17 April 2019 - Dublin Business School
CA2: Data Web Mining
Module code B8IT108
Student: Stefano Piras

Analyze the project Data file using the tools learned for missing data.

CONSIDERATIONS about the dataset:

- The Dataset contains 16 variables.
- Response variable has to be predicted using the other variables.
- X variables are the actual measurements (7 columns).
- Y variables are the categorical form of the X variables (7 columns).
- Group is the variable that might moderate the predictability of response by Xs or Ys: Group1 and Group2
- Response is 0 or 1
- We do not know the nature of the values.

Response	Group	Y1	Y2	Y3	Y4	Y5	Y6	Y7	X1	X2	X3	X4	X5	X6	X7
	1														
	2														

The original data has been divided in subsets:

- Group12YX contain the whole dataset
- Group1Y
- Group1X
- Group2Y
- Group2X
- Group1YX
- Group2YX
- Group12Y
- Group12X

The following instructions allow me to find out the number of missing data for each X variable (1-7):

```
Data$X1[is.na(Data$X1)]
```

```
Data$X2[is.na(Data$X2)]
Data$X3[is.na(Data$X3)]
Data$X4[is.na(Data$X4)]
Data$X5[is.na(Data$X5)]
Data$X6[is.na(Data$X6)]
Data$X7[is.na(Data$X7)]
```

From the last analytis I find out the number of missing values for each variable:
X1:4
X2:130
X3:131
X4:0
X5:4
X6:63
X7:24

CONSIDERATIONS:

- The total number of rows is 296
- The only variable that does not have missing data is the X4
- The variables with the higher number of missing values are X2 and X3, with 130 and 131 missing values, which means they are missing approximately 30% of their values.
- They are followed by x6 (63), X7 (24) and X1 and X5 with 4 each
- Ones I notice that X2 and X3 are the variables with the higher missing data, I assume that if the NA fields are filled with Mean, Median, KNN or linear regression, those 2 variables assume more importance and start playing a bigger role in the calculation of the response.
- For the set given, we do not have much information regarding the nature of the data. For this reason I assume that the missingness is MCRA, Missingness completely at random.

I will substitute the missing values with the mean, the median, KNN and linear regression and see if the percentage of the predictability of the model is good

CODE USED FOR MEAN AND MEDIAN

The following instruction allow me to find the mean of each value:

```
meanX1 <- mean(Data$X1[!is.na(Data$X1)])
```

The following instruction allow me to find the Median of each value:

```
medianX1 <- median(Data$X1[!is.na(Data$X1)])
```

Mean and Median of X2, X3, X6, X7 are the following:

meanX1: 301.3014
meanX2: 2908.602
meanX3: 5015.436
meanX5: 35.31654
meanX6: 3.836481
meanX7: 1353.129
medianX1:38
medianX2:126
medianX3: 192
medianX5: 19.3
medianX6: 3.6
medianX7: 653.25

I notice that the difference between the mean and the median for the X2 and X3 is extremely high.
This is due to the nature of their values that are very different with a high standard deviation.
In facts, by looking at the values I notice, beside the 30% of the missing data, also a high number of zeros and a few numbers of high values.
In the following lines i calculate the standard deviation for the X2 and X3 variables, just to have a prove that their value is high.

```
sdX2 <- sd(Data$X2[!is.na(Data$X2)])
sdX3 <- sd(Data$X3[!is.na(Data$X3)])
```

Standard Deviation for X2: 9293.693
Standard Deviation for X3: 16360.2

Given this data, I would expect, by substituting the missing values with the mean or the media, a very big change in the predictability of the models, **and not necessary in a better way**. Lets try.

Code Used for developing and plot the models

```
# Develop and plot the DT model:
DT_Model1 <-rpart(Response~., data=Data1)
DT_Model1
plot(as.party(DT_Model1))
print(DT_Model1)
```

```
attach(DT_Model1)
```

Code Used for the calculation of the predictability of the model

```
# Predict
PredModel1 = predict(DT_Model1, type="vector")
table(PredModel1)
table(PredModel1,Data1$Response)
table1<-table(PredModel1,Data1$Response)
PercModel1<-sum(diag(table1))/sum(table1)
PercModel1
```

THE FOLLOWING TABLE CONTAINS THE OUTPUT OF THE CALCULATION OF THE PREDICTABILITY FOR:

- Original data without substitution of missing values
- Missing values substituted with the mean
- Missing values substituted with the median

		Predictability
Without replacement	G12XY	0.8378378
	G12X	0.8378378
	G12Y	0.7736486
	G1XY	0.8020833
	G2XY	0.85
	G1X	0.8020833
	G1Y	0.7083333
	G2X	0.855
	G2Y	0.77
	G2X	0.855
	G2Y	0.77
		Predictability
Replacement with MEAN	G12XY	0.8344595
	G12X	0.8344595
	G12Y	0.7736486
	G1XY	0.8229167
	G2XY	0.86

	G1X	0.8229167
	G1Y	0.7395833
	G2X	0.855
	G2Y	0.77
		Predictability
Replacement with MEDIAN	G12XY	0.8344595
	G12X	0.8344595
	G12Y	0.7533784
	G1XY	0.8020833
	G2XY	0.84
	G1X	0.8020833
	G1Y	0.7083333
	G2X	0.84
	G2Y	0.765

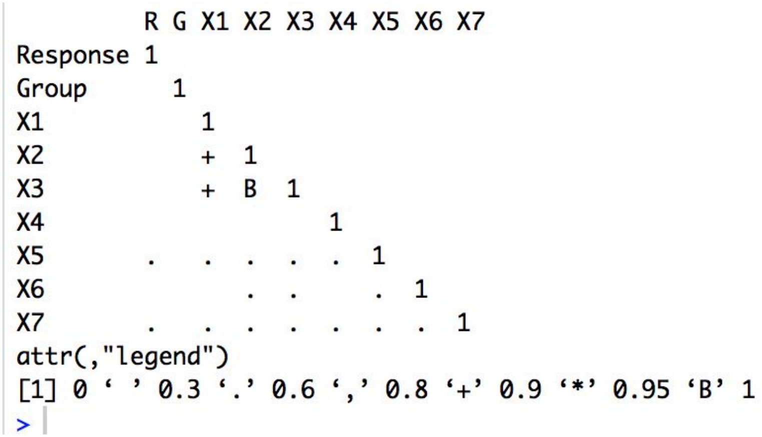
MISSING VALUES CALCULATED WITH LINEAR REGRESSION:

For simplicity purpose, in this case I will only analyze the columns Response, Group and X1-X7, because only the values X are related directly to the real world and the Ys only categorical values of it.

The most important part is to find the correlation between the variables. The code used is the following:

```
cor(Data)
cor(Data, use = "complete.obs")
symnum(cor(Data, use = "complete.obs"))

# From this analysis I find out that the highest correlation is for the X3, correlated to X2 at 95%
# and X2 is correlated to X1 at 90%. Since X1 has only 4 values missing, X2 130 and X3 131, I
# will infer the X2 from X1 and X3 from X2
```



```
         R G X1 X2 X3 X4 X5 X6 X7
Response 1
Group      1
X1           1
X2           +  1
X3           +  B  1
X4                    1
X5       .   .  .  .  .  1
X6              .  .     .  1
X7       .   .  .  .  .  .  .  1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
>
```

Predictability obtained inducting X2 from X1 and X3 from X2 is: **0.8378378**

KNN:
For the KNN imputation, the core instruction that allows imputing the missing values is the following:

Impute1 <- kNN(Data, variable= c ("X1", "X2", "X3", "X5", "X6", "X7"), k=5)

The predictability obtained is: 0.8243243

MIXED METHODS:
An other way to impute missing data into a dataset is a mixing the 4 methods.
In this way we can apply different logics. None of the will guarantee us that the result is perfect because still, any imputation done risks to worsen the quality of the data and make it reflect the real world even less than the original data with the missing values.
I tried to impute the missing data of X1 with KNN as the number of missing values is only 4.
Then, since the correlation between X2 and X1 is more than 80 % and X3 and X2 between 90 and 95%, I imputed the values using the linear regression.
For the X5, X6 and X7 I used the KNN values again.
The Predictability obtained is: 0.8145243

CONCLUSIONS:

Considered that X2 and X3 are the variables with the highest number of missing data but both highly correlated to X1, my conclusion is that we can simplify the model by ignoring them, as most of the plots drawn with the various models show.

The most important variable is X4. By imputing the missing data with the various models, this variable has not lost its importance, remaining in all the plots drown the first split.

In the end of this exercise, I am still very curious about the nature of the data. By not having this important information, I am not able to decide if the missing data is MAR, MCAR or MNAR. This information could have helped me to impute the data in a more reasonable way and with some logic.