Dublin Business School
Excellence Through Learning

Dublin Business School
Assessment Brief

Assessment Details

Module Title:	Advanced Data Analytics
Module Code:	
Module Leader:	Dr Shahram Azizi
Stage (if relevant):	
Assessment Title:	
Assessment Number (if relevant):	
Assessment Type:	
Restrictions on Time/Length :	Submission before deadline
Individual/Group:	
Assessment Weighting:	
Issue Date:	
Hand In Date:	
Planned Feedback Date:	
Mode of Submission:	Online

Question 1

Use in-built dataset 'airquality',

a) explore the general feature of dataset using appropriate R functions. **(5 Marks)**

b) perform data cleansing if required. **(5 Marks)**

c) consider 'Temp' attributes and compute the central and variational measures. **(10 Marks)**

d) apply boxplot technique to detect outlier of 'wind' attribute if any. **(10 Marks)**

(Total: 30 Marks)

Question 2

Use dataset available on http://users.stat.ufl.edu/~winner/data/nfl2008_fga.csv , then:

(a) Train the model using 80% of this dataset and suggest an appropriate GLM to model **homekick** to **togo, ydline** and **kicker** variables.

(5 Marks)

(b) Specify the significant variables on **homekick** at the level of α=0.05, and estimate the parameters of your model.

(5 Marks)

(c) Predict the test dataset using the trained model.

(5 Marks)

(d) Provide the confusion matrix and obtain the probability of correctness of predictions.

(10 Marks)

(Total: 25 Marks)

Question 3

Using Yahoo Finance API, select a specific stock market price, apply time series analysis, consider '***close price*** as your time series variable:

(a) Validate the assumptions using graphical visualization.

(5 Marks)

(b) Fit the optimized model for '***close price'*** and provide the coefficient estimates for the fitted model.

(5 Marks)

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

(5 Marks)

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

(10 Marks)

(Total: 25 Marks)

Question 4

Use dataset available on
http://users.stat.ufl.edu/~winner/data/nfl2008_fga.csv

1. Use LDA to classify the dataset into few classes so that at least 90% of information of dataset is explained through new classification. (**Hint**: model the variable "**qtr"** to variables "**togo**", "**kicker**", and "**ydline**"). How many LDs do you choose? Explain the reason.
   (5 Marks)

2. Apply PCA, 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)

3. Split the dataset into two sets of variables so that **X**=(togo,kicker,ydline) and **Y**=(distance, homekick). Apply canonical correlation analysis to find the cross-correlation between **X** and **Y**. What is the correlation between ***ydline*** and ***distance***?

(5 Marks)

4. Use K-means clustering analysis to identify the most important classes. How many classes do you select? Why?

(6 Marks)

(Total: 20 Marks)

Note: Technical support is available to student between **0930- 1700 hrs only**. There is no technical support after 1700 hrs. It is your responsibility to ensure that you allow time to troubleshoot any technical difficulties by uploading early on the due date.