HOMEWORK 3

Group 45

Prachi Patel

Jignasuben Vekariya

424-394-6096 (Tel. of Prachi)

857-262-0803 (Tel of Jignasu)

patel.prachi2@northeastern.edu

vekariya.j@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Prachi Patel

Signature of Student 2: J.R.Vekariya

Submission Date: 03/08/2022

Assignment 2:

Problem 1:

Cutoff=0

Confusion matrix:

	Predicted C1	Predicted C2
Actual C1	19 (n11)	0 (n12)
Actual C2	15 (n21)	0 (n22)

Sensitivity=(n11)/(n11+n12) **Specificity**=(n22)/(n22+n21)

=(19)/(19+0)=**1** =(0)/(0+15)=**0**

Cutoff=0.2

Confusion matrix:

	Predicted C1	Predicted C2
Actual C1	17 (n11)	2 (n12)
Actual C2	9 (n21)	6 (n22)

Sensitivity=(n11)/(n11+n12) **Specificity**=(n22)/(n22+n21)

=(17)/(17+2)=17/19**=0.894** =(6)/(6+9)=**0.4**

Cutoff=0.4

Confusion matrix:

	Predicted C1	Predicted C2
Actual C1	15 (n11)	4 (n12)
Actual C2	5 (n21)	10 (n22)

Sensitivity=(n11)/(n11+n12) **Specificity**=(n22)/(n22+n21)

=(15)/(15+4)=15/19**=0.789** =(10)/(10+5)=10/15=**0.666**

Cutoff=0.5

Confusion matrix:

	Predicted C1	Predicted C2

Actual C1	15 (n11)	4 (n12)
Actual C2	4 (n21)	11 (n22)

Sensitivity=(n11)/(n11+n12) **Specificity**=(n22)/(n22+n21)

=(15)/(15+4)=15/19**=0.789** =(11)/(4+11)=11/15=**0.733**

Cutoff=0.6

Confusion matrix:

	Predicted C1	Predicted C2
Actual C1	13 (n11)	6 (n12)
Actual C2	2 (n21)	13 (n22)

Sensitivity=(n11)/(n11+n12) **Specificity**=(n22)/(n22+n21)

=(13)/(13+6)=13/19**=0.684** =(13)/(13+2)=13/15=**0.866**

Cutoff=0.8

Confusion matrix:

	Predicted C1	Predicted C2
Actual C1	5 (n11)	14 (n12)
Actual C2	0 (n21)	15 (n22)

Sensitivity=(n11)/(n11+n12) **Specificity**=(n22)/(n22+n21)

=(5)/(5+14)=5/19**=0.263** =(15)/(15+0)=15/15=**1**

Cutoff=1.0

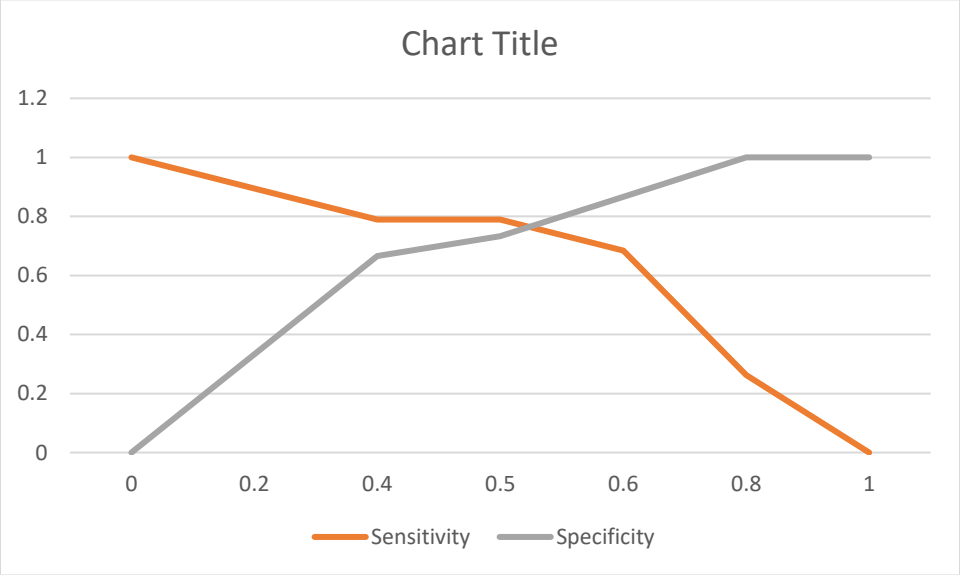
Confusion matrix:

	Predicted C1	Predicted C2
Actual C1	0 (n11)	19 (n12)
Actual C2	0 (n21)	15 (n22)

Sensitivity=(n11)/(n11+n12) **Specificity**=(n22)/(n22+n21)

=(0)/(0+19)=**0** =(15)/(15+0)=**1**

Cutoff	Sensitivity	Specificity
0	1	0
0.2	0.894	0.4
0.4	0.789	0.666
0.5	0.789	0.733
0.6	0.684	0.866
0.8	0.263	1
1	0	1



Part b:

MCC=(TP*TN)-(FP*FN)/$\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}$

= (n11*n22)-(n21*n12)/$\sqrt{(n11+n21)(n11+n12)(n22+n21)(n22+n12)}$

MCC for cutoff=0.5

= (11*15)-(4*4)/$\sqrt{(15)(15)(19)(19)}$ =(165-16)/$\sqrt{81225}$ =149/285=**0.5228**

MCC for cutoff=0.6

= (13*13)-(2*6)/$\sqrt{(15)(15)(19)(19)}$ =(169-12)/$\sqrt{81225}$ =157/285=**0.5508**

The default value(0.5) and the optimal cutoff value(0.6),both of them are positive and the models perform well. Higher the value(towards +1) better is the model ,the cutoff 0.6 MCC has a value 0.55 which is closer to 1 and performs better.

Problem 2

Part a:

C1: Diabetic retinopathy

C2: Normal conditions

Total number of records for the training data set: 100,000

DR=40000 Normal=60,000

Model:DR=38950 Normal:58500

Confusion matrix for the training data set:

	Predicted C1	Predicted C2
Actual C1	38950 (n11 or TP)	1050 (n21 or FN)
Actual C2	1500 (n21 or FP)	58500 (n22 or TN)

Error Rate=(FN+FP)/(TP+TN+FN+FP) OR (n12+n21)/(n11+n12+n21+n22)

Error rate=

(1050 + 1500)/(38950 + 1500 + 1050 + 58500) = 2550/100000 = **0.0255**

The error rate is **2.55%**

Sensitivity=(n11)/(n11+n12)

=38950/(38950+1050) = 38950/40000 = **0.97375**

The sensitivity is at 0.97

Specificity=(n22)/(n22+n21)

=58500/(58500 + 1500) = 58500/60000 = **0.975**

The specificity is 97.5%

Total number of records for the Validation data set: 10000

DR=3750 Normal=6250

Model: DR=2500 Normal=4975

Confusion matrix for the Validation data set:

	Predicted C1	Predicted C2
Actual C1	2500 (n11 or TP)	1250 (n21 or FN)
Actual C2	1275 (n21 or FP)	4975 (n22 or TN)

Error Rate:

Error Rate=(FN+FP)/(TP+TN+FN+FP) OR (n12+n21)/(n11+n12+n21+n22)

(1250 + 1275)/(2500 + 1250 + 1275 + 4975) = 2525 /10000 = **0.2525**

The error rate is **25.25%**

Sensitivity=(n11)/(n11+n12)

=2500/(2500 + 1250) = 2500/3750 = **0.66**

The sensitivity is **0.66**

Specificity=(n22)/(n22+n21)

=4975/(1275 + 4975) = 4975/6250 = 0.796

The specificity is **0.796**

Part b:

The training data set has a lower error value of 2.55% along with higher sensitivity and specificity values which indicates that the model does a better job at detecting the class of interest, which is to correctly detect patients with diabetic retinopathy as compared to the validation data set which does not perform as well in comparison and has a much higher error rate of 25.25%.Familiarity with the dataset may be one of the factors which cause the training dataset to perform better. But due to the high error rate in the validation data this may not be a very good model.

Problem 3:

C1: Positive Samples

C2: Negative Samples

Assuming that the sensitivity and specificity maintained for the validation set is going to be the same for training and the whole data set.

Sensitivity: 60%=0.6

Sensitivity=(n11)/(n11+n12)

0.6 = n11/(n11 + n12)

0.6n11 + 0.6n12 = n11

0.6n12 = 0.4n11

0.6/0.4 = n11/n12

n11= 1.5n12 ……..eq. 1

Specificity: 80%=0.8

Specificity=(n22)/(n22+n21)

0.8 = (n22)/(n22+n21)

0.8n21 + 0.8n22 = n22

0.2n22 = 0.8n21

n22/n21 = 0.8/0.2

n22/n21 = 4

n22= 4n21 ……..eq.2

Total samples in the dataset: 7000

Number of Positive Samples: 2800

Using this,

n11+n12 = 2800 ……..eq. 3

Using eq. 1 in eq. 3:

1.5n12 + n12 = 2800

2.5n12 = 2800

n12= 1120

Using n12 = 1120 in eq. 3

n11 = 1.5*1120

n11 = 1680

Negative Samples = Total Samples - Positive Samples = 7000 - 2800 = 4200

Hence we get,

4200 = n21 + n22 …….. eq. 4

Substituting eq.2 in eq.4

4200 = n21 + 4n21

4200 = 5n21

840 = n21

Substituting n21 in eq.2

n22 = 4*840 = 3360

Confusion Matrix for the total dataset:

	Predicted C1	Predicted C2
Actual C1	1680	1120
Actual C2	840	3360

confusion matrix for the validation set:

	Predicted C1	Predicted C2
Actual C1	1680*0.3=504(m11)	1120*0.3=336(m12)
Actual C2	840*0.3=252(m21)	3360*0.3=1008(m22)

Given that the prevalence is 8%

f1=No of C1 class in oversampled data/No of C1 class in field data

f1 = (504+336)/(7000*0.3*0.08) = 840/168 = 5

f2= No of C2 class in oversampled data/No of C2 class in field data

f2 = (252+1008)/(7000*0.3*0.92) = 1260/1932 = 0.65

Adjusted Confusion Matrix

	Predicted C1	Predicted C2
Actual C1	504/5 = 100.8	336/5 = 67.2
Actual C2	252/0.65 = 387.69	1008/0.65 = 1550.76

Adjusted Misclassification Rate = (387.69+67.2)/(7000*0.3)= 454.89/2100 = **0.216**

Precision = 100.8/(100.8+387.69) = 100.8/488.49 = **0.206**

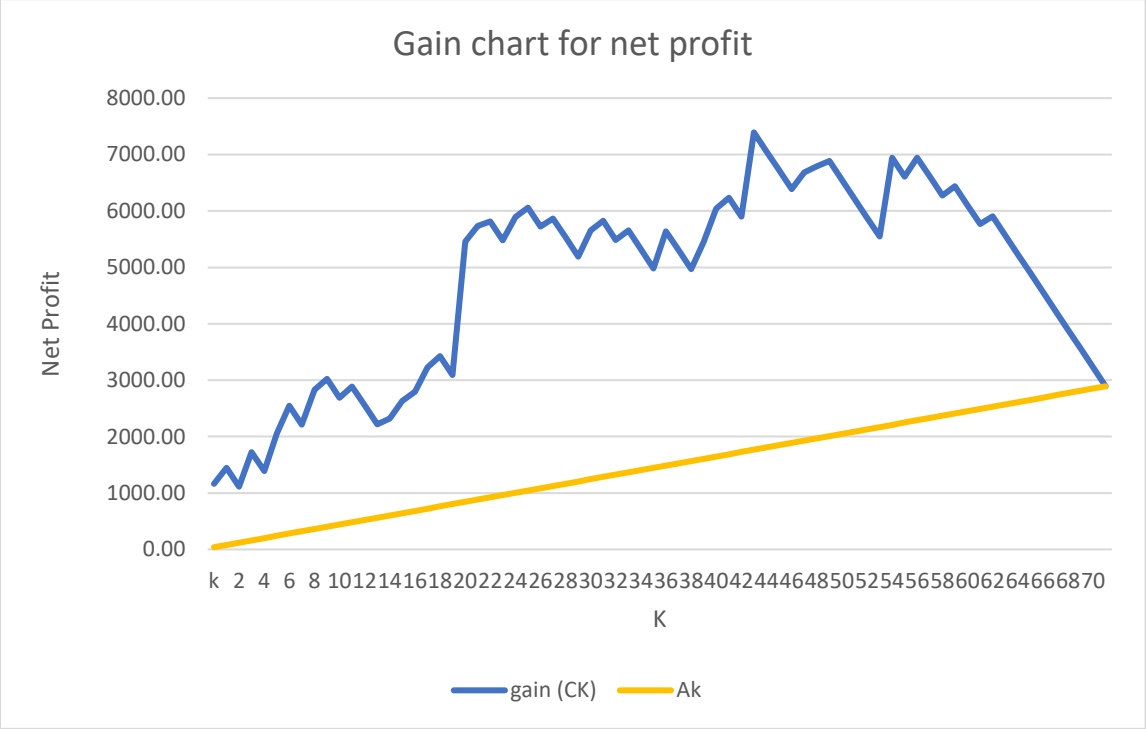
Recall = 100.8/(100.8+67.2) = 100.8/168 = **0.6**

Part b:

The unbalanced data can be dealt with by using it only to oversample training data and not the validation data, this way adjustments can be avoided and can be performed for multiple combinations of records.

Problem 4:

a.



b. The top 10 companies need to be targeted

Companies Name	**Probability of winning the opportunity**
Amazon.com	0.998
Charter Communications	0.996
AIG	0.988
MetLife	0.984
Citigroup	0.974
Bank of America	0.963

PepsiCo	0.963
Morgan Stanley	0.949
Ford Motor	0.913
Home Depot	0.91

The net profit=3021$ which on comparing to the baseline of 401.66$ gives a profit of 2619.34$ more on using the model.

c. On looking at the data we can see that the top 10 companies all are from the US.We can also see that they use GCP Big Query and Looker as the most popular Warehousing and BI tools. The cloud storage space for these companies has a range from 38TB to 858TB.We can see a common pattern when it comes to the type of tools that the use and the data they generate based on the cloud storage space that they occupy depending on the scale.