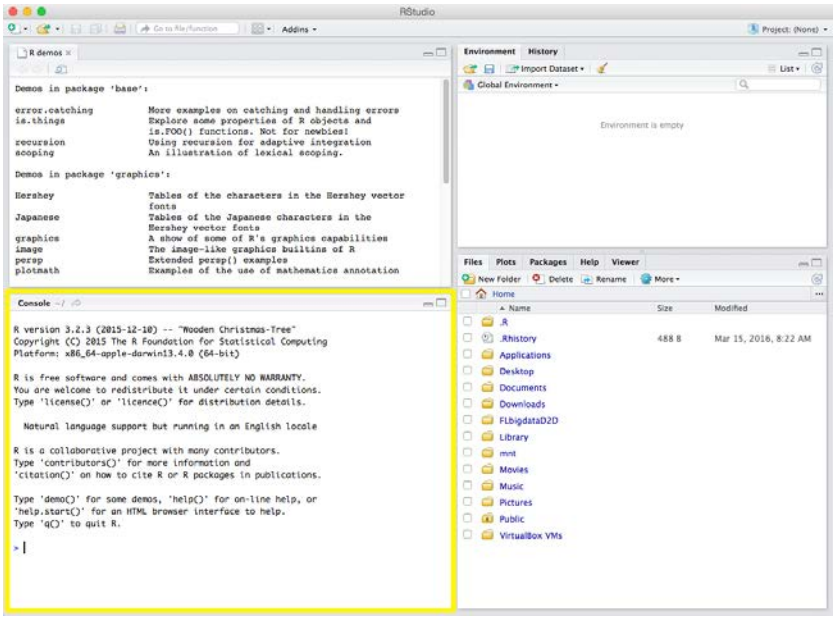
Instructions: Which customers will accept offers?

In this exercise you will use RStudio and H2O to create a logistic regression model using our bank customer dataset.

1) Open RStudio. Enter each command in the steps below, one line at a time, into your RStudio console.



2) Load H2O and start the H2O server locally.

```
library(h2o)
localH2o = h2o.init(ip = "127.0.0.1", port = 54321)
```

3) Import the dataset.

   a) Load the file path into a variable:

```
filePath = "~/FLbigdataStats/bank_customer_data.csv"
```

 © 2016 Queensland University of Technology | CRICOS No. 00213J

b) Load the dataset and save it to the local handle 'market_data':

```
market_data <- h2o.uploadFile(filePath,
                              destination_frame = "",
                              parse = T,
                              header = T,
                              sep = ",",
                              na.strings = c("unknown"),
                              progressBar = FALSE,
                              parse_type = "CSV")
```

Note: In our exercises the H2O server is local so the data ends up in our RAM. If the server were in the cloud the data would be stored there.

4) Inspect the data.

a) Print a summary of the data frame, fetched from the H2O server:

```
market_data
```

b) Fetch summary statistics for columns from the server:

```
summary(market_data)
```

c) Confirm 'y' is an appropriate response variable for this prediction problem.

d) Assess whether it's a continuous or categorical variable.

You will recall that a categorical variable has a set number of classes, whereas a continuous variable can take any value within a range or across infinity. This will influence the type of model and analysis that you do.

e) Inspect the 20 potential predictor variables (columns 1-20 in the data).

f) Assess whether these would be useful for this prediction problem.

Tip

Refer to the description of the dataset found in the .txt file alongside the .csv file

 © 2016 Queensland University of Technology | CRICOS No. 00213J

5) Use the H2O - R interface to build a predictive logistic regression model using the response variables you selected.

   a) Remove the 11th variable from the data set as this refers to the call length which is information we would not normally have available for prediction.

```
market_dataex1 <- market_data[,-11]
```

   b) Split the dataset into a training set and a validation set.

```
split_data <- h2o.splitFrame(market_dataex1, ratios=0.75)
train_data <- split_data[[1]]
validation_data <- split_data[[2]]
```

   c) Fit the predictive model.

```
glm_model = h2o.glm(x = 1:19,
                    y=20,
                    training_frame = train_data,
                    validation_frame = validation_data,
                    max_iterations = 100,
                    solver="L_BFGS",
                    family="binomial",
                    alpha = 1, #L2 regularisation
                    intercept = T)
```

6) Examine how the model performed.

```
summary(glm_model)
```

 © 2016 Queensland University of Technology | CRICOS No. 00213J