DATA 511 Intro to Data Science
Course Project 4
Prof Larose

Name: Chad Zeller

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.
- Contingency tables must have predicted values in the columns and actual values in the rows.

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.)

Project 4 continues the work you started in Project 3.
So that we all start out on the same page, let's all do Project 4 using the same data sets. Import the *proj3_tr* data set as *proj4.tr*. Import the *proj3_te* data set as *proj4.te*.
Set your seed to 12345 throughout.
Install and library the *plyr, caret* and *rattle* packages.
If necessary, delete the X variables.
The task is to predict the target variable *income*, based on the other variables.

Good luck!

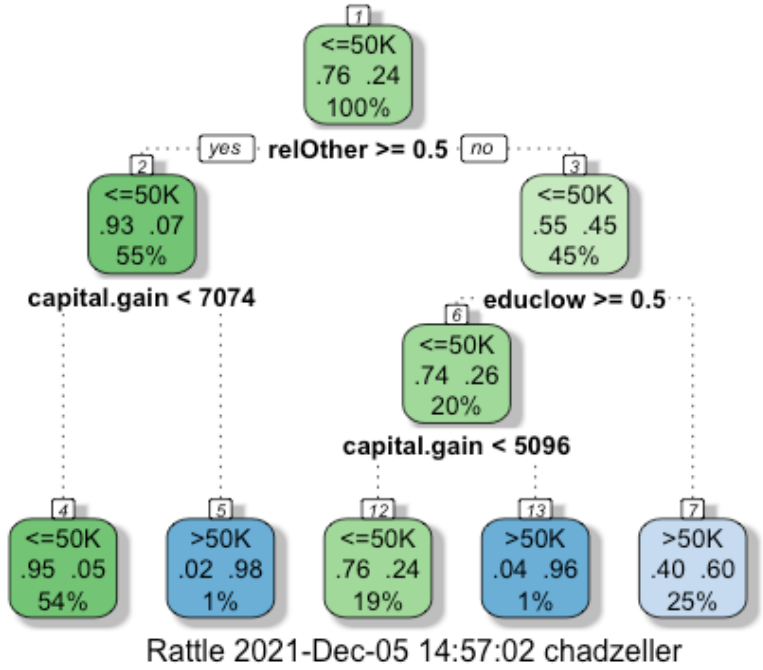
Prof Larose

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

   This project examines a set of 32,561 customers from our database, with the objective of predicting the customer's income based on variables capital gains, capital losses, education level, and relationship status using the CART (Classification and Regression Tree) classification model. Customers are classified as either high-income (income greater than $50,000 per year) or low-income (income of $50,000 or less per year). Approximately 76% of customers are classified as low-income and 24% are classified as high-income. The following key findings were observed.

   - The CART model accuracy is *82*% beat the baseline accuracy of *76*%.
   - While the all positive and all negative models did outperform the CART model in predicting positives and negatives respectively, the CART model had both the highest accuracy rate (82%) and the lowest error rate (18%) of all models analyzed.
   - The CART model *increased* the accuracy rate by 8.07% and *decreased* the error rate by 28.58% over the baseline model.
   - The decision tree has thirteen nodes and five decision rules. The decision rule of greatest interest is leaf node 7, which contains a strong majority (more than 90%) of all high-income customers. Leaf node 4 is the key decision node when considering low-income customers.
   - The overwhelming majority of high-income customers are married. The proportion of high-income earners who are married is well over *90%*, while those not married with a high-income are only 1% of the data set.
   - The vast majority of (at least 90%) of high-income customers have at least some college experience.
   - All ten folds were within the 4% of each other in accuracy. Therefore, we can conclude that this does not reflect significant evidence of overfitting.
   - The profile of a typical *high-income* customer is an individual who is married, has an income greater than $50,000 per year, and has an educational level of at least some college.
   - The profile of a typical *low-income* customer is an individual who is not married, has an income of $50,000 or less, and has capital gains totaling less than $7,074.

2. **Step 4 of the CMBM. Set your seed to 12345. 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 root node of this decision tree tells us that 76% of the records in the entire training set have low income (less than or equal to $50,000), with the remaining 24% have a high level of income (greater than $50,000).

The root node separates the records into two groups, based on the variable relationship status - married and not married. If the person is not married, the yes branch heading to node 2 is followed. If the person is married, the no branch to node 3 is followed.

If the person is married, their level of education is observed. If their education level is a high school diploma or less, the branch to node 7 is followed, while an education level of at least some college results in a branch to node 6, where capital gains would then be considered.

If the person is not married, their capital gains would be observed. If their capital gains are less than 7,074, you would proceed to the final leaf node.

3. **Step 5 of the CMBM. Check for overfitting. Provide the relevant output, highlighting the high and low values. Allow 4% variation in accuracy among your ten folds. Provide your conclusion regarding overfitting.**

	Accuracy	Kappa	Resample
1	0.8182934	0.5227196	Fold02
2	0.8194103	0.5266805	Fold01
3	0.8378378	0.5699534	Fold03
4	0.8187961	0.5301684	Fold06
5	0.8255528	0.5420153	Fold05
6	0.8169533	0.5210403	Fold04
7	0.8157248	0.5210187	Fold07
8	0.8095823	0.5115218	Fold10
9	0.8341523	0.5710471	Fold09
10	0.8390663	0.5688420	Fold08

Fold 8 achieved the highest degree of accuracy of 0.8391, while fold 10 had the lowest accuracy of 0.8096. All folds were within the 4% of each other in accuracy. Therefore, we can judge that this does not reflect significant evidence of overfitting and proceed to step 6.

4. **Step 6 of the CMBM. Apply the model to the test data set. Manually construct a nicely formatted contingency table of the results, showing the variable names and categories. In one sentence, state whether your model outperforms the baseline model.**

		Predicted Category		
		Low	**High**	**Total**
Actual Category	**Low**	TN = 10,630	FP = 1,730	TAN = 12,360
	High	FN = 1,193	TP = 2,727	TAP = 3,920
	Total	TPN = 11,823	TPP = 4,457	GT = 16,280

Accuracy = (10,360 + 2,727) / 16,280 = 82.05%

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

5. **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 the relative decrease in error rate your model achieved.**

	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	0.2408	0.7592	0.8205
Error Rate	0.7592	0.2408	0.1795
Sensitivity	1.0	0.0	0.6957
Specificity	0.0	1.0	0.8600

<u>All Positive Model</u>:

Sensitivity = 3,921 / (3,921 + 0) = <u>100</u>%
Specificity = 0 / (12,360) = <u>0</u>%

<u>All Negative Model</u>:

Sensitivity = 0 / (0 + 3,921) = <u>0</u>%
Specificity = 12,360 / (0 + 12,360) = <u>100</u>%

<u>CART Model</u>:

Sensitivity = 2,727 / (2,727 + 1,193) = <u>69.57</u>%
Specificity = 10,630 / (1,730 + 10,630) = <u>86</u>%

In the all positive model, 100% of the records were predicted positive, while 0% were predicted negative. The all positive model had the most true positives (TP), fewest false negatives (FN), fewest true negatives (TN), the most false positives (FP), the highest sensitivity rate, lowest specificity rate, lowest accuracy rate, and highest error rate.

In the all negative model, 0% of the records were predicted positive, while 100% were predicted negative. The all negative model had the most true negatives (TN), fewest false positives (FP), the most false negatives (FN), the fewest true positives (TP), the lowest sensitivity rate, and highest specificity rate.

In the CART model, 69.57% of the records were predicted positive, while 86% were predicted negative.

While the all positive and all negative models did outperform the CART model in predicting positives and negatives, the CART model had the highest accuracy rate and the lowest error rate. The all positive and all negative models performed best in sensitivity and specificity respectively.

The CART model *increased* the accuracy rate by (.8205 - .7592) / .7592 = 8.07%

The CART model *decreased* the error rate by (.2308 - .1795) / .1795 = 28.58%

6. **Provide an itemized list of all of the possible decision rules obtainable from your decision tree, including confidence and support. Make sure each decision rule takes the form of an English sentence. Select the decision rule of most interest to your client if your client is interested in contacting high-income customers.**

   There are five possible decision rules for this decision tree, represented in the leaf node numbers 4, 5, 7, 12 and 13.

   Leaf Node #4:

   - Antecedent - The customer is *not* married (branch to node 2) *and* has capital gains of less than $7,074 (branch to node 4).
   - Consequent - The customer has a low income ($50,000 or less).
   - Support - The proportion of all customers that make it to this leaf is 54%.
   - Confidence - Leaf node 4 reports that 96% of the records have a low income ($50,000 or less), so our confidence is 96%.

   Leaf Node #5:

   - Antecedent - The customer is *not* married (branch to node 2) *and* has capital gains of $7,074 or more (branch to node 5).
   - Consequent - The customer has a high income (more than $50,000)
   - Support - The proportion of all customers that make it to this leaf is 1%.
   - Confidence - Leaf node 5 reports that 98% of the records have a high income (more than $50,000), so our confidence is 98%.

   Leaf Node #7:

   - Antecedent - The customer *is* married (branch to node 3) *and* has an education level of at least some college (branch to node 7).
   - Consequent - The customer has a low income ($50,000 or less).
   - Support - The proportion of all customers that make it to this leaf is 25%.
   - Confidence - Leaf node 7 reports that 60% of the records have a high income (more than $50,000), so our confidence is 60%.

   Leaf Node #12:

   - Antecedent - The customer *is* married (branch to node 3), has an education level a high school diploma or less (branch to node 6), and has capital gains of less than $5,096 (branch to node 12).
   - Consequent - The customer has a high income (more than $50,000)
   - Support - The proportion of all customers that make it to this leaf is 19%.
   - Confidence - Leaf node 12 reports that 76% of the records have a low income ($50,000 or less), so our confidence is 76%.

Leaf Node #13:

- Antecedent - The customer *is* married (branch to node 3), has an education level a high school diploma or less (branch to node 6), and has capital gains of $5,096 or more (branch to node 13).
- Consequent - The customer has a high income (more than $50,000)
- Support - The proportion of all customers that make it to this leaf is 1%.
- Confidence - Leaf node 13 reports that 96% of the records have a high income (more than $50,000), so our confidence is 96%.

Assuming our client is interested in high-income customers, the decision rule of greatest interest is clearly the rule for leaf node 7. Leaf node 7 contains the vast majority of all high-income customers. The other two high-income leaf nodes (nodes 5 and 13) only account for 2% of the data set while node 7 accounts for 25%.

7. **Using your decision tree, decision rules, and EDA, develop <u>detailed</u> profiles of:**
   a. **High-income customer.**
   b. **Low-income customer.**

A typical *high-income* customer is an individual who is married, has an income greater than $50,000 per year, and has an educational level of at least some college.

- All high-income earners make in excess of $50,000 per year, as that is the determining factor for the high-income category.
- The high-income customers make up ~24% of all customers in the data set.
- The overwhelming majority of high-income customers are married. The proportion of high-income earners who are married is well over *90%*, with 26% of the overall sample being married with a high-income, while those not married with a high-income are only 1% of the data set.
- A strong majority (at least 90%) of high-income customers have at least some college experience.

A typical *low-income* customer is an individual who is not married, has an income of $50,000 or less, and has capital gains totaling less than $7,074.

- All low-income earners $50,000 per year or less, as that is the determining factor for the low-income category.
- The low-income customers make up ~76% of all customers in the data set.
- Nearly three-quarters of low-income customers are not married. Those low-income customers who are not married account for 54% of the data set, while low-income married customers account for only 19% of the overall data set.
- All low-income customers have capital gains of less than $7,074 or less per year.
- Of those low-income customers who *are* married, all have a high school diploma or less in educational attainment.

- **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 3 of Project 3.**
    b. **Problem 5**
    c. **Problem 6**
    d. **Problem 7**

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_Project4* (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_Project4_RScript.***

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