**SURGEON EFFICIENCY AND QUALITY METRICS REPORTING FEATURING A MACHINE LEARNING APPROACH: AN INTEGRATED SCORE CARD**

**Anthony S. Fargnoli, PhD**

**Background and Significance:**

Modern healthcare systems in the United States are under ever increasing pressure to improve quality, patient experience ratings, and reduces overall costs to maintain sustainable profit margins. Costs increases which have been passed on to payors and individual consumers has reached an unsustainable level. Health systems in response have focused on streamlining their best profit generating units to survive this new climate. Surgery, in particular cardiothoracic surgery departments have been one of the larger profit generating centers for health systems. In fact, some heart surgery departments account alone for greater than 40-50% net revenues. Cardiac surgery is a complex business, where strict quality standards, patient satisfaction scoring, and most recently managed care have determined key financial performance incentives. As the insurance and government program payors such as MediCare have evolved, heart surgery centers must meet certain quality and other performance objectives in order to operative and sustain robust profits to compensate for the losses in other hospital units. Thus, the role of department Chair in these centers to assess and implement changes in response to these metrics has never been more paramount.

Health systems in general have invested substantial capital in analytics and reporting system via various IT vendors which have dominated healthcare such as EPIC/Clarity and others. Quality and hospital operations divisions within these corporate units in health systems have been tasked with providing department Chairs and individual surgeons their specific impact on these metrics. Despite many years of progress and new metrics developed, there still remains a challenge in breaking down complex analytics and implementing the changes on the front line of care such that a surgeon can understand how he or she is performing on both a short and long term basis.

Department Chairs have cited these problems with utilizing the information they receive to adapt and utilize to change surgeon behavior in favor of adopting certain practices and techniques that would increase the value of care. Value is defined as achieving high quality patient outcome results at the lowest reasonable cost, while also achieving patient satisfaction scores. This balance is very difficult in a high pressure cardiac surgical environment where many factors can change these metrics, for better or worse, many are ineffective and not integrated as a department would envision as ideal.

This body of work represents a real case study for a large New York City cardiothoracic surgery department addressing the need for a novel approach to department reporting from previous analytic approaches that did not meet needs. A machine learning approach was offered and implemented, as well as current hospital metrics evaluated with data science tools. This contribution is one of many ongoing in cardiac surgery to understand the relationships between cost, quality, field established metrics and practices, and individual surgeon behaviors.

**Quality Metrics in Practice**

Health systems have both operations and quality departments that manage data reporting both locally to the service unit and national databases. This system has existed for years in cardiac surgery in an effort to improve quality. In terms of quality alone, any cardiac surgery department from the Society of Thoracic Surgeons (STS) database, can track its performance against its peers regionally and internationally. Field agreed upon metrics such as freedom from death within 30 days, stroke, and other complications have been standardized. Departments have quality officials handle this data and report it quarterly, and a Department Chair can disseminate this information and identify surgeons that are struggling to meet set objective quality per procedure. These are monitored closely since Medicare and other large payors now link revenue payments if certain quality milestones are reached. In other instances, programs can be penalized or closed if they do not meet basic quality.

**Efficiency Metrics in Practice**

Modeling hospitals as a business is a difficult problem, but a reality that must be managed prudently by upper management. Since surgeries generate higher proportions of net revenue, efficiency in these lines are demanded. Every hospital has a limited number of operating rooms that seek to maximize the efficiency or number of procedures per room. Other metrics such as reducing length of stay to free more beds to accommodate more patients in a timeline are also common. In the context of cardiac surgery, individual surgery times, length of stays per patient, any complications that increase length of stays or costs are now tracked closely. Hospital operation divisions manage this data and report it. The analytics involved are very sophisticated with many other metrics that seek to change surgeon behavior. This typically generates a conflict of interest, where surgeons and departments disagree with certain metrics that do not take into account the full scenario of per case basis and or create incentives to increase risk at the cost of quality. The operations and quality departments, by law are separated in the hospital but both interface with the performing surgery unit.

**Variable Direct Cost and Revenue**

The ability of any individual surgeon to understand how his or her procedures impact cost structure in the health system is an extremely difficult problem. Revenue derived from private insurance, changes in government payor fee schedules, changes in negotiated supplier device/drug prices, and many other factors such as case complexity can result in a cost structure that is highly variable and poorly understood since many of the elements are not transparent. Prudent health systems have adapted by monitoring select direct variable costs linked with a particular surgery and established a series of benchmarks to understand what a procedure costs to hold a procedure vs. the revenue it produces. Once quality is met as outlined above, the only way to maximize margin is to reduce the variable costs as best as possible. In this case study, variable direct costs were assessed per: Total operating room time used, supplies including closure devices/drugs, anesthesia times, and length of stay formulas where if an expected stay is violated additional charges hit the cost center.

**The Problem Statement**

**Cardiac surgery centers have demanded a reporting system that would integrate each of these categories of quality, efficiency, cost, patient satisfaction, into one readable report that is system specific for a group of practitioners.**

*Problems with existing systems:*

1. Heart surgeons are not IT professionals and do not understand complex dashboards or analytics in vendor provided software
2. Large and conflicted operations departments are focused on profit margin, and do not fully understand quality metrics or what changes a surgeon can make safely to increase efficiency and reduce costs
3. Surgeons tend to put their patients and reputation above adopting behaviors that would increase profit slightly; aggressive defensive behavior against non MD business and IT professionals with metrics
4. Major conflicts between department Chair and individual surgeons occur when conflicted or unrealistic conclusions are made from the data that is not relevant to actual practice
5. Surgeons are competitive perfectionists, and do not like to be ranked against peers that have different case loads or less risk patients.
6. Conflicts between the surgical unit, quality, and operations result in winner take all approach game of elimination as exists in all large health systems today

**Proposed solution:**

An integrated cost, quality, and efficiency data structure clearly communicated in a quality system surgeons understand and would adopt. Such a system would use machine learning approach and integrate data from each of these 3 sources reducing the need for separate conflicting reports.

**Case Study History and Problem Statement:**

One such prior attempt at an integrated data approach was tried by management at this particular health system with lackluster results. The primary complaint was that while the executive team thought the system would improve metrics, it was found out that it did not and surgeons did not understand the quantitative scoring.

A SEQI or “Surgeon efficiency quality index”, was developed, and generated a single score for each surgeon based on integrated metrics shown in the table below:



Figure 1. Overview of quality, efficiency and coronary artery bypass (CABG) open heart surgery specific metrics to score individual surgeons.



**Figure 2. The SEQI score rubric compilation of available assigned points per data field, maximum possible score 191 with higher rating meaning higher performance.**

The SEQI system implemented by hospital operations was not viewed favorably and did not achieve its objectives to provide surgeons with a number they could use to understand how their procedures performed. These problems were noted with a single score system:

1. A single SEQI score weights each surgeon against peers, but not national benchmarks of quality many of which data points were left out for selected biases.
2. The score flaws – redistribution of “winner take all” point distributions resulted in numbers that did not stratify individual surgeon performance on quality and cost.
3. The CABG specific procedure elements were motivated by behavior, and many non functional if 100% of a medication taken and nearly all procedures had it, didn’t add value to final scorings.
4. Much of the system measured noise, for example if one surgeon had difficult risk adjusted case quality defects vs. others that had easier or less cases, scores did not reflect it. Simply evaluated, the surgeon performing the most given procedures had an advantage since many of the formulas had flawed % percentage weighting and not a database standard.

**METHODS:**

A machine learning database approach with historical data integrated with a quality scoring system surgeons could buy in and understand was a proposed solution explored in this reporting. The Society of Thoracic Surgeons has established a Program rating star scale system of 1-3, which every two years rates a full program on its quality alone metrics. Surgeons however clearly understand this system, since the majority of cardiac surgery programs in the USA are two star out of three, with only 15% of programs reaching the highest honor of 3 star quality. 1 star programs are typically placed in probation to increase to 2 stars and else shut down by regulatory agencies if improvements are not made.

The overall strategy of the machine learning approach was based on the following elements:

1. STS quality elements of the 5 major categories of 30 day free mortality, freedom from stroke or neurological complications, no 30 day readmission or any cause, no and major internal chest infections or kidney failure. These are the major complication events that reduce payment.
2. Select efficiency metrics from the hospital operations reports including observed/expected length of stay, operating room time, discharge rates and others that impact time.
3. Utilizing direct variable costs specific to the procedure in terms of supplies, drugs, and time used for special equipment/staff; related to surgeon performance per case
4. Defining a new metric report system where each surgeon would receive a “Star” number distribution report on EACH of their cases, rather than a single score

The problem with SEQI was that it was essentially an aggregated compilation of many cases per term of time, but in reality much of what happens in terms of quality, cost and efficiency is very case specific. Therefore, surgeons can reflect back more accurately on a case if it is rated that way, rather than scored to an aggregate score that means nothing, nor would contribute to their change in approach.

**The STS Star system, we proposed could be revamped for individual case performance with machine learning classification.**

In this case example a meeting between Cardiac Surgery leadership, Operations & Financial committee, and Quality merged to define what would comprise a 3 star classification in terms of criteria. It was agreed:

1. Three Star Rating would include having perfect STS quality record with no adverse events + variable costs to not exceed 60-75th quartile. Efficiency scores in terms of length of stay and operating times would be less than 50 percentile of total cases. To model actual STS rating, since surgeons know that only 15% of centers reach this status, a maximum of around this target was sufficient for this high rating while all others meeting criteria, but exceeding this count would be categorized as a Two Star.
2. Two Star Rating – free from death and readmission within 30 days, but could include 1 or more quality issues if and only if costs were maintained below 50th percentile. If perfect quality but high costs, cases would be in this category. If one or more quality issues with high cost it would advance to borderline One Star decided by efficiency metrics. It was expected the majority of procedures would fall in this category.
3. One Star Rating – this indicates that the procedure had one of the following defects: A. Death in the operating room B. Death within 30 days post operation C. Re-operation for failed results D. Costs with an event in A B or C that exceeded 75th percentile plus efficiency metrics above 50 percentile.

Several key points were to penalize deaths and re-operations since these would increase other costs not shown by other metrics. This was not capture in the previous SEQI system and was paramount in the new approach.

Technical setup description for the machine learning approach:

1. The Star criteria was applied as the target feature and used to categorize over 900 previous cardiac surgeries performed by a group of 4 individual surgeons who had significant case experience to adjust for risk and experience factors.
2. A basic risk score 1-7 was implemented to adjust the machine for risk with basic medical history risk factors such that cases were not biased from case risk.
3. The procedure of interest is a well-controlled, isolated coronary artery bypass (isoCABG) case where the patients have little to no co-conditions and the procedure is expected to have less than 1% mortality.
4. The goal was to fit various machine learning multi-classification models and optimize parameters and scoring
5. Test train split strategy was adequate for 933 recent cases for 4 surgeons to assess by group and per individual.
6. Upper management also requested feature evaluation of a black box CMI/VDC metric that had a complex model that would compute variable costs as a function of revenue and a variety of other non disclosed metrics the hospital thought was prioritized to evaluate margins.
7. Management also had concern over its HCAPHS “Patient Experience” reporting score, where new market survey analysis found inconsistent data with scoring.

SQL, R and Python were used to complete the data science evaluation for this proposal.

**RESULTS**

A total of 3 data sources (Quality, Costs/Efficiency and Patient Experience reporting) were joined in SQL into one merged table by patient record and surgery date.

| **Surgeon** | **Year** | **Month** | **Age** | **Gender** | **Risk** | **30d\_Death** | **ReOperation** | **Prolonged** | **Infection** | **Renal** | **Stroke** | **STS\_Quality** | **HCAHPS** | **OR\_Time** | **LOS\_Ratio** | **CMI\_VDC** | **Total\_Costs** | **Stars** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ABC | 2018.0 | 1.0 | 80.0 | M | 2.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | Failed | 4.0 | 447.0 | 0.86 | 4062.1 | 60296.7 | One |
| XYZ | 2018.0 | 1.0 | 59.0 | M | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | Failed | 3.8 | 524.9 | 0.93 | 2605.0 | 70776.3 | One |
| XYZ | 2018.0 | 1.0 | 85.0 | F | 4.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | Failed | 3.8 | 286.5 | 0.80 | 5509.3 | 59157.3 | One |
| HJK | 2018.0 | 1.0 | 54.0 | M | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | Failed | 3.8 | 501.7 | 0.96 | 5584.2 | 57581.2 | One |
| DEF | 2018.0 | 1.0 | 69.0 | F | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | Yes | 4.0 | 375.1 | 0.61 | 5041.1 | 55579.6 | Two |

Table 1. Variables merged from 3 separate data sources including 2 hospital and one national quality database.

Surgeon and patients were de-identified. Month was used as a variable 1-12 for seasonal effects. Basic demographics of the patient such as age, gender were used for the Risk model field via computation of cardiac surgery risk factor CHAD2VASC score. This score is a 0-7 scale stratifying patients dying from cardiac surgery based on pre existing risk. The next five elements are binary and report if an STS quality adverse event of either:

A. Death within 30 day post surgery or on the day of surgery

B. Re-operation for any cause in length of stay

C. Surgical site infection, which carries a heavy financial penalty

D. Renal failure as defined by requiring dialysis and blood panels

E. Stroke incidence after surgery

STS Quality field was a binary variable used in the model as was a “Yes” success if and only if the patient experience was free from any of A-E defects.

Efficiency data points of total operating room time from initial cut to patient wheeled out of the room and the observed/expected length of stay ratio from the hospital logs were used. LOS ratio of 1 means the patient stayed as expected, less is better with 0.75 or lower ideal indicating the health system moved the patient out of a bed faster with the same quality.

Costs – Variable direct costs from the unblinded cardiac surgery department costs records adjusted for non surgery related expenses were implemented for “true cost per case” assessment. The “black box” CMI\_VDC metric cost was a test metric hospital operations gave for each case and wanted to determine if it would be in any way matched with quality and other metrics. It was not disclosed to the data science team how it worked, or how it was developed.

Patient Experience Quality Score (HCAPHS) – the health system developed an inpatient scale 0-4 questionaire from family members to report their family member’s experience focused on communication of important case updates during the patient’s length of stay. While this metric does not directly influence revenue or costs, it is something management would like to incentivize surgeons to track their scores and see if it would have an impact on other metrics.

**Exploratory Data Analysis:**

Since costs were of major concern in this initial analysis an initial describe:

Cases 933

mean $59458

std $14139

min $10451

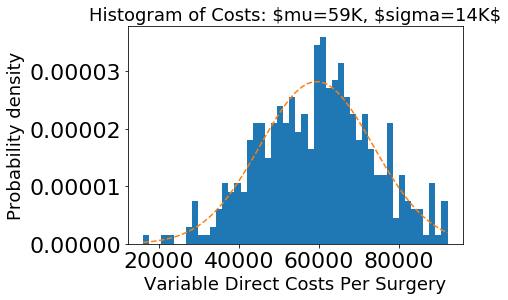
25% $49873

50% $59109

75% $69087

max $100638

The average procedure costs around $60,000. The team was not provided how much revenue each case generated, it was only understood that quality defects and inefficient would simultaneously increase costs and reduce revenue, hence double or single hit to margins.



**Figure 3. Histogram of Variable direct costs per case**

As shown in Figure 3, variable direct costs were roughly normally distributed as expected.

Since it was important to management to understand their quality ratings by surgeon the HACPHS score distribution were plotted by group and by individual surgeon:

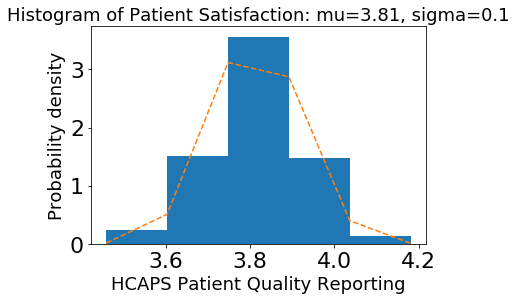
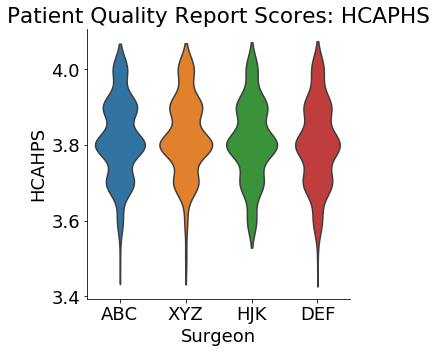


Figure 4. HCAPHS Patient Quality Distribution

The range of scoring was low between 3.4 -4.2 with the majority of scores at 3.8. Upon further analysis there was not much difference between individual surgeons:

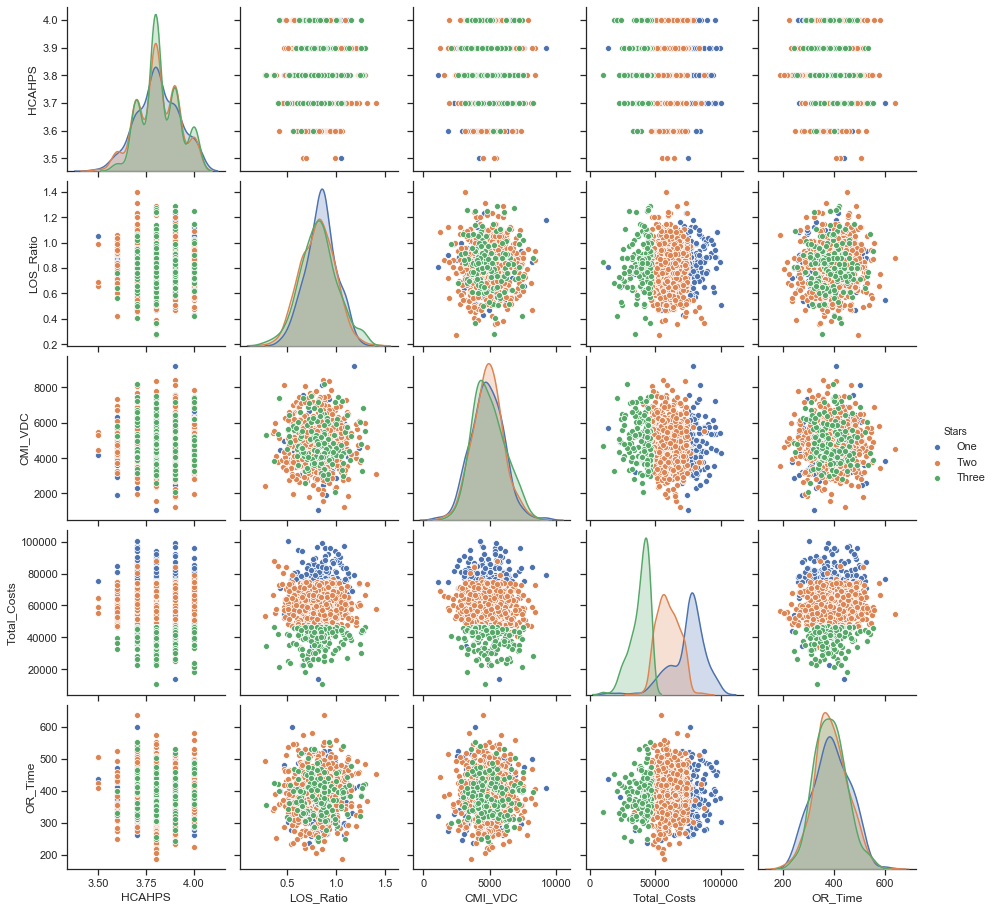


**Figure 5. Violin plot of HCAPHS distribution scores by surgeon.**

Based on the analysis two conclusions could be made: A. Surgeon “HJK” did not receive any low scores less than 3.5 which could indicate he or she performed the best of the group. B. The sensitivity of the measure was not sufficient.

The HCAPHS as suspected by analysts in the operations unit thought that the HCAPHS system did not have enough sensitivity to capture a wider range of patient experience. The data supports the recommendation to widen the scale beyond 0-5 and or ask more detailed questions with a natural language processing solution.

The next step was to plot all variables as a function of star rating to perform a baseline assessment of visual classification based on grouping:



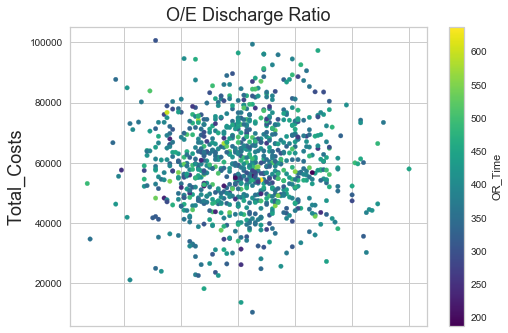
**Figure 6. Scatter plot distribution of model predictors on One (blue dots), Two (orange dots), and Three (green) Star heart procedure classifiers based on management rubric.**

Several interpretations:

1. HCAPHS quality scoring appeared to have somewhat higher trend at Three stars, but not definitive.
2. LOS\_Ratio efficiency metric against variable costs demonstrated a solid 3 tier classification in department variable costs, but not the CDI/VDC metric.
3. The CMI VDC did not have a solid trend in LOS\_Ratio, OR\_time efficiency metrics
4. OR\_time against total costs as a function of star demonstrated ability to classify

**Efficiency Analysis**

A 3 variable plot of LOS\_ratio and OR\_time against variable direct costs per OR time was developed:



**Figure 7. Discharge Ratio as a metric including LOS\_ratio and length of stay data was used to determine visual distribution of total costs on the y axis as function of OR\_time.**

It was determined here that operating room time is a difficult variable to model and predict other events that influence discharge. This was a point the surgical team made to management because often times there are elements such as unexpected complications, delays, other emergencies and teaching obligations that can extend OR\_time, but have no bearing on quality factors or the efficiency metrics achieved later in the discharge process.

Once additional count charts were submitted as preliminary analysis the corporate team was anticipating the machine learning fits to the 933 cases which served as the training data.

**Machine Learning Applications**

The overall strategy was to fit linear and more simplistic models and subsequently advance toward more advanced models. The target variable as described previously was the Categorized Star rating (i.e. One to Three) for each procedure. Each of the remaining elements were assigned as predictors with all categorical variables hot encoded.

After conducting test/train split sequence these were the models applied to the training and test set for selection and validation:

*Baseline models*

logit = linear\_model.LogisticRegression()

LDA = discriminant\_analysis.LinearDiscriminantAnalysis()

QDA = discriminant\_analysis.QuadraticDiscriminantAnalysis()

GNB = naive\_bayes.GaussianNB()

MNB = naive\_bayes.MultinomialNB()

BNB = naive\_bayes.BernoulliNB(binarize=1.5)

The printed results were revealing that indeed the data could be modelled with a LDA approach:

| **Model** | **Train /Test Scores** |
| --- | --- |
| **Logitistic** | 0.873995 | 0.909091 |
| **LDA** | 0.935657 | 0.909091 |
| **QDA** | 0.376676 | 0.390374 |
| **GNB** | 0.875335 | 0.860963 |
| **MNB** | 0.624665 | 0.631016 |
| **BNB** | 0.658177 | 0.652406 |

The LDA analysis was not ideal since it could not identify variations in certain cases therefore it was decided to advance to a random forest approach.

*Random Forest*

The random forest was fit readily to the data and yielded scores in the 83-87 range with cross validation indicating some potential issues with bias/variance. Despite initial attempts to Grid Search optimize, a shift toward boosting algorithms was implemented to compare if it was worth further investigating a more complex random forest.

*Gradient Boosting*

The initial results from the gradient boosting classifier were robust compared to the random forest.

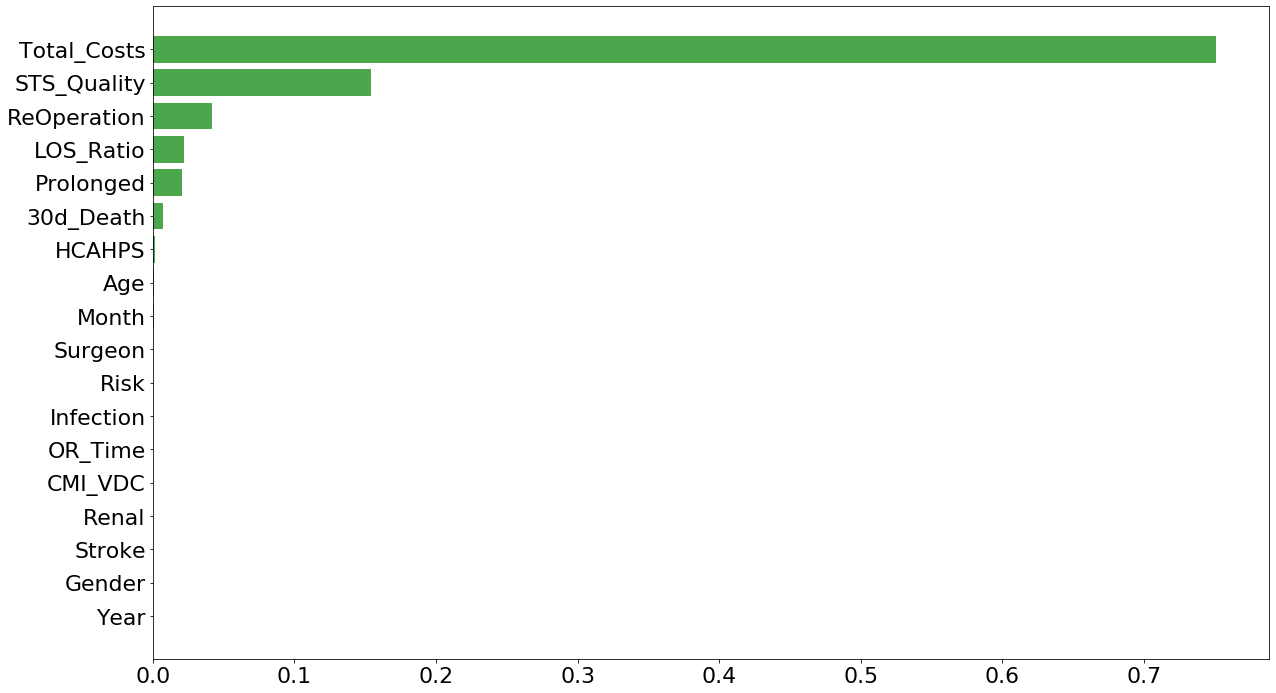
The Training Score is 1.000

The Testing Score is 0.984

This indicated that this model definitely adjusted to additional variations, however appeared that it could be overfit. After adjustment to the learning rate via GridSearhCV, the new model had an improved cross validation score with reduced variance. The profile shown below:

5 Fold Cross validation scores:([0.98, 0.99,0.99, 0.99, 1.00])

The final model selected was the Gradient Boosting Machine demonstrated above with a learning rate of 0.11, only a slight modification from the first iteration. This model met all of the criteria the committee deemed needed to run future analysis.



**Figure 8. Feature importance rank from the final model selection of gradient boosting.**

The total variable costs from the department was the primary feature deciding the majority of cases with by far the highest weighting followed by STS Quality and the need for any Re-operation. This validated what the committee wanted to see in the data since it was determined that scenarios with very high costs could not be awarded a Three star even if all other metrics were at the top tier. Vice Versa, a registered death with low costs or the need for re-operation would penalize a star rating. The previous SEQI scoring system failed to account for these scenarios, therefore gradient boosting machine met the needs.

It was interesting to find that HCAPHS, which was suspect as not having a role in quality or efficiency or costs wass validated with GBM. Other background variables such as month operated, age of the patient, any particular surgeon, and others did not have importance.

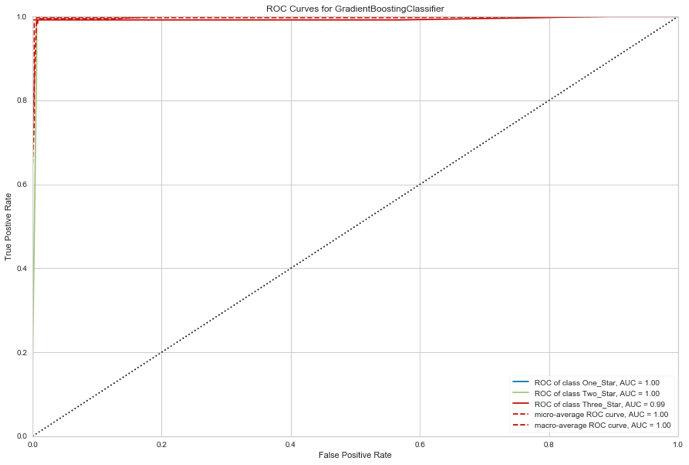


Figure 9. ROC Curves for the Gradient Boosting Classification of Procedure Star Ratings

The YellowBricks package in Python was utilized to construct the nearly perfect AUC/ROC curve demonstrating impeccable ability of the GBM to separate One, Two, and Three star procedures. Additional analysis shown in the table below validated excellent F1 scores and others of precision and recall, thus validating the model selection was superb.

Precision Recall F1-score Support

One 1.00 0.94 0.97 36

Three 0.97 1.00 0.98 29

Two 0.98 0.99 0.99 122

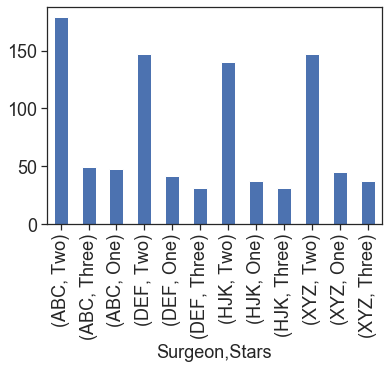
accuracy 0.98 187

macro avg 0.98 0.98 0.98 187

weighted avg **0.98 0.98 0.98** 187

*Model Performance on Additional Data*

The gradient boosting model was then applied to an additional 930 sample case list from 2019-2020 to re-validate results and plot additional insights from the proposed Star system. The goal at this point was to generate the sample report cards per surgeon.



**Figure 10. Surgeon report card by star rating**

Shown in the figure above is the number of procedures during 2019-early 2020 period each surgeon achieved in terms of one, two and three star cases. Not shown, a subreport of one star cases was generated such that the Department Chair could show the identified cause whether it was a typical STS quality or excess variable direct cost or a combination thereof. This system would be anticipated to movitate physician behavior to adopt measures to reduce future one star cases with identified cause during case reviews.

An additional simple score metric was the ratio of one to three star cases to track improvement as shown in the example below:

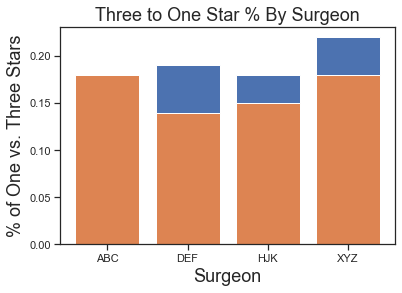


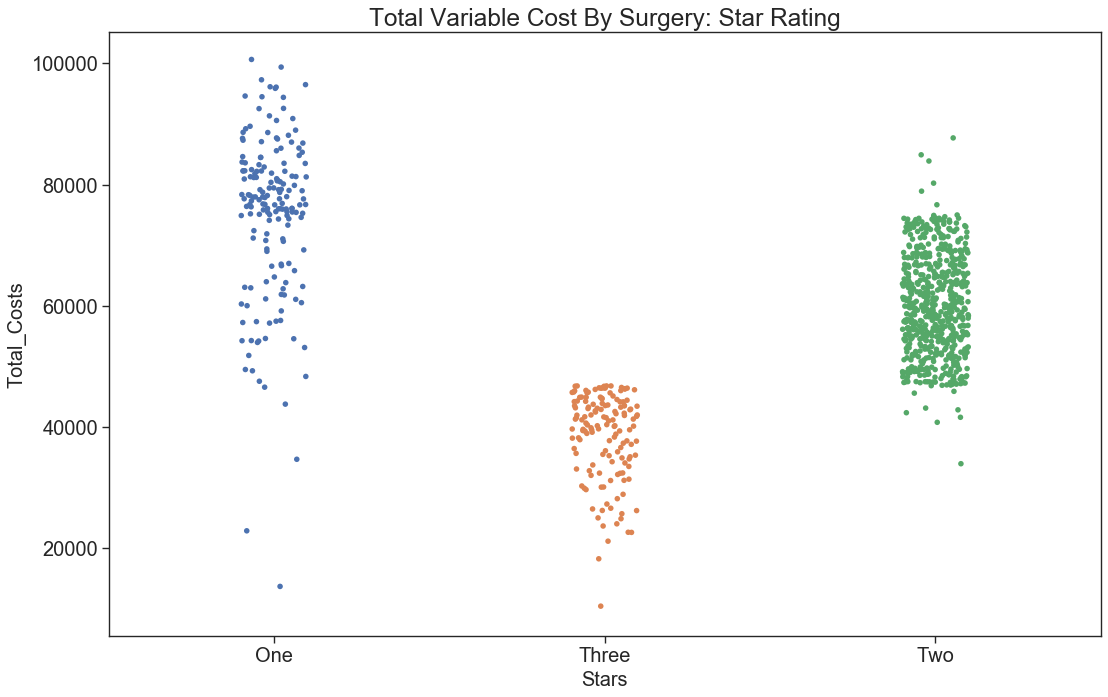
Figure 11. Per Surgeon report of % of Three star (orange) vs One Star (blue) procedures.

In this simple chart above, Surgeon ABC had more Three Star cases than One Star cases indicating the orange bar. This surgeon scored the best amongst peers as the other DEF, HJK and XYZ had slightly higher rates of One Star cases. This approach could identify quarterly or annual trends in improvement in this metric if there was an issue OR could be used to reallocate surgeons that typically perform two to three star ratings on one particular procedure, and reallocate them to another if the trend persists. Since this variable encompasses all metrics, there wouldn’t be a need for the Department Chair to analyze cost, quality and other reports separately.

**Conclusions and Future Direction**

This body of work was a compilation of a complex problem that was broken down systematically to identify how machine learning could be applied to solve a real world problem in healthcare. The methodology demonstrated the ability of machine learning to reduce several complex data sets and output into a readable format that three disparate stakeholders of surgeons, hospital executives, and quality program management professionals could agree to adopt.

The model however still has its limitations being too focused on variable direct costs. As shown in the final graphic below, it is clear that in the model and the data sets in decision making variable costs were the main driver. However, this was the reflection of management goals beyond quality, in that value based care demands exceptional quality all the time, but at reduced cost.



**Figure 12. Impact of variable direct costs as a function of star rating demonstrating clear stratifications.**

The future direction of this application is to repeat for per procedure type. Recall this initial iteration only performed the analysis for isolated coronary artery bypass procedures. Cardiac surgeons perform additional procedures that are more complex, including:

* Aortic valve replacements
* Minimally invasive robotic surgery valve repairs
* Heart transplants
* Mitral valve repairs
* Mitral valve replacements
* Aortic dissections
* Complex: Coronary artery bypass grafts multiple + valve replacements

Each procedure above has additional quality and cost considerations. Therefore, in a complied report per surgeon it would be expected the stars would be distributed per surgery and not as a whole. This system ideally would then permit hospital administration and department Chairs to identify system or surgeon specific problems per procedure.