Thoracic Surgery Data Set

From UCI Machine Learning

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 data set presents data of patients, attributes, and whether they survived within a one year time frame. The problem to solve is whether there is a way to determine post-operative life expectancy of lung cancer patients from patient attributes in the data set.

If there is a pattern to be recognized with the attributes and whether the patients do not survive the one year mark, this would help physicians and patients make a more educated decision on whether 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 a decision to follow through on surgery or decide to find alternative treatment methods or palliative care.

Not only would this influence physicians and patients, this information could be utilized by health insurance companies and national health organizations when it comes to making decisions on finances for thoracic surgery involving lung cancer. Also clinical researchers could consolidate any useful findings with other data research findings to search for new research areas.

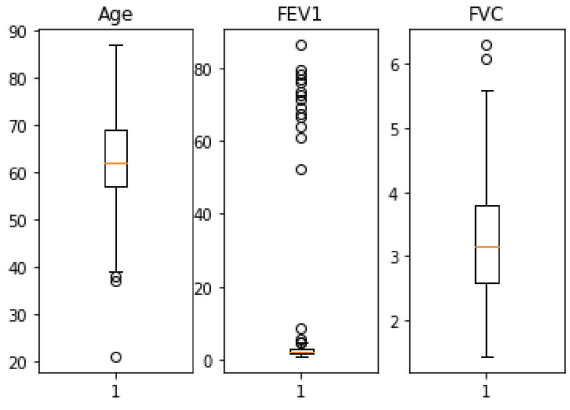
Data Collection and Wrangling

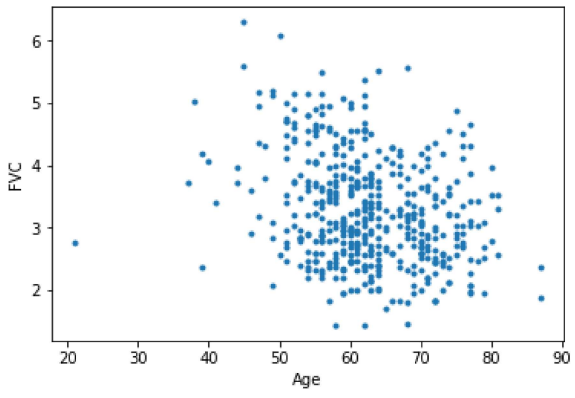
The original data set is from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data. According to the main repository site, the data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007-2011.

The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland. The main data set is in the form of a Weka ARFF file, so for analysis, I converted the file to CSV using a tool found at https://pulipulichen.github.io/jieba-js/weka/arff2csv/ .

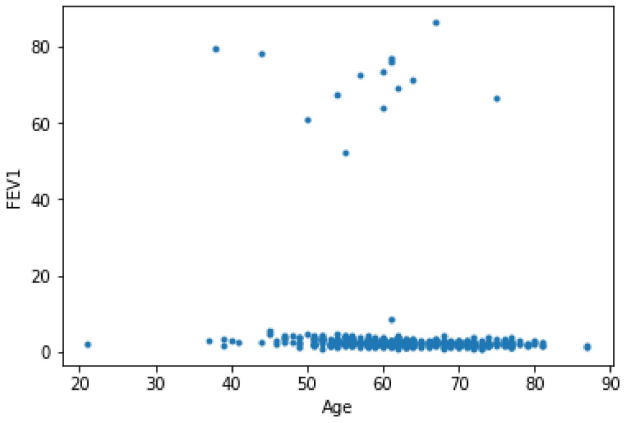
Analyzing the data set's info shows many columns as object strings for T and F values. These include PRE7, PRE8, PRE9, PRE10, PRE11, PRE17, PRE19, PRE25, PRE30, PRE32, and Risk1Yr. So, I converted the T and F object data types to 1 and 0 int data types in these columns. The columns DGN, PRE6, and PRE14 contains data in the form of a string with an int value attached. Reviewing the column data description, I concluded the string value was redundant and it will be more useful for analysis later on just utilizing the int value. So these three columns were adjusted to just have the int value as data type int. The id column was removed because it is not necessary and lacking in any useful description of each patient. The indices will suffice for identification of separate row values. The column names were renamed with more human readable words instead of the original codes.

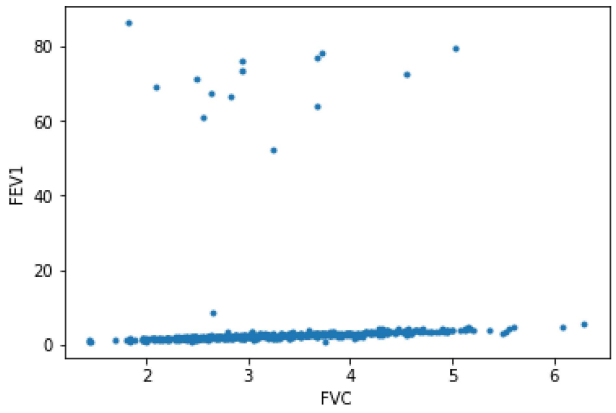
There are no missing values in the original data set to be dealt with. For outliers, the only numeric columns to be considered are PRE4, PRE5, and AGE. Data analysis with box plots and scatter plots reveal 16 noticable outliers.





The box plots reveal many outliers in the FEV1 column and one outlier at around 20 far outside the data range for Age. The two points in the FVC boxplot needs more investigation. With the scatter plots, it is more apparent that the Age outlier is noticeably outside the normal data group. Also the FEV1 outliers are also apparent in their distance from the normal data group.





Analysis of the data reveal, most data for FEV1 is below 8, so the other 15 points were considered outliers and removed from the data set. Majority of the data for Age ranges from 40-80 years old, so the one outlier at 20 was removed. Even with the removal of 16 outliers, the new data set contains 454 instances from the original 470, so the new data set should be sufficient in size for analysis.

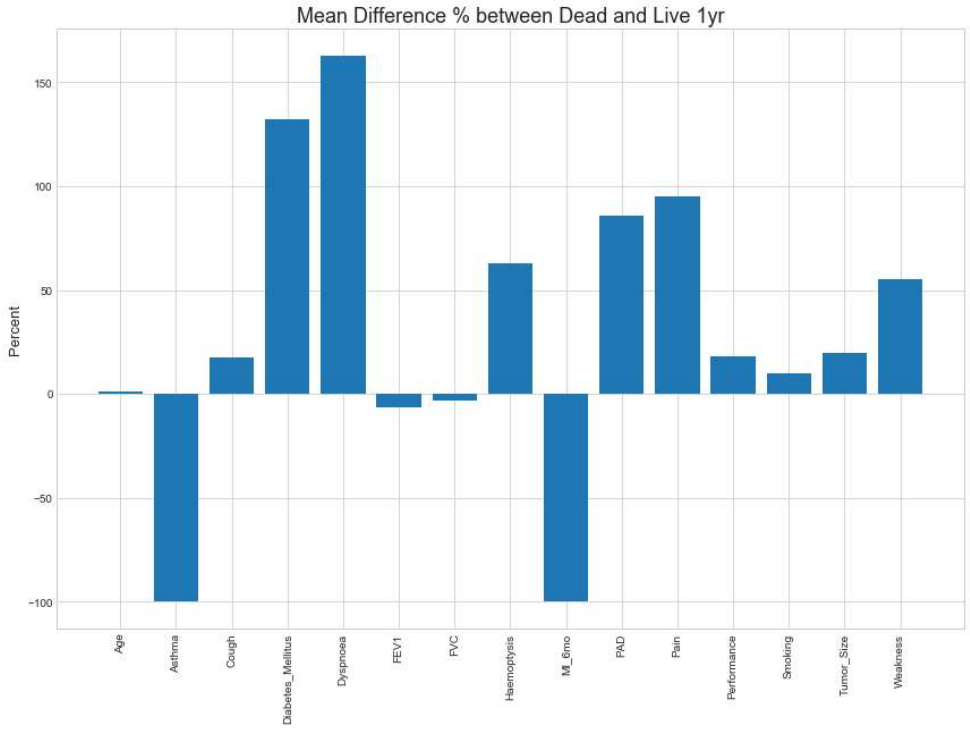
The code for these above procedure can be found at: https://github.com/sychi77/CapstoneProject1/blob/master/data/Data_Wrangling.ipynb .

Exploratory Data Analysis

Out of the 454 patients in the data set, 69 patients died in the 1 year time frame and 385 survived, which is a 15.20% death rate. The table below compares the different attributes and the two different classes of death and live in the 1 year period.

Attribute	Death in 1 year (Mean)	Live 1 year (Mean)
FVC	3.195072	3.304597
FEV1	2.383188	2.540805
Performance	0.913043	0.774026
Pain	0.101449	0.051948
Haemoptysis	0.202899	0.124675
Dyspnoea	0.115942	0.044156
Cough	0.797101	0.677922
Weakness	0.246377	0.158442
Tumor_Size	2.014493	1.683117
Diabetes_Mellitus	0.144928	0.062338
MI_6mo	0.000000	0.005195
PAD	0.028986	0.015584
Smoking	0.898551	0.815584
Asthma	0.000000	0.005195

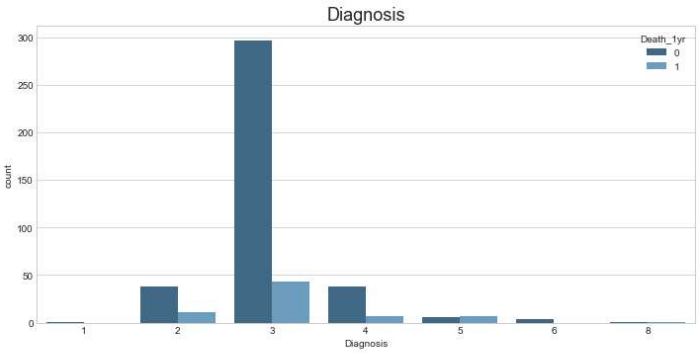
Looking at the means of the two different patient classes, there are features with significant differences and those with minor. However, to better compare the values between classes, a normalization step was performed for % differences.



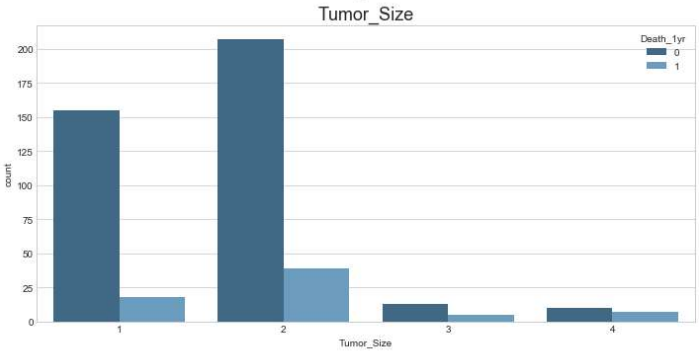
Looking at the graph, one can see easily compare the attributes to determine features of significance.

The most notable attributes for those who died are *Dyspnoea*, *Diabetes Mellitus*, *Pain*, *PAD*, and *Haemoptysis* (in decreasing order), indicating that for those who died, these features were strongly presented. *Asthma* and *MI of 6 months* have negative 100% values, and looking at the numerical values reveals that those who died did not exhibit asthma or MI. Although the mean differences are useful, further investigation of the number of instances of each attributes in combination with the mean differences helped improve our decision on what features to focus on.

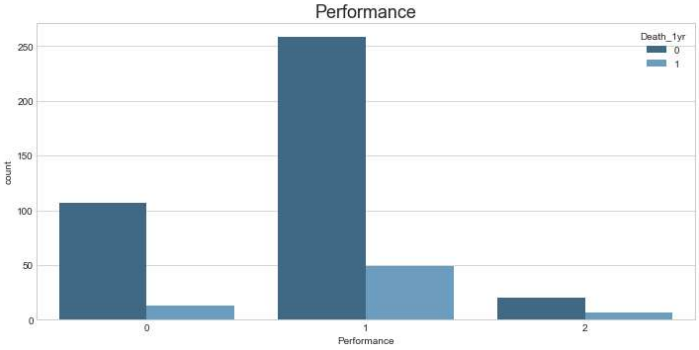
The overall count should be considered when comparing mean differences, because the lower count numbers will have larger fluctuations to small differences. The count of *Cough* and *Smoking* are most noteworthy indicating these conditions are strongly correlated to those patients who are to receive thoracic surgery for lung cancer, but the mean differences are a small positive value indicating more representation in the dead patients.



Referring to the figure on the right for the difference between live and death patients, there are noticeable trends in these categories.



For Diagnosis, the large majority of patients are in category 3. The other categories are relatively small while category 4, 2, and 5 should be considered for their counts in that order. The proportion



of live to dead at a glance seems to be similar for the diagnosis categories except for 5, where the death count is higher than the live count, which indicates this diagnosis is more fatal than the others even with surgery.

For Tumor Size, categories 1 and 2 are the majority. At a glance, the proportion of the dead to live generally increases with the tumor size ranging from 1 to 4, indicating the higher tumor size correlates to higher chance of death even with surgery. Category 4 tumor size is most even in its split between death and live patient data. Also looking at the dead to live mean difference graph, the dead had higher means indicating larger tumor sizes overall.

For Performance, categories are 1, 0, 2 in decreasing order of count. Performance 0 category reveals low death count and good proportion to live data, which makes sense since on the Zubrod scale 0 is good and 2 is poor. Category 1 and 2 display similar proportion to live and dead patients, but with category 1 having a majority of the count. Referring to the dead to live mean difference graph, the dead had higher means indicating the dead on average had poorer performance with a higher Zubrod score than the live.

All the observations above highlighted the trends and patterns in the attributes. However, to ascertain their significance, a hypothesis test will reveal what attributes are of significance and to focus on. The null hypothesis is that the 1 year live and death patients have the same mean, which is tested for each attribute. The test statistic is the mean difference between death and live patients with a significance level of 0.05. The resulting findings are:

- **Cannot Reject Null Hypothesis:** FVC, FEV1, Pain, Haemoptysis, Weakness, MI_6mo, PAD, Smoking, Asthma, Age
- **Reject Null Hypothesis:** Performance, Dyspnoea, Cough, Tumor_Size, Diabetes_Mellitus

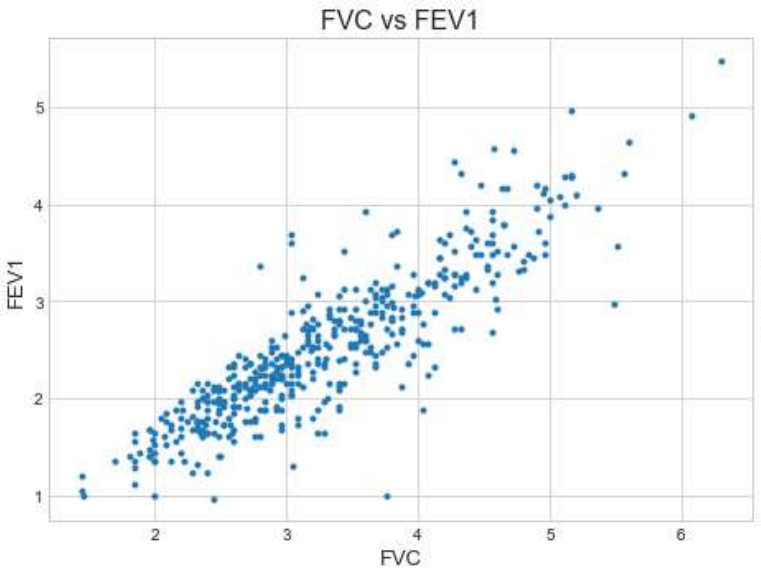
With the results above, the attributes of significance are those that rejected the null hypothesis. To highlight the trends for those that rejected the null hypothesis, the mean difference percentages are listed below.

Mean difference % for death in 1 year patients for attributes of significance:

- Performance = 17.96%
- Dyspnoea = 162.57%
- Cough = 17.58%
- Tumor_Size = 19.69%
- Diabetes_Mellitus = 132.49%

Proceeding forward, one key finding to consider in predictive modeling is the correlation between FVC and FEV1, which are both related to lung capacity.

As you can see, there is a strong positive linear correlation between FVC and FEV1. The calculated Pearson correlation coefficient is 0.89, which is very strong. If these attributes are needed in the machine learning model, combining these two features into one column as a ratio should be considered to reduce the effect of two columns being so highly correlated. The FEV1/FVC ratio, also called Tiffeneau-Pinelli index, will suffice for this feature engineering step.

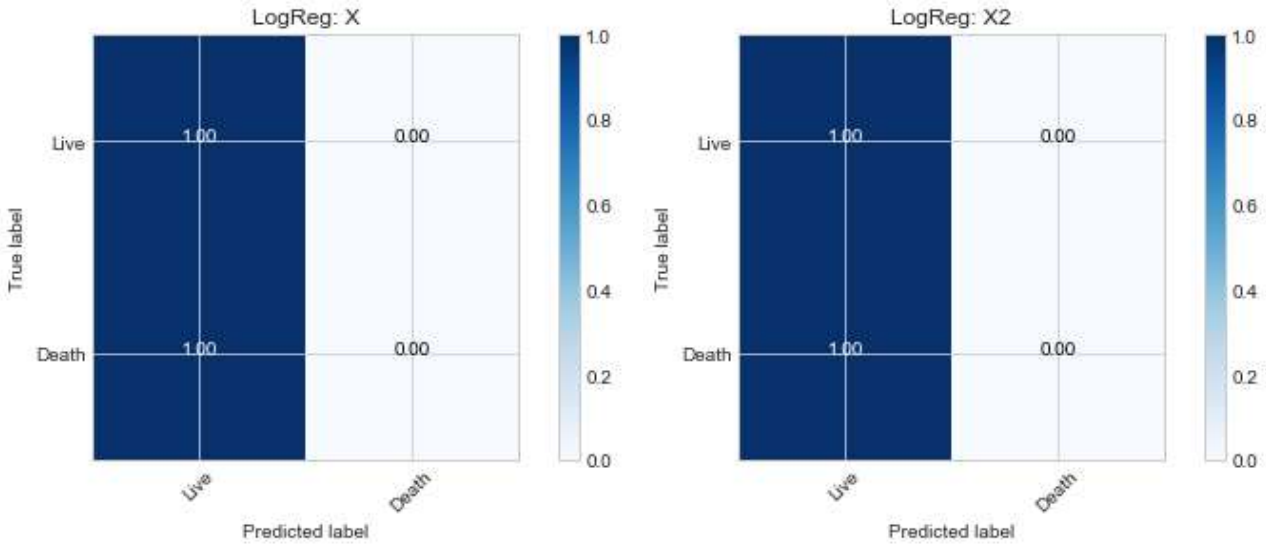


Machine Learning (Supervised Classification)

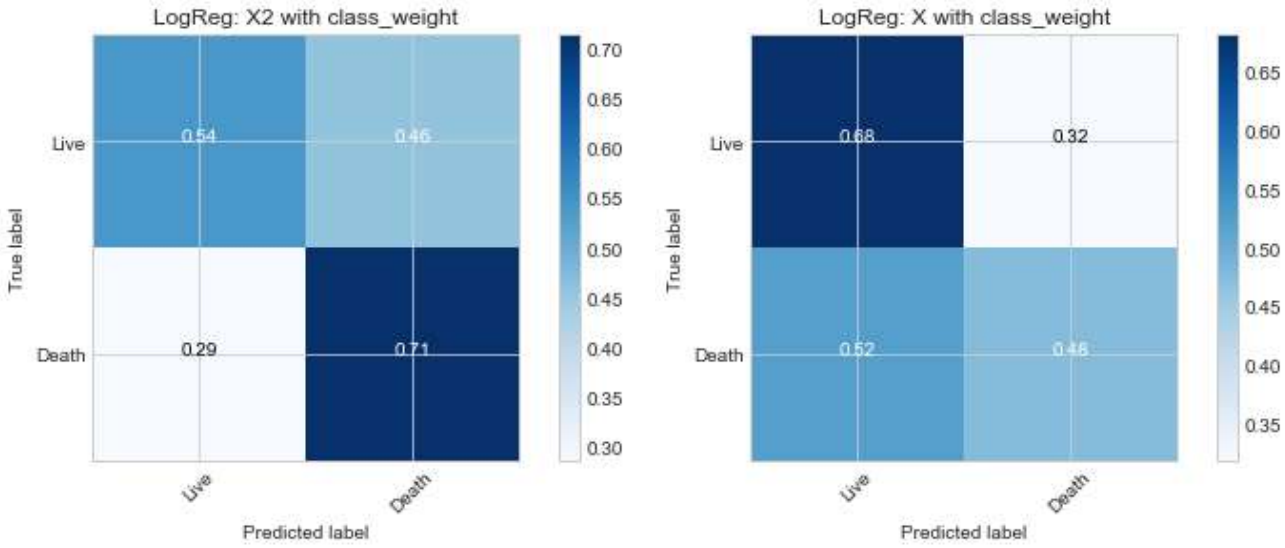
I focused on utilizing Logistic Regression and Random Forest Classifier for this supervised classification problem. From EDA and hypothesis testing to gather p values, I realized which attributes are significant in identifying the mean difference of those who lived and died in the 1 year period after surgery. I wanted focus the test on two different X data sets. The first data set drops the target variable, Death_1yr, and also the two attributes that shows little representation in the data itself, MI_6mo and Asthma. This data is referred to as X. The other data set only includes the attributes of significance concluded from the hypothesis testing in the EDA section: Performance, Dyspnoea, Cough, Tumor_Size, Diabetes_Mellitus. This data set is referred to as X2.

Since the data set is imbalanced and mostly live patients (85%), just predicting all live patients will give a high accuracy score ~85%. So for the model, accuracy will not be a good score method and instead I will look at average precision score, which summarizes the precision-recall curve. Also for the imbalance, there are couple options including downsampling, upsampling, or adjusting class weights to balance the classes. Since downsampling will create a small data set to work with and upsampling may complicate the data further, I will focus on adjusting the class weights.

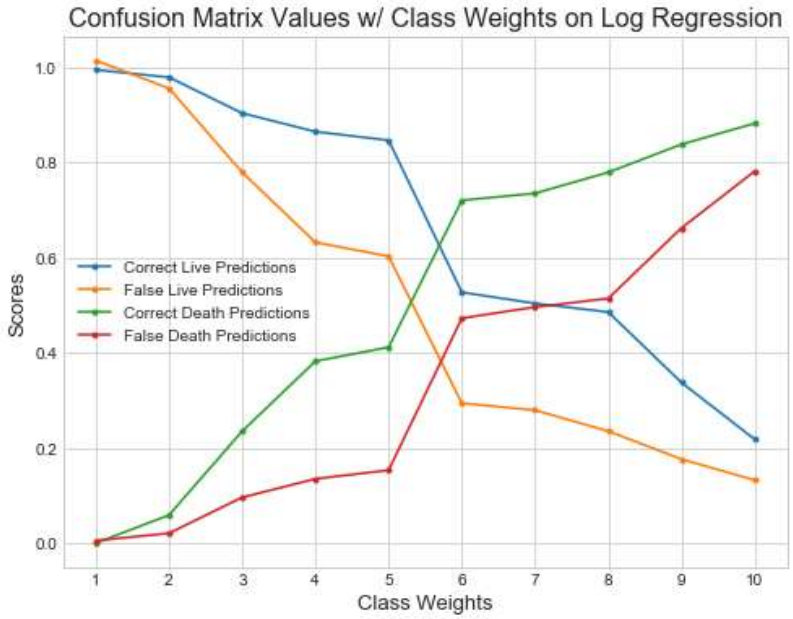
Logistic Regression



Since the data set is imbalanced with only 15% patient death, the results of the model without any class weight to offset this imbalance favors the live column in the confusion matrix. As you can see above, the model predicts mostly all live patients to maximize the accuracy score to 85%, the size of the live patient data, in both the X and X2 data sets.



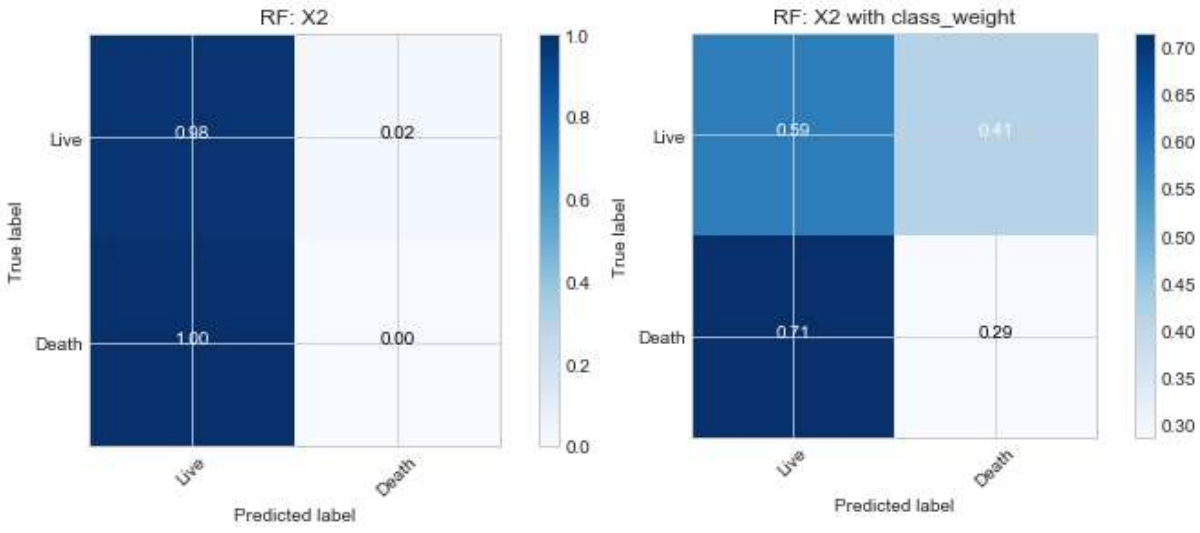
With the class weight parameter, the death prediction rate increases at the cost of live patient prediction, and also the accuracy. In order to see the effectiveness of the model for my purpose, the confusion matrix or classification report can be used to assess the death predictions. Also, the average precision score is a good summary of the precision-recall curve, which is useful in this case.

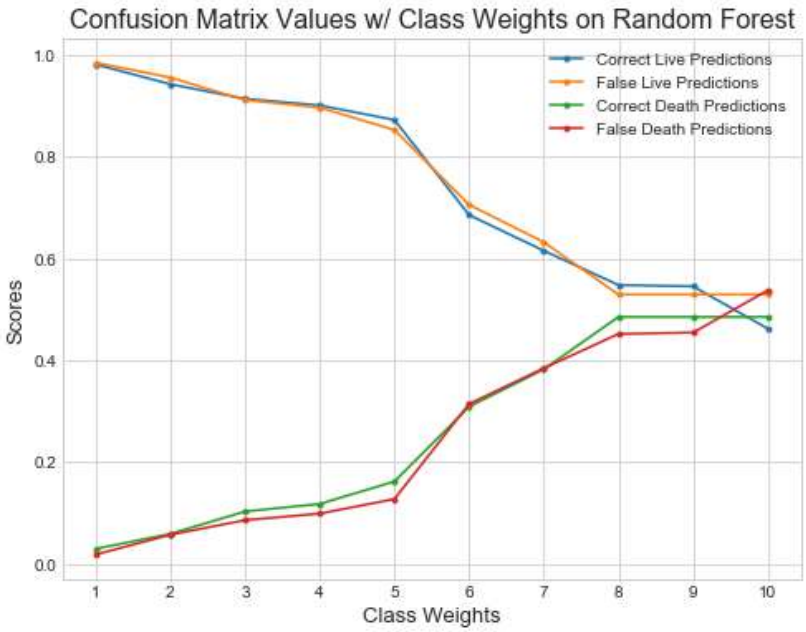


The class weight argument is set to 'balanced' to equalize the death to live ratio, which is 15 to 85. This argument can be altered with any ratio value and the effects can be seen in the graphs above. Although the correct death predictions increased with more class weight on the deaths, the false death predictions increased as well with decrease in correct live predictions. The influence of class weights can be seen in the graph above. Interesting to note that the score dips dramatically around the 5.67 value, which is the equalizing point for the ratio 15 to 85.

Random Forest Classifier

For the random forest classifier, I focused on utilizing the X2 data set to see performance comparative to the log regression model. Similar to the log regression, the random forest classifier heavily favors the live patient prediction without any class weight hyperparameters.





Similar to the Logistic Regression models, the Random Forest predicts deaths better with a class weight parameter to balance the data. The plot reveals the cost of correct live predictions and benefit of correct death predictions with differing class weights. It is interesting to note the different pattern this model takes compared to the log regression graphs above. Based solely on average precision, the log regression produces better results.

However, with hyperparameter tuning, the random forest classifier delivers higher average precision scores compared to what the log regression model did in any class weight value. It is interesting to note that GridSearchCV model notes the best parameter as having no class weight argument with the tested parameters in the report. So, there probably are combinations of hyperparameters that perform better than the models highlighted above with

class weights. For future optimization, more hyperparameter tuning can be done with a more in-depth parameter grid or utilizing RandomSearchCV if needed.

Proceeding Forward

This section displayed the initial steps for utilizing Logistic Regression and Random Forest Classifiers to this data set. Proceeding forward, there are several options to improve the models of this report and suggestions to how to take the next steps.

First, more data will improve the scope of the models. From analysis of this data set, it is clear that there is a significant overlap of attributes, so more patient data or perhaps creating a new data set with additional attributes could help better distinguish the differences and improve the model. If not new data recordings, there are probably similar data sets that have models that predict lung cancer deaths that could be of use in optimizing this model by using it in combination.

Another option is to optimize the models above with hyperparameter tuning. However, since the accuracy score is unreliable in determining positive death predictions, you would have to determine what score to maximize and minimize before proceeding forward in hyperparameter tuning. This scoring method could be the one used in this report, average precision score, or could be a custom scoring method created using the confusion matrix or the values in the classification report.

Finally, depending on the desired outcome considering false prediction costs, the models can be used in an ensemble method to maximize the outcome desired. The ensemble could be anything ranging from boosting methods or utilizing multiple different models. Again, the desired outcome will depend on the hospital or client and how they view the detriment of giving false positives and false negatives compared to the true predictions for live or death outcomes for patients; in other words, how they want to score the efficiency of the model.