Exam PA December 14, 2021 Project Statement

**IMPORTANT NOTICE – THIS IS THE DECEMBER 14, 2021 PROJECT STATEMENT. IF TODAY IS NOT DECEMBER 14, 2021, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.**

## General Information for Candidates

This examination has 11 tasks numbered 1 through 11 with a total of 100 points. The points for each task are indicated at the beginning of the task, and the points for subtasks are shown with each subtask.

Each task pertains to the business problem (and related data file) and data dictionary described below. Additional information on the business problem may be included in specific tasks—where additional information is provided, including variations in the target variable, it applies only to that task and not to other tasks. An .Rmd file accompanies this exam and provides useful R code for importing the data and, for some tasks, additional analysis and modeling. The .Rmd file begins with starter code that reads the data file into a dataframe. This dataframe should not be altered. Where additional R code appears for a task, it will start by making a copy of this initial dataframe. This ensures a common starting point for candidates for each task and allows them to be answered in any order.

The responses to each specific subtask should be written after the subtask and the answer label, which is typically ANSWER, in this Word document. Each subtask will be graded individually, so be sure any work that addresses a given subtask is done in the space provided for that subtask. Some subtasks have multiple labels for answers where multiple items are asked for—each answer label should have an answer after it. Where code, tables, or graphs from your own work in R is required, it should be copied and pasted into this Word document.

Each task will be graded on the quality of your thought process (as documented in your submission), conclusions, and quality of the presentation. The answer should be confined to the question as set. No response to any task needs to be written as a formal report. Unless a subtask specifies otherwise, the audience for the 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).

Prior to uploading your Word file, it should be saved and renamed with your five-digit candidate number in the file name. It is not required to upload your .Rmd file or other files used in determining your responses, as needed items from work in R will be copied over to the Word file as specified in the subtasks.

The Word file that contains your answers must be uploaded before the five-minute upload period time expires.

## Business Problem

*You work at XYZ, a large actuarial consulting firm. Your boss, B, is a Fellow of the Society of Actuaries with expertise in Predictive Analytics. Outside of work, B volunteers at an animal shelter that started operating in 2019. B recently convinced the decisionmakers at XYZ to take on the shelter as a pro bono (i.e., unpaid) client and put you in charge.*

*Animal shelters take in unwanted and lost dogs and cats. Some animals are reclaimed by owners, typically very soon. At “No Kill” shelters like the local one the unclaimed animals are housed until someone adopts them as a pet. Before the pandemic created a surge in demand for pets that emptied the local shelter, it housed an increasing population of animals because the demand for local adoptions was less than the flow of unclaimed animals into the shelter. To avoid returning to the same unsustainable situation, the shelter plans to start a transfer program whereby some animals are transferred to partner organizations in other locations where there is high demand for adopted pets. Transfers can help a shelter place many animals at once. They are a useful tool to manage shelter population levels (as opposed to a last resort for unadoptable animals). A transfer program can only transfer animals that the partner organization agrees to accept.*

*B has identified the following issues that the local shelter faces:*

* *Understanding the characteristics of animals included in transfer agreements would aid the local shelter in preparing to start such a program.*
* *An accurate estimate of the length of time between arrival at the shelter and placement (return to owner, adoption, or transfer) would aid the shelter in planning and budgeting. They want to estimate how long that animal will stay as each animal arrives.*

*B also created a dataset[[1]](#footnote-1) using public data from the Austin Animal Center (AAC) for you to use. AAC is a “No Kill” animal shelter in Austin, Texas. AAC has a robust transfer program and an excellent reputation. Your city is similar in size to Austin.*

*B has provided the following data dictionary and the dataset of 48,409 records derived from AAC data in a file called* Exam PA Animal Shelter Data.csv*.*

## Data Dictionary

|  |  |
| --- | --- |
| **Variable Name** | **Variable Values** |
| outcome | Adoption, Transfer, Return to Owner |
| stay | length of stay in days (0 to 1913) |
| animal | Cat, Dog |
| mf | Male, Female |
| age | age at intake in years ( -0.1 to 24) |
| in.month | 1 to 12 |
| in.year | 2013 to 2021 |
| out.month | 1 to 12 |
| out.year | 2015 to 2021 |
| in.reason | Owner Surrender, Public Assist, Stray |
| in.intact | 1 if able to have offspring, else 0 |
| out.intact | 1 if able to have offspring, else 0 |
| name | many values |
| breed | many values |
| color | many values |
|  |  |

**Comments**

The variables including “in.” in their names indicate conditions when an animal arrives at the shelter, and those including “out.” indicate conditions when an animal leaves the shelter, as indicated by outcome.

The data includes stays that ended in 2015 or afterwards.

Animals born at the center will have negative intake ages.

Many animals undergo a procedure during their stay that prevents offspring.

## Task 1 (*9 points*)

* 1. (*3 points*) Explain two reasons why predicting **stay** is not sufficient for addressing planning and budgeting for the animal shelter.

**ANSWER:**

* 1. (*1 point*) Identify the target variable for a second predictive model (in addition to the model predicting **stay**) that could be developed using the given data to provide a more complete picture for planning and budgeting.

**ANSWER:**

* 1. (*5 points*) Write a problem statement, written for a general audience, that incorporates both models for the purpose of planning and budgeting.

**ANSWER:**

## Task 2 (*7 points*)

Run the starter code and code for this task in the .Rmd file to see an overview of the data.

1. (*4 points*) Recommend two elements to add to the data dictionary to improve it. Justify your recommendations.

**ANSWER:**

1. (*3 points*)Describe two reasons why **name** cannot be used in a predictive model for the animal shelter.

**ANSWER:**

## Task 3 (*11 points*)

Your boss, B, has asked you to develop a simple visualization to get a sense of the impact of the COVID-19 pandemic on monthly cat adoptions in 2020. B is interested only in cats that were ultimately adopted and is only interested in total adoptions per month, not details such as breed, color, etc. Your assistant has completed some initial data work to summarize the monthly cat adoption data B asked about but is not sure how to create graphics using R.

Use the assistant’s data to complete the items below to assist B with the visualization. As B is interested in a quick, understandable visualization, complex formatting or color schemes are not required.

1. (*3 points*) Create a graph that shows the number of cat adoptions per month in 2020. Paste the code used to create the graph as well as the graph below.

**ANSWER:**

**Code:**

**Graph:**

1. (*2 points*)Explain to your assistant why the vertical axis on a bar graph should include zero while the axis for a scatterplot or line graph may or may not include zero.

**ANSWER:**



B identifies a large decrease in cat adoptions starting in March 2020 and asks the assistant to create a graph comparing cat adoptions each March from 2015 through 2020 to assess whether adoptions are always low in March or if this change is due to the pandemic. The assistant produces the graph above using some R code found online.

1. (*3 points*)Critique two aspects the assistant’s graph.

**ANSWER:**

1. (*3 points*)Recommend (but do not create) a graph that would be more effective in helping B make the desired assessment. Justify your recommendation.

**ANSWER:**

## Task 4 (*7 points*)

B is uncertain whether the entire range of dates should be used or only a subset after seeing the following graph:



1. (*3 points*)Describe, for a general audience, the effect on predictions for **stay** from including the earliest arrivals in the dataset.

**ANSWER:**

1. (*3 points*) Describe, for a general audience, one advantage and one disadvantage of including data from the pandemic era, March 2020 and forward.

**ANSWER:**

Your assistant wishes to use the “month” columns as predictors for length of stay in a GLM.

1. (*1 point*) Identify the issue with using both **in.month** and **out.month** to predict **stay**.

**ANSWER:**

## Task 5 (*12 points*)

B asks for explanations about the shelter operations and predictive modeling implications based on the graph below.



(*4 points for each observation*)Describe three observations from the graph that are important for both modeling and shelter operations. Discuss both the impact on shelter operations and the predictive modeling implications. Refer to the graph features that back up your observations.

**FIRST OBSERVATION ANSWER:**

**SECOND OBSERVATION ANSWER:**

**THIRD OBSERVATION ANSWER:**

## Task 6 (*6 points*)

B proposes fitting a classification model to distinguish 0-day stays and overnight stays and asks you to explore the data.

1. (*2 points*) Identify two significant differences between 0-day stays and overnight stays using the assistant’s code in the .Rmd file.

**FIRST ANSWER:**

**SECOND ANSWER:**

1. (*4 points*) Assess, for each of the differences you identified, whether it could be used to predict **stay** in this business problem.

**FIRST ANSWER:**

**SECOND ANSWER:**

## Task 7 (*8 points*)

You ask your assistant to try unsupervised learning techniques for exploring and better understanding the data. You assistant does k-means clustering on **age** and **stay** and plots results based on choices of *k* from 1 to 6, as shown in the .Rmd file. Run your assistant’s code on the data and inspect the output. Then do the following:

* 1. (*3 points*) Recommend what *k* should be. Justify your recommendation.

**ANSWER:**

* 1. (*2 points*) Assess your assistant’s claim that these specific clusters can help with the prediction of **outcome**.

**ANSWER:**

* 1. (*3 points*) Describe two practical differences between hierarchical clustering and k-means clustering. Do not implement hierarchical clustering.

**ANSWER:**

## Task 8 (*13 points*)

To reduce costly long-term stays, the animal shelter plans to display with each cat and dog available for adoption the number of days the pet has been available for adoption and the typical time to adoption for that sort of pet.

Believing that just using average or median times for the typical time across all pets would be inadequate for this purpose, B applies a generalized linear model (GLM) using a Gaussian distribution with identity link function on the public dataset to predict **stay**. Five such GLM models, differing only by their predictors, are set up.

Rather than rely on a single fitting for each model, each model is fit 200 times to directly observe how well that model will predict unseen data. Each fitting is called a trial for that model. For each of the 200 trials on a given model (set of predictors), the training data is a 20% random sample of adopted pets that came into AAC before 2020 (using **in.year**)and the test data was consistently all adopted pets from 2020 onwards.

The model performance is first measured individually for each record in the test data, calculating both the variance of the predicted results for **stay** and the square of the bias (the difference between the average predicted result and actual result for **stay**). Then, for each model, these record-level results are averaged across all test data records.

The process of fitting 200 trials, measuring performance for each record, and calculating the average performance over all records is repeated for each of the five models, each using the same random samples for the 200 trials, with the following results:

|  |  |  |
| --- | --- | --- |
| **Model formula** | **Mean Variance** | **Mean Squared Bias** |
| stay ~ animal\*age | 7.4 | 2799.2 |
| stay ~ animal\*age\_1 | 3.9 | 2791.3 |
| stay ~ animal\*age\_3 | 10.1 | 2735.4 |
| stay ~ animal\*age\_5 | 17.1 | 2760.0 |
| stay ~ animal\*age\_10 | 47.2 | 2787.4 |

The **age** predictor is the original numeric variable while the other four predictors involving age, e.g. **age\_1**, are categorical variables using age rounded down to the nearest integer, with the lowest age group always being “Under Age 1” and the highest age group being the age indicated and above. The levels for each categorical variable can be seen in the .Rmd code. Your assistant comes to you wanting to better understand B’s methodology.

1. (*2 points*) Explain what the variance and bias values indicate about the relative quality of predictions when comparing predictive models.

**ANSWER:**

1. (*2 points*)Calculate, for the first model listed, the typical errors up or down from the true value due separately to variance and bias for predictions of **stay**.

**ANSWER:**

1. (*2 points*) Explain how the mechanics of fitting the underlying linear model causes the variance to be higher for the model using **age** than it is for the model using **age\_1** even though they have the same number of coefficients.

**ANSWER:**

1. (*2 points*) Explain why bias (as calculated here) may not always decrease with additional degrees of freedom, as seen with the model that uses **age\_5** compared to that which uses **age\_3**.

**ANSWER:**

1. (*2 points*) Recommend which GLM should be used based on the above results. Justify your recommendation.

**ANSWER:**

1. (*3 points*) Calculate, after fitting the recommended GLM to all available data, a complete list of predicted stays based on animal and age for a general audience to use. Include the code used to fit the model (but not its output) in the space below.

**ANSWER:**

**Code used to fit model:**

**List of predicted stays:**

## Task 9 (*9 points*)

Your assistant builds a decision tree to predict whether the **outcome** is “Transfer” or not, resulting in the following tree.



B suggests that you include a new feature called kitten (**animal** = cat, **age** < 0.5) for your decision tree.

1. (3 points) Without creating a new decision tree, assess whether this variable will materially improve the decision tree.

**ANSWER:**

Due to high costs associated with transfers, the animal shelter would rather underpredict than overpredict the “Transfer” outcome while still successfully identifying “Transfer” outcomes. Your assistant produces the following confusion matrix from R, where TRUE corresponds to “Transfer”, using a cutoff of 0.25:

 Reference

Prediction TRUE FALSE

 TRUE 680 1392

 FALSE 1574 6034

1. (3 points) Recommend an evaluation metric based on the confusion matrix which supports the animal shelter’s goal. Calculate the metric using the confusion matrix above. Justify your recommendation.

**ANSWER:**

1. (3 points) Explain, for a general audience, why increasing the cutoff from 0.25 to 0.26 does not impact the confusion matrix from your assistant’s decision tree. Identify the smallest cutoff greater than 0.25 (rounded to 2 decimal places) that will result in a different value of the evaluation metric.

**ANSWER:**

## Task 10 (*9 points*)

B would like you to use a random forest to predict **stay**. B notes that it is important to the animal shelter that the predictions be equally likely to be too high or too low.

Your assistant fits two random forests to predict the length of stay. The first, rf1, predicts **stay** without a log transformation; the second, rf2, predicts log(**stay** + 1). The histograms below present the distribution of the actual **stay** variable followed by the distributions of the predicted values of **stay** from the two models.



1. (*3 points*) Explain why the histograms from appropriately fitted decision tree models would likely not be as smooth as those produced by a random forest, as above.

**ANSWER:**

1. (*3 points*)Explain why rf2 produces more predictions of shorter stays than rf1.

**ANSWER:**

Your assistant states that the random forest without the log transformation is preferable since the algorithm preserves the underlying distribution; i.e., predicted values more closely follow the distribution of the original dependent variable.

1. (*3 points*) Assess your assistant’s reasoning, including consideration of how the animal shelter will use the model.

**ANSWER:**

## Task 11 (*9 points*)

Your assistant creates a GLM called **glm\_start** on training data to predict whether an animal is adopted and then runs the drop1 function on it. Refer to the assistant’s code in the .Rmd file.

* 1. (*3 points*) Create a new model called **glm\_drop** based on the results of the drop1 function. Justify your predictor variables based solely on the drop1 results. Include the code that creates **glm\_drop** in the space below.

**ANSWER:**

**Code to create glm\_drop**:

**Justify your predictor variables:**

Your assistant also creates a model using LASSO, creating a model called **glm\_lasso**.

* 1. (*3 points*) Contrast the two methods, drop1 and LASSO, for selecting predictor variables.

**ANSWER:**

* 1. (*3 points*) Recommend whether to use **glm\_drop** or **glm\_lasso** based on AUC results and which predictors the models use. Justify your recommendation. Display the AUC results used in the space below.

**ANSWER:**

**AUC Results:**

**Recommendation and Justification:**

1. Adapted from Austin Animal Center Intakes (2021) and Austin Animal Center Outcomes (2021)  City of Austin, Texas Open Data Portal, <https://doi.org/10.26000/025.000002> and <https://doi.org/10.26000/025.000001>. [↑](#footnote-ref-1)