Midterm Exam - 2261

Started: Oct 9 at 6:41pm

Quiz Instructions

Hello,

Once you start the exam, the timer will start and run for 3 hours.

You will be given two files when you start, a Quarto file and a csv dataset. Download these to a local working directory. Although you will need to initially read, run and evaluate the work in the Quarto document in R-Studio, you will have to modify the Quarto document (edit it, making additions and changes) quite a lot with your own data wrangling and analysis in R-Studio to complete the exam. You will be asked to render it and to upload the Quarto file and its rendering at the end of the exam.

The exam consists of questions in Canvas, some of which require uploads of images and the last two questions are file uploads of the quarto-rendered MS Word (docx) file and your final version of your quarto file.

You are allowed to use the course materials (presentations, assignments, canvas materials and qmd files) and google search for R code *only as a last resort* as you should use the functions we covered in class; using functions that we did not cover in class for models and analysis will receive no credit when the exam is graded (they will not be graded)

You are NOT allowed to communicate or collaborate with each other or ask for anybody's help on the exam, it is entirely individual. Do not use any AI tools outside of those that may be embedded in Google Search, but note the caution with this approach stated above (in red).

Hello,

Here are the Quarto Notebook and the accompanying dataset you will need for this exam.

2251 - Data Mining Midterm.qmd (https://canvas.pitt.edu/courses/324587/files/21109282?wrap=1) (https://canvas.pitt.edu/courses/324587/files/21109282/download?download_frd=1)

driverschurn.csv (https://canvas.pitt.edu/courses/324587/files/21109380?wrap=1) (https://canvas.pitt.edu/courses/324587/files/21109380/download?download_frd=1)

Download both files to your working folder for this exam.

Although you need to initially read and evaluate the work in the Quarto document, you will also have to modify it (edit it, making additions and changes) quite a lot with your own data wrangling and analysis to complete the exam. You will be asked to render it, and to upload the Quarto file and its rendering at the end of the exam.

The exam consists of questions in Canvas, some of which require uploads of images and the last two questions are file uploads of the quarto-rendered MS Word (docx) file and your final version of your quarto file.

You are allowed to use the course materials (presentations, assignments, canvas materials and qmd files) and google search for R code.

You are allowed to use the course materials (presentations, assignments, canvas materials and qmd files) and google search for R code *only as a last resort* as you should use the functions we covered in class; using functions that we did not cover in class for models and analysis will receive no credit when the exam is graded (they will not be graded)

You are NOT allowed to communicate or collaborate with each other or ask for anybody's help on the exam, it is entirely individual. Do not use any AI tools outside of those that may be embedded in Google Search, but note the caution with this approach stated above (in red).

Begin now by stepping through the quarto file in R-Studio, reading the text and executing the R-code.

Then begin with the exam questions.

The exam timer is set to 3 hours and it has started and will not stop even if you exit the exam.

Question 1 5 pts

Please download and complete the Data Dictionary Table Template and upload it in the space below.

Data Dictionary Table Template.docx (https://canvas.pitt.edu/courses/324587/files/21109381?wrap=1) (https://canvas.pitt.edu/courses/324587/files/21109381/download?download_frd=1)

Upload

Choose a File

Question 2 5 pts

Why does Sandy Skeptic state that a linear regression is a poor model choice for Churn?

What types of models are more appropriate?

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

p 0 words </>

Question 3 5 pts

Reread the introductory paragraph regarding the problem and objective. Would the objective fit into an “explanatory” purpose or a “prediction” purpose?

○ Explanatory Purpose

○ Prediction Purpose

Question 4 10 pts

Based your answer above (explanatory or prediction), answer this question:

Will you need to have a training and testing split of the data for your analysis or not?

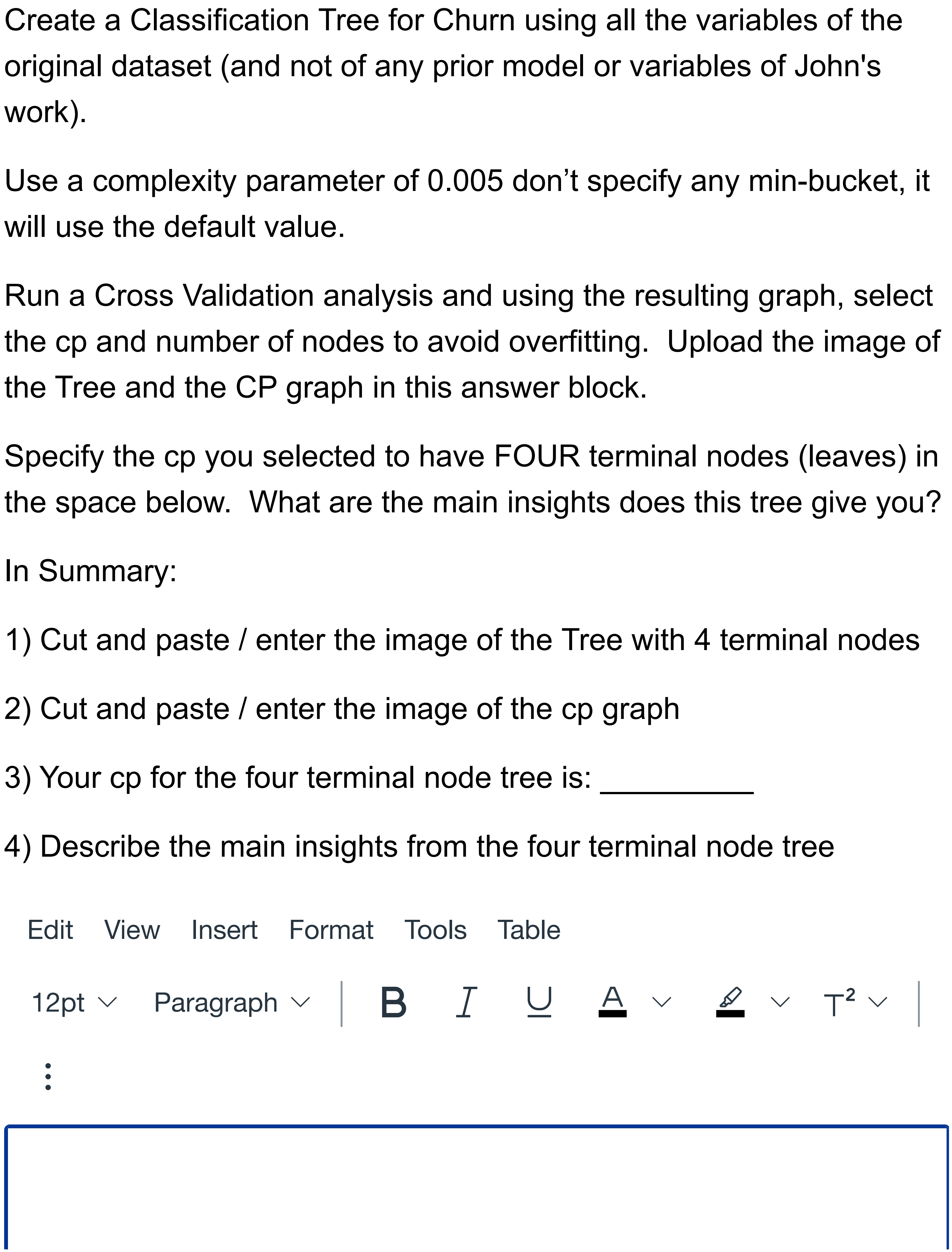
Please be sure to follow your answer to this question in all your analysis in R-studio. That is to say, if you answered that a split is needed, then use a split in your quarto work. Likewise, if you answered that you do not need a split of the dataset, then do not split the dataset (but use the entire dataset) in your quarto work for this exam.

○ No Split of the dataset is needed

○ The dataset should be split into training and testing.

Question 5 15 pts

Now you will start to code in the Quarto doc to complete many of the questions to follow. (Recall you will upload the Quarto doc when you finish.) Note that the original Quarto doc had only a one library() call, you may very well need to code in many other library() calls.



Create a Classification Tree for Churn using all the variables of the original dataset (and not of any prior model or variables of John's work).

Use a complexity parameter of 0.005 don’t specify any min-bucket, it will use the default value.

Run a Cross Validation analysis and using the resulting graph, select the cp and number of nodes to avoid overfitting. Upload the image of the Tree and the CP graph in this answer block.

Specify the cp you selected to have FOUR terminal nodes (leaves) in the space below. What are the main insights does this tree give you?

In Summary:

1) Cut and paste / enter the image of the Tree with 4 terminal nodes

2) Cut and paste / enter the image of the cp graph

3) Your cp for the four terminal node tree is: __________

4) Describe the main insights from the four terminal node tree

p

0 words

</>

Question 6 5 pts

Create a confusion matrix for this Classification Tree using confusiontMatrix() with the 1 being positive.

Cut and paste the Confusion Matrix and the key measures that reflect the model performance below.

Edit View Insert Format Tools Table

12pt Paragraph

p 0 words </>

Question 7 15 pts

Create a Logistic Regression for Churn using all the variables of the original dataset (and not of any prior model or datasets of John's work). The cutoff threshold should be 0.50.

Cut and paste the Coefficients from the model in the space below. Calculate the **exp(coefficient)** for these variables and cut and paste them in the space below.

Which variables (not including the intercept) are significant and how would you interpret the **exp(coefficient)** for these variables?

In Summary, after you complete the model and the **exp(coefficient) calculations**:

1) Cut and paste the model output with coefficients and their related data (that is to say: Est., Std. Error, z value, Pr(>|z|)

2) Cut and paste the **exp(coefficient) calculations**

3) Which variables are significant (do not include the intercept)?

4) How would you interpret the significant variables' **exp(coefficient)?** (do not include the intercept)

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

p 0 words </>

Question 8 5 pts

Create a confusion matrix for this Logistic Regression model, with a cutoff of 0.5, using confusiontMatrix() with the 1 being positive.

Cut and paste the Confusion Matrix and the key measures that reflect the model performance below.

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

p 0 words </>

Question 9 10 pts

Is there better cutoff for the Logistic Regression? Let's change the cutoff so that the Logistic Regression model has a new Sensitivity measure that is close to the Sensitivity of the Classification Tree. Here is the process to follow:

First, create the ROC curve graph and calculate the AUC, and cut and paste / enter it below.

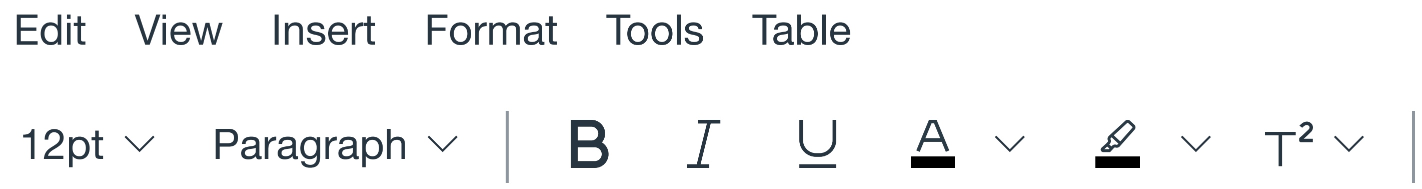
Based on the chart, select a cutoff that would have a Sensitivity that is close to the Classifaction Tree's Sensitivity (say within + / - 3%). You may need to try a few times to get the Sensitivity in the correct range.   Please redo the Log Reg with this cutoff and display the confusion matrix results.

In the answer box below:

1) Cut and paste the ROC graph and the AUC calculation

2) State which cutoff you used to get the Logistic Regressions' Sensitivity near that of the Classification Tree's Sensitivity.

3) Cut and paste Confusion Matrix and the key measures that reflect the model performance with the new cutoff.



p 0 words </>

Question 10 10 pts

Now you have a Classification Tree and Logistic Regression model that both have a similar sensitivities.

Which model do you prefer using, based on the confusion matrix performance metrics? Explain.

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

p 0 words </>

Question 11 10 pts

Now, considering the case competition's overall objective, how would you summarize the results you have developed in the context of churn? (Do this in the answer text box.)

What analysis and charts / reports would you use to support your discussion?

Edit View Insert Format Tools Table

12pt Paragraph

p 0 words </>

Question 12 5 pts

What are some concerns or improvement opportunities you may have with this overall analysis?

Said another way, if you wanted to have more confidence in your findings, what would you prefer to have or do for this analysis?

Give at least 2 ideas for full credit.

Edit View Insert Format Tools Table

12pt Paragraph **B** *I* U A T²

p 0 words </>

Question 13 0 pts

REQUIRED: Upload the Word DOCX file that is rendered from your final version of your Quarto document.

While there are no points with this upload, but it is required to grade your exam.

Upload

Choose a File

Question 14 0 pts

REQUIRED: Upload your final version of your Quarto file (with the results of the code execution).

While there are no points with this upload, but it is required to grade your exam.

Upload

Choose a File

Quiz saved at 6:55pm Submit Quiz