Dublin Business School

QQI

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS

MARCH 2020 EXAMINATIONS

Module Code:	**B8IT109**
Module Description:	**Advanced Data Analytics**
Examiner:	**Dr Shahram Azizi Sazi**
Internal Moderator:	**Mr Paul Laird**
External Examiner:	**Ms Catherine Mulwa**

Date: 30/03/2020
Time: 18:30-20:30

INSTRUCTIONS TO CANDIDATES

- *This is an open book-material exam, students are allowed to use their own laptop, lecture notes, code, and different websites to respond the questions.*
- *Please select four questions out of five questions. Explicitly specify your selected questions on your submission.*
- *R code and necessary outputs (i.e. graphs/plots/curves) need to be saved in word format and submit to Moodle.*

Question 1

Use **mtcars** dataset and consider **mpg** and **vs** as the attributes of interest.

a) Use the appropriate probability models to quantify the uncertainty in mpg and vs.
(5 Marks)

b) Estimate the parameters of your proposed models using the dataset.
(5 Marks)

c) Predict the future values of mpg and vs using (a) and (b).
(10 Marks)

d) Using (a), (b), find P(mpg > 90). **(5 Marks)**

(TOTAL: 25 Marks)

Question 2

In regression analysis, the **Boston** dataset is analysed in R and its output is as follows.

```
 Call:
lm(formula = medv ~ crim + zn + indus + chas + nox + rm, data =
Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-21.016  -3.420  -0.684   2.506  39.467

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.95464    3.21376  -5.587 3.81e-08 ***
crim         -0.17691    0.03459  -5.114 4.50e-07 ***
zn            0.02128    0.01385   1.537   0.1249
indus        -0.14365    0.06394  -2.247   0.0251 *
chas          4.78468    1.05909   4.518 7.81e-06 ***
nox          -7.18489    3.69353  -1.945   0.0523 .
rm            7.34159    0.41720  17.597  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.943 on 499 degrees of freedom
Multiple R-squared:  0.5874,    Adjusted R-squared:  0.5824
F-statistic: 118.4 on 6 and 499 DF,  p-value: < 2.2e-16
```

a) Using this output, specify the response and independent variables. **(5 Marks)**

b) Based on the output, which type of GLM is proposed for this analysis.
(5 Marks)

c) List the assumptions for your proposed regression model. **(5 Marks)**

d) Specify the significant independent variables on the response variable at the level of $\alpha = 0.05$. **(5 Marks)**

e) Using the output, find the optimal predictive model for the response variable.
(5 Marks)

(TOTAL: 25 Marks)

Question 3

Loading the package **'datasets',** use the dataset **'readingSkills'** and consider **nativeSpeaker** as the output variable.

a) Split the dataset into 80% as the train-set and 20% as the test-set. (use set.seed(104))

(2.5 Marks)

b) Apply Random Forest (RF) algorithm to train the classifier using train-set with 20 trees. **(5 Marks)**

c) Predict the test-set using the trained model of classifier. **(2.5 Marks)**

d) Provide the confusion matrix and obtain the accuracy. **(5 Marks)**

e) Redo parts b-d to apply either Naïve Bayes or Decision Tree. Which model does provide the higher accuracy? **(10 Marks)**

(TOTAL: 25 Marks)

Question 4

Use **data('EuStockMarkets')** to load the in-built dataset 'EuStockMarkets' in R, consider ***DAX*** as your time series variable:

(a) Validate the assumptions using graphical visualization.

(5 Marks)

(b) Fit the optimized ARIMA model for ***DAX*** and provide the coefficient estimates for the fitted model. **(10 Marks)**

(c) What is the estimated order for AR and MA?

(5 Marks)

(d) Forecast h=10 step ahead prediction of ***DAX*** on the plot of the original time series.

(5 Marks)

(Total: 25 Marks)

Question 5

Use dataset available on
http://users.stat.ufl.edu/~winner/data/hybrid_reg.csv

(a) Use LDA to classify the dataset into few classes so that at least 85% of information of dataset is explained through new classification. (**Hint**: model the output variable "**carclass_id"** to input variables "**msrp**", "**accelrate**", and "**mpg**"). How many LDs do you choose? Explain the reason. **(10 Marks)**

(b) Apply PCA to input variables, and identify the important principle components involving at least 90% of dataset variation. Explain your decision strategy? Plot principle components versus their variance (**Hint**: to sketch the plot use the Scree plot). **(5 Marks)**

(c) Use K-means clustering analysis to input variables and identify the most important classes. How many classes do you select? Why?

(5 Marks)

(d) Split the dataset into two sets of variables so that **X**=(msrp, mpgmpge) and **Y**=(accelrate, mpg). Apply canonical correlation analysis to find the cross-correlation between **X** and **Y**. What is the correlation between ***msrp*** and ***mpg***?

(5 Marks)
(Total: 25 Marks)

End of Examination