**Project #3 for MSDS 6372-402:**

**Analysis of Banking Data Using Logistic Regression and ROC curves**

**Date Due August 15, 2017**

**Submitted by:**

Kevin Okiah

James Hosker

James Park

**Course: MSDS 6372-402 (Mon, 8:30 to 10:00 PM)**

**Analysis of Banking Data Using Logistic Regression and ROC curves**

**Introduction & Problem Statement**

A common method to solicit bank deposit subscriptions in the banking industry is to call a prospective client with an offer of attractive interest rates through directed marketing campaigns. In order to conduct an efficient marketing campaign, contacts are targeted to achieve a successful result, which is a deposit subscription in this case. This paper describes the development and identification of a model that can explain the odds of success based on several explanatory variables related to each contact.

**Concerns and Limitations**

All data used in this study is from a data set donated to the University of California Irvine (UCI) Machine Learning Repository by S. Moro, P. Rita, and P. Cortez in 2014 [1]. In the interest of efficient computational time, a random data set of 4,119 (~10%) from the original 41,188 samples is used in this study that is provided by the website. Several of the explanatory factors involved in this study contained “unknown” entries. Some of the “unknowns” will be rationally interpreted by an assumption or through data inspection and some factors, along with the unknown levels, may be removed from the model due to lack of significance. This analysis is an observational study and therefore causality cannot be determined even though the data subset was randomly sampled from the original dataset.  The logistic model will, however, be insightful in directing marketing efforts to those individuals more likely to subscribe based on the data sample.

**Summary of the Data Collection**

The random subset of 4,119 samples consisted of the sixteen (16) explanatory variables (categorical, binary, and numerical) and one (1) binary response variables (see [Appendix A](#AppendixA) for full data set definitions). Only six explanatory variables were selected for this study based primarily on statistical significance from a logistic regression that used subscribed as the categorical response variable. The six variables selected may be viewed in **Table 1**, description of selected variables.

|  |
| --- |
| **Input variables (bank client data):**   1. **loan**: has personal loan? (binary: "yes","no")   # related with the last contact of the current campaign:   1. **contact**: contact communication type (categorical: "unknown","telephone","cellular") 2. **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec") 3. **duration**: last contact duration, in seconds (numeric) 4. **poutcome**: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success") 5. **subscribed** - has the client subscribed a term deposit? (binary: “1” for yes and “0” for no) |

Table 1: Description of Selected Variables

*While developing and experimenting with the model, the “****Subscribed****” variable was created from the original “Y” variable such that a “No” was indicated by a “0” and a “Yes” was indicated by a “1”.*  Initially the “ages” and “months” variables were consolidated into new variables called “Age\_Group” and “Quarter”. The “Age\_Group” variables lacked statistical significance so it was dropped from further investigations (p-values ranged from 0.34713 to 0.80863). Similarly, “Quarter” was found not be significant but “month” was significant. Additional information regarding the grouping efforts may be found in [Appendix B](#AppendixB) on variable grouping.

**Data Cleaning:** Initially, a tally of the term “unknown” was established by each factor to get a sense of where the data may need to be better understood, corrected, or further interpreted prior to analysis. The associated R code and output may be viewed in **Table 2** and[Appendix F](#AppendixF) for data cleaning.

|  |
| --- |
| **R-Code:** count\_unknown<-function(v){length(v[v=="unknown"])}  unknown <-sapply(my\_data, count\_unknown)  **R-Output:** |

Table 2: Tally of “unknown” terms in the training data set using R code

**Train and Test Cross-Validation Data Sets Created**

Prior to construction of the model, a training and test data set were created by randomly splitting the data set in half. The model was built using the “Train” data set and tested using the “Test” data set. The R code to create these two data sets can be seen in [Appendix C](#AppendixC) for creation of cross-validation data sets.

**Initial Data Exploration**

This study included the examination of various methods of exploratory data analysis including summary statistics, frequency tables, scatter plots, boxplots and histogram distributions by individual or paired explanatory factors that are highlighted by the “**Subscribed**” response factor. This provides a visual initial understanding of each factors relationship to the subscription response variable. Here, there are three factors of interest explored from the training data set used in the final model. For all factors that were explored, see [Appendix D](#AppendixD) for exploratory data analysis, including loan, education, month, default, housing, age, job type, pdays and previous. As noted in **Table 1**, some explanatory data sets have several unknowns including contact and poutcome. All R-code for this analysis can be found in [Appendix I](#AppendixI) and SAS code in [Appendix K](#AppendixK).

* **Contact:** In **Table 3**, many yeses to a subscription came from contacts that had cell phones (78%) compared contacts with telephones or unknown forms of contact. The yeses for a subscription using cellular phones were nearly 9% of all contacts made. In **Figure 1**, the type of contact shows that clearly cellular results in more customers that did subscribed. *It is noted that the total success rate was low at 11.4%, 257 out of 2260 in the training data set.*

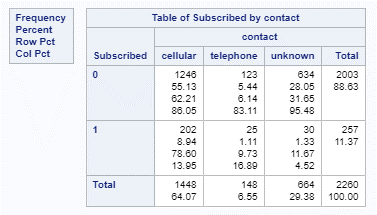
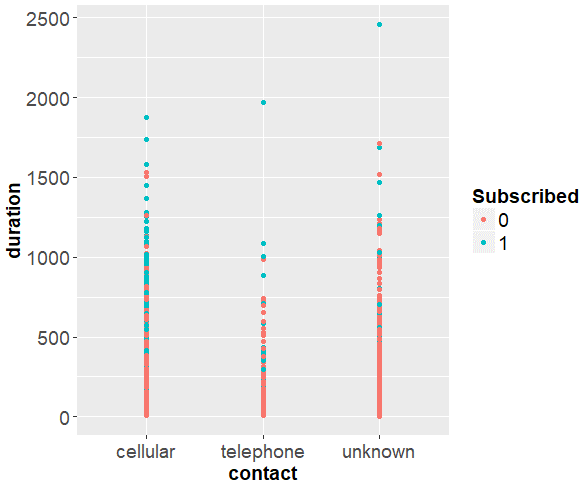
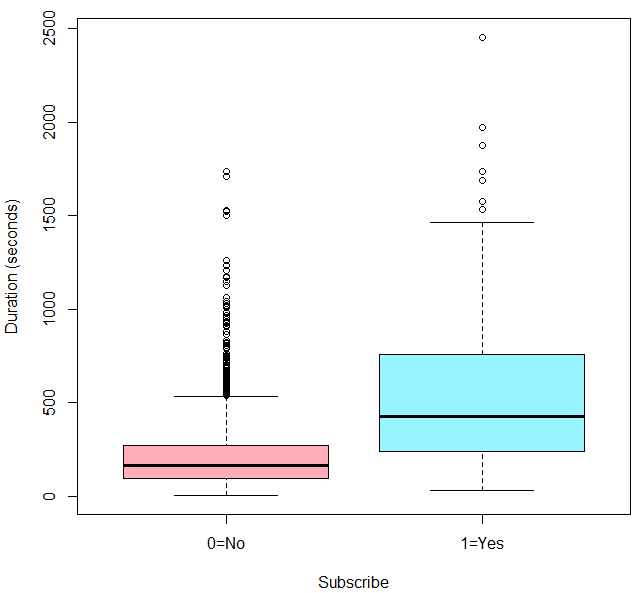
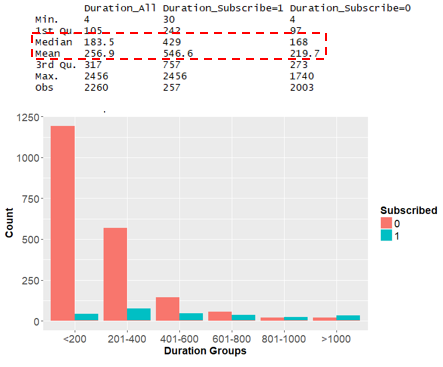
 

Table 3: SAS Frequency Table for Contact Figure 1: R Scatter Duration vs. Contact by Subscribed

* **Duration:** **Figure 2** shows the boxplot of duration vs. subscribed. The dashed red-line shows the mean for the entire training data set (256.9 sec). **Figure 3** shows the summary statistics of the duration of conversations and a graph of the duration group count by subscription (note that greater than 400 seconds has a higher likelihood of a yes to the subscription). The blue bars in the bar charts become equal or higher than the red bars after 800 second of conversation. From the stats table, the median and mean seconds on the phone for those that chose to subscribe (mean=546.6, median=429) was much higher than those that did not subscribe (mean=219.7, median=168). Subscriptions with a no response have a lower mean duration and boxplot distribution than subscriptions with a yes.

Mean All=

256.9

Figure 2: R Boxplot Duration vs. Subscribe Figure 3: R Duration Stats & Duration Group Count by Subscribe

* **poutcome:** In **Table 4**, if the prior campaign was successful (75 contacts), the likelihood of a subscription success was 66.7% no matter what the duration of the conversation while failure, unknown and other subscription hit ratios were lower. Nearly 19.5% of all subscriptions (50 out of 257 total subscription) came from customers where a prior campaign was successful. It is noted that poutcome had many unknowns from **Table 4** but here the focus is on the significance of prior campaign success. Although the unknown poutcome is 61.1% of all the subscriptions, the success ratio of unknown is much less at 8.6%, meaning a successful subscription outcome is less likely for unknowns. **Figure 4** shows similar results.

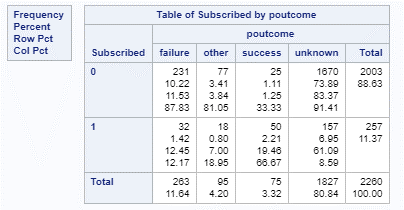
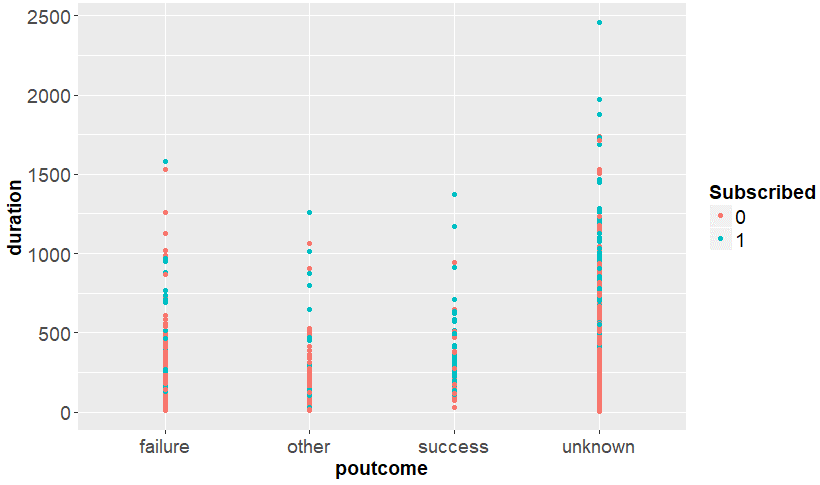
 

Table 4: SAS Frequency Table for Month Figure 4: R Scatter Duration vs. poutcome by Subscribed

Several visualizations (histograms and bar charts) were created for variables that ended up not being included in the model. In the interest of focusing on the remaining variables, this work is shown in [Appendix E](#AppendixE) for backup data exploration including age group and quarters.

**Logistic Regression Model for Banking Campaign Term Deposit Subscription**

In this section, the fitting of a multiple logistic regression model is discussed in order to predict the probability of a client subscribing to a bank term deposit.

**Variable Selection**

With any regression model, it is important to include all relevant variables. For the data, more variables generally produce a better model fit to the data. However, excessive variables always influence the coefficient in the model and can contribute to overfitting. A complicated model, including many insignificant variables, may result in less predictive power and it may often be difficult to interpret the results. The initial logistic regression model was run in R with the training data set as described in **Tables 1** and **2** of [Appendix G](#AppendixG) for table of full logistic regression. The “**Subscribed**” variable is the binary response factor and all of the other factors are explanatory. The associated output is in **Table 1** of[Appendix H](#AppendixH)for interim logistic regression(significant factors highlighted in red and with Asterisk \*). The following factors have significant p-values < 0.05: loan, education, poutcome, contact, month, housing and duration. However, through further exploratory and iterative analysis, the following factors were removed:

* **Remove housing:** On closer inspection of housing (see [Appendix D](#AppendixD) and [E](#AppendixE) for exploratory data), it was found that whether or not the contact has or does not have a housing loan results in close to a 60/40 split on receiving a yes to a subscription so it may not be that significant. In the full logistic regression in [Appendix G](#AppendixG) output, housing has p-value of 0.04.
* **Remove education:** A new interim model was run with explanatory variables of loan, education, poutcome, contact, month and duration as shown in **Table 1** of [Appendix H](#AppendixH). All education classifications were no longer significant, p-values 0.06 to 0.21 (purple highlight).
* **Remove month:** The month variable was subsequently removed through methods discussed in the following section titled Model Hypothesis Testing.

**Model Hypothesis Testing**

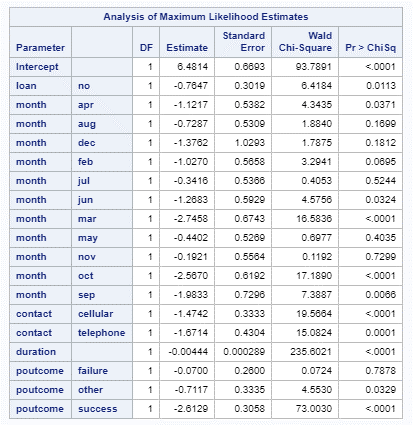
This analysis first establishes whether or not any of the predictors (explanatory variables) are significant by a **Global Test**. If the Global Test turns out to have any significant predictors, each parameter is then tested using **Individual Wald Chi-Square Tests**. Finally, the testing was concluded with the **Partial test**, which was reduced model minus the full model. In addition, all other parameters beyond the intercept were significant.

**Global Test:** In the first step, the null hypothesis is rejected where all of the predictor coefficients (𝛃’s) are zero. Per the SAS code and output in **Table 5**, the test indicates that the null hypothesis is rejected because at least one predictor is significant (p-values for all three tests are <0.0001).

|  |
| --- |
| **SAS-Code:**  proc logistic data=Train;  class month loan contact poutcome / param=ref;  model Subscribed = month loan contact duration poutcome / scale=none details lackfit influence;  output out=myoutput prepares=I p=probpred resdev=resdev reschi=reschi;  run; |
| **SAS-Output:** |

Table 5: Null Hypothesis Rejected (At Least one Predictor Coefficient is Not Zero).

**Individual Test:** Each parameter was then tested using the Wald Chi-Square test to determine their significance. **Table 6** shows an interim logistic regression that several parameters were identified as significant, thus rejecting the null that they are zero. The months of January (intercept ref), April, June, July, October and September were significant (p-values ranging from <0.0001 to 0.0371) as indicated by the red boxes in **Table 6**. Both forms of contact, “cellular” and “telephone” were significant (p-value <0.0001) in the blue boxes. The continuous factor, “duration”, was also found to be significant (p-value <0.0001) in the green box. Not having a loan is significant (p-value= 0.0113) in the yellowbox. Finally, the prior outcome (poutcome) factors “other” and “success” were determined to be significant (p-value of 0.0329 and < 0.0001, respectively) in the purple boxes.



Ref = “Jan”

Table 6: Individual Parameter Tests

**Reduced Model Tests:** This tested the null hypothesis that the reduced model (Intercept only) is the same as the full model (with Covariates). As shown in **Table 7**, removing month resulted in a lower AIC score and similar SC score, while simplifying the model and decreasing the degrees of freedom. The difference in Chi-square score for the reduced model and the full model is 0.704. This is difference of 1601.058 and 1601.762 in **Table 7** for the reduced model test.

|  |
| --- |
| With Month (p-value of 1 with 18 df difference). Do not reject null of the reduced model. Consider removal of the Month variable as model benefit may not be worth the 18 degrees of freedom. |
| Month Removed (p-value <0.0001 with 7 df difference). Reject null of the reduced model and use of the full model without the month variable. |
| Month and poutcome removed.(pvalue <0.0001 with 4 df difference). Reject null of the reduced model. However,the AIC/SC measures indicate that the model with only month removed is a better fit. |

Table 7: Reduced Model Test

**Logistic Regression Model**

The final Logistic model had “**Subscribed**” as the response variable and the following as explanatory variables: loan, contact, duration and poutcome. The logistic model was re-run with the reduced variables and the output is shown in **Table 8**. All R-code for this analysis can be found in [Appendix I](#AppendixI) and [J](#AppendixJ) and SAS code in [Appendix K](#AppendixK).

**R-code:** LogisticModelFinal\_train <- glm(Subscribed~loan+contact+duration+poutcome,

family=binomial(), data =Train)

summary(LogisticModelFinal\_train)

**R-Output:**

Call:

glm(formula = Subscribed ~ loan + contact + duration + poutcome,

family = binomial(), data = Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.7923 -0.3994 -0.2972 -0.1736 3.1364

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -5.6992341 0.3887238 -14.661 < 2e-16 \*\*\*

loanno 0.9214844 0.2927046 3.148 0.00164 \*\*

contactcellular 1.3216380 0.2496127 5.295 1.19e-07 \*\*\*

contacttelephone 1.6737079 0.3568878 4.690 2.74e-06 \*\*\*

duration 0.0042150 0.0002798 15.066 < 2e-16 \*\*\*

poutcomefailure 0.2079084 0.2391218 0.869 0.38459

poutcomeother 0.8394710 0.3124473 2.687 0.00721 \*\*

poutcomesuccess 2.9453448 0.2845218 10.352 < 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: 1601.1 on 2259 degrees of freedom

Residual deviance: 1119.8 on 2252 degrees of freedom

AIC: 1135.8

Number of Fisher Scoring iterations: 6

Table 8: R-code and Final Logistic Regression Model Output

From **Table 8** above, the generalized logistic model is listed in **Equation 1** below. Note the following reference terms are part of the intercept: , , and.

= + + + + + + +

Equation 1: General and Final Logit Model [2]

For practical interpretation, the logit function was exponentiated to yield the odds of a yes response at levels as shown in **Equation 2** below.

=

Equation 2: Expontiated Logit

**Interpratation of coefficients**

The odds that a client is subscribed for a term deposit , when all other explanatory levels are not considered ( ) is given by the intercept ( = 0.003349), which is the odds ratio shown in **Table 9**. However, the intercept does include reference factor levels (loan=yes, contact=unknown and poutcome=unknown) . This means the odds of a cleint not being subscribed for term deposit is ~300 times the odds of the client being subscribed if no other explanatory levels are considered other than the reference level factors.

For example, **loan** is binary variable indicating if a cleint has a personal loan. A client who has a personal loan (loan=yes) that gets subscribe to term deposit may be compared to a client who doesn’t have personal loan (loan=no). The odds ratio is estimated to be:

*= exp[]*

*= exp[] = 2.513 and [95% CI (1.457,4.614)]*

As shown in **Table 9** below for the odds ratio, loan equalled to “no” has an odds ratio of 2.513.

**R-Code:** exp(cbind(OddsRatio=coef(LogisticModelFinal\_train), confint(LogisticModelFinal\_train)))

**SAS-Code:**

proc logistic data=Train;

class loan(ref="no") contact(ref="cellular") poutcome(ref="success") / param=ref;

model Subscribed = loan contact duration poutcome / scale=none details lackfit influence;

output out=myoutput predprobs=I p=probpred resdev=resdev reschi=reschi;

run;

**R-Output: SAS-Output:**

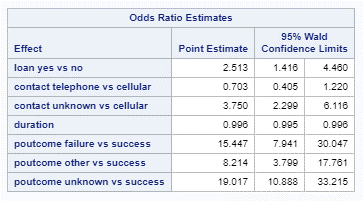
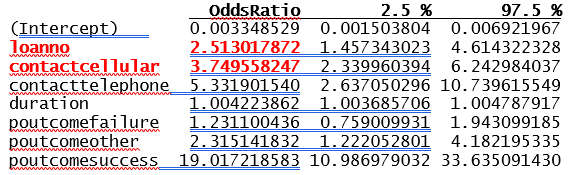


Table 9: R-Code and SAS Odds Ratio Output (some references are reversed in SAS)

1. So the odds of a client not carrying a personal loan but subscribing for the term deposit is 2.51 times the odds of a client carrying a personal loan and subscribing for the term deposit.
2. Simillarly, the odds of contact with a cellular phone and subscribing for a term deposit is 3.75 () times the odds of a contact with unknown method (reference) and subscribing for a term deposit.

**ROC Output Comparison**

The receiver operating characteristic (ROC) curves are commonly used to characterize the sensitivity/specificity tradeoffs for a binary classifier (logistic regression). **Figure 5** shows a comparison of ROC curves of the final model to a model with only duration as the explanatory variable and with only all categorical explanatory variables (poutcome, loan and contact). ROC using duration, poutcome, loan and contact has predicted probabilities that will be close to 0 or 1 and the area under the ROC curve is close to 1 at 0.876. With only duration, the area under the curve drops to 0.809 indicating duration explains much of the predicted probabilities for subscribed. Finally, using only the three categorical variables combined (poutcome, contact & loan), the area under the curves drops to 0.698, which still indicates the importance of these variables.

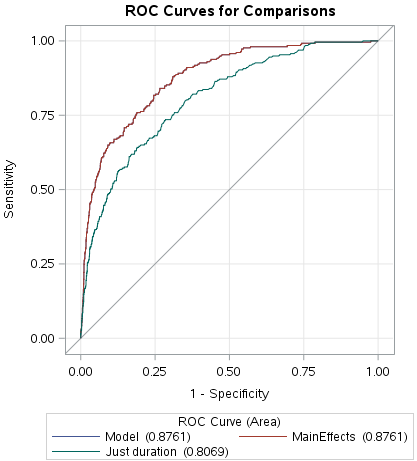
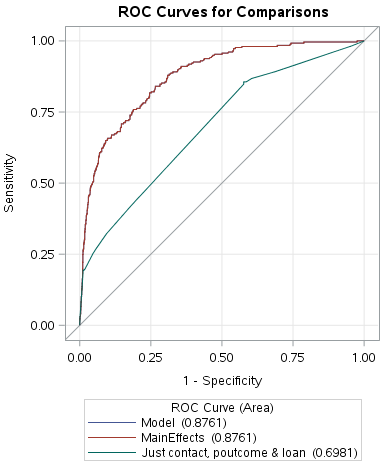
 

Figure 5: ROC curves for Full Logistic Model vs. Only Duration and vs. Contact, Poutcome and Loan (SAS)

**Model Validation**

To validate the performance of our model, the *predict()* function in R is used to compute probabilities on the test dataset. If the predicted probability was greater than 0.5, it is predicted a client would be subscribed to a term deposit.

The *table()* function was used to produce a confusion matrix in order to determine how many observations were correctly or incorrectly classified. The diagonal elements in the confusion matrix highlighted in **green** in **Table 10** indicate the correct predictions, whereas the off-diagonals represent the incorrect predictions. Our model correctly predicted that 75 clients would be subscribed to the term deposit and 1951 would not be subscribed for a total of 1951 + 75 = 2026 correct prediction. The *mean ()* function is used to compute the fraction of correct predictions for term deposit subscription. In our case, logistic regression model correctly predicted subscription to term deposit 89.6% of the times.

# Assessing the predictive ability of the model

fitted.probs <-predict(LogisticModelFinal\_train, Test,type='response')

fitted.predictions <- ifelse(fitted.results > 0.5,1,0) # significance level is 0.5

# check results of classfication

misClasificError <- mean(fitted.predictions != Test$Subscribed)

table(fitted.predictions,Test$Subscribed)

**fitted.predictions 0 1**

**0 1951 189**

**1 46 75**

print(paste('Model Accuracy is: ',1-misClasificError))

**[1] "Model Accuracy is: 0.896063688633348"**

Table 10: Model Predictive Ability

**Figure 6** below shows the ROC curve for the fitted model with the sensitivity on the x-axis and specificity plotted on the y-axis. The 45 degrees reference line is the line of non-discrimination and the area below it (=0.5) represents the classification occurring purely by chance. The graph shows that there is some scope for improvement for the predictive power of the model on both the test and training data. The AUC for training and test datasets are close indicating the model fits well

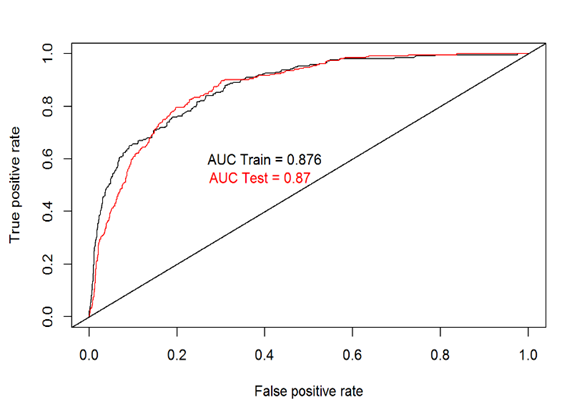


Figure 6: ROC Prediction Curves using Test Data

**Sampling Schema and Model Assumptions**

The data collection in this project was under a **Poisson** **sampling** scheme, meaning the sampling was unrestricted and the total sample size was not fixed as the data was collected over a time period from May 2008 to June 2013 for all of the phone contacts during that time. Given the scheme is Poisson, independence and homogeneity are then tested. [2] In addition, the odds ratio is the parameter that can be used to compare two groups of binary responses and the comparison of odds extends nicely to regression analysis. In practice, the odds ratio tends to remain more nearly constant over levels of confounding variables. [2]

**Model Assumptions**

Model assumption for our logistic regression are below [2]:

* **Homogeneity**: The distribution of responses across categories are likely similar for our data set. No evidence of lack of homogeneity was found in our data set during the data exploration. For this dataset, linearity may not be applicable since the analysis only had one continuous explanatory variable "duration" and several categorical explanatory variables. If the model had more than one continuous exploratory variable, linearity could have been checked by investigating the significance of their interactions term.
* **Independence:** Responses for combined categories of two categorical variables are independent (columns). Independence of the clients that were sampled in this campaign maybe assumed. However, the website for our data does mentions that some clients were contacted multiple times.
* **Goodness of fit:** The final model response is consistent with a stated probability distribution where parameters are specified but may have unknown parameter values.

**Residual Diagnostics**

Observations were inspected for high or low residuals and high or low leverage. In Figure 7 for the influence diagnostics, the Pearson residuals show somewhat constant variance between -5 to 5 and only a few points with high leverage (397 and 1912). **Figure 8** shows the predicted probability diagnostics where point 397 and 1912 show high leverage. In addition,[Appendix M](#AppendixM) shows there is not much variance via each explanatory variable.

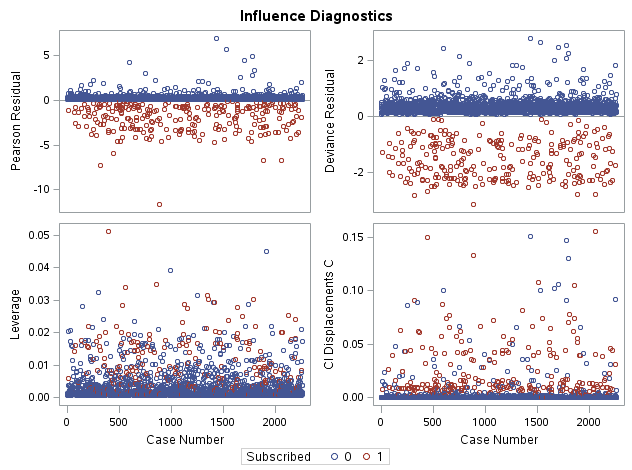
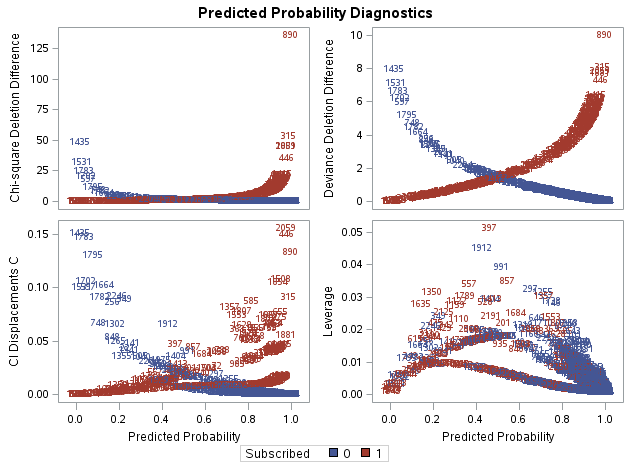
 

Figure 7: Influence Diagnostics Figure 8: Predicted Probability Diagnostics

**Conclusions**

* **Model:** A common method to solicit bank deposit subscriptions was analyzed. Efficient marketing campaign explanatory variables were determined that resulted in a success of a deposit subscription (response of Subscribed and explanatory of duration, loan, contact and poutcome). The final model was significant, resulting in a Likelihood Ratio, Score and Wald tests with Prob. Chi-Square < 0.0001. For the training data, ROC curve had an area of 0.876 out of 1 and for our test data, ROC predicted outcome had an area of 0.870. Finally, the logistic regression model correctly predicted subscription to term deposit 89.6% of the times using a test data set.
* **Recommendations:** Longer duration, contacts via a cellular phone, contacts without an existing loan, and prior campaign success were the most significant variables in predicting a subscription success. It is recommended to expand on techniques to extend the duration of the conversation; to increase contact with people that had cellular phone (more time to talk); to focus on those that do not have a loan with bank (perhaps extra cash to subscribe); and finally, to find more contacts where the prior campaign was successful (there were only contacted 75 individuals with a prior successful “poutcome” in the training set).
* **Future Analysis:** For future analysis, other random samples of the original data set could be analyzed to insure the results are consistent (or perhaps use all 45,000 contacts in the data set). Finally, other confounding variables can be investigated such as specific marketing techniques or benefits that may have been offered to some of the contacts.

**References:**

[1] Reference link data source/previous studies: University of California Irvine (UCI) Machine Learning Repository by S. Moro, P. Rita, and P. Cortez in 2014 <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

[2] Ramsey, Fred; Schafer, Daniel; *The Statistical Sleuth: A Course in Methods of Data Analysis*; Cengage Textbook. Kindle Edition., pp. 549-600.

**APPENDIX A: FULL DATA SET DEFINITIONS**

**Data Definitions (definitions of variables used in analyses)**

Input variables (bank client data):

1 - **age** (numeric)

2 - **job** : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student",

"blue-collar","self-employed","retired","technician","services")

3 - **marital** : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)

4 - **education** (categorical: "unknown","secondary","primary","tertiary")

5 - **default**: has credit in default? (binary: "yes","no")

6 - **balance**: average yearly balance, in euros (numeric)

7 - **housing**: has housing loan? (binary: "yes","no")

8 - **loan**: has personal loan? (binary: "yes","no")

# related with the last contact of the current campaign:

9 - **contact**: contact communication type (categorical: "unknown","telephone","cellular")

10 - **day**: last contact day of the month (numeric)

11 - **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - **duration**: last contact duration, in seconds (numeric)

# other attributes:

13 - **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - **previous**: number of contacts performed before this campaign and for this client (numeric)

16 - **poutcome**: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):

17 - **y** - has the client subscribed a term deposit? (binary: "yes","no")

**APPENDIX B: VARIABLE GROUPING**

A grouping of variables was attempted in order to investigate simplification of the model, or perhaps, find significance by a grouping rather than the original variables.

* The “**Age\_Group**” variable was created from the original “age” variable such that the ages were grouped in the following manner: <20, 20-30, 31-45, 46-65 , 65+.
* The “**Quarter**” variable was created from the original “”month” variable such that the months were grouped by traditional accounting quarters. For example Q1 is January, February, and March.

All associated R code and output may be viewed in **Table 1** – Grouping Variable Attempt

|  |
| --- |
|  |

Table 1 – Grouping Variable Attempt

**APPENDIX C: CREATION OF CROSS-VALIDATION DATA SETS**

**Below is the R code used to create the “train” and “test” data sets.**

###################################################

## Create Training and Test Data R Makefile

## Create by: Kevin Okiah

## James Hosker

## James Park

## SMU Course: MSDS6372-402

## Assignment: Project #3

## Date Created: 6-Aug-2017

## Description: pacman libraries amd saves R session

## into Analysis/Data/sessionInfo.txt

###################################################

#load the required packages using pacman

pacman::p\_load(pacman, plyr, dplyr, utils, ggplot2, readr,

scales, gridExtra, ROCR, gtools, glmnet,

stats, Rcmdr, RcmdrMisc)

###################################################

## Save R Session Info for Reference

###################################################

# write R session info to file for reference

# put all messages from session\_info into file

writeLines(capture.output(sessionInfo()),"Analysis/Data/sessionInfo.txt")

# load data

my\_data<-read.table('bank.csv', sep =";", header = TRUE)

# Get dimensions of my\_data

dim(my\_data)

# Create a new binary column Subscribed based on y

my\_data<-mutate(my\_data, Subscribed=ifelse(my\_data$y=='no',0,1))

my\_data$Subscribed<-as.factor(my\_data$Subscribed) # convert the Subscribed column to factors

# Check structure of the data

str(my\_data)

# Check change

head(my\_data)

# Remove Missing values are marked as unknown. count number of missing values by column

count\_unknown<-function(v){length(v[v=="unknown"])}

unknown <-sapply(my\_data, count\_unknown)

# print unknown

unknown

# creating a new varible agegroup

# agegroups = 19-20, 30-45, 45-65, 65++

my\_data$Age\_Group <- cut(my\_data$age, c(-Inf, 20, 30, 45, 65, Inf))

my\_data$dur\_group <- cut(my\_data$duration, c(-Inf, 200, 400, 600, 800, 1000, Inf))

levels(my\_data$dur\_group) <- c("<200","201-400", "401-600", "601-800", "801-1000", ">1000")

# Name the levels of 'Age\_Group' for readability

levels(my\_data$Age\_Group) <- c("<20","20-30", "31-45", "46-64", "65+")

# Check change

tail(my\_data)

# count by age group

my\_data%>%

group\_by(Age\_Group)%>%

summarise(Count = n())

## Subscribe is 0 for no and 1 for yes

## like proc freq in SAS

Age\_group.Stats <- my\_data%>%

group\_by(Age\_Group, Subscribed) %>%

summarize(Number = n(),

Percent\_Subscribed = 100\*(Number/nrow(my\_data)))

print(data.frame(Age\_group.Stats))

#create a new variable Quarters a derivative of months listing the four quarters in a year

# Quarter = Q1, Q2, Q3, Q4

my\_data<-mutate(my\_data,Quarter=as.factor(ifelse(my\_data$month =='jan'|my\_data$month =='feb'|my\_data$month =='mar','Q1',

ifelse(my\_data$month =='apr'|my\_data$month =='may'|my\_data$month =='jun','Q2',

ifelse(my\_data$month =='jul'|my\_data$month =='aug'|my\_data$month =='sep','Q3',

ifelse(my\_data$month =='oct'|my\_data$month =='nov'|my\_data$month =='dec','Q4','NA'))))))

# Check change

head(my\_data)

# print stats table

Quarter.Stats <- my\_data%>%

group\_by(Quarter, Subscribed) %>%

summarize(Number = n(),

Percent\_Subscribed = 100\*(Number/nrow(my\_data)))

print(data.frame(Quarter.Stats))

#create a test and train dataset but subseting the dataset into two halfs randomly selected for 50-50 cross validation

indexes = sample(1:nrow(my\_data), size=0.5\*nrow(my\_data)) # Random sample of 50% of row numbers created

head(indexes)

indexes

Train <- my\_data[indexes,] # Training data contains created indices

Test <- my\_data[-indexes,] # Test data contains the rest

# Save data used for SAS analysis

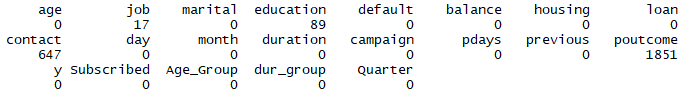
write.csv(my\_data, file = "Analysis/Data/CleanProj3.csv",row.names=FALSE)

write.csv(Train, file = "Analysis/Data/Train.csv",row.names=FALSE)

write.csv(Test, file = "Analysis/Data/Test.csv",row.names=FALSE)

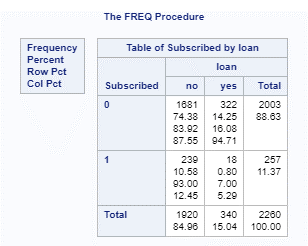
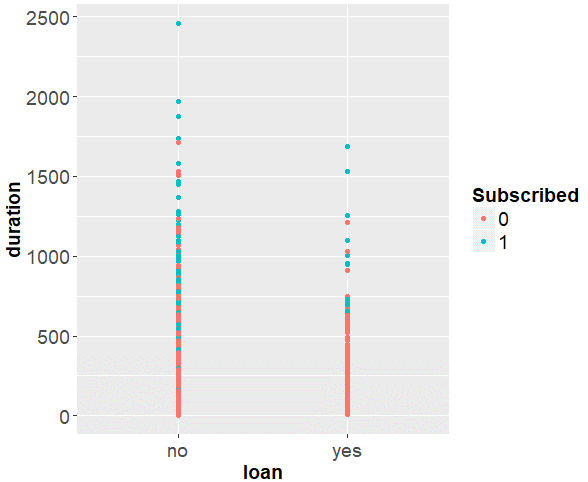
**APPENDIX D: Exploratory Data Analysis (Complete)**

We examined various methods of exploratory data analysis including summary statistics, frequency tables, scatter plots, boxplots and histograms distributions by individual or paired explanatory factors that are highlighted by the “Subscribed” outcome factor. This provides a visual initial understanding of each factors relationship to the subscription outcome. Here, we explore all relevant factors in the Training Data set. As noted in **Table 1**, some explanatory data sets have several unknowns including contact and poutcome. Note for the analysis below, a ***Subscribed*** *value of 1 is a yes to a subscription and a value of 0 is a no.*



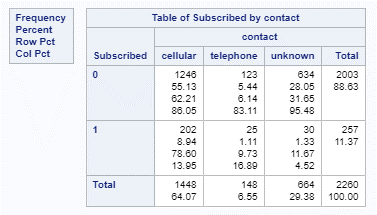
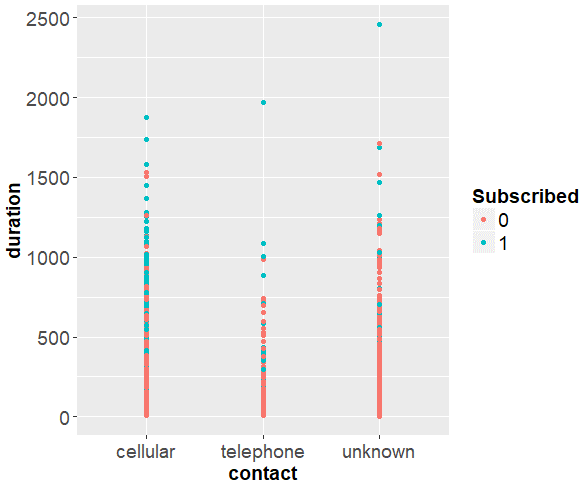
**Table 1.** Tally of “unknown” terms in the data set using R code

* **Loan:** In **Table 2**, a majority of the yeses to a subscription came from those that did not have a personal loan (93%) and only 5.3% of yeses to a subscription came from those that had loans. In **Figure 1**, the scatter plot shows higher subscriptions as the duration of the conversation continued.

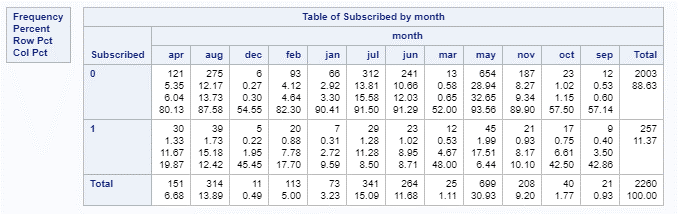
**Table 2: SAS Frequency Table for Loan Figure 1: R Scatter Duration vs. Loan by Subscribed**

* **Contact:** In **Table 3**, a majority of the yeses to a subscription came from those that had a cell phones (78%) compared to telephones or unknown forms of contact. The yeses for a subscription using cellular phones were nearly 9% of all contacts made. In **Figure 2**, the type of contact shows that clearly cellular results in more customers that did subscribed. *We note that he total success rate was low at 11.4%, 257 out of 2260 in our training data set of calls.*

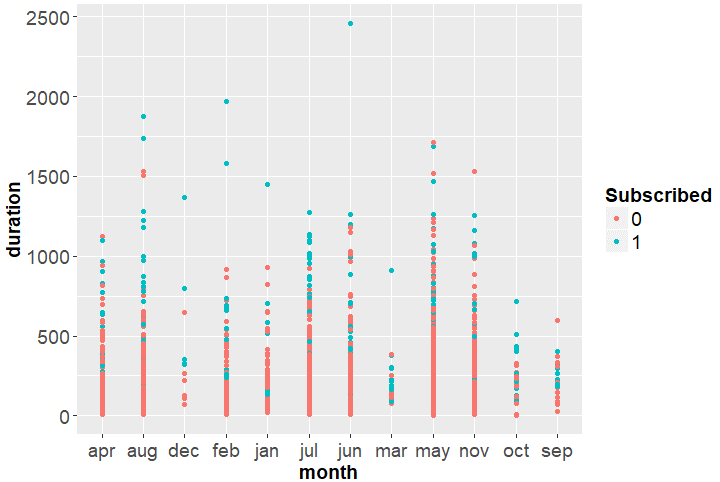
 

**Table 3: SAS Frequency Table for Contact Figure 2: R Scatter Duration vs. Contact by Subscribed**

* **Month:** In **Table 4**, a majority of the yeses to a subscription came in March and December with nearly 48% and 45% of contact, respectively. We note that there were a lower total number of contacts in March (25) and in December (11). However, Sep, Oct and Nov have higher hit ratios of 10%, 42% and 43% of contacts, respectively, resulting in yes to subscription for that month. The most yeses to subscription occurred in Aug, Jul, and May, but the hit ratio is lower due more contacts during that month. In **Figure 3**, yeses to subscriptions occurred in March, October and September with a smaller duration of the conversation. October and September may make sense since banks financial year-end is October or November, so this maybe a result of a year-end push by salespeople perhaps with other confounding variables like better terms, which some banks may do near year-end.

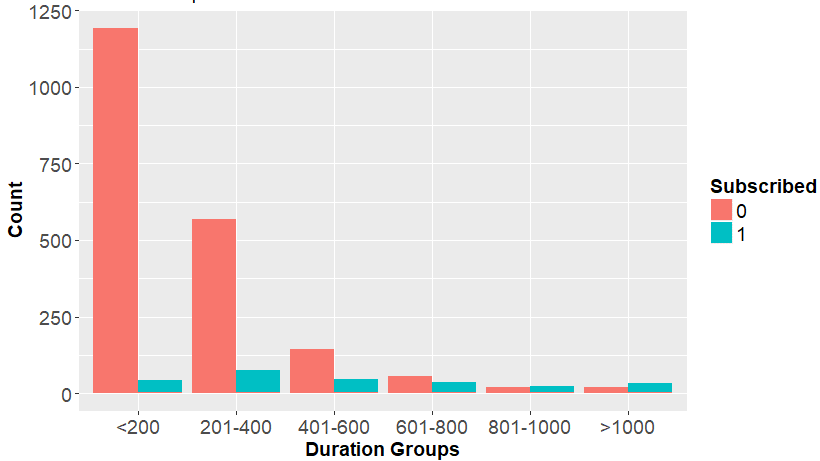
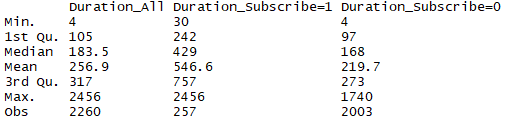


**Table 4: SAS Frequency Table for Month**

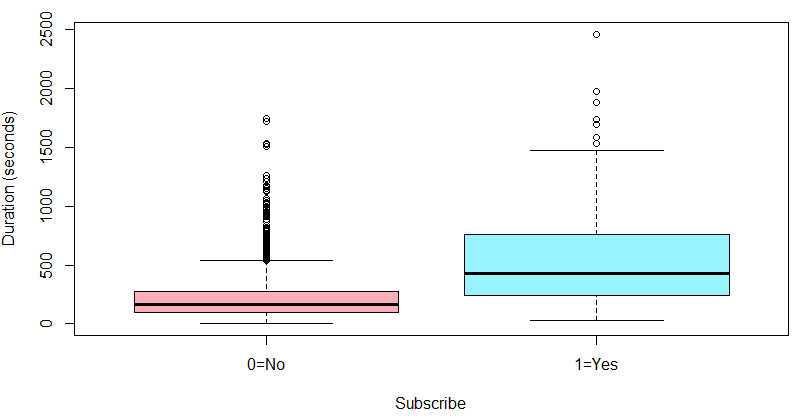
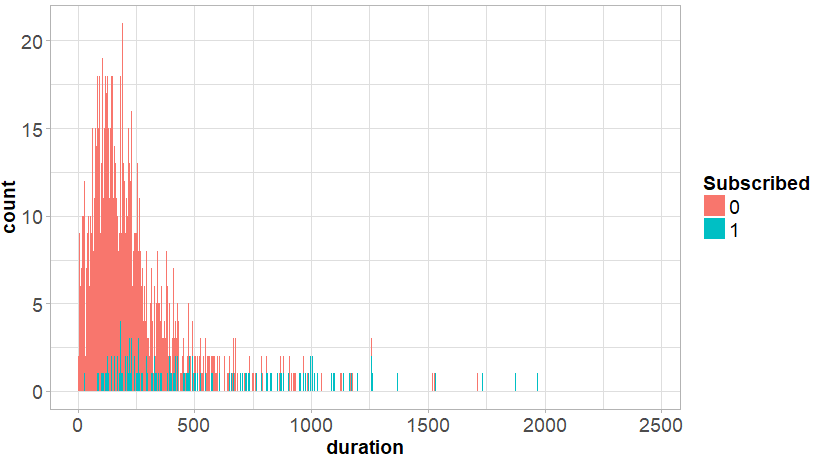


**Figure 3: R Scatter Duration vs. Month by Subscribed**

* **Duration:** **Table 5** and **Figure 5** shows that the summary statistics of the duration of the conversation and the boxplot of Duration vs. Subscribed, respectively. The median and mean seconds on the phone for those that chose to subscribe (mean=546.6, median=429) was much higher than those that did not subscribe (mean=219.7, median=168). **Figure 4** and **Figure 6** shows a graph of the duration group count and the duration histogram, respectively. **Figure 4** shows that greater than 400 seconds has a higher likelihood of a yes to the subscription. The blue bars in the bar charts become equal or higher than the red bars after 800 second of conversation.

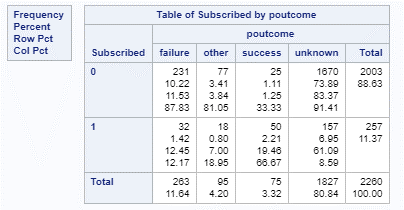
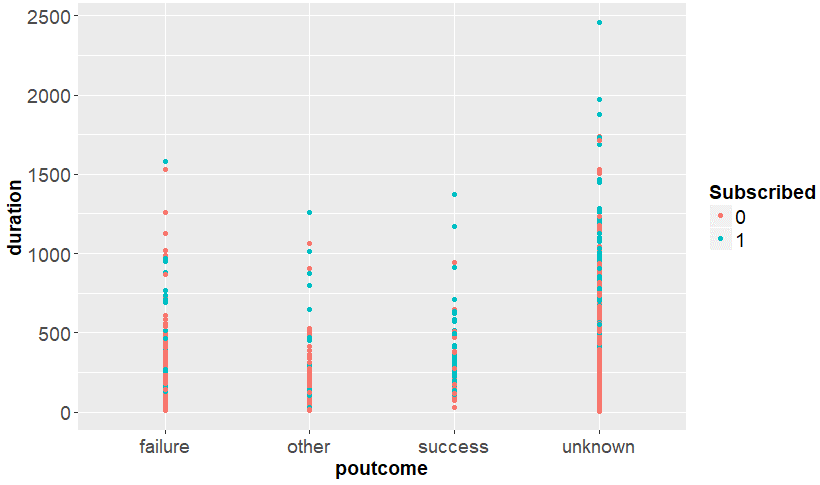


**Table 5: R Summary Stats on Duration Figure 4: Duration Group Count**

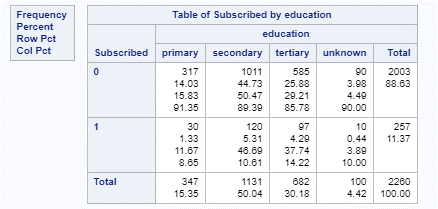
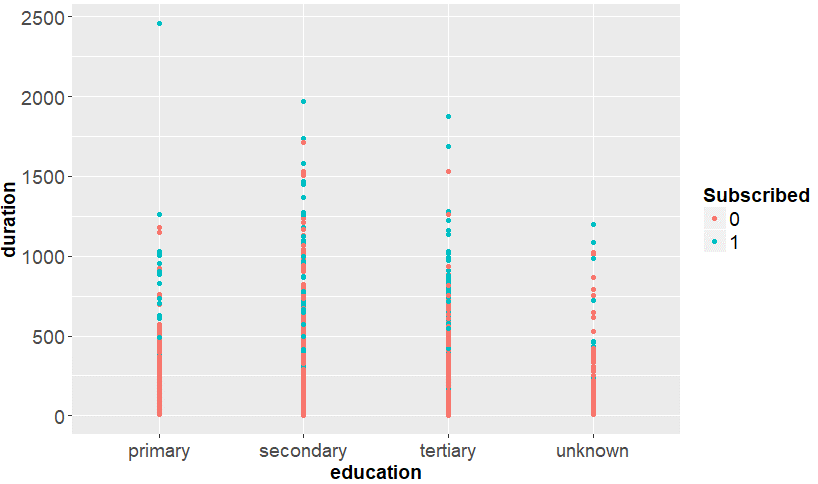
**Figure 5: R Summary Stats on Duration Figure 6: Duration Group Histogram**

* **poutcome:** In **Table 6**, if the prior campaign was successful, the likelihood of a subscription success was 66.7% no matter what the duration of the conversation while failure, unknown and other subscription hit ration were lower. Nearly 19.5% of all subscriptions (50 out of 257 total subscription) came from customers where a prior campaign was successful. We do note that poutcome had many unknowns from **Table 1** but here we are focused on the significance of prior campaign success. Although the unknown poutcome has a 61.1% of all the subscriptions, we note that the success ratio of unknown is much less at 8.6% meaning a successful subscription outcome is less likely for unknowns. **Figure 7** shows similar results.

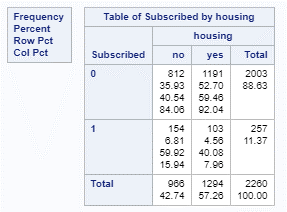
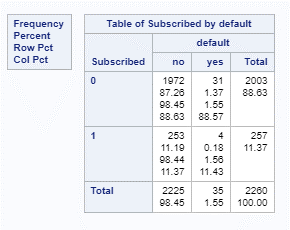
**Table 6: SAS Frequency Table for Month Figure 7: R Scatter Duration vs. poutcome by Subscribed**

* **Education:**  **Table 7** shows the that 14% of all those with a Tertiary education say yes to a subscription. There are 97 yeses from those with a tertiary education, which represents 37.74% of all yeses to a subscription. Those with secondary education represent 46.7% of all yeses to a subscription (120 out of 257) but it is only 10.8% of all those with a secondary education. **Figure 8** shows that more individuals with a secondary and tertiary education elected to subscribe and generally the conversation had a longer duration.

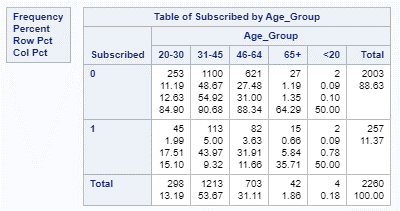
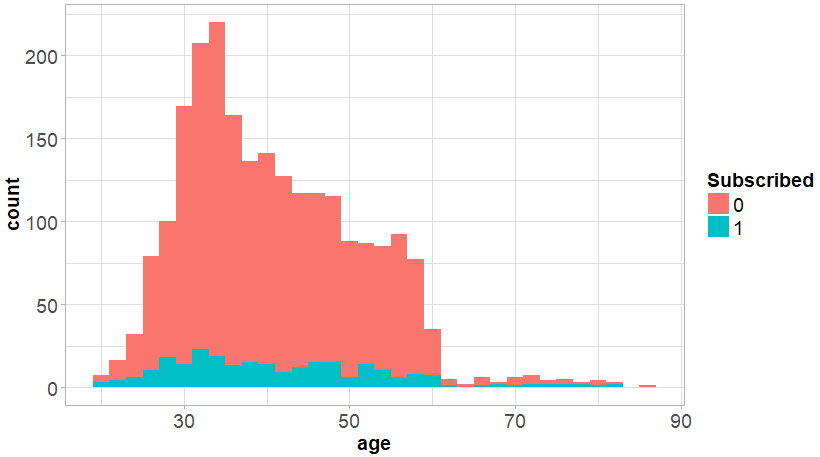
**Table 7: SAS Frequency Table for Education Figure 8: R Scatter Duration vs. education by Subscribed**

* **Default and Housing: Table 8** shows that most yeses to subscription came from those that did not have a current loan in default. In **Table 9**, we can see that there is a 60/40 split between those with and without housing loans, respectively.

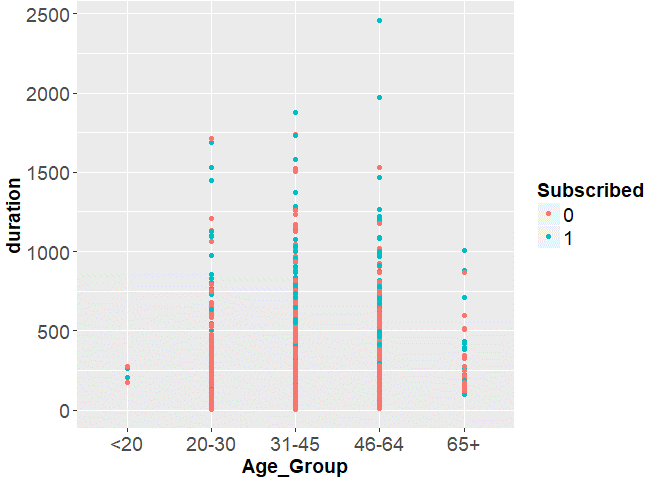


**Table 8: SAS Frequency Table for Education Table 9: SAS Frequency Table for Education**

* **Age: Table 10** shows that most yeses to subscription came from those mostly in the 31-45 (44%) and 46-64 (32%) age groups. **Figure 8** show the histogram of age and Figure 9 scatter plot shows a similar conclusion for the 31-45 and 46-64 age groups.

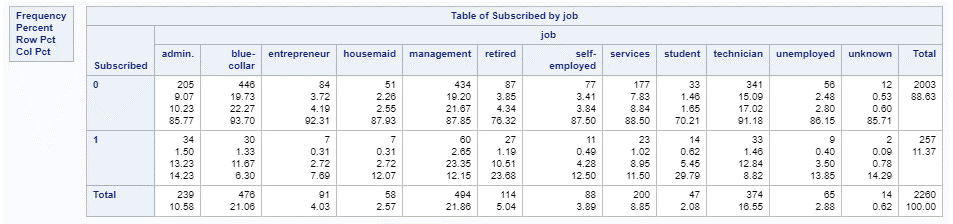
 

**Table 10: SAS Frequency Table for Education Figure 9: R Histogram of Age**

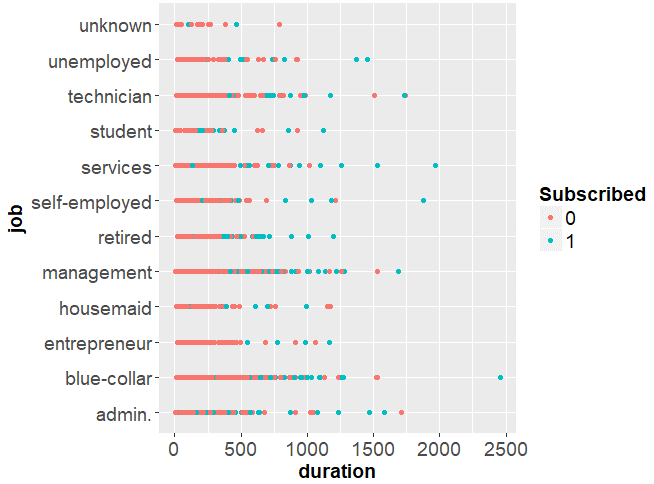
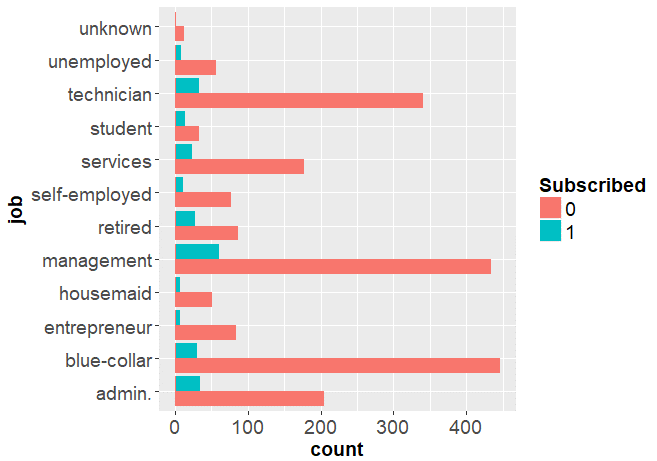


**Figure 10: R Scatter of Duration vs. Age Groups by Subscribed**

* **Job:** Not much looks attractive from type of jobs. In **Table 11** and **Figures 11 and 12**, management and retired seem to have the highest number of yeses within their respective groups.



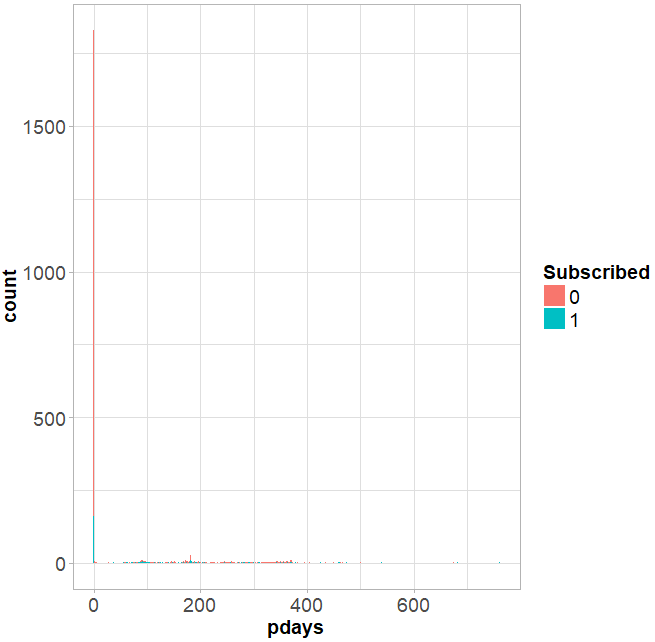
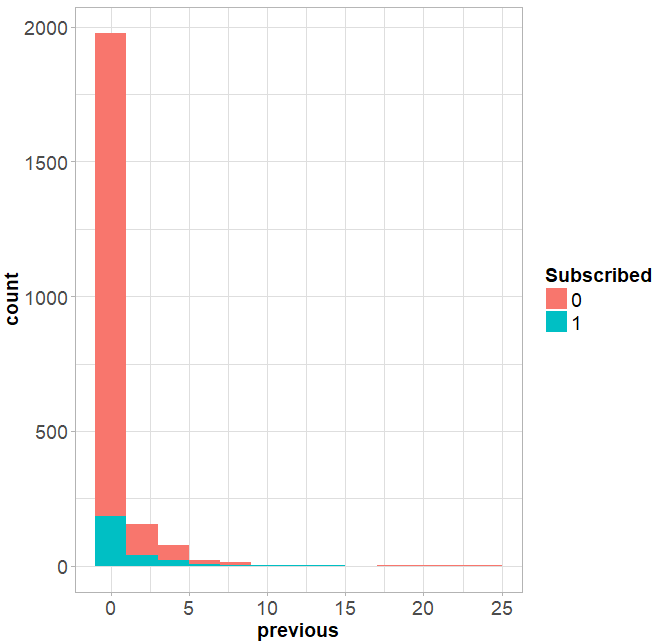
**Table 11: SAS Frequency Table for Jobs**



**Figure 11: R Bar Chart of Jobs vs. Count Figure 12. R Scatter of Duration vs. Jobs**

**By Subscribed By Subscribed**

* **pdays and previous:** In **Figure 13 and 14,** neither pdays (the number of days since last contact) nor previous (number of prior contacts), seem to show anything of interest. Both pdays and previous may not mean much, since zero days for both seems to dominate a majority of positive responses to subscriptions along with the negative responses.

**Figure 13: R Bar Chart of Pdays vs. Count Figure 14. R Bar Chart of Previous vs.Count**

**By Subscribed By Subscribed**

**APPENDIX E: BACKUP DATA EXPLORATION**

Several variables were initially explored, but ended up not being included into the model due to lack of statistical significance. However, the output may be interesting for future studies and also helps to visually understand various associations. Included in this section are Age/Count histograms, an Age\_Group histogram, Quarter/Count histogram, Count/Job bar chart, and Education/Count histogram.

|  |  |
| --- | --- |
| Age/Count  Histogram |  |
| Age\_Group Histogram |  |
| Quarter/ Count Histogram |  |
| Job/Count  Bar chart |  |
| Education/ Count Histogram |  |

**APPENDIX F: DATA CLEANING**

**Data Landscape (tallies, basic statistics, R code and output of the associated data)**

|  |
| --- |
|  |
| **Creating new variable Age\_Group** |

**APPENDIX G: TABLE OF FULL LOGISTIC REGRESSION**

Call:

glm(formula = Subscribed ~ Age\_Group + job + education + default +

housing + marital + balance + loan + contact + day + month +

duration + campaign + poutcome + pdays + previous, family = binomial(),

data = Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.8557 -0.3613 -0.2276 -0.1293 3.0431

Coefficients:

Estimate Std. Error z value Pr(>|z|)

**(Intercept) -5.072e+00 1.730e+00 -2.932 0.003363 \*\***

Age\_Group20-30 -1.494e+00 1.155e+00 -1.293 0.195944

Age\_Group31-45 -1.864e+00 1.185e+00 -1.573 0.115648

Age\_Group46-64 -1.546e+00 1.197e+00 -1.292 0.196510

Age\_Group65+ -2.069e+00 1.301e+00 -1.591 0.111597

jobadmin. -1.413e-02 9.376e-01 -0.015 0.987975

jobblue-collar -9.117e-01 9.418e-01 -0.968 0.333042

jobentrepreneur -6.723e-01 1.038e+00 -0.648 0.517208

jobhousemaid -6.673e-01 1.052e+00 -0.634 0.525855

jobmanagement -5.986e-01 9.277e-01 -0.645 0.518746

jobretired 3.519e-01 9.646e-01 0.365 0.715253

jobself-employed -4.869e-01 1.005e+00 -0.485 0.627997

jobservices -8.491e-02 9.534e-01 -0.089 0.929036

jobstudent 1.008e-01 1.019e+00 0.099 0.921251

jobtechnician -6.484e-01 9.297e-01 -0.697 0.485539

jobunemployed -8.596e-01 1.036e+00 -0.830 0.406728

educationprimary 1.003e+00 5.451e-01 1.839 0.065861 .

educationsecondary 8.744e-01 4.997e-01 1.750 0.080133 .

**educationtertiary 1.402e+00 5.245e-01 2.673 0.007507 \*\***

defaultyes 9.980e-01 6.581e-01 1.516 0.129416

**housingyes -4.249e-01 2.065e-01 -2.058 0.039584 \***

maritalmarried -4.047e-01 2.608e-01 -1.552 0.120677

maritalsingle -1.444e-01 3.125e-01 -0.462 0.644124

balance -2.647e-05 2.820e-05 -0.939 0.347918

**loanno 7.839e-01 3.144e-01 2.494 0.012643 \***

**contactcellular 1.340e+00 3.467e-01 3.866 0.000111 \*\*\***

**contacttelephone 1.558e+00 4.666e-01 3.339 0.000839 \*\*\***

day 1.170e-02 1.184e-02 0.989 0.322831

**monthapr 1.383e+00 5.661e-01 2.444 0.014536 \***

monthaug 9.071e-01 5.729e-01 1.583 0.113367

monthdec 1.727e+00 1.057e+00 1.634 0.102264

**monthfeb 1.238e+00 6.280e-01 1.971 0.048713 \***

monthjul 5.756e-01 5.681e-01 1.013 0.310995

**monthjun 1.455e+00 6.533e-01 2.227 0.025978 \***

**monthmar 2.718e+00 7.286e-01 3.731 0.000191 \*\*\***

monthmay 7.868e-01 5.739e-01 1.371 0.170353

monthnov 4.238e-01 5.857e-01 0.724 0.469257

**monthoct 2.534e+00 6.494e-01 3.902 9.56e-05 \*\*\***

**monthsep 1.830e+00 7.639e-01 2.395 0.016617 \***

**duration 4.762e-03 3.141e-04 15.158 < 2e-16 \*\*\***

campaign -1.023e-01 4.755e-02 -2.152 0.031404 \*

poutcomefailure 4.208e-01 4.499e-01 0.935 0.349613

**poutcomeother 1.121e+00 4.894e-01 2.291 0.021953 \***

**poutcomesuccess 2.912e+00 4.509e-01 6.459 1.06e-10 \*\*\***

pdays -4.244e-04 1.380e-03 -0.308 0.758451

previous -5.263e-02 4.982e-02 -1.056 0.290809

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

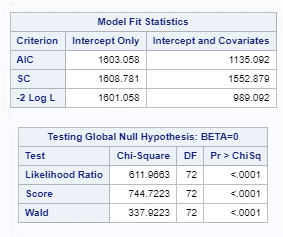
Null deviance: 1601.1 on 2259 degrees of freedom

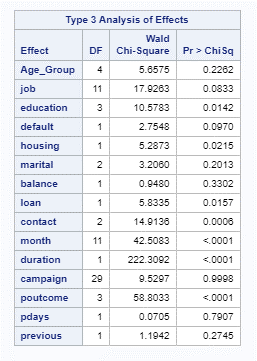
Residual deviance: 1000.1 on 2214 degrees of freedom

AIC: 1092.1

Number of Fisher Scoring iterations: 6

**Table 1: R-Logistic Regression of Full Model**





| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** |  | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | 1 | 7.2310 | 167.9 | 0.0019 | 0.9656 |
| **Age\_Group** | **20-30** | 1 | 1.4994 | 1.1657 | 1.6543 | 0.1984 |
| **Age\_Group** | **31-45** | 1 | 1.8729 | 1.1961 | 2.4520 | 0.1174 |
| **Age\_Group** | **46-64** | 1 | 1.5561 | 1.2094 | 1.6556 | 0.1982 |
| **Age\_Group** | **65+** | 1 | 2.0040 | 1.3127 | 2.3306 | 0.1269 |
| **job** | **admin.** | 1 | 0.3133 | 0.9757 | 0.1031 | 0.7481 |
| **job** | **blue-collar** | 1 | 1.2417 | 0.9807 | 1.6032 | 0.2055 |
| **job** | **entrepreneur** | 1 | 0.9608 | 1.0710 | 0.8049 | 0.3696 |
| **job** | **housemaid** | 1 | 0.9180 | 1.0832 | 0.7182 | 0.3967 |
| **job** | **management** | 1 | 0.8961 | 0.9710 | 0.8516 | 0.3561 |
| **job** | **retired** | 1 | -0.0312 | 0.9996 | 0.0010 | 0.9751 |
| **job** | **self-employed** | 1 | 0.7550 | 1.0439 | 0.5230 | 0.4696 |
| **job** | **services** | 1 | 0.3474 | 0.9916 | 0.1228 | 0.7261 |
| **job** | **student** | 1 | 0.1942 | 1.0562 | 0.0338 | 0.8541 |
| **job** | **technician** | 1 | 0.9697 | 0.9712 | 0.9968 | 0.3181 |
| **job** | **unemployed** | 1 | 1.2243 | 1.0747 | 1.2978 | 0.2546 |
| **education** | **primary** | 1 | -1.2072 | 0.5641 | 4.5795 | 0.0324 |
| **education** | **secondary** | 1 | -1.0787 | 0.5196 | 4.3102 | 0.0379 |
| **education** | **tertiary** | 1 | -1.6188 | 0.5462 | 8.7846 | 0.0030 |
| **default** | **no** | 1 | 1.0746 | 0.6474 | 2.7548 | 0.0970 |
| **housing** | **no** | 1 | -0.4790 | 0.2083 | 5.2873 | 0.0215 |
| **marital** | **divorced** | 1 | -0.1618 | 0.3138 | 0.2659 | 0.6061 |
| **marital** | **married** | 1 | 0.2578 | 0.2312 | 1.2432 | 0.2649 |
| **balance** |  | 1 | 0.000027 | 0.000027 | 0.9480 | 0.3302 |
| **loan** | **no** | 1 | -0.7575 | 0.3136 | 5.8335 | 0.0157 |
| **contact** | **cellular** | 1 | -1.3056 | 0.3447 | 14.3461 | 0.0002 |
| **contact** | **telephone** | 1 | -1.4667 | 0.4712 | 9.6871 | 0.0019 |
| **month** | **apr** | 1 | 0.3777 | 0.6334 | 0.3557 | 0.5509 |
| **month** | **aug** | 1 | 0.8747 | 0.6346 | 1.8996 | 0.1681 |
| **month** | **dec** | 1 | -0.0107 | 1.0553 | 0.0001 | 0.9919 |
| **month** | **feb** | 1 | 0.6744 | 0.6505 | 1.0748 | 0.2999 |
| **month** | **jan** | 1 | 1.6567 | 0.7529 | 4.8419 | 0.0278 |
| **month** | **jul** | 1 | 1.2543 | 0.6417 | 3.8210 | 0.0506 |
| **month** | **jun** | 1 | 0.3972 | 0.6821 | 0.3390 | 0.5604 |
| **month** | **mar** | 1 | -1.0335 | 0.7682 | 1.8099 | 0.1785 |
| **month** | **may** | 1 | 1.0144 | 0.6270 | 2.6173 | 0.1057 |
| **month** | **nov** | 1 | 1.3564 | 0.6444 | 4.4302 | 0.0353 |
| **month** | **oct** | 1 | -0.7446 | 0.7016 | 1.1263 | 0.2886 |
| **duration** |  | 1 | -0.00475 | 0.000319 | 222.3092 | <.0001 |
| **campaign** | **1** | 1 | -4.7572 | 167.9 | 0.0008 | 0.9774 |
| **campaign** | **2** | 1 | -4.5290 | 167.9 | 0.0007 | 0.9785 |
| **campaign** | **3** | 1 | -4.1267 | 167.9 | 0.0006 | 0.9804 |
| **campaign** | **4** | 1 | -4.5949 | 167.9 | 0.0007 | 0.9782 |
| **campaign** | **5** | 1 | -4.1410 | 167.9 | 0.0006 | 0.9803 |
| **campaign** | **6** | 1 | -3.8906 | 167.9 | 0.0005 | 0.9815 |
| **campaign** | **7** | 1 | -4.0585 | 167.9 | 0.0006 | 0.9807 |
| **campaign** | **8** | 1 | -4.4000 | 167.9 | 0.0007 | 0.9791 |
| **campaign** | **9** | 1 | 3.4190 | 171.8 | 0.0004 | 0.9841 |
| **campaign** | **10** | 1 | 2.8583 | 172.0 | 0.0003 | 0.9867 |
| **campaign** | **11** | 1 | 2.0098 | 172.5 | 0.0001 | 0.9907 |
| **campaign** | **12** | 1 | -4.0770 | 167.9 | 0.0006 | 0.9806 |
| **campaign** | **13** | 1 | -5.5565 | 167.9 | 0.0011 | 0.9736 |
| **campaign** | **14** | 1 | 1.9183 | 179.6 | 0.0001 | 0.9915 |
| **campaign** | **15** | 1 | 1.9725 | 178.9 | 0.0001 | 0.9912 |
| **campaign** | **16** | 1 | 1.8850 | 176.3 | 0.0001 | 0.9915 |
| **campaign** | **17** | 1 | -6.4500 | 167.9 | 0.0015 | 0.9693 |
| **campaign** | **18** | 1 | 1.0137 | 192.3 | 0.0000 | 0.9958 |
| **campaign** | **19** | 1 | 2.2086 | 188.3 | 0.0001 | 0.9906 |
| **campaign** | **20** | 1 | 1.9764 | 237.4 | 0.0001 | 0.9934 |
| **campaign** | **21** | 1 | 1.2920 | 205.1 | 0.0000 | 0.9950 |
| **campaign** | **22** | 1 | 2.6020 | 237.4 | 0.0001 | 0.9913 |
| **campaign** | **23** | 1 | 3.4560 | 237.4 | 0.0002 | 0.9884 |
| **campaign** | **24** | 1 | 0.4127 | 237.4 | 0.0000 | 0.9986 |
| **campaign** | **25** | 1 | 0.0342 | 237.4 | 0.0000 | 0.9999 |
| **campaign** | **28** | 1 | 2.1004 | 205.4 | 0.0001 | 0.9918 |
| **campaign** | **30** | 1 | 0.6093 | 237.4 | 0.0000 | 0.9980 |
| **campaign** | **32** | 1 | 1.2786 | 237.4 | 0.0000 | 0.9957 |
| **campaign** | **44** | 1 | 0.9769 | 237.4 | 0.0000 | 0.9967 |
| **poutcome** | **failure** | 1 | -0.3867 | 0.4494 | 0.7403 | 0.3896 |
| **poutcome** | **other** | 1 | -1.1379 | 0.4886 | 5.4226 | 0.0199 |
| **poutcome** | **success** | 1 | -2.9126 | 0.4536 | 41.2263 | <.0001 |
| **pdays** |  | 1 | 0.000365 | 0.00137 | 0.0705 | 0.7907 |
| **previous** |  | 1 | 0.0542 | 0.0496 | 1.1942 | 0.2745 |

**Table 2: SAS Output of Full Logistic Model**

**APPENDIX H: INTERIM LOGISTIC REGRESSION**

Call:

glm(formula = Subscribed ~ loan + contact + month + duration +

poutcome + education, family = binomial(), data = Train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.9000 -0.3686 -0.2609 -0.1541 3.0338

Coefficients:

Estimate Std. Error z value Pr(>|z|)

**(Intercept) -7.2611011 0.8171062 -8.886 < 2e-16 \*\*\***

**loanno 0.7870394 0.3045588 2.584 0.00976 \*\***

**contactcellular 1.4276998 0.3338892 4.276 1.9e-05 \*\*\***

**contacttelephone 1.7263598 0.4310605 4.005 6.2e-05 \*\*\***

**monthapr 1.0991040 0.5388091 2.040 0.04136 \***

monthaug 0.6665064 0.5323414 1.252 0.21056

monthdec 1.3504700 1.0279351 1.314 0.18892

monthfeb 0.9768797 0.5677560 1.721 0.08532 .

monthjul 0.3700955 0.5378193 0.688 0.49136

**monthjun 1.2573875 0.5937512 2.118 0.03420 \***

**monthmar 2.7138253 0.6769558 4.009 6.1e-05 \*\*\***

monthmay 0.4257985 0.5281102 0.806 0.42009

monthnov 0.1391995 0.5584086 0.249 0.80314

**monthoct 2.5103275 0.6210619 4.042 5.3e-05 \*\*\***

**monthsep 1.9328806 0.7329826 2.637 0.00836 \*\***

**duration 0.0045242 0.0002953 15.321 < 2e-16 \*\*\***

poutcomefailure 0.0929458 0.2613966 0.356 0.72216

**poutcomeother 0.6983769 0.3378245 2.067 0.03871 \***

**poutcomesuccess 2.6386904 0.3074052 8.584 < 2e-16 \*\*\***

**educationprimary 0.6170624 0.4940596 1.249 0.21168**

**educationsecondary 0.6978607 0.4551827 1.533 0.12524**

**educationtertiary 1.0389227 0.4614937 2.251 0.06437**

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

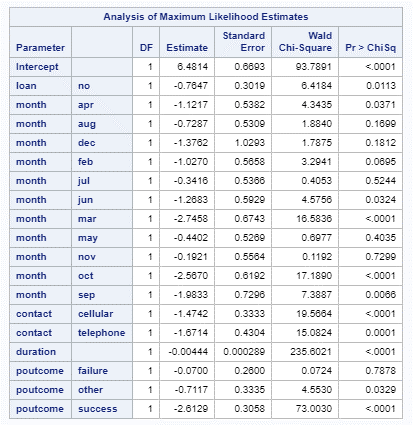
Null deviance: 1601.1 on 2259 degrees of freedom

Residual deviance: 1054.0 on 2238 degrees of freedom

AIC: 1098

Number of Fisher Scoring iterations: 6

**Table 1: R Output of Next Iteration of Logistic Model with Month and Education**



Ref = “Jan”

**Table 2: SAS Output of Next Iteration of Logistic Model with Month (ref=”Jan”)**

**APPENDIX I: R-CODE FOR MODEL ANALYSIS**

###################################################

## Banking\_Logistic.R is a R Makefile for

## Create by: Kevin Okiah

## James Hosker

## James Park

## SMU Course: MSDS6372-402

## Assignment: Project #3

## Date Created: 6-Aug-2017

## Description: pacman libraries amd saves R session

## into Analysis/Data/sessionInfo.txt

###################################################

#load the required packages using pacman

pacman::p\_load(pacman, plyr, dplyr, utils, ggplot2, readr,

scales, gridExtra, ROCR, gtools, glmnet,

stats, Rcmdr, RcmdrMisc)

###################################################

## Save R Session Info for Reference

###################################################

## write R session info to file for reference

## put all messages from session\_info into file

writeLines(capture.output(sessionInfo()),"Analysis/Data/sessionInfo.txt")

## my\_data read in Training Data

my\_data <- read.csv(file = "Analysis/Data/TrainFinal.csv", header=TRUE, sep=',')

my\_data$Subscribed<-as.factor(my\_data$Subscribed) # convert the Subscribed column to factors

Train <- read.csv(file = "Analysis/Data/TrainFinal.csv", header=TRUE, sep=',')

Test <- read.csv(file = "Analysis/Data/TestFinal.csv", header=TRUE, sep=',')

Train$Subscribed<-as.factor(Train$Subscribed) # convert the Subscribed column to factors

Test$Subscribed<-as.factor(Test$Subscribed) # convert the Subscribed column to factors

# Remove Missing values are marked as unknown. count number of missing values by column

count\_unknown<-function(v){length(v[v=="unknown"])}

unknown <-sapply(Train, count\_unknown)

# print unknown

unknown

# logistic model on training data

Train$contact <- relevel(Train$contact, ref='unknown')

Train$poutcome <- relevel(Train$poutcome, ref='unknown')

Train$job <- relevel(Train$job, ref='unknown')

Train$education <- relevel(Train$education, ref='unknown')

Train$loan <- relevel(Train$loan, ref='yes')

Train$default <- relevel(Train$default, ref='no')

Train$month <- relevel(Train$month, ref='jan')

##############################################

####### All Variables Logistic ##############

##############################################

LogisticModelFull\_train <- glm(Subscribed~Age\_Group+job+education+default+housing+

marital+loan+contact+day+month+duration+

campaign+poutcome+pdays+previous,

family=binomial(), data =Train)

summary(LogisticModelFull\_train)

##############################################

####### Reduced Variables Logistic ##########

##############################################

# logistic regression Final model on training data with reduced variables

# Z-score <2 removed since they are not significant

LogisticModelFinal\_train <- glm(Subscribed~loan+contact+month+duration+poutcome+education,

family=binomial(), data =Train)

summary(LogisticModelFinal\_train)

##############################################

####### Final Logistic Model ##########

##############################################

LogisticModelFinal\_train <- glm(Subscribed~loan+contact+duration+poutcome,

family=binomial(), data =Train)

## housing and default

summary(LogisticModelFinal\_train)

##############################################

################ Odds Ratio ##################

##############################################

exp(cbind(OddsRatio=coef(LogisticModelFinal\_train),confint(LogisticModelFinal\_train)))

##############################################

################ ALL Plots ###################

##############################################

## scatter of Loan vs. duration

ggplot(Train, aes(loan, duration, color = Subscribed)) + geom\_point() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

# Loan Group; position = "dodge" preserves total width of bars

Train <-mutate(Train, numloan=ifelse(Train$loan=='no',0,1))

ggplot(Train, aes(numloan, fill = Subscribed)) +

geom\_histogram(binwidth = 2)+ theme\_light() +

labs(title="Histogram of Loan Distribution") +

theme(plot.title = element\_text(hjust = 0.5),

axis.title.x = element\_text(size=20,face="bold"))

## scatter of contact vs. duration

ggplot(Train, aes(contact, duration, color = Subscribed)) + geom\_point() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## scatter of month vs. duration

ggplot(Train, aes(month, duration, color = Subscribed)) + geom\_point() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## scatter of poutcome vs. duration

ggplot(Train, aes(poutcome, duration, color = Subscribed)) + geom\_point() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## scatter of housing vs. duration

ggplot(Train, aes(housing, duration, color = Subscribed)) + geom\_point() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## scatter of default vs. duration

ggplot(Train, aes(default, duration, color = Subscribed)) + geom\_point() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## scatter of education vs. duration

ggplot(Train, aes(education, duration, color = Subscribed)) + geom\_point() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## scatter of job vs. Subscribed Count

ggplot(my\_data, aes(x=job,fill=Subscribed)) +

geom\_bar(position = "dodge")+stat\_count(mapping=aes(x=job,y=..prop..))+

labs(title = "Subscription by job")+ coord\_flip() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## scatter of job vs. duration

ggplot(Train, aes(job, duration, color = Subscribed)) + geom\_point() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold")) +

coord\_flip()

# histrogram of age distrubions color coded by subscription/y

ggplot(Train, aes(age, fill = Subscribed)) +

geom\_histogram(binwidth = 2)+ theme\_light() +

labs(title="Histogram of Age distribution") +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

# Age groups Spread; position = "dodge" preserves total width of bars

## scatter of education vs. duration

ggplot(Train, aes(Age\_Group, duration, color = Subscribed)) + geom\_point() +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## Boxplot of Age Group

## boxplot(Train[Train[,"Age\_Group"]=="<20","duration"],Train[Train[,"Age\_Group"]=="20-30","duration"],

## names=c("0=No", "1=Yes"),xlab="Subscribe", ylab = "Age",

## col=c("lightpink1","cadetblue1"))

# histrogram of pdays coded by subscription/y

ggplot(Train, aes(pdays, fill = Subscribed)) +

geom\_histogram(binwidth = 2)+ theme\_light() +

labs(title="Histogram of pdays distribution") +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

# histrogram of previous coded by subscription/y

ggplot(Train, aes(previous, fill = Subscribed)) +

geom\_histogram(binwidth = 2)+ theme\_light() +

labs(title="Histogram of previous distribution") +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

#######################################

## Summary Stat Table for duration ##

#######################################

t1 <- summary(Train$duration)

t1$Obs <- nrow(Train)

t2 <- summary(Train[Train[,"Subscribed"]==1,"duration"])

t2$Obs <- nrow(Train[Train[,"Subscribed"]==1,])

t3 <- summary(Train[Train[,"Subscribed"]==0,"duration"])

t3$Obs <- nrow(Train[Train[,"Subscribed"]==0,])

sumstats <- cbind(t1, t2, t3)

colnames(sumstats) <- c("Duration\_All", "Duration\_Subscribe=1","Duration\_Subscribe=0")

# print summary stat table

print(sumstats)

## crete bar chart of duration groups vs. count

dur\_group.Stats <- train%>%

group\_by(dur\_group, Subscribed) %>%

summarize(Number = n(),

Percent\_Subscribed = 100\*(Number/nrow(Train)))

## print(data.frame(dur\_group.Stats))

Train$dur\_group <- factor(Train$dur\_group, levels = c("<200","201-400", "401-600", "601-800", "801-1000", ">1000"))

ggplot(Train, aes(x=dur\_group,fill=Subscribed)) +

geom\_bar(position = "dodge")+stat\_count(mapping=aes(x=dur\_group,y=..prop..))+

labs(title = "Duration Group Count") + labs(x="Duration Groups", y="Count") +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## Histogram of duration

ggplot(Train, aes(duration, fill = Subscribed)) +

geom\_histogram(binwidth = 2)+ theme\_light() +

labs(title="Histogram of Duration Distribution") +

theme(axis.title.x = element\_text(size=14,face="bold"),

axis.title.y = element\_text(size=14,face="bold"),

axis.text.x = element\_text(size=14,face="plain"),

axis.text.y = element\_text(size=14,face="plain"),

legend.text=element\_text(size=14,face="plain"),

legend.title = element\_text(size= 14,face="bold"))

## Boxplot of duration

boxplot(Train[Train[,"Subscribed"]==0,"duration"],Train[Train[,"Subscribed"]==1,"duration"],

names=c("0=No", "1=Yes"),xlab="Subscribe", ylab = "Duration (seconds)",

col=c("lightpink1","cadetblue1"))

############################################

######### 95% Confidence Intervals ########

######### and Odds Ratio ########

############################################

confint(LogisticModelFinal\_train) # 95% CI for the coefficients

exp(coef(LogisticModelFinal\_train)) # exponentiated coefficients

exp(confint(LogisticModelFinal\_train)) # 95% CI for the coefficients

# Anovamodel

anova(LogisticModelFinal\_train,test="Chisq")

############################################

######### Plot ROC ########

############################################

# plotting the ROC plot

pacman::p\_load(ROCR)

fitted.results <-predict(LogisticModelFinal\_train,Train,type='response')

pred <- prediction(fitted.results, Train$Subscribed)

roc.perf <- performance(pred, measure = "tpr", x.measure = "fpr")

auc.train <- performance(pred, measure = "auc")

auc.train <- auc.train@y.values

# Plot ROC

plot(roc.perf)

abline(a=0, b= 1)

text(x = .40, y = .6,paste("AUC = ", round(auc.train[[1]],3), sep = ""))

# Assessing the predictive ability of the model

fitted.probs <-predict(LogisticModelFinal\_train, Test,type='response')

fitted.predictions <- ifelse(fitted.results > 0.5,1,0) # significance level is 0.5

# check results of classfication

misClasificError <- mean(fitted.predictions != Test$Subscribed)

table(fitted.predictions,Test$Subscribed)

print(paste('Model Accuracy is: ',1-misClasificError))

pred <- prediction(fitted.results, Test$Subscribed)

roc.perf <- performance(pred, measure = "tpr", x.measure = "fpr")

auc.train <- performance(pred, measure = "auc")

auc.train <- auc.train@y.values

# Plot ROC

plot(roc.perf)

abline(a=0, b= 1)

text(x = .40, y = .6,paste("AUC = ", round(auc.train[[1]],3), sep = ""))

# Save data used for SAS analysis

write.csv(my\_data, file = "Analysis/Data/CleanProj3.csv",row.names=FALSE)

**APPENDIX J: ADDITIONAL R-CODE FOR MODEL ANALYSIS**

###########################################

# Assessing the predictive ability of the model ROC #

###########################################

Test<-read.table('Test.csv', sep =",", header = TRUE)

pacman::p\_load(ROCR)

fitted.results <-predict(LogisticModelFinal\_train,Train,type='response')

fitted2.results <-predict(LogisticModelFinal\_train,Test,type='response')

pred <- prediction(fitted.results, Train$Subscribed)

pred2<-prediction(fitted2.results, Test$Subscribed)

roc.perf <- performance(pred, measure = "tpr", x.measure = "fpr")

roc2.perf <- performance(pred2, measure = "tpr", x.measure = "fpr")

x = cbind(roc.perf,roc2.perf)

auc.train <- performance(pred, measure = "auc")

auc.train <- auc.train@y.values

auc.test <- performance(pred2, measure = "auc")

auc.test <- auc.test@y.values

#Plot ROC

plot(roc.perf)

plot(roc2.perf, add = TRUE, col = 'red')

abline(a=0, b= 1)

text(x = .40, y = .6,paste("AUC Train = ", round(auc.train[[1]],3), sep = ""))

text(x = .39, y = .53,paste("AUC Test = ", round(auc.test[[1]],3), sep = ""), col = 'red')

####################################

**APPENDIX K: SAS FOR MODEL ANALYSIS**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Project #3

\* Course: MSDS 6372-4033 (Monday, 8:30 PM)

\* Source Code by: James Hosker

\* Kevin Okiah

\* James Park

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Generated Code (IMPORT) \*/

/\* Source File: CleanProj3.csv \*/

/\* Source Path: /home/jhosker0/sasuser.v94/MSDS6372/Project3 \*/

/\* Code generated on: 8/8/17, 5:04 PM \*/

%web\_drop\_table(WORK.IMPORT);

FILENAME REFFILE '/home/jhosker0/sasuser.v94/MSDS6372/Project3/CleanProj3.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.IMPORT;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

%web\_open\_table(WORK.IMPORT);

%web\_drop\_table(WORK.IMPORT1);

FILENAME REFFILE '/home/jhosker0/sasuser.v94/MSDS6372/Project3/TrainFinal.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.IMPORT1;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT1; RUN;

%web\_open\_table(WORK.IMPORT1);

%web\_drop\_table(WORK.IMPORT2);

FILENAME REFFILE '/home/jhosker0/sasuser.v94/MSDS6372/Project3/TestFinal.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.IMPORT2;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT2; RUN;

%web\_open\_table(WORK.IMPORT2);

data bankdata; set WORK.IMPORT;

if loan = 'yes' then numloan = 1;

else if loan = 'no' then numloan = 0;

else numloan = -1;

if default = 'yes' then numdefault = 1;

else if default = 'no' then numdefault = 0;

else numdefault = -1;

if housing = 'yes' then numhousing = 1;

else if housing = 'no' then numhousing = 0;

else numhousing = -1;

if contact = 'telephone' then numcontact = 1;

else if contact = 'cellular' then numcontact = 0;

else numcontact = -1;

if poutcome = 'success' then numpoutcome = 2;

else if poutcome = 'other' then numpoutcome = 1;

else if poutcome = 'failure' then numpoutcome = 0;

else numpoutcome = -1;

if month = 'jan' then nummonth = 1;

else if month = 'feb' then nummonth = 2;

else if month = 'mar' then nummonth = 3;

else if month = 'apr' then nummonth = 4;

else if month = 'may' then nummonth = 5;

else if month = 'jun' then nummonth = 6;

else if month = 'jul' then nummonth = 7;

else if month = 'aug' then nummonth = 8;

else if month = 'sep' then nummonth = 9;

else if month = 'oct' then nummonth = 10;

else if month = 'nov' then nummonth = 11;

else if month = 'dec' then nummonth = 12;

if Subscribed = '1' then numsubscribe = 1;

else numsubscribe = 0;

run;

data Train; set WORK.IMPORT1;

if loan = 'yes' then numloan = 1;

else if loan = 'no' then numloan = 0;

else numloan = -1;

if default = 'yes' then numdefault = 1;

else if default = 'no' then numdefault = 0;

else numdefault = -1;

if housing = 'yes' then numhousing = 1;

else if housing = 'no' then numhousing = 0;

else numhousing = -1;

if contact = 'telephone' then numcontact = 1;

else if contact = 'cellular' then numcontact = 0;

else numcontact = -1;

if poutcome = 'success' then numpoutcome = 2;

else if poutcome = 'other' then numpoutcome = 1;

else if poutcome = 'failure' then numpoutcome = 0;

else numpoutcome = -1;

if month = 'jan' then nummonth = 1;

else if month = 'feb' then nummonth = 2;

else if month = 'mar' then nummonth = 3;

else if month = 'apr' then nummonth = 4;

else if month = 'may' then nummonth = 5;

else if month = 'jun' then nummonth = 6;

else if month = 'jul' then nummonth = 7;

else if month = 'aug' then nummonth = 8;

else if month = 'sep' then nummonth = 9;

else if month = 'oct' then nummonth = 10;

else if month = 'nov' then nummonth = 11;

else if month = 'dec' then nummonth = 12;

if Subscribed = '1' then numsubscribe = 1;

else numsubscribe = 0;

run;

data Test; set WORK.IMPORT2;

if loan = 'yes' then numloan = 1;

else if loan = 'no' then numloan = 0;

else numloan = -1;

if default = 'yes' then numdefault = 1;

else if default = 'no' then numdefault = 0;

else numdefault = -1;

if housing = 'yes' then numhousing = 1;

else if housing = 'no' then numhousing = 0;

else numhousing = -1;

if contact = 'telephone' then numcontact = 1;

else if contact = 'cellular' then numcontact = 0;

else numcontact = -1;

if poutcome = 'success' then numpoutcome = 2;

else if poutcome = 'other' then numpoutcome = 1;

else if poutcome = 'failure' then numpoutcome = 0;

else numpoutcome = -1;

if month = 'jan' then nummonth = 1;

else if month = 'feb' then nummonth = 2;

else if month = 'mar' then nummonth = 3;

else if month = 'apr' then nummonth = 4;

else if month = 'may' then nummonth = 5;

else if month = 'jun' then nummonth = 6;

else if month = 'jul' then nummonth = 7;

else if month = 'aug' then nummonth = 8;

else if month = 'sep' then nummonth = 9;

else if month = 'oct' then nummonth = 10;

else if month = 'nov' then nummonth = 11;

else if month = 'dec' then nummonth = 12;

if Subscribed = '1' then numsubscribe = 1;

else numsubscribe = 0;

run;

proc sort data=Train;

by Subscribed;

ods graphics on;

/\*\*\*\* Frequency Tables of Explanatory Variables \*\*\*\*/

proc freq data=Train;

tables Subscribed\*loan;

tables Subscribed\*default;

tables Subscribed\*month;

tables Subscribed\*contact;

tables Subscribed\*education;

tables Subscribed\*housing;

tables Subscribed\*poutcome;

tables Subscribed\*job;

tables Subscribed\*Age\_Group;

tables Subscribed\*pdays;

tables Subscribed\*previous;

run;

/\* All Factors for Logistic Regression \*/

proc logistic data=Train;

class Age\_Group job education default housing marital loan contact

month campaign poutcome/ param=ref;

model Subscribed = Age\_Group job education default housing

marital balance loan contact month duration

campaign poutcome pdays

previous/ scale=none details lackfit influence;

output out=myoutput predprobs=I p=probpred resdev=resdev reschi=reschi;

run;

/\* Logistic Model with Month \*/

proc logistic data=Train;

class loan(ref="yes") month(ref="jan") contact(ref="unknown") poutcome(ref="unknown") / param=ref;

model Subscribed = loan month contact duration poutcome / scale=none details lackfit influence;

output out=myoutput predprobs=I p=probpred resdev=resdev reschi=reschi;

run;

/\* Final Logistic Model to match R odds table \*/

proc logistic data=Train;

class loan(ref="no") contact(ref="cellular") poutcome(ref="success") / param=ref;

model Subscribed = loan contact duration poutcome / scale=none details lackfit influence;

output out=myoutput predprobs=I p=probpred resdev=resdev reschi=reschi;

run;

/\* Logistic Model \*/

proc logistic data=Train plots(only label)=(phat leverage dpc);

class loan(ref="yes") contact(ref="unknown") poutcome(ref="unknown") / param=ref;

model Subscribed = loan contact duration poutcome / scale=none details lackfit influence;

output out=myoutput predprobs=I p=probpred resdev=resdev reschi=reschi;

run;

proc logistic data=Train;

class loan(ref="yes") contact(ref="unknown") poutcome(ref="unknown") / param=ref;

model Subscribed = loan contact duration poutcome / scale=none details lackfit influence;

output out=myoutput predprobs=I p=probpred resdev=resdev reschi=reschi;

run;

/\* ROCs \*/

proc logistic data=Train;

class loan contact poutcome;

model Subscribed(event='1')= loan contact duration poutcome / scale=none ctable pprob=.5 aggregate lackfit;

ROC 'MainEffects' loan contact duration poutcome;

ROC 'Just duration' duration;

roccontrast reference('Just duration') / estimate e;

/\* ROC 'Just contact, poutcome & loan' contact poutcome loan;

roccontrast reference('Just contact, poutcome & loan') / estimate e; \*/

/\*\*\*\*\* OTHER Tests \*\*\*\*\*\*/

/\* ROC 'Just contact' contact;

roccontrast reference('Just contact') / estimate e;

ROC 'Just contact/poutcome' contact poutcome;

roccontrast reference('Just contact/poutcome') / estimate e;

ROC 'Just loan/duration' loan duration;

roccontrast reference('Just loan/duration') / estimate e;

ROC 'Just housing/default' housing default;

roccontrast reference('Just housing/default') / estimate e;

ROC 'Just loan/contact' loan contact;

roccontrast reference('Just loan/contact') / estimate e;

ROC 'Just month' month;

roccontrast reference('Just month') / estimate e;

ROC 'Just duration' duration;

roccontrast reference('Just duration') / estimate e; \*/

run;

ods graphics off;

**APPENDIX M: ADDITONAL SAS FOR MODEL ANALYSIS**

We check to see if there are any observations that have high or low residuals and high or low leverage. In Figure 1 for the influence diagnostics, the Pearson residuals show somewhat constant variance between -5 to 5 and only a few points with high leverage (397 and 1912). Figure 2 shows the predicted probability diagnostics where point 397 and 1912 show high leverage.

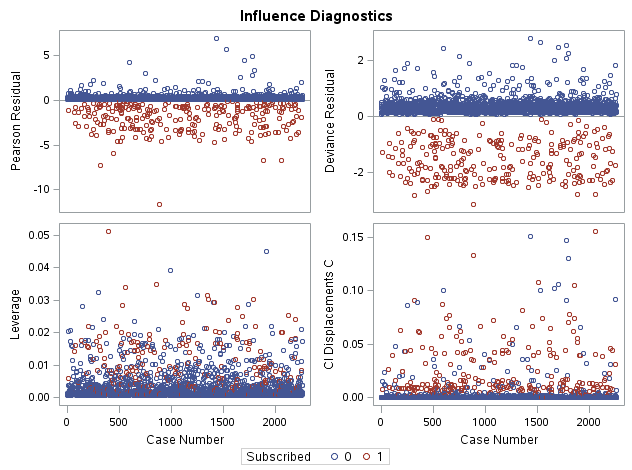
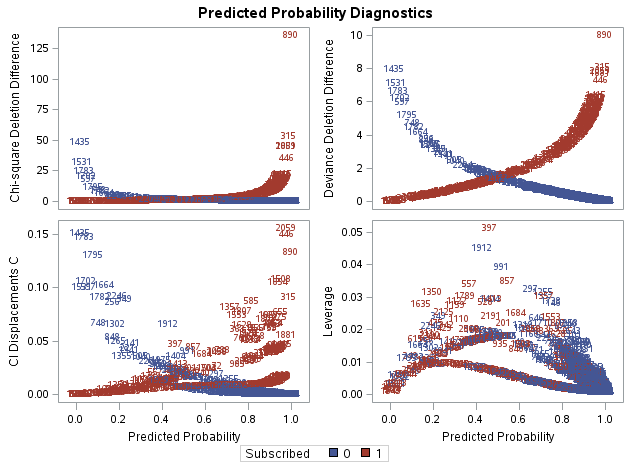
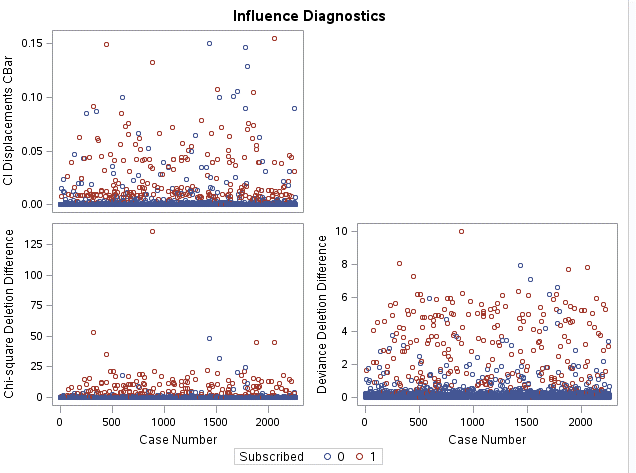
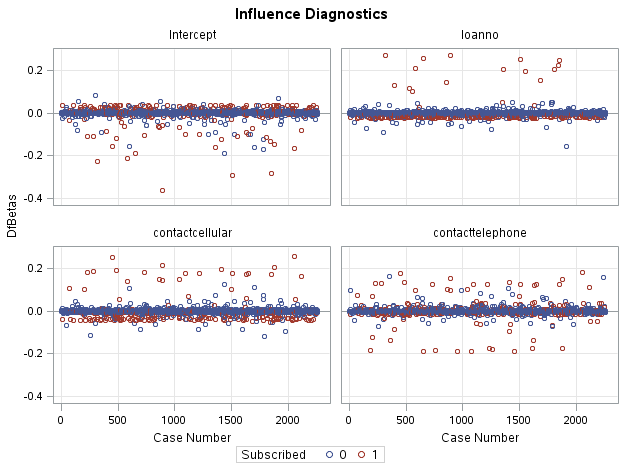
 

Figure 1: Influence Diagnostics Figure 2: Predicted Probability Diagnostics

Figure 3 shows we do not find too much variance via each explanatory variable.

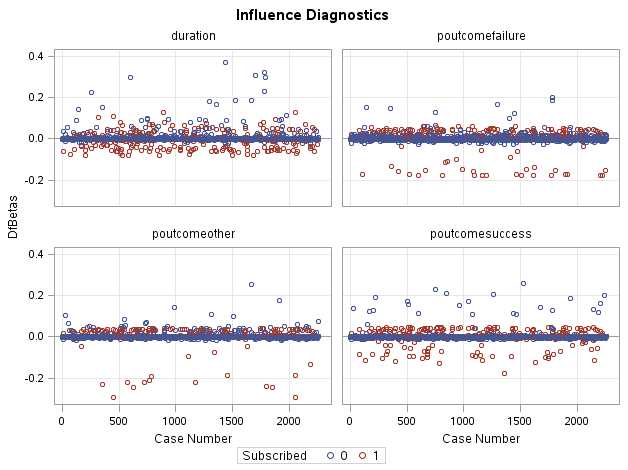


Figure 3: Influence of Each Explanatory Variable