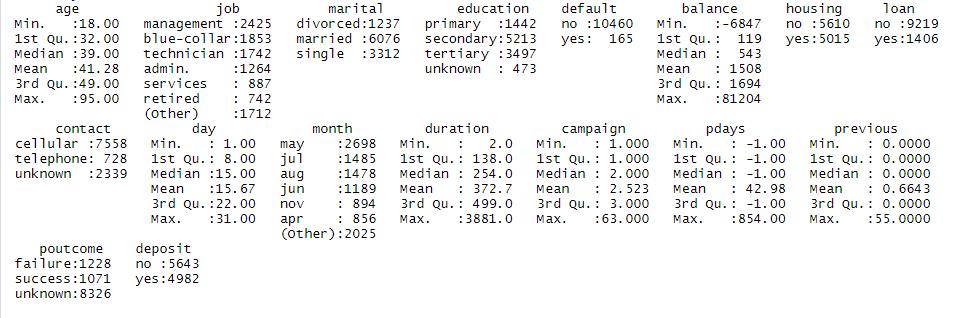
d=read.csv("D:/Old\_Data/math/Data science toseeh/Files/bank.csv")

summary(d)



هدف این ست که کل داده را به دو قسمت داده اموزشی و داده تست تقسیم کنیم. به این منظور 20 درصد داده ها را به طور تصادفی به عنوان داده تست در نظر میگیریم.

s=sample(nrow(d),floor(0.2\*nrow(d)),replace = FALSE)

داده تست روی سطر های تعیین شده و همه ستون ها تعریف میشود و داده اموزشی میشه بقیه سطرها به همرام تمام ستون ها. ابتدا ستون duration را حذف میکنم. چون در ادامه گفته شده که این ستون در نظر گرفته نشود.

d=d[,-12]

dtest=d[s,]

dtrain=d[-s,]

str(dtrain$day)

ویژگی مربوط به روز به صورت عدد در نظر گرفته شده و میانگین...برای ان محاسبه شده در حالیکه باید فاکتور باشد (روز اول، روز دوم...)

dtrain$day=factor(dtrain$day)

حال میخواهیم مدل را پیدا کنیم وپس از تعیین ویژگی های تاثیر گذار مدل را فیت کنیم و متغیر پاسخ را برای داده تست با استفاده از مدل فیت شده پیش بینی کنیم. سپس لیبل های واقعی متغیر پاسخ در داده تست را با لیبل های پیش بینی شده مقایسه کنیم و درصد تطبیق را پیدا کنیم.

در روش backward step از مدل کامل استفاده میکنیم. سپس در هر مرحله متغیری که حذف آن باعث میشه مقدار

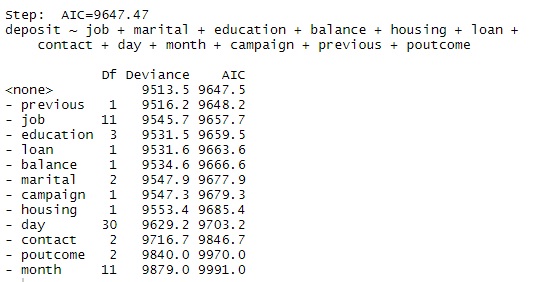
AIC مدل جدید(یعنی بدون یک متغیر تعیین شده) از AIC مدل کامل کمتر شود را حذف میکنیم. در پایان متغیر هایی که باقی میماند را در مدل قرار میدهیم.

چون متغیر پاسخ از نوع کیفی با دو سطح است (سرمایه گذاری کردن و سرمایه گذاری نکردن) پس توزیع آن برنولی است و از رگرسیون لاجیستیک استفاده میکنیم.

m=glm(deposit~.,family=binomial,data=dtrain)

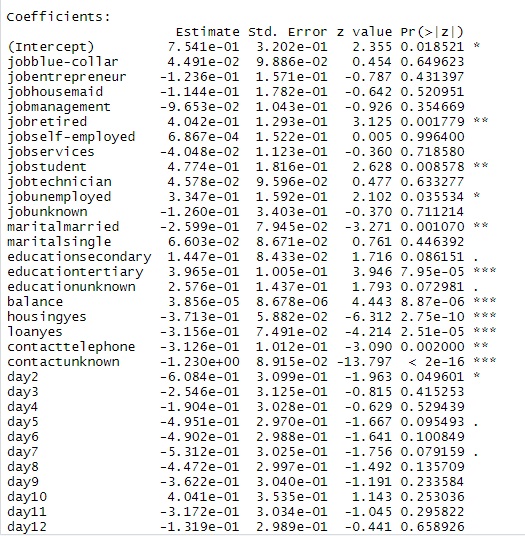
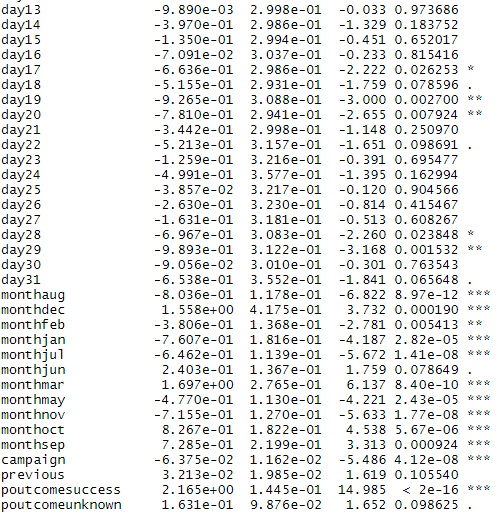
stb=step(m,direction ="backward")

output:



روی متغیر های تعیین شده در مرحله backward step مدل فیت شده است. این روش خود یک رگرسیون است پس:

summary(stb)



توصیف ضرایب: دو نمونه از ضرایب را توصیف میکنم.

ضریب jobbluecolor برابر است با 4.491e-02

متغیر شغل یک متغیر کیفی است با 12 سطح که سطح اول ان یعنی job admin به عنوان مبنا در نظر گرفته شده است. در حقیقت 11 متغیر به متغیر ها اصافه میکنیم که نشان دهنده 11 سطح شغل به غیر از ادمین هستند.

e\_i=(0,0,….,1,0,…0) نشان دهنده سطح i ام است و بردار تمام صفر یعنی فرد شغل ادمین داشته.

احتمال سپرده گذاری کردن به سپرده گذاری نکردن درjobbluecolor، exp(4.491e-02)=1.04 برابر احتمال سپرده گذاری کردن به سپرده گذاری نکردن job admin. پس در حقیقت این مقدار برای jobbluecolor ها بیشتر از job admin ها است. پس اگر در این کمپین سراغ افرادی با شغل jobbluecolor بروند بهتر است تا job admin

ها.

توصیف مقدار پی: چون از الفا 0.05 کمتر است پس این متغیر در توصیف پاسخ تاثیر دارد و همچنین از نظر تاثیر با متغیر ادمین متفاوت است. (فرض اچ صفر: یکیسان بودن تاثیر ادمین و جاب کالر روی پاسخ)

توصیف ضریب day3

ضریب ان -2.546e-01 است. یعنی احتمال سپرده گذاری کردن به نکردن افراد در روز سوم،

exp(-2.546e-01)=0.77 برابر احتمال سپرده گذاری کردن به نکردن افراد در روز اول است. (روز اول را به عنوان مبنا در نظر گرفته)

چون ضریب منفی بود احتمال سپرده گذاری کردن به نکردن افراد در روز سوم کمتر است از روز اول.

توصیف مقدار پی: چون از الفا 0.05 بیشتر است پس در توصیف متغیر پاسخ تاثیر گذار نیست و همچنین میزان تاثیر گذاری آن روی پاسخ با روز اول متفاوت است.( یعنی روز اول روی توصیف پاسخ تاثیر دارد.)

حال ابتدا متغیر روز در مدل تست را نیز به فاکتور تبدیل کرده و با استفاده از مدل فیت شده، متغیر پاسخ را پیش بینی می کنیم.

dtest$day=factor(dtest$day)

phat=predict(stb,dtest,type ="response")

head(phat)

output:

9642 2447 5216 195 10160

0.4280931 0.5821136 0.4016065 0.2740148 0.5105177

تمام مقادیری که از 0.5 بیشتر هستند را به عنوان سرمایه گذاری کردن در نظر گرفته و بقیه را به عنوان سرمایه گذاری نکردن.

yes\_no=function(x){

if (x>=0.5) {

x="yes"

{

else{

x="no"

{

return(x)

{

yhat=apply(as.matrix(phat), 1,yes\_no)

head(yhat)

output

9642 2447 5216 195 10160 7461

"no" "yes" "no" "no" "yes" "no"

درصد داده هایی از تست که لیبل واقعی متغیر پاسخ آن ها با لیبل پیش بینی شده یکسان است.

length(which(dtest$deposit==yhat))/length(dtest$deposit)

output: 0.71

یعنی از هر 100 داده 71 تا از آنها درست پیش بینی شده است.