Name: Ashish Verma Course:CS773 HW#3

**Solution(1)**

* Given Total number of documents: 12,000
* Documents retrieved by the system: 3,000
* Relevant documents among retrieved: 2,000
* Total relevant documents: 4,000

**TP** = Relevant documents that were retrieved = 2000

**FP** = Irrelevant documents that were retrieved. = 3000 – 2000 =1000

**FN** = Total relevant - Relevant retrieved = 4000 - 2000 = 2,000

**TN** = Total documents - (TP + FP + FN)= 12000 - (2000 + 1000 + 2000) = 7000

**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | Relevant | Irrelevant |
| Retrieved | TP=2000 | FP=1000 |
| Not Retrieved | FN=2000 | TN=7000 |

**Recall (Sensitivity)**: TP/TP+FN = 2000/2000+2000 = 0.5

**Precision: TP/TP+FP = 2000/3000=0.667**

**True Positive Rate (TPR) (same as Recall/Sensitivity)**: 0.5

**False Positive Rate (FPR)**: FP/FP+TN = 1000/1000+7000 = 0.125

**Sensitivity (Recall)**: 0.5

**Specificity**: TN/TN+FP = 7000/7000+1000 = 0.875

**Solution(2)**

Given

|  |  |  |
| --- | --- | --- |
| Sample# | Actual Class | Predicted probability of Yes |
| 1 | Yes | 0.95 |
| 2 | No | 0.7 |
| 3 | Yes | 0.95 |
| 4 | Yes | 0.4 |
| 5 | No | 0.75 |
| 6 | No | 0.65 |
| 7 | Yes | 0.99 |
| 8 | Yes | 0.98 |
| 9 | No | 0.55 |
| 10 | No | 0.97 |

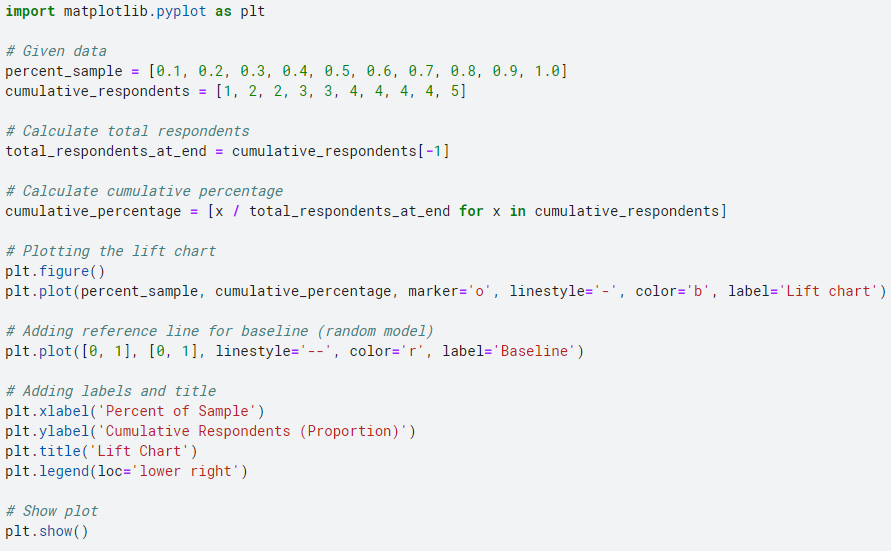
**Sort the data in the descending order of Prediction probability**

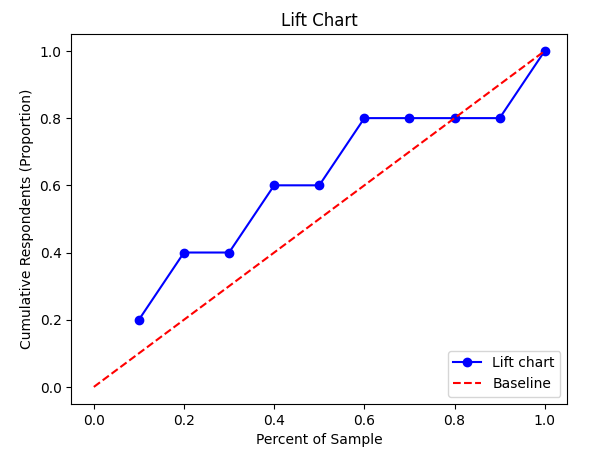
|  |  |  |
| --- | --- | --- |
| Sample# | Actual Class | Predicted probability of Yes |
| 7 | Yes | 0.99 |
| 8 | Yes | 0.98 |
| 10 | No | 0.97 |
| 1 | Yes | 0.95 |
| 3 | Yes | 0.95 |
| 5 | No | 0.75 |
| 2 | No | 0.7 |
| 6 | No | 0.65 |
| 9 | No | 0.55 |
| 4 | Yes | 0.4 |

**Calculate the %sample size and total respondents**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample# | Actual Class | Predicted probability of Yes | %Sample | #Total Respondents |
| 7 | Yes | 0.99 | 0.1 | 1 |
| 8 | Yes | 0.98 | 0.2 | 2 |
| 10 | No | 0.97 | 0.3 | 2 |
| 1 | Yes | 0.95 | 0.4 | 3 |
| 3 | Yes | 0.95 | 0.5 | 3 |
| 5 | No | 0.75 | 0.6 | 4 |
| 2 | No | 0.7 | 0.7 | 4 |
| 6 | No | 0.65 | 0.8 | 4 |
| 9 | No | 0.55 | 0.9 | 4 |
| 4 | Yes | 0.4 | 1 | 5 |

**Plot Lift Curve**

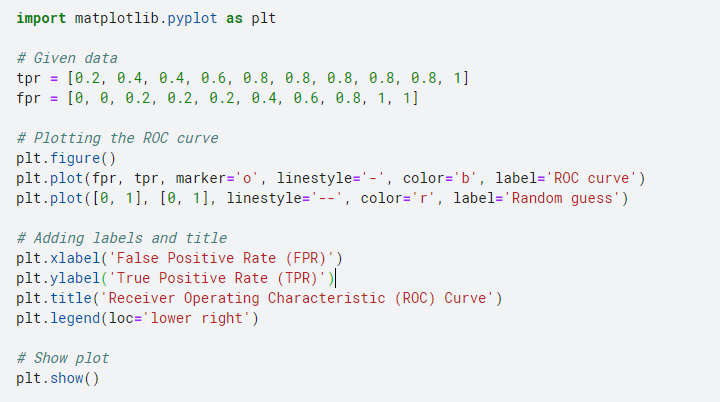
****

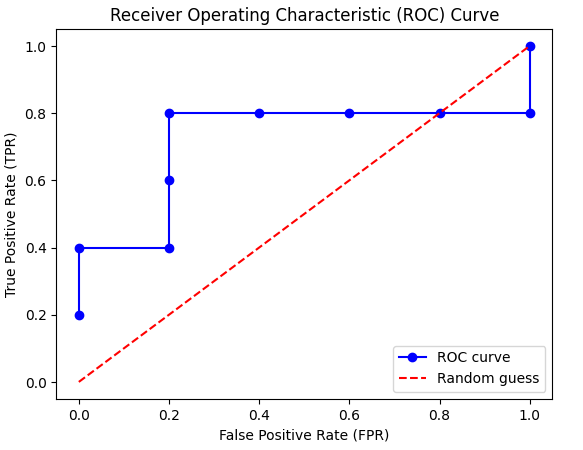


**Calculate the TPR and FPR for ROC Curve**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample# | Actual Class | Predicted probability of Yes | TP | FP | TPR | FPR | %FP | %TP |
| 7 | Yes | 0.99 | 1 | 0 | 0.2 | 0 | 0 | 0 |
| 8 | Yes | 0.98 | 2 | 0 | 0.4 | 0 | 0 | 20 |
| 10 | No | 0.97 | 2 | 1 | 0.4 | 0.2 | 0 | 40 |
| 1 | Yes | 0.95 | 3 | 1 | 0.6 | 0.2 | 20 | 40 |
| 3 | Yes | 0.95 | 3 | 1 | 0.8 | 0.2 | 20 | 60 |
| 5 | No | 0.75 | 4 | 2 | 0.8 | 0.4 | 40 | 80 |
| 2 | No | 0.7 | 4 | 3 | 0.8 | 0.6 | 60 | 80 |
| 6 | No | 0.65 | 4 | 4 | 0.8 | 0.8 | 80 | 80 |
| 9 | No | 0.55 | 4 | 5 | 0.8 | 1 | 100 | 80 |
| 4 | Yes | 0.4 | 5 | 5 | 1 | 1 | 100 | 100 |

**Plot ROC Curve**

****

****

**Solution(3)**

**Student 1: Actual grade = F**

Predicted probabilities: A: 0.1, B: 0.1, C: 0.2, D: 0.4, F: 0.2

𝑦=[0,0,0,0,1]

Quadratic Loss:

(0−0.1)2+(0−0.1)2+(0−0.2)2+(0−0.4)2+(1−0.2)2=0.01+0.01+0.04+0.16+0.64=0.86

Information Loss (Log Loss):−log\_2(0.2)= 2.32193

**Student 2: Actual grade = C**

Predicted probabilities: A: 0.2, B: 0.1, C: 0.5, D: 0.15, F: 0.05

𝑦=[0,0,1,0,0]

Quadratic Loss:

(0−0.2)^2+(0−0.1) ^2+(1−0.5) ^2+(0−0.15) ^2+(0−0.05) ^2=0.04+0.01+0.25+0.0225+0.0025=0.325

Information Loss (Log Loss): −log\_2(0.5)= 1

**Student 3: Actual grade = B**

Predicted probabilities: A: 0.3, B: 0.6, C: 0.15, D: 0.03, F: 0.02

𝑦=[0,1,0,0,0]

Quadratic Loss:

(0−0.3) ^2+(1−0.6) ^2+(0−0.15) ^2+(0−0.03) ^2+(0−0.02) ^2=0.09+0.16+0.0225+0.0009+0.0004=0.2738

Information Loss (Log Loss):−log\_2(0.6)= 0.7369

**Student 4: Actual grade = A**

Predicted probabilities: A: 0.7, B: 0.2, C: 0.05, D: 0.03, F: 0.02

𝑦=[1,0,0,0,0]

Quadratic Loss:

(1−0.7) ^2+(0−0.2) ^2+(0−0.05) ^2+(0−0.03) ^2+(0−0.02) ^2=0.09+0.04+0.0025+0.0009+0.0004=0.1338

Information Loss (Log Loss):−log\_2(0.7)= 0.5145

**Student 5: Actual grade = D**

Predicted probabilities: A: 0.1, B: 0.2, C: 0.1, D: 0.5, F: 0.1

𝑦=[0,0,0,1,0]

Quadratic Loss:

(0−0.1) ^2+(0−0.2) ^2+(0−0.1) ^2+(1−0.5) ^2+(0−0.1) ^2=0.01+0.04+0.01+0.25+0.01=0.32

Information Loss (Log Loss):−log\_2(0.5)= 1

**Solution(4)**

Given

|  |  |  |
| --- | --- | --- |
| Instance# | Actual Salary | Predicted Salary |
| 1 | 75 | 85 |
| 2 | 95 | 70 |
| 3 | 105 | 100 |
| 4 | 65 | 55 |
| 5 | 85 | 100 |
| 6 | 75 | 75 |
| 7 | 80 | 60 |
| 8 | 95 | 100 |
| 9 | 90 | 75 |
| 10 | 70 | 85 |

RMSE=1/n​∑​(Actuali​−Predictedi​)2​ from i=1 to n

Sum of squared errors = (85 – 75)^2+ (70 -95)^2+(100 – 105)^2+(55 – 65)^2+(100 -85)^2+(75 – 75)^2+(60 -80)^2+(100 – 95)^2 +(75 -90)^2+(85 – 70)^2= 100+625+25+100+225+0+400+25+225+225=1950

**Mean Squared Error** = 1950/10 = 195

**Root mean squared Error =sqrt(195)=13.96**

MAE=1/n​∑​∣Actuali​−Predictedi​∣, i = 1 to n

Sum of absolute error = (85 – 75)+ (70 -95)+(100 – 105)+(55 – 65)+(100 -85)+(75 – 75)+(60 -80)+(100 – 95)+(75 -90)+(85 – 70)=120

**Mean Absolute Error = 120/10 = 12**

**Compute Actual Mean** =Abar= 835/10 = 83.5

**Compute Predicted Mean** = Pbar=805/10 = 80.5

**Compute Deviation from mean**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Instance#** | **Actual Salary** | **Predicted Salary** | **Ai-Abar** | **Pi-Pbar** |
| 1 | 75 | 85 | -8.5 | 4.5 |
| 2 | 95 | 70 | 11.5 | -10.5 |
| 3 | 105 | 100 | 21.5 | 19.5 |
| 4 | 65 | 55 | -18.5 | -25.5 |
| 5 | 85 | 100 | 1.5 | 19.5 |
| 6 | 75 | 75 | -8.5 | -5.5 |
| 7 | 80 | 60 | -3.5 | -20.5 |
| 8 | 95 | 100 | 11.5 | 19.5 |
| 9 | 90 | 75 | 6.5 | -5.5 |
| 10 | 70 | 85 | -13.5 | 4.5 |

**Compute the products of the deviations**

(Ai -Abar)\*(Pi -Pbar)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Instance#** | **Actual Salary** | **Predicted Salary** | **Ai-Abar** | **Pi-Pbar** | **(Ai-Abar)\*(Pi-Pbar)** |
| 1 | 75 | 85 | -8.5 | 4.5 | -38.25 |
| 2 | 95 | 70 | 11.5 | -10.5 | -120.75 |
| 3 | 105 | 100 | 21.5 | 19.5 | 419.25 |
| 4 | 65 | 55 | -18.5 | -25.5 | 471.75 |
| 5 | 85 | 100 | 1.5 | 19.5 | 29.25 |
| 6 | 75 | 75 | -8.5 | -5.5 | 46.75 |
| 7 | 80 | 60 | -3.5 | -20.5 | 71.75 |
| 8 | 95 | 100 | 11.5 | 19.5 | 224.25 |
| 9 | 90 | 75 | 6.5 | -5.5 | -35.75 |
| 10 | 70 | 85 | -13.5 | 4.5 | -60.75 |

**Sum the products of the deviations** = −38.25−120.75+419.25+471.75+29.25+46.75+71.75+224.25−35.75−60.75= 1007.5

**SPA**=1007.5/9=111.95

#### Compute the sum of squares of the deviations

|  |  |  |
| --- | --- | --- |
| Instance | Ai-Abar | (Ai - Abar)^2 |
| 1 | -8.5 | 72.25 |
| 2 | 11.5 | 132.25 |
| 3 | 21.5 | 462.25 |
| 4 | -18.5 | 342.25 |
| 5 | 1.5 | 2.25 |
| 6 | -8.5 | 72.25 |
| 7 | -3.5 | 12.25 |
| 8 | 11.5 | 132.25 |
| 9 | 6.5 | 42.25 |
| 10 | -13.5 | 182.25 |

**SA** = 72.25+132.25+462.25+342.25+2.25+72.25+12.25+132.25+42.25+182.25=1452.5/n-1 = 161.39

|  |  |  |
| --- | --- | --- |
| Instance | (Pi- Pbar) | (Pi- Pbar)^2 |
| 1 | 4.5 | 20.25 |
| 2 | -10.5 | 110.25 |
| 3 | 19.5 | 380.25 |
| 4 | -25.5 | 650.25 |
| 5 | 19.5 | 380.25 |
| 6 | -5.5 | 30.25 |
| 7 | -20.5 | 420.25 |
| 8 | 19.5 | 380.25 |
| 9 | -5.5 | 30.25 |
| 10 | 4.5 | 20.25 |

**SP** = 20.25+110.25+380.25+650.25+380.25+30.25+420.25+380.25+30.25+20.25=2422.5/n-1 = 269.17

**R** = SPA/sqrt(SP)\*(SA) = 111.95/sqrt(161.39\*269.17)= 0.5371