ExamPA.net

Titanic Practice Exam Project Statement

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 ten 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 eleven 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 ten 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

You are an actuary in charge of expanding your company's market for life insurance. Your Chief Actuary wants to launch a niche product targeting people who go on luxury cruises. To better understand the type of customers that are likely to purchase, you have been tasked with conducting market research based on historical records from people who traveled on the Titanic. This product will be called Life Jacket™ and will provide short term life insurance during cruises.

Your Chief Actuary wants to understand what factors contribute to a person surviving in the event of an oceanic catastrophe such as occurred with the Titanic.

Your assistant has already gone through and conducted preliminary work to

- Remove missing values (or impute them where needed)
- Relevel factor variables so that the base level has the most observations
- Split the data into training and test
- Binarize factor variables
- Set up code for building models

Specific Tasks

1. (5 points) Explore the relationships between each of the variables and the target variable *survived*. Apply a transform to fare if necessary.

2. (7 points) Create a new variable called "title" which captures passenger's title.

   There are many titles which only a few people have. Simplify these titles and explain your reasoning as to why these titles may impact the survival probability. For example, all of the people with the title "Mlle" should have the clean title "Miss".

Title	Suggested Clean Title
Mlle	Miss
Ms/MS	Miss
Mme	Mrs
Lady	Miss
Dona	Miss
Capt	Officer
Col	Officer
Major	Officer
Dr	Officer
Rev	Officer
Dona	Officer
Sir	Officer
the Countess	Officer
Jonkheer	Officer

3. (2 points) Create a new variable called family_size which counts the number of family members that a passenger has on board (including themselves)

4. (10 points) Use Kmeans to detect outliers

Some of the data from the Titanic was of questionable quality. Some people had missing data filled in after the ship had sunk by contacting relatives, looking at social security records, tax info, and other public records. We don't trust this data and so we want to exclude it. We know that there are about 5-20 people who had missing values for age, fare, pclass, and family size, but we are unsure of the exact number.

One use of clustering algorithms is for detecting outliers. If a person has a very high or low value for certain variables, then they are dissimilar to the other people. The distance between points is used as a way of identifying these people.

Put each person into a cluster using kmeans. If a cluster has fewer than 5 people put into it, then say that these are outliers. There should be anywhere from 5 – 20 patients that are classified as outliers. Choose values for the number of centers and the number of starting values and justify these choices based on the data set and the goal of detecting these outliers.

- Any cluster with fewer than 10 people in it should be counted as an "outlier" cluster
- Once you assign clusters, create a new column called outlier_flag which identifies which clusters have fewer than 10 observations. Then remove the cluster feature (aka the cluster number column).

5. (3 points) Select which variables should be used in modeling.

   Your assistant was not sure which variables to keep, and so she has included everything except `passengerid` because she realized that this has no relevance to whether or not a person survived.

   Decide on which variables will be useful for predicting survival and delete the others.

6. (7 points) Fit a random forest for variable selection.

   Your assistant has set up a random forest to determine which variables are important but was unsure of what the `ntree` and `mtry` parameters do and what would be reasonable values to use. Explain what these are, choose values based on the Accuracy and AUC, and then record the top 7 most important variables. Do not spend a lot of time on this task. The purpose here is just to select features.

7. (7 points) Based on your selected random forest, choose the top 7 variables by importance.

   Note: The random forest package uses the mean increase in node impurity as importance. The other common importance measure is the increase in accuracy.

8. (4 points) Fit a logistic regression and interpret the signs of the coefficients.

What do the signs of the coefficients indicate? It is acceptable at this stage if some of the p-values are greater than 0.01.

9. (9 points) Using the model's coefficients, calculate the probabilities of a person of age 10, 30, and 60 surviving.

For the other variables that are in the model, use the average for continuous variables and the mode for factors. You may use the `predict()` function to check your work, but the calculations need to be shown explicitly for full credit. Using Excel may save time.

To convert from the log odds to the probability, use the fact that for p the probability of survival, z the linear predictor, $z = \log(p/(1-p))$.

10. (4 points) Validate the logistic regression and compare the AUC and Accuracy to the random forest in task 5

    For accuracy, use the cutoff of 0.5. Explain, prior to looking at the result, what you would expect given that the logit is using the top 7 most important predictors. Then, compare the AUC and Accuracy. What are some of the differences in model structure which may account for this discrepancy? Explain how these differences relate to the large p-values for some of the variables.

11. (6 points) Fit an elastic net with all pairwise interactions

    Run the code to create a model matrix which has all pairwise interactions. With 7 original variables there are 21 variables after adding the interaction terms. Set the value of alpha so that you can perform variable selection.

    Bonus question: For a model with p variables, how many variables will there be after adding all pairwise interactions?

12. (6 points) Choose a value of lambda that results in about 5-10 variables remaining. Interpret the graphs. Hint: Look at ?plot.glmnet as well as the model's coefficients.

13. (6 points) Compare the AUC with the Logit's AUC

14. (24 points) Executive Summary

    The executive summary should use non-technical language to summarise the steps that you performed. It should include a problem statement, coherent explanation of all of steps 1-13, and insights into how these findings can Life Jacket™ a successful product.

Data Dictionary

Variable	Definition
passengerid	passengerid
survived	Survived Y/N
pclass	Ticket class
name	Name
sex	male, female
age	Age
sibsp	# of siblings
parch	# of parents or children aboard the Titanic
ticket	Number fare
fare	Cost of ticket
cabin	Cabin number
embarked	Port of Embarkation: C=Cherbourg, Q=Queenstown, S=Southampton