MSDS660_Week7_Assignment_APeetz

2022-11-25

Adam Peetz

MSDS660 Week 7 Assignment

Regis University

Dr. Siripun Sanguansintukul

December 4th 2022

Logistic Regression for Sale Response

Data: The data used in this notebook is marketing data provided by the Hult International School of Business.

Objective: Predict a customer's response to a marketing campaign (i.e. 1 if customer accepted the offer in the last campaign, 0 otherwise).

```
#set working directory
setwd("C:\\Users\\adamg\\Documents\\MSDS_660\\Week_7")

#load libraries
library(tidyverse)
library(data.table)
library(car)
library(caTools)
library(readr)
library(caret)
library('fastDummies')
library('ggpubr')
library(MASS)
library(pROC)

# load data
data <- read_csv("marketing.csv",show_col_types = FALSE)
# convert data to table
df <-as.data.table(data)
```

Cleaning Data/Feature Engineering

Income

The income variable starts as a character column. The $ sign needs to be removed from the start of each income so it can be treated as a numerical value. The Income column also contains several outliers that need to be removed. One couple has stated they have an income of 666,666 dollars which is a suspicious number. Rows containing outlying income, such as the 666,666 row, will be removed from the dataset prior to analysis.

Feature Selection

Some features such as customer ID are unique for each row. These unique identifying features will not correlate to other features in the dataset and will be dropped from the model. Features corresponding to purchase counts, amounts, and customer demographics will be kept in the model.

Dummy Variables

There are several categorical variables in the model. Education, marital status, and a customer's country all need to be transformed into numerical values before they can be evaluated by the model. These features will be one-hot encoded into sparse binary matrices using the dummy_col() function in the code below.

```
#remove NA from column
df<- df[-which(is.na(df$Income)), ]

#remove $ signs
df$Income <- parse_number(df$Income)

# subset
df_1 <- df %>% dplyr::select(Education, Income, Kidhome, MntWines, MntFruits, MntMeatProducts, MntFishP

# One hot encoding categorical variables
dum_df_1 <- dummy_cols(df_1,
                       select_columns=c('Education','Marital_Status', 'Country'),
                       remove_selected_columns = TRUE)

# remove outliers
Q <- quantile(dum_df_1$Income, probs=c(.25, .75), na.rm = TRUE)
iqr <- IQR(dum_df_1$Income, na.rm = TRUE)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
dum_df_1<- subset(dum_df_1, dum_df_1$Income > (Q[1] - 1.5*iqr) & dum_df_1$Income < (Q[2]+1.5*iqr))
```

Train Test Split

After cleaning and feature selection, the dataset is broken into training and test sets to provide data sets for the development of the model. A seed is set here to ensure reproducibility of the results.

```
# set seed
set.seed(1)

# create train test split
samp <- sample.split(dum_df_1$Response, SplitRatio = 0.8)
train <- subset(dum_df_1, samp == TRUE)
test <- subset(dum_df_1, samp == FALSE)
```

Model #1, Using All Available Data:

An initial model is created using all data except for -Education_PHD, -Marital_Status_YOLO, and -Country_US. These three features were shown to have multicollinearity issues by the model and have been removed. Creating an initial model allows feature selection by stepwiseAIC in the next step.

```
# generate model
model <- glm(Response ~ . -Education_PhD -Marital_Status_YOLO -Country_US, data = train, family = "bino
```

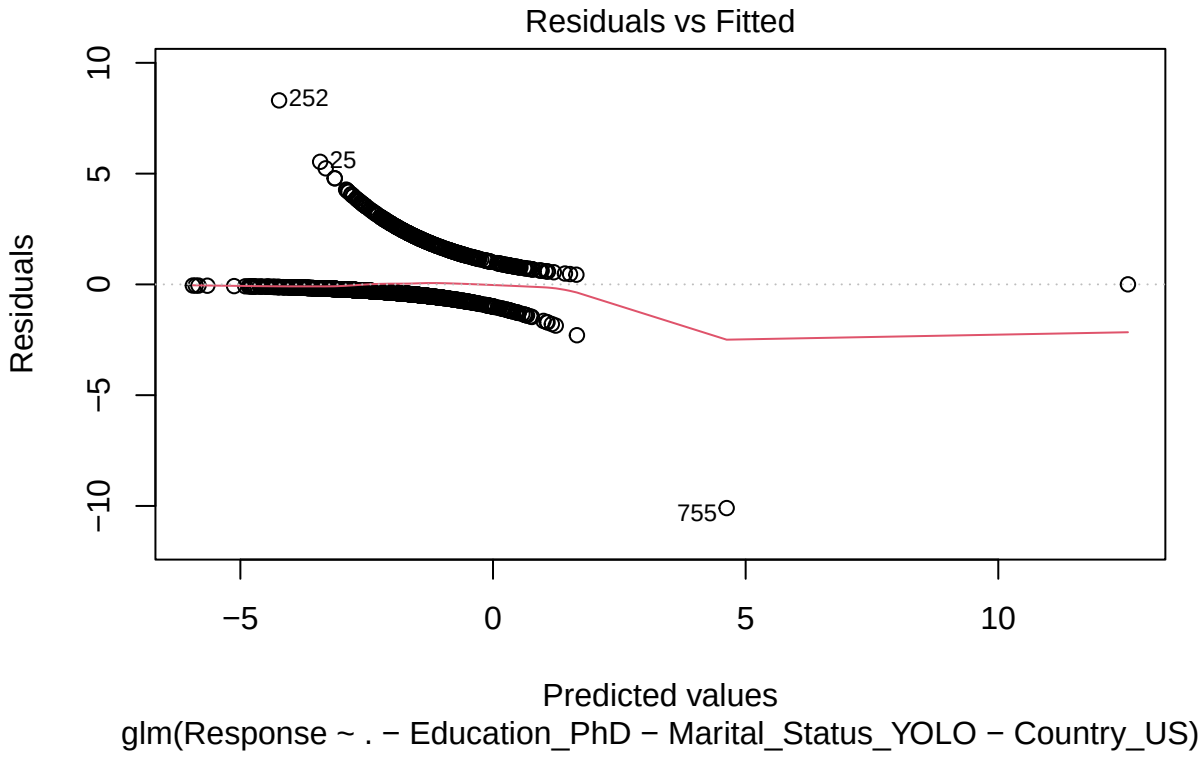
```r
# Summary
summary(model)
```

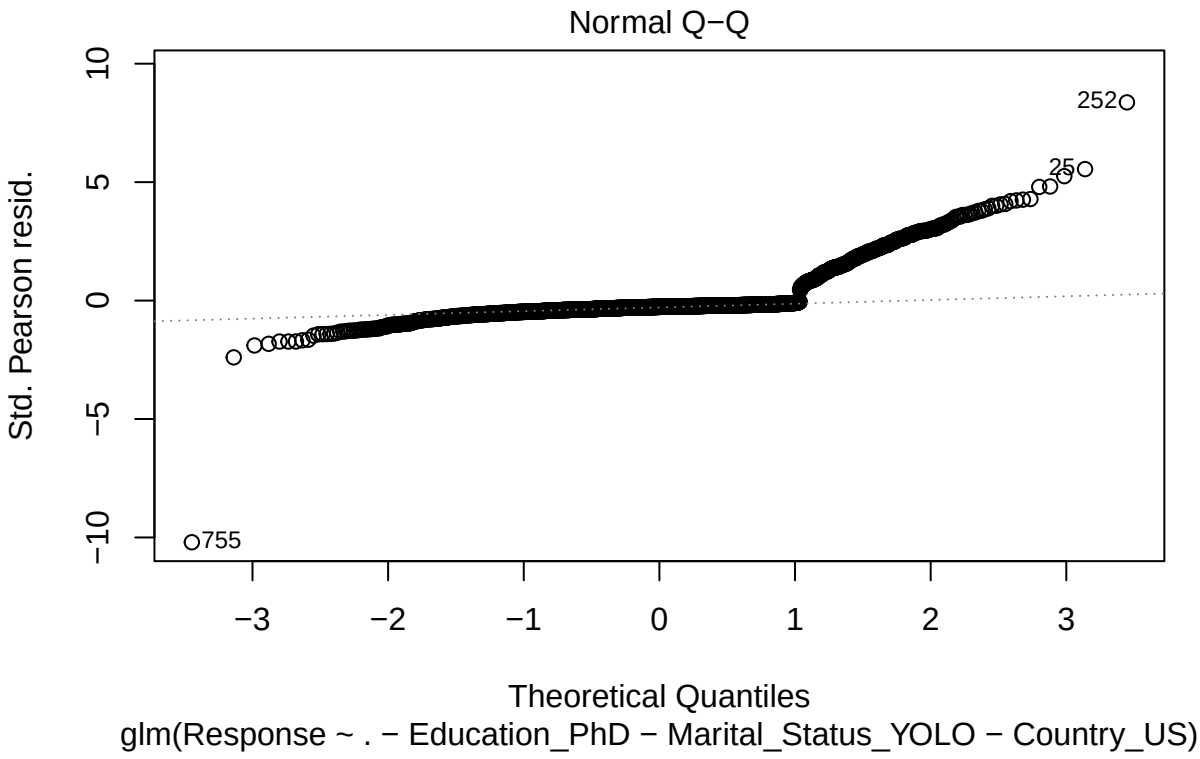
```
##
## Call:
## glm(formula = Response ~ . - Education_PhD - Marital_Status_YOLO -
##     Country_US, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0444  -0.5365  -0.3686  -0.2573   2.9148
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.231e+01  3.247e+02   0.038  0.96975
## Income                   -6.217e-06  7.060e-06  -0.881  0.37853
## Kidhome                   5.109e-01  2.026e-01   2.522  0.01167 *
## MntWines                  1.405e-03  3.109e-04   4.520 6.17e-06 ***
## MntFruits                 8.952e-04  2.284e-03   0.392  0.69511
## MntMeatProducts           2.005e-03  4.841e-04   4.141 3.45e-05 ***
## MntFishProducts          -1.528e-03  1.729e-03  -0.884  0.37686
## MntSweetProducts          7.567e-04  2.161e-03   0.350  0.72621
## MntGoldProds              2.412e-03  1.495e-03   1.613  0.10673
## NumDealsPurchases         4.304e-02  4.266e-02   1.009  0.31310
## NumCatalogPurchases       1.445e-01  3.857e-02   3.745  0.00018 ***
## NumStorePurchases        -1.851e-01  3.220e-02  -5.750 8.93e-09 ***
## NumWebPurchases           9.783e-02  3.080e-02   3.177  0.00149 **
## `Education_2n Cycle`     -4.960e-01  3.180e-01  -1.560  0.11882
## Education_Basic          -1.939e+00  1.045e+00  -1.856  0.06349 .
## Education_Graduation     -6.054e-01  1.911e-01  -3.169  0.00153 **
## Education_Master         -4.288e-01  2.305e-01  -1.861  0.06278 .
## Marital_Status_Absurd    -1.401e+01  3.247e+02  -0.043  0.96558
## Marital_Status_Alone     -1.406e+01  3.247e+02  -0.043  0.96546
## Marital_Status_Divorced -1.487e+01  3.247e+02  -0.046  0.96348
## Marital_Status_Married  -1.573e+01  3.247e+02  -0.048  0.96136
## Marital_Status_Single   -1.475e+01  3.247e+02  -0.045  0.96377
## Marital_Status_Together -1.580e+01  3.247e+02  -0.049  0.96118
## Marital_Status_Widow    -1.487e+01  3.247e+02  -0.046  0.96347
## Country_AUS               1.186e+00  5.304e-01   2.235  0.02540 *
## Country_CA                7.242e-01  5.105e-01   1.419  0.15595
## Country_GER               1.092e+00  5.565e-01   1.962  0.04980 *
## Country_IND              -3.294e-02  6.049e-01  -0.054  0.95657
## Country_ME                2.545e+00  1.348e+00   1.888  0.05901 .
## Country_SA                7.963e-01  4.981e-01   1.599  0.10990
## Country_SP                1.055e+00  4.704e-01   2.242  0.02494 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.8  on 1765  degrees of freedom
## Residual deviance: 1219.7  on 1735  degrees of freedom
## AIC: 1281.7
##
```

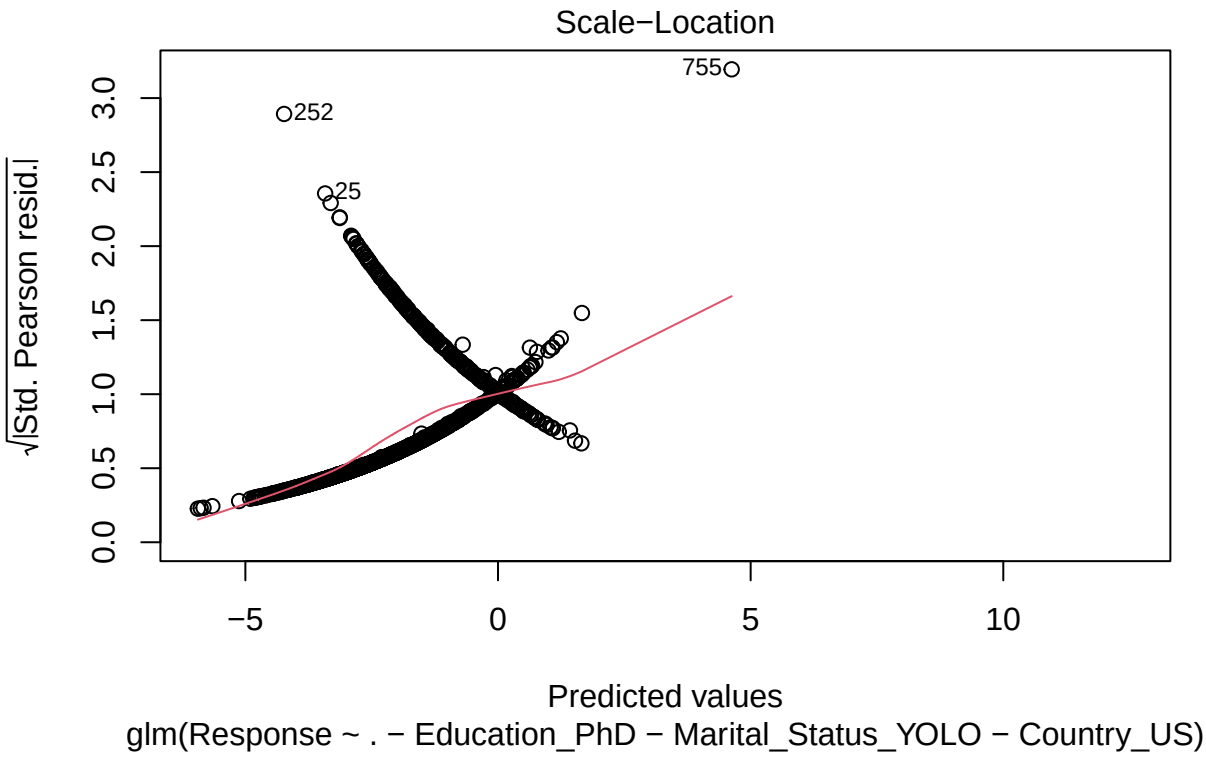
```
## Number of Fisher Scoring iterations: 11
```

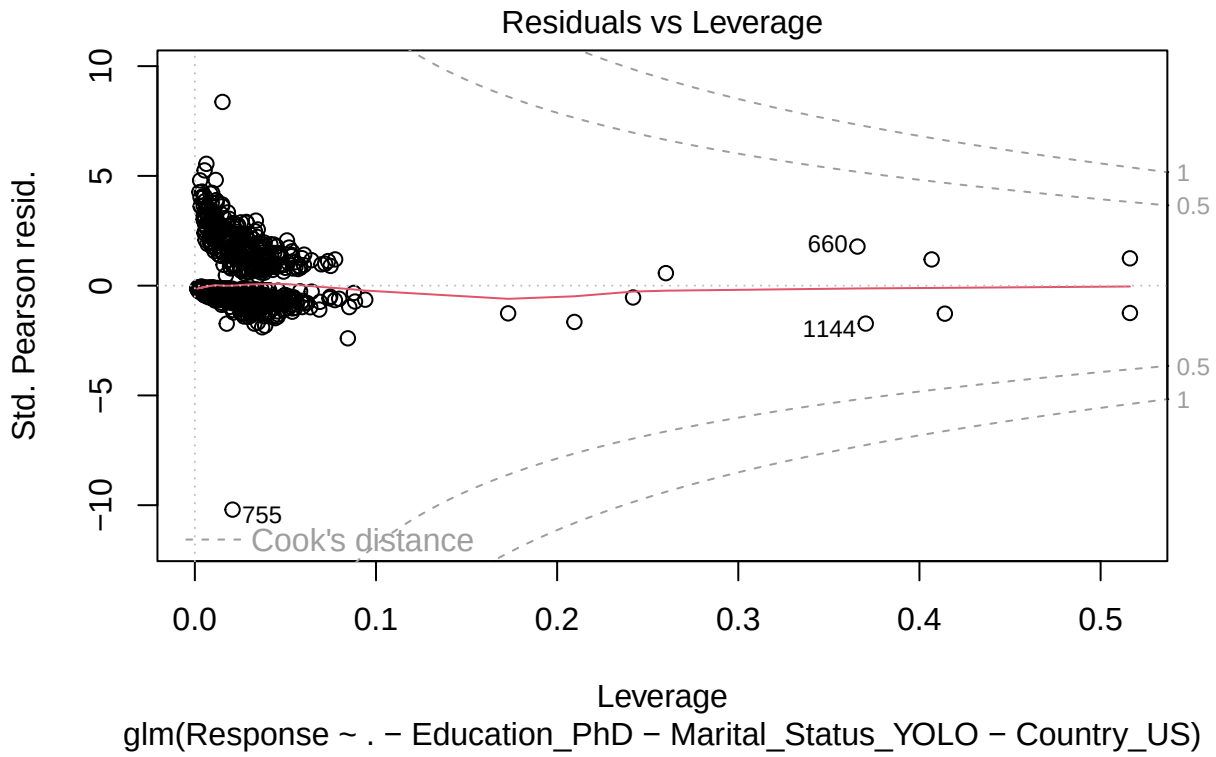
```
# check data
plot(model)
```

```
## Warning: not plotting observations with leverage one:
##   82
```




Normal Q−Q
Std. Pearson resid.
10
5
0
−5
−10
252
25
755
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(Response ~ . − Education_PhD − Marital_Status_YOLO − Country_US)


Scale−Location
755
252
25
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−5
0
5
10
Predicted values
glm(Response ~ . − Education_PhD − Marital_Status_YOLO − Country_US)




```
# Check for collinearity
vif(model)
```

```
##                   Income                  Kidhome                 MntWines
##             4.029931e+00             2.020619e+00             2.633323e+00
##                MntFruits            MntMeatProducts            MntFishProducts
##             1.904359e+00             2.696552e+00             1.927475e+00
##            MntSweetProducts             MntGoldProds         NumDealsPurchases
##             1.859357e+00             1.412955e+00             1.395286e+00
##         NumCatalogPurchases         NumStorePurchases           NumWebPurchases
##             2.643034e+00             1.861915e+00             1.537931e+00
##      `Education_2n Cycle`           Education_Basic       Education_Graduation
##             1.299184e+00             1.046346e+00             1.667782e+00
##           Education_Master      Marital_Status_Absurd       Marital_Status_Alone
##             1.394577e+00             5.151321e+04             6.508509e+04
## Marital_Status_Divorced    Marital_Status_Married     Marital_Status_Single
##             2.295984e+06             4.201726e+06             3.970803e+06
## Marital_Status_Together      Marital_Status_Widow              Country_AUS
##             3.108033e+06             9.031706e+05             3.783697e+00
##                Country_CA               Country_GER              Country_IND
##             4.803379e+00             3.027639e+00             2.293075e+00
##                Country_ME               Country_SA               Country_SP
##             1.142990e+00             5.590021e+00             1.008493e+01
```

Feature Selection by Stepwise AIC

StepwiseAIC is a method for selecting the best combination of features for the model. It does this by sequentially testing different combinations of features and returns the best combination based on its AIC score.

```
# Perform stepwiseAIC
stepAIC(model, direction = 'both')
```

```
## Start:  AIC=1281.69
## Response ~ (Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + NumDealsPurchases +
##     NumCatalogPurchases + NumStorePurchases + NumWebPurchases +
##     `Education_2n Cycle` + Education_Basic + Education_Graduation +
##     Education_Master + Education_PhD + Marital_Status_Absurd +
##     Marital_Status_Alone + Marital_Status_Divorced + Marital_Status_Married +
##     Marital_Status_Single + Marital_Status_Together + Marital_Status_Widow +
##     Marital_Status_YOLO + Country_AUS + Country_CA + Country_GER +
##     Country_IND + Country_ME + Country_SA + Country_SP + Country_US) -
##     Education_PhD - Marital_Status_YOLO - Country_US
##
##                           Df Deviance    AIC
## - Country_IND              1   1219.7 1279.7
## - MntSweetProducts         1   1219.8 1279.8
## - MntFruits                1   1219.8 1279.8
## - Income                   1   1220.5 1280.5
## - MntFishProducts          1   1220.5 1280.5
## - NumDealsPurchases        1   1220.7 1280.7
## <none>                         1219.7 1281.7
## - Country_CA               1   1221.9 1281.9
## - Marital_Status_Absurd    1   1221.9 1281.9
## - Marital_Status_Alone     1   1222.2 1282.2
## - `Education_2n Cycle`     1   1222.2 1282.2
## - MntGoldProds             1   1222.2 1282.2
## - Country_SA               1   1222.6 1282.6
## - Education_Master         1   1223.2 1283.2
## - Country_ME               1   1223.5 1283.5
## - Country_GER              1   1223.8 1283.8
## - Marital_Status_Single    1   1224.1 1284.1
## - Marital_Status_Widow     1   1224.3 1284.3
## - Marital_Status_Divorced  1   1224.3 1284.3
## - Country_AUS              1   1225.3 1285.3
## - Education_Basic          1   1225.6 1285.6
## - Country_SP               1   1225.9 1285.9
## - Marital_Status_Married   1   1225.9 1285.9
## - Marital_Status_Together  1   1226.0 1286.0
## - Kidhome                  1   1226.1 1286.1
## - Education_Graduation     1   1229.6 1289.6
## - NumWebPurchases          1   1230.0 1290.0
## - NumCatalogPurchases      1   1233.8 1293.8
## - MntMeatProducts          1   1237.1 1297.1
## - MntWines                 1   1240.2 1300.2
## - NumStorePurchases        1   1255.7 1315.7
##
## Step:  AIC=1279.69
```

```
## Response ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + NumDealsPurchases +
##     NumCatalogPurchases + NumStorePurchases + NumWebPurchases +
##     `Education_2n Cycle` + Education_Basic + Education_Graduation +
##     Education_Master + Marital_Status_Absurd + Marital_Status_Alone +
##     Marital_Status_Divorced + Marital_Status_Married + Marital_Status_Single +
##     Marital_Status_Together + Marital_Status_Widow + Country_AUS +
##     Country_CA + Country_GER + Country_ME + Country_SA + Country_SP
##
##                           Df Deviance    AIC
## - MntSweetProducts         1   1219.8 1277.8
## - MntFruits                1   1219.8 1277.8
## - Income                   1   1220.5 1278.5
## - MntFishProducts          1   1220.5 1278.5
## - NumDealsPurchases        1   1220.7 1278.7
## <none>                         1219.7 1279.7
## - Marital_Status_Absurd    1   1221.9 1279.9
## - Marital_Status_Alone     1   1222.2 1280.2
## - `Education_2n Cycle`     1   1222.2 1280.2
## - MntGoldProds             1   1222.2 1280.2
## - Education_Master         1   1223.2 1281.2
## + Country_IND              1   1219.7 1281.7
## - Country_ME               1   1223.7 1281.7
## - Country_CA               1   1223.8 1281.8
## - Marital_Status_Single    1   1224.1 1282.1
## - Marital_Status_Widow     1   1224.3 1282.3
## - Marital_Status_Divorced  1   1224.3 1282.3
## - Country_SA               1   1225.3 1283.3
## - Education_Basic          1   1225.6 1283.6
## - Marital_Status_Married   1   1225.9 1283.9
## - Kidhome                  1   1226.1 1284.1
## - Marital_Status_Together  1   1226.1 1284.1
## - Country_GER              1   1226.2 1284.2
## - Country_AUS              1   1228.9 1286.9
## - Education_Graduation     1   1229.6 1287.6
## - NumWebPurchases          1   1230.0 1288.0
## - Country_SP               1   1233.6 1291.6
## - NumCatalogPurchases      1   1233.8 1291.8
## - MntMeatProducts          1   1237.2 1295.2
## - MntWines                 1   1240.2 1298.2
## - NumStorePurchases        1   1255.7 1313.7
##
## Step:  AIC=1277.81
## Response ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntGoldProds + NumDealsPurchases + NumCatalogPurchases +
##     NumStorePurchases + NumWebPurchases + `Education_2n Cycle` +
##     Education_Basic + Education_Graduation + Education_Master +
##     Marital_Status_Absurd + Marital_Status_Alone + Marital_Status_Divorced +
##     Marital_Status_Married + Marital_Status_Single + Marital_Status_Together +
##     Marital_Status_Widow + Country_AUS + Country_CA + Country_GER +
##     Country_ME + Country_SA + Country_SP
##
##                           Df Deviance    AIC
## - MntFruits                1   1220.0 1276.0
```

```
## - Income                   1   1220.5 1276.5
## - MntFishProducts          1   1220.5 1276.5
## - NumDealsPurchases        1   1220.8 1276.8
## <none>                         1219.8 1277.8
## - Marital_Status_Absurd    1   1222.1 1278.1
## - `Education_2n Cycle`     1   1222.3 1278.3
## - Marital_Status_Alone     1   1222.3 1278.3
## - MntGoldProds             1   1222.3 1278.3
## - Education_Master         1   1223.3 1279.3
## + MntSweetProducts         1   1219.7 1279.7
## + Country_IND              1   1219.8 1279.8
## - Country_ME               1   1223.9 1279.9
## - Country_CA               1   1224.0 1280.0
## - Marital_Status_Single    1   1224.2 1280.2
## - Marital_Status_Widow     1   1224.4 1280.4
## - Marital_Status_Divorced  1   1224.4 1280.4
## - Country_SA               1   1225.4 1281.4
## - Education_Basic          1   1225.7 1281.7
## - Marital_Status_Married   1   1226.1 1282.1
## - Kidhome                  1   1226.1 1282.1
## - Marital_Status_Together  1   1226.2 1282.2
## - Country_GER              1   1226.2 1282.2
## - Country_AUS              1   1229.2 1285.2
## - Education_Graduation     1   1229.6 1285.6
## - NumWebPurchases          1   1230.6 1286.6
## - Country_SP               1   1233.8 1289.8
## - NumCatalogPurchases      1   1234.0 1290.0
## - MntMeatProducts          1   1237.7 1293.7
## - MntWines                 1   1240.3 1296.3
## - NumStorePurchases        1   1255.8 1311.8
##
## Step:  AIC=1276.04
## Response ~ Income + Kidhome + MntWines + MntMeatProducts + MntFishProducts +
##     MntGoldProds + NumDealsPurchases + NumCatalogPurchases +
##     NumStorePurchases + NumWebPurchases + `Education_2n Cycle` +
##     Education_Basic + Education_Graduation + Education_Master +
##     Marital_Status_Absurd + Marital_Status_Alone + Marital_Status_Divorced +
##     Marital_Status_Married + Marital_Status_Single + Marital_Status_Together +
##     Marital_Status_Widow + Country_AUS + Country_CA + Country_GER +
##     Country_ME + Country_SA + Country_SP
##
##                           Df Deviance    AIC
## - MntFishProducts          1   1220.6 1274.6
## - Income                   1   1220.7 1274.7
## - NumDealsPurchases        1   1221.0 1275.0
## <none>                         1220.0 1276.0
## - Marital_Status_Absurd    1   1222.3 1276.3
## - `Education_2n Cycle`     1   1222.4 1276.4
## - Marital_Status_Alone     1   1222.5 1276.5
## - MntGoldProds             1   1222.8 1276.8
## - Education_Master         1   1223.5 1277.5
## + MntFruits                1   1219.8 1277.8
## + MntSweetProducts         1   1219.8 1277.8
## - Country_ME               1   1224.0 1278.0
```

```
## + Country_IND              1   1220.0 1278.0
## - Country_CA               1   1224.2 1278.2
## - Marital_Status_Single    1   1224.5 1278.5
## - Marital_Status_Widow     1   1224.6 1278.6
## - Marital_Status_Divorced  1   1224.7 1278.7
## - Country_SA               1   1225.6 1279.6
## - Education_Basic          1   1225.8 1279.8
## - Marital_Status_Married   1   1226.3 1280.3
## - Kidhome                  1   1226.3 1280.3
## - Country_GER              1   1226.4 1280.4
## - Marital_Status_Together  1   1226.4 1280.4
## - Country_AUS              1   1229.3 1283.3
## - Education_Graduation     1   1229.6 1283.6
## - NumWebPurchases          1   1230.8 1284.8
## - Country_SP               1   1233.9 1287.9
## - NumCatalogPurchases      1   1234.5 1288.5
## - MntMeatProducts          1   1239.7 1293.7
## - MntWines                 1   1240.3 1294.3
## - NumStorePurchases        1   1255.8 1309.8
##
## Step:  AIC=1274.6
## Response ~ Income + Kidhome + MntWines + MntMeatProducts + MntGoldProds +
##     NumDealsPurchases + NumCatalogPurchases + NumStorePurchases +
##     NumWebPurchases + `Education_2n Cycle` + Education_Basic +
##     Education_Graduation + Education_Master + Marital_Status_Absurd +
##     Marital_Status_Alone + Marital_Status_Divorced + Marital_Status_Married +
##     Marital_Status_Single + Marital_Status_Together + Marital_Status_Widow +
##     Country_AUS + Country_CA + Country_GER + Country_ME + Country_SA +
##     Country_SP
##
##                            Df Deviance    AIC
## - Income                    1   1221.4 1273.4
## - NumDealsPurchases         1   1221.6 1273.6
## <none>                          1220.6 1274.6
## - Marital_Status_Absurd     1   1223.0 1275.0
## - MntGoldProds              1   1223.0 1275.0
## - Marital_Status_Alone      1   1223.1 1275.1
## - `Education_2n Cycle`      1   1223.5 1275.5
## + MntFishProducts           1   1220.0 1276.0
## - Education_Master          1   1224.3 1276.3
## - Country_ME                1   1224.5 1276.5
## + MntFruits                 1   1220.5 1276.5
## + MntSweetProducts          1   1220.5 1276.5
## + Country_IND               1   1220.6 1276.6
## - Country_CA                1   1224.8 1276.8
## - Marital_Status_Single     1   1225.0 1277.0
## - Marital_Status_Widow      1   1225.2 1277.2
## - Marital_Status_Divorced   1   1225.2 1277.2
## - Country_SA                1   1226.0 1278.0
## - Education_Basic           1   1226.7 1278.7
## - Marital_Status_Married    1   1226.8 1278.8
## - Country_GER               1   1226.9 1278.9
## - Marital_Status_Together   1   1227.0 1279.0
## - Kidhome                   1   1227.0 1279.0
```

```
## - Country_AUS              1   1229.8 1281.8
## - Education_Graduation     1   1231.0 1283.0
## - NumWebPurchases          1   1231.3 1283.3
## - Country_SP               1   1234.5 1286.5
## - NumCatalogPurchases      1   1234.5 1286.5
## - MntMeatProducts          1   1239.7 1291.7
## - MntWines                 1   1241.7 1293.7
## - NumStorePurchases        1   1256.9 1308.9
##
## Step:  AIC=1273.43
## Response ~ Kidhome + MntWines + MntMeatProducts + MntGoldProds +
##     NumDealsPurchases + NumCatalogPurchases + NumStorePurchases +
##     NumWebPurchases + `Education_2n Cycle` + Education_Basic +
##     Education_Graduation + Education_Master + Marital_Status_Absurd +
##     Marital_Status_Alone + Marital_Status_Divorced + Marital_Status_Married +
##     Marital_Status_Single + Marital_Status_Together + Marital_Status_Widow +
##     Country_AUS + Country_CA + Country_GER + Country_ME + Country_SA +
##     Country_SP
##
##                           Df Deviance    AIC
## - NumDealsPurchases        1   1222.9 1272.9
## <none>                         1221.4 1273.4
## - MntGoldProds             1   1223.8 1273.8
## - Marital_Status_Absurd    1   1223.9 1273.9
## - Marital_Status_Alone     1   1224.0 1274.0
## - `Education_2n Cycle`     1   1224.3 1274.3
## + Income                   1   1220.6 1274.6
## + MntFishProducts          1   1220.7 1274.7
## - Education_Master         1   1225.1 1275.1
## - Country_ME               1   1225.4 1275.4
## + MntFruits                1   1221.4 1275.4
## + Country_IND              1   1221.4 1275.4
## + MntSweetProducts         1   1221.4 1275.4
## - Country_CA               1   1225.7 1275.7
## - Marital_Status_Single    1   1225.8 1275.8
## - Marital_Status_Widow     1   1226.0 1276.0
## - Marital_Status_Divorced  1   1226.0 1276.0
## - Country_SA               1   1226.8 1276.8
## - Education_Basic          1   1226.9 1276.9
## - Marital_Status_Married   1   1227.6 1277.6
## - Marital_Status_Together  1   1227.8 1277.8
## - Country_GER              1   1227.8 1277.8
## - Kidhome                  1   1228.0 1278.0
## - Country_AUS              1   1230.5 1280.5
## - NumWebPurchases          1   1231.5 1281.5
## - Education_Graduation     1   1232.0 1282.0
## - NumCatalogPurchases      1   1234.6 1284.6
## - Country_SP               1   1235.4 1285.4
## - MntMeatProducts          1   1240.1 1290.1
## - MntWines                 1   1242.0 1292.0
## - NumStorePurchases        1   1263.6 1313.6
##
## Step:  AIC=1272.93
## Response ~ Kidhome + MntWines + MntMeatProducts + MntGoldProds +
```

```
##     NumCatalogPurchases + NumStorePurchases + NumWebPurchases +
##     `Education_2n Cycle` + Education_Basic + Education_Graduation +
##     Education_Master + Marital_Status_Absurd + Marital_Status_Alone +
##     Marital_Status_Divorced + Marital_Status_Married + Marital_Status_Single +
##     Marital_Status_Together + Marital_Status_Widow + Country_AUS +
##     Country_CA + Country_GER + Country_ME + Country_SA + Country_SP
##
##                            Df Deviance    AIC
## <none>                          1222.9 1272.9
## + NumDealsPurchases         1   1221.4 1273.4
## - Marital_Status_Absurd     1   1225.5 1273.5
## - MntGoldProds              1   1225.6 1273.6
## - Marital_Status_Alone      1   1225.6 1273.6
## + Income                    1   1221.6 1273.6
## - `Education_2n Cycle`      1   1225.9 1273.9
## + MntFishProducts           1   1221.9 1273.9
## - Education_Master          1   1226.4 1274.4
## - Country_ME                1   1226.9 1274.9
## + MntSweetProducts          1   1222.9 1274.9
## + Country_IND               1   1222.9 1274.9
## + MntFruits                 1   1222.9 1274.9
## - Country_CA                1   1227.1 1275.1
## - Marital_Status_Single     1   1227.6 1275.6
## - Marital_Status_Widow      1   1227.7 1275.7
## - Marital_Status_Divorced   1   1227.8 1275.8
## - Country_SA                1   1228.3 1276.3
## - Education_Basic           1   1228.5 1276.5
## - Country_GER               1   1229.2 1277.2
## - Marital_Status_Married    1   1229.4 1277.4
## - Marital_Status_Together   1   1229.5 1277.5
## - Country_AUS               1   1231.8 1279.8
## - Kidhome                   1   1232.5 1280.5
## - Education_Graduation      1   1233.5 1281.5
## - NumWebPurchases           1   1235.7 1283.7
## - NumCatalogPurchases       1   1236.5 1284.5
## - Country_SP                1   1236.7 1284.7
## - MntMeatProducts           1   1240.2 1288.2
## - MntWines                  1   1243.5 1291.5
## - NumStorePurchases         1   1263.7 1311.7
##
## Call:  glm(formula = Response ~ Kidhome + MntWines + MntMeatProducts +
##     MntGoldProds + NumCatalogPurchases + NumStorePurchases +
##     NumWebPurchases + `Education_2n Cycle` + Education_Basic +
##     Education_Graduation + Education_Master + Marital_Status_Absurd +
##     Marital_Status_Alone + Marital_Status_Divorced + Marital_Status_Married +
##     Marital_Status_Single + Marital_Status_Together + Marital_Status_Widow +
##     Country_AUS + Country_CA + Country_GER + Country_ME + Country_SA +
##     Country_SP, family = "binomial", data = train)
##
## Coefficients:
##             (Intercept)                 Kidhome                MntWines
##               12.210072                0.593935                0.001310
##         MntMeatProducts             MntGoldProds      NumCatalogPurchases
```

```
##               0.001735                0.002412                0.134827
##       NumStorePurchases          NumWebPurchases     `Education_2n Cycle`
##               -0.186583                0.104387                -0.520222
##         Education_Basic     Education_Graduation         Education_Master
##               -1.851078               -0.611085                -0.426193
##   Marital_Status_Absurd     Marital_Status_Alone  Marital_Status_Divorced
##              -14.313438              -14.179961               -14.953588
##  Marital_Status_Married    Marital_Status_Single  Marital_Status_Together
##              -15.823616              -14.840882               -15.889755
##    Marital_Status_Widow             Country_AUS              Country_CA
##              -14.954534                1.176717                0.739840
##             Country_GER              Country_ME              Country_SA
##                1.088658                2.580748                0.795147
##              Country_SP
##                1.061453
##
## Degrees of Freedom: 1765 Total (i.e. Null);  1741 Residual
## Null Deviance:       1497
## Residual Deviance: 1223  AIC: 1273
```

Model #2: Features Selected by Stepwise AIC

StepwiseAIC recommends a model that contains the number of kids in each home, the amounts spent on wines and meat, the number of purchases a customer has made, and a select combination of demographic variables for a customer's education, marital status, and country. A list of all included features is shown in the cell below.

```
# create glm model
model2 <-  glm(formula = Response ~ Kidhome + MntWines + MntMeatProducts +
    NumCatalogPurchases + NumStorePurchases + NumWebPurchases +
    Education_Basic + Education_Graduation + Marital_Status_Divorced +
    Marital_Status_Married + Marital_Status_Single + Marital_Status_Together +
    Marital_Status_Widow + Country_AUS + Country_GER + Country_SA +
    Country_SP, family = "binomial", data = train)
```

Confusion Matrix for Predictions on Training

A confusion matrix can be generated to show how the model predicted against the ground truth labels of the dataset. Shown below for the training set, the model was able to achieve 85.56% accuracy. This high accuracy is misleading and is a result of predictions for class 0 in a dataset that contains many examples of class 0. The model had a much harder time predicting for the minority class. This prediction accuracy does not significantly exceed the null information rate of 84.84%.

```
# generate predictions
trainpreds <- predict(model2, type = 'response', train)

# Round prediction values at 0.5 cutoff threshold
trainp <- factor(trainpreds >= 0.5,labels = c('0', '1'))

# plot confusion matrix
trainCM <- confusionMatrix(trainp,as.factor(train$Response))
trainCM
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction    0    1
##          0 1460  215
##          1   40   51
##
##                Accuracy : 0.8556
##                  95% CI : (0.8383, 0.8717)
##     No Information Rate : 0.8494
##     P-Value [Acc > NIR] : 0.2437
##
##                   Kappa : 0.2263
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9733
##             Specificity : 0.1917
##          Pos Pred Value : 0.8716
##          Neg Pred Value : 0.5604
##              Prevalence : 0.8494
##          Detection Rate : 0.8267
##    Detection Prevalence : 0.9485
##       Balanced Accuracy : 0.5825
##
##        'Positive' Class : 0
##
```

Confusion Matrix for Predictions on Test

A confusion matrix for predictions against the test dataset reveals the same pattern shown for the training data. 85.07% accuracy, which is a result of predictions for the majority class in a test set that contains many examples of that class.

```
# generate predictions
testpreds <- predict(model2, type = 'response', test)

# round predictions around 0.5 threshold
testp <- factor(testpreds >= 0.5, labels = c('0', '1'))

# generate confusion matrix
testCM <- confusionMatrix(testp, as.factor(test$Response))
testCM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 366  57
##          1   9  10
##
##                Accuracy : 0.8507
##                  95% CI : (0.814, 0.8826)
##     No Information Rate : 0.8484
##     P-Value [Acc > NIR] : 0.4797
```

```
##
##                   Kappa : 0.1775
##
##  Mcnemar's Test P-Value : 7.238e-09
##
##             Sensitivity : 0.9760
##             Specificity : 0.1493
##          Pos Pred Value : 0.8652
##          Neg Pred Value : 0.5263
##              Prevalence : 0.8484
##          Detection Rate : 0.8281
##    Detection Prevalence : 0.9570
##       Balanced Accuracy : 0.5626
##
##        'Positive' Class : 0
##
```

ROC Curve and Threshold

If this model had been used to extend offers to customers, it would have only sent 19 offers, and missed 57 customers who are likely to say yes to the deal. This is not ideal. Fortunately, the model's tendency to predict positive outcomes can be adjusted by fine-tuning the prediction cutoff threshold used in the model.

The first step in threshold adjustment is to plot the model's ROC curve. An ROC curve displays the tradeoff between true positive and false positive predictions at different threshold settings. The point on the curve closest to the top left corner is considered the ideal threshold setting and can be automatically calculated as demonstrated in the code below.

```
# Create a Roc curve and results for the Test data
test_roc_curve <- roc(test$Response, testpreds)
```

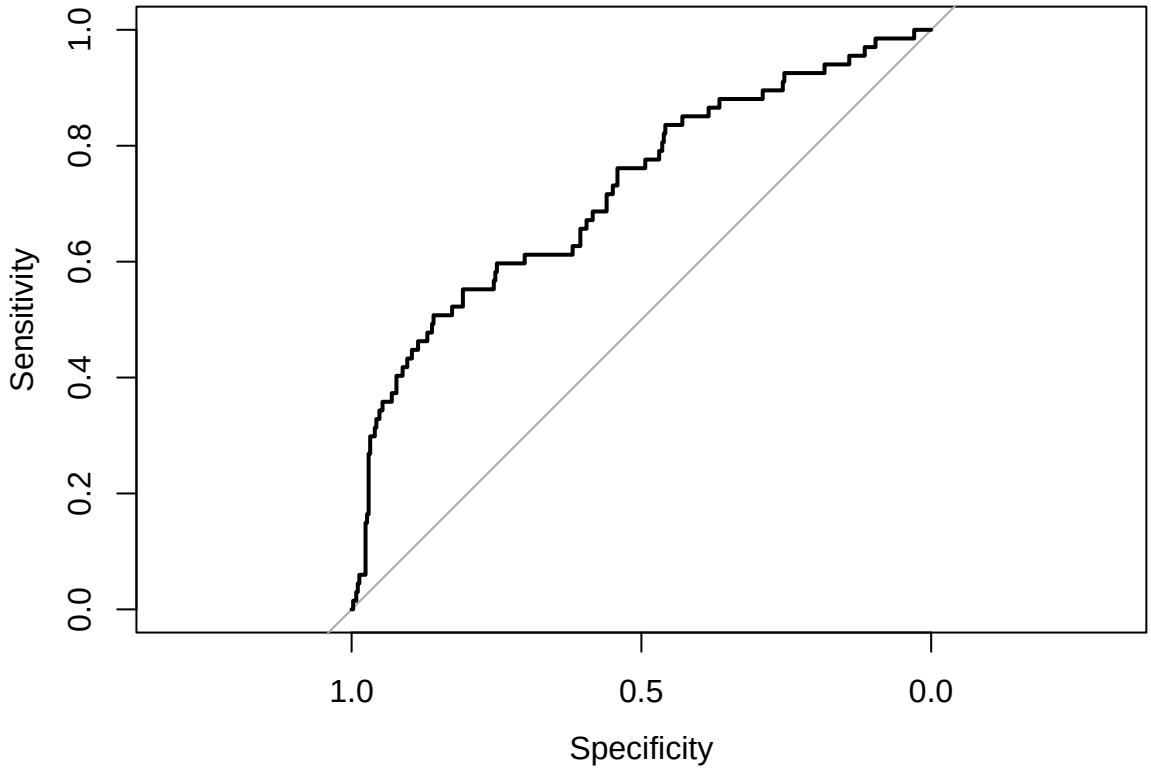
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
test_roc_curve
```

```
##
## Call:
## roc.default(response = test$Response, predictor = testpreds)
##
## Data: testpreds in 375 controls (test$Response 0) < 67 cases (test$Response 1).
## Area under the curve: 0.7217
```

```
plot(test_roc_curve)
```




```
# set the threshold
thresh <- coords(roc=test_roc_curve, x = 'best', best.method = 'closest.topleft', transpose=TRUE)

# display threshold
thresh
```

```
##   threshold specificity sensitivity
##   0.1574166   0.7493333   0.5970149
```

Modifying predictions with a fine-tuned Threshold

The ROC curve recommends a threshold setting of about 0.15. Adjusting the model's predictions around this threshold lowers the model's overall accuracy from 85.07% to 72.62% but boosts the number of true positive predictions. While the overall accuracy has decreased, the model now recommends sales to 40 out of 67 receptive customers, whereas the untuned model only recommended 10.

```
# generate predictions and ground truth labels
rounded_preds <- as.factor(as.integer(testpreds > thresh[1]))
targets <- as.factor(as.integer(test$Response))

# orient data for confusion matrix
postResample(pred = rounded_preds, obs = targets)
```

```
##  Accuracy     Kappa
## 0.7262443 0.2455210
```

```r
# generate confusion martrix
confusionMatrix(rounded_preds, targets)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 281  27
##          1  94  40
##
##                Accuracy : 0.7262
##                  95% CI : (0.6821, 0.7673)
##     No Information Rate : 0.8484
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2455
##
##  Mcnemar's Test P-Value : 1.973e-09
##
##             Sensitivity : 0.7493
##             Specificity : 0.5970
##          Pos Pred Value : 0.9123
##          Neg Pred Value : 0.2985
##              Prevalence : 0.8484
##          Detection Rate : 0.6357
##    Detection Prevalence : 0.6968
##       Balanced Accuracy : 0.6732
##
##        'Positive' Class : 0
##
```

Conclusion

After cleaning and preparing the data for analysis, a logistic regression model was able to correctly predict the response of a customer around 85% of the time. This outcome did not significantly improve over the null information rate of 84%.

This accuracy is a result of a class imbalance in this data set. Future research could attempt to resolve this imbalance by under sampling the majority class or creating synthetic data points for the minority to attempt to balance the distribution of the dataset. This may improve the model's ability to predict for the minority class.

Fine tuning the threshold of the model made it more likely to predict positive outcomes. This results in more customers being sent offers who are likely to say yes to the deal. Threshhold adjustment improved true positive sale offers from 10/67 to 40/67.

The tradeoff between false and true positive predictions made by threshold adjustment needs to be considered in the context of the model. In a medical context, false positives could result in unnecessary medical procedures. In marketing, offering sales to people who do not want them is probably harmless and may even result in additional sales from unexpected customers who are receptive to the offer. The company should proceed with deployment of the threshold adjusted model to maximize its sales.

References

Hult International Business School. (n.d.). marketing data . dataset. retrieved 10/22/22 from https://worldclass.regis.edu/d2l/le/content/297311/Home

MSDS660. (2022). Statistical Methods and Experimental Design. Taught by Dr. Siripun Sanguansintukul.