ASSIGNMENT 2

ENGIN 242 - Applications in Data Analytics

Aditya Peshin
Enrollment Number: 3035280249

Q1. Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2= undergrad GPA, and Y= receive an A. We fit a logistic regression and produce estimated coefficient, β0=−6, β1=0.05, β2=1.

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

(b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

Ans)

Part (a)

The probability that the student gets an A can be found by substituting his grades in the logistic regression model.

$$P(Y = 1 \mid X) = \frac{1}{1 - e^{\beta^T X}}$$

$$= \frac{1}{1 - e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}$$

$$= \frac{1}{1 - e^{-6 + 0.05 * 40 + 1 * 3.5}}$$

$$= 0.3775$$

Part (b)

To increase the probability of the student getting an A to 50%, we can calculate the new log odds required.

Log (P(X) / 1 - P(X)) = β0 + β1 * X1 + β2 * X2

LHS:

New Log_e(odds) = ln (0.5 / (1 - 0.5)) = 0

RHS:

New equation = β0 + β1 * $X1_{new}$ + β2 * X2 = -6 + 0.05 * $X1_{new}$ + 1 * 3.5

Equating the two we get:

0 = 0.05 * $X1_{new}$ - 2.5

$X1_{new}$ = 2.5 / 0.05 = **50 hours**

This is the minimum number of study hours required to have probability of getting an A to 50%.

Q2. Suppose that we wish to predict whether a given stock will issue a dividend this year (“Yes” or “No”) based onX, last year’s percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was X= 10, while the mean for those that didn’t was X= 0. In addition, the variance of X for these two sets of companies was $\sigma^2 = 36$. Finally, 80 % of companies issued dividends.
Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was X = 4 last year.

Ans.

(On the next page, PTO ->)

Q2. We need to apply Linear Discriminant analysis for $p=1$ to find the probability that the company will issue dividends.

Step 1: Developing the equation of probability.

We know, from Bayes theorem

$$P_K(x) = \frac{\pi_K \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{1}{2\sigma^2}(x-\mu_K)^2\right)}}{\sum_{l=1}^{K} \pi_l \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{1}{2\sigma^2}(x-\mu_K)^2\right)}} \quad -(1)$$

In this expression, we know that the value of the denominator is $P(x)$, which is independent of 'K' and can be treated as a constant. Combining the constants together, we get:

$$P_K(x) = C \cdot \pi_K \cdot e^{-\left(\frac{1}{2\sigma^2}(x-\mu_K)^2\right)}$$

Taking $\log_e()$ on both sides, we get:

$$\ln(P_K(x)) = \ln(C) + \ln(\pi_K) - \frac{1}{2\sigma^2}(x-\mu_K)^2$$

$$= \ln(C) + \ln(\pi_K) - \frac{x^2}{2\sigma^2} + x \cdot \frac{\mu_K}{\sigma^2} - \frac{\mu_K^2}{2\sigma^2}$$

$$= \ln(c) + \ln(\pi_k) - \frac{x^2}{2\sigma^2} + x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

$$= \underbrace{\ln(c) - \frac{x^2}{2\sigma^2}}_{\text{Constants w.r.t 'k'}} + \ln(\pi_k) + x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

$$\ln(P_k(x)) = c' + \underbrace{\ln(\pi_k) + x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}}_{\text{Expression for } \delta_k(x)} \quad - (2)$$

which we maximize with respect to different 'k' to see which class the obs. belongs to.

This expression is important as we cannot directly evaluate $\delta_k(x)$ to find $P_k(x)$, we must also calculate the c' term to get the true probability.

Calculating c':

$$c' = \frac{-x^2}{2\sigma^2} + \ln\left(\frac{\frac{1}{\sqrt{2\pi}\sigma}}{\pi_0 \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\left(\frac{1}{2\sigma^2}(x-\mu_0)^2\right)} + \pi_1 \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\left(\frac{1}{2\sigma^2}(x-\mu_1)^2\right)}}\right)$$

where $x = 4$, $\pi_0 = 0.2$, $\pi_1 = 0.8$
$\sigma^2 = 36$, $\sigma = 6$, $\mu_0 = 0$, $\mu_1 = 10$

(PTO →)

Substituting the values, we get:

$$C' = \frac{-(4)^2}{2(36)} + \ln\left(\frac{\frac{1}{\sqrt{2\pi}\times 6}}{\frac{1}{\sqrt{2\pi}\times 6}\left(0.2\times e^{-\left(\frac{1}{2\times 36}(4-0)^2\right)} + 0.8\times e^{-\left(\frac{1}{2\times 36}(4-10)^2\right)}\right)}\right)$$

Removing common terms

$$= \frac{-2}{9} + \ln\left(\frac{1}{0.2\times e^{-(4^2/72)} + 0.8\times e^{-(6^2/72)}}\right)$$

$$= \frac{-2}{9} + 0.437928$$

$$C' = 0.21570 \qquad - (3)$$

Evaluating $S_1(x)$

$$S_1(x) = \ln(0.8) + 4\times\frac{10}{36} - \frac{10^2}{2(36)}$$

$$S_1(x) = -0.50092 \qquad - (4)$$

Thus, substituting (3), (4) in (2), we get

$$\ln(P_1(x)) = 0.21570 - 0.500902 = -0.28522$$

$$\Rightarrow P_1(x) = 0.75184 \simeq 75.2\ \%$$

This is the psobabilety with which the company will issue a dividend. (for $X = 4$)

Q3. **Framingham Heart Study**

Part a)

i) The logistic regression model is (for seed set as 144):

$$P(Y=1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\beta_n X_n)}}$$

Where β_0 = -8.495 is the intercept
β_1 = 0.429 is the coefficient for male
β_2 = 0.064 is the coefficient for age
β_3 = -0.120 is the coefficient for education (High School/GED)
β_4 = -0.077 is the coefficient for education (Some College/ vocational school)
β_5 = 0.064 is the coefficient for education (Some High School)
β_6 = 0.103 is the coefficient for currentsmoker
β_7 = 0.017 is the coefficient for cigsperday
β_8 = -0.107 is the coefficient for BPMeds
β_9 = 0.936 is the coefficient for prevelantstroke
β_{10} = 0.244 is the coefficient for prevelantHyp
B_{11} = -0.005 is the coefficient for diabetes
β_{12} = 0.001 is the coefficient for totChol
β_{13} = 0.016 is the coefficient for sysBP
β_{14} = -0.007 is the coefficient for diaBP
β_{15} = 0.004 is the coefficient for BMI
β_{16} = $-8.3E^{-7}$ is the coefficient for heartRate
β_{17} = 0.008 is the coefficient for glucose

The model uses this equation to come up with a value for the probability that a person with health metrics ‘X’ will develop Coronary Heart Disease.

ii) The most important factors, according to the model are:

- Age
- SysBP

Other important factors identified, are:

- Male
- cigsPerDay
- Glucose

Let us consider one of the important factors, (Age). With every year increase in the age of the person, the log (predicted odds of them developing Coronary Heart Disease in the next ten years) increases by 0.064

iii) The ideal switching point is when the expected costs associated with taking the medicine are equal to the expected costs of not taking the medicine.

Expected costs of not taking the medicine are:

E(no meds) = 500000 * p + 0 * (1-p)

E(take meds) = 560000 * p/4 + 60000 * (1 - p/4)

Equating the two, we get:

500000 * p = 560000 * p/4 + 60000 * (1 - p/4)

500000 * (p - p/4) = 60000

3p/4 = 6/50

p = 6/50 * 4/3 = 4/25 = 0.16

This is the ideal value of p for which we should consider giving the patients medicine

iv) The Confusion Matrix created is as follows

	Prediction		
Reality		Doesn't develop	Develops
	Doesn't develop	637 (TN)	293 (FP)
	Develops	56 (FN)	111 (TP)

The accuracy of the model is = (637 + 111) / (637 + 293 + 111 + 56) = 0.6818

- This metric refers to the model's ability to make a correct prediction. In other words, the model predicts the right outcome (either developing the disease or not) 68.18% of the time, over the entire test dataset.

The True Positive Rate (TPR) is = 111 / (111+56) = 0.6646

- This metric refers to the fraction of people that the model correctly identifies as developing the disease, out of all the people who actually end up developing the disease. In other words, the model identifies 66.46% of the people who end up with the disease after 10 years.

The False Positive Rate (FPR) is = 293 / (637 + 293) = 0.3150

- This metric refers to the fraction of the people the model incorrectly assumes will get the disease, out of the people who don't end up getting the disease. In other words, 31.50% of the people who will not end up being affected by the disease are incorrectly predicted to develop the disease.

v) Calculating the estimated economic cost per patient. The loss matrix is

	Prediction		
Reality		Doesn't develop	Develops
	Doesn't develop	0 (TN)	60000 (FP)
	Develops	500000 (FN)	560000 (TP)

Total cost incurred is,

T.C = 0 * (True negatives) + 50000 * (False Negatives) + 60000 (False Positives)
+ 560000 * (True Positives)
= 0 + 500000 (56) + 60000 (293) + 560000(111)
= $107,740,000

Cost per patient = 107740000 / 1097 = $98,213.30

The earlier assumption is not reasonable, as it does not take into consideration the finding that taking the medicine reduces the possibility of contracting the disease.
Now, if we consider that taking the medicine reduces the probability of developing the disease to p/4, then only 1/4th of the people who receive the medicine will end up developing the disease.

- TP_{new} = 111 / 4 = 27.75 ≃ 28
- FP_{new} = 293 + (111 - 28) = 376

The new confusion matrix becomes:

	Prediction		
Reality		Doesn't develop	Develops
	Doesn't develop	637 (TN)	376 (FP)
	Develops	56 (FN)	28 (TP)

With the new total cost becoming,

TC = 0 * (True negatives) + 50000 * (False Negatives) + 60000 (False Positives)
+ 560000 * (True Positives)
= 0 + 500000 (56) + 60000 (376) + 560000(28)
= $66,240,000

Expected cost per patient = 66240000 / 1097 = $60,382.86

vi) Considering the simple baseline model, which recommends against medication for all patients, we have the confusion matrix as: (when applied to the test dataset)

	Prediction		
Reality		Doesn't develop	Develops
	Doesn't develop	930 (TN)	- (FP)
	Develops	167 (FN)	- (TP)

The accuracy of this model is = 930/1097 = 84.77%
The True Positive Rate is = 0 / 930 = 0%
The False Positive Rate is = 0 / 167 = 0%

The total expected economic cost = 0 * 930 + 500000 *167 = $ 83,500,000
The expected economic cost per patient is = 83500000 / 1097 = $76,116.68

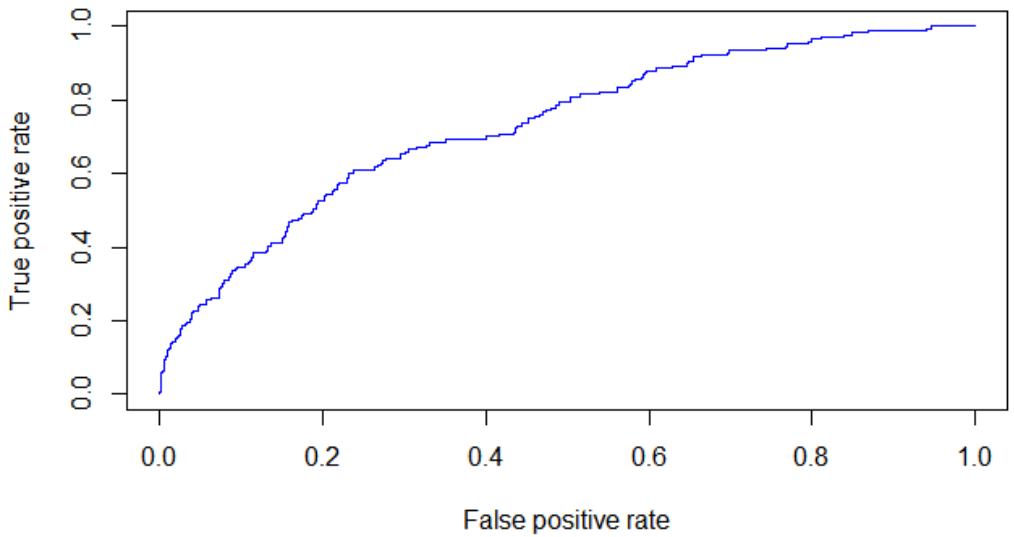
Although this model has a higher accuracy than the model we use, it accrues a higher cost per patient than our earlier model. If the objective is to minimize the prediction error, then the baseline model is better, however, if the intent is to minimize patient costs then the previous model is better.

vii) The new patient has the following characteristics:

Female, age 51, college education, currently a smoker with an average of 20 cigarettes per day. Not on blood pressure medication, has not had a stroke, but has hypertension. Not diagnosed with diabetes; total Cholesterol at 220. Systolic/diastolic blood pressure at 140/100, BMI at 31, heart rate at 59, glucose level at 78.

Feeding this data into the model, we arrive at a probability of 0.156 that this person will develop CHD in the next ten years. As our value for p is 0.16 and the probability of her developing the disease is lower than the threshold, the physician will not recommend medications to this lady.

b) Plotting the ROC curve of the model, we get:



The ROC curve is a means of understanding how varying the threshold affects the True Positive Rate and the False Positive Rate.

- In simpler terms, this graph captures the tradeoff between setting a very strict and a very low threshold for the prediction. If we try to minimize the number of people incorrectly classified as prone to develop the disease (strict threshold), we compromise on the number of people incorrectly classified as safe from the disease. Also, if we try to minimize the number of people which the algorithm incorrectly classifies as safe from the disease (low threshold), then we compromise on the number of people incorrectly classified as prone to the disease.

The ROC curve is characteristic to a particular model, and if it hugs the top left of this graph, then we can say that the model accurately describes the situation.

- An interesting observation is that the ROC curve does not hug the top left corner, which means that either the logistic regression model is not good enough to describe the data, or that there are some unknown and unaccounted features that influence the probability of getting CHD ten years from now. This means that we should try to improve the quality of the model, so that we can get better predictions.

The area under the curve (AUC) is found to be 0.7335

c)
Now, in this new case where the patient decides whether he should take the medication, we have to determine the co-payment cost (C), so as to motivate the patients to ‘self-select’ in the same optimal strategy identified earlier.

Now, the patients will choose the option where the expected cost to them is lesser. Expected cost of getting treatment(or forgoing it) can be calculated as,
EC(treatment) = (300k + C) * p/4 + C * (1 - p/4)
EC(no treatment) = (300k) * p + 0 * (1 - p)

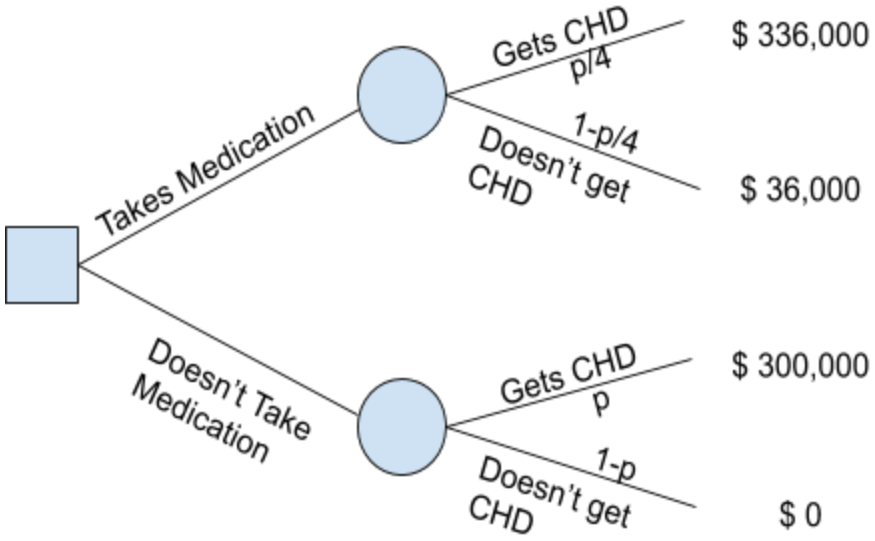
Now, equating the two to get decision point, we have

300k * p/4 + Cp/4 + C - Cp/4 = 300k * p
300k * p/4 + C = 300k * p
C = 300k * p - 75k * p
C = 225k * p

If we want users to select in a manner that is consistent with the earlier case, we must set the value of p to 0.16 to achieve the same decision boundary.

C = 225k * 0.16 = 36k

So the new decision tree is:



This will result in the same 'optimal strategy' being opted for by the customer.

d)
The ethical concerns I have regarding this analysis are:

- The quantification of the quality of human life has been done to 'objective standards', however each person values their lives more than the objective value they add to society. How do we capture this mathematically?
- Keeping the value of human life constant, the value of p only depends on the cost of the medicine. Different medicines will yield different threshold criteria, does this mean that the main factor influencing whether the patient receives medication or not is the cost of the medicine? Ten years is a long time, it may be possible that the costs of the medicine go down by then. How do we factor this into our analysis? (By medicine I mean treatment/medicine)

In order to better capture the range of values that the optima can take, we can perform a sensitivity analysis and identify how changing the value of the "cost of medicine" and "economic equivalent of human life" leads to different optima.

Another ethical question regarding the analysis is:

- In both cases (no insurance vs insurance), by deciding the value of 'p' or 'C', we are effectively taking the choice out of the patient's hand and asking them to effectively choose the option we want them to choose. Does the patient have no say in the decision making process? Manipulating human behaviour in order to reach organizational goals (of the insurance companies) is something that may not be ethically acceptable.

Q3. Code:

```
# Homework 2 Q 3

# loading required libraries
library(dplyr)
library(ggplot2)
library(caTools)
library(GGally)
library(ROCR)

# reading in the file
framingham <- read.csv("framingham.csv")
str(framingham)

# setting the factor variables as factors
framingham$TenYearCHD <- as.factor(framingham$TenYearCHD)
framingham$male <- as.factor(framingham$male)
framingham$currentSmoker <- as.factor(framingham$currentSmoker)
framingham$BPMeds <- as.factor(framingham$BPMeds)
framingham$prevalentStroke <- as.factor(framingham$prevalentStroke)
framingham$prevalentHyp <- as.factor(framingham$prevalentHyp)
framingham$diabetes <- as.factor(framingham$diabetes)
framingham$education <- as.factor(framingham$education)

# splitting the data into training and testing
set.seed(144)
split <- sample.split(framingham$TenYearCHD, 0.7)

framingham.train <- filter(framingham, split == TRUE)
framingham.test <- filter(framingham, split == FALSE)

# using logistic regression to make model
mod1 <- glm(TenYearCHD ~ . , data = framingham.train, family = binomial)
summary(mod1)

# making a prediction on the test data
pred_CHD = predict(mod1 , newdata = framingham.test, type = "response")

# confusion matrix for training data
table (framingham.test$TenYearCHD, pred_CHD>= 0.16)

# confusion matrix for baseline model
```

```
table (framingham.test$TenYearCHD, pred_CHD>= 1)

# Probability that the new patient has the disease 10 years from now
new_patient <- 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(mod1, new_patient, type = "response")

# Plotting a ROC curve for the model
pred <- prediction(pred_CHD, framingham.test$TenYearCHD)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf, col = "Blue")
auc <- performance(pred, measure = "auc")
auc <- auc@y.values[[1]]
auc     # to print value of AUC
```