Exam PA December 12, 2019 Project Statement

General information for candidates

This assignment has two components. One is a statement of the business problem to be addressed. The other is a list of tasks to be done. Your report will consist of responses to eleven specific tasks followed by an executive summary. The audience for the task responses is the examination grading team. Hence, technical language can be used, unless the task explicitly asks for a different approach. Each task will be graded individually, so be sure any work that addresses a given task is done within the writeup for that task. The final item in your report is an executive summary written for an audience **not** familiar with analytics concepts.

This document and the report template indicate the points assigned to each of the twelve components. The total is 100 points. Each of the first eleven tasks will be graded on the quality of your thought process, added or modified code, and conclusions. The executive summary will be graded on the quality of the presentation and clarity of communication. Note that a component of the grading of the first eleven tasks will also relate to the quality of the exposition.

At times you will be instructed to include specific output (typically tables or graphs) in your response. These should not be the only times you display output in your response.

At a minimum you must submit your completed report template and an Rmd file that supports your work. Graders expect that your Rmd code can be run from beginning to end. The code snippets provided should either be commented out or adapted for execution. Ensure that it is clear where in the code each of the tasks is addressed. Documentation of your thought process and conclusion for each task should be completely contained within your Word report. The Rmd code should be clear, may contain commentary, and support your work.

You may submit other files as needed to support your work. In addition to Word (.docx) and RStudio (.Rmd) files, you may also submit Excel files (.xlsx or .csv). There is a limit of 10 files and no file can be larger than 25MB.

Business Problem

You are an actuary at MEB Insurance and have been asked to assist the marketing department. They have been collecting data on past policyholders. For each policyholder they have determined if they are "high value" or "low value" based on profitability. They would like to be able predict for each prospective policyholder if they will be high or low value.[1]

[1] Depending on the regulatory environment, an approach like this could be viewed as inappropriate. You should not be concerned with this issue when performing your analysis or writing your report.

Your task is to use the available data[2] to construct a model that will accurately predict if a potential customer will be high or low value. While predictive accuracy is more important to the marketing department than understanding the relationships of the predictors to the target variable, the marketing department will have more confidence in your work if the results make sense.

MEB has indicated that this analysis will only be applied to individuals age 25 and older.

Your assistant has done some preliminary analyses, which are scattered throughout the supplied Rmd file.

Specific Tasks

The tasks are intended to be done in order with results from one task informing work in later tasks. Graders will look for the solution to a given task within that task's area in the report and Rmd file.

In all cases you should justify the choices you make in your report.

1. (*12 points*) Examine each variable and make appropriate adjustments.

   Examine each predictor variable other than cap_gain both on its own and with respect to value_flag. Make appropriate adjustments. Do not make any adjustment to the cap_gain variable at this time.

   There should be no further variable adjustments unless specifically requested.

2. (*10 points*) Construct a classification tree.

   Employ cost-complexity pruning to construct the best possible single tree. Use an ROC and a confusion matrix on the test set to evaluate the tree. Your assistant has provided two tree building approaches:

   - Build a tree by setting various parameters. Determine the best tree by varying the parameters that are set in the code. Any parameter not set in the code is to be left at its default value.
   - Build a tree by using cross validation to set the cp value. Explain how cross validation selects the cp parameter and compare that approach to the trial-and-error approach used to build the first tree.

   After building both trees, recommend one of them to remain in consideration for your final model.

   For the recommended tree, do the following:

   - Provide the following output in your report (other output may also be provided):
     - List the variables that are actually used in making splits.
     - Display the tree as a figure.
     - Display output that shows all splits in list form.

[2] The data are adapted from the "Adult Data Set" contributed by R. Kohavi and B. Becker to the UCI Machine Learning Repository. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

- Interpret the tree in a way that would explain it to the marketing department should it be selected as the final model.
- Based on the tree, recommend one interaction to be used when constructing a GLM in a later task.

3. (*6 points*) Construct a random forest.

   Construct a random forest model. Use the parameters provided in the code chunk to ensure that excess run time is not used. Evaluate the model against the test set.

   Include a variable importance plot or table in your report. Provide an interpretation.

4. (*4 points*) Recommend a tree model.

   Recommend either your single classification tree from Task 2 or your random forest from Task 3 as the best choice (so far) for your final model. Provide and interpret the ROC and the AUC value for your selected tree.

   Your choice should be based on the business problem as well as the output.

5. (*5 points*) Consider a boosted tree.

   Do not construct a boosted tree. Describe both a random forest and a boosted tree and explain the similarities and differences.

6. (*5 points*) Convert cap_gain to a factor variable.

   Prior to constructing a GLM, create a factor variable based on cap_gain using intervals. Note that both the number of categories and the boundaries in the provided code are illustrative examples, not guidelines. Use the output from the tree constructed in Task 2 to inform your choice of boundaries. Use this factor variable for the GLM.

   Explain why this new variable may be more useful in a GLM than cap_gain as provided.

7. (*3 points*) Select a distribution and link function for a GLM.

   Determine a distribution and link function to use. Justify your choice of distribution based on the business problem and data and use only that combination for all further work. Do not justify by trying various combinations and then selecting the one that performs best.

8. (*10 points*) Select variables to use in the GLM.

   Use an appropriate (with justification) variable selection method to construct and evaluate a GLM (but do not conduct a regularized regression). Be sure to include the interaction identified in Task 2. For variable selection, select from AIC or BIC, forward or backward selection, and use an approach that retains or drops factor variables in their entirety (as opposed to dropping individual factor levels).

   After completing the process, include the following in your report:

   - The variables selected in your final model and their coefficients.

- Compare your selected variables to those used in the Task 2 tree and the important variables from the Task 3 random forest. Comment on any differences.
- Discuss the advantages and disadvantages of using binarization on the factor variables to allow dropping individual levels. Do not perform this analysis.

9. (*5 points*) Select the final model

Recommend which model should be used, the tree selected in Task 4 or the final GLM from Task 8. Justify your choice.

10. (*10 points*) Select the cutoff probability to maximize expected profit.

In cooperation with MEB's finance department, the marketing department has determined the following profits or losses that will result from the possible outcomes of using your model. The units are arbitrary.

- Market the policy to an applicant who turns out to be high value – profit of 50
- Market the policy to an applicant who turns out to be low value – loss of 25
- Do not market to an applicant – loss of 5

Your assistant has provided code that allows you to determine the expected profit based on a selected cutoff. Use the model selected in Task 9 for your calculations and demonstrate that you have obtained the cutoff that maximizes expected profit (two decimal places are sufficient). Perform calculations on the test set.

After determining the cutoff, write an explanation for marketing that explains the notion of a cutoff and how the confusion matrix leads to your estimated expected profit. This explanation should be more detailed than what will go into your executive summary (which should contain only a brief description and focus on the results).

11. (*10 points*) Write model demo for marketing

Marketing has asked for a demonstration of how your model is to be used with examples of cases that predict high value and cases that predict low value. Your assistant has prepared some sample cases that can be run through your model. You may need to adjust some of them to obtain illustrative examples that would be of interest to marketing.

Write, in language appropriate for marketing, the illustration and demonstration they are looking for. This demonstration should be more detailed than what will go into your executive summary (which could include an example).

The sample cases are provided here and in your report template in case you wish to include them in your report.

age	education_num	marital_status	occupation	cap_gain	hours_per_week	score
39	10	Married-civ-spouse	Group 3	0	40	60
53	10	Married-civ-spouse	Group 3	0	40	60
39	13	Married-civ-spouse	Group 3	0	40	60

39	10	Never-married	Group 3	0	40	60
39	10	Married-civ-spouse	Group 5	0	40	60
39	10	Married-civ-spouse	Group 3	2000	40	60
39	10	Married-civ-spouse	Group 3	0	50	60
39	10	Married-civ-spouse	Group 3	0	40	64

12. (*20 points*) Executive summary

Your executive summary should reflect the information provided and work from Tasks 1-11 as relevant to the marketing department. Your executive summary should include a problem statement, discussion of data, a coherent explanation and justification of your recommended model, and conclusions.

Data Dictionary

age	The age of the prospective policyholder	An integer from 17 to 90
education_num	An indicator of the amount of education – it is not the number of years of education, but a higher number does mean more years	An integer from 1 to 16
marital_status	For married, AF means alternative form while civ means civil	A factor variable with seven levels
occupation	Occupations have been grouped into five categories. There is no indication regarding what they mean. A sixth group represents cases where the occupation is unknown.	A factor variable with six levels
cap_gain	Capital gains recorded on investments	An integer from 0 to 99,999
hours_per_week	The number of hours worked per week	An integer from 1 to 99
score	A proprietary "insurance score" developed by MEB.	A real number with two decimal places
value_flag	In indicator a policyholder being High or Low value	A factor variable with two levels