IE 242

Assignment 2.

Q.1) X_1 = hours studied .
X_2 = undergrad GPA
Y = recieve an A (binary) (1 if A)
$\hat{\beta}_0 = -6 \quad \hat{\beta}_1 = 0.05 \quad \hat{\beta}_2 = 1$

a) We know that,
Probability in logistic regression is given by,

$$P_r(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}$$

Plugging in $\hat{\beta}$ values & X values we get,

$$P_r(Y=1|X) = \frac{1}{1+e^{-(-6+0.05\times 40+1\times 3.5)}}$$

$$= \frac{1}{1+e^{-(-6+2+3.5)}}$$

$$= \frac{1}{1+e^{-(-0.5)}} = \frac{1}{1+e^{0.5}}$$

$$= \underline{\underline{0.3775}}$$

b) We have, $P_r(Y=1|X) = 0.5$.

$$\therefore 0.5 = \frac{1}{1+e^{-(-6+0.05x_1+3.5)}}$$

$$\therefore 0.5 + 0.5(e^{-(0.05x_1-2.5)}) = 1.$$

$$\therefore 0.5 + 0.5e^{(2.5-0.05x_1)} = 1$$

$e^{(2.5-0.05x_1)} = 1$

Taking $\log_e$ on both sides,

$2.5 - 0.05x_1 = 0$

$\therefore x_1 = \dfrac{2.5}{0.05}$

$\therefore \boxed{x_1 = 50 \text{ hours}}$

Q.2) Let,

$$Y = \begin{cases} 2 & \text{if stock will issue dividend} \\ 1 & \text{otherwise.} \end{cases}$$

X = Last year's % profit of company

$\bar{X} = 10$ for companies that issued dividend.

$\bar{X} = 0$ " " " didn't " " (let's call it

$\hat{\sigma}^2 = 36$

% of companies that issued dividend = 80%

= 0.8

$X \sim$ Normal

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

We know that 80% companies issued dividents, i.e. $Y=2$ for 80% companies.

$\therefore Pr(Y=2) = 0.8 = \pi_2$

$\therefore \pi_1 = Pr(Y=1) = 0.2$

Using Bayes's theorem,
We calculate posterior prob distribution of Y=1 given X=x,

$$Pr(Y=2|X=x) = \frac{\pi_2 f_2(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

~~$\therefore Pr(Y=1|X=x) = \frac{0.8 \times e^{-(x-10)^2/2\times 36}}{0.2 \times e^{-(x-0)^2/2\times 36}}$~~

~~$= 4 \cdot e^{-\frac{(x-10)^2+x^2}{2\times 36}}$~~

~~$\odot\ \cancel{4 \cdot e^{-10/72}}$~~

~~$= 4 \cdot e^{\frac{\cancel{-x^2}+20x-100+\cancel{x^2}}{72}}$~~

~~$= 4 \cdot e^{\frac{(20x-100)}{72}}$~~

~~$= 4 \cdot \exp\left(\frac{20x-100}{72}\right)$~~

$$\therefore Pr(Y=2|X=x) = \frac{0.8 \times e^{-(x-10)^2/72}}{0.2e^{-x^2/72} + 0.8e^{-(x-10)^2/72}}$$

$\therefore$ Now if $X=4$,

$$Pr(Y=2|X=4) = 4\left(\frac{e^{-(-6)^2/72}}{e^{-16/72} + 4 \cdot e^{-(-6)^2/72}}\right)$$

$$= 4\left(\frac{e^{-0.5}}{e^{-16/72} + 4\cdot e^{-0.5}}\right)$$

$$= 0.75185$$

$\therefore$ Pr (Company will issue dividend given that last year profit = 4 %) = $\underline{\underline{75.18\ \%}}$

Question 3

Part a

Below are the results from the logistic regression model taking into account all the features. As is evident from the model, ‘age’ is the most significant feature affecting the likelihood of getting a CHD in the next 10 years. The other important risk factors in descending order of significance are Systolic Blood Pressure, Gender (Male or not), glucose levels and Number of Cigarettes smoken per day.

```
> summary(mod)

Call:
glm(formula = TenYearCHD ~ ., family = "binomial", data = heart.train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.5660  -0.5879  -0.4221  -0.2890   2.8298  

Coefficients:
                                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)                            -8.495e+00  8.528e-01  -9.962  < 2e-16 ***
male                                    4.292e-01  1.306e-01   3.285 0.001020 ** 
age                                     6.495e-02  7.964e-03   8.156 3.47e-16 ***
educationHigh school/GED               -1.209e-01  2.131e-01  -0.567 0.570495    
educationSome college/vocational school -7.719e-02  2.311e-01  -0.334 0.738329    
educationSome high school               6.404e-02  1.970e-01   0.325 0.745147    
currentSmoker                           1.033e-01  1.868e-01   0.553 0.580275    
cigsPerDay                              1.739e-02  7.411e-03   2.346 0.018961 *  
BPMeds                                 -1.072e-01  2.835e-01  -0.378 0.705326    
prevalentStroke                         9.369e-01  5.912e-01   1.585 0.113042    
prevalentHyp                            2.443e-01  1.691e-01   1.445 0.148444    
diabetes                               -5.921e-03  3.903e-01  -0.015 0.987897    
totChol                                 1.872e-03  1.351e-03   1.386 0.165826    
sysBP                                   1.678e-02  4.791e-03   3.502 0.000462 ***
diaBP                                  -7.463e-03  7.847e-03  -0.951 0.341558    
BMI                                     4.455e-03  1.546e-02   0.288 0.773177    
heartRate                              -8.383e-07  4.995e-03   0.000 0.999866    
glucose                                 8.356e-03  2.721e-03   3.071 0.002132 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2185.3  on 2560  degrees of freedom
Residual deviance: 1930.7  on 2543  degrees of freedom
AIC: 1966.7

Number of Fisher Scoring iterations: 5
```

i.

Above, you can see the results of predictions made by the model on the test set. The equation of the fitted logistic regression model is as follows:

$$P(Y = 1|X) = \frac{1}{1 + e^{-w}}$$

Where, $w = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_{17} x_{17}$

The β values corresponding to individual features are given in the ‘Estimate’ column in the summary of the model in the figure shown above. The *x* values represent the numerical values of the independent variables or the features.

ii.

As is evident from the model, ‘age’ is the most significant feature affecting the likelihood of getting a CHD in the next 10 years. The other important risk factors in descending order of significance are Systolic Blood Pressure, Gender (Male or not), glucose levels and Number of Cigarettes smoked per day.

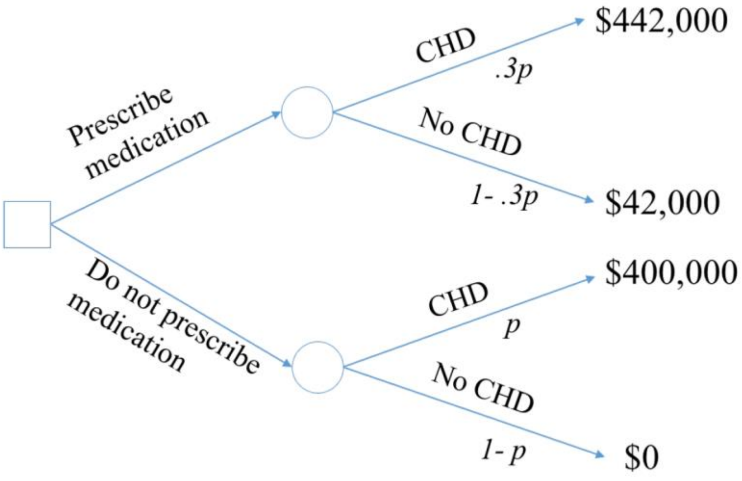
Let’s consider the variable ‘age’ since it is the most significant risk factor of all. The value of its estimated coefficient β from the model is +0.06495. The positive sign indicates that the odds of a person getting a CHD in the next 10 years increase as the age of the person increases.

The factorial increase in odds of getting a CHD in the next 10 years for a unit increase in age is equal to $e^{0.06495} = 1.0671$.

Thus for every 1 year increase in age of a person, the odds of getting a CHD in next 10 years are multiplied by 1.0671. In other words, the odds increase by 6.71%.

iii.

The decision tree representing the costs with and without treatment along with corresponding probabilities is given below.



Based on the decision tree, we can equate the situations where we prescribe medication or not as follows:

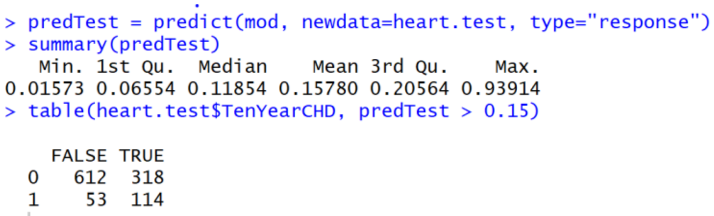
$$442000 * 0.3p + (1 - 0.3p) * 42{,}000 = 400{,}000p + 0 * (1 - p)$$

Solving above equation, **we get threshold p value as 0.15.**

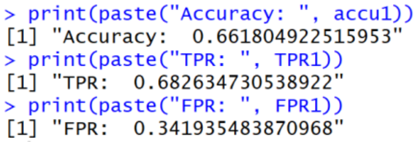
Thus **we prescribe medication if the probability of getting a CHD in next 10 years is greater than 0.15.**

iv.

The statistical data, confusion matrix, accuracy, True Positive Rate and False Positive Rate for predictions made on Test set with threshold probability value of 0.15 are shown below.



```
> predTest = predict(mod, newdata=heart.test, type="response")
> summary(predTest)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.01573 0.06554 0.11854 0.15780 0.20564 0.93914 
> table(heart.test$TenYearCHD, predTest > 0.15)
   
    FALSE TRUE
  0   612  318
  1    53  114
```



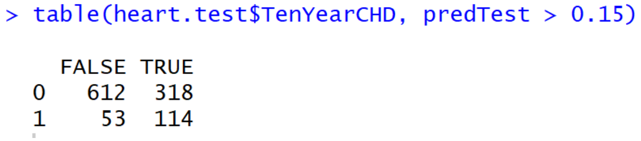
```
> print(paste("Accuracy: ", accu1))
[1] "Accuracy:  0.661804922515953"
> print(paste("TPR: ", TPR1))
[1] "TPR:  0.682634730538922"
> print(paste("FPR: ", FPR1))
[1] "FPR:  0.341935483870968"
```

We can see that the accuracy of the model is 66.18% which is not very good. From an economic and ethical standpoint, we need a model with better accuracy. The false negative predictions are only 53 which is also a concern and need to be countered. If there are more false negative predictions, it is essentially bad for patients and would also raise ethical issues.

The model correctly predicts non-risk patients equal to 612. It also correctly predicts patients at risk which are equal too 114. The model predicts wrongly that 53 patients would not get CHD when they actually did and also predicts incorrectly for 318 patients that they are ate risk for CHD when in fact they didn't get CHD.

v.

In the first case, we assume that the medication does not affect the risk of a person getting CHD. So in this case, we can directly calculate the cost based on the confusion matrix with threshold value of p = 0.15. The confusion matrix and cost calculations are shown below:



```
> table(heart.test$TenYearCHD, predTest > 0.15)
   
    FALSE TRUE
  0   612  318
  1    53  114
```

$$Economic\ Cost\ per\ person = \frac{(0 * True\ Negatives) + (42{,}000 * False\ Positives) + (400{,}000 * False\ Negatives) + (442{,}000 * True\ Positives)}{Total\ Number\ of\ people}$$

$$Economic\ Cost\ per\ person = \frac{(0 * 612) + (42{,}000 * 318) + (400{,}000 * 53) + (442{,}000 * 114)}{612 + 318 + 53 + 114} \approx \$77{,}433$$

Thus, we have expected economic cost equal to **$77,433 per person.**

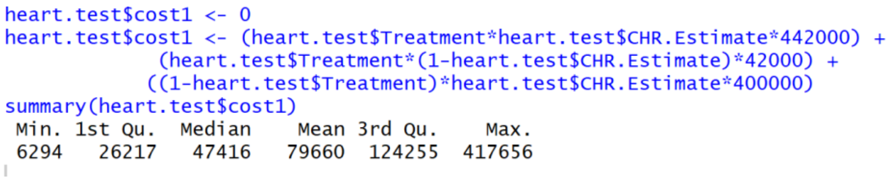
Alternatively, we can also calculate expected economic cost in first case by taking into account the probability of each prediction as follows:

$$Expected\ cost\ per\ person = \frac{1}{n}\sum_{i=1}^{n}[(T_i * p_i * 442000) +\ (T_i * (1 - p_i) * 42000) +\ ((1 - T_i) * p_i * 400000)]$$

Where, $T_i = \begin{cases} 1, & if\ Medication\ is\ prescribed \\ 0, & if\ Medication\ is\ not\ prescribed \end{cases}$

$p_i = Risk\ probability\ value\ predicted\ by\ the\ model$

$n = Number\ of\ predictions(people)in\ test\ set$



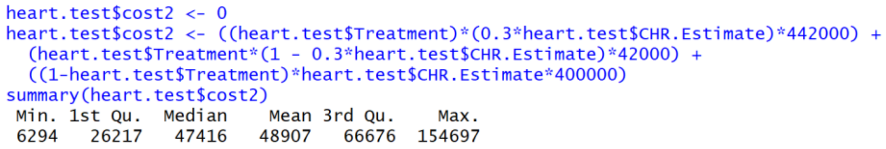
```
heart.test$cost1 <- 0
heart.test$cost1 <- (heart.test$Treatment*heart.test$CHR.Estimate*442000) +
             (heart.test$Treatment*(1-heart.test$CHR.Estimate)*42000) +
            ((1-heart.test$Treatment)*heart.test$CHR.Estimate*400000)
summary(heart.test$cost1)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 6294   26217   47416   79660  124255  417656 
```

The Treatment column contains a binary variable whether treatment should be provided (Value = 1) or not (Value = 0) based on threshold p=0.15. The 'CHR.Estimate' columns the risk probability values as predicted by the model. Using the results from our model in R, we can get the Estimated economic cost per person equal to $79,660 per person.

The assumption would make sense only if clinical trials of the treatment have been proved to be ineffective. In that case, it is unnecessary for people to spend extra money on the treatment. However, if the medication does perform as per the claims, the assumption does not make sense to use the same risk values in calculating economic costs even if the medication is prescribed.

In case 2, we must account for the reduced risk of CHD in case the medication is prescribed. We know that the risk of CHD reduces by 70% if medication is prescribed. Thus, we must account for that in our calculations of expected economic costs. Thus the equation for calculating expected economic costs is modified as follows.

$$Expected\ cost\ per\ person = \frac{1}{n}\sum_{i=1}^{n}[(T_i * (0.3 * p_i) * 442000) + \ (T_i * (1 - 0.3 * p_i) * 42000) + \ ((1 - T_i) * p_i * 400000)]$$

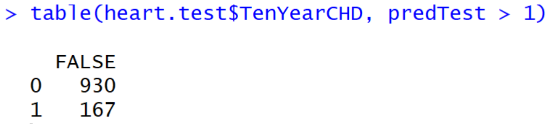


```
heart.test$cost2 <- 0
heart.test$cost2 <- ((heart.test$Treatment)*(0.3*heart.test$CHR.Estimate)*442000) +
  (heart.test$Treatment*(1 - 0.3*heart.test$CHR.Estimate)*42000) +
  ((1-heart.test$Treatment)*heart.test$CHR.Estimate*400000)
summary(heart.test$cost2)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 6294   26217   47416   48907   66676  154697 
```

All the notations are the same as before. Based on this equation and results obtained from R, We get the expected economic cost per person to be **$48,907 per person**.

vi.

The baseline model predicts that none of the patients are at risk for CHD in the next 10 years. This is essentially equivalent to the threshold value of *p* being equal to 1.0. Thus, using p=1, we get the confusion matrix and performance of baseline model as follows:



```
> table(heart.test$TenYearCHD, predTest > 1)

    FALSE
  0   930
  1   167
```

$Accuracy = \frac{930}{930+167}$ = 0.8478 = 84.78%

True Positive Rate = 0

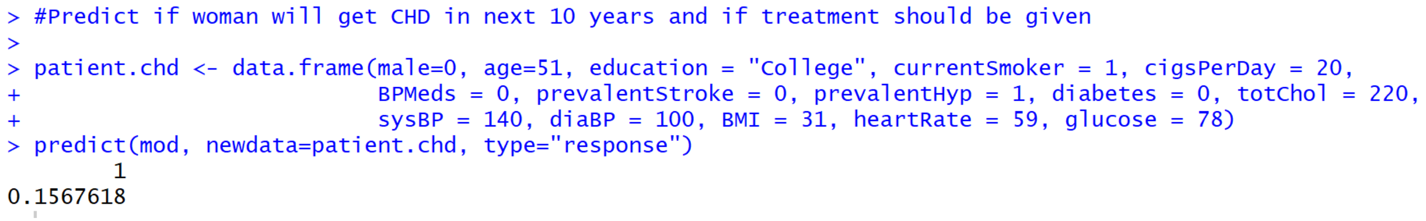
False Positive Rate = 0

$$Expected\ economic\ cost\ per\ person = \frac{(0 * 930) + (400{,}000 * 167)}{(930 + 167)} = \$\mathbf{60{,}893.34}$$

Thus, we can see that even though the baseline model has a better accuracy in predicting CHD, in terms of economic efficiency, our model performs better since it has a much lower economic cost per person provided that the medication is actually effective as per the claims.

vii.

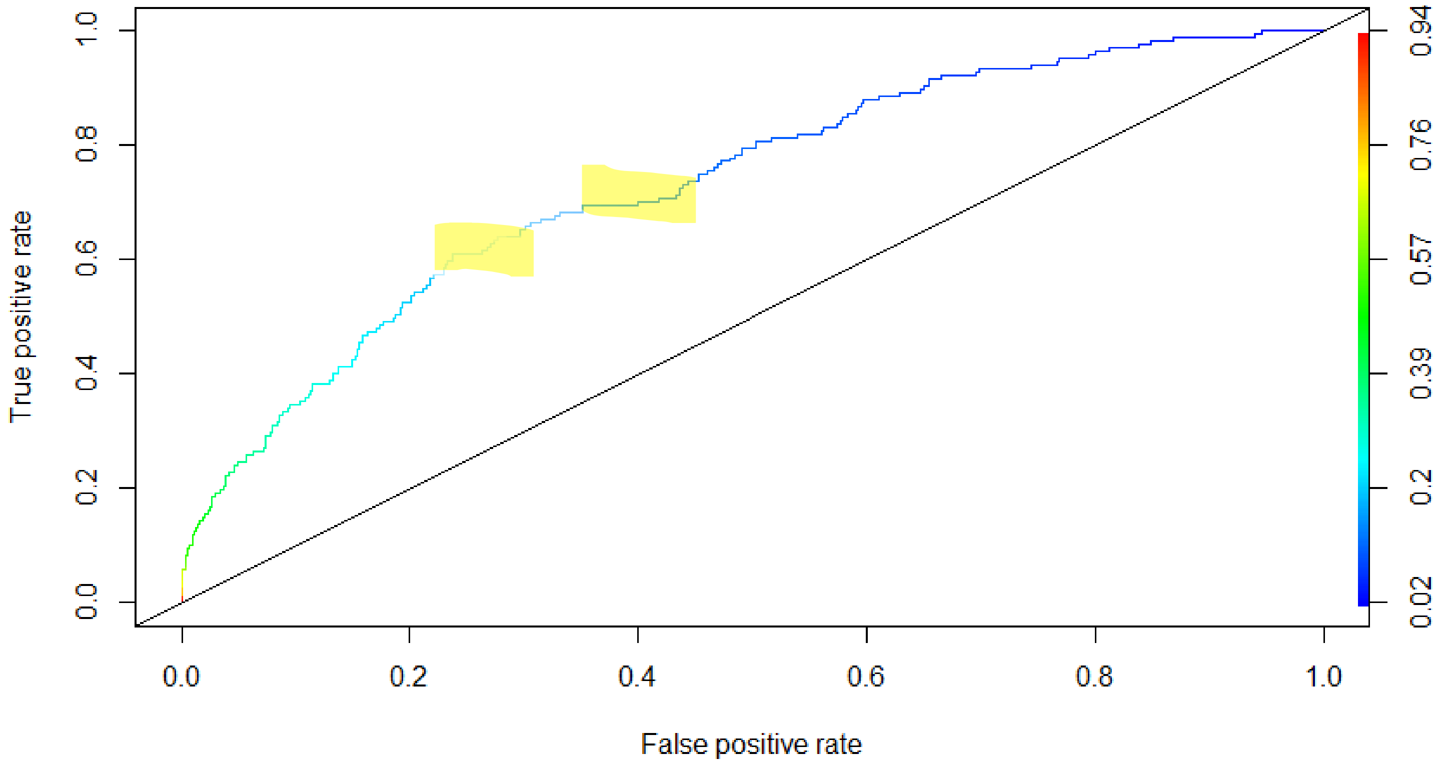
The data for the patient is given in the question. If we plug in the values of the data in our model, we get results as follows:



```
> #Predict if woman will get CHD in next 10 years and if treatment should be given
>
> patient.chd <- data.frame(male=0, age=51, education = "College", currentSmoker = 1, cigsPerDay = 20,
+                           BPMeds = 0, prevalentStroke = 0, prevalentHyp = 1, diabetes = 0, totChol = 220,
+                           sysBP = 140, diaBP = 100, BMI = 31, heartRate = 59, glucose = 78)
> predict(mod, newdata=patient.chd, type="response")
        1
0.1567618
```

Thus we can see that the risk probability of the woman getting a CHD in next 10 years is 0.1567618 which is slightly higher than our threshold risk of p=0.15. **Thus, according to our threshold from the decision tree, the physician should prescribe the treatment to the woman.**

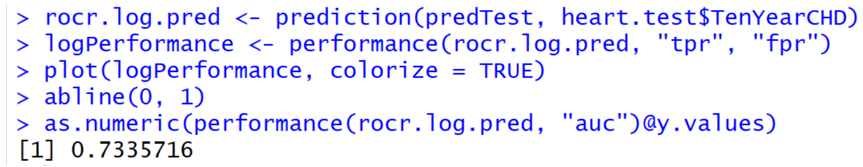
Part B:



The ROC curve for our logistic regression model is as shown above. We can use the ROC curve to determine the ideal threshold we need to use so that our model would perform predictions as per our priorities. We can infer from the curve that for the medication under consideration in our analysis, there are certain points (like the ones highlighted in yellow) where the slope of the tangent to the curve is zero or almost zero. This indicates that with an increase/decrease in threshold risk, the true positive rate is constant.

Under the scenario in consideration, we can afford to have a bit higher false positive rate since we know that the medication can reduce the risk of patients getting a CHD in next 10 years by 70%. Thus, we can choose a point based on our priorities in one of these highlighted regions which has a lower threshold. Also, from a business standpoint, we would prefer to have a higher false positive rate also since the company could advertise it as the prevention of CHD was caused by its medication.

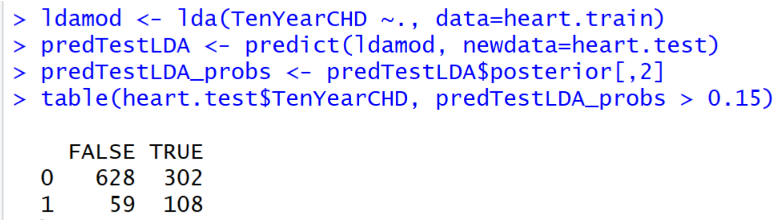
The area under the curve AUC is 0.7335716. (R result shown below)



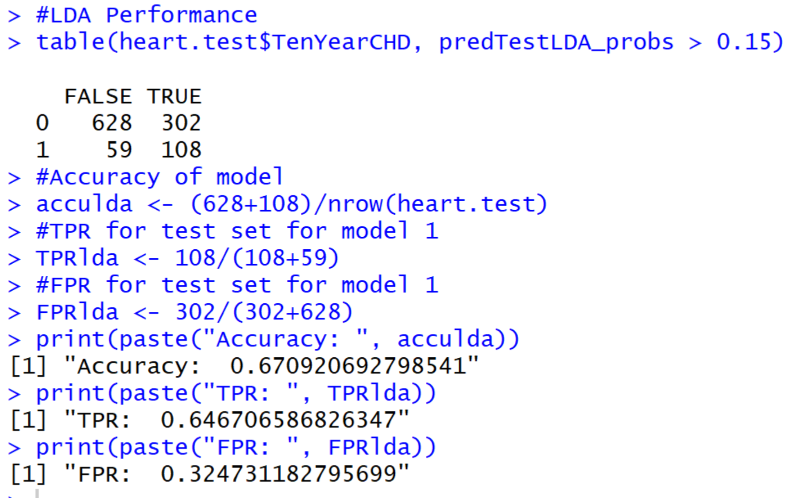
```
> rocr.log.pred <- prediction(predTest, heart.test$TenYearCHD)
> logPerformance <- performance(rocr.log.pred, "tpr", "fpr")
> plot(logPerformance, colorize = TRUE)
> abline(0, 1)
> as.numeric(performance(rocr.log.pred, "auc")@y.values)
[1] 0.7335716
```

Part C:

The results obtained from the LDA model are as follows:



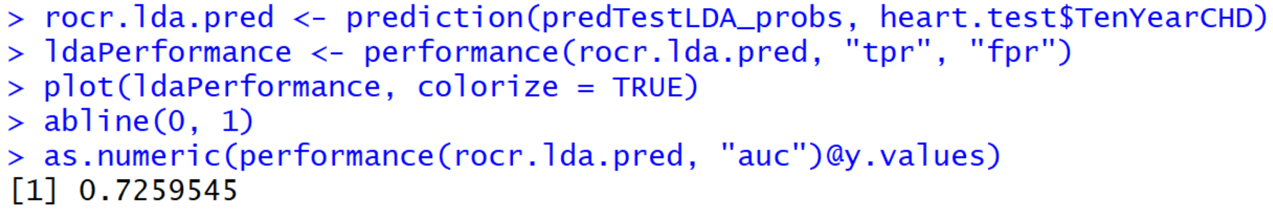
```
> ldamod <- lda(TenYearCHD ~., data=heart.train)
> predTestLDA <- predict(ldamod, newdata=heart.test)
> predTestLDA_probs <- predTestLDA$posterior[,2]
> table(heart.test$TenYearCHD, predTestLDA_probs > 0.15)

    FALSE TRUE
  0   628  302
  1    59  108
```



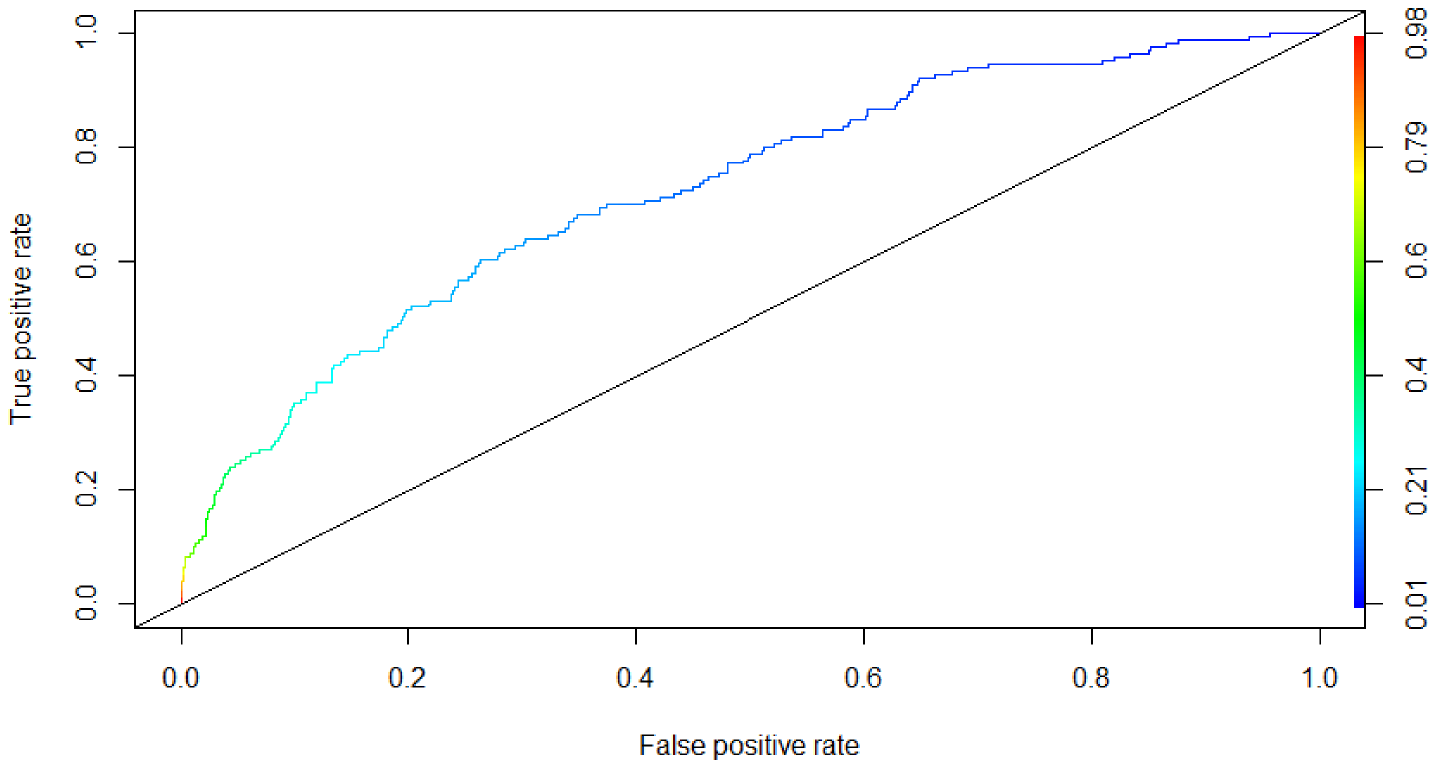
```
> #LDA Performance
> table(heart.test$TenYearCHD, predTestLDA_probs > 0.15)

    FALSE TRUE
  0   628  302
  1    59  108
> #Accuracy of model
> acculda <- (628+108)/nrow(heart.test)
> #TPR for test set for model 1
> TPRlda <- 108/(108+59)
> #FPR for test set for model 1
> FPRlda <- 302/(302+628)
> print(paste("Accuracy: ", acculda))
[1] "Accuracy:  0.670920692798541"
> print(paste("TPR: ", TPRlda))
[1] "TPR:  0.646706586826347"
> print(paste("FPR: ", FPRlda))
[1] "FPR:  0.324731182795699"
```

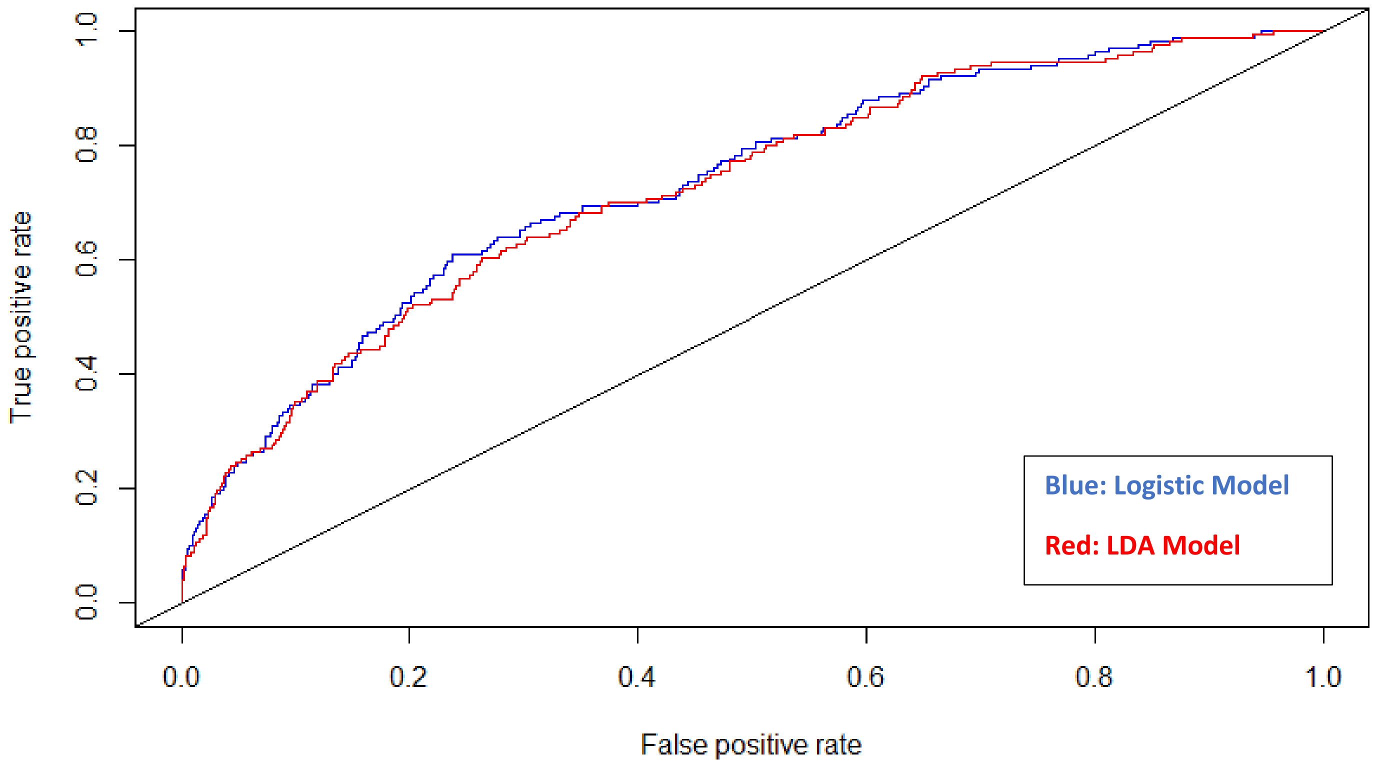
We can see that the accuracy of the LDA model is better than the logistic regression model. However, the True positive rate for the LDA model is lower too which is bad. The False positive rate is also lower for LDA which is good.

The ROC curve and AUC for the LDA model as well as the combined ROC curves for LDA and Logistic models are given below:



```
> rocr.lda.pred <- prediction(predTestLDA_probs, heart.test$TenYearCHD)
> ldaPerformance <- performance(rocr.lda.pred, "tpr", "fpr")
> plot(ldaPerformance, colorize = TRUE)
> abline(0, 1)
> as.numeric(performance(rocr.lda.pred, "auc")@y.values)
[1] 0.7259545
```


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.98
0.79
0.6
0.4
0.21
0.01


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Blue: Logistic Model
Red: LDA Model

The area under curve (AUC) for the LDA Model applied to test set is 0.7259545.

Based on ROC plot of LDA and comparing the 2 plots of LDA and Logistic regression, **we can infer that logistic regression model is a better choice for our analysis since the area under the curve is slightly greater for the logistic model.** Also logistic model is more computationally inexpensive than the LDA model and is also simpler and easier.

Part D:

The model predicts the risk of people getting a CHD in the next 10 years and based on that, recommends if the medication to reduce the risk should be prescribed or not. The ethical concerns that may be raised in this analysis are that if we predict a person will not have CHD and do not prescribe medication, but still the person gets CHD (False Negative predictions), it is a heavy economic and psychological burden on the patient ($400,000 vs $42,000 if we prescribed and prevented CHD). It also would pose fatal health problems which may have been prevented with the medication. Thus, it may backfire on our decisions and have a negative impact on the physician or company who would use this model.

There are 2 ways in which we could counter that. First, we could introduce heavy penalties/costs for unfavorable predictions (like False negatives) and essentially train the model to optimize the total penalties to minimize them. This will ensure that our model makes less predictions that are unfavorable thus avoiding the ethical issues.

Another way in which we can tackle this is by reducing the risk threshold for prescribing the medication. This will cause the predictions to be overall inclined towards positives and recommend us to prescribe medication. This may not be the most economic way based on costs per person, but it would tackle the ethical issues arising out of false negative predictions and would also keep our conscience clear!

In some cases it may also cause problems if the patients are prescribed the medication but still get a CHD in future. This may pertain to lifestyle and other factors related to the patient. Such issues should be mitigated by letting the patient know about the statistics and the effectiveness of the treatment as well as supplementary habits that would help the treatment be effective. This should be done at the time of prescribing the treatment so that the patient is aware and would lessen the chances of an ethical issue in the future.

R Script:

```
#install.packages(c("caTools", "ROCR", "dplyr", "ggplot2", "GGally"))


#Load packages

library(dplyr)

library(ggplot2)

library(GGally)

library(caTools)

library(ROCR)


#Read Data

heart <- read.csv("framingham.csv")
```

```
set.seed(144)

#Split data into train and test set
split = sample.split(heart$TenYearCHD, SplitRatio = 0.7)
heart.train <- filter(heart, split == TRUE)
heart.test <- filter(heart, split == FALSE)

#Train dataset statistics
#table(heart.train$TenYearCHD)

#ggscatmat(heart.train, alpha = 0.8)

#Logistic regression model

mod <- glm(TenYearCHD ~., data=heart.train, family="binomial")
summary(mod)

#Applying the model to test set
#Break-even p=0.15, i.e. Only reasonable to prescribe medication if p>0.15

predTest = predict(mod, newdata=heart.test, type="response")
summary(predTest)

table(heart.test$TenYearCHD, predTest > 0.15)
#Accuracy of model
accu1 <- (612+114)/nrow(heart.test)
#TPR for test set for model 1
TPR1 <- 114/(114+53)
#FPR for test set for model 1
FPR1 <- 318/(318+612)
TPR1
```

```
print(paste("Accuracy: ", accu1))
print(paste("TPR: ", TPR1))
print(paste("FPR: ", FPR1))

#Logistic regression model 2: Dropping heartRate

mod2 <- glm(TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay + BPMeds + prevalentStroke +
           prevalentHyp + totChol + sysBP + diaBP + BMI + diabetes + glucose, data=heart.train, family="binomial")
summary(mod2)

#Logistic regression model 3: Dropping diabetes

mod3 <- glm(TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay + BPMeds + prevalentStroke +
            prevalentHyp + totChol + sysBP + diaBP + BMI + glucose, data=heart.train, family="binomial")
summary(mod3)

#Logistic regression model 4: Dropping BMI

mod4 <- glm(TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay + BPMeds + prevalentStroke +
            prevalentHyp + totChol + sysBP + diaBP + glucose, data=heart.train, family="binomial")
summary(mod4)

#Logistic regression model 5: Dropping education

mod5 <- glm(TenYearCHD ~ male + age + currentSmoker + cigsPerDay + BPMeds + prevalentStroke +
            prevalentHyp + totChol + sysBP + diaBP + glucose, data=heart.train, family="binomial")
summary(mod5)

#Logistic regression model 6: Dropping BPMeds

mod6 <- glm(TenYearCHD ~ male + age + currentSmoker + cigsPerDay + prevalentStroke +
            prevalentHyp + totChol + sysBP + diaBP + glucose, data=heart.train, family="binomial")
```

```
summary(mod6)

#Logistic regression model 7: Dropping currentSmoker

mod7 <- glm(TenYearCHD ~ male + age + cigsPerDay + prevalentStroke +
            prevalentHyp + totChol + sysBP + diaBP + glucose, data=heart.train, family="binomial")
summary(mod7)

#Logistic regression model 8: Dropping diaBP

mod8 <- glm(TenYearCHD ~ male + age + cigsPerDay + prevalentStroke +
            prevalentHyp + totChol + sysBP + glucose, data=heart.train, family="binomial")
summary(mod8)

#Logistic regression model 9: Dropping prevalentHyp

mod9 <- glm(TenYearCHD ~ male + age + cigsPerDay + prevalentStroke +
            totChol + sysBP + glucose, data=heart.train, family="binomial")
summary(mod9)

#Logistic regression model 10: Dropping totChol

mod10 <- glm(TenYearCHD ~ male + age + cigsPerDay + prevalentStroke +
            sysBP + glucose, data=heart.train, family="binomial")
summary(mod10)

#Logistic regression model 11: Dropping prevalentStroke

mod11 <- glm(TenYearCHD ~ male + age + cigsPerDay + sysBP + glucose, data=heart.train, family="binomial")
summary(mod11)

#Applying the model to test set
```

```
#Break-even p=0.15, i.e. Only reasonable to prescribe medication if p>0.15

predTest = predict(mod, newdata=heart.test, type="response")
summary(predTest)

table(heart.test$TenYearCHD, predTest > 0.15)
#Accuracy of model
accu1 <- (612+114)/nrow(heart.test)
#TPR for test set for model 1
TPR1 <- 114/(114+53)
#FPR for test set for model 1
FPR1 <- 318/(318+612)
print(paste("Accuracy: ", accu1))
print(paste("TPR: ", TPR1))
print(paste("FPR: ", FPR1))

predTest2 = predict(mod11, newdata=heart.test, type="response")
summary(predTest2)

table(heart.test$TenYearCHD, predTest2 > 0.15)
#Accuracy of model
accu2 <- (607+111)/nrow(heart.test)
#TPR for test set for model 1
TPR2 <- 111/(111+56)
#FPR for test set for model 1
FPR2 <- 323/(323+607)
print(paste("Accuracy: ", accu2))
print(paste("TPR: ", TPR2))
print(paste("FPR: ", FPR2))

#a.v
#Assuming treatment to be provided if p>0.15
```

```
#We add a column to dataset if treatment should be provided or not.
heart.test$Treatment <- 0
heart.test[predTest > 0.15,]$Treatment <- 1

#Add column to estimate expected cost for patients if treatment doesn't affect CHR risk
heart.test$CHR.Estimate <- predTest
heart.test$cost1 <- 0
heart.test$cost1 <- (heart.test$Treatment*heart.test$CHR.Estimate*442000) +
            (heart.test$Treatment*(1-heart.test$CHR.Estimate)*42000) +
           ((1-heart.test$Treatment)*heart.test$CHR.Estimate*400000)
summary(heart.test$cost1)
heart.test$cost2 <- 0
heart.test$cost2 <- ((heart.test$Treatment)*(0.3*heart.test$CHR.Estimate)*442000) +
  (heart.test$Treatment*(1 - 0.3*heart.test$CHR.Estimate)*42000) +
  ((1-heart.test$Treatment)*heart.test$CHR.Estimate*400000)
#write.csv(heart.test, "hearttestcosts.csv")

#Performance of baseline model
table(heart.test$TenYearCHD, predTest > 1)

#a.vii
#Predict if woman will get CHD in next 10 years and if treatment should be given

patient.chd <- data.frame(male=0, age=51, education = "College", currentSmoker = 1, cigsPerDay = 20,
                        BPMeds = 0, prevalentStroke = 0, prevalentHyp = 1, diabetes = 0, totChol = 220,
                        sysBP = 140, diaBP = 100, BMI = 31, heartRate = 59, glucose = 78)
predict(mod, newdata=patient.chd, type="response")

# Question 3: Part b
rocr.log.pred <- prediction(predTest, heart.test$TenYearCHD)
logPerformance <- performance(rocr.log.pred, "tpr", "fpr")
plot(logPerformance, colorize = TRUE)
```

```
abline(0, 1)
as.numeric(performance(rocr.log.pred, "auc")@y.values)

#Question 3: Part c
library(MASS)
ldamod <- lda(TenYearCHD ~., data=heart.train)
summary(ldamod)

predTestLDA <- predict(ldamod, newdata=heart.test)
predTestLDA_probs <- predTestLDA$posterior[,2]

#LDA Performance
table(heart.test$TenYearCHD, predTestLDA_probs > 0.15)
#Accuracy of model
acculda <- (628+108)/nrow(heart.test)
#TPR for test set for model 1
TPRlda <- 108/(108+59)
#FPR for test set for model 1
FPRlda <- 302/(302+628)
print(paste("Accuracy: ", acculda))
print(paste("TPR: ", TPRlda))
print(paste("FPR: ", FPRlda))

#LDA ROCR
rocr.lda.pred <- prediction(predTestLDA_probs, heart.test$TenYearCHD)
ldaPerformance <- performance(rocr.lda.pred, "tpr", "fpr")
plot(ldaPerformance, colorize = TRUE)
abline(0, 1)
as.numeric(performance(rocr.lda.pred, "auc")@y.values)

#Combined Logistic-LDA ROCR plots
plot(logPerformance, col="blue")
plot(ldaPerformance, col="red", add=TRUE)

abline(0,1)
```