Thoracic Surgery for Lung Cancer Patients

Predicting 1 Year Patient Outcomes

Robert Lewis

PROBLEM STATEMENT:

Advances in technology have made the early detection of lung cancer a reality for thousands of individuals at risk. If caught at its earliest stages, this deadly disease can frequently be cured or sent into remission. Alone, or in combination with other cancer therapies, surgery offers the best chance of stopping the cancer from growing and metastasizing[1]. Even so, lung cancer continues to claim more lives in the United States than any other single cancer[2].

Thoracic surgery for the treatment of lung cancer can be further broken down into several common procedures that vary based on the type and severity of the cancer. They include: lobectomy, wedge resections, segmentectomy, pneumonectomy, and metastasectomy[1]. While the 5-year survival rate for lobectomy patients is fairly high, at roughly 70%, other procedures do not fare as well, and all come with a relatively high 30-day mortality rate post-surgery[3].

While a number of patient metrics are evaluated, the current 'gold standard' are the patient's Forced Expiratory Volume in 1 second (FEV1) and Forced Vital Capacity (FVC). As useful as these metrics are, there is at-present no definitive threshold which can say with certainty whether a patient will have a positive or negative outcome from the surgical procedure. The goal of this project is to see if we can better predict a lung cancer patient’s post-surgery survival rate, given their various pre-operative health conditions.

DATA WRANGLING:

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 to 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.

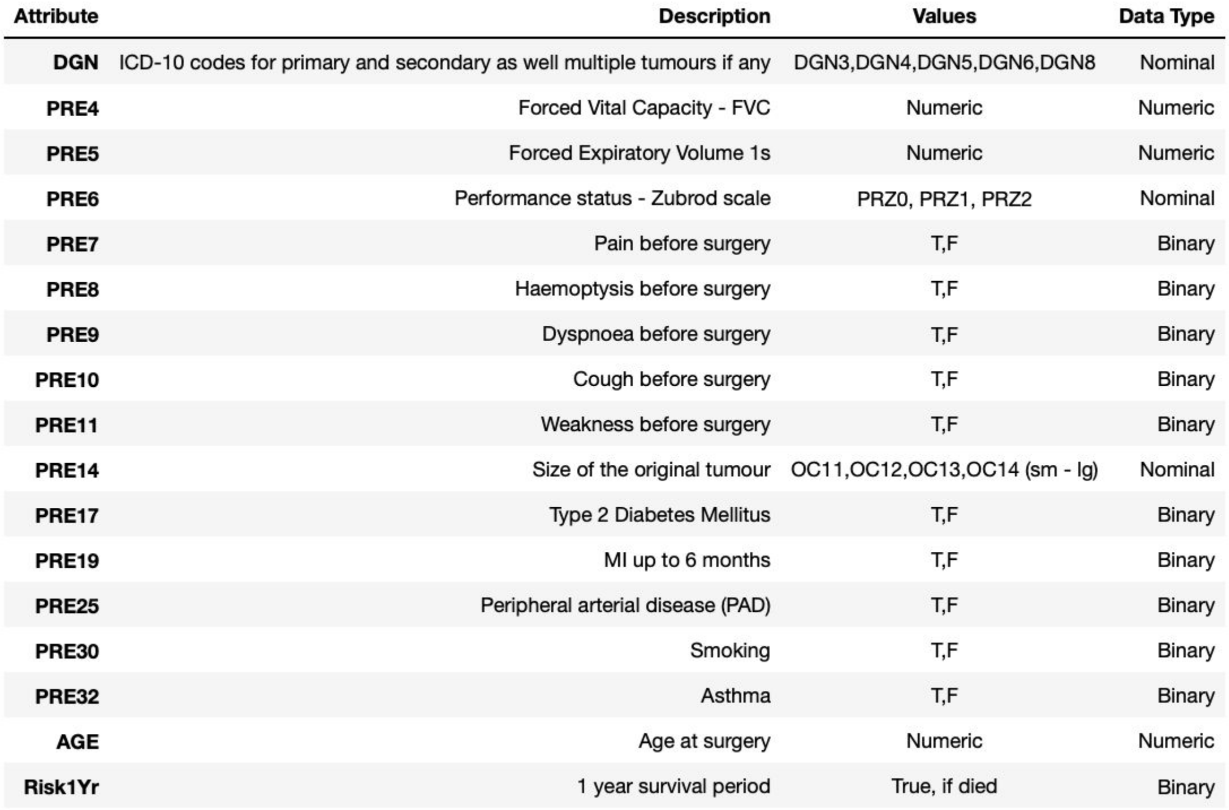
The actual raw data file was obtained through the UCI Machine Learning Repository (Thoracic_Data_Set)

This data set contained 470 observations of patients who underwent thoracic surgery for lung cancer. Each observation contained 16 features that related to an individual's preoperative health condition and an additional column that related to our target of whether that person had survived 1 year post-surgery or not. The target variable was encoded as a binomial with a 0 representing alive and a 1 if they were deceased.

Overall, the data set was relatively clean, as it contained no missing values, however, additional cleaning was necessary to get it to a state of use for modeling. The file was imported as an Attribute-Relation File Format (.ARFF) in which strings are encoded as byte objects. To make it easier to read, a function was applied that decodes these objects and converts them to uint8 data types.

Because 11 of the feature columns are of a binary T/F type, which is less desirable than numeric data for analysis and modeling, these values were replaced with either a T = 1 or F = 0. The final two steps of cleaning involved using an Ordinal Encoder on the nominal 'PRE6' and 'PRE14' features and to verify whether there were any missing or null values.

EXPLORATORY DATA ANALYSIS:



Attribute	Description	Values	Data Type
DGN	ICD-10 codes for primary and secondary as well multiple tumours if any	DGN3,DGN4,DGN5,DGN6,DGN8	Nominal
PRE4	Forced Vital Capacity - FVC	Numeric	Numeric
PRE5	Forced Expiratory Volume 1s	Numeric	Numeric
PRE6	Performance status - Zubrod scale	PRZ0, PRZ1, PRZ2	Nominal
PRE7	Pain before surgery	T,F	Binary
PRE8	Haemoptysis before surgery	T,F	Binary
PRE9	Dyspnoea before surgery	T,F	Binary
PRE10	Cough before surgery	T,F	Binary
PRE11	Weakness before surgery	T,F	Binary
PRE14	Size of the original tumour	OC11,OC12,OC13,OC14 (sm - lg)	Nominal
PRE17	Type 2 Diabetes Mellitus	T,F	Binary
PRE19	MI up to 6 months	T,F	Binary
PRE25	Peripheral arterial disease (PAD)	T,F	Binary
PRE30	Smoking	T,F	Binary
PRE32	Asthma	T,F	Binary
AGE	Age at surgery	Numeric	Numeric
Risk1Yr	1 year survival period	True, if died	Binary

Table 1: Data Set Description

Given that our problem is a classification by nature and that the majority of our features are binary, the bulk of the EDA will revolve around distributions, outliers and correlations observed.

The first step to understanding our data was to check how the target classes are distributed. Of the 470 patients in the raw data, 400 survived and 70 perished -- giving us an imbalance in the data of approximately 85% majority class to 15% minority class. This will have an effect later on when choosing what classification algorithms to use.

I next looked at the distribution of the categorical features included in the data. These correspond to: Pain, Haemoptysis, Shortness of Breath, Coughing, Weakness, Diabetes, Heart Attack w/n 6 mo., PAD, Smoker and Asthma and can be found in Table 1.

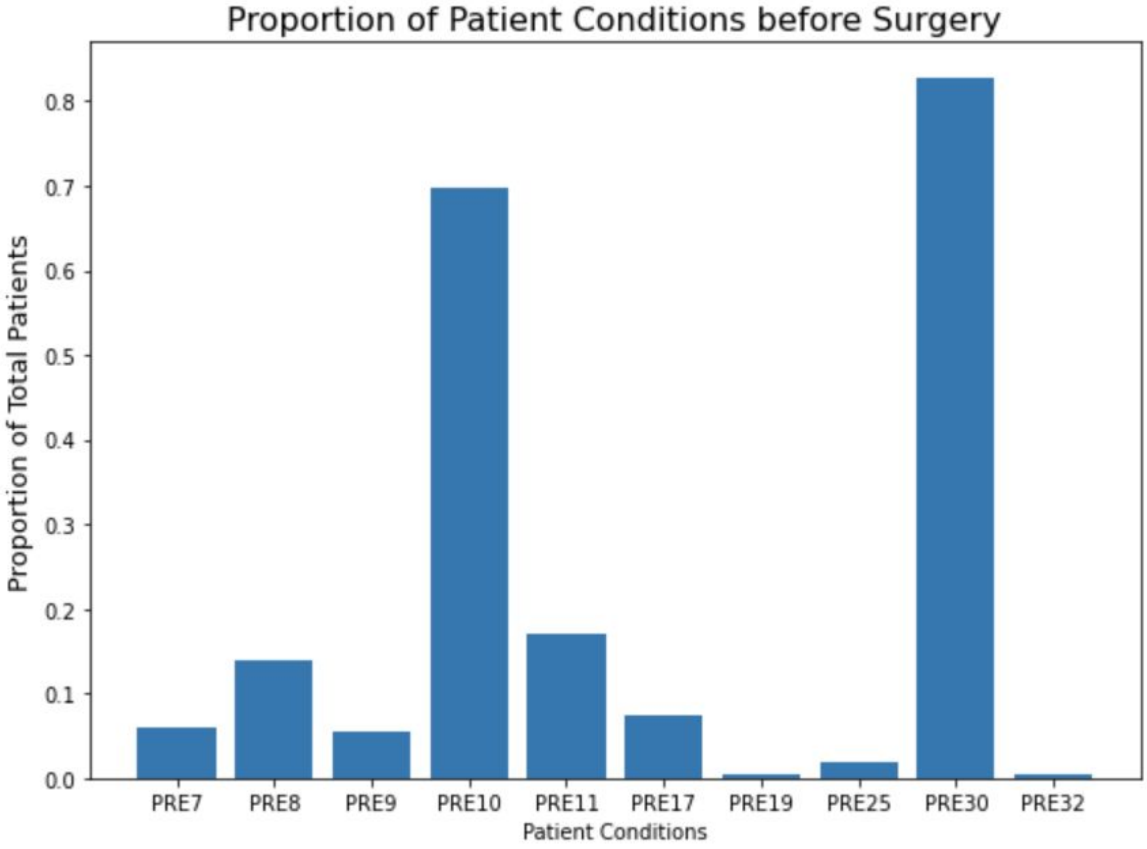


Fig 1: Proportion of Patient Conditions before Surgery

As seen in the plot above, the overwhelming majority of patients (both living and deceased) came into their surgical procedure experiencing a cough (PRE10) with a prior history of smoking (PRE30). Given that our cohort is related to lung cancer patients, this is not surprising. It also means that these two features are unlikely to give us any important information in regards to predicting a person’s most likely outcome after surgery.

With an idea of how our categorical features are distributed, I turned my focus to the continuous data. These features include: forced vital capacity (FVC), volume of air expelled after 1 second (FEV1) and the patient ages.

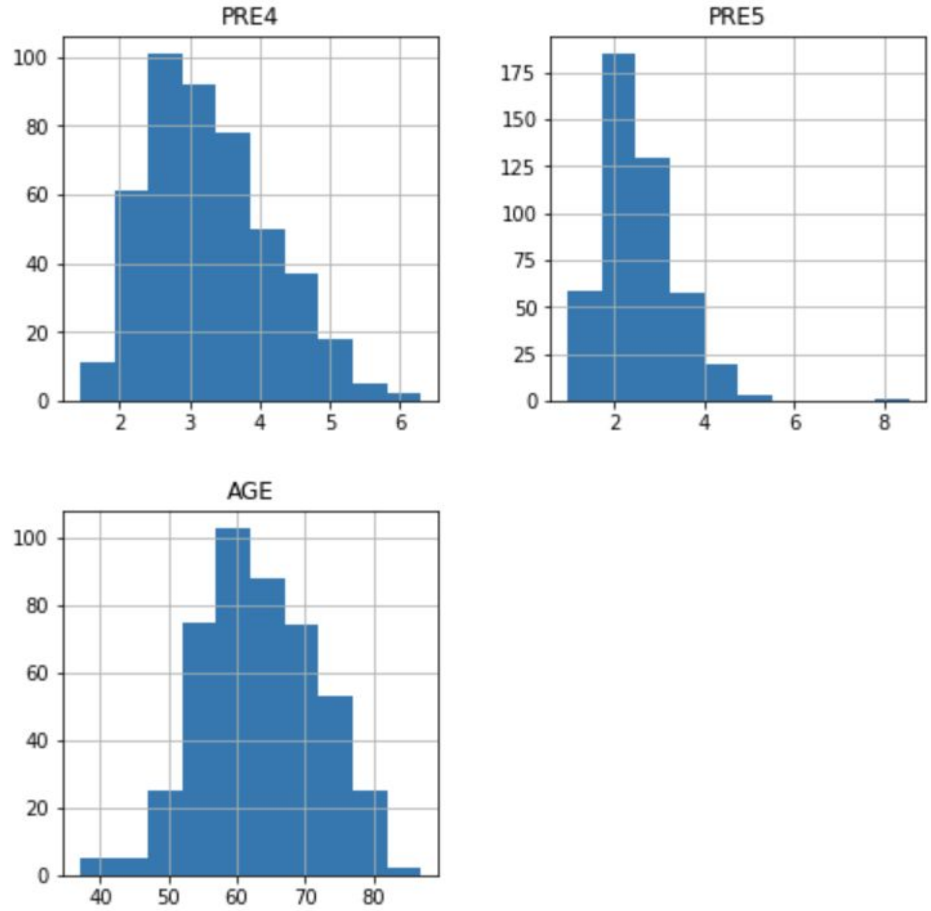
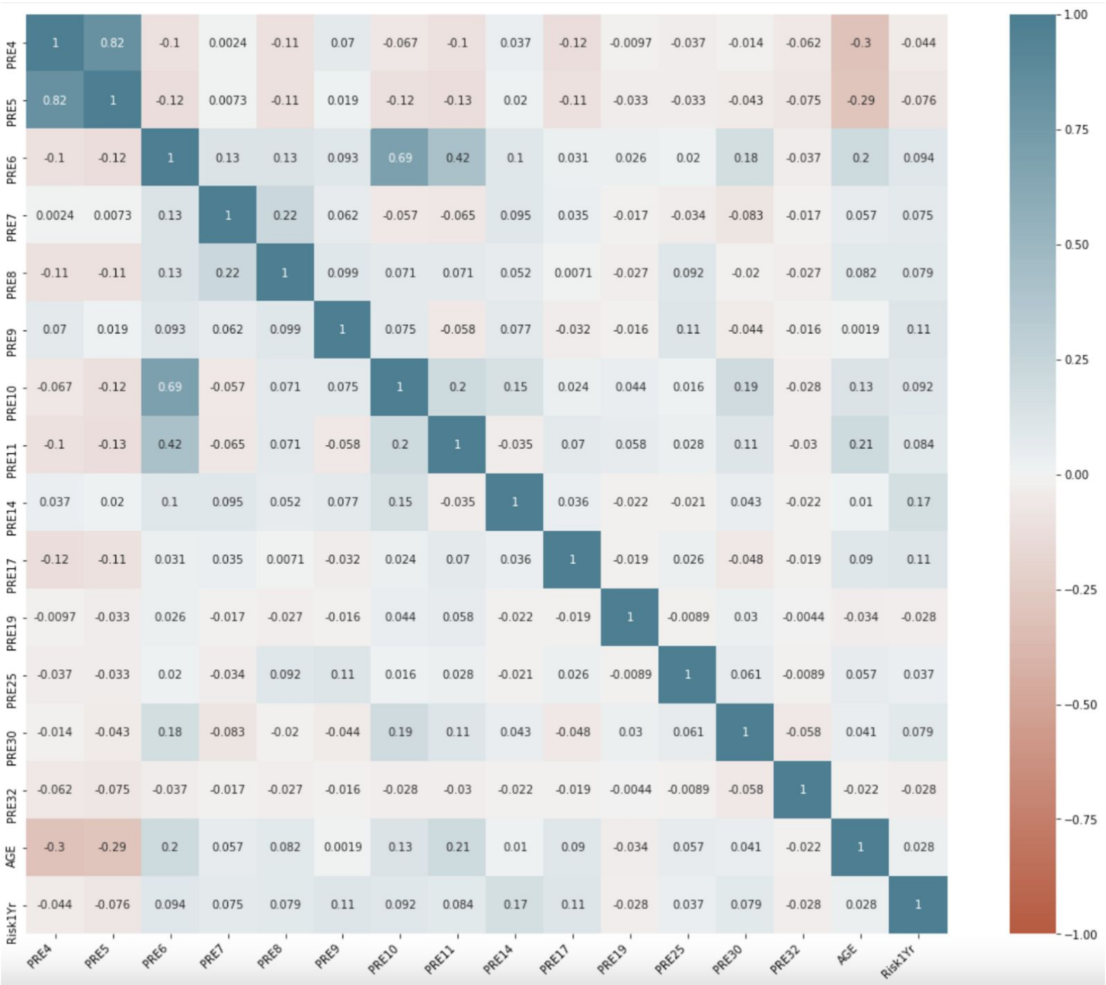


Fig 2: Distributions of Continuous Variables

From the distributions, a few interesting things can be observed. The distribution of the FVC is right skewed, however, the small range gives the impression that there is nothing atypical in the data. Our FEV1 and Age distributions, however, paint a bit of a different picture. Both of these features appear to have outliers present. In order to pull them out, and ultimately remove them, I first plotted box-and-whisker plots of each feature to visualize the outliers. I then removed the rows from the original dataframe where these features fell outside 3 standard deviations.



	PRE4	PRE5	PRE6	PRE7	PRE8	PRE9	PRE10	PRE11	PRE14	PRE17	PRE19	PRE25	PRE30	PRE32	AGE	Risk1Yr
PRE4	1	0.82	-0.1	0.0024	-0.11	0.07	-0.067	-0.1	0.037	-0.12	-0.0097	-0.037	-0.014	-0.062	-0.3	-0.044
PRE5	0.82	1	-0.12	0.0073	-0.11	0.019	-0.12	-0.13	0.02	-0.11	-0.033	-0.033	-0.043	-0.075	-0.29	-0.076
PRE6	-0.1	-0.12	1	0.13	0.13	0.093	0.69	0.42	0.1	0.031	0.026	0.02	0.18	-0.037	0.2	0.094
PRE7	0.0024	0.0073	0.13	1	0.22	0.062	-0.057	-0.065	0.095	0.035	-0.017	-0.034	-0.083	-0.017	0.057	0.075
PRE8	-0.11	-0.11	0.13	0.22	1	0.099	0.071	0.071	0.052	0.0071	-0.027	0.092	-0.02	-0.027	0.082	0.079
PRE9	0.07	0.019	0.093	0.062	0.099	1	0.075	-0.058	0.077	-0.032	-0.016	0.11	-0.044	-0.016	0.0019	0.11
PRE10	-0.067	-0.12	0.69	-0.057	0.071	0.075	1	0.2	0.15	0.024	0.044	0.016	0.19	-0.028	0.13	0.092
PRE11	-0.1	-0.13	0.42	-0.065	0.071	-0.058	0.2	1	-0.035	0.07	0.058	0.028	0.11	-0.03	0.21	0.084
PRE14	0.037	0.02	0.1	0.095	0.052	0.077	0.15	-0.035	1	0.036	-0.022	-0.021	0.043	-0.022	0.01	0.17
PRE17	-0.12	-0.11	0.031	0.035	0.0071	-0.032	0.024	0.07	0.036	1	-0.019	0.026	-0.048	-0.019	0.09	0.11
PRE19	-0.0097	-0.033	0.026	-0.017	-0.027	-0.016	0.044	0.058	-0.022	-0.019	1	-0.0089	0.03	-0.0044	-0.034	-0.028
PRE25	-0.037	-0.033	0.02	-0.034	0.092	0.11	0.016	0.028	-0.021	0.026	-0.0089	1	0.061	-0.0089	0.057	0.037
PRE30	-0.014	-0.043	0.18	-0.083	-0.02	-0.044	0.19	0.11	0.043	-0.048	0.03	0.061	1	-0.058	0.041	0.079
PRE32	-0.062	-0.075	-0.037	-0.017	-0.027	-0.016	-0.028	-0.03	-0.022	-0.019	-0.0044	-0.0089	-0.058	1	-0.022	-0.028
AGE	-0.3	-0.29	0.2	0.057	0.082	0.0019	0.13	0.21	0.01	0.09	-0.034	0.057	0.041	-0.022	1	0.028
Risk1Yr	-0.044	-0.076	0.094	0.075	0.079	0.11	0.092	0.084	0.17	0.11	-0.028	0.037	0.079	-0.028	0.028	1

Fig 3: Heatmap of Intrafeature Correlations

A heatmap was produced to observe any correlations that exist between the features and our target variable, and amongst the features themselves. The reason being that highly correlated features create a redundancy of information that can lead to poorer predictions later on in the modeling stage. A threshold of |0.5| was used to determine whether a feature was useful or should be removed. Rather unsurprisingly, the two breathing measurements were positively correlated. As were PRE6 and PRE10 which represent the Zubrod score (a measurement of a patient’s overall wellness) and the presence of coughing, respectively.

Model Selection:

For this project, I selected 3 classification models on which to train my data for its predictions: Weighted Logistic Regression, Decision Tree classifier and Random Forest classifier. Given that our project relates to a medical situation, where a predicted outcome of the positive class means that an individual is less likely to survive post surgery, it was also important to pick the appropriate metric. Because false negatives far outweigh false positives for this project, I chose to focus on Recall to minimize false negatives.

Results of Modeling

Model 1: Weighted Logistic Regression

A baseline model was first constructed using no weighting and using all of the features given to gauge how weighting affected model performance. Unsurprisingly, due to the imbalanced data, our model made every prediction in favor of the majority class. That is, it predicted ‘Alive’ for every prediction -- giving an 81% Accuracy , but 0% Recall.

After removing features that were highly correlated and using a GridSearch to find optimal weights, our model was able to bring its recall up to just over 50%. This was not a promising outcome and so further models were tested.

Model 2: Decision Tree Classifier

Decision trees are popular for classification problems because they don’t require scaling, can handle data of mixed types and gives more interpretability than some other models.

Our baseline model with unlimited depth resulted in overfitting, but again gave a point of comparison. Once tuned (max_depth=6, criterion=Gini) our model’s performance gave us a mediocre accuracy, but a high recall of around 92%. This would mean that while not suited to give definitive predictions on a patient’s post-op outcome, it would be able to tell which individuals would be far less optimal candidates.

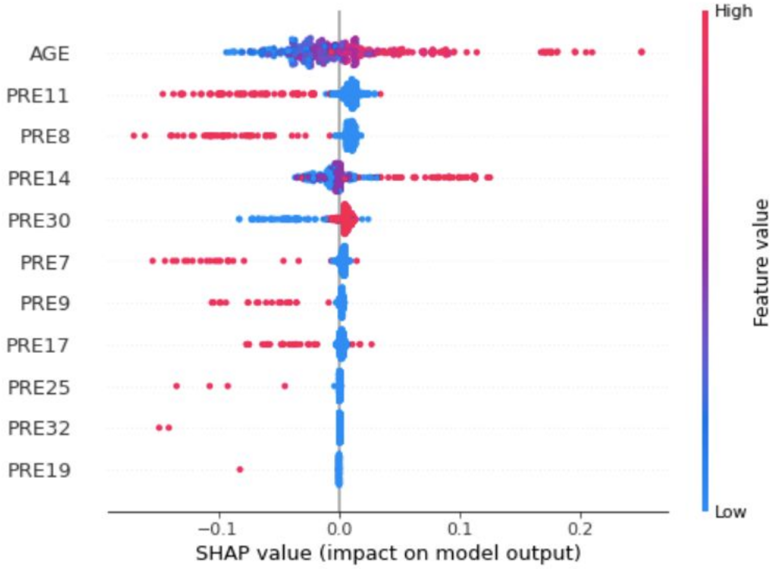
Model 3: Random Forest Classifier

For much of the same reasons that decision tree classifiers are popular, random forest classifiers expand on this by using an ensemble method. The algorithm creates multiple decision trees and then makes a final prediction by averaging the individual tree’s outcomes.

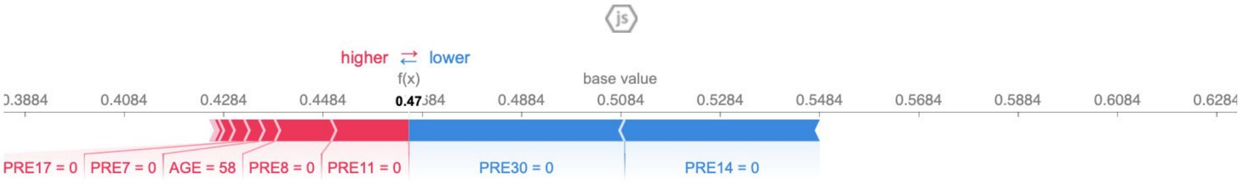
In order to give the model the best shot, in addition to removing some features, synthetic minority oversampling technique (SMOTE) was applied before training to lift the minority class in the training data to a 50/50 split with the majority.

After some tuning, the model gave a recall of 86% with an improved accuracy and precision over our decision tree classifier.

Figure 5: Shapley Additive Explanation values for Random Forest Classifier



Due to the nature of how random forest classifiers work, they don't carry the same level of interpretability as individual decision trees. For this reason, we looked at Shapley additive explanation values (SHAP values) to understand how the model was making its decisions. In brief, Shapley values break down a model to see how much individual features contributed to the model's prediction. Even better, they allow for localized analysis of individual observations (rows in our test data). This allows the user of the model to pick out cases where the model performance was low and make adjustments for further optimization. Below are examples of the output for the prediction of a survivor and non-survivor:



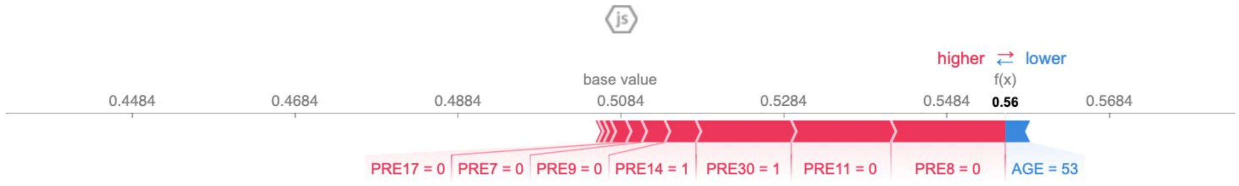


Fig 6: Force Plots of a model prediction of zero (top) vs. a positive prediction (bottom)

The advantage of the force plots above is that they pack a lot of information about how our classifier is working into an easily digestible visual. For the two examples above, our model gave the upper prediction a 53% chance of survival with their absence of smoking and small tumor size leading contributors to this decision. Conversely, the individual below was given a 56% chance of death based on a combination of several variables including being a smoker and having a larger tumor. It is important to note that these are far from perfect results and that what these Shapley force plots do is provide guidance on next steps to take to improve our classifier.

Having compared the three models that were tested, it appears that the best results came from the Decision Tree model with a max depth of 6 and it had the highest recall among those tested.

Moving Forward:

This project really opened my eyes to the ways in which machine learning can be used in the medical field to better improve patient care. With even relatively basic models, you can gain greater insight into how a patient's complicated medical history could be used to make more data driven decisions.

There are several avenues which I would like to explore further if given the time. I would like to see if there would be a way to collect more data. The more data available to train the model with, the more robust it would be. This could be increasing the number of patients, having additional features or being able to segment patients based on the specific type of surgery they had.

Additionally, there are other algorithms and hyperparameter tuning that could be tested to see if their performance on this data set gave any improvement.

Works Cited:

1. “Thoracic Surgery.” *NorthShore*, NorthShore University Health System, 2020, www.northshore.org/thoracic-surgery/procedures/lung-resection/.
2. “Common Cancer Sites - Cancer Stat Facts.” *SEER*, National Cancer Institute: SEER, 2020, seer.cancer.gov/statfacts/html/common.html.
3. Elsevier. "'Aggressive' surgery is best treatment option for early stage lung cancer: Patients who undergo lobectomy for the disease live longer." ScienceDaily. ScienceDaily, 30 November 2017. <www.sciencedaily.com/releases/2017/11/171130122808.htm>.