**Cost Sensitive Analysis**

From the analysis, we understand that Yes (<30 + >30) versus NO prediction is better. Therefore considering this kind of binary classification…

The total dataset contained 101765 patient encounters

During preprocessing the instances having some missing features values were removed leaving us with 98053 instances of patient encounters.

The cost per readmission = α = 10591

The cost per *Special Diagnosis* for Patients predicted as Yes (<30 or >30) Readmissions = β = 2409

Confusion Matrix Format=>

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix Format  Total = 98053 instances  Test Data = 23053 instances  Train Data = 75000 instances | | Predicted Class | |
| YES (<30+>30) | NO |
| Actual Class | YES (<30+>30) = 10675 instances | TP | FN |
| NO = 12378 instances | FP | TN |

Saved Cost Matrix => Cost without Analysis – Cost with Predictions

Hence choosing the threshold to maximize the Saved Cost Matrix, we can get the best cost saved by each algorithm.

Choosing=>

α = 10591

β = 2409

α - β = 8182

We get the Saving Cost Matrix according to which the cost has t be maximized =>

Below is the table representing all the algorithms with the respective cost saved by each of them.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Cost Saved for Test Set (in Million USD) | Cost Saved for Total set (in Million USD) | Respective Confusion Matrix |
| Naïve Bayes | 58.726 | 249.783563 |  |
| Bayes Network | 58.808 | 250.1323396 |  |
| Random Forest | 59.425 | 252.7566705 |  |
| Adaboost-Trees | 58.298 | 247.9631195 |  |
| Neural Network | 58.963 | 250.7916123 |  |

Conslusion:

Looks like we are not rewarding proper classification of NO, leading very low recall on NO. So we might have to consider some cost ﻵ, which would be saved for if we predict that this person is not getting readmitted to hospital.