Assignment 3 Chunlei Zhou

Q1

(a)

OR(abuse==1|mothalc==1)=$\frac{odds(abuse == 1|mothalc == 1)}{odds(abuse == 1|mothalc == 0)}$

$$= \frac{p(abuse == 1|mothalc == 1)/(1 - p(abuse == 1|mothalc == 1)}{p(abuse == 1|mothalc == 0)/(1 - p(abuse == 1|mothalc == 0)}$$

It is obvious that the odds ratio OR (abuse==1|mothalc==1) does not depend on the value of fathalc.

OR(abuse==1|{mothalc==1 and fathalc==1},{mothalc==0 and fathalc ==0})

$$= \frac{odds(abuse == 1|mothalc == 1 \; and \; fathalc == 1)}{odds(abuse == 1|mothalc == 0 \; and \; fathalc == 0)}$$

$$= \frac{p(mothalc == 1 \; and \; fathalc == 1| \; abuse == 1)/p(mothalc == 1 \; and \; fathalc == 1|abuse ==}{p(mothalc == 0 \; and \; fathalc == 0| \; abuse == 1)/p(mothalc == 0 \; and \; fathalc == 0|abuse ==}$$

$$= \frac{p(mothalc == 1|abuse == 1)p(mothalc == 0|abuse == 0)}{p(mothalc == 1|abuse == 0)p(mothalc == 0|abuse == 1)}$$

$$* \frac{p(fathalc == 0|abuse == 0)p(fathalc == 0|abuse == 0)}{p(fathalc == 1|abuse == 0)p(fathalc == 0|abuse == 1)}$$

=OR(abuse==1|mothalc==1) * OR(abuse==1|fathalc==1)

(b)

There are three cases for

OR(abuse==1|{mothalc==1 or fathalc==1},{mothalc==0 and fathalc ==0})

1st:

OR_1 = OR(abuse==1|{mothalc==1 and fathalc==1},{mothalc==0 and fathalc ==0})

Denote

{mothalc==1 and fathalc==1} as A and {mothalc==0 and fathalc ==0} as B

Then we get

$$\eta_A = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

And

$$\eta_B = \beta_0$$

So we get

$$\mathrm{OR}_1 = e^{\eta_A - \eta_B} = e^{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2}$$

Since X_1 *and* X_2 are identity metrics so we can have

$$\mathrm{OR}_1 = e^{\beta_1 + \beta_2 + \beta_3} = e^{0.63}$$

So we have

$$\beta_1 + \beta_2 + \beta_3 = 0.63$$

2$^{\text{nd}}$:

OR_2 = OR(abuse==1|{mothalc==1 and fathalc==0},{mothalc==0 and fathalc ==0})

Denote

{mothalc==1 and fathalc==0} as A and {mothalc==0 and fathalc ==0} as B

Then we get

$$\eta_A = \beta_0 + \beta_1 X_1$$

And

$$\eta_B = \beta_0$$

So we get

$$\mathrm{OR}_2 = e^{\eta_A - \eta_B} = e^{\beta_1 X_1}$$

Since X_1 is an identity metrics so we can have

$$\mathrm{OR}_2 = e^{\beta_1} = e^{0.63}$$

So we have

$$\beta_1 = 0.63$$

3$^{\text{rd}}$:

OR_3 = OR(abuse==1|{mothalc==0 and fathalc==1},{mothalc==0 and fathalc ==0})

Denote

{mothalc==0 and fathalc==1} as A and {mothalc==0 and fathalc ==0} as B

Then we get

$$\eta_A = \beta_0 + \beta_2 X_2$$

And

$$\eta_B = \beta_0$$

So we get

$$\mathrm{OR}_2 = e^{\eta_A - \eta_B} = e^{\beta_2 X_2}$$

Since X_2 is an identity metrics so we can have

$$\mathrm{OR}_2 = e^{\beta_2} = e^{0.63}$$

So we have

$$\beta_2 = 0.63$$

It is easy to get that

$$\beta_3 = -0.63$$

(c)

(i)

```
                          2.5 %     97.5 %
(Intercept) 0.0984387 0.09126191 0.1061799
mothalc     1.4697085 1.10368787 1.9571141
fathalc     1.6551649 1.40537665 1.9493498
```

The odds ratios OR(abuse == 1 j mothalc == 1) and OR(abuse == 1 j fathalc == 1) from model (1), with approximate 95% confidence intervals are 1.469709 and 1.655165 respectively.

(ii)

```
Call:
glm(formula = yabuse ~ mothalc * fathalc, family = "binomial",
    data = alcohol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6518  -0.4332  -0.4332  -0.4332   2.1966

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.31877    0.03898 -59.491  < 2e-16 ***
mothalc          0.39696    0.19933   1.992   0.0464 *
fathalc          0.50618    0.08742   5.791 7.02e-09 ***
mothalc:fathalc -0.02558    0.29282  -0.087   0.9304
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6349.8 on 9821 degrees of freedom
Residual deviance: 6303.8 on 9818 degrees of freedom
AIC: 6311.8

Number of Fisher Scoring iterations: 5

Based on the summary, there is no evidence that $\beta_3 \neq 0$.

Q2

(a)

$$\text{Odds(Document is Authentic | X)} = \frac{P(X|\text{Document is Authentic})}{p(X|\text{Document is Forged})} * \frac{P(\text{Document is Authentic})}{1-p(\text{Document is Authentic})}$$

$$\mathrm{P}(X|\text{Document is Authentic}) = \frac{n!}{n_1!n_5!n_9!} * p_{A1}^{n1}p_{A5}^{n5}p_{A9}^{n9}$$

$$\mathrm{P}(X|\text{Document is Forged}) = \frac{n!}{n_1!n_5!n_9!} * p_{F1}^{n1}p_{F5}^{n5}p_{F9}^{n9}$$

So,

$$\text{Odds(Document is Authentic | X)} = \frac{\frac{n!}{n_1!n_5!n_9!}*p_{A1}^{n1}p_{A5}^{n5}p_{A9}^{n9}}{\frac{n!}{n_1!n_5!n_9!}*p_{F1}^{n1}p_{F5}^{n5}p_{F9}^{n9}} * \frac{\pi_A}{1-\pi_A} = \frac{p_{A1}^{n1}p_{A5}^{n5}p_{A9}^{n9}}{p_{F1}^{n1}p_{F5}^{n5}p_{F9}^{n9}} * \frac{\pi_A}{1-\pi_A}$$

(b)

$$\log[\text{Odds (Document is Authentic | X)}]$$

$$= n_1\log p_{A1} + n_5\log p_{A5} + n_9\log p_{A9} - n_1\log p_{F1} - n_5\log p_{F5} - n_9\log p_{F9} + \log\pi_A - \log(1-\pi_A)$$

$$= n_1\log(p_{A1}/p_{F1}) + n_5\log(p_{A5}/p_{F5}) + n_9\log(p_{A9}/p_{F9}) + \log\left(\frac{\pi_A}{1-\pi_A}\right)$$

So,

$$a = \log(p_{A1}/p_{F1});\ b = \log(p_{A5}/p_{F5});\ c = \log(p_{A9}/p_{F9});\ d = \log\left(\frac{\pi_A}{1-\pi_A}\right).$$

(c)

We have

$$p_{A1} = 0.72,\ p_{A5} = 0.17,\ p_{A9} = 0.11,\ p_{F1} = p_{F5} = p_{F9} = 1/3.$$

Given

$$X = (n_1, n_5, n_9) = (7,5,8) \text{ and } \pi_A = \frac{1}{2}$$

We can get

$$\text{Odds(Document is Authentic | X)} = \frac{p_{A1}^{n1}p_{A5}^{n5}p_{A9}^{n9}}{p_{F1}^{n1}p_{F5}^{n5}p_{F9}^{n9}} * \frac{\pi_A}{1-\pi_A} = 1.09 * 10^{-3}$$

Since

$$\text{Odds(Document is Authentic} \mid \text{X}) = \frac{P(\text{Document is Authentic}|\text{X})}{1-p(\text{Document is Authentic}|\text{X})}$$

We can have

$$P(\text{Document is Authentic}|\text{X}) = 1.09 * 10^{-3}$$

$$P(\text{Document is Forged}|\text{X}) = 1 - P(\text{Document is Authentic}|\text{X}) = 0.9989$$

Q3

(a)

Report the P-values:

[1] "npreg : P-value = 3.20684823540759e-06"
[1] "glu : P-value = 2.01071777376076e-28"
[1] "bp : P-value = 1.02775734163267e-05"
[1] "skin : P-value = 2.26930291623554e-09"
[1] "bmi : P-value = 1.02588422797911e-11"
[1] "ped : P-value = 7.38261979346636e-08"
[1] "age : P-value = 3.45737086135437e-17"

This suggests that the full model should be used to build an accurate classifier.

(b)

Please see code.

(c)

CE = $(n_{21}+n_{12})/(n_{11}+n_{12}+n_{21}+n_{22})$

sens =$n_{22}/(n_{22} + n_{12})$

spec =$n_{11}/(n_{11}+n_{21})$

Please refer to the code for the function.

(d)

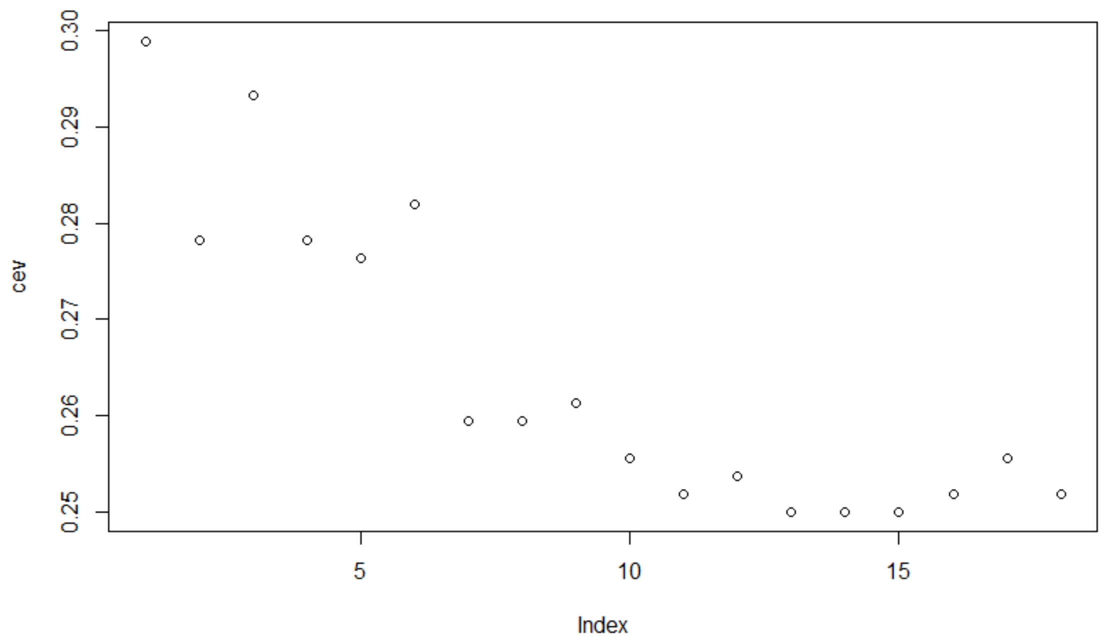
After building the KNN classifier accordingly, I found the K that minimized CE is 25.
The minimum CE = 0.25.

(ii)

Summary Statistics

	CE	sens	spec
KNN	0.2500000	0.4576271	0.8957746
LDA	0.2105263	0.5988701	0.8845070
QDA	0.2124060	0.6384181	0.8270270

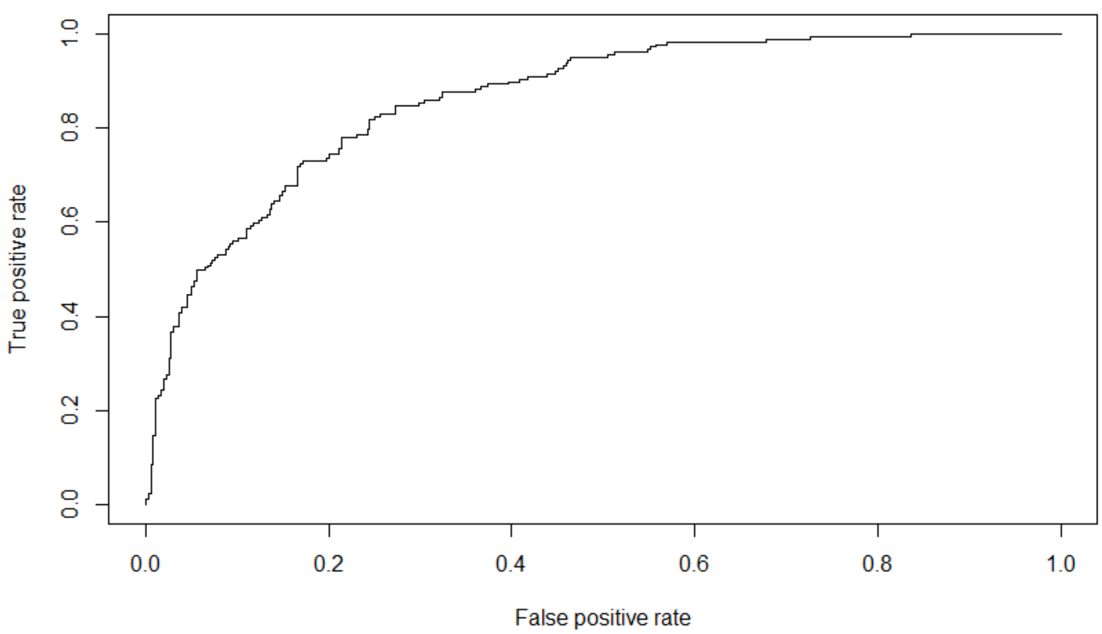
Based on the table, it can be claimed that there is no single model better than others.



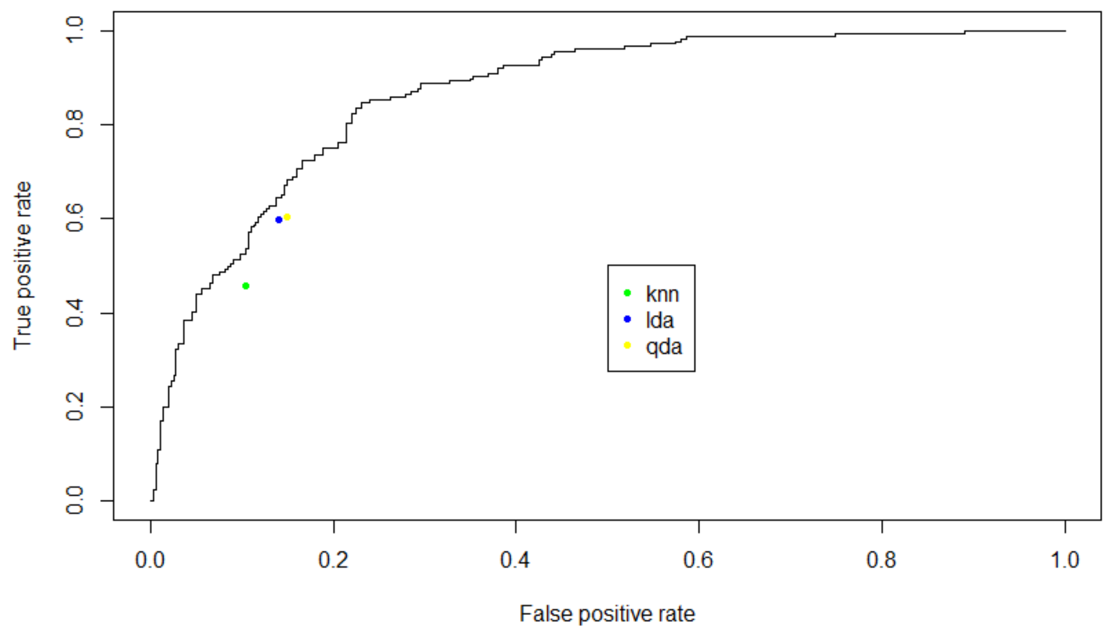
(e)

(i) Please refer to the code.

(ii) ROC curve

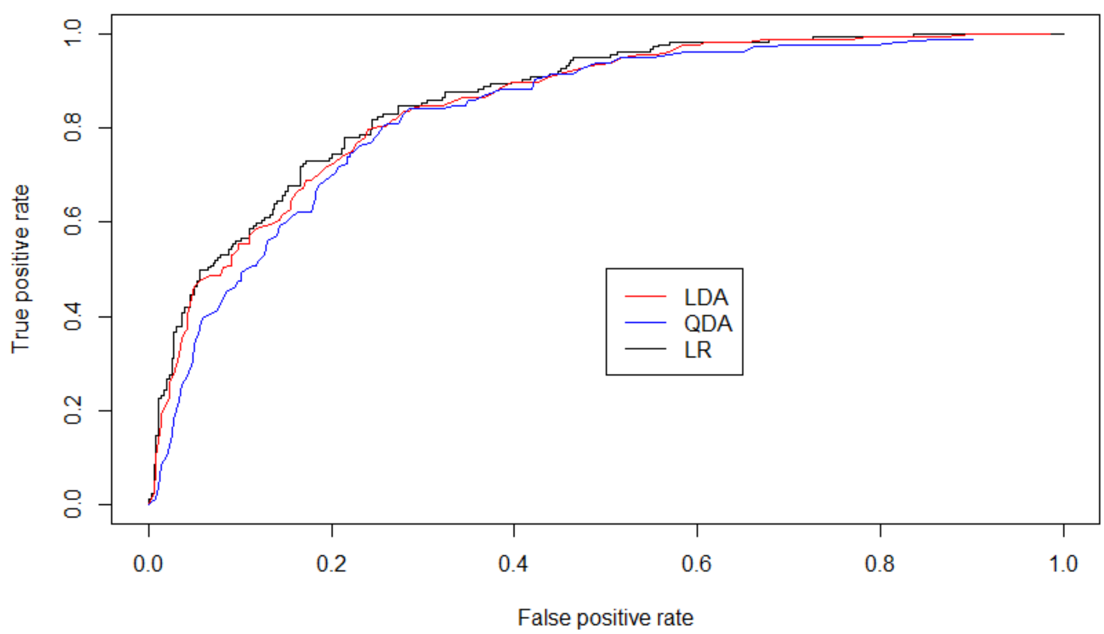


(iii) ROC curve



The plot shows a trade-off and based so that can help programmers to choose the model that make the most efficient use of the data.

(iv) Neither LDA nor QDA seems preferable.

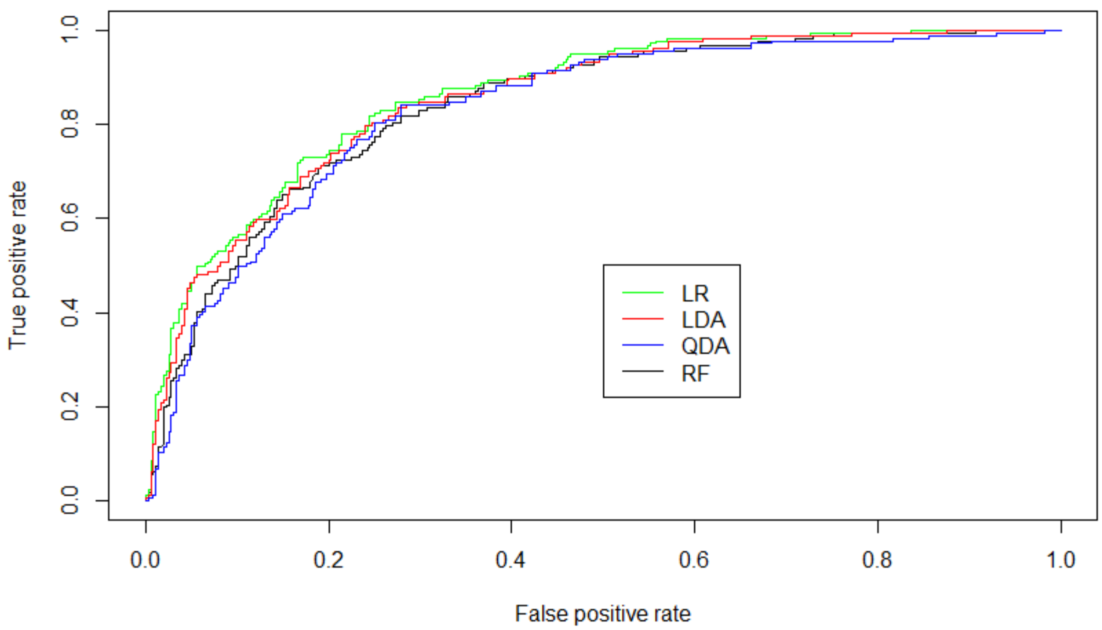


Q4

(a)

Please refer to the code.

(b)



This form of classifier does not offer any advantage over LDA or QDA for this application.