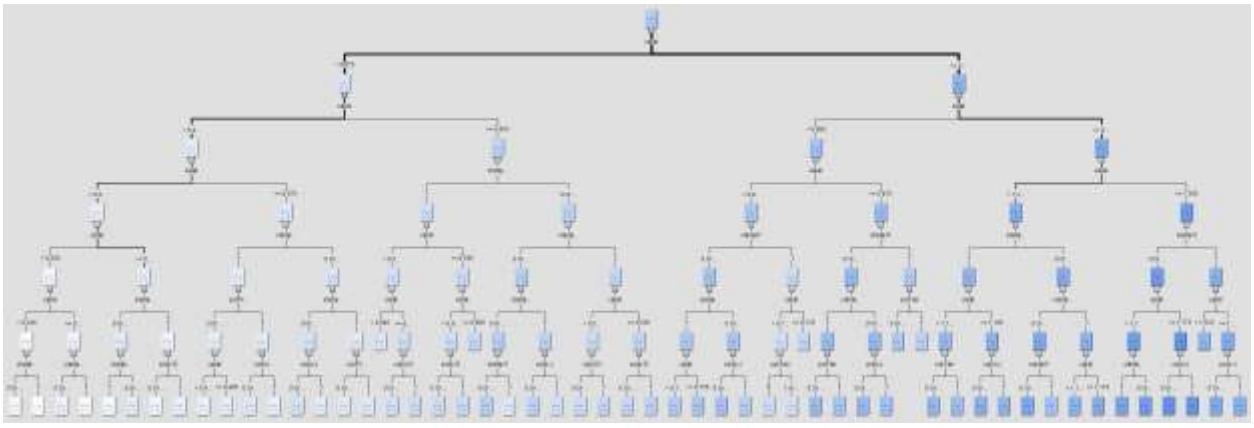
The f value is 111545 is really higher than other two regression , and the Adj R2 is 87 %

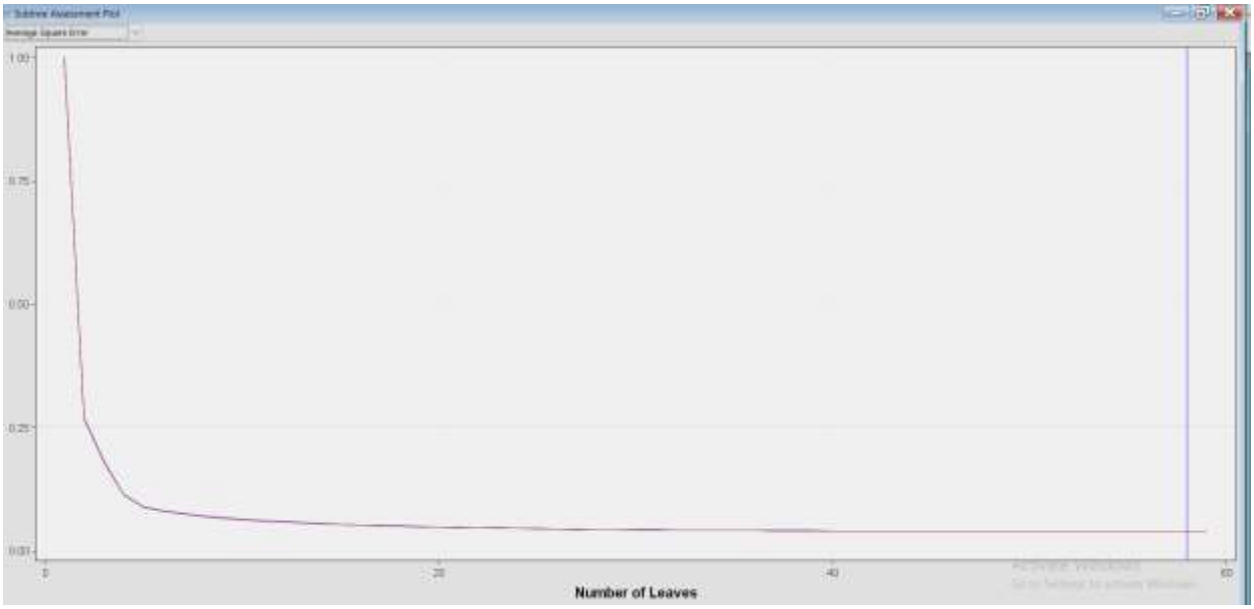
The Average square error is .12

Decision Tree :



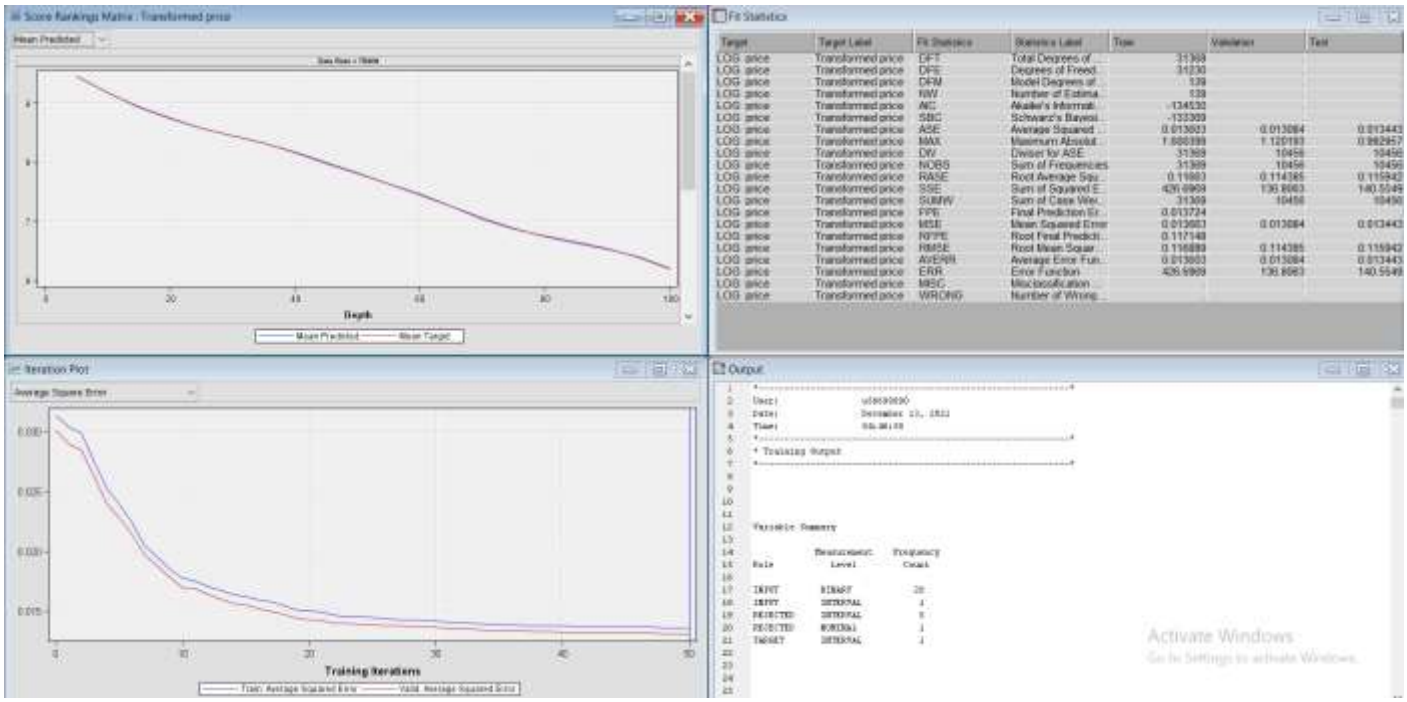
Is perfectly matches the target variable and error is much lower than regression , the max level is 6 and max branches is 2.





There is no Over or unbefitting in the data and number of leaves is close to 60.

ANN:



The ANN is showing perfect model compare to others and has the ASE of 0.1 , it perfectly matches the target.

.. Property	Value
Training Technique	Default
Maximum Iterations	50
Maximum Time	5 Minutes
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1

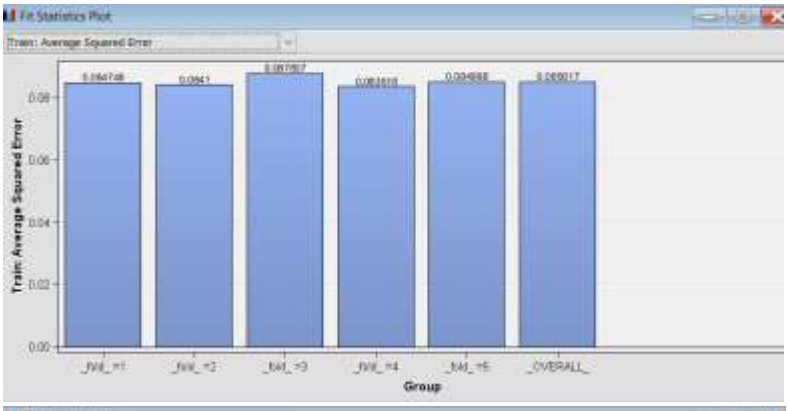
Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	6
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default

Default

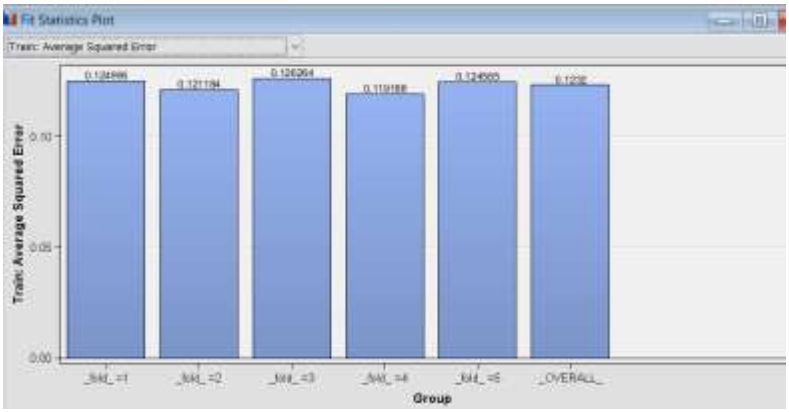
Number of hidden unite is defined 6 , and number of iteration is Max 50

Ensemble Models and Cross Validation of each model :

All the ensemble models of each model have almost similar result of each indivisual models except Ann which perform slightly better. Cross Validation of each model shows that in all folds the ASE are same and validate the model itself.



Regression (without feature selection)



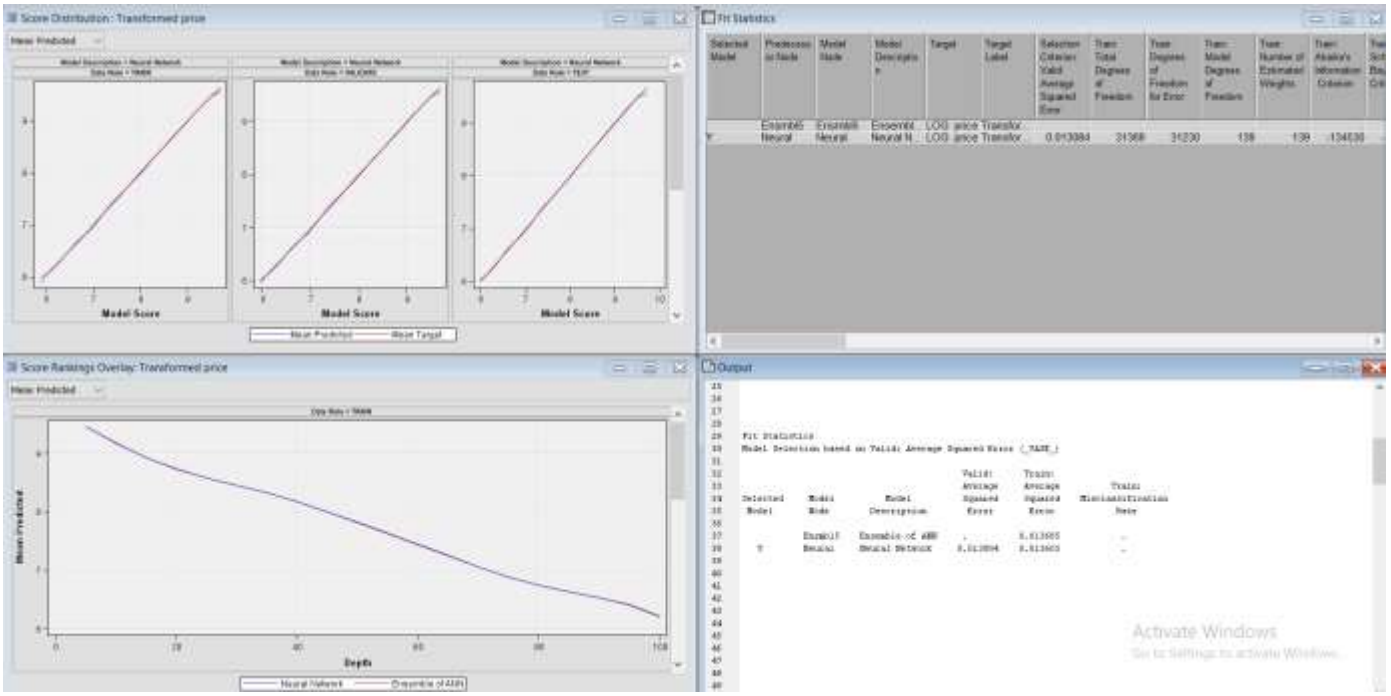
stepwise

ANN

Decision Tree

Final Model:

Ensemble of the ANN is the winner.



Predict the price of new diamonds using ensemble of the ANN

Column1	carat	cut	cut_ord	color	clarity	clarity_ord	log price	price
1246	0.21	Premium	4	E	SI2	2	5.85677626	349.5953
366	0.23	Very Good	3	E	VVS2	6	6.27386261	530.5226
654	0.23	Very Good	3	F	VVS1	7	6.29780409	543.3774
755	0.23	Very Good	3	F	VS1	4	6.12729505	458.1951
1125	0.23	Good	2	E	VVS2	6	6.27473466	530.9855
1202	0.23	Good	2	G	VVS1	7	6.24363412	514.7257
1228	0.23	Very Good	3	E	VVS2	6	6.27386261	530.5226
1254	0.23	Very Good	3	E	VVS1	7	6.34046512	567.06
1295	0.23	Very Good	3	D	VS2	5	6.19311045	489.3659
1746	0.23	Very Good	3	H	VVS1	7	6.14411712	465.9681
1909	0.23	Good	2	E	VVS1	7	6.32131203	556.3024
2513	0.23	Very Good	3	E	VVS1	7	6.34046512	567.06
2530	0.23	Good	2	E	VVS2	6	6.27473466	530.9855
2706	0.23	Ideal	5	E	SI2	2	5.871258	354.6949
327	0.24	Ideal	5	E	VVS2	6	6.39928053	601.4122
667	0.24	Ideal	5	G	VS1	4	6.19351818	489.5655
670	0.24	Very Good	3	E	VVS1	7	6.38565208	593.2715
906	0.24	Very Good	3	E	VVS2	6	6.31764705	554.2673
927	0.24	Very Good	3	E	VS1	4	6.20292342	494.1917
1081	0.24	Good	2	F	VVS1	7	6.32100351	556.1308
1280	0.24	Good	2	F	VS1	4	6.16780286	477.1366
1749	0.24	Very Good	3	J	VVS2	6	5.89503682	363.2302
1882	0.24	Ideal	5	F	IF	8	6.52887397	684.6269
2158	0.24	Very Good	3	F	VVS1	7	6.34224842	568.0721
65	0.25	Very Good	3	E	VVS2	6	6.36175402	579.2615