Exam PA June 21 Project Statement

IMPORTANT NOTICE – THIS IS THE JUNE 21 PROJECT STATEMENT. IF TODAY IS NOT JUNE 21, 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. Alternating 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 21 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

Your small consulting firm, consisting of you and your less experienced assistant, is approached by a travel agency that has multiple offices across Canada and specializes in leisure vacations. To manage its increasing call volume, the travel agency plans to direct callers to its new and improved website, where visitors can fill out a form to request its overnight vacation planning services. The travel agency will then contact the prospective clients as time allows while servicing existing clients, and it plans to use the form to decide which clients to prioritize. The travel agency benefits more when their clients spend more on travel, so its priority is to call clients who are expected to spend more on vacation.

The travel agency specifically wants advice on what information to include on the form and how to use that information to automatically prioritize prospective clients. It is sure to include the origin and destination, from which the travel distance can be automatically calculated. Beyond that, it wants to include as few items as possible to encourage completion of the form and have more possible clients to contact.

The travel agency prides itself on operating with the highest ethical standards. Due to privacy concerns, it requests that you use a public dataset available on trips taken by Canadian travelers. It provides you with records from this data,[1] *incorporating the data cleaning another firm had done for this travel agency and business problem. However, the travel agency terminated that relationship due to what it perceived was poor quality work. The travel agency notes that some of the data is not applicable to its situation, and the other firm continued to give them inadequate responses on other matters. It passes along the data in a file called* June 21 Data.csv *and provides the following data dictionary.*

Data Dictionary

Variable Name	Definition	Values
Q	Calendar quarter of trip	Q1, Q2, Q3, Q4
ProvO	Trip province of origin	Province names
Distance	Distance traveled in trip, in km	6 - 4996
Duration	Number of nights spent on trip	1 - 81
Reason	Main reason for the trip	vacation: holiday, leisure, or recreation, visit: visit friends or relatives
Age	Age of adult survey respondent	Six age bins, as labeled
Gender	Gender of adult survey respondent	Female, Male
HHI	Household income, in Canadian $	Four income bins, as labeled
Others	Number of other persons that accompanied the respondent on the trip	0 - 21
Mode	Main mode of transportation	car, plane
Cost	Total spending on trip, in Canadian $	0 - 19200

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 applies for all tasks that follow the information and is labelled accordingly. Items that further explain a specific task are in plain text. Your reasoning and justifications should appear in your Word report.

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

Additional information for Task 1 and beyond

To avoid a misunderstanding with the travel agency, you decide to write a formal business problem based on the information given above and review it with the travel agency prior to starting your analytical work.

1. (*5 points*) Define the business problem.

Write a formal problem statement for a general audience, the travel agency, that translates its vague request into a business problem that can be analyzed with predictive analytics. Be sure to consider data and implementation challenges when defining the problem. The problem statement should be about one-quarter page long and should not exceed half a page.

Additional information for Task 2 and beyond

The travel agency agrees with your problem statement and explains that it fired the prior firm when it was not convinced that the data selection was appropriate. The prior firm disclosed the following information regarding data sources and sampling used to create the public dataset but could not satisfactorily explain its impact.

Data sources

Responding to this survey is voluntary.

Data are collected directly from survey respondents.

Selected households either receive an invitation letter in the mail or an electronic invitation (email). The letter explains who, from the household, is selected to participate in the survey using the age selection method as described below under Sampling. A household may receive up to two mailed or three emailed reminders. The average time required to complete the survey is 15 minutes.

Note that due to the COVID-19 pandemic, collection of data for the first quarter of 2020 took place for only the months of January and February 2020.

Sampling

For the first stage, determining the selected dwellings, the sample is first allocated among the provinces using the cube root of the number of dwellings in each province. For each province, the sample is then allocated according to income level in relation to the square root of the sum of the incomes. Within each stratum, by province and income level, dwellings will be sorted by postal code, and a systematic sample will be drawn. This will allow the different regions of each province to be represented in the sample.

For the second stage, one adult per selected dwelling will be randomly chosen using a selection method based on the age of household members. This method randomly chooses one adult for dwellings with up to six adults. Depending on the number of adults living in the dwelling, the adult will be selected from the oldest, the second oldest, the third oldest, the youngest, the second youngest, or the third youngest adult.

2. (*7 points*) Outline modeling impacts of data sources and sampling information.

Describe briefly for a general audience, the travel agency, three impacts the disclosed information could have on the quality of the predictive modeling.

Additional information for Task 3 and beyond

The travel agency had also learned that the prior firm had access not only to provincial information but also more specific regional information about the trip origin. It cannot understand why such valuable information would be thrown away. It shows you a comparison of the provincial and regional information (open June 21 Regional Data.xlsx *to see it) and asks you to defend the exclusion of the regional information.*

You and your assistant discuss the issue privately to make sure you have an appropriate response.

3. (*10 points*) Explain modeling impacts of high-dimensional and granular data.

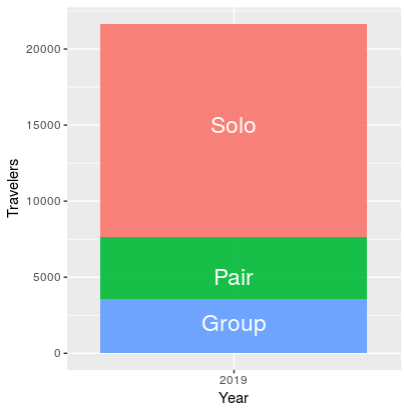
Perform the following:

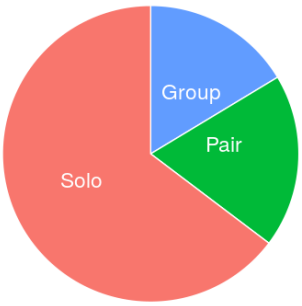
- Define the difference between dimensionality and granularity for a categorical variable.
- Identify whether the provincial or regional information is more granular.
- Describe how high-dimensional categorical variables can influence predictive modeling using a generalized linear model (GLM).
- Describe how high-dimensional categorical variables can influence predictive modeling when using a tree-based model.

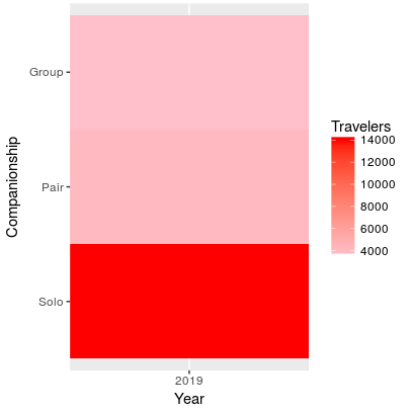
Additional information for Task 4 and beyond

The travel agency relents and accepts only using the provincial information.

Another incident with the prior consulting firm involved the following graphs, different versions of the same information.







Your contacts at the travel agency did not feel they could share and explain these graphs with their superiors and are hoping you can make a more presentable graph. You decide this is a good teaching moment for your assistant.

4. (*9 points*) Improve the graph.

- Describe to your assistant the necessary features for an effective graph of the relevant values.
- Critique each of the three graphs, including which features do and do not help to clearly depict the relative magnitude of the three values in each.
- Produce a more appropriate graph for the values that incorporates your advice. The data is provided directly in the .Rmd file. Justify your choice of visualization.

Additional information for Task 5 and beyond

Satisfied with your responses so far, the travel agency asks you to continue your work, with a plan to meet after the data is filtered to select the applicable observations and variables.

5. (*7 points*) Filter the data to fit the business problem.

Complete the following steps to prepare the data based on the business problem, including consideration of how the model will be used. Do not take additional data preparation or exploration steps. All but one point in this section is assigned to your explanations for what is removed.

- Remove from the data observations that should not be used to predict **Cost**. Explain and justify your decisions.
- Remove from the data variables that should not be used to predict **Cost**. Do not remove **Duration**, **Distance**, or **Others**; also, do not remove **Cost**, the target variable. Explain and justify your decisions.

Additional information for Task 6 and beyond

The travel agency is satisfied with your decisions. In the meeting where you reviewed the data, the travel agency describes how the prior firm had planned to use something called hierarchical clustering in the analysis. It wants you to experiment with hierarchical clustering and recommend whether this technique is worth pursuing.

*Your assistant does some prototyping work, as shown in the .Rmd file based on the data filtering you did earlier. The clustering is based on **Duration** and **Distance**, two variables the travel agency is sure to include in its form. The assistant is worried because the clustering does not look like it is providing useful results.*

6. (*11 points*) Assess your assistant's hierarchical clustering work.

Run your assistant's code on the data after the filtering from the prior task and inspect the output. Then do the following:

- Define what is meant by Height, as seen when the dendrogram is plotted.
- Explain, for your assistant, why the clustering seems to ignore **Duration** and only separate observations by **Distance**.
- Describe two approaches that would help the clustering balance **Duration** and **Distance** in its groupings.
- Recommend which of the two approaches to use. Justify your recommendation but do not implement it.

Additional information for Task 7 and beyond

After more work, you and your assistant conclude that hierarchical clustering is not going to be helpful to the travel agency. With a target of cost, the two of you also determine that running a generalized linear model (GLM) with a gamma distribution and without weights or offsets would make sense.

The travel agency then asks about something else the prior firm had trouble explaining to it: how to be sure that the model produced is predictive.

7. (*8 points*) Recommend a model selection method.

Without running any code, do the following using language appropriate for a general audience, the travel agency:

- Explain the differences among the following three model selection methods, assuming they are used separately:
    - Akaike Information Criterion (AIC)
    - 80%/20% train/test split
    - 5-fold cross-validation
- Recommend one of the above model selection methods (or a combination of methods) for the travel agency work. Justify your recommendation.

Additional information for Task 8 and beyond

The travel agency likes your explanation on validating the model and then wants to discuss further exactly what to predict. They pitch two alternatives: 1) predicting the cost in excess of $500, so that bigger spenders could more accurately prioritized, or 2) predicting the cost assuming that it is proportional to the number of nights, so that a relationship with bigger spenders could be established even if the current trip is short. You and your assistant discuss the modeling ramifications of these alternatives.

8. (*8 points*) Discuss modeling implications of the alternatives.

Run the code provided in the .Rmd file and consider the resulting data visualizations when discussing the modeling implications in the following, but do not run any models.

- Recommend a more appropriate choice of distribution than gamma for modeling the cost in excess of $500. Justify your recommendation.
- Explain the difference between weights and offsets when applied to a GLM, and then recommend which is more appropriate for implementing the assumption that the cost is directly proportional to the number of nights. Justify your recommendation.

Additional information for Task 9 and beyond

After further discussion, the travel agency determines that the target going forward should simply be whether a client is predicted to spend at least $500, as management now expects every such client be called the same day regardless of how much in excess of $500 they might spend.

The agency also mentions that it was intrigued by something called "elastic net" that the other firm had been talking about. You agree to include elastic net in your modeling work. Privately, you explain elastic net to your assistant, who then questions how adding something to the objective function you are trying to minimize would improve model performance.

9. (*4 points*) Explain the reasoning behind elastic net regression.

Explain, in terms of bias and variance, how the elastic net technique can improve predictive power. Including specific details about the hyperparameters will not earn any points.

Additional information for Task 10 and beyond

*Your assistant seems to understand and builds the code to run elastic net regression using just **Duration**, **Distance**, and **Others** as predictor variables on some train and test data (regardless of the distribution recommendation above). When comparing the results to a non-regularized GLM with the same predictors, however, the assistant finds virtually no impact from the elastic net regression and requires your help.*

10. (*10 points*) Improve the elastic net regression setup.

Perform the following steps to complete the elastic net regression:

- Explain why the elastic net regression had virtually no impact in the assistant's setup.
- Determine, based on the test metric described in the .Rmd file, the best value for alpha among {0, 0.25, 0.5, 0.75, 1} by repeatedly applying *cv.glmnet* using default settings.
- Fit the entire data set using the chosen hyperparameters and then interpret the predicted impact of **Distance** on the target variable **BigCost**.

Additional information for Task 11 and beyond

Just after you finish up your elastic net modeling, one of your contacts at the travel agency calls you and says they just read something about a random forest being a good predictive model and want you to work on that as well. They also request that the modeling now include all applicable data so they can evaluate what to include on the website form.

Your assistant starts work with the randomForest *method but is not sure exactly what the settings signify. The assistant soon returns to you with three models run on the training set with the following settings:*

randomForest setting	***rf_model_1***	***rf_model_2***	***rf_model_3***
Mtry	*<default>*	*ncol(data.rf.train) - 1*	*<default>*
Replace	*FALSE*	*<default>*	*<default>*

In addition, the assistant has remembered that AUC would be a good test metric for comparing the models but is not confident how to interpret the AUC results.

11. (*8 points*) Evaluate the random forest models.

Do the following to help the assistant:

- Describe the differences between the three model settings.
- Explain how the models determine their predictions from multiple component trees.
- Recommend which settings to use based on AUC. Justify your recommendation.

Additional information for Task 12

The travel agency contacts you and, having learned how soon decisions on the website need to be made, asks for a recommendation on what data to ask for as soon as possible, using whatever model is handy. You decide to use the chosen random forest model on the latest target ***BigCost*** *and use variable importance and additional model runs to make a recommendation.*

12. (*13 points*) Recommend data elements to collect.

Do the following, using language appropriate for a general audience, the travel agency. The response to this task should be approximately one page long.

- Display the variable importance data for the final model of the previous task, prior to selecting variables below.
- Explain what the variable importance results represent, including a sense of how they are determined.
- Improve model performance by selecting variables—run additional random forest models on the training data using the same settings and only varying which predictors are included in the model formula, using AUC as a test metric.

- Recommend which data elements the travel agency should collect, considering both model performance and other aspects of the business problem. Justify your recommendation.