Final R Programming Edureka Certificate Project

Olufemi George

20 April 2017

R Markdown

This is my final Edureka Project for the R Programming Certficate Firstly, i will load up the libraries i need for data wrangling, machine learning and visualization

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.3.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.3.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.3.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.3.3
```

```
## Loading required package: lattice
```

Import the project dataset

```
df.retail <- read.csv("Retail_Case_Study_Data.csv")
```

Take out the ID

```
df.retail = df.retail[,-1]
```

Convert spend category and dependent variables to Factor

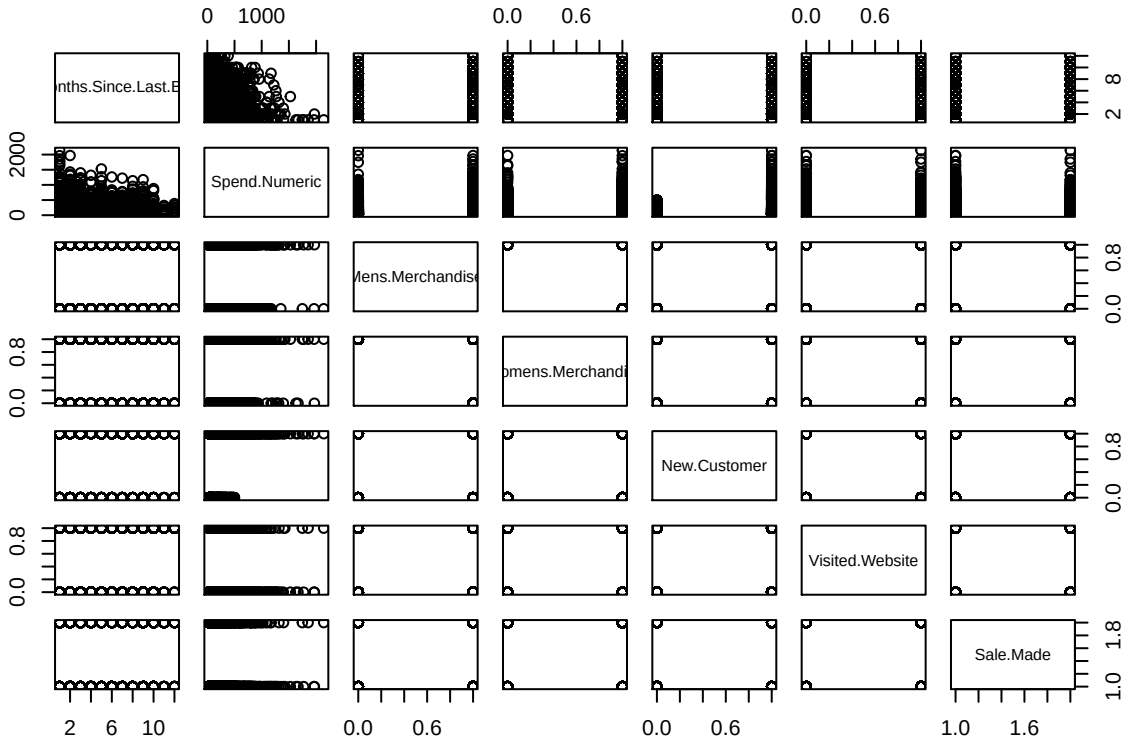
```
df.retail$Spend.Category<-as.factor(df.retail$Spend.Category)
df.retail$Sale.Made<-as.factor(df.retail$Sale.Made)
str(df.retail)
```

```
## 'data.frame':    1747 obs. of  10 variables:
##  $ Months.Since.Last.Buy: int  1 1 2 1 1 1 9 1 2 1 ...
##  $ Spend.Category       : Factor w/ 7 levels "1) $0 - $100",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ Spend.Numeric         : num  30 30 30 30 30 ...
##  $ Mens.Merchandise      : int  1 1 0 1 0 1 0 0 1 0 ...
##  $ Womens.Merchandise    : int  0 0 1 0 1 0 1 1 0 1 ...
##  $ Area                  : Factor w/ 3 levels "Rural","Surburban",..: 2 2 3 1 1 1 2 3 1 2 ...
##  $ New.Customer          : int  1 1 1 0 1 1 0 1 1 1 ...
##  $ Purchase.Channel      : Factor w/ 3 levels "Multichannel",..: 2 3 2 2 3 2 3 3 2 3 ...
##  $ Visited.Website       : int  0 0 1 0 0 0 1 0 1 0 ...
##  $ Sale.Made             : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

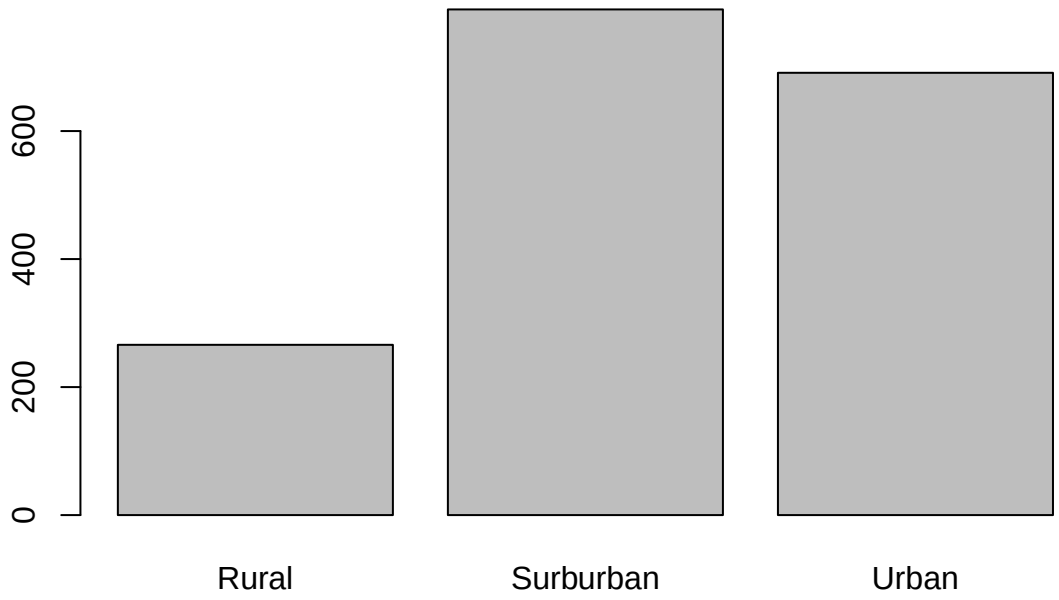
see if any correlation between numeric variables

```
pairs(df.retail[,-c(2,6,8)])
```



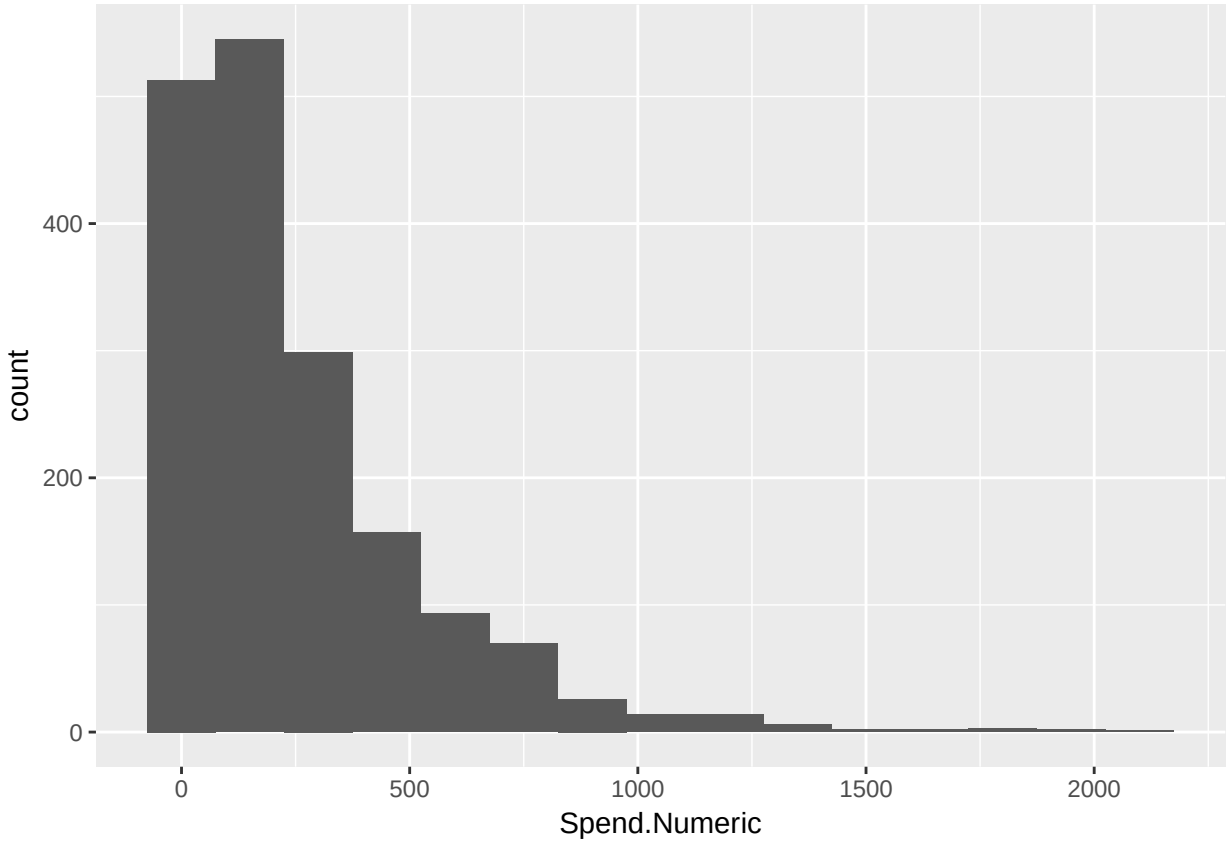
Univariate analysis

```
barplot(xtabs(~df.retail$Area))
```

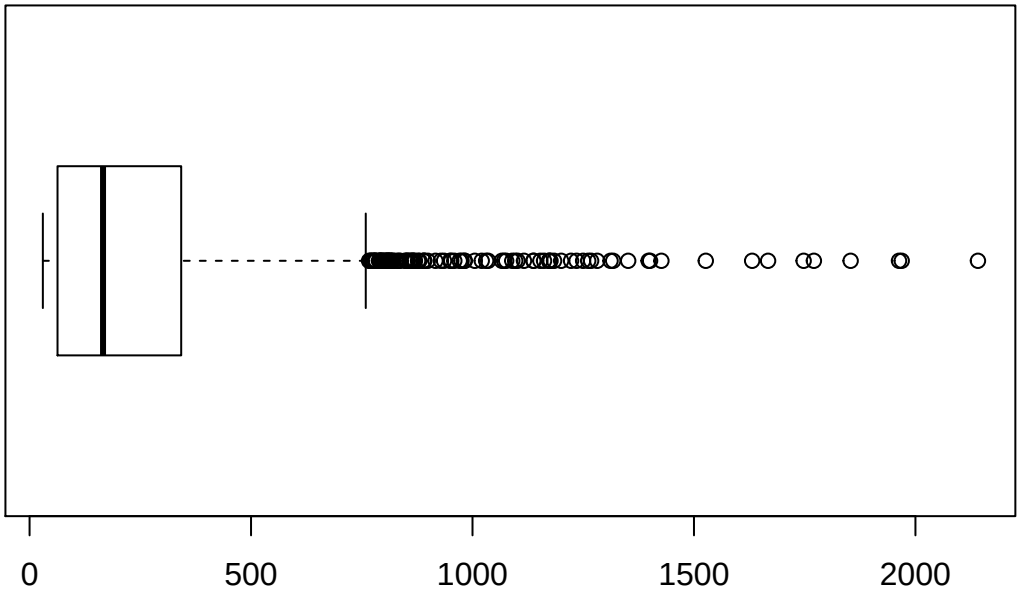


Univariate analysis using ggplot2

```
ggplot(df.retail, aes(x=Spend.Numeric)) +
  geom_histogram(binwidth = 150)
```

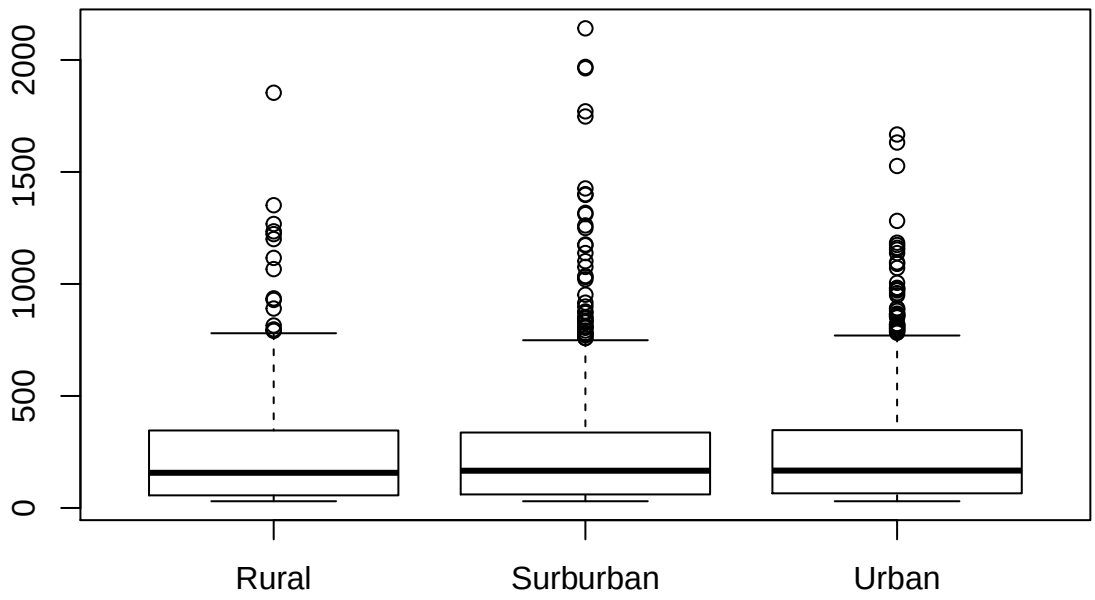



```
boxplot(df.retail$Spend.Numeric, horizontal = TRUE)
```



Side by side boxplot

```
boxplot(df.retail$Spend.Numeric~ df.retail$Area)
```



Split data for training and testing datasets

```
validationIndex <- createDataPartition(df.retail$Sale.Made, p=0.80, list=FALSE)
validation <- df.retail[-validationIndex,]
dataset <- df.retail[validationIndex,]
```

Build the model regressor using all independent variables

```
classifier= glm(formula=Sale.Made~.,
                family=binomial,
                data=df.retail)
summary(classifier)
```

```
##
## Call:
## glm(formula = Sale.Made ~ ., family = binomial, data = df.retail)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5195  -0.5994  -0.4641  -0.3525   2.2949
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.852134   0.400401  -4.626 3.73e-06 ***
## Months.Since.Last.Buy          -0.042406   0.020100  -2.110  0.03488 *
## Spend.Category2) $100 - $200   -0.561612   0.203689  -2.757  0.00583 **
## Spend.Category3) $200 - $350   -0.971529   0.297778  -3.263  0.00110 **
## Spend.Category4) $350 - $500   -0.832353   0.445339  -1.869  0.06162 .
```

```
## Spend.Category5) $500 - $750   -1.568851   0.633624  -2.476  0.01329 *
## Spend.Category6) $750 - $1,000 -1.093183   0.881328  -1.240  0.21483
## Spend.Category7) $1,000 +      -1.434958   1.386927  -1.035  0.30084
## Spend.Numeric                   0.000741   0.001018   0.728  0.46661
## Mens.Merchandise                0.002689   0.227466   0.012  0.99057
## Womens.Merchandise             -0.072224   0.225960  -0.320  0.74925
## AreaSurburban                   0.304667   0.194784   1.564  0.11779
## AreaUrban                       0.295917   0.197286   1.500  0.13363
## New.Customer                    0.332291   0.143022   2.323  0.02016 *
## Purchase.ChannelPhone           0.025889   0.234090   0.111  0.91194
## Purchase.ChannelWeb            -0.080189   0.235396  -0.341  0.73336
## Visited.Website                 2.024617   0.135021  14.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1832.5  on 1746  degrees of freedom
## Residual deviance: 1567.1  on 1730  degrees of freedom
## AIC: 1601.1
##
## Number of Fisher Scoring iterations: 4
```

Eliminate variables with low significance

```
classifier= glm(formula=Sale.Made~Months.Since.Last.Buy+Spend.Category+Spend.Category+Spend.Numeric+Mens
                family=binomial,
                data=df.retail)
summary(classifier)
```

```
##
## Call:
## glm(formula = Sale.Made ~ Months.Since.Last.Buy + Spend.Category +
##     Spend.Category + Spend.Numeric + Mens.Merchandise + Womens.Merchandise +
##     Area + New.Customer + Purchase.Channel + Visited.Website,
##     family = binomial, data = df.retail)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5195  -0.5994  -0.4641  -0.3525   2.2949
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.852134   0.400401  -4.626 3.73e-06 ***
## Months.Since.Last.Buy          -0.042406   0.020100  -2.110  0.03488 *
## Spend.Category2) $100 - $200   -0.561612   0.203689  -2.757  0.00583 **
## Spend.Category3) $200 - $350   -0.971529   0.297778  -3.263  0.00110 **
## Spend.Category4) $350 - $500   -0.832353   0.445339  -1.869  0.06162 .
## Spend.Category5) $500 - $750   -1.568851   0.633624  -2.476  0.01329 *
## Spend.Category6) $750 - $1,000 -1.093183   0.881328  -1.240  0.21483
## Spend.Category7) $1,000 +      -1.434958   1.386927  -1.035  0.30084
## Spend.Numeric                   0.000741   0.001018   0.728  0.46661
## Mens.Merchandise                0.002689   0.227466   0.012  0.99057
## Womens.Merchandise             -0.072224   0.225960  -0.320  0.74925
## AreaSurburban                   0.304667   0.194784   1.564  0.11779
```

```
## AreaUrban                       0.295917   0.197286   1.500  0.13363
## New.Customer                    0.332291   0.143022   2.323  0.02016 *
## Purchase.ChannelPhone           0.025889   0.234090   0.111  0.91194
## Purchase.ChannelWeb            -0.080189   0.235396  -0.341  0.73336
## Visited.Website                 2.024617   0.135021  14.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1832.5  on 1746  degrees of freedom
## Residual deviance: 1567.1  on 1730  degrees of freedom
## AIC: 1601.1
##
## Number of Fisher Scoring iterations: 4
```

Remove mens merchandise and Purchase Channels

```
classifier= glm(formula=Sale.Made~Months.Since.Last.Buy+Spend.Category+Spend.Category+Spend.Numeric+Wome
                family=binomial,
                data=df.retail)
summary(classifier)
```

```
##
## Call:
## glm(formula = Sale.Made ~ Months.Since.Last.Buy + Spend.Category +
##     Spend.Category + Spend.Numeric + Womens.Merchandise + Area +
##     New.Customer + Visited.Website, family = binomial, data = df.retail)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4962  -0.5977  -0.4660  -0.3523   2.2898
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.8755247  0.2481627  -7.558  4.1e-14 ***
## Months.Since.Last.Buy          -0.0429304  0.0200819  -2.138  0.03254 *
## Spend.Category2) $100 - $200   -0.5644309  0.2035844  -2.772  0.00556 **
## Spend.Category3) $200 - $350   -0.9709942  0.2907951  -3.339  0.00084 ***
## Spend.Category4) $350 - $500   -0.8302783  0.4407901  -1.884  0.05962 .
## Spend.Category5) $500 - $750   -1.5609923  0.6268780  -2.490  0.01277 *
## Spend.Category6) $750 - $1,000 -1.0895738  0.8762542  -1.243  0.21370
## Spend.Category7) $1,000 +      -1.4401263  1.3836654  -1.041  0.29797
## Spend.Numeric                   0.0007521  0.0010180   0.739  0.46000
## Womens.Merchandise             -0.0733223  0.1296548  -0.566  0.57172
## AreaSurburban                   0.3069984  0.1945714   1.578  0.11461
## AreaUrban                       0.3010422  0.1969726   1.528  0.12643
## New.Customer                    0.3306822  0.1429065   2.314  0.02067 *
## Visited.Website                 2.0221205  0.1344197  15.043  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1832.5  on 1746  degrees of freedom
```

```
## Residual deviance: 1567.7  on 1733  degrees of freedom
## AIC: 1595.7
##
## Number of Fisher Scoring iterations: 4
```

Take out womens merchandise

```
classifier= glm(formula=Sale.Made~Months.Since.Last.Buy+Spend.Category+Spend.Category+Spend.Numeric+Wome
                family=binomial,
                data=df.retail)
summary(classifier)
```

```
##
## Call:
## glm(formula = Sale.Made ~ Months.Since.Last.Buy + Spend.Category +
##     Spend.Category + Spend.Numeric + Womens.Merchandise + Area +
##     New.Customer + Visited.Website, family = binomial, data = df.retail)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4962  -0.5977  -0.4660  -0.3523   2.2898
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.8755247  0.2481627  -7.558  4.1e-14 ***
## Months.Since.Last.Buy          -0.0429304  0.0200819  -2.138  0.03254 *
## Spend.Category2) $100 - $200   -0.5644309  0.2035844  -2.772  0.00556 **
## Spend.Category3) $200 - $350   -0.9709942  0.2907951  -3.339  0.00084 ***
## Spend.Category4) $350 - $500   -0.8302783  0.4407901  -1.884  0.05962 .
## Spend.Category5) $500 - $750   -1.5609923  0.6268780  -2.490  0.01277 *
## Spend.Category6) $750 - $1,000 -1.0895738  0.8762542  -1.243  0.21370
## Spend.Category7) $1,000 +      -1.4401263  1.3836654  -1.041  0.29797
## Spend.Numeric                   0.0007521  0.0010180   0.739  0.46000
## Womens.Merchandise             -0.0733223  0.1296548  -0.566  0.57172
## AreaSurburban                   0.3069984  0.1945714   1.578  0.11461
## AreaUrban                       0.3010422  0.1969726   1.528  0.12643
## New.Customer                    0.3306822  0.1429065   2.314  0.02067 *
## Visited.Website                 2.0221205  0.1344197  15.043  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1832.5  on 1746  degrees of freedom
## Residual deviance: 1567.7  on 1733  degrees of freedom
## AIC: 1595.7
##
## Number of Fisher Scoring iterations: 4
```

Take out spend

```
classifier= glm(formula=Sale.Made~Months.Since.Last.Buy+Spend.Category+Spend.Category+Area+New.Customer+
                family=binomial,
                data=df.retail)
summary(classifier)
```

```
##
## Call:
## glm(formula = Sale.Made ~ Months.Since.Last.Buy + Spend.Category +
##     Spend.Category + Area + New.Customer + Visited.Website, family = binomial,
##     data = df.retail)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5053  -0.5989  -0.4702  -0.3534   2.2662
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.86389    0.23262  -8.013 1.12e-15 ***
## Months.Since.Last.Buy          -0.04349    0.02005  -2.169  0.03009 *
## Spend.Category2) $100 - $200   -0.49554    0.18003  -2.753  0.00591 **
## Spend.Category3) $200 - $350   -0.81919    0.19020  -4.307 1.65e-05 ***
## Spend.Category4) $350 - $500   -0.56443    0.22986  -2.456  0.01407 *
## Spend.Category5) $500 - $750   -1.15795    0.27225  -4.253 2.11e-05 ***
## Spend.Category6) $750 - $1,000 -0.51532    0.34865  -1.478  0.13940
## Spend.Category7) $1,000 +      -0.49590    0.41920  -1.183  0.23682
## AreaSurburban                   0.30616    0.19438   1.575  0.11525
## AreaUrban                       0.30005    0.19678   1.525  0.12732
## New.Customer                    0.32767    0.14280   2.295  0.02176 *
## Visited.Website                 2.01782    0.13421  15.035  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1832.5  on 1746  degrees of freedom
## Residual deviance: 1568.6  on 1735  degrees of freedom
## AIC: 1592.6
##
## Number of Fisher Scoring iterations: 4
```

Take out Area

```
classifier= glm(formula=Sale.Made~Months.Since.Last.Buy+Spend.Category+Spend.Category+New.Customer+Visi
                family=binomial,
                data=df.retail)
summary(classifier)
```

```
##
## Call:
## glm(formula = Sale.Made ~ Months.Since.Last.Buy + Spend.Category +
##     Spend.Category + New.Customer + Visited.Website, family = binomial,
##     data = df.retail)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4733  -0.5895  -0.4703  -0.3568   2.2773
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.61438    0.17548  -9.200  < 2e-16 ***
```

```
## Months.Since.Last.Buy          -0.04161    0.01997  -2.084  0.03719 *
## Spend.Category2) $100 - $200   -0.48475    0.17988  -2.695  0.00704 **
## Spend.Category3) $200 - $350   -0.81686    0.19012  -4.297 1.74e-05 ***
## Spend.Category4) $350 - $500   -0.54093    0.22845  -2.368  0.01789 *
## Spend.Category5) $500 - $750   -1.13354    0.27086  -4.185 2.85e-05 ***
## Spend.Category6) $750 - $1,000 -0.50188    0.34766  -1.444  0.14886
## Spend.Category7) $1,000 +      -0.48966    0.41701  -1.174  0.24030
## New.Customer                    0.32687    0.14276   2.290  0.02205 *
## Visited.Website                 2.00234    0.13345  15.004  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1832.5  on 1746  degrees of freedom
## Residual deviance: 1571.4  on 1737  degrees of freedom
## AIC: 1591.4
##
## Number of Fisher Scoring iterations: 4
```

Run the predictions against the unseen test data

```
prob_pred=predict(classifier,type='response',newdata=validation[-10])
y_pred=ifelse(prob_pred>0.5,1,0)
```

Create confusion matrix and determine model accuracy

```
confusionMatrix(validation[,10],y_pred)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 252  21
##          1  65  11
##
##                Accuracy : 0.7536
##                  95% CI : (0.7049, 0.7979)
##     No Information Rate : 0.9083
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0857
##  Mcnemar's Test P-Value : 3.538e-06
##
##             Sensitivity : 0.7950
##             Specificity : 0.3438
##          Pos Pred Value : 0.9231
##          Neg Pred Value : 0.1447
##              Prevalence : 0.9083
##          Detection Rate : 0.7221
##    Detection Prevalence : 0.7822
##       Balanced Accuracy : 0.5694
##
##        'Positive' Class : 0
##
```

Model accuracy is 77 percent