MATH 4397, Intro to Data Science & Machine Learning, Exam # 2.

Your name:

Guidelines:

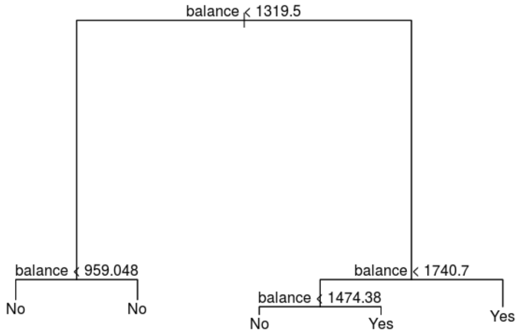
- Closed-book, no unauthorized help allowed.
- Write your answers in the blanks between questions. Be very brief and to the point.
- Use the last sheet as scratch paper.
- DON'T write final answers on scratch paper - those won't be graded. If you run out of space in the blanks between questions - write on the side on the worksheets. Once again - DON'T use scratch paper for final answers.
- Each problem is worth 40 points, each point is worth 1% towards your Exam 2 grade. Hence you get potentially get 120% for this exam.

PROBLEM #1.

Working for a credit card company, we would like to predict whether a customer will default on his/her payment ($default = Yes/No$). We look at such predictors as customer's credit card balance ($balance$), annual income ($income$), student status ($student = Yes/No$).

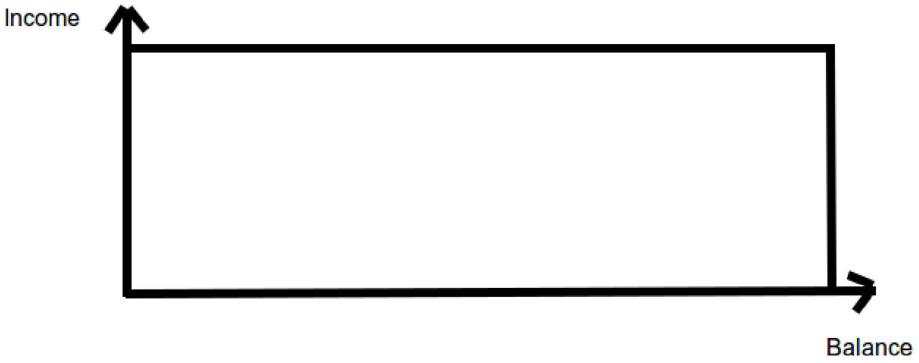
1. What' the type of the response variable? Hence, is it regression or classification task?

2. The following decision tree resulted from fitting response $default$ to the three aforementioned predictors:



   (a) Which is the only predictor that plays a role in predicting $default$?

   (b) Draw a predictor space segmentation corresponding to this tree (round split values to integers). Write the tree-predicted Yes/No in the middle of each region. Hint: all of your splits will be dictated by just one of these two predictors.



(c) Interpret the following summary of a terminal node:

```
5) balance>959.048   276    181.20   No   (0.898551, 0.101449) *
```

How many customers are in this node? What is the range of their *balance* values? What is the proportion of customers that defaulted ($default = Yes$) in this node? What is the overall $default$ prediction for this node - Yes or No?

3. (a) In very few words: What is meant by tree pruning? Why do we do it?

(b) Having performed cost-complexity pruning, we got:

```
$size                  # Tree sizes (no. of terminal nodes in a tree).
[1] 5 4 2 1
$dev                   # Corresponding CV errors.
[1] 128 122 155 267
```

What tree size would you pick? Why (few words)?

(c) Having pruned the tree down to 4 terminal nodes, we get the same test classification error - 10% - for both the large tree (with all 5 terminal nodes) and the pruned tree (with 4 nodes).

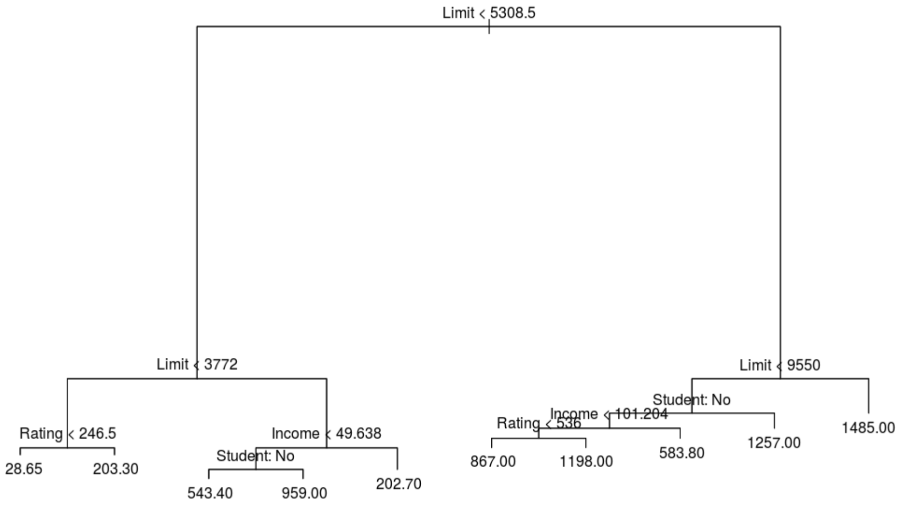
Given the fact that the terminal node eliminated during pruning did not affect the classification performance - which node was it? Either circle it on the tree plot in part 2, or provide it's split criteria here.

PROBLEM #2.

We, once again, are working for a credit card company, but this time we would like to predict customer's exact credit card balance (*balance*). We will use a total of 11 predictors, including income, limit, gender, marriage status, and others.

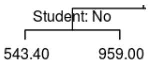
1. What's the type of the response variable? Hence, is it classification or regression task?

2. Fitted decision tree for response *balance* onto 11 predictors:



   (a) List all the predictors that appear in the tree.

   (b) How do we interpret (in plain English) this terminal node:



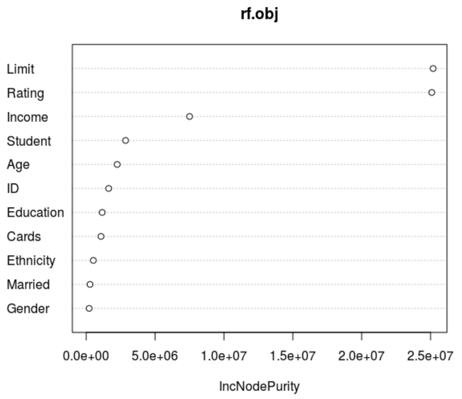
   What is 543.40? What is 959.00?

3. (a) The mean squared prediction error from single decision tree was 35618.33. What's the name of the approach that can potentially improve the prediction compared to just using a single decision tree? Proceed to outline the three main steps (as we did in class, in review) of this approach in order to obtain predictions.

(b) The resulting prediction test error of the approach we described in part 3(a) is 6407. Is it larger or smaller than the one we mentioned for a single tree?

(c) Random forests apply a tweak to the method we discuss in parts 3(a, b) - what tweak is it (in a few words)? What does this tweak try to achieve (in 2-3 words)?

(d) The variable importance measures of fitted random forest are provided below:



List four most important variables.

Does that list fully correspond to your answer in part 2(a) (Yes/No)?

PROBLEM #3 (a couple of disjoint questions).

1. (a) Name two approaches of estimating a test error of a statistical model.

   (b) Explain how k-fold Cross-Validation is implemented as we did in class, in review (three steps, second step has multiple substeps). No need for precise formulas, just the ideas of actions performed and names of quantities being calculated.

2. Presume you are asked to estimate a median height for a population of UH students, while only having access to a random sample $\mathbf{x} = (x_1, \ldots, x_{500})$ of 500 students. You are able to obtain sample median $med(\mathbf{x})$, but you need to evaluate its standard error.

   (a) Name the resampling method that will help you evaluate that standard error.

   (b) Given this sample $\mathbf{x}$ and statistic of interest $\hat{\alpha}(\mathbf{x}) = med(\mathbf{x})$, provide three steps of this method as described in class, in review.

3. Presume that for a single linear neuron model with input variables $x_1, \ldots, x_3$, you are given the following parameter values:

   - weights $w_1 = 0.5, w_2 = -0.5, w_3 = -0.2$,
   - bias $b = -1$.

   Then proceed to

   (a) Draw a mathematical model of this linear neuron that takes an arbitrary input vector $\mathbf{x} = (x_1, x_2, x_3)$.

   (b) Calculate the linear neuron output for the case of $x_1 = 2, x_2 = -4, x_3 = 5$. Show your work.