Exam PA June 22 Project Statement

IMPORTANT NOTICE – THIS IS THE JUNE 22 PROJECT STATEMENT. IF TODAY IS NOT JUNE 22, 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, which are written in plain text. Additional information is in italics and applies to all tasks that come after it. Note that while the same business context applies for all tasks, the target variable may change from one task to the next, as indicated.

Your report will consist of responses to each of the specific tasks. Each task will be graded individually, so be sure any work that addresses a given task is done within the writeup for that task. Unless a task specifies otherwise, the audience for the task responses is the examination grading team and technical language can be used. When "for a general audience" is specified, write for an audience **not** familiar with analytics acronyms (e.g., RMSE, GLM, etc.) or analytics concepts (e.g., log link, binarization).

This document and the report template indicate the points assigned to each of the tasks. The total is 100 points. Each task will be graded on the quality of your thought process (as documented in your submission), conclusions, and quality of the presentation. No response to any task needs to be written as a formal report.

At a minimum, you must submit your completed report template and .Rmd file that supports your work. Please include June 22 and your candidate number (never your name) in your file names. Graders expect that your .Rmd file can be run from beginning to end. The .Rmd file should be clear and show all the code used to support your work. The code provided should either be commented out or adapted for execution as necessary. Make sure it is clear where in the code each of the tasks is addressed. Your thought process and conclusion for each task should be completely documented within your Word report.

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 work at ABC, a Canadian travel insurance provider that sells trip cancellation insurance, which covers pre-paid expenditures for a trip (flights, hotels, etc.) in the event of trip cancellation. The product is increasingly attractive to travelers as the cost of a trip increases.

You are part of a newly formed predictive modeling team in the marketing department at ABC to help the organization increase sales and profitability of their high-end product, which covers trips with at least $1,000 in covered costs. Recently, your manager recommended using a public data source to help

with your overall predictive modeling efforts – the Canadian National Travel Survey.[1] *A model built on this data may provide insight relevant to the marketing department regarding which travelers are in the target market for the high-end trip cancellation insurance. Your manager has provided the following data dictionary and 10,000 records from selected columns of the survey data in a file called* June 22 Data.csv*.*

Data Dictionary

Variable Name	Definition	Values
Distance	Distance traveled in trip, in km	6 - 4996
Duration	Number of nights spent on trip	0 - 81
Reason	Main reason for the trip	vacation: holiday, leisure, or recreation, visit: visit friends or relatives
Age	Age of adult survey respondent	1: 19-24 2: 25-34 3: 35-44 4: 45-54 5: 55-64 6: 65+
Others	Number of other persons that accompanied the respondent on the trip	0 - 52
Mode	Main mode of transportation	car, plane
Cost	Total spending on trip, in Canadian $	0 - 60170

Specific Tasks

The tasks are intended to be done in order with results from one task informing work in subsequent tasks. Graders will look for the solution to a given task only within that task's area in the report and, where applicable, the .Rmd file. Additional information given below in italics and not indented applies for all tasks that follow the information and is labelled accordingly. Items that further explain a specific task are indented and in plain text. Your reasoning and justifications should appear in your Word report.

Additional information for Task 1 and beyond

In addition to using the National Travel Survey, your manager has asked your team to propose other datasets that may prove useful in your predictive modelling work.

Your assistant has discovered the following datasets:

a. *Audio recordings from your customer service call center*
b. *Publicly available social media profiles and images of policyholders*
c. *Demographic information on the policyholder*
d. *Expected trip itinerary information of policyholders*

[1] Adapted from Statistics Canada, National Travel Survey, 2019. This does not constitute an endorsement by Statistics Canada of this product.

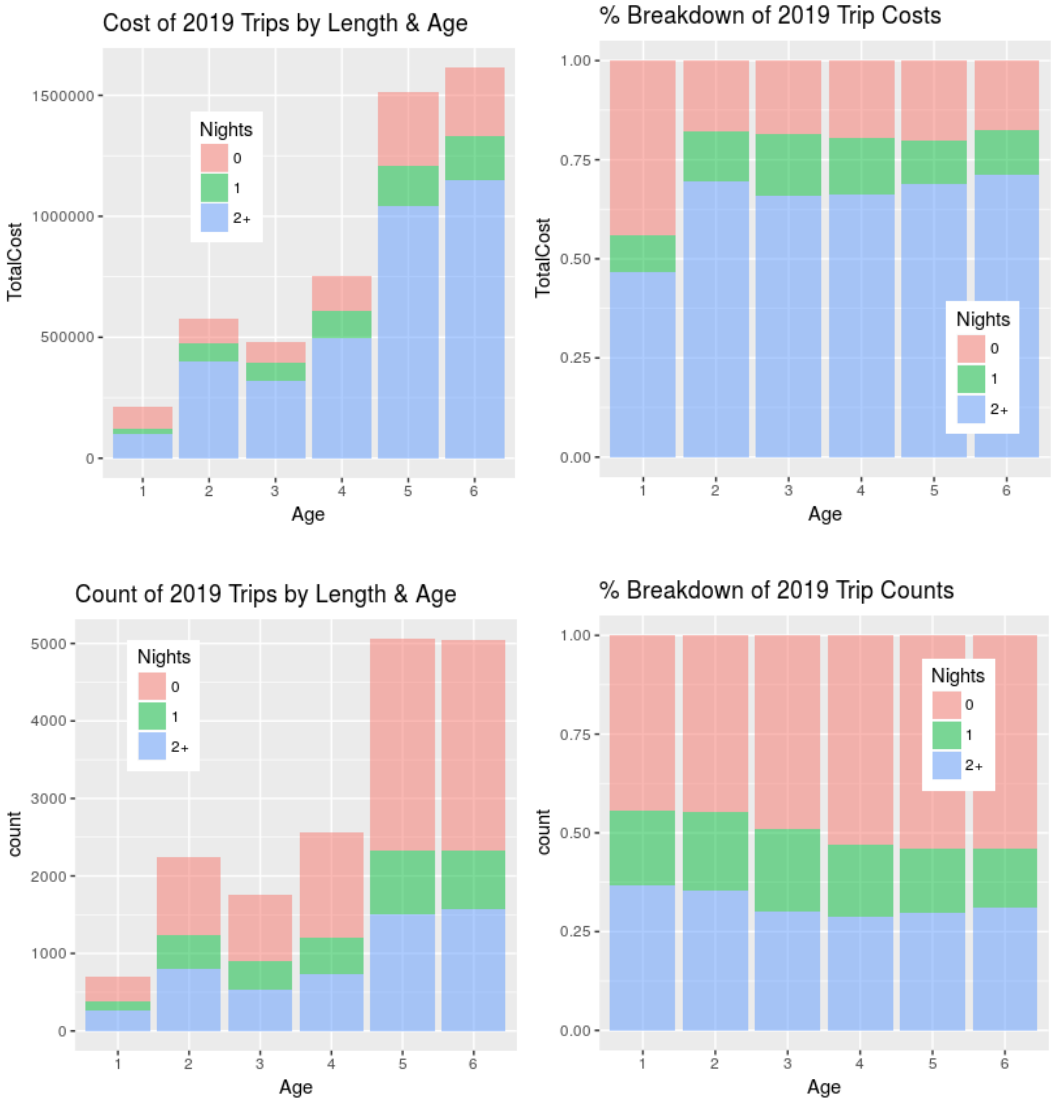
1. (*12 points*) Assess the data sources.

Perform the following:

- Classify each of the four data sources as structured, unstructured, or semi-structured. Justify your classifications.
- Assess, for a general audience, your manager, the tradeoffs of using each of the above data sources for this business problem. Your assessment should include considerations beyond the structure of the data but should not exceed one page.

Additional information for Task 2 and beyond

Before starting your modeling work you asked your assistant to summarize the data. As part of this work, your assistant produced the following four graphs on a similar but larger dataset.



2. (*12 points*) Interpret the graphs.

Describe for a general audience, your manager, three specific conclusions from interpreting one or more of the graphs above. For each conclusion, identify the specific graph or combination of graphs that supports it and describe which features of the graph(s) led to your conclusion.

Additional information for Task 3 and beyond

*Your assistant continues to explore the data. When reviewing your assistant's work, you notice that the **Cost** variable requires some attention before modeling proceeds. Your assistant's initial two graphs (which you can see if you run the code in the .Rmd file) depicting the variable do not fully reveal the issues with **Cost**. You alter one of the graphs to get a better view and then make a first edit on **Cost** (also in the .Rmd file). After creating a second graph, you also make a second edit on **Cost**. You then produce a third and final graph that conveys some important features of **Cost**. You meet with your assistant to discuss these graphs and edits, which are found in the .Rmd file.*

3. (*12 points*) Explain work on the Cost variable.

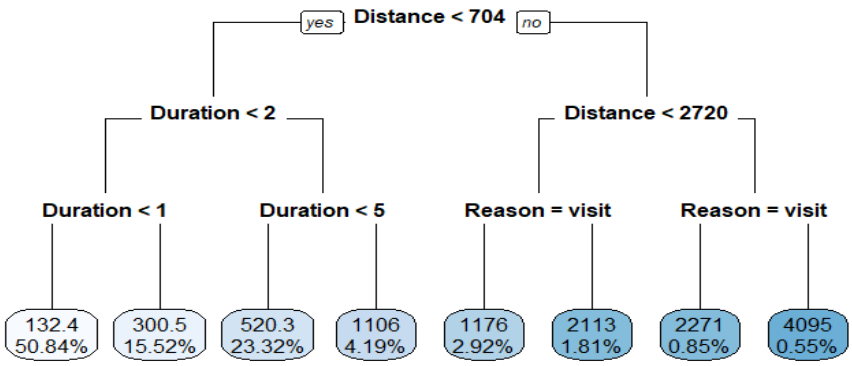
After running and considering the code in the .Rmd file, perform the following:

- Critique your assistant's two graphs.
- Justify your first edit.
- Justify your second edit.
- Discuss important features of **Cost** as shown by the final graph, including how the modeling implications of **Cost** will vary for a decision tree or generalized linear model (GLM) in which **Cost** is the regression target variable.

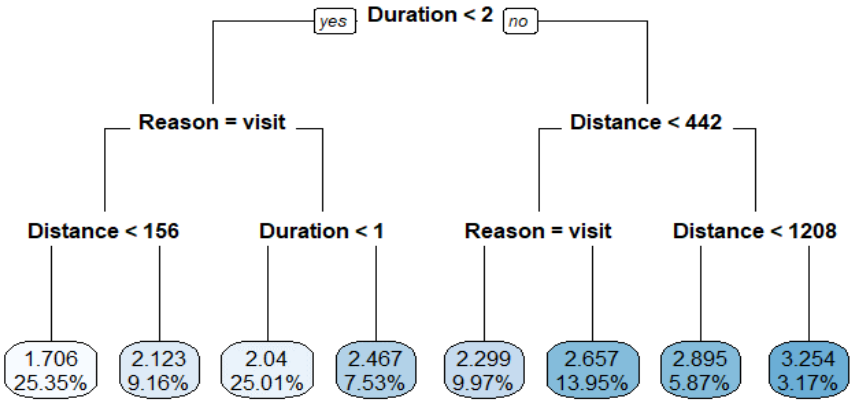
Additional information for Task 4 and beyond

Having benefited from your explanation, your assistant builds two regression trees that differ only by the form of the dependent variable, one using ***Cost*** *and the other using* $\log_{10}$***Cost****. The resulting trees (no code is given) are below. Your assistant is surprised by how different they are and is not sure which one to use. Just then, your manager stops by and wants to know more about what the interesting diagrams might mean for the trip cancellation insurance product.*

Tree 1



Tree 2



4. (*7 points*) Consider regression trees.

Perform the following, writing for a general audience, your manager:

- Describe the important differences between the two trees and explain why they occur.
- Recommend which tree is preferable for informing the trip cancellation insurance product. Justify your recommendation.

Additional information for Task 5 and beyond

Your manager is satisfied with your explanation and leaves. Your assistant, moving to another topic, is worried about highly correlated variables in the data and has built a correlation matrix for the numeric variables. Your assistant has also read about principal components analysis (PCA) and decided to apply it.

5. (*8 points*) Explore correlations and PCA.

After running and considering the code in the .Rmd file, assess the choices made by the assistant for both the correlation matrix and PCA, including implications for predictive modeling work and practical advice for your assistant. Do not alter the assistant's work.

Additional information for Task 6 and beyond

After trying out PCA, your assistant does not understand how PCA, an unsupervised learning technique for dimension reduction, can be used to generate features for a supervised predictive model. The assistant wonders whether running a generalized linear model (GLM) with LASSO regularization would be a preferable alternative for dimension reduction.

6. (8 *points*) Discuss dimension reduction techniques.

Perform the following:

- Explain how PCA can be used to develop features for a supervised predictive model.
- Compare PCA and LASSO regularization on both how they perform dimension reduction and their interpretability

Additional information for Task 7 and beyond

*You and your assistant decide not to use any dimension reduction technique and agree to predict **Cost** using a GLM. Your assistant focuses on the **Age** variable and is considering two transformations to replace age: 1) convert it to a factor variable with six levels, or 2) replace each value with the average age for each age band.*

7. (*9 points*) Consider two transformations of the Age variable.

Perform the following—no coding is required:

- Describe how the GLM results will differ between using each of the two suggested transformations.
- Explain the difference between the two transformations in terms of bias and variance.
- Describe, for each transformation, one distinct advantage it has over the other transformation besides that relating to bias and variance.

Additional information for Task 8 and beyond

Your manager returns, having thought more about the regression trees shown earlier, and wonders whether you should instead build a classification model that predicts whether the cost is at least $1000. The manager is unsure about which will better help ABC understand the market for its trip cancellation insurance.

8. (*8 points*) Consider two models and their respective targets.

Discuss advantages and disadvantages for each of the two models above (a GLM with Cost as the target variable, or the classification model mentioned by your manager) for informing marketing about potential buyers of the trip cancellation insurance product. Do not build either model. Aim for one-half page and do not write more than one page.

Additional information for Task 9 and beyond

Having heard your analysis, your manager wants to move things along and decides that your models going forward should predict whether the cost is at least $1000. Before leaving, your manager also lets you know that the legal department is uncomfortable with including age and asks you to remove it from your future models.

*Your assistant gets to work, naming the new variable **HighCost** and running a logistic GLM on training data. The assistant notes that **HighCost** is unbalanced, with slightly less than 10% of the observations being a positive instance but assures you that the unbalanced target is not causing any problems in this case, however, because the model has over 90% accuracy and the AUC is over 0.9, both excellent results.*

9. (*5 points*) Explain the problem with unbalanced classes.

After running the assistant's code, explain how fitting the unbalanced **HighCost** has led to a poor model for informing marketing about the trip cancellation insurance product despite the excellent accuracy and AUC.

Additional information for Task 10 and beyond

You then discuss undersampling and oversampling for addressing the problem.

10. (*5 points)* Discuss undersampling and oversampling.

Describe how undersampling and oversampling work to address the problem of unbalanced classes, including the impact these techniques have on the predictions.

Additional information for Task 11

After this discussion, you and your assistant decide to apply oversampling. Right when you are finishing up, your manager stops by again and is interested in the confusion matrix.

11. (*14 points)* Implement oversampling and explain the confusion matrix.

Perform the following, being sure to include the relevant confusion matrices in your Word document:

- Explain why oversampling must be applied after splitting train and test data.
- Implement oversampling and compare, for your assistant, the most applicable confusion matrix metrics(s) between the original and oversampled models.
- Explain for a general audience, your manager, both the confusion matrix and the most applicable metrics for this business problem using only the results from the oversampled model.
- Describe for a general audience, your manager, what additional information from the marketing department would help you and your assistant refine the model, perhaps with a different cutoff value than one-half.