Exam PA December 13, 2021 Project Statement

**IMPORTANT NOTICE – THIS IS THE DECEMBER 13, 2021 PROJECT STATEMENT. IF TODAY IS NOT DECEMBER 13, 2021, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.**

## General Information for Candidates

This examination has 10 tasks numbered 1 through 10 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 (*6 points*)

You assess the applicability of the data for addressing the business problem.

1. (*3 points*) Describe, for a general audience, two advantages of using data from the Austin animal shelter in addressing this business problem.

**ANSWER:**

1. (*3 points*) Describe, for a general audience, two disadvantages of using data from the Austin animal shelter in addressing this business problem.

**ANSWER:**

## Task 2 (*9 points*)

The shelter has asked for a model that predicts length of stay. Your assistant is eager to start and wants to put all the variables into a predictive model. After explaining the value of exploratory data analysis, you have your assistant make a series of graphs of **stay** against all the other variables (except **name**, **breed**, and **color**). Run the starter code and the code for task 2 in the .Rmd file to see the graphs.

(*3 points for each decision*) Explain to your assistant three important modeling decisions you can make based on one or more of the graphs. Copy and paste the graph(s) used in each response below. Do not apply these modeling decisions for future tasks.

**FIRST ANSWER:**

**Graph(s):**

**Explanation:**

**SECOND ANSWER:**

**Graph(s):**

**Explanation:**

**THIRD ANSWER:**

**Graph(s):**

**Explanation:**

## Task 3 (*14 points*)

Your assistant notices several instances in **name** where weight is provided instead of a name and creates a new variable called **name.weight** for such instances. You explore the new variable.

1. (*4 points per observation*) Provide two observations based on bivariate analysis of **name.weight**. Include a graph supporting each of your observations in your answers below.

**FIRST ANSWER:**

**Graph:**

**Observation:**

**SECOND ANSWER:**

**Graph:**

**Observation:**

The animal shelter tracks animal weight but the records may be incomplete and will need to be manually inputted from paper copies.

1. (*3 points*) Describe two advantages of collecting animal weight within the context of the business problem.

**ANSWER:**

1. (*3 points*) Describe two disadvantages of collecting animal weight within the context of the business problem.

**ANSWER:**

## Task 4 (*11 points*)

Your assistant notes that it may be a good idea to develop a feature for a generalized linear model (GLM) based on whether the pet is a purebred or of a mixed breed.

1. (*2 points*) Explain the challenges the **breed** variable presents when considering it as a predictor variable in a GLM.

**ANSWER:**

Your assistant creates a new variable, **mix**, based on **breed**, that indicates whether an animal is a mixed breed or purebred (not mixed breed). Run the code provided by your assistant and perform bivariate analysis of the new **mix** variable with other variables.

1. (*5 points*)Describe an important finding that causes **mix** not to be usable as a predictive variable when using of all the data. Include a table or graph that illustrates the finding.

**ANSWER:**

**Table or graph:**

**Important finding:**

1. (*4 points*)Describe two additional approaches for using the information in the original **breed** variable for a GLM.

**ANSWER:**

## Task 5 (*9 points*)

You ask your assistant to conduct principal component analysis (PCA) on some of the numerical variables in the data with the intention of creating features for the target of **stay**. Your assistant returns with the PCA coding as shown in the .Rmd file and suggests that you only keep the first principal component of the PCA to minimize dimensionality. Run your assistant’s code on the data and inspect the output. Then do the following:

* 1. (*2 points*) Assess the tradeoff between the assistant’s suggestion, to keep only the first principal component, and including additional principal components.

**ANSWER:**

* 1. (*4 points*) Critique your assistant’s PCA and make improvements to address the issues discussed. Include your revised PCA code (but not the output) in the space below.

**ANSWER:**

**Critique:**

**Revised code:**

* 1. (*3 points*) Interpret the loadings of the first two principal components of your improvedPCAand discuss whether they would be sensible features for predicting **stay**. Include the loadings in the space below.

**ANSWER:**

**Loadings:**

**Interpretation:**

## Task 6 (*10 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 Poisson distribution with log link function on the public dataset to predict **stay** + 1 (allowing the model performance measures described below to work on stays of zero days). 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 and the square of the bias (the difference between the average predicted and actual result). The predicted and actual results used are not **stay** + 1 but log(**stay** + 1), or the level of the linear predictor, so that the variance and average are less affected by the skewness of **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 ~ age | 0.0010 | 1.5805 |
| stay ~ animal | 0.0011 | 1.4856 |
| stay ~ age + animal | 0.0016 | 1.4786 |
| stay ~ age + animal + mf | 0.0021 | 1.4792 |
| stay ~ age + animal + black | 0.0019 | 1.4767 |

The **black** predictor is derived from the **color** variable, being 1 if “Black” and 0 otherwise. 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** + 1.

**ANSWER:**

* 1. (*3 points*) State two reasons why bias, as calculated here, may not always decrease with additional degrees of freedom, as seen with the model that adds **mf**.

**ANSWER:**

* 1. (*3 points*) State which predictors should be selected based on the above data, putting them in order from more to less predictive. Explain your selection and ranking.

**ANSWER:**

**Predictors to select from more to less predictive:**

**Explanation:**

## Task 7 (*8 points*)

The animal shelter has asked for a simple model to identify animals that can be transferred. Your assistant builds decision trees for this purpose and performs cost complexity pruning on the tree.

1. (*3 points*) Explain, for a general audience, what cost complexity pruning does.

**ANSWER:**

Your assistant produces the table below with the error rate, sensitivity, and specificity for the training and test data on the unpruned and pruned trees:

 ErrorRate Sensitivity Specificity

unpruned.train 0.219 0.107 0.984

pruned.train 0.225 0.052 0.992

unpruned.test 0.232 0.073 0.974

pruned.test 0.229 0.039 0.988

1. (*3 points*) Recommend whether the unpruned or pruned tree would be preferable to use. Justify your recommendation in the context of the business problem.

**ANSWER:**

1. (*2 points*)Explain why neither the unpruned nor the pruned tree will be a satisfactory model for the local animal shelter to use.

**ANSWER:**

## Task 8 (*13 points*)

Your assistant seeks to fit a boosted tree to identify animals that can be transferred.

Your assistant wants to set eta as high as possible to reduce runtime.

* 1. (*3 points*) Explain why you disagree with your assistant about setting eta as high as possible.

**ANSWER:**

Your assistant then needs help setting eta.

* 1. (*3 points*) Explain cross validation and how it can be used to set the eta hyperparameter.

**ANSWER:**

The assistant uses the eta found via cross validation and calculates an AUC of 0.628 for the boosted tree, claiming that this demonstrates excellent model performance.

* 1. (*3 points*)Explain, for a general audience, what AUC represents.

**ANSWER:**

* 1. (*2 points*) Critique your assistant’s assessment of the model performance.

**ANSWER:**



You present the above ROC curve to B, who then circles two points on the curve to investigate further.

1. (*2 points*) Explain how the same model can be used to produce both points on the ROC curve.

**ANSWER:**

## Task 9 (*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:**

## Task 10 (*11 points*)

An accurate estimate of the length of stay would aid the shelter in planning and budgeting. The shelter is uncertain whether the effects of the pandemic will continue into 2022. Therefore, they would like to predict **stay** under two different assumptions:

1. Activity returns to pre-pandemic levels (prior to March 2020)
2. Activity remains at the levels seen since March 2020

Your assistant creates a feature called **pandemic** to distinguish data as described above and wants to build a GLM including this feature but is unsure what distribution to use.

Your assistant says that because we only care about the predicted mean, the choice of distribution does not matter as long as its domain includes all values of the target variable.

1. (*3 points*) Explain why your assistant’s statement is incorrect.

**ANSWER:**

Your assistant runs two models with different distributions, one using Poisson and one using gamma, using the log link function for each, and presents you the deviance residual plots below:

Poisson:



Gamma:



1. (*3 points*) Describe what the deviance residual plot for the gamma distribution shows about the model fit.

**ANSWER:**

1. (*2 points*) State which distribution is preferable. Explain your choice.

**ANSWER:**

1. (*3 point*s) Fit a GLM on all data using your preferred distribution and its model formula as seen in the deviance residual plots above. Provide the code that fits the GLM (but not the output) in the space below. State the coefficient for **pandemic** and calculate the effect of being in a pandemic on length of stay based on the GLM.

**ANSWER:**

**Code to fit GLM:**

**Coefficient for pandemic:**

**Effect of being in a pandemic on length of stay:**

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)