Problem #1

1a) X is restricted to the set {0,1} (i.e categorical) and $P_{X,Y}$ is given by distribution in Table1.

Outcome	Probability
(X=0, Y=0)	0.1
(X=0, Y=1)	0.2
(X=1, Y=0)	0.4
(X=1, Y=1)	0.3

I know there is way to do this through deriving but I want to plug the values instead.

Given $R(h) = Pr(Y \neq h(X))$. Derive largest ϵ^* value for both cases. explain.

→ Also X is restricted

we do Y=0 because of ↓

So two cases first where we want to compare values: → $R(H) = Pr(Y = h(X)$

→ This when Y=0 → so we do these value when combination Y=0 with X.

$P(Y=0) \Rightarrow P(X=0, Y=0) + P(X=1, Y=0) = \underline{0.1 + 0.4 = 0.5}$

Now same with Y=1

> based on table 1.

$P(Y=1) \Rightarrow P(X=0, Y=1) + P(X=1, Y=1) = \underline{0.2 + 0.3 = 0.5}$.

This finding tells us there is no majority class,

so what we need to do is take max value of Y=0, Y=1

→ Answer value for ϵ^* but can be 0.6 larger ϵ^*

$(Y=0) = \max(P(X=0, Y=0) + P(X=1, Y=0) = \max(0.1, 0.4) = \boxed{0.4}$

$(Y=1) = \max(P(X=0, Y=1) + P(X=1, Y=1) = \max(0.2, 0.3) = \underline{0.3}$

$\max(0.3, 0.4) = \boxed{0.4}$ also 0.6 larger

lower

So basically the largest possible error rate for any classifier, would be the same error rate or the largest one because regardless x value you get 0.5. So the lower ϵ^* would be 0.4, largest 0.6

b) The marigned distribution of Y is given by Pr(Y=0) : 0.3 and Pr (Y=1) .0 7 Given that Y=0, the distribution of X normal with mean 5 and variance 2 Given that Y=1, the Aribution of X is normal with mean -3 and variance 2.

Given: Y=0, X normal with mean 5 and variance 2.

Y=1, X normal with mean -3 and variance 2

Both and Code for CDF.

CDF for Y=0 → normal distribution

$X \sim N(5,2)$

↳ CDF $\underline{\Phi(5,2)}$

CDF for Y=1 $X \sim N(-3,2)$

↳ CDF $\underline{\Phi(-3,2)}$

I am going to use statistical definition for standard normal distribution which is denoted by Φ.

$$\Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{u^2}{2}\right\} du$$

so what I am going to do to solve for this to use built in function from stats, and do the CDF value for both. In Python and by hand.

Y=0 | Y=1

$Pr(Y=0, X \leq 1) = Pr(Y=0) \times Pr(X \leq 1 | Y=0)$

$= 0.3 * \Phi\left(2 < \frac{1-5}{\sqrt{2}}\right)$

$= 0.3\,(0.00233)$

$Pr(Y=1, X>1) = Pr(Y=1) * Pr(X>1 | Y=1)$

$= 0.7 \times \Phi\left(2 > \frac{1+3}{\sqrt{2}}\right)$

$= 0.7\,(0.00233)$

$R(h) = Y=0 \Rightarrow 0.0062097$

$R(\hat{h}) = Pr(X \neq h(x)) = Pr(Y=0, x>1) +$ Previous value $Pr(Y=1, X \leq 1)$

$= 0.3 \times (1 - \Phi(-2.83) + 0.7\,\Phi(2.83)$

$= 0.933192$

classifier would be 0.3 or 0.7 where Y=0 or Y=1 regardless value of X. Thats why error rate of 0.3 and 0.7, so between 0.006 209 and 0.933

so ϵ^* 0.933192 for largest

Problem #7: consider binary classification problem where $X \in \mathbb{R}^p$ and $Y \in \{0,1\}$ for a fixed $x \in R^p$, suppose that $Pr(Y=1 \mid X=x) = p$ for some $p \in [0,1]$. Consider prediction problem where there is a loss $L_{FN} > 0$ associated with predicting $Y=0$ when the actual outcome is $Y=1$, and another loss $L_{FP} > 0$ associated with predicting $Y=1$ when the actual outcome is $Y=0$. Threshold $Y=1$ if $p \geq \bar{p}$ and predicting $Y=0$. What is value of $\bar{p}$?

Given: If you predict $Y=0$, when the actual is $Y=1 \Rightarrow L_{FN}$

If you predict $Y=1$, when the actual is $Y=0 \Rightarrow L_{FP}$

Lets do two case → This to find the expected loss for a given threshold of $\bar{p}$

One where $Y=0$ and when $p \geq \bar{p}$ ← P_hat

↳ $E[\text{loss} \mid P \geq \bar{p}] = L_{FN} * Pr(Y=1 \mid X=x)$

Because when we do False Negative it is $Pr(Y=1 \mid X=x)$ when $p \geq \bar{p}$.

Second: where $Y=1$ when $p < \bar{p}$

↳ $E[\text{loss} \mid P < p] = L_{FP} \times Pr(Y=0 \mid X=x)$ when $p < \bar{p}$

Okay now we write the given information into a equation.

$$Pr(Y=1 \mid X=x) = p$$

when the expected loss $L(p)$ defined as:

$$L(p) = p * L_{FN} + (1-p) * L_{FP}$$

Now simply take derivative respect to p, to minimize to find the threshold.

$$\frac{d'L(p)}{d'(p)} = p * L_{FN} + (1-p) \times L_{FP}$$

$$\boxed{\bar{p} = \frac{L_{FP}}{(L_{FN} + L_{FP})}}$$

back for more → work to explain

The other way is to use the same expected value function

$$L(P) = P * L_{FN} + (1-P) * L_{FP}$$

↳ rerrang the equation to just find P which is what we want.

$$L(P) = P * L_{FN} + \underbrace{(1-P) * L_{FP}}_{*\text{load}} \leftarrow$$

$$= P * L_{FN} + 1L_{FP} - P L_{FP}$$ ← I think I mmg keep it split

$P * L_{FN} = L_{FP}(1-P)$ ← now both side have P so

simply ↓

$$\underline{PL_{FN}} = L_{FP} - \underline{PL_{FP}}$$

$$PL_{FN} + L_{FP} = L_{FP}$$

↓

now split P out again:

$$P(L_{FN} + L_{FP}) = L_{FP}$$

solve for P

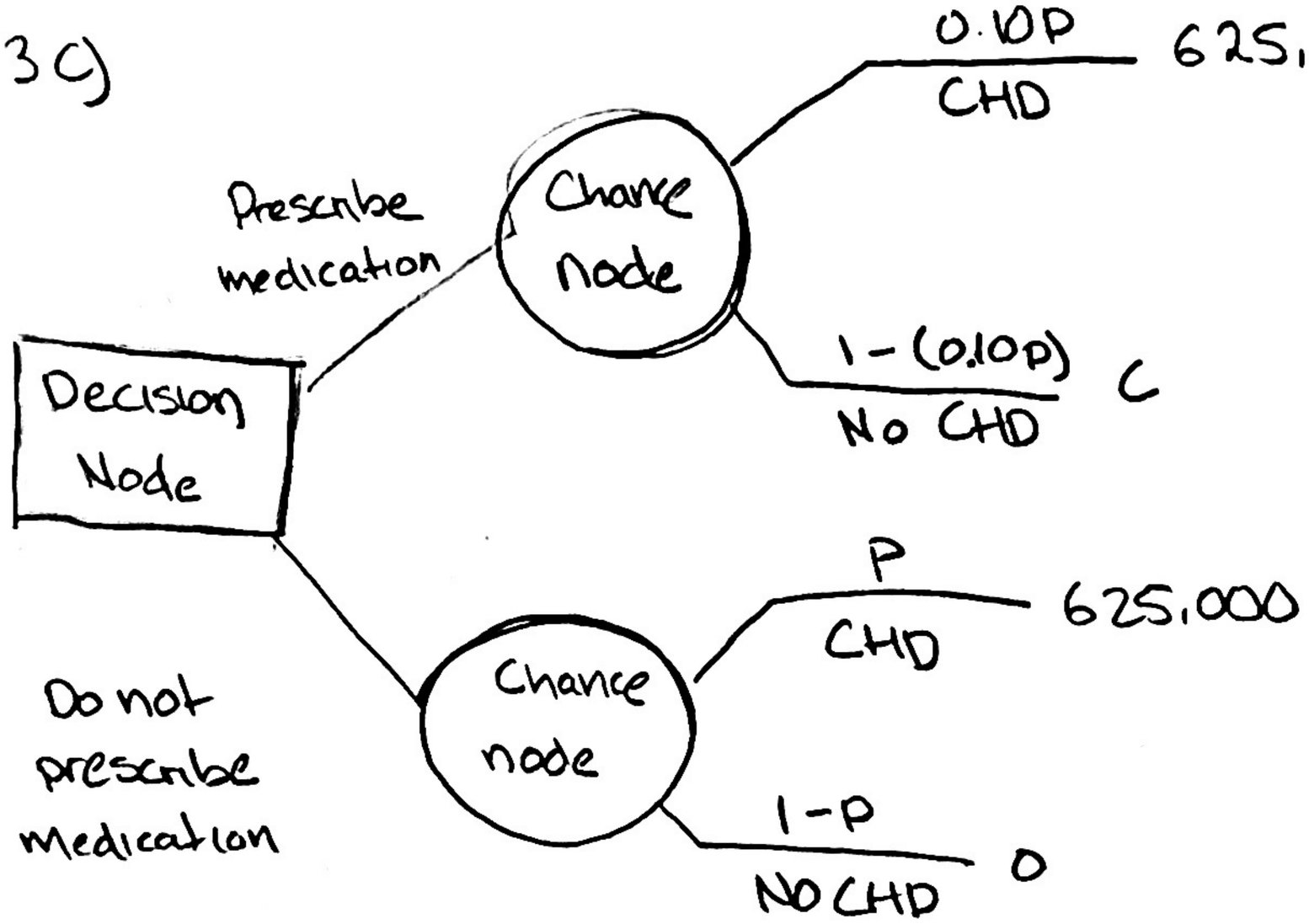
$$\frac{P(L_{FN} + L_{FP})}{(L_{FN} + L_{FP})} = \frac{L_{FP}}{(L_{FN} + L_{FP})}$$

$$\boxed{P = \frac{L_{FP}}{(L_{FN} + L_{FP}).}}$$

This tells us that $P \geq \bar{P}$ is the expected loss predicting $Y=0$ is greater than the expected loss of predicting $Y=1$.

The other $Y=0$ would be if $P \leq \bar{P}$.

3c)



$$\frac{\bar{P}}{10} \times (625{,}000 + C) + \left(1 - \frac{\bar{P}}{10}\right) \times C = 625000 \times \bar{P}$$

$\bar{P} = .11052937$ → Given from 3 Part (Pii°)

same value

$$\frac{0.11052937}{10} \times (625000 * C) + \left(1 - \frac{\bar{P}}{10}\right) \times C = 625000 \times \bar{P}$$

$$0.011052937 \times (625000 + C) + (0.988947) = 690808.5625$$

$$6908.085 + 0.011052937C + (0.988947C = 690808.5625$$

$$- 6908.085$$

$$- 6908.085$$

↳ I messed up on Math cause my value is to great., Messed on decimal place

$$0.011105 (625000 + C) + (0.988895C) = 625000 \times 0.1105$$

↳ $6{,}906 + 0.01105C + 0.988895C) = 69.06$

-6.906

-6.906

$1C = 61.4229$

62157=C
Friend got that*

$C = 61422.9$ is the money that insurance charges now

Imported Libarires

In [111]:
```
import numpy as np
import pandas as pd
import matplotlib as plt
import statsmodels.formula.api as smf
import sympy as sp
from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt
from scipy.stats import norm
```

Q1b

In [112]:
```
from scipy.stats import norm

# Parameters for Y = 0 (mean 5, variance 2)
mean_Y0 = 5
variance_Y0 = 2

# Parameters for Y = 1 (mean -3, variance 2)
mean_Y1 = -3
variance_Y1 = 2

# Calculate CDF values at a given threshold t for both scenarios
t = 0  # You can replace this with the actual threshold value
cdf_Y0 = norm.cdf(t, loc=mean_Y0, scale=variance_Y0)
cdf_Y1 = norm.cdf(t, loc=mean_Y1, scale=variance_Y1)

print("CDF for Y = 0:", cdf_Y0)
print("CDF for Y = 1:", cdf_Y1)
```

```
CDF for Y = 0: 0.006209665325776132
CDF for Y = 1: 0.9331927987311419
```

Q3. Part (i)

What is the fitted logistic regression model? Do not simply copy the results of your code, but instead state the equation used by the model to make predictions. Use all features from Table 1 to build your model.

In [96]:
```
framingham_train = pd.read_csv("framingham_train.csv")
framingham_train.head(3)
```

Out[96]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totC
0	1	59	Some college/vocational school	0	0	0	0	1	0	
1	0	43	High school/GED	1	15	1	0	1	0	
2	0	48	Some high school	0	0	0	0	1	0	

In [97]:
```
framingham_test = pd.read_csv("framingham_test.csv")
framingham_test.head(3)
```

Out[97]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol
0	0	48	High school/GED	1	25	0	0	0	0	250
1	0	58	High school/GED	1	20	0	0	1	0	231
2	0	37	High school/GED	1	20	0	0	0	0	164

In [98]:
```
logreg = smf.logit(formula = "TenYearCHD ~ male + age + education + "
                   " currentSmoker + cigsPerDay + BPMeds + prevalentStroke + "
                   " prevalentHyp + diabetes + totChol + sysBP + diaBP + BMI + "
                   " heartRate + glucose" , data = framingham_train).fit()

print(logreg.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.371879
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:             TenYearCHD   No. Observations:                 2560
Model:                          Logit   Df Residuals:                     2542
Method:                           MLE   Df Model:                           17
Date:                Tue, 03 Oct 2023   Pseudo R-squ.:                  0.1102
Time:                        19:07:02   Log-Likelihood:                -952.01
converged:                       True   LL-Null:                       -1069.9
Covariance Type:            nonrobust   LLR p-value:                 1.627e-40
==============================================================================================
====================
                                                 coef    std err          z      P>|z|
[0.025      0.975]
----------------------------------------------------------------------------------------------
--------------------
Intercept                                     -8.0533      0.855     -9.423      0.000
-9.728      -6.378
education[T.High school/GED]                   0.0041      0.221      0.018      0.985
-0.428       0.436
education[T.Some college/vocational school]    0.1267      0.242      0.524      0.601
-0.348       0.601
education[T.Some high school]                  0.1930      0.205      0.940      0.347
-0.209       0.595
male                                           0.5124      0.133      3.855      0.000
0.252       0.773
age                                            0.0637      0.008      7.830      0.000
0.048       0.080
currentSmoker                                  0.0608      0.191      0.318      0.750
-0.314       0.435
cigsPerDay                                     0.0190      0.008      2.507      0.012
0.004       0.034
BPMeds                                         0.1631      0.279      0.584      0.559
-0.385       0.711
prevalentStroke                                0.7908      0.570      1.387      0.166
-0.327       1.909
prevalentHyp                                   0.2797      0.166      1.682      0.093
-0.046       0.606
diabetes                                      -0.0086      0.378     -0.023      0.982
-0.750       0.733
totChol                                        0.0027      0.001      1.975      0.048
2.11e-05       0.005
sysBP                                          0.0133      0.005      2.832      0.005
0.004       0.022
diaBP                                         -0.0066      0.008     -0.810      0.418
-0.023       0.009
BMI                                            0.0150      0.016      0.964      0.335
-0.015       0.045
heartRate                                     -0.0056      0.005     -1.084      0.279
-0.016       0.005
glucose                                        0.0054      0.003      1.979      0.048
5.27e-05       0.011
==============================================================================================
====================
```

Answer Part(i)

It should be noted the intercept for our model is -8.0533, which tells us that it may be an "inadquent model, as negative intercept could be a sign that your logistic regression model is not well-specified for the data. It's possible that the chosen predictor variables are not capturing the underlying relationships effectively, or the model assumptions are not met." (logestic regression interpretor, Google). For most part all the other variables are positive, besides 3 but they have very small values, and high p-values, which tells us they aren't significant, this was the reason why I think our model is an inadquent model. It should be noted that some variables that have high coefficent values when predicting for 10 year CHD, are male,age,prevlentStroke and prevlentHyp. If I had to remake the model I would make the model based on that.

$$Pr(Y = 1|X) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_i)}}$$

P(Y = 1) = 1 / (1 + e^-(β_0 + β_1X_1 + β_2X_2 + ... + β_pX_p))

Here I am writing it in equation of each beta and their correspodning beta values

Logit(P(TenYearCHD)) = -8.0533β_0 + 0.0041β_1 * education[T.High school/GED] +

0.1267β_2 * education[T.Some college/vocational school] +

0.1930β_3 * education[T.Some high school] + 0.5124β_4 * male +

0.0637β_5 * age + 0.0608β_6 * currentSmoker + 0.0190β_7 * cigsPerDay +

0.1631β_8 * BPMeds + 0.7908β_9 * prevalentStroke + 0.2797β_{10} * prevalentHyp

-0.0086β_{11} * diabetes + 0.0027β_{12} * totChol + 0.0133β_{13} * sysBP -

0.0066β_{14} * diaBP + 0.0150β_{15} * BMI - 0.0056β_{16} * heartRate + 0.0054β_{17} * glucose

Q3. Part (ii)

What are the most important risk factors for 10-year CHD risk identified by the model? Pick one of these variables and describe its impact on a patient's predicted odds of developing CHD in the next 10 years.

a) Variables important: male, age, sysBP, glucose, totChol, prevalentHYP, are the most important risk factors for the 10-year CHD risk as they have the lowest p-value significanying. Male and Age have the lowest p-value of 0.000, making them very signficant and also having high coefficants. Then you have sysBP, with pvalue of 0.005, and then we have glucose with p-value with 0.048, and then totCHOld with p-value of 0.048. and PrevalentHYP p-value of 0.093.

b) one variable that I am picking its impact on patients predicted off of develiping CHD in next 10 years would be that of age (also because has low p-value and strong coeffient) (exp(0.0637) ≈ 1.065) which gives you ≈ 1.065, when keeping every other variable constant, and only adding one extra year of age. This tells us that increase odds of developing CHD in 10 years by 1 year is 1.065.

Q3. Part (iii)

Suppose that you wish to determine the optimal strategy for assigning which new patients receive the medication. Given your colleague's analysis of the costs and benefits associated with the recently approved treatment, identify a threshold value of p, call it $\bar{p}$, such that it is optimal to prescribe the medication to a patient if and only if their 10-year CHD risk exceeds $\bar{p}$

In [99]:
```
# Define the variable p ### If you want to look at my tree I have it in my google Doc. Showi
p = sp.symbols('p') ## I couldn't figure out how to do p_hat, otherwise this would look cool
equation = 1050000 * (0.1 * p) + 95000 * (1 - 0.1 * p) - 955000 * p
solutions = sp.solve(equation, p)
print("p_hat is :", solutions)
```

```
p_hat is : [0.110529377545084]
```

In [100]:
```
p_hat = 0.110529377545084
```

The p_hat value would be 0.11052. This is the most optimal point where you would prescribe the medication to the patient if they exceed 0.11052.

Q3 Part(iv)

Describe the test set performance of the logistic regression model, using the threshold identified in part (iii) to separate patients into those who are at high risk for CHD (risk exceeding the threshold $\bar{p}$) and those who are at low risk for CHD (risk below the threshold $\bar{p}$). State the model's accuracy, True Positive Rate (TPR), and False Positive Rate (FPR), and briefly

In [101]:
```
logreg_test_2 = smf.logit(formula = "TenYearCHD ~ male + age + education + "
                  " currentSmoker + cigsPerDay + BPMeds + prevalentStroke + "
                  " prevalentHyp + diabetes + totChol + sysBP + diaBP + BMI + "
                  " heartRate + glucose" , data = framingham_train).fit()

print(logreg_test_2.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.371879
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:             TenYearCHD   No. Observations:                 2560
Model:                          Logit   Df Residuals:                     2542
Method:                           MLE   Df Model:                           17
Date:                Tue, 03 Oct 2023   Pseudo R-squ.:                  0.1102
Time:                        19:07:14   Log-Likelihood:                -952.01
converged:                       True   LL-Null:                       -1069.9
Covariance Type:            nonrobust   LLR p-value:                 1.627e-40
==============================================================================================
                                                 coef    std err          z      P>|z|
[0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                                     -8.0533      0.855     -9.423      0.000
-9.728      -6.378
education[T.High school/GED]                   0.0041      0.221      0.018      0.985
-0.428       0.436
education[T.Some college/vocational school]    0.1267      0.242      0.524      0.601
-0.348       0.601
education[T.Some high school]                  0.1930      0.205      0.940      0.347
-0.209       0.595
male                                           0.5124      0.133      3.855      0.000
0.252       0.773
age                                            0.0637      0.008      7.830      0.000
0.048       0.080
currentSmoker                                  0.0608      0.191      0.318      0.750
-0.314       0.435
cigsPerDay                                     0.0190      0.008      2.507      0.012
0.004       0.034
BPMeds                                         0.1631      0.279      0.584      0.559
-0.385       0.711
prevalentStroke                                0.7908      0.570      1.387      0.166
-0.327       1.909
prevalentHyp                                   0.2797      0.166      1.682      0.093
-0.046       0.606
diabetes                                      -0.0086      0.378     -0.023      0.982
-0.750       0.733
totChol                                        0.0027      0.001      1.975      0.048
2.11e-05       0.005
sysBP                                          0.0133      0.005      2.832      0.005
0.004       0.022
diaBP                                         -0.0066      0.008     -0.810      0.418
-0.023       0.009
BMI                                            0.0150      0.016      0.964      0.335
-0.015       0.045
heartRate                                     -0.0056      0.005     -1.084      0.279
-0.016       0.005
glucose                                        0.0054      0.003      1.979      0.048
5.27e-05       0.011
==============================================================================================
```

In [102]:
```
y_prob = logreg.predict(framingham_test)
y_pred = pd.Series([1 if x > 0.110529377545084 else 0 for x in y_prob], index = y_prob.index
```

In [103]:
```
from sklearn.metrics import confusion_matrix

y_test = framingham_test['TenYearCHD']
cm = confusion_matrix(y_test, y_pred)
print ("Confusion Matrix : \n", cm)
```

```
Confusion Matrix :
 [[495 423]
 [ 31 149]]
```

In [104]:
```
## accuracy formula lab3 or lab4 # = (TN+TP)/total
accuracy = (495 + 149)/(495 + 423 + 31 + 149)
print('The accuracy is:', accuracy)

## accuracy formula lab3 or lab4 # = (TP)/(FN + TP)

TPR_logit = (149)/(149 + 31)
print('TPR is: ', TPR_logit)

FPR_logit = 423/(423 + 495) # FPR = FP/(FP+TN)
print('FPR is: ', FPR_logit)
```

```
The accuracy is: 0.5865209471766849
TPR is:  0.8277777777777777
FPR is:  0.46078431372549017
```

Briefly explained

Based on the model we can say that we predict accuracy for 0.58% patients. Our TPR is 83% which tells that group of our patients would have CHD in ten years, this means that we are correctly predicting our patients would get CHD in ten year time. And for FPR which is 46% which tells who wouldn't have CHD in ten years, so basically faslely predicted them developing CHD in 10 years.

Q3 Part (v)

If patients are prescribed the medication using the strategy implied by the model, use the test set data to provide an estimate(s) for the expected economic cost per patient. You should first report your estimate assuming that the CHD outcomes in the test set are not affected by the treatment decision. Is this assumption reasonable? You should then adjust your estimate in a way that takes into account the fact that the treatment decision impacts a patient's risk of developing CHD. (Hint: keep in mind that this dataset was collected before the option of prescribing the medication was even considered.)

[TN FP]

[FN TP]

In [61]:
```
#When the treatment decision doesn't affect the CHD outcome
#1050000 keep misstypign this value
total = (495 + 423 + 31 + 149)
expected_cost_per_patient = ((1050000 * 149) + (955000 * 31) + (95000 * 423)) / total
print('The expected cost per patient is:', expected_cost_per_patient)
```

```
The expected cost per patient is: 206047.35883424408
```

In [105]:
```
## Rebuilt the model Only TP and FP should change
#[TN FP + (TP * 0.9)] tn stays the same value
#[FN (TP * 0.9 )- TP] Fn stays the same value
np.round(149*0.9)
TN = 495
FP = 423 + np.round(149*0.9)
FN = 31
TP = (149) - np.round(149*0.9)
print(TN,FP)
print(FN,TP)
```

```
495 557.0
31 15.0
```

In [106]:
```
#Assuming 90% of CHD patients who take the treatment are cured
#adjusted_TP = np.round(149*0.9) ## --> you get 134
##print(adjusted_TP) ## this value is 134

adjusted_expected_cost_per_patient = ((955000 * 31) + (95000 * 557) + ( 1050000 * 15 )) / 10
print('The adjusted expected cost per patient is:', adjusted_expected_cost_per_patient)
```

```
The adjusted expected cost per patient is: 89499.08925318762
```

No the assumption isn't reasonable, because prescribing medications changes/impacts the patients chance of getting/developing CHD. Its hard to put in words but its like you have 90% chance of the medicine working, where they prevented the CHD but now you need to be able to inditify the people that have in the 10 years. Its simply because people that take medication treatment won't get CHD, so in that sense 90% of CHD partients would be cured in the next ten year if they take the medication. It should be noted that the oringla expcted cost was about 200k dolalrs while the ajusted is 85K, this something you would do if you need to save money, but also to see how much the insurance should cost so it evens out.

Q3 Part (vi)

Consider a simple baseline model that predicts none of the patients are at high risk for CHD and therefore does not recommend treatment for any of the patients. Describe the test set performance of the baseline model in terms of accuracy, TPR, and FPR, as well as expected economic cost per patient.

In [93]:
```
y_base = pd.Series([1 if x > 1 else 0 for x in y_prob], index = y_prob.index)
y_test = framingham_test['TenYearCHD']
cm = confusion_matrix(y_test, y_base)
cm
```

Out[93]:
```
array([[918,   0],
       [180,   0]], dtype=int64)
```

TN FP

FN TP

In [90]:
```
# One of doing it:
accuracy_base = 918/1098
print('The accuracy is:', accuracy_base)

TPR_logit_base = 0/(180) #TPR = TP/P = TP/(TP+FN)
print('TPR is:' , TPR_logit_base)

FPR_logit_base = 0/(918) #FPR = FP/N = FP/(FP+TN)
print('FPR is:', FPR_logit_base)
```

```
The accuracy is: 0.8360655737704918
TPR is: 0.0
FPR is: 0.0
```

In [113]:
```
## This code from lab03, adjusted for this model to get accuracy
default_false_test = np.sum(framingham_test['TenYearCHD'] == 0)
default_true_test = np.sum(framingham_test['TenYearCHD'] == 1)
ACC_test = default_false_test / (default_false_test + default_true_test)
print(ACC_test)

TPR = 0                          # TPR = TP/P = TP/(TP+FN)
FPR = 0                          # FPR = FP/N = FP/(FP+TN)
print("TPR:" ,TPR)
print("FPR:" ,FPR)
```

```
0.8360655737704918
TPR: 0
FPR: 0
```

In [87]:
```
expected_cost_base_line_model = (180 * 955000)/ 1098
#all patients didn't get prescribed
print("expected cost for base line model in $",expected_cost_base_line_model)
```

```
expected cost for base line model in $ 156557.37704918033
```

The new accuracy would be 0.836065. This showing that our baseline set acuracy is better than the logestic regression where we had threshold of phat, however we should note that both TPR and FPR are 0. Basically we are assuming in this model that no one needs to the treatment and it does help save the expected cost but I wouldn't implement this model, as its an ethical cocern (part D). The expected cost is 156557.37704918033 dollars with baseline which is lower than our threshold model.

Q3 Part (vii)

Use an example to explain how to use the model in a real clinical setting. Suppose a new patient arrives, and the physician accesses the patient's electronic medical records and retrieves the following about the patient:

Female, age 39, GED education, currently a smoker with an average of 6 cigarettes per day. Currently not on blood pressure medication, has not had stroke and is not hypertensive. Currently diagnosed with diabetes; total Cholesterol at 230. Systolic/diastolic blood pressure at 110/50, BMI at 28, heart rate at 72, glucose level at 80.

What is the predicted probability that this patient will experience CHD in the next ten years? Based on your calculated $\bar{p}$ threshold from part (iii) from the decision tree, should the physician prescribe the preventive medication for this patient?

In [120]:
```
part_vii = pd.DataFrame(data = {'education' : ['High school/GED'], 'male' : [0], 'age' :[39],
                                'currentSmoker' : [1], 'cigsPerDay' : [6],'BPMeds' : [0],
                                'prevalentStroke' : [0],'prevalentHyp' : [0], 'diabetes' : [
                                'totChol': [230], 'sysBP' : [110], 'diaBP' : [50],'BMI' : [28
                                'heartRate' : [72], 'glucose' : [80]})

logreg.predict(part_vii)
```

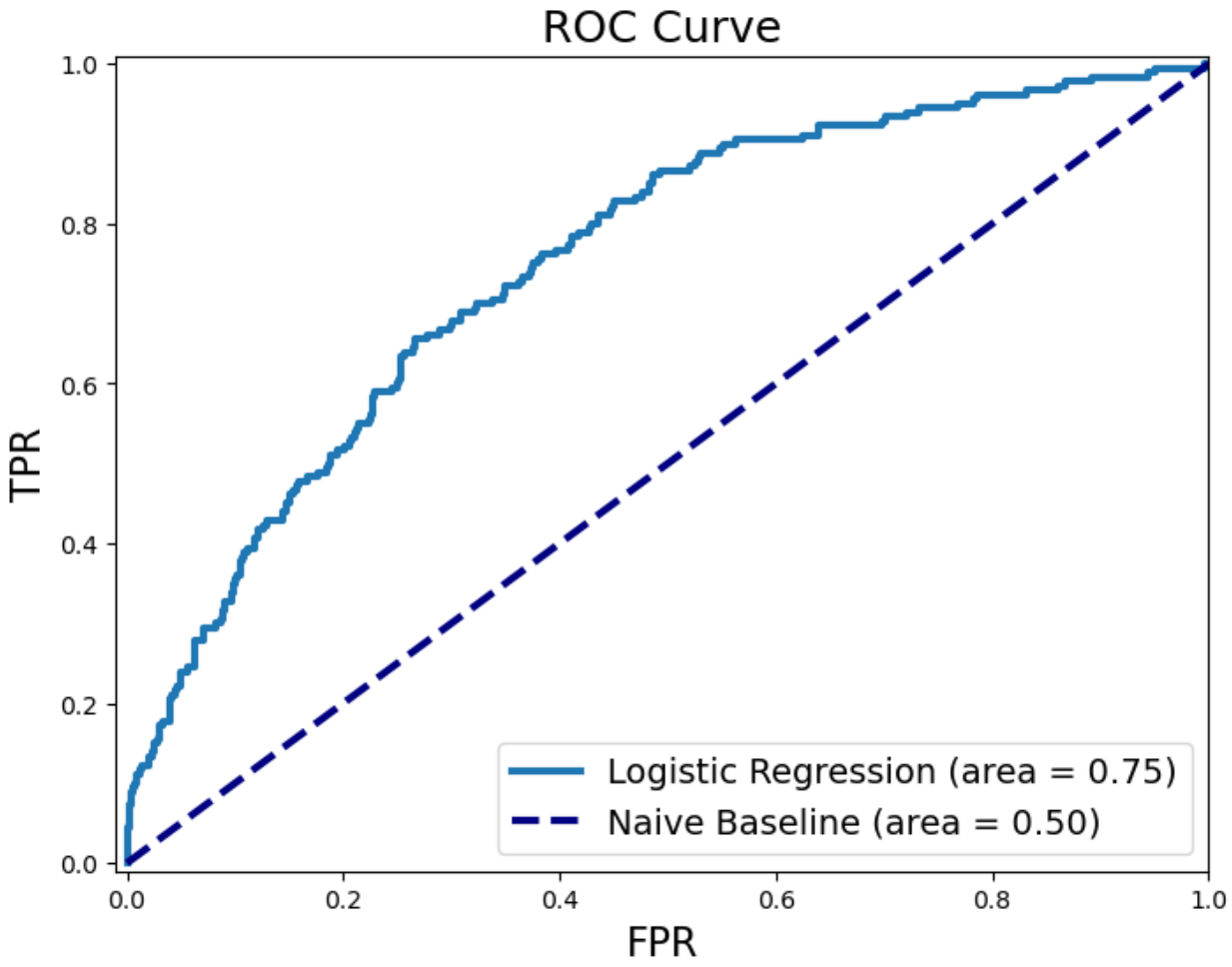
Out[120]:
```
0    0.039055
dtype: float64
```

The probability we get is 0.039 based on our fitted logistic regression model. Since this value is less than our threshold value of phat, we wouldn't prescribe her medication.

Q3 Part B

(15 points) Show the ROC curve for your logistic regression model on the test set and describe how this curve may be helpful to decision-makers looking to further study the medication you have considered so far in this homework as well as other possible medications for preventing CHD. Describe one interesting observation implied by examining the ROC curve. What is the area under the curve (AUC) for your model in the test set?

In [109]:
```
## Code from lab or online I forgot.
fpr, tpr, _ = roc_curve(y_test, y_prob)
roc_auc = auc(fpr, tpr)
plt.figure(figsize=(8, 6))
plt.title('ROC Curve', fontsize=18)
plt.xlabel('FPR', fontsize=16)
plt.ylabel('TPR', fontsize=16)
plt.xlim([-0.01, 1.00])
plt.ylim([-0.01, 1.01])
plt.plot(fpr, tpr, lw=3, label='Logistic Regression (area = {:0.2f})'.format(roc_auc))
plt.plot([0, 1], [0, 1], color='navy', lw=3, linestyle='--', label='Naive Baseline (area = 0
plt.legend(loc='lower right', fontsize=14)
plt.show()
```



Our logistic regression AUC is 0.75. Since the AUC value is at 0.75, it should tell us that our model can classify correctly at a given threshold if picked correctly, basically are able to determine if someone will get CHD in 10 years or won't. like this tells us that it iwll predict 75% correctly for true positives and 25% for false positive cases.

Q3 Part C

Rather than explicitly dictating which patients should receive the medication, let us consider letting patients decide for themselves. Suppose that if a patient has health insurance, the treatment costs for CHD (including the proposed medication) will be covered by their insurance company. However, a patient will still incur an equivalent cost of $625,000 for decreased quality of life if they develop CHD. Disregarding other factors such as side effects

of the medication, if there were no insurance co-payment then it should be clear that every patient would always choose to receive the medication because it would cost them nothing and it would lower their risk of CHD. Thus let us consider setting a co-payment value C – the amount that each patient would have to pay in order to receive the medication – in order to provide an incentive for some patients to forego the treatment while others would choose to receive the treatment. What value of C should the insurance company charge as a co-payment for the medication in order that the patients would “self select” in a manner that is consistent with the previously examined “optimal strategy” discussed in part (a) above?

I did this on the PDF which I am attaching

Q3 Part D

Are there any aspects of the analysis performed thus far that raise ethical concerns? If so, suggest at least one way that this analysis could be changed to address such concerns

Answer

Ethical concerns arise in the context of education. What if, in the future, insurance companies begin to discriminate against individuals with low levels of education or those who have had a stroke? This discrimination could result in insurance companies refusing to accept them as clients due to the perceived higher liability based on expected costs. We conduct analyses on expected losses, and the primary motivation behind these analyses is to save money while ensuring that individuals do not receive unnecessary medication.

For example, I am aware that insurance companies often charge higher premiums or even deny coverage to individuals who are prediabetic. This situation could potentially lead to a similar outcome where insurance companies are unwilling to assist individuals with a high probability of developing coronary heart disease within the next 10 years.