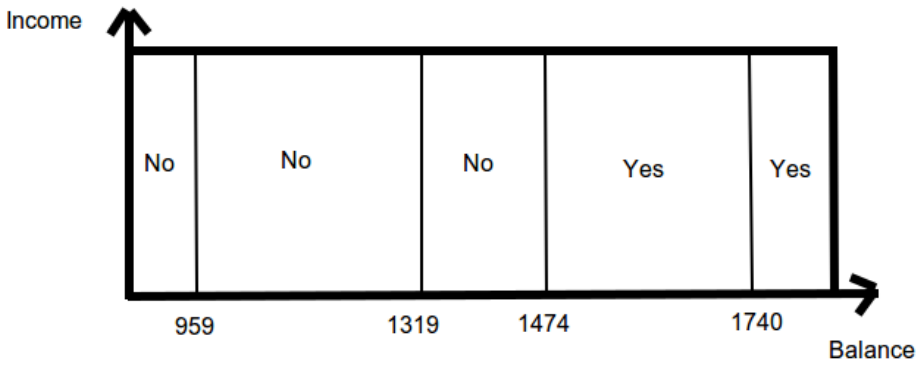
MATH 4397, Intro to Data Science & Machine Learning, Exam # 2, Solutions.

PROBLEM #1.

1. Qualitative, categorical, factor - either would work. It is a classification task.

2. (a) *balance* predictor.

   (b)



   (c) There are 276 customers in this node. Their range of *balance* values is $balance > 959.048$, or $balance \in (959.048, 1319.5)$ to be more precise, but I counted both answers. Proportion of defaulted customers is 0.101449. The overall prediction is $default = No$, the customer won't default.

3. (a) Tree pruning consists of selecting subtrees of the large tree to minimize the CV error. It is done to avoid overfitting.

   (b) Optimal tree size is 4, because it corresponds to smallest CV (122).

   (c) It is the terminal node corresponding to $balance < 959.048$ split.

PROBLEM #2.

1. It is quantitative, continuous, numerical - either would work. Hence, it is a regression task.

2. (a) Predictors: *Limit*, *Rating*, *Student*, *Income*.

   (b) For non-students we predict a 543.40 credit balance (543.40 is the average credit balance of all non-students in this node).

   For students, we predict 959 credit balance (959 is the average credit balance of all students in this node).

3. (a) Bagging. See exam review for the steps.

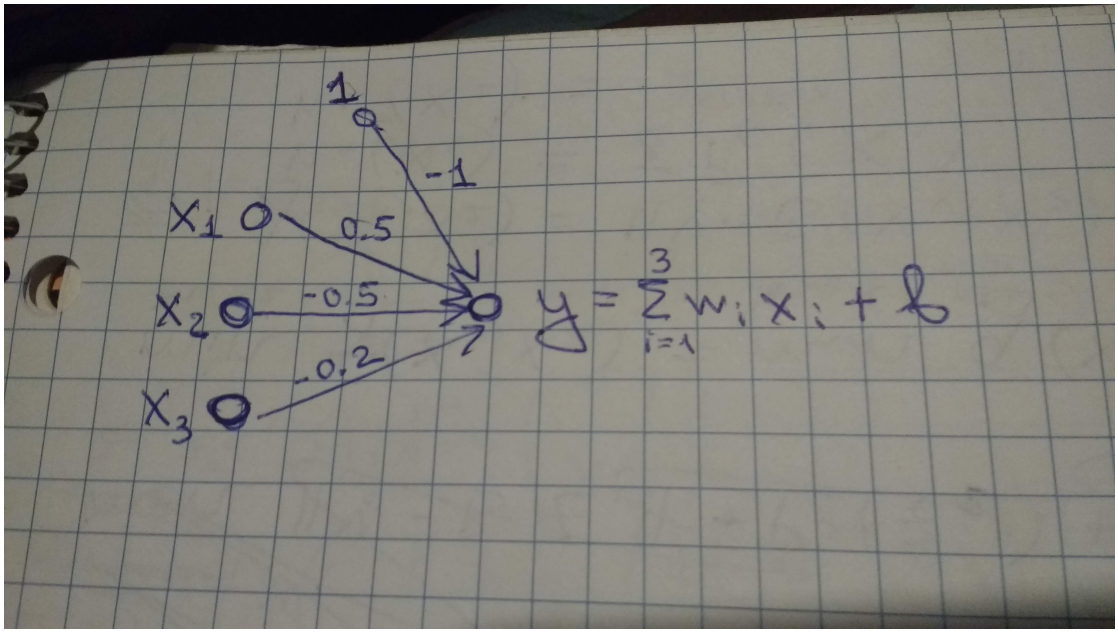
   (b) Smaller.

   (c) Random forests look at random subsets of variables for tree splits. This tweak tries to decorrelate the trees, reduce the estimate variance.

   (d) Four most important predictors: *Limit*, *Rating*, *Income*, *Student*. Yes, it fully corresponds to the list in part $2(a)$.

PROBLEM #3 (a couple of disjoint questions).

1. (a) 1) Validation set approach. 2) K-fold Cross-Validation approach.
   (b) See the exam review for these steps.
2. (a) Bootstrap.
   (b) See these steps in the review.
3. (a)



(b) $\hat{y} = \sum_{i=1}^{3} w_i x_i + b = 0.5 \times 2 + (-0.5) \times (-4) + (-0.2) \times 5 + (-1) = 1 + 2 - 1 - 1 = 1$