Exam PA June 16, 2020 Project Statement

IMPORTANT NOTICE – THIS IS THE JUNE 16 PROJECT STATEMENT. IF TODAY IS NOT JUNE 16, 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 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 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 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. 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

Your actuarial consulting firm has been hired by the large healthcare system Merged and Acquired Clinics and Hospitals (MACH) to help the hospital executives gain a better understanding of the factors that drive inpatient length of stay. This insight will allow the administrators to better understand and manage patient needs.

MACH has provided data[1] on historical inpatient encounters for patients with diabetes. Each encounter includes the length of stay in hospital, measured in days. Your task is to use the available data to identify and interpret the factors that relate to inpatient stay duration.

To help get you started, your assistant has done some preliminary analyses, which are scattered throughout the supplied Rmd file. The data dictionary at the end of this document describes the available variables.

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*) Edit the data for missing and invalid data.

    Your assistant has provided generic code to address missing or invalid data. You can modify and use some or all of this code or add your own to edit the data. The code provided can perform the following:

    - Change numeric variable type to factor variable type
    - Remove variables / columns
    - Remove observations / rows
    - Combine / reassign variable levels to address missing / invalid values or reduce levels of variables

    Identify variables with missing or invalid values. For each, describe the issues you encountered and how you addressed these issues. Provide the rationale for your decisions and make the corresponding changes to the data frame.

2. (*15 points*) Explore the data.

    Investigate the distribution of the target variable and three variables that are most likely to predict the target (include among these at least one numeric and one categorical variable). Do not discuss the other variables in your response. For each of the four variables, perform the following:

    - Show key descriptive statistics.
    - Create visual representations (e.g., graphs, charts).
    - Explain, if not the target variable, why you selected this variable and how the variable relates to the target variable.
    - State your observations from the descriptive statistics and visual representations. Limit your comments to one page (statistics and visual representations will not count towards this page limit).

[1] The data are adapted from the "Diabetes 130-US hospitals for years 1999-2008" dataset contributed by Strack, DeShazo, Gennings, Olmo, Ventura, Cios, and Clore (2014) to the UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

3. (*4 points*) Consider two data issues.

- Identify a variable that may have potential ethical concerns. Discuss the considerations related to using this variable in a model for this business problem. Regardless of any concerns, continue to use this variable in subsequent analyses.
- Your assistant informed you that an additional variable that counts the number of laboratory procedures during the hospital stay is available. Explain whether such a variable should be included in your model and justify your position.

4. (*6 points*) Write a data summary for your actuarial manager.

Summarize your work performed in Tasks 1 through 3. Limit your response to one page, including any visualizations.

- Describe the process used to prepare and validate the data.
- Describe data issues encountered and how you addressed these issues.
- Provide a data summary. Include one or two key visualizations in your data summary.

5. (*8 points*) Perform Principal Components Analysis (PCA).

Perform PCA on the selected numeric variables and create a new feature based on the first principal component (`PC1`) using the code provided by your assistant.

- Describe PCA. Limit your response to a half page.
- State the advantages and disadvantages of using PCA for this problem.
- Interpret the output, including the loadings on `PC1`.
- Discuss whether other principal components aside from `PC1` should be used in the model. Regardless of your response, do not use the other principal components in subsequent tasks.

After completing Task 5, run the code to split the data into training and test data sets.

In the following tasks, use the Pearson goodness-of-fit statistic (formula shown below) to evaluate the models you create. The calculation is provided in the code.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{\hat{y}_i}$$

6. (*10 points*) Construct a decision tree.

- Describe what pruning does and why it might be considered it for this business problem.
- Construct an unpruned regression tree using the code provided.
- Review the complexity parameter table and plot 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.

- Calculate and compare the Pearson goodness-of-fit statistic on the test set for both trees (original and pruned).
- Interpret the entire pruned tree (all leaves) in the context of the business problem.

7. (*7 points*) Construct a generalized linear model (GLM).

   The code provided specifies a Poisson distribution with a log link function. Do not change these assumptions.

   - Consider the gamma and binomial distributions. Explain whether these distributions are reasonable alternatives to the Poisson distribution for this data and business problem.
   - Fit a first GLM with all original variables included prior to employing PCA. Then fit a second GLM with the PCA variable created in Task 5 included (and without the numeric variables used to produce the PCA variable). It is not necessary to display the output of these GLMs in the report.
   - Evaluate the Pearson goodness-of-fit statistic on the test set for both GLMs.
   - Select one of these GLMs to proceed with in the tasks that follow. Justify your recommendation based on the results and the business problem.

8. (*4 points*) Perform feature selection with lasso regression.

   Run a lasso regression using the code chunk provided. The code will need to be modified to reflect your decision in Task 7 regarding the PCA variable.

   - List the features used in the resulting model and calculate the Pearson goodness-of-fit statistic on the test set.
   - Compare the resulting model to the GLM selected in Task 7 and recommend which model is more appropriate for this business problem.

9. (*7 points*) Discuss the bias-variance tradeoff.

   - Define bias and variance and describe the bias-variance tradeoff.
   - Explain how lasso regression seeks to address the tradeoff.
   - Explain how splitting the data into training and test sets and calculating a metric such as the Pearson goodness-of-fit statistic on the test set seeks to address the tradeoff.

10. (4 *points*) Consider the final model.

    Your supervisor has decided the best model to share with the client is the GLM selected in Task 7. Do not assume that the best response in Task 8 is to select this model.

    - List one advantage and one disadvantage of employing a GLM for this business problem versus a decision tree.
    - List one advantage and one disadvantage of employing a GLM that does not remove features for this business problem versus feature selection using lasso regularization.

11. (*7 points*) Interpret the model for the client.

Run the recommended GLM from Task 7 on the full dataset.

- Copy the model output into your response.
- Interpret the coefficients for one categorical variable and one numeric variable to describe how these features relate to the target. The interpretations should be written in language appropriate for the client.

12. (*20 points*) Write an executive summary for the client.

Your executive summary should reflect the information provided and work from Tasks 1-11 as relevant to the hospital executives. Your executive summary should include a problem statement, discussion of the data, a coherent explanation and justification of the recommended model, and conclusions. You may include any particularly relevant visualizations to supplement your writing.

Data Dictionary

Name	Description	Values
`days`	Number of days between admission into and discharge from hospital	Integer 1 - 14
`gender`	Patient gender	Female, Male
`age`	Patient age (in ten-year age bands)	[0, 10), [10, 20), …, [90, 100)
`race`	Patient race	AfricanAmerican, Asian, Caucasian, Hispanic, Other
`weight`	Patient weight (in twenty-five pound weight bands)	[0, 25), [25, 50), …, [175, 200)
`admit_type_id`	Identifier corresponding to the type of hospital admission	1 = Emergency 2 = Urgent 3 = Elective 4 = Not Available
`metformin`	Indicates whether upon admission, *metformin* was prescribed or there was a change in the dosage	Up = dosage was increased Down = dosage was decreased Steady = dosage did not change No = drug was not prescribed
`insulin`	Indicates whether upon admission, *insulin* was prescribed or there was a change in the dosage	Up = dosage was increased Down = dosage was decreased Steady = dosage did not change No = drug was not prescribed
`readmitted`	Indicates whether patient had been readmitted after an inpatient stay in the twelve months preceding the encounter	<30 = patient was readmitted in less than 30 days >30 = patient was readmitted in more than 30 days No = no record of readmission
`num_procs`	Number of procedures performed in the twelve months preceding the encounter	Integer 0 - 6

num_meds	Number of distinct medications administered in the twelve months preceding the encounter	Integer 1 - 67
num_ip	Number of inpatient visits of the patient in the twelve months preceding the encounter	Integer 0 - 21
num_diags	Number of diagnoses entered to the system in the twelve months preceding the encounter	Integer 1 - 16