# Exam PA December 7, 2020 Project Statement

## IMPORTANT NOTICE – THIS IS THE DECEMBER 7 PROJECT STATEMENT. IF TODAY IS NOT DECEMBER 7, 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**

The ABC Bike Sharing company provides locked bikes throughout town. Users pick up a bike using an access code and leave it at their destination. ABC has asked you to develop a model to predict bike

sharing usage. The company is interested in predicting the number of bike rentals in a given hour to help them with the distribution of bikes.[1]

**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*) Select factor variables.

   Your assistant has cleaned and provided data.

   - Select which variables should be treated as factor variables. Justify your selection.
   - Convert the selected variables into factor variables, renaming and releveling the factors as needed. Do not change the number of levels.
   - Run the summary function on your revised data and provide the output in your report.

2. (*3 points*) Consider a new variable.

   Your assistant has informed you that a new workday variable is available. The value of the workday variable is 0 for weekends or holidays and 1 for non-holiday weekdays.

   - Describe one advantage and one disadvantage of including the workday variable in your model.

3. (*12 points*) Write an overview of the data for your actuarial manager.

   Limit your response to one page, including any visualizations.

   - Describe the target and predictor variables.
   - Include one univariate variable analysis and one bivariate variable analysis.
   - Include key descriptive statistics and one or two key visualizations that best indicate what is going on in the data.

4. (*6 points*) Select an interaction to consider for your model.

   - Explain what an interaction is.
   - Explain why you selected this interaction.
   - Include descriptive statistics and visualizations to support your reasoning in the report.

[1] The data are adapted from the "Bike Sharing" dataset contributed by Hadi and Joao (2013) to the UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

5. (*9 points*) Perform a k-means cluster analysis.

   Your assistant has provided code to perform k-means cluster analysis on `temp` and `humidity`.

   - Describe k-means cluster analysis.
   - Interpret the elbow plot.
   - Select a value for k based on the elbow plot. Justify your selection.
   - Create a new variable from the two variables used to form clusters based on the value of k that you selected.

   Consider the graphs of clusters based on the temp and humidity variables.

   - Recommend whether the new variable based on clustering should replace temp and humidity. Justify your recommendation based on the graphs and the potential implications for the predictive model.

6. (*7 points*) Construct a decision tree.

   Your assistant has provided code to split the data into test and training sets.

   - Run the code as provided by the assistant.

   Your assistant has set up the code to generate a decision tree on the training data and perform cost-complexity pruning.

   - Construct a pruned regression tree with four terminal nodes using the code provided.
   - Describe, for each split in the tree, why the characteristics of each resulting node makes sense as it relates to the drivers of bike usage.

7. (*12 points*) Construct a boosted decision tree.

   - Explain what boosting does and why it might be considered for this business problem.

   Your assistant has provided code to run a boosted decision tree with 1000 trees, shrinkage of 0.01 or 0.1, and depth of 4.

   - Plot the partial dependence plot for each of the two most important factors using shrinkage of 0.01 and explain each plot.
   - Run the tree with the two given shrinkage parameters. Describe the differences in results and explain how the change to the shrinkage parameter causes the observed differences.
   - Explain how a boosted decision tree can predict negative bike rentals in this business problem when a decision tree cannot.

8. (*10 points*) Compare distribution choices for a generalized linear model (GLM).

   Your assistant has prepared code to fit two GLMs to the data, without the interaction and cluster variables, using the Poisson and gamma distributions, each with its canonical link function.

   - Explain whether each choice for distribution and link function is reasonable for this data and business problem.

   Do *each* of the following for *both* models, the one with the Poisson distribution *and* the one with the gamma distribution:

   - Fit the GLM using the given distribution and canonical link function on the training data without the interaction or the cluster variable.
   - Run the summary function for the model and include the summary in your report.
   - Explain what information the model provides regarding the effect of being in summer compared to winter.
   - Explain what information the model provides regarding the effect of temperatures of 10 degrees Celsius compared to 20 degrees Celsius.

9. (*6 points*) Evaluate the interaction term.

   - Explain why the interaction term was not needed when fitting a decision tree.

   To evaluate the interaction, your assistant set up code using the Poisson distribution with log link function.

   - Run the code to fit the two GLMs with this setup, one without an interaction term and one with the interaction term selected in Task 4.
   - Run the summary function for each model and include the summary in the report.
   - Recommend whether the interaction term should be included in your GLM model. Justify your recommendation.

10. (*4 points*) Evaluate the cluster variable in the GLM.

    To evaluate the cluster variable, your assistant set up code similar to Task 9 but where the cluster variable is included, the variables producing the cluster variable are excluded, and the interaction term is excluded.

    - Run the code to fit the GLM with the cluster variable.
    - Run the summary function for this model and include the summary in your report.
    - Recommend whether the cluster variable should be included in your GLM in place of the variables that produced it. Justify your recommendation.

11. (*6 points*) Select the final model to present to the client.

    The following models are to be considered: a decision tree (Task 6), boosted decision tree (Task 7), GLM with all original variables using Poisson (Task 8), GLM with all original variables using gamma (Task 8), GLM with an interaction term (Task 9), and GLM with clustered variables (Task 10).

    - Recommend a final model to present to the client. Justify your recommendation, including how it addresses the business problem.

    To help the client understand the value of your recommendation, do the following analysis and explain in terms the client will understand:

    - Fit an intercept-only GLM using the code provided and include the summary in the report.
    - Explain how your model represents an improvement over using a model without any predictors, that is, with just an intercept term.

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

    Your executive summary should reflect the information provided and work from previous Tasks as relevant to the business problem. 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.

    The executives are interested in the value added by the model. Explain what your model adds to an analysis of the business problem.

**Data Dictionary**

| Variable | Description | Variable Values |
|---|---|---|
| season | Season | Integer from 1 to 4<br>1 – Winter<br>2 – Spring<br>3 – Summer<br>4 – Fall |
| year | Year | Integer from 0 to 1<br>0 – 2011<br>1 – 2012 |
| hour | Hour | Integer from 0 to 23 |
| holiday | Whether the day is a holiday | Integer from 0 to 1<br>0 – not holiday<br>1 – holiday |
| weekday | Day of the week | Integer from 0 to 6<br>0 – Sunday<br>1 – Monday<br>2 – Tuesday<br>3 – Wednesday<br>4 – Thursday<br>5 – Friday<br>6 – Saturday |
| weathersit | Weather situation | Integer from 1 to 3<br>1 – Clear or partly cloudy<br>2 – Mist<br>3 – Rain or Snow |
| temp | Normalized temperature in Celsius.<br>The values are derived via (t – t_min) / (t_max – t_min), t_min = -9, t_max = +39 | Numeric, two decimal places |
| humidity | Normalized humidity. The values are divided by 100 (max). | Numeric, three decimal places |
| windspeed | Normalized windspeed. The values are divided by 67 (max). | Numeric, four decimal places |
| bikes_per_hour | Count of rental bikes in each hour | Integer |