DATA 245 MACHINE LEARNING

FALL 2020

Prediction of Post Surgery Life Expectancy in Lung Cancer Patients

Submitted To: Prof. Shih Yu Chang

Submission Date: 12/13/2020

Submitted By:

Group 2

Samiksha Pandey
HimaBindu Nayani
Varunika Raja
Venkata Sai Kusuma Sindhoora Vankayala Siva

Demo Link

https://www.youtube.com/watch?v=5lHAP6d7nQ0&feature=youtu.be

Objective

The amount of data generated at the touch of a finger tip these days in the world is astronomical. However, the vast majority of this data is often rendered useless and is in fact more expensive to get rid of then it is to keep. That is where Data Scientists and Analysts alike come into the picture as they hold the ability to sift through all this data and find the parts that they can extract meaning from in order to gain insights which in return lead to profit for all parties involved.

Data Scientists employ the tools of Machine Learning algorithms onto data and generate models that can be fine tuned to provide highly accurate predictions and insights. This work is fastly being adopted in a myriad of different fields such as Business, Engineering, Medicine, and even Commercial applications.

In this project, we hoped to combine our knowledge gained so far on Machine Learning and how to build different models to address a problem that currently exists and requires the help of data scientists like us in order to be resolved. The field we chose is medicine, particularly datasets containing patient information regarding their behavior after going through intensive Thoracic Surgery. More specifically, the datasets show each patient's relevant metric levels as descriptive features and the target feature is the status of their survival 1 year post-op.

Unfortunately, the commonly observed pattern is that even if the medical history of patients are carefully recorded and their health is closely monitored, their survival rates long after receiving the surgery is not always good. This need not be the result of poorly performed surgery, but rather that there is an underlying pattern in their metrics resulting in death when in the right combination for each patient. This is not detected prior to surgery as it initially implies no risk to receive the go ahead on the patient receiving the surgery. Only cases detected in the early stages survive. Thoracic Cancer Surgery increases life expectancy. But, Post-surgery deaths are high. So, still continuously monitoring with tests, the cancer cells can be impossible to detect but still reside in the body. Or maybe they'll go somewhere and come back which needs to be detected early is very necessary to increase the survival rate. However invasive thoracic surgery

can be extremely taxing and could trigger the pattern even more leading to fatal outcomes months or even years later[1].

Our goal is to build a model that can accurately classify these patients into their likely survival ability in order to improve medical practices and urge doctors to reconsider their procedures or think of new innovative ways to reduce negative outcomes. This will not only improve best practices for surgery, but also it will surely improve the quality of life.

Approach

As with all good machine learning projects, best practice for setting up the entire procedure follows the cross industry standard procedure for data mining (CRISP-DM) method as shown in Figure 1. The steps of this method include: business understanding, data understanding, data preparation, modeling, evaluation, deployment [2].

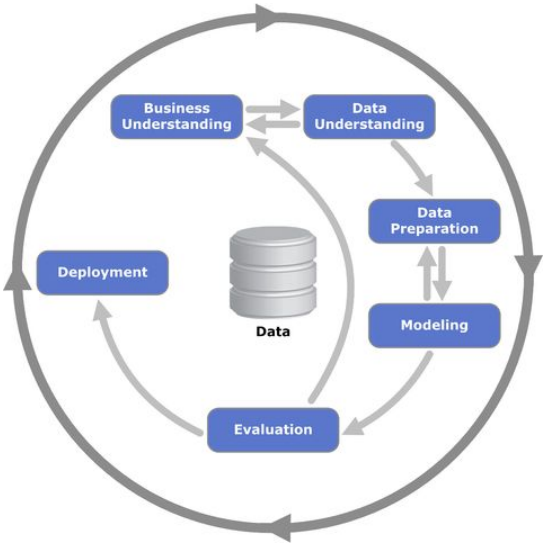

Figure 1. CRISP-DM Method Diagram

First, we gain knowledge and insight on our domain, which in our case is Thoracic Surgery. This includes an overview of the procedure and the requirements a patient must meet in order to receive it. This also entails understanding the meaning of the commonly associated terms and metric levels so that we can translate their impact when used in regards to the data.

The next step is understanding the data. The dataset we obtained for this project is from the UCI Repository. The Thoracic Surgery Survival Rates dataset contains 470 instances for a total of 18 attributes. This will be incorporated into our model as the training data set.

Data preparation is the next step in the CRISP-DM method. The main factors to check for in preparation is to ensure there are no missing values in our data and that there is no noise or outliers that could affect the accuracy and results. Once the cleaning of this dataset is complete, it is now ready to use in our model.

For our model, we wanted to choose one that would best serve the purpose of performing a binary classification. Instead of just training and using one model, we instead decided to use a handful of models that we learned about in this course such as Random Forest, K-Nearest Neighbors, Naive Bayes, Decision Tree, Logistic Regression, and Support Vector Machines. We also employed ensemble techniques in modeling to serve as good comparisons.

To evaluate our models, We generated the accuracy, F1-scores, and AUC scores of each model. The purpose of having multiple models was so that we could compare and decide which one was best suited for this problem type. The resulting scores could evaluate each model against itself based on a set threshold as well as be evaluated in its performance compared to the other models that were utilized.

Data Exploration

The thoracic surgery survival rates dataset is taken from patients who underwent major lung surgeries for primary lung cancer in the years 2007-2011. It was collected retrospectively at the Wroclaw Thoracic Surgery Centre [3]. The dataset contains a total of 470 instances for 18 attributes as shown in Figure 2. There is a good mix of continuous and categorical descriptive features in this dataset. The target feature in this dataset is the survival status of a patient 1 year post-op. The chosen models are going to learn from the training data and then be able to predict if the patients in the test data set are going to survive or not.

Attributes	Type	Description
ID	Continuous	Patient's ID
DGN	Categorical	Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)
PRE4	Continuous	Forced Vital Capacity - FVC (numeric)
PRE5	Continous	Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
PRE6	Categorical	Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)
PRE7	Categorical	Pain before surgery (T,F)
PRE8	Categorical	Haemoptysis before surgery (T,F)
PRE9	Categorical	Dyspnoea before surgery (T,F)
PRE10	Categorical	Cough before surgery (T,F)
PRE11	Categorical	Weakness before surgery (T,F)
PRE14	Categorical	T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)
PRE17	Categorical	Type 2 DM - diabetes mellitus (T,F)
PRE17	Categorical	MI up to 6 months (T,F)
PRE19	Categorical	PAD - Peripheral Arterial Diseases (T,F)
PRE30	Categorical	Smoking (T,F)
PRE32	Categorical	Asthma (T,F)
AGE	Continuous	Age at the time of surgery (numeric)
Risk1Y	Target	1 year survival period - (T)rue value if died (T,F)

Figure 2. Thoracic Surgery Dataset Attributes

Before the data can be used however, it must go through a preparation step. This means that it must first be cleaned and only viable features and instances should be kept. First, we went through and ensured that there were no missing values. Then, we checked for the presence of outliers that could skew the models and alter results from being more accurate. We found 3 fields that contained outliers. As shown in Figure 3 and 4, the Patient's Age field contained outliers on its lower end and the Forced Capacity and Forced Expiration fields contained outliers on the higher end. We also ensured there were no duplicates in the data. After performing this cleaning, we collected descriptive stats and created visuals in order to better show the basic descriptive statistics of the dataset. The code shows this visualization in the forms of box plots and histograms.

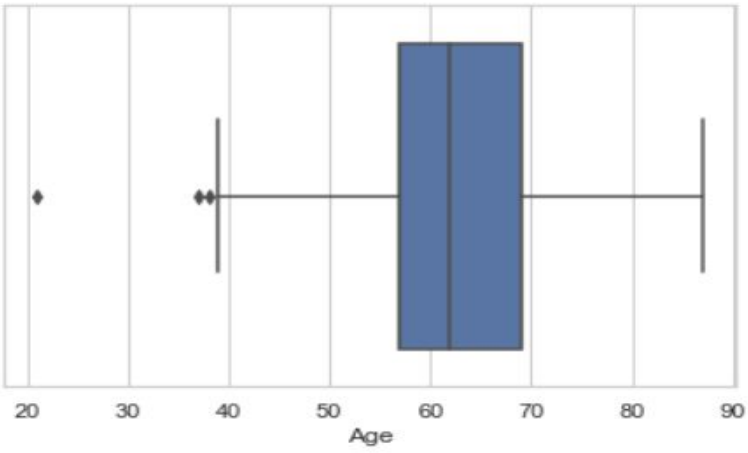

Figure 3. Boxplot for Patient Age with Outliers

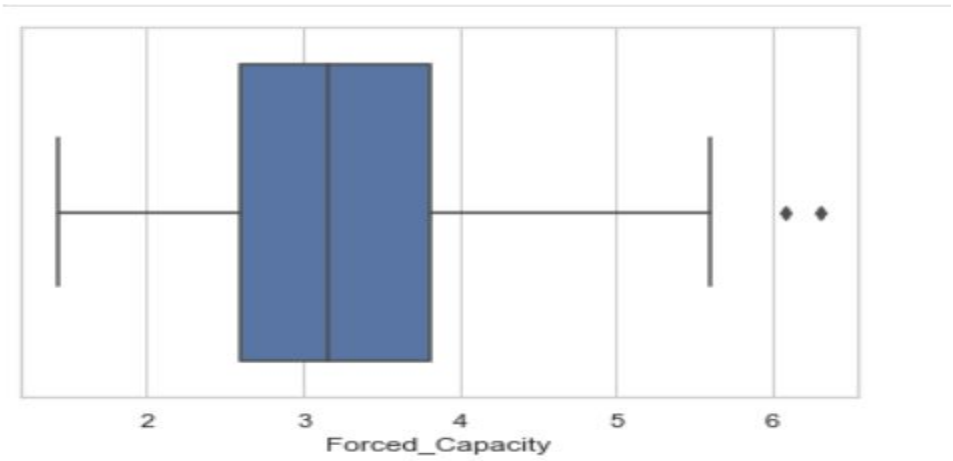

Figure 4. Boxplot for Forced Capacity with Outliers

Since the dataset we obtained did not come with a designated test dataset, we split the dataset into both train and test data and employed the trained models on the test data where we did not know the survival status in order to generate predictions to evaluate. This is an application of Supervised Learning. We also decided to use all the given descriptive feature fields and did not do any extensive feature selection as it was evident that each descriptive feature held equivalent weights in determining the survival status. This assumption was made based on the description given to us by the UCI Machine Learning Repository's explanation of the dataset.

Model Implementation

The six models we chose were: Random Forest, K-Nearest Neighbors, Naive Bayes, Decision Tree, Logistic Regression, and Support Vector Machines. The code for each model is shown consecutively in Python script and was done on the Jupyter Notebooks platform. The results of the predictions of each model are shown in Figure 5.

	Accuracy	Mean Cross validation score	Mean precision score	Mean Recall score	Mean F1 score
KNN	0.85	0.84	0	0	0
SVM	0.85	0.84	0	0	0
Decision Tree	0.79	0.84	0.02	0.014	0.019
Naive Bayes	0.22	0.25	0.16	0.98	0.28
Random Forest	0.85	0.85	0	0	0
Logistic Regression	0.85	0.85	0	0	0

Figure 5. Results- Model Comparison

But our target feature is heavily imbalanced. Above results are without handling this issue and we can see that apart from accuracy all the other scores are close to 0 and models are overfitting to majority class. To handle this issue we applied various sampling techniques on our dataset and evaluated above 6 models with balanced data. Below are the sampling techniques we used:

1. SMOTE: Synthetic Minority Oversampling Technique is one of the commonly used oversampling techniques. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesises new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class[4].
2. Borderline SMOTE: One of the extensions of SMOTE. involves selecting those instances of the minority class that are misclassified, such as with a k-nearest neighbor classification model, and only generating synthetic samples that are difficult to classify.

3. Adaptive Synthetic Sampling(ADASYN): It generates synthetic samples inversely proportional to the density of the examples in the minority class. It is designed to create synthetic examples in regions of the feature space where the density of minority examples is low, and fewer or none where the density is high[5].
4. Borderline SMOTE SVM: It fits an SVM to the dataset and uses the decision boundary as defined by the support vectors as the basis for generating synthetic examples, again based on the idea that the decision boundary is the area where more minority examples are required[5].
5. Random Oversampling: As the name suggests it randomly synthesizes instances of minority class in the dataset.
6. K-means SMOTE: It applies k means clustering before performing oversampling using SMOTE.
7. Near Miss Algorithm: It is an undersampling technique. It aims to balance class distribution by randomly eliminating majority class examples. When instances of two different classes are very close to each other, we remove the instances of the majority class to increase the spaces between the two classes [4].

Model Evaluation

After we balanced our target feature using different oversampling techniques. We got best results using KNN with Random Oversampling. We did stratified k fold evaluation and used grid search for parameter tuning. K and cv value was set to 5. Figure 6 highlights the accuracy score for each model and sampling technique. KNN with Random Oversampling gave best results in terms of accuracy and other scores as well. We got a 0.9 cross validation score and better f1 scores as well.

Oversampling and Undersampling Methods - Accuracy Scores							
ML Models	SMOTE	Adapative Synthetic Sampling (ADASYN)	Borderline SMOTE - SVM	Borderline SMOTE	Kmeans SMOTE	Random Oversampling	Near Miss - Undersampling
K-Nearest	0.6914	0.7021	0.7127	0.6914	0.734	0.734	0.3723
Support Vector Classifier	0.5212	0.5106	0.5957	0.5319	0.5425	0.6702	0.3617
Decision Tree Classifier	0.5638	0.4893	0.7659	0.4893	0.5	0.4787	0.5106
Naïve Bayes	0.2021	0.2021	0.2446	0.2446	0.2021	0.2234	0.5106
Random Forests	0.5425	0.5531	0.7127	0.5531	0.5319	0.5851	0.5212
Logistic Regression	0.5531	0.5212	0.617	0.5744	0.5744	0.5531	0.5319

Figure 6. Oversampling and Undersampling methods-Accuracy Scores

Additionally, we created an ensemble model using our top 3 models from evaluation above. We created a voting classifier using knn, svm and random forest. Weights for each were set to 3,2 and 1 respectively based on their individual performance. However this resulted in lowering the AUC score. The script for the ensemble techniques and their results are still included in the Python code in our Jupyter Notebook with comments.

Impact

This project not only allowed us to apply all the machine learning principles we have learned throughout the entirety of this semester in DATA 245, but also it allowed us to address a problem and try our hand at implementing a solution in order to resolve it.

Thoracic Cancer is one of the more serious and fatal types of Lung Cancer. The survival of this illness lies at only a meager 56%. That is why Thoracic Surgery is such a vital procedure in the field of medicine. When detected in its earlier stages, the effects of this cancer can either be severely reduced or taken away entirely upon the performance of the surgery. However, such a powerful procedure is weakened when patients end up dying regardless a few months or even a year later due to post-op complications. If this negative outcome could be better foreseen so that post-op care could be altered in order to avoid it, this would result in a great increase in survival rates of patients who can now live cancer free.

Our project is paramount in ensuring the betterment of the quality of life of these surgical patients. With our models and their comparisons, highly accurate and effective predictions about whether patients will survive post-surgery could be predicted even beforehand. This could then be taken into consideration before the patient receives the surgery and alter post-op care to be far more efficient, thus leading to better survival rates. Surgeons are already performing magic in the operating in order to combat this deadly disease and save their patients, but we believe that with our solution, we can help make sure that their efforts do not go to waste.

References

1. Alberg AJ, Brock MV, Samet JM (2016). "Chapter 52: Epidemiology of lung cancer". *Murray & Nadel's Textbook of Respiratory Medicine* (6th ed.). Saunders Elsevier. ISBN 978-1-4557-3383-5.
2. Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000); *CRISP-DM 1.0 Step-by-step data mining guides*. https://the-modeling-agency.com/crisp-dm.pdf
3. Marek Lubicz , Konrad Pawelczyk , Adam Rzechonek , Jerzy Kolodziej (2013, November); *Thoracic Surgery Data Data Set*; UCI Machine learning Repository; https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data
4. *GeeksforGeeks.*(2019, June20). Retrieved from https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/
5. Brownlee, J. (2020, January 24). Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/data-sampling-methods-for-imbalanced-classification/