Thoracic Surgery Survival Prediction using Boosted Random Forest and Support Vector Machines

Michael Mendes

Introduction and Background

Thoracic surgery, specifically lung resections, is a standard procedure performed on patients with lung issues including tumors of the lung, diseases such as emphysema or tuberculosis, lung abscesses, and collapse of lung tissue. It involves the removal of some portion of the lung as well as diseased surrounding tissue. Thoracic surgery incidence is on the rise in conjunction with lung cancer incidence as well as a decrease in air quality in heavily populated areas. Asthma and other lighter breathing conditions can be exacerbated by decrease in air quality that may lead to surgical intervention as an option. Lung cancer, in particular, caused 1.6 million deaths worldwide in 2012 and has continued to increase across the globe (de Groot 2018). While lung cancer incidence and mortality in the US is on the decline thanks to increased awareness of the dangers of smoking, other areas of the globe face an opposite trend. Surgery is a primary treatment option in the case of lung cancers, since it involves removing the diseased tissue before metastasis can occur.

While even older studies of lung resection mortality report fairly low numbers, only 58 over the course of 4 years in one study in Tokyo, Japan, mortality resulting from surgical complications is an ever-present concern when deciding the course of a patient's treatment (Watanabe 2004). In this investigation, we are looking to find the optimal predictors for patient mortality post lung resection surgery. The hypotheses for this investigation are as follows:

Null Hypothesis 1

There are no predictors or set of predictors that provide a method to discern patient survival 1 year after thoracic surgery.

Alternative Hypothesis 1

There is some predictor or set of predictors that provide an accurate means of discerning a patient's survival 1 year after thoracic surgery.

Null Hypothesis 2

There is no optimal model for the prediction of lung resection surgery analysis; all machine learning approached to predicting surgery outcomes are similar and ineffective.

Alternative hypothesis 2

There is some optimal and effective model for predicting thoracic surgery survival with prediction results better than other models.

This investigation will use support vector machine models and random forests to decide if there are any variables in patient metadata that optimally and practically predict the incidence of mortality post lung cancer resection. The random forest model will be used to select the most relevant variables from the data set, while the support vector machine will be developed to use all the variables available to construct a hyperplane capable of separating survival one year post operation from mortality prediction.

The data set used by this study was collected between 2007 and 2011 in Wroclaw Poland at the Wroclaw Thoracic Surgery Center (Maciej 2014). It contains information obtained both pre and post operatively by survey. Survey results contained a total of 139 possible predictors and 1200 records. This has been cut to a total of 470 patients with completion of all fields and 17 total predictors deemed most relevant to the study by the support vector machine model provided by Maciej, 2014. Through the use of another support vector machine model and random forests, we look to further divide the data into the most accurate predictor of patient mortality and the most efficient model for predicting mortality. These models could serve as a decision support system for physicians when deciding whether or not a patient should undergo a lung resection procedure.

For boosted random forest selection of relevant features, the original data set underwent no preprocessing (besides the transfer of true/false values to binary numerical values) since forest methods are able to process categorical data regardless of ordinal or nominal status. id was removed as a possible predictor, primarily since relative importance was rated fairly high (0.2, the highest predictor value). Shrinkage value of 0.1, 0.01, and 0.001 were used for this random forest.

Methods

Support vector machine models were used with polynomial, linear, and radial kernels of differing cost and degree to find the model with the least error. 10-fold cross validation was used to evaluate these models to select the optimal degree and cost.

There were a few issues with the dataset that required preprocessing and consideration of results obtained. a few of the variables including PRE19 and DGN1 proved to have such low numbers of positive values that their use as a predictor is deemed negligible. PRE19 indicates a history of myocardial infarction for the patient in question; only 3 of the 470 records actually had a “true” value for this field. A full description of each predictor in the dataset follows:

Response variable: Risk1Y - Risk of death less than 1-year after the resection occurred.
This will be a Boolean variable (1 for death within a year of treatment, 0 for survival).

Feature Description:
NOTE: For all Boolean values, the dataset will be changed for analysis so that TRUE=1 and FALSE=0

id: the identification number for each patient record

This is an arbitrary value assigned to each anonymized record.

DGN: 8 groups (DGN1 - DGN8) of diagnosis classes consisting of unique sets of ICD10 codes corresponding to the patients' primary diagnosis and secondary tumors if present, qualitative data.

No information regarding the ICD10 codes contributing to each class were available from initial literature review; contact with the Institute of Tuberculosis and Pulmonary Disease in Warsaw has been initiated for possible retrieval of original diagnosis code values.

PRE4: Forced Vital Capacity (FVC), quantitative data.

Forced vital capacity is the amount of air that can be forcibly pushed from the lungs as quickly as possible after inhaling as much as possible. This value is measured in liters by a device called a pneumotach. A pneumotach screen measures the volume of air exhaled by measuring the pressure drop during exhalation. A measurement is taken and recorded after at least 3 measures within 150 mL of each other are taken.

PRE5: Forced Expiratory Volume (FEV1), quantitative data.

The maximal amount of air that a patient can exhale within 1 second measured in liters. This measures pulmonary obstruction, and it is commonly used with FVC to measure amount of obstruction incurred by a given diagnosis. A few of the values in the dataset do not make sense, and consideration will be taken in removing these records from the analysis. For example, a few of the patient records have FEV1 values higher than their FVC.

PRE6: Performance Status of the patient, scaled data.

The performance status for a patient is measured on the Zubrod (also known as the ECOG scale for the Eastern Cooperative Oncology Group) scale. For this data set, data between scale value 0 - 2 is used, though the scale has 5 levels total. The three scale levels included in this study are as follows:

0 - normal activity

1 - Symptomatic but still ambulatory and self-sufficient

2 - Ambulatory more than 50% of the time with some occasional third party assistance for activities

PRE7: Pain incurred pre-operation, Boolean data.

This is a Boolean measure of whether or not the patient was in pain pre-operation

PRE8: Hemoptysis before surgery, Boolean data.

This is a Boolean measure of hemoptysis, or coughing up blood, pre-operation.

PRE9: Dyspnea before surgery, Boolean data.

Dyspnea is a fairly broad term encompassing any difficulty with breathing. It is often used as a synonym for shortness of breath, though it can also be described as a tightening of the chest or a feeling of suffocation.

PRE10: Cough symptom before surgery, Boolean data.

This is a Boolean measure of whether or not the patient experienced coughing regularly before surgery. The data contains no measure of severity of coughing or frequency.

PRE11: Weakness experienced before surgery, Boolean data.

This is a Boolean measure of whether or not the patient experienced weakness pre-operation. Weakness is often an accompanying symptom with shortness of breath and lung obstruction.

PRE14: Size of the original tumor on the TNM scale, scale data.

This attribute corresponds to the T of the TNM cancer staging system. It is the size of the main tumor found in the patient. In our study, these range from OC11 to OC14, corresponding to T1 through T4.

T1 - tumor of size 3cm or less

T2 - tumor size between 3cm and 7cm

T3 - tumor size greater than 7cm, or the tumor that is specifically invading one of several lung-peripheral areas

T4 - This level has no size parameters, but refers to the extent of metastasis the tumor has undergone. Tumors that have invaded the mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, vertebral body, or carina are included. Tumors that have formed a separate nodule in an ipsilateral lobe are also included.

PRE17: Type 2 Diabetes Mellitus, Boolean data.

This is a Boolean measure of diabetes mellitus type 2 before the operation. Diabetes mellitus type 2 has been associated with higher risk of mortality in a surgical setting for several types of surgery including cardiac and orthopedic procedures.

PRE19: Myocardial infarction occurrence within the 6 months preceding surgery, Boolean data.

This Boolean attribute records the occurrence of myocardial infarction within the 6 months leading to thoracic surgery. Myocardial infarction (also known as heart attack), as well as peripheral arterial disease have been shown to be complicating factors in surgical outcomes and recovery rate.

PRE25: Peripheral arterial disease (PAD) occurrence, Boolean data.

If the patient has been diagnosed with PAD, this Boolean attribute will be a 1, 0 if negative.

PRE30: Smoking status, Boolean data.

This Boolean attribute records if the patient has ever smoked before, and has no information on pack-years smoked or type of tobacco product smoked.

PRE32: Asthma diagnosis, Boolean data.

If the patient has a diagnosis of asthma, this attribute will be a 1. Asthma causes pulmonary obstruction on its own in moderate to severe cases.

AGE: Patient's age, quantitative data.

This is the patient's age measured in years at the time of the surgery.

DGN, PRE14, and PRE6 each had to be preprocessed to numerical data before accurate SVM models could be produced. DGN represents the specific ICD10 combination of diagnosis codes for each patient. While neither Maciej 2014 or the Wroclaw Thoracic Surgery Center could provide an accurate description of these diagnosis, the role of the actual diagnosis of a patient and their survival of lung resection surgery is without question important. Prognosis of cancers can be widely varied, and some of these diagnoses may have been ordinal, though in this investigation the 8 possible levels are treated as nominal values. One hot encoding was used to change these nominal values into a more machine learning algorithm friendly version. Inclusion of these dummy variables expanded the feature space and increased bias in the model, as well as possible over-fitting of the data. An unfortunate side-effect of this was that optimization of the polynomial SVM model took too long to be computationally feasible (around 2 hours per each cross-validated attempt at optimal cost and degree). PRE14 is a measure of the size of the tumor that each patient had, and thus could be easily translated to an ordinal number scale from 1 to 4 from OC11 to OC14. PRE6 was a measure of the ambulatory ability of the patient, and thus was easily transferred to an ordinal 0 to 2 scale from PRZ1 to PRZ2.

Results and Conclusion

As expected, the boosted random forest model predicted DGN as the most relevant predictor across each attempted shrinkage value. Shrinkage values of 0.1 and 0.01 provided mixed results with DGN, PRE4, PRE5, and PRE14. With a shrinkage of 0.1, age was also close behind PRE14 with a relative influence score of 11.31, whereas PRE14 was 11.98. From prior assumptions, age of the patient was considered to be a primary predictor since the aging has profound affects on surgical recovery rate (Watanabe 2014). However, as the shrinkage value decreased, age was removed from the list of primary predictors and DGN and PRE14 were shown to be the highest relative influencers, with relative influence scores of 35.62 and 32.18 with a shrinkage of 0.001. This also follows previous assumptions about predictors. PRE14 tracked the size and grade of the tumor involved in the patients' diagnoses, and it goes without saying that the larger the area of lung removed, the less likely a patient is to survive the procedure by loss of lung capacity. DGN was the leading predictor in each boosted random forest attempt, and it is regrettable that further specification as to the individual diagnosis codes could not be retrieved. We expected a small shrinkage value to provide good results, since our feature set was fairly expansive.

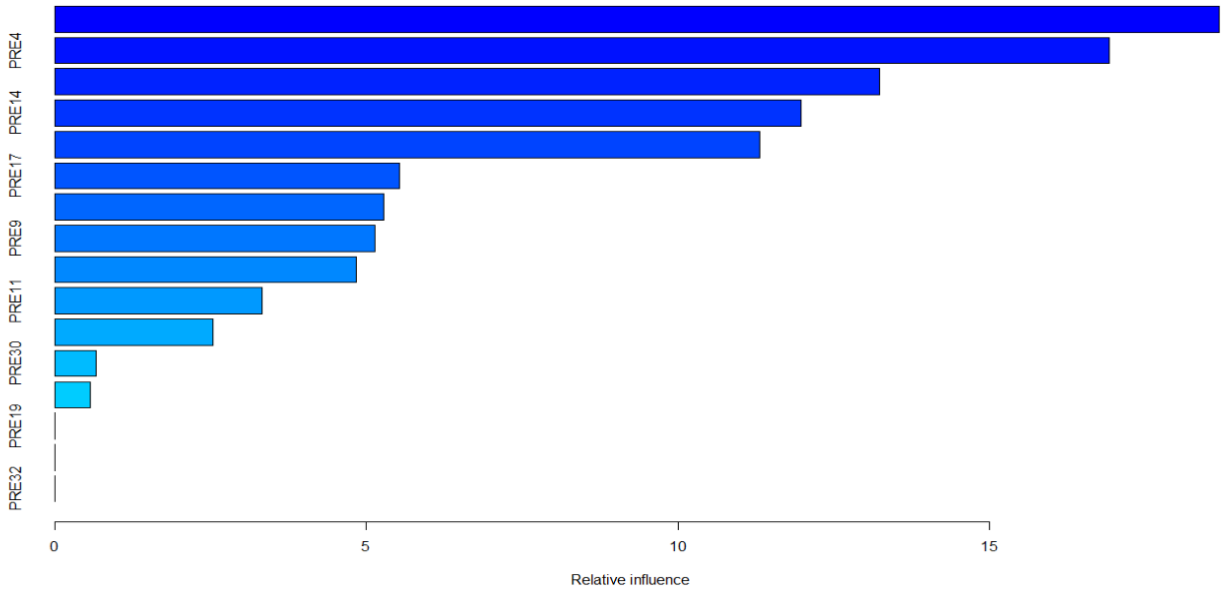

Figure 1: Relative Influence of Predictors in a boosted random forest with shrinkage 0.1

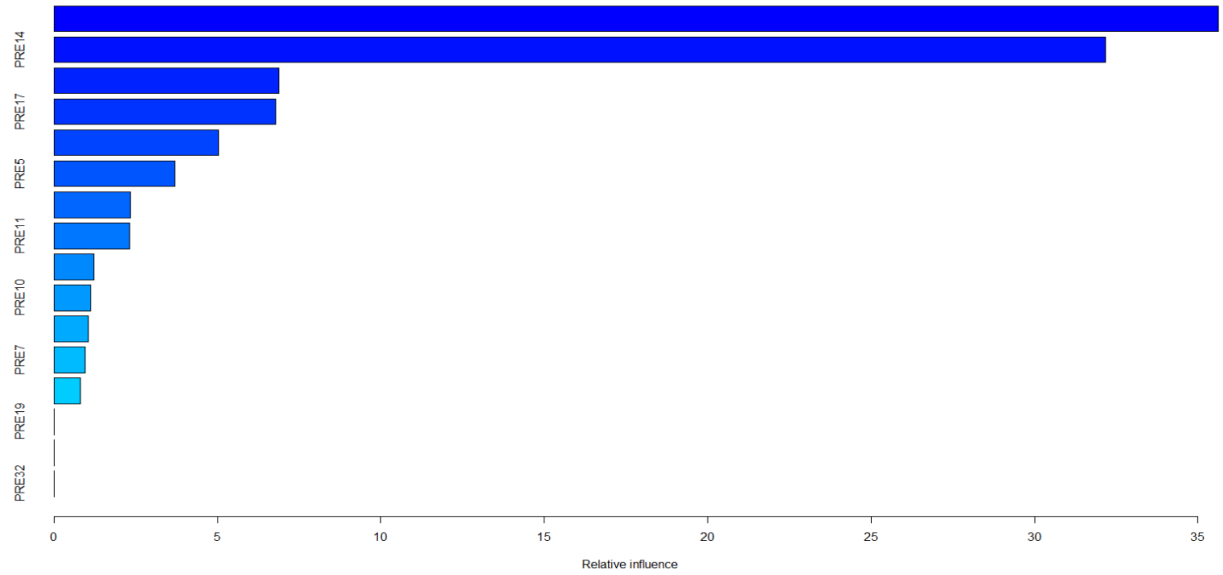

Figure 2: Relative Influence of predictors in a boosted random forest model with shrinkage 0.001

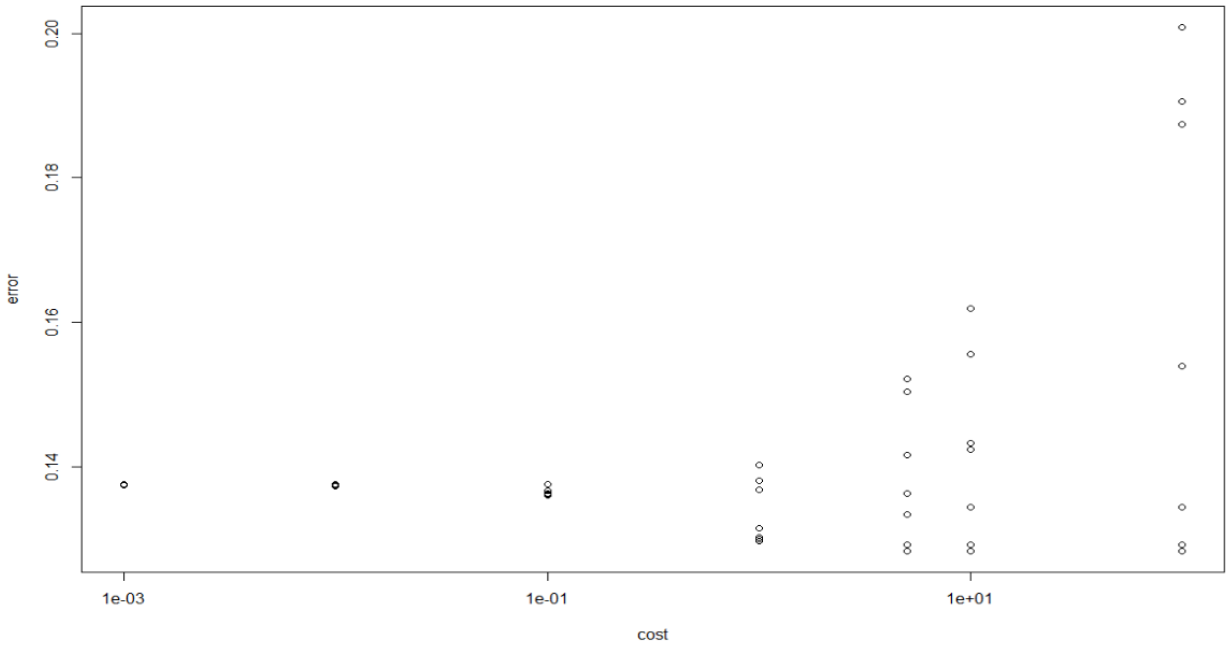

Figure 3: Error vs. cost performance for a SVM with a radial kernel, for cost selection

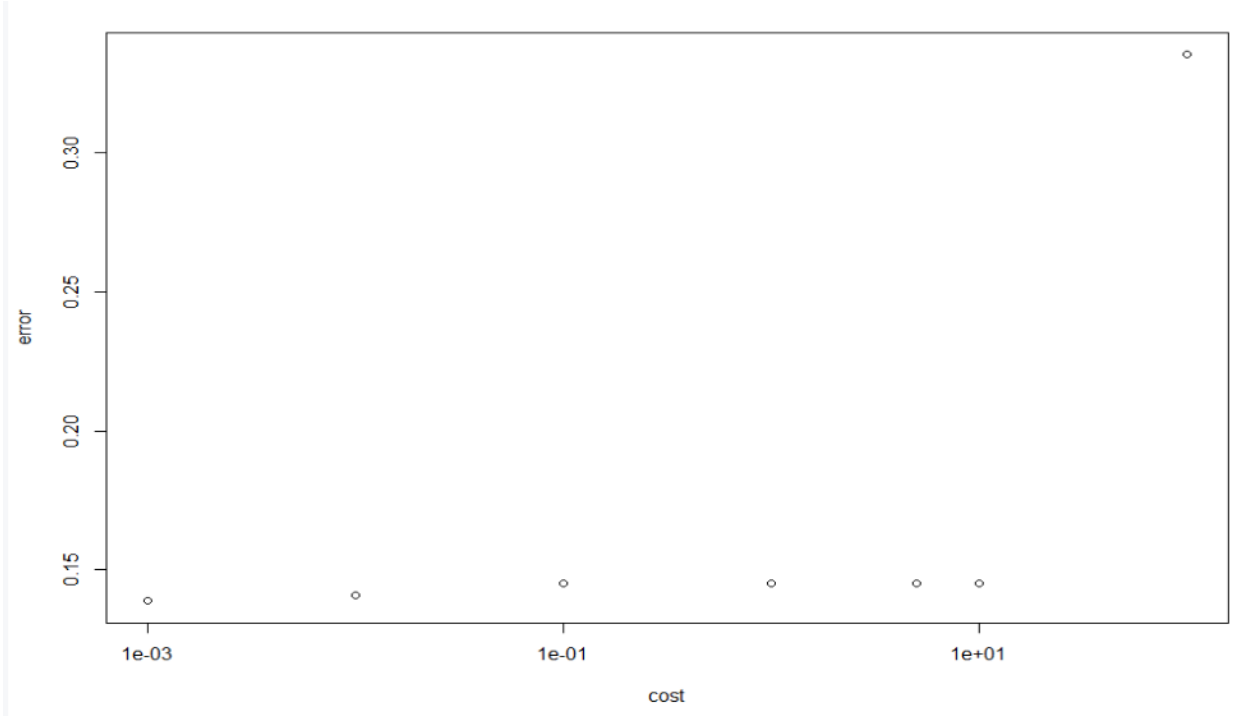

Figure 4: Error vs. cost analysis of a SVM with a linear kernel

The support vector model approaches performed fairly well, providing error rates of 0.139 for a linear kernel and 0.128 for a radial kernel. The linear kernel used a cost of 0.001 and the radial kernel used a gamma of 100 and a cost of 5 after optimization through cross-validation. One issue when using data of

this type is bias in the data itself. Following clinical data reported in Watanabe 2004, most of these results showed survival. This means that part of the reason the support vector machine performed well could be because of the bias towards survival. a machine that predicted all survival would also perform well. One solution to this is to try a boosted SVM machine, as was tried in the study conducted by Maciej 2014. The polynomial kernel simply took too long to optimize, and further studies should include a polynomial SVM evaluation with higher computing power, or with fewer features. Part of the problem with one hot encoding was this expansion of the feature set. An SVM analysis using the diagnoses as factors instead of transformed values may provide faster, more accurate results. Part of the problem was that the feature set was already fairly expansive, and extending it from 17 to 24 features was computationally expensive. The cost of 5 that achieved the highest performance in the radial kernel meant that the hyperplane's margin was small, so there were less support vectors involved in the hyperplane's calculation. The gamma for this machine, however, is considered large at 100, so this model may be prone to overfitting and high variance when it comes to more test cases. This large gamma also means that each support vector had less of an influence on the hyperplane's shape.

References

de Groot PM, Wu CC, Carter BW, Munden RF. The epidemiology of lung cancer. Transl Lung Cancer Res. 2018;7(3):220–233. doi:10.21037/tlcr.2018.05.06

Watanabe, Shun-ichi et al. Recent results of postoperative mortality for surgical resections in lung cancer. The Annals of Thoracic Surgery 2004;78(3):999-1002.

Maciej Z, Tomczak JM, Lubicz M, Witek J (2014) Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. In: Applied soft computing, vol 14, Elsevier, pp 99-108.