The Pennsylvania State University

The Graduate School

Department of Industrial and Manufacturing Engineering

PREDICTIVE ANALYTICS OF PATIENT HOSPITAL ADVERSE EVENTS AFTER COLORECTAL SURGERY

A Thesis in

Industrial Engineering and Operations Research

by

Sijia Guo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2015

The thesis of Sijia Guo was reviewed and approved* by the following:

Soundar Kumara
Allen E. Pearce/Allen M. Pearce Professor of Industrial and Manufacturing Engineering
Thesis co-adviser

Guodong Pang
The Harold and Inge Marcus Career Assistant Professor of Industrial and Manufacturing Engineering
Thesis co-adviser

Venkataraman Shankar
Professor of Civil and Environmental Engineering
Thesis Reader

David A. Nembhard
Associate Professor and Graduate Program Coordinator of Industrial and Manufacturing Engineering
Chair of Graduate Program

*Signatures are on file in the Graduate School

ABSTRACT

Patients can face adverse events after colorectal surgery in a hospital. These adverse events include 30-day readmission, 30-day mortality, prolonged length of stay and complications after surgery. Predicting these events based on clinical and demographic data is critical for pre and post-operative surgical intervention. In this thesis, we investigate predictive analytics for adverse events for patients undergoing colorectal surgery. For the design, development and implementation of predictive analytics we use real life data provided by a prominent hospital system located in the Northeastern region of Pennsylvania. To protect the proprietary aspects of the data, all the variables in the dataset are de-identified. The longitudinal dataset consists of 8150 original records and 322 attributes from August 2006 to October 2014. In addition, the data related to the provider's behaviors with respect to colorectal surgery is also used.

The first step in this thesis addresses data cleaning and filling in the missing values. We use several statistical methods to perform these tasks. As the attribute set is large using subjective as well as statistical means we reduce the dimensionality to 120 attributes. We investigate four methodologies, including Naïve Bayes, Random Forest, Gradient Boosting Method and Logistic Regression for predictive analytics. We construct regression models with the four methodologies, the models are either tree-based or classifying models. After the models are constructed, we conduct comparative analysis of the methodologies based on certain performance criteria, and select the most accurate model for further study. We find that Gradient Boosting Method (GBM) has the best performance. We examine the most important predictors in the selected model, look for predictor features, and investigate intrinsic implications behind the predictors. Our conclusions point to the fact that patient's health condition and surgery information are the most important factors leading to adverse events. We suggest areas of future research.

Most of the past work in colorectal surgical adverse event prediction deals with retrospective data collected from several hospitals across several years. In addition, a fairly large number of specific inputs from surgeons are also used. This puts considerable burden on the hospital staff to collect data. Our model is developed using the data from only one hospital, with the general inputs from the physicians and surgeons that are normally entered. This simplifies the data collection and usage. The results we have obtained are comparable to the national statistics from the earlier models. However, we need to study several hospitals to validate our model and evaluate its efficacy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I want to give my sincere gratitude to Dr. Soundar Kumara for his continuous support and encouragement for my thesis, for his knowledge, patience and passion in scientific research. Dr. Kumara had given me guidance throughout the whole process of completing this piece of work. I could not have gone this far without his strong support.

I would like to give thanks to the co-advisor of my thesis Dr. Guodong Pang, for his insightful comments and suggestions, and also for the hard questions that incented me to widen my research from various perspectives.

I also want to express my gratitude to Dr. Shankar from Civil Engineering department for his support and guidance, and for sharing his truthful and illuminating views on a number of issues related to the thesis.

I want to thank my lab mates in for the stimulating and intellectual discussions; in particular, I am grateful to Cheng-Bang Chen and Deepak Agrawal for enlightening me in the research.

Last but not the least, I want to thank my parents dear Mr. Weiguo Guo and Mrs. Xiaoping Jian, for their support and guidance in my research and my life. I also want to express my gratitude to my boyfriend Chaoyi Wang, for his encouragements. Financial support for this research is gratefully acknowledged.

Chapter 1

Introduction

1.1 Motivation

In recent years, the delivery of healthcare systems has become a hot research issue. People are having more access to efficient healthcare systems, and millions are gaining coverage due to the Affordable Care Act. The U.S. healthcare system has been ever expanding, and healthcare related industries are rising in a fast pace. According to the "Report to the President on Better Health Care and Lower Costs: Accelerating Improvement Through Systems Engineering" in 2014, millions of Americans signed up for insurance coverage, and millions gained access to Medicaid. With the wide expansion, healthcare has gradually turned its emphasis to quality instead of quantity, that is, to not only ensure healthcare access to people, but also to provide the requisite quality. Healthcare expenditures are approaching one fifth of the total economy, whereas rising costs did not lead to better health, and patients are not receiving better quality of care (Smith et al., 2013).

In order to guarantee the effect of the Affordable Care Act, performance improvement measures should be taken. Healthcare should place emphasis on a series of integrated services provided for individuals, families and communities. Recent studies have shown that over one quarter of Medicare patients experienced certain types of harms in hospitals (Levinson, 2010), and even more experienced medical errors. Hospitals are now working hard to improve quality of care, and make rational use of resources. When patients have complications or prolonged length of stays after surgeries, they would suffer from days lost from work, discomfort, and the

inconvenience of commuting between hospital and their homes. While these inefficiencies and harms are preventable, measures should be taken to improve the healthcare system.

Systems Engineering is one effective way that has been implemented in many areas, which explores optimization in industrial systems. While systems engineering works well in many areas, there is also considerable systems engineering engagement in the surgical and anesthesiology world, it is applicable to solve prediction, scheduling and resource allocation issues. This will help hospitals as in current times hospitals are crowded and strained by financial pressure. In order to accomplish this, hospitals need to know adverse events that may affect patients ahead. The problem of post-surgical complications are prevalent in hospitals. Post-surgery infections, readmissions, mortalities, prolonged length of stays and complications are identifiable and rectifiable to improve post-surgical quality of life of the patient.

This thesis explores the rates of adverse events after colorectal surgery, and identifies the important factors that impact these rates. Although the problem is not hard to define, the solution may depend on many independent factors and their interactions, and patient readmissions and complications is a stochastic process with no explicit distribution which makes the predication process more complex. In this thesis with data provided by a local hospital, internal relationship between the data is analyzed and the best predictive analytics method is found. The results are presented intuitively via data visualization tools, and analysis on the predictor variables is conducted.

1.2 Research Problem & Research Objectives

The objective of this study is to identify patients who are at high risk of adverse events, including post-surgery readmissions, mortalities, prolonged length of stays and complications. We build regression models to produce the outcomes. Four regression models are built and

validated. Along with the models predictor importance charts are also constructed. We analyze the underlying factors for the outcomes.

We use regression models instead of classification models because we want to produce the outcomes in the forms of probability, and manually set a threshold to classify the probabilities. The four desired outcomes are all categorical, with readmission, mortality, complication as binomial dependent variables and length of stay as multinomial dependent variable. The four models selected all have good performance with high dimensional data, which are flexible and robust at handling non-linear relationships, and are in practical use in many areas.

After the models are constructed, we apply performance criteria to select the model with the best performance. We apply two criteria for binomial and multinomial regression models respectively. We conduct in-depth analysis of variable importance for the best performance model. We also examine the underlying features behind influential predictors.

1.3 Introduction to Methodology

We use four algorithms for constructing the regression models, Naïve Bayes Method, Random Forest, Gradient Boosting Model and Logistic Regression. We will run these algorithms in 'R' with the appropriate packages. The parameters are set through trials of tests and an optimized set of parameters are selected for each model.

The original data is provided by a hospital in the Northeastern region of Pennsylvania. It is a real life dataset collected on a daily basis over eight years. The raw dataset will be cleaned and organized in order to be eligible to be used in models. The dataset is split into training and testing sets. The models are learned through training set and are validated on the testing set. The dataset is split multiple times in order to reduce variance and bias, which is the process of bootstrapping samples.

After the models are validated respectively performance criteria are applied to select the best performance model, performance criteria include Receiver Operating Characteristic (ROC) curve and Cohen's Kappa statistic (see section 2.4). Variable importance charts are also produced along with the models.

1.4 Contributions and uniqueness of Research

This thesis is more applied from the sense that it uses existing theoretical models for predictions on real-life data from a hospital system. Previous models were mostly built from retrospective data from many hospitals with excessive amounts of attributes. Although the models achieve great performance, the attributes are hard to collect. This study aims at solving this issue by using data from only one hospital, and the attributes in the data are general physician inputs which are readily available in many hospitals. The model performance in this study is also satisfying, thus it is an excellent candidate for adaption in any hospital.

Another important feature of this study is that we predict four types of adverse events instead of one, and we used four methodologies instead of using only one method. Previous studies mostly predict only one adverse event, usually either readmission or mortality. In our study inspired by the American College of Surgeon Risk Calculator, we added complication and length of stay as target outcomes.

1.5 Organization of the Thesis

Chapter 1 is a brief introduction to the thesis, the definition of problem and the past research work. Chapter 2 focuses on literature review that includes a review of the surgery in this study, and prediction of adverse events. Chapter 3 is a detailed description of problem in this

study, including problem features, data source and data features. Chapter 4 discusses data pre-processing. Detailed descriptions of data cleaning and organization procedures are reported, and a process flow chart is used to illustrate the methodology. Chapter 5 represents the methodologies used in this study. The results are posted and briefly interpreted, performance charts and variable importance charts are presented. Chapter 6 deals with the analysis of results, including analysis for model performance, usual outcome and predictor importance. Chapter 7 concludes the study. Chapter 8 gives a few limitations and directions for promising directions of future work.

Chapter 2

Literature Review

In this chapter, we briefly review the background literature related to: Colorectal cancer, Prediction of adverse events, Predictive analytics and Model performance criteria.

2.1 Introduction to colorectal surgery

Colorectal surgery deals with disorders in the colon and rectum areas in the body. The surgery is for disorders such as colorectal cancer, hemorrhoids, fistulas, constipation conditions, Crohn's disease, anal injuries, etc. Colorectal cancer surgery is one of the most frequently performed surgery, as colorectal cancer is currently the third most common tumor type world worldwide (NCI, retrieved 07 April 2015). Types of treatments for colorectal disorders include hemorrhoidectomy, colectomy, colostomy, etc. Diagnostic procedures for colorectal disorders include colonoscopy, proctoscopy, sigmoidoscopy, etc. As the number of people having colorectal disorders is increasing every year, healthcare providers are putting emphasis on colorectal surgery to prevent adverse events, including post-operative complications, mortality, readmission and prolonged length of stay.

In recent years pre-operative care has considerably improved with more effective anesthetic techniques, and advanced medicine to reduce surgical stress. Bowel preparation to clean patient's bowel tract by chemical matter before surgery is an example of pre-operative procedures (Zmora, 2001). Post-operative programs have also been developed to promote better outcome. Programs such as oral feeding and mobilization are examples of such procedures. Nevertheless, a study by Kehlet (2008) has shown that elective colorectal surgery is still

associated with a complication rate of 20-30% and an average with a post-operative length of stay of 8-12 days.

Many countries across the world are putting efforts on reducing adverse events. Other than improving surgery techniques, they are also working on the prediction for potential adverse events. Predicting adverse events can help in better allocation, scheduling and utilization of hospital resources. This in turn will help in improving quality of care and reducing costs.

2.2 Prediction of Adverse Events

The quality of surgery is an important criterion to assess the quality of healthcare providers. Nowadays hospitals are focusing not only on improving the surgical procedures, but also on the quality of care after surgery. Even when a surgery is well performed, adverse events can still happen, and it is therefore important to have post-surgical care. When adverse events happen they not only cause inconvenience to both the hospital and the patient, it also contributes to lost time of clinicians, usage of critical resources which otherwise could have been used effectively. These in turn increase healthcare costs.

Every healthcare provider wants to reduce number of post-operative adverse events, for which prediction is the basic step. This is the major objective in this study as well. The initial target is to predict 30-day readmission rate, which indicates whether a patient is likely to be readmitted within 30 days after colorectal cancer surgery. The reason for using 30-day interval is that most readmission cases happen within 30 days after the patient's discharge. A report by Centers for Medicare & Medicaid Services (CMS) points out "during 2003 and 2004, almost one-fifth of Medicare beneficiaries-over 2.3 million patients-were re-hospitalized within 30 days of discharge". The readmission costs for Medicare is more than 17 billion annually (Horwitz et al, 2012). Thus, 30-day readmission is a strong indicator of the lack of communication, surgery

effectiveness and insufficient follow-up care. CMS chose to measure readmissions within 30 days, because readmissions during longer periods (1 year) may be affected by other reasons outside hospital's range of responsibility. Long-term readmissions may be caused by accidents, patient's habits, or medications (http://www.medicare.gov/hospitalcompare/Data/30-day-measures.html, accessed on June, 2015). US government is currently taking measures to penalize hospitals with high readmission rates. Moreover, healthcare payers are refusing to pay for preventable post-operative adverse events, and most of the times it is at their discretion for deeming "preventable".

In the current challenging times with financial pressure and shortage of resources, many organizations are researching on applying operations research techniques in healthcare. The National Surgical Quality Improvement Program initiated by American College of Surgeons is an example, which centers on patient safety, reduction of morbidity and mortality, and they have developed a risk calculator for patients (https://www.facs.org/~/media/files/quality%20programs/nsqip/nsqipinfobook1012.ashx, accessed on June 15, 2015). The risk calculator is a web-based tool, which uses around 20 indicators related to patient and hospital provider's behaviors to predict the risk of postoperative adverse events. Common postoperative complications include surgical site infection (SSI), kidney failure, urinary tract problem, etc. The risk calculator is an innovative way of using an user-friendly interface to give an estimate of risks. Surgeons and patients can access the webpage from their homes and offices. The inputs include patient's age, sex, smoking history, health condition, chronic disease history, etc. With these information hospitals can have an estimate of patient's postoperative conditions (http://riskcalculator.facs.org/PatientInfo/PatientInfo, accessed on July 1, 2015).

In this particular study, instead of estimating adverse events for a wide range of surgical operations, we focus on colorectal surgery only. Similar to all other kinds of surgeries,

identifying risk factors and making better healthcare plan for colorectal surgery is crucial to healthcare providers, although measures have been taken to reduce adverse events, colorectal surgery still has a post-operative mortality rate of around 5% (Arnaud et al., 2005). Many existing studies are based on retrospective studies of post-operative patient records. Some studies focus on a single variable that affects post-operative events, and they analyze the effects of the single variable through controlled trials, such as smoking cessation programs and exercise programs (Lars Tue and Torben Jørgensen, 2003). In this thesis, we analyze pre-operative indicators and see how they impact adverse events. The following section is an introduction to algorithms used in prediction.

2.3 Predictive analytics

Predictive analytics include a wide set of statistical techniques that vary from machine learning, data mining to data modeling that make forecasts about future events using current and historical information (Nyce et al, 2007). When predictive analytics are used in healthcare, they can be applied in patient census forecast, hospital readmission forecast, operation rooms (OR) scheduling, nurse staffing, nursescheduling etc. With the help of statistical techniques, hospitals are getting a more accurate sense in arranging resources, instead of based on previous experience and intuition. The essential idea of predictive analytics lies in finding relationships between explanatory variables and the response variables from the past, and using this relationship to predict the unknown outcome. However, the accuracy of predicted results depend greatly on the quality of data analysis, assumptions, and the dataset itself (Siegel, 2013).

2.3.1 Logistic Regression

Logistic regression is a statistical method developed by D. R. Cox in 1958 (Cox, 1958). It is a probability model used to predict binomial or multinomial outcomes, based on known explanatory variables. The core of logistic regression is finding out the parameters in the model, and the outcomes are modeled as a function of independent variables. In binomial models, the results are often represented as '0' and '1', which stand for two contrary facts. The independent variables can be either continuous or categorical.

The essential logic in logistic regression lies in logistic function. The logistic function can turn inputs of any real number into categorical outcomes. Using σ_t to represent the logistic function, we have

$$\sigma_t = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

We will have outcome as a linear function of explanatory. Suppose we have $P(x)$ as the outcome probability that ranges from 0 to 1, and we have β_0 and β_1 as intercept and regression coefficient in the linear relationship. Then the logistic function is expressed as

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

There have been many in-depth research works on logistic regression. The method has applications in a broad range of industries. When applied in healthcare, it may be used to predict the risk of a patient getting a certain kind of disease, such as diabetes, high blood pressure, cancer, etc. Based on historical medical records of the patient and their symptoms (gender, height, weight, body mass index, blood cholesterol level, blood hemoglobin level, white blood cells level, lifestyle and habits), logistic regression model can be constructed.

2.3.2 Naive Bayes

Naive Bayes (NB) is an algorithm for classification problems. It is a supervised learning technique based on Bayesian interpretation of probability. NB complies with the independence assumption that each feature is independent of any other else (McCallum and Nigam, 1998), by these assumptions, the model employed is

$$P(x_1, x_2, \dots, x_n|y) = \prod_{i=1}^{n} P(x_i|y)$$

By Bayes' theorem, we get the following formula

$$P(y|x_1, x_2, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

Finally, the decision rule is obtained by applying the maximum a posteriori probability (MAP) estimation to the formula

$$\hat{y} = \operatorname{argmax} P(y) \prod_{i=1}^{n} P(x_i|y)$$

We can easily make a prediction for a new instance using this algorithm. Although the NB assumption is very strong, the algorithm has worked well in many problems. However, the NB estimator requires a relatively large training set to make reliable predictions, which results in a big limitation on its efficacy in real-world problems, even though the current world is a "Big Data" world.

2.3.3 Random Forest

Random forest is a learning method for both classification and regression problems, the concept is to construct a "forest" of decision trees, and output the mean prediction of the trees. The method combines the idea of bagging and random selection of variables and tree subsets.

The idea of bagging is selecting random samples from the training set repeatedly and fitting decision tree to the samples. The number of samples depends on particular situations. Given a set of n independent observations $k_1, k_2, k_3, \dots, k_n$, each with variance σ^2, the variance of the mean of observations is σ^2/n, which implies bagging can reduce variance. Suppose we generate B subsets from the training set, we get

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x)$$

This process is bagging. Random forest is built upon the concept of bagging. Instead of building tree subsets with all of the predictor variables, random forest draws only part of predictor variables each time, a fresh sample of m predictors is used in each tree split. Suppose we have p predictor variables. Usually we draw $m = \sqrt{p}$ predictor variables each time. Random Forest is the process of de-correlating the trees and thereby smoothing the variance of bagging. Random Forest method often works better than traditional regression methods (James et al. 2013).

2.3.4 Gradient Boosting Method

Gradient Boosting Method (GBM) is a concept that also includes decision trees and bagging. Instead of combining the "forest" and taking the average of the outcomes, boosting works in a way to let the trees grow sequentially. In Random Forest the trees are independent, but in boosting each tree is grown based on previous grown trees, and fit in a modified version of the training dataset (James et al. 2013). We set a shrinkage parameter λ, it allow more shaped trees to attack the residuals. Fitting a tree $\hat{f}^b$ to the training data, and keep updating the tree with shrinkage parameter, we have

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

At the same time residuals should also be updated. In the end the output model is

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

Random Forest and GBM have been used in high dimensional data consisting of a large sample of records (i.e., patients) and variables. Both techniques are flexible and do not require distributional assumptions compared to traditional parametric models. They are capable of modeling complex, non-linear relationships of continuous predictor variables and multi-way interactions.

2.4 Performance Criteria

When using different models to produce desired outcomes, there should be a standard criterion to compare across model performances. The Receiver Operating Characteristic (ROC)

curve is an effective evaluation measure for categorical outcomes. By comparing the Area Under Curve (AUC) statistic, we can have an idea model performance on the testing set.

2.4.1 The ROC curve

The ROC curve demonstrates the tradeoff between sensitivity and specificity, which is also related to Type I error and Type II error. The ROC curve has a diagonal line, and the curve is above the line. The closer the curve is to the diagonal line, the less accurate the model predicts. If the curve is in the shape of a fully stretched bow, then the model has a good performance. The area under the curve (AUC) is the measure of accuracy, in R we use the package *'proc'* to produce ROC curve.

An AUC of 1 represents a perfect test, and an AUC of .5 represents a worthless test that means the prediction is even worse than randomly predicting by chance. The traditional scoring system for AUC is as follows (http://gim.unmc.edu/dxtests/roc3.htm, accessed on 10 July, 2015),

- .90-1 = excellent
- .80-.90 = good
- .70-.80 = fair
- .60-.70 = poor
- .50-.60 = fail

2.4.2 The Kappa Statistic

Although the AUC is a convincing statistic for measuring performance on binomial outcomes, for multinomial outcomes we use the Kappa statistic. Cohen's kappa coefficient is a

statistic that measures multi-rater agreement for qualitative (categorical) items. The equation for the statistic is

$$k = \frac{pr(a)\text{-}pr(e)}{1 - pr(e)}$$

$pr(a)$ represents the relative observed agreement among raters, whereas $pr(e)$ represents the hypothetical probability of chance agreement. If the predicted and observed have a total match, then kappa statistic is 1. If the predicted is even worse than random guessing by chance, then the kappa statistic is 0. There is no standard interpretation for Kappa statistic, according to a study by Landis and Koch (1977), the scoring system is as follows,

- <0 = no agreement
- 0-0.2 = slight
- 0.21-0.40 = fair
- 0.41-0.60 = moderate
- 0.61-0.80 = substantial
- 0.80-1.0 = almost perfect

The Kappa statistic command in R is under the package *'irr'*, and the command is *'kappa2'*.

Chapter 3

Problem Description and Discussion

In this chapter, we describe the problem in detail, and give a brief summary of the dataset. We focus on the problem objective, methodology and performance criteria in the thesis. The data description section includes a summary of data size, data source and data features, several variables are briefly discussed.

3.1 Problem Statement

We have discussed the significance of colorectal surgery in chapter 2. We identify the following as the adverse events after such a surgery:

1. 30-Day Readmission Rate
2. 30-Day Mortality Rate
3. Length of Stay
4. Complication Observed vs. Expected Rate

The first two target outcomes are expressed as rates between 0 and 1. As "Length of Stay" is the number of days after surgery, we grouped the values into four categories (see section 4.2.2). Complication indicates a secondary disease happening after surgery, such as surgical site infection, bleeding, pain, nausea, etc. As a complication recorded in the dataset is a ratio of observed and expected, this outcome has been transformed into binary and expressed as rates between 0 and 1 (see section 4.2.2).

These outcomes are measured from patient's first discharge from the hospital, not from the second discharge from a transferred facility.

Our objectives in this research are:

1. To design, develop and implement a predictive analytics model to predict the patients at high risk of adverse events;
2. To identify the statistically significant variables that help in the prediction.

The steps in the design, development, implementation and analysis of the predictive analytics model are:

1. To organize the raw dataset in order to make it useful for modeling. The raw dataset may include predictor variables with excess amount of missing values or predictors with too many categories, in such cases statistical analysis techniques are implemented. The dataset is separated into training and testing sets. The testing dataset is set aside until the models are constructed.
2. After the data is cleaned and well prepared, an initial model is developed using traditional demographic variables (e.g., gender, age) as well as a unique collection of measurements available from the electronic health record (EHR) of a hospital (e.g., existing comorbidities, medication use, laboratory values, and initial treatment strategies). Four different algorithms are used for data modeling; namely, Naïve Bayes Method, Random Forest, GBM and Logistic Regression. Comparison of the four algorithms is conducted, each of the algorithms produces a variable importance chart, and overall result is analyzed and conclusions are made on the most significant predictors.
3. Estimate the AUC statistic for four algorithms, while using the fewest number of parameters. The AUC is a quantitative measure of the discrimination ability of the model to correctly classify patients. The results of the model are used to

construct a risk evaluation of adverse events. The properties of the risk evaluation are described using AUC. Kappa statistic are calculated for multinomial regression, which is be used to compare the performance across different methodologies.

3.2 Data Description

The dataset in this study is a collection of unique measurements from a hospital located in the Northeastern part of Pennsylvania. The hospital belongs to a physician-led health system, and the data is collected on a daily basis ranging from August 18, 2006 to October 07, 2014. There are a total of 8150 observations and 322 predictor variables in the dataset. The data is extracted from hospital's EHR system, and the patients' information is de-identified. A data dictionary is also included for interpretation of variables.

As an initial step we classified the predictor variables into 3 groups: very important, moderate, and not important. The ranking is arrived at subjectively using literature and personal understanding. This is an initial grouping of importance. After regression models are constructed variable importance charts are produced along with the models (see chapter 5). By comparing the importance chart and subjectively grouped importance table we can decide which insignificant predictors can be removed.

The predictor variables have also been grouped according to their features in order to have a better understanding of the dataset. The grouping is based on the information provided by the hospital and on previous literature reviews. The data features are as follows:

Table 3-1. Predictor variables features.

Feature	Variables
Socio-demographic Information	PAT_BMI, PAT_AGE, GENDER, COUNTY
Admission Information	ADMISSION_QUARTER, EMERGENCY_CASE
Health Condition Information	WOUND_CLASS, ASA_RATING
Surgery Information	COLORECTAL_SURGEON, PANEL1_LENGTH
Discharge Information	DISCH_DISPOSITION, DISCHARGE_DATE
Healthcare Utilization	LENGTH_OF_STAY, NUM_LABS
Social Support	BILLING_PAYOR
Diagnosis Information	ENCOUNTER_REASON, PREVIOUS_ENCOUNTERS

The first four columns of dataset contain patient identity information, including surgery ID, patient ID, patient medical record number (MRN) ID, and patient encounter ID (see table 3-2). Among the four ID columns, patient encounter ID was recorded based on each encounter, each patient might have multiple surgeries during one encounter, and each LOG ID represents a unique current surgery. Consequently we use LOG ID as the primary key for dataset.

Table 3-2. First four columns in dataset.

LOG_ID_DEID	PAT_ID_DEID	PAT_MRN_ID_DEID	PAT_ENC_CSN_ID_DEID
107586796	76927035	64056017	65170273
55707145	21045926	10820167	74951802
45212489	9528745	92948041	27773297
41502356	104733386	75690692	120968069
18910977	122159089	75015432	20058199
111154664	101639193	4230550	84690043
113332865	115316089	118224418	104182019
40256505	13563279	454363	106271557
34349071	74620583	79389112	113312450

In the next chapter we describe in details of the data preparation process, which is the first step in the design, development, implementation and analysis of the predictive analytics models. We explain the data cleaning process, data categorization process and missing data handling process.

Chapter 4

Data Preparation

In this chapter, we discuss the data preparation process. The data is provided by a hospital in the Northeastern part of Pennsylvania, and it has been retrieved from hospital EHR. The dataset contains records of eight years from 2006 to 2014. We will discuss how the data is cleaned, and how we handled missing data. After data cleaning we conducted data organizing, including previous information transformation and variable categorization. A flow chart is displayed which visualizes the whole data preparation process (see figure 4-1).

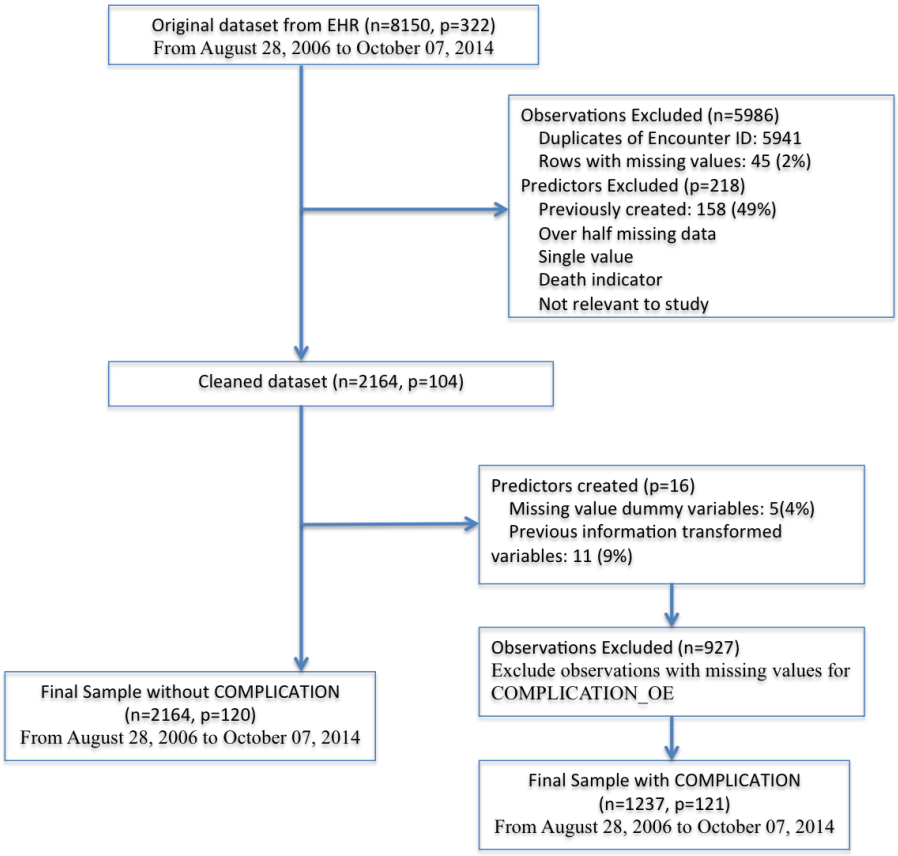

Figure 4-1. Data preparation process.

4.1 Data Cleaning

In the dataset there are many attributes that are not related to the outcomes, or have too many missing values. In order to make the data useful

for constructing prediction models, cleaning and data organizing are done as initial steps, and data pre-processing is necessary in all kinds of statistical analysis.

4.1.1 Organizing Useful Data

In the dataset there are 8150 observations and 322 predictor variables in total (see figure 4-1). Some of the variables are known to have certain kinds of relationship with the outcomes. Some are yet to be explored. From the dataset we have extracted useful information, and removed some information that are considered not relevant. The process is as follows (see figure 4-1),

- Among the 322 predictors, 158 have no definitions in the data dictionary. They are self-created variables from previous modeling use. These are excluded from this study.
- 17 variables have over 50% of missing values. In such cases missing values cannot be imputed because it will cause strong bias in regression. For example consider the variables “Second Complaint”, “Third Complaint” ,…, “Tenth Complaint”; as these variables contain over 90% of missing data, they are removed.
- 10 variables have only one single value for every observation. They do not contribute to regression and therefore are excluded.
- 8 variables are considered not having much relevance to this study, such as “Patient’s Previous Admission Time” which records the admission time for all of a patient’s

previous surgeries, and “Previous ED triage time” which records the time patient previously received triage. They do not contribute to regression and thus removed.

- 6 variables are endogenous variable as they already contain information on the outcomes when recorded, for example the variable “Death Indicator” indicates whether the patient is currently dead, and the variable “surgeon’s 90 day readmission rate” that includes the readmission information on the current surgery, such variables would cause strong bias in regression and are therefore removed.
- 6 variables are duplicates of other variables, such as “Operation Procedure ID” which is a duplicate of the variable “Operation Procedure Name”. These variables are removed.
- 10 variables have less than 30% of missing data, and do not have much variance in the observations (standard deviation less than ½ of mean). The missing values will be imputed with mean or mode.
- 2 variables are newly created by subtraction of 3 existing variables, e.g. “Admission to Discharge” is calculated from “Admission Date” and “Discharge Date”.
- “Patient ID” and “Patient MRN ID” (see table 3-2) are removed. “Patient Encounter ID” is kept for reference. “LOG ID” is used as the primary key.
- While “LOG ID” is the primary key, we will only keep the unique observations. There are 2629 out of 8150 unique records in total. We have removed a part of observations that are complete duplicates, and the other part has been transformed into useful current information (see section 4.2.1).
- 3 variables have only a small amount of missing data (< 2%). The records that contain the missing data corresponding to these 3 variables are removed. 130 records are removed.

- For the variable “Discharge Disposition”, some records contain information on the outcomes, especially on mortality (see table 4-1). 84 records are removed corresponding to the endogenous values.
- After the above cleaning phase, we are left with 2415 unique records, and 117 predictor variables (see figure 4-1).

Table 4-1. Example of Discharge Disposition.

LOG_ID_DEID	BMI	AGE	DISCHARGE_DISPOSITION
111064387	27.1	64	In Hospital, Other Death, No Autopsy
119098883	26.6	33	Hospice, Home Routine Care
115263460	21.6	75	In Hosp- Post-Op Dth, No Autop
4577408	14.5	88	In Hospital, Other Death, Autopsy Unknown
14271889	34.2	32	Coroner- Postsurg >48H, AUT UN
12069567	28.5	60	In Hospital, Other Death, Autopsy

4.1.2 Handling Missing Data

When organizing and categorizing data we often encounter difficulty with missing values. There are several types of missing data such as data missing completely at random (MCAR), missing not at random (NMAR), etc. In this study the missing data are mostly missing completely at random, which means missing data depend neither on explanatory variables nor response variables. There are several ways to deal with missing values; the most common methods are as follows:

- List wise deletion—delete the entire row or column that contains missing values.
- Mean/Mode substitution—Impute missing data with variable mean or mode.
- Dummy variable adjustment—Create a new binary variable besides the original variable, the new binary variable indicates whether the original variable contains missing values.

For a variable x_i, if a certain value $x_{i,j}$ under this attribute is missing, we transform the variable into a binary variable, and replace with an indicator variable $x_i^{'}$, defined as:

$$x_{i,j}^{'} = \begin{cases} 1, & if\ x_{i,j}\ is\ missing \\ 0, & Otherwise \end{cases} \quad \forall i,j$$

Table 4-2. Example of dummy variable adjustment.

Original Variable	New Variable
1	1
2	1
Null	0
4	1
Null	0
6	1
Null	0
7	1

Table 4-2 is an example of creating a binary variable, which indicates whether the original variable contains missing values.

- Regression substitution—Impute missing value with predicted value from variable regression, but this may weaken variance and overestimate model fit (Humphries, 2013).

In this dataset, exclusion has already been done in the previous section (see figure 4-1), we used mean/mode imputation for variables that have only a small amount of missing data (< 5%), and we have also created missing data binary variables to indicate whether the original variables have been imputed with mean/mode. Through this process we arrived at 109 variables in total (see figure 4-1).

4.2 Data Management

After data cleaning, we have a clear set of data that can be readily used; however, there are still problems with categorical variables that have too many categories (e.g., Patient Complaints, Surgery Start Hour). As we know in regression, categorical values are transformed into dummy variables. If we have 240 different patient complaints, 239 dummy variables will be created, which is twice more than original predictor variables. Thus for fast computation and convenience, we need to manage the data and sort them into groups. We use 3 types of data management techniques in this study which are explained in the following section. .

4.2.1 Previous information transformation

The original dataset contains 8150 observations in which we use patient encounter ID as the primary key. After analyzing the data we have found that there are only 2209 unique values, the reason for this is that we have the data in two parts, current visits and previous visits, current visits information are all duplicates while previous visits are unique, therefore we want to delete the duplicates for current information while transforming previous information into current information. We illustrate this concept in the following table (Table 4-3).

Table 4-3. Example of the original dataset.

ENC_ID	BMI	AGE	PREV_ADMSN	SVC	ENC_REASON_NAME_1	NUM_LABS
1121525	23.92	41	11/11/09 9:26	General Surgery (GMCGLS)	ABDOMINAL PAIN	24
1121525	23.92	41	3/18/09 17:09	Emergency Med (GMCE/R)	CAT SCAN	
1121525	23.92	41	8/4/06 5:26	General Surgery (GMCGLS)	ABDOMINAL PAIN	9
1121525	23.92	41	7/20/08 12:43	Emergency Med (GMCE/R)	ABDOMINAL PAIN	4
1121525	23.92	41	8/3/06 5:10	Emergency Med (GMCE/R)	PAIN	5
1121525	23.92	41	11/11/09 9:26	General Surgery (GMCGLS)	ABDOMINAL PAIN	24
1121525	23.92	41	3/18/09 17:09	Emergency Med (GMCE/R)	CAT SCAN	
1121525	23.92	41	8/3/06 5:10	Emergency Med (GMCE/R)	PAIN	5
1121525	23.92	41	7/20/08 12:43	Emergency Med (GMCE/R)	ABDOMINAL PAIN	4
1121525	23.92	41	8/4/06 5:26	General Surgery (GMCGLS)	ABDOMINAL PAIN	9

It can be seen that the first three columns are duplicate records for the encounter ID 112525 (see table 4-1), they represent the current visit and duplicates are removed. The following 4 columns indicate the patient's previous visits, 2 types of transformations are used in this case.

1. Sum up previous information counts, the variable PREVIOUS_ADMISSION_TIME (see table 4-1) can be treated as counts, the sum of the counts is stored in a new variable PREVIOUS_VISITS (see table 4-2), in this case the number is 10. In the last column - number of lab tests, it is clear that each previous visit for a single patient has multiple lab tests, we sum all the tests up and create a new variable PREVIOUS_NUMBER_OF_LAB_TESTS (see table 4-2), in this case the value is 84.
2. Group categorical values and turn them in to continuous variables of counts, we filter the values in which 30-day readmission is positive, and find the matching frequency for the categorical variable, take the top 2 or 3 categories and group the rest as "others", then make these categories as dummy variables. For example in the variable "SVC" which means patient's service line, a plot of frequency chart for positive readmission rates is produced

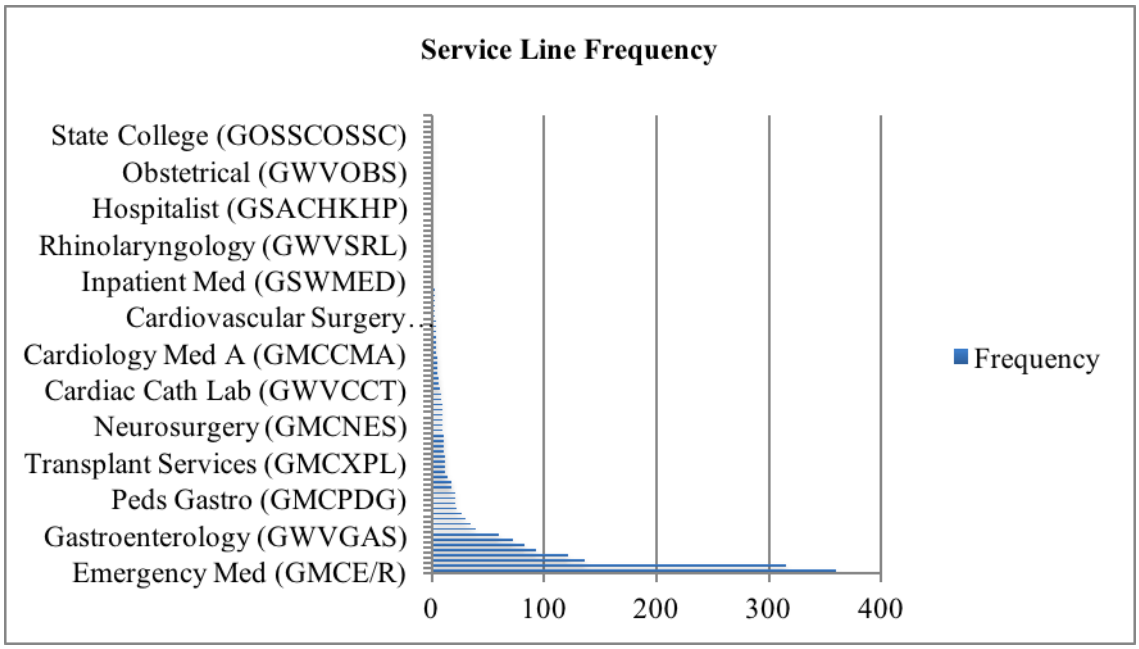

Figure 4-2. Service line frequency distribution.

From the chart we pick the variable(s) with the highest frequency associated with positive 30-day readmissions, in this way the variable can be more relevant to the outcomes in regression. In the case illustrated above we pick top 2 frequency values, 'Emergency Med' and 'Gastroenterology', we make them into continuous variables by summing up the values in the dataset, and the data table for ID 1121525 is turned in to a single row (see table 4-2)

Table 4-4 Transformed data record

ENC_ID	BMI	AGE	PREV_VISITS	PREV_SVC_GMCE/R	PREV_SVC_GMCGAS	NUM_LABS
1121525	23.92	41	10	6	0	84

After transformation 11 more variables have been created, and we have a complete dataset with 2209 observations and 120 predictor variables (see figure 4-1).

4.2.2 Grouping categorical values

The above transformation turned previous data in to current summary, and reduced categories for variables with several categories, but the transformation is based on statistical analysis. Another type of transformation is based on the distribution and feature of a variable. Without statistical analysis, we produce a frequency chart of the values and see if there are any patterns in the distribution. If there is an obvious pattern then we group the values according to pattern, for example, the frequency chart for the variable SURGERY_START_HOUR is as follows,

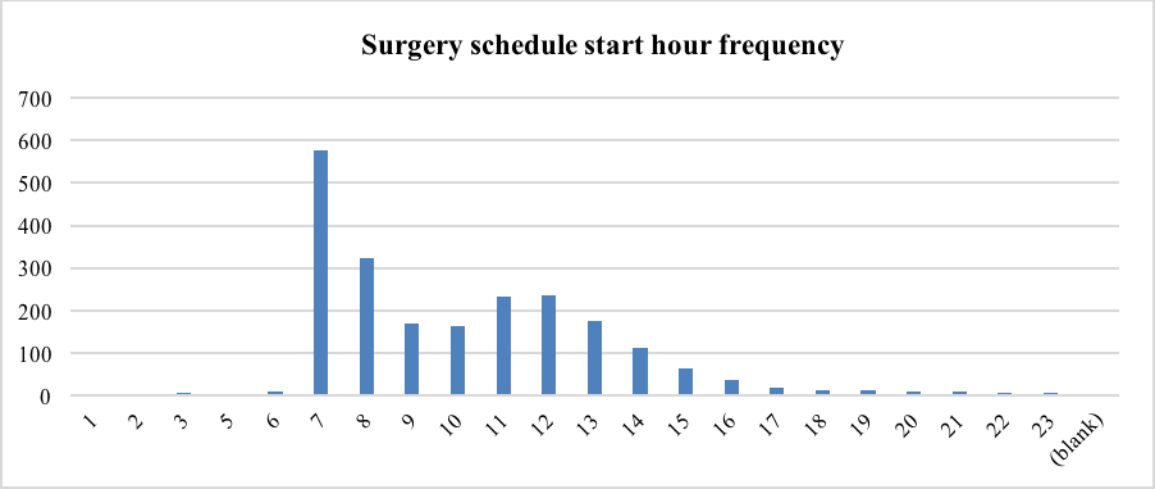

Figure 4-3. Surgery schedule start hour frequency distribution.

There is a clear trend in the chart, 7:00 am is the most frequent categorical value (see figure 4-3), therefore this variable will be grouped in the following way

- 7:00-12:00 as morning
- 13:00-17:00 as afternoon
- 18:00-6:00 as evening

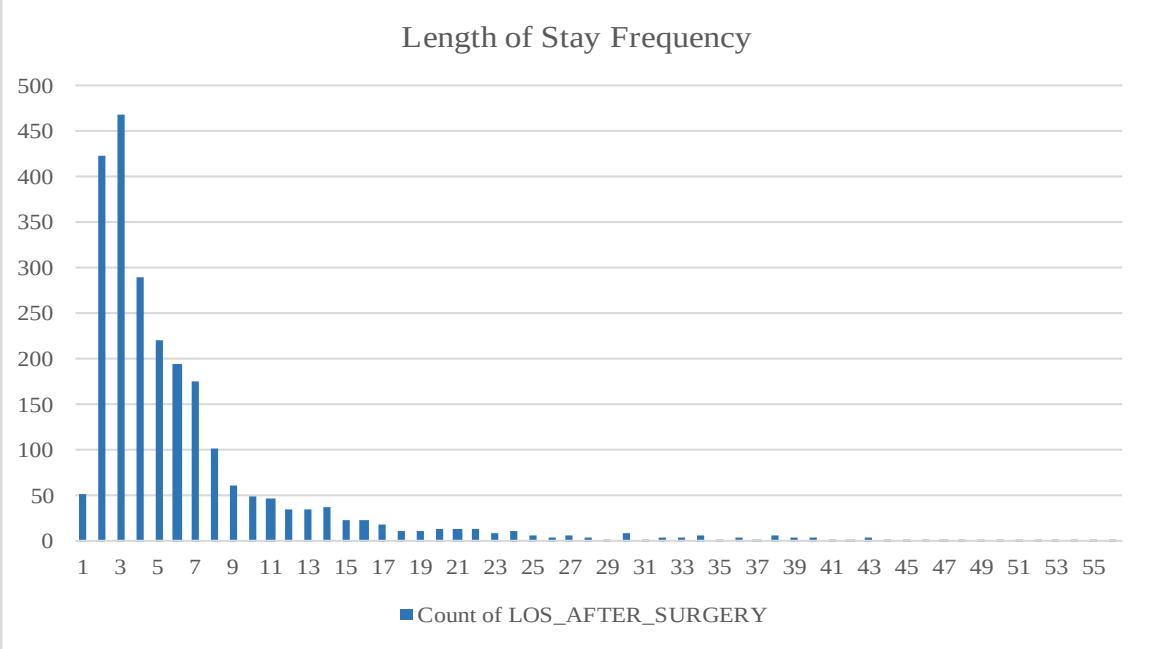

Figure 4-4. Length of stay frequency.

One of the outcomes we are aiming to predict is "Length of Stay" after surgery, since this variable is continuous and the methodologies we use in this study are focused on classifiers, they are being discretized.

From the chart we find a pattern (see figure 4-4), the distribution is typically log-normal, it can be seen that length of stay of 1-11 days has high frequency and the days after that have relatively low frequency. The variable "Length of Stay" is discretized the following way

- 1-3 days as short
- 4-17 days as medium
- 18-87 days as long

The other outcome we are aiming to predict is complications after surgery, we produced a pie chart of its frequency distribution (see figure 4-5), there are 1001 observations that have no complications, and 236 have different complication rates, therefore we group this variable as binary, either with or without complications.

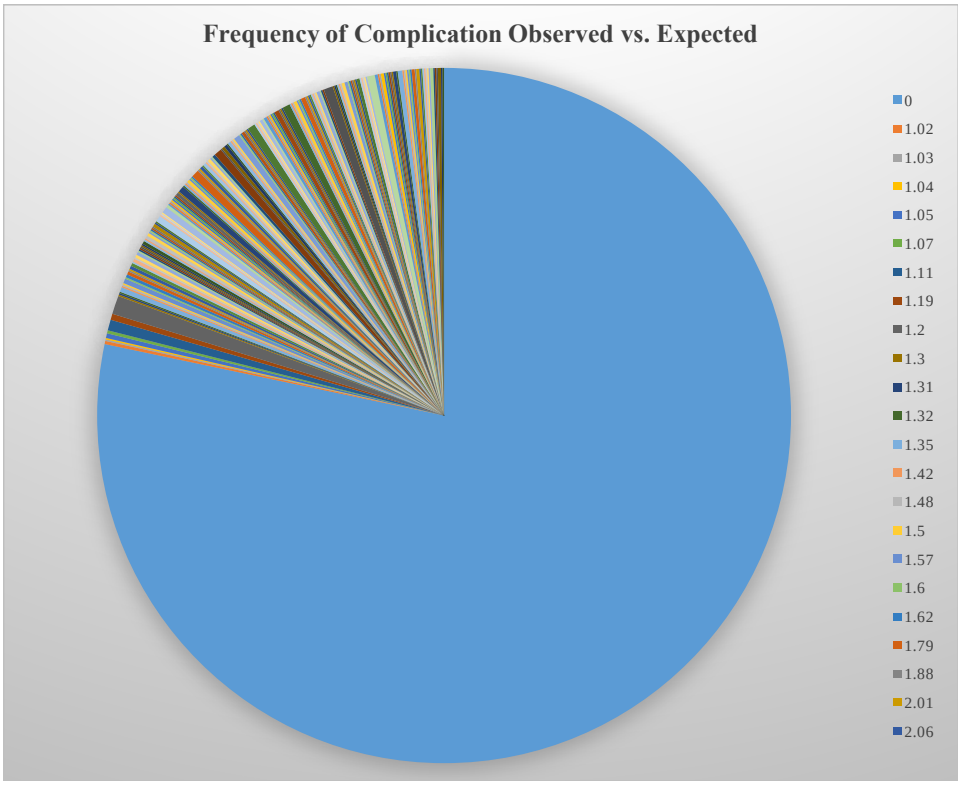

Figure 4-5. Complication frequency pie chart.

Other variables such as DISCHARGE DATE, ADMISSION DATE, SURGERY DATE all exhibit quarterly trend so we grouped them as 4 quarters and created a variable named "Number of Days between Scheduled to Surgery Date".

After data management we have a clean dataset that is ready for modeling, for the rows that still have several missing values, the rows have been removed. In the final stage we have 2164 observations and 120 predictor variables (see figure 4-1). Since we need to predict complication rate, there are over half missing data, we delete the missing rows, and create a separate set for complication prediction, we have 1237 observations and 121 predictors (COMLICATION_OE included).

Chapter 5

Methodology

In this chapter we construct regression models with four methodologies; Naïve Bayes Method, Random Forest, Gradient Boosting Method (GBM) and Logistic Regression. ROC curves are produced in each methodology, and variable importance charts are displayed. We discuss ROC curves and our analyses.

5.1 Validation dataset

In this section, we discuss the different methodologies used to build regression models. Fundamentally there are two types of analytic methods: regression and classification. In regression model the output is continuous. In classification model, the output is categorical, if we want a binary result, the output will be separated into 2 categories. If we use software packages to run methodologies the software will automatically set a threshold and output binary result, and this might not be the outcome we desire as we want to manually decide the threshold. Therefore, in our study we will use regression methods that produces continuous outcomes between 0 and 1.

Following the data preparation chapter we will construct regression models, but the first step is to split data into testing set and training set. The 2 sets are randomly selected subsets from the original dataset by randomly splitting it. The training set is a subset used for learning. We use the training set to run regression and fit a model and use the testing set to validate the model. The reason why we need to split dataset is that the error rate of training set is usually biased (smaller than the true error rate), and the role of testing set is to test how well the trained model

performs. Training and testing sets are separated usually in a ratio of 8:2 or 7:3, in our study we use 75% of the original dataset as training subset and 25% as testing subset.

We will run regression analysis in software 'R', which randomly splits the dataset. One thing to note is that although the dataset is randomly split, the training result can also be biased, we can coincidently select a training set that is good fit and output very poor testing results. In order to reduce variance and bias, we will split the dataset and train the model multiple times, each dataset is split 5 times, we will take the average of testing set outcome and analyze the results. This is the concept of bootstrap sampling.

Bootstrapping is a method for random sampling with replacement, it allows assigning accuracy such as variance, bias, confidence intervals, error to sample estimates (Efron, 1993). The reason we use bootstrap sampling is its simplicity, it derives estimates of standard errors and confidence intervals for complex parameters, it is also a good way to control and check the stability of the outcomes. In software 'R' we use the command '*set.seed ()*' to separate dataset multiple times and implement bootstrap sampling method.

5.2 Performance Importance

After assessing the performance of each model, predictor performance should also be analyzed, which predictor has large impacton the outcome is an interesting topic not only to researchers, but also to healthcare providers, it can be a guide to healthcare providers on how to adjust their resources, and what aspects need improvement.

All of the predictors are plugged into the models, a set of most important predictors are produced, the greater the importance value is, the larger impact the predictor has on the outcome. In logistic regression, variable importance is calculated based on odds ratio of predictors, the most important variable has the largest odds ratio. In 'R' we use the *'varImp'* command under

the package *'caret'* to output predictor importance in logistic regression. We use *'importance'* command under the package *'randomForest'* to output predictor importance in random forest. In GBM the importance is automatically calculated as *'relative influence'*.

5.2.1 Naïve Bayes Method

Naive Bayes is one of the simplest classification algorithms, it tries to classify outcomes based on the probabilities of previously seen attributes, assuming that attributes are independence.

We use Naïve Bayes Method to predict 4 desired outputs: 30-day readmission rate, mortality rate, complication rate and length of stay. The first 3 outputs are expressed as probabilities between 0 and 1 (can be slightly above 1 and below 0), to test the performance of the model for classifying outcomes we will use the ROC curve. For "Length of Stay", Kappa statistic is used to measure model performance for this variable. After we run the model in R with the package *'e1071'*, we used the testing set to predict the outcomes.

1. 30-Day Readmission Rate

The outcome is produced in two columns that indicate the probability of being 0 and 1 respectively, in each row the 2 values add up to 1. We have 541 rows in the testing set and ROC curve is used to measure model performance.

The AUC is 0.68 in this case, which is acceptable but not too perfect, it might not be sufficient for real life use. According to the criteria in section 2.4.1, the model performance is poor.

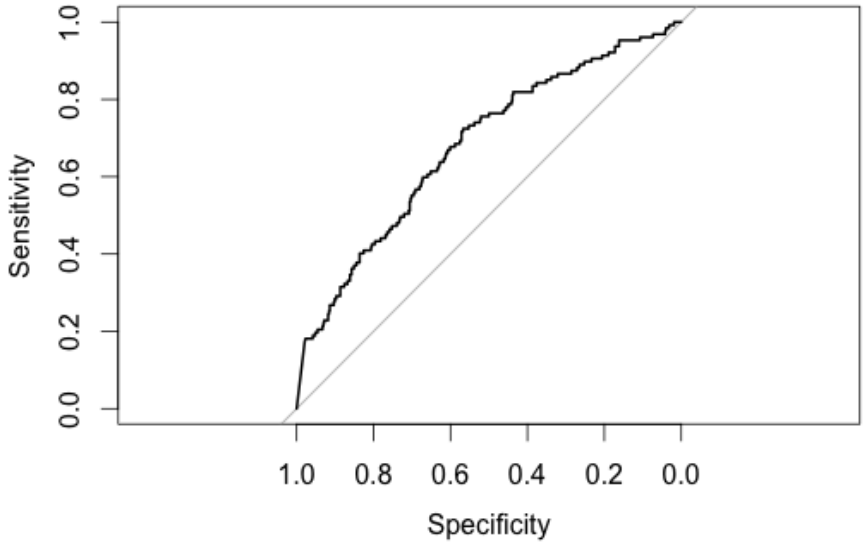

Figure 5-1. ROC curve for "30-Day Readmission Rate" in Naïve Bayes Method.

2. 30-Day Mortality Rate

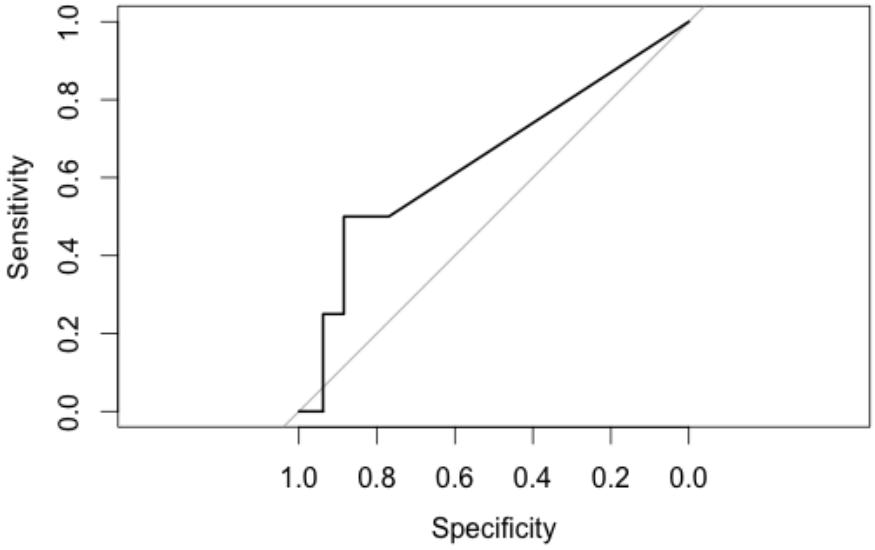

Figure 5-2. ROC curve for "30-Day Mortality Rate" in Naïve Bayes Method.

The AUC is 0.65, which is poor.

3. Length of Stay

We have discretized the variable "Length of Stay" and the output is categorical with three levels, we use Fleiss's Kappa to measure model performance, in 'R' the command is *'kappam.fleiss'*. The kappa statistic is -0.26, which means it even worse than random guessing, this model is unsatisfactory.

4. Complication Observed vs. Expected Rate

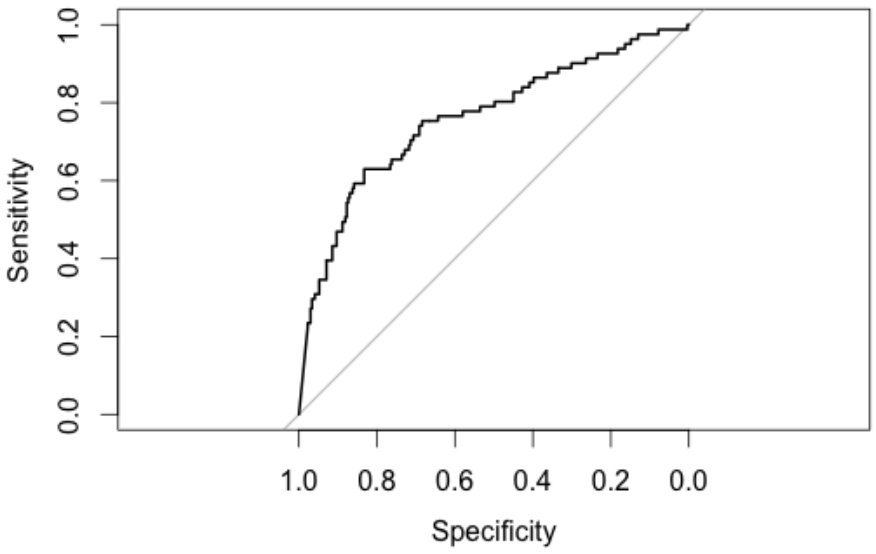

Figure 5-3. ROC curve for "Complication Rate" in Naïve Bayes Method.

The AUC is 0.76, since it is between 0.7 and 0.8, the model performance is fair.

5.2.2 Random Forest

Random forest is a technique using bootstrapping and growing trees, the model produces a collection of decision trees that form a black box, the outcome will be produced from the black box. Random Forest usually produces good results.

By trial and error in changing parameter values we obtain the best set of parameters, we set the number of predictors in each tree at 50, and the total number of trees grown to 500.

1. 30-day readmission rate

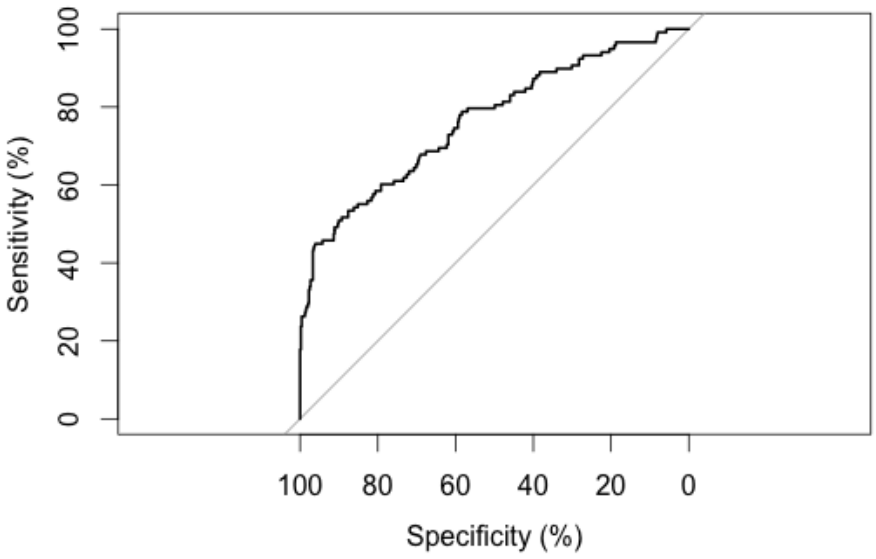

Figure 5-4. ROC curve for "30-Day Readmission Rate" in Random Forest.

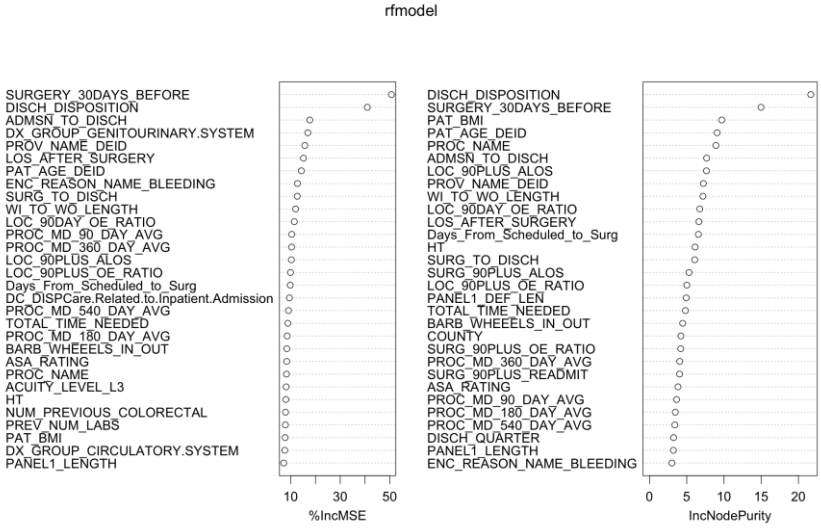

Figure 5-5. Variable Importance for "30-Day Readmission Rate" in Random Forest.

The AUC is 0.77, which is fair, variable importance will be examined further.

In can be seen from figure 5-5 that whether the patient had a colorectal surgery before, discharge disposition are very strong predictors for readmission rate. Patient's socio-demographic information such as BMI, height and age are important as well.

2. 30-Day Mortality Rate

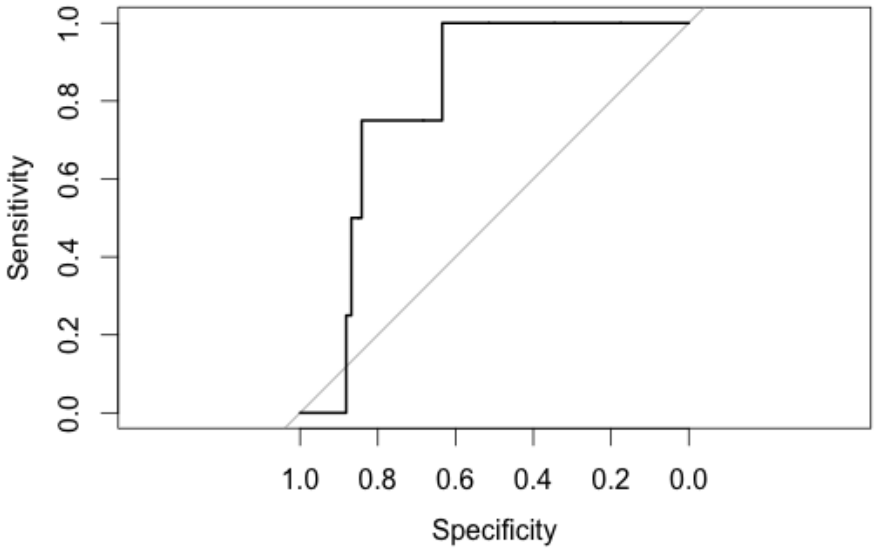

Figure 5-6. ROC curve for "30-Day Mortality Rate" in Random Forest.

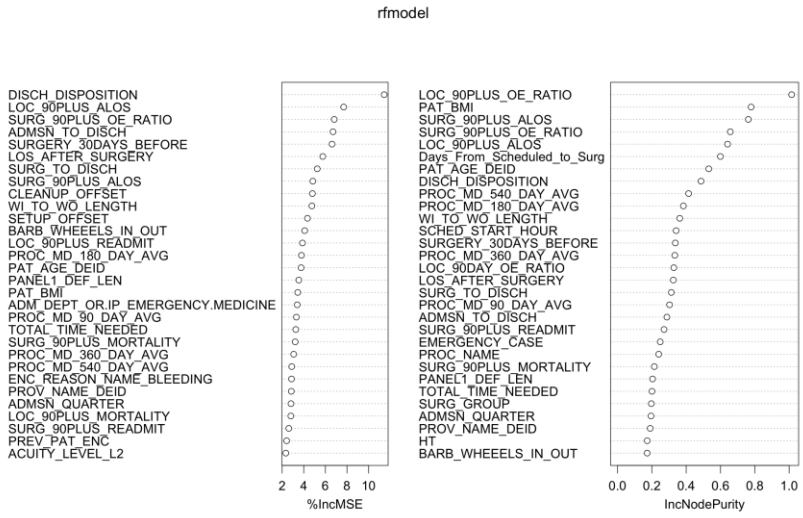

Figure 5-7. Variable Importance for "30-Day Mortality Rate" in Random Forest.

The AUC is 0.81, which is strong. This model might be sufficient for practical use.

The variable importance chart showed that discharge disposition, location's average Observed vs. Expected ratio, surgeon's average length of stay, and admission to discharge date are very strong predictors, due to the fact that their node purity is very high comparing to others.

3. Length of Stay

The Kappa statistic produced is 0.53, which implies that it is a fair model. We categorized this variable into three groups according to its distribution, there is a clear Poisson trend in the distribution, therefore the accuracy is satisfactory when each predicted value falls into 1 of 3 bins.

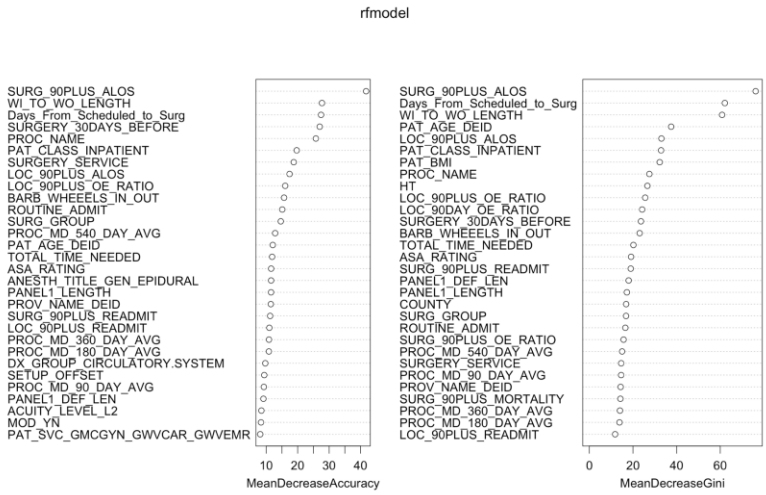

Figure 5-8. Variable Importance for "Length of Stay" in Random Forest.

The variable importance chart showed that the surgeon's average length of stay up to 90 days prior to current surgery, and wheel in to wheel out length are strong predictors.

4. Complication Observed vs. Expected Rate

The AUC is 0.88, which is satisfying. We can see from figure 5-8 that the length between admission to discharge date, discharge disposition and length of stay after surgery are strong predictors as they have high values of mean decrease Gini indexes.

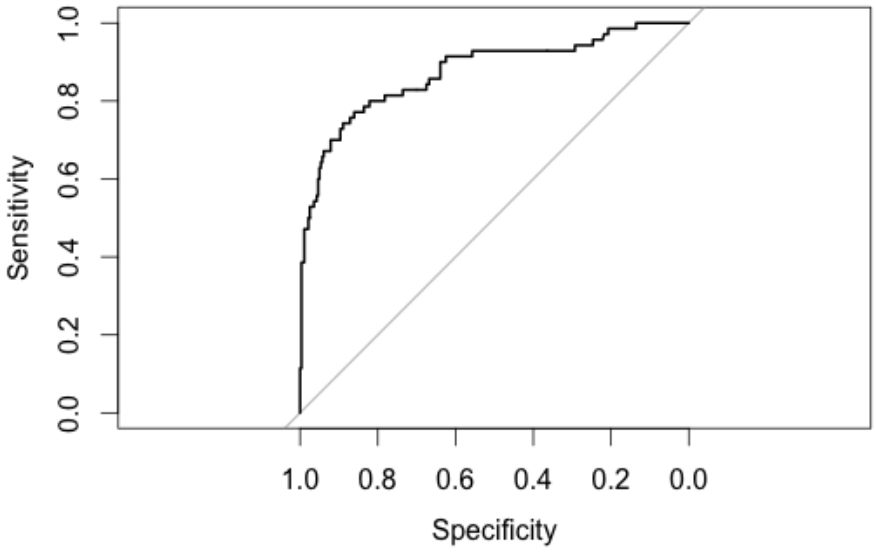

Figure 5-9. ROC curve for "Complication Rate" in Random Forest.

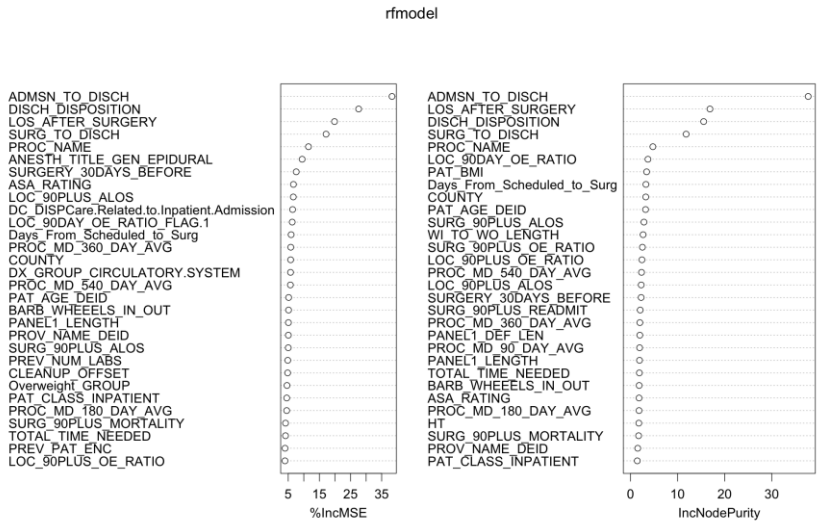

Figure 5-10. Variable Importance for "Complication rate" in Random Forest.

5.2.3 Gradient Boosting Method

GBM is similar to random forest in that it also uses decision trees, but instead of building a "forest", GBM gradually improves the existing model with a shrinkage parameter. In our model we set the distribution as Gaussian, number of trees to fit at 5000, the maximum depth of variable interactions at 5, and the shrinkage parameter (learning rate) at 0.001.

1. 30-Day Readmission Rate

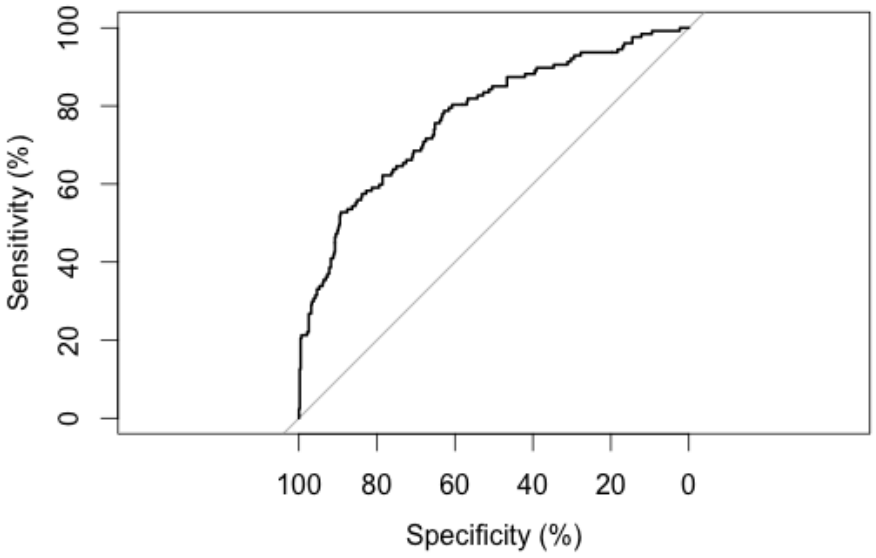

Figure 5-11. ROC curve for "30-Day Readmission Rate" in GBM.

The AUC is 0.77, which is the best performance model so far.

From Table 5-1 we find discharge disposition and whether the patient had a colorectal surgery before are strong predictors for patient readmission rate, because their values of relative influence are high. The AUC shown in Figure 5-11 is 0.85, which is the best performance model so far.

Table 5-1. Variable Importance for "30-Day Readmission Rate" in GBM.

Relative Influence	
DISCH_DISPOSITION	39.34
SURGERY_30DAYS_BEFORE	20.82
PROC_NAME	5.83
PROV_NAME_DEID	5.65
NUM_PREVIOUS_COLORECTAL	5.25
ADMITTING_DIAGNOSIS_Malignant.neoplasm	3.10
ADMSN_TO_DISCH	2.45
PROC_MD_360_DAY_AVG	1.64
LOS_AFTER_SURGERY	1.59
SURGERY_SERVICE	1.41
Days_From_Scheduled_to_Surg	1.34
ASA_RATING	1.34
ENC_REASON_NAME_BLEEDING	1.29
WEEKDAY_OF_SURGERY_SUNDAY	1.13
SURG_TO_DISCH	1.09
COUNTY	0.92
ENC_REASON_NAME_PAIN	0.66
LOC_90DAY_OE_RATIO	0.61
ACUITY_LEVEL_L2	0.57

2. 30-Day Mortality Rate

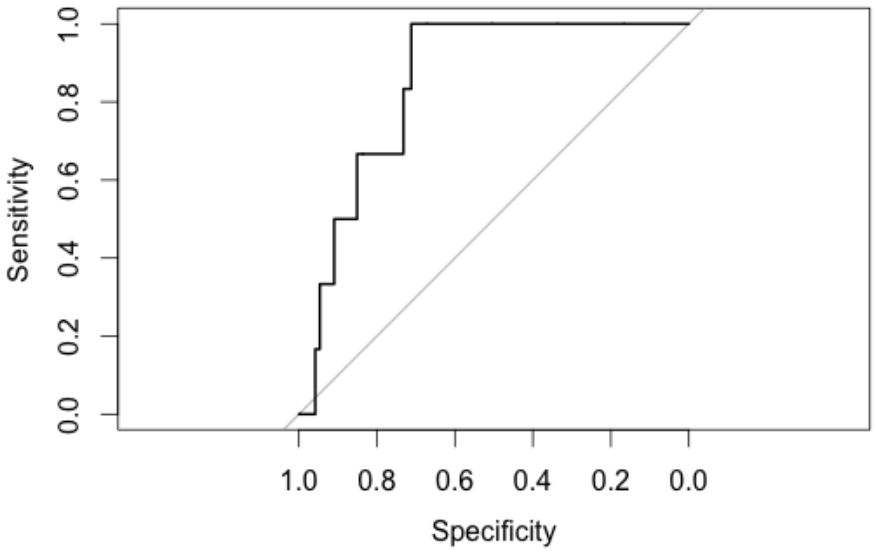

Figure 5-12. ROC curve for "30-Day Mortality Rate" in GBM.

Table 5-2. Variable Importance "30-Day Mortality Rate" in GBM.

Relative Influence	
DISCH_DISPOSITION	31.21
PROC_NAME	11.16
SURG_90PLUS_OE_RATIO	8.66
SCHED_START_HOUR	8.50
PAT_AGE_DEID	5.03
ASA_RATING	4.61
PROC_MD_360_DAY_AVG	4.35
Days_From_Scheduled_to_Surg	3.66
EMERGENCY_CASE	3.17
WI_TO_WO_LENGTH	2.86
TOTAL_TIME_NEEDED	2.55
SURG_90PLUS_ALOS	2.54
PROC_MD_540_DAY_AVG	2.53
PANEL1_LENGTH	2.02
LOC_90PLUS_OE_RATIO	1.76
Three_fiveft_GROUP	0.82
PROC_MD_180_DAY_AVG	0.53
ADMSN_QUARTER	0.47
BARB_WHEEELS_IN_OUT	0.46

From table 5-2 we can see that the discharge disposition is a strong predictor, it has relative influence value of 31.21.

3. Length of Stay

The Fleiss Kappa statistic is 0.45, which is fair. In this model, surgeon's average length of stay is a strong predictor for patient length of stay.

Table 5-3. Variable Importance for "Length of Stay" in GBM.

Relative Influence	
SURG_90PLUS_ALOS	22.13
Days_From_Scheduled_to_Surg	16.58
SURGERY_30DAYS_BEFORE	14.74
PAT_CLASS_INPATIENT	12.15
WI_TO_WO_LENGTH	6.63
PROC_NAME	6.48
ROUTINE_ADMIT	4.10
ASA_RATING	3.09
PAT_AGE_DEID	2.25
LOC_90PLUS_OE_RATIO	1.88
BARB_WHEEELS_IN_OUT	1.62
PROV_NAME_DEID	1.49
SURG_90PLUS_MORTALITY	1.34
SURGERY_SERVICE	1.01
NUM_PREVIOUS_COLORECTAL	0.87
SURG_GROUP	0.60
TOTAL_TIME_NEEDED	0.55
ANESTH_TITLE_GEN_EPIDURAL	0.55
HIGH_VOL_SURG	0.40

4. Complication Observed vs. Expected Rate

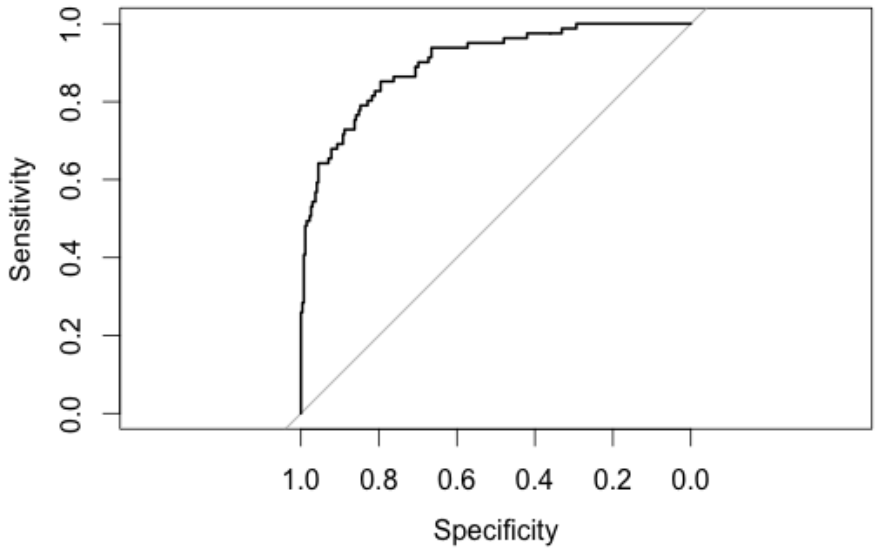

Figure 5-13. ROC curve for "Complication Rate" in GBM.

The AUC is 0.90, which is the best performance so far, further analysis will be conducted in later chapters.

Table 5-4. Variable Importance for "Complication Rate" in GBM.

Relative Influence	
ADMSN_TO_DISCH	58.57
DISCH_DISPOSITION	20.46
SURG_TO_DISCH	4.92
LOS_AFTER_SURGERY	4.48
PROC_NAME	2.81
PROV_NAME_DEID	1.74
ASA_RATING	0.82
LOC_90DAY_OE_RATIO	0.67
SURGERY_30DAYS_BEFORE	0.64
TOTAL_TIME_NEEDED	0.52
WI_TO_WO_LENGTH	0.50
SURG_90PLUS_READMIT	0.50
TRNS_DIFF_MEDSURG_HOSP_ADMIT	0.44
PROC_MD_180_DAY_AVG	0.35
BARB_WHEEELS_IN_OUT	0.29
NUM_PREVIOUS_COLORECTAL	0.27
ANESTH_TITLE_GEN_EPIDURAL	0.26
PAT_AGE_DEID	0.25
SURG_90PLUS_ALOS	0.18

The length between admission to discharge date and discharge disposition are the strongest predictors for complication rate, they have relative influence values os 58.57 and 20.46, which are way above other predictors.

5.2.4 Logistic Regression

Logistic regression is a method that uses a series of predictors to predict a binomial or multinomial outcome. Logistic regression makes use of either continuous or categorical data. In order to output the result as categorical, logistic regression takes the natural logarithm of the odds of the dependent variable to create a continuous format variable by transforming dependent

variable. This transformation works as a connection for continuous and categorical data. The predicted outcome is converted back into predicted odds by the inverse of the natural logarithm.

1. **30-day Readmission Rate**

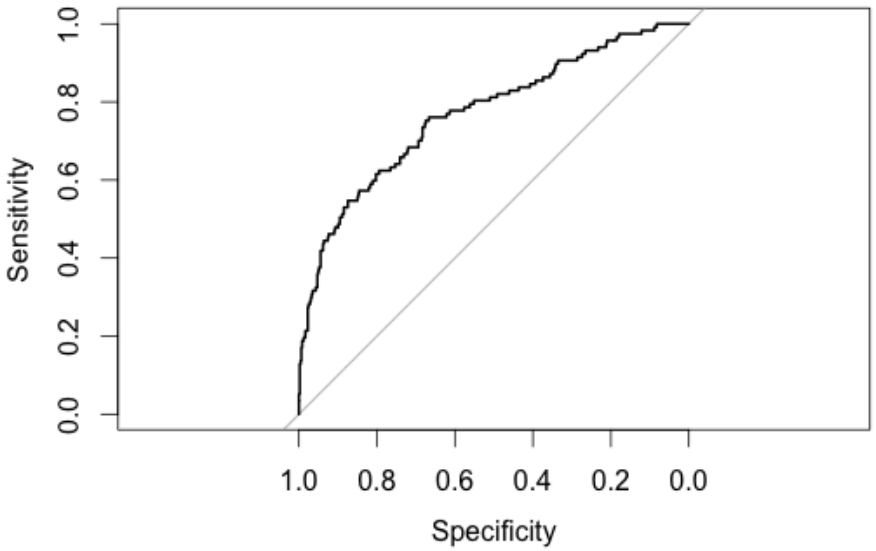

Figure 5-14. ROC curve for "30-Day Readmission Rate" in Logistic Regression.

In this model an AUC of 0.77 is achieved, which indicates fair performance.

From Table 5-5 we can see that in logistic regression model the most important predictors are almost the same as those displayed in other algorithms, and from Figure 5-15 we can observe that the AUC value is 0.70, which is acceptable.

Table 5-5. Variable Importance for "30-Day Readmission Rate" in Logistic Regression.

Predictor Importance	
SURGERY_30DAYS_BEFORE	8.84
DISCH_DISPOSITION	5.86
ADM_DEPT_OR.IP_EMERGENCY.MEDICINE	3.65
SURGERY_SERVICE	3.30
PROC_NAME	3.16
WEEKDAY_OF_SURGERY_SUNDAY	3.05
MS_DRG_392	3.02
DISCH_QUARTER	3.00
PAT_SVC_GMCGYN_GWVCAR_GWVEMR	2.84
ADMSN_QUARTER	2.73
ENC_REASON_NAME_BLEEDING	2.67
PAT_BMI	2.59
PROV_NAME	2.58
NUM_PREVIOUS_COLORECTAL	2.54
ASA_RATING	2.43
NEWBORN_ADMIT	2.19
SURG_90PLUS_READMIT	2.14
WOUND_CLASS	2.06
ENC_REASON_NAME_PAIN	2.06

2. 30-Day Mortality Rate

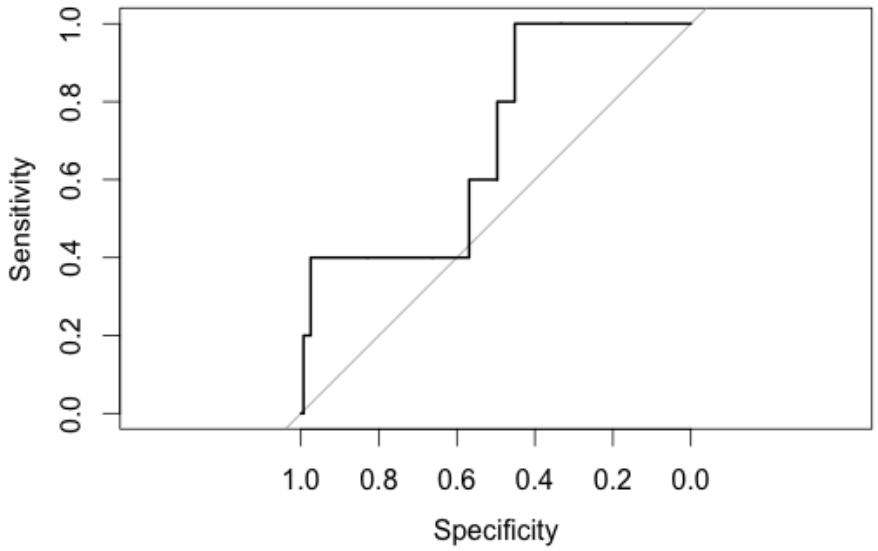

Figure 5-15. ROC curve for "30-Day Mortality Rate" in Logistic Regression.

Table 5-6. Variable Importance for "30-Day Mortality Rate" in Logistic Regression.

Predictor Importance	
SURG_GROUP	2.78
SCHED_START_HOUR	2.47
DISCH_DISPOSITION	2.29
DISCH_QUARTER	2.29
WEEKDAY_OF_SURGERY_MONDAY	2.01
HIGH_VOL_SURG	2.00
PROC_NAME	1.91
NUM_PREVIOUS_COLORECTAL	1.88
PREV_NUM_LABS	1.75
PAT_CLASS_Observation	1.72
SURG_QUARTER	1.71
ASA_RATING	1.69
PAT_AGE_DEID	1.66
PROV_NAME_DEID	1.65
PREV_NUM_CTS	1.63
ADMSN_TO_DISCH	1.59
DX_GROUP_CIRCULATORY.SYSTEM	1.53
COLORECTAL_SURGEON	1.51
ENC_REASON_NAME_PAIN	1.48

From the table we find that none of the predictors are significantly performing better. The importance indices gradually decrease at the same magnitude, indicating that this is a balanced model, removing a few predictors would not cause a significant change in AUC. The strong predictors in this case are surgeon groups (based on the number of colorectal surgeries they performed), surgery start hour and discharge disposition.

3. Length of Stay

The kappa statistic is 0.46, which is moderate. Results in Table 5-7 are also different from previous models, service line, and colorectal surgery indicator are strong predictors. This implies that the type of surgery and patient's physical conditions influence the rate of adverse events. Figure 5-16 shows the AUC value is 0.83, which is good.

Table 5-7. Variable Importance for "Length of Stay" in Logistic Regression.

Predictor Importance	
SURGERY_SERVICE	9.49
NEWBORN_ADMIT	7.43
SURGERY_30DAYS_BEFORE	6.71
COUNTY	6.57
Un1ft_GROUP	5.70
ADMSN_QUARTER	5.10
COPD_FLAG	5.04
SURG_QUARTER	4.88
AGE_GROUP_TEENAGER	4.64
PANEL1_IS_COMB_Y	4.31
REFERRED_PCP_ADMIT	4.16
LOC_90PLUS_READMIT	4.08
ASA_RATING	3.98
ANESTH_TITLE_GEN_EPIDURAL	3.86
Six_SevenFt_GROUP	3.81
PROC_NAME	3.81
ROUTINE_ADMIT	3.55
EMERGENCY_CASEU	3.39
SrCitizen_GROUP1	3.31

4. Complication Observed vs. Expected rate

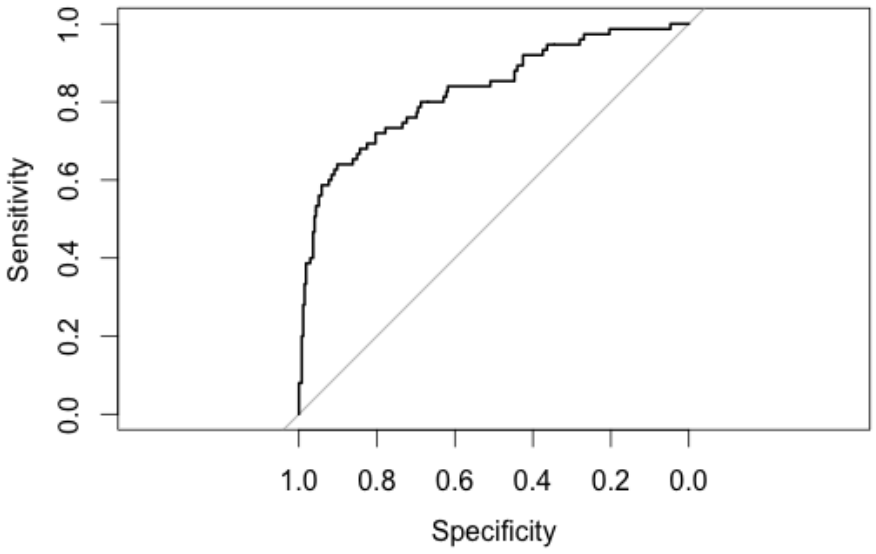

Figure 5-16. ROC curve for "Complication Rate" in Logistic Regression.

Table 5-8. Variable Importance for "Complication Rate" in Logistic Regression.

Predictor Importance	
ADMSN_TO_DISCH	4.02
ANESTH_TITLE_GEN_EPIDURAL	3.41
NEWBORN_ADMIT	3.23
PAT_CLASS_Emergency	2.76
PROC_MD_180_DAY_AVG_FLAG	2.56
PROC_NAME	2.55
DC_DISto.Hospital.Rehab_Home.with.IV.Care	2.53
Normal_GROUP	2.52
PROV_NAME_DEID	2.51
AGE_GROUP_TEENAGER	2.44
GEISINGER_ED_ADMIT	2.37
DISCH_DISPOSITION	2.33
PAT_CLASS_INPATIENT	2.33
TRNS_DIFF_MEDSURG_HOSP_ADMIT	2.30
LOC_90DAY_OE_RATIO_FLAG	2.28
SCHED_START_HOUR	2.25
ENC_REASON_NAME_PAIN	2.12
COUNTY	2.08
PAT_SVC_GMCGYN_GWVCAR_GWVEMR	2.06

In this model the strongest predictors are the length between admission to discharge date, whether the anesthesia is general/epidural, and whether the admission source is a newborn admission, the strong predictors are different from those in other methods. Overall the performance statistics for logistic regression are all slightly lower than tree based models, thus in real life applications, tree based models might be more appropriate for this kind of healthcare data.

Chapter 6

Result and Analysis

In the previous chapter, four methodologies were used to produce four desired outcomes. In this chapter, we summarize the results and conduct further analysis. In the first section, results are summarized and the model with the best performance is selected for analysis. In the subsequent section, variable importance is examined and its intrinsic implications are evaluated.

6.1 Model Results

Table 6-1. Model results.

	Naïve Bayes	Random Forest	GBM	Logistic Regression
30-day Readmission Rate[$]	0.68	0.77	0.77	0.77
30-day Mortality Rate[$]	0.65	0.81	0.85	0.70
Length of Stay[T]	0	0.53	0.45	0.46
Complication Rate[$]	0.76	0.88	0.90	0.83
Note: [$]: AUC; [T]: Kappa				

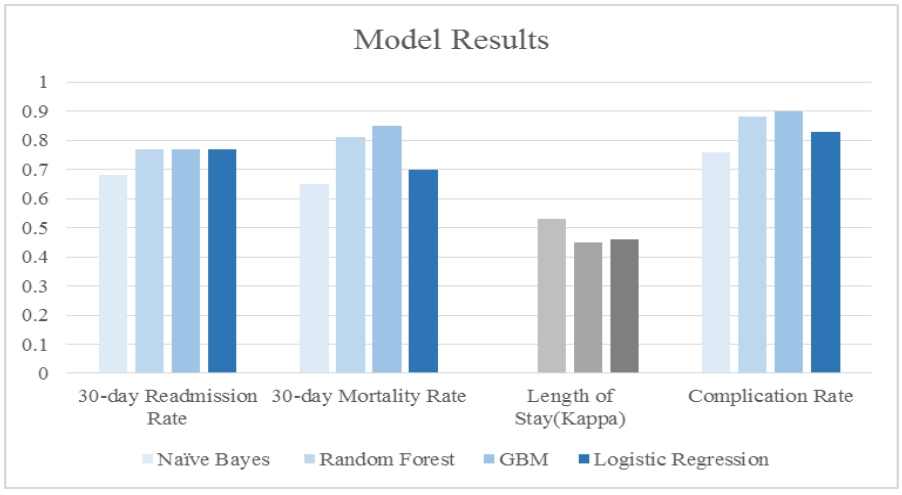

Figure 6-1. Model results bar plot.

The Kappa statistic produced in Naïve Bayes Method for length of stay is negative, which means that it is even worse than randomly guessing the results, thus it is rounded to 0 in this case.

Figure 6-1 visualizes the performance of each model for each outcome. "Complication rate" has good results. We created two bins for this variable according to the clear breaking point in the trend, and thus the highest AUC reached 0.90. The results for "30-Day Readmission Rate" and "30-Day Mortality Rate" are satisfactory, as they do not have a clear trend and are more stochastic, but the AUCs are both above 0.75, indicating that the best performance model is useful in hospitals. The models produced poor results for "Length of Stay". The highest Kappa statistic is 0.53, which is moderate, but not perfect. Classification methods might not be sufficient for this variable.

From figure 6-1 we can see that GBM has the best overall performance, although it's Kappa statistic for "Length of Stay" is slightly lower than Random Forest Model. Considering the other three outcomes, GBM out-performs, and thus it is chosen for further analysis.

6.2 Analysis of Predictor Importance

GBM had the best performance in this study. Therefore, we use variable importance output by GBM to analyze the desired outcomes. Only the top 10 predictors that are most important are analyzed.

6.2.1 30-Day Readmission Rate

Table 6-2. Important Predictors for "30-Day Readmission Rate".

Relative Influence	
DISCH_DISPOSITION	39.34
SURGERY_30DAYS_BEFORE	20.82
PROC_NAME	5.83
PROV_NAME_DEID	5.65
NUM_PREVIOUS_COLORECTAL	5.25
ADMITTING_DIAGNOSIS_Malignant.neoplasm	3.10
ADMSN_TO_DISCH	2.45
PROC_MD_360_DAY_AVG	1.64
LOS_AFTER_SURGERY	1.59
SURGERY_SERVICE	1.41

The AUC for this model is 0.77, which is persuasive and good. From table 6-1 it can be seen that the variable "DISCH_DISPOSITION" is the strongest predictor. It records information of where the patient was discharged to DISCH_DISPOSITION categories are as follows:

Table 6-3. Discharge Disposition Categories.

Care Related to Inpatient Admission
Disch/Trans Home, IV Provider - DEAC 4/13/08
Disch/Trans to Home with IV Care
LTACH
Routine Discharge
Tran/Discharge to Another Acute Facility - DEAC 4/13/08
Tran/Discharge to Home Health - DEAC 4/13/08
Tran/Discharge to Other Facility - DEAC 4/13/08
Tran/Discharge to Psychiatric Hospital
Tran/Discharge to Skilled Nursing Facility
Trans or Disch to Correlated Institution
Trans To Rehabilitation Center
Trans to Short-Term General Hospital
Trans/Discharge to Hospital Owned Rehab

According to their frequencies related to readmission rate these are grouped into the following:,

- Transfer to short-term general hospital
- Transfer/discharge to hospital owned rehabilitation center
- Transfer/discharge to correlated institution
- Long-Term Acute Care Hospital (LTACH)
- Transfer/discharge to skilled nursing facility
- Transfer/discharge to home with IV care
- Transfer/discharge to rehabilitation center
- Others

Most of the patients are discharged routinely or been provided care as inpatients. They do not have necessary connection with readmission rate, and therefore are grouped into the category "others". It can been where the patient was discharged to, the care they receive at that facility, and their health condition before discharge are strong indicators of readmission rate. If a patient received sufficient care at their discharge institution, their readmission rates are likely to be low.

The second strong predictor is "SURGERY_30DAYS_BEFORE" which indicates if the patient had a colorectal surgery 30 days before. If a patient had a colorectal surgery 30 days before, then it is possible that their sickness would relapse, or the previous surgery is not satisfactory that they have to receive it again. Another interpretation is that a patient having multiple colorectal surgeries within a month might have severe colorectal disorders. They are not in good health conditions, and thus the odds of readmission for current surgery would increase.

Other predictors such as surgery service line, length of stay after surgery, procedure name, provider name and number of previous colorectal surgeries the patient had have less impact on the outcome. If a patient has many colorectal surgeries before, or the patient has a long

length of stay, then it is likely that the patient is not in good health condition and the illness is severe, and thus readmission rate would increase.

The variable "ADMITTING_DIAGNOSIS_Malignant.neoplasm" is a transformed variable from previous surgery information, it records how many times a patient was previously diagnosed with Malignant Neoplasm, the more times the patient was diagnosed with this symptom, the more likely the patient has severe illness, and it would affect readmission rate.

6.2.2 30-Day Mortality Rate

The AUC in GBM model 0.85, indicating it is a strong model. The strongest predictor is "Discharge Disposition", which is the same as the 30-Day Readmission Rate. The care a patient receives after their discharge is associated with mortality rate, also discharge disposition reflects patient health condition information, patients not in good conditions would be discharged to intensive care unit (ICU) , and thus the mortality associated with ICU would be high.

Table 6-4. Importance predictors for "30-Day Mortality Rate".

Relative Influence	
DISCH_DISPOSITION	31.21
PROC_NAME	11.16
SURG_90PLUS_OE_RATIO	8.66
SCHED_START_HOUR	8.50
PAT_AGE_DEID	5.03
ASA_RATING	4.61
PROC_MD_360_DAY_AVG	4.35
Days_From_Scheduled_to_Surg	3.66
EMERGENCY_CASE	3.17
WI_TO_WO_LENGTH	2.86

The second strong indicator is "PROC_NAME" which records the name of operation procedure. It has over 200 categories and has been grouped as follows,

It is to our knowledge that different procedures aim at different disorders, and the corresponding mortality rate would be different.

The third strongest predictor is "SURG_90PLUS_OE_RATIO" which indicates surgeon's average observed vs. expected ratio up to 90 days prior to the patient's surgery date. "SURG_90PLUS_OE_RATIO" is a predictor that represents surgeon's surgical ability, the quality of surgery would affect patient mortality rate.

Table 6-5. Procedure Name Categories.

PROC_NAME
Others
Incision And Drainage Complex Post Operative Wound Infect
Laparoscopic Total Colectomy With Ileostomy Or Ileoproctostomy
Colectomy Total With Ileostomy
Laparoscopic Partial Colectomy Removal Of Terminal Ileum
Enterectomy Small Bowel Resection
Colectomy Total With Proctectomy And Ileostomy
Colectomy With End Colostomy
Colectomy Abdominal And Transanal Approach
Drainage Of Abdomen Abscess Open
Laparoscopic Total Colectomy With Colon Proctectomy Ileostomy
Colectomy With Coloproctostomy And Colostomy
Closure Enterostomy Large Intestine Resection And Anastomosis

Other predictors such as patient age, ASA rating are strong indicators because they record information on patient health condition. Surgery start hour also has influence on mortality rate, because surgery start hour might reflect the quality of surgery.

6.2.3 Length of Stay

The Kappa statistic in the best performance model is 0.45, which is not satisfactory, and not very persuasive. The strongest predictor is "SURG_90DAY_ALOS", which is surgeon's average length of stay 90 days prior to the patient's surgery date. It can be seen that surgeon's skill has a strong influence on patient length of stays. Longer lengths of stay may point to less skilled surgeon. Another understanding can be that a surgeon having a high length of stay record is a surgeon that performs surgeries for severe morbidity, so that the current length of stay tends to be long.

Table 6-6. Important predictors for "Length of Stay".

Relative Influence	
SURG_90PLUS_ALOS	22.13
Days_From_Scheduled_to_Surg	16.58
SURGERY_30DAYS_BEFORE	14.74
PAT_CLASS_INPATIENT	12.15
WI_TO_WO_LENGTH	6.63
PROC_NAME	6.48
ROUTINE_ADMIT	4.10
ASA_RATING	3.09
PAT_AGE_DEID	2.25
LOC_90PLUS_OE_RATIO	1.88

"Days from Scheduled to Surgery" is also a strong predictor, as the days contain information on patient health condition and morbidity, if the illness is severe or emergent, the days from scheduled to surgery would reflect the patient health condition, thus affecting the length of stay after surgery.

Whether a patient had colorectal surgery 30 days before also reflects patient's health condition, if a patient has to have two surgeries within a month, then the patient's health condition might not be perfect, and the length of stay tend to be longer.

6.2.4 Complication Observed vs. Expected Rate

The AUC is 0.90, which is near perfect. The strongest predictor is "ADMSN_TO_DISCH" which records the length of days between patient admission and discharge, if the length is long, the patient might not be in good health condition and thus likely to have more complications.

Table 6-7. Importance predictors for "Complication Rate".

Relative Influence	
ADMSN_TO_DISCH	58.57
DISCH_DISPOSITION	20.46
LOS_AFTER_SURGERY	4.48
PROC_NAME	2.81
PROV_NAME_DEID	1.74
ASA_RATING	0.82
LOC_90DAY_OE_RATIO	0.67
SURGERY_30DAYS_BEFORE	0.64
TOTAL_TIME_NEEDED	0.52
WI_TO_WO_LENGTH	0.50

"Discharge Disposition" is also important as the care patients receive at discharge facilities have different care levels, which would affect patient complication rate.

Length of stay after surgery is strong as well, the longer a patient stays in the hospital, the more likely the patient to have post site infections, bleeding, pain, weakness, and the patient's health condition may not be so well. Therefore, length of stay is strongly associated with complication rate.

Different procedures and providers affect complication rate. Total time needed for surgery and wheel to wheel out length is associated with surgery type, quality, features, which influence the complication rate.

Chapter 7

Conclusions and Discussion

In this chapter we draw conclusions from the entire study, provide some healthcare insights, and make a few recommendations. In the first section we analyze the overall model performance, and the reason why the best model out-performs other models. Finally, we address variable importance.

7.1 Model Performance

After the reported model performance analysis, we conclude that the models in the decreasing order of performance are: GBM, Random Forest, Logistic Regression and Naïve Bayes Method. The performance of a model depends on many factors

1. The features of data used in training data set. In this study the dataset is from EHR from a hospital, the features of healthcare data, economics data, military data and many others in different areas are all different, thus the best performance model may only do well in certain areas. Healthcare data usually have many predictors, among which there are many categorical variables, and they are often collected over a longer time period.
2. The size of data used in training impacts performance. Larger size dataset and smaller size dataset have different best performance models. When there are many interaction terms in a dataset, simple linear regression is not sufficient.

3. The features of models. Each model has its unique specialty, its advantages and disadvantages, tree based models are different from classification models. Thus to choose the best model, many trials should be performed with different algorithms.
4. The parameters in the model being set. Usually the parameters are set manually, such as the learning rate in GBM, the number of trees in Random Forest etc. Different sets of parameters can have impact on performance.
5. The size of subsets. Larger training sets tend to impact performance positively. There are many techniques such as bootstrapping, cross-validation, leave-one-out method can all help to reduce variance and bias. Thus, how to choose training set is an interesting and important topic.

In this study GBM and Random Forest had the best performance, and GBM is slightly better than Random Forest, which showed that tree methods work better than classifying methods for the healthcare data under use.

GBM is a technique that uses decision trees to output results. In Random Forest the sub-trees are run in parallel, while in GBM the sub-trees are run sequentially trial after trial. If the dataset size is very large, then Random Forest would run faster than GBM, but in this study, the data size is moderate, so the speed of model execution is similar. The reason why GBM runs better than Random Forest is that it gradually improves the results instead of producing a mass of results.

When comparing GBM with logistic regression, the prevailing opinion is that logistic regression usually produces more robust results, while decision tree techniques have a risk of over fitting the data, but when a dataset contains many interaction terms, GBM will out-perform logistic regression. In this dataset there are many interactions between predictors, because it is a healthcare dataset, many healthcare indexes and indicators have intrinsic connections, either two-

way or multi-way. In our limited study we can conclude that due to the interaction complexity GBM may be performing better.

Naïve Bayes is a simple classifier method that uses Bayes probability equation. Sometimes it is not very robust, especially when there are many predictors, using one equation to compute the relationship parameters will induce bias. This perhaps may be the reason why Naïve Bayes has the worst performance in this study.

7.2 Variable Importance

From the previous chapter variable importance has been analyzed for different outcomes, the most significant factors are related to surgeon's information, such as surgeon's average length of stay up to 90 days prior to surgery. Therefore, it can be concluded that surgeon's ability greatly affects the rate of adverse events. If a hospital wants to reduce patients adverse events they should focus on the fundamental cause, improving the quality of the surgeon and the skill of surgeons.

Another interpretation can be that different surgeons may be operating on diverse cases. For example, if a skilled surgeon is operating mainly on advanced stage cancer patients, the chance of survival may be low; compared to routine surgery performed by even a less skilled surgeon. Therefore we cannot argue that skill and adverse event occurrence are directly related.

Discharge disposition is also a significant factor. Where the patient was discharged to is greatly associated with the chances of encountering adverse events, if the patient was routinely discharged or received regular inpatient care, then the patient is likely to have a good recovery, and it is unlikely for the patient to be readmitted. In addition, the care patients receive after discharge is different, if a patient is well attended and taken care of in discharge facilities then

they are less likely to have adverse events. Thus hospitals should focus on the care they provide at discharge centers in order to improve post-surgical care quality.

Whether a patient had colorectal surgery 30 days before is a strong predictors for the outcomes. If a patient had colorectal surgery 30 days before, then the underlying illness is more likely to relapseand the patient may not be in good health condition, thus adverse events are likely to take place. To reduce adverse events in this aspect, patients should raise awareness of colorectal disorders, get regular body examinations, and engage in intervention programs.

Factors related to patient's socio-demographic aspects are also important. Patient's BMI, height, age, previous number of lab tests, CT scans, MRIs, previous number of surgeries all have substantial impact on the outcomes. In order to reduce adverse events early intervention programs, healthy lifestyle programs should be provided for patients.

Factors related to the surgery such as cut to close length, length of stay, procedure name, total time needed also affect the outcomes. Healthcare providers should focus on improving the quality of surgery to reduce adverse event rates.

Overall, in this study, with the available data set, good results are produced, the predictors are eligible for future use of prediction, and the models are also sufficient. There are yet more areas to be explored but having a basic idea of adverse events can be a great motivation and inspiration for healthcare providers.

Chapter 8

Limitations and future work

The models constructed in this study showed good performance, and therefore can be used in hospitals, However these models have limitations and further can be improved,

1. The fundamental limitation is data collection, in this study the dataset relates to a span of eight years from a single hospital, some of the data are missing, others are not displayed in the dataset, because a patient has to meet certain criteria in order for their information to be recorded in the database. In prediction problems the larger and more complete the dataset, the stronger the model will be. For a more persuasive result, more observations should be collected from different hospitals.
2. In data preparation process many missing values are being deleted, otherwise imputed, this will cause bias in the training model, better operations need to be used in order to deal with missing values, or more complete dataset should be collected from hospitals.
3. In data preparation process, many categorical variables have hundreds of categories, they have been grouped according to their frequency distributions and statistical analysis, but for a more precise result, we should consult the hospital on grouping of variables according to their knowledge of variable features.
4. Four models have been constructed for the outcomes, for a more complete study, more algorithms such as Support Vector Machine; Artificial Neural Network could be tried.

5. The outcomes “30-day Readmission Rate”, “30-day Morbidity Rate” and “Complication Observed vs. Expected Rate” are all in the forms of continuous numbers between 0 and 1, if a binary outcome is needed, a threshold should be used to separate the continuous number, how the threshold should be set is the decision made by healthcare provider. Involving more physicians in this study will result in better threshold definition.
6. In this study variable importance is decided by common sense. In order to have a more complete interpretation of results, we should consult and communicate with the hospital to validate the reasons behind the variable importance.

Based on the limitations and our observations we suggest the following future directions for the study:

1. More data should be collected from other hospitals distributed across the nation, accordingly we can have dataset that covers a more complete collection of socio-demographic, surgery, and healthcare provider information. This will result in the model being more robust and unbiased.
2. Other than the four models used in this study, more algorithms can be used to produce the outcome, and through more trials we can see if there is a better algorithm for this problem.
3. Although we had the hospital researchers and physicians involved with this research, more communication and continuous feedback may help improve model development, implementation and validation.
4. Some of the grouping of variables in this study are conducted subjectively and based on personal understanding. Grouping in a more objective and scientific way will help the study..

5. A threshold needs to be decided in order to produce binary outcome. In the future the choice of threshold should be made depending on the tolerance for type I and type II errors.
6. In the methodology chapter we took training set and testing set in a portion of 7.5:2.5 from the fixed original dataset. In the future we should try cross-validation and take random samples of training and testing sets for a thousand or more times, this would greatly reduce model bias.
7. In the future we could conduct duration analysis for complication rate and length of stay, these are not censored within 30-day range. Duration analysis deals with the instantaneous probability of duration end given that the duration since an event has not ended. This technique deals with time as a continuous measure. In this thesis, since a 30-day survey period is in effect, some measurements such as complication rate or length of stay may be right censored especially if their durations exceed 30 days. Right censoring occurs when the true end time is not known; rather, the end time is identified as the end of survey time.

References

Alves, A., Panis, Y., Mathieu, P., Mantion, G., Kwiatkowski, F., and Slim, K. (2005). "Postoperative mortality and morbidity in French patients undergoing colorectal surgery: results of a prospective multicenter study." *Archives of Surgery*, 140.3, 278-283.

Counihan, TC., and Favzza, J. (2009). "Fast-track colorectal surgery." Clinics in Colon and Rectal Surgery, 22(1), 60-72.

Cox, David R. (1958). "The regression analysis of binary sequences." *Journal of the Royal Statistical Society.* Series B (Methodological), 215-242.

Efron, B., Tibshirani, R. (1993). An Introduction to the Bootstrap.

Gareth, J., Witten, D., Hastie, T., and Tibshinari, R. (2013). An introduction to statistical learning.

Holdren, J.P. (2014). "Report to the President Better Health Care and Lower Costs: Accelerating Improvement Through Systems Engineering." Executive Office of the President President's Council of Advisors on Science and Technology.

Horwitz, L., Partovian, C., Lin, Z., Herrin, J., Grady, J., Conover, M., Montague, J., Dilaway, C., Bartczak, K., Ross, J., Bernheim, S., Drye, E., and Krumholz, M.H. (2011). "Hospital-wide (all-condition) 30-day risk-standardized readmission measure." Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. 10.

Humphries, M. (2013). "Missing Data & How to Deal: An overview of missing data." Population Research Center. University of Texas. Recuperado.

Landis, J.R., Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics*, 33 (1), 159–174.

Levinson, D. (2010). "*Adverse events in hospital: National incidence Medicare beneficiaries.*" Department of Health and Human Services-USA, office of inspector general, 1–7.

McCallum, A., and Nigam, K. (1998). “A comparison of event models for naive bayes text classification.” AAAI-98 workshop on learning for text categorization, 752, 41-48.

Nyce, C. (2007). "Predictive analytics white paper." American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America.

Siegel, E. (2013). Predictive analytics: The power to predict who will click, buy, lie, or die.

Smith, M., Saunders, R., Stuckhardt, L., and McGinnis, M.J. (2013). Best care at lower cost: the path to continuously learning health care in America.

Sørensen, L.T., and Jørgensen, T. (2003). "Short-term pre-operative smoking cessation intervention does not affect postoperative complications in colorectal surgery: a randomized clinical trial." *Colorectal Disease*, 5.4, 347-352.

Zmora, O., Pikarsky, A.J., and Wexner, S.D. (2001). "Bowel preparation for colorectal surgery". *Disease of the Colon & Rectum*, 44 (10), 1537-49.

http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-diagnosed Retrieved 07 April 2015.

https://www.facs.org/~/media/files/quality%20programs/nsqip/nsqipinfobook1012.ashx, Retrieved on June 15, 201

http://gim.unmc.edu/dxtests/roc3.htm, Retrieved 10 June, 2015

http://www.medicare.gov/hospitalcompare/Data/30-day-measures.html, Retrieved 15 June, 2015.

http://riskcalculator.facs.org/PatientInfo/PatientInfo, Retrieved on July 1, 2015

Appendix A

Naïve Bayes code in R

```
data<-read.csv("Myfinaldata.csv",nrows =-1,header =TRUE  )
COMPdata<-read.csv("Myfinaldata-Complications.csv",nrows =-1,header =TRUE  )
LOSdata<-read.csv("MyfinaldataLOS.csv",nrows =-1,header =TRUE  )
##Load data into R##
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98)
for(i in L){ data[,i]<-as.factor(data[,i])}
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98,121)
for(i in L){ COMPdata[,i]<-as.factor(COMPdata[,i])}
for(i in L){LOSdata[,i]<-as.factor(LOSdata[,i])}
##Turn some of the variables into categorical variables##
nbdataREAD<-data[-c(1,2,72)]
nbdataMORT<-data[-c(1,2,71)]
nbdataLOS<-LOSdata[-c(1,2,71,72)]
nbdataCOMP<-COMPdata[-c(1,2,71,72)]
##Remove the ID columns for computation, remove the correlated columns##
set.seed(1)
train.idx<-sample(nrow(nbdataCOMP),ceiling(nrow(nbdataCOMP)*0.75),replace=F)
train<-nbdataCOMP[train.idx,]
test<-nbdataCOMP[-train.idx,]
##Create testing and training sets##
train$COMPLICATION_OE<-as.numeric (train$COMPLICATION_OE)
test$COMPLICATION_OE<-as.numeric (test$COMPLICATION_OE)
##Turn outcomes into numerical outcomes##
library (e1071)
nbmodel<-naiveBayes(formula = COMPLICATION_OE ~ .,data = train,type="raw")
nbdatahat <- predict(nbmodel, newdata=test,type="raw")
yhat = data.frame(nbdatahat)
y <- test[c("COMPLICATION_OE")]
##Implement Naive Bayes Method to build model and validate on testing set##
library(pROC)
y<-data.frame(y)
y[,1]<-as.numeric(y[,1])
yhat[,1]<-as.numeric(yhat[,1])
roc(y$COMPLICATION_OE,yhat$X2,auc=TRUE,plot=TRUE,percent=TRUE,CI=TRUE)
##Produce ROC curve##
x<-data.frame(y,yhat)
library(psych)
cohen.kappa(x)
library(irr)
kappam.fleiss(x)
##Produce Kappa statistic##
```

Appendix B

Random Forest code in R

```
data<-read.csv("Myfinaldata.csv",nrows =-1,header =TRUE  )
COMPdata<-read.csv("Myfinaldata-Complications.csv",nrows =-1,header =TRUE  )
LOSdata<-read.csv("MyfinaldataLOS.csv",nrows =-1,header =TRUE  )
##Load dataset into R##
##turn in to categorical variables##
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98)
for(i in L){ data[,i]<-as.factor(data[,i])}
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98,121)
for(i in L){COMPdata[,i]<-as.factor(COMPdata[,i])}
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98)
for(i in L){LOSdata[,i]<-as.factor(LOSdata[,i])}
##Turn some variables into categorical variables##
rfdataREAD<-data[-c(1,2,72)]
rfdataMORT<-data[-c(1,2,71)]
rfdataLOS<-LOSdata[-c(1,2,71,72)]
rfdataCOMP<-COMPdata[-c(1,2,71,72)]
##Remove ID columns and correlated columns##
set.seed(1)
train.idx<-sample(nrow(rfdataCOMP),ceiling(nrow(rfdataCOMP)*0.75),replace=F)
train<-rfdataCOMP[train.idx,]
test<-rfdataCOMP[-train.idx,]
##Create testing and training sets##
library(randomForest)
set.seed(1)
rfmodel=randomForest(train$COMPLICATION_OE~.,data=train, mtry=50,ntree=500,importance=TRUE)
rfdatahat=predict(rfmodel, newdata=test)
var.imp1<-importance(rfmodel)
var.imp<-varImpPlot(rfmodel)
##Run model and validate on testing set, produce variable importance chart##
library(pROC)
yhat<-data.frame(rfdatahat)
y=test[,c("COMPLICATION_OE")]
y<-data.frame(y)
roc(y$y,yhat$rfdatahat,auc=TRUE,plot=TRUE,percent=TRUE,CI=TRUE)
##Produce ROC curve##
```

Appendix C

Gradient Boosting Method code in R

```
data<-read.csv("Myfinaldata.csv",nrows =-1,header =TRUE  )
COMPdata<-read.csv("Myfinaldata-Complications.csv",nrows =-1,header =TRUE  )
LOSdata<-read.csv("MyfinaldataLOS.csv",nrows =-1,header =TRUE  )
##Load dataset in R##
##turn in to categorical variables##
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98)
for(i in L){data[,i]<-as.factor(data[,i])}
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98,121)
for(i in L){COMPdata[,i]<-as.factor(COMPdata[,i])}
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98)
for(i in L){LOSdata[,i]<-as.factor(LOSdata[,i])}
##Formatting variables##
rfdataREAD<-data[-c(1,2,72)]
rfdataMORT<-data[-c(1,2,71)]
rfdataLOS<-LOSdata[-c(1,2,71,72)]
rfdataCOMP<-COMPdata[-c(1,2,71,72)]
##Date preparation##
set.seed(1)
train.idx<-sample(nrow(rfdataCOMP),ceiling(nrow(rfdataCOMP)*0.75),replace=F)
train<-rfdataCOMP[train.idx,]
test<-rfdataCOMP[-train.idx,]
##Create training and testing sets##
library(gbm)
set.seed(1)
gbmmodel=gbm(train$COMPLICATION_OE~.,data=train,n.trees=5000,distribution="gaussian",interactio
n.depth=5)
summary(gbmmodel)
gbmdatahat=predict(gbmmodel, newdata=test,n.trees=5000)
##Run regression and validate on testing set##
library(pROC)
yhat<-data.frame(gbmdatahat)
y=test[,c("COMPLICATION_OE")]
y<-data.frame(y)
roc(y$y,yhat$gbmdatahat,auc=TRUE,plot=TRUE,percent=TRUE,CI=TRUE)
relative.influence(gbmmodel)
##Create Roc curve and produce variable importance##
```

Appendix D

Logistic Regression code in R

```
data<-read.csv("Myfinaldata.csv",nrows =-1,header =TRUE  )
COMPdata<-read.csv("Myfinaldata-Complications.csv",nrows =-1,header =TRUE  )
LOSdata<-read.csv("MyfinaldataLOS.csv",nrows =-1,header =TRUE  )
##Load data in R##
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98)
for(i in L){data[,i]<-as.factor(data[,i])}
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98,121)
for(i in L){COMPdata[,i]<-as.factor(COMPdata[,i])}
L<-c(1,2,6,7,9,10,11,12,13,14,15,16,18,19,21,32,33,34,35,36,37,38,39,40,41:72,78,
    81,85,88,91,92,93,94,98)
for(i in L){LOSdata[,i]<-as.factor(LOSdata[,i])}
##Data formatting##
lrdataREAD<-data[-c(1,2,72)]
lrdataMORT<-data[-c(1,2,71)]
lrdataLOS<-LOSdata[-c(1,2,71,72)]
lrdataCOMP<-COMPdata[-c(1,2,71,72)]
##Data preparation##
set.seed(5)
train.idx<-sample(nrow(lrdataCOMP),ceiling(nrow(lrdataCOMP)*0.75),replace=F)
train<-lrdataCOMP[train.idx,]
test<-lrdataCOMP[-train.idx,]
##Data subsetting##
lrmodel<-glm(formula =  train$COMPLICATION_OE ~ .,data = train)
lrmodelstep<-step(lrmodel,direction = "both", steps = 5000)
lrdatahat <- predict(lrmodel, newdata=test, type="response")
yhat = data.frame(lrdatahat)
y <- test[c("COMPLICATION_OE")]
##Run logistic and stepwise regression##
library(pROC)
y<-data.frame(y)
roc(y$COMPLICATION_OE,yhat$lrdatahat,auc=TRUE,plot=TRUE,percent=TRUE,CI=TRUE)
##Produce ROC curve##
```