Exam PA June 18, 2020 Project Statement

IMPORTANT NOTICE – THIS IS THE JUNE 18 PROJECT STATEMENT. IF TODAY IS NOT JUNE 18, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.

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 twelve 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 thirteen components. The total is 100 points. Each task 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. Note that a component of the grading of the first twelve 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. 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 ABC Insurance and have been asked to assist the marketing department. They have collected data on a completed marketing campaign for a particular insurance product. The campaign consisted of calling prospective customers. For each call a variety of data was collected, including whether the prospect purchased the product. They would like you to advise them on future campaigns so that their efforts will be most productive in terms of purchases.

Your task is to use the available data[1] to construct a model that will generate insights into the productivity of marketing calls for the same insurance product. Your model will not be used to predict the outcome of future calls. Instead, the marketing department wants advice regarding possible adjustments they can incorporate into future marketing efforts.

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. (*8 points*) Explore the data.

   Your assistant has built the following charts: a histogram for `irate`, bar charts for `month` and `edu_years`, a boxplot of the `age` distribution by different values of `job`, and a graph showing the proportion purchasing by `age` and `month`.

   - Run the code to make them.
   - Include them in your report.
   - Discuss the impact of what each chart shows for your later modeling.

2. (*3 points*) Consider the education variable.

   Your assistant had created a new numeric variable, `edu_years`, from a categorical variable using the following rules to map each category to a number.
   - basic.4y: 4
   - basic.6y: 6
   - basic.9y: 9
   - High school: 12
   - Illiterate: 1
   - Professional Course: 13
   - University degree: 16

   Justify the choice of using a numerical variable instead of a categorical variable. Discuss the implications of that choice when running a generalized linear model (GLM) compared to running a tree model.

3. (*5 points*) Handle missing values.

   - Check the missing value proportion for all the variables.

[1] The data are adapted from the "Bank Marketing" dataset contributed by Moro, Cortez, and Rita (2014) to the UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

- Also consider the relation between `purchase` and missing values for each variable that has missing values. For each variable that has missing values, your assistant has created a table showing the mean response for missing values and the mean response for when the variable has values.
- Based on the results of these tables determine which of the following methods you want to use for each variable. Justify your choices and then implement them so no missing values remain.
    - Remove rows
    - Remove columns
    - Convert missing values to "unknown" (this only can be done for factor variables)
    - Impute the missing value using mean (this only can be done for numeric variables)

4. (*3 points*) Investigate correlations.

- Create a correlation coefficient matrix for all the numeric variables in the dataset.
- Among these pairwise correlations, determine which correlations concern you in building GLM and tree models. The response may differ by model.
- State a method other than principal components analysis (PCA) that can be used to handle the correlated variables. Do not implement this method.

5. (*7 points*) Conduct a principal components analysis (PCA).

Your assistant suggested PCA might be a good way to handle correlated variables.

- Perform a PCA on the following variables: `CPI`, `CCI`, `irate`, and `employment`.

After running the PCA, do the following:

- Create a bi-plot of the first two principal components and explain the loadings of the first two principal components.
- Explain why it is appropriate not to consider `age` and `edu_years` for the PCA.
- Only the first principal component will be used. Run the code to attach it to the data frame.

At this point, run the code to split the data into train and test sets.

6. (*5 points*) Create a generalized linear model (GLM).

- Fit a first GLM using a binomial distribution with a logit link function on the training dataset. Use `age` as the only explanatory variable for `purchase`.
- Run the summary function on the first GLM and provide the output in your report.
- Fit a second GLM using a binomial distribution with a logit link function on the training dataset. Use all the explanatory variables, including the first principal component, and excluding the four original variables used for the PCA.
- Run the summary function on the second GLM and provide the output in your report.

The marketing department thought `age` would be a significant feature in determining whether a prospective customer would make the purchase.

- The significance of `age` is different in the two GLMs. Explain why this is the case.

7. (*8 points*) Select features using stepwise selection.

   Your assistant has suggested that adding the square of age as a feature may improve the model.

   - Explain, without running any models, why this is or is not a reasonable suggestion. Regardless of your conclusion, add the square of age to subsequent models when asked to.

   Some of the features may lack predictive power and lead to overfitting. Best subset selection fits separate GLMs for each possible combination of features and selects the best combination.

   - Discuss the advantages and disadvantages of forward or backward selection compared to best subset selection.
   - Perform selection using the stepAIC function to determine which features should be retained. The code employs backward selection using AIC. Explain, without running any models, why this is or is not a reasonable approach based on the business problem. Regardless, employ this approach for the analysis.
   - Include the square of age variable when performing the backward selection.
   - Run the summary function on the GLM selected by this process. Include it in your report and list the variables selected.

8. (*12 points*) Evaluate the model.

   - Evaluate the model from Task 7 on both the train and test data. As part of your evaluation, output the AUC for both sets of data.
   - Explain what an AUC of near 0, 0.5, and almost 1 indicates.
   - Compare the selected features as they relate to the data analysis performed in earlier tasks. Comment on how any results are different than expected from the data exploration in Task 2 and the second GLM in Task 6.
   - Interpret the coefficients from the model from Task 7 in a manner that will provide useful information to the marketing department. (This part is worth 5 points)

9. (*9 points*) Investigate a shrinkage method.

   Elastic net could also be used in feature selection.

   - Explain how elastic net performs feature selection.
   - Use elastic net with alpha equal to 0.5 to create a regularized regression. Use the same variables as the second GLM in Task 6 plus the square of age.
   - Explain why using alpha = 0 is not appropriate if the goal is feature selection.
   - Evaluate this model on both the train and test data. As part of your evaluation, output the AUC for both sets of data.
   - List the features selected and compare the selected features to those using stepwise regression in Task 7. Comment on any differences.

10. (*7 points*) Construct a decision tree.

Note that the code does not use the constructed PCA variables or the square of age. Instead, three of the underlying variables are used.

- Justify the choice of using underlying variables instead of the constructed PCA variables.
- Two of them, `irate` and `employment`, are very highly correlated. Your assistant recommends that you only use `irate`. Justify the elimination of `employment`.
- Explain why adding age squared will not improve the decision tree.

Use the parameters provided in the code chunk to ensure that excess run time is not used.

- Run the code without changes.
- Plot the ROC curve and calculate the AUC on the train and test sets. Discuss any differences in these between the train and test sets and what they signify.

11. (*9 points*). Employ cost-complexity pruning to construct a smaller tree.

- Review the complexity parameter table for this tree. State the optimal complexity parameter and the number of leaves that will result if the tree is pruned using that value.
- Prune the tree using a complexity parameter that will result in eight leaves. If eight is not a possible option, select the largest number less than eight that is possible. Insert the selected complexity parameter in the code, replacing XX, and create tree2 by pruning tree1. This is the only code alteration for this task.
- Plot the ROC curve and calculate the AUC on the train and test sets.
- Discuss any differences in these between this tree and the previous tree in task 10 and what these differences signify.
- Include the pruned tree diagram.
- Consider the two leaves from the pruned tree that account for the largest proportion of the sample. State two insights they give, if any, about marketing the insurance product.

12. (*4 points*) Choose a model.

Select one of your models from tasks 7, 9, 10, or 11 to use for advising the marketing department. Explain your choice.

13. (*20 points*) Write an executive summary.

Your executive summary should reflect the information provided and work from Tasks 1-12 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. Be sure to include recommendations regarding future marketing efforts from the model chosen in Task 12.

Data Dictionary

Name	Description	Values
`age`	Age of the prospective customer	Integer from 17 to 98
`edu_years`	Years of education	Integer from 1 to 16

marital	Marital status	Factor with levels divorced, married, single
job	Occupation category	Factor with eleven levels
housing	Indicates whether the prospect has a housing loan	Factor with levels no, yes
loan	Indicates whether the prospect has a consumer loan	Factor with levels no, yes
month	The month of the marketing call	Factor with twelve levels
weekday	The day of the week of the marketing call	Factor with five levels
phone	The type of phone the prospect uses	Factor with levels cellular, landline
CPI	Consumer Price Index, at the time of the call	Numeric from 92.20 to 94.77
CCI	Consumer Confidence Index, at the time of the call	Numeric from -50.8 to -26.9
irate	Short term interest rate, at the time of the call	Numeric from 0.634 to 5.045
employment	Number of employees of ABC Insurance, at the time of the call	Numeric from 4964 to 5228
purchase	Indicator for purchase	Integer (1 for purchase, or 0 for no purchase)