Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

VII SEMESTER B.TECH
END SEMESTER EXAMINATIONS, NOVEMBER 2019
SUBJECT: PROGRAM ELECTIVE VI : ADVANCED DATA SCIENCE [CRA 4012]
REVISED CREDIT SYSTEM
(28/11/2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL** the questions.
- Missing data, if any, may be suitably assumed.

1A. Explain procedure and considerations for K-fold and Random subsampling cross validation approaches. **5**

1B. Create a shiny application that takes a numeric value 'n' between 1 and 500, a colour from a list of colours and a main title. Use these inputs to create a histogram of selected colour using random data from any distribution. **3**

1C. What is hard thresholding? How is it done? **2**

2A. Explain the following with an example: **5**

i. Two types of layouts that can be used in shiny package.
ii. Types of input controls in Shiny package.

2B. Explain the following commands with respect to forecasting in R: **3**

i. getSymbols("TICKER", src="google", from=date, to=date)
ii. window(ts, start=1, end=6)
iii. ets(train, model="MMM")

2C. Justify the statement "Unsupervised prediction is effectively an exploratory technique". **2**

3A. What is the significance of using Swirl in R. Describe with an example, the eight types of questions supported by swirlify package. **5**

3B. Given the output of **nearZeroVar**(training,saveMetrics=TRUE), find the variables which are not useful to construct a prediction model with justification to the answer. **3**

```
##               freqRatio percentUnique zeroVar   nzv
## year           1.017647     0.33301618   FALSE FALSE
## age            1.231884     2.85442436   FALSE FALSE
## sex            0.000000     0.04757374    TRUE  TRUE
## maritl         3.329571     0.23786870   FALSE FALSE
## race           8.480583     0.19029496   FALSE FALSE
## education      1.393750     0.23786870   FALSE FALSE
## region         0.000000     0.04757374    TRUE  TRUE
## jobclass       1.070936     0.09514748   FALSE FALSE
## health         2.526846     0.09514748   FALSE FALSE
## health_ins     2.209160     0.09514748   FALSE FALSE
## logwage        1.011765    18.83920076   FALSE FALSE
## wage           1.011765    18.83920076   FALSE FALSE
```

3C. What is the significance of a box plot and heatmaps? How do you create box plot and **2**

heatmaps using plotly in R ?

4A. Explain Linear Regression model with explanation to each term in the model. Write R code to perform the following:

i. Build a linear model *m1* to predict *y* using predictor variable *x*
ii. Print summary of the model
iii. Predict outcome variable for new data *df*
iv. Calculate RMSE for new data *df*

5

4B. Describe gvisTable function in googleVis package. Write R code to combine geoChart, gvisTable and motionChart together using googleVis package. **3**

4C How does R support object oriented programming? Explain. **2**

5A. Given testing, training and validation data, and outcome variable *y* perform the following operations using R

- train the data using glm model
  glm.fit ←
- train the data random forest model using train control parameters method as cv and number 3
  rf.fit ←
- use these models to predict the results on the testing set
  glm.pred.test ←
  rf.pred.test ←
- combine the prediction results and the true results into new data frame
  combinedTestData ←
- run a Generalized Additive Model (gam) model on the combined test data
  comb.fit ←

5

5B. What is a leaflet? Write R code to draw map with all the cities represented as circles proportional to the population of the city. [Assume your own dataset] **3**

5C. Write R code to develop the following features using Shiny package

i. Create a dropdown menu to select a dataset
ii. Display the data in the data table.

2

****