Life Expectancy Post Thoracic Surgery

Engineering College, Ajmer

Submitted By:
Priyanka Jain
Jaskirat Singh

Contents

Abstract

Operative mortality rates have been a topic of great interest among surgeons, patients, lawyers, and health policy administrators. Postoperative respiratory complications are the most common fatality following any type of thoracic surgery. The exact incidence is most contingent upon the preoperative health and lung function of the patient, and we would like to explore and understand how those conditions can drive these complications. One particular metric that has been used to quantify mortality rates in the past has been the thirty day mortality rate. This metric, however, may not be entirely comprehensive because many patients die shortly after this time period or become very weak, having to be taken to another facility before passing away there. As a result, many of these deaths are severely underreported. The scope of our project is to examine the mortality of patients within a full year after the surgery. More specifically, we are examining the underlying health factors of patients that could potentially be a powerful predictor for surgically related deaths.

1 Thoracic Surgery Data Set

1.1 Problem Statement

Patients who receive thoracic surgery for lung cancer do so with the expectation that their lives will be prolonged for a sufficient amount of time afterwards. This dataset presents data of patients, attributes and wheather they survived within a one year time frame the problem to solve is weather there is a way to determine post operative life expectancy of lung cancer patients from patient attributes in the dataset.

If there is pattern to be recognized with the attributes and weather the patients do not survive the one year mark, this would help physicians and patients make a more educated decision on weather they should proceed forward with surgery. If physicians feel the surgery will only hinder the patients quality of life with a recognized high risk of death within a one year time frame, then both parties can make decision to follow through on surgery or decide to find alternative treatment methods or palliative care.

1.2 Data Collection

The data was compiled by Marek Lubicz, Konrad Pawelczyk, Adam Rzechonek, and Jerzy Kolodziej at Polands Wroclaw Thoracic Surgery Centre for patients that were victims of severe lung resections for primary lung cancer in 2007 to 2011. The database is part of the National Lung Cancer Registry and is administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.

1.3 Features

The data is represented as follows: the rows are the patients (470 training examples) and the columns are the features (16 features and the true or false labelling). our feature set includes both continuous and classification data regarding to the patients health conditions at the time of the surgery. Each patient has 16 variables associated with them. Some of the continuous data includes a patients forced vital capacity, the maximum volume their lungs exhaled, size of original tumor, and age at surgery. In addition we have several classification features such as presence of pain before surgery, haemoptysis before surgery, cough before surgery, whether the patient is a smoker, whether the patient has asthma, and a few others.

Analyzing the data sets info shows many columns as object strings for T and F values. These include DGN, PRE6, PRE7, PRE8, PRE9, PRE10, PRE11, PRE14, PRE17, PRE19, PRE25, PRE30, PRE32. So, I converted the T and F object data types to 1 and 0 int data types in these columns using one hot encoding. The id column was removed because it is not necessary and lacking in any useful description of each patient. The column names were renamed with more human readable words instead of the original codes like id, diagnosis, forced capacity, forced expiration, zubrod scale, pain, haemoptysis, dyspnoea, cough, weakness, size of tumour, diabetas, mi 6months, pad, smoker, asthmtic, age.

1.4 Labels

The training examples are labeled with ground truth, we know whether the given patient lived or died. A false label indicatesthat the patient lived 1 year past the surgery, while a true label indicates the patient died within 1 year after the surgery. So, I converted the T and F object data types to 1 and 0 int data types in the last columnswhich is label.

2 Algorithms

Algorithms that we used are follows:
Naive Bayes
SVM
Logistic Regression
k folds cross validation
Bootstraping
Random Forest
Feature Selection

2.1 Naive Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, every pair of features being classified is independent of each other.In our problem we use Bernoulli Naive Bayes.

2.2 Bernoulli Naive Bayes

Bernoulli Naive Bayes: In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks, where binary term occurrence(a word occurs in a document or not) features are used rather than term frequencies(frequency of a word in the document).

2.3 SVM

In machine learning, Support vector machine(SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It is mostly used in classification problems. In this algorithm, each data item is plotted as a point in n dimensional space (where n is number of features), with the value of each feature being the value of a particular coordinate. Then, classification is performed

by finding the hyper-plane that best differentiates the two classes.
In addition to performing linear classification, SVMs can efficiently perform a non linear classification, implicitly mapping their inputs into high dimensional feature spaces.

2.4 Logistic Regression

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X.

2.5 Random Forest

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging.

2.6 Bootstrapping

The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. It is used in applied machine learning to estimate the skill of machine learning models when making predictions on data not included in the training data.

2.7 k folds cross validation

In this method, we split the dataset into k number of subsets(known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

2.8 Feature Selection

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

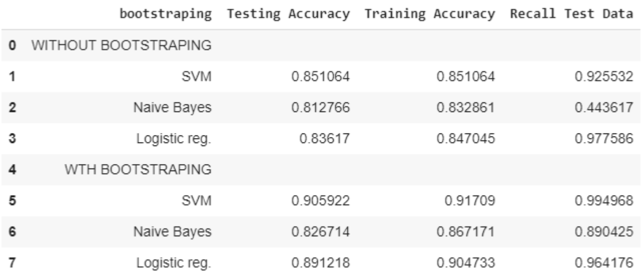
3 Process

3.1 Naive Bayes, SVM, and Logistic Regression approach

Initially, we simply ran Naive Bayes, SVM, and Logistic Regression to obtain a better understanding of our data. We trained our algorithms on the full data set and obtained our testing values using using k-folds cross validation (setting k equal to 10).And get train and test data accuracy and recall.

We then implemented a bootstrapping approach on our three algorithms Naive Bayes, SVM, and Logistic Regression.And get train and test data accuracy and recall.Then we compare the result.

3.2 Accuracies and Recalls table for Different Classifiers:



	bootstraping	Testing Accuracy	Training Accuracy	Recall Test Data
0	WITHOUT BOOTSTRAPING			
1	SVM	0.851064	0.851064	0.925532
2	Naive Bayes	0.812766	0.832861	0.443617
3	Logistic reg.	0.83617	0.847045	0.977586
4	WTH BOOTSTRAPING			
5	SVM	0.905922	0.91709	0.994968
6	Naive Bayes	0.826714	0.867171	0.890425
7	Logistic reg.	0.891218	0.904733	0.964176

Figure 1: Note: All results are based on dataset size of 470

3.3 Accuracies graph for Different Classifiers:

Reffering to the graph on the below for the different classifiers accuracies with bootstrapping.

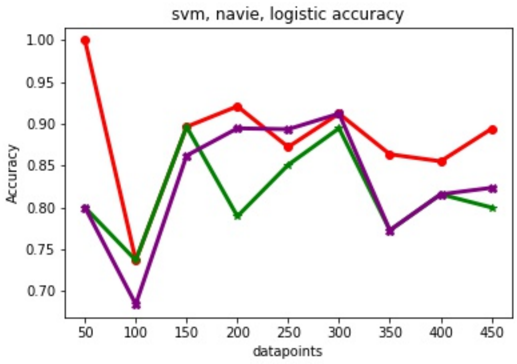


Figure 2: Testing Accuracy with Bootstrapping

3.4 Recalls graph for Different Classifiers:

Reffering to the graph on the below for the different classifiers recalls with bootstrapping.

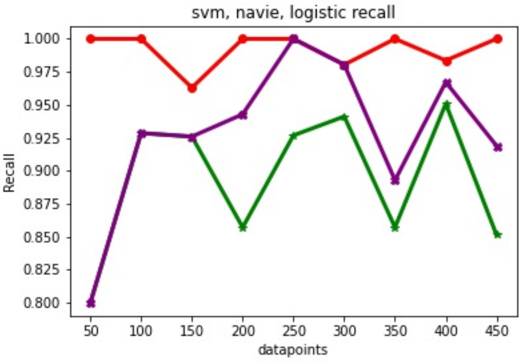


Figure 3: Testing Recall with Bootstrapping

3.5 Random Forest approach:

Our next approach to tackle our variance problem was the random forest. Using the tree classification algorithm allows us to average multiple deep decision trees that would be trained on different parts of the same training set. This approach comes at the expense of a slightly higher bias (and potentially some loss of interpretability) but will ultimately improve the final performance of the model. The classification tree algorithm works very well when you have mixed categories of continuous and binary features. By taking random subsets of features (examining all of the possible split points), the algorithm can make a decision on which feature is the best and pick that one, while still accounting for uncertainty. We then combined our bootstrapping approach to incorporate the random forest algorithm.

First we run Random Forest without Bootstrapping and obtained our testing values using using k-folds cross validation (setting k equal to 10). And obtained accuracy and recall.

3.6 Accuracy graph for Random forest without bootstrapping:

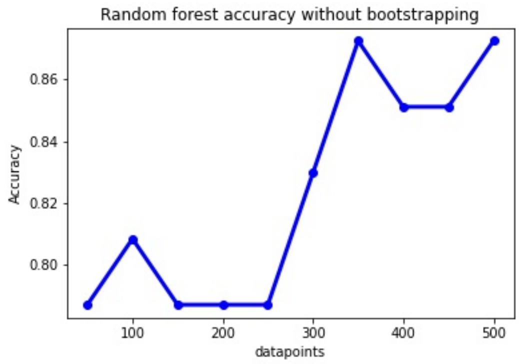


Figure 4: Testing Accuracy without Bootstrapping

3.7 Recall graph for Random forest without bootstrapping:

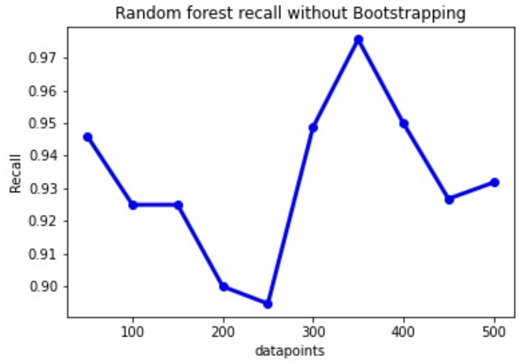


Figure 5: Testing Recall without Bootstrapping

Now repeat the process with bootstrapping and obtained accuracy and recall.

3.8 Accuracy graph for Random forest with bootstrapping:

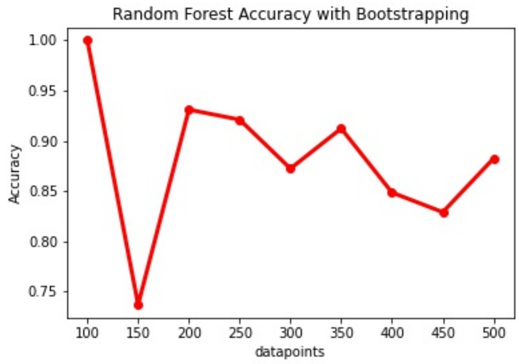


Figure 6: Testing Accuracy with Bootstrapping

3.9 Recall graph for Random forest with bootstrapping:

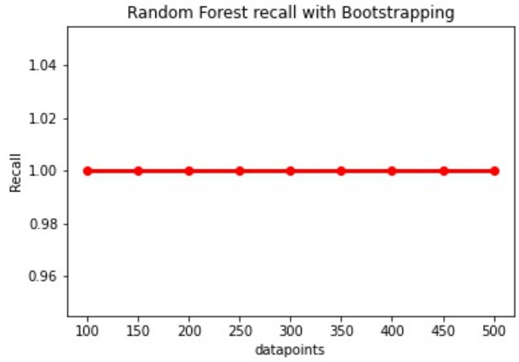
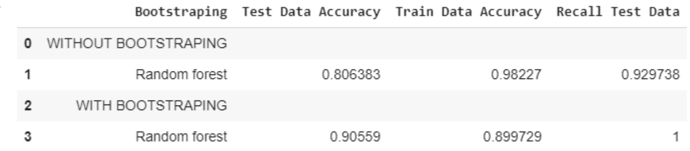


Figure 7: Testing Recall with Bootstrapping

Then apply feature selection to reduce the features and get most important features.

3.10 Accuracy and Recall table for Random forest with/without bootstrapping:

Compare the accuracy and recall for random forest with or without bootstrapping.



	Bootstraping	Test Data Accuracy	Train Data Accuracy	Recall Test Data
0	WITHOUT BOOTSTRAPING			
1	Random forest	0.806383	0.98227	0.929738
2	WITH BOOTSTRAPING			
3	Random forest	0.90559	0.899729	1

Figure 8: Note: All results are based on dataset size of 470

4 Front End

For front end we use
Flask
Html
Css

4.1 Flask

Flask is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application.
All html files contained in template engines. Using a template engine will save you a lot of time when creating your application but also when updating and maintaining it.

4.2 Html

HTML Stands for Hypertext Markup Language. HTML is the language used to create webpages. "Hypertext" refers to the hyperlinks that an HTML page may contain. Markup language refers to the way tags are used to define the page layout and elements within the page.HTML is used to create the actual content of the page, such as written text. HTML describes the structure of a Web page. HTML consists of a series of elements. HTML elements tell the browser how to display the content. HTML elements are represented by tags. HTML tags label pieces of content such as heading, paragraph, table, and so on.Browsers do not display the HTML tags, but use them to render the content of the page.

Using html we print all accuracies and recalls and show graphs.

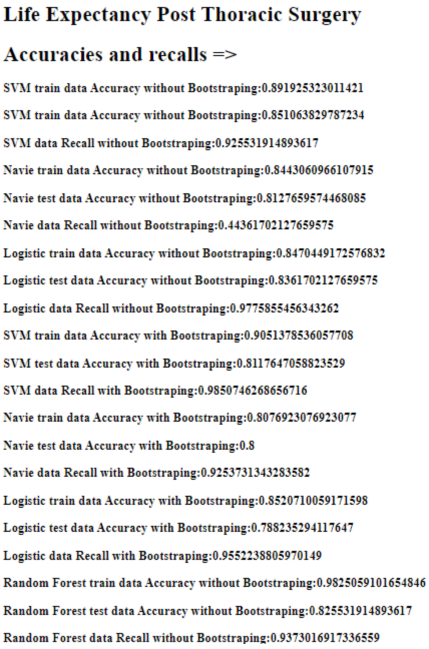
4.3 CSS

CSS stands for Cascading Style Sheets. It is a style sheet language which is used to describe the look and formatting of a document written in markup language. It is generally used with HTML to change the style of web pages.CSS is responsible for the design or style of the website, including the layout, visual effects and background colour. CSS describes how HTML elements are to be displayed on screen, paper, or in other media.CSS saves

a lot of work. It can control the layout of multiple web pages all at once. External stylesheets are stored in CSS files.
we use css to give margin , height and weidth to div which stores graph figures.

4.4 Accuracies an Recalls

Here all the accuracies and recalls of train and test data for different classifiers with and without bootstrapping.



Life Expectancy Post Thoracic Surgery

Accuracies and recalls =>

SVM train data Accuracy without Bootstraping:0.891925323011421

SVM train data Accuracy without Bootstraping:0.851063829787234

SVM data Recall without Bootstraping:0.925531914893617

Navie train data Accuracy without Bootstraping:0.8443060966107915

Navie test data Accuracy without Bootstraping:0.8127659574468085

Navie data Recall without Bootstraping:0.44361702127659575

Logistic train data Accuracy without Bootstraping:0.8470449172576832

Logistic test data Accuracy without Bootstraping:0.8361702127659575

Logistic data Recall without Bootstraping:0.9775855456343262

SVM train data Accuracy with Bootstraping:0.9051378536057708

SVM test data Accuracy with Bootstraping:0.8117647058823529

SVM data Recall with Bootstraping:0.9850746268656716

Navie train data Accuracy with Bootstraping:0.8076923076923077

Navie test data Accuracy with Bootstraping:0.8

Navie data Recall with Bootstraping:0.9253731343283582

Logistic train data Accuracy with Bootstraping:0.8520710059171598

Logistic test data Accuracy with Bootstraping:0.788235294117647

Logistic data Recall with Bootstraping:0.9552238805970149

Random Forest train data Accuracy without Bootstraping:0.9825059101654846

Random Forest test data Accuracy without Bootstraping:0.825531914893617

Random Forest data Recall without Bootstraping:0.9373016917336559

Figure 9: Accuracies an Recalls

4.5 Graph

Here all the accuracies and recalls graph for different classifiers with and without bootstrapping.

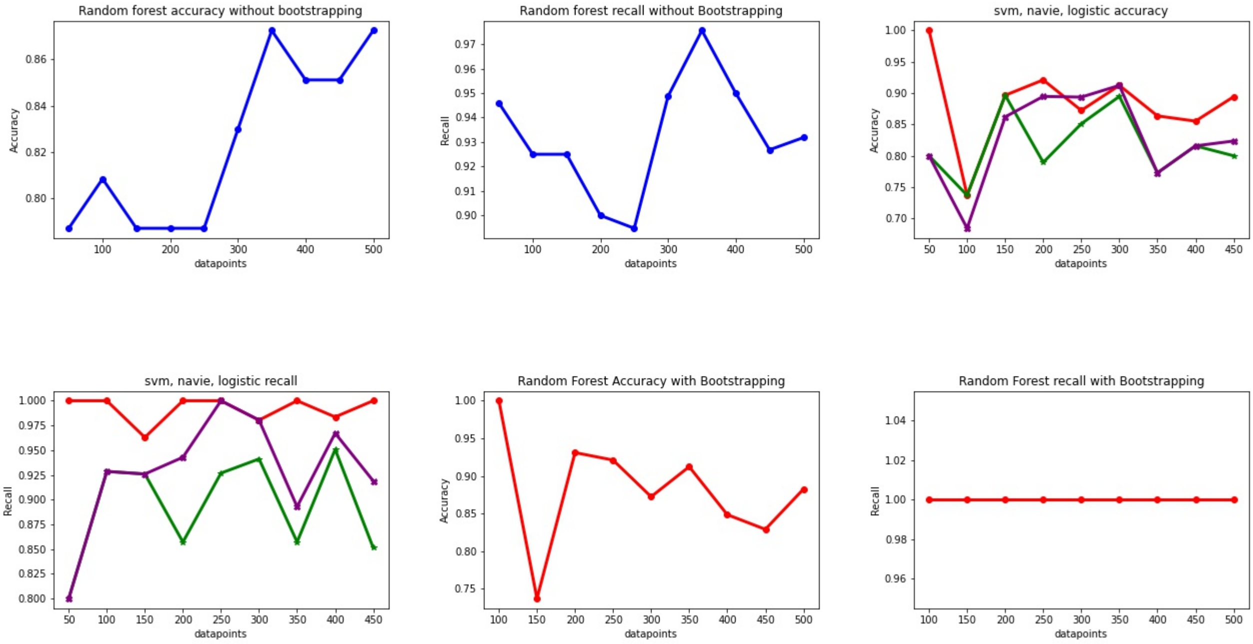


Figure 10: Different graphs of accuracy and recall for different classifiers

5 Conclusion

The desired outcome will depend on the hospitals or client and how they view the detriment of giving accuracies and recalls compared to the true predicitions for live or death outcomes for patients. In other words, how they want to score the efficiency of the model.

By the end of all of our iterations and improvements, we were able to achieve fairly good results with the different classifiers and bootstrapping. These results have large implications in the medical field. An analysis similar to ours could be performed before a patient goes in for surgery to see how high risk they are, which could be crucial information.