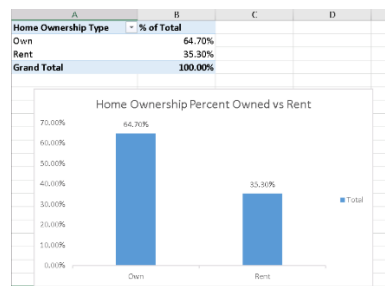July 1, 2018

Course: CIS570 – Business Intelligence
Name: Robert Palumbo
Assignment: BI Project Part2 - Submission Answers
Due Date: Sunday, July 1 @ 11:59pm

4. Do preliminary analysis on the Customers table data (if necessary, use a pivot table) to answer the following questions:

   a. How many cases (i.e., records or rows) are in the Customers table? **3187 rows**

   b. What percentage of customers own homes, and what percentage rent?

   Owned: 2062 / 3187 = 64.7
   Rented: 1125 / 3187 = 35.3

	A	B	C	D
Home Ownership Type		% of Total		
Own		64.70%		
Rent		35.30%		
Grand Total		100.00%		

Home Ownership Percent Owned vs Rent

70.00%	64.70%	
60.00%		
50.00%		
40.00%		35.30%
30.00%		■ Total
20.00%		
10.00%		
0.00%	Own	Rent

   c. Identify any one variable that could be a good predictor of home ownership (i.e., own or rent). Provide evidence for your choice.

   I my opinion, Num Children is a good predictor of home ownership. As shown in the following table, there is a 7x increase in total number of children for those that do own a home versus those that rent. Based on average, if you have a child then you are highly likely to own a home versus rent one.

A	B	C	D
Home Ownership	Customers Count	Num Children	AvgNumKids
Own	2062	2218	1.075654704
Rent	1125	312	0.277333333
Grand Total	3187	2530	

   d. What is the range (i.e., minimum and maximum) of values for Number of TVs?

   Minimum TVs: 0
   Maximum TVs: 10

A	B
NumTVs Descriptive Stats	
Mean	2.089359015
Standard Error	0.021760189
Median	2
Mode	1
Standard Deviation	1.224578698
Sample Variance	1.499592988
Kurtosis	3.247671321
Skewness	1.377213343
Range	10
Minimum	0
Maximum	10
Sum	6617
Count	3167

e. What is the mean (i.e., average) age? **Average age is: 33.12716.**

A	B
Age Descriptive Statistics	
Mean	33.12716
Standard Error	0.128267
Median	32
Mode	32
Standard Deviation	7.22978
Sample Variance	52.26972
Kurtosis	0.321658
Skewness	0.735457
Range	42
Minimum	20
Maximum	62
Sum	105245
Count	3177

8. Use the Mining Model Viewer and Mining Accuracy Chart to answer the following questions (Note: Set the Slider to show all 5 levels):

a) Identify the node that has customers with the highest probability of owning homes.

What is the probability? 94%
What is the rule for this node? (Num Cars >= 2 and Age >=32 and Marital Status = 'Married')

Mining Legend	▾ ☐ ✕
High Low	
Total Cases: 839	

Value	Ca...	Probability	Histogram
✔ Missing	0	0.00%	
✔ Own	789	94.00%	
✔ Rent	50	6.00%	

Num Cars >= 2 and Age >= 32 and Marital Status = 'Married'

b) Identify the node that has customers with the highest probability of renting.

What is the probability? 99.93%
What is the rule for this node? Num Cars < 2 and Age < 28 and Marital Status = 'Never Married' and Internet Connection = 'No Internet Connection'

Mining Legend	▾ ☐ ✕
High Low	
Total Cases: 27	

Value	Ca...	Probability	Histogram
✔ Missing	0	0.00%	
✔ Own	0	0.07%	
✔ Rent	27	99.93%	

Num Cars < 2 and Age < 28 and Marital Status = 'Never Married' and Internet Connection = 'No Internet Connection'

c) What are the weakest and strongest predictors of Home Ownership?

Weakest Predictor?   Weakest indicator is: Internet Connection
Strongest Predictor? Strongest indicator is: Num Cars

d) Complete the following table:

Counts for Customers on Home Ownership:

Predicted	Own (Actual)	Rent (Actual)
Own	418	103
Rent	96	179

e) Based on the values in the table above, what percentage of the total numbers of cases in the test data, does the Decision Tree model predict correctly?

Total Test Cases: 418 + 103 + 96 + 179 = 796
True Positives: 418
True Negatives: 179
Correctly Predicted: 418 + 179 = 597
Percent Correct: (597 / 796) = **75%**

f) Set the Predict Value in the Input Selection Tab (within the Mining Accuracy Tab) to "Rent". At 30% of the overall population.

NOTE: I could not get the slide bar to drag back and forth on the plot line to easily select what I believed to be the 30% mark. These are the values I think equate to 30%.

What percentage of the target population is correctly predicted by:

a) the ideal model? 84.68%
b) the Decision Tree model? 57.8%
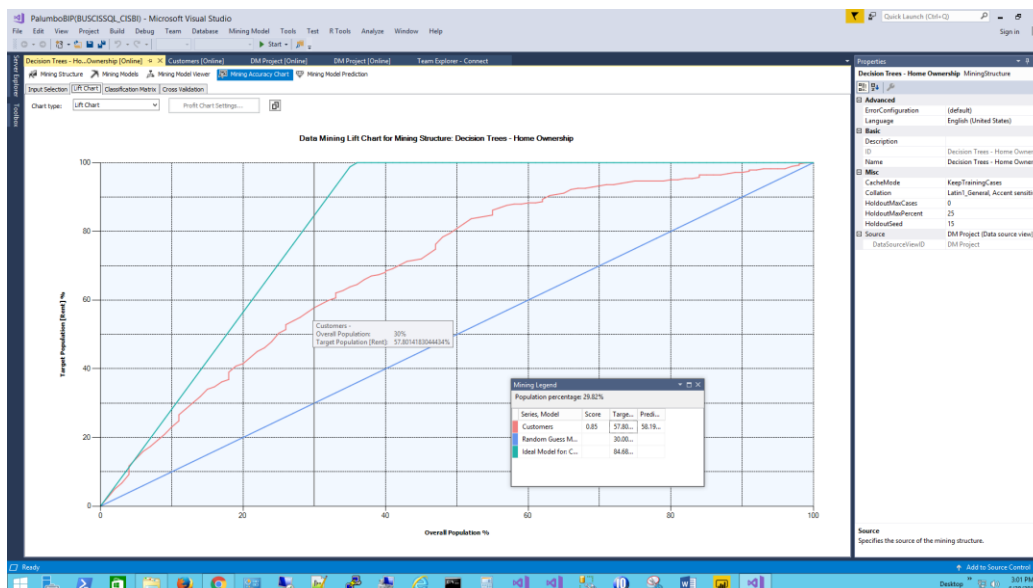
g) Set the Predict Value in the Input Selection Tab (within the Mining Accuracy Tab) to "Own". At 65% of the overall population.

NOTE: I could not get the slide bar to drag back and forth on the plot line to easily select what I believed to be the 65% mark. These are the values I think equate to 65%.

What percentage of the target population is correctly predicted by:

   a) the ideal model? 100%
      b) the Decision Tree model? 81.128%
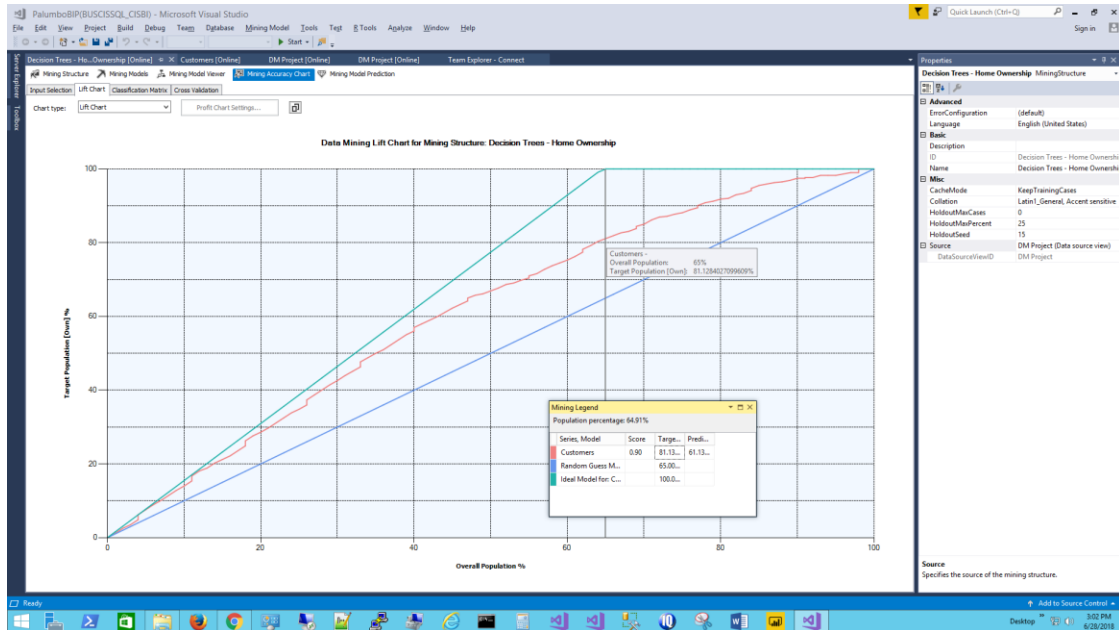
h) Ram Realty will pay you $3 for each emailed solicitation that reaches a renter. Assume there are 30,000 people on an email list, the fixed cost for the email campaign is $2,500 and it costs $1 per email.

NOTE: I could not get the slide bar to drag back and forth on the plot line to easily select what I believe is the maximum profit point. Thus, my closest select is marked at what is shown in the image below.

If your objective is to maximize profit, to how many people should you email the solicitation?

My maximum profit is marked at 51.75% of the population. With a population of 30,000 that would equate to: 0.5175 * 30,000 = 15,525 renters

What will be your profit? $8,583.42 (input selection -> predict value = 'Rent')
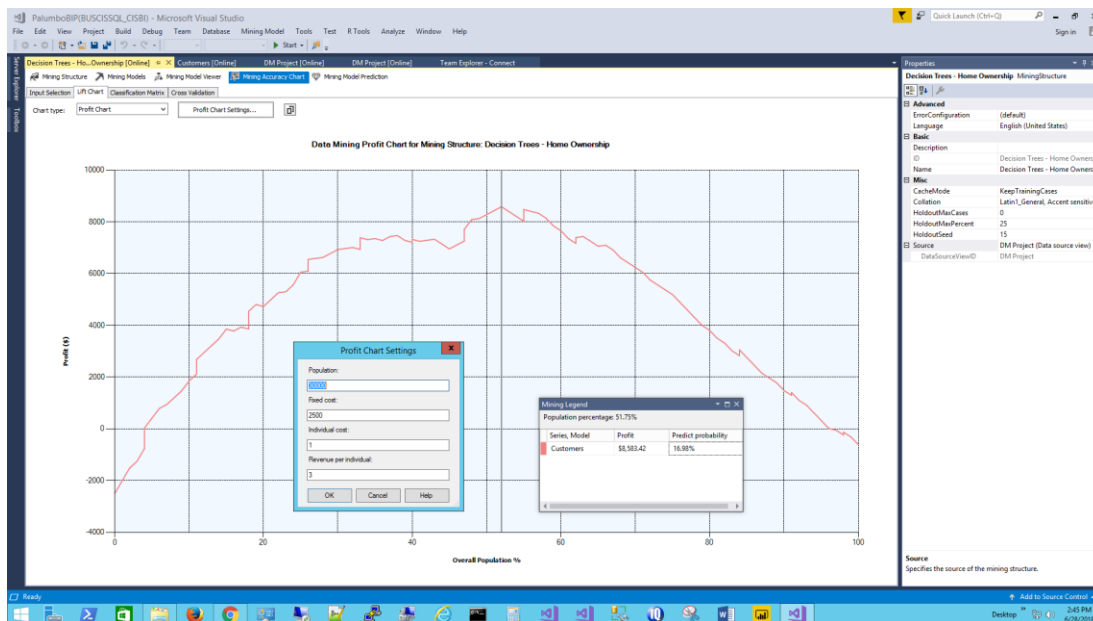
Bonus Question: If you followed the model's recommendation, how many renters in the population (i.e., 30,000 people) will receive your email solicitation?

If I understand how to interpret lift correctly for model performance then for the same 51.75% of the total population shown in the profit chart, the ideal model at 51.75% in the lift chart would suggest ALL 30,000 renters would have received the email solicitation as shown in the following lift chart: (if I do not understand this correctly then my answer is unfortunately wrong)