HAKAN SONMEZ
MBA/LGO 2018 - Section-B

15.071 Analytics Edge - Homework Assignment # 3

Problem-2: Preventing Hospital Readmissions

Part-A.) The following two insights are observed after some initial exploratory analysis,:
i.) The overall readmission percentage is **11.16%** =(11357)/(11357+90409)

```
> table(readmission$readmission)
    0     1
90409 11357
```

ii.) The majority (53.76%) of the recorded patients are Female. (by calculating 54708/(54708+47055)= 53.76%)

```
> table(readmission$gender)
Female   Male
 54708  47055
```

Part-B.) Break-even probability (p) is calculated by equating costs associated with not using intervention equals to costs associated with using intervention method and solving for p:
→ p*25000=(0.7*p)*26000+(1-0.7*p)*1000
→Solving for p → **p=0.1333333**

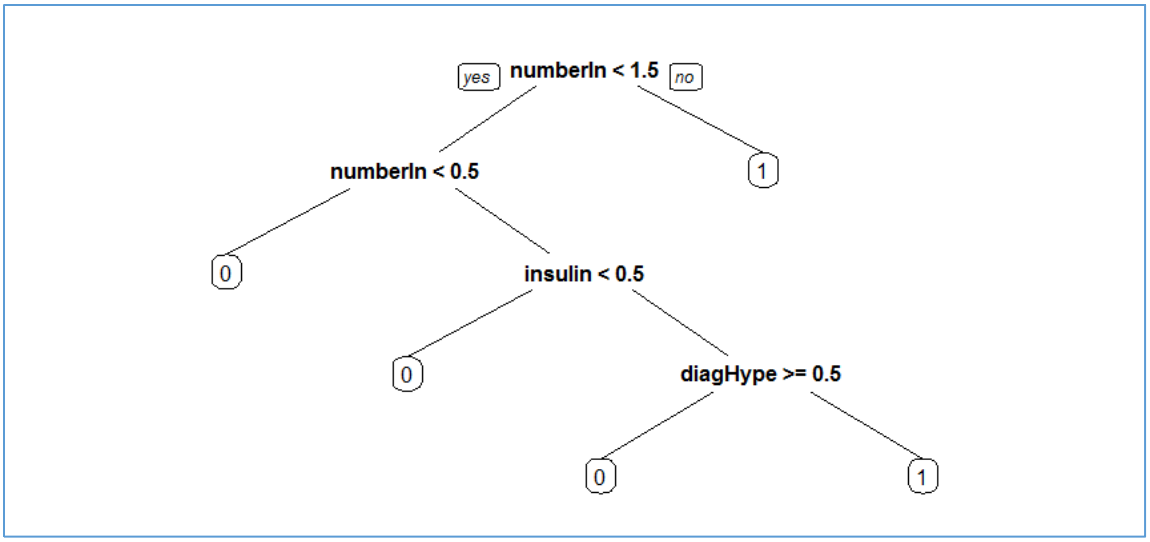
Then we assign the losses for False-Positive and False negative events:
L_{FP}=1
L_{FN}=(1-p)/p = (1-0.133333)/0.133333 = **6.5**

Resulting Loss Matrix:

		Model Prediction	
		0	**1**
Realized	**0**	0	1
	1	6.5	0

Part-C.
i.) The image of the fitted tree is copied below.
As seen on the tree, the following variables are used for selection: “NumberIn”, “Insulin” and “DiagHype”.
The subsets of patients selected for intervention are:
Subset-1: All patients that have NumberIn higher than 1.5
Subset-2: Patients who have NumberIn values between 0.5 and 1.5, and have insulin level higher than 0.5 and have diagHype lower than 0.5.
These subsets are consistent with my intuition, as I expect the Number of Inpatient visit to be the most significant indicator of whether a patient will be readmitted to the hospital. I also intuitively expect Insulin treatment and Hypertension to be very significant predictors as they are very critical health metrics.
(The CART model summary output is on the Appendix)

Fitted classification tree:



ii.) Applying the CART model on the test set results in the following confusion matrix:

```
> table(test$readmission,PredictTest)
   PredictTest
        0     1
  0 24573  7070
  1  2381  1594
```

The interpretation:

		Model Prediction	
		No readmission	**Readmission**
Reality	**Not readmitted**	24573	7070
	Readmitted	2381	1594

- Number of patients selected for intervention = 7070 + 1594 = **8864**
- Associated costs for applying the intervention = 8864 * $1000 = **$8.864M**
- Estimated number of readmissions prevented = 1594 * 0.3 = **478.2**
- Monetary benefit of reduction in admission readmissions = $25K*478.2 = $11,955K = **$11.955M**
- Total cost with intervention = 8864*$1000 +(1594*0.7)*$25000 + 2381*$25000= $96284000 = **$96.284M**
- Under current practice, total cost with no intervention= (2381+1594)*$25000 = 3975*$25K = **$99.375M** (Baseline model cost)
- Total net monetary benefit of applying interventions = $99.375M **-** $96.284M = **$3.091M**

We observe that on the test set, a savings of 3.11% (= 3.091M/99.375M) is achieved by using our prediction model compared to using the baseline model.

Part-D.) Assuming the same confusion matrix and CART model developed in part c.)

- For fixed e=0.3, calculating the range of intervention cost, c:
  99.375M > = 8864*c + (1594*0.7)*$25000 + 2381*$25000 → **max c=$1348.71**
  Range of c = (0, $1348.71)

This means as long as cost of intervention is below $1384.7, the intervention method is profitable for the hospital.

- For fixed c=$1000, calculating the range of readmission reduction factor, e:
  99.375M > = 8864*$1000 + 1594*(1-e)*$25000 + 2381*$25000 → max (1-e)= ~0.777
  → **min e=~ 0.23**
  Range of e = (0.23, 1)

As long as the reduction in readmission factor, e is above ~0.23, the intervention method is profitable for the hospital.

Part-E.) We can achieve the budget limitation by modifying our initial loss matrix that we feed into our CART model. Basically, we want our model to apply less interventions by predicting less readmissions. That's achieved by giving a similar value the $L_{FN,}$ the loss for false-negative event.

By trial, I obtain the following loss matrix
L_{FP} = 1
L_{FN} = 3.65

Feeding the loss matrix to the training set results in the following

```
> tree6 = rpart(readmission ~ ., data=train, method="class",cp=0.002, minbuck
et=25, parms=list(split="information", loss=matrix(c(0,1,3.65,0), byrow=TRUE,
nrow=2)))
> PredictTrain6 = predict(tree6, newdata = train, type="class")
> table(train$readmission,PredictTrain6)
   PredictTrain6
        0     1
  0 56178  2588
  1  6338  1044
```

We observe that the new confusion matrix results in 5.5% intervention constraint by calculating:
Predicted readmissions/ Total patients= (2588+1044) / (2588+1044+56178+6338) = 0.0549 → 5.5%

Applying the CART model prediction on test set:

```
> PredictTest6 = predict(tree6, newdata = test, type="class")
> table(test$readmission,PredictTest6)
   PredictTest6
        0     1
  0 30189  1454
  1  3440   535
```

We obtain the following result from the new confusion matrix under budget constraint:

- Number of interventions = 1454 + 535 = **1989** → Less than 8864 found in part c.)
- Number of readmission prevented = 535*0.3 = **160.5** → Less than 478.2 found in part c.)
- New Total Cost= 3440*$25000 + 535*0.7*($25000)+ (1454+535)*$1000 = $97351500= **$97.3515M** → Higher cost than $96.284M found with no budget constraint in part. C)

- Increase in the total cost due to budget limitation= $97.3515M - $96.284M =~ **$1.0675 m**

We observe that the budget constraint has indeed increased the total cost about $1.0675M for the hospital. This cost increase happened due to the fact that we failed to predict many patients who would be re-admitted to the hospital, and readmission costs more than intervention.

I would argue that budget constraint has caused the opposite of its intended purpose by increasing the total cost for hospital , therefor the constraint should be removed.

Part-F.)

i.) Creating a logistic regression model and applying it to the test set results in the following confusion matrix:

```
> PredictTest2=predict(logisticmodel1, newdata = test, type="response")
> table(test$readmission,PredictTest2 > 0.133333)
    FALSE  TRUE
  0 22182  5188
  1  2238  1226
```

We notice that the logistic regression did not make a prediction for every test set observation becausethere are some patients with missing data fields and logistic regression can not make a prediction for these patients with missing data. This can be prevented by removing the patient (rows) with the missing fields from the initial .csv file.

ii.) In terms of model interpretability, CART has several advantages over logistic regression model:

- CART has simpler rules to predict the readmissions
- Logistic regression results in a larger number of significant independent variables which makes it harder to interpret and thus CART model has more transparency
- CART makes a prediction for every observation in the test set
- CART can deliver nonlinear predictions

Note: The summary output of the logistic regression model is on the appendix.

iii.) It’s easier to implement a targeted intervention using CART model compared to the Logistic Regression Model.

- Using CART model, the doctor needs to look only at a few variables, NumberIn”, “Insulin” and “DiagHype” to decide if a patient needs intervention. Basically by just looking at the values of these variables, the doctor can move down the decision tree branches to decide in a few seconds and can also explain it to the patient easily.

- Using Logistic regression mode, the doctor should enter the values of many more significant variables into the regression equation to come up with a probability value for readmission and compare it to the threshold value to decide whether to apply intervention or not.

Part-**G.)** There can be several biases introduced by using the published information instead of piloting at the particular institution:

- A bias to geographical region where the published data is collected
- A bias due to the time period the published data is collected.
- A bias due to the size of the varying size of hospitals where the data is collected.

These biases can be removed by using only a subset of the data collected more recently from hospitals which are in the same geographical region and have similar size with our hospital

The model can be improved after pilot study in the patient population by checking if there are any significant outliers that contradict your model and modifying your model accordingly. After large scale launch, the analysis can be improved creating a new CART model based on the large data set collected from your own patient population and comparing it to the original car model.

APPENDIX:

R-Code and Outputs:

```
### HW3 Problem-2:Preventing Hospital Readmissions
library(caTools)
library(ROCR)
library(dplyr)
library(ggplot2)
library(rpart)
library(rpart.plot)

###Part-A
readmission=read.csv("readmission.csv")
summary(readmission)
str(readmission)
table(readmission$readmission)
11357/(11357+90409)
table(readmission$gender)
54708/(47055+54708)

###Part-B
p=0.13333
p*25000
(0.7*p)*26000+(1-0.7*p)*1000
LFP=1
LFN=(1-p)/p
LFN

###Part-C
set.seed(144)
idx = sample.split(readmission$readmission, 0.65)
readm.train = readmission[idx,]
readm.test = readmission[!idx,]
train = readmission[idx,]
test = readmission[!idx,]

table(train$readmission)
table(test$readmission)

tree5 = rpart(readmission ~ ., data=train, method="class",cp=0.002, minbucket=25,
parms=list(split="information", loss=matrix(c(0,1,6.5,0), byrow=TRUE, nrow=2)))
```

```
prp(tree5)

PredictTest = predict(tree5, newdata = test, type="class")
PredictTrain = predict(tree5, newdata = train, type="class")

table(test$readmission,PredictTest)
7070+1594
cost1=8664*1000
0.3*1594
savings1=478.2*25000-8864000
savings1
478.2*25000
2381+1594
25000*3975
8864000 + (1594*0.7)*25000 + 2381*25000
99375000 - 96284000

###Part-D
c=1348.71
8864*c + (1594*0.7)*25000 + 2381*25000
e=0.23
8864*1000 + 1594*(1-e)*25000 + 2381*25000


###PART-E
tree6 = rpart(readmission ~ ., data=train, method="class",cp=0.002, minbucket=25,
parms=list(split="information", loss=matrix(c(0,1,3.65,0), byrow=TRUE, nrow=2)))

PredictTrain6 = predict(tree6, newdata = train, type="class")
table(train$readmission,PredictTrain6)

PredictTest6 = predict(tree6, newdata = test, type="class")
table(test$readmission,PredictTest6)

###Part-F
logisticmodel1 = glm(readmission ~., data=train, family="binomial")
summary(logisticmodel1)

PredictTest2=predict(logisticmodel1, newdata = test, type="response")
table(test$readmission,PredictTest2 > 0.133333)
```

Summary of CART Tree:

```
> summary(tree5)
> tree5
n= 66148

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 66148 47983.0 0 (0.88840177 0.11159823)
   2) numberInpatient< 1.5 56692 34807.5 0 (0.90554223 0.09445777)
     4) numberInpatient< 0.5 44005 24316.5 0 (0.91498693 0.08501307) *
     5) numberInpatient>=0.5 12687 10491.0 0 (0.87278316 0.12721684)
      10) insulin< 0.5 5741  4166.5 0 (0.88834698 0.11165302) *
      11) insulin>=0.5 6946  5973.0 1 (0.85991938 0.14008062)
```

```
        22) diagHypertension>=0.5 586   390.0 0 (0.89761092 0.10238908) *
        23) diagHypertension< 0.5 6360  5447.0 1 (0.85644654 0.14355346) *
   3) numberInpatient>=1.5 9456  7429.0 1 (0.78563875 0.21436125) *
```

Summary of Logistic Regression Model for Part-F:

```
> summary(logisticmodel1)

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

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1811  -0.5023  -0.4440  -0.3882   2.9711  

Coefficients:
                                                                           Estimate Std. Error z 
value Pr(>|z|)    
(Intercept)                                                              -4.7244568  1.0164419  -
4.648 3.35e-06 ***
raceAsian                                                                 0.0424720  0.1734704   
0.245 0.806582    
raceCaucasian                                                             0.0063839  0.0352543   
0.181 0.856303    
raceHispanic                                                              0.0330812  0.1051079   
0.315 0.752963    
raceOther                                                                -0.1010392  0.1209671  -
0.835 0.403570    
genderMale                                                                0.0190782  0.0275367   
0.693 0.488417    
age[10-20)                                                                1.3982047  1.0287509   
1.359 0.174106    
age[20-30)                                                                1.9941125  1.0120770   
1.970 0.048802 *  
age[30-40)                                                                1.9385754  1.0095807   
1.920 0.054835 .  
age[40-50)                                                                1.9025740  1.0084051   
1.887 0.059199 .  
age[50-60)                                                                1.8122362  1.0081393   
1.798 0.072240 .  
age[60-70)                                                                1.9799822  1.0080099   
1.964 0.049501 *  
age[70-80)                                                                2.0654469  1.0079605   
2.049 0.040449 *  
age[80-90)                                                                2.0950404  1.0081489   
2.078 0.037700 *  
age[90-100)                                                               2.0249487  1.0108054   
2.003 0.045145 *  
numberOutpatient                                                         -0.0014449  0.0101398  -
0.142 0.886688    
numberEmergency                                                           0.0426712  0.0117504   
3.631 0.000282 ***
numberInpatient                                                           0.2498577  0.0087082  
28.692  < 2e-16 ***
acarbose                                                                 -0.0720021  0.2463777  -
0.292 0.770101    
chlorpropamide                                                           -1.1010519  0.7223273  -
1.524 0.127431    
glimepiride                                                               0.0006945  0.0605151   
0.011 0.990843    
```

```
glipizide                                                                 0.0687520  0.0414754  
1.658 0.097386 .  
glyburide                                                                 0.0482002  0.0460959  
1.046 0.295723    
glyburide.metformin                                                       0.0181953  0.1573429  
0.116 0.907937    
insulin                                                                   0.1315041  0.0284786  
4.618 3.88e-06 ***
metformin                                                                -0.0662166  0.0367974  -
1.799 0.071941 .  
nateglinide                                                               0.0653582  0.1486441  
0.440 0.660158    
pioglitazone                                                             -0.0409460  0.0532906  -
0.768 0.442277    
repaglinide                                                               0.1147794  0.0992885  
1.156 0.247673    
rosiglitazone                                                            -0.0621601  0.0583197  -
1.066 0.286492    
admissionTypeEmergency                                                    0.1244904  0.0572207  
2.176 0.029584 *  
admissionTypeUrgent                                                       0.0791817  0.0470389  
1.683 0.092312 .  
admissionSourceCourt/Law Enforcement                                      0.8073943  0.8022623  
1.006 0.314225    
admissionSourceEmergency Room                                            -0.0064256  0.1330925  -
0.048 0.961494    
admissionSourceHMO Referral                                               0.7057519  0.2796902  
2.523 0.011625 *  
admissionSourcePhysician Referral                                         0.0752034  0.1315496  
0.572 0.567543    
admissionSourceTransfer from a hospital                                   0.0043418  0.1496703  
0.029 0.976857    
admissionSourceTransfer from a Skilled Nursing Facility (SNF)            -0.0305850  0.1854315  -
0.165 0.868991    
admissionSourceTransfer from another health care facility                -0.1207597  0.1652150  -
0.731 0.464825    
admissionSourceTransfer from critial access hospital                     -9.3615537 79.7689088  -
0.117 0.906576    
admissionSourceTransfer from hospital inpt/same fac reslt in a sep claim  0.2967166  0.8145262  
0.364 0.715648    
numberDiagnoses                                                           0.0277096  0.0085562  
3.239 0.001201 ** 
diagAcuteKidneyFailure                                                   -0.0883478  0.0672661  -
1.313 0.189046    
diagAnemia                                                               -0.1486359  0.0832772  -
1.785 0.074289 .  
diagAsthma                                                               -0.2053919  0.0922900  -
2.226 0.026047 *  
diagAthlerosclerosis                                                     -0.0618160  0.0510689  -
1.210 0.226109    
diagBronchitis                                                           -0.1271321  0.0679114  -
1.872 0.061203 .  
diagCardiacDysrhythmia                                                   -0.0944747  0.0450614  -
2.097 0.036031 *  
diagCardiomyopathy                                                        0.0307902  0.0816969  
0.377 0.706260    
diagCellulitis                                                           -0.1381889  0.0702121  -
1.968 0.049049 *  
diagCKD                                                                   0.1444991  0.0610555  
2.367 0.017948 *  
diagCOPD                                                                 -0.0119680  0.0562430  -
0.213 0.831490    
diagDyspnea                                                              -0.2605385  0.0737956  -
3.531 0.000415 ***
diagHeartFailure                                                          0.0772871  0.0355486  
2.174 0.029696 *  
```

```
diagHypertension                                                   -0.2330388  0.0500119  -
4.660 3.17e-06 ***
diagHypertensiveCKD                                                 0.1382797  0.0540500
2.558 0.010517 *
diagIschemicHeartDisease                                            0.0218024  0.0886141
0.246 0.805653
diagMyocardialInfarction                                            0.0163324  0.0709396
0.230 0.817913
diagOsteoarthritis                                                  0.1148768  0.0983564
1.168 0.242821
diagPneumonia                                                      -0.2793876  0.0650327  -
4.296 1.74e-05 ***
diagSkinUlcer                                                       0.1827372  0.0679263
2.690 0.007140 **
timeInHospital                                                      0.0159751  0.0052705
3.031 0.002437 **
numLabProcedures                                                   -0.0015929  0.0007981  -
1.996 0.045939 *
numNonLabProcedures                                                -0.0363386  0.0099958  -
3.635 0.000278 ***
numMedications                                                      0.0094421  0.0021623
4.367 1.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 40464  on 57274  degrees of freedom
Residual deviance: 38934  on 57210  degrees of freedom
  (8873 observations deleted due to missingness)
AIC: 39064

Number of Fisher Scoring iterations: 10
```