CSPA Exam 3 - Predictive Modeling - Methods and Techniques

Practice Exam Fall 2020

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.

6) 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.

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.

15) 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 on 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.

16) 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.

17) Use of Excel functions (for example SUM, AVERAGE, SUMPRODUCT, etc.) is allowed and encouraged for efficiency but not required.

18) You must clearly specify any additional assumptions you have made to answer the question.

19) 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.

20) DO NOT use named ranges as they may not copy over correctly to the graders.

21) DO NOT use Visual Basic code. It will not be provided to the graders.

22) DO NOT use cell comments. Content in cell comments will not be graded.

23) DO NOT include links to other sheets; linked values in candidate answers will not carry over correctly to the grading files.

24) 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 (R1 to R5)

25) The text of the questions is in the Excel workbook.

26) **The R questions are "Do any 4 of 5". Please indicate in cell B1 of the workbook which questions you intend to have graded by marking them "Finished".**

27) **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.**

28) 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.

29) 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.

30) 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 question4.R in the same folder as the Rproj file. Click on it to open the script.

31) 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.

32) 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.

33) 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.**

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 changed 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.

CSPA Exam 3: Predictive Modeling - Methods and Techniques

Sep-18

Candidates must sign below to confirm acknowledgement of the following:

Candidates must not give or receive assistance of any kind during the examination. Any cheating, any attempt to cheat, assisting others to cheat, or participating therein, or other improper conduct will result in the Casualty Actuarial Society and the Canadian Institute of Actuaries disqualifying the candidate's paper, and such other disciplinary action as may be deemed appropriate within the guidelines of the CAS Policy on Examination Discipline.

Candidate Signature:

(sign here by typing your full name)

After your exam, please log on to the CAS website to complete the Exam Survey.
The Syllabus & Exam Committee values your feedback. Thank you.

This tab will NOT be graded.

This tab will NOT be graded

Question	Points	Status	MC Answer
1	10	Incomplete	0
2	10	Incomplete	0
3	10	Incomplete	0
4	10	Incomplete	0
5	10	Incomplete	0
6	10	Incomplete	0
7	10	Incomplete	0
8	10	Incomplete	0
9	10	Incomplete	0
10	10	Incomplete	0
11	10	Incomplete	0
12	10	Incomplete	0
13	50	Incomplete	
14	40	Incomplete	
15	50	Incomplete	
16	50	Incomplete	
17	40	Incomplete	
18	30	Incomplete	
19	40	Incomplete	
20	20	Incomplete	
21	20	Incomplete	
22	20	Incomplete	
23	60	Incomplete	
24	20	Incomplete	
25	30	Incomplete	
26	20	Incomplete	
27	20	Incomplete	
28	30	Incomplete	
29	20	Incomplete	
30	20	Incomplete	
31	50	Incomplete	
32	20	Incomplete	
33	20	Incomplete	
34	20	Incomplete	
R1	200	Incomplete	
R2	200	Incomplete	
R3	200	Incomplete	
R4	200	Incomplete	
R5	200	Incomplete	

1 Points 10

Incomplete

Consider a dataset with n observations. The dependent variable is y_i and the predictors are $x_1, x_2, \ldots, x_k$.
We have fitted a linear regression of the form $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon$.

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_i 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

Answer:

2 Points 10

Incomplete

Consider the following statements regarding the bootstrap methodology.

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.

Please indicate which of the above statements are correct by selecting the appropriate option below

A. None of I, II, or III are true
B. I and II only
C. I and III only
D. II and III only
E. The answer is not given by (A), (B), (C), or (D)

Answer:

3
Points
10

Incomplete

The following list of statements is relevant to regression and classification trees.
Choose from the list which set of statements are true.

I. A split in a classification tree cannot yield two terminal nodes that have the same predicted value.
II. When building a classification tree the cross entropy is preferable to the Gini index when evaluating the quality of a particular split since it is more sensitive to node purity.
III. When pruning a classification tree the classification error rate is preferable to the Gini index or the cross-entropy when evaluating the quality of a particular split if prediction accuracy of the final pruned tree is the goal.
IV. For a classification tree the residual sum of squares should not be used as a criterion for making binary splits.
V. We are often interested not only in the class prediction corresponding to a particular terminal node region, but also in the class proportions among the training observations that fall into that region.

A. I, II, III, IV, V
B. I, II
C. I, II, III
D. III, IV, V
E. IV, V

Answer:

4
Points
10

Incomplete

Determine which of the following are key assumptions for drawing causal inference from observation studies.

Answer:

I. Subjects are randomly assigned between test and control groups.
II. There is independence, conditional on measured covariates, between the group assignment and the outcome variable.
III. Treatments applied to one unit do not affect the outcome of another unit.

A. None of the above
B. I and II only
C. I and III only
D. II and III only
E. The answer is not given by (a), (b), (c), or (d)

5 Points 10

Incomplete

Consider the following statements (labelled I to VI) regarding the rank ordering of the predictive value of

personality tests,
reference checks, and
cognitive tests

in determining a prospective employee's job performance as discussed in chapter six "Ineligible to Serve" of O'Neil's book Weapons of Math Destruction.

I. Cognitive tests are less predictive than personality tests
II. Personality tests are less predictive than cognitive tests
III. Reference checks are less predictive than personality tests
IV. Personality tests are less predictive than reference checks
V. Reference checks are less predictive than cognitive tests

Select which combination of the above statements is correct:

A. I and III only
B. II and IV only
C. III and V only
D. I and II only
E. none of the above options are correct

Answer:

6 Points 10

Incomplete

Consider the following statements:

I. The process of evaluating a model's performance is known as model assessment.
II. The process of selecting the proper level of flexibility for a model is known as model selection.
III. The bootstrap provides a measure of accuracy of a given statistical learning method.

Determine which of the above statements are true.

A. I, II only
B. I, III only
C. II, III only
D. I, II and III
E. None of the above

Answer:

7 Incomplete

Points 10

A third degree regression spline is to be fitted to some data. Judgement suggests using 5 knots to get a good approximation to the data.

How many parameters do we need to estimate?

A. 6
B. 7
C. 8
D. 9
E. 10

Answer:

8
Points
10

Incomplete

The following list of statements is relevant to regression and classification trees.
Choose from the list which set of statements are true.

I. Trees can be displayed graphically, and are easily interpreted even by a non-expert.
II. Trees cannot easily handle qualitative predictors without the need to create dummy variables.
III. Trees generally do not have the same level of predictive accuracy as other regression and classification approaches discussed in the syllabus.
IV. Trees may outperform classical approaches when there is a non-linear and complex relationship between the features and the response.

A. I, II, III
B. I, II, IV
C. I, III, IV
D. II, III, IV
E. III, IV

Answer:

9 Points 10

Incomplete

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?

Answer:

10 Points
10

Incomplete

The CRISP-DM Data Mining process has six phases as diagrammed below. Two of the phases have been filled in.

Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6
	Data Under-standing		Modeling		

Which of the following is true?

A. Phase 1 is Cost Benefit Analysis
B. Phase 6 is Project Validation
C. Phase 3 is Exploratory Data Analysis
D. Phase 1 is Business Understanding and Phase 6 is Deployment
E. A, B, C, or D are true

Answer:

11
Points
10

Incomplete

A dataset contains consumer spending at the household level, which has been sampled for each county in Pennsylvania.

Consider the following statements:

A data scientist has fit both a “no pooling model” and “multilevel model” to the data.
The data scientist will observe that the _____ model _____ the variation among counties and tend to make the individual counties more different than they are.

Choose the most appropriate response from the following choices:

A. Multilevel; overstates
B. No pooling; understates
C. No pooling; overstates
D. Multilevel; understates
E. No pooling; corrects

Answer:

12 Points 10

Incomplete

Consider an ordinary least squares regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon$$

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

Fill in the blanks to make the following statement true:

Using Type I errors equal to 5% and 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.

A. normal, large (equal or greater than 1.96)
B. lognormal, small (less than 1.96%)
C. t, small (less than 2)
D. t, large (greater than or equal to 2)
E. lognormal, large (greater than or equal to 1.96)

Answer:

13
Points
50

Incomplete

The first phase of an analytics project, as described in the CRISP-DM protocols, is an understanding of the business objectives. Outline phase 1 for the following project:

You have been informed by the management of a personal lines company, that the loss ratio for personal lines automobile in state X has dramatically increased in the past year due to suspected fraud and abuse. You are asked to lead a team of modelers to help address the problem. Describe what activities you would be involved in in developing the business understanding of the project.

14
Points
40

Incomplete

Identify and give a brief description of the four types of missing data mechanisms.

15 Points 50

Incomplete

This question consists of five parts.

Part 1. Describe what is meant by “binary classification”.

Part 2. Describe why is it not preferred to fit a binary response variable with a linear regression?

Part 3. Describe the “odds ratio” and its relationship to logistic regression.

Part 4. 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)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.373105   0.319779  -4.294 1.76e-05 ***
Pclass_1     2.175104   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.
Sex – gender of the passenger.
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)

Compare the odds of survival for female passengers compared to male passengers.

Part 5. 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

16 Points 50

Incomplete

When modeling count data (namely Y_i), we may consider using a zero-inflated Poisson model

Part A. When should you consider using a zero-inflated Poisson regression?

Part B. What are the two components (sub-models) of a zero-inflated Poisson model?
Please explain each component (sub-model)

Part C. What is the probability of observing a zero count?
Please note the pdf of a poisson distribution is p(x;mu) = e^(-mu) * mu^x/x!,
where x is the actual number of successes that result from the experiment, and e is approximately equal to 2.71828, and mu is the mean.
If needed, please use _ for subcript and ^ for superscript

Part D. Show how to calculate the conditional expectation of Y_i , given the sub-model expectations

Part E. Show how to calculate the conditional variance of Y_i , given the sub-model expectations

17
Points
40

Incomplete

When confronted with a multilevel data structure one typically starts by fitting some simple classical regressions and then works ones' way up to a full multilevel model. Identify and describe the four natural starting points using classical regression.

18
Points
30

Incomplete

Briefly explain how the validation set approach, k-fold cross validation, and leave-one-out cross-validation are related to each other.

19
Points
40

Incomplete

Part A. List and briefly describe two sampling methods that can be used to overcome poor performance from a classification algorithm due to an unbalanced data set.

Part B. For each of the sampling methods listed in your answer to Part a, briefly describe a drawback of the method.

20 Points

20

Incomplete

Briefly explain why modelers, using ordinary least squares, are usually not satisfied with prediction accuracy and interpretation.

21
Points
20

Incomplete

Briefly state why when using polynomial regression modelers typically keep the degree of their polynomials less than or equal to 4.

22
Points
20

Incomplete

Briefly explain the procedure known as “bagging” and give an argument as to why “bagging” reduces the variance of the predictions.

23
Points
60

Incomplete

In the context of supervised and unsupervised learning:

Part A. Briefly describe two differences between unsupervised learning and supervised learning.

Part B. Give two situations in which unsupervised learning would be preferred to supervised learning.

Part C. Give two situations in which supervised learning would be preferred to unsupervised learning.

24
Points
20

Incomplete

A modeler wishes to conduct an experiment. There are four treatments, A, B, C, and D to be tested and there are 16 experimental units, labeled 1 through 16. The modeler performs the assignment of treatments to experimental units using the following procedure:

As each experimental unit is encountered, the modeler assign the treatment as follows:

If the seconds hand on his wrist watch is between 1 and 15 seconds treatment A is assigned, if it is between 16 and 30 seconds, then treatment B is assigned, if it is between 31 and 45 seconds, then treatment C is assigned, and if it is between 46 and 60 seconds, treatment D is assigned.

Briefly explain why this assignment procedure should not be used.

25 Points
30

Incomplete

A data scientist has conducted an A/B test to see if a newly designed webpage would increase the conversion rate for an online retail company XYZ. The data scientist wanted to fix the sample size in advance in order to avoid the repeated significance testing errors.

Part A. Briefly define power in the context of A/B testing.

Part B. Briefly define significance in the context of A/B testing.

Part C. Briefly describe why "repeated significance testing errors" is a problem when conducting an A/B test?

26
Points
20

Incomplete

In chapter six “Ineligible to Serve” of Weapons of Math Destruction O’Neil contrasts the predictive models built by professional sports teams and the models used by large corporations in their hiring processes. She mentions three key factors that make personality tests in hiring departments weapons of math destruction.

Please list these three factors and briefly contrast them against their use in professional sports predictive models.

27

Incomplete

Points

20

Consider the validation set approach for measuring test error.

Briefly describe one advantage and one disadvantage of this approach.

28 Points 30

Incomplete

A binary classifier has been fitted to a dataset and the following confusion matrix has been calculated based on a holdout dataset.

		Predicted Class	
		Positive	Negative
Actual Class	Positive	475	116
	Negative	93	841

Based on this confusion matrix, calculate the following metrics:

(a) Accuracy
(b) Precision
(c) Sensitivity
(d) Specificity

29
Points
20

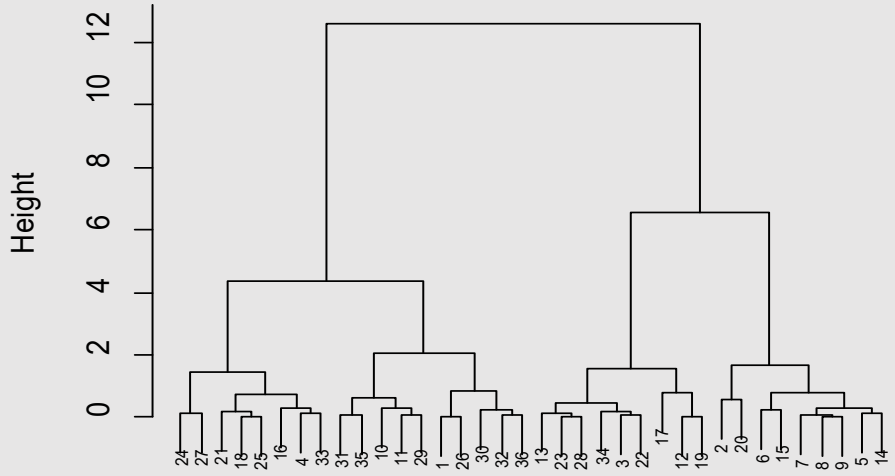
Incomplete

In the context of decision trees, briefly explain what it means to say that they suffer from high variance.

30
Points
20

Incomplete

Consider the following dendrogram built using hierarchical clustering with complete linkage.



Determine how many clusters there are in this data and justify your answer.

31 Incomplete

Points 50

Part A. Using the table provided below, calculate the support, confidence and lift of the association rule that purchases of both product J and K lead to purchases of product L.

Part B. Given your answers to Part A, briefly interpret the support of this association rule.

Part C. Given your answers to Part A, briefly interpret the confidence of this association rule.

Part D. Given your answers to Part A, briefly interpret the lift of this association rule.

	Products			
Transaction	**J**	**K**	**L**	**M**
1	1	1	1	0
2	0	0	0	1
3	1	0	0	1
4	1	1	1	1
5	0	1	0	1
6	1	0	1	0
7	0	1	0	0
8	1	1	0	1
9	1	1	1	0
10	1	1	1	1

32
Points
20

Incomplete

An alternative to using matching methods for causal analysis from observational data is to use regression analysis to control for imbalances in covariate values between test and control units.

Give two reasons why matching methods are preferable and justify each.

33
Points
20

Incomplete

A new start-up company is offering a service to hiring departments that claims to be able to identify candidates that are more likely to remain longer with the company, be more productive, and fit the company's culture.

They claim to be able to do this by analyzing candidate's "social data;" that is, the information candidates leave on the internet; such as, blogs, participation in discussion forums, etc…

Briefly discuss two drawbacks of using such data in a model to identify candidates.

34
Points
20

Incomplete

Cathy O'Neil, in her book Weapons of Math Destruction, discusses how in the earlier days of insurance actuaries developed predictive models from data for large groups or pools of customers.

Discuss how the application of predictive models in insurance causes the departure from the original goals of insurance to share costs over groups or pools of customers.

R1 Incomplete

Points
200

COMPLETE 4 OF THE 5 R QUESTIONS AND MARK THEM "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:

(a) Bodily injury per claimant limit
(b) Bodily injury per accident limit
(c) Property damage limit
(d) 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	1,000		
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.

Part A. Display a summary of the survey data

Part B. Perform an analysis to estimate the preferences of the respondents

Part C. What would be the highest rated option?

R2 Incomplete

Points
200

COMPLETE 4 OF THE 5 R QUESTIONS AND MARK THEM "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 dataset 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.

Part A. 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?

Part B. 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?

Part C. Based on the results above, which of the two models would you recommend? Why?

R3 Incomplete

Points
200

COMPLETE 4 OF THE 5 R QUESTIONS AND MARK THEM "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'. Inforamtion about other variables may be found in the file: "SAheart.info.txt"

Before answering the questions, split the "SAheart.data.csv" file into training and hold out datasets.
Use the first 375 **rows** for the training dataset and the remaining rows as hold out.
NOTE THAT RECORD NUMBER 262 DOES NOT EXIST IN THE DATA.

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

Part B. Using the model created in Part A produce a prediction for each record and plot the prediction on a histogram by the values of 'chd'.

Part C. Score the hold out data set using the model in Part A. Create a confusion matrix and determine the sensitivity and specificity of the model.

R4 Incomplete

Points
200

COMPLETE 4 OF THE 5 R QUESTIONS AND MARK THEM "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.

Market1 is a market research data collected from 150 customers. The data will be used compute association rule statistics. The data contains demographic and product selection information. It has three fruit choice variables (with levels 1 = yes, 0 = no)

apple
orange
pear

The demographic variables is the language(s) spoken by the customer. It also has three variables (with levels 1 = yes, 0 = no)

English (eng)
Chinese (chin)
Japanese (jap)

Part A. Load the workspace Q4.Rdata containing the Market1 data. The data is in the variable 'market1'. Estimate the Pearson correlation matrix for all six variables. Use three decimal places.

Part B. Using type I errors equal to 5%, determine which pair(s) of variables are statistically significant. Hint: use cor.test function. You may use the help command to get more information on this function.

Part C. Compute the support for apples.

Part D. Develop an R function support() to determine the support of two variables that occur together (that is, both are 1 for the same transaction). Use support() to estimate the support for apple -> pear.
Hint: support <- function(X, Y) {
…
return(…) }

Part E. Use the support() function to develop a function market.analysis() with output equal to the association rules: support, confidence, and lift for two sets of variables.
Hint: market.analysis <- function(X,Y) {
supp = support(X, Y)
…
return(list(support = …, confidence = …, lift = …))}

Part F. Use market.analysis() to estimate the support, confidence, and lift of (apple -> pear) and (pear -> apple).
Explain the similarity and difference between the outputs.

R5 Incomplete

Points
200

COMPLETE 4 OF THE 5 R QUESTIONS AND MARK THEM "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.

The Blackmore data frame has 945 rows and 4 columns. Blackmore data on exercise histories of 138 teenaged girls hospitalized for eating data disorders and 98 control subjects.

Load the Blackmore dataset from Q5.RData. (5 points)

Part A. Do a summary and scatterplot of all the column in the data. (15 points)

Part B. The response variable is "exercise" (hr per week). Build a linear regression model with all other predictor variables and test the significance of all predictor variables. (20 points)

Part C. Choose a final linear regression model from Part B based on linear model statistical output. Display the final optimal sub-model. (25 points)

Part D. Update the final model in Part C with a mixed effects model (random intercept and random slope -age). In your linear mixed effects function include the statement:
control=lmerControl(check.conv.singular = .makeCC(action = "ignore", tol = 1e-4))
Without this statement, the function may not converge
Refit the model with a mixed effects model and maximum likelihood (ML) method. Explain the difference between residual maximum likelihood (REML) and ML based on the output of the two model fits. [Hint: use library(lmer)] (50 points)

Part E. Assuming that a value less than 2 is the critical t-value, refit the REML mixed effects models on all fixed parameters with t-value greater than 2, use the likelihood ratio test to determine the final mixed effects models. (50 points)

Part F. Use the likelihood ratio test to compare the mixed effects models fitted by REML in Part E and the final selection linear regression in Part C. Explain how you will select your final model between a fixed effect model and a mixed effects model. (40 points)