DATA 511 Intro to Data Science
Course Project 3
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.)

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.
Impute the data values for capital.gain = 99999, just like you did for Project 1.

No Executive Summary is required for this project.

Good luck!

Prof Larose

1. **Pain in the Drain Data Prep.**
   a. **Provide the summary of *income*, and mention the proportion of high income. On your own (not reproduced here), 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". The categories Assoc-acdm, Assoc-voc, and prof-school should belong to the high group. Delete *education*. Provide a prop.table of *educ* against *income*, rounded to two decimal places. Make sure the table uses column proportions, and *educ* is in the columns. Clearly describe the relationship between *educ* and *income*, using a dramatic statistic.**

Income (Proportion)	
$50K or Less	**More Than $50K**
75.92%	24.08%

There are 24,720 records in the low-income category, defined as having an income of $50,000 or less. From the proportions table, these low-income records account for just over three-quarters of all records. There are 7,841 records in the high-income category, defined as those with an income of over $50,000, accounting for just under one-quarter of all records.

		Education Level	
		High	**Low**
Income	$50K or less	56.88%	85%
	More than $50K	43.12%	15%

In the above table comparing income and level of education, there clearly appears to be a positive correlation between higher income and higher level (at least a college degree) of education, whereas those with less than a college degree worth of education tend to earn less. Of those with a high level of education, 57% earn $50,000 or less, while 43% make more than $50,000. Of those low level of education, more than 8 out of 10 earn less than $50,000

   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. Provide a prop.table of *rel* against *income*, rounded to two decimal places. Make sure the table uses column proportions, and *rel* is in the columns. Clearly describe the relationship between *rel* and *income*, using a dramatic statistic.**

   c.

		Relationship Status	
		Husband and Wife	**Other**
Income	$50K or less	54.86%	93.38%
	More than $50K	45.14%	6.62%

From the table of relationship status against income, there appears to be clear evidence that a husband & wife relationship has a far greater likelihood at having an income of more than $50,000 than does those in a different relationship status (other).

Those in a husband/wife relationship are split fairly evenly, with 55% making $50,000 or less, while 45% make in excess of $50,000. Among those with a different relationship status, more than 9 in 10 make $50,000 or less, while less than 7% make $50,000 or more.

2. **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.**

Test Data Set	
$50K or Less	More Than $50K
12,360	3,921

Training Data Set	
$50K or Less	More Than $50K
12,360	3,920

In both the test and training data sets, the breakdown of those making $50,000 or less and those making more than $50,000 is approximately equal. In both data sets, 12,360 are in the $50,000 or less category, while ~3,920 are in the greater than $50,000 income category. Considering the records were partitioned 50/50, this is the expected output.

   b. **Validate the partition for the imputed *capital.gain* (that is, *cg.imp*) and for *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.**

Kruskal-Wallis for Capital Gains:

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

Kruskal-Wallis for Capital Losses:

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

c. **Again, set your seed to 12345. 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_pt_table
X-squared = 1.465e-05, df = 1, p-value = 0.9969
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01017792  0.01013825
sample estimates:
   prop 1         prop 2
0.3229531.  0.3229730
```

Relationship:

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

data:  rel_pt_table
X-squared = 0.0036854, df = 1, p-value = 0.9516
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01114928  0.01047936
sample estimates:
   prop 1    prop 2
0.4531663 0.4535012
```

d. **What is your conclusion regarding the partition?**

The box plots are very similar and the p-values for both the capital gains (.2975) and capital losses (.9514) from the Kruskal-Wallis test are much larger than the .05 level of significance. Therefore, we can conclude that the partition for capital gains and losses is validated.

From the two-sample test for equality of proportions (Chi-Square Test), we see p-values of .9969 for education and .9516 for relationship status. Both are far above the .05 significance level, so we can conclude that the partition for education and relationship status variables is validated.

3. Step 2 of the CMBM. Establish baseline model performance, using the training data set. Manually construct the two baseline contingency tables, nicely formatted, using MS Word. Make sure to include the variable names and categories. Use the accuracy metric. What accuracy will our model have to beat?

All Positive Model:

		Predicted Income		
		False	True	Total
Actual Income	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 = (TN + TP) / GT = (0 + 3,921) / 16,281 = 24.08%

All Negative Model:

		Predicted Income		
		False	**True**	**Total**
Actual Income	**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 = (TN + TP) / GT = (12,360 + 0) / 16,281 = 75.92%

The classification model that we develop will have to beat its accuracy of *75.92%*.

Well done!

Deliverables:

1. **Save your completed Word document as a pdf file, named *Doe_Jane_Project3* (if your name is Jane Doe, with last name first!). 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.