Portuguese Banking Institution

Project Report

Dana 4820 -Predictive Analytics (Qualitative variable)

Ayushi Singh
Lien Pham
Priya Yadav
Sai Prasanna

1. Project Abstract

The project is about a Portuguese bank that consists of data related to the characteristics of clients such as their ages, marital status, education, current balances, and others. The data also consists of the marketing campaign activities applied to a client, specifically the result of the previous campaign, the number of contacts for that client in the campaign, and other related information.

2. Research question

To find out the focus groups who are most likely to deposit based on their characteristics such as age, education, marital status, credit loan, etc., and based on the result of the marketing campaign, the number of bankers contacting the client in the campaign and others related.

3. Data Introduction

4512 people of data related to the direct marketing campaign of a Portuguese banking institution. The marketing campaigns were based on phone calls with the aim of assessing if the client would subscribe to a bank term deposit or not.

3.1. Variable description

Variable	Type	Description
age	Numerical	Range (19, 87)
job	Categorical (12)	‘admin.’, ‘blue-collar’, ‘entrepreneur’, ‘housemaid’, ‘management’, ‘retired’, ‘self-employed’, ‘services’, ‘student’, ‘technician’, ‘unemployed’, ‘unknown’
marital	Categorical (2)	yes/no
education	Categorical (4)	(‘primary’, ‘secondary’, ‘tertiary’, ‘unknown’)
default	Categorical (2)	has credit in default? yes/no
balance	Numerical	average yearly balance, in euros | Range (-3313, 71188)
housing	Categorical (2)	has housing loan (yes/no)
loan	Categorical (2)	has personal loan? (’no’, ’yes’) related with last contact of current campaign.
day	Categorical	last contact day of the month 1 (1-31)
month	Categorical	last contact month of year (Jan- Dec)
campaign	Numerical	number of contacts performed during this campaign and for this client (1, 50)
pdays	Numerical	Number of days passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted) | Range (-1, 871)
previous	Numerical	number of contacts performed before this campaign and for this client (‘Success’, ‘failure’, ‘other’, ‘unknown’)
poutcome	Categorical (2)	outcome of the previous marketing campaign (yes/no)
Y	Categorical (2)	Output variable - has the client subscribed a term deposit? (‘yes’, ‘no’)

4. Methodology

To approach the research question, we plan to do logistic regression on the data set. Since, the response variable in our data set is "y", which is of the form binary variable (yes/no), and the predicting variables are either quantitative or categorical. Hence best model fitting can be achieved by logistic regression.

4.1. Initial Variable selection

4.1.1. Checking significance using chi-square test and t-test

To check the significance of explanatory variables on the response variable, based on variable type, we can run either a chi-square test or a two-sample t-test. If through these tests, the variable shows either different means in two samples or association with the response variable; it indicates that variable does have a significant impact on the response variable.

Two sample test is applicable to numerical variables like age, balance, campaign, pdays, and previous. In this test since the response variable is categorical with 2 categories, we divide the numerical variable into two groups to first check the variance in two groups, and based on variance we then finally check whether the mean of two groups is the same or not. If the two groups' mean are not the same, it means it has a significant effect on the response variable. Statistically, the hypothesis for the t-test would be as follows:

>check variance
var.test(groupA, groupB)
H(0) = Variances of group A is equal to variance of group B
H(a) = Variance of group A is not equal to variance of group B

>t_test
H(0) = The difference in group means is zero
H(a) = The difference in group means is different from zero

Similarly, we have ***Chi-Square*** test. It is applicable to *categorical variables* like marital, default, housing, loan, poutcome, education, and job. In this test, we run the test on the response variable and categorical explanatory variable. If two variables are not independent of each other it means they show significant association. Statically, following is the hypothesis for chi-square test is as follows:

Chi-square test-
H(0) = The variables are independent of each other
H(a) = The variables are not independent of each other (shows association)

Below is the summarization of the test conducted on all variables:

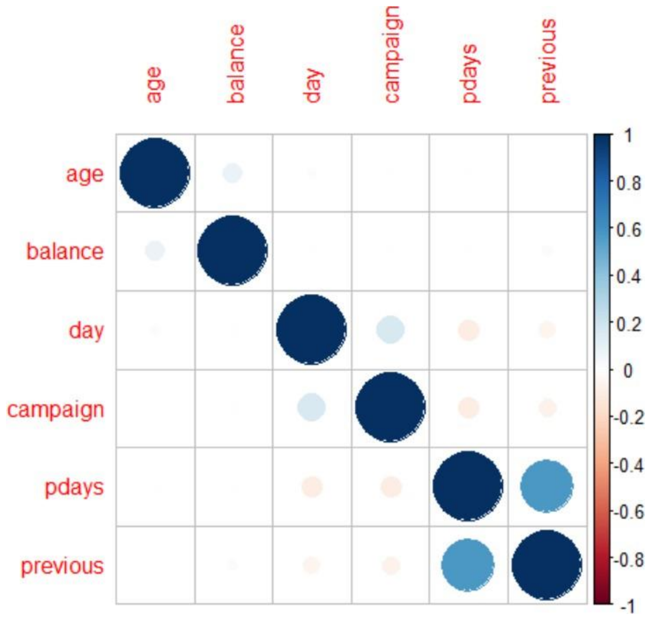
Variable	**Test conducted**	**Test Result**
age	2 sample T-test	Based on p-value the variance and mean of two groups are different
balance	2 sample T-test	Based on p-value the variance and mean of two groups are different
campaign	2 sample T-test	Based on p-value the variance and mean of two groups are different

pdays	2 sample T-test	Based on p-value the variance and mean of two groups are different
previous	2 sample T-test	Based on p-value the variance and mean of two groups are different
job	Chi-square test	Based on p-value the variables show association
marital	Chi-square test	Based on p-value the variables show association
education	Chi-square test	Based on p-value the variables show association
default	Chi-square test	Based on p-value the variables show association
housing	Chi-square test	Based on p-value the variables show association
loan	Chi-square test	Based on p-value the variables show association
day	Chi-square test	Based on p-value the variables show association
month	Chi-square test	Based on p-value the variables show association
poutcome	Chi-square test	Based on p-value the variables show association

4.1.2. Correlation check

	age	**balance**	**day**	**campaign**	**pdays**	**previous**
age	1.0000	0.0838	-0.0179	-0.0051	-0.0089	-0.0035
balance	0.0838	1.0000	-0.0087	-0.0100	0.0094	0.0262
day	-0.0179	-0.0087	1.0000	0.1607	-0.0944	-0.0591
campaign	-0.0051	-0.0100	0.1607	1.0000	-0.0931	-0.0678
pdays	-0.0089	0.0094	-0.0944	-0.0931	1.0000	0.5776
previous	-0.0035	0.0262	-0.0591	-0.0678	0.5776	1.0000

As you can see from the table above and the correlation plot below there are no highly correlated variables that can significantly affect the response variable hence no variable to is excluded based on correlation.



4.1.3. Multicollinearity check

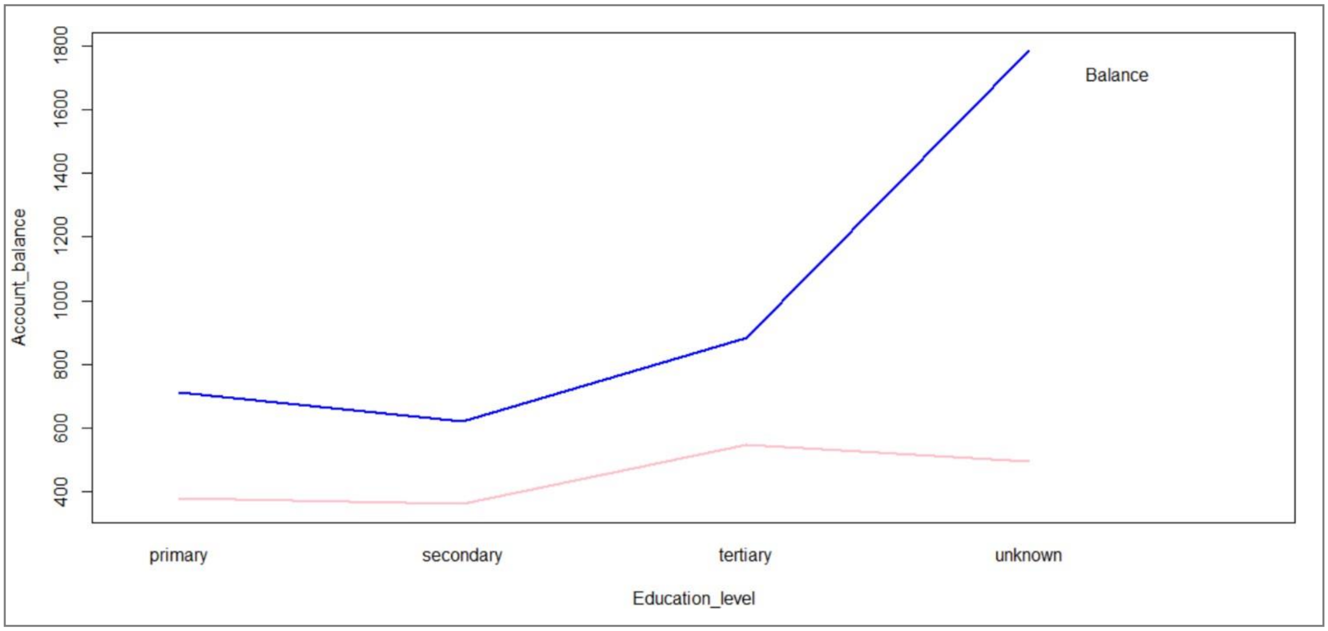
	GVIF	Df
age	2.060359	1
job	4.412429	11
marital	1.420881	2
education	2.343245	3
default	1.02456	1
balance	1.059505	1
housing	1.410971	1
loan	1.048124	1
day	1.333652	1
month	2.662009	11
campaign	1.122278	1
pdays	3.732927	1
previous	1.870314	1
poutcome	5.33572	3

In the multicollinearity check also, we didn't receive any variables with VIF higher than 10. So, again we don't have any variable to exclude at this stage either.

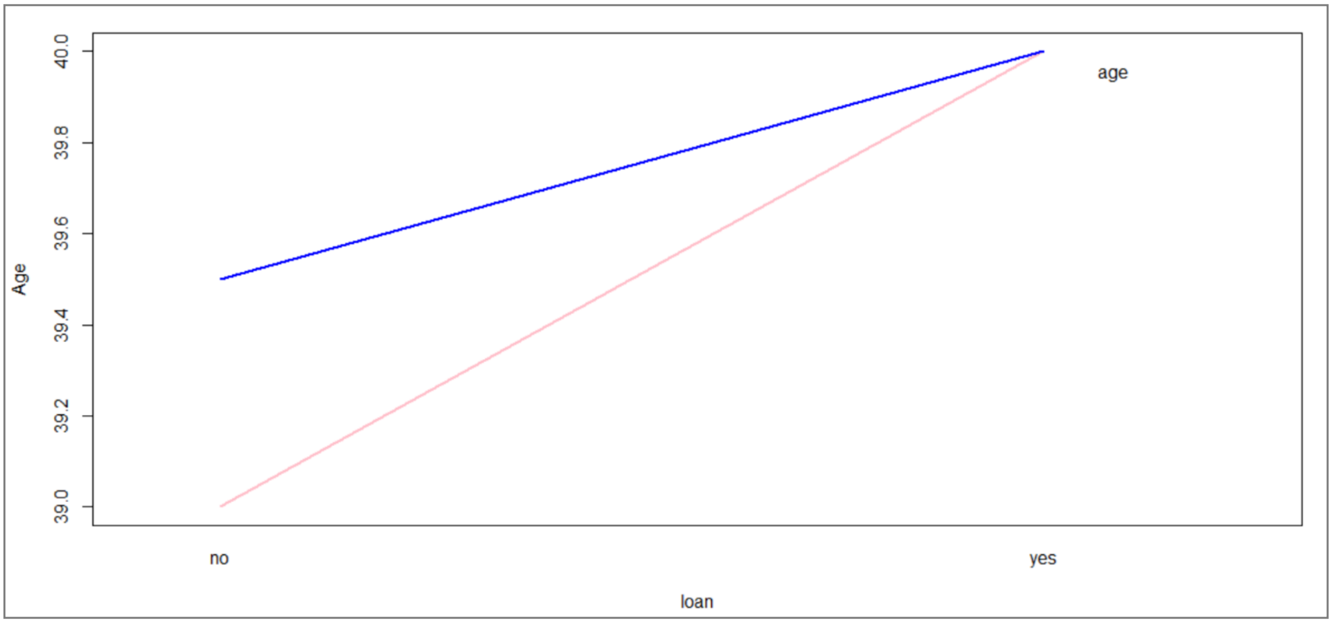
4.2. Checking Interaction

Since the data set has categorical variables (factors), we took the function as a median for the interaction plots. Following are the few interactions we checked randomly based on our logical understanding of the relationship between the variables.

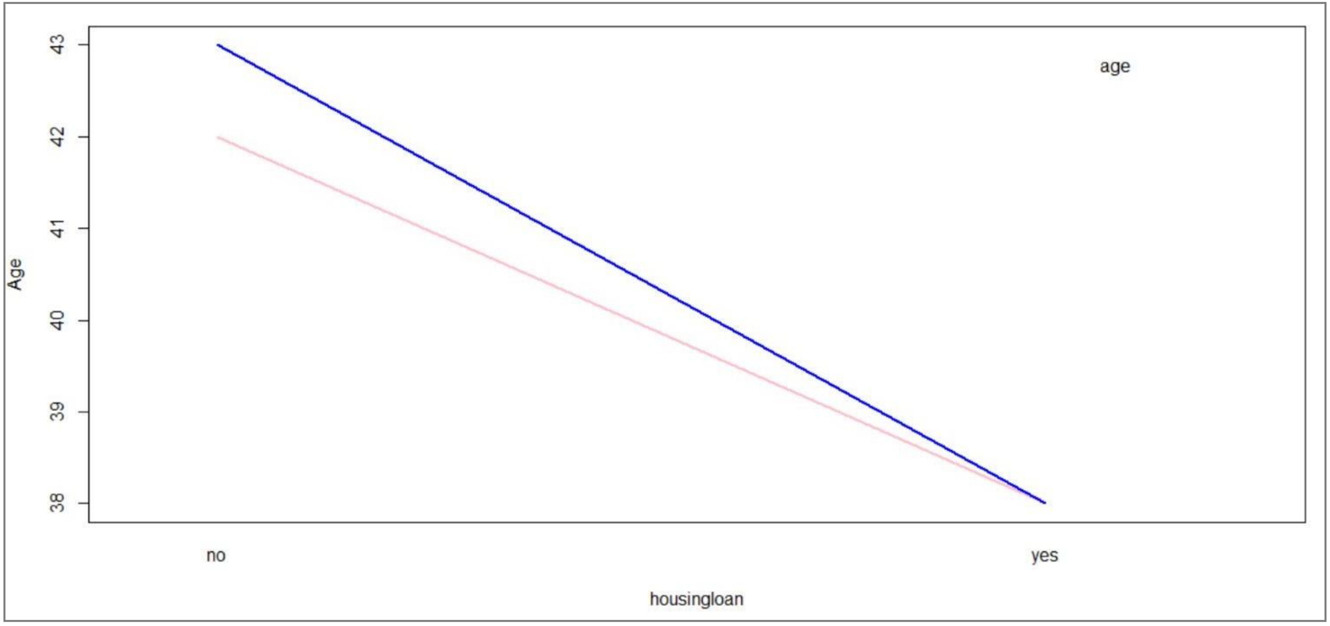
i) Education vs Balance



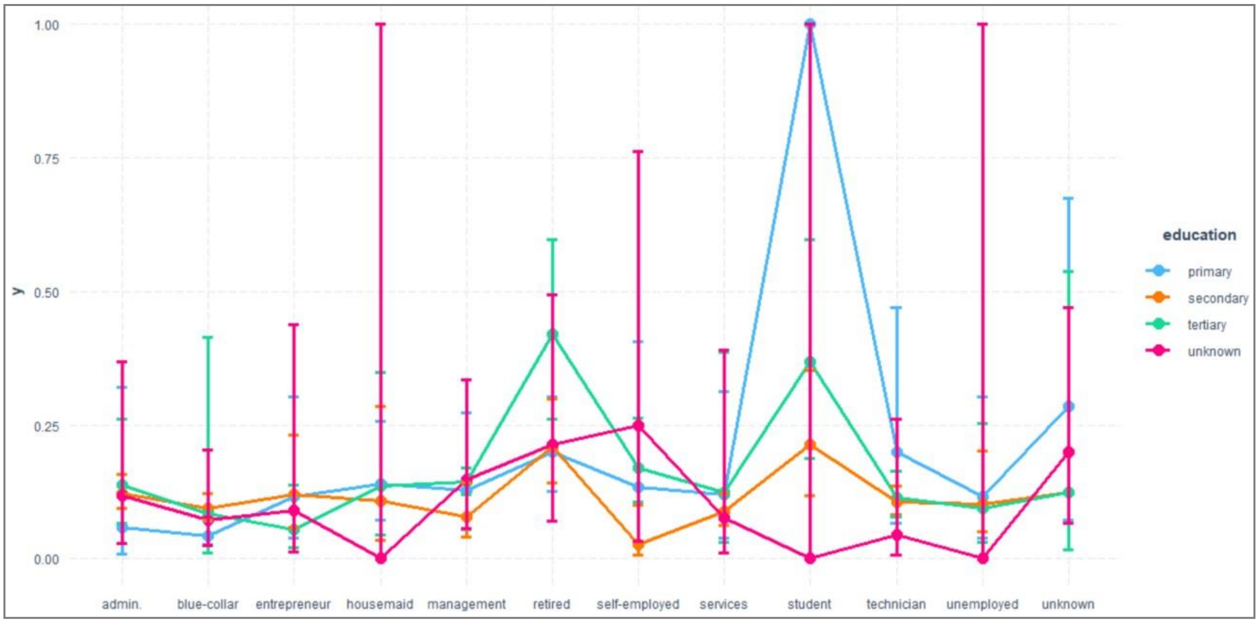
ii) Age vs personal loan



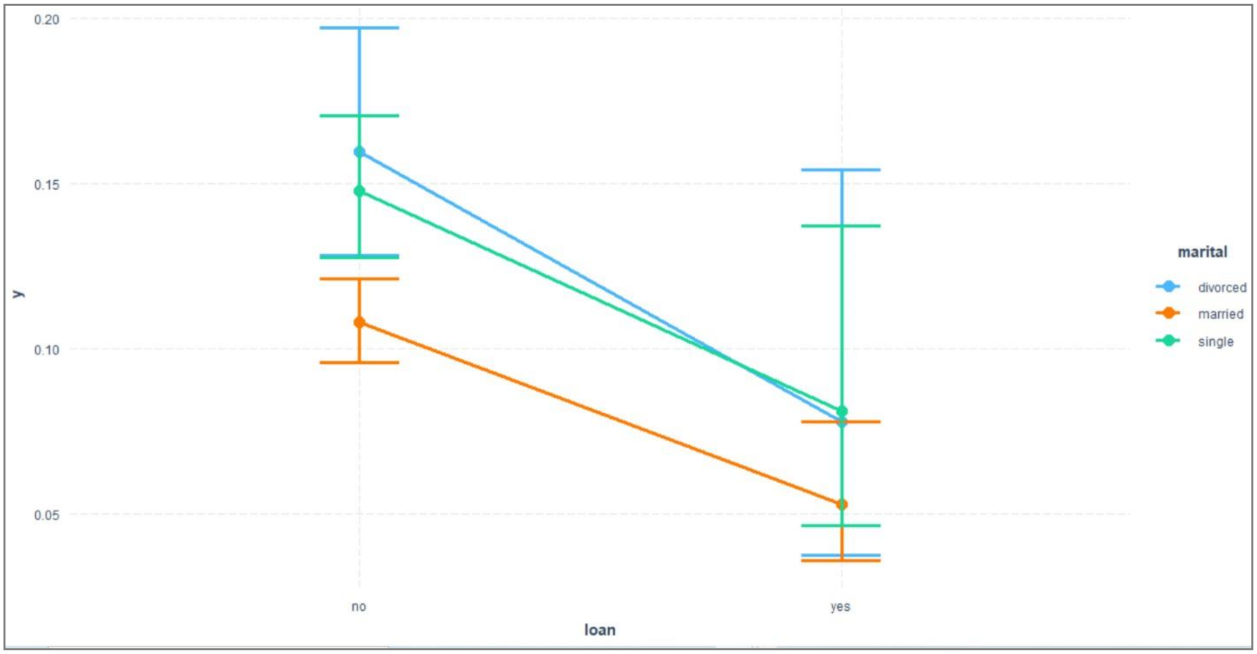
iii) Age vs Housing Loan



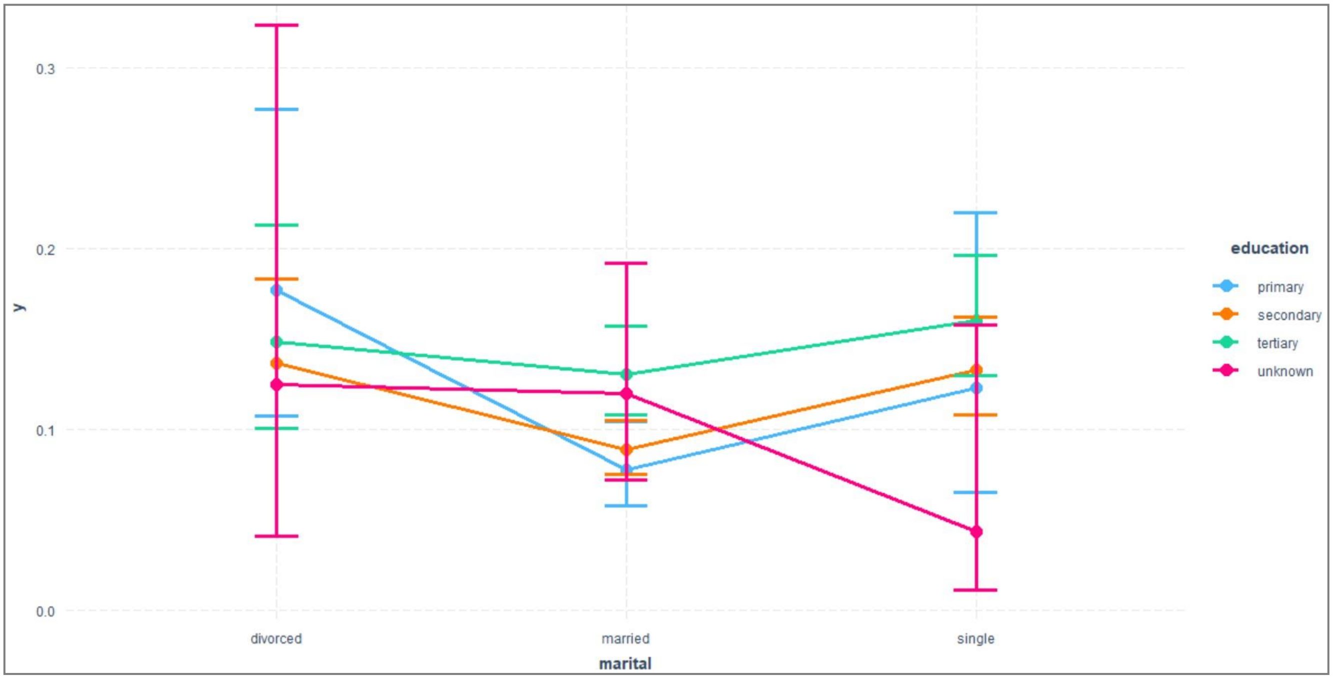
iv) Job vs Education



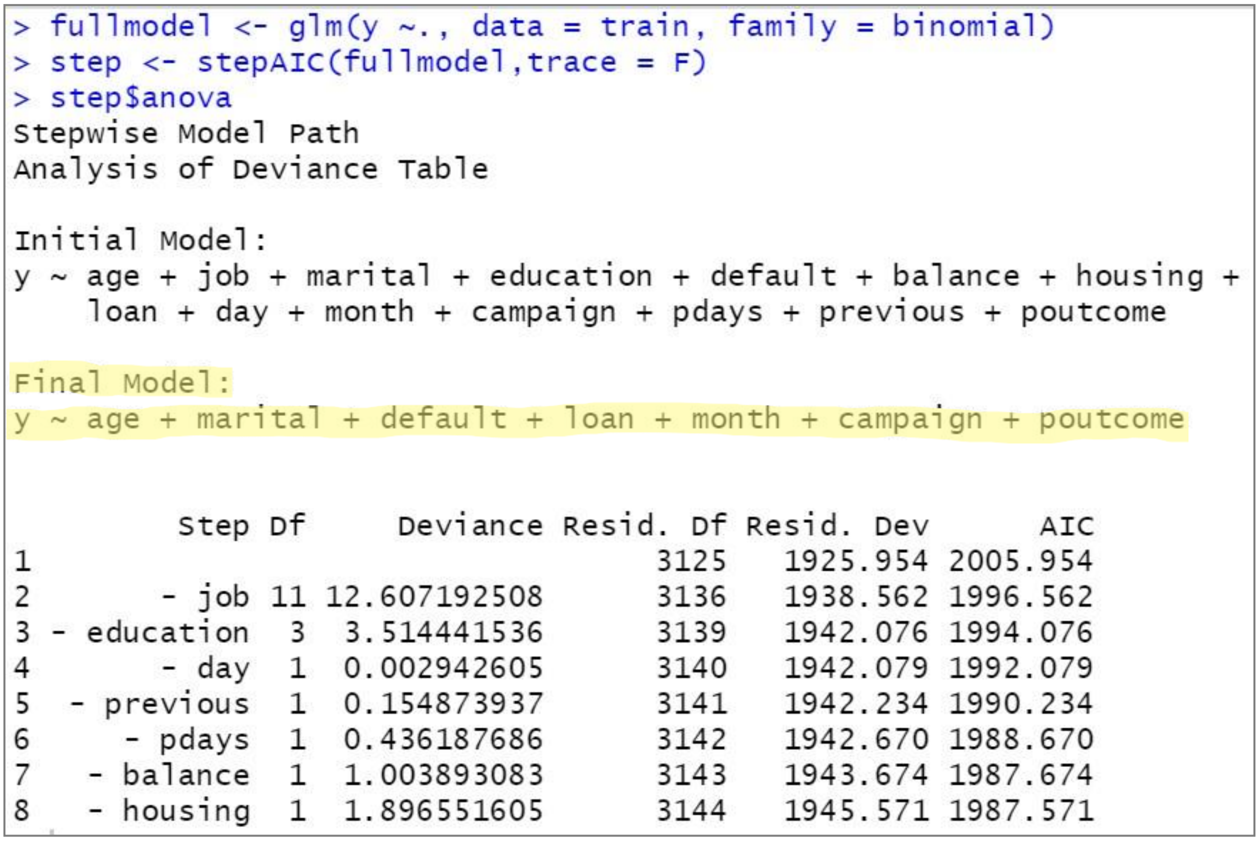
v) Personal loan vs Marital



vi) Education vs Marital



4.3. Stepwise regression on training data set



```
> fullmodel <- glm(y ~., data = train, family = binomial)
> step <- stepAIC(fullmodel,trace = F)
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
y ~ age + job + marital + education + default + balance + housing +
    loan + day + month + campaign + pdays + previous + poutcome

Final Model:
y ~ age + marital + default + loan + month + campaign + poutcome


         Step Df     Deviance Resid. Df Resid. Dev      AIC
1                                  3125   1925.954 2005.954
2       - job 11 12.607192508      3136   1938.562 1996.562
3 - education  3  3.514441536      3139   1942.076 1994.076
4       - day  1  0.002942605      3140   1942.079 1992.079
5  - previous  1  0.154873937      3141   1942.234 1990.234
6     - pdays  1  0.436187686      3142   1942.670 1988.670
7   - balance  1  1.003893083      3143   1943.674 1987.674
8   - housing  1  1.896551605      3144   1945.571 1987.571
```

After removing “default” variable during initial variable selection, we have run stepwise regression on our full model. As we can see after running stepwise regression the full model has been reduced to the final model with only 7 explanatory variables as highlighted.

4.4. Comparing model with and without interactions

Based on variable selections and stepwise regression the final model we achieved

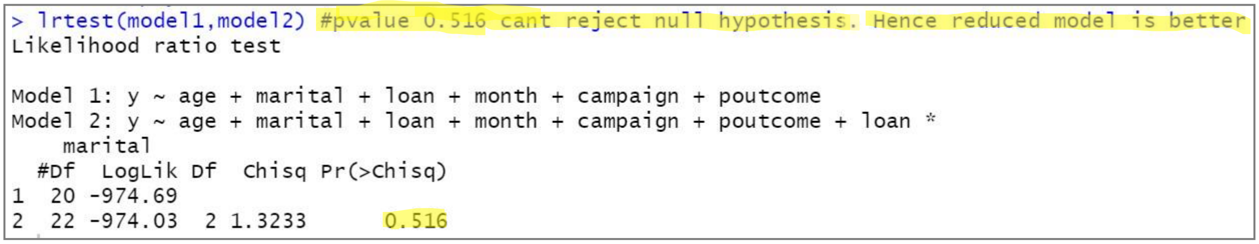
Final Model:

y ~ age + marital + default + loan + month + campaign + poutcome

Based on interactions we performed, let’s observe the effect of the interaction of loan & marital on the above model. Let’s test the significance effect based on log-likelihood ratio test and the wald test.

4.4.1. Loglikelihood ratio test

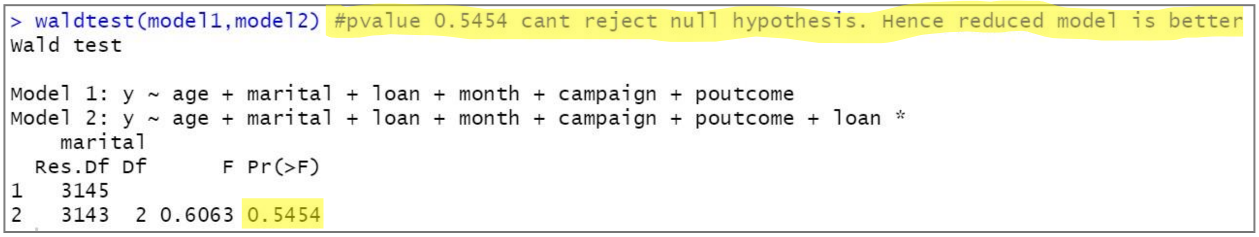
Hypothesis for the test:
H(0) - reduced model is better - model1
H(a) - full model is better - model2



```
> lrtest(model1,model2) #pvalue 0.516 cant reject null hypothesis. Hence reduced model is better
Likelihood ratio test

Model 1: y ~ age + marital + loan + month + campaign + poutcome
Model 2: y ~ age + marital + loan + month + campaign + poutcome + loan *
    marital
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  20 -974.69
2  22 -974.03  2 1.3233      0.516
```

4.4.2. Wald test

Hypothesis for the test:
H(0) - reduced model is better - model1
H(a) - full model is better - model2



```
> waldtest(model1,model2) #pvalue 0.5454 cant reject null hypothesis. Hence reduced model is better
Wald test

Model 1: y ~ age + marital + loan + month + campaign + poutcome
Model 2: y ~ age + marital + loan + month + campaign + poutcome + loan *
    marital
  Res.Df Df      F Pr(>F)
1   3145
2   3143  2 0.6063 0.5454
```

From both the wald test and loglikelihood test we can conclude that the reduced model is better, i.e. the one without the interaction.

4.5. Classification report and ROC of the two models

Logistic regression for model 1

```
> model1 <- glm(y ~ age + marital + loan + month + campaign + poutcome, data
= train, family = binomial)
> summary(model1)

Call:
glm(formula = y ~ age + marital + loan + month + campaign + poutcome, 
    family = binomial, data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.2001  -0.4785  -0.3904  -0.3046   2.9930  

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -1.576534   0.399398  -3.947 7.90e-05 ***
age              0.012948   0.005958   2.173 0.029773 *  
maritalmarried  -0.536528   0.178349  -3.008 0.002627 ** 
maritalsingle   -0.129255   0.207534  -0.623 0.533408    
loanyes         -0.852143   0.231878  -3.675 0.000238 ***
monthaug        -0.396233   0.242184  -1.636 0.101823    
monthdec         0.497621   0.643825   0.773 0.439574    
monthfeb        -0.434370   0.301911  -1.439 0.150226    
monthjan        -0.824694   0.366543  -2.250 0.024454 *  
monthjul        -0.646235   0.252305  -2.561 0.010427 *  
monthjun        -0.625762   0.258626  -2.420 0.015539 *  
monthmar         1.127341   0.412182   2.735 0.006237 ** 



monthsep        -0.123052   0.466708  -0.264 0.792042    



poutcomeunknown -0.049028   0.204130  -0.240 0.810191    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2263.0  on 3164  degrees of freedom
Residual deviance: 1949.4  on 3145  degrees of freedom
AIC: 1989.4

Number of Fisher Scoring iterations: 5
```

Interpretation

- If the age increase by one, the odds of client who will likely to deposit will increase e^(0.012948) = 1.01

- The odds of ‘deposit’ for people who are married is only 0.58 (e^-0.536528) the odds of people who are divorced
- The odds of ‘deposit’ for people who has loan is only 0.42 (e^-0.852143) the odds of people who are don’t have loans
- The odds of deposit for the successful campaign (‘poutcomesuccess’) is 12 (e ^2.491701) times the odds of unsuccessful campaigns. However, the odds of deposit for other outcome of the campaign is 2.4 times the odds of unsuccessful campaign
- The odds of deposit for the last month of contact in March is 3.1 (e ^1.127341) the odds of the last month of contact in December
- The betas (coefficients) for the last month of contact in Jan, July, Jun, May, November are negative, that means the odds of deposit of these months are less than the odds of deposit in December. For example, the odds of deposit for a person who is lastly contacted in Jan only 0.43 times the odds of deposit for a person who is contacted in December

Logistic regression for model 2

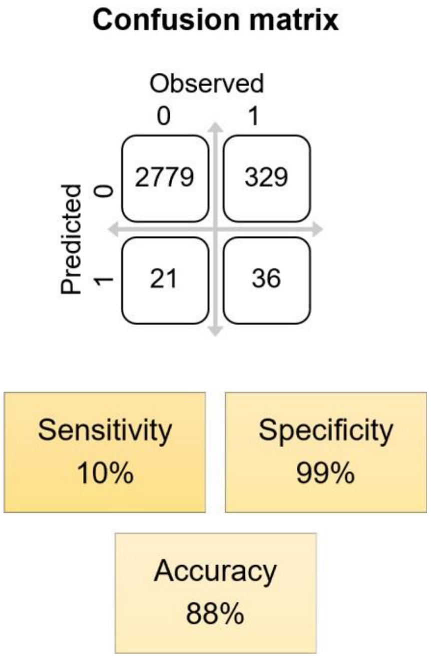
```
model2 <- glm(y ~ age + marital + loan + month + campaign + poutcome + loan*
marital, data = train, family = binomial)
> summary(model2)

Call:
glm(formula = y ~ age + marital + loan + month + campaign + poutcome + 
    loan * marital, family = binomial, data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1997  -0.4773  -0.3870  -0.3039   2.9418  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -1.531020   0.401396  -3.814 0.000137 ***
age                     0.012938   0.005964   2.169 0.030065 *  
maritalmarried         -0.595980   0.185197  -3.218 0.001290 ** 
maritalsingle          -0.167773   0.213804  -0.785 0.432625    
loanyes                -1.446168   0.672179  -2.151 0.031440 *  
monthaug               -0.402079   0.242506  -1.658 0.097314 .  
monthdec                0.482465   0.642631   0.751 0.452794    
monthfeb               -0.447696   0.302356  -1.481 0.138689    
monthjan               -0.834032   0.366470  -2.276 0.022855 *  
monthjul               -0.655648   0.252822  -2.593 0.009506 ** 
monthjun               -0.632379   0.258950  -2.442 0.014602 *  
monthmar                1.136921   0.411479   2.763 0.005727 ** 
monthmay               -1.215044   0.230238  -5.277 1.31e-07 ***
monthnov               -0.865438   0.288968  -2.995 0.002745 ** 
monthoct                1.143424   0.356616   3.206 0.001344 ** 
monthsep               -0.133340   0.466790  -0.286 0.775143    
campaign               -0.054543   0.027136  -2.010 0.044431 *  
poutcomeother           0.898721   0.288947   3.110 0.001869 ** 
poutcomesuccess         2.500209   0.294910   8.478  < 2e-16 ***
poutcomeunknown        -0.041560   0.204295  -0.203 0.838800    
```

```
maritalmarried:loanyes  0.773857   0.730401   1.059 0.289373
maritalsingle:loanyes   0.492158   0.826712   0.595 0.551630
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2263.0 on 3164 degrees of freedom
Residual deviance: 1948.1 on 3143 degrees of freedom
AIC: 1992.1

Number of Fisher Scoring iterations: 5
```

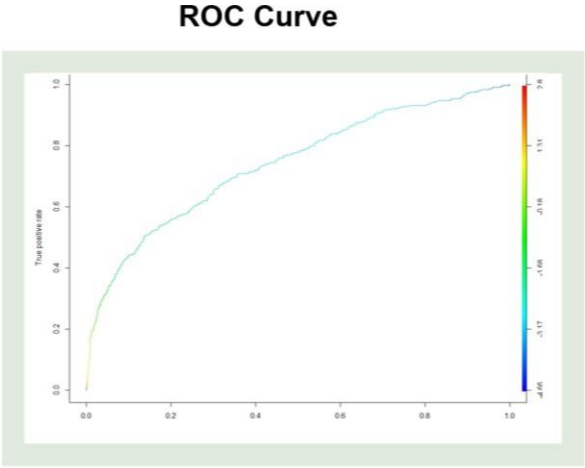
Interpretation:

Essentially, the variables such as ‘age’, ‘marital status’ and others are significant and same as the reduced model (model without interaction)
The interaction ‘maritalstatus*loan” is not significant as the p-value is greater than 0.05

Model 1 accuracy



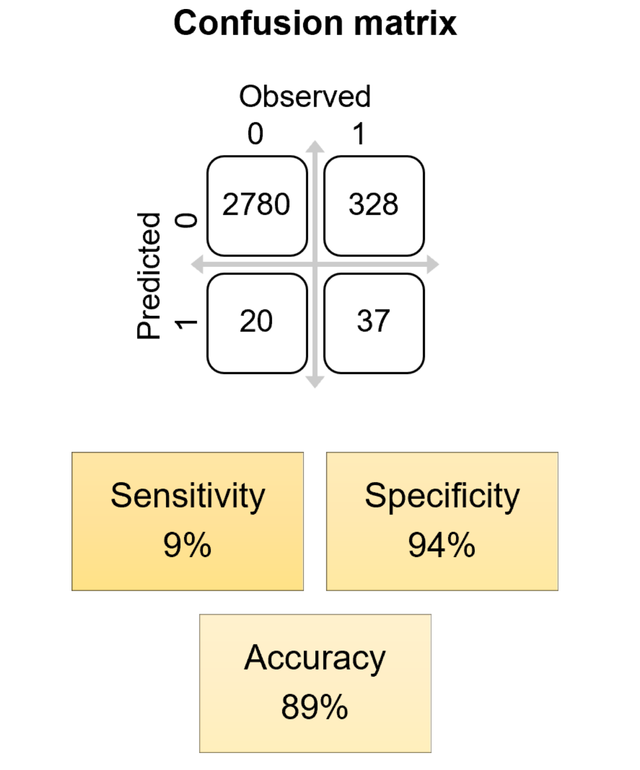
		Observed 0	Observed 1
Predicted	0	2779	329
Predicted	1	21	36

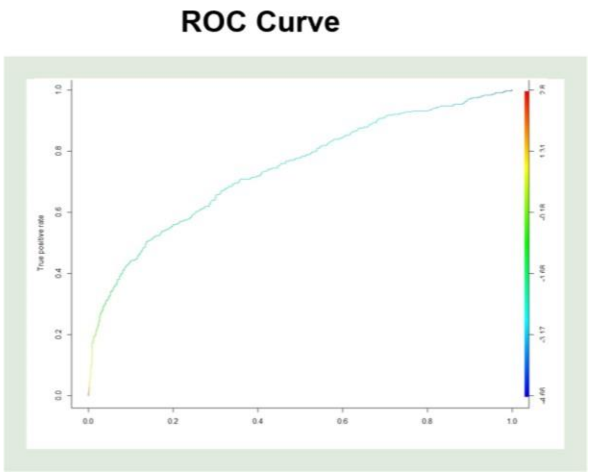


Interpretation:

The accuracy of model 1 is 88%, sensitivity is 10% and specificity is 99%

Model 2 accuracy





Interpretation:

The accuracy of model 1 is 89%, sensitivity is 10% and specificity is 99%

4.6. Lack of fit test for two models

We use Hosmer – Lemeshow to test for ungroup data

The steps are:

1. Estimate the probability of $y_i = 1$ for each observation.
2. Sort the observations into groups (usually 10) with increasing probability.
3. For each group compute the number of observations multiplied by the average probability. This gives the expected number of $y_i = 1$'sin this group.
4. Compute the expected number of $y_i = 0$'s too
5. Compute a chi-square test statistic
6. This test statistic is χ^2_{10-2} distributed

Model 1 **Hosmer – Lemeshow test for ungroup data:** Ho: model fits H1: model does not fit X-squared = 10.296, df = 8, p-value = 0.2448 Failed to reject Ho, therefore data fits model well	**Model 2** **Hosmer – Lemeshow test for ungroup data:** Ho: model fits H1: model does not fit X-squared = 9.493, df = 8, p-value = 0.3024 Failed to reject Ho, therefore data fits model well

5. Conclusion

The logistic regression model is used to predict the deposit probability of clients at a Portuguese bank.
The six variables of the fitted model: age, marital status, loan, last months of contact (Jan, March, June, July, November), campaign, outcome of the marketing campaign are significant to predict the likelihood of deposit.

The model without interaction is more appropriate to be used for prediction with the accuracy rate of 88%, specificity of 99% and sensitivity of 10%

According to the Hosmer – Lemeshow test, the data fits both models well

Based on the above interpretation of logistic regression, our ***recommendations*** for the bank are:

- Focus on the group of people who are divorced since they are not tied – knot so it is likely that they would deposit than other groups
- Focus on contacting people in December since the clients may receive bonus at the end of the year, they may have free money to deposit
- Focus on the clients who don’t have loans since they still have to pay for loans (houses, cars…) so they are less likely to deposit money
- Focus on a successful marketing campaign instead of failed one or the one with other outcomes

Appendix (R Code)

```
library(tidyverse)
library(dplyr)
library(MASS)
library(caTools)
library(lmtest)
library(caret)
library(ROCR)
library(interactions)
library(ggplot2)
library(ResourceSelection)

setwd("~/Langara/Sem 3/DANA-4820/project")
df <- read.csv('bank_data.csv')
View(df)

#as.factor categorical variable and response variable
df$y <- as.factor(df$y)
df$marital <- as.factor(df$marital)
df$default <- as.factor(df$default)
df$housing <- as.factor(df$housing)
df$loan <- as.factor(df$loan)
df$poutcome <- as.factor(df$poutcome)
df$education <- as.factor(df$education)
df$job <- as.factor(df$job)



#summary
summary(df)

#4.1 t test on age, balance, campaign, pdays, previous

#1. check variance
###var.test(group1, group2) #if p-value less then H0 rejected - variance of 2 groups not equal
###H(0) = Variances of group A is equal to variance of group B
###H(a) = Variance of group A is not equal to variance of group B
#2. t_test
###H(0) = The difference in group means is zero
###H(a) = The difference in group means is different from zero



group1_bal <- df$balance[df$y == 'yes']
group2_bal <- df$balance[df$y == 'no']
var.test(group1_bal, group2_bal) ###p value very less hence H0 rejected - variance of 2groups not equal
t.test(group1_bal, group2_bal, var.equal=F) ###p value very less hence H0 rejected - means of 2groups not equal
```

```
group1_age <- df$age[df$y == 'yes']
group2_age <- df$age[df$y == 'no']
var.test(group1_age, group2_age) ###p value very less hence H0 rejected - variance of 2groups not equal
t.test(group1_age, group2_age, var.equal=F) ###p value less hence H0 rejected - means of 2 groups not equal


group1_camp <- df$campaign[df$y == 'yes']
group2_camp <- df$campaign[df$y == 'no']
var.test(group1_camp, group2_camp) ###p value very less hence H0 rejected - variance of 2groups not equal
t.test(group1_camp, group2_camp, var.equal=F) ###p value very less hence H0 rejected - means of 2groups not equal


group1_pdays <- df$pdays[df$y == 'yes']
group2_pdays <- df$pdays[df$y == 'no']
var.test(group1_pdays, group2_pdays) ###p value very less hence H0 rejected - variance of 2groups not equal
t.test(group1_pdays, group2_pdays, var.equal=F) ###p value very less hence H0 rejected - means of 2groups not equal


group1_prev <- df$previous[df$y == 'yes']
group2_prev <- df$previous[df$y == 'no']
var.test(group1_prev, group2_prev) ###p value very less hence H0 rejected - variance of 2groups not equal
t.test(group1_prev, group2_prev, var.equal=F) ###p value very less hence H0 rejected - means of 2groups not equal


#4.2 chi-square test- Categorical variable: marital, default, housing, loan, poutcome, education, job
# It is generally to check association between explanatory and response variable
# hence lets check between housing and loan, housing and y, loan and y
# Hyposthesis
#H(0) = The variables are independent of each other
#H(a) = The variables are not independent of each other - associated

#marital
chisq.test(df$marital,df$y) ###p value less then 0.05 hence H0 rejected marital and y are not independent

#default
chisq.test(df$default,df$y) ###p value greater then 0.05 hence H0 accepted default and y are independent

#housing_loan
chisq.test(df$housing,df$y) ###p value less then 0.05 hence H0 rejected housing and y are not independent

#personal_loan
chisq.test(df$loan,df$y) ###p value less then 0.05 hence H0 rejected loan and y are not independent
```

```
#poutcome
chisq.test(df$poutcome,df$y) ###p value less then 0.05 hence H0 rejected loan and y are not independent

#education
chisq.test(df$education,df$y) ###p value less then 0.05 hence H0 rejected loan and y are not independent

#job
chisq.test(df$job,df$y) ###p value less then 0.05 hence H0 rejected loan and y are not independent

#month
chisq.test(df$month,df$y) ###p value less then 0.05 hence H0 rejected loan and y are not independent

#day
chisq.test(df$day,df$y) ###p value less then 0.05 hence H0 rejected loan and y are not independent

#######Now based on chisq test and t.test only "default" variable is insignificant#######

#5.1 correlation check
nums <- sapply(df, is.numeric)
numvar <- names(nums[nums == T])
cor(df[,numvar])
corrplot::corrplot(cor(df[,numvar]))
### we can see slight correlation between pdays and previous but not greater than 0.8


#5.2 Multicollinearity check
model <- glm(y~.,family=binomial(link='logit'),data = df)
car::vif(model)
### no Multicollinearity issue nothing greater than 10

###So the only insignificant variable to drop is only "default"
#dropping insignificant variable
df$default <- NULL

#6. Interaction plot

#interaction.plot(df$education,df$y,df$balance,median)
#check if deposit and education have interaction

interaction.plot(x.factor = df$education, #x-axis variable
        trace.factor = df$y, #variable for lines
        response = df$balance, #y-axis variable
        fun = median, #metric to plot
        ylab = "Account_balance",
        xlab = "Education_level",
        col = c("pink", "blue"),
        lty = 1, #line type
        lwd = 2, #line width
        trace.label = "Balance")
```

```
str(df)

interaction.plot(x.factor = df$loan, #x-axis variable
        trace.factor = df$y, #variable for lines
        response = df$age, #y-axis variable
        fun = median, #metric to plot
        ylab = "Age",
        xlab = "loan",
        col = c("pink", "blue"),
        lty = 1, #line type
        lwd = 2, #line width
        trace.label = "age")

interaction.plot(x.factor = df$housing, #x-axis variable
        trace.factor = df$y, #variable for lines
        response = df$age, #y-axis variable
        fun = median, #metric to plot
        ylab = "Age",
        xlab = "housingloan",
        col = c("pink", "blue"),
        lty = 1, #line type
        lwd = 2, #line width
        trace.label = "age")

df$balance= as.numeric(df$balance)

fit <- glm(y ~ marital * loan, family ="binomial", data = df)
summary(fit)
cat_plot(fit, pred = loan, modx = marital,geom = "line", interval = TRUE) ###clear interaction

fit2 <- glm(y ~ education * job, family ="binomial", data = df)
cat_plot(fit2, pred = job, modx = education, geom = "line", interval = TRUE) ###high interaction

fit3 <- glm(y ~ education * marital, family ="binomial", data = df)
cat_plot(fit3, pred = marital, modx = education, geom = "line", interval = TRUE) ###high interaction

#7. data split
set.seed(1234) # is used so that each time we get the same data set after splitting
sample_size<- sample.split(df$y, SplitRatio = 7/10) #Splitting the dataset into 70/30 ratio
train <- subset(df, sample_size==T)
test <- subset(df, sample_size==F)
nrow(train) #3165
nrow(test) #1356

##8. STEPWISE REGRESSION STARTS HERE...

# Full model ref-https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4842399/
fullmodel <- glm(y ~., data = train, family = binomial)
```

```
summary(fullmodel)

step <- stepAIC(fullmodel,trace = F)

step$anova

#final selected variables are: Final Model-> y ~ age + marital + loan + month + campaign + poutcome

#9.comparing two models final model and final model + potential interaction
#comparison using Likelihood ratio test ref-https://api.rpubs.com/tomanderson_34/lrt
#H(0) <- reduced model is better -model1
#H(a) <- full model is better -model2
model1 <- glm(y ~ age + marital + loan + month + campaign + poutcome, data = train, family = binomial)
summary(model1)

model2 <- glm(y ~ age + marital + loan + month + campaign + poutcome + loan*marital, data = train,
family = binomial)
summary(model2)

A <- logLik(model1)
B <- logLik(model2)

teststat <- -2 * (as.numeric(A)-as.numeric(B))
p.val <- pchisq(teststat, df = 2, lower.tail = FALSE) #0.5160002

##or simply
lrtest(model1,model2) #pvalue 0.516 cant reject null hypothesis. Hence reduced model is better

#------------------comparison using Wald test---------------------------------------------------------
#comparison using Likelihood ratio test
#H(0) <- reduced model is better -model1
#H(a) <- full model is better -model2
waldtest(model1,model2) #pvalue 0.5454 cant reject null hypothesis. Hence reduced model is better
#Wald test

#Model 1: y ~ age + marital + loan + month + campaign + poutcome
#Model 2: y ~ age + marital + loan + month + campaign + poutcome + loan * marital
#Res.Df Df      F Pr(>F)
#1   3145
#2   3143  2 0.6063 0.5454

###10.classification report for both models ref- https://daviddalpiaz.github.io/r4sl/logistic-regression.html

#Predictions and classification report of model1
predictions_test1 <- predict(model1, train)
summary(predictions_test1)

pred1<- ifelse(predictions_test1>0.5, "yes","no")
Conf.mat_test1<- confusionMatrix(table(pred1, train$y), positive = "yes")
```

```
c(Conf.mat_test1$overall["Accuracy"],
  Conf.mat_test1$byClass["Sensitivity"],
  Conf.mat_test1$byClass["Specificity"])

#Accuracy Sensitivity Specificity
#0.88941548  0.09863014  0.99250000

#Predictions and classification report of model2
predictions_test2 <- predict(model2, train)
predictions_test2

pred2<- ifelse(predictions_test2>0.5, "yes","no")
Conf.mat_test2<- confusionMatrix(table(pred2, train$y), positive = "yes")
Conf.mat_test2
c(Conf.mat_test2$overall["Accuracy"],
  Conf.mat_test2$byClass["Sensitivity"],
  Conf.mat_test2$byClass["Specificity"])

#Accuracy Sensitivity Specificity
#0.8900474   0.1013699   0.9928571

###11.ROC Curve for both models
#model1

library(ROCit)
library(dlstats)    # for package download stats
library(pkgsearch)


rocpred1<- prediction(predictions_test1, train$y)
perf <- performance(rocpred1,"tpr","fpr")
plot(perf,colorize=TRUE)

#model2
rocpred2<- prediction(predictions_test2, train$y)
perf2 <- performance(rocpred2,"tpr","fpr")
plot(perf2,colorize=TRUE)

###12.Lack of fit test (Hosmer-Lemshow test- ungrouped data) and Interpret the coefficient.

#The Hosmer-Lemeshow test involves a few steps.
# The steps are:
#1. Estimate the probability of yi = 1 for each observation.
#2. Sort the observations into groups (usually 10) with increasing probability.
#3. For each group compute the number of observations multiplied by the average probability. This gives the expected number of yi = 1'sin this group.
#4. Compute the expected number of yi = 0's too
#5. Compute a ??2 test statistic
#6. This test statistic is ??2 10???2 distributed
```

```
hl1 <- hoslem.test(model1$y, fitted(model1),g=10)
hl1

#data:  model1$y, fitted(model1)
#X-squared = 10.296, df = 8, p-value = 0.2448

#This gives p=0.25, indicating no evidence of poor fit.
#This is good, since here we know the model is indeed correctly specified.
#We can also obtain a table of observed vs expected, from our hl object:

cbind(hl1$observed,hl1$expected)
#                   y0  y1  yhat0      yhat1
#[0.00934,0.0401] 304  13 307.5053   9.494663
#(0.0401,0.0497]  304  12 301.5608  14.439235
#(0.0497,0.0616]  304  14 300.4608  17.539207
#(0.0616,0.0726]  292  23 293.8767  21.123277
#(0.0726,0.0847]  294  24 292.9424  25.057567
#(0.0847,0.0971]  295  20 286.5135  28.486500
#(0.0971,0.112]   278  38 283.2184  32.781596
#(0.112,0.136]    286  31 277.8747  39.125322
#(0.136,0.185]    259  57 267.1269  48.873063
#(0.185,0.943]    184 133 188.9204 128.079573


# Similarly for MODEL 2
hl2 <- hoslem.test(model2$y, fitted(model2),g=10)
#data:  model2$y, fitted(model2)
#X-squared = 9.493, df = 8, p-value = 0.3024

#This gives p=0.3, indicating no evidence of poor fit.
#This is good, since here we know the model is indeed correctly specified.
#We can also obtain a table of observed vs expected, from our hl object:

cbind(hl2$observed,hl2$expected)
#                y0  y1    yhat0      yhat1
#[0.0101,0.0414] 303  15 308.0516   9.948367
#(0.0414,0.0488] 306   9 300.6434  14.356605
#(0.0488,0.0599] 305  13 300.8052  17.194772
#(0.0599,0.0718] 292  23 294.2112  20.788829
#(0.0718,0.084]  293  25 293.1938  24.806241
#(0.084,0.0961]  290  25 286.7295  28.270503
#(0.0961,0.111]  283  33 283.3865  32.613459
#(0.111,0.137]   285  32 277.7679  39.232053
#(0.137,0.187]   260  56 266.6855  49.314541
#(0.187,0.946]   183 134 188.5254 128.474632


#final logistic regression (conclusion)
```

```
model <- glm(y ~ age + marital + loan + month + campaign + poutcome, data = train, family = binomial)
summary(model)
```