Machine Learning-Assisted Prediction of Surgical Mortality of Lung Cancer Patients

Sidra Xu
21SidraX@students.harker.org
The Harker School

Abstract

Operative mortality rates are currently not predictable; there are no tools to help health professionals determine prognoses and assist patients in making informed decisions. To address this issue, a machine learning algorithm was developed in this study to predict operative mortality for lung cancer patients, which provides accurate information as to whether or not a lung-cancer patient should undergo surgery based on this mortality prediction. We implemented a comparison of six different computational models: Naïve Bayes, Support Vector Machine, Random Forest, Adaptive Boost, Extreme Gradient Boost, and Artificial Neural Network. Of the six algorithms tested, XGBoost, a powerful yet relatively new algorithm that is just beginning to make its presence in the field of medical research, produced the best results, giving a testing-accuracy over 97% when aided with data balancing and feature engineering, a 10% increase over previous research. The strength of this project is that we demonstrate an accurate big data-driven, machine learning approach to predictive analytics in perioperative care, which can be used to improve surgical risk prediction for patients, healthcare professionals, and hospitals. Although our current example utilizes only data on thoracic surgery for lung cancer, this concept of machine learning-assisted decision-making can be expanded to incorporate more patients, more diseases, and more institutions in the future.

1 Introduction

With 250 million surgeries being performed every day around the world [1], accurate prediction of surgical risk is important for clinical decision-making and for guiding the allocation of health care resources. However, operative mortality rates are currently not predictable; there is no data available in a systematic format to help health professionals predict prognoses and help patients make informed decisions. Postoperative complications are the main reason for increased costs in surgery, and patients who develop complications use a disproportionately larger share of available resources for a hospital. According to the US National Library of Medicine, between 1992 and 2003, the average cost of operative death on lung cancer patient was $38,088 [2]. Moreover, a study conducted by the British Medical Journal found that doctors were only able to predict life expectancy accurately 19.7% of the time [3]. As a result, many patients are uncertain if surgery is a viable option for them. To address this issue, we aim to develop machine learning algorithms that accurately predict operative mortality of patients, with lung cancer patients as an example, to use in an intelligent system that allows for data analysis and optimization. The application of such systems is expected to significantly improve decision-making procedures for patients, hospitals, and healthcare professionals in the near future.

2 Methods

2.1 Data

Publicly available data from the UC Irvine Machine Learning Repository was used to train and test the algorithms. The database is part of the National Lung Cancer Registry and contains data compiled at Poland's Wroclaw Thoracic Surgery Centre [4]. It is collected from patients that underwent major lung resection surgery for primary lung cancer between 2007 and 2011. The data is represented as follows: each of the 470 rows represents one patient and each of the 17 columns represents one feature. One of the 17 features is a true/false label, indicating whether the given patient lived or died within a year after surgery. The 17 features include both class data, such as the presence of pain or cough before surgery, and continuous data, such as a patient's forced vital capacity, size of the original tumor, and age.

2.2 Data Preprocessing

The main challenge of this study is the highly imbalanced data: only 70 out of 470 instances, i.e. 14.9% of the data, are associated with the positive label (death within one year). This problem is resolved with a bootstrapping algorithm, SMOTE (Synthetic Minority Over-sampling Technique). Other techniques are also experimented with, including ROS (Random Over Sampler). Another challenge lies in the number of features available for in-depth analysis of the data. It was shown in a previous study that there was much difficulty in drawing correlations between the various features. This issue is addressed using feature engineering. A number of new features are created in this study to better represent the underlying relationships among various features of the dataset and thus improved model performance [5].

2.2.1 Data Balancing

Data balancing algorithms, SMOTE and ROS, are implemented to compare with the base case without data balancing. SMOTE generates new samples through interpolation of nearby samples [6]. ROS, on the other hand, generates new samples in the under-represented class by randomly sampling with replacement the current available samples until the classes are balanced [7].

2.2.2 Feature Engineering

To derive important features that could be used for creating new features, a feature ranking model is conducted on the original dataset. Subsequently the top three most relevant features, FVC, FEV, and Age (Figure 1) are used to create nine new features in total, by performing various basic mathematical calculations.

2.3 Machine Learning Algorithms

Six candidate machine learning models, Naïve Bayes, Supporting Vector Machine, Random Forest, Adaptive Boosting, Extreme Gradient Boost, and Artificial Neural Network (a multilayer perceptron with three layers), were trained on the dataset. Model parameter tuning was performed via 10-fold cross-validation with the goal of optimizing accuracy. Although accuracy is used as the major performance metric for model comparison, we also generated a receiver operating characteristic curve (ROC) and precision-recall curve (PRC), and calculated the areas under both curves (AUROC and AUPRC, respectively). These metrics are considered informative for evaluating the discriminatory ability of binary classifiers, and AUPRC is particularly powerful for imbalanced datasets [8].

3 Results and Discussions

Shown in Table 1 is a summary of prediction data obtained in this study. Both training and testing accuracy are listed to provides a sense of how well the models are able to generalize from training data to test data and to what extent it is overfitting to the training data. Among six different predictive models examined, XGBoost has provided the highest accuracy, AUROC, and AUPRC, with an improvement of nearly 8% in accuracy, 5% in AUROC, and 5% in AUPRC than the next best algorithm, namely Adaptive Boosting. Both boosting models have excellent predictive power, with

both AUROC and AUPRC scores above 90%, and they are more robust than Random Forest and Supporting Vector Machine because their training and test accuracy closely match each other. Results with Random Forest and Supporting Vector Machine, though achieving around 80% accuracy, suffer from over-fitting which would be problematic when employing the algorithm on patients outside of the training dataset. The performance of Naïve Bayes is only slightly better than random guess. And the three-layer simple neural network has given an accuracy above 84%, but an AUROC of 58% shows the relatively poor skill of the model in distinguishing between the two classes in this project.

Table 1: Training and testing accuracy of algorithms with data balancing and feature engineering

Model	Training Accuracy	Testing Accuracy	AUROC	AUPRC
Naive Bayes	0.5500	0.5059	0.7220	0.7140
SVM	0.9984	0.8436	0.9690	0.8847
Random Forest	0.9047	0.8189	0.8704	0.8808
Adaptive Boosting	0.9000	0.8943	0.9469	0.9439
XGBoost	0.9990	0.9731	0.9989	0.9985
Artificial Neural Networks	0.8590	0.8404	0.5828	-

The success of gradient boosting algorithms demonstrated in this study contrasts with the previous research, where Supporting Vector Machine was used and an accuracy of only 87% was obtained [9]. The previous researchers also noted that their models struggled with over-fitting, with testing accuracy at 87% but training accuracy at 99% [9]. This is similar to what has been found in this work with Supporting Vector Machine as shown in Table 1. This difference in performance between the two types of algorithms suggests that gradient boosting models might be better choices of predictive models for datasets similar to what is used in this study. These models implement gradient boosted decision trees, where new models are created to predict the residuals or errors of prior models and subsequently added to prior models to produce the final prediction [10]. Although proven successful across many domains, they are still very new to the field of medical research in comparison to other algorithms. Another finding of this study is that the artificial neural network does not seem to be a desirable model choice. This is likely due to the relatively small size of the data in this project as artificial neural networks generally perform well with large data size. Now that more and more data is becoming available, the power of artificial neural networks should be revisited considering their huge success in other fields such as computer vision and natural language processing.

In addition to the type of algorithms, data balancing and feature engineering have also played important roles in improving model performance, as demonstrated in Table 2 using XGBoost, as an example. The testing accuracy of XGBoost is improved by more than 10% after data balancing and by another 2% after feature engineering, reaching a final testing accuracy of 97%. For data balancing, ROS, in comparison with SMOTE, has produced better results across all five algorithms tested.

Table 2: Testing accuracy of XGBoost under various pre-processing procedures

Pre-processing Procedure	Testing Accuracy
Base Case	0.8553
Data Balancing	0.9588
Data Balancing and Feature Engineering	0.9731

As stated earlier, the aim of this study is to develop an intelligent decision-support system for patients and healthcare professionals. It is therefore important that the output of such a system is interpretable to them before they make the final decision. Toward this goal, we have performed further analysis of the model that performs best and are able to identify both critical features that are used by the algorithm in making its classification and negligible features that do not make any significant impact. Shown in Figure 1 and 2 are the ranking of features according to their importance before and after feature engineering, respectively, which reveals that many of the newly engineered features are of significant relevance to the prediction results. Such information is not only instrumental to the decision makers, but also useful in providing better directions to data collection and analysis and thus help with health care cost savings.

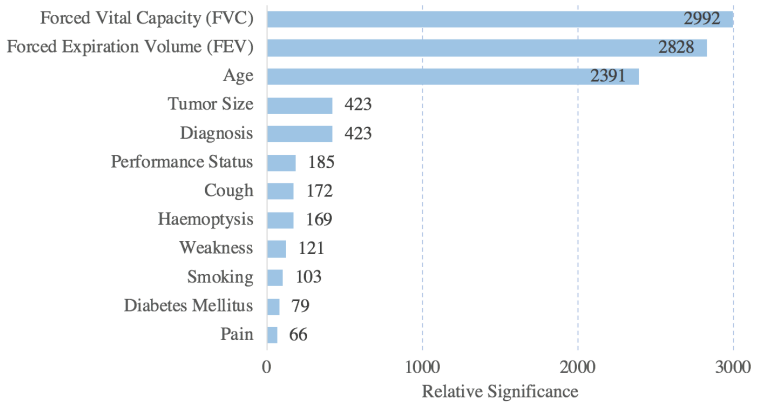

Figure 1: Feature importance ranking (top 12) - base case

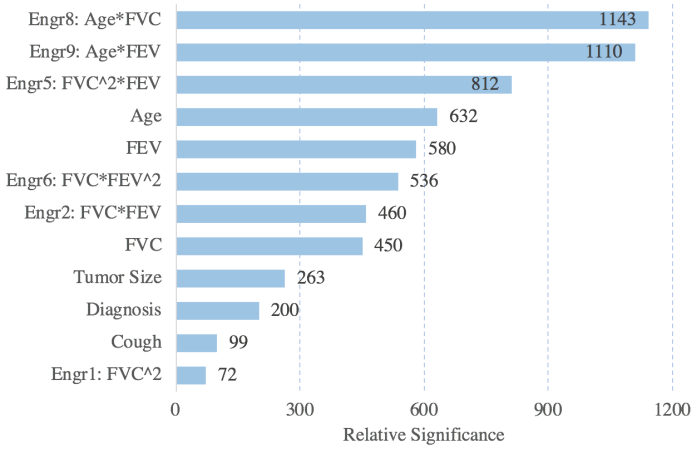

Figure 2: Feature importance ranking (top 12) - with feature engineering

In summary, a robust and accurate machine learning model based on XGBoost for prediction of surgical mortality of lung cancer patients is developed in this study. Combined with data balancing and feature engineering, this model is capable of providing a prediction accuracy up to 97%, an improvement of 10% compared with previous studies. While the model is developed on a single set of data, it can be updated either real-time or periodically as new data is acquired. And although we focus on only one disease at the moment, the same methodology can be applied to incorporate more diseases and more hospitals in the future to benefit more people. In light of the lack of tools nationwide for doctors and patients to make informed decisions about surgery, this research is an important step to fill the gap.

On the other hand, the output of a machine-learning algorithm, such as what we propose here, should be interpreted with extreme caution, especially when the life a patient is on the line. The prediction result should only be used in a supportive role in assisting doctors and patients when a surgical decision is made. To make this research and others alike practically applicable, it is thus proposed that further research should look into the possibility of re-framing surgical mortality problems into multi-class or multi-label classification, or even regression (e.g. the number of years that a patient survives after a surgery) projects, where more specific and relevant information can be provided, instead of just binary classification. This will be possible as more and more data is becoming available. Additionally, depending on the algorithm used, it would be more informative to provide doctors and patients with detailed information on how the algorithm makes its decision (e.g. the decision tree used in XGBoost) rather than providing a black-box output.

References

[1] Weiser TG, Haynes AB, Molina G, et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. Lancet. 2015;385(suppl 2):S11.

[2] Cipriano, L. (2011) Lung cancer treatment costs, including patient responsibility, by stage of disease and treatment modality, 1992–2003. *Value Health* **14**(1):41-52.

[3] Christakis, N. (2000) Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *British Medical Journal* **320**(7233):469-473.

[4] UCI Machine Learning Repository - Thoracic Surgery Data Data Set, https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data.

[5] Zheng, A., Casari, A. (2018) Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. 1st edn. O'Reilly, Sebastopol.

[6] Chawla, N. (2002) SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**:321-357.

[7] Lemaitre, G. (2016) Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **7**:1-5.

[8] Saito T, Rehmsmeier M. (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 10:e0118432.

[9] Abdulhamid, A. (2014). Life Expectancy Post Thoracic Surgery. Stanford University: CS 229.

[10] Chen, T. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785-794.