A Basic Classifier to Prospectively Identify Dissatisfied Patients at the Time of Admission

CS613 Machine Learning - McWelling Todman

Due: March 19^{th}, 2019

1 Abstract

Demographic and clinical data have proven useful predictors of patient satisfaction performance in multiple contexts. I develop a basic classifier to prospectively identify inpatients who are likely to be dissatisfied at the time of discharge. Model performance is poor (AUC=0.24295 — Accuracy 57.2%), however there is lots of opportunity for future work involving a more rigorous feature selection process.

2 Background

Hospitals in the United States that receive reimbursement for treating Medicare and Medicaid patients participate in a regulatory incentive program called Value-Based Purchasing (VBP), which is administered by the Center for Medicare and Medicaid Services (CMS).[1] Through this program 2% medicare revenue is withheld from all participating hospitals, from which anywhere between 0-4% can be earned back by performing well on the measures CMS uses to assess quality. One component of quality as measured by CMS is "Person and Community Engagement," which is captured using the Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey. Performance on this survey accounts for one quarter of the revenue withheld by CMS (0.5%) each federal fiscal year. For large health systems with sizeable Medicare and/or Medicaid populations, the financial implications of HCAHPS performance can be quite significant. The survey is administered to all inpatients discharged from the hospital, however there are a few exceptions: mortalities, minors (under age 18), psychiatric patients, discharges to skilled nursing facilities or rehab, law enforcement, and patient subject to state-level regulations. CMS measures performance on the various questions using a methodology called "Top Box Percentage," which is essentially a simplification of the net promoter score concept.[2] For each question on the survey, one or more of the possible responses a patient can select is considered a "Top Box" response. Top Box responses are generally the most positive of the available choices. For example, of the possible answer to the question "Did nurses treat you with courtesy and respect?": Never, Sometimes, Usually, Always—"Always" is designated as the Top Box response. The binary nature of the Top Box measurement system lays the foundations for a very straightforward classification problem. The objective of this study is to develop classification model that at the time of admission will prospectively identify the likelihood of whether or not a patient will be satisfied with the care they received at the time of discharge. An early warning system of this variety would provide hospital administrators with enough lead time to check-in with the patient and perform service recovery if needed prior to discharge, which would hopefully result in greater satisfaction and more Top Box responses on the survey.

3 Related Work

A survey of current literature reveals that there have been a decent number of studies analyzing HCAHPS survey data as it relates to patient/provider demographics or scanning for interactions between questions. On balance, much of important work that has been done in this area to-date frames the inquiry in a similar fashion to how an economist might study preferences, centering descriptive and diagnostic analytic techniques to retrospectively gain insight into discrepancies in the perception of care among patient sub-populations. There appear to be very few—if any—studies involving the predictive modeling of HCAHPS scores at the patient level. Despite the dearth of studies in which the authors attempt to prospectively estimate satisfaction on the HCAHPS survey, previous work in the field provides valuable insight into which sorts of demographic factors should be considered during the feature selection process. For instance:

1. A 2011 Hopkins study was able to show that a composite feature, composed of race, gender, age, and education can be used to explain variation in patient perception of care. The authors were able to show that patient-provider social discordance (large differences in the composite measure) was associated with lower patient perceptions of care: lower overall rating of office visit, lower likelihood to recommend the practice to a friend.[3]

2. A 2015 study by McFarland, Ornstein, and Holcombe showed that at the hospital level, bed count and percent of patient population with a primary language other than English negatively correlate with HCAHPS scores, whereas education level and percentage of patient population to identify as white were positively correlated with HCAHPS scores.[4]

3. A 2016 study by Li, Lee, Glicksberg, Radbill, and Dudley surfaced a number risk factors to satisfaction as reported by HCAHPS among surgical inpatients, such as self-evaluation of health, education level, Asian, White, treatment by oncology division, and being prescribed a new medication.[5]

4. A 2008 study by Jha, Orav, Zheng, and Epstein published in NEJM found that a high ratio of nurses to patient-days was predictor of top quartile performance on the HCAHPS survey (P¡0.001).[6]

In addition to formal publications, Ohio State University's Wexler Medical Center developed a classification scheme using the conditional probability of a patient providing a top box response based on gender, age group, and hospital service. This technique was replicated by the Predictive Healthcare team at Penn Medicine: University of Pennsylvania Health System.[6]

4 Theoretical Underpinnings

I opted for a logistic regression model optimized using batch gradient ascent because it is fairly straightforward and is not computationally expensive to train. Because there has not been a lot of work done in this area, it makes sense to use a start with a basic linear classifier. The core mathematical components of the logistic regression model are as follows:

1. Dataset X

2. Labels Y

3. Learning rate λ

4. Sigmoid function $g(z) = \frac{1}{1+e^{-z}}$, is differentiable and tends toward 0 as z decreases and tends towards 1 as z increases.

5. Likelihood estimate $l(y|X, \theta) = \Pi_{t=1}^{N}(g(x, \theta))^{y_t}(1 - g(x, \theta))^{1-y_t}$

6. Log Likelihood estimate $l(y|X, \theta) = \Sigma_{t=1}^{N} y_t \ln(g(x_t, \theta)) + (1 - y_t)\ln(1 - g(x_t, \theta))$

7. Cost Function (Average Gradient of Log Likelihood) $\frac{\delta l}{\delta \theta} = \frac{1}{N} X^T(Y - g(X, \theta))$

8. Update Function $\theta = \theta + \frac{\lambda}{N} X^T(Y - g(X, \theta))$

When training a logistic regression model, the goal is to find a vector θ such that your log likelihood estimate is maximized. There is no closed form solution to this problem, meaning an iterative approach must be taken. During each iteration of the training process, a log likelihood estimate is recorded and the θ vector is updated using the update function. Assuming all goes according the plan, your theta parameter will slowing inch along the gradient and converge around a global maximum when the change in log likelihood estimates approaches a reasonably small number (e.g. $1E - 23$).

5 Formalized Approach

The assumption or hypothesis to be tested in this study is whether a broad selection of features, many of which have demonstrated predictive power in realtion to the HCAHPS performance, can be used to reliably predict whether or not a patient will be satisfied with the care they received upon discharge. In order to test this hypothesis, I use batch gradient descent to train a single-node logistic regression classifier. My input data was a combination of HCAHPS responses and demographic characteristics of inpatients discharged from an acute care, tier-1 trauma center located in Philadelphia, Pennsylvania between January 2015 and December of 2018.

I use the "Rate Hospital on a scale of 0-10" question from the HCAHPS survey as a classification target. CMS scores this question on a binary scale in which responses of "9" or "10" are considered "Top Box" responses, allowing for conversion to a simple binary response variable. The question is generally used as a proxy for overall satisfaction because it exhibits non-trivial correlation with all other questions on the survey.

HCAHPS PATIENT-LEVEL CORRELATIONS*

	Communication with Nurses	Communication with Doctors	Responsiveness of Hosp. Staff	Comm. About Medicines	Cleanliness of Hospital Env.	Quietness of Hospital Env.	Discharge Information	Care Transition	Hospital Rating	Recommend the Hospital
Communication with Nurses	1	0.53	0.56	0.50	0.39	0.32	0.27	0.43	0.64	0.58
Communication with Doctors		1	0.38	0.44	0.27	0.26	0.28	0.40	0.52	0.47
Responsiveness of Hosp. Staff			1	0.41	0.35	0.32	0.20	0.35	0.51	0.45
Comm. About Medicines				1	0.33	0.29	0.35	0.45	0.48	0.43
Cleanliness of Hospital Env.					1	0.28	0.18	0.27	0.41	0.36
Quietness of Hospital Env.						1	0.13	0.25	0.35	0.29
Discharge Information							1	0.30	0.30	0.28
Care Transition								1	0.47	0.45
Hospital Rating									1	0.76
Recommend the Hospital										1

*Patient-level Pearson correlations of rescaled linear means of HCAHPS measures, for patients discharged between July 2016 and June 2017 (3.0 million completed surveys).

Note: All correlations are significant at p<0.001.

Figure 1: Correlation Among HCAHPS Questions

In an ideal situation, feature selection would have been a careful and methodical process in which subject matter experts were consulted, a causal model was constructed, and features were added or dropped based on Information Gain. Unfortunately this was not possible due to time constraints. The features that were included in the model were selected because of their accessibility in the EHR system, availability at the time patient admission, and demonstrative predictive power as outlined in the related work section. These features can be bucketed into several thematic groups: language, marriage, race, age, gender, payor (proxy for SES), admit disposition, and clinical service. The features are as follows:

1 Response Variable / Classification Target (1 binary target)

a) OVERALL1_TB_N - Binary, HCAHPS measure with an 11-point scale (0–10) converted to 1 (0–8) or 0 (9–10).

2 Explanatory Variables (43 binary features)

a) LANG_ENGLISH - Binary, patient prefers english communication
b) MARRIED_YN - Binary, patient is married
c) RACE_WHITE - Binary, patient self-reported being White
d) RACE_BLACK - Binary, patient self-reported being Black
e) RACE_OTHER - Binary, patient self-reported another racial background
f) AGE_0_17 - Binary, patient age between 0 and 17 Y.O.
g) AGE_18_34 - Binary, patient age between 18 and 34 Y.O.
h) AGE_35_49 - Binary, patient age between 35 and 49 Y.O.
i) AGE_50_64 - Binary, patient age between 50 and 64 Y.O.
j) AGE_65_79 - Binary, patient age between 65 and 79 Y.O.
k) AGE_80_PLUS - Binary, patient age >= 80 Y.O.
l) GENDER_FEMALE - Binary, patient gender identity is Female
m) GENDER_MALE - Binary, patient gender identity is Male
n) GENDER_OTHER - Binary, patient gender identity does not conform to Male/Female binary
o) PAYOR_MEDICAID - Binary, patient primary payor is Medicaid
p) PAYOR_PRIVATE - Binary, patient primary payor is a private insurer
q) PAYOR_MEDICARE - Binary, patient primary payor is Medicare, medicare advantage, etc.
r) PAYOR_SELFPAY - Binary, patient pays out of pocket
s) PAYOR_OTHER - Binary, patient primary payor is atypical
t) ED_ADMIT - Binary, patient was admitted to hospital through the emergency department
u) ADM_CLINIC - Binary, patient was referred to the hospital from a clinic
v) ADM_NONHEALTHCARE - Binary, the patient came to the hospital from a non-healthcare point of origin
w) ADM_TRANSFER - Binary, the patient was transferred from another healthcare facility
x) SVC_MED - Binary, primary hospital service caring for patient was medicine
y) SVC_CVM - Binary, primary hospital service caring for patient was cardiovascular medicine
z) SVC_ORT - Binary, primary hospital service caring for patient was orthopaedics
aa) SVC_HSP - Binary, primary hospital service caring for patient was hospitalists
ab) SVC_GER - Binary, primary hospital service caring for patient was geriatrics
ac) SVC_FAM - Binary, primary hospital service caring for patient was family medicine
ad) SVC_NSU - Binary, primary hospital service caring for patient was neurosurgery
ae) SVC_GIS - Binary, primary hospital service caring for patient was GI surgery
af) SVC_URO - Binary, primary hospital service caring for patient was urology
ag) SVC_CSU - Binary, primary hospital service caring for patient was cardiac surgery
ah) SVC_PUL - Binary, primary hospital service caring for patient was pulmonary medicine
ai) SVC_POD - Binary, primary hospital service caring for patient was podiatry
aj) SVC_THO - Binary, primary hospital service caring for patient was thoracic surgery
ak) SVC_VAS - Binary, primary hospital service caring for patient was vascular surgery
al) SVC_PLS - Binary, primary hospital service caring for patient was plastic surgery
am) SVC_CRS - Binary, primary hospital service caring for patient was colorectal surgery
an) SVC_SUR - Binary, primary hospital service caring for patient was general surgery
ao) SVC_HEM - Binary, primary hospital service caring for patient was hemotology/oncology
ap) SVC_NEU - Binary, primary hospital service caring for patient was neurology

aq) SVC_OTHER - Binary, primary hospital service caring for patient was an atypical service

I used a randomized hold-out approach for validation, leaving 33% of the data for testing (N=1564) and using the majority of the observations for training the model (N=3178) with a learning rate of $\lambda = 0.01$. The training process involved minimizing the negative log likelihood to find (what I hope) is a global minimum.

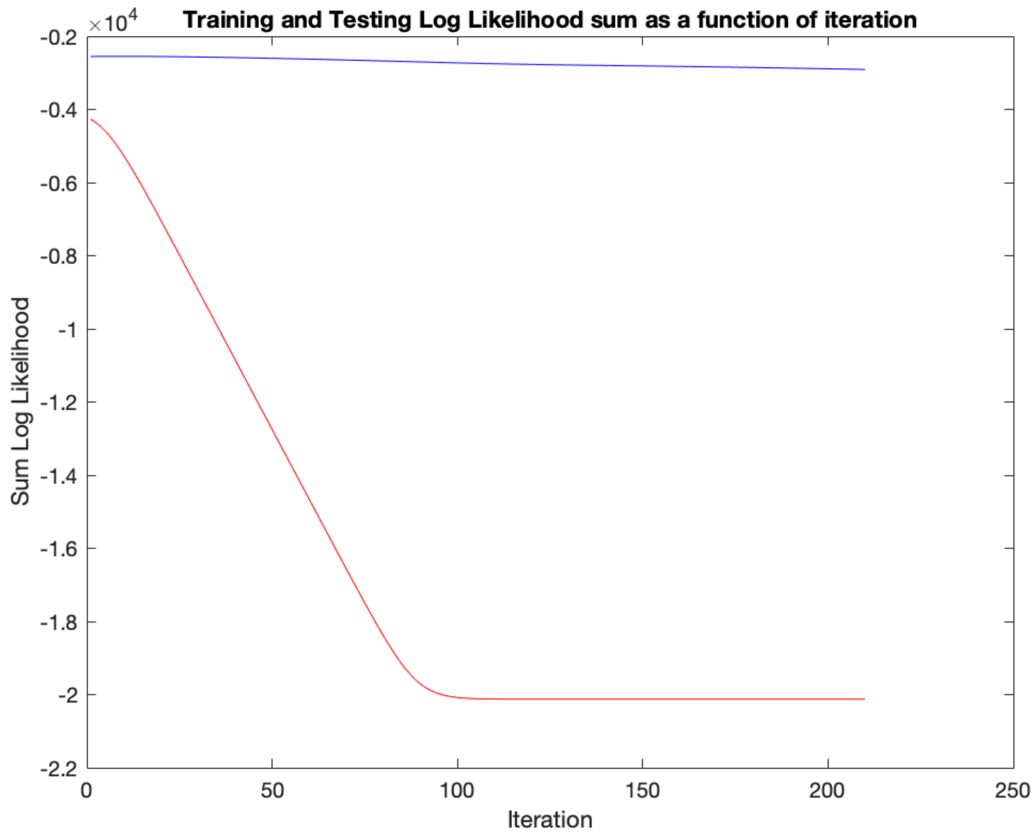

Figure 2: Training Process

6 Evaluation

The logistic regression model did not perform particularly well (AUC=0.24295) when transitioning from the testing to the training dataset. A sample classification threshold of 0.75 was chosen ($satisfied < 0.75 \geq unsatified$) to illustrate performance. The model struggled to reliably identify positive cases (unsatisfied patient), exhibiting low performance across the board. Even if an appropriate cost function were to be developed, which prioritized maximizing true positives, the false positive rate would climb to rapidly for the predictions to be actionable in an operational context.

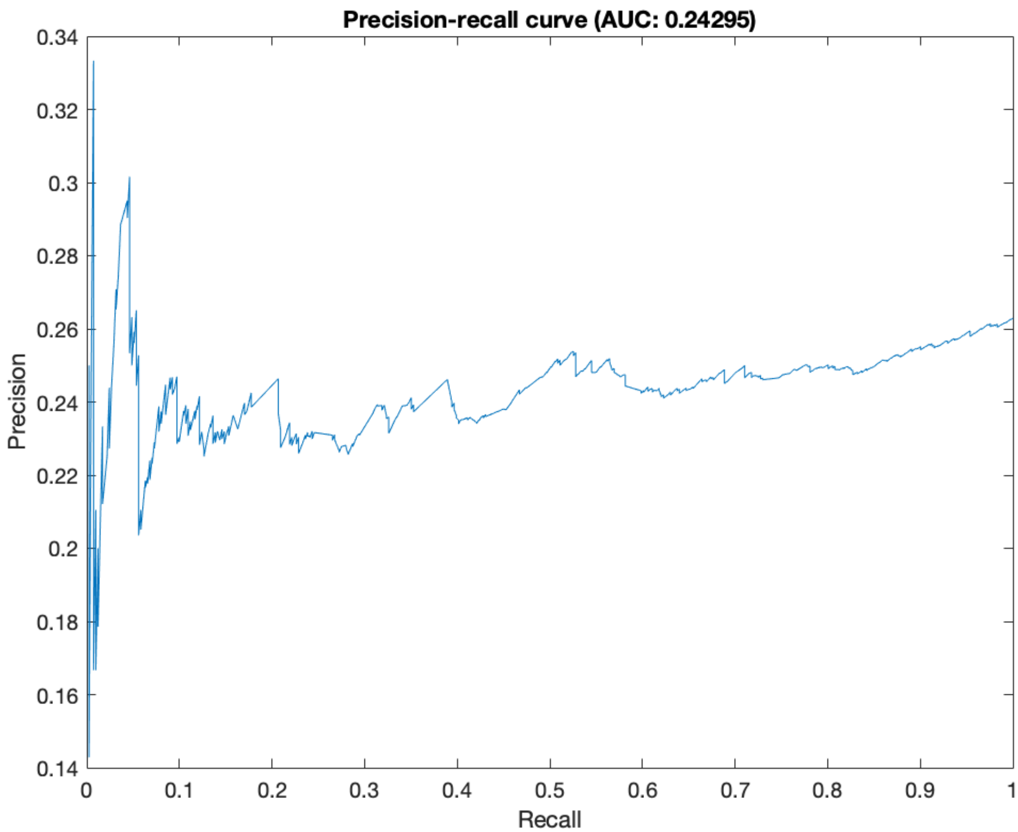

Figure 3: PR Curve

Summary at Classification Threshold = 0.75	
Precision	0.229473684210526
Recall	0.265206812652068
F-measure	0.246049661399549
Accuracy	0.572890025575448

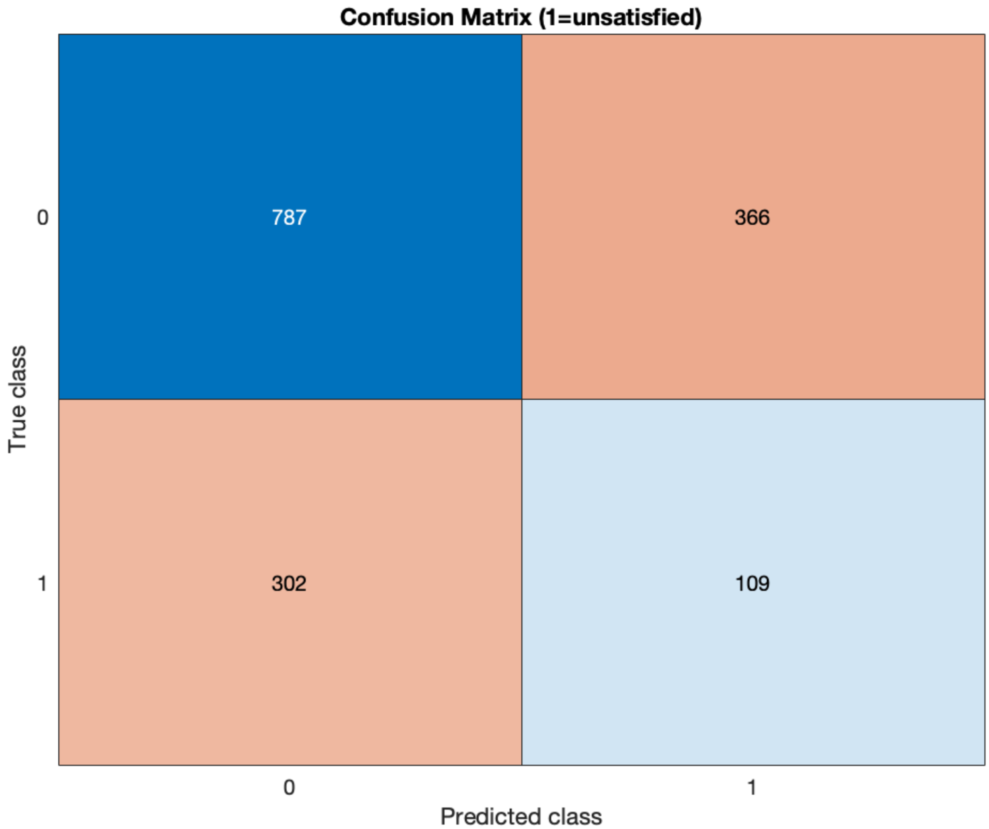

Figure 4: Confusion Matrix

Model Parameters [θ]:

$$INTERCEPT : \theta_0 = -0.3$$
LANG_ENGLISH: $\theta_1 = 4.9$
MARRIED_YN: $\theta_2 = 0.1$
RACE_WHITE: $\theta_3 = 0.4$
RACE_BLACK: $\theta_4 = 0.1$
RACE_OTHER: $\theta_5 = 0.6$
AGE_0_17: $\theta_6 = 298175707235692$
AGE_18_34: $\theta_7 = 0.9$
AGE_35_49: $\theta_8 = -0.3$
AGE_50_64: $\theta_9 = 0.7$
AGE_65_79: $\theta_{10} = 1.1$
AGE_80_PLUS: $\theta_{11} = 0.2$
GENDER_FEMALE: $\theta_{12} = 0.3$
GENDER_MALE: $\theta_{13} = 1$
GENDER_OTHER: $\theta_{14} = 298175707235691$
PAYOR_MEDICAID: $\theta_{15} = 1.3$
PAYOR_PRIVATE: $\theta_{16} = -0.3$
PAYOR_MEDICARE: $\theta_{17} = 0.7$
PAYOR_SELFPAY: $\theta_{18} = 298175707235691$
PAYOR_OTHER: $\theta_{19} = 298175707235692$
ED_ADMIT: $\theta_{20} = 0.1$
ADM_CLINIC: $\theta_{21} = 1.5$
ADM_NONHEALTHCARE: $\theta_{22} = -0.2$

ADM_TRANSFER: $\theta_{23} = 0.3$
SVC_MED: $\theta_{24} = 0.3$
SVC_CVM: $\theta_{25} = 1.1$
SVC_ORT: $\theta_{26} = -0.5$
SVC_HSP: $\theta_{27} = 1$
SVC_GER: $\theta_{28} = 0.6$
SVC_FAM: $\theta_{29} = -0.4$
SVC_NSU: $\theta_{30} = 0.5$
SVC_GIS: $\theta_{31} = 0.4$
SVC_URO: $\theta_{32} = 0.4$
SVC_CSU: $\theta_{33} = 0.8$
SVC_PUL: $\theta_{34} = 0.9$
SVC_POD: $\theta_{35} = 1.1$
SVC_THO: $\theta_{36} = 1.4$
SVC_VAS: $\theta_{37} = 1.2$
SVC_PLS: $\theta_{38} = 0.5$
SVC_CRS: $\theta_{39} = -0.2$
SVC_SUR: $\theta_{40} = 1.6$
SVC_HEM: $\theta_{41} = 1.6$
SVC_NEU: $\theta_{42} = 2.4$
SVC_OTHER: $\theta_{43} = 1$

7 Conclusion

The model was not able to provide reliable estimates of whether a patient would be dissatisfied with their care at the time of discharge (Accuracy = 57.3%). While there were likely many limiting factors, the key point of failure was the lack of rigor in the feature evaluation and selection process. Logistic regression is fairly straightforward to implement and train, but will not account for feature interactions and performance suffers when irrelevant features are pulled in.

8 Future Work/Extensions

As I see it, there are two options for future work in this area: 1) build a rigorous causal model using all relevant data and iteratively refine the feature set with a group of expert clinical/administrative stakeholders, or rather than predicting with the intent to intervene, calculate an Observed/Expected ratio to be used in a performance management context. Tuning the model to get the correct ratio of dissatisfied patients might prove far easier than predicting exactly which patients were dissatisfied. Another more practical alternative could be to circumvent the prediction process altogether by collecting real-time patient feedback during the hospitalization.

9 Bibliography

1 https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Hospital_VBPurchasing_Fact_Sheet_ICN907664.pdf

2 https://www.netpromoter.com/know/

3 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3217162/

4 https://onlinelibrary.wiley.com/doi/abs/10.1002/jhm.2371

5 https://www.ncbi.nlm.nih.gov/pubmed/27228056

6 https://www.nejm.org/doi/full/10.1056/NEJMsa0804116

7 https://www.hcahpsonline.org/globalassets/hcahps/summary-analyses/correlations/report_april_2018_corrs_pain_removed.pdf

8 Figure 5 - 2013 UHC conference - Ohio State Wexler Approach

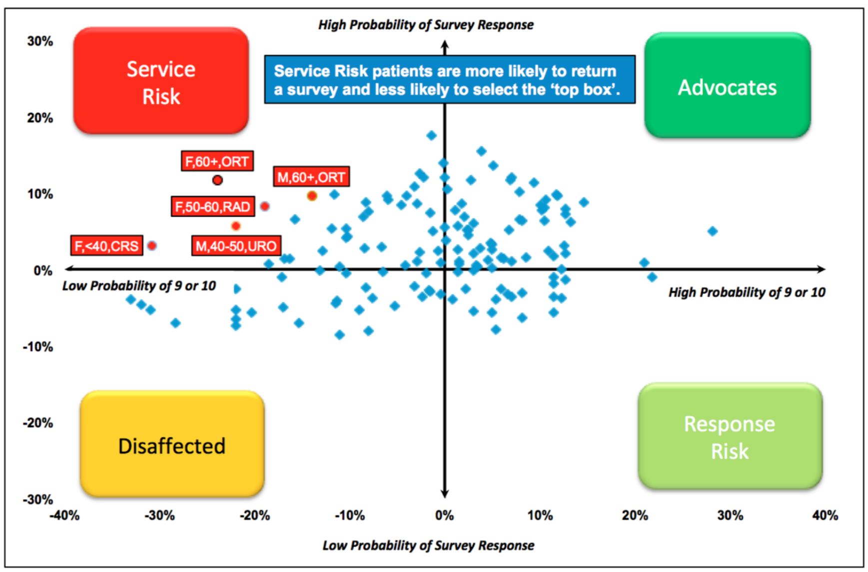

Figure 5: Confusion Matrix