Predicting Net Promoter Score (NPS) to Improve patient Experience at Manipal Hospitals

(Harvard Business Case Study)

Group Members:

- Suhail Shaikh
- Vikalp Mehta

Q.1). What is the business problem in this case and how is this business problem converted into an analytics problem?

Answer: MHE employed manual system to take patient's feedbacks which account for around 30% coverage of total registrations in month. This delayed the real time implementation of patient's feedback and updating of action taken back to patients.

A method employed to convert manual feedback into real time feedback gives tangibility to process and an opportunity to deploy analytics solutions to data gathered. Now the problem is much able to be viewed from lens of analysis of data gathered and can be solved from predicting the net promoter score.

Q.2). What is the extent of missing information is survey data? What implication could it have on model building? Use KNN to fill in missing information.

Not to be attempted as per Instructor.

Q.3). How can we estimate sensitivity and specificity of three-class problem? Provide the formulas.

Answer: For a multi class classification problem, we can deploy one vs all approach to find sensitivity and specificity. For example, consider following matrix. In this problem we use referencing method to determine sensitivity and specificity.

Prediction	Detractor	Passive	Promotor
Detractor	21	9	1
Passive	11	50	26
Promotor	12	58	176

In this example sensitivity of Detractor is TP(Detractor)/(TP(Detractor)+FN(Detractor) where

TP(Detractor) = 21

FN(Detractor)= E (detractor and passive) +E(Detractor + promoter)

FN(Detractor) = 9+1 =10

So, sensitivity = 21/31 = .67

Similarly, Specificity is TN(Detractor)/TN(Detractor)+FP(Detractor)

TN(Detractor) = 50+176+58+26= 310

FP(Detractor) = E(Passive and Detractor) + E(Promoter and Detractor)

FP(Detractor) = 11+12 = 23

Specificity = 310/(310+23) = .93

Q.4). What is quasi-complete separation? Which variables in the Manipal Hospital dataset are leading to quasi-complete separation?

Answer: Quasi complete separation leads to the inability of a predictor variable to separate response variable in clear cut classes. Or it can separate predictor class up to some extent.

In following example, we can see that for P1(12) we have y as 0 and 1. While for P1 < 12 and P1 > 12 Y is 0 and 1 respectively. It means that for all P1= 12 we have achieved quasi separation in data.

Y	P1	P2
0	10	2
0	10	0
0	12	-3
1	12	2
1	14	1
1	15	8
1	17	1
1	18	4
1	19	2
1	20	4

We come to know about this situation in Data set when we get error message like "fitted probabilities numerically 0 or 1 occurred" in R.

We can keep or remove variables from data as desired to build the model.

When we encountered this problem in our Data, we found out following variables causing quasi and complete separation.

Marital Status Country EM_NURSING Bed Category State

EM_DOCTOR DOC_ATTITUDE NS_NURSESATTITUDE OVS_OVERALLSTAFFATTITUDE

CE_NPS.

Q.5). What is orthogonal polynomial coding and how is it implemented in contrasting ordinal variables?

Answer: When using regression, categorical variables of k categories are passed as k-1 sequence of variables. Regression coefficients of these k-1 set corresponds to set of linear hypotheses. To understand this phenomenon, we will have to dive deeper in vector and its projections on planes.

Multiple Regression assumes that there is no collinearity among variables but, it is not the case. Orthogonal polynomial regression helps treating its vectors in an Orthogonal plane to keep relationship between them independent.

90-degree angle between two vectors in plane illustrate independent relation between them. If angle decreases from 90 degree, they are positively related and vice versa. To handle collinearity among variables we try to keep plane of vectors orthogonal using gram Schmidt model. This model helps generate kth order polynomial and transfer model to linear, quadratic and cubic easily. This contrast method is always used for Ordinal Variables in which labels are equally spaced.

Q.6). How can we convert a multi-class problem to a binary classification problem when the objective is to understand the detractors among the group? Apply logistic regression on a binary classification problem. Use all variables except the ones identified as leading to quasi-complete separation (use step-wise regression to build the model). Keep this in mind that you need to convert the attributes of survey questionnaires to ordinal variables before building the logistic model.

Answer:

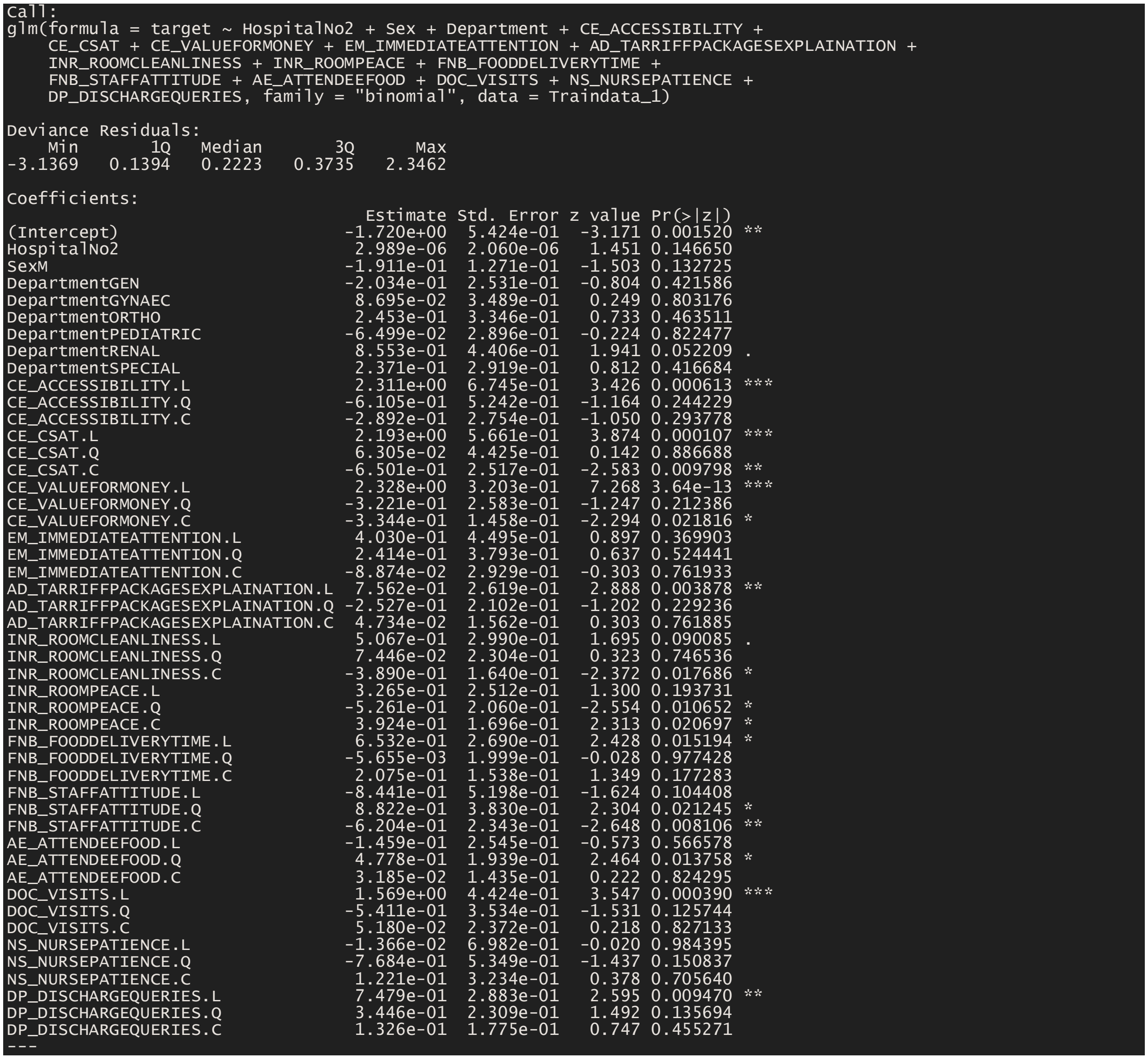
There are many methods to perform transformation to binary from multiple class. Most prominent one's are

1. Dummy Variable: This method employs transferring predictor class in n-1 columns and keeping values of class only in that column and set rest to zero. This helps training classifier for each binary class.
2. Reference Method: This method employs 1 to 1 and 1 vs all method of transformation of multiclass to binary class. We keep two binary class and rows to those classes and train classifiers on them.

We have used reference method one vs all to perform this binary classification. Increase in probability for predicted class means that class is moving up on NPS score.

After converting all the survey response columns to ordinal factor form, we have run logistic model to find out variables which are causing quasi complete separation.

After removing those variables, we have deployed Step wise model to do feature selection and build model. Step wise regression build on logistic model will give us the significant variables. Passing in the ordered factors has resulted in feature selection in the form of Orthogonal Polynomials which suggest that which polynomial form is significant with the predicted class. For example, value for money has the positive linear relation with predicted class. An increase in value for money will increase the probability of NPS score.

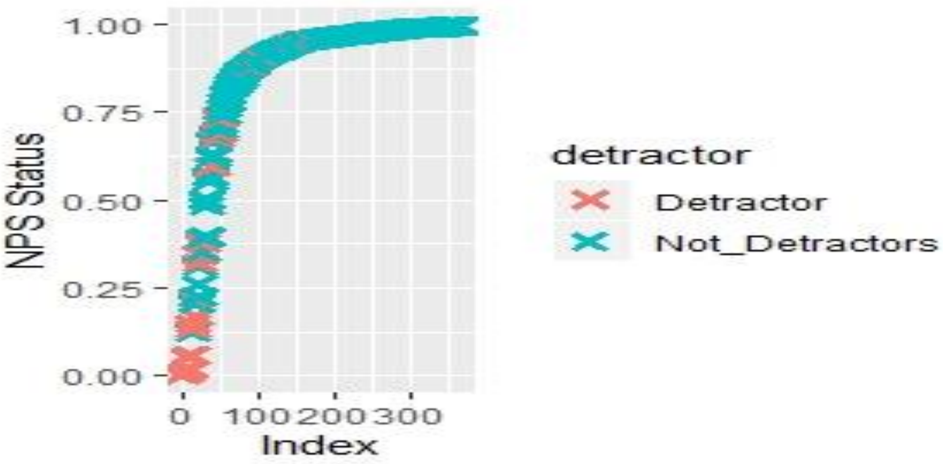


```
Call:
glm(formula = target ~ HospitalNo2 + Sex + Department + CE_ACCESSIBILITY +
    CE_CSAT + CE_VALUEFORMONEY + EM_IMMEDIATEATTENTION + AD_TARRIFFPACKAGESEXPLAINATION +
    INR_ROOMCLEANLINESS + INR_ROOMPEACE + FNB_FOODDELIVERYTIME +
    FNB_STAFFATTITUDE + AE_ATTENDEEFOOD + DOC_VISITS + NS_NURSEPATIENCE +
    DP_DISCHARGEQUERIES, family = "binomial", data = Traindata_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1369   0.1394   0.2223   0.3735   2.3462

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.720e+00  5.424e-01  -3.171 0.001520 **
HospitalNo2                       2.989e-06  2.060e-06   1.451 0.146650
SexM                             -1.911e-01  1.271e-01  -1.503 0.132725
DepartmentGEN                    -2.034e-01  2.531e-01  -0.804 0.421586
DepartmentGYNAEC                  8.695e-02  3.489e-01   0.249 0.803176
DepartmentORTHO                   2.453e-01  3.346e-01   0.733 0.463511
DepartmentPEDIATRIC              -6.499e-02  2.896e-01  -0.224 0.822477
DepartmentRENAL                   8.553e-01  4.406e-01   1.941 0.052209 .
DepartmentSPECIAL                 2.371e-01  2.919e-01   0.812 0.416684
CE_ACCESSIBILITY.L                2.311e+00  6.745e-01   3.426 0.000613 ***
CE_ACCESSIBILITY.Q               -6.105e-01  5.242e-01  -1.164 0.244229
CE_ACCESSIBILITY.C               -2.892e-01  2.754e-01  -1.050 0.293778
CE_CSAT.L                         2.193e+00  5.661e-01   3.874 0.000107 ***
CE_CSAT.Q                         6.305e-02  4.425e-01   0.142 0.886688
CE_CSAT.C                        -6.501e-01  2.517e-01  -2.583 0.009798 **
CE_VALUEFORMONEY.L                2.328e+00  3.203e-01   7.268 3.64e-13 ***
CE_VALUEFORMONEY.Q               -3.221e-01  2.583e-01  -1.247 0.212386
CE_VALUEFORMONEY.C               -3.344e-01  1.458e-01  -2.294 0.021816 *
EM_IMMEDIATEATTENTION.L           4.030e-01  4.495e-01   0.897 0.369903
EM_IMMEDIATEATTENTION.Q           2.414e-01  3.793e-01   0.637 0.524441
EM_IMMEDIATEATTENTION.C          -8.874e-02  2.929e-01  -0.303 0.761933
AD_TARRIFFPACKAGESEXPLAINATION.L  7.562e-01  2.619e-01   2.888 0.003878 **
AD_TARRIFFPACKAGESEXPLAINATION.Q -2.527e-01  2.102e-01  -1.202 0.229236
AD_TARRIFFPACKAGESEXPLAINATION.C  4.734e-02  1.562e-01   0.303 0.761885
INR_ROOMCLEANLINESS.L             5.067e-01  2.990e-01   1.695 0.090085 .
INR_ROOMCLEANLINESS.Q             7.446e-02  2.304e-01   0.323 0.746536
INR_ROOMCLEANLINESS.C            -3.890e-01  1.640e-01  -2.372 0.017686 *
INR_ROOMPEACE.L                   3.265e-01  2.512e-01   1.300 0.193731
INR_ROOMPEACE.Q                  -5.261e-01  2.060e-01  -2.554 0.010652 *
INR_ROOMPEACE.C                   3.924e-01  1.696e-01   2.313 0.020697 *
FNB_FOODDELIVERYTIME.L            6.532e-01  2.690e-01   2.428 0.015194 *
FNB_FOODDELIVERYTIME.Q           -5.655e-03  1.999e-01  -0.028 0.977428
FNB_FOODDELIVERYTIME.C            2.075e-01  1.538e-01   1.349 0.177283
FNB_STAFFATTITUDE.L              -8.441e-01  5.198e-01  -1.624 0.104408
FNB_STAFFATTITUDE.Q               8.822e-01  3.830e-01   2.304 0.021245 *
FNB_STAFFATTITUDE.C              -6.204e-01  2.343e-01  -2.648 0.008106 **
AE_ATTENDEEFOOD.L                -1.459e-01  2.545e-01  -0.573 0.566578
AE_ATTENDEEFOOD.Q                 4.778e-01  1.939e-01   2.464 0.013758 *
AE_ATTENDEEFOOD.C                 3.185e-02  1.435e-01   0.222 0.824295
DOC_VISITS.L                      1.569e+00  4.424e-01   3.547 0.000390 ***
DOC_VISITS.Q                     -5.411e-01  3.534e-01  -1.531 0.125744
DOC_VISITS.C                      5.180e-02  2.372e-01   0.218 0.827133
NS_NURSEPATIENCE.L               -1.366e-02  6.982e-01  -0.020 0.984395
NS_NURSEPATIENCE.Q               -7.684e-01  5.349e-01  -1.437 0.150837
NS_NURSEPATIENCE.C                1.221e-01  3.234e-01   0.378 0.705640
DP_DISCHARGEQUERIES.L             7.479e-01  2.883e-01   2.595 0.009470 **
DP_DISCHARGEQUERIES.Q             3.446e-01  2.309e-01   1.492 0.135694
DP_DISCHARGEQUERIES.C             1.326e-01  1.775e-01   0.747 0.455271
---
```

Accuracy for this model is around 82%.



Q.7): Compare the results of ensemble methods (Random Forest and Ada Boost) when applied to a multi-class classification problem vis-a-vis a binary classification problem.

We applied Random Forest and Ada boost to both sets of Binary and multi class data and found following as a result.

	Accuracy			Sensitivity	
Model	RF	Adaboost		RF	Adaboost
Normal					
Multiclass	67.86	67.2		47.73	47.73
Binary	90.11	90.65		40.9	43.18

Assumption: We have taken Detractors as positive class in order to identify them in real time. This will improve ability of MHE staff to predict Detractors and take appropriate actions to improve NPS score.

In order to improve NPS accurately we will need to reduce False Negative rate or increase sensitivity so that actual count of detractors can be identified.

After running Random Forest and Ada Boost on binary and multi class, we found that sensitivity is better in Adaboost and Random Forest and its around 47% for both cases. Since data is quite imbalanced, we need to deploy some resampling techniques to improve out result. Sensitivity score for Multiclass classification is better than Binary class.

Q.8). Check the effect of balancing methods (under-sampling, over-sampling, and SMOTE (Synthethic Minority Oversampling) on the performance of ensemble methods.

After running Random Forest and Ada boost on Normal data, we tried to transform imbalance in data via deploying techniques like Up sampling, SMOTE and down sampling. These techniques help remove imbalance in predictor class.

We ran Random Forest and Ada Boost for binary and Multiple class for each of the re sampling technique. And found results as follow.

	Accuracy			Sensitivity	
Model	RF	Adaboost		RF	Adaboost
Normal					
Multiclass	67.86	67.2		47.73	47.73
Binary	91.12	89		45.5	38

Upsample					
Multiclass	66.48	62.36		47.72727	59
Binary	90.66	85.16		40.91	68.18182
Downsample					
Multiclass	59.07	61.53		65.91	54.55
Binary	73.9	76.92		70.46	65.91
SMOTE					
Multiclass	60.99	62.63		52.27	52.27
Binary	89.84	90.38		47.73	56.82

Assumption: We have taken Detractors as positive class in order to identify them in real time. This will improve ability of MHE staff to predict Detractors and take appropriate actions to improve NPS score.

We can interpret following from above.

1. We observed that sensitivity for Random Forest multiclass was highest for Down sample technique and it reached to 66% in Downsample technique from 47% in Normal dataset without sampling.
2. For Binary class Random Forest, it reached to 70% from 40%. A good jump in sensitivity score which is most desired in our assumed case.
3. Though sensitivity has improved for each of the resampling technique, but Down Sampling has improved it Drastically. A huge increase in terms of 30 percentage points.

Q.9). What should be the strategy for using the model to improve patient experience in the hospital and reduce proportion of detractors?

Answers: In order to respond in real time, we need to look upon factors which are most influential one's to predictions. For example, we have significant variables like value for money which has linear relationship with target variable. Which suggest that increase in the satisfaction index of this variable will increase the probability of Non-Detractor in target variable. Few of the positive significance variables are CE_Accessibility, CS_STAT, DOC_Visit, Foof Delivery time and AD_tarrifpackages derived from model.

One of the possible strategies could be to check NPS score periodically and keep correcting the areas of concerns in real time. We can build a predictive model based upon boosting techniques to predict class of the patient in real time and this will help us to taking that patient to desire class by implementing measures for the correct problem variable.