Exam PA December 8, 2020 Project Statement

IMPORTANT NOTICE – THIS IS THE DECEMBER 8 PROJECT STATEMENT. IF TODAY IS NOT DECEMBER 8, 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 eleven 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 eleven 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

Your actuarial consulting firm has been hired by a national retail firm to help them understand the factors that affect pedestrian activity. They have noted that their sales are proportional to pedestrian activity near their existing stores. When considering new locations across the U.S. for stores, insight into how weather affects pedestrian activity would be useful. For existing stores, it would be useful to understand how pedestrian activity varies by time of day and day of the week when deciding on which hours to be open.

The data for this study consists of hourly counts of pedestrians that were collected in 2017-2019. The exam data is derived from a public dataset provided by NYC Open Data (https://opendata.cityofnewyork.us/).

Some cleaning was performed in advance. The data dictionary at the end of this document describes the available variables. It is not necessary for you to paste the dictionary into your report.

Your goal is to identify and interpret factors that relate to higher or lower counts of `pedestrians`.

To get you started, your assistant has done some preliminary analyses, which are scattered throughout the supplied Rmd file. Your assistant also supplied code to perform various tasks as indicated in the Rmd file. There is no assurance that your assistant has made the best choices in each code chunk.

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. (*7 points*) Explore the variables.

    Your assistant has supplied code that allows you to consider each variable alone and in relation to `pedestrians`. Use the graphical displays and summary statistics to form preliminary conclusions regarding which variables are likely to have significant predictive power.

    - Hypothesize which variable is least likely to have a significant contribution to your model. Justify your opinion.
    - Hypothesize which variable is most likely to have a significant contribution to your model. Justify your opinion.

2. (*7 points*) Reduce factor levels.

    Your assistant has reduced the factor levels of `weekday` since the box plots of several weekdays were similar.

    - Run the assistant's code to consolidate `weekday`.

    Your assistant also recommends reducing the number of factor levels of `weather` and prepared some code to do so but requires your direction.

    - Evaluate the relevant issues of the `weather` variable with regard to the business problem.
    - Then, reduce the number of factor levels to five or fewer. Justify your choices.

3. (*11 points*) Modify the hour and temperature variables.

Your assistant has prepared code defining three different versions of the hour variable:

i. `hour_1`: A numeric integer variable with 17 values, ranging from 6 to 22 (representing 6 a.m.–10 p.m.)
ii. `hour_2`: A factor variable with 17 levels, also ranging from 6 to 22 (representing 6 a.m.–10 p.m.)
iii. `hour_3`: A numeric variable that measures time as the absolute number of hours from hour 14 (hour 14 is 2 p.m.). This results in values from 0 to 8.

- Recommend which hour variable to use for a decision tree. Justify your recommendation, including consideration of the business problem.
- Recommend which hour variable to use for a generalized linear model (GLM). Your recommendation may or may not be different from that for a decision tree. Justify your recommendation, including consideration of the business problem.
- Run the code to create the `hour_tree` and `hour_glm` variables as recommended and remove the other variables related to the hour of the day.

Your assistant has also prepared code defining two different versions of the temperature variable:

i. `temperature_1`: A numeric variable of degrees Fahrenheit at each hour
ii. `temperature_2`: A numeric variable of predicted daily average temperature

- Recommend which temperature variable to use for both a decision tree and a GLM. Your recommendation must be the same for both. Justify your recommendation, including consideration of the business problem.
- Run the code to create the `temperature_new` variable as recommended and remove the other variables related to temperature.

4. (*9 points*) Consider transformations of the target variable.

Your assistant notes that the target variable `pedestrians` is skewed and therefore recommends transforming it.

- Explain how transforming the target variable has an effect when fitting a decision tree model.

Your assistant provides code for several choices for transformations, supplying code for these:

i. `pedestrians` (no transformation)
ii. Log of `pedestrians`
iii. Square of `pedestrians`
iv. Square root of `pedestrians`
v. Inverse of `pedestrians`

- Recommend which transformation of the target variable from the list above to use for a decision tree model. Justify your recommendation.

Note to candidates: You will not use a transformed target variable in most modeling tasks in this exam. You will use the untransformed, original `pedestrians` as the target variable unless you are directed otherwise.

5. (*8 points*) Build two trees.

- Run the chunk of code that partitions the dataset into train/test/holdout subsets.

Your assistant has prepared two tree models. One uses the untransformed `pedestrians` as the target variable. The other uses the square root of `pedestrians` as the target.

- Run the provided code to fit each tree.
- For each tree, state the predicted pedestrian count for the following levels of the original variables.
  a) At `hour` = 6, `temperature` = 65, `temp_forecast` = 55, `weather` = cloudy, `precipitation` = 0, `weekday` = Thursday
  b) At `hour` = 15, `temperature` = 75, `temp_forecast` = 86, `weather` = rain, `precipitation` = 0.010000, `weekday` = Saturday

Your assistant notes that the RMSE of the model using the transformed target is the lower of the two and recommends that model.

- Critique your assistant's reasoning.

6. (*3 points*) Consider a random forest.

Your assistant recommends not using a random forest because it requires too much computing power.

- Describe two other considerations when deciding whether to use this method.

7. (*8 points*) Fit a generalized linear model (GLM).

- Explain why Poisson is a reasonable distribution choice for this problem.
- State two other reasonable choices of distribution. Explain why each choice is reasonable.
- Run the code for a GLM, using a Poisson distribution with log link function.
- Run the code for a second GLM model, using a different distribution than Poisson, still using a log link function.
- Recommend which GLM to use. Justify your recommendation, including consideration of the business problem.

8. (*6 points*) Consider an interaction.

Your assistant has prepared code to help explore for a suitable interaction.

- Recommend an interaction term, where at least one of the two variables is a factor variable, to include in the GLM. Justify your recommendation, including consideration of the business problem.
- Train a GLM using your recommended interaction term and your choice of distribution from Task 7.
- Interpret the resulting coefficient(s) for the interaction term.

9. (*9 points*) Select features.

Some features may lack predictive power and lead to overfitting. You will select features using BIC for your model from Task 8, including your selected interaction term. Forward and backward selection are among the available techniques for feature selection.

- Describe forward and backward selection and state the difference between the two techniques.
- Recommend which technique to use. Justify your recommendation.
- Determine which features should be retained, using only the recommended technique. Run the stepAIC function (from the MASS package) to make this determination.
- Interpret the results of the first step of the stepAIC function output.

10. (*8 points*) Recommend a model.

- Recommend a GLM for this business problem from among those fit in tasks 7, 8, and 9. Justify your recommendation both quantitatively, using RMSE as the test metric, and qualitatively.
- Recommend a model for this business problem between the GLM just recommended and the decision tree using the untransformed target variable from Task 5. Justify your recommendation both quantitatively, using RMSE as the test metric, and qualitatively.
- Run the recommended model on the full dataset for use in the executive summary, regardless of the results of Task 11.

11. (*4 points*) Validate the model selection.

- Assess the model selection performed in Task 10 quantitatively by using the holdout data.

12. (*20 points*) Executive summary.

Your executive summary should reflect the information provided and work from Tasks 1-11 as relevant to the client. Your executive summary should include a problem statement and a coherent explanation of all the steps leading to your recommended model and conclusions. Interpret the results of your final model in a manner that will provide useful information to client.

Data Dictionary

pedestrians	The count of pedestrians during one hour starting at the indicated time	Integer 0 - 4330
hour	Time at beginning of measuring hour	Integer 6 - 22
weather	Hourly weather condition	Factor with eleven categories
temperature	Hourly temperature in degrees Fahrenheit	Integer 2 - 97
precipitation	Hourly precipitation in inches	Numeric with six decimal places, 0 - 0.554300
weekday	Day of the week	Factor with seven categories
temp_forecast	Predicted daily average temperature	Numeric