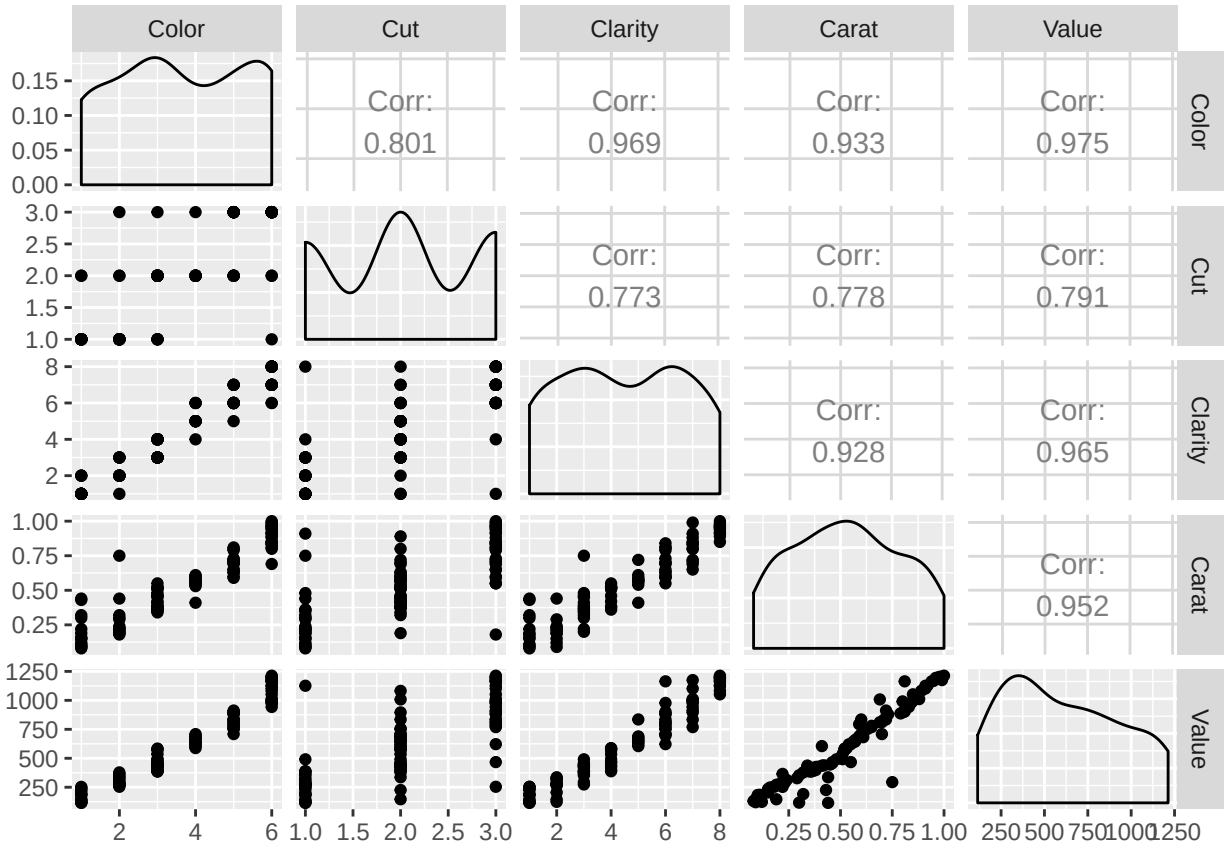
Exam 2

Cody Frisby

11/13/2017

1.

Below is a plot of the `Diamonds` dataset. As can be seen, **Carat** is highly correlated with **Value** among other variables.



Model table is below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-64.942549	17.19645	-3.7765086	0.0002786
Color	96.194119	16.03415	5.9993277	0.0000000
Cut	1.702045	13.30199	0.1279542	0.8984586
Clarity	29.811706	11.45695	2.6020637	0.0107661
Carat	358.430314	68.85559	5.2055366	0.0000011

A.

Cut is not a significant predictor of **Value** and should NOT be included in the model.

B.

Color is the most significant predictor in the model ($p < 0.000001$).

C.

The model explains 96.7040882% of the variation in **Value**. This is the R^2 value of the model. We can interpret it as the amount of variation in Y that is explained by X.

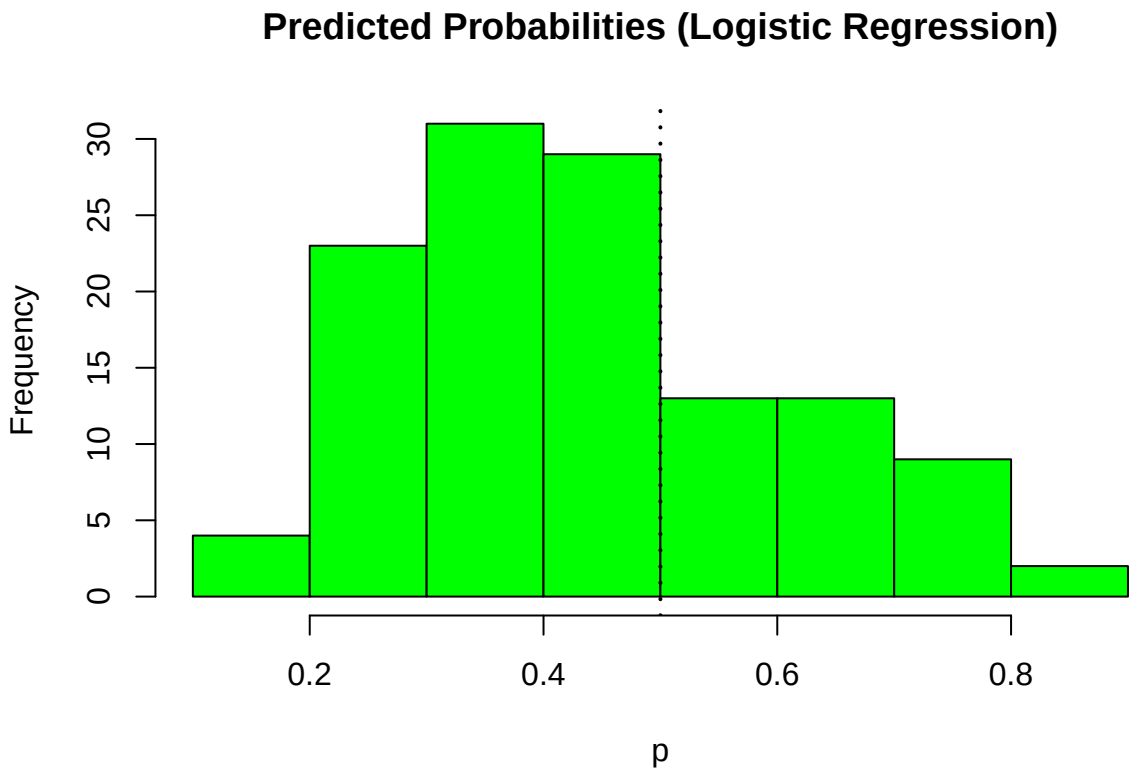
D. Predictions

The preditions for the table of values is below.

Color	Cut	Clarity	Carat	predictedValue
4	3	3	0.65	647.3549
5	1	8	0.55	853.3604
7	3	7	0.94	1159.1289

2.

Below is a histogram of our predicted probabilities for the "**InvestmentMarketing**" dataset. I've drawn a vertical line at 0.5 which can be thought of as our cutoff. All the values to the right of the line would be our "yes" class.



A.

Counting all the values that meet the criteria ($0.5 \geq$) the total number is 37.

B.

The largest predicted value is 0.8450046. This means, according to our model, that there is a 0.8450046 probability that they will use investment services.

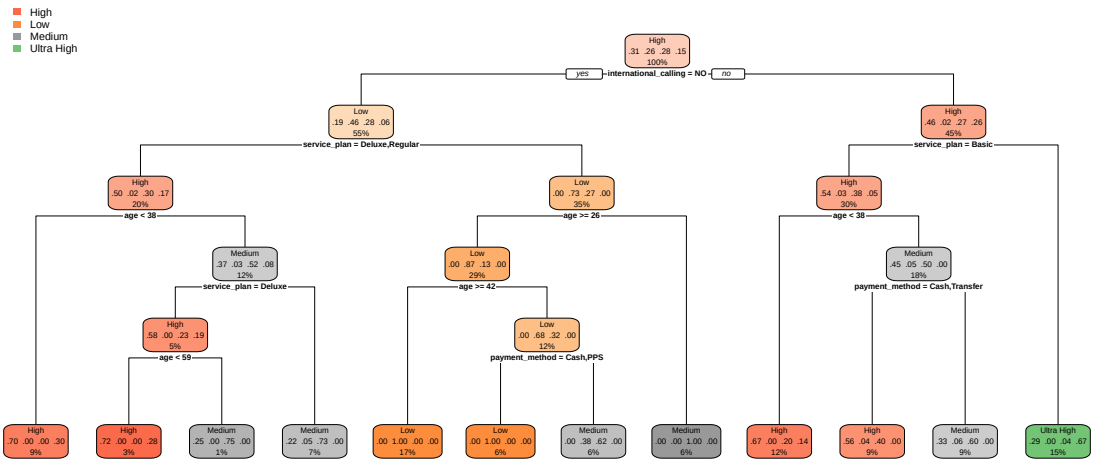
C.

The table below shows the coefficients of the model with their probabilities. As can be seen, the worst predictor is **region**.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.6040438	0.5983053	-2.6809789	0.0073407
age	0.0144510	0.0128132	1.1278165	0.2593974
sex	0.6223115	0.2525679	2.4639377	0.0137420
region	-0.0217950	0.1062455	-0.2051384	0.8374640
income	0.0000394	0.0000150	2.6360381	0.0083880
children	-0.1608996	0.1176948	-1.3670928	0.1715962
car	-0.2120860	0.2500121	-0.8483029	0.3962693
personal_loan	-0.3307559	0.2732410	-1.2104917	0.2260903
mortgage	-0.1862119	0.2657417	-0.7007252	0.4834745

3. Using a Decision Tree, predict the spending group for the following individuals:

For this tree I used $cp = 0.01$ using the training dataset to build the tree.



A. 44 year old man who pays for his Deluxe, non-international account using a credit card.

Ultra-High is the predicted class.

B. High school student who pays cash for a Basic account with no added features

High is the predicted class.

C. 17 year old who uses bank transfer to pay for a Basic account, but has added on features including international calling, voice messaging, call forwarding.

Medium is the predicted class.

4.

A.

For the predictions of each class the counts for each class are shown in the table below.

High	Low	Medium	UltraHigh
5	2	1	0

B.

Below is a table with the predicted probabilites for each class using a naive bayes model. As we can see, **ID182710** ($catagory = Low$) has the largest probability.

	High	Low	Medium	Ultra High
ID182710	0.0981106	0.6492650	0.2466314	0.0059930
ID182711	0.2597828	0.0061792	0.4356713	0.2983667
ID182712	0.6445454	0.0001216	0.0952133	0.2601196
ID182713	0.5073473	0.0281416	0.4478163	0.0166948
ID182714	0.1312979	0.5330971	0.3299499	0.0056551
ID182715	0.4390724	0.0004018	0.2372534	0.3232724
ID182716	0.4934884	0.0018658	0.1153727	0.3892731
ID182717	0.4020849	0.2456343	0.3402386	0.0120422

C.

The most likely group (among men and women) appears to be women, but only slightly. Below is the table values for the conditional probabilities for variable **sex**.

	F	M
High	0.4698795	0.5301205
Low	0.5000000	0.5000000
Medium	0.4362416	0.5637584
Ultra High	0.5308642	0.4691358

5.

Like I did in the homework, I first standardize the variables by diving each element of the variable by the range, i.e.

$$\frac{x_i}{range(X)}$$

where x_i is the ith element of the Xth variable.

The predicted number of each catagory is in the table below.

High	Medium	Low	Lowest
4	26	12	1

The person(s) with the largest probability of being promoted are **3835921**, **6439891**, **8903741**, **7787991**. These people have probability of close to 1 for $class = High.Priority$.