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

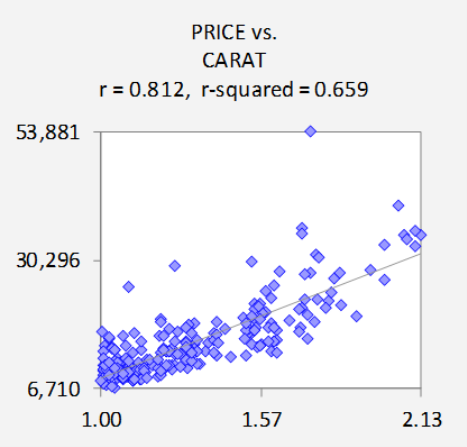
DESCRIPTIVE STATS AND ANALYSIS

MODEL

Correlations and Squared Correlations -vs- PRICE (n=246)		
Variable	Correlation	Squared
PRICE	1.000	1.000
CARAT	0.812	0.659
CLARITY.Eq.FL	0.184	0.034
CLARITY.Eq.IF	0.234	0.055
CLARITY.Eq.VS1	0.112	0.012
CLARITY.Eq.VS2	-0.074	0.005
CLARITY.Eq.VVS1	-0.080	0.006
CLARITY.Eq.VVS2	-0.153	0.024
COLOR.Eq.D	0.200	0.040
COLOR.Eq.E	0.106	0.011
COLOR.Eq.F	-0.027	0.001
COLOR.Eq.G	-0.056	0.003
COLOR.Eq.H	-0.158	0.025
HeartsXArrows	-0.033	0.001
HxA_CrownAngle_34to35	0.051	0.003
HxA_LowerGirdle_76to78	0.092	0.009
HxA_PavillionAngle_406to409	0.018	0.000
HxA_StarFacets_45to50	0.036	0.001
HxA_TableSize_54to57	0.008	0.000
Vendor.Eq.BlueNile	-0.135	0.018
Vendor.Eq.BrianGavin	0.015	0.000
Vendor.Eq.CraftedByInfinity	0.209	0.044
Vendor.Eq.EnchantedDiamonds	0.057	0.003
Vendor.Eq.JamesAllen	-0.072	0.005
Vendor.Eq.WhiteFlash	-0.045	0.002

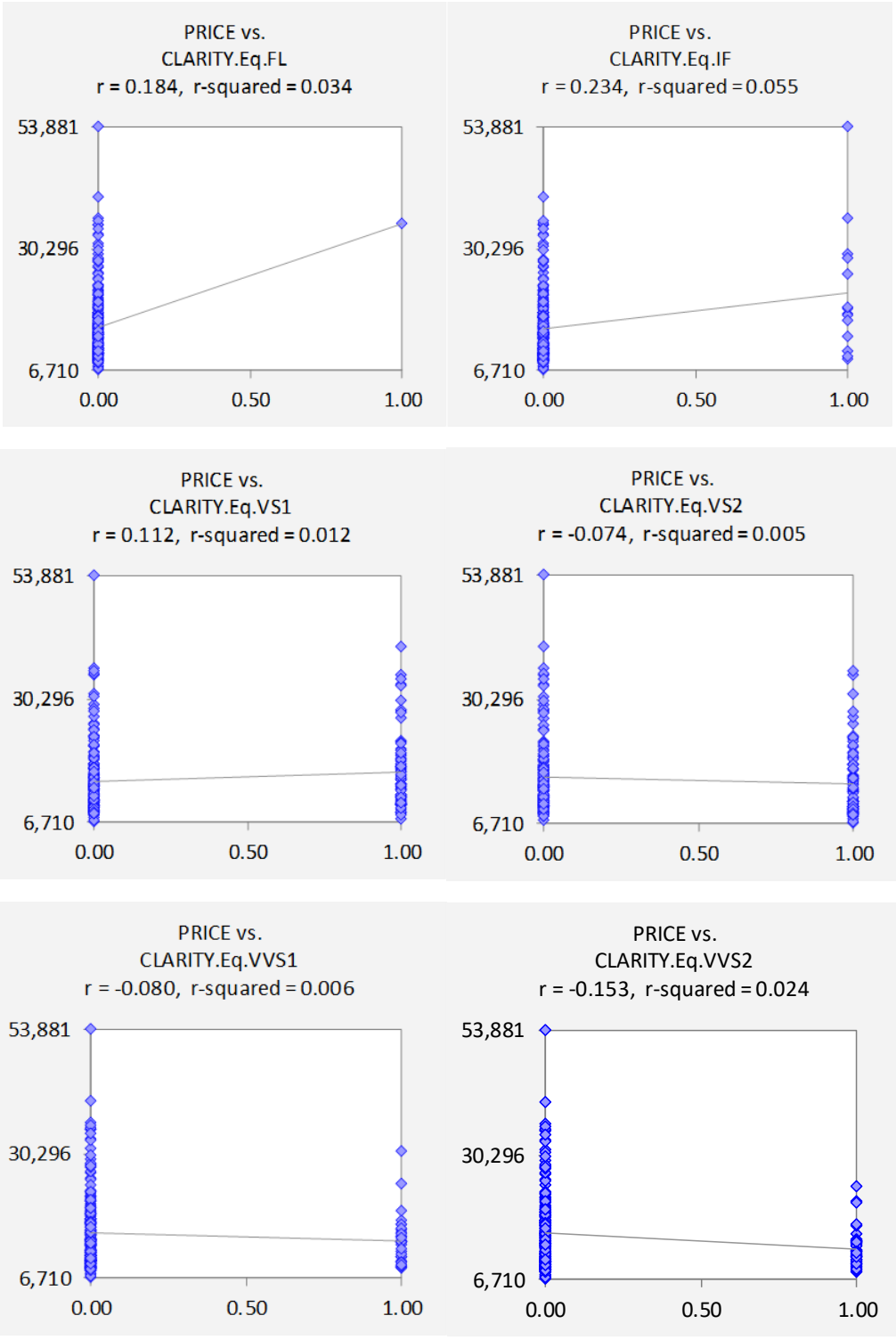
The highest correlation of price was with carat for this dataset (0.812). After carat, clarity and colour were important factors. However, carat had the highest correlation with price. Hence, it was important to see the graph of price vs carat to know the nature of the graph.

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



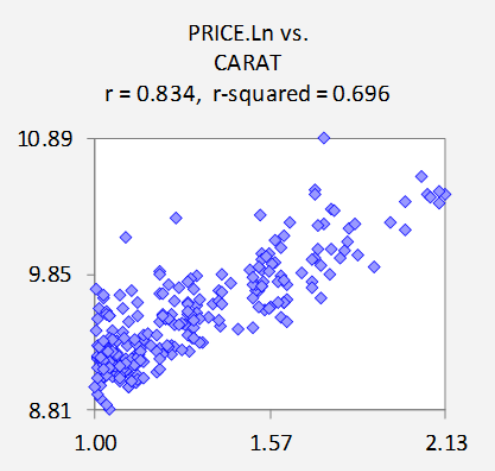
From the graphs of Descriptive stats, we can see that FL, IF have greater price as compared to VS1, VS2, VVS1, VVS2 proving the fact that FL, IF diamonds are premium quality with greater clarity which is affected on the price.

COLOR

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

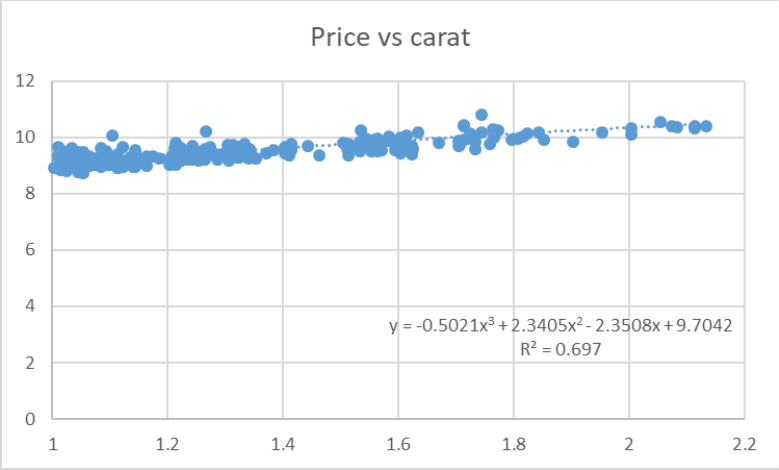
CARAT CORRELATION WITH PRICE

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.



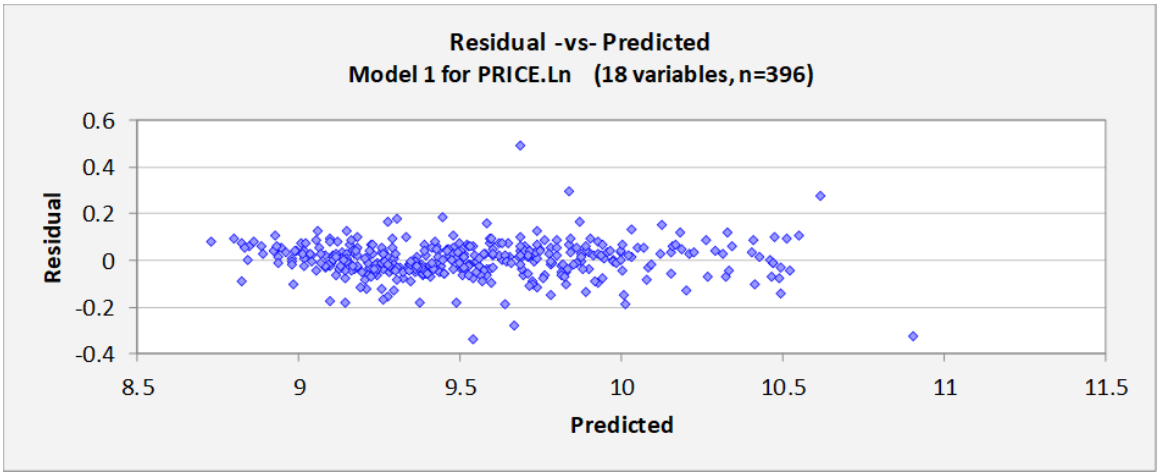
MODEL TRIALS- PART 3

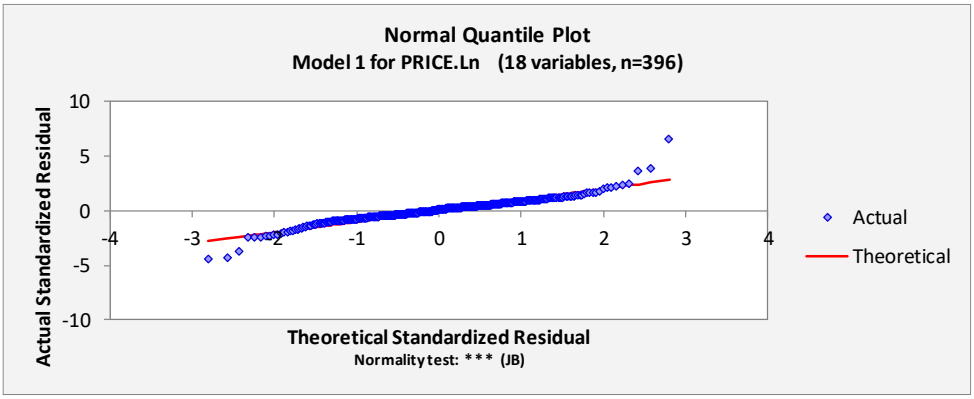
HeartXArrows and variables including HxA are the ones on which cut depends.

Model 1

Coefficient Estimates: Model 1 for PRICE.Ln (18 variables, n=396)

Variable	Coefficient	Std.Err.	t-statistic	P-value	Lower95%	Upper95%	VIF	Std. Coeff.
Constant	7.655	0.028	276.054	0.000	7.601	7.710	0.000	0.000
CARAT	1.291	0.015	88.009	0.000	1.263	1.320	1.163	0.915
CLARITY.Eq.FL	0.354	0.080	4.453	0.000	0.198	0.511	1.046	0.044
CLARITY.Eq.IF	0.214	0.021	10.373	0.000	0.173	0.254	1.397	0.118
CLARITY.Eq.VS1	-0.079	0.014	-5.775	0.000	-0.106	-0.052	2.530	-0.089
CLARITY.Eq.VS2	-0.146	0.014	-10.138	0.000	-0.174	-0.118	3.217	-0.175
CLARITY.Eq.VVS1	0.107	0.016	6.845	0.000	0.076	0.138	1.777	0.088
COLOR.Eq.D	0.511	0.013	37.929	0.000	0.484	0.537	1.422	0.436
COLOR.Eq.E	0.368	0.013	27.836	0.000	0.342	0.394	1.352	0.312
COLOR.Eq.F	0.255	0.012	20.390	0.000	0.230	0.279	1.401	0.233
COLOR.Eq.G	0.122	0.011	11.427	0.000	0.101	0.143	1.485	0.134
HeartsXArrows	-0.005037	0.011	-0.448	0.654	-0.027	0.017	1.385	-0.005082
HxA_True	0.046	0.011	3.976	0.000	0.023	0.068	2.151	0.056
TableClarity	-0.001703	0.010	-0.170	0.865	-0.021	0.018	1.641	-0.002100
Vendor.Eq.BlueNile	-0.002743	0.012	-0.230	0.818	-0.026	0.021	2.198	-0.003288
Vendor.Eq.BrianGavin	0.078	0.016	5.014	0.000	0.048	0.109	1.251	0.054
Vendor.Eq.CraftedByInfinity	0.125	0.014	9.275	0.000	0.099	0.152	1.518	0.110
Vendor.Eq.EnchantedDiamonds	-0.178	0.020	-9.096	0.000	-0.217	-0.140	1.485	-0.107
Vendor.Eq.JamesAllen	-0.129	0.020	-6.376	0.000	-0.168	-0.089	1.158	-0.066





Regression Statistics: Model 1 for Price.Ln (21 variables, n=1500)

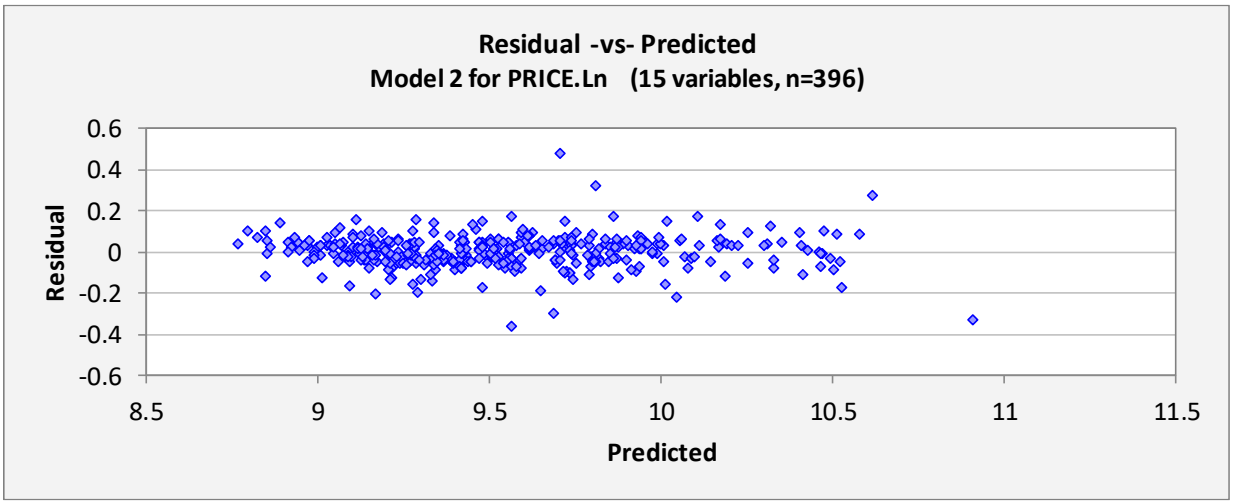
R-Squared	Adj.R-Sqr.	Std.Err.Reg.	Std. Dev.	# Fitted	# Missing	t(2.50%,1478)	Conf. level
0.972	0.971	0.120	0.705	1500	0	1.962	95.0%

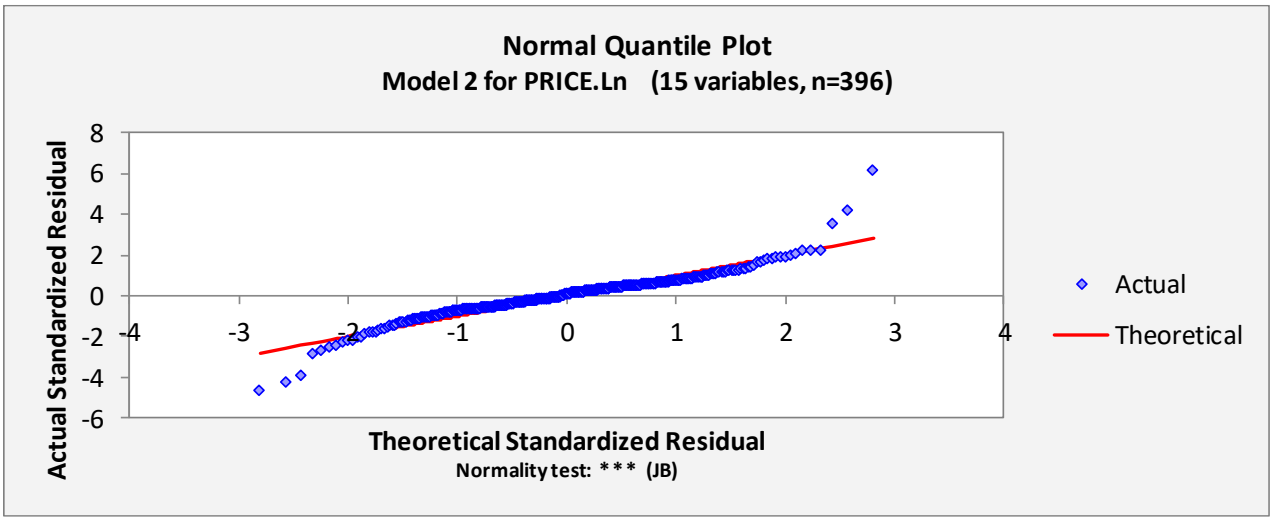
The model included all variables. However, certain variables with insignificant p-values were removed and the model was rerun. Moreover, the normal quantile plot shows deviation for very low and very high values of prices. The residual vs predicted plot shows variance in the errors. The errors increase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

Model 2

Coefficient Estimates: Model 2 for PRICE.Ln (15 variables, n=396)

Variable	Coefficient	Std.Err.	t-statistic	P-value	Lower95%	Upper95%	VIF	Std. Coeff.
Constant	7.668	0.024	323.200	0.000	7.622	7.715	0.000	0.000
CARAT	1.299	0.015	88.548	0.000	1.270	1.328	1.123	0.920
CLARITY.Eq.FL	0.344	0.081	4.263	0.000	0.185	0.502	1.038	0.043
CLARITY.Eq.IF	0.207	0.021	10.013	0.000	0.166	0.247	1.356	0.114
CLARITY.Eq.VS1	-0.075	0.014	-5.535	0.000	-0.102	-0.048	2.413	-0.084
CLARITY.Eq.VS2	-0.142	0.014	-10.509	0.000	-0.169	-0.116	2.746	-0.171
CLARITY.Eq.VVS1	0.104	0.016	6.666	0.000	0.073	0.135	1.701	0.085
COLOR.Eq.D	0.507	0.014	37.192	0.000	0.481	0.534	1.410	0.433
COLOR.Eq.E	0.362	0.013	27.526	0.000	0.337	0.388	1.293	0.307
COLOR.Eq.F	0.254	0.013	20.073	0.000	0.229	0.278	1.387	0.232
COLOR.Eq.G	0.123	0.011	11.357	0.000	0.102	0.144	1.482	0.136
Vendor.Eq.BlueNile	-0.024	0.011	-2.269	0.024	-0.045	-0.003186	1.651	-0.029
Vendor.Eq.BrianGavin	0.085	0.016	5.447	0.000	0.055	0.116	1.225	0.059
Vendor.Eq.CraftedByInfinity	0.135	0.013	10.252	0.000	0.109	0.161	1.391	0.119
Vendor.Eq.EnchantedDiamonds	-0.199	0.018	-10.921	0.000	-0.235	-0.163	1.245	-0.119
Vendor.Eq.JamesAllen	-0.119	0.020	-5.882	0.000	-0.159	-0.079	1.126	-0.061





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.

Regression Statistics: Model 7 for PRICE.Ln (15 variables, n=246)

R-Squared	Adj.R-Sqr.	Std.Err.Reg.	Std. Dev.	# Fitted	# Missing	t(2.50%,230)	Conf. level
0.977	0.975	0.062	0.391	246	200	1.970	95.0%

Coefficient Estimates: Model 7 for PRICE.Ln (15 variables, n=246)

Variable	Coefficient	Std.Err.	t-statistic	P-value	Lower95%	Upper95%	VIF	Std. Coeff.
Constant	8.262	0.035	236.928	0.000	8.194	8.331	0.000	0.000
CARAT.Power.3	-0.225	0.021	-10.737	0.000	-0.266	-0.183	94.506	-1.050
CARAT.Sqr	0.958	0.048	20.112	0.000	0.864	1.051	95.195	1.973
CLARITY.Eq.VS1	-0.086	0.013	-6.422	0.000	-0.113	-0.060	2.527	-0.103
CLARITY.Eq.VS2	-0.168	0.013	-12.604	0.000	-0.195	-0.142	2.592	-0.204
CLARITY.Eq.VVS1	0.099	0.015	6.662	0.000	0.070	0.128	1.659	0.086
COLOR.Eq.D	0.524	0.014	37.225	0.000	0.496	0.551	1.456	0.452
COLOR.Eq.E	0.360	0.013	27.261	0.000	0.334	0.386	1.382	0.322
COLOR.Eq.F	0.253	0.012	20.584	0.000	0.229	0.277	1.441	0.248
COLOR.Eq.G	0.120	0.011	11.020	0.000	0.098	0.141	1.591	0.140
Dummy_for_FL_IF	0.227	0.019	11.732	0.000	0.189	0.266	1.399	0.140
Vendor.Eq.BlueNile	-0.052	0.012	-4.350	0.000	-0.075	-0.028	2.230	-0.065
Vendor.Eq.BrianGavin	0.064	0.016	3.957	0.000	0.032	0.096	1.438	0.048
Vendor.Eq.CraftedByInfinity	0.128	0.014	9.315	0.000	0.101	0.155	1.804	0.126
Vendor.Eq.EnchantedDiamonds	-0.224	0.018	-12.208	0.000	-0.260	-0.188	1.481	-0.149
Vendor.Eq.JamesAllen	-0.117	0.020	-5.956	0.000	-0.155	-0.078	1.249	-0.067

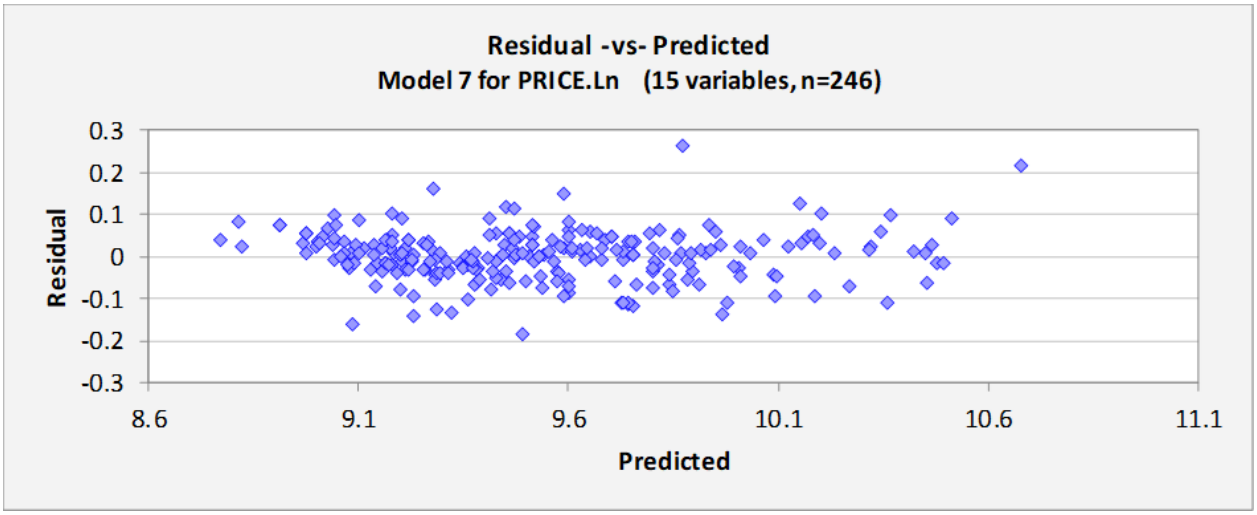
Analysis of Variance: Model 7 for PRICE.Ln (15 variables, n=246)

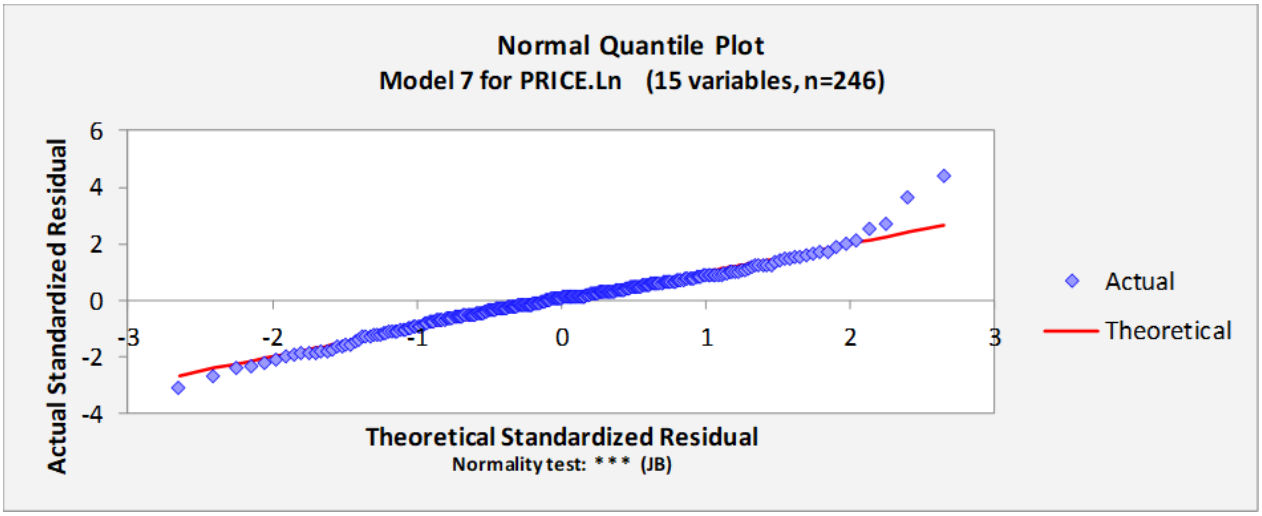
Source	Deg.Freedom	Sum Squares	Mean Square	F-statistic	P-value		Mean
Regression	15	36.557	2.437	644.048	0.000		9.537
Residual	230	0.870	0.003784				
Total	245	37.427					

Error Distribution Statistics: Model 7 for PRICE.Ln (15 variables, n=246)

	Mean Error	RMSE	MAE	Minimum	Maximum	MAPE	Normality
Fitted (n=246)	0.000	0.059	0.044	-0.185	0.261	0.5%	* * * (JB)

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.





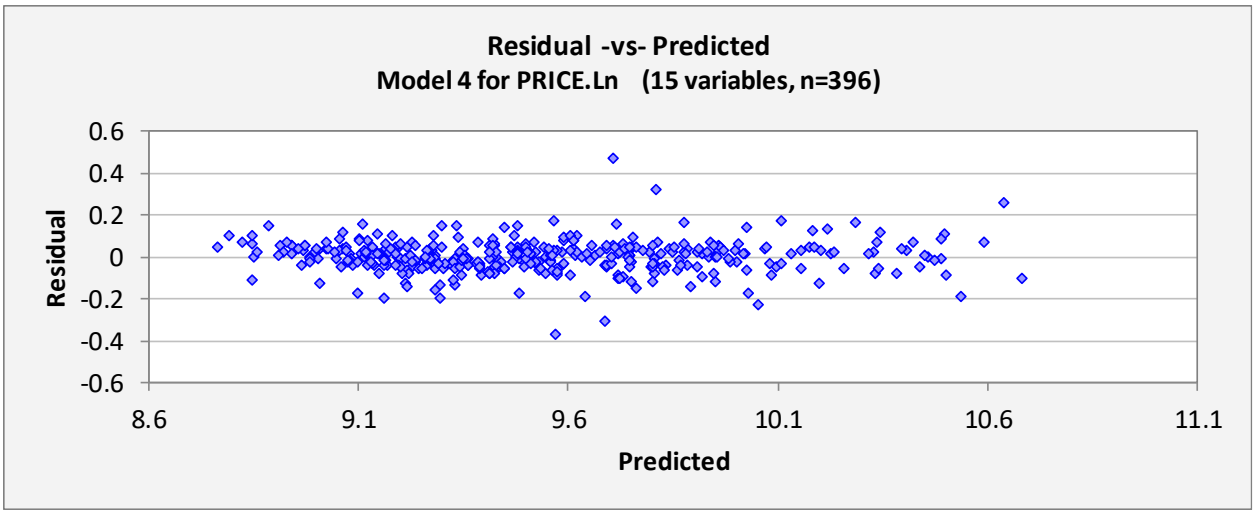
In the residual vs predicted plot, the variance remains almost constant over the entire x-axis. This shows that the errors are close to i.i.d. errors with almost constant variance. However, we can see a slight increase for higher price range values.

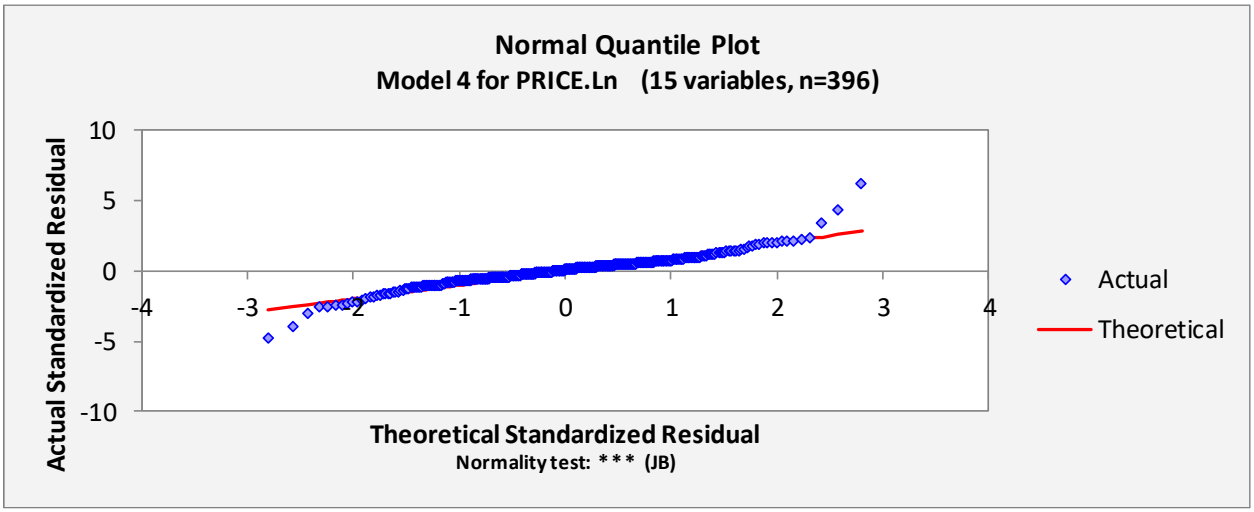
From the normal quantile plot, we can clearly see that the model misses forecast for very low values and misses by a huge extent for higher price range values. We tried to incorporate various factors including segregation of clarity and color based on their superiority, however the fit for higher price 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

Obs#	Forecast	StErrFcst	Lower50%F	Upper50%F	StErrMean	Lower50%M	Upper50%M
347	9.081	0.079	9.028	9.134	0.014	9.072	9.091
348	9.669	0.079	9.616	9.722	0.015	9.659	9.679
349	8.993	0.080	8.939	9.048	0.021	8.979	9.008
350	9.370	0.079	9.317	9.423	0.013	9.361	9.379
351	9.385	0.079	9.332	9.438	0.015	9.375	9.395
352	9.788	0.079	9.734	9.841	0.014	9.778	9.797
353	10.046	0.079	9.993	10.099	0.015	10.036	10.056
354	9.304	0.079	9.250	9.357	0.016	9.293	9.314
355	9.823	0.079	9.769	9.876	0.017	9.811	9.834
356	9.386	0.080	9.332	9.440	0.019	9.373	9.399
357	9.989	0.080	9.935	10.043	0.019	9.976	10.002
358	9.430	0.078	9.377	9.483	0.013	9.421	9.438
359	10.399	0.080	10.345	10.453	0.018	10.387	10.411
360	9.686	0.079	9.633	9.739	0.013	9.677	9.695
361	9.532	0.080	9.479	9.586	0.018	9.520	9.544
362	9.372	0.079	9.319	9.425	0.014	9.362	9.381
363	9.314	0.079	9.260	9.367	0.015	9.303	9.324
364	10.147	0.079	10.093	10.200	0.016	10.136	10.158
365	9.769	0.079	9.716	9.823	0.016	9.758	9.780
366	9.898	0.078	9.845	9.951	0.012	9.890	9.906
367	9.148	0.079	9.095	9.202	0.014	9.139	9.158
368	9.341	0.079	9.288	9.394	0.015	9.331	9.351
369	9.471	0.080	9.417	9.525	0.019	9.458	9.483
370	9.059	0.078	9.006	9.112	0.012	9.051	9.067
371	9.846	0.079	9.793	9.899	0.014	9.836	9.855
372	9.272	0.079	9.219	9.325	0.014	9.263	9.282
373	9.716	0.079	9.663	9.770	0.015	9.706	9.727
374	9.956	0.080	9.902	10.010	0.020	9.943	9.970
375	9.721	0.079	9.667	9.774	0.015	9.711	9.731
376	9.961	0.080	9.907	10.016	0.022	9.947	9.976
377	9.539	0.078	9.486	9.592	0.012	9.531	9.547
378	9.475	0.079	9.422	9.529	0.015	9.465	9.485
379	9.068	0.080	9.014	9.121	0.019	9.055	9.080
380	9.195	0.079	9.142	9.248	0.014	9.186	9.204
381	9.282	0.079	9.229	9.335	0.013	9.273	9.291
382	9.428	0.079	9.375	9.482	0.013	9.420	9.437
383	9.604	0.078	9.551	9.657	0.011	9.597	9.612
384	10.387	0.079	10.333	10.440	0.016	10.376	10.397
385	9.564	0.080	9.509	9.618	0.021	9.550	9.578
386	9.399	0.078	9.346	9.452	0.013	9.391	9.408
387	9.084	0.078	9.031	9.137	0.011	9.077	9.092
388	9.731	0.079	9.677	9.784	0.015	9.720	9.741
389	9.873	0.079	9.820	9.926	0.014	9.864	9.882
390	9.267	0.079	9.214	9.320	0.015	9.257	9.277
391	9.385	0.079	9.332	9.438	0.014	9.376	9.394
392	8.996	0.079	8.943	9.049	0.014	8.986	9.005
393	10.796	0.092	10.733	10.858	0.050	10.762	10.829
394	9.444	0.080	9.390	9.498	0.018	9.432	9.456
395	9.740	0.079	9.686	9.793	0.015	9.729	9.750
396	9.071	0.078	9.018	9.124	0.011	9.063	9.078





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

Logistic Regression Statistics: Model 1 for Vendor.Eq.EnchantedDiamonds (17 variables, n=346)

R-squared (McFadden)	Adj.R-Sqr.	RMSE	Mean	# Fitted	# Missing	Critical z	Conf. level
0.911	0.711	0.082	0.072	346	1	1.960	95.0%

Logistic Regression Coefficient Estimates: Model 1 for Vendor.Eq.EnchantedDiamonds (17 variables, n=346)

Variable	Coefficient	Std.Err.	z-statistic	P-value	Lower95%	Upper95%	VIF	Std. coeff.
Constant	-165.773	3,524	-0.047	0.962	-7,072	6,741		
CARAT	244.230	170.738	1.430	0.153	-90.409	578.870	7.887	38.411
CLARITY.Eq.FL	137.256	5,441	0.025	0.980	-10,526	10,801	1.112	4.068
CLARITY.Eq.IF	54.283	38.899	1.395	0.163	-21.958	130.524	1.461	6.104
CLARITY.Eq.VS1	-31.728	22.340	-1.420	0.156	-75.513	12.058	2.433	-8.096
CLARITY.Eq.VS2	-38.541	26.885	-1.434	0.152	-91.235	14.153	2.948	-10.273
CLARITY.Eq.VVS1	-3.463	3.553	-0.975	0.330	-10.427	3.501	1.811	-0.643
COLOR.Eq.D	90.862	61.604	1.475	0.140	-29.881	211.604	2.412	17.188
COLOR.Eq.E	43.041	29.371	1.465	0.143	-14.525	100.607	1.882	7.993
COLOR.Eq.F	58.704	40.199	1.460	0.144	-20.085	137.494	1.667	12.190
COLOR.Eq.G	10.123	6.909	1.465	0.143	-3.419	23.665	1.611	2.518
HeartsXArrows	-29.668	19.644	-1.510	0.131	-68.168	8.833	1.397	-6.615
HxA_CrownAngle_34to3	-2.817	3.304	-0.853	0.394	-9.293	3.658	1.716	-0.768
HxA_LowerGirdle_76to7	-45.577	3,523	-0.013	0.990	-6,950	6,859	1.096	-1.908
HxA_PavillionAngle_406	10.244	9.229	1.110	0.267	-7.845	28.333	1.174	1.938
HxA_StarFacets_45to5(	60.232	43.373	1.389	0.165	-24.778	145.242	1.215	16.439
HxA_TableSize_54to57	-68.074	48.899	-1.392	0.164	-163.914	27.766	1.052	-10.415
PRICE	-0.009512	0.006565	-1.449	0.147	-0.022	0.003355	8.299	-36.690

Analysis of Deviance: Model 1 for Vendor.Eq.EnchantedDiamonds (17 variables, n=346)

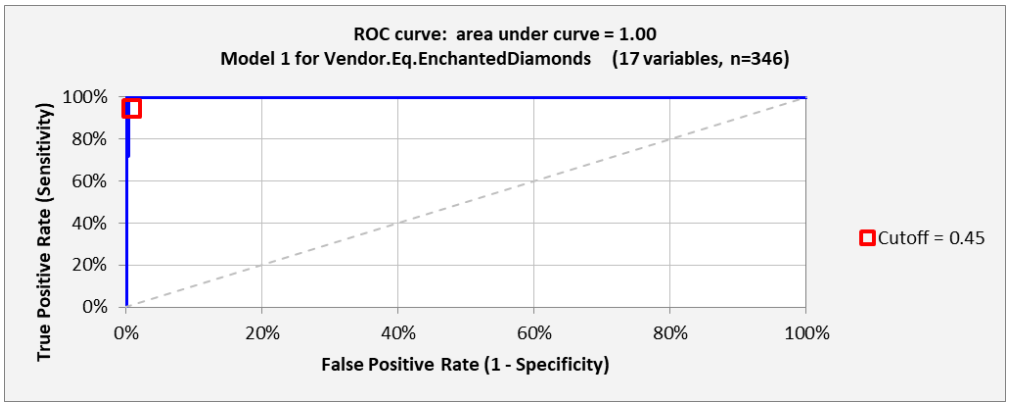
Correlation Matrix of Coefficient Estimates : Model 1 for Vendor.Eq.EnchantedDiamonds (17 variables, n=346)

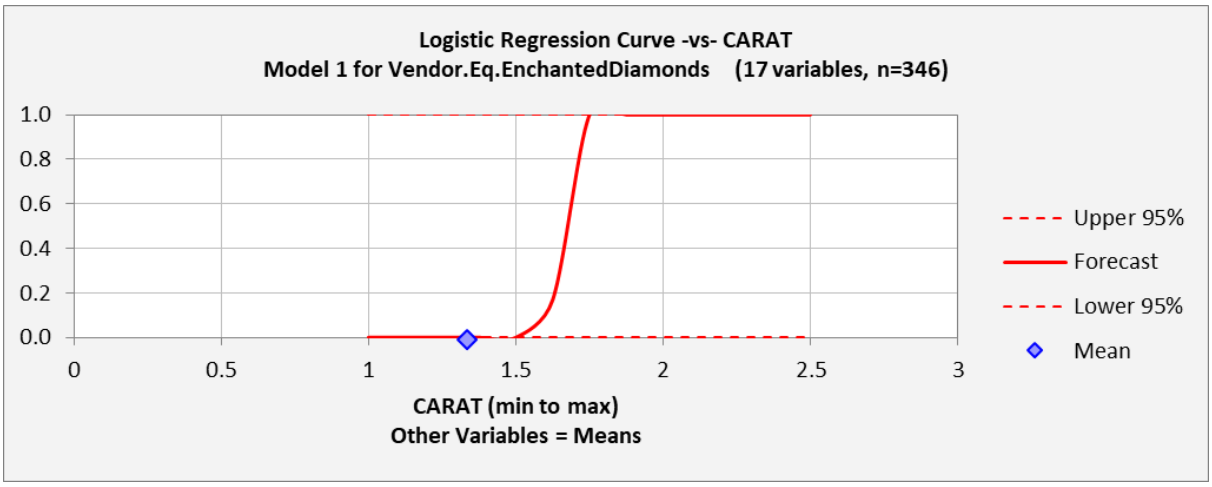
Classification Table: Model 1 for Vendor.Eq.EnchantedDiamonds (17 variables, n=346)

Cutoff value for prediction of 1: 0.45 RMSE = 0.082

Actual:	Predicted: # 0	Predicted: # 1	Total
# 0	320	**1**	321
# 1	1	24	25
Total	321	25	346
Percent correct =	99.4%	True positive rate =	96.0%

Actual:	Predicted: % 0	Predicted: % 1	Total
% 0	92%	0%	93%
% 1	0%	7%	7%
Total	93%	7%	100%
True negative rate =	99.7%		



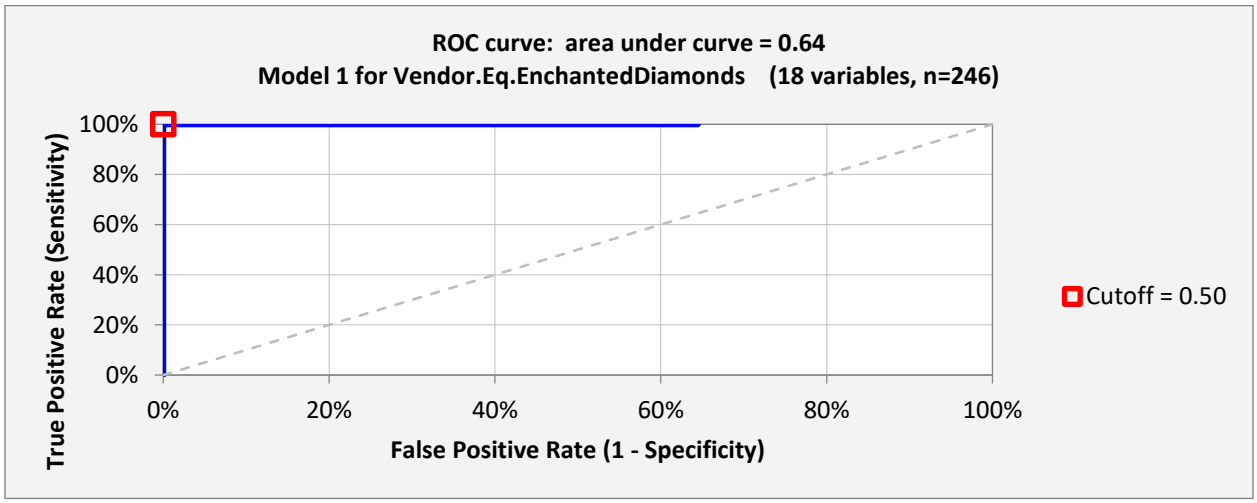


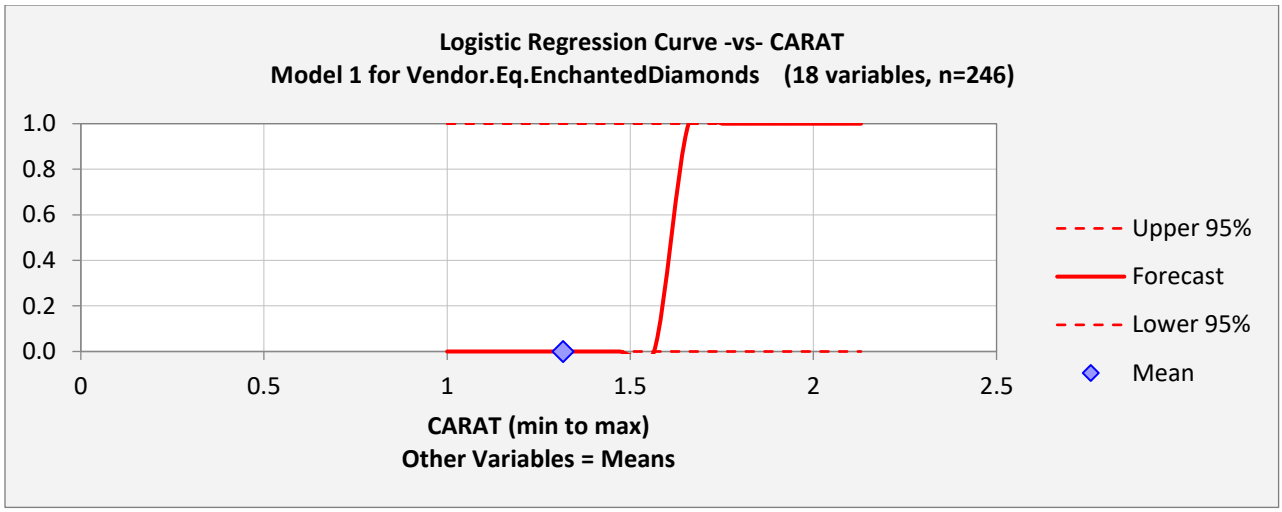
This model has a very good r-squared and adj-r-squared. However, none of the parameters are significant with p-value less than 0.05. Hence, even if the correlation is good, the model does not perform well. Moreover, the ROC curve has an area under the curve of 1 which seems unrealistic. There is a very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





Logistic Regression Statistics: Model 1 for Vendor.Eq.EnchantedDiamonds (18 variables, n=246)

R-squared (McFadden)	Adj.R-Sqr.	RMSE	Mean	# Fitted	# Missing	Critical z	Conf. level
1.000	0.705	0.000	0.073	246	1	1.960	95.0%

Logistic Regression Coefficient Estimates: Model 1 for Vendor.Eq.EnchantedDiamonds (18 variables, n=246)

Variable	Coefficient	Std.Err.	z-statistic	P-value	Lower95%	Upper95%	VIF	Std. coeff.
Constant	-1,883	1,795	-1.049	0.294	-5,401	1,634		
CARAT	1,704	1,108	1.537	0.124	-468.281	3,877	7.786	260.625
CLARITY.Eq.FL	928.665	2,095	0.443	0.658	-3,177	5,035	1.162	32.644
CLARITY.Eq.IF	359.812	241.389	1.491	0.136	-113.302	832.925	1.592	46.051
CLARITY.Eq.VS1	-311.267	205.839	-1.512	0.130	-714.705	92.171	2.410	-79.741
CLARITY.Eq.VS2	-395.850	260.729	-1.518	0.129	-906.868	115.169	2.716	-103.401
CLARITY.Eq.VVS1	-83.126	62.432	-1.331	0.183	-205.491	39.238	1.754	-15.651
COLOR.Eq.D	557.173	362.056	1.539	0.124	-152.443	1,267	2.589	103.546
COLOR.Eq.E	196.145	5,397	0.036	0.971	-10,382	10,774	2.105	37.854
COLOR.Eq.F	334.577	220.055	1.520	0.128	-96.723	765.877	1.827	70.837
COLOR.Eq.G	100.430	69.519	1.445	0.149	-35.826	236.686	1.758	25.244
HeartsXArrows	-142.306	97.399	-1.461	0.144	-333.204	48.592	1.436	-30.907
HxA_CrownAngle_34to3	19.153	38.071	0.503	0.615	-55.465	93.772	1.780	5.213
HxA_LowerGirdle_76to7	-672.485	1,383	-0.486	0.627	-3,382	2,037	1.149	-33.362
HxA_PavillionAngle_406	196.514	134.200	1.464	0.143	-66.514	459.541	1.234	37.000
HxA_StarFacets_45to5(	487.797	320.912	1.520	0.129	-141.179	1,117	1.220	133.737
HxA_TableSize_54to57	-527.396	344.926	-1.529	0.126	-1,203	148.645	1.051	-75.875
Standardized_Price	-405.165	264.526	-1.532	0.126	-923.628	113.297	8.142	-223.379
Super_Ideal_Diamonds	42.297	5,398	0.008	0.994	-10,538	10,623	1.284	9.034

Analysis of Deviance: Model 1 for Vendor.Eq.EnchantedDiamonds (18 variables, n=246)

Correlation Matrix of Coefficient Estimates : Model 1 for Vendor.Eq.EnchantedDiamonds (18 variables, n=246)

Classification Table: Model 1 for Vendor.Eq.EnchantedDiamonds (18 variables, n=246)

Cutoff value for prediction of 1: 0.50 RMSE = 0.000

Actual:	Predicted: # 0	Predicted: # 1	Total
# 0	228	**0**	228
# 1	0	18	18
Total	228	18	246
Percent correct =	100.0%	True positive rate =	100.0%

Actual:	Predicted: % 0	Predicted: % 1	Total
% 0	93%	0%	93%
% 1	0%	7%	7%
Total	93%	7%	100%
True negative rate =	100.0%		

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

Logistic Regression Coefficient Estimates: Model 3 for Vendor.Eq.EnchantedDiamonds (14 variables, n=246)

Variable	Coefficient	Std.Err.	z-statistic	P-value	Lower95%	Upper95%	VIF	Std. coeff.
Constant	-134.491	94.562	-1.422	0.155	-319.830	50.847		
CARAT	90.557	61.570	1.471	0.141	-30.118	211.232	5.364	13.850
CLARITY.Eq.IF	17.875	13.265	1.348	0.178	-8.124	43.873	1.483	2.288
CLARITY.Eq.VS1	-13.243	9.489	-1.396	0.163	-31.841	5.355	2.378	-3.393
CLARITY.Eq.VS2	-19.845	13.371	-1.484	0.138	-46.052	6.362	2.657	-5.184
CLARITY.Eq.VVS1	-5.471	4.296	-1.274	0.203	-13.891	2.949	1.665	-1.030
COLOR.Eq.D	29.364	19.207	1.529	0.126	-8.282	67.009	1.677	5.457
COLOR.Eq.F	14.274	9.773	1.461	0.144	-4.881	33.428	1.275	3.022
COLOR.Eq.G	3.711	3.519	1.054	0.292	-3.187	10.608	1.303	0.933
HeartsXArrows	-8.081	4.749	-1.701	0.089	-17.389	1.228	1.384	-1.755
HxA_PavillionAngle_406	11.724	9.731	1.205	0.228	-7.348	30.796	1.270	2.207
HxA_StarFacets_45to5(	28.043	19.694	1.424	0.154	-10.556	66.642	1.169	7.688
HxA_TableSize_54to57	-31.407	20.377	-1.541	0.123	-71.345	8.530	1.191	-4.519
HxA_True	-0.426	3.900	-0.109	0.913	-8.069	7.218	1.927	-0.118
Standardized_Price	-20.638	13.886	-1.486	0.137	-47.854	6.578	5.286	-11.378

Analysis of Deviance: Model 3 for Vendor.Eq.EnchantedDiamonds (14 variables, n=246)

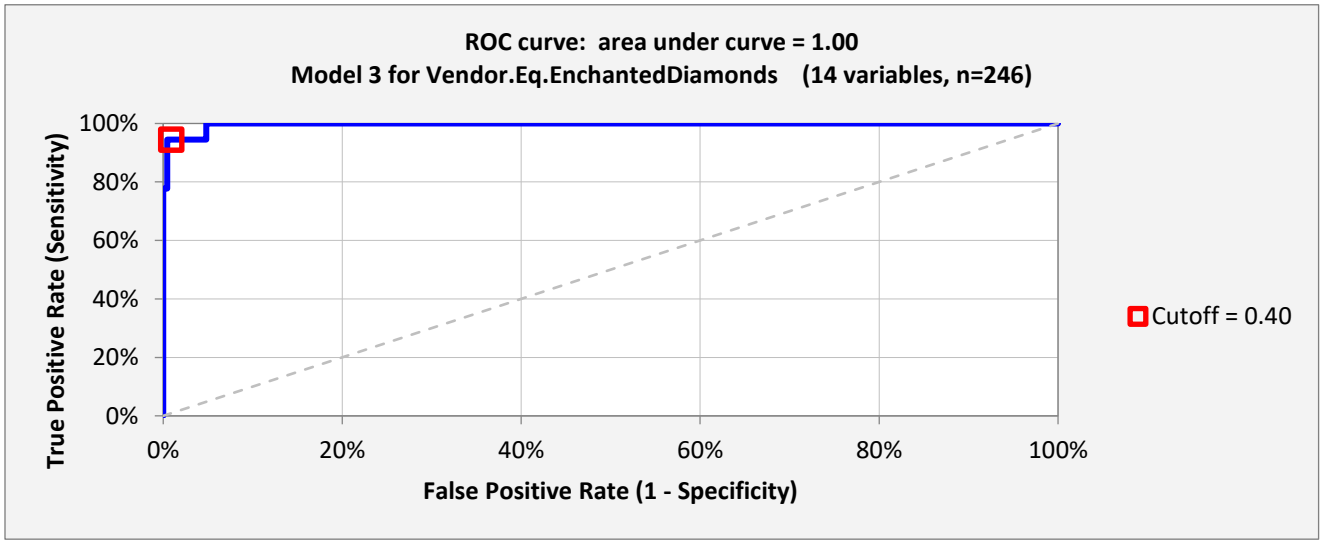
Correlation Matrix of Coefficient Estimates : Model 3 for Vendor.Eq.EnchantedDiamonds (14 variables, n=246)

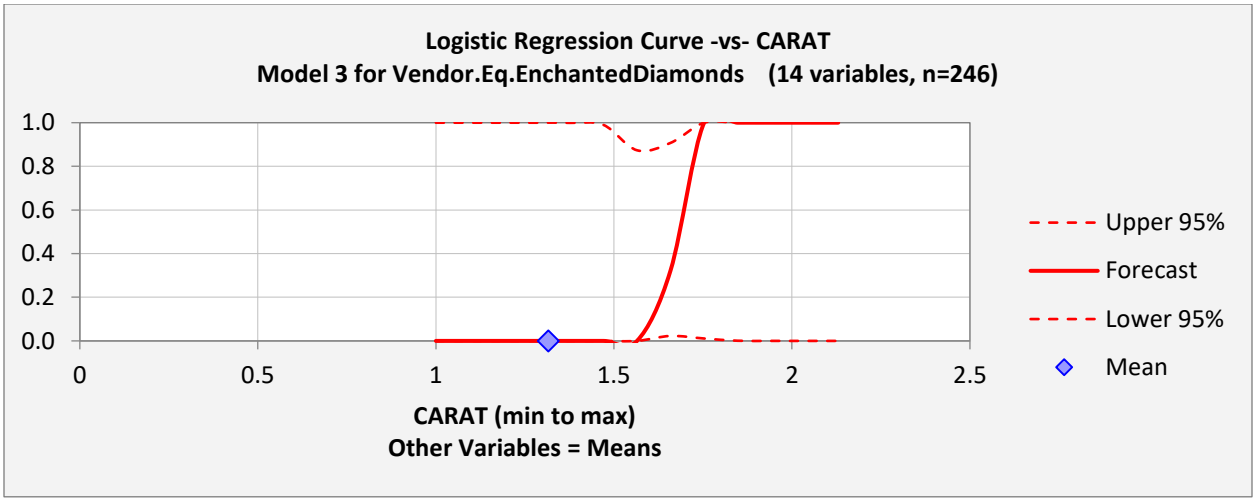
Classification Table: Model 3 for Vendor.Eq.EnchantedDiamonds (14 variables, n=246)

Cutoff value for prediction of 1: 0.40 RMSE = 0.110

Actual:	Predicted: # 0	Predicted: # 1	Total
# 0	226	**2**	228
# 1	1	17	18
Total	227	19	246
Percent correct =	98.8%	True positive rate =	94.4%

Actual:	Predicted: % 0	Predicted: % 1	Total
% 0	92%	1%	93%
% 1	0%	7%	7%
Total	92%	8%	100%
True negative rate =	99.1%		





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

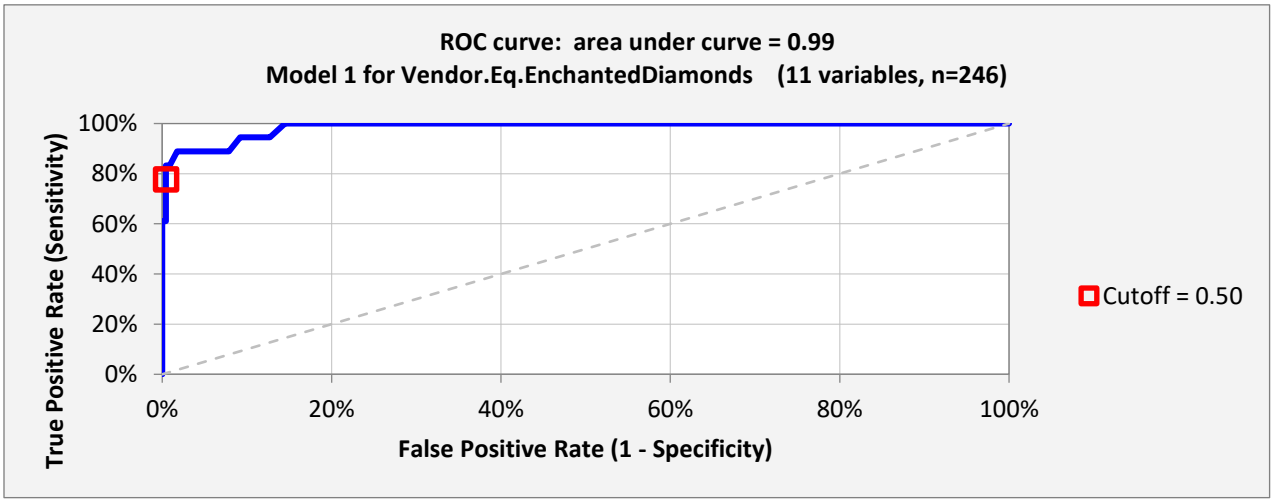
By further removing the dependent variables, the model was performing worse, hence a different approach was to be tried.

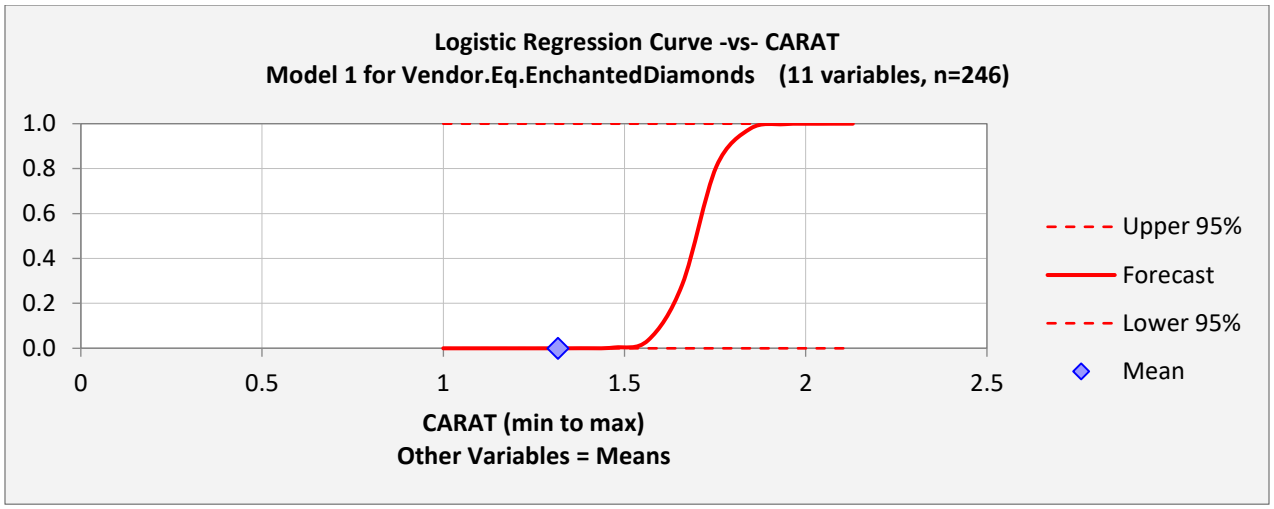
Interesting Fact

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.





Logistic Regression Statistics: Model 1 for Vendor.Eq.EnchantedDiamonds (11 variables, n=246)

R-squared (McFadden)	Adj.R-Sqr.	RMSE	Mean	# Fitted	# Missing	Critical z	Conf. level
0.748	0.562	0.134	0.073	246	100	1.960	95.0%

Logistic Regression Coefficient Estimates: Model 1 for Vendor.Eq.EnchantedDiamonds (11 variables, n=246)

Variable	Coefficient	Std.Err.	z-statistic	P-value	Lower95%	Upper95%	VIF	Std. coeff.
Constant	-46.674	5,927	-0.008	0.994	-11,664	11,571		
CARAT	25.174	8.755	2.875	0.004	8.014	42.334	5.410	3.850
Clarity_without_FL_IF	-7.261	2.411	-3.012	0.003	-11.987	-2.536	1.289	-0.960
D	9.746	3.488	2.795	0.005	2.911	16.582	1.795	1.811
E_F	2.824	1.930	1.463	0.143	-0.959	6.607	1.401	0.728
HeartsXArrows	-4.186	1.583	-2.645	0.008	-7.288	-1.084	1.400	-0.909
HxA_CrownAngle_34to3	-1.934	1.296	-1.492	0.136	-4.475	0.607	1.502	-0.526
HxA_LowerGirdle_76to7	15.747	5,927	0.003	0.998	-11,602	11,633	1.082	0.781
HxA_PavillionAngle_406	1.742	1.613	1.081	0.280	-1.418	4.903	1.187	0.328
HxA_StarFacets_45to5(	8.782	2.744	3.201	0.001	3.404	14.159	1.153	2.408
HxA_TableSize_54to57	-8.463	2.486	-3.404	0.001	-13.336	-3.590	1.031	-1.218
Standardized_Price	-7.164	2.614	-2.741	0.006	-12.287	-2.041	5.827	-3.950

Analysis of Deviance: Model 1 for Vendor.Eq.EnchantedDiamonds (11 variables, n=246)

Correlation Matrix of Coefficient Estimates : Model 1 for Vendor.Eq.EnchantedDiamonds (11 variables, n=246)

Classification Table: Model 1 for Vendor.Eq.EnchantedDiamonds (11 variables, n=246)

Cutoff value for prediction of 1: 0.50 RMSE = 0.134

Actual:	Predicted: # 0	Predicted: # 1	Total
# 0	227	1	228
# 1	4	14	18
Total	231	15	246
Percent correct =	98.0%	True positive rate =	77.8%

Actual:	Predicted: % 0	Predicted: % 1	Total
% 0	92%	0%	93%
% 1	2%	6%	7%
Total	94%	6%	100%
True negative rate =	99.6%		

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.

Logistic Regression Statistics: Model 6 for Vendor.Eq.EnchantedDiamonds (8 variables, n=346)

R-squared (McFadden)	Adj.R-Sqr.	RMSE	Mean	# Fitted	# Missing	Critical z	Conf. level
0.780	0.680	0.125	0.072	346	0	1.960	95.0%

Logistic Regression Coefficient Estimates: Model 6 for Vendor.Eq.EnchantedDiamonds (8 variables, n=346)

Variable	Coefficient	Std.Err.	z-statistic	P-value	Lower95%	Upper95%	VIF	Std. coeff.
Constant	155.725	47.254	3.295	0.001	63.109	248.342		
CARAT	27.590	8.201	3.364	0.001	11.516	43.663	8.059	4.339
Clarity_without_FL_IF	-6.582	1.917	-3.434	0.001	-10.339	-2.826	1.154	-0.763
D	9.854	2.949	3.342	0.001	4.074	15.633	2.153	1.864
E_F	4.563	2.156	2.116	0.034	0.337	8.788	1.651	1.155
HeartsXArrows	-4.458	1.314	-3.393	0.001	-7.033	-1.883	1.034	-0.994
HxA_StarFacets_45to5(	6.368	1.681	3.789	0.000	3.074	9.661	1.038	1.738
HxA_TableSize_54to57	-6.591	1.835	-3.592	0.000	-10.188	-2.995	1.034	-1.008
PRICE.Ln	-19.917	6.052	-3.291	0.001	-31.778	-8.056	8.747	-4.380

Analysis of Deviance: Model 6 for Vendor.Eq.EnchantedDiamonds (8 variables, n=346)

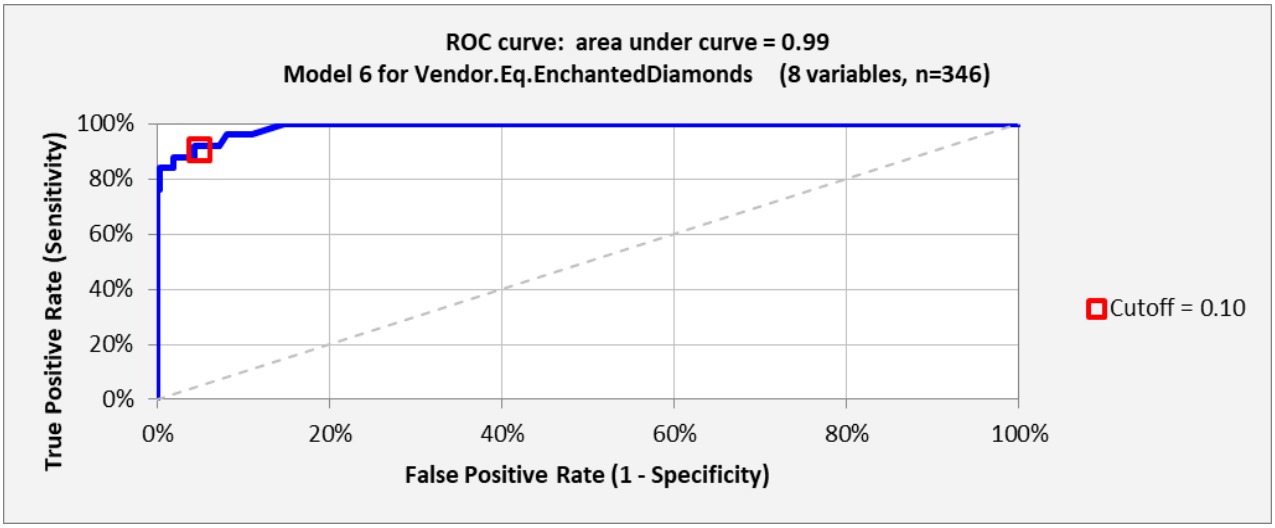
Correlation Matrix of Coefficient Estimates : Model 6 for Vendor.Eq.EnchantedDiamonds (8 variables, n=346)

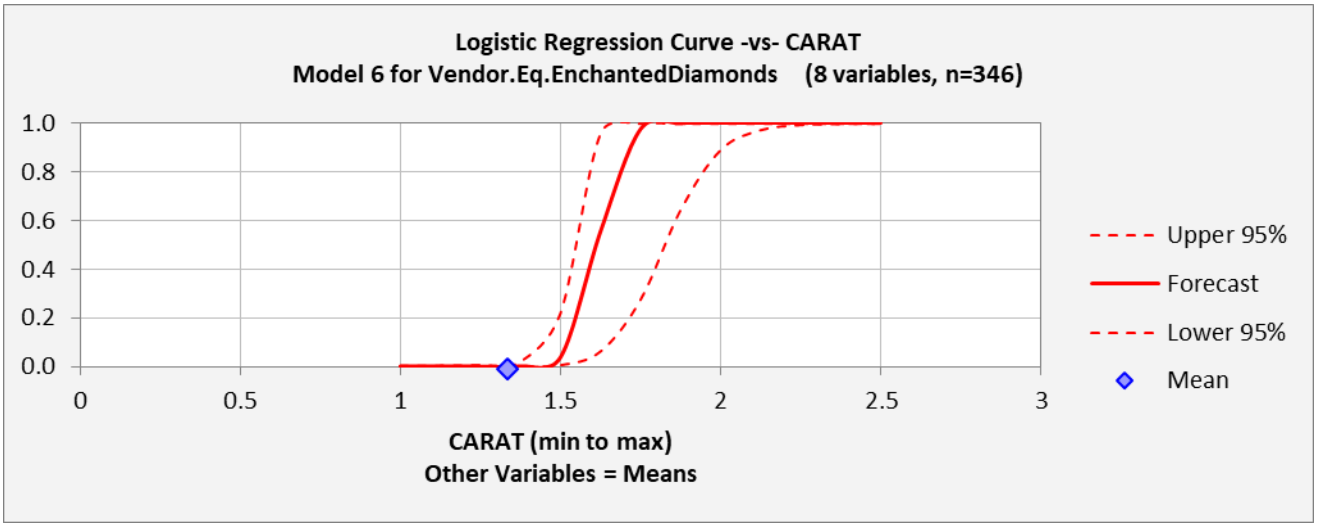
Classification Table: Model 6 for Vendor.Eq.EnchantedDiamonds (8 variables, n=346)

Cutoff value for prediction of 1: 0.10 RMSE = 0.125

Actual:	Predicted: # 0	Predicted: # 1	Total
# 0	307	**14**	321
# 1	2	23	25
Total	309	37	346
Percent correct =	95.4%	True positive rate =	92.0%

Actual:	Predicted: % 0	Predicted: % 1	Total
% 0	89%	4%	93%
% 1	1%	7%	7%
Total	89%	11%	100%
True negative rate =	95.6%		





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.

R-squared (McFadden)	Adj.R-Sqr.	RMSE	Mean	# Fitted	# Missing	Critical z	Conf. level
0.700	0.616	0.147	0.076	396	50	1.960	95.0%

Logistic Regression Coefficient Estimates: Model 1 for Vendor.Eq.EnchantedDiamonds (8 variables, n=396)

Variable	Coefficient	Std.Err.	z-statistic	P-value	Lower95%	Upper95%	VIF	Std. coeff.
Constant	-10.529	3.309	-3.182	0.001	-17.015	-4.044		
CARAT	17.535	4.344	4.037	0.000	9.021	26.048	6.030	2.785
Clarity_without_FL_IF	-5.507	1.365	-4.035	0.000	-8.181	-2.832	1.174	-0.616
D	6.022	1.478	4.075	0.000	3.125	8.918	1.749	1.184
E_F	2.022	1.327	1.524	0.128	-0.579	4.623	1.374	0.508
HeartsXArrows	-3.352	0.848	-3.954	0.000	-5.013	-1.690	1.045	-0.750
HxA_StarFacets_45to5(	5.841	1.358	4.301	0.000	3.180	8.503	1.033	1.592
HxA_TableSize_54to57	-5.526	1.231	-4.490	0.000	-7.938	-3.114	1.033	-0.866
PRICE	-0.000709	0.000187	-3.803	0.000	-0.001075	-0.000344	6.368	-2.754

Analysis of Deviance: Model 1 for Vendor.Eq.EnchantedDiamonds (8 variables, n=396)

Correlation Matrix of Coefficient Estimates : Model 1 for Vendor.Eq.EnchantedDiamonds (8 variables, n=396)

Classification Table: Model 1 for Vendor.Eq.EnchantedDiamonds (8 variables, n=396)

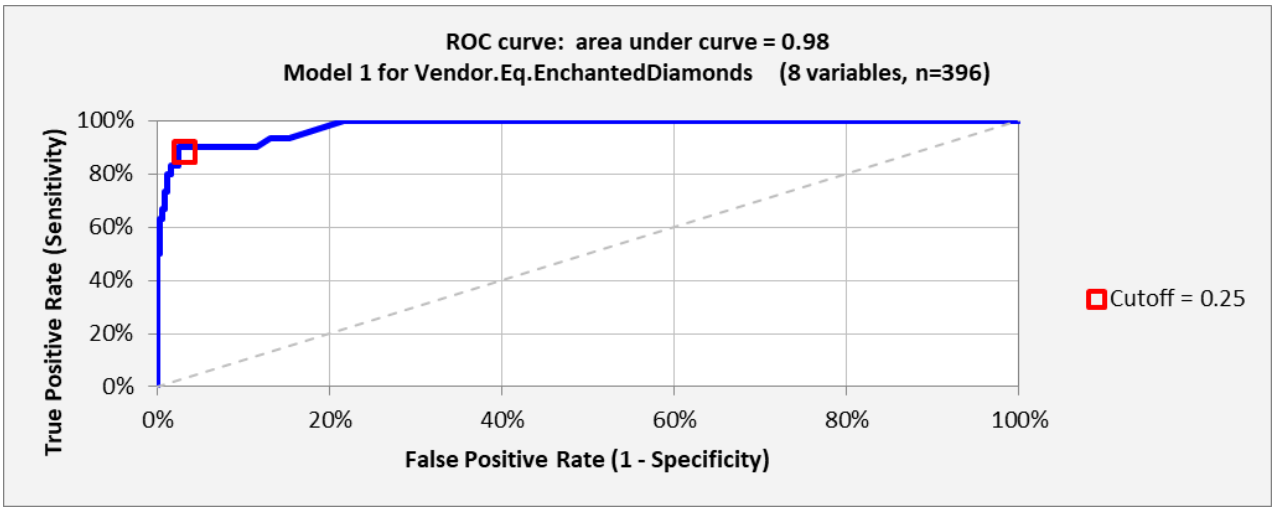
Cutoff value for prediction of 1: 0.25 RMSE = 0.147

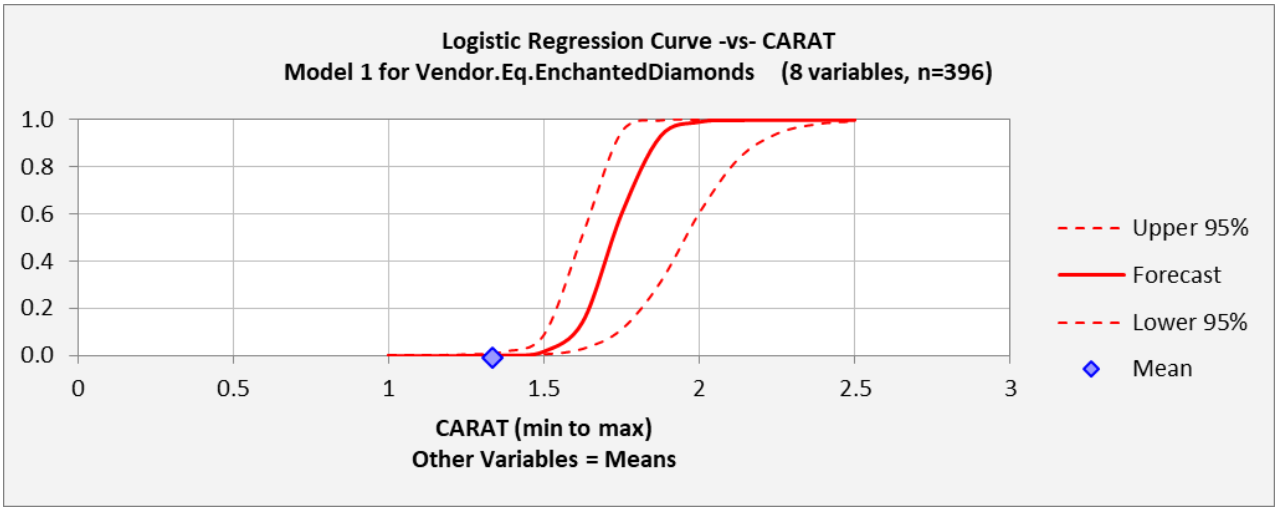
Actual:	Predicted: # 0	# 1	Total
# 0	356	**10**	366
# 1	3	27	30
Total	359	37	396

Percent correct = 96.7% True positive rate = 90.0%

Actual:	Predicted: % 0	% 1	Total
% 0	90%	3%	92%
% 1	1%	7%	8%
Total	91%	9%	100%

True negative rate = 97.3%





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.

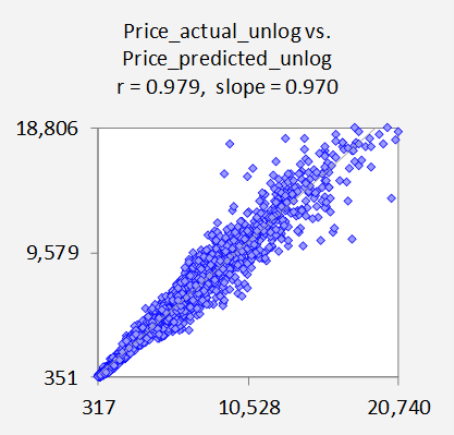
Question 4:

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.

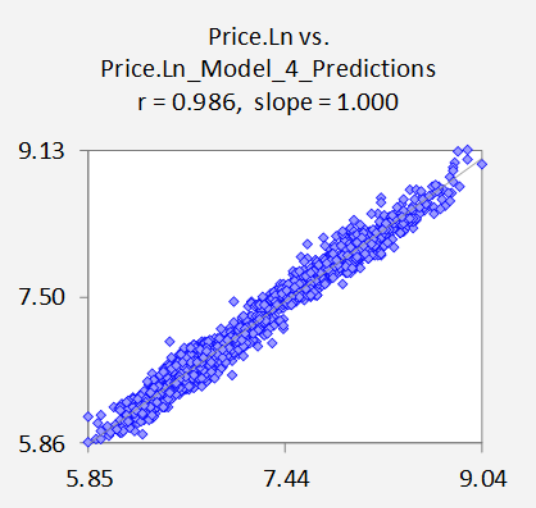
However, in the residual vs predicted plot for large values, the errors vary downwards and hence we can say that the model is overpredicting the data. The larger model also deviates from normality at smaller range values of carat.

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.

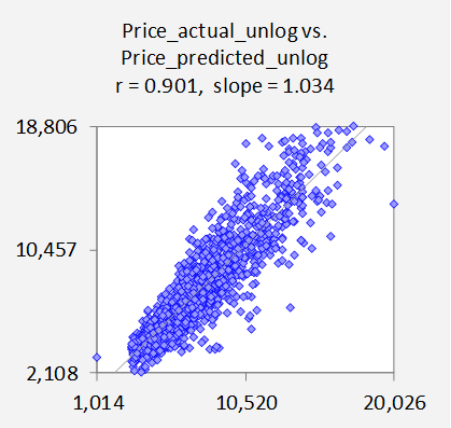
Question 5:



For the entire data set



For small diamonds



For large diamonds