DATA 511 Intro to Data Science
Course Project 3
Prof Larose

Name: Danielle Senechal

The project instructions are shown in bold. This is to distinguish the instructions from your work. Your work should be in not-bold.

- Work neatly. Aim for a professional-looking presentation. Make sure that you interpret all of your results, using the language shown in the notes.
- Make sure all graphs and tables fit neatly on the page.
- Neither add nor delete pages.

For all text output, surround it with a text box. In Word, select the text output, then Insert > Text Box > Draw Text Box.

Apart from this document, which you will save as a pdf and submit, you must submit your R script, containing the code you used to solve the problems. The R script should be neat and easily understandable by people who are not you. It should be well-annotated, describing what you are doing so that anyone could understand it.

This Project is brand new, and may have typos, errors, etc, that I have missed. Please report these to me asap. For this and other reasons, this Project is subject to change at any time (though of course I will be reasonable.)

Import the *proj3_income* data set as *proj3*. Set your seed to 12345. Install and library the *plyr, caret* and *rattle* packages.
Delete the variable *occupation*. `(proj3$occupation <- NULL)`
The task is to predict the target variable *income*, based on the other variables.

Good luck!

Prof Larose

1. **Insert your Executive Summary here. (A strategy for this is given at the end.)**

The objective of developing the following classification model is to predict the target variable income based on capital gains, capital losses, education level, and marital status. The two income categories for customers will be high-income individuals (those who make greater than $50,000) and low-income individuals (those who make less than or equal to $50,000). In the data set, 75.92% of customers in this data set have low incomes, and 24.08% have high incomes. The classification model chosen to predict type of income is a CART model (Classification and Regression Tree).

The following was observed:

- The resulting decision tree has 13 nodes, with 5 leaf nodes. The first node determines the marital status of a customer. If unmarried, the education level and/or the capital gains determine income. If married, the income is determined by the amount of capital gains.
- There is no evidence of overfitting in this model, each fold has an accuracy within 4% of all the other folds.
- The accuracy of the CART models 82.05%, which beat the baseline model which had an accuracy of 75.92%.
- While the CART model was outperformed by the baseline models in predicting true/false positive and negatives, its accuracy of 82.05% and error rate of 17.95% outperformed the baseline models.
- The all positive model outperformed the CART model in sensitivity, and the all negative model outperformed the CART model in specificity, both having 100%.
- The CART model improved accuracy by 8.06% and decreased the error rate by 28.58%.
- This tree has five decision rules. The best decision rule for finding high-income customers would be the rule that leads to leaf node 7, and the best for finding low-income customers would be the rule for leaf node 4.
- Leaf node 7 (high-income decision rule) contains 25% of all customers, and leaf node 4 (low-income decision rule) contains 54% of all customers in the data set.
- Leaf node 7 has a 60% confidence, and leaf node 4 has 95% confidence.
- The typical profile for a high-income customer is an individual who is married and has some college education or more. 25% of all customers in the data set fall into this profile.
- The typical profile of a low-income customer is an unmarried individual with less than 7.074 in capital gains. 54% of all customers in the data set fall into this category.
- The other profiles for high-income only have 1% of all customers in each, and the other profile for low-income customers only has 19% of all customers. The profiles with higher support are more useful when trying to find and contact high or low-income customers.

A future direction for this model would be to refine it using a different amount of data for testing and training, or to try to use a different classification model to see if the profiles developed would be similar to the profiles made by the CART Model.

2. **Pain in the Drain Data Prep.**
   a. **Provide the summary of *income*, and mention the proportion of high income. On your own, observe a table of *education* (columns) against *income.* Reclassify *education* into *educ*, consisting of two categories only, *low* and *high.* Reclassify "Some-college" and up as "high", and "HS-grad" down as "low". Delete *education*. Provide a prop.table of *educ* against income, rounded to two decimal places. Comment.**

   The following table shows the amount of records for each type of income, less than or equal to $50,000, and greater than $50,000. There are 32,561 records in total. 24,720 people have incomes less than or equal to $50,000, and 7,841 have incomes greater than $50,000.

<=50K	>50K
24,720	7,841

   When comparing education against income, it can be concluded from the table below that of all those who have completed some college or more, 66.74% (11,885) have incomes less than or equal to $50,000, and 33.26% (5,922) have incomes greater than $50,000. For those who have at most a high school degree, 86.99% (12,835) have incomes less than or equal to $50,000, and 13.01% (1,919) have incomes greater than $50,000. This suggests that an individual will more likely have a higher income if they have completed some education past a high school degree.

	High	Low
<=50K	66.74%	86.99%
>50K	33.26%	13.01%

   b. **On your own, observe a table of *relationship* against *income.* Reclassify *relationship* into *rel*, consisting of two categories only, *HusWife* and *Other.* Reclassify "Husband" and "Wife" as "HusWife", and the other categories as "Other". Delete *relationship.* Provide a prop.table of *rel* against income, rounded to two decimal places. Comment.**

   The table below shows that for all people who have a husband/wife 54.86% (8,098) have incomes less than or equal to $50,000, and 45.14% (6,663) have incomes greater than $50,000. For everyone else, 93.38% (16,622) have incomes less than or equal to $50,000 and 6.62% have incomes greater than $50,000. This suggests that it is more likely an individual will have a higher income if they have a husband or a wife (are married) than if they do not.

	HusWife	Other
<=50K	54.86%	93.38%
>50K	45.14%	6.62%

3. **Step 1 of the CMBM. Partition the data set, into a training set *proj3.tr* and a test set *proj3.te*. Do so, so that each data set contains 50% of the records.**
   a. **Provide a summary of each data set's *income* variable, and comment.**

Training data:

<=50K	>50K
12,360	3,921

Testing data:

<=50K	>50K
12,360	3,920

The following table shows the number of records in each category. In the training data, 12,360 individuals have incomes less than or equal to $50,000, and 3,921 individuals have incomes higher than $50,000. In the testing data set, 12,360 individuals have incomes less than or equal to $50,000, and 3,920 individuals have incomes higher than $50,000.

   b. **Validate the partition for *capital.gain* and *capital.loss*. Do the boxplots, but do not include them here, to save space. Provide the Kruskal-Wallis test results for each, with the p-values in bold red.**

Capital-gains:

```
        Kruskal-Wallis rank sum test

data:  proj3.all$`capital-gain` by as.factor(part)
Kruskal-Wallis chi-squared = 1.0854, df = 1, p-value = 0.2975
```

Capital-losses:

```
        Kruskal-Wallis rank sum test

data:  proj3.all$`capital-loss` by as.factor(part)
Kruskal-Wallis chi-squared = 0.003709, df = 1, p-value = 0.9514
```

Since both p-values are above the significance level of 0.05, the partition is validated on the capital-gains and capital-losses variables.

c. Validate the partition for *educ* and *rel.* Provide the prop.test results for each, with the p-values in red bold.

Education:

```
        2-sample test for equality of proportions without continuity correction

data:  educ.part.table
X-squared = 2.3448, df = 1, p-value = 0.1257
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.019432890  0.002385891
sample estimates:
   prop 1    prop 2
0.4961532 0.5046767
```

Relationship:

```
        2-sample test for equality of proportions without continuity correction

data:  rel.part.table
X-squared = 0.0036854, df = 1, p-value = 0.9516
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01124726  0.01057145
sample estimates:
   prop 1    prop 2
0.4998306 0.5001685
```

Since both p-values are above the significance level of 0.05, the partition is validated on both the education and relationship variables.

d. What is your conclusion regarding the partition?

Upon observing the boxplots for capital-gains and capital-losses, the distribution is very similar. Thus, the Kruskal-Wallis test can be performed. The two p-values produced by this test are 0.2975 and 0.9514, and since both of these p-values are higher than the significance level of 0.05, the partition can be validated on the capital-gains and capital-losses variables. To validate the education and relationship variable, a 2-sample test for equality of proportions without continuity correction was ran on both variables and produced p-values of 0.1257 and 0.9516 respectively. These two p-values also fall above 0.05, and the partition can be validated on the education and relationship variables.

4. **Step 2 of the CMBM. Establish baseline model performance, using the training data set. Provide the two baseline contingency tables, nicely formatted. Use the accuracy metric.**

All Positive:

Predicted Category				
Actual Category		*False*	*True*	**Total**
	False	TN = 0	FP = 12,360	TAN = 12,360
	True	FN = 0	TP = 3,921	TAP = 3,921
	Total	TPN = 0	TPP = 16,281	GT = 16,281

$$Accuracy = \frac{TN + TP}{GT} = \frac{0 + 3{,}921}{16{,}281} = 24.08\%$$

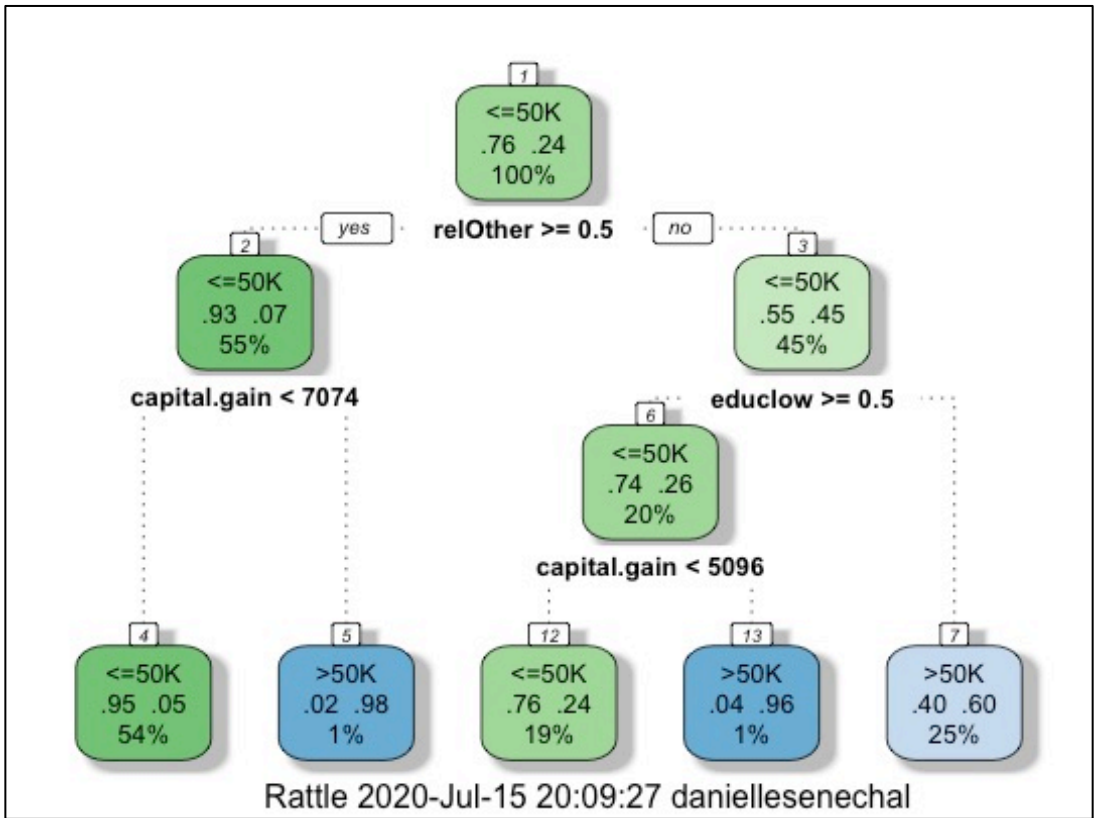
All Negative:

Predicted Category				
Actual Category		*False*	*True*	**Total**
	False	TN = 12,360	FP = 0	TAN = 12,360
	True	FN = 3.921	TP = 0	TAP = 3,921
	Total	TPN = 16,281	TPP = 0	GT = 16,281

$$Accuracy = \frac{TN + TP}{GT} = \frac{12{,}360 + 0}{16{,}281} = 75.92\%$$

5. **Step 4 of the CMBM. Develop your CART model, using 10-fold cross-validation. Use *method = "rpart2"*. Provide the decision tree here, along with a couple of sentences of description.**

The following decision tree was made using the training data from the project3_income data set.



There are 13 nodes in this tree, starting with the root node which states that 76% of the records in the training data set have low-incomes (less than or equal to $50,000) and the remaining 24% have high-incomes (greater than $50,000).
The root node splits customers into married and not married categories, so if the customer is unmarried, the yes branch is followed, and if they are married, the no branch is followed. If the customer is unmarried, their capital gains are then observed and the next branch depends on if they have more or less than 7,074 in capital gains, which will lead to the final leaf. For those who are married, their education is observed, and if they have less than a high school degree, they branch to leaf node 7. If a customer has some or more college, their capital gains are observed, and the final leaf depends on if they make more or less than 5,096 in capital gains.

6. **Step 5 of the CMBM. Check for overfitting. Allow 4% variation in accuracy among your ten folds. Provide your conclusion regarding overfitting.**

	Accuracy	Kappa	Resample
1	0.8372236	0.5715806	Fold02
2	0.8224816	0.5319959	Fold01
3	0.8237101	0.5444142	Fold03
4	0.8170657	0.5357643	Fold06
5	0.8304668	0.5600655	Fold05
6	0.8200246	0.5348898	Fold04
7	0.8212531	0.5215266	Fold07
8	0.8249386	0.5361369	Fold10
9	0.8298526	0.5499172	Fold09
10	0.8095823	0.4984098	Fold08

It can be seen that Fold 2 achieved the highest accuracy of 0.8372, and Fold 8 had the lowest accuracy at 0.8096. Using a 4% variation in accuracy, it can be judged that there is not significant evidence of overfitting, because all folds were within 4% of each other in accuracy.

7. **Step 6 of the CMBM. Apply the model to the test data set. Provide a nicely formatted contingency table of the results. In one sentence, state whether your model outperforms the baseline model.**

Predicted Category				
Actual Category		*False*	*True*	**Total**
	False	TN = 10,630	FP = 1,730	TAN = 12,360
	True	FN = 1,193	TP = 2,727	TAP = 3920
	Total	TPN = 11,823	TPP = 4,457	GT = 16,280

$$Accuracy = \frac{TN + TP}{GT} = \frac{10{,}630 + 2{,}727}{16{,}280} = 82.05\%$$

The final model achieved an accuracy of 82.05%, which beats the baseline accuracy of 75.92%.

8. **Step 8 of the CMBM. Construct a table comparing the evaluation metrics for the all-positive model, the all-negative model, and your CART model. Include the contingency table results in this table. The included metrics should be *accuracy, error rate, sensitivity,* and *specificity.* For each line in the table, indicate in green which model did best and highlight in red which model did worst. <u>Fully discuss your results.</u> Report how much your model decreased the error rate by.**

	All Positive Model	All Negative Model	CART Model
TN	0	12,360	10,630
FP	12,360	0	1,730
FN	0	3,921	1,193
TP	3,921	0	2,727
Accuracy	0.2308	0.7592	0.8205
Error Rate	0.7592	0.2308	0.1795
Sensitivity	1.0000	0.0000	0.6957
Specificity	0.0000	1.0000	0.8600

<u>All Positive</u>:

$$Sensitivity = \frac{TP}{TP + FN} = \frac{3{,}921}{3{,}921 + 0} = 100.00\%$$

$$Specificity = \frac{TN}{FP + TN} = \frac{0}{12{,}360 \ + \ 0} = 0.00\%$$

100% of the positive records were predicted positive, and 0% of the negative records were predicted negative by this model.

<u>All Negative</u>:

$$Sensitivity = \frac{TP}{TP + FN} = \frac{0}{0 + 3{,}921} = 0.00\%$$

$$Specificity = \frac{TN}{FP + TN} = \frac{12{,}360}{0 \ + \ 12{,}360} = 100.00\%$$

0% of the positive records were predicted positive, and 100% of the negative records were predicted negative by this model.

<u>CART Model</u>:

$$Sensitivity = \frac{TP}{TP + FN} = \frac{2{,}727}{2{,}727 + 1{,}193} = 69.57\%$$

$$Specificity = \frac{TN}{FP + TN} = \frac{10{,}630}{1{,}730 \ + \ 10{,}630} = 86.00\%$$

69.57% of the positive records were predicted positive, and 86.00% of the negative records were predicted negative by this model.

The all negative model had the most true negative and the least false positives, while the all positive model had the most true positives and the least false negatives. The two models outperformed the CART model in these categories. The CART model however did outperform the all positive and all negative model with the highest accuracy of 82.05% and the lowest error rate of 17.95%. In regard to sensitivity and specificity, the all positive model had a perfect sensitivity of 1.0, and the all negative model had a perfect specificity of 1.0, both which outperform the CART model.

The CART model improved the accuracy by $\frac{0.8205-0.7592}{0.7603} = 8.06\%$, and decreased the error rate by $\frac{0.2308-0.1795}{0.1795} = 28.58\%$.

9. **Provide an itemized list of all of the possible decision rules obtainable from your decision tree, including confidence and support. Select the model of most interest to your client if your client is interested in contacting high-income customers.**

There are five possible decision rules for this tree.

Leaf Node 4:

- *Antecedent*: Customer is not married (branch to Leaf Node 2) and has less than 7,074 in capital gains (branch to Leaf Node 4).
- *Consequent*: Customer has income less than or equal to $50,000.
- *Support*: The proportion of all customers that make it to this node is 54%.
- *Confidence*: Leaf Node 4 reports that 95% of the records have incomes below or equal to $50,000, meaning the confidence is 95%.

If the customer is not married and has capital gains less than 7,074, then the customer has an income less than or equal to $50,000, with 54% support and 95% confidence.

Leaf Node 5:

- *Antecedent*: Customer is not married (branch to Leaf Node 2) and has more than 7,074 in capital gains (branch to Leaf Node 5).
- *Consequent*: Customer has income greater than $50,000.
- *Support*: The proportion of all customers that make it to this node is 1%.
- *Confidence*: Leaf Node 5 reports that 98% of the records have incomes greater than $50,000, meaning the confidence is 98%.

If the customer is not married and has capital gains more than 7,074, then the customer has an income more than $50,000, with 1% support and 98% confidence.

Leaf Node 7:

- *Antecedent*: Customer is married (branch to Leaf Node 3) and has an education of some college or more (branch to Leaf Node 7).
- *Consequent*: Customer has income greater than $50,000.
- *Support*: The proportion of all customers that make it to this node is 25%.
- *Confidence*: Leaf Node 7 reports that 60% of the records have incomes greater than $50,000, meaning the confidence is 60%.

If the customer is married and has an education of some college or more, then the customer has income greater than $50,000, with 25% support and 60% confidence.

Leaf Node 12:

- *Antecedent*: Customer is married (branch to Leaf Node 3), has a high school degree or lower (branch to Leaf Node 6), and has capital gains less than 5,096 (branch to Leaf Node 12).
- *Consequent*: Customer has an income less than or equal to $50,000.
- *Support*: The proportion of all customers that make it to this node is 19%.
- *Confidence*: Leaf Node 12 reports that 76% of the records have incomes less than or equal to $50,000, meaning the confidence is 76%.

If the customer is married, has a high school degree or lower, and capital gains less than 5,096, then the customer has an income of less than or equal to $50,000, with 19% support and 76% confidence.

Leaf Node 13:

- *Antecedent*: Customer is married (branch to Leaf Node 3), has a high school degree or lower (branch to Leaf Node 6), and has capital gains more than 5,096 (branch to Leaf Node 13).
- *Consequent*: Customer has an income greater than $50,000.
- *Support*: The proportion of all customers that make it to this node is 1%.
- *Confidence*: Leaf Node 13 reports that 96% of the records have incomes greater than $50,000, meaning the confidence is 96%.

If a customer is married, has a high school degree or lower, and capital gains more than 5,096, then the customer has an income greater than $50,000, with 1% support and 96% confidence.

The best decision rule of this model if interested in contacting high interest customers would be the decision rule for Leaf Node 7. The support for this rule is 25%, while the other two Leaf Nodes with high incomes only have 1% support. 25% of all customers in this data set fit this profile, versus 1% of customers who fit each of the other two decision rules.

10. Using your decision tree, decision rules, and EDA, develop detailed profiles of:

a. High-income customer.

b. Low-income customer.

High-income customer:

A high-income customer from this data set typically is married and has some college education or more.

Other profiles for high income customers include being unmarried with more than 7,074 in capital gains or being married with a high school degree and more than 5,096 in capital gains. These are less common than the first profile, because support for these two profiles are each 1%, while the first profile has 25% support. This means that 25% of all customers in this data set fit into the first profile, and only 1% fit the second profile and 1% fit the third profile. The first profile should be used when searching for high-income customers because it will contain the highest percentage of high-income customers.

Low-income customer:

A low-income customer from this data set typically is unmarried, with less than 7,074 in capital gains.

Another profile for a low-income customer is an individual who is married, has some or more college education, and has less than 5,096 in capital gains. This profile only has 19% support, while the first one has 54% support. This means that 54% of individuals in this dataset fall into this profile, and 19% fit into the other profile. The first profile should be used when finding low-income customers because it will contain the highest percentage of low-income profiles.

Craft your Executive Summary as follows.

Your boss makes more money than you. He or she has little time for the arcane details of all the data prep and other work you did to produce your report. Your boss is only interested in RESULTS.

A good executive summary should consist of the following.

1. **A quick summary of the *Objective* of the analysis, especially for what the client is interested in. Also include the original proportion of high-income customers.**

2. **Bullet points with explanations of your most salient results. I think you can make good bullet points with your results from the following problem numbers:**
    a. **Problem 5**
    b. **Problem 7**
    c. **Problem 8**
    d. **Problem 9**
    e. **Problem 10**

3. **What NOT to include in your Executive Summary is anything about data prep, unless it affects managerial policy.**

4. **Brief mentioning of next steps.**

5. **And, whatever you do, do not exceed one page!** 😉

==========

Well done!

<u>Deliverables:</u>

1. **Save your completed Word document as a pdf file, named *Doe_Jane_Project3* (if your name is Jane Doe, <u>with last name first!</u>). Because of virus issues, no Word documents will be accepted.**
2. **Your well-annotated R script, named *Doe_Jane_Project3_RScript.***

Do NOT zip these two files together. Rather, make two separate submissions using the Project Submissions Tool.