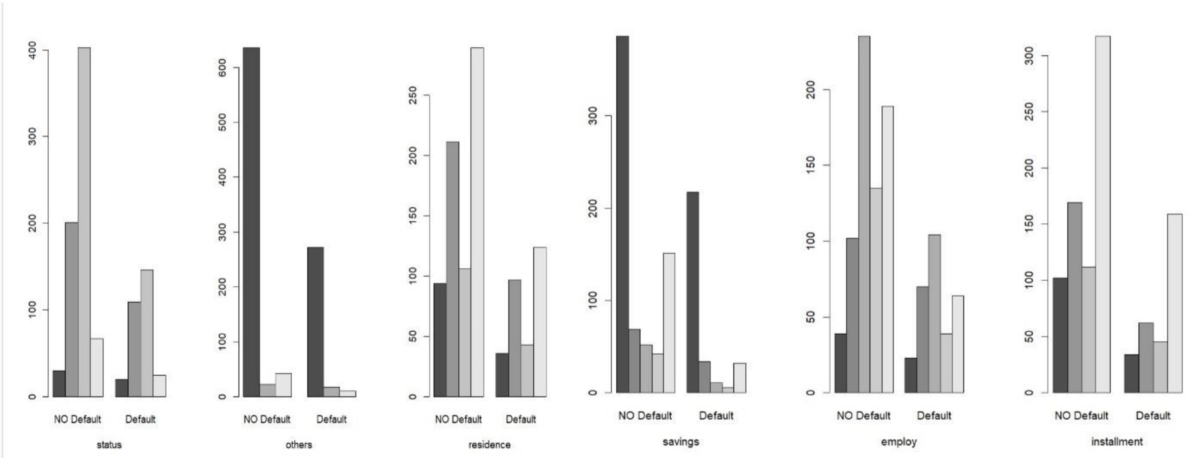
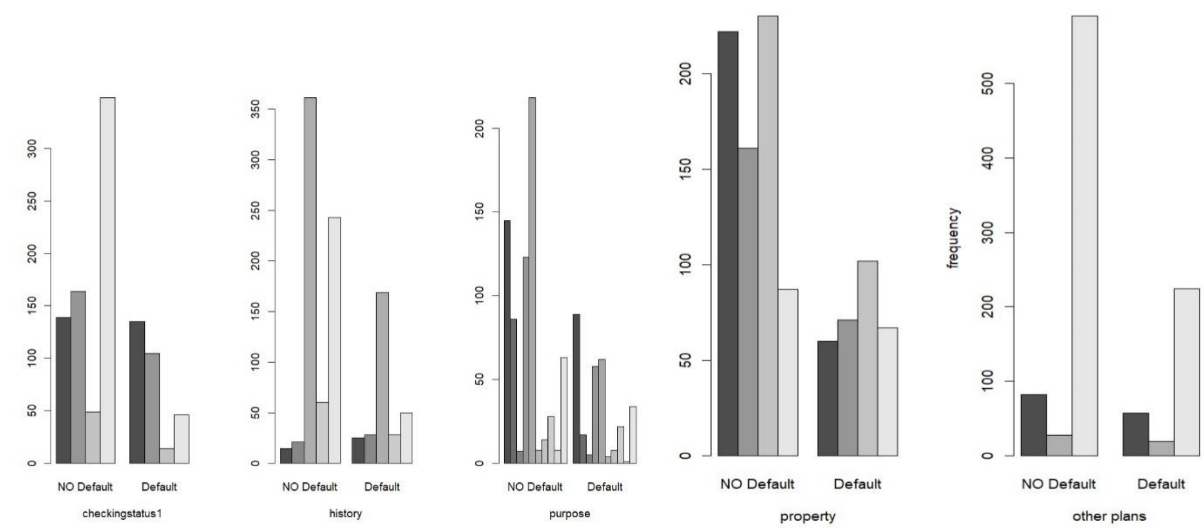
Ashish Mani Acharya MiniProject3 Stat 6340

Section1

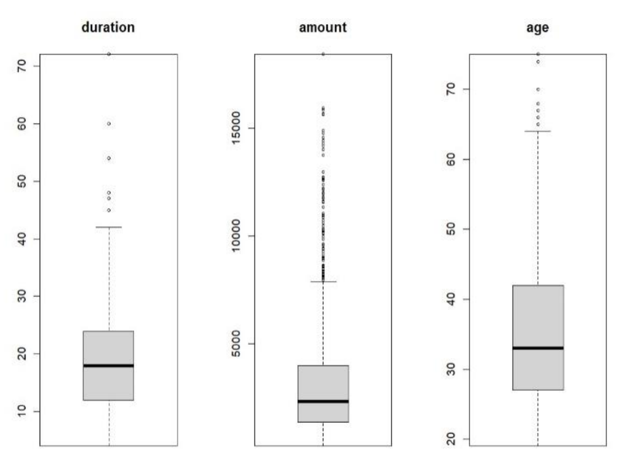
Q1

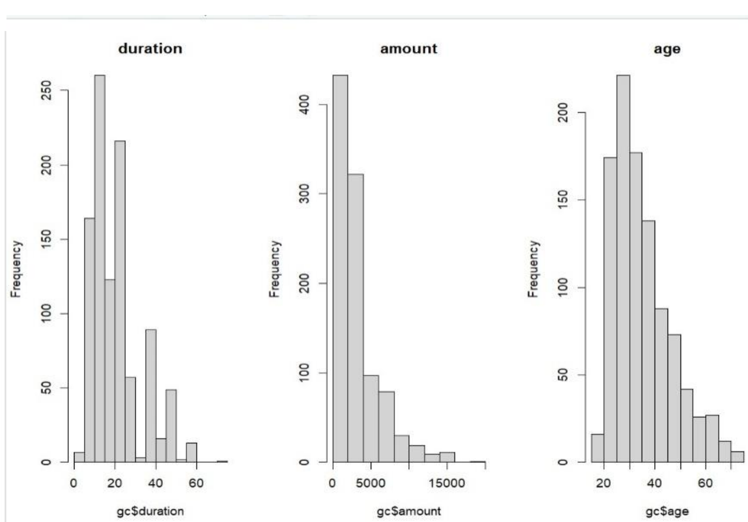
Q1A) We have 21 variables and 1000 observations. The Default here is used as the response variable. We have 3 quantitative variables while rest are qualitative including few of those ones that seem numeric but are not continuous and hence are treated as qualitative. Such variables are installment,residence,cards and liable. Here I have used Barplot for qualtitative and Boxplot and Histogram for quantitative variables.





QUANTITATIVE DATA





Ashish Mani Acharya MiniProject3 Stat 6340

We can see that the both quantitative and qualitative data are hard to analyze by plots as they have different scale and distributions. However, under surficial observation we may find that the quantitative data seem to be left skewed .

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          8.683e-01  1.012e+00   0.858 0.391077
checkingstatus1A12  -3.834e-01  2.194e-01  -1.748 0.080499 .
checkingstatus1A13  -9.739e-01  3.717e-01  -2.620 0.008794 **
checkingstatus1A14  -1.780e+00  2.358e-01  -7.547 4.45e-14 ***
duration             2.801e-02  9.448e-03   2.965 0.003028 **
historyA31           1.690e-01  5.614e-01   0.301 0.763407
historyA32          -5.672e-01  4.434e-01  -1.279 0.200816
historyA33          -9.496e-01  4.780e-01  -1.987 0.046964 *
historyA34          -1.496e+00  4.452e-01  -3.360 0.000780 ***
purposeA41          -1.660e+00  3.792e-01  -4.379 1.19e-05 ***
purposeA410         -1.485e+00  7.871e-01  -1.887 0.059123 .
purposeA42          -7.481e-01  2.635e-01  -2.839 0.004519 **
purposeA43          -8.743e-01  2.498e-01  -3.501 0.000464 ***
purposeA44          -5.109e-01  7.745e-01  -0.660 0.509496
purposeA45          -1.603e-01  5.528e-01  -0.290 0.771819
purposeA46           1.130e-01  3.976e-01   0.284 0.776171
purposeA48          -1.931e+00  1.180e+00  -1.637 0.101692
purposeA49          -6.888e-01  3.377e-01  -2.040 0.041386 *
amount               1.233e-04  4.502e-05   2.740 0.006153 **
savingsA62          -3.638e-01  2.916e-01  -1.248 0.212176
savingsA63          -3.664e-01  4.021e-01  -0.911 0.362218
savingsA64          -1.460e+00  5.422e-01  -2.693 0.007075 **
savingsA65          -9.732e-01  2.661e-01  -3.657 0.000255 ***
employA72            6.662e-02  4.396e-01   0.152 0.879539
employA73           -2.293e-01  4.212e-01  -0.544 0.586261
employA74           -7.634e-01  4.596e-01  -1.661 0.096714 .
employA75           -2.213e-01  4.236e-01  -0.523 0.601303
installment2         2.641e-01  3.094e-01   0.854 0.393281
installment3         6.260e-01  3.414e-01   1.834 0.066671 .
installment4         9.369e-01  3.047e-01   3.075 0.002106 **
statusA92           -2.616e-01  3.885e-01  -0.673 0.500728
statusA93           -8.427e-01  3.813e-01  -2.210 0.027106 *
statusA94           -3.764e-01  4.576e-01  -0.823 0.410748
othersA102           4.329e-01  4.127e-01   1.049 0.294219
othersA103          -9.828e-01  4.264e-01  -2.305 0.021160 *
residence2           7.613e-01  2.994e-01   2.543 0.010985 *
residence3           5.246e-01  3.359e-01   1.562 0.118342
residence4           3.885e-01  3.029e-01   1.282 0.199687
propertyA122         2.698e-01  2.551e-01   1.058 0.290201
propertyA123         1.607e-01  2.387e-01   0.673 0.500881
propertyA124         7.367e-01  4.294e-01   1.716 0.086230 .
age                 -1.279e-02  9.317e-03  -1.373 0.169876
otherplansA142      -8.884e-02  4.166e-01  -0.213 0.831134
otherplansA143      -6.475e-01  2.403e-01  -2.694 0.007056 **
housingA152         -4.573e-01  2.364e-01  -1.934 0.053071 .
housingA153         -6.303e-01  4.854e-01  -1.299 0.194111
cards2               4.050e-01  2.456e-01   1.649 0.099170 .
cards3               2.741e-01  6.087e-01   0.450 0.652475
cards4               4.550e-01  1.072e+00   0.424 0.671370
jobA172              4.416e-01  6.867e-01   0.643 0.520167
jobA173              4.694e-01  6.625e-01   0.709 0.478594
jobA174              3.691e-01  6.708e-01   0.550 0.582115
liable2              2.628e-01  2.518e-01   1.044 0.296625
teleA192            -2.848e-01  2.031e-01  -1.402 0.160870
foreignA202         -1.461e+00  6.265e-01  -2.333 0.019658 *
---
```

Q1b) We know that under likelihood ratio test, H0: You should use the nested model. Ha: You should use full model.
Hence I chose the value based on p value obtained through p value obtained likelihood ratio (ChiSq)tests. Hence I begin with dropping predictors based on their significance which is determined by their p values as shown in this adjoining diagram where * stands for significant variables. After dropping insignificant variables I did Chisq test and observed that reduced 12 Var model is as good as full model. Then I started dropping Var one by one based on ChiSq Test. When Model reached 11 Var then all Var seemed significant. Hence, I used that model as the final model.

The final model is based on selection of var and their coeffcients through chisq test and p values obtained here

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.218e+00  6.349e-01   1.918 0.055083 .
checkingstatus1A12 -4.102e-01  2.098e-01  -1.956 0.050517 .
checkingstatus1A13 -1.062e+00  3.588e-01  -2.961 0.003069 **
checkingstatus1A14 -1.776e+00  2.264e-01  -7.845 4.32e-15 ***
duration             2.822e-02  8.848e-03   3.190 0.001424 **
historyA31          -1.400e-01  5.205e-01  -0.269 0.787970
historyA32          -8.610e-01  4.050e-01  -2.126 0.033499 *
historyA33          -9.826e-01  4.610e-01  -2.132 0.033041 *
historyA34          -1.588e+00  4.270e-01  -3.718 0.000201 ***
purposeA41          -1.558e+00  3.613e-01  -4.312 1.62e-05 ***
purposeA410         -1.562e+00  7.559e-01  -2.066 0.038855 *
purposeA42          -6.561e-01  2.491e-01  -2.635 0.008424 **
purposeA43          -8.946e-01  2.395e-01  -3.735 0.000188 ***
purposeA44          -5.656e-01  7.434e-01  -0.761 0.446742
purposeA45          -1.740e-01  5.392e-01  -0.323 0.747009
purposeA46           1.853e-01  3.852e-01   0.481 0.630566
purposeA48          -2.123e+00  1.226e+00  -1.732 0.083293 .
purposeA49          -8.041e-01  3.229e-01  -2.490 0.012763 *
amount               1.115e-04  4.095e-05   2.722 0.006482 **
savingsA62          -2.679e-01  2.739e-01  -0.978 0.328005
savingsA63          -4.355e-01  3.934e-01  -1.107 0.268263
savingsA64          -1.338e+00  5.048e-01  -2.651 0.008018 **
savingsA65          -9.707e-01  2.554e-01  -3.800 0.000145 ***
installment2         2.005e-01  2.995e-01   0.669 0.503352
installment3         5.738e-01  3.297e-01   1.740 0.081837 .
installment4         8.609e-01  2.888e-01   2.980 0.002878 **
statusA92           -1.280e-01  3.659e-01  -0.350 0.726463
statusA93           -7.686e-01  3.599e-01  -2.136 0.032700 *
statusA94           -2.798e-01  4.345e-01  -0.644 0.519525
othersA102           5.378e-01  4.007e-01   1.342 0.179578
othersA103          -1.022e+00  4.123e-01  -2.479 0.013181 *
otherplansA142      -1.397e-01  4.016e-01  -0.348 0.727974
otherplansA143      -6.526e-01  2.333e-01  -2.797 0.005155 **
foreignA202         -1.306e+00  6.252e-01  -2.089 0.036676 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ashish Mani Acharya MiniProject3 Stat 6340

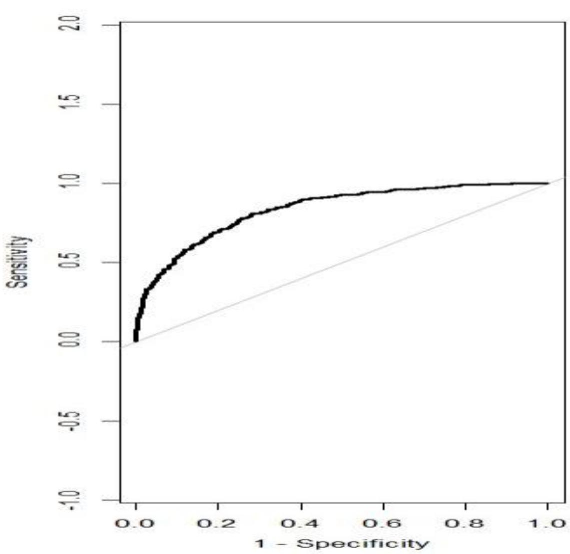
Q1C)bsed on above calculations the final model would be $\mathrm{p(x) = (1 + exp(-f_x))^{-1}}$,

where $\mathrm{f_x = 0.865 - 0.411 * I(checkingstatus1 = A12) - 1.069 * I(checkingstatus1 = A13) - 1.779 * I(checkingstatus1 = A14) + 0.02805 * duration - 0.1361 * I(history = A31) - 0.8588 * I(history = A32) - 0.98 * I(history = A33) - 1.587 * I(history = A34) - 1.556 * I(purpose = A41) - 1.575 * I(purpose = A410) - 0.6612 * I(purpose = A42) - 0.8968 * I(purpose = A43) - 0.5635 * I(purpose = A44) - 0.1782 * I(purpose = A45) + 0.18 * I(purpose = A46) - 2.133 * I(purpose = A48) - 0.8093 * I(purpose = A49) + 0.0001116 * amount - 0.2671 * I(savings = A62) - 0.4271 * I(savings = A63) - 1.331 * I(savings = A64) - 0.9677 * I(savings = A65) + 0.3038 * installment - 0.1243 * I(status = A92) - 0.7705 * I(status = A93) - 0.2786 * I(status = A94) + 0.5323 * I(others = A102) - 1.022 * I(others = A103) - 0.1417 * I(otherplans = A142) - 0.6536 * I(otherplans = A143) - 1.29 * I(foreign = A202)}$

We can see that the misclassification rate for the model is 21.9%

I have taken first two variable's coefficients; checkingstatus1 and duration for explanations. If the checkingstatus1 predictor value is A13, borrower is $1 - \exp(-1.069) = 34.33\%$ LESS LIKELY to default than the borrower with checkingstatus1 predictor value is A12. Similarly per unit increase in duration will make someone exp(.02805) -1=.028 or2.8 % more likely to default

Q2

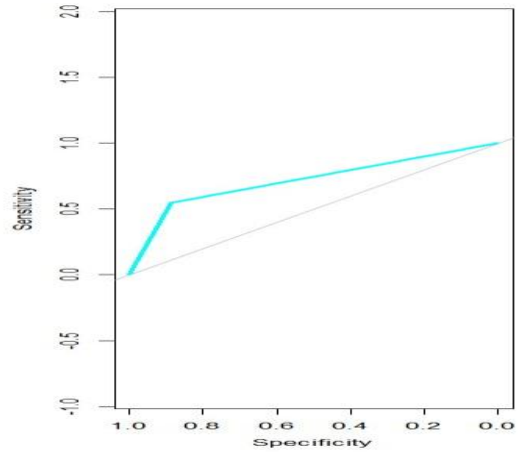


Q2A) if we include all predictors then logistic regression would have training error rate of 21.4 % while its sensitivity ,specificity and AUC(Area under curve under ROC) is 53.33%,89.43 % and 83.38% respectively.

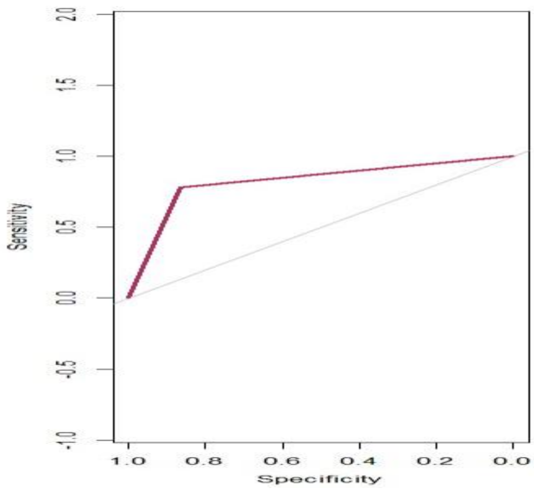
Q2B) LOOCV through manual calculations is 24.9%.

Q2C) We obtained the same test error as manual calculations as LOOCV through caret package and mentioning the cost function is 24.9 % . Hence, the results match as asked by the question.

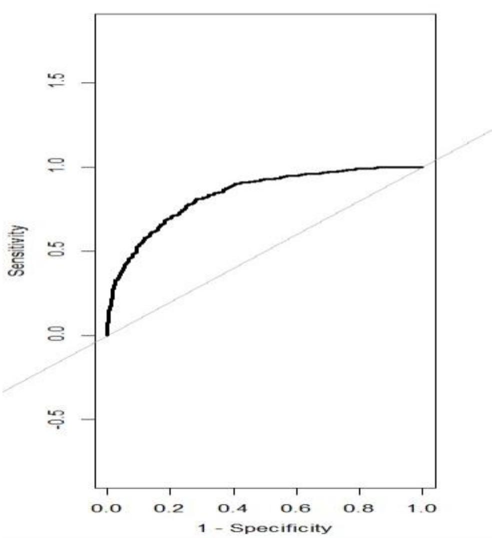
Ashish Mani Acharya MiniProject3 Stat 6340



Q2D) LDA has training error rate of 21.7 % while its sensitivity ,specificity and AUC(Area under curve under ROC) is 54% ,87.85%,and 70.93% respectively.



Q2E) QDA has training error rate of 17.7% while its sensitivity ,specificity and AUC(Area under curve under ROC) is 76.67%, 84.71%, and 82.29% respectively.



Q2G) logistic regression model in Q1 has training error rate of 21.6% while its sensitivity ,specificity and AUC(Area under curve under ROC) is 53.33%,89.14% % and 82.39% respectively.

Q2H)Based on LOOCV and AUC , QDA seems best. Here is overall comparison of classifiers through Confusion matrix.ROC curves have been attached with the question itself.

Ashish Mani Acharya MiniProject3 Stat 6340

	0	1
0	619	136
1	81	164

LDA

	0	1
0	606	66
1	94	234

QDA

	0	1
0	624	135
1	76	165

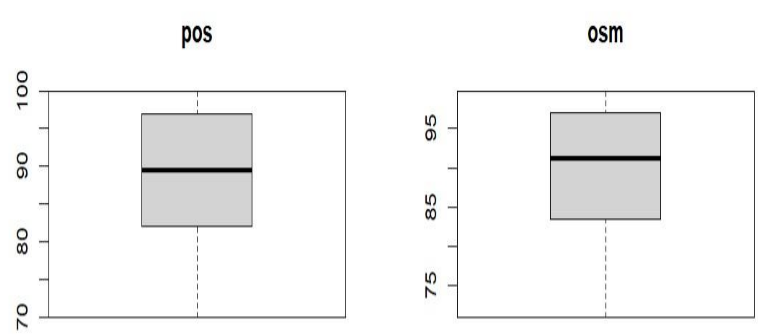
LR full Model

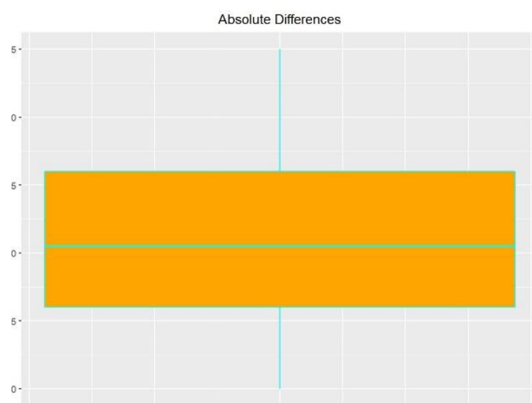
	0	1
0	624	135
1	76	165

LR reduced model

Q3

Q3A) Since we are here to estimate the differences I think the boxplot of differences is appropriate here.





Q3B) Provide a point estimate ˆθ of θ, appropriate estimates of bias and standard error of the estimate, and a 95% confidence interval for θ. Interpret the results.

here natural estimate would the 90th quantile of the absolute differences is , ˆθ(theta hat) = 2.

Similarly bias = -1.5775, Standard Error = 1.210859, 95% CI lower estimate=-1.5675 and 95% CI upper estimate=- 2.2225

Q3C) Using our own bootstrap method(nb=1000) we got bias estimate of 0.00246999999999975 (since this is negligible we can say bootstrap estimator to be an unbiased estimator) whereas SE=.1311613. 95% upper confidence bound for θ is 2.2 % which means we are 95% certain that the TDI of the two methods is below 2.2.

Q3D) the boot package got us similar results as in 3.c getting a bias of 0.00243 and standard error of 0.1281914. The 95% upper confidence bound for θ =2.2 similar to 3c.

Q3E) we can conclude that about 90% of the measurements of these two methods will agree within a difference in percent saturation of hemoglobin with oxygen of at most 2.2% with a 5% uncertainty, thus these methods are interchangeable.

Ashish Mani Acharya MiniProject3 Stat 6340

Section 2

```
library(MASS)
library(boot)
library(caret)
library(ISLR)
library(pROC)

gc <- read.csv("C:/Users/alexk/OneDrive/Desktop/stat 6340/miniproject 3/germangc.csv", header=TRUE)

#for ease of typing german gc has been shortened to gc#

#exploratory analysis of the data#
View(gc)
dim(gc)
head(gc)
#########################Q1####################################
# factoring Categorical Variables #
gc$Default = as.factor(gc$Default)
gc$checkingstatus1 = as.factor(gc$checkingstatus1)
gc$history = as.factor(gc$history)
gc$purpose = as.factor(gc$purpose)
gc$savings = as.factor(gc$savings)
gc$employ = as.factor(gc$employ)
gc$status = as.factor(gc$status)
gc$others = as.factor(gc$others)
gc$property = as.factor(gc$property)
gc$otherplans = as.factor(gc$otherplans)
gc$housing = as.factor(gc$housing)
gc$job = as.factor(gc$job)
gc$tele = as.factor(gc$tele)
```

Ashish Mani Acharya MiniProject3 Stat 6340

```
gc$foreign = as.factor(gc$foreign)

gc$liable = as.factor(gc$liable)



#boxplots#

par(mfrow = c(1, 3))

boxplot(gc$duration,main="duration" )

boxplot(gc$amount,main="amount")

boxplot(gc$age,main="age")



#histogram#

hist(gc$duration,main="duration" )

hist(gc$amount,main="amount")

hist(gc$age,main="age")



#making barplots#

table(gc$checkingstatus1)

t1 = xtabs(~checkingstatus1+Default, data = gc)

barplot(t1, xlab="checkingstatus1",names.arg = c("NO Default", "Default"), beside = TRUE)


table(gc$history)

t1 = xtabs(~history+Default, data = gc)

barplot(t1, xlab = "history",

        names.arg = c("NO Default", "Default"), beside = TRUE)

table(gc$purpose)

t1 = xtabs(~purpose+Default, data = gc)

barplot(t1, xlab = "purpose",

        names.arg = c("NO Default", "Default"), beside = TRUE)

table(gc$savings)

t1 = xtabs(~savings+Default, data = gc)

barplot(t1, xlab = "savings",

        names.arg = c("NO Default", "Default"), beside = TRUE)

table(gc$employ)
```

Ashish Mani Acharya MiniProject3 Stat 6340

```
t1 = xtabs(~employ+Default, data = gc)

barplot(t1, xlab = "employ",

        names.arg = c("NO Default", "Default"), beside = TRUE)

table(gc$installment)

t1 = xtabs(~installment+Default, data = gc)

barplot(t1, xlab = "installment",

        names.arg = c("NO Default", "Default"), beside = TRUE)

table(gc$status)

t1 = xtabs(~status+Default, data = gc)

barplot(t1, xlab = "status",

        names.arg = c("NO Default", "Default"), beside = TRUE)

table(gc$others)

t1 = xtabs(~others+Default, data = gc)

barplot(t1, xlab = "others",

        names.arg = c("NO Default", "Default"), beside = TRUE)

table(gc$residence)

t1 = xtabs(~residence+Default, data = gc)

barplot(t1, xlab = "residence",

        names.arg = c("NO Default", "Default"), beside = TRUE)

table(gc$property)

t1 = xtabs(~property+Default, data = gc)

barplot(t1, xlab = "property",

        names.arg = c("NO Default", "Default"), beside = TRUE)


table(gc$otherplans)

t1 = xtabs(~otherplans+Default, data = gc)

barplot(t1, xlab = "other plans", ylab = "frequency",

        names.arg = c("NO Default", "Default"), beside = TRUE)


#Q 1 B#

fit.b1 = glm(Default ~ ., family = binomial, data =  gc) #using all the predictors#

summary(fit.b1)

 dropped.variables = c(12,14,17,18,19,20);  gc2 =  gc[, - dropped.variables]#dropped on the basis of significance based
on  p value of coefficients#

fit.b2 = glm(Default ~ ., family = binomial, data =  gc2)
```

Ashish Mani Acharya MiniProject3 Stat 6340

```
summary(fit.b2)

 dropped.variables = c(8,12);  gc3 =  gc2[, - dropped.variables]#dropped on p values#

fit.b3 = glm(Default ~ ., family = binomial, data =  gc3)

summary(fit.b3)

anova(fit.b3, fit.b1, test = "Chisq")

#since P value is .07264 we can use this reduced model in other words reduced model is as good as full model#


fit.b3.1 = glm(Default ~ ., family = binomial, data = gc3[, -2]); anova(fit.b3.1, fit.b3,test = "Chisq")

fit.b3.2 = glm(Default ~ ., family = binomial, data = gc3[, -3]); anova(fit.b3.2, fit.b3,test = "Chisq")

fit.b3.3 = glm(Default ~ ., family = binomial, data = gc3[, -4]); anova(fit.b3.3, fit.b3,test = "Chisq")

fit.b3.4 = glm(Default ~ ., family = binomial, data = gc3[, -5]); anova(fit.b3.4, fit.b3,test = "Chisq")

fit.b3.5 = glm(Default ~ ., family = binomial, data = gc3[, -6]); anova(fit.b3.5, fit.b3,test = "Chisq")

fit.b3.6 = glm(Default ~ ., family = binomial, data = gc3[, -7]); anova(fit.b3.6, fit.b3,test = "Chisq")

fit.b3.7 = glm(Default ~ ., family = binomial, data = gc3[, -8]); anova(fit.b3.7, fit.b3,test = "Chisq")

fit.b3.8 = glm(Default ~ ., family = binomial, data = gc3[, -9]); anova(fit.b3.8, fit.b3,test = "Chisq")

fit.b3.9 = glm(Default ~ ., family = binomial, data = gc3[, -10]); anova(fit.b3.9, fit.b3,test = "Chisq")

fit.b3.10 = glm(Default ~ ., family = binomial, data = gc3[, -11]); anova(fit.b3.10, fit.b3,test = "Chisq")

fit.b3.11 = glm(Default ~ ., family = binomial, data = gc3[, -12]); anova(fit.b3.11, fit.b3,test = "Chisq")

fit.b3.12 = glm(Default ~ ., family = binomial, data = gc3[, -13]); anova(fit.b3.12, fit.b3,test = "Chisq")


gc4 = gc3[, -12] # col 12 or residence dropped based on p value #


fit.1.1 = glm(Default ~ ., family = binomial, data = gc4[, -2]); anova(fit.1.1, fit.b3.11,test = "Chisq")

fit.1.2 = glm(Default ~ ., family = binomial, data = gc4[, -3]); anova(fit.1.2, fit.b3.11,test = "Chisq")

fit.1.3 = glm(Default ~ ., family = binomial, data = gc4[, -4]); anova(fit.1.3, fit.b3.11,test = "Chisq")

fit.1.4 = glm(Default ~ ., family = binomial, data = gc4[, -5]); anova(fit.1.4, fit.b3.11,test = "Chisq")

fit.1.5 = glm(Default ~ ., family = binomial, data = gc4[, -6]); anova(fit.1.5, fit.b3.11,test = "Chisq")

fit.1.6 = glm(Default ~ ., family = binomial, data = gc4[, -7]); anova(fit.1.6, fit.b3.11,test = "Chisq")

fit.1.7 = glm(Default ~ ., family = binomial, data = gc4[, -8]); anova(fit.1.7, fit.b3.11,test = "Chisq")

fit.1.8 = glm(Default ~ ., family = binomial, data = gc4[, -9]); anova(fit.1.8, fit.b3.11,test = "Chisq")

fit.1.9 = glm(Default ~ ., family = binomial, data = gc4[, -10]); anova(fit.1.9, fit.b3.11,test = "Chisq")

fit.1.10 = glm(Default ~ ., family = binomial, data = gc4[, -11]); anova(fit.1.10, fit.b3.11,test = "Chisq")

fit.1.11 = glm(Default ~ ., family = binomial, data = gc4[, -12]); anova(fit.1.11, fit.b3.11,test = "Chisq")


# as per p values all of them are less than .005 hence all are needed for this model #
```

Ashish Mani Acharya MiniProject3 Stat 6340

```
#hence 11 var model is the best model#

summary(fit.b3.11)
#Q1C#
prediction = predict(fit.b3.11, data = gc, type = "response");
model.prediction = ifelse(prediction>= 0.5, 1, 0)
mean(model.prediction != gc$Default) # misclassification for model is 21.9#
################Q2##############################################################
#Q2A#
pred.fit.b1 = predict(fit.b1, data = gc, type = "response"); # 20 var model
pred.this.model = ifelse(pred.fit.b1 >= 0.5, 1, 0)
mean(pred.this.model != gc$Default)#finding miscalculation rate .21#
p = table(pred.this.model, gc$Default)
p # confusion matrix
p[2,2]/(p[1,2]+p[2,2]) # finding sensitivity 0.55#
p[1,1]/(p[1,1]+p[2,1]) # finding specificity .8912#

#for reduced model#
pred.fit.b3.11 = predict(fit.b3.11, data = gc, type = "response"); # 11 var model
pred.this.model = ifelse(pred.fit.b3.11 >= 0.5, 1, 0)
mean(pred.this.model != gc$Default)#finding miscalculation rate .21#
p2 = table(pred.this.model, gc$Default)
p2 #confusion matrix
library(pROC)
ROC.fit.b3.11 = roc(response = gc$Default, pred.fit.b3.11, levels = c("0", "1"))
plot(ROC.fit.b3.11, legacy.axes = T)

#Q2B#
pred.2b <- sapply(1:nrow(gc), FUN = function(i){
  fit <- glm(gc$Default ~ ., family = binomial, data = gc[-i,])
  prob <- predict(fit, gc[i,], type = "response")
  pred <- ifelse(prob >= 0.5, 1, 0)
  return(pred)})
loocverror=1 - mean(pred.2b == gc$Default)
loocverror #.249
```

```
accuracy=1-loocverror

accuracy #.751#


#Q2C#

library(boot)


cost <- function(r, pi = 0) mean(abs(r-pi) > 0.5)#defining a Cost function#

 loocv.error.bootpackage = cv.glm(gc,fit.b1,cost=cost)$delta[1]

 loocv.error.bootpackage #.079#

 loocv.accuracy.bootpackage=1-loocv.error.bootpackage

 loocv.accuracy.bootpackage #.249#


 #Q2D#

 model_final.lda<-lda(Default ~., data = gc)

 lda.pred.g <- predict(model_final.lda, gc)



 misclassification=mean(lda.pred.g$class != gc$Default)

 misclassification #0.217 misclassification rate#


 #Confusion matrix

 lda.cf = table(lda.pred.g$class, gc$Default)

 lda.cf

 sensitivity(lda.cf)#0.8785714

 specificity(lda.cf)#0.54



 ##ROC Curve#

 roc.lda<-roc(gc$Default,as.numeric(lda.pred.g$class), direction=("<"), levels=c(0,1))



 plot(roc.lda, col="CYAN")
```

```
 auc.lda= auc(gc$Default, as.numeric(lda.pred.g$class))
 auc.lda #Area under the curve: 0.7093#

#LDA with LOOCV method #
 train.control <- trainControl(method = "LOOCV")

 lda.model <- train(Default~., data = gc, method = 'lda',
                    trControl = train.control)

 plot(roc.lda, legacy.axes = T)
#Q2E#

 model_final.qda<-qda(Default ~., data = gc)
 qda.pred.gc <- predict(model_final.qda, gc)


 mean(qda.pred.gc$class != gc$Default) #Overall MisClassifications rate is .177#

 #Confusion matrix
 qda.cf = table(qda.pred.gc$class, gc$Default)
 qda.cf
 sensitivity(qda.cf)#0.0.8471429
 specificity(qda.cf)# 0.7666667


 ##ROC Curve#
 roc.qda<-roc(gc$Default,as.numeric(qda.pred.gc$class), direction=("<"), levels=c(0,1))

 roc.qda#area under roc curve or AUC is 82.29%

 plot(roc.qda, col="Maroon")
```

```
 auc.qda= auc(gc$Default, as.numeric(lda.pred.g$class))

 auc.qda #Area under the curve: 0.8069#


 #QDA with LOOCV method #

 train.control <- trainControl(method = "LOOCV")


 qda.model <- train(Default~., data = gc, method = 'qda',

                  trControl = train.control)

#Q2f#

#KNN#



 library(caret)



 # Define training control

 train.control <- trainControl(method = "LOOCV",classProbs = TRUE,

                             summaryFunction = twoClassSummary)

 # Train the model

 model.knn <- train(as.factor(Default ~.,) data = gc, method = "knn",

                  tuneGrid = expand.grid(k=seq(1, 100, by = 1)),

                  preProcess = c("center", "scale"),

                  trControl = train.control)

 # Summarize the results

 model.knn1 <- train(Default ~., data = gc, method = "knn",

                   tuneGrid = expand.grid(k=24),

                   preProcess = c("center", "scale"),

                   trControl = train.control)



 print(model.knn)

 plot(model)


################################Q3####################################

library(ggplot2)
```

Ashish Mani Acharya MiniProject3 Stat 6340

```
library(plotly)

library(ggpubr)

library(MASS)

library(caret)

library(boot)

library(e1071)


dim(oxygen_saturation)

str(oxygen_saturation)

head(oxygen_saturation)

oxygen_saturation = oxygen.saturation <- read.delim("C:/Users/alexk/OneDrive/Desktop/stat 6340/miniproject 3/oxygen saturation.txt")

 View(oxygen.saturation)



D=oxygen_saturation [,1]- oxygen_saturation [,2]

abs_D=abs(D)

abs_D


gbox=ggplot ()+ geom_boxplot(fill='orange', color="cyan",aes(y=abs_D))+

  ggtitle("Absolute Differences")+ theme(plot.title=element_text(hjust =0.5))+

  theme(axis.title.y=element_blank (),axis.title.x=element_blank (),

        axis.text.x=element_blank (),

        axis.ticks.x=element_blank ())

print(gbox)


####Question 3.b####

#Natural estimate =  90% quantile

theta_hat=quantile(abs_D,0.9)[[1]]

print(theta_hat)

theta.hat = quantile(abs_D, 0.9) # point estimate 90th quantile=1.99

mean(oxygen_saturation$abs_D)-theta.hat # bias=-1.5775

sd(oxygen_saturation$abs_D) # Standard Error 1.210859

quantile(abs_D, 0.025) # 95% CI lower estimate=-1.5675

quantile(abs_D, 0.975)#2.2225
```

Ashish Mani Acharya MiniProject3 Stat 6340

```
#Q3C#

#Number of bootstrap samples is 1000#

nb =1000

#Generating Bootstrap Samples#



set.seed(1)

x <- .Random.seed

rnorm(5)

boot_sample= replicate (nb , sample(abs_D, replace=TRUE),simplify = FALSE)

#Finding Bootstrap Estimate


boot_estimates =sapply(boot_sample , function(x){quantile(x ,0.9)[[1]]})

#bias estimate

print(paste("Bias estimate:",mean(boot_estimates )-theta_hat))

#std error

print(paste("Std error estimate:",sd(boot_estimates )))


print(paste("95% upper bound:",sort(boot_estimates )[ceiling(.95*nb )]))


#Q3D#

quantile.fn=function(x,indices)

{

  quantile(x[indices ] ,0.9)[[1]]

}

#Generating bootstrap estimates

set.seed(5)

theta.boot=boot(abs_D,quantile.fn ,nb)

print(theta.boot)


print(paste("95% upper bound:",sort(theta.boot$t)[ceiling(.95*nb )]))
```

Ashish Mani Acharya MiniProject3 Stat 6340