Case Study Report

Vishwa Koppisetti

January 18, 2019

I. Executive Summary

This case study focuses on targeting through telemarketing phone calls to sell long-term deposits. Within a campaign, the human agents execute phone calls to a list of clients to sell the deposit either through inbound or outbound calls Thus, the goal is to predict whether a client will subscribe a term deposit or not.

The data is related with direct marketing campaigns of a Portuguese banking institution. The dataset set contains 4119 observations of 20 input variables. The output variable is binary: 'yes' 'no'. indicating whether a client has subscribed or not.

The dataset is searched for any missing values of categorical, numeric fields and any outliers as a part of data cleaning

Boxplots and summary statistics are analysed for predictor variables to study their distribution and check for any outliers.

New variable pdaydummy is created for pdays so that '999' value assumed as numeric value doesn't bias the model. Appendix C

The correlation matrix is examined for any multicollinearity between the continuous predictors. Appendix D

Boxplots between output and various continuous predictors are analysed to see individual effect on the output variable.

For evaluation purposes, the data is split into 70% train and 30% test set. The training data is used for feature and model selection. The test data is used for measuring the prediction capabilities of the selected data-driven model.

The predictor duration is removed while fitting the logistic model since it is given that this attribute highly affects the output target and be discarded if the intention is to have a realistic predictive model.

The model is initially fit without duration and pdays and with pdaydummy on the train set, Appendix E. StepAIC is performed for stepwise model selection. The resulting model is used for prediction on test data and accuracy is calculated with help of confusion matrix.

The predictor that is highly correlated is removed from the model to check any improvement in model accuracy.

II. The Problem

A. Introduction/Background

Marketing selling campaigns constitute a typical strategy to enhance business. Companies use direct marketing when targeting segments of customers by contacting them to meet a specific goal. Centralizing customer remote interactions in a contact center eases operational management of campaigns. Such centers allow communicating with customers through various channels, telephone (fixed-line or mobile) being one of the most widely used. Marketing operationalized through a contact center is called telemarketing due to the remoteness characteristic. Contacts can be divided into inbound and outbound, depending on which side triggered the contact (client or contact center), with each case posing different challenges (e.g., outbound calls are often considered more intrusive). Technology enables rethinking marketing by focusing on maximizing customer lifetime value through the evaluation of available information and customer metrics, thus allowing us to build longer and tighter relations in alignment with business demand. Also, it should be stressed that the task of selecting the best set of clients, i.e., that are more likely to subscribe a product, is considered NP-hard

B. Purpose of study/importance of study/statement of problem

Assessment of a real problem of bank telemarketing to sell long-term deposits
The purpose is to predict whether a client will subscribe a term deposit or not, given some client information, telemarketing attributes, certain social and economic features

C. Questions to be answered/conceptual statement of hypotheses

Are the predictors statistically significant or not
Is the model significant and resulting in a useful prediction
Are there any missing values or outliers
Is there multicollinearity
Are the linear assumptions met

D. Outline of remainder of report (brief)

Procedure followed to predict the response including data cleaning and preprocessing and other recommendations

III. Review of Related Literature

A. Existing methodologies used in this area.

In this area, we can use four binary classification data mining models: logistic regression (LR), decision trees (DTs), neural network (NN) and support vector machine (SVM).

The LR is a popular choice that operates a smooth nonlinear logistic transformation over a multiple regression model and allows the estimation of class probabilities. Due to the additive linear combination of its independent variables the model is easy to interpret. Yet, the model is quite rigid and cannot model adequately complex nonlinear relationships.

The DT is a branching structure that represents a set of rules, distinguishing values in a hierarchical form.
The SVM classifier transforms the predictor space into a high m-dimensional feature space. Then, the SVM finds the best linear separating hyperplane, related to a set of support vector points, in the feature space.

IV. Methodology

A. Identification, classification and operationalization of variables.
The data set has 4119 observations of 21 variables
There are 20 independent variables and 1 binary dependent variable
The Classification and operationalization of variables in Appendix A

B. Statements of hypotheses being tested and/or models being developed.
Null hypothesis: None of the predictors have significant effect on the response. The coefficients of the predictors are 0.
Is the model significant and resulting in a accurate prediction

C. Sampling techniques, if full data is not being used.
The data is split into 70% train and 30% test set, Appendix B

D. Data collection process, including data sources, data size, etc. Primary/secondary?

This study considers real data collected from a Portuguese retail bank, containing 4119 records. The data source is secondary data that has the Bank's client details and some information from telemarketing phone calls by human agents to these clients

Each record included the output target and candidate input features. These include telemarketing attributes and client information (e.g., age). These records were enriched with social and economic influence features (e.g., unemployment variation rate), by gathering external data from the central bank of the Portuguese Republic statistical web site.

E. Modeling analysis/techniques used
Logistic Regression is used for binary response and continuous/categorical predictors

F. Methodological assumptions and limitations.

The Logistic Regression is a popular choice that operates a smooth nonlinear logistic transformation over a multiple regression model and allows the estimation of class probabilities. Due to the additive linear combination of its independent variables (x), the model is easy to interpret. Yet, the model is quite rigid and cannot model adequately complex nonlinear relationships.

V. Data

A. Data cleaning

There are no missing values, but age predictor variable has few outliers to be removed. Appendix C

B. Data preprocessing

The predictor, emp.var.rate that is highly correlated with other predictors is removed from the model to check any improvement in model accuracy. Appendix F. The accuracy has improved by a very small percentage, 0.25%. the increment is due to couple more of 'yes' responses being classified correctly with removal of emp.var.rate predictor

The other highly correlated predictors were not significantly improving the accuracy

Data limitations

There were no missing values but there certainly are unknown fields for variables such as job, marital, education, default, housing and loan Which indicates that particular attribute data couldn't be collected for a client. This limits model performance

VI. Findings (Results)

A. Results presented in tables or charts when appropriate

Qualitative Analysis of relationship between response and continuous predictors through boxplots shows that all have an effect on the response variable. The frequencies from summary statistics were observed for categorical predictors, Appendix D.

Predictor age from its Boxplot distribution, Appendix C, is shown to have outliers 1.5IQR above the 3rd quartile which were removed
Other predictors like pdays have misleading outliers as '999' were majority and other values were incorrectly detected as outliers but that was resolved by converting it into dummy variable.
The outliers detected for campaign and previous were also important information to be retained hence not removed

Checking multicollinearity between multiple continuous predictors Appendix D

emp.var.rate is highly positively correlated with euribor3m , nr.employed and cons.price.idx as well.
euribor3m is highly positively correlated with nr.employed
cons.price.idx is somewhat correlated with euribor3m

B. Results reported with respect to hypotheses/models.

We use logistic regression since the response is binary : 'yes' or 'no'

The model is initially fit without duration and pdays and with pdaydummy on the train set, Appendix E. StepAIC is performed for stepwise model selection.

Few predictors were significant such as contact via telephone, month of aug, dec, march, emp.var.rate, cons.price.idx, euribor3m had a significant impact on the response variable since the p-value below 0.05 . Except contacttelephone and emp.var.rate rest all have a positive relation with the response. One of the interpreatation is The odds of a client subscribing to deposit is more when calls are made during month of aug dec and march.

The resulting model is used for prediction on test data and accuracy is calculated with help of confusion matrix to be 91.5%.

Furthermore, the predictor, emp.var.rate that is highly correlated with other predictors is removed from the model to check any improvement in model accuracy. Appendix F.

The accuracy has improved by a very small percentage, 0.25%. the increment is due to couple more of 'yes' responses being classified correctly with removal of emp.var.rate predictor

The other highly correlated predictors were not significantly improving the accuracy

C. Factual information kept separate from interpretation, inference and evaluation.

Focusing on the case studied of bank telemarketing, it is difficult to financially quantify costs, since long term deposits have different amounts, interest rates and subscription periods. Moreover, human agents are hired to accept inbound phone calls, as well as sell other non-deposit products. In addition, it is difficult to estimate intrusiveness of an outbound call (e.g., due to a stressful conversation). Nevertheless, we highlight that current bank context favors more sensitive models: communication costs are contracted in bundle packages, keeping costs low; and more importantly, the 2008 financial crisis strongly increased the pressure for Portuguese banks to increase long term deposits. Hence, for this particular bank it is better to produce more successful sells even if this involves loosing some effort in contacting non-buyers.

VII. Conclusions and Recommendations

I would recommend a larger dataset to get better idea of the impact of the predictor variables on the response. Though this case study has used logistic regression there are alternative methodologies such as Decision Trees, Neural Networks and Support vector machines which are flexible, extend to non-linear applications when decision boundary is non-linear and may give more accurate predictions

Appendix A

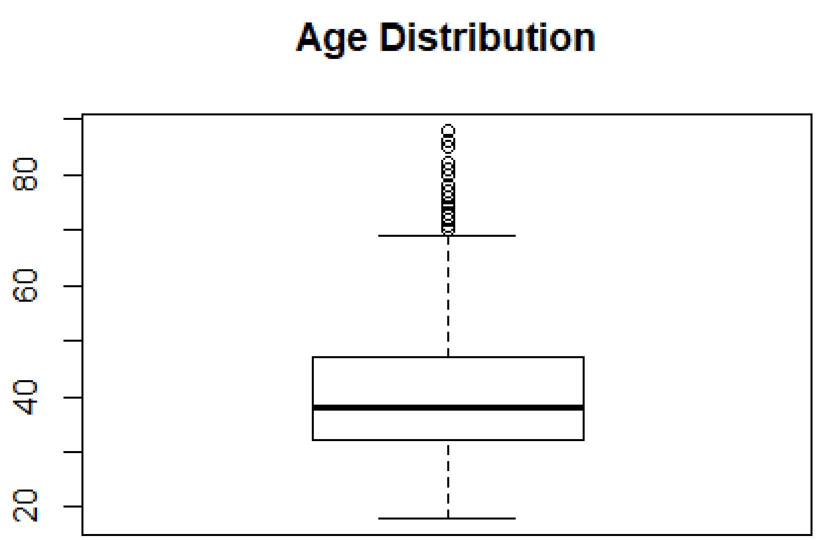
```
BankData = read.csv("/Users/prady_000/Documents/Vishwa/MSDA/DAA/Bank Marketin
g Casestudy/bank-additional.csv", sep= ';',header=TRUE)
str(BankData)

## 'data.frame':    4119 obs. of  21 variables:
##  $ age           : int  30 39 25 38 47 32 32 41 31 35 ...
##  $ job           : Factor w/ 12 levels "admin.","blue-collar",..: 2 8 8 8
1 8 1 3 8 2 ...
##  $ marital       : Factor w/ 4 levels "divorced","married",..: 2 3 2 2 2 3
3 2 1 2 ...
##  $ education     : Factor w/ 8 levels "basic.4y","basic.6y",..: 3 4 4 3 7
7 7 7 6 3 ...
##  $ default       : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 1 1 1 2 1
2 ...
##  $ housing       : Factor w/ 3 levels "no","unknown",..: 3 1 3 2 3 1 3 3 1
1 ...
##  $ loan          : Factor w/ 3 levels "no","unknown",..: 1 1 1 2 1 1 1 1 1
1 ...
##  $ contact       : Factor w/ 2 levels "cellular","telephone": 1 2 2 2 1 1
1 1 1 2 ...
##  $ month         : Factor w/ 10 levels "apr","aug","dec",..: 7 7 5 5 8 10
10 8 8 7 ...
##  $ day_of_week   : Factor w/ 5 levels "fri","mon","thu",..: 1 1 5 1 2 3 2
2 4 3 ...
##  $ duration      : int  487 346 227 17 58 128 290 44 68 170 ...
##  $ campaign      : int  2 4 1 3 1 3 4 2 1 1 ...
##  $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
##  $ previous      : int  0 0 0 0 0 2 0 0 1 0 ...
##  $ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2
2 1 2 2 1 2 ...
##  $ emp.var.rate  : num  -1.8 1.1 1.4 1.4 -0.1 -1.1 -1.1 -0.1 -0.1 1.1 ...
##  $ cons.price.idx: num  92.9 94 94.5 94.5 93.2 ...
##  $ cons.conf.idx : num  -46.2 -36.4 -41.8 -41.8 -42 -37.5 -37.5 -42 -42 -3
6.4 ...
##  $ euribor3m     : num  1.31 4.86 4.96 4.96 4.19 ...
##  $ nr.employed   : num  5099 5191 5228 5228 5196 ...
##  $ y             : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Appendix B

```
set.seed(1)

train <- sample(1:nrow(Bank2),nrow(Bank2)*.7,rep=FALSE)
test <- -train
```

Appendix C

```
Bank=BankData

Bank$pdaydummy = ifelse(Bank$pdays==999,0,1)

boxplot(Bank$age, main = "Age Distribution")
```



Checking for any possible outliers

```
remove_outliers <- function(x, na.rm = TRUE, ...) {
    qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
    H <- 1.5 * IQR(x, na.rm = na.rm)
    y <- x
    y[x < (qnt[1] - H)] <- NA
    y[x > (qnt[2] + H)] <- NA
    y
}

z=lapply(Bank[1],remove_outliers)
Bank1=cbind(z,Bank[-1])
```

Checking for any missing values of numeric variables

```
Bank2=na.omit(Bank1)
```

Checking for any missing values of categorical variables

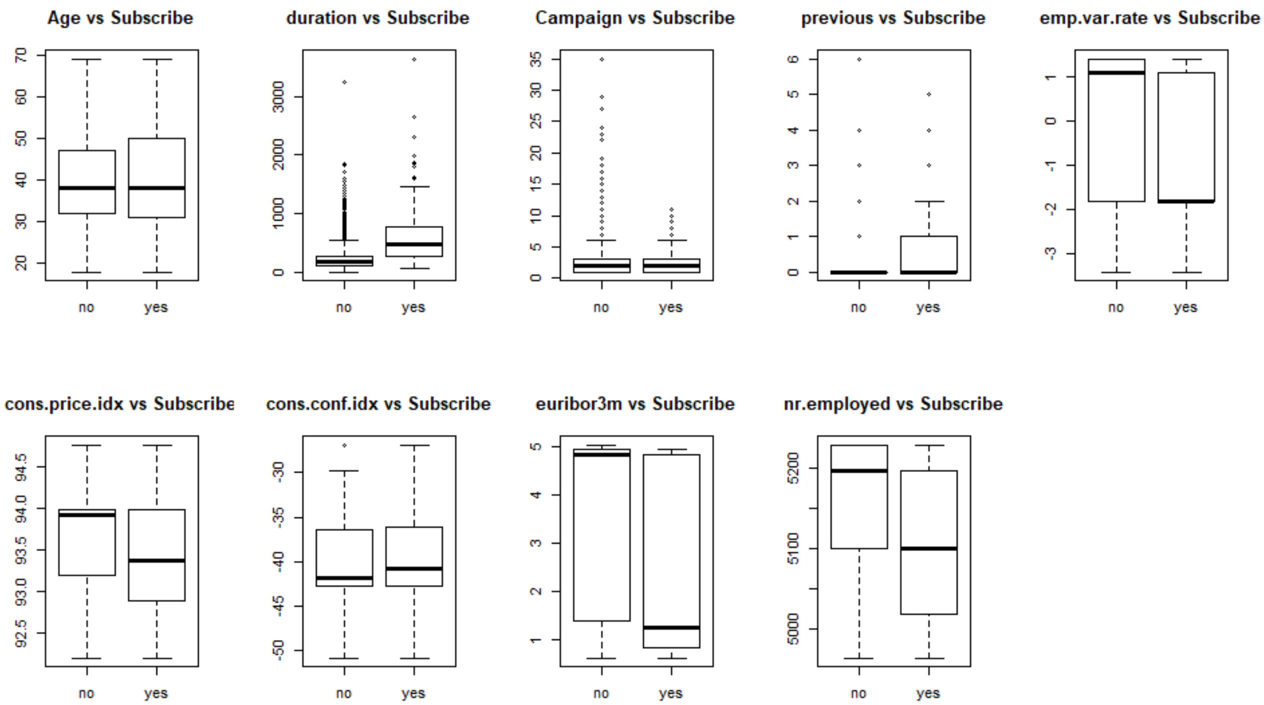
```
Bank2=subset(Bank2, !is.nan(Bank2$job))
Bank2=subset(Bank2, !is.nan(Bank2$marital))
Bank2=subset(Bank2, !is.nan(Bank2$education))
Bank2=subset(Bank2, !is.nan(Bank2$default))
Bank2=subset(Bank2, !is.nan(Bank2$housing))
Bank2=subset(Bank2, !is.nan(Bank2$loan))
Bank2=subset(Bank2, !is.nan(Bank2$contact))
Bank2=subset(Bank2, !is.nan(Bank2$month))
Bank2=subset(Bank2, !is.nan(Bank2$day_of_week))
Bank2=subset(Bank2, !is.nan(Bank2$poutcome))
```

Appendix D

```
summary(Bank2)

##       age                     job            marital
##  Min.   :18.00   admin.       :1012   divorced: 436
##  1st Qu.:32.00   blue-collar  : 883   married :2482
##  Median :38.00   technician   : 691   single  :1151
##  Mean   :39.76   services     : 393   unknown :  11
##  3rd Qu.:47.00   management   : 323
##  Max.   :69.00   self-employed: 159
##                  (Other)      : 619
##                education       default         housing           loan
##  university.degree  :1256   no     :3283   no     :1822   no     :3316
##  high.school        : 917   unknown: 796   unknown: 105   unknown: 105
##  basic.9y           : 573   yes    :   1   yes    :2153   yes    : 659
##  professional.course: 532
```

```
##  basic.4y            : 409
##  basic.6y            : 228
##  (Other)             : 165
##       contact          month      day_of_week    duration
##  cellular :2618   may    :1375   fri:763     Min.   :   0.0
##  telephone:1462   jul    : 708   mon:843     1st Qu.: 103.0
##                   aug    : 625   thu:851     Median : 180.0
##                   jun    : 529   tue:832     Mean   : 256.5
##                   nov    : 444   wed:791     3rd Qu.: 316.2
##                   apr    : 214               Max.   :3643.0
##                   (Other): 185
##     campaign          pdays          previous           poutcome
##  Min.   : 1.000   Min.   :  0.0   Min.   :0.0000   failure    : 445
##  1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000   nonexistent:3501
##  Median : 2.000   Median :999.0   Median :0.0000   success    : 134
##  Mean   : 2.545   Mean   :962.5   Mean   :0.1833
##  3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000
##  Max.   :35.000   Max.   :999.0   Max.   :6.0000
##
##   emp.var.rate     cons.price.idx  cons.conf.idx      euribor3m
##  Min.   :-3.4000   Min.   :92.20   Min.   :-50.80   Min.   :0.635
##  1st Qu.:-1.8000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.344
##  Median : 1.1000   Median :93.75   Median :-41.80   Median :4.857
##  Mean   : 0.1065   Mean   :93.58   Mean   :-40.53   Mean   :3.647
##  3rd Qu.: 1.4000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961
##  Max.   : 1.4000   Max.   :94.77   Max.   :-26.90   Max.   :5.045
##
##   nr.employed     y          pdaydummy
##  Min.   :4964   no :3648   Min.   :0.00000
##  1st Qu.:5099   yes: 432   1st Qu.:0.00000
##  Median :5191              Median :0.00000
##  Mean   :5168              Mean   :0.03676
##  3rd Qu.:5228              3rd Qu.:0.00000
##  Max.   :5228              Max.   :1.00000
##
```

```r
par(mfrow=c(2,5))

boxplot(age ~ y, data = Bank2, main = "Age vs Subscribe")
boxplot(duration ~ y, data = Bank2, main = "duration vs Subscribe")
boxplot(campaign ~ y, data = Bank2, main = "Campaign vs Subscribe")
boxplot(previous ~ y, data = Bank2, main = "previous vs Subscribe")
boxplot(emp.var.rate ~ y, data = Bank2, main = "emp.var.rate vs Subscribe")
boxplot(cons.price.idx ~ y, data = Bank2, main = "cons.price.idx vs Subscribe
")
boxplot(cons.conf.idx ~ y, data = Bank2, main = "cons.conf.idx vs Subscribe")
boxplot(euribor3m ~ y, data = Bank2, main = "euribor3m vs Subscribe")
boxplot(nr.employed ~ y, data = Bank2, main = "nr.employed vs Subscribe")
```



checking multicollinearity between continuous predictors

```
cor(Bank2[c(1,11,12,14,16,17,18,19,20)])

##                         age    duration     campaign    previous
## age             1.000000000  0.03874379 -0.003877032  0.00324792
## duration        0.038743786  1.00000000 -0.085429115  0.02548905
## campaign       -0.003877032 -0.08542911  1.000000000 -0.09057378
## previous        0.003247920  0.02548905 -0.090573776  1.00000000
## emp.var.rate    0.031497063 -0.02806244  0.174530737 -0.42342083
## cons.price.idx  0.015599620  0.01588511  0.146907671 -0.18160671
## cons.conf.idx   0.080883248 -0.03407881  0.008904608 -0.05626981
## euribor3m       0.041180001 -0.03089473  0.156872584 -0.46264139
## nr.employed     0.027887459 -0.04276668  0.158310413 -0.51640572
##                emp.var.rate cons.price.idx cons.conf.idx   euribor3m
## age              0.03149706     0.01559962   0.080883248  0.04118000
## duration        -0.02806244     0.01588511  -0.034078810 -0.03089473
## campaign         0.17453074     0.14690767   0.008904608  0.15687258
## previous        -0.42342083    -0.18160671  -0.056269807 -0.46264139
## emp.var.rate     1.00000000     0.75914689   0.216638588  0.97085895
## cons.price.idx   0.75914689     1.00000000   0.063638642  0.66516012
## cons.conf.idx    0.21663859     0.06363864   1.000000000  0.29656128
## euribor3m        0.97085895     0.66516012   0.296561283  1.00000000
## nr.employed      0.89984386     0.48531711   0.124957398  0.94305640
##                nr.employed
## age             0.02788746
```

```
## duration       -0.04276668
## campaign        0.15831041
## previous       -0.51640572
## emp.var.rate    0.89984386
## cons.price.idx  0.48531711
## cons.conf.idx   0.12495740
## euribor3m       0.94305640
## nr.employed     1.00000000
```

Appendix E

```
mod1 = glm(formula = y ~ .-pdays-duration, data = Bank2[train, ], family = bi
nomial)
summary(mod1)

##
## Call:
## glm(formula = y ~ . - pdays - duration, family = binomial, data = Bank2[tr
ain,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0568  -0.4088  -0.3229  -0.2506   2.9835
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.076e+02  1.262e+02  -1.644 0.100151
## age                           8.528e-03  8.328e-03   1.024 0.305858
## jobblue-collar               -2.654e-01  2.622e-01  -1.012 0.311478
## jobentrepreneur              -4.074e-01  4.165e-01  -0.978 0.327944
## jobhousemaid                 -5.464e-01  5.246e-01  -1.042 0.297530
## jobmanagement                -6.256e-01  3.063e-01  -2.043 0.041097 *
## jobretired                   -4.430e-01  3.778e-01  -1.173 0.240982
## jobself-employed             -5.381e-01  4.015e-01  -1.340 0.180120
## jobservices                  -3.495e-01  2.899e-01  -1.205 0.228019
## jobstudent                    2.550e-02  4.180e-01   0.061 0.951361
## jobtechnician                -1.411e-01  2.240e-01  -0.630 0.528714
## jobunemployed                -1.242e-02  3.730e-01  -0.033 0.973449
## jobunknown                   -3.991e-01  7.840e-01  -0.509 0.610694
## maritalmarried               -2.189e-01  2.271e-01  -0.964 0.335199
## maritalsingle                -3.298e-01  2.673e-01  -1.234 0.217224
## maritalunknown               -4.417e-02  1.308e+00  -0.034 0.973061
## educationbasic.6y             2.093e-01  3.835e-01   0.546 0.585273
## educationbasic.9y            -1.560e-01  3.230e-01  -0.483 0.629197
## educationhigh.school         -9.621e-02  3.117e-01  -0.309 0.757601
## educationilliterate          -1.247e+01  5.354e+02  -0.023 0.981418
## educationprofessional.course  2.601e-01  3.326e-01   0.782 0.434260
## educationuniversity.degree    9.599e-02  3.129e-01   0.307 0.759015
```

```
## educationunknown              -3.498e-01  4.506e-01  -0.776 0.437491    
## defaultunknown                -3.500e-02  2.124e-01  -0.165 0.869117    
## defaultyes                    -1.036e+01  5.354e+02  -0.019 0.984568    
## housingunknown                -6.312e-01  4.988e-01  -1.265 0.205725    
## housingyes                    -1.410e-01  1.392e-01  -1.013 0.311082    
## loanunknown                           NA         NA      NA       NA    
## loanyes                        5.495e-02  1.853e-01   0.296 0.766853    
## contacttelephone              -1.049e+00  2.863e-01  -3.665 0.000248 ***
## monthaug                       2.860e-01  4.334e-01   0.660 0.509366    
## monthdec                       1.505e+00  7.278e-01   2.068 0.038593 *  
## monthjul                      -2.183e-02  3.481e-01  -0.063 0.949995    
## monthjun                      -1.698e-01  4.328e-01  -0.392 0.694875    
## monthmar                       2.268e+00  5.608e-01   4.045 5.24e-05 ***
## monthmay                      -2.616e-01  2.910e-01  -0.899 0.368595    
## monthnov                      -5.735e-01  4.123e-01  -1.391 0.164267    
## monthoct                      -1.075e-01  5.224e-01  -0.206 0.836893    
## monthsep                       1.213e-02  6.179e-01   0.020 0.984341    
## day_of_weekmon                -3.076e-02  2.184e-01  -0.141 0.888013    
## day_of_weekthu                 1.342e-01  2.160e-01   0.621 0.534428    
## day_of_weektue                 3.716e-02  2.218e-01   0.168 0.866973    
## day_of_weekwed                 1.509e-01  2.233e-01   0.676 0.499020    
## campaign                      -6.415e-02  4.051e-02  -1.583 0.113311    
## previous                       3.559e-02  1.885e-01   0.189 0.850236    
## poutcomenonexistent            4.490e-01  3.154e-01   1.424 0.154564    
## poutcomesuccess               -7.529e-03  7.596e-01  -0.010 0.992091    
## emp.var.rate                  -1.328e+00  4.764e-01  -2.789 0.005291 ** 
## cons.price.idx                 1.986e+00  8.355e-01   2.377 0.017436 *  
## cons.conf.idx                  3.568e-02  2.788e-02   1.280 0.200718    
## euribor3m                      2.316e-01  4.328e-01   0.535 0.592559    
## nr.employed                    3.909e-03  1.019e-02   0.384 0.701268    
## pdaydummy                      1.541e+00  7.612e-01   2.025 0.042883 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2024.2  on 2855  degrees of freedom
## Residual deviance: 1565.7  on 2804  degrees of freedom
## AIC: 1669.7
## 
## Number of Fisher Scoring iterations: 12
```

```r
library(MASS)
mod2 = stepAIC(mod1,direction="backward",trace=F)
summary(mod2)
```

```
## 
## Call:
## glm(formula = y ~ contact + month + campaign + previous + emp.var.rate + 
##     cons.price.idx + euribor3m + pdaydummy, family = binomial, 
```

```
##     data = Bank2[train, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0864  -0.3921  -0.3300  -0.2693   2.8300
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.786e+02  3.203e+01  -5.578 2.43e-08 ***
## contacttelephone -9.065e-01  2.497e-01  -3.630 0.000283 ***
## monthaug          6.227e-01  3.025e-01   2.058 0.039561 *
## monthdec          1.605e+00  6.437e-01   2.494 0.012641 *
## monthjul          9.583e-02  3.292e-01   0.291 0.770949
## monthjun         -1.239e-01  3.539e-01  -0.350 0.726230
## monthmar          2.223e+00  4.654e-01   4.776 1.79e-06 ***
## monthmay         -2.332e-01  2.591e-01  -0.900 0.368141
## monthnov         -6.042e-01  3.571e-01  -1.692 0.090654 .
## monthoct         -3.047e-04  4.115e-01  -0.001 0.999409
## monthsep          3.526e-01  4.062e-01   0.868 0.385368
## campaign         -6.613e-02  4.022e-02  -1.644 0.100129
## previous         -1.723e-01  1.183e-01  -1.456 0.145437
## emp.var.rate     -1.499e+00  3.360e-01  -4.461 8.15e-06 ***
## cons.price.idx    1.866e+00  3.349e-01   5.573 2.50e-08 ***
## euribor3m         5.860e-01  2.586e-01   2.266 0.023437 *
## pdaydummy         1.530e+00  2.881e-01   5.311 1.09e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2024.2  on 2855  degrees of freedom
## Residual deviance: 1589.4  on 2839  degrees of freedom
## AIC: 1623.4
##
## Number of Fisher Scoring iterations: 6
```

```
LRPredProb = predict.glm(mod2,newdata= Bank2[test,], type= "response")
LRPredsub = ifelse(LRPredProb >= 0.5,"yes","no")

caret::confusionMatrix(Bank2$y[test],as.factor(LRPredsub))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  1091   26
##        yes   78   29
##
##                Accuracy : 0.915
##                  95% CI : (0.898, 0.9301)
```

```
##     No Information Rate : 0.9551
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3175
##  Mcnemar's Test P-Value : 5.705e-07
##
##             Sensitivity : 0.9333
##             Specificity : 0.5273
##          Pos Pred Value : 0.9767
##          Neg Pred Value : 0.2710
##              Prevalence : 0.9551
##          Detection Rate : 0.8913
##    Detection Prevalence : 0.9126
##       Balanced Accuracy : 0.7303
##
##        'Positive' Class : no
##
```

Appendix F

removing `emp.var.rate`

```
mod3 = glm(formula = y ~ .-pdays-duration-emp.var.rate, data = Bank2[train, ]
, family = binomial)
summary(mod3)

##
## Call:
## glm(formula = y ~ . - pdays - duration - emp.var.rate, family = binomial,
##     data = Bank2[train, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0627  -0.4137  -0.3255  -0.2495   2.9107
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   60.685680  81.967580   0.740  0.45908
## age                            0.009245   0.008319   1.111  0.26643
## jobblue-collar                -0.277693   0.261507  -1.062  0.28828
## jobentrepreneur               -0.384560   0.414166  -0.929  0.35314
## jobhousemaid                  -0.539480   0.522110  -1.033  0.30148
## jobmanagement                 -0.598540   0.304704  -1.964  0.04949 *
## jobretired                    -0.402453   0.375107  -1.073  0.28332
## jobself-employed              -0.534545   0.401598  -1.331  0.18317
## jobservices                   -0.327901   0.289700  -1.132  0.25769
## jobstudent                     0.140987   0.411444   0.343  0.73185
## jobtechnician                 -0.145441   0.223755  -0.650  0.51569
## jobunemployed                 -0.001779   0.371370  -0.005  0.99618
## jobunknown                    -0.314974   0.781325  -0.403  0.68685
```

```
## maritalmarried               -0.220830   0.226851  -0.973  0.33032
## maritalsingle                -0.325082   0.266421  -1.220  0.22240
## maritalunknown               -0.004330   1.289705  -0.003  0.99732
## educationbasic.6y             0.237966   0.381296   0.624  0.53256
## educationbasic.9y            -0.158986   0.323122  -0.492  0.62270
## educationhigh.school         -0.094329   0.310711  -0.304  0.76144
## educationilliterate         -12.355792 535.411458  -0.023  0.98159
## educationprofessional.course  0.255905   0.332243   0.770  0.44116
## educationuniversity.degree    0.107058   0.312127   0.343  0.73160
## educationunknown             -0.336149   0.451759  -0.744  0.45682
## defaultunknown               -0.040038   0.211816  -0.189  0.85007
## defaultyes                  -10.341883 535.411424  -0.019  0.98459
## housingunknown               -0.669432   0.503564  -1.329  0.18372
## housingyes                   -0.149146   0.138770  -1.075  0.28248
## loanunknown                         NA         NA      NA       NA
## loanyes                       0.052968   0.184442   0.287  0.77397
## contacttelephone             -0.819118   0.261408  -3.133  0.00173 **
## monthaug                     -0.348882   0.374058  -0.933  0.35098
## monthdec                      1.136856   0.708696   1.604  0.10868
## monthjul                      0.050497   0.344021   0.147  0.88330
## monthjun                      0.636004   0.323057   1.969  0.04899 *
## monthmar                      1.660077   0.525183   3.161  0.00157 **
## monthmay                     -0.530621   0.273318  -1.941  0.05221 .
## monthnov                     -0.572335   0.416209  -1.375  0.16910
## monthoct                     -0.421431   0.524215  -0.804  0.42144
## monthsep                     -0.815180   0.550959  -1.480  0.13899
## day_of_weekmon               -0.041890   0.217870  -0.192  0.84753
## day_of_weekthu                0.135920   0.215497   0.631  0.52822
## day_of_weektue                0.038646   0.221092   0.175  0.86124
## day_of_weekwed                0.138646   0.222970   0.622  0.53406
## campaign                     -0.066290   0.040572  -1.634  0.10228
## previous                      0.053121   0.192062   0.277  0.78210
## poutcomenonexistent           0.456232   0.317398   1.437  0.15060
## poutcomesuccess               0.105920   0.763647   0.139  0.88969
## cons.price.idx                0.045362   0.457277   0.099  0.92098
## cons.conf.idx                 0.025972   0.027702   0.938  0.34847
## euribor3m                     0.110054   0.438725   0.251  0.80193
## nr.employed                  -0.012868   0.008324  -1.546  0.12211
## pdaydummy                     1.485533   0.766136   1.939  0.05250 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2024.2  on 2855  degrees of freedom
## Residual deviance: 1573.4  on 2805  degrees of freedom
## AIC: 1675.4
##
## Number of Fisher Scoring iterations: 12
```

```
library(MASS)
mod4 = stepAIC(mod3,direction="backward",trace=F)
summary(mod4)

##
## Call:
## glm(formula = y ~ contact + month + campaign + nr.employed +
##     pdaydummy, family = binomial, data = Bank2[train, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0242  -0.3894  -0.3380  -0.2399   2.7409
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      51.319826   5.277894   9.724  < 2e-16 ***
## contacttelephone -0.533394   0.196128  -2.720 0.006536 **
## monthaug          0.137350   0.285086   0.482 0.629959
## monthdec          1.457856   0.621356   2.346 0.018964 *
## monthjul          0.363858   0.292651   1.243 0.213751
## monthjun          0.793471   0.292449   2.713 0.006664 **
## monthmar          1.713681   0.465577   3.681 0.000233 ***
## monthmay         -0.479762   0.253832  -1.890 0.058748 .
## monthnov         -0.270424   0.303807  -0.890 0.373402
## monthoct          0.043499   0.388179   0.112 0.910777
## monthsep         -0.215306   0.413764  -0.520 0.602814
## campaign         -0.066135   0.039811  -1.661 0.096668 .
## nr.employed      -0.010359   0.001037  -9.991  < 2e-16 ***
## pdaydummy         1.370852   0.244088   5.616 1.95e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2024.2  on 2855  degrees of freedom
## Residual deviance: 1601.1  on 2842  degrees of freedom
## AIC: 1629.1
##
## Number of Fisher Scoring iterations: 6

LRPredProb = predict.glm(mod4,newdata= Bank2[test,], type= "response")
LRPredsub = ifelse(LRPredProb >= 0.5,"yes","no")

caret::confusionMatrix(Bank2$y[test],as.factor(LRPredsub))

## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##         no  1092   25
```

```
##        yes   76   31
##
##                Accuracy : 0.9175
##                  95% CI : (0.9006, 0.9323)
##     No Information Rate : 0.9542
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3408
##  Mcnemar's Test P-Value : 6.519e-07
##
##             Sensitivity : 0.9349
##             Specificity : 0.5536
##          Pos Pred Value : 0.9776
##          Neg Pred Value : 0.2897
##              Prevalence : 0.9542
##          Detection Rate : 0.8922
##    Detection Prevalence : 0.9126
##       Balanced Accuracy : 0.7443
##
##        'Positive' Class : no
##
```