# Assignment 2 - Logistic Regression

*Eoin Flynn*

*5 March 2018*

Bond University
Data Science

# Contents

# Data

```
library(RMySQL)
```

```
## Loading required package: DBI
```

```
USER <- "root"
PASSWORD <- "A13337995"
HOST <- "localhost"
DBNAME <- "world"

statement <- "Select * from world.customerChurn"
db <- dbConnect(MySQL(), user = USER, password = PASSWORD, host = HOST,
    dbname = DBNAME, port = 3306)
customerDataset <- dbGetQuery(db, statement = statement)
dbDisconnect(db)
```

```
## [1] TRUE
```

```
# Loops through and changes all relevant rows to factors and
# returns the dataset post modification
setRowAsFactor <- function(dataset, columns) {
    for (column in columns) {
        dataset[, column] <- as.factor(dataset[, column])
    }
    return(dataset)
}

customerDataset <- setRowAsFactor(customerDataset, c("gender",
    "SeniorCitizen", "Partner", "Dependents", "PhoneService",
    "MultipleLines", "InternetService", "OnlineSecurity", "OnlineBackup",
    "DeviceProtection", "TechSupport", "StreamingTV", "StreamingMovies",
    "Contract", "PaperlessBilling", "PaymentMethod", "Churn"))

# Drop the columns that will not be needed
customerDataset = customerDataset[, -which(names(customerDataset) %in%
    c("customerID"))]
```

## Split Data

We will now split our data into test and training sets. The purpose of this is to create a sample of data that the model has never seen before in order to gauge its accuracy. The training set will consist of 80% of the data while the remaining 20% will constitute the test set.

```
suppressMessages(library(caTools))

# Set the seed to reproducability
set.seed(12216)

# Create our two datasets
sample <- sample.split(customerDataset, SplitRatio = 0.8)

train_df <- subset(customerDataset, sample == TRUE)
test_df <- subset(customerDataset, sample == FALSE)
```