Peer-graded Assignment: Final Assignment

You passed!
Congratulations. You earned 10 / 10 points.

Data Science Methodology – Hospitals

Which topic did you choose to apply the data science methodology to? **(2 marks)**

Hospitals

Next, you will play the role of the client and the data scientist.

Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. **(3 marks)**

You are required to:

1. Describe the problem, related to the topic you selected.

Nosocomial infections or health care-associated infections (HAIs) are infections acquired while in the hospital for another condition. “These infections lead to the loss of tens of thousands of lives and cost the U.S. health care system billions of dollars each year.” https://health.gov/hcq/prevent-hai.asp “The U.S. Department of Health and Human Services (HHS) announced new targets for the national acute care hospital metrics for the National Action Plan to Prevent Health Care-Associated Infections: Road Map to Elimination (HAI Action Plan) in October 2016. The targets use data from calendar year 2015 as a baseline — and are in effect for a 5-year period from 2015 to 2020. These new targets replace the previous targets that expired in December 2013. These target goals for reduction of health care-associated infections (HAIs) are ambitious, but achievable.” https://health.gov/hcq/prevent-hai-measures.asp.

National acute care hospital metrics for the HAI Action Plan are collected for several HAIs. This assignment is limited to infections acquired for central line-associated bloodstream infections (CLABSIs). (A Central line (CL) is defined as “an intravascular catheter that terminates at or close to the heart, OR in one of the great vessels that is used for infusion, withdrawal of blood, or hemodynamic monitoring.” https://www.cdc.gov/nhsn/pdfs/pscmanual/4psc_clabscurrent.pdf).

The fictious hospital for this assignment has met the previous HHS target goals, and now is required to meet or exceed the new, more ambitious goals for 2015 to 2020.

The 2020 target (from the 2015 baseline) for CLABSI is a 50% reduction. https://health.gov/hcq/prevent-hai-measures.asp

2. Phrase the problem as a question to be answered using data.

Is the infection rate for CLABSIs in our hospital ICU on track to meet or exceed the HHA goals for 2015 to 2020?

Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with. **(5 marks)**:

1. Analytic Approach
2. Data Requirements
3. Data Collection
4. Data Understanding and Preparation
5. Modeling and Evaluation

1. Analytic Approach: Use a decision tree classification system to determine risk of CLABSI in the ICU. Factors that contribute to patients getting a CLABSI would be nodes in the tree. The probability of each factor or node contributing to the outcome can be determined leading to the predicted risk of getting a CLABSI.

2. Data Requirements: The sample population of patients includes all patients in the ICU. Collect initial data on all patients that enter the ICU, even those who do not have a central line as it may be necessary to insert one during treatment. One record per patient that includes all data collected, per the CDC guidelines (https://www.cdc.gov/nhsn/pdfs/pscmanual/4psc_clabscurrent.pdf), upon initial admission to the ICU through discharge from the ICU, is required.

3. Data Collection: Include information about patient demographics such as age, the presence of a central line, blood work and other tests that indicate the presence of pathogens, symptoms, site(s) of infection and relevant procedures both past and present.

4. Data Understanding and Preparation: Since the hospital has been involved with HHA national acute care hospital metrics in the past, including the application of CDC guidelines, many of the data elements have already been collected, automated, and used to access performance. With the new more ambitious targets the existing data and system will need to be evaluated to determine if more data is required, if data is no longer relevant, or if other modifications are required. Also, the CDC guidelines need to be revisited to see if there were any changes that need to be addressed. And finally, feedback on the system in place requires evaluation to determine updates or changes that need to be made.

   To understand the data, statistics on the data such as mean, median, minimum, maximum, and standard deviation need to be calculated. Pairwise correlations on the data are important to understand which data items are highly correlated versus not correlated at all. Data Visualization such as histograms, and scatter plots are important to evaluate the distribution of data values.

To prepare the data, use the CDC guidelines to define central line, define what the indications are for an HAI, and define the relevant time period for data collection.

All transaction records for each patient are to be aggregated into one record. These records include lab reports while in the ICU, physician records before admittance to the ICU and while there, vitals reports while in the ICU. Also included in the aggregation of data is demographic information.

In addition, the HAI Action Plan as well as the CDC guidelines need to be evaluated to determine if more data items are required, or if data are required in a different way. A literature review may also be conducted to learn of new advancements in the field which may be advantageous to track, and therefore, new data items added to the record. All this data needs to be merged into a table with columns of patient attributes and one record or row per patient. The dependent variable is the presence of a CLABSI, either yes or no, for each patient. The patient records were randomly divided into a training set and a test set to train and test the model.

Through data understanding and preparation it can be determined if there are extraneous data variables, missing data variables, or data variables that can be consolidated.

5. Modeling and Evaluation: The outcome of interest is CLABSI equals ‘yes’. The chosen model is evaluated to determine the proportion of correctly classifying patients as having a CLABSI. The goodness of the model is determined by the proportion of false positives and false negatives the model predicts. A false positive (type I error) is when the patient does not have CLABSI, but the model determines the patient does – the model says 'yes', when it should say 'no'. A false negative (type II error) is when the model says the patient does not have a CLABSI when in fact the patient does – the model says 'no', when it should say 'yes'. The goal is to minimize both. Since the goal is to reach a lower rate of CLABSI a model that incorrectly increases the rate is less desirable since the hospital met its goal but the model says it did not and therefore may incur some penalty. Also undesirable is a model that makes it appear that the hospital has met the target or come close to meeting the target when it has not, meaning more patients are actually getting CLABSIs than the model shows. To determine the optimal model, the ROC curve can be used to visualize which model best classifies CLABSIs given the true positive rate versus the false positive rates calculated by each model.