Numeric and Binary targets Forecasting Models

Fujie Mei Sergio Delgado Mario Wang

June 1, 2023

Contents

1 Introduction

In today's competitive business landscape, effective marketing campaigns play a crucial role in driving customer engagement and maximizing business outcomes. To optimize campaign performance, it is essential to understand the factors that influence key metrics such as the duration of client calls and the likelihood of a positive response. Predictive modeling techniques, such as linear regression and logistic regression, provide valuable insights into these factors and enable organizations to make data-driven decisions for campaign optimization.

This project aims to analyze a dataset from a marketing campaign and develop predictive models to estimate the duration of client calls and predict whether a client will respond positively or negatively. By leveraging linear regression for call duration prediction and logistic regression for response prediction, we can uncover the underlying patterns and variables that significantly impact these outcomes.

The project workflow begins by constructing initial regression models using various predictor variables. To refine the models, variable selection techniques will be employed, such as assessing variable significance and addressing multicollinearity using Variance Inflation Factor (VIF) analysis. By iteratively evaluating and eliminating variables, we can identify the subset of predictors that contribute most significantly to the target variables.

Once the optimal models are identified, they will be further validated using appropriate evaluation metrics and techniques. The performance of the models will be assessed based on criteria such as model fit, goodness-of-fit measures, and predictive accuracy. Validation helps ensure the robustness and reliability of the chosen models, enhancing their practical utility for real-world marketing campaign scenarios.

The outcomes of this project have the potential to provide valuable insights for marketers, enabling them to optimize campaign strategies, allocate resources effectively, and improve customer engagement. By accurately predicting call duration and client responses, organizations can make informed decisions to enhance campaign effectiveness, drive customer conversions, and ultimately achieve their marketing objectives.

Through this project, we will showcase the power of predictive modeling techniques in marketing analytics and highlight the practical benefits of utilizing linear and logistic regression models for campaign optimization. By combining statistical analysis with real-world marketing data, we aim to contribute to the field of marketing analytics and provide actionable insights for businesses seeking to improve their marketing campaign performance.

2 Loading data and deleting columns

We will delete the columns we said that won't contained too many errors to be analyzeable.

```
df<-read.csv2("clean_data.csv")
df$X<-NULL
df$pdays<-NULL
df$previous<-NULL
df$errVar<-NULL
names(df)
```

```
##  [1] "age"              "job"              "marital"
##  [4] "education"        "housing"          "loan"
##  [7] "contact"          "month"            "day_of_week"
## [10] "duration"         "campaign"         "poutcome"
## [13] "emp.var.rate"     "cons.price.idx"   "cons.conf.idx"
## [16] "euribor3m"        "nr.employed"      "y"
## [19] "Age_group"        "Campaign_contacts" "mout"
```

```
vars_con = c("age","campaign","emp.var.rate","cons.price.idx","cons.conf.idx","euribor3m","nr.employed")
vars_dis = c("job","marital","education","housing","loan","contact","month","day_of_week","Age_group","C
vars_res= c("y","duration")
df$y<-factor(df$y)
head(df)
```

```
##   age         job marital           education housing loan   contact month
## 1  41      admin. married   university.degree      no   no  cellular   jul
## 2  35 blue-collar married               basic      no   no telephone   may
## 3  30  technician  single   university.degree     yes   no  cellular   aug
## 4  29 blue-collar married               basic     yes   no  cellular   apr
## 5  30 blue-collar married               basic      no   no  cellular   jul
## 6  40  technician  single professional.course     yes   no  cellular   may
##   day_of_week duration campaign    poutcome emp.var.rate cons.price.idx
## 1         mon     1360        3 nonexistent          1.4         93.918
## 2         wed      622        3 nonexistent         -1.8         92.893
## 3         mon      720        1 nonexistent          1.4         93.444
## 4         thu     1042        2 nonexistent         -1.8         93.075
## 5         tue      623        2 nonexistent          1.4         93.918
## 6         fri      317        1     failure         -1.8         92.893
##   cons.conf.idx euribor3m nr.employed   y Age_group Campaign_contacts    mout
```

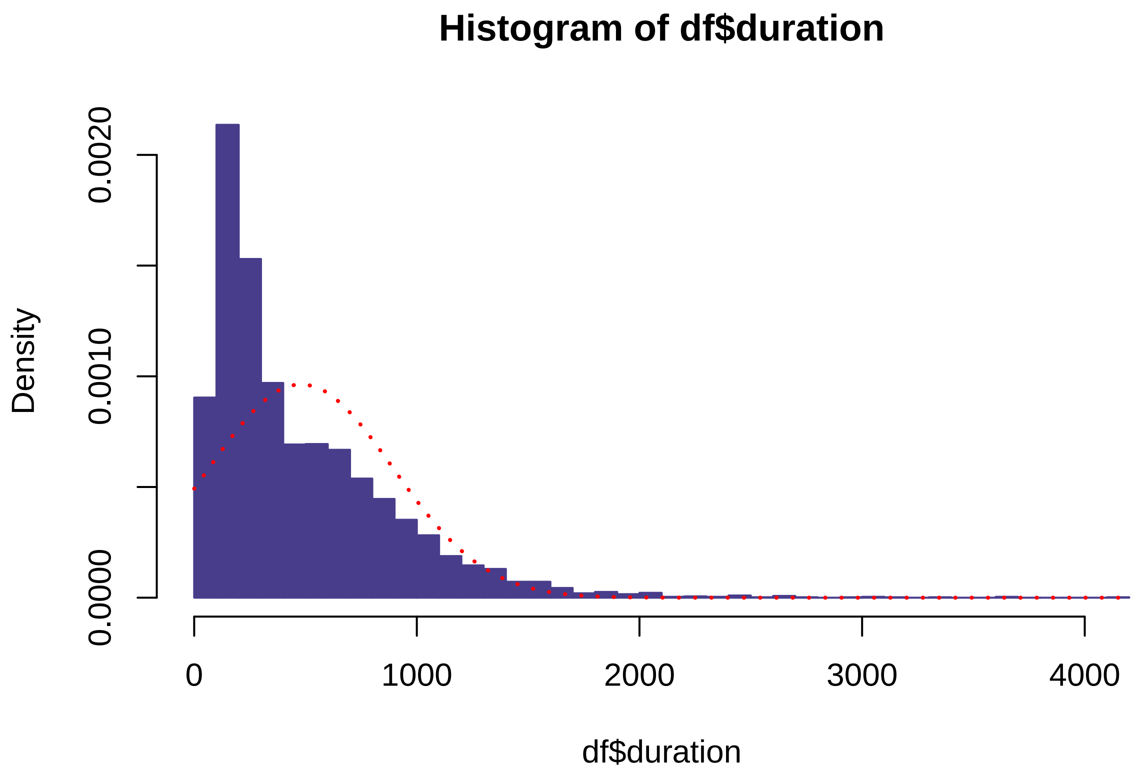
```
## 1          -42.7      4.960      5228.1 yes      30-50         Infrequent YesMOut
## 2          -46.2      1.281      5099.1 yes      30-50         Infrequent YesMOut
## 3          -36.1      4.965      5228.1 yes      30-50         Infrequent YesMOut
## 4          -47.1      1.435      5099.1 yes      20-30         Infrequent YesMOut
## 5          -42.7      4.962      5228.1 yes      30-50         Infrequent YesMOut
## 6          -46.2      1.259      5099.1 yes      30-50         Infrequent YesMOut
```

3 Target variable normality

Before we begin to start modelling for our linear model with our numerical target, we should consider the normality of this.

3.1 Normality

```
hist(df$duration,50,freq=F,col="darkslateblue",border = "darkslateblue")
mm<-mean(df$duration);ss<-sd(df$duration)
curve(dnorm(x,mean=mm,sd=ss),col="red",lwd=2,lty=3, add=T)
```




```
shapiro.test(df$duration)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$duration
## W = 0.83982, p-value < 2.2e-16
```

We see that the target total_amount is not normally distributed for the following reasons:

- graph: there is no symmetry in the plot

- shapiro: we see that the p-value is too large to accept the assumption that target.total_amount is normally distributed

3.1.1 Symmetry

```
skewness(df$duration)
```

```
## [1] 1.877425
```

Normal data should have 0 skewness: we see that our data is left skewed (1.877425).

4 Numerical target modelization

4.1 Numerical explicative variables

```
(length(vars_con))
```

```
## [1] 7
```

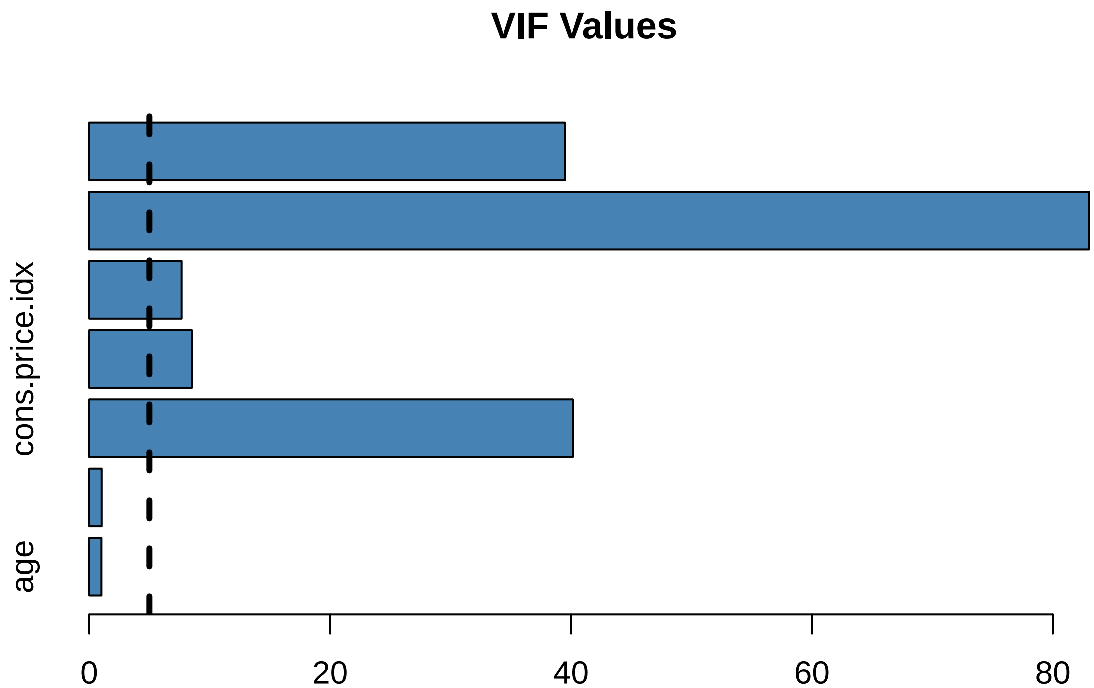
The first step is deciding the number of explicatives variables . We have many methods including condes, PCA, correlation…if we have a great amount of numerical variables but since it's not our case (we can see that there are only 7) we can use all and decide with the model created which are the best ones to use. We will start using lm to create our model and from there we can discard the ones which are irrelevant, then we use AIC and BIC methods to affirm it.

```
m1<-lm(duration~.,data=df[,c("duration",vars_con)])
summary(m1)
```

```
##
## Call:
## lm(formula = duration ~ ., data = df[, c("duration", vars_con)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -797.3 -198.7  -90.6   95.6 3325.8
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.073e+05  5.298e+03 -20.249  < 2e-16 ***
## age            -5.008e-01  4.899e-01  -1.022 0.306767
## campaign        1.022e+01  3.792e+00   2.696 0.007038 **
## emp.var.rate   -6.371e+01  2.270e+01  -2.807 0.005027 **
## cons.price.idx  1.143e+02  3.015e+01   3.790 0.000152 ***
## cons.conf.idx   1.266e+01  3.065e+00   4.130 3.69e-05 ***
## euribor3m      -6.006e+02  2.931e+01 -20.491  < 2e-16 ***
## nr.employed     1.932e+01  6.509e-01  29.684  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 345.8 on 4992 degrees of freedom
## Multiple R-squared:  0.3023, Adjusted R-squared:  0.3014
## F-statistic:   309 on 7 and 4992 DF,  p-value: < 2.2e-16
```

```
vif_values<-vif(m1)
#create horizontal bar chart to display each VIF value
barplot(vif_values, main = "VIF Values", horiz = TRUE, col = "steelblue")
```

```r
#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```



In our initial model, we observe that the variable "age" lacks statistical significance. Additionally, a careful examination of the Variance Inflation Factors (VIFs) reveals the presence of exceptionally high values, particularly for the variable "euribor3m." As a result, we will exclude "euribor3m" from subsequent model iterations to assess its impact on model performance.

It is worth noting that the explanatory power of the current model, as measured by the coefficient of determination (R-squared), is relatively low, standing at 30%. This indicates that the model accounts for only a moderate proportion of the total variability in the response variable.

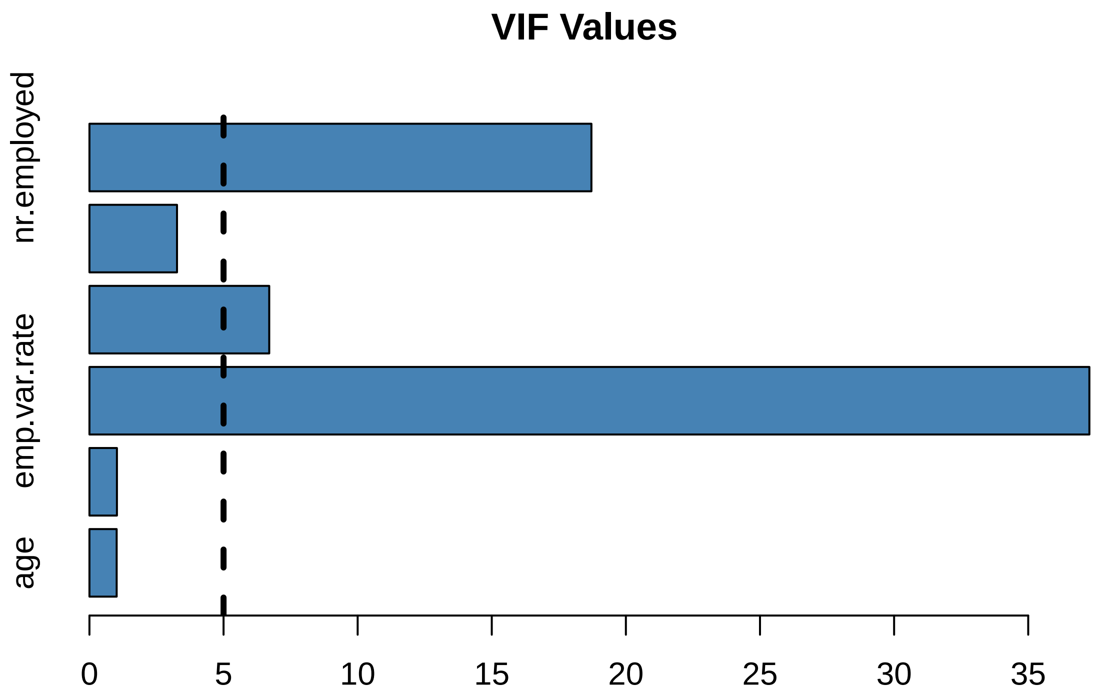
Moving forward, VIFs above a threshold value of 5 will be regarded as high, aligning with the guidelines provided by the R VIF function documentation. This threshold helps identify potential issues of multicollinearity among the predictor variables, thereby aiding in the selection of more reliable and robust models.

```r
m2<-lm(duration~age+campaign+emp.var.rate+cons.price.idx+cons.conf.idx+nr.employed,data=df[,c("duration
summary(m2)
```

```
##
## Call:
## lm(formula = duration ~ age + campaign + emp.var.rate + cons.price.idx +
##     cons.conf.idx + nr.employed, data = df[, c("duration", vars_con)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -835.5 -210.9  -96.1  123.3 3475.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.475e+04  4.104e+03  -8.468  < 2e-16 ***
## age            -7.151e-01  5.099e-01  -1.402    0.161
## campaign        1.779e+01  3.929e+00   4.529 6.08e-06 ***
```

```
## emp.var.rate   -1.880e+02  2.278e+01  -8.253  < 2e-16 ***
## cons.price.idx -1.710e+02  2.784e+01  -6.141 8.83e-10 ***
## cons.conf.idx  -3.494e+01  2.082e+00 -16.783  < 2e-16 ***
## nr.employed     9.647e+00  4.665e-01  20.681  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 360.1 on 4993 degrees of freedom
## Multiple R-squared:  0.2437, Adjusted R-squared:  0.2427
## F-statistic: 268.1 on 6 and 4993 DF,  p-value: < 2.2e-16
```

```
vif_values<-vif(m2)
#create horizontal bar chart to display each VIF value
barplot(vif_values, main = "VIF Values", horiz = TRUE, col = "steelblue")

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```



After further analysis, we have decided to remove the variable "emp.var.rate" from our model. The decision was based on the observation that this variable exhibits a high Variance Inflation Factor (VIF). VIF is a measure of multicollinearity, and a high VIF indicates a strong correlation between the variable and other predictors in the model.

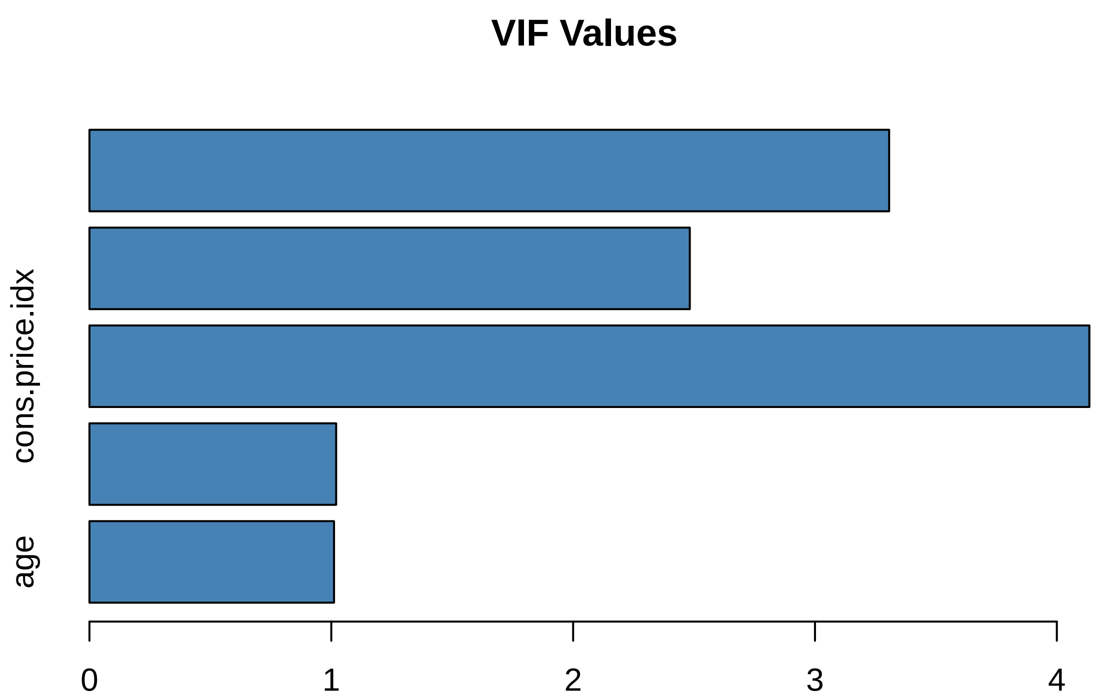
By removing "emp.var.rate," we aim to mitigate the issue of multicollinearity and improve the stability and interpretability of our model. Multicollinearity can lead to unreliable coefficient estimates and difficulties in interpreting the individual effects of correlated predictors.

```
m3<-lm(duration~age+campaign+cons.price.idx+cons.conf.idx+nr.employed,data=df[,c("duration",vars_con)])
summary(m3)
```

```
##
## Call:
## lm(formula = duration ~ age + campaign + cons.price.idx + cons.conf.idx +
##     nr.employed, data = df[, c("duration", vars_con)])
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max 
## -911.5 -218.1  -98.7  129.2 3466.4 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -3736.6326  1660.0255  -2.251   0.0244 *  
## age               -0.9556     0.5125  -1.864   0.0623 .  
## campaign          15.6083     3.9463   3.955 7.75e-05 ***
## cons.price.idx  -313.2049    22.0136 -14.228  < 2e-16 ***
## cons.conf.idx    -43.3474     1.8274 -23.721  < 2e-16 ***
## nr.employed        6.1541     0.1974  31.178  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 362.5 on 4994 degrees of freedom
## Multiple R-squared:  0.2333, Adjusted R-squared:  0.2326 
## F-statistic:   304 on 5 and 4994 DF,  p-value: < 2.2e-16
```

```r
vif_values<-vif(m3)
#create horizontal bar chart to display each VIF value
barplot(vif_values, main = "VIF Values", horiz = TRUE, col = "steelblue")

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```



Upon further analysis, it becomes evident that by removing the variable “emp.var.rate” from our model, all remaining predictor variables exhibit statistical significance, as indicated by their p-values being less than 0.05. However, it is worth noting that the variable “age” still fails to attain significance. As a result, we will proceed to eliminate “age” from our model.

Additionally, to address concerns of multicollinearity, we observe that all Variance Inflation Factors (VIFs) are below the threshold of 5. This suggests that the predictor variables do not suffer from substantial intercorrelation issues.

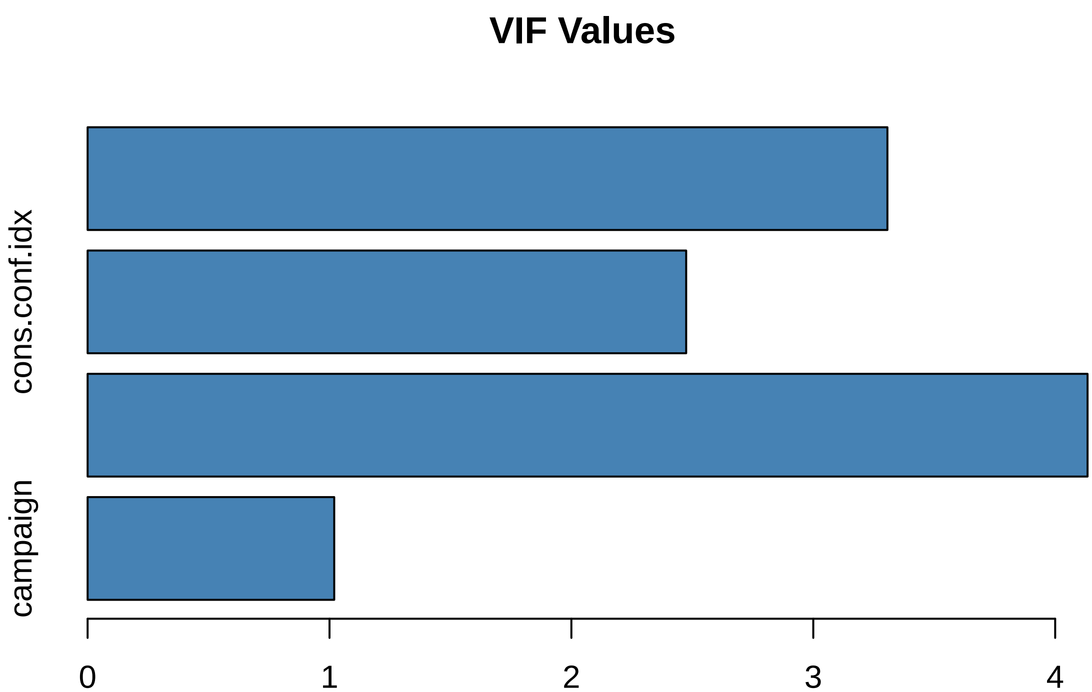
Therefore, our subsequent step involves assessing the performance of an alternative model, which excludes the variable "age." By evaluating this model, we aim to determine the impact of removing "age" on the overall model performance and effectiveness.

```
m4<-lm(duration~campaign+cons.price.idx+cons.conf.idx+nr.employed,data=df[,c("duration",vars_con)])
summary(m4)
```

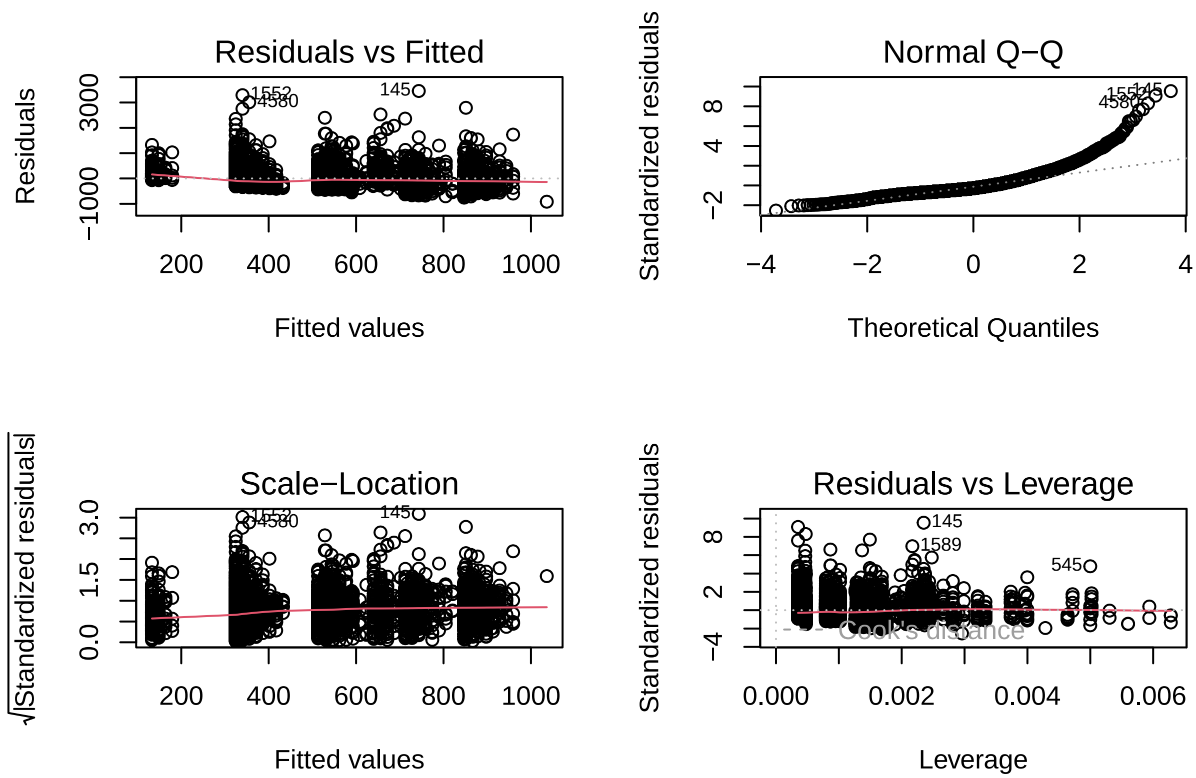
```
##
## Call:
## lm(formula = duration ~ campaign + cons.price.idx + cons.conf.idx +
##     nr.employed, data = df[, c("duration", vars_con)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -917.2 -218.3  -98.7  130.2 3455.6
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3742.3160  1660.4341  -2.254   0.0243 *
## campaign          15.4208     3.9460   3.908 9.43e-05 ***
## cons.price.idx  -313.7387    22.0172 -14.250  < 2e-16 ***
## cons.conf.idx    -43.5382     1.8249 -23.857  < 2e-16 ***
## nr.employed        6.1561     0.1974  31.180  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 362.6 on 4995 degrees of freedom
## Multiple R-squared:  0.2328, Adjusted R-squared:  0.2322
## F-statistic: 378.9 on 4 and 4995 DF,  p-value: < 2.2e-16
```

In this case, we can see that all of our variables are statistically significant and the vif's values fall into acceptable range so we will decide to use all the variables of these model even though the R2 isn't the highest.

```
vif_values<-vif(m4)
#create horizontal bar chart to display each VIF value
barplot(vif_values, main = "VIF Values", horiz = TRUE, col = "steelblue")

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```




```r
par(mfrow=c(2,2))
plot(m4)
```



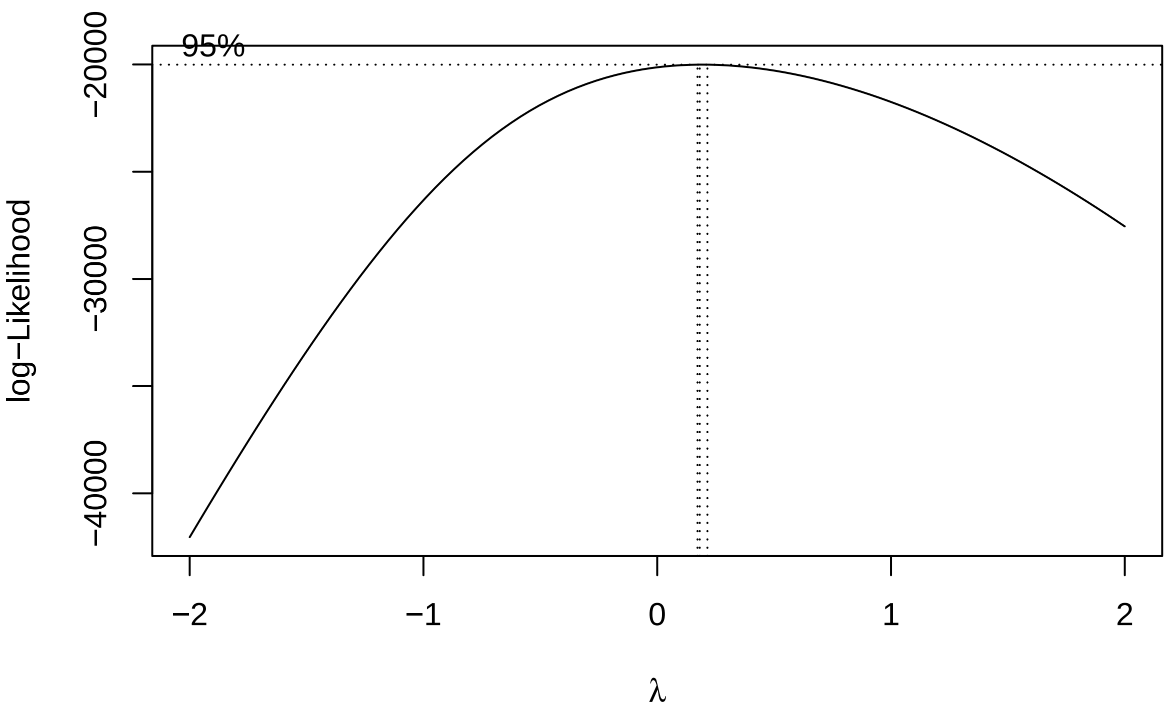
To examine the normality assumption of our data, we have conducted an analysis and found that it is not met. In order to address this issue, we propose using the Box-Cox transformation, which allows us to determine the optimal power transformation to achieve normality. By applying the Box-Cox function to our target variable, "duration," we have obtained an estimated lambda ( ) value that is close to 0.

Based on this finding, we will proceed with a log-transformation of the "duration" variable in conjunction with our predictor variables (regressors). This transformation aims to normalize the distribution of the "duration" variable and improve the suitability of our data for linear regression modeling.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
boxcox(duration~campaign+cons.price.idx+cons.conf.idx+nr.employed ,data=df[,c("duration",vars_con)])
```

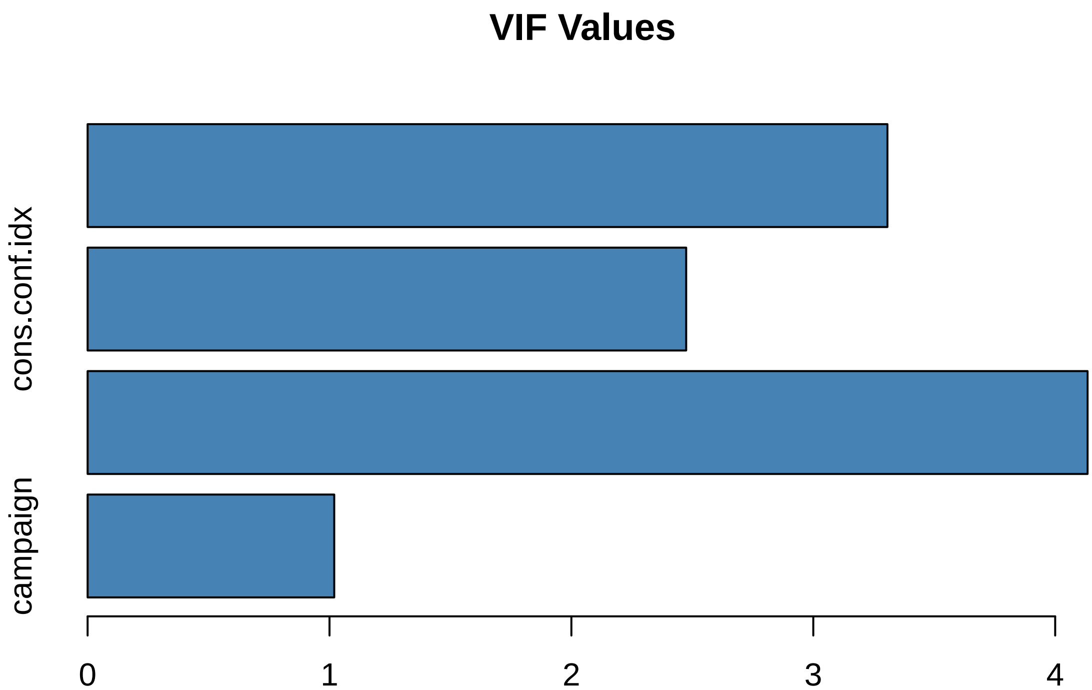



```
m5 <-
lm(log(duration)~campaign+cons.price.idx+cons.conf.idx+nr.employed,df[,c("duration",vars_con)]);
summary(m5)
```

```
##
## Call:
## lm(formula = log(duration) ~ campaign + cons.price.idx + cons.conf.idx +
##     nr.employed, data = df[, c("duration", vars_con)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0716 -0.4527  0.0269  0.5059  2.7606
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.0012440  3.6314661   2.479   0.0132 *
## campaign        0.0211887  0.0086301   2.455   0.0141 *
## cons.price.idx -0.8355017  0.0481529 -17.351   <2e-16 ***
## cons.conf.idx  -0.1004966  0.0039913 -25.179   <2e-16 ***
```

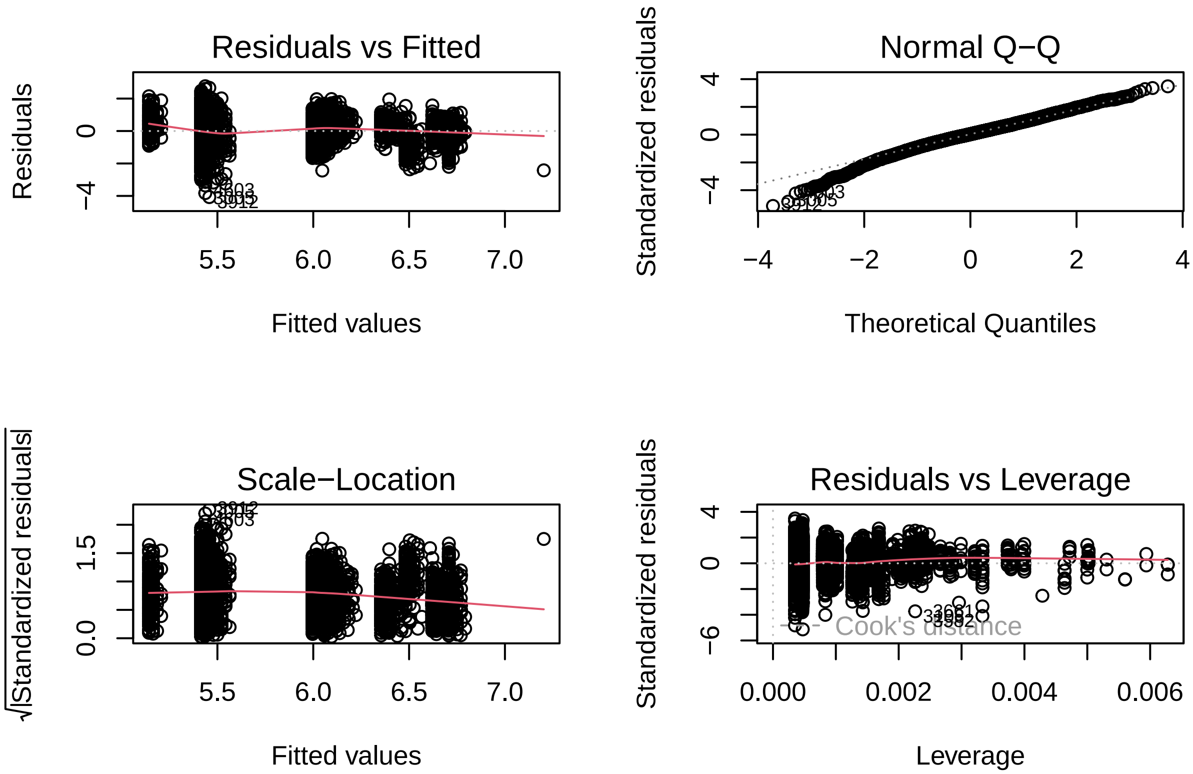
```
## nr.employed     0.0137290  0.0004318  31.795   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7929 on 4995 degrees of freedom
## Multiple R-squared:  0.2627, Adjusted R-squared:  0.2622
## F-statistic:   445 on 4 and 4995 DF,  p-value: < 2.2e-16
```

```r
vif_values<-vif(m5)
#create horizontal bar chart to display each VIF value
barplot(vif_values, main = "VIF Values", horiz = TRUE, col = "steelblue")

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```



We can see that the quality stays the same, even R2 went up and the vifs are in acceptable ranges, we want to check the normality now to see if it has improved.

```r
par(mfrow=c(2,2))
plot(m5)
```



As we can see, the dot's follow the normality line so we can assume that it complies with this assumption.

So far we have seen 5 models, the first one with all the numerical variables included, the second one with numerical variables excluded from using VIF, the third one we also excluded another variable using VIF, the fourth one we withdraw age because it was not significant and the vif values were ok and finally the 5th model normalizing our target variable. Now we are going to compare them:

- Model 1
  - Coefficient of determination = 30.23%
  - 5 VIFs: 5/7
- Model 2
  - Coefficient of determination = 24.27%
  - 5 VIFs: 4/6
- Model 3
  - Coefficient of determination = 23.33%
  - VIFs: 0/5
- Model 4
  - Coefficient of determination = 23.28%
  - 5 VIFs: 0/4
- Model 5
  - Coefficient of determination = 26.27%
  - VIFs: 0/4

  We can see that models 1 and 5 have the highest R2 value and between those two the best one is model5 because none of its variables have high vif, models 2,3 and 4 have similar R2 values and their VIF's are comparable but it can't be the model 5 so it's the one we will keep for now.

4.2 Modelization with factors

We will first make a condes to see which categorical variables are the most influential with respect to our target duration to see which ones we will choose for our model.

```r
condes(df[,c("duration",vars_dis)],1,proba=0.05)
```

```
##
## Link between the variable and the categorical variable (1-way anova)
## =============================================
##                             R2       p.value
## month             0.2054752461 1.182391e-242
## contact           0.1221648631 1.251448e-143
## day_of_week       0.0065961120  1.159340e-06
## Campaign_contacts 0.0025668220  3.385297e-04
## Age_group         0.0029161318  5.630996e-03
## loan              0.0008230113  4.251226e-02
##
## Link between variable abd the categories of the categorical variables
## ================================================================
##                                 Estimate       p.value
## contact=cellular               147.49958 1.251448e-143
## month=jul                      367.12692 3.230283e-101
## month=aug                      314.42614  1.543596e-49
## month=nov                      246.68788  4.195651e-23
## month=jun                      132.02683  7.434551e-17
## Campaign_contacts=Frequent      44.83664  3.385297e-04
## job=self-employed              105.41710  1.767133e-03
## day_of_week=wed                 33.62693  2.930958e-03
## marital=single                  16.77553  2.831054e-02
## day_of_week=thu                 21.86174  3.410484e-02
## day_of_week=fri                 24.61238  3.616968e-02
## loan=yes                        16.88335  4.251226e-02
## loan=no                        -16.88335  4.251226e-02
## marital=married                -13.50531  4.068143e-02
## day_of_week=tue                -33.60887  5.429733e-03
## month=oct                     -233.75748  2.740875e-03
## Age_group=NA                   -61.53969  1.269264e-03
## Campaign_contacts=Infrequent   -44.83664  3.385297e-04
## day_of_week=mon                -46.49218  6.010119e-05
## month=mar                     -232.05113  2.097783e-07
## contact=telephone             -147.49958 1.251448e-143
## month=may                     -155.01486 9.806608e-148
```

Upon examining the statistical significance of the categorical variables, we have determined that the factors with the smallest p-values are "contact" and "month." Therefore, for the sake of simplicity, we will proceed with these variables for our modeling purposes.

However, considering that the variable "month" consists of numerous levels, we acknowledge the potential complexities it may introduce to the modeling process. To facilitate a more manageable and streamlined analysis, we will undertake a regrouping or re-categorization of the "month" variable. This regrouping will involve combining certain levels to create broader categories that retain meaningful information while reducing the overall number of levels.

```r
# Months to groups
df$f.influentMonth <- 3
# 1 level - mar-may
aux<-which(df$month %in% c("apr","jun","aug"))
df$f.influentMonth[aux] <-1
# 2 level - jun-ago
```

```r
aux<-which(df$month %in% c("sep","may","jul"))
df$f.influentMonth[aux] <-2
# 3 level - aug-feb
aux<-which(df$month %in% c("mar","dec","oct","nov"))
df$f.influentMonth[aux] <-3
df$f.influentMonth<-factor(df$f.influentMonth,levels=1:3,labels=c("apr-ju
n-aug","sep-may-jul","mar-dec-oct-nov"))
levels(df$f.influentMonth)<-paste0("f.influentMonth.",levels(df$f.influentMonth)) # Hacemos las etiquet
summary(df$f.influentMonth)
```

```
##   f.influentMonth.apr-ju\nn-aug     f.influentMonth.sep-may-jul
##                            1070                            3571
## f.influentMonth.mar-dec-oct-nov
##                             359
```

Since we have campaign as both categorical and numerical factors, we will model with both of them with our new categorical variables to see which is better to use, the numerical or the categorical one using AIC criteria because our model isn't too complex. We can see that AIC is smaller in m6, with numerical campaign, so is the go-to model for us.

```r
m6<-lm(log(duration)~campaign+cons.price.idx+cons.conf.idx+nr.employed+contact+f.influentMonth,data=df)
m7<-lm(log(duration)~Campaign_contacts+contact+cons.price.idx+cons.conf.idx+f.influentMonth+contact,data
AIC(m6,m7)
```

```
##    df      AIC
## m6  9 11552.68
## m7  8 12043.30
```

```r
summary(m6)
```

```
##
## Call:
## lm(formula = log(duration) ~ campaign + cons.price.idx + cons.conf.idx +
##     nr.employed + contact + f.influentMonth, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0171 -0.4456  0.0028  0.4649  2.8019
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                     7.6898157  5.3530130   1.437
## campaign                                        0.0080698  0.0084092   0.960
## cons.price.idx                                 -0.8958105  0.0872423 -10.268
## cons.conf.idx                                  -0.1006825  0.0050515 -19.931
## nr.employed                                     0.0151294  0.0006737  22.458
## contacttelephone                               -0.1516672  0.0596047  -2.545
## f.influentMonthf.influentMonth.sep-may-jul     -0.1597367  0.0301231  -5.303
## f.influentMonthf.influentMonth.mar-dec-oct-nov -0.8393331  0.0585239 -14.342
##                                                Pr(>|t|)
## (Intercept)                                       0.151
## campaign                                          0.337
## cons.price.idx                                  < 2e-16 ***
## cons.conf.idx                                   < 2e-16 ***
## nr.employed                                     < 2e-16 ***
## contacttelephone                                  0.011 *
```

```
## f.influentMonthf.influentMonth.sep-may-jul     1.19e-07 ***
## f.influentMonthf.influentMonth.mar-dec-oct-nov  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7675 on 4992 degrees of freedom
## Multiple R-squared:  0.3098, Adjusted R-squared:  0.3088
## F-statistic:   320 on 7 and 4992 DF,  p-value: < 2.2e-16
```

We see that campaign have a p-value>0.05 will drop them and our final variables will be the ones left.

```
m7<-lm(log(duration)~cons.price.idx+cons.conf.idx+nr.employed+contact+f.influentMonth,data=df)
summary(m7)
```

```
##
## Call:
## lm(formula = log(duration) ~ cons.price.idx + cons.conf.idx +
##     nr.employed + contact + f.influentMonth, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0088 -0.4464  0.0014  0.4643  2.8021
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                     7.7544371  5.3525471   1.449
## cons.price.idx                                 -0.9012036  0.0870604 -10.351
## cons.conf.idx                                  -0.1011624  0.0050266 -20.125
## nr.employed                                     0.0152137  0.0006679  22.778
## contacttelephone                               -0.1489161  0.0595353  -2.501
## f.influentMonthf.influentMonth.sep-may-jul     -0.1594497  0.0301214  -5.294
## f.influentMonthf.influentMonth.mar-dec-oct-nov -0.8453029  0.0581919 -14.526
##                                                Pr(>|t|)
## (Intercept)                                      0.1475
## cons.price.idx                                  < 2e-16 ***
## cons.conf.idx                                   < 2e-16 ***
## nr.employed                                     < 2e-16 ***
## contacttelephone                                 0.0124 *
## f.influentMonthf.influentMonth.sep-may-jul     1.25e-07 ***
## f.influentMonthf.influentMonth.mar-dec-oct-nov  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7675 on 4993 degrees of freedom
## Multiple R-squared:  0.3096, Adjusted R-squared:  0.3088
## F-statistic: 373.2 on 6 and 4993 DF,  p-value: < 2.2e-16
```

```
vif(m7)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## cons.price.idx  14.425141  1        3.798044
## cons.conf.idx    4.189900  1        2.046924
## nr.employed      8.445500  1        2.906114
## contact          7.229838  1        2.688836
## f.influentMonth  2.197984  2        1.217604
```

We can see that after dropping campaign all the variables remaining are significant and the vif values falls into the acceptable range so we will use this model.

4.3 Interacctions

```
m8<-lm(log(duration)~(cons.price.idx+cons.conf.idx+nr.employed+contact+f.influentMonth)^2,data=df)
anova(m8)
```

```
## Analysis of Variance Table
##
## Response: log(duration)
##                                   Df  Sum Sq Mean Sq   F value    Pr(>F)
## cons.price.idx                     1  191.80  191.80  348.1434 < 2.2e-16 ***
## cons.conf.idx                      1  270.05  270.05  490.1694 < 2.2e-16 ***
## nr.employed                        1  653.58  653.58 1186.3287 < 2.2e-16 ***
## contact                            1   78.82   78.82  143.0756 < 2.2e-16 ***
## f.influentMonth                    2  124.70   62.35  113.1770 < 2.2e-16 ***
## cons.price.idx:cons.conf.idx       1   71.16   71.16  129.1656 < 2.2e-16 ***
## cons.price.idx:nr.employed         1   36.68   36.68   66.5872 4.201e-16 ***
## cons.price.idx:contact             1   49.15   49.15   89.2180 < 2.2e-16 ***
## cons.price.idx:f.influentMonth     2   22.14   11.07   20.0937 2.034e-09 ***
## cons.conf.idx:nr.employed          1    2.28    2.28    4.1300 0.0421833 *
## cons.conf.idx:contact              1    4.01    4.01    7.2837 0.0069816 **
## cons.conf.idx:f.influentMonth      1    0.26    0.26    0.4676 0.4941148
## nr.employed:contact                1    7.34    7.34   13.3311 0.0002637 ***
## nr.employed:f.influentMonth        1    1.10    1.10    1.9968 0.1576951
## contact:f.influentMonth            2    2.49    1.25    2.2637 0.1040725
## Residuals                       4981 2744.18    0.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To do the interactions we will choose cons.priceidx-influent-month as factor-covariate interaction and contact-nr.employed as 2-factor interaction.

```
m9<-lm(log(duration)~cons.price.idx*f.influentMonth+cons.conf.idx+nr.employed+contact*nr.employed,data=
summary(m9)
```

```
##
## Call:
## lm(formula = log(duration) ~ cons.price.idx * f.influentMonth +
##     cons.conf.idx + nr.employed + contact * nr.employed, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9851 -0.4164 -0.0029  0.4275  2.8259
##
## Coefficients:
##                                                                 Estimate
## (Intercept)                                                   -6.321e+01
## cons.price.idx                                                 1.297e-01
## f.influentMonthf.influentMonth.sep-may-jul                     9.217e+01
## f.influentMonthf.influentMonth.mar-dec-oct-nov                -1.758e+00
## cons.conf.idx                                                 -5.678e-02
## nr.employed                                                    1.068e-02
## contacttelephone                                               1.163e+01
```

```
## cons.price.idx:f.influentMonthf.influentMonth.sep-may-jul     -9.873e-01
## cons.price.idx:f.influentMonthf.influentMonth.mar-dec-oct-nov  1.405e-02
## nr.employed:contacttelephone                                  -2.335e-03
##                                                               Std. Error
## (Intercept)                                                    8.783e+00
## cons.price.idx                                                 1.299e-01
## f.influentMonthf.influentMonth.sep-may-jul                     7.022e+00
## f.influentMonthf.influentMonth.mar-dec-oct-nov                 1.402e+01
## cons.conf.idx                                                  6.002e-03
## nr.employed                                                    7.591e-04
## contacttelephone                                               5.999e+00
## cons.price.idx:f.influentMonthf.influentMonth.sep-may-jul      7.513e-02
## cons.price.idx:f.influentMonthf.influentMonth.mar-dec-oct-nov  1.504e-01
## nr.employed:contacttelephone                                   1.161e-03
##                                                               t value Pr(>|t|)
## (Intercept)                                                    -7.197 7.09e-13
## cons.price.idx                                                  0.999   0.3181
## f.influentMonthf.influentMonth.sep-may-jul                     13.126  < 2e-16
## f.influentMonthf.influentMonth.mar-dec-oct-nov                 -0.125   0.9003
## cons.conf.idx                                                  -9.461  < 2e-16
## nr.employed                                                    14.069  < 2e-16
## contacttelephone                                                1.938   0.0526
## cons.price.idx:f.influentMonthf.influentMonth.sep-may-jul     -13.141  < 2e-16
## cons.price.idx:f.influentMonthf.influentMonth.mar-dec-oct-nov   0.093   0.9256
## nr.employed:contacttelephone                                   -2.010   0.0445
##
## (Intercept)                                                   ***
## cons.price.idx
## f.influentMonthf.influentMonth.sep-may-jul                    ***
## f.influentMonthf.influentMonth.mar-dec-oct-nov
## cons.conf.idx                                                 ***
## nr.employed                                                   ***
## contacttelephone                                              .
## cons.price.idx:f.influentMonthf.influentMonth.sep-may-jul     ***
## cons.price.idx:f.influentMonthf.influentMonth.mar-dec-oct-nov
## nr.employed:contacttelephone                                  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7507 on 4990 degrees of freedom
## Multiple R-squared:  0.3398, Adjusted R-squared:  0.3386
## F-statistic: 285.4 on 9 and 4990 DF,  p-value: < 2.2e-16
```

```r
m10<-lm(log(duration)~campaign+contact*f.influentMonth+nr.employed,data=df)
m11<-lm(log(duration)~campaign*f.influentMonth+contact+nr.employed,data=df)
```

```r
AIC(m9,m10)
```

```
##     df      AIC
## m9  11 11334.05
## m10  9 11453.30
```

```r
AIC(m9,m11)
```

```
##     df      AIC
## m9  11 11334.05
```

```
## m11  9 11934.33
```

```r
AIC(m10,m11)
```

```
##     df      AIC
## m10  9 11453.30
## m11  9 11934.33
```

```r
vif(m9,type="predictor")
```

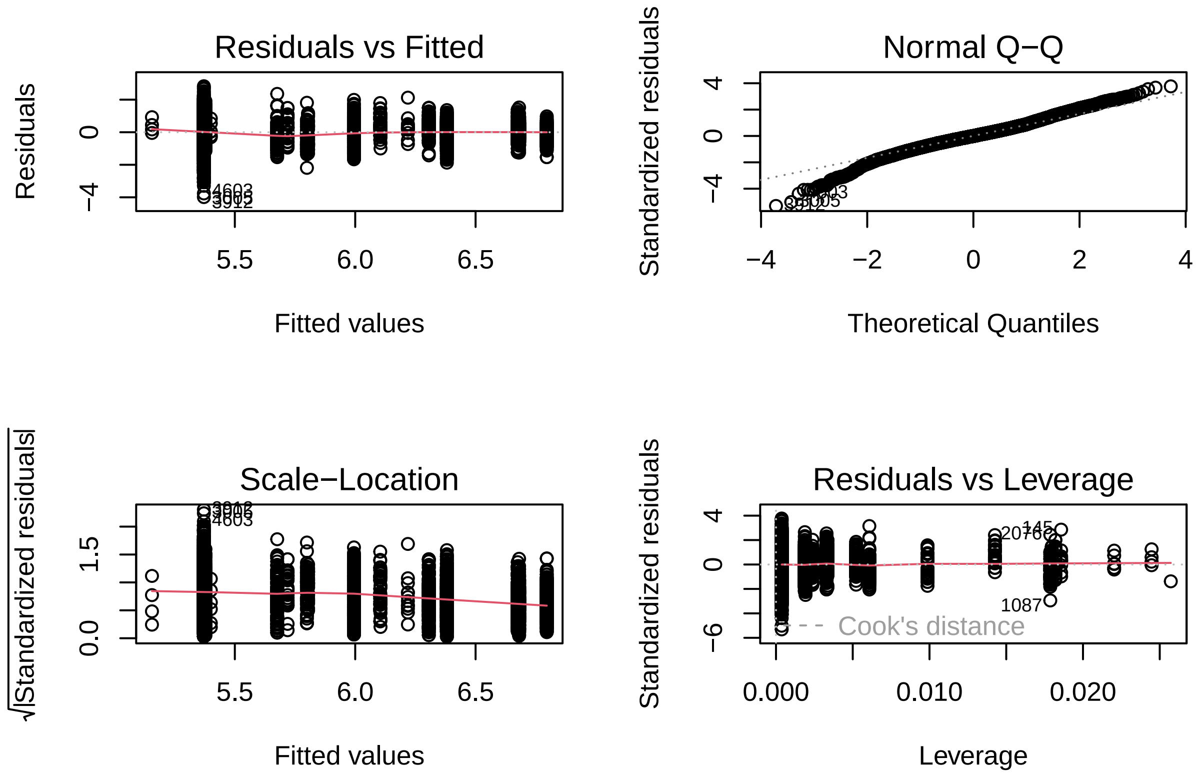
```
## GVIFs computed for predictors

##                      GVIF Df GVIF^(1/(2*Df))  Interacts With
## cons.price.idx  38.624833  5        1.441075 f.influentMonth
## f.influentMonth 38.624833  5        1.441075  cons.price.idx
## cons.conf.idx    6.243172  1        2.498634              --
## nr.employed     44.001353  3        1.878932         contact
## contact         44.001353  3        1.878932     nr.employed
##                                                 Other Predictors
## cons.price.idx                  cons.conf.idx, nr.employed, contact
## f.influentMonth                 cons.conf.idx, nr.employed, contact
## cons.conf.idx   cons.price.idx, f.influentMonth, nr.employed, contact
## nr.employed            cons.price.idx, f.influentMonth, cons.conf.idx
## contact                cons.price.idx, f.influentMonth, cons.conf.idx
```

We want to compare all the interactions, including them all in a single model or one interaction at a time and from the AIC criteria the best one seems to be the model 9, having the the two interactions at the same time, because it has the lowest AIC value. So we think that this is the best model so far in our modelling process and we will proceed to validate it.

4.4 Validation

After selecting the best model, Model 9, which incorporates both numerical and categorical factors along with their interactions, we will now proceed with the crucial step of model validation. Model validation aims to assess the performance and reliability of the chosen model on unseen data, ensuring its generalizability and usefulness in real-world scenarios.

```r
par(mfrow=c(2,2))
plot(m9)
```




```
par(mfrow=c(1,1))
```

We want to verify that our models complies with the linear regressions assumptions, we will cover all four of them seeing the graph above:

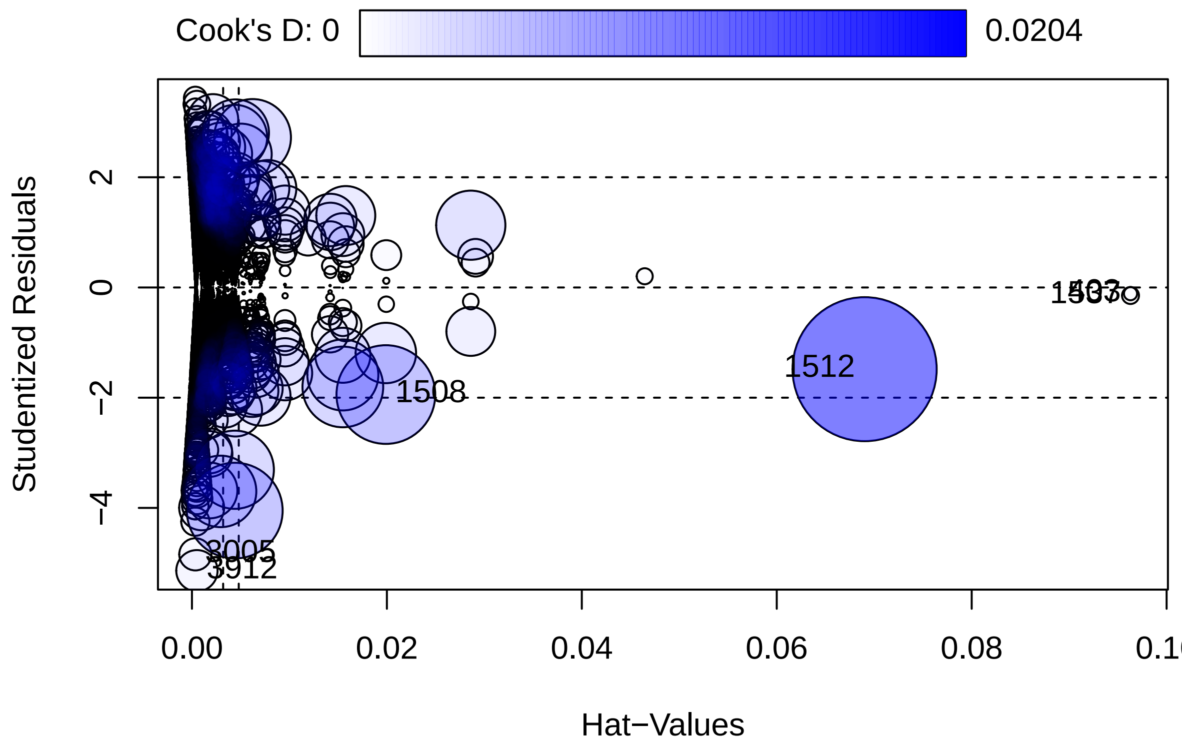
- Normality: Normality: From the Residual vs Fitted graph, we observe that the data points closely follow a straight line with minor deviations. This indicates that the residuals are approximately normally distributed. Hence, we can reasonably assume that our model satisfies the normality assumption.
- Linearity: In the Residual vs Fitted graph, the red line representing the model's fitted values aligns closely with the dotted line. This suggests that the relationship between the predictors and the response variable is adequately captured by a linear relationship. Thus, we can conclude that our model meets the linearity assumption.
- Homoscedasticity: Analyzing the Scale-Location graph, we observe that the residuals do not exhibit any discernible pattern, such as a cone-shaped or fan-shaped dispersion. This lack of a clear pattern indicates that the variability of the residuals is consistent across different levels of the predictor variables. As a result, we can assume that our model fulfills the homoscedasticity assumption.
- Independence: In the Residual vs Fitted graph, the scattered points appear to be randomly distributed across the plot without displaying any noticeable pattern or trend. This randomness suggests that the residuals are not systematically related to each other, supporting the assumption of independence. Therefore, we can infer that our model satisfies the independence assumption.

All in all, we can see that our model is valid because it complies with all the four assumptions of a linear regression model.

4.5 Lack of fit observations and influence data

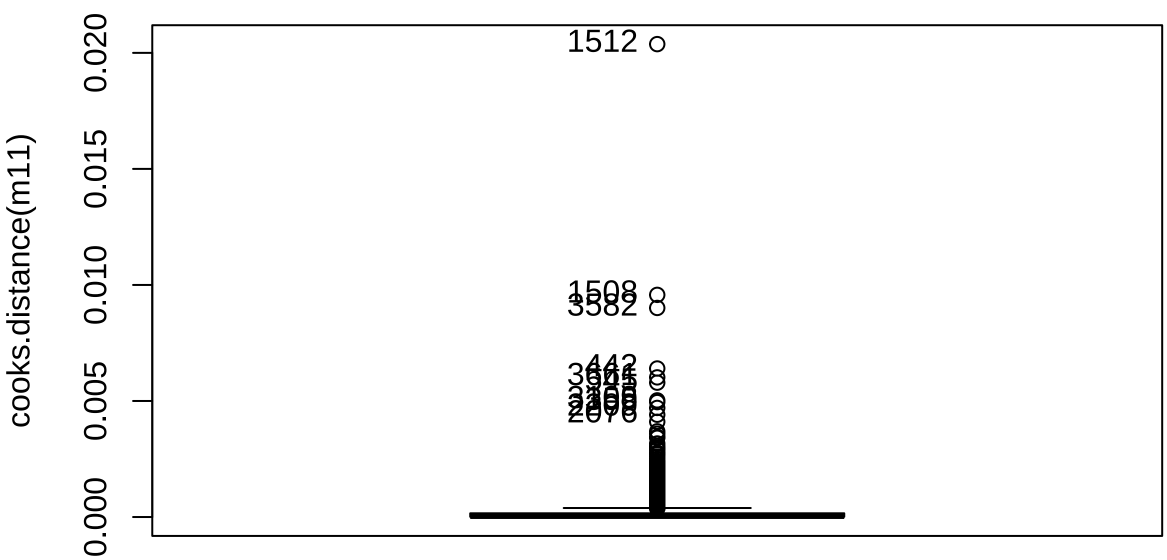
Now we will discuss the lack of fit observations and influence data .

```
par(mfrow=c(1,1))
influencePlot(m11)
```




```
##           StudRes          Hat        CookD
## 403  -0.1128683 0.0962959346 0.0001697154
## 1508 -1.9429471 0.0198942446 0.0095729357
## 1512 -1.4830595 0.0690161631 0.0203765886
## 1537 -0.1456746 0.0962959346 0.0002827121
## 3005 -4.8425575 0.0003486725 0.0010178412
## 3912 -5.1396164 0.0005094378 0.0016744738
```

```
Boxplot(cooks.distance(m11))
```



```
##  [1] 1512 1508 3582  442 3661  545 3108  359 2209 2076
```

Based on the influential plot and Cook's distance, we have identified three individuals in the dataset who exert a significant influence on the model. The influential plot provides a visual representation of the influence of each observation on the model's fit, while Cook's distance quantifies the impact of each observation on the overall model performance.

To ensure the robustness and reliability of our model, we have decided to remove these three influential individuals from the dataset. By excluding these observations, we aim to mitigate their disproportionate influence, which could potentially affect the model's coefficients and predictions.

```
which(row.names(df)==1512)
```

```
## [1] 1512
```

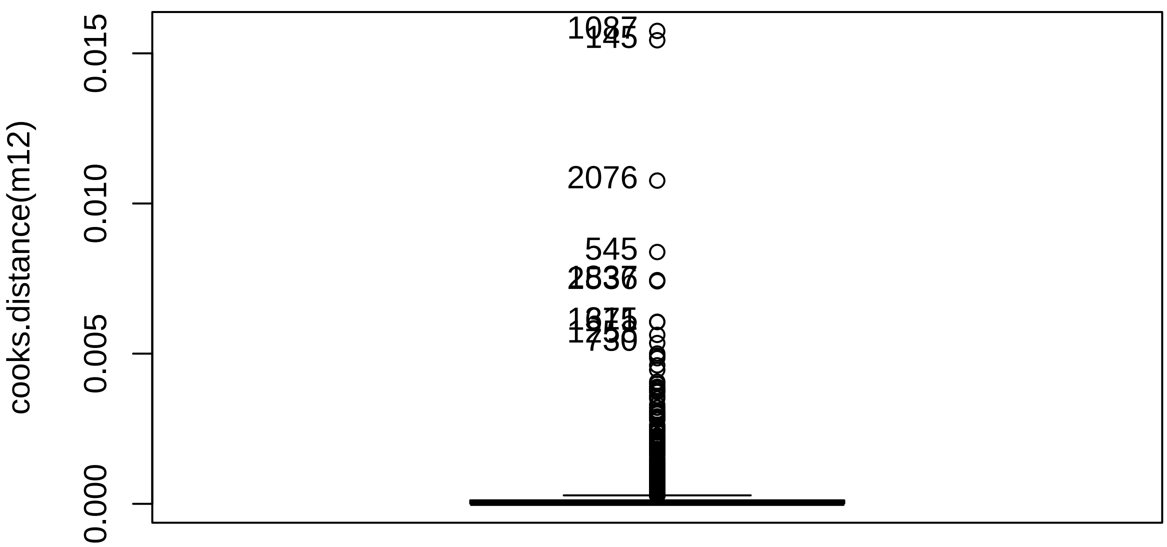
```
which(row.names(df)==1508)
```

```
## [1] 1508
```

```
which(row.names(df)==3582)
```

```
## [1] 3582
```

```
m12<-lm(log(duration)~cons.price.idx*f.influentMonth+cons.conf.idx+nr.employed+contact*nr.employed,data=
Boxplot(cooks.distance(m12))
```




```
##  [1] 1087  145 2076  545 1837 2536  315 1671 1258  730
```

```
summary(m12)
```

```
##
## Call:
## lm(formula = log(duration) ~ cons.price.idx * f.influentMonth +
##     cons.conf.idx + nr.employed + contact * nr.employed, data = df[,
##     c(-1512, -1508, -3582)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9851 -0.4164 -0.0029  0.4275  2.8259
##
## Coefficients:
##                                                                  Estimate
```

```
## (Intercept)                                                -6.321e+01
## cons.price.idx                                              1.297e-01
## f.influentMonthf.influentMonth.sep-may-jul                  9.217e+01
## f.influentMonthf.influentMonth.mar-dec-oct-nov             -1.758e+00
## cons.conf.idx                                              -5.678e-02
## nr.employed                                                 1.068e-02
## contacttelephone                                            1.163e+01
## cons.price.idx:f.influentMonthf.influentMonth.sep-may-jul  -9.873e-01
## cons.price.idx:f.influentMonthf.influentMonth.mar-dec-oct-nov  1.405e-02
## nr.employed:contacttelephone                               -2.335e-03
##                                                            Std. Error
## (Intercept)                                                 8.783e+00
## cons.price.idx                                              1.299e-01
## f.influentMonthf.influentMonth.sep-may-jul                  7.022e+00
## f.influentMonthf.influentMonth.mar-dec-oct-nov              1.402e+01
## cons.conf.idx                                               6.002e-03
## nr.employed                                                 7.591e-04
## contacttelephone                                            5.999e+00
## cons.price.idx:f.influentMonthf.influentMonth.sep-may-jul   7.513e-02
## cons.price.idx:f.influentMonthf.influentMonth.mar-dec-oct-nov  1.504e-01
## nr.employed:contacttelephone                                1.161e-03
##                                                            t value Pr(>|t|)
## (Intercept)                                                 -7.197 7.09e-13
## cons.price.idx                                               0.999   0.3181
## f.influentMonthf.influentMonth.sep-may-jul                  13.126  < 2e-16
## f.influentMonthf.influentMonth.mar-dec-oct-nov              -0.125   0.9003
## cons.conf.idx                                               -9.461  < 2e-16
## nr.employed                                                 14.069  < 2e-16
## contacttelephone                                             1.938   0.0526
## cons.price.idx:f.influentMonthf.influentMonth.sep-may-jul  -13.141  < 2e-16
## cons.price.idx:f.influentMonthf.influentMonth.mar-dec-oct-nov   0.093   0.9256
## nr.employed:contacttelephone                                -2.010   0.0445
##
## (Intercept)                                                ***
## cons.price.idx
## f.influentMonthf.influentMonth.sep-may-jul                 ***
## f.influentMonthf.influentMonth.mar-dec-oct-nov
## cons.conf.idx                                              ***
## nr.employed                                                ***
## contacttelephone                                           .
## cons.price.idx:f.influentMonthf.influentMonth.sep-may-jul  ***
## cons.price.idx:f.influentMonthf.influentMonth.mar-dec-oct-nov
## nr.employed:contacttelephone                               *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7507 on 4990 degrees of freedom
## Multiple R-squared:  0.3398, Adjusted R-squared:  0.3386
## F-statistic: 285.4 on 9 and 4990 DF,  p-value: < 2.2e-16
```

Deleting the influential individuals we improved the R2 up to 33.98%.

In conclusion, our modeling process began by considering all numerical variables and subsequently selecting the most significant ones. We then proceeded to analyze the categorical variables, identifying the most influential factors. By incorporating interactions between variables, we constructed our best model, which achieved an R-squared value of approximately 34%.

Throughout this process, we ensured that the selected variables exhibited acceptable Variance Inflation Factors (VIFs), indicating minimal multicollinearity. Moreover, we carefully assessed the model's adherence to the assumptions of linear regression and found that it complied with all of them, including normality, linearity, homoscedasticity, and independence.

This final model represents a significant improvement in explaining the variability in the target variable compared to the initial model. With an R-squared of nearly 34%, it suggests that approximately 34% of the variation in the response variable can be attributed to the selected predictors.

Overall, our comprehensive modeling approach, which involved systematic variable selection, incorporation of interactions, and rigorous assessment of model assumptions, has resulted in a robust and interpretable model. This model provides valuable insights into the relationships between the predictors and the target variable, enabling us to make more accurate predictions and informed decisions in the context of the marketing campaign dataset.

5 Binary target modelization

We start by splitting our sample in a training sample and a testing sample, for accomplishing this we randomly select 25% of the sample to create the testing set and the rest for the training.

```
set.seed(19101990)
sam <-sample(1:nrow(df),0.75*nrow(df))
dfw<-df[sam,]
dft<-df[-sam,]
```

5.1 Modelling with numerical variables

To begin our modeling process, we initially focus on the numerical variables within the dataset. Conducting a comprehensive analysis, we aim to identify the most significant variables that have a substantial impact on the target variable.

To achieve this, we perform a condes analysis. This analysis involves examining the relationships between each numerical predictor variable and the target variable. By calculating various statistical measures such as correlation coefficients, p-values from hypothesis tests, and effect sizes, we gain insights into the strength and significance of these associations.

Following the condes analysis, we will select the most significant numerical variables based on their statistical importance and relevance to our research objectives. These selected variables will serve as the foundation for further modeling steps, including feature engineering, model building, and assessment of model performance.

```
catdes(dfw[,c("y",vars_con,"duration")],1)
```

```
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                      Eta2       P-value
## cons.price.idx 0.391830839  0.000000e+00
## cons.conf.idx  0.560633457  0.000000e+00
## euribor3m      0.317663806 1.891462e-313
## duration       0.304718211 3.821545e-298
## emp.var.rate   0.296889945 5.013259e-289
## nr.employed    0.124066957 5.726030e-110
## age            0.011535113  4.285126e-11
## campaign       0.001093139  4.291421e-02
##
```

```
## Description of each cluster by quantitative variables
## ====================================================
## $no
##                     v.test Mean in category Overall mean sd in category
## cons.conf.idx    45.845554       -36.400000  -39.8149067   0.000000e+00
## cons.price.idx   38.327194        93.994000   93.6885347   0.000000e+00
## euribor3m        34.509732         4.856070    3.9746123   8.686026e-04
## emp.var.rate     33.362260         1.100000    0.3391467   0.000000e+00
## nr.employed      21.566804      5191.000000 5173.9592533   0.000000e+00
## age               6.576104        40.972467   39.8690667   8.852829e+00
## campaign         -2.024396         1.970942    2.0160672   1.257462e+00
## duration        -33.799239       240.796256  480.7064000   2.111737e+02
##                  Overall sd       p.value
## cons.conf.idx     4.4194581  0.000000e+00
## cons.price.idx    0.4728706  0.000000e+00
## euribor3m         1.5154698 5.731832e-261
## emp.var.rate      1.3531093 4.837739e-244
## nr.employed      46.8802848 3.682779e-103
## age               9.9552427  4.829358e-11
## campaign          1.3225328  4.292947e-02
## duration        421.1425227 2.023610e-250
##
## $yes
##                     v.test Mean in category Overall mean sd in category
## duration         33.799239      705.9788004  480.7064000    444.1107507
## campaign          2.024396        2.0584387    2.0160672      1.3795013
## age              -6.576104       38.8329886   39.8690667     10.7870009
## nr.employed     -21.566804     5157.9582213 5173.9592533     61.0960447
## emp.var.rate    -33.362260       -0.3752844    0.3391467      1.5799085
## euribor3m       -34.509732        3.1469354    3.9746123      1.7431463
## cons.price.idx -38.327194       93.4017068   93.6885347      0.5135017
## cons.conf.idx   -45.845554      -43.0214581  -39.8149067      4.0791522
##                  Overall sd       p.value
## duration        421.1425227 2.023610e-250
## campaign          1.3225328  4.292947e-02
## age               9.9552427  4.829358e-11
## nr.employed      46.8802848 3.682779e-103
## emp.var.rate      1.3531093 4.837739e-244
## euribor3m         1.5154698 5.731832e-261
## cons.price.idx    0.4728706  0.000000e+00
## cons.conf.idx     4.4194581  0.000000e+00
```

We can see that all variables have p-values < 0.05 so we will choose all of them.

```r
gm1<-glm(y ~
duration +
nr.employed +
euribor3m +
emp.var.rate +
campaign +
age+
cons.price.idx+
cons.conf.idx
, family = binomial, data = dfw[,c("y",vars_con,"duration")])
```

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(gm1)
```

```
##
## Call:
## glm(formula = y ~ duration + nr.employed + euribor3m + emp.var.rate +
##     campaign + age + cons.price.idx + cons.conf.idx, family = binomial,
##     data = dfw[, c("y", vars_con, "duration")])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.2030  -0.1437   0.0000   0.0000   3.1808
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     8.589e+04  1.046e+06   0.082    0.935
## duration        5.777e-03  3.951e-04  14.623   <2e-16 ***
## nr.employed    -6.699e+00  3.012e+02  -0.022    0.982
## euribor3m       1.185e+03  1.423e+02   8.332   <2e-16 ***
## emp.var.rate   -5.626e+02  9.442e+03  -0.060    0.952
## campaign        7.409e-02  9.929e-02   0.746    0.456
## age            -8.511e-03  1.418e-02  -0.600    0.548
## cons.price.idx -6.513e+02  1.475e+04  -0.044    0.965
## cons.conf.idx  -1.363e+02  8.724e+02  -0.156    0.876
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5194.89  on 3749  degrees of freedom
## Residual deviance:  472.48  on 3741  degrees of freedom
## AIC: 490.48
##
## Number of Fisher Scoring iterations: 25
```

Based on the summary of our analysis, we have determined that out of the numerical variables, only "duration" and "euribor3m" demonstrate statistical significance in relation to our target variable. As a result, we will proceed with including only these two variables in our modeling process.

By selecting "duration" and "euribor3m" as our predictors, we aim to build a simplified yet effective model that focuses on the most influential numerical factors in predicting the target variable. This streamlined approach not only reduces the complexity of the model but also ensures that we concentrate our efforts on the variables that have the greatest impact on the outcome of interest.

```
gm2<-glm(y ~
duration +
euribor3m
, family = binomial, data = dfw[,c("y",vars_con,"duration")])
summary(gm2)
```

```
##
## Call:
## glm(formula = y ~ duration + euribor3m, family = binomial, data = dfw[,
##     c("y", vars_con, "duration")])
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.5040  -0.3571   0.0013   0.0938   2.8699
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.8594404  1.0699067   7.346 2.04e-13 ***
## duration     0.0075029  0.0002733  27.452  < 2e-16 ***
## euribor3m   -2.4635534  0.2231780 -11.039  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5194.9  on 3749  degrees of freedom
## Residual deviance: 1737.2  on 3747  degrees of freedom
## AIC: 1743.2
##
## Number of Fisher Scoring iterations: 9
```

```
vif(gm2)
```

```
##  duration euribor3m
##  1.026079  1.026079
```

We see that from summary, after only keeping those variables they are still significant, the residual variance is significatly lower than the null deviance and the vif's are equal to one so it seems like a good model so far.

5.2 Including factors

As we did with the numerical variables, now we will do the same for the factors using catdes, and we can see that all of them have relation to the target so we will use them all and discard from there.

```
catdes(dfw[,c("y",vars_dis)],1)
```

```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## =================================================================================
##                  p.value df
## contact     0.000000e+00  1
## month       0.000000e+00  8
## Age_group   1.675553e-24  4
## marital     1.010629e-19  2
## education   3.759076e-17  4
## job         3.711092e-16 10
## day_of_week 1.060943e-13  4
## housing     3.678538e-05  1
##
## Description of each cluster by the categories
## =============================================
## $no
##                              Cla/Mod     Mod/Cla     Global       p.value
## month=may                   75.63515 100.0000000 64.0266667  0.000000e+00
## contact=telephone           79.78910 100.0000000 60.6933333  0.000000e+00
## marital=married             53.57143  68.5572687 61.9733333  7.430167e-16
```

```
## education=basic             56.45412  39.9779736 34.2933333  1.180019e-12
## job=blue-collar             57.06638  29.3502203 24.9066667  1.080680e-09
## day_of_week=mon             56.32603  25.4955947 21.9200000  2.944095e-07
## Age_group=30-50             51.20411  71.4207048 67.5466667  8.927218e-07
## day_of_week=tue             55.25000  24.3392070 21.3333333  1.352624e-05
## housing=no                  51.89665  51.9823789 48.5066667  3.692704e-05
## Age_group=40-60             55.59105  19.1629956 16.6933333  8.604192e-05
## job=services                54.21053  11.3436123 10.1333333  1.748912e-02
## job=housemaid               59.77011   2.8634361  2.3200000  3.295057e-02
## job=retired                 35.25180   2.6982379  3.7066667  1.475816e-03
## Age_group=10-20              0.00000   0.0000000  0.3466667  1.790679e-04
## housing=yes                 45.15795  48.0176211 51.4933333  3.692704e-05
## day_of_week=thu             41.78168  18.3370044 21.2533333  2.279424e-05
## day_of_week=wed             40.83095  15.6938326 18.6133333  8.227871e-06
## job=admin.                  41.95652  21.2555066 24.5333333  6.016139e-06
## Age_group=NA                21.42857   0.9911894  2.2400000  2.850217e-07
## job=student                 18.51852   0.8259912  2.1600000  1.875671e-08
## month=oct                    0.00000   0.0000000  0.9066667  1.448475e-10
## education=university.degree 38.06510  23.1828194 29.4933333  1.732907e-16
## Age_group=20-30             30.97166   8.4251101 13.1733333  3.616735e-17
## marital=single              36.09756  20.3744493 27.3333333  1.265543e-20
## month=mar                    0.00000   0.0000000  2.4533333  1.193206e-27
## month=nov                    0.00000   0.0000000  3.6800000  1.770764e-41
## month=aug                    0.00000   0.0000000  5.0400000  4.125043e-57
## month=jun                    0.00000   0.0000000  6.8266667  3.752721e-78
## month=jul                    0.00000   0.0000000  8.3200000  3.357427e-96
## month=apr                    0.00000   0.0000000  8.7200000 4.142529e-101
## contact=cellular             0.00000   0.0000000 39.3066667  0.000000e+00
##                                 v.test
## month=may                          Inf
## contact=telephone                  Inf
## marital=married               8.063235
## education=basic               7.107689
## job=blue-collar               6.097014
## day_of_week=mon               5.126991
## Age_group=30-50               4.913922
## day_of_week=tue               4.351414
## housing=no                    4.125911
## Age_group=40-60               3.926916
## job=services                  2.376260
## job=housemaid                 2.132685
## job=retired                  -3.179397
## Age_group=10-20              -3.746852
## housing=yes                  -4.125911
## day_of_week=thu              -4.235600
## day_of_week=wed              -4.459167
## job=admin.                   -4.525821
## Age_group=NA                 -5.133091
## job=student                  -5.623094
## month=oct                    -6.410706
## education=university.degree  -8.239252
## Age_group=20-30              -8.424701
## marital=single               -9.311067
## month=mar                   -10.896847
```

```
## month=nov                     -13.490838
## month=aug                     -15.926869
## month=jun                     -18.714765
## month=jul                     -20.812178
## month=apr                     -21.347173
## contact=cellular                    -Inf
##
## $yes
##                                Cla/Mod   Mod/Cla     Global       p.value
## contact=cellular             100.00000 76.215098 39.3066667  0.000000e+00
## month=apr                    100.00000 16.907963  8.7200000 4.142529e-101
## month=jul                    100.00000 16.132368  8.3200000  3.357427e-96
## month=jun                    100.00000 13.236815  6.8266667  3.752721e-78
## month=aug                    100.00000  9.772492  5.0400000  4.125043e-57
## month=nov                    100.00000  7.135471  3.6800000  1.770764e-41
## month=mar                    100.00000  4.756980  2.4533333  1.193206e-27
## marital=single                63.90244 33.867632 27.3333333  1.265543e-20
## Age_group=20-30               69.02834 17.631851 13.1733333  3.616735e-17
## education=university.degree  61.93490 35.418821 29.4933333  1.732907e-16
## month=oct                    100.00000  1.758014  0.9066667  1.448475e-10
## job=student                   81.48148  3.412616  2.1600000  1.875671e-08
## Age_group=NA                  78.57143  3.412616  2.2400000  2.850217e-07
## job=admin.                    58.04348 27.611169 24.5333333  6.016139e-06
## day_of_week=wed               59.16905 21.354705 18.6133333  8.227871e-06
## day_of_week=thu               58.21832 23.991727 21.2533333  2.279424e-05
## housing=yes                   54.84205 54.756980 51.4933333  3.692704e-05
## Age_group=10-20              100.00000  0.672182  0.3466667  1.790679e-04
## job=retired                   64.74820  4.653568  3.7066667  1.475816e-03
## job=housemaid                 40.22989  1.809721  2.3200000  3.295057e-02
## job=services                  45.78947  8.996898 10.1333333  1.748912e-02
## Age_group=40-60               44.40895 14.374354 16.6933333  8.604192e-05
## housing=no                    48.10335 45.243020 48.5066667  3.692704e-05
## day_of_week=tue               44.75000 18.510858 21.3333333  1.352624e-05
## Age_group=30-50               48.79589 63.908997 67.5466667  8.927218e-07
## day_of_week=mon               43.67397 18.562565 21.9200000  2.944095e-07
## job=blue-collar               42.93362 20.734230 24.9066667  1.080680e-09
## education=basic               43.54588 28.955533 34.2933333  1.180019e-12
## marital=married               46.42857 55.791107 61.9733333  7.430167e-16
## month=may                     24.36485 30.248190 64.0266667  0.000000e+00
## contact=telephone             20.21090 23.784902 60.6933333  0.000000e+00
##                                 v.test
## contact=cellular                   Inf
## month=apr                    21.347173
## month=jul                    20.812178
## month=jun                    18.714765
## month=aug                    15.926869
## month=nov                    13.490838
## month=mar                    10.896847
## marital=single                9.311067
## Age_group=20-30               8.424701
## education=university.degree   8.239252
## month=oct                     6.410706
## job=student                   5.623094
## Age_group=NA                  5.133091
```

```
## job=admin.                    4.525821
## day_of_week=wed               4.459167
## day_of_week=thu               4.235600
## housing=yes                   4.125911
## Age_group=10-20               3.746852
## job=retired                   3.179397
## job=housemaid                -2.132685
## job=services                 -2.376260
## Age_group=40-60              -3.926916
## housing=no                   -4.125911
## day_of_week=tue              -4.351414
## Age_group=30-50              -4.913922
## day_of_week=mon              -5.126991
## job=blue-collar              -6.097014
## education=basic              -7.107689
## marital=married              -8.063235
## month=may                         -Inf
## contact=telephone                 -Inf
```

```r
gm3<-glm(y ~
duration +
euribor3m+
contact+
f.influentMonth+
marital+
education+
job+
day_of_week+
housing+
Age_group
,family = binomial, data = dfw)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
Anova(gm3)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                 LR Chisq Df Pr(>Chisq)
## duration          651.93  1  < 2.2e-16 ***
## euribor3m          48.43  1  3.423e-12 ***
## contact           399.78  1  < 2.2e-16 ***
## f.influentMonth   514.47  2  < 2.2e-16 ***
## marital             1.38  2    0.50184
## education           2.93  4    0.57014
## job                15.78 10    0.10596
## day_of_week         8.03  4    0.09062 .
## housing             0.01  1    0.92894
## Age_group           0.97  3    0.80765
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova we see that only contact and month are the signifcant variables so we will only keep those.

```
gm4<-glm(y ~
duration +
euribor3m+
contact+
f.influentMonth
, family = binomial, data = dfw)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Anova(gm4)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                 LR Chisq Df Pr(>Chisq)
## duration          668.29  1  < 2.2e-16 ***
## euribor3m          49.59  1  1.894e-12 ***
## contact           421.09  1  < 2.2e-16 ***
## f.influentMonth   529.18  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(gm4)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## duration        1.00423  1        1.002113
## euribor3m       1.00423  1        1.002113
## contact         1.00000  1        1.000000
## f.influentMonth 1.00000  2        1.000000
```

From the anova we can see that this model all the variables are significant and all the vif values are in acceptable ranges so it seems like a good model so far.

5.3 Interactions

Now we are going to see the interactions of our model, we will see all the interactions and choose factor-factor and covariate-factor for further modelling.

```
gm5<-glm(y ~
(duration +
euribor3m+
contact+
f.influentMonth)^2
, family = binomial, data = dfw)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Anova(gm4)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                 LR Chisq Df Pr(>Chisq)
## duration          668.29  1  < 2.2e-16 ***
## euribor3m          49.59  1  1.894e-12 ***
## contact           421.09  1  < 2.2e-16 ***
## f.influentMonth   529.18  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the interactions have very big values from factor-factor and covariate factor, but there isn't anything which can be done from previous models because the stats proved us the they were the most significant values and dont have collinearity, so we will proceed to choose two interaction either way to see how they perform. We will choose duration:contact and euribor3m:f.influentMonth.

```
gm6<-glm(y ~
duration*contact +
euribor3m*f.influentMonth
, family = binomial, data = dfw)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Anova(gm6)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                            LR Chisq Df Pr(>Chisq)
## duration                     668.29  1  < 2.2e-16 ***
## contact                      421.09  1  < 2.2e-16 ***
## euribor3m                     49.59  1  1.894e-12 ***
## f.influentMonth              529.18  2  < 2.2e-16 ***
## duration:contact               0.00  1     0.9995
## euribor3m:f.influentMonth      0.00  2     1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gm7<-glm(y ~
duration*contact +
euribor3m+f.influentMonth
, family = binomial, data = dfw)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
gm8<-glm(y ~
duration+contact +
euribor3m*f.influentMonth
, family = binomial, data = dfw)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
AIC(gm6,gm7)
```

```
##     df      AIC
## gm6  9 700.0342
## gm7  7 696.0342
```

```
AIC(gm6,gm8)
```

```
##     df      AIC
## gm6  9 700.0342
## gm8  8 698.0342
```

```
AIC(gm7,gm8)
```

```
##     df      AIC
## gm7  7 696.0342
## gm8  8 698.0342
```

```
Anova(gm7)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                  LR Chisq Df Pr(>Chisq)
## duration           668.29  1  < 2.2e-16 ***
## contact            421.09  1  < 2.2e-16 ***
## euribor3m           49.59  1  1.894e-12 ***
## f.influentMonth    529.18  2  < 2.2e-16 ***
## duration:contact     0.00  1          1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(gm7)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif

##                          GVIF Df GVIF^(1/(2*Df))
## duration         5.668212e+07  1     7528.752796
## contact          1.003431e+01  1        3.167699
## euribor3m        1.004230e+00  1        1.002113
## f.influentMonth  1.000000e+00  2        1.000000
## duration:contact 5.668213e+07  1     7528.753268
```

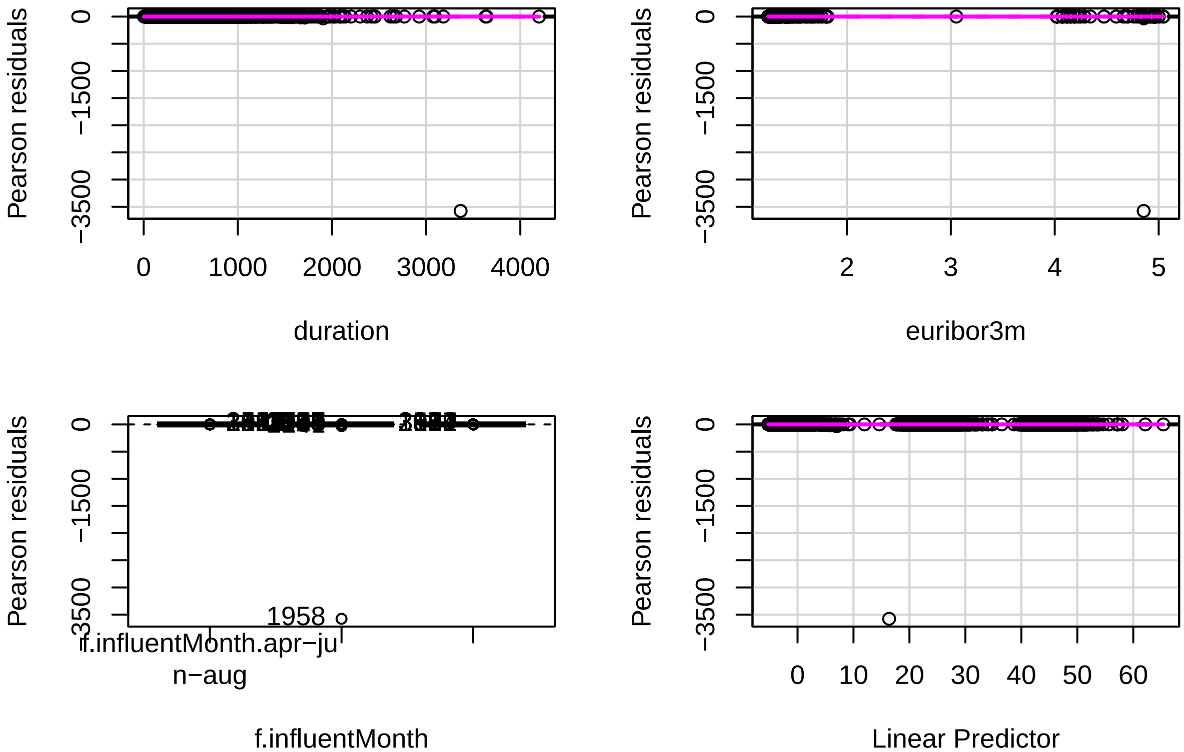
We can see that this model is unacceptable because the vif's are very high in 2 of the 5 variables. We tried the same thing for the other interactions model but the result is the same. So our best model so far is gm4 which includes 2 factors and 2 numerical variables without interactions

5.4 Validation

Now we will proceed to validate the best model we got, GM4.

```
residualPlots(gm4)
```

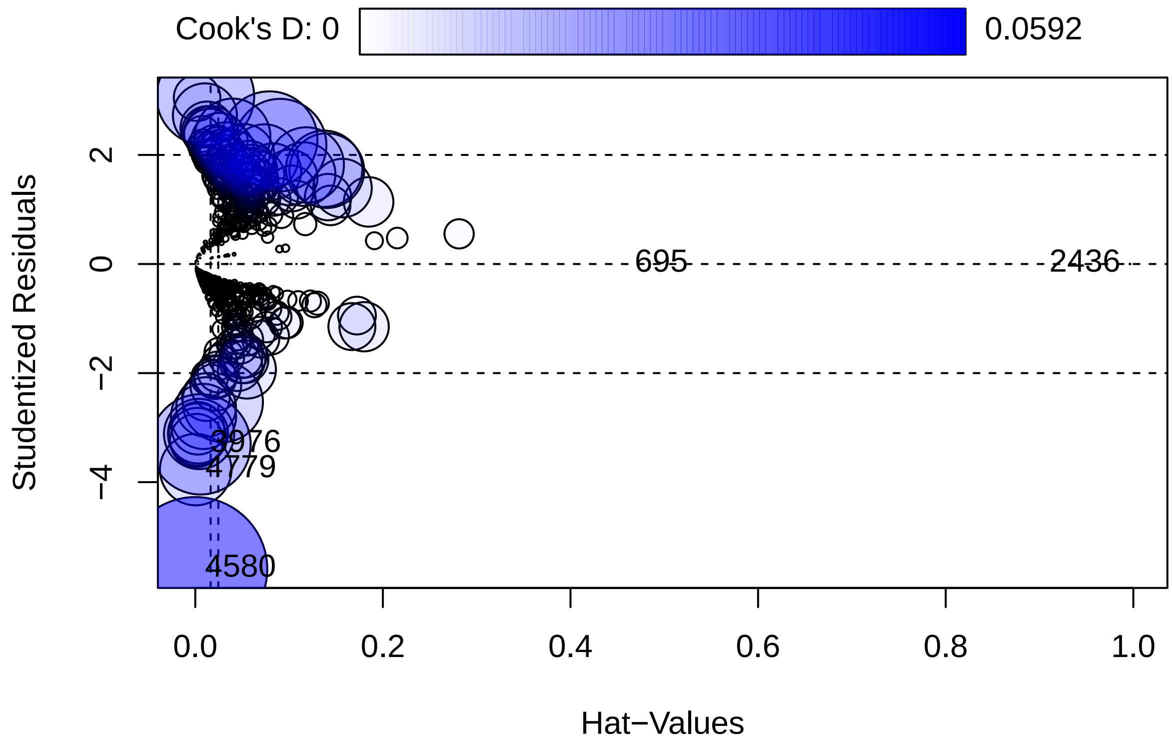
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```




```
##                  Test stat Pr(>|Test stat|)
## duration            87.536          <2e-16 ***
## euribor3m        -8545.141               1
## f.influentMonth
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from residual plots that there aren't really any influential individual apart from one in whichh we will check it with the influence plot.

```
influencePlot(gm3)
```




```
##            StudRes          Hat        CookD
## 4779 -3.769871e+00 4.474091e-04 1.460522e-02
## 2436  3.342670e-04 9.964642e-01 1.045952e-06
## 4580 -5.593324e+00 6.942648e-07 5.919568e-02
## 3976 -3.306446e+00 5.574801e-03 2.876078e-02
## 695   7.196431e-05 5.355414e-01 1.359196e-10
```

We can see two major individuals who influences quite a lot, 4580 and 4779 so we will delete them for our model.

```
gm4<-glm(y ~
duration +
euribor3m+
contact+
f.influentMonth
, family = binomial, data = dfw[c(-4580,-4779),])
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(gm4)
```

```
##
## Call:
## glm(formula = y ~ duration + euribor3m + contact + f.influentMonth,
##     family = binomial, data = dfw[c(-4580, -4779), ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.7209  -0.1843   0.0000   0.0000   2.8628
##
## Coefficients:
##                                                    Estimate Std. Error z value
```

```
## (Intercept)                                     4.569e+01  1.538e+03   0.030
## duration                                        6.421e-03  3.567e-04  18.003
## euribor3m                                      -1.657e+00  4.456e-01  -3.718
## contacttelephone                               -2.142e+01  8.943e+02  -0.024
## f.influentMonthf.influentMonth.sep-may-jul     -2.147e+01  1.251e+03  -0.017
## f.influentMonthf.influentMonth.mar-dec-oct-nov  2.648e+00  2.442e+03   0.001
##                                                Pr(>|z|)
## (Intercept)                                    0.976294
## duration                                        < 2e-16 ***
## euribor3m                                      0.000201 ***
## contacttelephone                               0.980889
## f.influentMonthf.influentMonth.sep-may-jul     0.986304
## f.influentMonthf.influentMonth.mar-dec-oct-nov 0.999135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5194.89  on 3749  degrees of freedom
## Residual deviance:  682.03  on 3744  degrees of freedom
## AIC: 694.03
##
## Number of Fisher Scoring iterations: 21
```

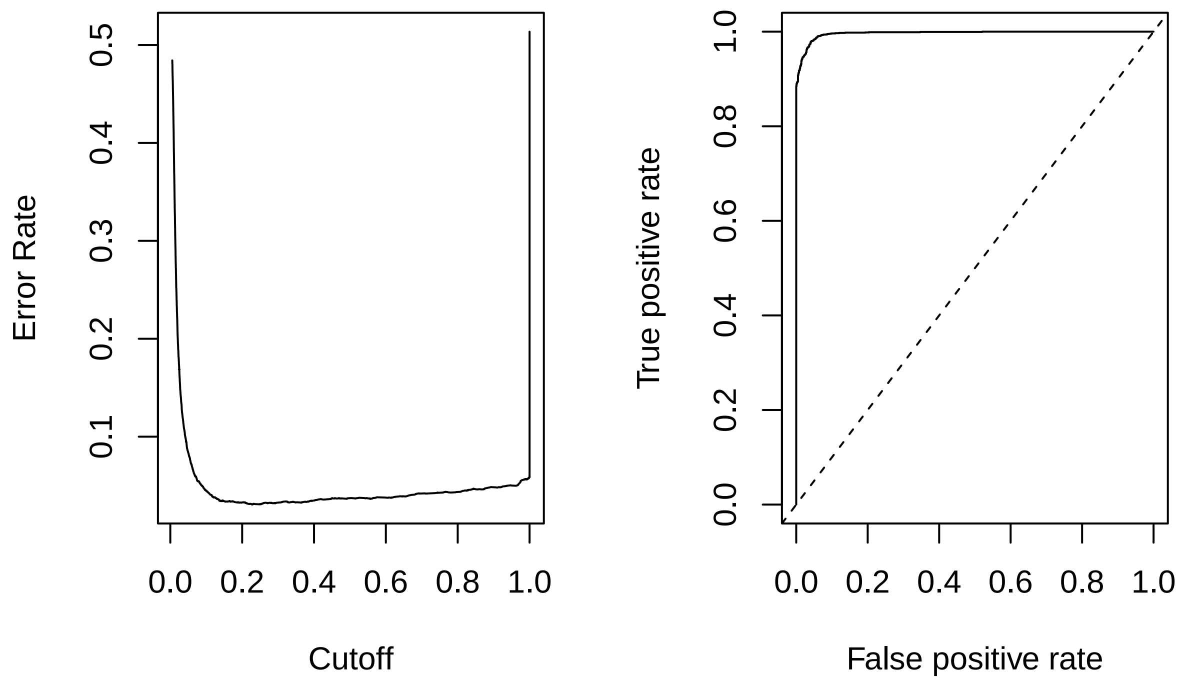
```
Anova(gm4)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                 LR Chisq Df Pr(>Chisq)
## duration          668.29  1  < 2.2e-16 ***
## euribor3m          49.59  1  1.894e-12 ***
## contact           421.09  1  < 2.2e-16 ***
## f.influentMonth   529.18  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from anova that the regressors are significant and the residual deviance is way lower compared to the null deviance so the model seems valid.

```
dataroc<-prediction(predict(gm5, type="response"),dfw$y)
par(mfrow=c(1,2))
plot(performance(dataroc,"err"))
plot(performance(dataroc,"tpr","fpr"))
abline(0,1,lty=2)
```



From the ROC curves we can see that ours falls into excellent category from the slides we've seen in class.But in the other graph we see something strange happening when cutoff=1, but apart from that seems quite good as well.

```
fittedSamplesTest=predict(gm5, newdata=dft, type="response")
fittedTest=ifelse(fittedSamplesTest<0.5,"No","Yes" )
ConfMatTest=table(dft$y,fittedTest)
ConfMatTest
```

```
##      fittedTest
##        No Yes
##   no  571  13
##   yes  27 639
```

```
accuracy = (ConfMatTest[1,1]+ConfMatTest[2,2])/sum(ConfMatTest)
error_rate = (ConfMatTest[1,2] + ConfMatTest[2,1])/sum(ConfMatTest)
sensibilty = ConfMatTest[2,2]/(ConfMatTest[2,2]+ ConfMatTest[2,1])
specificity = ConfMatTest[1,1]/(ConfMatTest[1,1]+ ConfMatTest[1,2])
```

```
accuracy*100
```

```
## [1] 96.8
```

```
error_rate*100
```

```
## [1] 3.2
```

```
sensibilty*100
```

```
## [1] 95.94595
```

```
specificity*100
```

```
## [1] 97.77397
```

We have an accuracy of 96.8%. We have a recall of 95.3% which means that the positive results of this confusion table is very accurate. We can see that we have 571 + 13 positive observations, from which 571 of them have been correctly classified. Now, we are going to do the same, but for the negative results (specificity). We can see that only a 97.77% of specificity, which is an ecellent result. 639 of the 27 + 639 negative observations have been classified as negative so it's very precise. To conclude, we see that the error rate is only of 3.2% which is amazing.

In conlusion, the results suggest that the model exhibits a remarkable level of accuracy and precision in both positive and negative predictions. With a high accuracy rate, strong recall for positive instances, and excellent specificity for negative instances, the model demonstrates its effectiveness in correctly classifying observations.

It is important to note that these performance metrics should be interpreted in the context of the specific problem and dataset being analyzed. However, based on the provided information, the model's performance appears to be impressive, with a low error rate indicating its reliability and efficacy.