PGDM / PGDM(IB) 2016-18

Predictive Analytics

Subject Code- DM-342 / IB-341

Time: 2 hrs. 30 mins. **Term III, End Term Exam.** 29th March 2017

For this examination you have to work on your laptop, using R. Your answers as well as analysis outputs will have to be written on the Answer Script provided.

Marks will be awarded for clarity and completeness in answers.

Section A

Instructions for this section
From Google Drive to which you have been given access, pick up the comma separated dataset *'HousePrices.csv'*. Questions in this part are based on this dataset.
Answer any 3 questions .... **(3 x 5)**

A1. What is the proportion of 'Brick' houses in this dataset?

A2. The number of Brick and non-Brick houses by Neighborhood as shown below is given by the *table* command

```
##        Neighborhood
## Brick East North West
##    No   26    37   23
##    Yes  19     7   16
```

Create this table and find the total number of 'Brick' houses, and the number of houses in the 'East'

A3. Explain the following command. What does the resultant output mean?

```
r[r[,2]==c(min(r$Price),max(r$Price)),8]

## [1] North West
## Levels: East North West
```

A4. Find out the average price of houses in the different neighborhoods. What command can be used to find the neighborhood where the average price is the highest?

A5. Write a set of commands as follows:

(i) Write a command to append a column named *price-per-sqft* to the dataset.

(ii) Display the values for the new column for the first four (using the head function) records

(iii) Find the average price-per-sqft in each of the neighborhoods

30

PGDM / PGDM(IB) 2016-18
Predictive Analytics
Subject Code- DM-342 / IB-341
Time: 2 hrs. 30 mins. **Term III, End Term Exam.** 29th March 2017

Section B

Instructions for this section
From Google Drive to which you have been given access, pick up the comma separated dataset *'LoanData.csv'*. Questions in this part are based on this dataset.
Answer any 2 questions .... **(2 x 10)**

B1. In the loan data, *Borrower.Rate* is dependent on the Loan *Status*, *Credit.Grade* and the Loan *Amount*. For Credit Grading, AA is the highest or best grade (most eligible for a loan), followed by A, B, C, D, E, HR (High Risk) and NC grades.
The following is the output of the regression model created to predict the Borrower.Rate

Model 1

```
l <- read.csv('LoanData.csv',header=T)
attach(l)
l1 <- lm(Borrower.Rate ~ Status + Amount)
summary(l1)

##
## Call:
## lm(formula = Borrower.Rate ~ Status + Amount)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.244718 -0.047053  0.001193  0.049418  0.253393
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.007e-01  1.340e-03 149.816   <2e-16 ***
## StatusDefault  6.459e-02  7.721e-03   8.365   <2e-16 ***
## StatusLate     5.209e-02  3.667e-03  14.203   <2e-16 ***
## Amount        -2.311e-06  1.999e-07 -11.564   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06637 on 5607 degrees of freedom
## Multiple R-squared:  0.06876,Adjusted R-squared:  0.06826
## F-statistic:   138 on 3 and 5607 DF,  p-value: < 2.2e-16
```

PGDM / PGDM(IB) 2016-18
Predictive Analytics
Subject Code- DM-342 / IB-341

Time: 2 hrs. 30 mins. **Term III, End Term Exam.** 29th March 2017

Model 2

```
l2 <- lm(Borrower.Rate ~ Status + Amount + Credit.Grade)
summary(l2)

##
## Call:
## lm(formula = Borrower.Rate ~ Status + Amount + Credit.Grade)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.259076 -0.022797 -0.000583  0.029488  0.220238
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.238e-02  2.408e-03  38.364  < 2e-16 ***
## StatusDefault    3.662e-02  5.094e-03   7.188 7.42e-13 ***
## StatusLate       2.157e-02  2.441e-03   8.836  < 2e-16 ***
## Amount           2.591e-06  1.448e-07  17.895  < 2e-16 ***
## Credit.GradeAA  -1.695e-02  2.950e-03  -5.744 9.74e-09 ***
## Credit.GradeB    3.046e-02  2.812e-03  10.832  < 2e-16 ***
## Credit.GradeC    6.111e-02  2.611e-03  23.407  < 2e-16 ***
## Credit.GradeD    9.863e-02  2.596e-03  37.994  < 2e-16 ***
## Credit.GradeE    1.389e-01  2.566e-03  54.143  < 2e-16 ***
## Credit.GradeHR   1.405e-01  2.603e-03  53.976  < 2e-16 ***
## Credit.GradeNC   1.247e-01  5.797e-03  21.512  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04354 on 5600 degrees of freedom
## Multiple R-squared:  0.5996,Adjusted R-squared:  0.5989
## F-statistic: 838.8 on 10 and 5600 DF,  p-value: < 2.2e-16
```

(i) Comment on the two models. Which one is better and why?

(ii) Explain the significance of the dummy variable Credit.Grade in Model 2 and its influence on the Borrower.Rate.

(iii) What is the predicted *Borrower.Rate* for Status='Current', Credit Grade = 'AA' and Amount=2500.

(iv) How many linear regression equations can be read from Model 2 summary?

31

B2. What command would you use to plot a scatter diagram for Borrower.Rate (y-axis) versus Amount (x-axis)?
What command would you use to plot the scatter diagram with different plot characters for the different Status indicators?
(Hint: *pch* is used for plot character; ifelse to select different plot characters for different *Status*)
What command(s) would you use to superimpose 3 parallel regression lines - no interaction between *Status* and *Amount*?

B3. Create a regression model of Borrower.Rate on Status, Amount and the interaction between Status and Amount. Write down the three different equations for the three different Status levels and interpret the same.

Section C

Instructions for this section
From Google Drive to which you have been given access, pick up the comma separated dataset *'DirectMarketing.csv'*.
The mandatory question in this section is based on this dataset. .... **(1 x 15)**

C1. Create a Regression Model to predict Amount Spent, field *AmountSpent.*

(i) Examine the distribution of AmountSpent using histogram (R function *hist()*).
Find out the minimum, maximum and average value of *AmountSpent.*
How many records have *AmountSpent* value as outlier, where outlier is defined as $value > \mu + 2\sigma, or\ value < \mu - 2\sigma$.

(ii) Find out the average *AmountSpent* for the three *Age* categories

(iii) Create a regression model (without interaction) for *AmountSpent.*
Consider significant 'factors' and numeric variables. Justify your model.

(iv) Now create a second model with an additional variable *Salary*Age.*
Is this an improved model? Why or why not?