# Diamond Project

# Deeksha Dave

# Anushashna FNU

# PART 1

# Price Prediction

## APPROACH

* Training set of 246 and test set of 50 predictions was taken after splitting the data. The descriptive stats and model were run for the training set and validated on test set. Later after improving the accuracy of model, the model was used to predict the price.
* Dummies were created for Vendor, Colour, Clarity.
* The relationship of price with various parameters was seen in descriptive statistics
* The average prices of vendors for every cut was observed using pivot tables
* The changes in price for every change in cut, colour, clarity was seen using the pivot tables.

## BACKGROUND RESEARCH ABOUT DIAMOND INDUSTRY

* Price greatly depends on the carat of the diamonds. Hence, it is important to see the price fluctuation with respect to size of diamonds. Usually there is lesser variation in price for smaller carat diamonds. As the carat increases, the price fluctuation would increase. General perception is that there is greater fluctuation for 3 carat round diamonds than smaller carat round diamonds. 4 carats and above show higher price fluctuations.
* Limited Volatility: We need to also consider the DeBeers factor and economic climate. When the diamond price is higher, people invest in larger stocks and hence they would be willing to pay higher prices.
* It is important to follow the type of diamond and observe the price fluctuation
* Cut is an important factor contributing to the prices. There is around 10-20% price difference between excellent ad good category of diamonds.
* Fluorescence is an important factor determining the prices. We need to see the percentage difference of strong and no fluorescence. Diamonds with fluorescence is 10-20% cheaper than diamonds with no fluorescence.
* Grade of diamonds would make a price difference ranging from 5%-30%
* Polish and symmetry can affect the price of diamonds by 7-10%
* The 6 major factors affecting the price of diamonds are: carat weight, cut, color, clarity, shape, AGS or GIA certification

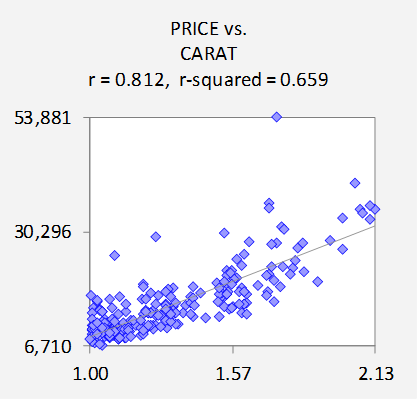
## DESCRIPTIVE STATS AND ANALYSIS

## MODEL



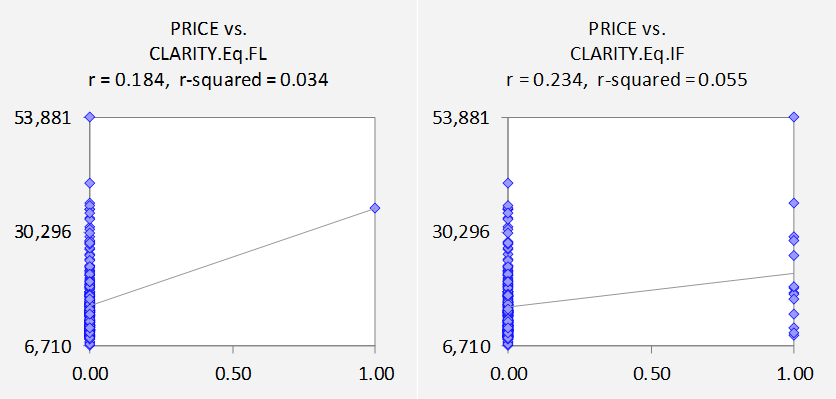
The highest correlation of price was with carat for this dataset (0.812). After carat, clarity and colour were important factors. However, carat had the highest correlation with price. Hence, it was important to see the graph of price vs carat to know the nature of the graph.

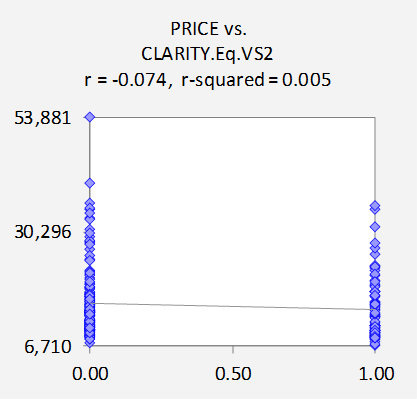
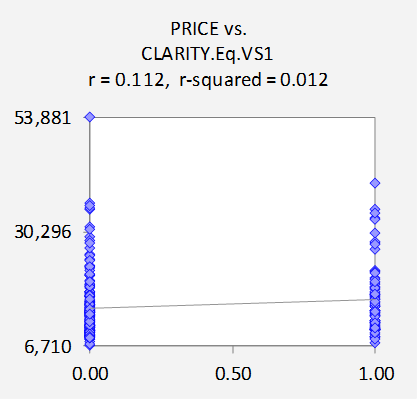
### CARAT

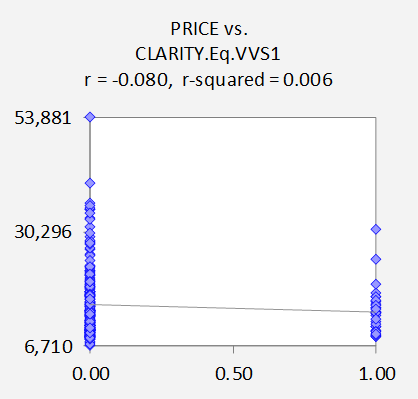


As seen from the Price vs carat graph, the graph curves at the lower end and is not completely linear. Hence, some transformation is need for a completely linear correlation.

### CLARITY







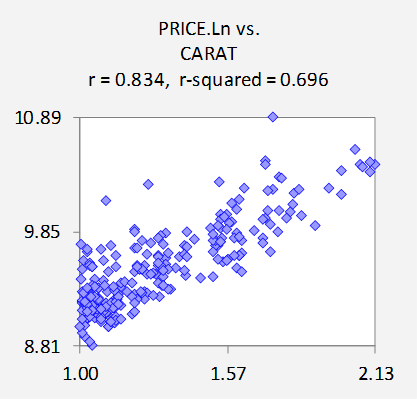
From the graphs of Descriptive stats, we can see that FL, IF have greater price as compared to VS1, VS2, VVS1, VVS2 proving the fact that FL, IF diamonds are premium quality with greater clarity which is affected on the price.

### COLOR

Not very strong correlation of color and price were seen in this dataset.

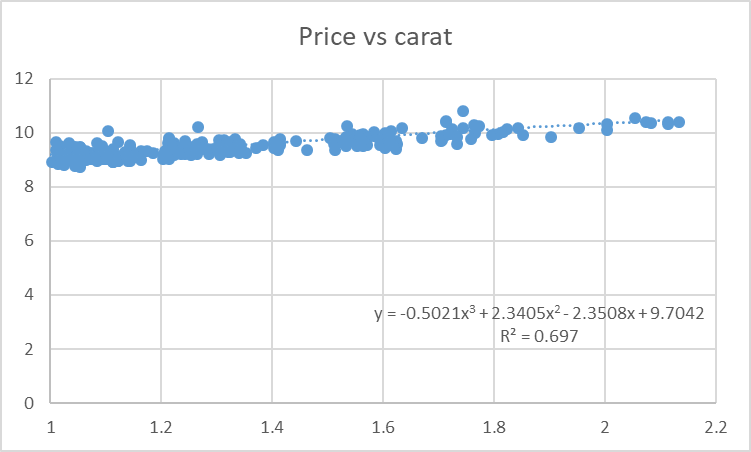
### CARAT CORRELATION WITH PRICE

Due to non-linear pattern seen between the price and carat, some transformation in data was needed. Hence, since Log transformation is most common, we decided to go ahead with log transformation of price to achieve the linear correlation. Moreover, as seen in price vs carat graph earlier, the variance was not constant and was increasing along the x-axis. All these factors point towards log transformation.



After log transformation, the graph of price vs carat was much more linear with higher correlation. (r-squared of 0.696 as compared to 0.659).

To increase the correlation between the price and carat, various trendline models were fitted on the data. The polynomial relation between carat and log price gave the best correlation. Thus, polynomial function of order 3 was taken.

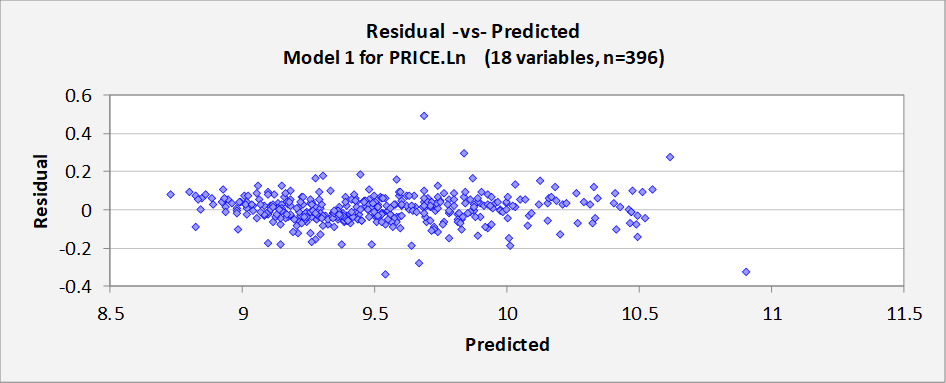


### MODEL TRIALS- PART 3

HeartXArrows and variables including HxA are the ones on which cut depends.

#### Model 1









The model included all variables. However, certain variables with insignificant p-values were removed and the model was rerun. Moreover, the normal quantile plot shows deviation for very low and very high values of prices. The residual vs predicted plot shows variance in the errors. The errors increase as the value on the x-axis increases. This suggests a better model could be made.

#### Model 2







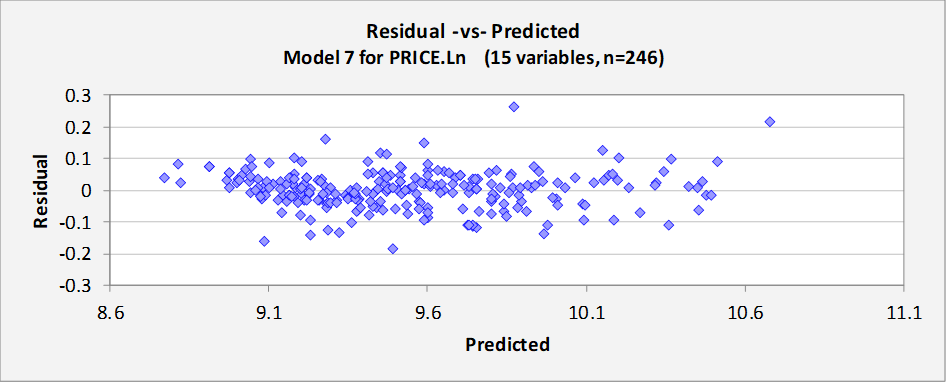
This model predicted worse than the earlier model with RMSE almost like the earlier model and lower r-squared. The residual vs predicted plot seemed better than the earlier model, however the normal quantile plot showed huge deviations for lower and higher price range data. Carat still has the strongest correlation. Hence, to increase the linearity of carat, the next model takes into consideration polynomial fitting of log price with carat for better linear fit. After various iterations, we came up with the model below

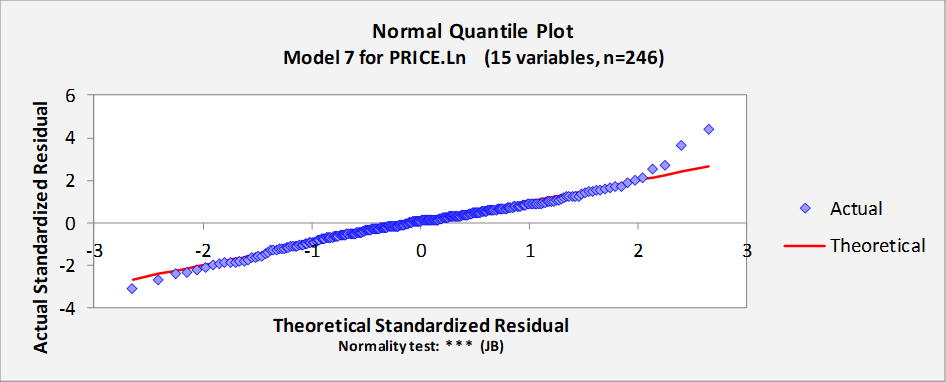
#### Model 7

Here, we created dummies separately for IF, FL clarity and other grades of clarity due to large difference in correlation among IF, FL and others. We run the regression model using polynomial equation for carat and carat^2 and carat^3 were more important than carat itself. Hence, removing the non-significant variables, the following model was created.



There was an improvement in the RMSE from the previous model. (0.077 to 0.059). The r-squared and adjusted r-squared values increased as compared to the earlier models.



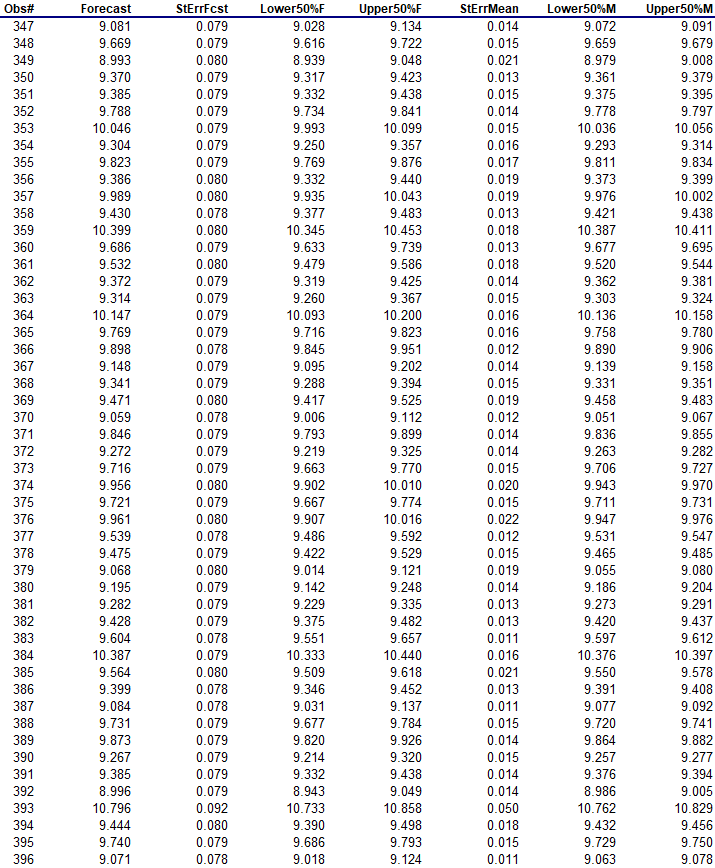


In the residual vs predicted plot, the variance remains almost constant over the entire x-axis. This shows that the errors are close to i.i.d. errors with almost constant variance. However, we can see a slight increase for higher price range values.

From the normal quantile plot, we can clearly see that the model misses forecast for very low values and misses by a huge extent for higher price range values. We tried to incorporate various factors including segregation of clarity and color based on their superiority, however the fit for higher price values did not give an accurate value. Hence, we need to have more information on the rarity of the diamonds and the availability of the diamonds which also affect the price of diamonds. Rare diamonds demand for a premium price which could not be incorporated within this model. However according to us this was the best model and was used for forecasting of prices.

PREDICTION OF PRICES ON ENTIRE DATASET

The table below forecasted for log price. The Price\_Forecast column gives the prediction of prices unlogged. The RMSE was 0.078 which is much







The plot of residual vs predicted shows almost constant variance over the x-axis. However, for the forecast, there is a certain curving pattern within the graph for lower price range values. Hence, different models for high and low-price range could be the next logical step to improve the predictions of the model. For higher prices, the error increases suggesting that model does not predict well for very high prices.

The normal quantile plot shows deviation from normality at very low and very high prices. Hence, we could make different models for higher and lower price ranges or an extra parameter is needed to segregate the very low and very high prices so that they can be accurately predicted.

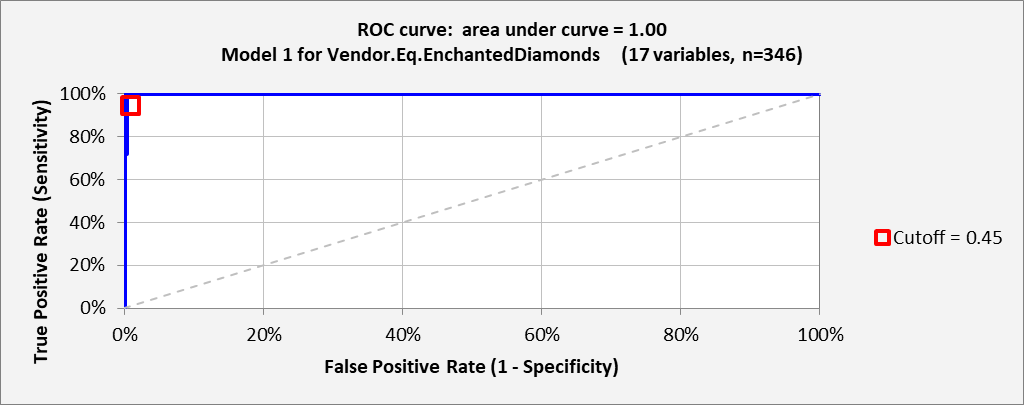
# VENDOR PREDICTION – PART 4

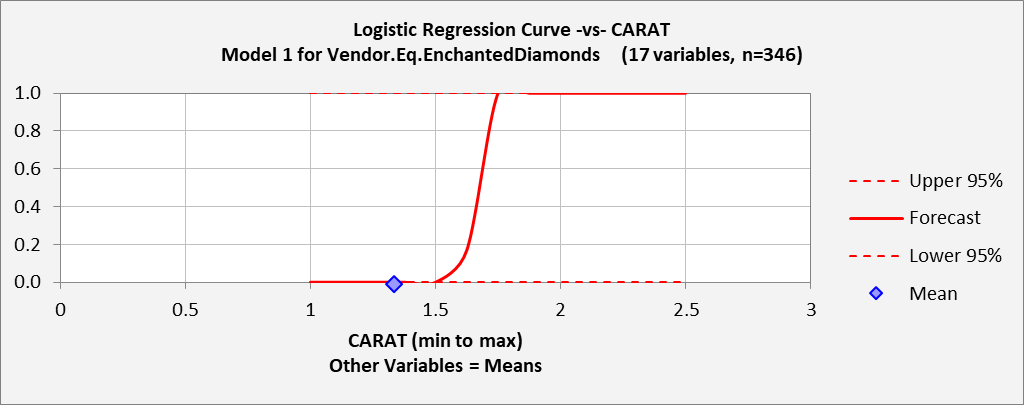
In the vendor prediction, we need to carry out Logistic Regression. The dependent variable taken was Vendor. Enchanted. Diamond dummy variable.

Cut is dependent on HeatxArrows and HxA variables.

#### Model 1







This model has a very good r-squared and adj-r-squared. However, none of the parameters are significant with p-value less than 0.05. Hence, even if the correlation is good, the model does not perform well. Moreover, the ROC curve has an area under the curve of 1 which seems unrealistic. There is a very high true positive and true negative rate which shows a possibility of over fitting the data.

Also, in the Logistic Regression Curve, the confidence limits are very wide and unrealistic. The confidence limits are close to 0 and 1 for all data points and do not follow the curve. This shows huge uncertainty in the forecast. Hence, improvement of the model is needed.

#### Model 2

To improve the logistic Regression results, we though of standardizing the price. Due to 3 order difference in price and dummy variables, maybe the regression model was not performing well. Hence, keeping everything the same, standardised price was taken as the dependent variable.



Here too, the ROC curve area is 1 and the confidence limits are close to 0 and 1. Hence, we decided to eliminate the less important variables in the dependent variables to improve the model. Here as well, none of the variables were statistically significant

#### Model 3



Here, the r-square improved. However, the area under ROC curve was still 1 and the confidence limits for the forecast were extremely unreasonable.

By further removing the dependent variables, the model was performing worse, hence a different approach was to be tried.

#### Interesting Fact

While playing with the data, we realized a flaw in data. All the diamonds with **Excellent** cut were sold by Enchanted Diamonds only and no other Vendor. Each Vendor sold only one type of cut, hence the easiest way to check if a vendor is Enchanted Diamond was to check if the cut was Excellent cut or not. Hence, we decided that the type of diamond was very important and decided to group various parameters including color, clarity, cut into premium and the rest for more accurate predictions

#### Model 4

In this model, we classified clarity in group of IF, FL and others. Moreover, we also classified the color variables in groups such as D, E-F, G-H.



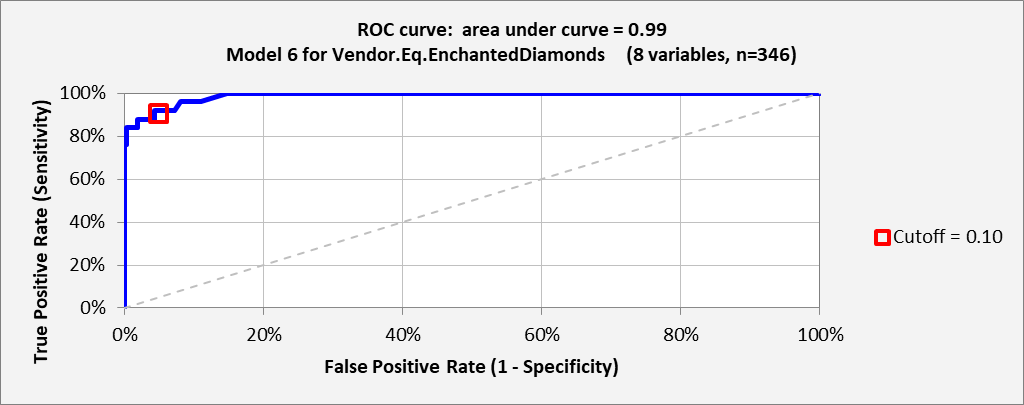
In this model, we can see that the variables have become significant with most variables p-values less than 0.05 and t-stats greater than 2. Hence, there was an improvement in the model. However, the r-squared was lower than the earlier models. However, the confidence limits are still unrealistic.

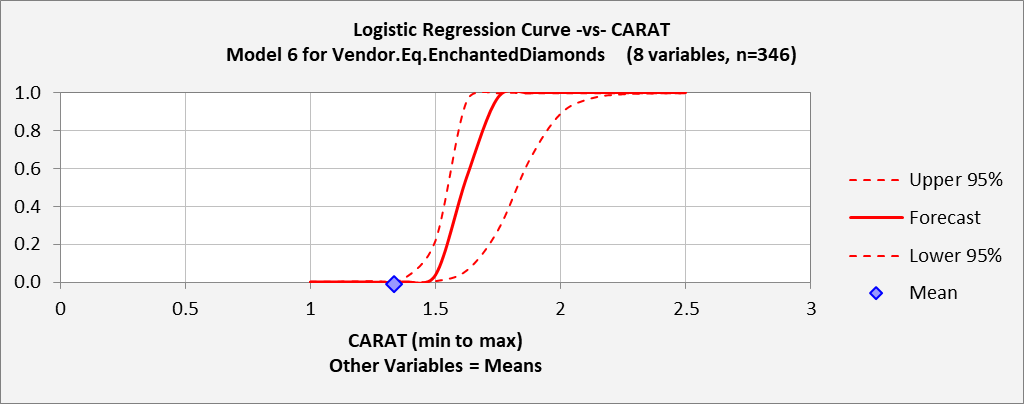
Similar model was continued while removing the non-significant dependent variables to improve the model. We also tried the model for Price, Standardized Price and log price which is present in the Excel Sheet.

#### Model 5

In this model, log price was considered as the dependent variable.





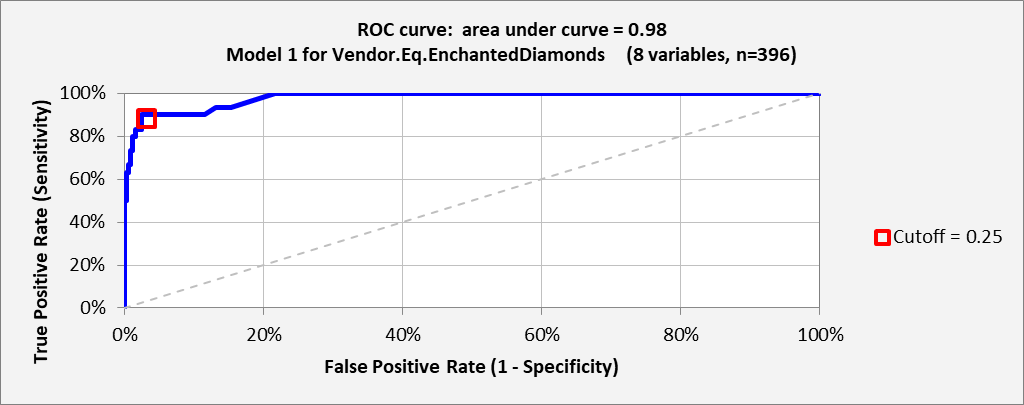


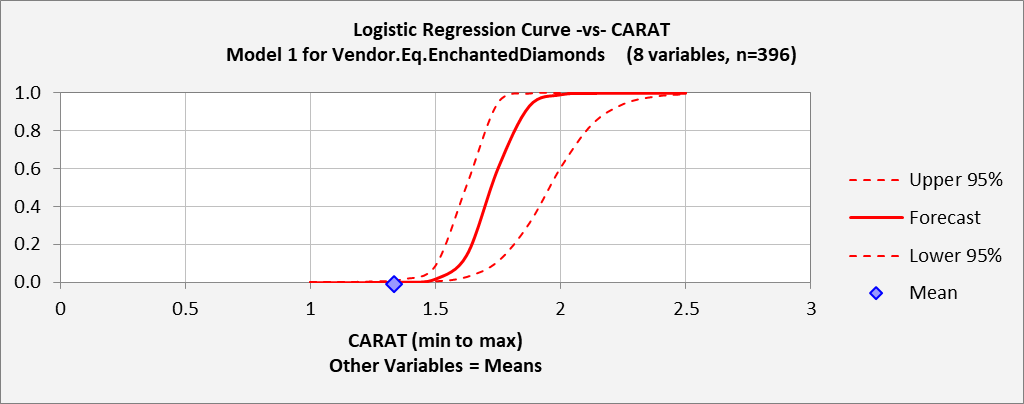
In this model, we can see a strong improvement in the confidence limits of the Forecast. The confidence limits move along with the forecast. Moreover, all the dependent variables are significant with p-values less than 0.05. The area under the ROC curve is also 0.98. Hence, this could be a good model for prediction.

#### FINAL MODEL

In this model, we considered Price as the dependent variable.







In this model, we can see all variables are significant. The r-squared and adj r-squared are lower than the earlier model. However, the confidence limits seem reasonable.

Both the models are considered for the final forecast model.

#### CUT-OFF VALUE

To decide the cut-off value, we need to see the trade-offs of the true positive and true negative rates. True positive rates are important because if we predict the vendor is Enchanted Diamond and the Vendor is not Enchanted Diamond, the customer would not get the desired Excellent quality of the cut and would end up getting different quality of diamond or paying high prices. Hence, true positive rate is important.

In True Negative rates, if we predict that the vendor is Enchanted Diamonds is not the Vendor, but turns out to be Enchanted Diamonds, the desired Vendor as well as quality and price won’t be available leading to customer dissatisfaction.

Hence, both true positive and true negative rates are important to decide the cut-off. We kept our cut-off of 0.25 for Final Model and 0.10 for the log price Model 5 based on the highest value of true positive and true negative rates.

### FORECAST

In the Excel sheet, the probabilities and binary answer of whether the Vendor is Enchanted Diamond or not are calculated.

Based on the forecast, the price model forecasted 7 Enchanted Diamonds out of the dataset of 50. Based on the number of Excellent cuts, there were total 9 Excellent cuts which means our model predicted 7 out of 9 Enchanted Diamonds with 78% accuracy.

The Log price model performed worse for forecasts. It overpredicted the number of Enchanted Diamonds predicting 11 Enchanted Diamond Vendors and including Sig-Ideal cut in Enchanted Diamonds Predictions instead of only Excellent cut predictions.

# PART 2

Question 1:

Small diamonds:

For small diamonds, the Color variables have a strong positive correlation on the price of diamond, while clarity has a strong negative correlation with the prices. In clarity, only IF category has a positive correlation. This could be maybe due to the premium quality demanding higher prices. The Depth and Length have strong positive correlations while width has a slight negative correlation. Cut was also significant but not as much of positive correlation as color. Carat also had very less positive correlation with price as compared to color.

Large diamonds:

For large diamonds, carat was insignificant. Color and clarity have strong positive correlation with the price. Width also had a strong positive correlation with the price, while depth has a relatively less positive correlation with the price.

Entire data:

For the entire data, carat and color had strong correlation with the price while clarity has a strong negative correlation on price excluding If category. Depth and length had strong negative correlation with the price. However, width did not matter much in predicting the price.

We can notice above that carat was very important for the entire data set but not very important for small and large data sets due to the classification of data sets based on carat. Hence, there was not much variation between the large and small datasets in carat. However, color had a huge impact on predicting the price of the diamonds. For larger diamonds, clarity had strong positive while for smaller diamonds, clarity had strong negative correlation. This tells us that there are greater imperfections in smaller diamonds as compared to larger diamonds. Hence, as the diamond gets larger, there would be lesser imperfections leading to higher prices. The coefficient of color was similar for small and large carat but different for entire dataset. The coefficient of clarity was similar for entire dataset and small but were very different for large. Depth, width coefficients were not same for any of the data sets

Question 2:

The dimensions length, width and depth have a strong correlation on the price. This is because most people also look at the size of the diamond to see the size vs price trade-off. Width is important because it is the most visible part of the diamond. The size of the diamond depends on the cut of the diamond which will determine how the diamond will reflect light and brilliance of diamond. The better proportion of the stone, the better brilliance and sparkle of the diamond. Hence, the dimensions of the diamond are very important. Cut quality is a major factor contributing the price of diamond and hence the dimensions of the diamond are important. The cut also determines the size vs carat. Hence, a proper cut is very important which makes the dimensions important.

Question 3:

The standard error of regression for the new dataset for large values was 0.188 in log units while the standard error of regression for Tommy Lam’s dataset for larger diamonds was 0.062. Hence, the model predicting the data based on Tommy’s data was more accurate as compared to this data. This could be since the earlier data had Vendor’s data included and each vendor sold a cut of diamond. Hence, Vendor was a more significant in determining the price since it also included the cut variable in it which is an important predictor of price. Hence, Vendor data is more reliable in predicting prices. However, the difference in error is not very significant as the depth, width and length are also factors on which cut is dependent. Vendor is also more familiar for common man to decide than make it based on the dimensions.

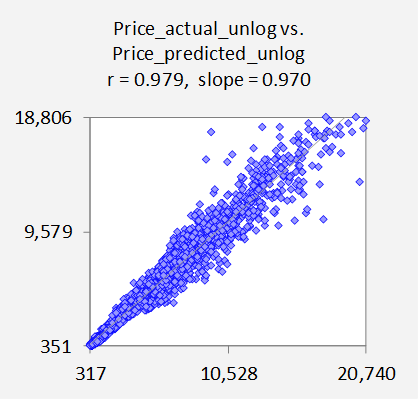
Question 4:

If we look at the Residual vs Predicted graph for small diamonds, there is a good forecast for smaller values, but as the carat increases, the variance in the errors seem to increase. Hence, it is not very effective for higher values within the small dataset values. However, there is no such trend in the residual vs predicted plot. From the normal quantile plot, we can see that there is deviation from normality at higher values.

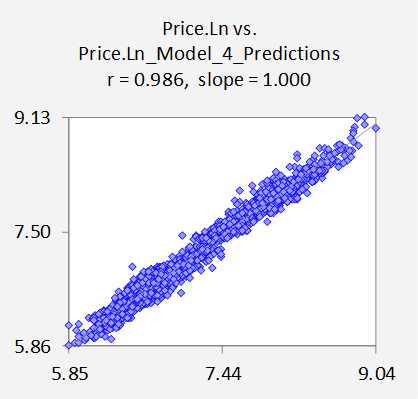
However, in the residual vs predicted plot for large values, the errors vary downwards and hence we can say that the model is overpredicting the data. The larger model also deviates from normality at smaller range values of carat.

In the entire data set including small plus large carat, it predicts well for mid-range but there is an increase in the variance of errors for larger values and slight deviation for smaller values as seen from the normal quantile plot and residual vs predicted plot.

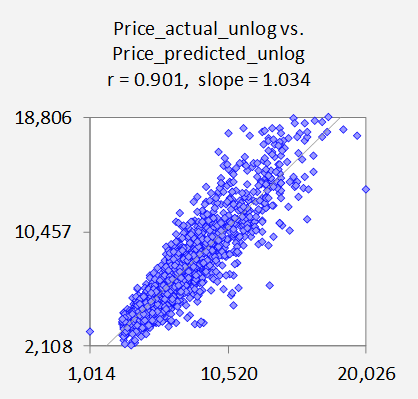
Question 5:



For the entire data set



For small diamonds



For large diamonds