FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

END OF SEMESTER EXAMINATION

MODULE CODE: MA4128

SEMESTER: Spring 2018

MODULE TITLE: Advanced Data Modelling

DURATION OF EXAM: 2.5 hours

LECTURER: Kevin O'Brien

GRADING SCHEME: 100 marks
60% of total module marks

EXTERNAL EXAMINER: Prof. A Marshall

INSTRUCTIONS TO CANDIDATES

This paper is comprised of five questions, each worth 25 marks. Attempt any four questions.
Scientific calculators approved by the University of Limerick can be used.
Statistical tables are provided at back of exam paper.

1. (a) (4 Marks) Showing your working, use the Dixon Q Test test that there is no outlier present in the following data set.

$$\{112, 167, 140, 129, 125, 139, 117, 135, 131, 119\}$$

(b) The following statistical procedure is based on this dataset.

$$\{6.98, 8.49, 7.97, 6.64, 8.80, 8.48, 5.94, 6.94, 6.89, 7.47, 7.32, 4.01\}$$

```
> grubbs.test(x, two.sided=T)

Grubbs test for one outlier

data:  x
G = 2.4093, U = 0.4243, p-value = 0.05069
alternative hypothesis: lowest value 4.01 is an outlier
```

(i) (2 Marks) Describe the purpose of this procedure. State the null and alternative hypotheses.

(ii) (1 Mark) Write the conclusion that follows from the code output above.

(iii) (1 Mark) State any relevant assumptions for this procedure.

(c) Consider the following inference procedure performed on data set X.

```
> shapiro.test(X)

Shapiro-Wilk normality test

data:  X
W = 0.84987, p-value = 2.143e-13
```

(i) (2 Marks) Describe the purpose of this procedure. Include in your answer how the outcome of the procedure is to be interpreted.

(ii) (3 Marks) What is the null and alternative hypotheses for this test? Write the conclusion that follows from this procedure.

This question is continued on the next page.

(iii) (2 Marks) A subsequent procedure is reported below. Describe what was attempted in the procedure, and the outcome. Suggest a possible reason for this outcome.

```
> X <- log(X)
>
> shapiro.test(logX)

Shapiro-Wilk normality test

data:  X
W = NaN, p-value = NA
```

(d) The following questions relate to Missing Data.

(i) (2 Marks) What is Missing Data? Discuss the implications of Missing Data in the context of a statistical analysis.

(ii) (3 Marks) Compare and contrast the following types of missing data: Missing At Random, Missing Not At Random, Missing Completely at Random.

(iii) (5 Marks) Describe the technique of Multiple Imputation.

2. (a) (3 Marks) With reference to the table below, define each of the following appraisal metrics in the context of a binary classification procedure (*1 Mark for each*).
   (i) Accuracy
   (ii) Precision
   (iii) Recall

	Predicted Negative	Predicted Positive
Observed Negative	True Negative	False Positive
Observed Positive	False Negative	True Positive

(b) (2 Marks) What is the F-score? Explain why the F-score is considered a more informative measure of performance than the Accuracy score.

$$\text{Hint:} \qquad \text{F-score} = \frac{2 \times P \times R}{P + R}$$

(c) (3 Marks) Calculate the following appraisal metrics using the below table of outcomes for binary classification *(1 Mark for each).*
   (i) Recall,
   (ii) Precision,
   (iii) F-score.

	Predict Negative	Predict Positive
Observed Negative	9710	90
Observed Positive	80	120

(d) (3 Marks) What is a Receiver Operator Character (ROC) curve? Explain its function in the context of a binary classification procedure, how it is determined, and the means of interpreting the curve. Support your answer with sketches.

(e) Answer the following questions relating to the SPSS output on the next page. In this analysis, we wish to predict whether or not a person has a saving's account, based on the following demographic variables.

- Age
- Socio-economic Status
- Sector within city
- Disease Status

There are three possible outcomes for socio-economic status.

$$\{1 = \text{Upper}, 2 = \text{Middle}, 3 = \text{Lower}\}$$

There are three possible outcomes for socio-economic status.

$$\{1 = \text{Inner City}, 2 = \text{Inner Suburbs}, 3 = \text{Outer Suburbs}\}$$

The Disease status variable is a binary variable, with 1 indicating the presence of some sort of illness.

This question is continued on the next page.

(i) (1 Mark) Describe how Wald's Test was used to refine the initial model. Make reference to relevant figures in the output.

(ii) (2 Marks) What is a logit? How can you transform a logit into a probability?

(iii) (1 Mark) State the regression equation for the final logistic regression model.

(iv) (4 Marks) What information is contained in the column labeled `Exp(B)`? For the initial model, interpret the figures from this column for both ***Socioeconomic status*** and ***Sector within city***. As part of your answer, comment on the 95% confidence intervals for both.

(v) (2 Marks) Predict the outcome for the following case: a 55 year old person from the upper socio-economic category residing in the outer suburbs.

(vi) (2 Marks) Predict the outcome for the following case: a 25 year old person from the middle socio-economic category residing in the inner suburbs.

(vii) (2 Marks) What is a dummy variable? Explain how it is used in Logistic Regression. Support your answer with an example.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	.037	.010	13.802	1	.000	1.038
	Socioeconomic Status	-.941	.200	22.242	1	.000	.390
	Sector within city	.732	.356	4.221	1	.040	2.078
	Disease Status	-.120	.390	.095	1	.758	.887
	Constant	.157	.699	.051	1	.822	1.170
Step 2[a]	Age	.036	.010	14.030	1	.000	1.037
	Socioeconomic Status	-.944	.199	22.377	1	.000	.389
	Sector within city	.703	.343	4.189	1	.041	2.020
	Constant	.183	.694	.070	1	.792	1.201

This question is continued on the next page.

Variables in the Equation

		95% C.I.for EXP(B)	
		Lower	Upper
Step 1[a]	Age	1.018	1.058
	Socioeconomic Status	.264	.577
	Sector within city	1.034	4.177
	Disease Status	.413	1.904
	Constant		
Step 2[a]	Age	1.018	1.057
	Socioeconomic Status	.263	.575
	Sector within city	1.030	3.959
	Constant		

a. Variable(s) entered on step 1: Age, Socioeconomic Status, Sector within city, Disease Status.

3. (a) (4 Marks) Compute the following distance metrics between the cases A, and B, described below. (*1 Mark for each*).
   (i) Euclidean Distance
   (ii) Squared Euclidean Distance
   (iii) Manhattan Distance
   (iv) Chebyshev Distance

$$A = \{5, 9, 2, 11, 4\}$$
$$B = \{3, 6, 9, 4, 7\}$$

   (v) (1 Mark) Explain why the squared Euclidean distance may be used in preferences in to the Euclidean Distance.

   (b) The following questions relate to Hierarchical Clustering.
   (i) (1 Mark) Distinguish between agglomerative and divisive hierarchical clustering techniques.
   (ii) (3 Marks) Why do you standardize variables before carrying out a cluster analysis. Support your answer with an example.
   (iii) (3 Marks) Describe the process of Ward's Linkage in the context of cluster analysis.
   This question is continued on the next page.

(iv) (9 Marks) Describe any three of the following linkage methods. Support your answer with sketches (*3 Marks for each*).

- Nearest Neighbour Linkage
- Furthest Neighbour Linkage
- Centroid Linkage
- Average Linkage

(v) (2 Marks) In the context of cluster analysis, What is the chaining effect? Give a brief description, supporting your answer with sketches.

(vi) (2 Marks) Describe how a dendrogram would assist in the interpretation of a hierarchical clustering solution. Support your answer with a sketch.

4. (a) The following questions relate to K-means Clustering.

   (i) (1 Mark) Compare and contrast k-means clustering and hierarchical clustering in terms of the number of clusters determined.

   (ii) (7 Marks) Explain the process of k-means clustering, starting with an initial cluster allocation. You may work on the basis of a two-cluster solution. Support your answer with several sketches.

   (iii) (2 Marks) Describe a graphical procedure to assist in determining the appropriate number of clusters.

   (iv) (2 Marks) For a 4 cluster k-means solution, interpret the ANOVA table below.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Net profit	495,145	3	1419,744	3	,349	,795
Own funds	2878,202	3	2537,200	3	1,134	,460
Assets	842788,443	3	9987,138	3	84,387	,002
Client deposits	634017,636	3	35643,498	3	17,788	,021
Loans	957411,333	3	37401,709	3	25,598	,012

(b) The following topics relate to techniques for predictive modeling count techniques.

   (i) (2 Marks) What does a Poisson regression model? State any assumptions that must be checked before it can be used.

   (ii) (1 Mark) The R Code output given below is used to predict the number of awards won by students.

   - Information is provided on which of the three school programs the student takes part in (*General*, *Vocational* or *Academic*).
   - Also we are given the mathematics test score.

   State the mathematical formula used to predict the number of awards won.
   You can denote ***progAcademic***, ***progVocational*** *and* ***math*** *as* x_1,x_2 *and* x_3 *respectively.*

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.2471     0.6585   -7.97  1.6e-15 ***
progAcademic    1.0839     0.3583    3.03   0.0025 **
progVocational  0.3698     0.4411    0.84   0.4018
math            0.0702     0.0106    6.62  3.6e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This question is continued on the next page.

(iii) (2 Marks) Use the model in Part (ii) to predict the number of awards won by a general program student, with a maths score of 55.

(iv) (2 Marks) Use the model in Part (ii) to predict the number of awards won by an academic program student, with a maths score of 75.

(v) (1 Mark) Describe the circumstances whereby Negative Binomial Regression Models would be used instead of Poisson Models.

(vi) (3 Marks) What is Zero Inflation? Explain the modeling process for a Zero Inflated Model. Give an example of Zero-Inflated Count Process. *Support your answer with a sketch, if necessary.*

(vii) (2 Marks) What is Zero Truncation? Give an example of a Zero Truncated Count Process.

5. (a) The following questions relate to multicollinearity in the context of multiple regression analysis.

(i) (1 Mark) Define multicollinearity.

(ii) (2 Marks) State two ways in which a multiple regression analysis could be affected by severe multicollinearity.

(iii) (2 Marks) State two ways of formally diagnosing the severity of multicollinearity, making reference to how both should be used to make decisions about the data.

(b) The following questions relate to Principal Component Analysis.

(i) (2 Marks) What is the purpose of a principal component analysis?

(ii) (1 Marks) Principal Component Analysis is a Dimensionality Reduction technique. Explain what this term means.

(iii) (4 Marks) What is meant by the "true" dimension of the data? How does an analyst determine the appropriate number of principal components to retain, making reference to three different approaches.

(iv) (3 Marks) The Kaiser-Meyer-Olkin (KMO) statistic is used to measure a certain characteristic of the data. What is this characteristic? Explain how the KMO statistic should be interpreted.

(v) (2 Marks) Briefly describe the Bartlett Test for Sphericity, with reference to the null and alternative hypotheses, and how those statements relate to the purpose of the test.

(c) The following questions relate to model selection and validation in the context of multiple regression analysis.

(i) (1 Mark) Explain the purpose of variable selection procedures.

This question is continued on the next page.

(ii) (3 Marks) Compare and contrast the following variable selection procedures (*1 Mark for each*).

- Forward Selection
- Backward Elimination
- Stepwise Regression

(iii) (1 Mark) Explain how the *Akaike information criterion* is used to compare two models fitted for the same data.

(iv) (1 Mark) Explain why the adjusted R^2 value may differ in value from the corresponding multiple R^2 value for the same fitted model.

(v) (3 Marks) Describe model validation in the model-building process, with particular emphasis on the standard data partition.

(d) (5 Marks) For the computer code output on the following pages, there are 7 iterations of a model fitting process. The response variable is denoted Y, and the predictor variables are denoted $V1$ to $V10$. Describe what the process is doing, stating how and what conclusion is reached.

Iteration 1

```
Start:  AIC = -1107.36
Y ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10

       Df Sum of Sq     RSS      AIC
- V5    1   0.00005 0.91210 -1109.35
- V4    1   0.00012 0.91217 -1109.33
- V2    1   0.00241 0.91446 -1108.81
- V7    1   0.00517 0.91722 -1108.18
- V3    1   0.00576 0.91781 -1108.05
- V1    1   0.00784 0.91989 -1107.58
<none>              0.91205 -1107.36
- V8    1   0.00972 0.92177 -1107.15
- V6    1   0.01340 0.92544 -1106.32
- V9    1   0.02434 0.93639 -1103.88
- V10   1   0.72914 1.64118  -987.16
```

Iteration 2

```
Step:  AIC = -1109.35
Y ~ V1 + V2 + V3 + V4 + V6 + V7 + V8 + V9 + V10

Df Sum of Sq     RSS      AIC
- V4    1   0.00038 0.91248 -1111.26
- V2    1   0.00238 0.91448 -1110.80
- V7    1   0.00512 0.91722 -1110.18
- V3    1   0.00594 0.91804 -1110.00
- V1    1   0.00779 0.91989 -1109.58
<none>              0.91210 -1109.35
- V8    1   0.01000 0.92210 -1109.08
- V6    1   0.01761 0.92971 -1107.37
+ V5    1   0.00005 0.91205 -1107.36
- V9    1   0.02431 0.93641 -1105.87
- V10   1   0.72945 1.64155  -989.11
```

Iteration 3

```
Step:  AIC = -1111.26
Y ~ V1 + V2 + V3 + V6 + V7 + V8 + V9 + V10

Df Sum of Sq     RSS      AIC
- V2    1   0.00224 0.91472 -1112.75
- V7    1   0.00534 0.91782 -1112.04
- V3    1   0.00679 0.91927 -1111.72
- V1    1   0.00753 0.92001 -1111.55
<none>              0.91248 -1111.26
- V8    1   0.01006 0.92254 -1110.98
+ V4    1   0.00038 0.91210 -1109.35
- V6    1   0.01739 0.92987 -1109.33
+ V5    1   0.00031 0.91217 -1109.33
- V9    1   0.02411 0.93659 -1107.83
- V10   1   0.73044 1.64292  -990.94
```

Iteration 4

```
Step:  AIC = -1112.75
Y ~ V1 + V3 + V6 + V7 + V8 + V9 + V10

Df Sum of Sq     RSS      AIC
- V3    1   0.00459 0.91930 -1113.71
- V1    1   0.00529 0.92001 -1113.55
- V7    1   0.00578 0.92050 -1113.44
<none>              0.91472 -1112.75
- V8    1   0.00970 0.92442 -1112.55
+ V2    1   0.00224 0.91248 -1111.26
- V6    1   0.01737 0.93209 -1110.84
+ V4    1   0.00023 0.91448 -1110.80
+ V5    1   0.00017 0.91455 -1110.79
- V9    1   0.02429 0.93900 -1109.30
- V10   1   0.73427 1.64898  -992.17
```

Iteration 5

```
Step:  AIC = -1113.71
Y ~ V1 + V6 + V7 + V8 + V9 + V10

Df Sum of Sq     RSS      AIC
- V7    1   0.00772 0.92703 -1113.97
<none>              0.91930 -1113.71
- V8    1   0.00938 0.92869 -1113.60
+ V3    1   0.00459 0.91472 -1112.75
- V1    1   0.01641 0.93571 -1112.03
+ V4    1   0.00120 0.91810 -1111.98
+ V5    1   0.00016 0.91914 -1111.74
+ V2    1   0.00003 0.91927 -1111.72
- V9    1   0.02333 0.94263 -1110.50
- V6    1   0.02597 0.94527 -1109.91
- V10   1   0.73279 1.65209  -993.78
```

Iteration 6

```
Step:  AIC = -1113.97
Y ~ V1 + V6 + V8 + V9 + V10

Df Sum of Sq     RSS      AIC
- V8    1   0.00403 0.93106 -1115.07
<none>              0.92703 -1113.97
+ V7    1   0.00772 0.91930 -1113.71
+ V3    1   0.00653 0.92050 -1113.44
+ V4    1   0.00159 0.92544 -1112.33
- V1    1   0.01678 0.94381 -1112.24
+ V5    1   0.00047 0.92656 -1112.07
+ V2    1   0.00011 0.92692 -1111.99
- V6    1   0.01831 0.94534 -1111.90
- V9    1   0.02020 0.94723 -1111.48
- V10   1   0.72507 1.65210  -995.78
```

Iteration 7

```
Step:  AIC = -1115.07
Y ~ V1 + V6 + V9 + V10

Df Sum of Sq     RSS      AIC
<none>              0.93106 -1115.07
+ V3    1   0.00539 0.92568 -1114.27
+ V8    1   0.00403 0.92703 -1113.97
+ V7    1   0.00238 0.92869 -1113.60
- V9    1   0.01644 0.94750 -1113.42
+ V4    1   0.00128 0.92978 -1113.35
+ V5    1   0.00012 0.93095 -1113.09
+ V2    1   0.00009 0.93097 -1113.09
- V1    1   0.01857 0.94963 -1112.96
- V6    1   0.02851 0.95958 -1110.79
- V10   1   0.72205 1.65311  -997.65
```

```
Call:
lm(formula = Y ~ V1 + V6 + V9 + V10, data = Sonar2)

Coefficients:
(Intercept)           V1           V6           V9          V10
0.03883      0.44476      0.21224     -0.16043      0.91507
```

Tables

Critical Values for Dixon Q Test

N	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	0.941	0.97	0.994
4	0.765	0.829	0.926
5	0.642	0.71	0.821
6	0.56	0.625	0.74
7	0.507	0.568	0.68
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568
11	0.392	0.444	0.542
12	0.376	0.426	0.522
13	0.361	0.41	0.503
14	0.349	0.396	0.488
15	0.338	0.384	0.475
16	0.329	0.374	0.463