

Sample Test

Predictive Modeling – Methods and Techniques

For March 2019 Exam Sitting

CSPA Exam 3 - Predictive Modeling - Methods and Techniques Practice Exam

This Practice Exam will not be graded. Some of the instructions below will refer to grading. These instructions were intentionally left in the Practice Exam Instructions to simulate "exam-day" instructions.

Instructions Relating to the Virtual Exam Environment

- 1) Once this exam begins it will be available for up to FOUR hours. If you take a break, the exam timer will not stop.
- 2) With the exception of questions entitled "R question...", please answer the questions in the Excel workbook. (Details below.)
- 3) For the R questions, you will save all of your work in an R script (.R) file that is provided for that question. (Details below.)
- 4) Because you are able to choose which R questions you want graded, it is very important to indicate this by modifying cell B1 on the relevant sheets of the Excel workbook.
- 5) Do not save exam files under different names from those they already have. Only the original files will be graded.
Control and Alt keyboard shortcuts may not work in the virtual environment. They have not been intentionally turned off, but these features may work differently in the virtual environment than they do normally, and functionality may vary for different types of computers. Some people have found that Ctrl-C (copy) and Ctrl-V (paste) work while Ctrl-Page Down (switch tabs) does not. In fact, for them using Ctrl-Page Down creates an unusual situation where additional tabs are grouped with the current tab until the Ctrl key is pressed again. Candidates may want to avoid using this shortcut.
- 6)

Instructions Relating to the Multiple Choice and Free Answer Questions

- 7) Each question is asked on a single sheet, with the sheet name matching the question number (e.g. Question 1 is on sheet "1"). The question number is also shown in cell A1 on each question sheet.
- 8) On each question sheet, the exam question is provided in a protected grey area; while you may modify the formatting within this area, you may not change the content of the area, insert any rows/columns, or delete any rows/columns. **If the content or cell range of the grey area is changed in any way, your answer to that question will not be graded.**
- 9) In the event that you accidentally delete a question or question sheet, there is a read-only copy of this workbook available on the desktop. You can copy question sheets back in from that workbook if you accidentally delete from here. Please then copy and paste any work you may have done for that question to the sheet you have copied in.
- 10) For each question, the number of points for the full question is indicated in cell A3. The number of points for each subpart may be indicated in some cases.
- 11) In cell B1 of each question sheet you have the option to identify the status of your answer as "Incomplete", "Finished", or "Review". Any selections made will also appear in the Point Grid sheet. With the exception of the R questions, these selections are optional and solely for your benefit, and they will not be provided to the graders.
- 12) Candidates can change the size of the Excel content by changing the zoom slider in the lower right corner of Excel. Multiple sheets can be adjusted at the same time by selecting them before zooming.
- 13) DO NOT use "Clear Formats" or "Clear All" to remove cell contents. Doing so will lock the cell. Instead, use "Clear Contents" or just delete the contents of the desired cell.
- 14) Enter answers in the white space below or to the right of the grey question box. Any cell content beyond Row 200 or Column AZ will NOT be graded.

- The answer should be concise and confined to the question as posed. When a specified number of items are requested, do not offer more items than requested. For example, if you are requested to provide three items, only the first three responses will be graded. Also, for multiple choice questions, only the choice will be graded and not any work that may have been required to get there. In other words, there is no partial credit for multiple choice questions. Please ensure that you clearly indicate a choice, which should always be A, B, C, D, or E, and be sure that that indication is NOT in the grey area of the sheet.
- In order to receive full credit or to maximize partial credit on mathematical and computational questions, you must clearly outline your approach in either verbal or mathematical form, showing calculations where necessary. It is not necessary to state the formula verbally if the calculation is made directly in the cell. While Excel tools may be available to assist in calculations, candidates should ensure there is sufficient documentation of their work.
- Use of Excel functions (for example SUM, AVERAGE, SUMPRODUCT, etc.) is allowed and encouraged for efficiency but not required.
- You must clearly specify any additional assumptions you have made to answer the question.
- Only work shown on the question sheets will be graded; a copy of the sheets will be provided to the graders in Excel such that the graders can consider both the formula entered in a cell and the result of that formula. An optional "Scratch" sheet is available for candidates to use for side work. Any contents included on the Point Grid or the Scratch sheets will not be provided to the graders.
- DO NOT use named ranges as they may not copy over correctly to the graders.
- DO NOT use Visual Basic code. It will not be provided to the graders.
- DO NOT use cell comments. Content in cell comments will not be graded.
- DO NOT include links to other sheets; linked values in candidate answers will not carry over correctly to the grading files.
- Cell contents do not need to be printer-friendly. Text within a cell can extend beyond what can be seen on the screen.

Instructions Relating to the R Questions (10-12)

- The text of the questions is in the Excel workbook.
- The R questions are "Do any 2 of 3". Please indicate in cell B1 of the workbook which questions you intend to have graded by marking them "Finished".**
- When R questions are graded, they will be sorted primarily in order "Finished", "Review", and "Incomplete", and within each category by question number. The first four (4) questions when sorted in this order will be graded**
- To start the R questions in general, sign into RStudio in the remote version of Chrome using the ID and password provided in the Notepad document. If the browser is not already set to RStudio, please click on the RStudio button on the TaskBar.
- To start a given R question, please go to File...Open Project, and open the folder for the given question and click on the .Rproj file that then appears.
- The upper-left pane within RStudio will contain a script, with a few lines already in it that will load the relevant data and packages. Do not remove or modify these lines. After executing them, you will add whatever code you need to answer the question. **If this script does not appear when the project opens (this will probably be the case the first time you open each project), there will be a file under the FILES tab in the lower right pane with a name like "question46.R" in the same folder as the Rproj file. Click on it to open the script.**
- Answer the question by appending code and comments to the script and running the script. The grader will run your code in order. To run only the lines you have recently entered, you can select them with your mouse and click on the "->Run" button at the top of the script page
- Questions also call for interpretation and commentary. Please insert your interpretation and commentary as comments in your script. As a reminder, comments in R begin with a # and extend to the end of the line.

- Question reviewers will only rely on information contained in your script to grade your answer. They must be able to run that script to recreate your answer, so be sure that your script records every relevant action you have taken. If you execute lines at the console, be sure to copy them to the script if they are necessary for your code to run properly. For example, if you create an object or a variable from the console and then reference that object or variable in your script, the script will not run later for the grader, since that object or variable will never have been created. **Candidates are strongly encouraged to run their script top to bottom (preferably after having cleared objects from the environment) to ensure that it will run as intended for the grader.**
- 33)
- 34) When you have completed a question, or wish to switch to working on a different R question, use "File...Close Project". You will be prompted to save changes to your script file. You should do so. You may also wish to use "File...Save As" (but do NOT change the filename) while working to save changes specifically to the script.
- 35) The environment is set up so that only one RStudio session may be open at a time, so you must Close Project on one R question to work on a different one.

1

Incomplete

Points

10

Fill in the Blank

Consider an ordinary least squares regression

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki} + e$$

with a training dataset $(y_i, x_{1i}, x_{2i}, \dots, x_{ki})$ for $i = 1, 2, \dots, n$ where y_i is the logarithm of the actual response variable.

Fill in the blanks to make the following statement true:

Under the null hypothesis that $\beta_j = 0$, the z-score (z_j) has a _____ distribution, and hence a _____ (absolute) value will lead to the acceptance of this null hypothesis.

- I. normal, large
- II. lognormal, small
- III. t, small
- IV. t, large
- V. lognormal, large

2

Incomplete

Points

15

Multiple Choice

Consider a dataset with n observations. The dependent variable is y and the predictors are x_1, x_2, \dots, x_k . We have fitted a linear regression of the form $y_i = b_0 + b_1 x_{1i} + \dots + b_k x_{ki}$.

Thinking about observations that may have a large influence on the regression coefficients consider the following statements:

- I. The hat value, h_i , summarizes the potential influence of y on all of the fitted values.
- II. An outlier among Cook's D statistic is an observation that exerts substantial influence on the regression coefficients.
- III. Studentized residuals follow a t-distribution with $n + k - 2$ degrees of freedom.

Select the correct statements from the following choices:

- A. All statements (I, II, and III) are true
- B. Only I and II are true
- C. Only I and III are true
- D. Only II and III are true
- E. All statements (I, II, and III) are false

3

Incomplete

Points

15

Multiple Choice

The following list of statements is relevant to the bootstrap methodology

Choose from the list which set of statements are true

- I. In practice, it is usual to generate bootstrap samples from the original population
- II. When sampling observations from a data set in order to generate bootstrap samples the sampling is done without replacement so that the same observation will not occur twice in a bootstrapped sample
- III. If we randomly select n observations for a bootstrap sample from a dataset with N observations it is necessary that $n = N$

- A. I
- B. II
- C. III
- D. None

4

Incomplete

Points
15

Multiple Choice

Identify which of the following is not a threat to external validity

- A. The process of assigning subjects to treatment and control groups
- B. Having a unique and non-generalizable environment for experiment
- C. Dropouts
- D. Non-participation
- E. The answer is not given by (A), (B), (C), or (D)

5

Incomplete

Points

15

Multiple Choice

Identify which of the following is not an important consideration for the feasibility of running a business experiment.

- A. Does the experiment have a testable prediction?
- B. Is management committed to acting on the outcome?
- C. What is the required sample size?
- D. Can the organization feasibly conduct the experiment at the test locations for the required duration?
- E. Can we randomize the treatments?

Consider the following potential criticisms of the use of credit scores in insurance

- I. Credit-score based pricing is more likely to impact the poor, uneducated, and recent immigrants
- II. How a person manages their finances is not related to their driving behavior
- III. Too many levels of price, with too broad a range, are generated by credit scores
- IV. In some cases bad credit can outweigh bad driving history like a drunk-driving conviction

By selecting an option below, indicate which of the above statements are criticisms given by O'Neil

- A. I only
- B. I and III only
- C. I and IV only
- D. I, II and IV only
- E. The answer is not given by (A), (B), (C), or (D)

7	Incomplete
---	------------

Points

30

Briefly explain why "repeated significance testing errors" can be a problem in A/B testing.

8	Incomplete
---	------------

Points

30

Short Answer

Ted is building a predictive model and he is planning to use a cross-validation model validation approach in the choice of model which includes decisions such as the type of the model and feature selection. Since Ted is using a cross-validation approach to model validation he decides that he does not require a holdout sample for the build of his model.

Describe the appropriateness of Ted's model validation approach.

Long Answer

a.

Describe what is meant by "binary classification".

b.

Describe why it is not preferred to fit a binary response variable with a linear regression.

c.

Describe the "odds ratio" and its relationship to logistic regression.

d.

The figure below shows the results of a logistic regression model that predicts the odds of survival for passengers on the Titanic.

```
Call:
glm(formula = Survived ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3804  -0.6562  -0.4300   0.6392   2.3950

Coefficients: (3 not defined because of singularities)
(Intercept)  -1.373105  0.319779  -4.294  1.76e-05 ***
Pclass_1      2.173104  0.359365  -6.053  1.42e-09 ***
Pclass_2      1.302268  0.271680  -4.793  1.64e-06 ***
Pclass_3      NA         NA         NA         NA
Sex_female    2.677814  0.226863  11.804  < 2e-16 ***
Sex_male      NA         NA         NA         NA
Age           -0.031671  0.008945  -3.540  0.000399 ***
SibSp         -0.248975  0.123365  -2.018  0.043570 *
Parch         -0.091603  0.141950  -0.645  0.518718
Fare          -0.001397  0.003179  -0.440  0.660254
Embarked_C    0.431447  0.271693  1.588  0.112288
Embarked_Q    0.533193  0.369337  1.444  0.148837
Embarked_S    NA         NA         NA         NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pclass Describes the passenger class. Categorical variable with 3 levels. Pclass_3 is the base level.

Sex Gender of the passenger. Sex_male is the base level.

Age Age of passenger

SibSp Number of siblings/spouses aboard

Parch Number of parents/children aboard

Fare Passenger fare in British Pounds

Embarked Port of Embarkation (C = Cherbourg, Q=Queenstown, S=Southampton). Emabrked_S is the base level.

Compare the odds of survival for female passengers compared to male passengers holding all other variables constant.

e.

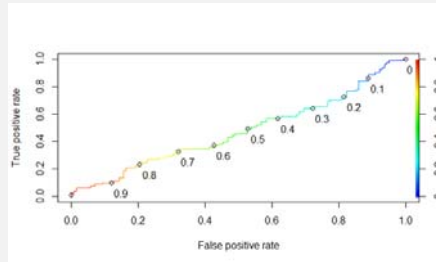
Using the logistic model above, determine the predicted survival probability for a Titanic passenger with the following characteristics:

Variable	Value
PClass	1
Sex	Male (base level)
Age	35
SibSp	0
Parch	0
Fare	50
Embarked	S

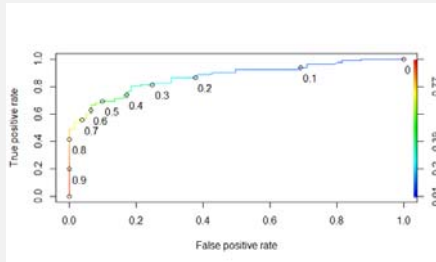
f.

Below are ROC curves for three models that have been fit to the Titanic data. Order these three models from the best fit to the worst fit.

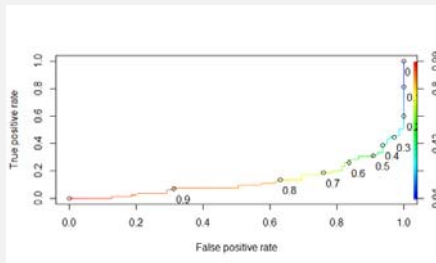
Model 1



Model 2



Model 3



Points

200

COMPLETE 2 OF THE 3 R QUESTIONS. MARK 2 OF THE R QUESTIONS AS "FINISHED" AND LEAVE THE OTHERS AS "INCOMPLETE".

ANSWER THE QUESTION IN THE RSTUDIO PROJECT. ANY DATASETS NEEDED FOR THE QUESTION WILL BE AVAILABLE IN THE RSTUDIO PROJECT.

An insurance company is interested in expanding into personal automobile insurance in a new state.

The table below shows options that the company is considering offering prospective insurance.

The options are for the following policy features:

- * Bodily injury per claimant limit
- * Bodily injury per accident limit
- * Property damage limit
- * Personal Injury Protection (PIP) (in states where both no-fault and tort options are available, PIP is required when the no-fault option is selected for liability indicating that if the insured is injured, the insured will receive payments under the PIP option)

Bodily Injury Limit	Bodily Injury per Accident Limit	Property Damage Limit	PIP
20	100	15	Yes
100	500	200	No
500	1,500		
1,000			

Values in (000)s

The insurance company hires a firm to administer surveys to a sample of potential policyholders from the state to collect information that will help management. The survey presents a number of options to the survey respondents and asks them to rate them from 1 to 10. The data collected from the survey is in the file `conjoint.data.CSV`. The rating is recorded in the variable "rating". BILim, BIAcc, PDLim, and PIP are the BI limit, BI per Accident Limit, Property Damage Limit, and PIP selection for each option rated. The data contains multiple observations per individual, as is typical in a conjoint analysis.

Note: remember to convert all predictor variables to factors, as variables such as Bodily Injury Limit will be read in as numeric variables.

- a.
Display a summary of the survey data
- b.
Perform an analysis to estimate the preferences of the respondents
- c.
What would be the highest rated option?

11

Incomplete

Points

200

COMPLETE 2 OF THE 3 R QUESTIONS. MARK 2 OF THE R QUESTIONS AS "FINISHED" AND LEAVE THE OTHERS AS "INCOMPLETE".

ANSWER THE QUESTION IN THE RSTUDIO PROJECT. ANY DATASETS NEEDED FOR THE QUESTION WILL BE AVAILABLE IN THE RSTUDIO PROJECT.

Your data set contains 10,000 observations of a binary response variable and six predictor variables. The data is split into a training and test sets with an 80/20 split.

- 1) Fit the data with a logistic regression model.
 - a) Use the model to make predictions for the test set.
 - b) How many observations in the test set were predicted correctly?
 - c) What is the accuracy of the model when applied to the test set?
- 2) Fit the data with a KNN model. Before constructing the model, run `set.seed(1)`.
 - a) Use the model to make predictions for the test set.
 - b) How many observations in the test set were predicted correctly?
 - c) What is the accuracy of the model when applied to the test set?
- 3) Based on the results above, which of the two models would you recommend? Why?

12

Incomplete

Points

200

COMPLETE 2 OF THE 3 R QUESTIONS. MARK 2 OF THE R QUESTIONS AS "FINISHED" AND LEAVE THE OTHERS AS "INCOMPLETE".

ANSWER THE QUESTION IN THE RSTUDIO PROJECT. ANY DATASETS NEEDED FOR THE QUESTION WILL BE AVAILABLE IN THE RSTUDIO PROJECT.

This question will use data about the incidence of coronary heart disease in South Africa. Code to read in data is given in the RStudio project. The data files are located in the RStudio project. The response variable is `chd`. Information about other variables may be found in the file:"SAheart.info.txt".

a.

Using linear discriminant analysis, construct a model to predict whether an individual has coronary heart disease. Produce summary output for the model.

b.

Using the model you created in Part 1. Form class predictions for each of the sample observations. When doing this, you only need to form predictions of 1 or 0 for the response `chd`. You do not need to calculate the probabilities of class membership.

c.

What percentage of your predictions from Part 2 were correct? Part 3 - What percentage of your predictions from Part 2 were correct?

d.

Using the model you created in Part 1, make predictions about the new data contained in the file "SAheart_new.csv".