Homework 4

Group 3
Nathalia Negri
Troy Yang

305-318-3446
781-812-4931

negri.n@husky.neu.edu
yang.tr@husky.neu.edu

Percentage of Effort Contributed by Student 1:________50%________

Percentage of Effort Contributed by Student 2:________50%________

Negri

Signature of Student 1:____________ ________________

Signature of Student 2:____________ ___________________

Submission Date:________June 14, 2017__________

Question 8.1

a. Online:

	Column					
	0		1		Total Count	Total Count
Row Labels	Count of CreditCard	Count of Personal Loan	Count of CreditCard	Count of Personal Loan		
0	**860**	**860**	**1247**	**1247**	**2107**	**2107**
0	781	781	1115	1115	1896	1896
1	79	79	132	132	211	211
1	**371**	**371**	**522**	**522**	**893**	**893**
0	329	329	471	471	800	800
1	42	42	51	51	93	93
Grand Total	**1231**	**1231**	**1769**	**1769**	**3000**	**3000**

b. 51/522 = 9.8%

c. Online:

Count of Personal Loan	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
0	1110	1586	2696
1	121	183	304
Grand Total	**1231**	**1769**	**3000**

Credit Card:

Count of Personal Loan	Column Labels		
Row Labels	**0**	**1**	**Grand Total**
0	1896	800	2696
1	211	93	304
Grand Total	**2107**	**893**	**3000**

d.
i. P(CC=1|Loan=1) = 93/304 = 30.59%
ii. P(Online=1|Loan=1) =1 83/304 = 60.2%
iii. P(Loan=1) = 304/3000 = 10.13%
iv. P(CC=1|Loan=0) = 800/2696 = 29.7%
v. P(Online=1|Loan=0) = 1586/2696 = 58.83%
vi. P(Loan=0) = 1 - P(Loan=1) = 89.87%
e. P(Loan=1|CC=1, Online=1) =

P(CC=1|Loan=1) * P(Online=1|Loan=1) * P(Loan=1) /
[P(CC=1|Loan=1) * P(Online=1|Loan=1) * P(Loan=1) +
P(CC=1|Loan=0) * P(Online=1|Loan=0) * P(Loan=0)]

= .3059 * 0.602 *.1013 / [.3059 * 0.602 *.1013 + 0.297 * 0.5883 * 0.8987]
= 0.106= 10.6%

f. This value is less than 1% larger than the value obtained in part (b). The exact classifier is more accurate, although the naive one is supposed to be an approximation for it, as long as the predictor values are independent of one another.

g.

0	1	0.89344	0.10656	-1.73924	1	1

The entries needed in order to calculate the probability are those that have credit cards and are online. The result of the XL Miner Naive Bayes Calculation for P(Loan =1 | CC =1, Online = 1) was 10.656%,

Question 8.2

a. If no other information is available, the prediction should be injury = yes. That is because injury = yes is the majority class in the given dataset.

b.

i. INJURY as a function of TRAF_CON and WEATHER_CON:

Row Labels	Column Labels no Count of TRAF_CON_R	Count of WEATHER_R	yes Count of TRAF_CON_R	Count of WEATHER_R	Total Count of TRAF_CON_R	Total Count of WEATHER_R
0	**6**	**6**	**3**	**3**	**9**	**9**
1	1	1	2	2	3	3
2	5	5	1	1	6	6
1	**2**	**2**			**2**	**2**
1	1	1			1	1
2	1	1			1	1
2	**1**	**1**			**1**	**1**
1	1	1			1	1
Grand Total	**9**	**9**	**3**	**3**	**12**	**12**

ii. INJURY = YES AND TRAN_CON = 0 AND WEATHER = 1 P = 2/3
INJURY = YES AND TRAN_CON = 1 AND WEATHER = 1 P = 0
INJURY = YES AND TRAN_CON = 2 AND WEATHER = 1 P = 0
INJURY = YES AND TRAN_CON = 0 AND WEATHER = 2 P = 1/6
INJURY = YES AND TRAN_CON = 1 AND WEATHER = 2 P = 0

INJURY = YES AND TRAN_CON = 2 AND WEATHER = 2   P = 0

iii.

TRAF_CON_R	WEATHER_R	INJURY	CLASSIFICATION
0	1	yes	yes
0	2	no	no
1	2	no	no
1	1	no	no
0	1	no	yes
0	2	yes	no
0	2	no	no
0	1	yes	yes
0	2	no	no
0	2	no	no
0	2	no	no
2	1	no	no

The three rows were classified as yes because TRAN_CON = 0 AND WEATHER = 1 had a prior probability of ⅔ to be INJURY = YES, which is above our cutoff of 0.5. The rest were classified as INJURY = NO as they fell below our threshold.

iv.

P(WEATHER = 1 | INJURY = 1) * P(TRAF = 1 | INJURY = 1) * P(INJURY = 1) /
[P(WEATHER = 1 | INJURY = 1) * P(TRAF = 1 | INJURY = 1) * P(INJURY = 1) +
P(WEATHER = 1 | INJURY = 0) * P(TRAF = 1 | INJURY = 0) * P(INJURY = 0)]

→ 0/[0+0.25] = **0%**

v.

Workbook	8.2_Accidents.xlsx
Worksheet	Sheet2
Range	$A$1:$F$13

Cutoff probability value for success (UPDATABLE)	0.5	Updating the value here will NOT update value in summary report

Predicted Class	Actual Class	Prob. for 0	Prob. for 1(success)	Log PDF	RAF_CON_	WEATHER_R
0	1	0.614035	0.385965	-1.35058	0	1
0	0	0.806806	0.193194	-1.06399	0	2
0	0	0.877437	0.122563	-1.99521	1	2
0	0	0.731707	0.268293	-2.3732	1	1
0	0	0.614035	0.385965	-1.35058	0	1
0	1	0.806806	0.193194	-1.06399	0	2
0	0	0.806806	0.193194	-1.06399	0	2
0	1	0.614035	0.385965	-1.35058	0	1
0	0	0.806806	0.193194	-1.06399	0	2
0	0	0.806806	0.193194	-1.06399	0	2
0	0	0.806806	0.193194	-1.06399	0	2
0	0	0.645161	0.354839	-2.65279	2	1

The classifications using XL Miner were the same as the classifications in (ii.). From (ii.):
INJURY = YES AND TRAN_CON = 1 AND WEATHER = 1 P = 0
XL Miner classified all of the records as INJURY = 0, which is the equivalent as the exact bayes.

c.
i.
We can include the following predators: HOUR_I_R, ALIGN_I, WRK_ZONE, WKDY_I_R, INT_HWY, LGTCON_I_R, REL_JCT_I_R, REL_RWY_R, SPD_LIM, SUR_CON, TRAF_CON_R, TRAF_WAY, and WEATHER_R.

ii.

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	

Confusion Matrix		
	Predicted Class	
Actual Clas	yes	no
yes	1785	1247
no	1408	1560

iii.

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	

Confusion Matrix		
	Predicted Class	
Actual Clas	yes	no
yes	1209	810
no	961	1020

Error Report			
Class	# Cases	# Errors	% Error
yes	2019	810	40.11887
no	1981	961	48.51085
Overall	4000	1771	44.275

Performance	
Success Class	yes
Precision	0.557143
Recall (Sensitivity)	0.598811
Specificity	0.514891
F1-Score	0.577226

iv.

Error on naive = 49.1%
Error on naive bayes classifier = 44.275%

Percent improvement = 49.1 - 44.275 = **4.825% smaller error**

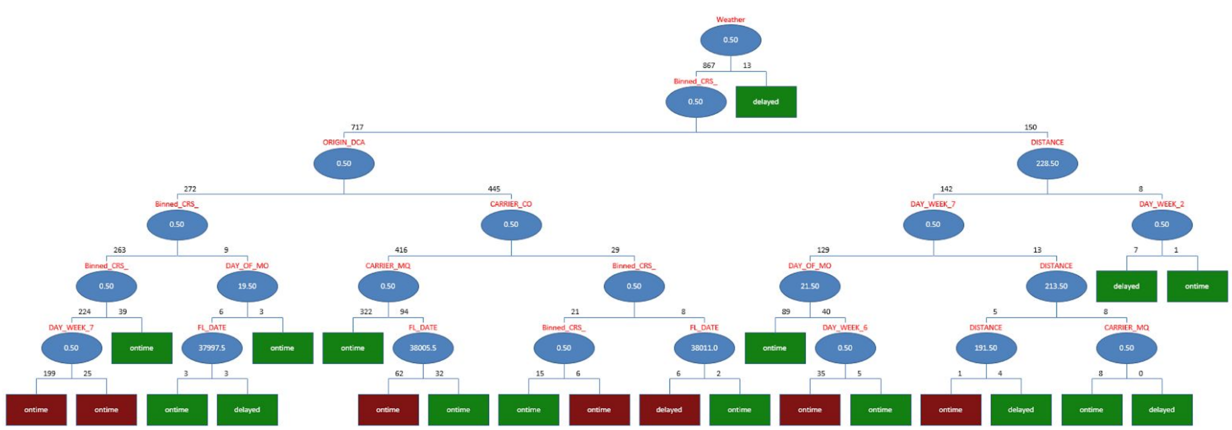
v.

	10	0.00067	10	0.001313
	15	0.006705	15	0.004595
	20	0.00771	20	0.003938
	25	0.119343	25	0.098129
	30	0.080121	30	0.098786
	35	0.200134	35	0.228421
	40	0.096212	40	0.10896
SPD_LIM	45	0.126048	45	0.14342
	5	0.00067	5	0.000328
	50	0.043245	50	0.031506
	55	0.156889	55	0.127995
	60	0.043916	60	0.051854
	65	0.071405	65	0.068264
	70	0.041904	70	0.027896
	75	0.005028	75	0.004595

The probability is zero, because at a speed of 5 MPH there were essentially no cases in which an injury occurred..

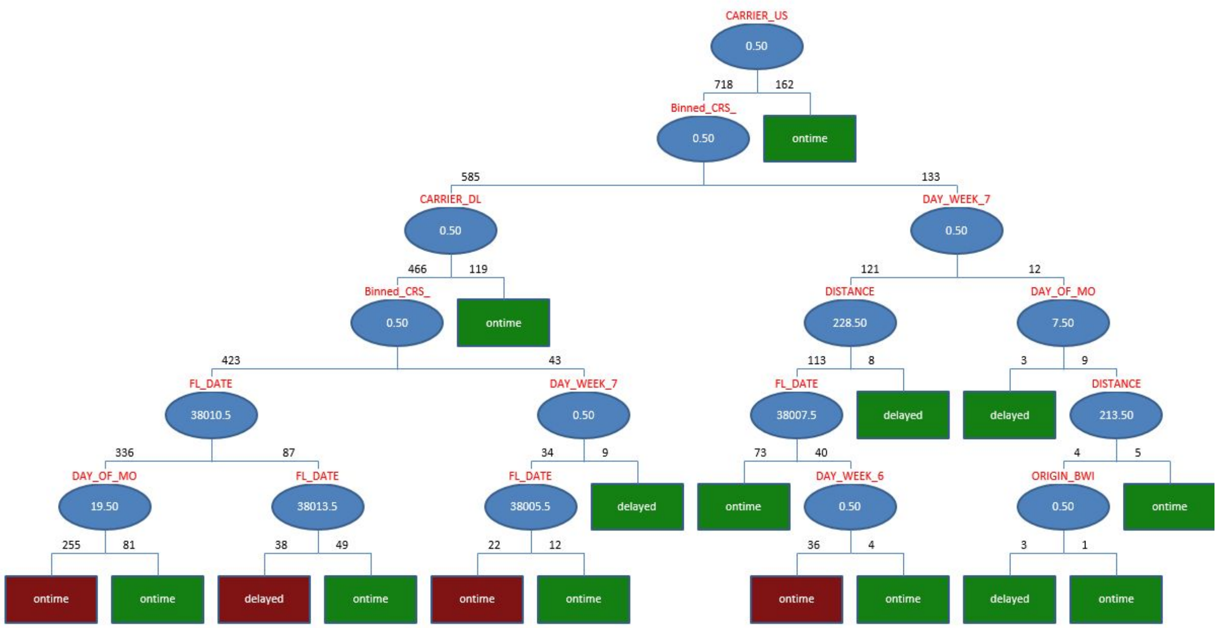
Question 9.2

a.



b. No we cannot use this tree. That is because we need additional information such as the day of the month, as well as the carrier. This information should be available in practice, as flight tickets always include flight date airline information. It appears that there is no redundant information.

c.
i.



The rules are as shown in the best pruned tree above. Based on the values of the predictor variables, one follows the rules of the tree down the splits.

ii. Classification and regression trees make predictions based on the values of predictor variables. We thought that this was most similar multiple linear regression, because the numerical estimation from a multiple linear regression model is based the values of the predictors.

iii. According to the full tree, the top three predictors appear to be the day of the week, departure time, and if you are flying with US or DL.

iv. That is because the best-pruned tree is the smallest tree in the pruning sequence with error within one standard error of the minimum error tree.

v. Using the best pruned tree is better, because the full tree, including the top levels, overfits the training data to complete accuracy, whereas the best pruned tree does not.

vi. The classification tree's failure to find a good predictive model can be a result of the limited number of levels generated by the full tree - 7 levels. However, one potential issue with adding more levels would be overfitting the training data. In order to combat this, the data miner can check the RMSE on the validation data.

Question 9.3

a.

i.

Full-Grown Tree Rules (Using Training Data)

#Decision Nodes	631	#Terminal Nodes	632

Level	NodeID	ParentID	SplitVar	plitValue/Se	Cases	LeftChild	RightChild	PredVal	Node Type
0	0	N/A	Age_08_04	32.5	718	1	2	10846.95	Decision
1	1	0	HP	113	100	3	4	18238.27	Decision
1	2	0	Age_08_04	57.5	618	5	6	9650.945	Decision
2	3	1	Age_08_04	17.5	89	7	8	17482.04	Decision
2	4	1	Age_08_04	5.5	11	9	10	24356.82	Decision
2	5	2	KM	127349	219	11	12	11490.34	Decision
2	6	2	Age_08_04	68.5	399	13	14	8641.351	Decision
3	7	3	tomatic_air	0.5	30	15	16	19001.77	Decision
3	8	3	uarterly_Ta	222	59	17	18	16709.31	Decision
3	9	4	CD_Player	0.5	2	19	20	31887.5	Decision
3	10	4	Age_08_04	20.5	9	21	22	22683.33	Decision
3	11	5	Age_08_04	45	205	23	24	11716.76	Decision
3	12	5	uarterly_Ta	68	14	25	26	8175	Decision
3	13	6	HP	96.5	199	27	28	9281.809	Decision
3	14	6	KM	157348	200	29	30	8004.095	Decision

Age_08_04, HP and KM appear to be the most important predictors when it comes to the cars price.

ii.

Training Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
0	0	0

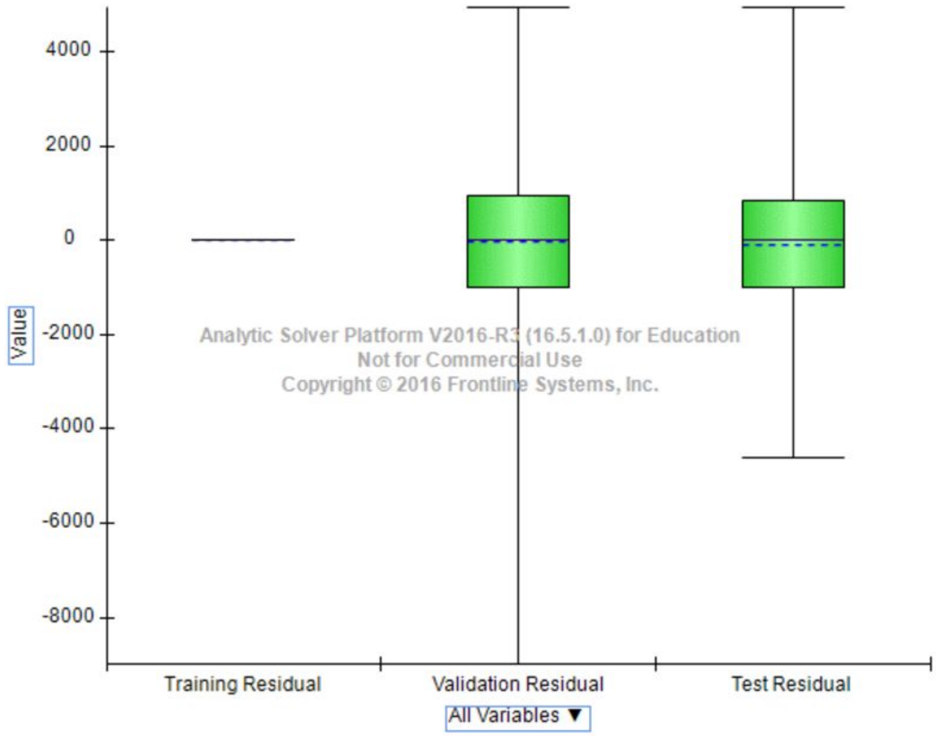
Validation Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
906867530	1450.552	-46.0232

Test Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
556721776	1392.766	-98.6202

The training data RMS error, as expected, is zero. The validation RMS Error and the test RMS error were quite similar, and were 1450.552 and 1392.766 respectively. The reason is because the regression tree was built on the training data itself, and tested on both the validation and test data (why they are somewhat similar).



iii.

If we make a best pruned tree instead of a full tree, then our predictions will not be equal to the actual values itself.

iv.

If we used the full tree instead of the best-pruned tree to score the validation set, the error would go down (increased predictive performance). That is because the pruning process trades off misclassification error in the validation data set against the number of decision nodes in the pruned trees, to arrive at a tree that captures the patterns (excluding noise) in the training data.

b.

i.

Full-Grown Tree Rules (Using Training Data)

#Decision Nodes	577	#Terminal Nodes	578

Level	NodeID	ParentID	SplitVar	plitValue/Se	Cases	LeftChild	RightChild	PredVal	Node Type
0	0	N/A	Age_08_04	57.5	862	1	2	10.30394	Decision
1	1	0	Age_08_04	42.5	370	3	4	15.14054	Decision
1	2	0	Age_08_04	68.5	492	5	6	6.666667	Decision
2	3	1	Age_08_04	32.5	195	7	8	17.48718	Decision
2	4	1	KM	119287.5	175	9	10	12.52571	Decision
2	5	2	KM	103729	253	11	12	8.573123	Decision
2	6	2	KM	76075.5	239	13	14	4.648536	Decision
3	7	3	KM	21858	112	15	16	18.86607	Decision
3	8	3	ered_Wind	0.5	83	17	18	15.62651	Decision
3	9	4	HP	103.5	160	19	20	13.16875	Decision
3	10	4	KM	154160	15	21	22	5.666667	Decision
3	11	5	uarterly_Ta	78.5	208	23	24	9.120192	Decision
3	12	5	KM	142260.5	45	25	26	6.044444	Decision
3	13	6	Airco	0.5	106	27	28	5.490566	Decision
3	14	6	arantee_Per	4.5	133	29	30	3.977444	Decision
4	15	7	tomatic_air	0.5	44	31	32	19.43182	Decision
4	16	7	uarterly_Ta	92.5	68	33	34	18.5	Decision
4	17	8	el_Type_Die	0.5	30	35	36	14.73333	Decision
4	18	8	KM	180249	53	37	38	16.13208	Decision
4	19	9	uarterly_Ta	66.5	55	39	40	11.81818	Decision
4	20	9	Age_08_04	56.5	105	41	42	13.87619	Decision
4	21	10	Age_08_04	46	10	43	44	7.6	Decision
4	22	10	Doors	3.5	5	45	46	1.8	Decision
4	23	11	KM	34557	120	47	48	7.966667	Decision
4	24	11	Age_08_04	60.5	88	49	50	10.69318	Decision

The structure and the number of levels in both trees were pretty similar. However, one of the top three predictors was different, as powered windows was one of the top predictors when price was binned. They are different because the classification tree is classifying records into 20 bins, which are more general, while the regression tree is making exact numerical estimation.

ii.
The regression tree predicts that the price would be $9,550, while the classification tree put that record in the $6,900 - $7,400 bin.

iii.
As we said in (ii.), some of the top predictors were quite similar, but as one continues down each tree, the structure of the tree, as well as the predictors of the tree differed more and more. Using the average of the bin, the difference between the two predictions is is $2,400, which is quite large (the prediction from the regression tree is 130% times larger than the one from the classification tree). The advantages of a regression tree in this instance, is that you are trying to predict the exact price, a numerical amount, so the regression tree is tailored to estimate precise numerical values, such as price. Using a classification tree requires one to also bin the outcome variable. Using a classification tree to predict numerical values is like trying to use a fork to eat soup.

10.1

a.

i. Logit = -14.7207 + 89.8321(TotExp/Assets) +8.3712(TotLns&Lses/Assets)

ii. Odds = e^(-14.7207 + 89.8321(TotExp/Assets) +8.3712(TotLns&Lses/Assets))

iii. P = 1 / (1+ e^-(-14.7207 + 89.8321(TotExp/Assets) +8.3712(TotLns&Lses/Assets))

b.
i. Logit = 0.1835
ii. Odds = 1.2014
iii. Probability = 0.5458
iv. Classification = Financially Weak (1)

c.
For odds:
.5 = odds/(1 + odds)
Odds = 1

For logit:
logit = e^(logit)ln(1)
logit = 0

d.

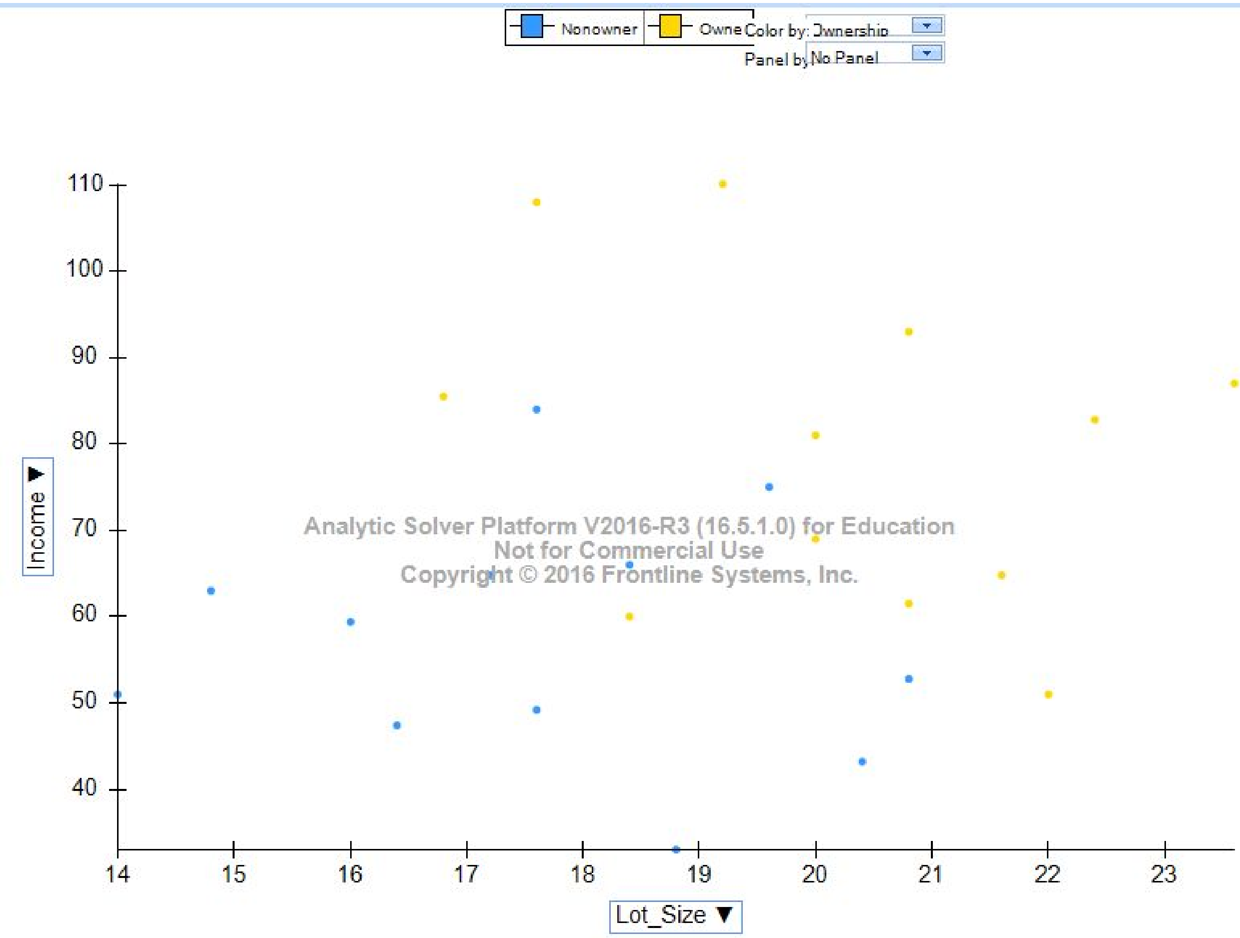
Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	-14.72072258	6.674731002	4.863968127	0.027423	4.04456E-07	8.42E-13	0.194273134
TotExp/Assets	89.83209519	47.77978313	3.534880281	0.060091	1.03177E+39	0.022047	Inf
TotLns&Lses/Assets	8.371184736	5.7787247	2.098504196	0.147443	4320.751983	0.052083	358441858.8

i.
As long as holding assets remain constant, for every extra dollar of expenditure, the bank is 1.03^39 times more likely to be classified as weak by the model.
ii.
As long as holding assets remain constant, for every extra dollar of loans, the bank is 4320.75 times more likely to be classified as weak by the model.

e.
In order to minimize the expected cost of misclassification, the cutoff should be decreased, as that would lead to more banks being classified as weak, and reduce the number of strong banks being misclassified as weak.

10.3
a. 50% of the records in the data are owners (12).

b.



Based on the scatterplot, the owners appear to have a higher level of income

c.

20/24 = 83.33 %

Confusion Matrix		
	Predicted Class	
Actual Clas	Owner	Nonowner
Owner	10	2
Nonowner	2	10

d.

In order to increase the number of correctly classified non owners, we should increase the cutoff. This will classify less records as owners, leading to less misclassified non owners.

e.

odds = e ^ (-25.9336+0.1108(60)+0.9636(20)) = e ^ -0.0136 = 0.9865

f.
Given a cutoff of 0.5, that record would be classified as an nonowner.

g.
Cutoff for odds = 1
Odds = 1 = e ^ -25.9336 + 0.1110(Income) + 0.9636(Lot Size)
= e ^ -25.9336 + 0.1110(x) + 0.9636(16)

X >= 94.7387387

The minimum income that a 16,000 sqft lot size owner should have before being classified as an owner is approximately $94.7387387k.