Health Costs Practice Exam 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 ten specific tasks followed by an executive summary. The audience for the task responses is the examination grading team. Hence, technical language can be used. 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 components. The total is 100 points. Each task will be graded on the quality of your thought process and conclusions. The executive summary will be graded on the quality of the presentation. Note that a component of the grading of the first ten tasks will also relate to the quality of the exposition, but these sections need not be written as formal reports.

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, 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

Freedom Health is looking accelerate their underwriting process by building a predictive model. In the past, they have used a GLM, but their VP of Data Science believes that a GBM would offer higher predictive power. They have hired your consulting firm to conduct a comparative analysis using their historical claims data. Additionally, they want to understand what demographic factors indicate that a policy is likely to file large claims.

The data set consists of prior year's claims, along with patient demographic information. Your assistant has already begun this analysis and has

- Removed outliers
- Removed missing values
- Releveled factors to those with the most observations
- Run hierarchical clustering
- Fit a single GLM
- Fit a GBM, including tuning the hyper parameters

There are no assurances that your assistant has made the best choices in each code chunk, although they have left comments on the sections where new changes are definitely needed. On the exploratory sections, tasks 1-2, looking at additional summary statistics and graphs may be helpful.

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. (5 points) Write a high-level summary of the data. Include the number of records and number of variables.

2. (5 points) Explore the target variable charges and its relationship to smoker and bmi.

3. (9 points) Engineer additional features in order to improve the predictive power of the models.

- Create an age bucket feature which creates a separate level for
    - Age < 24
    - 24 <= Age <= 36
    - 36 < Age <= 50
    - Age > 50
- Age-household-size ratio
    - (1 + the number of children)/age
- log_household_size which is the log of 1 + number of children.
- Remove the original children variable

4. (8 points) Examine hierarchical clustering of age and bmi.

   Your assistant has set up code to run hierarchical clustering using the Euclidean distance metric on age and bmi. Only the first 50 records are used in order to make the graphs easier to read. Rescale both variables prior to clustering and then select a cutoff height. Explain why scaling is necessary, and what the cutoff threshold controls. You do not need to add the cluster column to the data or use it in the next tasks.

5. (5 points) Select an interaction based on the results from Task 2

Explain what an interaction effect is. Example code is provided in Task 5 which shows how to add an interaction between age and sex. There is no guarantee that this is the best choice of interaction. If appropriate, make adjustments so that your choice of interaction is included. Only interact two variables.

6. (15 points) Fit two GLMs.

   A glm consists of two key assumptions:

   a) A random component: the response family of claims.
   b) A link function: the way that the mean of the claims distribution, $\mu = E[Y]$, is related to the linear predictor, $z = \mathbf{X}\beta$.

   Choose the best two link functions and response family combinations (two different models) using all variables. Base your decision off of the MAE on the test set.

7. (10 points) Compare the two models.

   Compare the advantages and disadvantages of using either of these two link function and response family combinations. The following information on the domain of the mean of the response distribution and the range of the link function may be helpful. Based on this comparison, choose one of the two models.

Distribution	Range
Gaussian	$(-\infty, +\infty)$
Binomial	$\{0,1\}$
Gamma	$(0, +\infty)$
Inverse Gaussian	$(0, +\infty)$
Poisson	$\{0,1,2,3,4,5\}$

Name	Link Function	Mean	Range of Mean
Identity	$z = \mu$	$\mu = z$	$(-\infty, +\infty)$
Log	$z = \log(\mu)$	$\mu = \exp(z)$	$(0, +\infty)$
Inverse	$z = 1/\mu$	$\mu = \frac{1}{z}$	$(-\infty, +\infty)$
Inverse Squared	$z = 1/\mu^2$	$\mu = \frac{1}{\sqrt{z}}$	$(0, +\infty)$
Square root	$z = \sqrt{\mu}$	$\mu = z^2$	$(0, +\infty)$

8. (5 points) Examine the residual plots and determine if the model is a good fit.

9. (10 points) Fit a GBM using the given parameters.

   Your assistant has already run a grid search to find the optimal values of four hyper-parameters. For each of these parameters, describe what it does and whether a higher value would increase

or decrease the likelihood that the model overfits. Overfitting means that the model is too sensitive to random noise in the training data (or that the performance on the training data is good but the performance on the test data is poor).

Example: the n.trees parameter

"A gradient boosting machine is an ensemble method of decision trees. Each tree attempts to correct for the errors of the prior tree. The n.trees parameter controls how many boosting iterations to use. A higher value means that more trees are used, and hence overfitting becomes easier as n.trees increases."

10. (5 points) Compare the GLM to the GBM based on the MAE on the test set. Explain the advantages and disadvantages to using the GBM. Based on these findings, which model would you recommend for this problem?

11. (5 points) Interpret the variable importance and partial dependence plots of age, bmi, and smoker status.

12. (23 points) Executive Summary

    Your executive summary should reflect the information provided from tasks 1-10 as relevant to Freedom Health in non-technical language.

Data Dictionary

Variable	Definition
`age`	Age of the policyholder
`sex`	M or F
`bmi`	Body Mass Index: weight divided by height
`children`	Number of children
`smoker`	Smoker status. Yes or No
`region`	Geographic region
`charges`	Annual medical claims for this policy