Logo

Description automatically generated

**Predicting customer characteristics for a term deposit product in a retail bank.**

Icon

Description automatically generated with medium confidence

**By:**

**TATHAGATA CHAKRABORTY**

**(MASTER OF SCIENCE, BUSINESS ANALYTICS)**

**STUDENT NUMBER: 40379173**

**MODULE CODE: MGT\_7177**

**MODULE NAME: STATISTICS FOR BUSINESS**

Logo, company name

Description automatically generated

**CONTENTS**

**Pg no.**

* **Introduction -------------------------------------------- 3**
* **Background --------------------------------------------- 3**
* **Methodology ------------------------------------------- 4**
* **Discussion and Conclusion ------------------------- 11**
* **References ----------------------------------------------- 12**
* **Appendix ------------------------------------------------ 13**

**INTRODUCTION**

Organizations use marketing campaigns as a basic kind of outsourcing in order to strengthen the financial position of their companies and get an edge over their competitors (Vaughn and Wang, 2008). Administrative governance of campaigns is facilitated by connecting with the customers through remote communication hubs. Such call centres enable customers to communicate with them using a range of channels, such as cell phones or telephones. Because of its remoteness, the promotion of a product upon a contact centre is known as telemarketing (Frempong and Jayabalan, 2017; Moro et al., 2014). With the use of technology, marketing can be rethought with a focus on maximising customer lifetime value through the analysis of accessible data and customer KPIs, allowing us to forge closer, more lasting relationships that are in line with customer demand (Moro et al., 2014). The retail bank in question had recently run a telemarketing campaign in order to gain customers to subscribe to their term deposit scheme. The customer data obtained from the internal departments enlists the various customer characteristics along with their decision to subscribe to this product. These characteristics have been minutely observed and modelled in this report using logistic regression technique in order to develop a model which can benefit the marketing departments in understanding the customer behaviour and efficiently target their advertisement or modulate their campaigns accordingly.

**BACKGROUND**

The most reliable sources of credit and revenue for the banking industry are term deposit accounts (Zhuang et al., 2018). Direct marketing and mass campaigns are the two most common types of marketing strategies used by businesses to advertise their services and/or products (Ling and Li, 1998) . While direct marketing efforts are carried out with a specified target audience in mind, mass campaigns are directed to the general, indiscriminate public. Direct marketing strategies are more successful than mass advertising, which receive less than 1% of favourable responses, according to a study conducted by Ou et al. (2003).

The dataset obtained from the bank provides a range of customer attributes relating to their eventual decision of subscribing to the term deposit scheme. The marketing campaign used only cell phone and telephone as mode of communication. The dataset encompasses a wide array of characteristics including education of the customers, age, duration of the conversation during campaign, type of employment, day and month of contact, and outcome of the previous campaign for that customer. Apart from these qualitative attributes, the dataset also provides insights into broader economic indicators which includes consumer price index, consumer confidence index, euribor3m and employment variation rate, which have been taken into consideration during the formulation of the regression model for this analysis.

Logistic regression is a popular technique used in analysis when the dependent variable is a categorical parameter of two groups (Field et al., 2012). Frempong and Jayabalan (2017) had performed logistic regression in determining the efficiency of direct marketing techniques of retail banks for their products and services. They were able to obtain an accuracy of 83.5% in their analysis using logistic regression, which was quite an optimistic figure given the complexity of the dataset.

Considering the case at hand, five distinct customer attributes have been chosen for the formulation of respective hypotheses : duration, housing, default, age, euribor3m. These hypotheses have been tested using appropriate bivariate testing measures and the results of which shall help us realise the dependencies and relationship between the variables in consideration. The hypotheses have been listed below.

**H1** **:** **There is a substantial difference in mean of duration between customers who have subscribed and those who have not. :** Hou et al. (2022) in their analysis, concluded that the length of the call has a big impact on the client's decision since the longer the final call, the more interested the client will be in the term deposit. Thus, duration can be a defining factor in building the final model.

**H2 : There is significant correlation between having a housing loan and subscribing to term deposit for the customers. :** Having a housing loan for customers might affect their decision to enter into another financial commitment. As emphasised by Aron et al. (2012), that housing market is one of the key indicators of economic liberalisation. Thus, whether or not a customer decides to save up on his finances or increase his propensity towards spending can be affected by any association or commitment to an existing house loan (Fernandez-Cuergedo and Muellbauer, 2006).

**H3 : There is significant correlation between credit default and subscription of term deposit. :** Heide and Neuer (2010) in their paper have discussed how banks try to minimise their associations in schemes and transactions which involve high credit risks. Customers with a history in credit default are thus less likely to be considered as a potential client in this venture.

**H4 : There is significant difference in the mean age of the customers who subscribed to term deposit and the ones who did not. :** Age influences customers' willingness to purchase a term deposit product (Hou et al., 2022). This may be because older bank customers are more prone than younger ones to subscribe to term deposits.

**H5 : There is a notable difference in mean of euribor3m between customers who have subscribed and those who have not. :** Euribor3m is a significant predictor of average interbank interest rates in the Eurozone because consumers are more ready to spend their money on financial products when interest rates are high (Hou et al., 2022; Ilham et al., 2019).

**METHODOLOGY**

In this analysis, logistic regression technique has been eventually used to model the data for better understanding of the parameters that affect customer subscription. Upon initial analysis, it was observed that out of 46,563 customers, only 4640 customers had subscribed i.e roughly 10% of the targeted customers subscribed to the term deposit. In the next stages, the dataset shall be prepared for summarisation and further analysis to understand what trends affect customer decision.

**Data Preparation :**  The dataset was summarised using the summary function and skim function.

A picture containing text

Description automatically generatedIt was observed that the age variable had certain outlier values and the “pdays” variable contained 40 missing values, as shown in figure 1(a) and 1(b).

A screenshot of a computer

Description automatically generated with low confidence

**1(a)** **1(b)**

The variable pdays displays the number of days that has passed since the customer was last contacted. The missing pdays values were replaced with “999”, which indicates that the client was not previously contacted.

The age variable was plotted to identify the outliers. [Figure2].

Chart, box and whisker chart

Description automatically generatedThe age for the customers were restricted between 18 and 85 years and the outliers were removed.

Certain categorical variables were also factorised in order to make grouping operations and visualisations more concise.

**Figure 2**

**Data Visualisation :**  Any statistical analysis is impossible to complete without the use of data visualisation (Waskom, 2021). A scientist will be able to comprehend their own data and share their discoveries with others by using effective visualisations (Tukey, 1977).

Chart, box and whisker chart

Description automatically generatedIt was observed that the median employee variation rate was very low for customers who had subscribed to term deposit in comparison to the ones who had not.

Employment variation rate serves as a macroeconomic indicator (Chen et al., 2014). It has a negative effect on customers' purchasing decisions, which means a significant change in the employment rate makes **Figure 4**

customers less likely to sign up for a term deposit. This makes sense and shows that a steady employment rate denotes a steady economic environment, in which bank clients are more confident to make financial investments (Chen et al., 2014).

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generatedThe median duration was observed to be higher with unmarried customers who had subscribed to the product. The customers whose marital status was unknown also had a higher median duration of conversation during campaign as depicted in Figure 5.

**Figure 5**

It was also observed that for customers with no credit default the maximum number of subscriptions were between the age of 30 and 40 years old. [See Figure 6]

**Figure 6**

Chart, box and whisker chart

Description automatically generatedChart, bar chart

Description automatically generated

**Figure 7(a) Figure 7(b)**

The duration of phone call remains consistent irrespective of the contact (Figure 7(a)). Additionally, it was observed that most subscribers were contacted over cellular medium than telephone (Figure 7(b)).

The visualisations have given important insights which have been utilised in enhancing the models. Furthermore, the hypotheses test results have been discussed in the following sub-section to get a clearer idea on the dependencies of the predictors.

**Hypotheses test results :**

**H1 :** T-test was performed to find the difference in mean of duration of conversation between two group of customers, who subscribed and the ones who did not.

Text

Description automatically generated

P-value is **<0.05** which proves that the difference in mean duration between these two groups are significant, thus the null hypotheses can be rejected. A high magnitude of t-value **(-55.393)** also indicates a significant difference in the mean.

Text

Description automatically generated**H2 :** Chi-squared test was performed to determine the correlation between having a housing loan and subscribing to term deposit. The p-value **(0.015)** is <0.05 thus there is a significant relation between housing and subscribed and null hypotheses can be rejected. The chi-squared value **(5.9126)** is >3.84, which proves that there is a strong relationship between housing and subscribed.

**H3 :** Chi-squared test was performed to determine the correlation between default and subscribed.

Text

Description automatically generatedThe p-value is <0.05, thus there is a significant relation between default and subscribed, and the null hypotheses can be rejected. A high chi-squared value **(412.84)** also proves that the null hypotheses can be rejected.

**H4 :**  T-test was performed to check the significance of difference in the mean age between the customers who subscribed to term deposit and the ones who did not.

Text

Description automatically generated

The p-value is <0.05 which rejects the null hypotheses and proves that the difference in mean age across two groups is not equal to zero. The t-statistic has a moderate magnitude which proves that there is some difference in the mean in ages across two subscription groups.

**H5 :** T-test was performed to find the difference in mean of euribor3m between two group of customers, who subscribed and the ones who did not. The interest rate at which a number of European banks lend money to one another in euros with a 3-month duration is known as the 3-month Euribor interest rate.

Text

Description automatically generated

The p-value is <0.05 thus the null hypotheses can be rejected, and it can be proved that the difference in mean euribor3m is significant across two groups. The t-statistic also has a high magnitude which proves that there is significant difference in mean across two groups of subscription.

In the next sub-section, logistic regression technique have been used to devise and describe a model that predicts the probability of a customer subscribing to the term deposit scheme.

**Regression model testing and discussion :** When there are several explanatory variables, odds ratio is calculated using logistic regression. With the exception of the binomial response variable, the process is very similar to multiple linear regression. The outcome is how each variable affects the odds ratio of the observed important event (Sperandei, 2014) .

Three regression models were tested, and the final model (model3) with the best fitting possibility has been discussed in detail. The remainder models have been briefly overviewed. The assumptions were tested for multicollinearity, homoscedasticity, residual distribution, auto-correlation and outlier influence, and none of the models violate the assumption tests. [See Appendix]

**Model1 :**  This model was created by considering duration, age and default as predictor variables.

Text

Description automatically generated

Null deviance : 23034 Residual deviance : 19143.

Null deviance is greater than the residual deviance, which is an optimistic result for the fit of the model.

Text

Description automatically generatedThe pseudoR2 was calculated for this model and it was found to be:

Hosmer and Lemeshow – 0.192

Cox and Snell – 0.126

Nagelkerke – 0.25

Only moderate correlation can be explained

using this model, as can be interpreted from the R2  values.

The odds ratio for all predictors in this model apart from default(“yes”) is greater than 1 which indicates a direct proportionality between the predictors and target variable.

Log likelihood : -9571.51 Deviance statistic : 19143 AIC : 19151 Mean vif : 1.001

**Model 2 :** This model was created by considering duration, age, consumer price index, euribor3m, pdays and marital status as predictor variables.

Text

Description automatically generated

Null deviance : 23034 Residual deviance : 14788

Text

Description automatically generatedThe pseudoR2 was calculated for this model and it was found to be:

Hosmer and Lemeshow – 0.358

Cox and Snell – 0.22

Nagelkerke – 0.44

The R2 values for this model are much better than the previous model as it explains more correlation between the predictors and target variable.

The odds ratio for all predictors in this model except euribor3m is greater than 1 which indicates a direct proportionality between the predictors and target variable.

Log likelihood : -7393.841 Deviance statistic : 14788 AIC : 14806 Mean vif : 1.001

**Model 3 :**  This model was created by considering duration, age, consumer price index, nr.employed, contact, campaign, poutcome marital status as predictor variables.

Text

Description automatically generated with medium confidenceNull deviance : 23034

Residual deviance : 14387

Text

Description automatically generated

The pseudoR2 was calculated for this model and it was found to be:

Hosmer and Lemeshow – 0.378

Cox and Snell – 0.233

Nagelkerke – 0.462

The R2 values for this model are better than the previous models as it explains more correlation between the predictors and target variable.

The odds ratio for all predictors in this model except contact.telephone is greater than 1 which indicates a direct proportionality between the predictors and target variable.

Log likelihood : -7193.468 Deviance statistic : 14387 AIC : 14411 Mean vif : 1.25

**DISCUSSION AND CONCLUSION**

Table

Description automatically generatedThree models were tested in order to find the best fit and accuracy for the given data set.Model 3 has the lowest AIC value of all the three models, the pseudo R2 and deviance statistics for model 3 also show how it would be a better fit in comparison to the other models.

Also, only 749 observations i.e 2.2% data lie outside 1.96, as was discovered upon analysing the standardised residuals which proves that this model is a good fit. This model also showed a predictive accuracy of 90.87% . Mode of communication was observed to be more effective for cellular phones and customer with a higher duration of conversation had greater chances of subscription.

Sensitivity analysis is essential for determining a predictive model's robustness as well as for revealing important details about the model's structure and dependence on the input variables (Saltelli et al., 2000).However, this model has a moderate kappa value of 0.42 which risks the reliability of this model in terms of application. From inspection it can be concluded that a correlation of 0.51 between consumer price index and employment rate of the bank might contribute to this factor.

**REFERENCE LIST**

Aron, J., Duca, J.V., Muellbauer, J., Murata, K. and Murphy, A., 2012. Credit, housing collateral, and consumption: Evidence from Japan, the UK, and the US. *Review of Income and Wealth*, *58*(3), pp.397-423.

Asare-Frempong J., and Jayabalan, M., 2017, September. Predicting customer response to bank direct telemarketing campaign. In *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)* (pp. 1-4). IEEE.

Chen, J., Han, Y., Hu, Z., Lu, Y., & Sun, M. (2014). Who will subscribe a term deposit? Columbia University, Available at : <http://www.columbia.edu/~jc4133/ADA-Project.pdf>

Fernandez-Corugedo, E. and Muellbauer, J., 2006. Consumer credit conditions in the United Kingdom. *London: Centre for Banking Studies, Bank of England, and Nuffield College*.

Field, A., Miles, J. and Field, Z., (2012). Discovering statistics using R. *Great Britain: Sage Publications, Ltd*, *958*.

Heyde, F. and Neyer, U., 2010. Credit default swaps and the stability of the banking sector. *International Review of Finance*, *10*(1), pp.27-61.

Hou, S., Cai, Z., Wu, J., Du, H. and Xie, P., 2022. Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription. *International Journal of Business Analytics (IJBAN)*, *9*(1), pp.1-12.

Ilham, A., Khikmah, L. and Iswara, I.B.A.I., 2019, March. Long-term deposits prediction: a comparative framework of classification model for predict the success of bank telemarketing. In *Journal of Physics: Conference Series* (Vol. 1175, No. 1, p. 012035). IOP Publishing.

Ling, X. and Li, C. (1998) ‘Data Mining for Direct Marketing: Problems and Solutions’. Proceedings of the 4th KDD conference, AAAI Press, pp.73–79.

Moro S., Cortez P., and Rita P., 2014. "A data-driven approach to predict the success of bank telemarketing", Decision Support Systems, vol. 62, pp. 22-31.

Ou, C., Liu, C., Huang, J. & Zhong, N. (2003) ‘On Data Mining for Direct Marketing’. Proceedings of the 9th RSFDGrC conference, 2639, pp.491–498.

Palaniappan, S., Mustapha, A., Foozy, C.F.M. and Atan, R., 2017. Customer profiling using classification approach for bank telemarketing. *JOIV: International Journal on Informatics Visualization*, *1*(4-2), pp.214-217.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). Sensitivity analysis in practice: A guide to assessing scientific models. John Wiley and Sons.

Sperandei, S., 2014. Understanding logistic regression analysis. *Biochemia medica*, *24*(1), pp.12-18.

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley publication.

Vaughn B. K. and Wang Q., 2008. "Classification based on tree-structured allocation rules", The Journal of Experimental Education, vol. 76, no. 3, pp. 315-340.

Waskom, M.L., 2021. Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), p.3021.

Zhuang, Q.R., Yao, Y.W. and Liu, O., 2018. Application of data mining in term deposit marketing. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 2).

**APPENDIX**

setwd("C:/Users/Tathagata/OneDrive/Desktop/MGT\_7177\_A2/R\_A2")

library(readxl)

data<- read\_excel("banksv.xlsx")

library(tidyverse)

library(dplyr)

library(caret)

library(skimr)

library(ggplot2)

install.packages("psych", dependencies = T)

library(psych)

install.packages("car")

library(car)

#Data Exploration

View(data)

summary(data) #summarising data

skim(data) # summarising data using skim function

qplot(mapping = aes(x=as.factor(data$subscribed),y=data$age),data,geom=c("boxplot","point"), xlab = "Subscibed", ylab = "Age (in years)", color = data$subscribed)

#plotting the age for each subscription group

table(data$subscribed)

table(datacln$month) #tabulating all categorical variables to gain insight on the existing levels.

table(datacln$day\_of\_week)

table(datacln$default)

table(datacln$housing)

table(datacln$loan)

table(datacln$job)

table(datacln$contact)

table(datacln$previous)

table(datacln$poutcome)

table(datacln$subscribed)

table(datacln$cons.conf.idx)

table(datacln$marital)

#Data Cleaning

datacln <- data %>% replace(is.na(.),999) #replacing na values with 999

datacln <- datacln %>% filter(age>18&age< 85) #removing age outliers from the dataset

datacln$month <- as.factor(datacln$month) #factorising the categorical variables for ease of computation

levels(datacln$month)[7] <- "mar"

datacln$day\_of\_week <- as.factor(datacln$day\_of\_week)

datacln$default <- as.factor(datacln$default)

levels(datacln$default)[1] <- "no"

datacln$housing <- as.factor(datacln$housing)

datacln$loan <- as.factor(datacln$loan)

datacln$job <- as.factor(datacln$job)

datacln$contact <- as.factor(datacln$contact)

datacln$previous <- as.factor(datacln$previous)

datacln$poutcome <- as.factor(datacln$poutcome)

datacln$marital <- as.factor(datacln$marital)

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

datacln$education <- as.factor(datacln$education)

datacln$default <- as.factor(datacln$default)

View(datacln)

str(datacln)

skim(datacln)

#visualisation of trends

ggplot(data = datacln, mapping= aes(x=datacln$subscribed, y=datacln$emp.var.rate,color=datacln$subscribed))+ geom\_boxplot() + labs(x="Subscribed", y = "employee variation rate ") + ggtitle("Subscribed and Employee Variation rate")

ggplot(data = datacln, mapping= aes(x=datacln$subscribed,y=datacln$duration, color=datacln$subscribed)) + geom\_boxplot() + facet\_wrap(~default) + labs(x="Subscribed", y = "Duration (in seconds) ") + ggtitle("Duration and Subscribed, by marital status")

ggplot(data = datacln, mapping= aes(x=datacln$subscribed,y=datacln$age, color=datacln$subscribed)) + geom\_boxplot() + facet\_wrap(~marital)+ labs(x=" Subscribed", y = "Age(in years)") + ggtitle("Subscription by Age and marital status")

ggplot(data = datacln, mapping= aes(y=datacln$duration, x=datacln$subscribed, fill=datacln$subscribed)) + geom\_boxplot() + facet\_wrap(~contact) +labs(x="Subscribed", y = "Duration (In seconds") + ggtitle("Duration and subscription by contact")

ggplot(data = datacln, mapping= aes(x=datacln$subscribed, fill=datacln$subscribed)) + stat\_count(width = 0.5) + facet\_wrap(~contact) +labs(x="Subscribed", y = "Count") + ggtitle("Subscription by contact")

ggplot(data = datacln, mapping= aes(x=datacln$age, fill=subscribed)) + stat\_count(width=0.5) + facet\_wrap(~default) + labs(x="Age (in years) ", y = "Count") + ggtitle("Age distribution by subscription and default")

##Hypotheses bivariate analysis.

#H1 -- Substantial diff in mean of duration between y&n customers

describeBy(datacln$duration, group = datacln$subscribed)

t.test(duration ~ subscribed, data = datacln, conf.level= 0.95, na.action=na.exclude) #performing t-test

#H2 -- High correlation between customers having a housing loan and y&n term deposit.

table(datacln$housing,datacln$subscribed)

chisq.test(datacln$housing,datacln$subscribed, correct = F) #performing chi-squared test

pchisq(5.9126,1,lower.tail = F)

#H3 -- strong relation between default and subscription

table(datacln$default,datacln$subscribed)

chisq.test(datacln$default,datacln$subscribed, correct = F) #performing chi-squared test

pchisq(412.84,1,lower.tail = F)

#H4 -- Strong relationship between age and subscription

describeBy(datacln$age, group = datacln$subscribed)

t.test(age ~ subscribed, data = datacln, conf.level= 0.95, na.action=na.exclude) #performing t-test

#H5 -- Significant difference in mean euribor3m between the customers who subscribed and who didnot.

describeBy(datacln$euribor3m, group = datacln$subscribed)

t.test(euribor3m ~ subscribed, data = datacln, conf.level= 0.95, na.action=na.exclude) #performing t-test

##Partitioning the data

set.seed(40379173)

index <- createDataPartition(datacln$subscribed,p=0.8,list=F)

train <- datacln[index,]

test <- datacln[-index,]

logisticPseudoR2s <- function(LogModel) { #preparing function for calculating pseudoR2 for models

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")

}

##model1

fm1 <- subscribed ~ duration + age + default

model1 <- glm(fm1, data = train, family="binomial")

summary(model1)

logisticPseudoR2s(model1) #calculating pseudoR2 for model

exp(model1$coefficients) #finding odd ratio of predictors

logLik(model1) #calculating log likelihood

devst1 <- -2 \* logLik(model1) #calculating deviance statistics

vif(model1) #calculating variance inflation factor

mean(vif(model1))

plot(model1) #plotting the model to check for assumptions

c1 <- cooks.distance(model1)

sum(c1>1) #checking for cook's distance

train$Standardisedresiduals <- rstandard(model1) #standardising residuals

train$Studentisedresiduals <- rstudent(model1)

sum(train$Standardisedresiduals > 1.96)

prediction <- predict(model1, test, type="response") #checking predictive accuracy of the model

pred\_data <- as.factor(ifelse(prediction > 0.5,"yes","no"))

postResample(pred\_data,test$subscribed)

confusionMatrix(data = pred\_data,test$subscribed)

cor(train$cons.price.idx,train$nr.employed)

##model2

fm2 <- subscribed ~ duration + age + cons.price.idx + euribor3m + pdays + marital

model2 <- glm(fm2, data = train, family="binomial")

summary(model2)

logisticPseudoR2s(model2)

exp(model2$coefficients)

logLik(model2)

devst2 <- -2 \* logLik(model2)

vif(model2)

mean(vif(model2))

plot(model2)

c1 <- cooks.distance(model2)

sum(c1>1)

train$Standardisedresiduals <- rstandard(model3)

train$Studentisedresiduals <- rstudent(model3)

sum(train$Standardisedresiduals > 1.96)

prediction <- predict(model2, test, type="response")

pred\_data <- as.factor(ifelse(prediction > 0.5,"yes","no"))

postResample(pred\_data,test$subscribed)

confusionMatrix(data = pred\_data,test$subscribed)

cor(train$cons.price.idx,train$nr.employed)

##model3

fm3 <- subscribed ~ duration + age + cons.price.idx + nr.employed + marital + contact + campaign + poutcome

model3 <- glm(fm3, data = train, family="binomial")

summary(model3)

logisticPseudoR2s(model3)

exp(model3$coefficients)

logLik(model3)

devst3 <- -2 \* logLik(model3)

vif(model3)

mean(vif(model3))

plot(model2)

c1 <- cooks.distance(model1)

sum(c1>1)

train$Standardisedresiduals <- rstandard(model3)

train$Studentisedresiduals <- rstudent(model3)

sum(train$Standardisedresiduals > 1.96)

prediction <- predict(model3, test, type="response")

pred\_data <- as.factor(ifelse(prediction > 0.5,"yes","no"))

postResample(pred\_data,test$subscribed)

confusionMatrix(data = pred\_data,test$subscribed)

cor(train$cons.price.idx,train$nr.employed)

##Responses of all three models have been tabulated using the stargazer function

install.packages("stargazer")

library(stargazer)

stargazer(model1,model2,model3,type="text",title="Model1: Results",align = T,out="model.html")