HAKAN SONMEZ
MBA/LGO 2018 - Section-B

15.071 Analytics Edge - Homework Assignment # 2

Problem-2: Framingham Heart Study
Part-a: Memo To Policymakers

To: Department of Health and Human Services
Date: March 3, 2017
Re: Predicting CHD development at early stage and decreasing related healthcare costs

---

In this report we would like to provide you our predictive analysis on the Framingham Heart Study results and our recommendations on how to improve the early detection of Coronary Heart Disease (CHD) by classifying the patients into low-risk and high-risk groups and then, by using this information, decreasing healthcare costs by prescribing a special medication to those patients who are in high-risk group for CHD.

After using our analytics tools, we identified the following 5 variables to be the most significant and impactful in predicting a patient who would suffer CHD within 10 years of initial examination:

1. **Gender (Being male)**
2. **Age**
3. **Cigarettes consumed per day (cigsPerDay)**
4. **Systolic blood pressure (sysBP)**
5. **Blood glucose Level (Glucose)**

We observe that being male and having higher values for other 4 parameters increase the probability of developing CHD in next 10 years. We use notation P(Y=1) to express the probability of a patient developing CHD in 10 years. This probability can be estimated using the logistics regression formula below based on our model:

$$P(Y = 1) = \frac{1}{1+e^{-[-8.6298+0.4145(Male)+0.067(Age)+0.0162\ (cigsPerDay)+0.0173(sysBP)+0.0078(Glucose)]}}$$

This formula excludes variables that do have significant impact on the prediction according to our findings. We can also express the impact of these parameters on the estimated probability using odd ratios for each parameter.
For example, the impact of age variable can be expressed as follows: Keeping all other variables the same, for every increase of 1 year of age, the odds of developing CHD in 10 years is multiplied by a factor of 1.067 (which equals to $e^{0.067}$). For example, by using this odds ratio, we can tell that if a patient is 10 years older than another patient with all other conditions being the same, then older patient is 1.95 ($=e^{(0.067x10)}$) times more likely to experience CHD in the next ten years.

We can define an optimal strategy by assessing which patients should receive the special medication by assessing the potential cost savings from medication and the costs of not using the medication. After our calculations (included in Appendix), we find a threshold value of **p=0.0734** which indicates that if the probability of a patient having CHD in next ten years is higher than 0.734, then that patient should be

given the preventive medication for cost savings. The equation we use to obtain **p=0.0374** is included in the appendix of this report.

Using this threshold value of 0.0734 in conjunction with applying our prediction model on a test group of 1207 patients in the Framingham Study, we found the following predication quality metrics:
Accuracy = 0.42 (42%) Number of Correct Prediction / Total predictions
TPR =0.896 (89.6%) Correctly predicted positives/ All positives
FPR =0.665 (66.5%) Falsely predicted positives / All negatives
Note that: Calculations leading to these results are included in the appendix.

Also by applying our prediction model on the test group, we estimated the total expected economic costs associated with CHD and medication will be limited to **$17.72M** which is much lower than the expected economic cost using a simple baseline model of assuming none of the patients will develop CHD and thus no patient should be given the preventive medication.

We also developed a baseline model to compare the potential cost benefits our prediction model. Our baseline model predicts none of the patients are at the risk for CHD and thus no patient is given the preventive medication. We find the following quality metrics for the baseline prediction model.

Accuracy= 0.8475 (84.7%)
TPR and FPR are both 0 for baseline model, since our baseline model assumes no patient is at risk for CHD disease (i.e. no positive outcomes predicted).

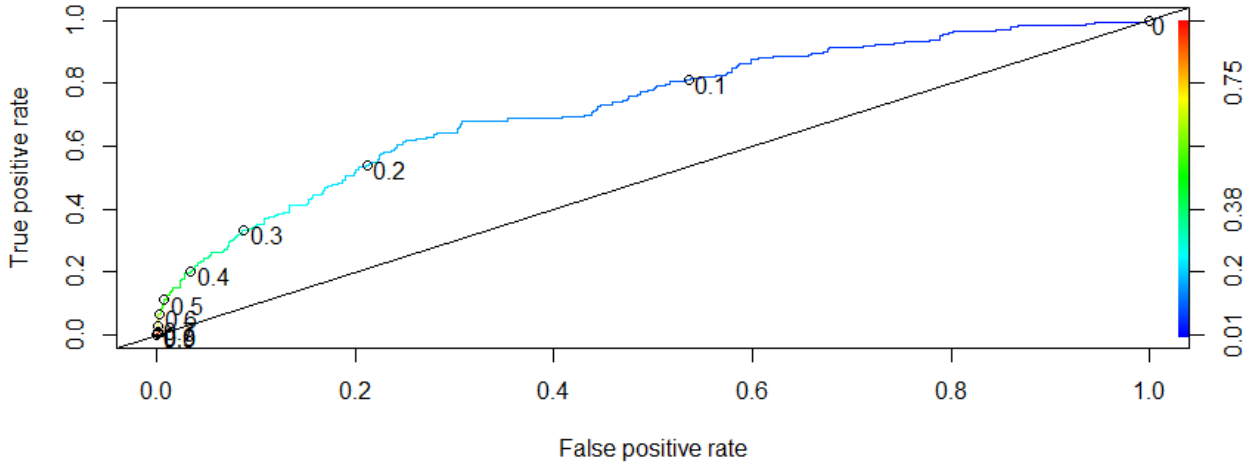
Our calculations after applying this baseline prediction model on our test group, lead to a total expected cost of **$36.8M** the baseline model.
Comparing the expected cost associated with our advanced prediction model and the expected costs associated with baseline model, we observe that our prediction model has significant savings over the baseline model. Savings through the are estimated to be around **$19.07M** (= $36.8M - $17.72M). Thus we strongly recommend using our advanced prediction model over the baseline model.

One other area we looked into was determining an optimal co-payment value for the medication in order that right patients will choose to be treated using the medication and thus control the treatment costs covered by the insurance company. Obviously the co-payment value will **be lower than $8300** which is the full price of medication. Assuming a patient would be willing to pay as much as the expected value of quality-of-life savings he would have by getting the medication and all the treatment costs are covered by medical insurance, we calculated the optimal copayment value to be around **$4148** for the right patients to choose to be treated using the medication. (Calculation is in the appendix.)

Finally, to illustrate a typical application of our model, we assume a male 57-year old patient with a specific personal set of information, (the detailed information is included in the appendix), visits a physician. Our electronic systems will inform the physician that specific patient has a probability of **0.297 (29.7%)** developing CHD in next 10 years (which is higher than our initially calculated threshold value of 0.0734) and thus the patient should be prescribed the preventive medication.

-End of memo-

Part-b: ROC Curve



The Area Under Curve (AUC) = **0.7277**

```
>  as.numeric(performance(ROCRpred, "auc")@y.values)
[1] 0.7277084
```

ROC curve gives a visual representation of how successful our model is in correctly predicting future CHD cases in terms of highest TPR achievable for a given FPR. If the Area under curve is higher and closer to 1, it means our model is doing a better job. Another way to interpret the ROC curve is that, ROC curve gives us highest TPR value achieved for a given (accepted) FPR value. For example from the curve above, we see that, if we set p value to have FPR around 0.2, we can achieve TPR close 0.6. Or for a given FPR=0.66, we can achieve TPR around 0.9.
By comparing the ROC curve with the straight baseline crossing the graph, we can see how better our model does compared to the baseline model. ROC curve being higher above the the baseline curve means the prediction model is more effective.

Part-c: Ethical Concerns

One aspect that raises ethical concerns to me is the determination of a copayment value by the insurance company. If the company does not fully inform the patients about the risks they face by not getting the preventive medicine and just focuses on reducing the costs for the company, that raises ethical concerns. One way to address this concern is the fact that, the company should be very transparent with the patients and should share all the prediction model info with the patients so they can decide whether to get and pay

the copayment for the preventive medicine or not. The analysis can be done with assumption that patients make decision by knowing all the potential risks and their probabilities.

Part-d: Additional variables to gather to improve the model

I would suggest to gather those two additional pieces of data to improve the prediction model:

- Current marital status: married or single: Marital status is a significant lifestyle factor that may potentially impact the CHD prediction.
- Alcohol consumption (weekly): Just like cigarette consumption per day is a significant variable, I would expect alcohol consumption (per week) to be a significant life-style related variable

Part-e.: Logistic regression model with 3 variables

I have chosen these 3 variables Male (Gender), Age, CigsPerDay to create my new model. When I apply the model on the test data, I obtain the following prediction quality metrics:

Accuracy = 0.346 (34.6%)
TPR =0.913 (91.3%)
FPR =0.755 (77.5%)

The formula for the new logistic regression model is below:

$$P(Y = 1) = \frac{1}{1+e^{-[-6.559+0.298(Male)+0.087(Age)+0.0185\ (cigsPerDay)]}}$$

As quality metrics collected on the test group indicate, this model does worse that our initial model as expected with a lower Accuracy value and a higher FPR value. However it still does better than the baseline model in reducing the expected costs.

The confusion metrics obtained by this model and the R code to calculate the quality parameters is copied below:

```
> PredictTest2=predict(model2, newdata=test, type="response")
> table(test$TenYearCHD, PredictTest2 > 0.0734)

    FALSE TRUE
  0   250  773
  1    16  168
> table(test$TenYearCHD, PredictTest2 > 0.0734)/nrow(test)

        FALSE       TRUE
  0 0.20712510 0.64043082
  1 0.01325601 0.13918807
```

```
> New_Accuracy=0.20712510 + 0.13918807
> New_Accuracy
[1] 0.3463132

> New_TPR=168/(168+16)
> New_TPR
[1] 0.9130435

> New_FPR=773/(773+250)
> New_FPR
[1] 0.7556207
```

APPENDIX

- **Calculation of p threshold probability:**

We solve the following equation to derive the value of p:

(p/2.3)*($208300)+(1-p/2.3)*($8300) = p*($200000)
p= 0.0734231

- **Calculation of Quality Metrics for Logistic Regression Prediction Model on Test Group:**

Accuracy =(342+165)/(342+165+19+681) = 0.4200497
TPR=165/(165+19) = 0.8967391
FPR=681/(681+342)= 0.6656891

```
> table(test$TenYearCHD,PredictTest > 0.0734)

             (Prediction)
              FALSE TRUE
  Reality: 0:   342  681
           1:    19  165
```

- **Calculation of Expected Economic Costs using Prediction model on Test Group:**

Total Expected Cost =(342+19)*(0.0734)*($200,000)+ (342+19)*(1-0.0734)*(0$)
(681+165)*(0.0734/2.3)*($208300)+(681+165)*(1-0.074/2.3)*($8300) = **$17.72M**

Or alternatively:

Total Expected Cost = (681+165)*($8300)+(681+165)*(0.0734/2.3)($200000) +
(342+19)*(0.0734)*($200000) + (342+19)*(1-0.0734)*($0) = **$17.72M**

- **Calculation of Quality Metrics for Baseline Prediction Model on Test Group:**

Accuracy = 1023/(184+1023) = 0.8475559
TPR= 0/(184+0) = 0
FPR= 0/(2013+0)= 0

Baseline Model Confusion Matrix:

```
    FALSE TRUE
  0   1023  0
  1    184  0
```

- **Calculation of Expected Economic Costs using Baseline Model on Test Group:**

Total Expected Cost = 1023*(0$) + 184*($200,000) = $36.8M

• Calculation of copayment value for the patient (expected quality-of-life savings for the patient obtained by using preventive medicine):

(0.0734)*100000 - (0.0734/2.3)*(100000) = $4148

- **R Code:**

```
### Problem-2: Framingham Heart Study
### Part-a: Memo to a Policy Maker

### Part-a: i-ii-iii.)
framingham= read.csv("framingham.csv")
library(caTools)
library(ROCR)
library(dplyr)
library(ggplot2)
set.seed(144)
split = sample.split(framingham$TenYearCHD, SplitRatio = 0.67)
table(split)
train = framingham[split == TRUE,]
test = framingham[split == FALSE,]
table(train$TenYearCHD)
table(train$TenYearCHD)/nrow(train)
model1 = glm(TenYearCHD ~., data=train, family="binomial")
summary(model1)

### Part-a: iv.) Test-set performance
PredictTrain=predict(model1, newdata = train, type="response")
table(train$TenYearCHD,PredictTrain > 0.0734)
PredictTest=predict(model1, newdata = test, type="response")
table(test$TenYearCHD,PredictTest > 0.0734)
table(train$TenYearCHD,PredictTrain > 0.0734)/nrow(train)
table(test$TenYearCHD,PredictTest > 0.0734)/nrow(test)
table(test$TenYearCHD,PredictTest > 0.0734)
Accuracy=(342+165)/(342+165+19+681)
Accuracy
TPR=165/(165+19)
TPR
FPR=681/(681+342)
FPR
```

```
### Part-a: v.) Baseline Model
summary(test$TenYearCHD==1)
summary(test$TenYearCHD==0)
BaselineAccuracy=1023/(184+1023)
BaselineAccuracy

### Part-a: vii.) Prediction on New Patient
play.obs = data.frame(male=1, age=57, education="College",currentSmoker=1,
cigsPerDay=15, BPMeds=0, prevalentStroke=0,
                      prevalentHyp=1 ,diabetes=0, totChol=220, sysBP=140, diaBP=100,
BMI=32, heartRate=63, glucose=81)

predict(model1, newdata=play.obs, type="response")

### Part-b: ROC Curve ####
ROCRpred = prediction(PredictTest,test$TenYearCHD)
ROCCurve = performance(ROCRpred, "tpr", "fpr")
plot(ROCCurve)
abline(0,1)
plot(ROCCurve, colorize=TRUE, print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,0.7))
as.numeric(performance(ROCRpred, "auc")@y.values)

### Part-e: Updated Logistic Regression Model with 3 variables ###
model2 = glm(TenYearCHD ~ male + age + cigsPerDay, data=train, family="binomial")
summary(model2)

PredictTrain2=predict(model2, newdata = train, type="response")
table(train$TenYearCHD,PredictTrain2 > 0.0734)
table(train$TenYearCHD,PredictTrain2 > 0.0734)/nrow(train)

PredictTest2=predict(model2, newdata=test, type="response")
table(test$TenYearCHD, PredictTest2 > 0.0734)
table(test$TenYearCHD, PredictTest2 > 0.0734)/nrow(test)

New_Accuracy=0.20712510 + 0.13918807
New_Accuracy
New_TPR=168/(168+16)
New_TPR
New_FPR=773/(773+250)
New_FPR
```