Candidate Number:

ISSU0053 Data Science and Big Data Analytics

UCL International Summer School for Undergraduates 2019

**Assessment II Computer Practical under examination conditions (50%)**

**Timing Friday 19th July 10am – 12:30pm**

**Examination conditions:**

* **The session is open book, so you can consult your notes, textbooks and programming websites as you work.**
* **You must work on the UCL desktop.**
* **All work submitted must be your own. All forms of communication and messaging with other students is strictly prohibited, and violations will be dealt with in accordance with UCL policy.**
* **At the end of the assessment you will have a short time to collate the files you have produced and upload them for marking.**

**Notes:**

Marks for questions are indicative, and a grading curve may be applied to generate a final grade.

All code used to complete the tasks must be submitted in R script files or R notebook files in Rmd format. Partial marks will be awarded for code sections that have been completed but are non-functional.

You should work in one file per section, so at the end of the exam upload either e.g.

sectionA.R, sectionB.R, sectionC.R sectionD.R, sectionE.R

or sectionA.Rmd, sectionB.Rmd, sectionC.Rmd, sectionD.Rmd, sectionE.Rmd

It is recommended you use comments / notes in your files to indicate which part of the question you are attempting, e.g.

# A1

plot(…)

# A2

summary(boston)

**Section (A)**

*Start a new R file for your answer to this question.*

The file cafe.csv contains data on the sales of items in a café that opens on Mondays to Fridays and is closed at the weekend.

Write R code to:

A1. Load the contents of the file into a data frame.

(1 mark)

A2. Count the number of rows in the data frame.

(1 mark)

A3. Display the names of the columns that have been loaded.

(1 mark)

A4. Edit the data frame so that. column sales\_usd is renamed to sales

(1 mark)

A5. Drop all rows with NA values and calculate the number of rows removed.

(2 marks)

A6. Edit the data frame to remove the location column.

(1 mark)

A7. The column day stores the weekday as an integer using the following mapping:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| mon | tues | wed | thurs | fri |

Adjust the data frame so that day uses the names as given above in a factor type column.

(3 marks)

A8. Add a column drinks that totals the number of drinks (juice/soda/coffee) sold each day.

(2 marks)

(12 marks)

**Section (B)**

*Start a new R file for your answer to this question.*

Load the following file into R. This contains the data frame: cafe\_data. This contains the modified version of the data frame you worked on in Section (A) without the drinks column.

load("SectionB.Rda")

Write R code to find:

B1. The total number of sales made summed over all the days recorded in the dataset.

(1 mark)

B2. The highest number of coffees sold in a single day.

(1 mark)

B3. The mean value of sales on each Monday.

(1 mark)

B4. The mean value of sales on each Friday.

(1 mark)

B5. Perform a t-test to test the hypothesis that the average daily sales differ between a Monday and a Friday.

(1 mark)

You should see the following result.

Welch Two Sample t-test

data: ...

t = 2.0599, df = 12.671, p-value = 0.06057

alternative hypothesis: ...

95 percent confidence interval:

-1.510121 60.081233

sample estimates:

mean of x mean of y

122.3456 93.0600

B6. Interpret the result against the hypothesis based on a significance level of 0.05.

(1 mark)

B7. Write R code to examine the dataset and display the maximum and minimum temperatures over the recorded period.

(1 mark)

B8. Make the following fits:

model 1. predict sales of chips using temperature

model 2. predict sales of coffee using temperature

model 3. predict sales of soda using temperature

(2 marks)

B9. Display the fits using screenreg.

(1 mark)

============================================

Model 1 Model 2 Model 3

--------------------------------------------

(Intercept) 11.99 \*\*\* 45.31 \*\*\* 11.35 \*

(2.48) (3.39) (4.59)

temp -0.07 -0.57 \*\*\* 0.44 \*\*\*

(0.06) (0.08) (0.10)

--------------------------------------------

R^2 0.03 0.55 0.28

Adj. R^2 0.01 0.54 0.27

Num. obs. 47 47 47

RMSE 5.49 7.52 10.17

============================================

\*\*\* p < 0.001, \*\* p < 0.01, \* p < 0.05

B10. Explain (with reference to the p-values) how we can interpret the value and significance of the coefficients associated with chips, coffee and soda.

(3 marks)

B11. Use the fit to predict of the number of coffees sold when the temperature is i) 30 degrees and ii) 80 degrees.

(2 marks)

(15 marks)

**Section (C)**

*Start a new R file for your answer to this question.*

Start by loading the cafe\_data: load("SectionB.Rda")

In this question we will try to predict sandwich sales based on the other column values.

C1. Write code to display the frequency table for daily sandwich sales.

(1 mark)

C2. Build a linear regression model to predict sandwich sales from the other columns using forward stepwise selection until an optimal model is reached.

(3 marks)

If you are unable to complete this question. You can generate the resulting fit using the code below (also found in the code\_example.R file):

lm\_step = lm(sandwich ~ day + cookies + temp + sandwich\_waste + wraps\_waste + sales + wraps, data=cafe\_data)

C3. Examine the diagnostic plots associated with this model.

Why is row 8 highlighted in the final plot?

(1 mark)

C4. Write an R command that displays the data associated with this row.

(1 mark)

C5. Do you feel this row should be considered an outlier to the rest of the dataset?

Explain your answer.

(1 mark)

C6. Write R code that uses the fitted residuals to calculate the RSS value for the fit.

(You should find this to be 71.129).

(1 mark)

C7. Calculate the estimated RMSE of the fit using

(1 mark)

C8. The number 47 in the formula above refers to the number of rows.

What does the other number refer to?

(1 mark)

C9. Which of the terms MSE, RMSE, TSS and RSS is of most use to someone considering the accuracy of a prediction produced by the model?

(1 mark)

C10. Use R to build a simple model that predicts sandwich sales using day as a single predictor.

(1 mark)

This should generate a table of coefficients like:

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)   3.5556     0.5871   6.056  3.3e-07 \*\*\*

daytues       3.1444     0.8093   3.885 0.000356 \*\*\*

daywed        1.4444     0.8093   1.785 0.081509 .

daythurs      1.7444     0.8093   2.156 0.036896 \*

dayfri       -1.1806     0.8559  -1.379 0.175077

C11. Examine the coefficents. How can we interpret the meaning of these values?

(2 marks)

C12. Use R to compare the performance of this model with the larger model using an ANOVA test.

(1 mark)

This should produce the following table:

Analysis of Variance Table

Model 1: sandwich ~ day

Model 2: sandwich ~ day + cookies + ...

Res.Df RSS Df Sum of Sq F Pr(>F)

1 42 130.297

2 36 71.129 6 59.168 4.9911 0.0008236 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

C13. Interpret the result of the ANOVA test with reference to the coefficients of the additional predictors included in the larger model.

(1 mark)

C14. Explain why even through the performance of the larger model is better it is of no practical purpose in predicting sandwich sales.

(1 mark)

C15. Why do we use cross validation methods to compare models when we already have estimates like R2, Adjusted-R2, and estimated RMSE to measure performance?

(2 marks)

C16. When performing k-fold validation using cvFit we can provide an argument R.

Explain what this argument does, and why it is useful to make use of it.

(2 marks)

C17. When performing k-fold on this data set what is the maximum number of folds we could use?

(1 mark)

C18. Perform k-fold validation on the following models with R=20 and 5 folds. Include a suitable command to ensure that your analysis is reproducible.

(4 marks)

model 1. predict number of sandwich sales based on weekday

model 2. predict number of sandwich sales based on weekday and temperature

You should expect to obtain results that are similar (but not identical) to:

|  |  |  |
| --- | --- | --- |
| **Predictors** | **RMSE** | **Standard Error** |
| day | 1.873878 | 0.03238197 |
| day and temperature | 1.848384 | 0.0495979 |

C19. Which of these models is best to use in this case? Explain your reasoning.

(1 mark)

(27 marks)

**Section (D)**

*Start a new R file for your answer to this question.*

Load in the diabetes data set db\_all using the command.

load("SectionD.Rda")

D1. Make two histogram plots showing the distributions of the entries in columns insulin, and log\_insulin

(2 marks)

D2. Based on these plots would insulin or log\_insulin be better to use as a predictor in a logistic regression or LDA classification analysis? Explain your answer.

(2 marks)

D3. Write the R command that could be used to create column log\_insulin based on the values in the insulin column.

(1 mark)

D4. Use the data in db\_all to create a plot similar to the one shown below:

(6 marks)



allocation of marks for:

1 mark – displaying the correct data

1 mark – matching x and y axis limits

1 mark – matching x and y axis labels

1 mark – matching title

1 mark – matching legend

1 mark – colouring points according

to diabetes category

(red: positive & black: negative)

You may use ggplot if you wish.

***PLEASE CHECK RESOURCE FOLDER FOR COLOUR VERSION OF PLOT!***

D5. Look at the plot. Do you think it will be possible to use insulin and glucose to predict whether a person will be in the positive group, i.e. developed diabetes? Explain your answer.

(2 marks)

In order to run both logistic regression and K-nearest neighbours methods we will create columns that store rescaled predictors for log\_insulin and glucose.

This can be done by applying the scale function:

db\_all$glucose\_rs <- scale(db\_all$glucose)

db\_all$log\_insulin\_rs <- scale(db\_all$log\_insulin)

D6 . Write R code that displays the mean and standard deviation of the rescaled columns, and comment on how the columns are now scaled.

(2 marks)

To run our analysis we will create a dataframes db\_train containing a random sample containing 80% of the full dataset. The rest of the data will be used as a validation set and stored in data frame db\_test.

D7. Write R commands that shows how to create these data frames from db\_all.

(3 marks)

(18 marks)

**Section (E)**

*Start a new R file for your answer to this question.*

In order to ensure consistency load in the db\_train and db\_test data frames that have been generated for you:

load("SectionE.Rda")

E1. Perform a logistic regression to predict the diabetes class based on the log\_insulin\_rs and glucose\_rs values.

(2 marks)

E2. Use the resulting model to make predictions for the test data set, predicting diabetes if the model predicts more than a 50% or 0.5 probability for a positive diagnosis.

(2 marks)

E3. Create a confusion matrix in the following format. You should find the following result:

(2 marks)

actual

predicted neg pos

neg 46 19

pos 2 12

E4. How many total cases of positive diabetes onset were there in the test dataset?

(1 mark)

E5. How many of these cases of diabetes onset were correctly predicted by the model?

(1 mark)

E6. Hence determine the true positive rate TPR:

(1 mark)

E7. Make a new set of predictions, placing people in the positive category if they are predicted to have more than a 20% or 0.2 probability of being in the group of people developing diabetes.

(1 mark)

actual

predicted neg pos

neg 32 7

pos 16 24

E8. Show that this increases the true positive rate (TPR).

(1 mark)

E9. Suggest a reason why this could be advantageous.

(1 mark)

E10. In what respect has making this change decreased our measured performance?

(1 mark)

We will now try to classify the cases using k-nearest neighbours method.

E11. From db\_train and db\_test create the following:

db\_x\_train a data frame containing rescaled predictors for glucose and log(insulin)

db\_y\_train a column vector containing diabetes class.

db\_x\_test a data frame containing rescaled predictors for glucose and log(insulin)

db\_y\_test a column vector containing diabetes class.

(4 marks)

*If you cannot complete E11. you can load these objects from the SectionE\_KNN.Rda file.*

E12. Apply the method of KNN to predict the class of the test data set with 5 nearest neighbours.

(3 marks)

**Note.** The knn function requires the db\_y\_train to be 1D object. Depending on the way you create db\_y\_train you may end up with a 2D data frame that consists of a single column. In this case you can convert it using command:

db\_y\_train <- unlist(db\_y\_train).

E13. Calculate the misclassification rate.

(1 mark)

E14. The following code creates a sequence of 48 numbers running from 1 to 300:

k\_vals = c( 1:20, seq(30,300,10) )

Write code to test the KNN method using these values for K, and plot the resulting misclassification rate against K.

(4 marks)

E15. Code a method to repeat the KNN analysis resampling the train and test groups several times (start with ~10 repeats) in order to generate a measure of averaged performance.

Plot the resulting improved performance estimates vs K, including +/- 1 standard error limits.

(7 marks)

(32 Marks)