Assignment 2

Data Analysis Project

Darren Rowland

MSc Business Analytics

MGT7177 Statistics for Business

Student No: 40350685

Contents

[1.0 Introduction 3](#_Toc92488337)

[2.0 Data Formatting & Quality Issues 4](#_Toc92488338)

[3.0 Data Exploration, Descriptive Statistics & Visualisations 6](#_Toc92488339)

[4.0 Regression Analysis 9](#_Toc92488340)

[5.0 Prediction & Accuracy 14](#_Toc92488341)

[6.0 Findings/Written Report Incl Intro 15](#_Toc92488342)

[7.0 References 16](#_Toc92488343)

[8.0 R Code 17](#_Toc92488344)

[Appendix 1 : Summary complete dataset – Virgin Data 33](#_Toc92488345)

[Appendix 2 : GLM Complete Model 34](#_Toc92488346)

[Appendix 5 : Data Format 35](#_Toc92488347)

# Introduction

Understanding and measuring implicit cognition is providing marketing researchers with the ability to probe automatic processes without the reliance on conscious deliberation Nevid (2010). These efforts by many marketers to record and understand the attitudes of consumers is not a new concept, it dates to the mid twentieth century with publications like that by Vance Packard ‘The Hidden Persuaders’ where the authors investigate the efforts to monitor the unconscious motivations in the influence of thought processes and purchasing decisions.

McKinsey (2011) reports that companies are capturing trillions of bytes of data on customers, with some citizens views as an intrusion of privacy and regard this accumulation of data with deep suspicion. They report that the world is on a cusp of tremendous innovation as the rate of change big data is bringing is at an inflection point and will expand greatly as technology trends accelerate and converge.

The ability to capture and store data has greatly increased with the development technology and cheaper storage facilities. As XXXX (XXXX) outlines in his book ‘The story of big data’, historically in the 1960s and 1980s computer scientists’ limitations were on technology and data. What has changed is the reduction in the cost of digital storage and the rapid increase in generated data from personal phones and computers to internet sensors and scientific sensors. Where once with the limited storage capacity this data may have been discarded to make way for new data now all the data can be stored cheaply.

Netto & Slongo (2019) state how analysing data for businesses now is becoming more of a necessity than just a differential and how the investment in analytical and big data skills must grow. If these skills aren’t found within marketing departments they will be recruited from elsewhere which reinforces the importance for marketing graduates and professionals to be rehearsed with working with metrics. In the past big organisation executives don’t understand fully the role of marketing, its value as a department or its responsibility or influence for business growth McDonald (2009) but more recently the understanding has grown which would change this mind-set.

Using data relating to Customer Lifetime Value (CLV) it is possible now for machines to learn a model relating to the customer characteristics to CLV which then can predict in test the outcome of the CLV of that customer. Research has on customer’s attitudes, emotional responses and preferences so much that neuromarketing is helping understand the brain activity as an underlying response indication of products like brand identities and movie trailers. Nevid (2010)

This study will use historical banking data in order to understand the correlations between variables and the possibility of achieving certain outcomes in order to predict if a client will subscribe to a term deposit financial product.

The data supplied from previous marketing campaign shows numerous variables and information on the customers. Becoming familiar with the data supplied and for the purpose of this study it would be suggested that one could potentially use certain variables to predict the potential of a customer subscribing to a term deposit. Without in depth analysis of the data it could be assumed that the customer most likely to be subscribe to this financial would be that who may have extra expendable cash. It could be assumed that; without social & economic market data being factored, the customer is more likely to be in a well-paid job, the customer may have a professional education and could be married. This is an initial idea but it would be remiss to think that these are the only customers in a financial position to subscribe to this product.

The above assumption may be naive and therefore the ability to use historical data on all customers in a given marketing campaign is invaluable in terms of visualising the data, understanding the banks customers and also to make predictions for future campaigns.

The data contains useful variables on the customer like marital status, job, housing loan, loan and default on loans. Also data captured includes social and economic data at that time of customer contact. This data therefore can used to understand the term deposit rates based on euribor rates and also the use of consumer confidence index and employment variable rates which may give an idea of the economy at that time.

The hypothesis below will be test throughout this study to understand if the data has relationship on the dependent variable or it doesn’t.

H0 – Null Hypothesis – there is no relationship between Subscription and other variables

H1 – a relationship exists between Subscription and other variables

# Data Formatting & Quality Issues

Exploring the data that has been received it is evident issues exist in the data quality & formatting. These issues will be resolved in order to progress with the data analysis and data modelling. To carry out the formatting tasks both datasets test & train are joined using rbind() function in order to reduce the workload required to clean two data frames. On completion the data will be split back into 80% train 20% test for the regression model.

From summary of the data all variables are showing as either numeric or character variables. In reality this is not the case, from the data dictionary it is compared that all character variables shown should be set as factors (see list of variables in Table 1 and variable types required). The mutate() function is a very quick way to convert these (See appendix XX), lapply can also be used as an alternative for a given range ie. data[3:11].

Figure 1 : Summary Formatted & Cleaned Data



# Data Exploration, Descriptive Statistics & Visualisations

The dataset supplied is concerned with retail bank client personal information and other variables including social & economic performance. The dataset contains 41,188 observations and 22 variables with numeric and categorical classifications. The data is based on historical information captured on the performance of a telephone & cellular marketing campaign to sell subscriptions to a term deposit financial product.

Table 1 : Dataset Variable Classification

|  |  |  |
| --- | --- | --- |
| **Column1** | **Variable** | **Variable Class** |
| 1 | ID | Numeric |
| 2 | age | Numeric |
| 3 | job | Categorical |
| 4 | marital\_status | Categorical |
| 5 | education | Categorical |
| 6 | default | Categorical (Binary) |
| 7 | housing | Categorical (Binary) |
| 8 | loan | Categorical (Binary) |
| 9 | contact | Categorical |
| 10 | month | Categorical |
| 11 | day\_of\_week | Categorical |
| 12 | duration | Numeric |
| 13 | campaign | Numeric |
| 14 | pdays | Numeric |
| 15 | previous | Numeric |
| 16 | poutcome | Categorical |
| 17 | emp\_var\_rate | Numeric |
| 18 | cons\_price\_idx | Numeric |
| 19 | cons\_conf\_idx | Numeric |
| 20 | euribor3m | Numeric |
| 21 | nr\_employed | Numeric |
| 22 | subscribed | Categorical (Binary) |

**Measures of Association**

Given the desired requirements of the study to predict whether a client subscribes to a financial product or not it is imperative that the data is explored to understand and relationships and how strong the correlation between the dependent variable and any independent variables. Using the renamed ‘subscribed’ variable a number of association tests will be carried out to analyse relationships.

To understand the associations of variables with the dependent outcome variable Table 2 was constructued using the Pearsons Chi-Squared Test to account for characteristic variables and Table 3 was constructed using the T-Test.

Table 2 : Pearsons Chi-Squared Test

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Pearson's Chi-Squared Test** | | | | |  |
|  | Constant Dependent Variable : 'subscribed' | |  |  |  |  |
|  |  |  |  |  |  |  |
|  | **Independent Variable** | **X-squared** | **df** | **p-value** | |  |
|  | age | 1658.2 | 79 | < | 2.20E-16 |  |
|  | marital | 122.66 | 3 | < | 2.20E-16 |  |
|  | job | 961.24 | 11 | < | 2.20E-16 |  |
|  | education | 193.11 | 7 | < | 2.20E-16 | \* |
|  | default | 406.58 | 2 | < | 2.20E-16 | \* |
|  | housing | 5.6845 | 2 |  | 0.05829 |  |
|  | loan | 1.0466 | 2 |  | 0.5926 |  |
|  | contact | 863.27 | 1 | < | 2.20E-16 |  |
|  | month | 3101.1 | 9 | < | 2.20E-16 |  |
|  | day\_of\_week | 26.145 | 4 |  | 2.96E-05 |  |
|  | poutcome | 4230.5 | 2 | < | 2.20E-16 |  |
|  |  |  |  |  |  |  |

Table 3 : t-test

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Welch Two Sample t-test** | | | | | | | |  |
|  | Constant Dependent Variable : 'subscribed' | | | |  |  |  |  |  |
|  |  |  |  |  |  |  | **Sample Estimates** | |  |
|  |  | **t** | **df** | **p-value** | **95 percent conf int** | | **Mean No** | **Mean Yes** |  |
|  | age | -4.7757 | 5258.4 | 1.84E-06 | -1.4121351 | -0.5901921 | 39.91198 | 40.91315 |  |
|  | duration | -59.681 | 4938.1 | < 2.2e-16 | -322.5814 | -302.0627 | 218.4159 | 530.738 |  |
|  | pdays | 32.251 | 4744 | < 2.2e-16 | 180.4023 | 203.7544 | 984.1139 | 792.0356 |  |
|  | previous | -28.126 | 4908.8 | < 2.2e-16 | -0.3854126 | -0.3351845 | 0.1323739 | 0.4926724 |  |
|  | emp.var.rate | 59.137 | 5665.6 | < 2.2e-16 | 1.433185 | 1.531463 | 0.2488755 | -1.2334483 |  |
|  | cons.price.idx | 24.06 | 5472.5 | < 2.2e-16 | 0.2288268 | 0.2694238 | 93.60379 | 93.35467 |  |
|  | cons.conf.idx | -8.6365 | 5258.3 | < 2.2e-16 | -0.9856585 | -0.620966 | -40.5931 | -39.78978 |  |
|  | euribor3m | 62.58 | 5729.2 | < 2.2e-16 | 1.635467 | 1.741246 | 3.811491 | 2.123135 |  |
|  | nr.employed | 60.975 | 5298.3 | < 2.2e-16 | 78.44475 | 83.65647 | 5176.167 | 5095.116 |  |
|  |  |  |  |  |  |  |  |  |  |

**Visualisations**

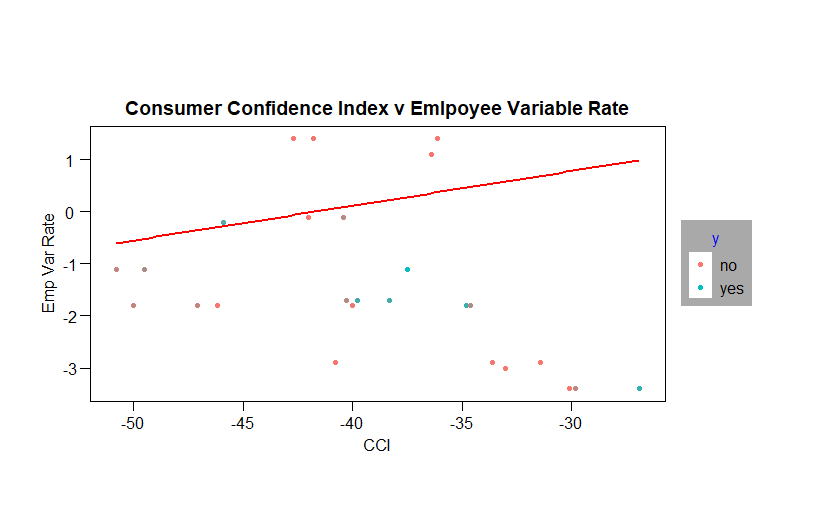
Figure 2

Visualisation 1



Figure 3

Visualisation 2



# Regression Analysis

Methodology

Throughout this study the data mining tasks which needed to be carried out was 1. Data Extraction & Binding 2. Data Cleaning & Quality Issues 3. Descriptive Statistics with Visualisations 4. Data Exploration & Initial Analysis 5. Data Splitting 80% Train 20% Test 6. Model Development & Accuracy 7. Prediction on Test Data. The overall data analysis will look at the fit of the model and try to make the most accurate prediction.

**Pseudo r Squared**

As shown below the variation between the r^2 values of the train dataset and the r^2 values of the test dataset differ very little and therefor would suggest confidence in an accurate model. The model3 that has been developed accounts for 27.6% outcome variability.

Table 4a: data=train Table 4b: data=test

**Odds Ratio**

Below table outlines the odds ratios of the predictor variables on the outcome of the dependent variable.

Table 5 :



**Akaike information criterion (AIC)**

It can be seen from the below summaries of the logistic regression model2 & model3 that the AIC (Akaike information criterion) reduces as the model is developed with more predictor variables. This outlines that the residual values have decreased and therefore shows that there is a decrease in the error of the model. According to AIC the best fit model is that which explains greatest variation with least independent variables.

Table 6 : Summary Model2 Table 7 : Summary Model3

**Variance Inflation Factor (vif)**

No variance inflation factor =greater than 10 which is good

Table 8 : VIF outcomes for model3



Noticing the vif of the emp.var.rate, poutcome and euribor3m very high the model was modified to remove these as predictor variables. The accuracy changes and kappa changes are minimal but there are significant reductions in the vif of the predictor variables as hoped. See Table 9 below.

Table 9 : VIF Updated model3 to remove social economic variables



**Cooks Distance**

Any cooks distances above 1 then that observation will be having a negative effect on your regression line

Table 10 : Summary Cooks Distance



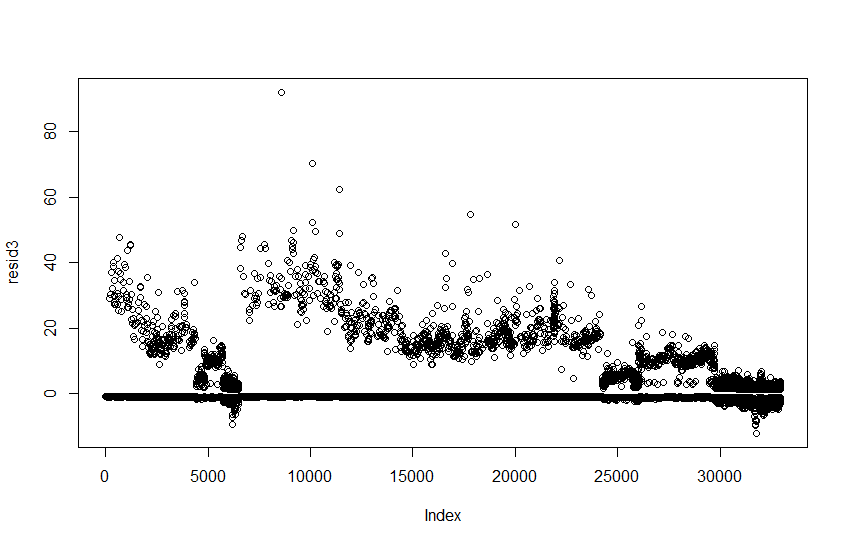
**Linearity of the Logit**

Testing the linearity of the logit on the continuous variables will whether the violates the assumptioons or not. The variables age, pdays, cons.conf.idx and nr.employed were used as these are continuous. R Studio provided a warning that the cons.conf.idx is giving NaNs outputs and when this is used in the model to check the assumptions an error is seen so for the logit this has been removed. The interaction of the logits doesn’t affect the model significantly so this would not violate the assumptions.

**Residuals**

Model3a has a total of 373 residuals greater than 1.96 which is less than 5% of the data and would suggest a good fitting model. Although as can be seen from the residuals they are not randomly spaced out which is a cause for concern.

Figure 4 : Plot Model3 Residuals



**Durban Watson Test**

When the DWT is ran on the model a positive autocorrelation can be seen which indicates a rejection of the null hypothesis.

Table 11 : Durban Watson Test Result



# Prediction & Accuracy

The model developed obtained a 90% accuracy and a kappa value of 0.2939 which translates to have a fair probability. This kappa value could be increased to increase the confidence of the outcome.

Table 12 : Prediction Model3



Table 13 : Confusion Matrix Model3



# Findings/Written Report Incl Intro

This investigative research study is based on a marketing department dataset from a retail bank which includes 22 variables and 41,188 observations. The aim is to analyse the available data to understand which factors pertaining to the personal data of a customer most influences the chances of customer subscribing to a financial product; in this case a term deposit.

A Logistic Regression Model then has been built to support the marketing department make predictions on the outcome of the marketing campaign. Three models in total have been built; the model1 using the full dataset for initial review of the input data.

From the analysis of the data it can be seen that many of the variables of the dataset have statistically significant influence on the dependent variable as can be seen from the p values in the tables with the chi-squared test and the t-test. Given the influence of these variables it was paramount to be able to build an accurate logistic regression model to make predictions on the outcome of the marketing campaign for the banks term deposit product.

The assumptions at the introduction were quickly tested on the complete joined dataset to get an understanding of the statistically significant variables. As can be seen in Appendix 2 the jobblue-collar, jobretired and jobstudent all show to have a statistical significance. The blue collar job and retired could have been captured in the initial predictions whereas student was overlooked as having limited cash flow to subscribe. As well as that university degree and number employed have statistical significance. All others showing statistical significance could be assumed and these were used to influence the accuracy of the model.

From the accuracy and kappa values it the model shows to be accurate, one concern is the residuals. The plot of the residuals does not follow a random sequence which should be seen, some experimentation was done with the model but the issues it would seem lie with the dependent variable. This requires further analysis to understand what the issue is and why this is violating the assumptions when all other assumptions show confidence in the model.

The social and economic variables as outlined in the study created major issues in the variance inflation factor as they are all strongly correlated. All except consumer confidence index and number employed which do not have negative impact on the model were removed.

This study proves that with limited data how an accurate model is built and can be used to predict customers’ decision on purchases. In industry thanks to the applications of models and algorithms like this one, the marketing team can spend more time on more demanding tasks resulting in a cost effective and efficient marketing department Gentsch (2019).

# References

Carla Freitas Silveira Netto, Luiz Antonio Slongo (2019) Marketing metrics, Big Data and the role of the marketing department

Davenport, Thomas H. (2014) Big data at work : dispelling the myths, uncovering the opportunities (1st Edition). USA: Harvard Business School Publishing Corporation.

Gentsch, Peter. 2019. AI in Marketing Sales and Service. (1st Edition) Springer Nature Switzerland AG

Gijs Overgoor, Manuel Chica, William Rand, Anthony Weishampel (2019) Letting the Computers Take Over: Using AI to Solve Marketing Problems. California Management Review

Jeffrey S. Nevid (2010) Introduction to the Special Issue: Implicit Measures of Consumer Response—The Search for the Holy Grail of Marketing Research. Psychology & Marketing.

McKinsey Global Institute. 2011. Big data: The next frontier for innovation, competition, and productivity [website] Accessed by : <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation> [accessed January 2022]

Se´rgio Moro, Paulo Cortez, Paulo Rita (2014) Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. The Natural Computing Applications Forum 2014

# R Code

install.packages("readxl")

install.packages("dplyr")

library(readxl)

library(dplyr)

install.packages("ggplot2")

install.packages("psych")

install.packages("Hmisc")

install.packages("ggthemes")

install.packages("extrafont")

install.packages("caret")

install.packages("car")

library(ggplot2)

library(tidyverse)

library(psych)

library(Hmisc)

library(ggthemes)

library(extrafont)

library(scales)

library(caret)

library(car)

setwd("F:/!MSc Business Analytics/MGT7177 Statistics for Business/!\_Assignment 2")

getwd()

train <- read\_excel("bank\_train.xlsx")

test <- read\_excel("bank\_test.xlsx")

data <- rbind(test, train, deparse.level = 1)

summary(data)

classTable <- t(data.frame(lapply(data, class)))

classTable1 <- t(data.frame(lapply(data, class)))

### convert multiple variables to factor ###

data <- data %>% mutate\_if(is.character, as.factor)

summary(data$contact)

## another way to change all characters to factors is

# data[3:11] <- lapply(data[3:11], factor)

# summary(data[3:11])

### AGE ###

summary(data$age)

sum(data$age > 100)

data$age[data$age > 100] <- mean(data$age)

data$age[data$age < 16] <- mean(data$age)

summary(data)

### contact ###

summary(data$contact)

data$contact[data$contact == "mobile"] <- "cellular"

data$contact <- droplevels(data$contact)

### month ###

### arrange months in order

data$month = factor(data$month, levels(data$month)[c(6,1,7,5,4,2,10,9,8,3)])

summary(data$month)

### days of week ###

### change and add 'Friday' data to 'fri'

### remove 'Friday' from the data

### arrange days of the week in order

summary(data$day\_of\_week)

data$day\_of\_week[data$day\_of\_week == 'Friday'] <- 'fri'

data$day\_of\_week <- droplevels(data$day\_of\_week)

data$day\_of\_week = factor(data$day\_of\_week, levels(data$day\_of\_week)[c(2,4,5,3,1)])

summary(data$day\_of\_week)

### duration ###

summary(data$duration)

outliersDuration <- boxplot(data$duration)$out

summary(outliersDuration)

sum(outliersDuration > 1800)

sum(data$duration > 1800)

data$duration[data$duration > 1800] <- mean(data$duration)

boxplot(data$duration)$out

### campaign ###

summary(data$campaign)

### Summary pdays ###

### remove 999 from data, replace with NA - 999 not previously contacted

summary(data$pdays)

sum(data$pdays == 999)

sum(data$pdays == 0)

sum(data$pdays == 'NA') #0 NA in original data

summary(data$pdays)

### Employment Variation Rate ###

summary(data$emp.var.rate)

plot(data$emp.var.rate)

barplot(data$emp.var.rate)

### Consumer Price Index ###

summary(data$cons.price.idx)

plot(data$cons.price.idx)

sum(data$cons.price.idx < 92)

sum(data$cons.price.idx > 95)

sum(data$cons.price.idx == 149)

### Remove outlier values from the CPI and apply mean value

outlierCPI <- boxplot(data$cons.price.idx)$out

data$cons.price.idx[data$cons.price.idx > 100] <- mean(data$cons.price.idx)

hist(data$cons.price.idx, breaks = 50)

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

### Consumer Confidence Index ###

summary(data$cons.conf.idx)

hist(data$cons.conf.idx, data$emp.var.rate, breaks = 20)

### Euribor 3 Month rate ###

summary(data$euribor3m)

plot(data$euribor3m)

plot(data$euribor3m ~ data$cons.conf.idx)

### Number of Employees ###

summary(data$nr.employed)

plot(data$nr.employed ~ data$campaign)

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

### subscribed ###

### change the name of variable y to 'subscribed'

summary(data$y)

names(data)[names(data) == 'y'] <- "subscribed"

summary(data$subscribed)

summary(data)

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

########## Measures of Association ##########

#### Correlation #####

### CHI Sq Test ###

chisq.test(data$subscribed, data$marital, correct = F)

chisq.test(data$subscribed, data$job, correct = F)

chisq.test(data$subscribed, data$education, correct = F)

summary(data$education)

chisq.test(data$subscribed, data$default, correct = F)

summary(data$default)

chisq.test(data$subscribed, data$housing, correct = F)

chisq.test(data$subscribed, data$loan, correct = F)

chisq.test(data$subscribed, data$contact, correct = F)

chisq.test(data$subscribed, data$month, correct = F)

chisq.test(data$subscribed, data$day\_of\_week, correct = F)

chisq.test(data$subscribed, data$poutcome, correct = F)

t.test(data$age ~ data$subscribed, data=data)

t.test(data$duration ~ data$subscribed, data=data)

t.test(data$pdays ~ data$subscribed, data=data)

t.test(data$previous ~ data$subscribed, data=data)

t.test(data$emp.var.rate ~ data$subscribed, data=data)

t.test(data$cons.price.idx ~ data$subscribed, data=data)

t.test(data$cons.conf.idx ~ data$subscribed, data=data)

t.test(data$euribor3m ~ data$subscribed, data=data)

t.test(data$nr.employed ~ data$subscribed, data=data)

## why use spearmans below???

cor(data$age, data$emp.var.rate, use = "complete.obs", method = "spearman")

cor(data$age, data$emp.var.rate, use = "complete.obs", method = "pearson")

cor.test(data$emp.var.rate, data$cons.conf.idx)

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

#### Visualisations ###

plot()

plot(data$euribor3m ~ data$nr.employed)

plot(data$euribor3m ~ data$emp.var.rate)

plot(data$age ~ data$subscribed)

scatter.smooth(data$age, data$subscribed)

plot(data$duration ~ data$campaign)

plot(data$campaign, data$nr.employed)

barplot(data$day\_of\_week, data$campaign)

plot(data$campaign, data$day\_of\_week)

hist(data$day\_of\_week)

plot(lm(data$day\_of\_week))

summary(data$day\_of\_week)

plot(data$euribor3m)

hist(data$subscribed ~ data$job)

summary(data$job)

### Visualisation 1 ###

### Age ~ Emp Var Rate ###

ggplot(data=data) +

geom\_point(mapping = aes(x = age, y = emp.var.rate, col=subscribed)) +

labs(title="Customer Age ~ Employee Variable Rate",

x="Age (years)",

y="Emp Var Rate") +

theme(legend.title=element\_text(color="blue"),

legend.background = element\_rect(fill = "darkgray"),

legend.position = "right")+

facet\_wrap(~ subscribed)

### Visualisation 2 ###

### cons.conf.idx ~ emp.var.rate ###

ggplot(data=data) +

geom\_point(mapping = aes(x = cons.conf.idx, y = emp.var.rate, color= subscribed, alpha =0.5)) +

geom\_smooth(mapping = aes(x = cons.conf.idx, y = emp.var.rate), method = lm, se=F, col="Red") +

labs(title="Consumer Confidence Index v Emlpoyee Variable Rate",

x="CCI",

y="Emp Var Rate") +

theme\_par() +

theme(legend.title=element\_text(color="blue"),

legend.background = element\_rect(fill = "darkgray"),

legend.position = "right") +

scale\_alpha(guide = "none")

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

### Skewness & Kurtosis of the variable ###

psych::describe(data$age, na.rm=T, interp=F, skew=T)

psych::describe(log(data$age))

psych::describe(data$emp.var.rate, na.rm=T, interp=F, skew=T)

qqplot(data$age, data$cons.conf.idx)

psych::describe(data$emp.var.rate)

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

################ Logistic Regression ################

################# Model 1 ####################

formula1 <- subscribed ~ age + marital + job + cons.conf.idx + euribor3m + nr.employed

formula1 <- subscribed ~ .

model1 <- glm(formula1, data=data, family ="binomial")

summary(model1)

## evaluate model assumptions

train$predictProbabilities <- fitted(model1)

summary(train$predictProbabilities)

predictions <- predict(model1, test, type ="response")

head(predictions)

summary(predictions)

class\_pred <- as.factor(ifelse(predictions > 0.5, "Yes", "No"))

summary(class\_pred)

exp(model1$coefficients)

summary(data$marital)

summary(data$job)

### Pseudo R Sqaured ###

logisticPseudoR2s <- function(LogModel) {

dev <- LogModel$deviance

nullDev <- LogModel$null.deviance

modelN <- length(LogModel$fitted.values)

R.l <- 1 - dev / nullDev

R.cs <- 1- exp ( -(nullDev - dev) / modelN)

R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))

cat("Pseudo R^2 for logistic regression\n")

cat("Hosmer and Lemeshow R^2 ", round(R.l, 3), "\n")

cat("Cox and Snell R^2 ", round(R.cs, 3), "\n")

cat("Nagelkerke R^2 ", round(R.n, 3), "\n")

}

logisticPseudoR2s(model1)

#Source: Field, 2012

# check how many residuals

resid <- model1$residuals

plot(resid)

std\_resid <- rstandard(model1)

sum(std\_resid > 1.96) # must be less than 5% of the data

summary(std\_resid) #check if there are any std residual above 3

large\_resid <- std\_resid > 2 | std\_resid < -2

sum(large\_resid)

library(car)

vif(model1)

mean(vif(model1))

1/(vif(model1))

train$cook <- cooks.distance(model1)

cooks <- cooks.distance(model1) #check for observatiosn having any undue influence on the model

sum(cooks > 1)

# add standardised residuals to the data

data$standardisedResid <- rstandard(model1)

sum(train$standardisedResiduals > 1.96) #should be less than 5% of the data

################ Split Data ################

set.seed(123)

### create a 80/20 partition

trainIndex <- createDataPartition(data$subscribed, p=0.8,

list = FALSE,

times = 1)

train <- data[trainIndex,]

test <- data[-trainIndex,]

summary(test$job)

summary(train$marital)

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

################# Model 2 ####################

formula2 <- subscribed ~ age + marital + job + cons.conf.idx + euribor3m + nr.employed

formula2 <- subscribed ~ age + marital + job + cons.conf.idx + euribor3m + nr.employed

model2 <- train(formula2, data=train, method="glm")

model2a <- glm(formula2, data=train, family ="binomial")

predictions2 <- predict(model2, test)

summary(predictions2)

summary(model2)

## evaluate model assumptions

train$predictProbabilities <- fitted(model2)

summary(train$predictProbabilities)

exp(model2$coefficients)

postResample(predictions2, test$subscribed)

confusionMatrix(data = predictions2, test$subscribed)

### Pseudo R Sqaured ###

logisticPseudoR2s <- function(LogModel) {

dev <- LogModel$deviance

nullDev <- LogModel$null.deviance

modelN <- length(LogModel$fitted.values)

R.l <- 1 - dev / nullDev

R.cs <- 1- exp ( -(nullDev - dev) / modelN)

R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))

cat("Pseudo R^2 for logistic regression\n")

cat("Hosmer and Lemeshow R^2 ", round(R.l, 3), "\n")

cat("Cox and Snell R^2 ", round(R.cs, 3), "\n")

cat("Nagelkerke R^2 ", round(R.n, 3), "\n")

}

logisticPseudoR2s(model2)

#Source: Field, 2012

# check how many residuals

resid2 <- model2$residuals

plot(resid2)

resid2a <- model2a$residuals

plot(resid2a)

std\_resid2 <- rstandard(model2)

sum(std\_resid2 > 1.96) # must be less than 5% of the data

summary(std\_resid2) #check if there are any std residual above 3

large\_resid2 <- std\_resid2 > 2 | std\_resid2 < -2

sum(large\_resid2)

vif(model1)

train$cook <- cooks.distance(model2)

cooks <- cooks.distance(model2) #check for observatiosn having any undue influence on the model

sum(cooks > 1)

# add standardised residuals to the data

data$standardisedResid <- rstandard(model2)

sum(train$standardisedResiduals > 1.96) #should be less than 5% of the data

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

################# Model 3 ####################

formula3 <- subscribed ~ age + job + month + day\_of\_week + contact + campaign + pdays + cons.conf.idx + nr.employed

formula3 <- subscribed ~ age + job + month + day\_of\_week + contact + campaign + pdays + cons.conf.idx + nr.employed + train$ageLogInt + train$pdaysLogInt + train$emplLogInt

model3 <- train(formula3, data=train, method="glm", na.action=na.omit)

model3a <- glm(formula3, data=train, family ="binomial")

summary(model3a)

predictions3 <- predict(model3, test)

summary(predictions3)

predictions3a <- predict(model3a, test)

summary(predictions3a)

class\_pred3a <- as.factor(ifelse(predictions3a > 0.5, "Yes", "No"))

summary(class\_pred3a)

summary(model3)

## evaluate model assumptions

postResample(predictions3, test$subscribed)

confusionMatrix(data = predictions3, test$subscribed)

## evaluate model assumptions

train$predictProbabilities <- fitted(model3)

summary(train$predictProbabilities)

exp(model3a$coefficients)

vif(model3a)

mean(vif(model3a))

### Pseudo R Sqaured ###

logisticPseudoR2s <- function(LogModel) {

dev <- LogModel$deviance

nullDev <- LogModel$null.deviance

modelN <- length(LogModel$fitted.values)

R.l <- 1 - dev / nullDev

R.cs <- 1- exp ( -(nullDev - dev) / modelN)

R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))

cat("Pseudo R^2 for logistic regression\n")

cat("Hosmer and Lemeshow R^2 ", round(R.l, 3), "\n")

cat("Cox and Snell R^2 ", round(R.cs, 3), "\n")

cat("Nagelkerke R^2 ", round(R.n, 3), "\n")

}

logisticPseudoR2s(model3a)

#Source: Field, 2012

# check how many residuals

resid3 <- model3a$residuals

plot(resid3)

std\_resid3 <- rstandard(model3a)

sum(std\_resid3 > 1.96) # must be less than 5% of the data

summary(std\_resid3) #check if there are any std residual above 3

large\_resid2 <- std\_resid2 > 2 | std\_resid2 < -2

sum(large\_resid2)

vif(model1)

train$cook <- cooks.distance(model3a)

cooks <- cooks.distance(model3a)

#check for observatiosn having any undue influence on the model

sum(cooks > 1)

summary(cooks)

# add standardised residuals to the data

data$standardisedResid <- rstandard(model2)

sum(train$standardisedResiduals > 1.96) #should be less than 5% of the data

durbinWatsonTest(model3a)

### Linearity of the Logit ###

train$ageLogInt <- log(train$age) \* train$age

train$pdaysLogInt <- log(train$pdays) \* train$pdays

train$cLogInt <- log(train$cons.conf.idx) \* train$cons.conf.idx

train$emplLogInt <- log(train$nr.employed) \* train$nr.employed

### confidence interval ###

exp(confint(model3a))

### odds ratio ###

exp(model3a$coefficients)

####################### END ################################

# Appendix 1 : Summary complete dataset – Virgin Data

# Appendix 2 : GLM Complete Model

# Appendix 5 : Data Format

Age

agetest <- data$age[data$age > 90]

agetest

91 92 95 120 120 170 91 98 98 94 92 92 92

agetest <- data$age[data$age > 90]

agetest

data$age[data$age > 100] <- NA



summary(data$loan)



Contact

summary(data$contact)



data$contact[data$contact == "mobile"] <- "cellular"

data$contact <- droplevels(data$contact)



Duration

sum(data$duration > 1800)

[1] 101

Summary(data$y)



names(data)[names(data) == 'y'] <- "subscribed"

data$subscribed <- as.factor(data$subscribed)

summary(data$subscribed)



summary(data$month)



data$month = factor(data$month, levels(data$month)[c(6,1,7,5,4,2,10,9,8,3)])

summary(data$month)



summary(data$day\_of\_week)



data$day\_of\_week[data$day\_of\_week == 'Friday'] <- 'fri'

data$day\_of\_week <- droplevels(data$day\_of\_week)

data$day\_of\_week = factor(data$day\_of\_week, levels(data$day\_of\_week)[c(2,4,5,3,1)])

summary(data$day\_of\_week)



Consumer Price Index

Only 4 observations of total 41,188 outside the range 92-95 and these observations all = 149.

outlierCPI <- boxplot(data$cons.price.idx)$out

sum(data$cons.price.idx < 92 )



sum(data$cons.price.idx >95 )



sum(data$cons.price.idx == 149)



Replace with mean CPI

data$cons.price.idx[data$cons.price.idx > 100] <- mean(data$cons.price.idx)

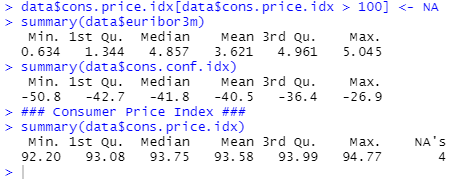


Consumer Confidence Index

shows there being 447 outliers in the variable though all these observations are -26.9, even with these observations only accounting for 1.09% of the data they will be retained.

outlierCCI <- boxplot(data$cons.conf.idx)$out

mean(outlierCCI) : -26.9



Summary(data$previous)

