modelling\_HR\_Data.R

Mrinal

Thu Aug 16 02:01:42 2018

HR\_Data=read.csv("E:\\R\_Lectures\\dr. mohit\\PROJECT\\HR\_Data.csv")  
View(HR\_Data)  
?ChickWeight

## starting httpd help server ...

## done

#################Creation of training and testing dataset. So that we can test the created model on both the sets to look at the training and testing accuracy in order to judge that how effective over modelled prediction structure will be on unseen data.  
l=sample(1:length(HR\_Data$left),floor(0.7\*length(HR\_Data$left)))  
train\_data=HR\_Data[l,]  
test\_data=HR\_Data[-l,]  
  
  
###################Modelling on train data.  
lm\_train=glm(left~.,data=train\_data,family=binomial)  
summary(lm\_train)

##   
## Call:  
## glm(formula = left ~ ., family = binomial, data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2366 -0.6610 -0.4038 -0.1301 3.0707   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.4556001 0.2294388 -6.344 2.24e-10 \*\*\*  
## satisfaction\_level -4.0916977 0.1169310 -34.992 < 2e-16 \*\*\*  
## last\_evaluation 0.6836517 0.1775785 3.850 0.000118 \*\*\*  
## number\_project -0.3319875 0.0256016 -12.967 < 2e-16 \*\*\*  
## average\_montly\_hours 0.0048813 0.0006162 7.921 2.35e-15 \*\*\*  
## time\_spend\_company 0.2744785 0.0187597 14.631 < 2e-16 \*\*\*  
## Work\_accident -1.4822023 0.1062254 -13.953 < 2e-16 \*\*\*  
## promotion\_last\_5years -1.4840161 0.3140713 -4.725 2.30e-06 \*\*\*  
## departmenthr 0.2078171 0.1577280 1.318 0.187649   
## departmentIT -0.1059797 0.1453958 -0.729 0.466060   
## departmentmanagement -0.3133887 0.1892182 -1.656 0.097675 .   
## departmentmarketing 0.0906774 0.1578413 0.574 0.565640   
## departmentproduct\_mng -0.1611226 0.1546916 -1.042 0.297610   
## departmentRandD -0.7215195 0.1801978 -4.004 6.23e-05 \*\*\*  
## departmentsales -0.0300152 0.1227766 -0.244 0.806867   
## departmentsupport 0.0178546 0.1309889 0.136 0.891579   
## departmenttechnical 0.1354456 0.1270634 1.066 0.286438   
## salarylow 1.8490837 0.1512258 12.227 < 2e-16 \*\*\*  
## salarymedium 1.3184163 0.1522272 8.661 < 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: 11443.8 on 10498 degrees of freedom  
## Residual deviance: 8981.4 on 10480 degrees of freedom  
## AIC: 9019.4  
##   
## Number of Fisher Scoring iterations: 5

#Asteriks against every entry in the table genrated by the summary signifies the importance of coresponding variables on the result.  
  
#Training accuracy.  
lm\_train.prob =predict (lm\_train,type ="response")  
lm\_train.pred=rep("0",length(train\_data$left))  
lm\_train.pred[lm\_train.prob>0.5]="1"  
table(lm\_train.pred,train\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 7466 1623  
## 1 568 842

accu\_train=(sum((lm\_train.pred==train\_data$left)/length(train\_data$left)))\*100  
accu\_train

## [1] 79.13135

#Testing accuracy.  
lm\_test.prob =predict(lm\_train,test\_data,type ="response")  
lm\_test.pred=rep("0",length(test\_data$left))  
lm\_test.pred[lm\_test.prob>0.5]="1"  
table(lm\_test.pred,test\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 3153 709  
## 1 241 397

accu\_test=(sum((lm\_test.pred==test\_data$left)/length(test\_data$left)))\*100  
accu\_test

## [1] 78.88889

#  
print(paste("tarining accuracy:",accu\_train,"% testing accuracy:",accu\_test,"%" ))

## [1] "tarining accuracy: 79.1313458424612 % testing accuracy: 78.8888888888889 %"

##################Modelling data only on the basis of variables with p values less than '2e-16'.  
lm\_train1=glm(left~.-department-promotion\_last\_5years-average\_montly\_hours-last\_evaluation ,data=train\_data,family=binomial)  
summary(lm\_train1)

##   
## Call:  
## glm(formula = left ~ . - department - promotion\_last\_5years -   
## average\_montly\_hours - last\_evaluation, family = binomial,   
## data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0877 -0.6684 -0.4257 -0.1522 2.8049   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.75524 0.18482 -4.086 4.38e-05 \*\*\*  
## satisfaction\_level -3.88275 0.11090 -35.012 < 2e-16 \*\*\*  
## number\_project -0.18426 0.01991 -9.255 < 2e-16 \*\*\*  
## time\_spend\_company 0.27133 0.01819 14.916 < 2e-16 \*\*\*  
## Work\_accident -1.50072 0.10526 -14.257 < 2e-16 \*\*\*  
## salarylow 1.93376 0.14868 13.006 < 2e-16 \*\*\*  
## salarymedium 1.38176 0.14979 9.224 < 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: 11443.8 on 10498 degrees of freedom  
## Residual deviance: 9160.5 on 10492 degrees of freedom  
## AIC: 9174.5  
##   
## Number of Fisher Scoring iterations: 5

#We can quite clearly see that the p value of every variable has moved more close to 0.  
  
#Training accuracy  
lm\_train1.probs =predict (lm\_train1,type ="response")  
lm\_train1.pred=rep("0",length(train\_data$left))  
lm\_train1.pred[lm\_train1.probs>0.5]="1"  
table(lm\_train1.pred,train\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 7433 1512  
## 1 601 953

accu\_train1=(sum((lm\_train1.pred==train\_data$left)/length(train\_data$left)))\*100  
accu\_train1

## [1] 79.87427

#Testing accuracy.  
lm\_test1.prob =predict(lm\_train1,test\_data,type ="response")  
lm\_test1.pred=rep("0",length(test\_data$left))  
lm\_test1.pred[lm\_test1.prob>0.5]="1"  
table(lm\_test1.pred,test\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 3143 648  
## 1 251 458

accu\_test1=(sum((lm\_test1.pred==test\_data$left)/length(test\_data$left)))\*100  
accu\_test1

## [1] 80.02222

#  
print(paste("tarining accuracy:",accu\_train1,"% testing accuracy:",accu\_test1,"%" ))

## [1] "tarining accuracy: 79.8742737403562 % testing accuracy: 80.0222222222222 %"

#We have quite clearly seen the improvement in train and test accuracy than what it was before.  
  
  
  
###############Modelling data after adding the interaction term(time\_spend\_company:promotion\_last\_5years) obtained on the basis of EDA.  
lm\_train3=glm(left~.-department-promotion\_last\_5years-average\_montly\_hours-last\_evaluation+time\_spend\_company:promotion\_last\_5years,data=train\_data,family=binomial)  
summary(lm\_train3)

##   
## Call:  
## glm(formula = left ~ . - department - promotion\_last\_5years -   
## average\_montly\_hours - last\_evaluation + time\_spend\_company:promotion\_last\_5years,   
## family = binomial, data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1404 -0.6677 -0.4203 -0.1420 3.3175   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) -0.71720 0.18393 -3.899  
## satisfaction\_level -3.89206 0.11116 -35.013  
## number\_project -0.18880 0.01998 -9.448  
## time\_spend\_company 0.28655 0.01845 15.534  
## Work\_accident -1.49893 0.10570 -14.181  
## salarylow 1.87190 0.14854 12.602  
## salarymedium 1.33577 0.14975 8.920  
## time\_spend\_company:promotion\_last\_5years -0.38196 0.07439 -5.135  
## Pr(>|z|)   
## (Intercept) 9.65e-05 \*\*\*  
## satisfaction\_level < 2e-16 \*\*\*  
## number\_project < 2e-16 \*\*\*  
## time\_spend\_company < 2e-16 \*\*\*  
## Work\_accident < 2e-16 \*\*\*  
## salarylow < 2e-16 \*\*\*  
## salarymedium < 2e-16 \*\*\*  
## time\_spend\_company:promotion\_last\_5years 2.83e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11443.8 on 10498 degrees of freedom  
## Residual deviance: 9114.5 on 10491 degrees of freedom  
## AIC: 9130.5  
##   
## Number of Fisher Scoring iterations: 5

#Training accuracy  
lm\_train3.probs =predict (lm\_train3,type ="response")  
lm\_train3.pred=rep("0",length(train\_data$left))  
lm\_train3.pred[lm\_train3.probs>0.5]="1"  
table(lm\_train3.pred,train\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 7437 1461  
## 1 597 1004

accu\_train3=(sum((lm\_train3.pred==train\_data$left)/length(train\_data$left)))\*100  
accu\_train3

## [1] 80.39813

#Testing accuracy.  
lm\_test3.prob =predict(lm\_train3,test\_data,type ="response")  
lm\_test3.pred=rep("0",length(test\_data$left))  
lm\_test3.pred[lm\_test3.prob>0.5]="1"  
table(lm\_test3.pred,test\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 3143 621  
## 1 251 485

accu\_test3=(sum((lm\_test3.pred==test\_data$left)/length(test\_data$left)))\*100  
accu\_test3

## [1] 80.62222

#  
print(paste("tarining accuracy:",accu\_train3,"% testing accuracy:",accu\_test3,"%" ))

## [1] "tarining accuracy: 80.3981331555386 % testing accuracy: 80.6222222222222 %"

#Both train and test accuracy has improved as compared to lm\_train1.  
  
###########Modelling data after adding the interaction term(average\_montly\_hours:salary) obtained on the basis of EDA.  
lm\_train4=glm(left~.-department-promotion\_last\_5years-average\_montly\_hours-last\_evaluation+average\_montly\_hours:salary,data=train\_data,family=binomial)  
summary(lm\_train4)

##   
## Call:  
## glm(formula = left ~ . - department - promotion\_last\_5years -   
## average\_montly\_hours - last\_evaluation + average\_montly\_hours:salary,   
## family = binomial, data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2539 -0.6637 -0.4147 -0.1464 2.9020   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.3206363 0.5503498 0.583 0.560  
## satisfaction\_level -4.0006792 0.1135013 -35.248 < 2e-16  
## number\_project -0.3016658 0.0240189 -12.559 < 2e-16  
## time\_spend\_company 0.2664234 0.0183117 14.549 < 2e-16  
## Work\_accident -1.4916371 0.1056836 -14.114 < 2e-16  
## salarylow 0.2036473 0.5593742 0.364 0.716  
## salarymedium -0.3024834 0.5665484 -0.534 0.593  
## average\_montly\_hours:salaryhigh -0.0029125 0.0027480 -1.060 0.289  
## average\_montly\_hours:salarylow 0.0058556 0.0007470 7.838 4.56e-15  
## average\_montly\_hours:salarymedium 0.0056084 0.0008488 6.608 3.90e-11  
##   
## (Intercept)   
## satisfaction\_level \*\*\*  
## number\_project \*\*\*  
## time\_spend\_company \*\*\*  
## Work\_accident \*\*\*  
## salarylow   
## salarymedium   
## average\_montly\_hours:salaryhigh   
## average\_montly\_hours:salarylow \*\*\*  
## average\_montly\_hours:salarymedium \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11443.8 on 10498 degrees of freedom  
## Residual deviance: 9066.6 on 10489 degrees of freedom  
## AIC: 9086.6  
##   
## Number of Fisher Scoring iterations: 5

#Training accuracy  
lm\_train4.probs =predict (lm\_train4,type ="response")  
lm\_train4.pred=rep("0",length(train\_data$left))  
lm\_train4.pred[lm\_train4.probs>0.5]="1"  
table(lm\_train4.pred,train\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 7454 1650  
## 1 580 815

accu\_train4=(sum((lm\_train4.pred==train\_data$left)/length(train\_data$left)))\*100  
accu\_train4

## [1] 78.75988

#Testing accuracy.  
lm\_test4.prob =predict(lm\_train4,test\_data,type ="response")  
lm\_test4.pred=rep("0",length(test\_data$left))  
lm\_test4.pred[lm\_test4.prob>0.5]="1"  
table(lm\_test4.pred,test\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 3158 721  
## 1 236 385

accu\_test4=(sum((lm\_test4.pred==test\_data$left)/length(test\_data$left)))\*100  
accu\_test4

## [1] 78.73333

#  
print(paste("tarining accuracy:",accu\_train4,"% testing accuracy:",accu\_test4,"%" ))

## [1] "tarining accuracy: 78.7598818935137 % testing accuracy: 78.7333333333333 %"

#Both train and test accuracy has decreased as compared to lm\_train1.  
  
  
  
  
##########Modelling data after adding the interaction term(time\_spend\_company:salary) obtained on the basis of EDA.  
lm\_train5=glm(left~.-department-promotion\_last\_5years-average\_montly\_hours-last\_evaluation+time\_spend\_company:salary,data=train\_data,family=binomial)  
summary(lm\_train5)

##   
## Call:  
## glm(formula = left ~ . - department - promotion\_last\_5years -   
## average\_montly\_hours - last\_evaluation + time\_spend\_company:salary,   
## family = binomial, data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4644 -0.6567 -0.4173 -0.1583 2.8841   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.62662 0.34319 1.826 0.0679 .   
## satisfaction\_level -3.94113 0.11214 -35.145 < 2e-16 \*\*\*  
## number\_project -0.19746 0.02011 -9.820 < 2e-16 \*\*\*  
## time\_spend\_company -0.04715 0.08108 -0.581 0.5609   
## Work\_accident -1.54268 0.10729 -14.379 < 2e-16 \*\*\*  
## salarylow 0.06180 0.34433 0.179 0.8576   
## salarymedium 0.45913 0.34465 1.332 0.1828   
## time\_spend\_company:salarylow 0.47816 0.08567 5.581 2.39e-08 \*\*\*  
## time\_spend\_company:salarymedium 0.21800 0.08523 2.558 0.0105 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 11443.8 on 10498 degrees of freedom  
## Residual deviance: 9091.1 on 10490 degrees of freedom  
## AIC: 9109.1  
##   
## Number of Fisher Scoring iterations: 5

#Training accuracy  
lm\_train5.probs =predict (lm\_train5,type ="response")  
lm\_train5.pred=rep("0",length(train\_data$left))  
lm\_train5.pred[lm\_train5.probs>0.5]="1"  
table(lm\_train5.pred,train\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 7425 1587  
## 1 609 878

accu\_train5=(sum((lm\_train5.pred==train\_data$left)/length(train\_data$left)))\*100  
accu\_train5

## [1] 79.08372

#Testing accuracy.  
lm\_test5.prob =predict(lm\_train5,test\_data,type ="response")  
lm\_test5.pred=rep("0",length(test\_data$left))  
lm\_test5.pred[lm\_test5.prob>0.5]="1"  
table(lm\_test5.pred,test\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 3160 693  
## 1 234 413

accu\_test5=(sum((lm\_test5.pred==test\_data$left)/length(test\_data$left)))\*100  
accu\_test5

## [1] 79.4

#  
print(paste("tarining accuracy:",accu\_train5,"% testing accuracy:",accu\_test5,"%" ))

## [1] "tarining accuracy: 79.0837222592628 % testing accuracy: 79.4 %"

#Both train and test accuracy has decreased as compared to lm\_train1.  
  
  
################Modelling after adding interaction term(time\_spend\_company:satisfaction\_level:salary).  
lm\_train2=glm(left~.-department-promotion\_last\_5years-average\_montly\_hours-last\_evaluation+time\_spend\_company:satisfaction\_level:salary ,data=train\_data,family=binomial)  
summary(lm\_train2)

##   
## Call:  
## glm(formula = left ~ . - department - promotion\_last\_5years -   
## average\_montly\_hours - last\_evaluation + time\_spend\_company:satisfaction\_level:salary,   
## family = binomial, data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3195 -0.6381 -0.3060 -0.0701 3.0340   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 5.74391 0.33033  
## satisfaction\_level -14.26889 0.43854  
## number\_project -0.11216 0.02082  
## time\_spend\_company -1.22288 0.06321  
## Work\_accident -1.54283 0.11021  
## salarylow 0.95570 0.21143  
## salarymedium 1.02532 0.21353  
## satisfaction\_level:time\_spend\_company:salaryhigh 2.09556 0.11414  
## satisfaction\_level:time\_spend\_company:salarylow 2.63566 0.10329  
## satisfaction\_level:time\_spend\_company:salarymedium 2.32274 0.09872  
## z value Pr(>|z|)   
## (Intercept) 17.388 < 2e-16 \*\*\*  
## satisfaction\_level -32.537 < 2e-16 \*\*\*  
## number\_project -5.386 7.20e-08 \*\*\*  
## time\_spend\_company -19.346 < 2e-16 \*\*\*  
## Work\_accident -13.999 < 2e-16 \*\*\*  
## salarylow 4.520 6.18e-06 \*\*\*  
## salarymedium 4.802 1.57e-06 \*\*\*  
## satisfaction\_level:time\_spend\_company:salaryhigh 18.360 < 2e-16 \*\*\*  
## satisfaction\_level:time\_spend\_company:salarylow 25.518 < 2e-16 \*\*\*  
## satisfaction\_level:time\_spend\_company:salarymedium 23.528 < 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: 11443.8 on 10498 degrees of freedom  
## Residual deviance: 8310.1 on 10489 degrees of freedom  
## AIC: 8330.1  
##   
## Number of Fisher Scoring iterations: 6

#We can quite clearly see that the p value of every variable has moved more close to 0.  
  
#Training accuracy  
lm\_train2.probs =predict (lm\_train2,type ="response")  
lm\_train2.pred=rep("0",length(train\_data$left))  
lm\_train2.pred[lm\_train2.probs>0.5]="1"  
table(lm\_train2.pred,train\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 7424 1186  
## 1 610 1279

accu\_train2=(sum((lm\_train2.pred==train\_data$left)/length(train\_data$left)))\*100  
accu\_train2

## [1] 82.89361

#Testing accuracy.  
lm\_test2.prob =predict(lm\_train2,test\_data,type ="response")  
lm\_test2.pred=rep("0",length(test\_data$left))  
lm\_test2.pred[lm\_test2.prob>0.5]="1"  
table(lm\_test2.pred,test\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 3148 521  
## 1 246 585

accu\_test2=(sum((lm\_test2.pred==test\_data$left)/length(test\_data$left)))\*100  
accu\_test2

## [1] 82.95556

#  
print(paste("tarining accuracy:",accu\_train2,"% testing accuracy:",accu\_test2,"%" ))

## [1] "tarining accuracy: 82.8936089151348 % testing accuracy: 82.9555555555556 %"

#We have quite clearly seen the improvement in train and test accuracy as compared to lm\_train5.  
  
  
#############Modelling after adding interaction term(time\_spend\_company:satisfaction\_level:salary & #Modelling after adding interaction term(time\_spend\_company:satisfaction\_level:salary).  
   
  
lm\_train6=glm(left~.-department-average\_montly\_hours-last\_evaluation+time\_spend\_company:promotion\_last\_5years+time\_spend\_company:satisfaction\_level:salary ,data=train\_data,family=binomial)  
summary(lm\_train6)

##   
## Call:  
## glm(formula = left ~ . - department - average\_montly\_hours -   
## last\_evaluation + time\_spend\_company:promotion\_last\_5years +   
## time\_spend\_company:satisfaction\_level:salary, family = binomial,   
## data = train\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3296 -0.6364 -0.3029 -0.0682 3.2896   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 5.70282 0.33209  
## satisfaction\_level -14.23668 0.44088  
## number\_project -0.11539 0.02087  
## time\_spend\_company -1.21003 0.06367  
## Work\_accident -1.53351 0.11046  
## promotion\_last\_5years 0.77969 0.91865  
## salarylow 0.96942 0.21204  
## salarymedium 1.04241 0.21419  
## time\_spend\_company:promotion\_last\_5years -0.51619 0.25025  
## satisfaction\_level:time\_spend\_company:salaryhigh 2.10837 0.11577  
## satisfaction\_level:time\_spend\_company:salarylow 2.62670 0.10387  
## satisfaction\_level:time\_spend\_company:salarymedium 2.31743 0.09958  
## z value Pr(>|z|)   
## (Intercept) 17.172 < 2e-16 \*\*\*  
## satisfaction\_level -32.292 < 2e-16 \*\*\*  
## number\_project -5.529 3.22e-08 \*\*\*  
## time\_spend\_company -19.005 < 2e-16 \*\*\*  
## Work\_accident -13.883 < 2e-16 \*\*\*  
## promotion\_last\_5years 0.849 0.3960   
## salarylow 4.572 4.83e-06 \*\*\*  
## salarymedium 4.867 1.13e-06 \*\*\*  
## time\_spend\_company:promotion\_last\_5years -2.063 0.0391 \*   
## satisfaction\_level:time\_spend\_company:salaryhigh 18.211 < 2e-16 \*\*\*  
## satisfaction\_level:time\_spend\_company:salarylow 25.287 < 2e-16 \*\*\*  
## satisfaction\_level:time\_spend\_company:salarymedium 23.273 < 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: 11443.8 on 10498 degrees of freedom  
## Residual deviance: 8285.1 on 10487 degrees of freedom  
## AIC: 8309.1  
##   
## Number of Fisher Scoring iterations: 6

#We can quite clearly see that the p value of every variable has moved more close to 0.  
  
#Training accuracy  
lm\_train6.probs =predict (lm\_train6,type ="response")  
lm\_train6.pred=rep("0",length(train\_data$left))  
lm\_train6.pred[lm\_train6.probs>0.5]="1"  
table(lm\_train6.pred,train\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 7422 1185  
## 1 612 1280

accu\_train6=(sum((lm\_train6.pred==train\_data$left)/length(train\_data$left)))\*100  
accu\_train6

## [1] 82.88408

#Testing accuracy.  
lm\_test6.prob =predict(lm\_train6,test\_data,type ="response")  
lm\_test6.pred=rep("0",length(test\_data$left))  
lm\_test6.pred[lm\_test6.prob>0.5]="1"  
table(lm\_test6.pred,test\_data$left,dnn = c('predicted','true'))

## true  
## predicted 0 1  
## 0 3143 521  
## 1 251 585

accu\_test6=(sum((lm\_test6.pred==test\_data$left)/length(test\_data$left)))\*100  
accu\_test6

## [1] 82.84444

#  
print(paste("tarining accuracy:",accu\_train6,"% testing accuracy:",accu\_test6,"%" ))

## [1] "tarining accuracy: 82.8840841984951 % testing accuracy: 82.8444444444444 %"

#Test and train accuracy both have increased as compared to lm\_train2.  
  
###########Checking corelation between used variables.  
cor(train\_data[,c("satisfaction\_level","number\_project","time\_spend\_company","Work\_accident")])

## satisfaction\_level number\_project time\_spend\_company  
## satisfaction\_level 1.00000000 -0.13439233 -0.09735549  
## number\_project -0.13439233 1.00000000 0.19806384  
## time\_spend\_company -0.09735549 0.19806384 1.00000000  
## Work\_accident 0.05193200 -0.00794519 0.00852857  
## Work\_accident  
## satisfaction\_level 0.05193200  
## number\_project -0.00794519  
## time\_spend\_company 0.00852857  
## Work\_accident 1.00000000

#There is no significant amount of corelation among checked vairables.