Outcomes after Thoracic Surgery for Patients with Lung Cancer

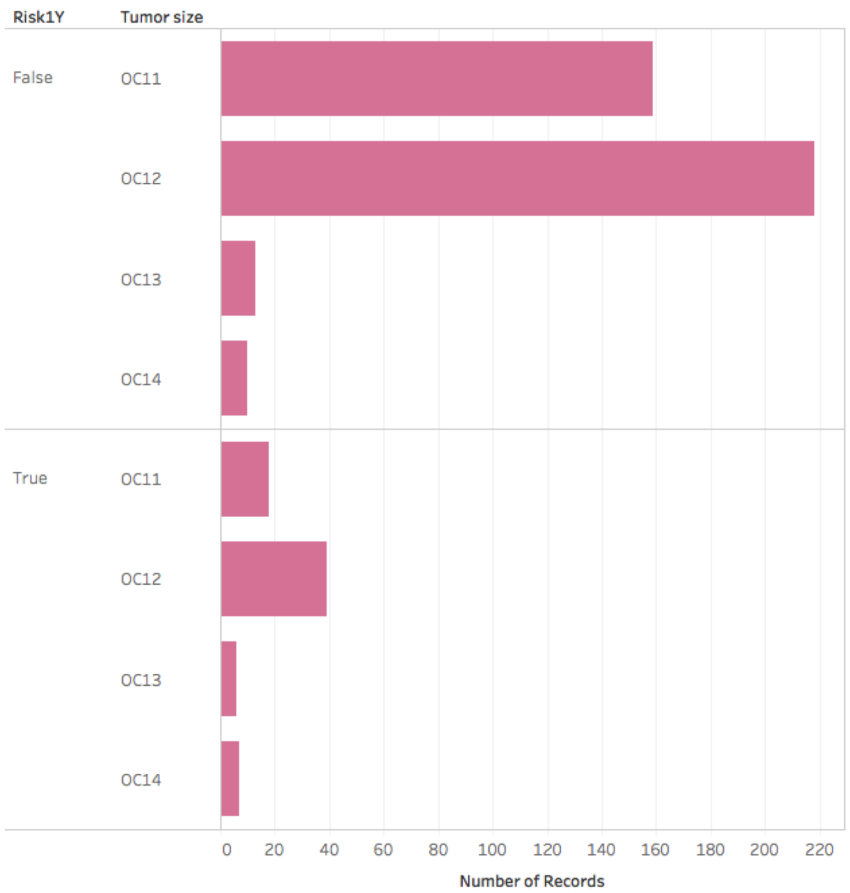
Capstone 2

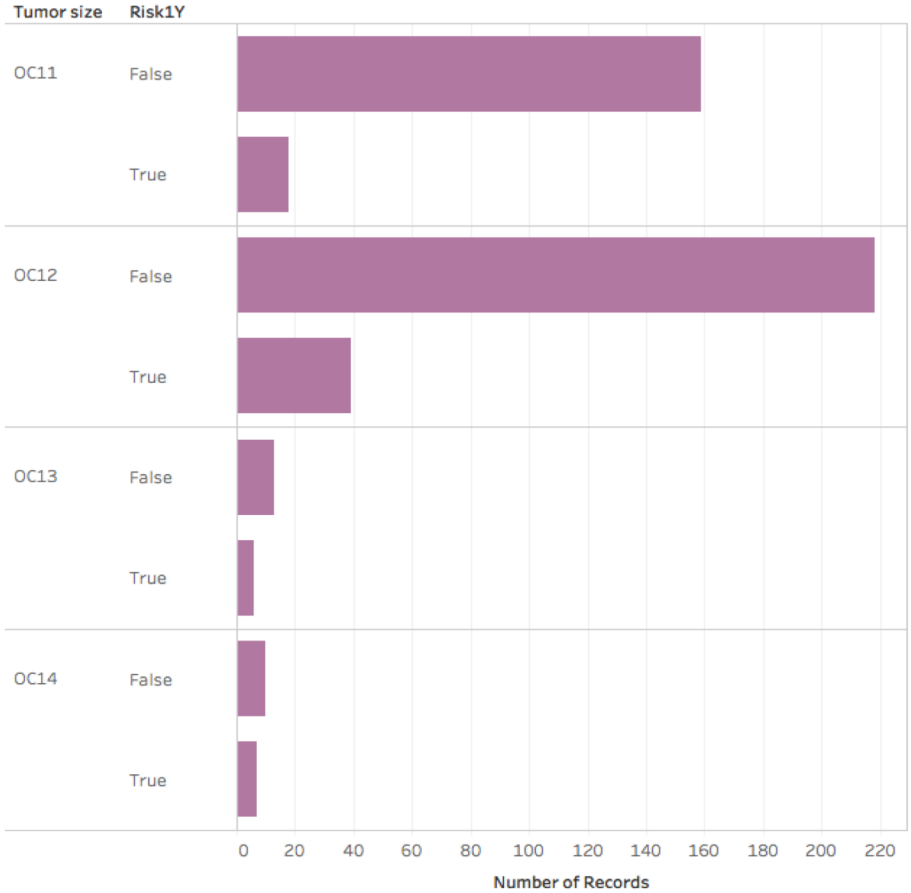
V. Moore

The Problem

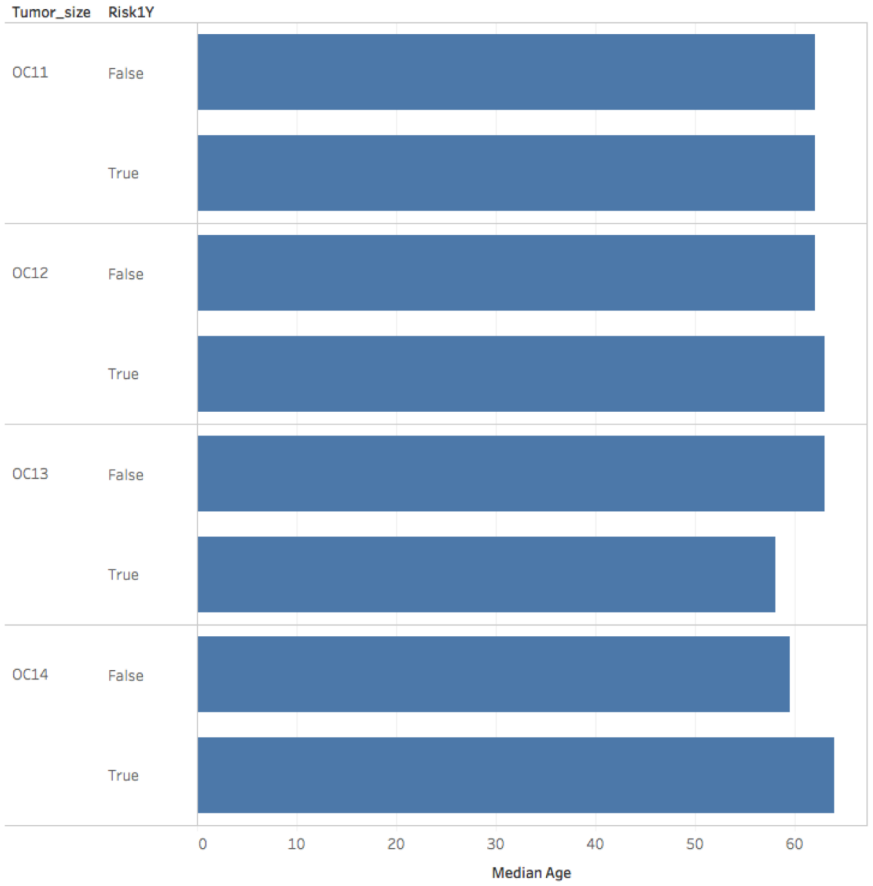
- Surgical resection for lung cancer carries risks.
- How do we determine risk of mortality following this surgery?
- Answers could be useful to patients and anyone involved in patient care.
- We look at one-year post-surgery patient survival data from the Wroclaw Thoracic Surgery Centre in Poland, available from the UC Irvine Machine Learning Repository, collected 2007-2011 from 470 patients, of whom 70 did not survive.
- Target variable is “Risk1Y”, positive for mortality at one year after surgery.
- Features include diagnosis code (DGN), forced vital capacity (FVC), forced expiratory volume in first second (FEV1), Zubrod performance score, presence of pain, presence of haemoptysis, presence of dyspnea, presence of cough, presence of weakness, tumor size, presence of type II diabetes (T2DM), history of myocardial infarction (MI), presence of peripheral arterial disease (PAD), presence of smoking, presence of asthma, and age.

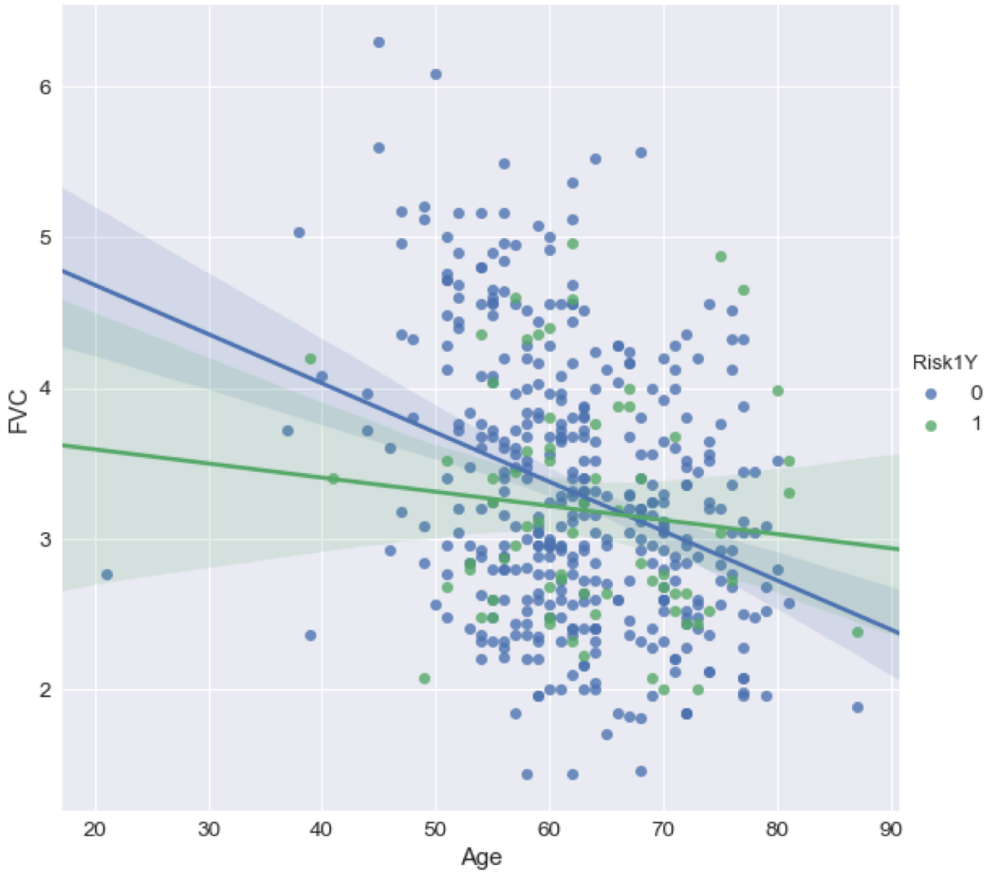
Some patterns of mortality risk



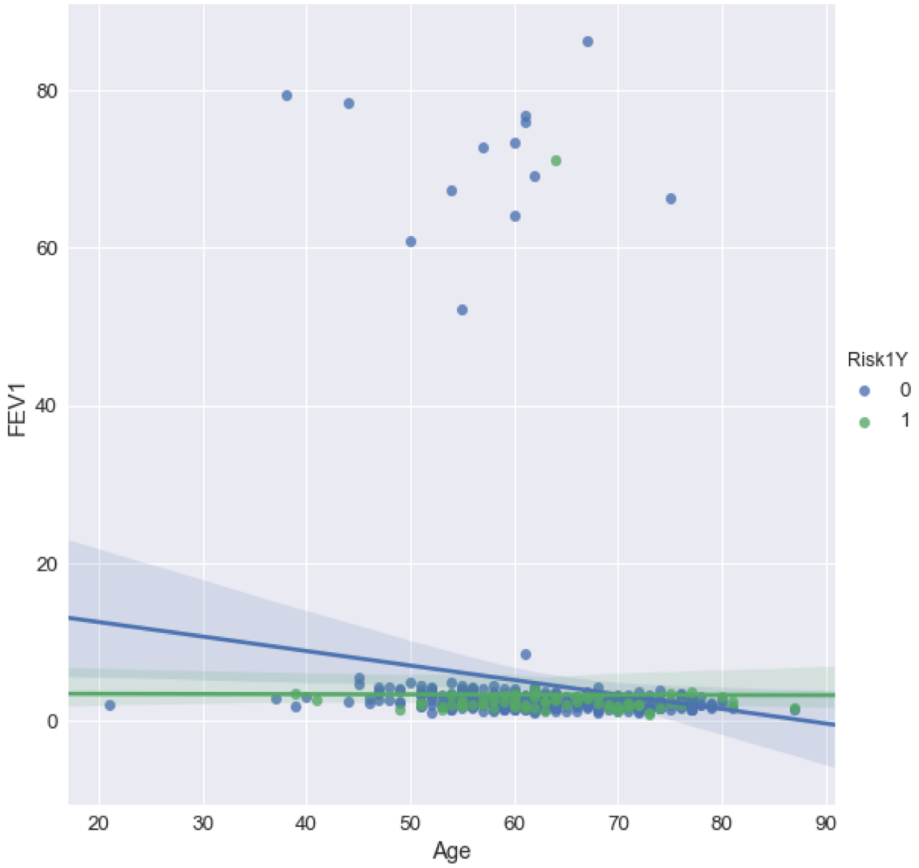


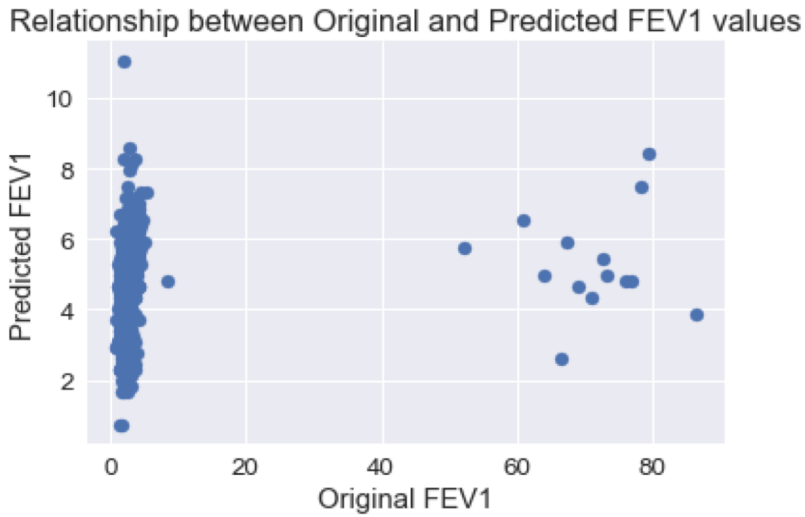
Patterns in data by age

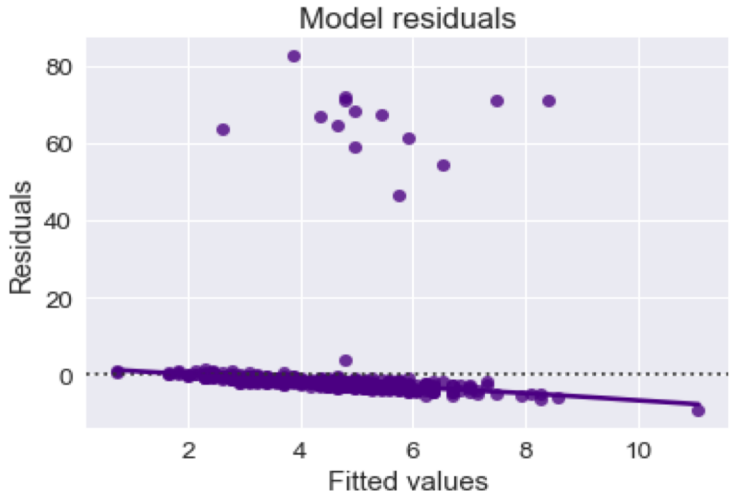




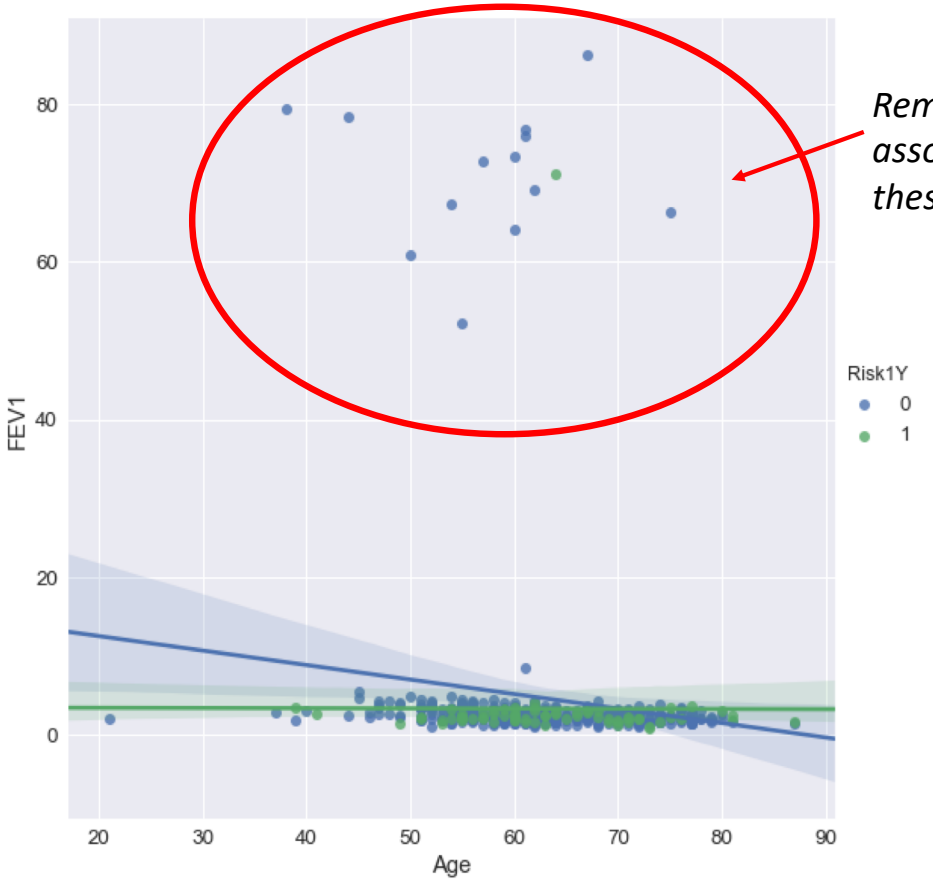
Outlier treatment



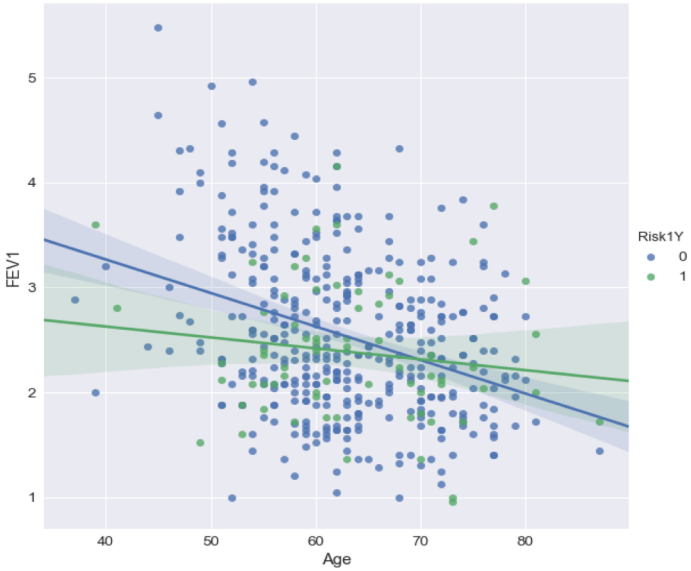


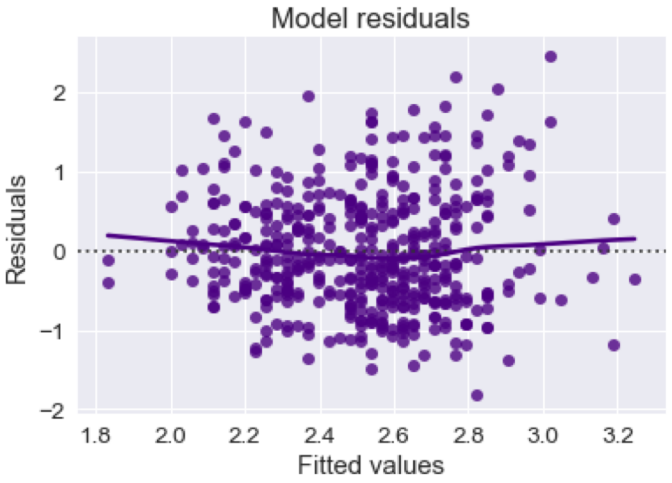


Outlier treatment

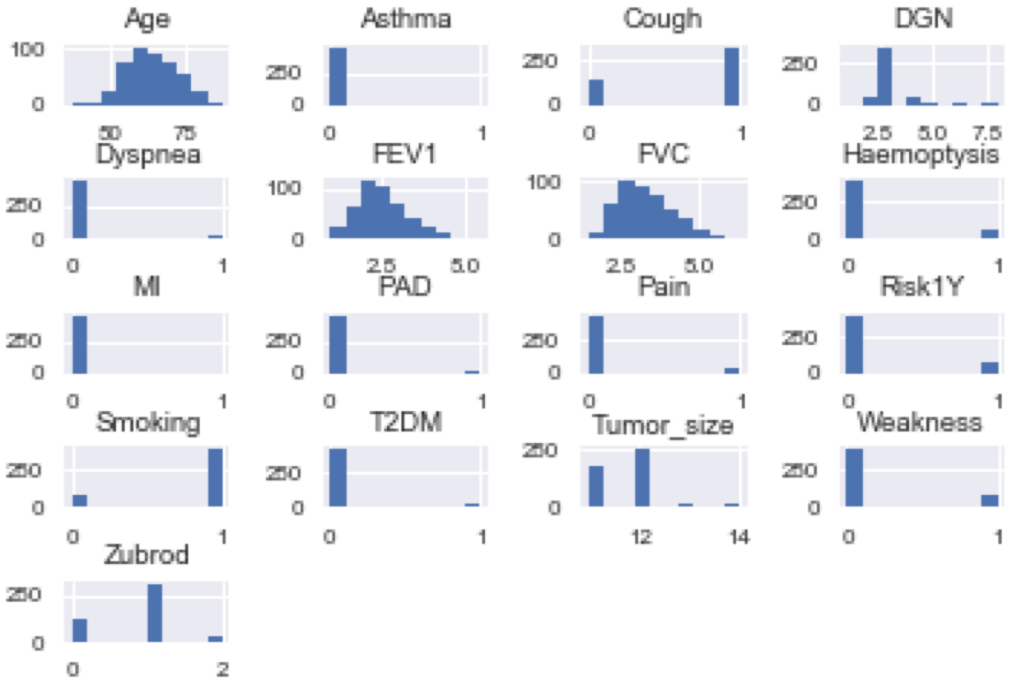


Removal of entries associated with these points





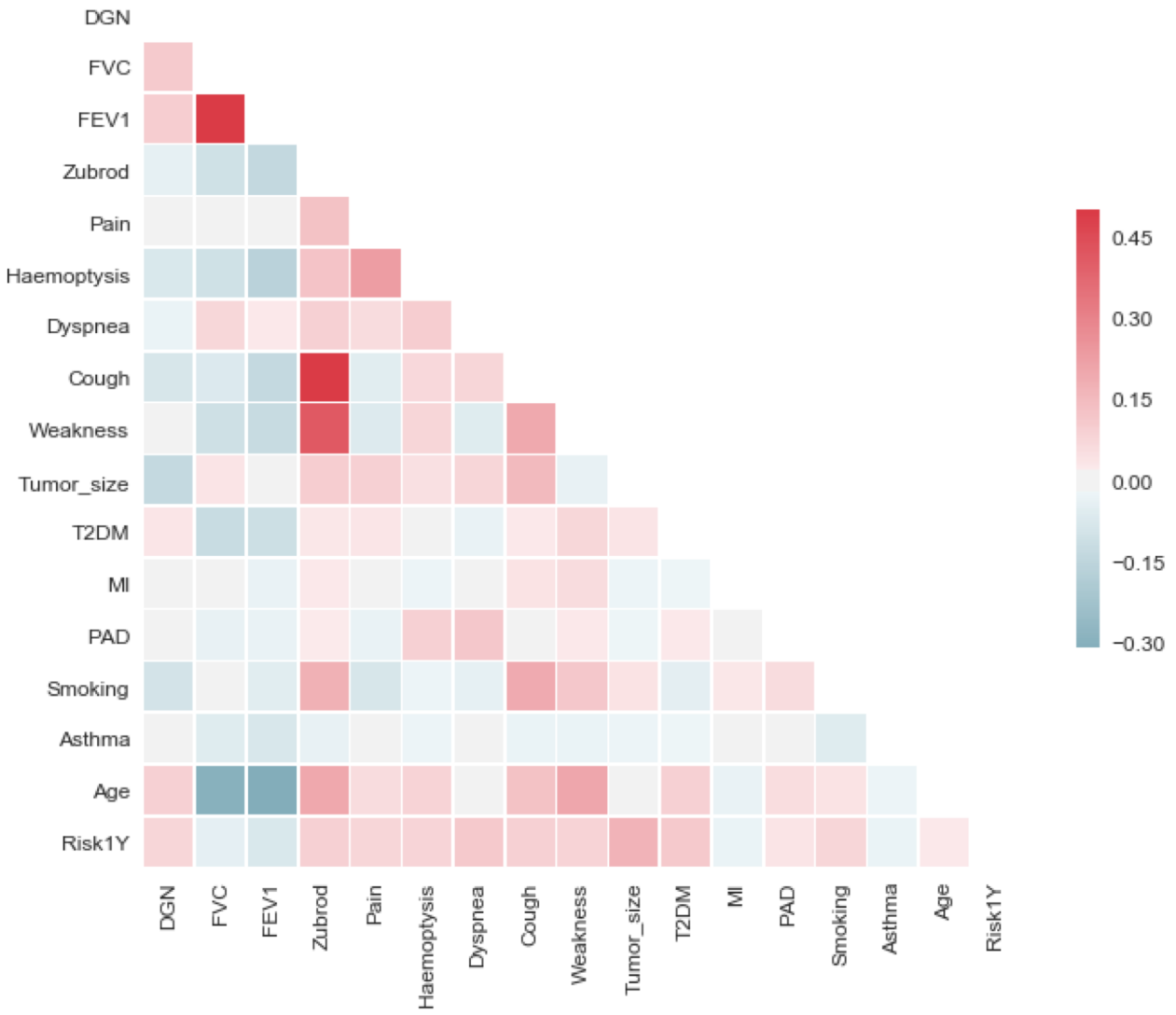
Quick view of histograms after removal of outliers



- FEV1 is distributed as expected without outlier entries.
- The three continuous variables (age, FVC, FEV1) appear normally distributed or close to being so.
- The discrete variables show that most patients carry the healthier characteristic for most variables.
- However, presence of cough is more common than absence is, as is smoking, and Zubrod is at 1 for most (rather than the optimal 0).
- The most common diagnosis is DGN3, though it is not defined what diagnosis this is.
- The target variable is “Risk1Y”, with 0 meaning survival and 1 being non-survival at one year after surgery.

Correlations

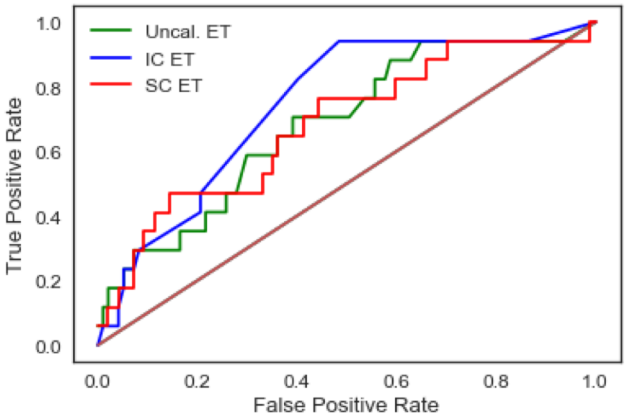
- Not many strong correlations among features or the target
- FVC and FEV1 are highly correlated (Pearson's r = 0.88, p=0.0), but this is not unexpected given that FEV1 is a component of FVC measurement.
- Cough and weakness are associated with Zubrod score here, but these relationships are not unexpected either.

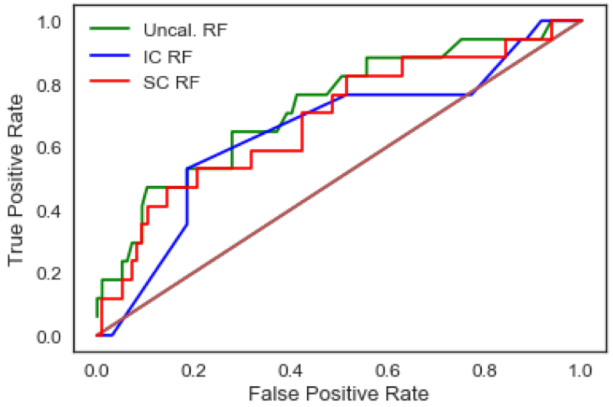


Machine learning & imbalance

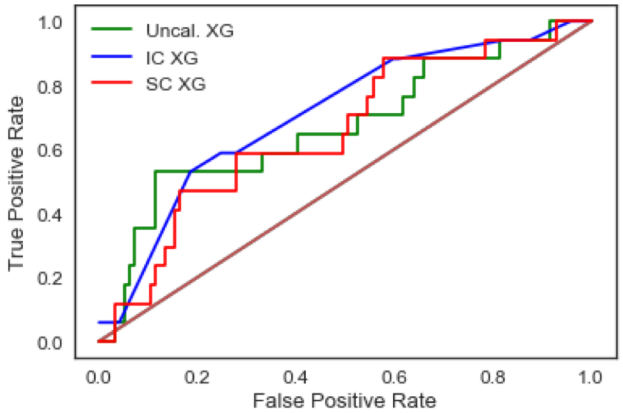
- The target class of the dataset is imbalanced with 400 survivors and 70 non-survivors.
- Attempts to address this included multiple treatments (SMOTE, down-, and upsampling treatments of training data), but ultimately analysis relied on stratification by class upon train-test-splitting, with samples weighted by target class
- Isotonic and sigmoid calibrations were also tested with multiple classifiers

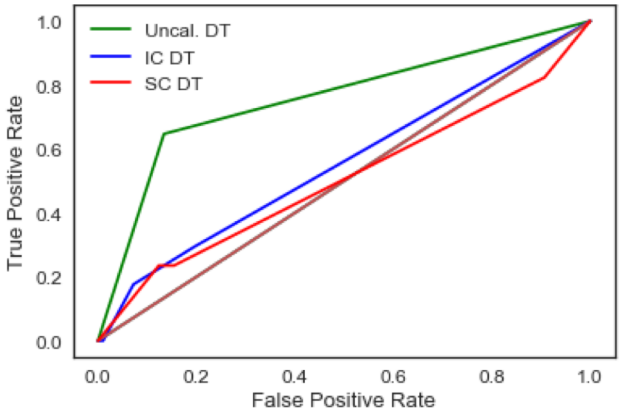
ROC curves with 4 classifiers and calibration





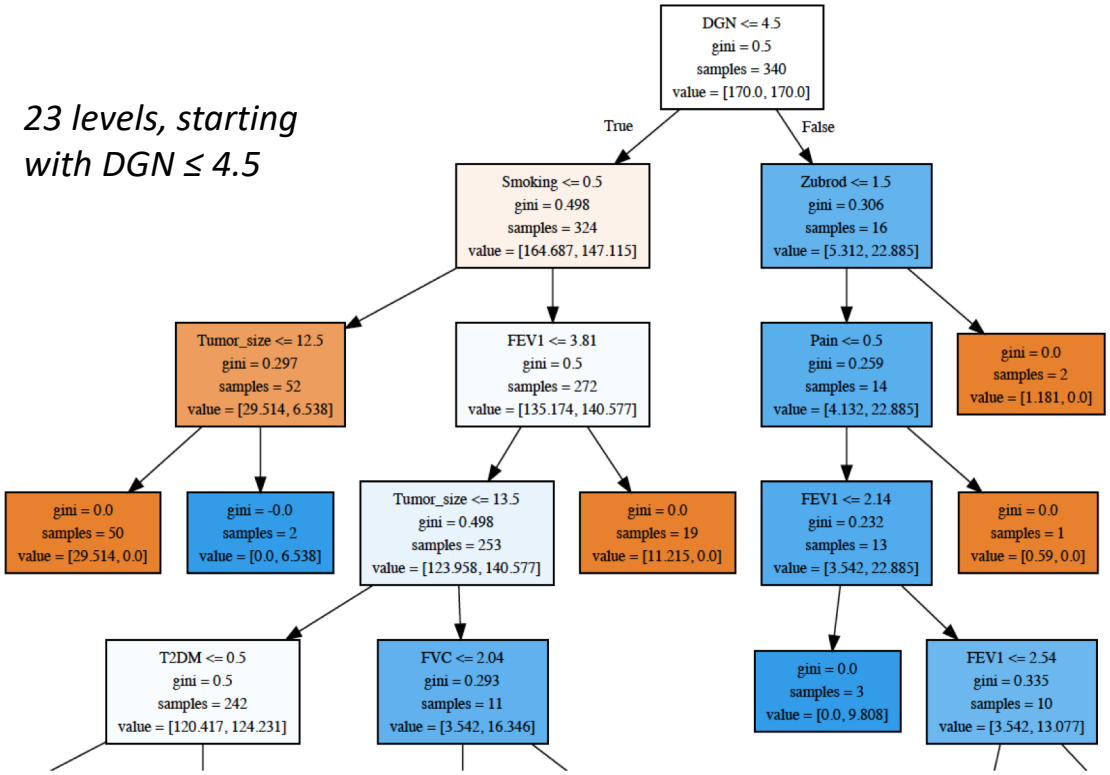
Key:
"Uncal." = uncalibrated
"IC" = isotonic calibration
"SC" = sigmoid calibration
"ET" = extra trees classifier
"RF" = random forest classifier
"XG" = XGBoost classifier
"DT" = decision tree classifier





Decision tree

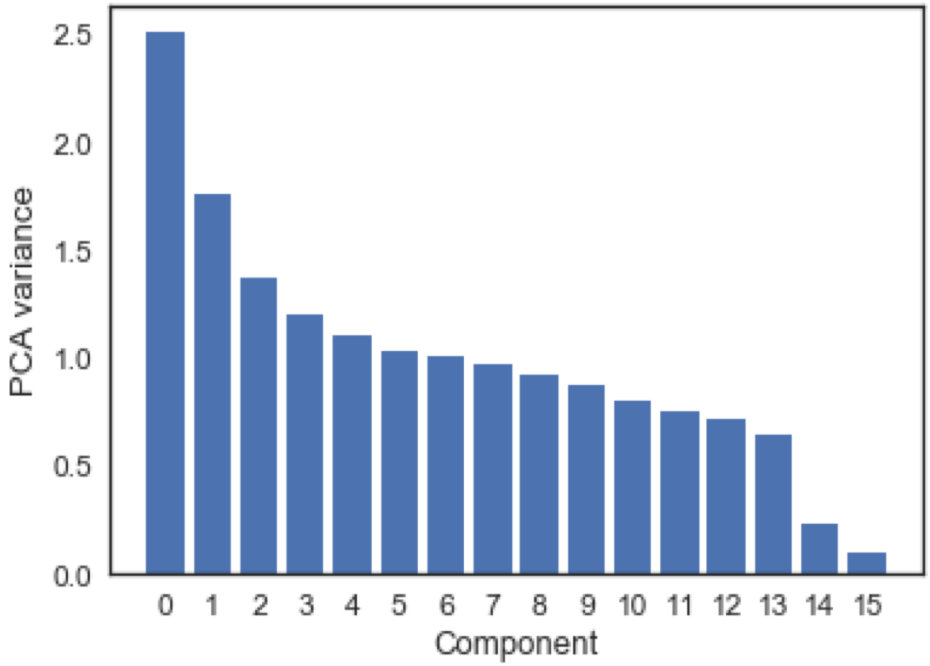
23 levels, starting with DGN ≤ 4.5



	Score	Class	Precision	Recall	F1-score	Support
Extra trees (isotonic calib.):						
Brier score	0.229					
AU-ROC	0.7216					
		0	0.94	0.60	0.73	97
		1	0.25	0.76	0.38	17
		Avg/total	0.83	0.62	0.68	114
Random forest (no calib.):						
Brier score	0.324					
AU-ROC	0.7247					
		0	0.85	1.00	0.92	97
		1	0.00	0.00	0.00	17
		Avg/total	0.72	0.85	0.78	114
XGBoost (isotonic calib.):						
Brier score	0.233					
AU-ROC	0.7089					
		0	0.91	0.73	0.81	97
		1	0.28	0.59	0.38	17
		Avg/total	0.82	0.71	0.75	114
Decision tree (no calib.):						
Brier score	0.277					
AU-ROC	0.722					
		0	0.92	0.86	0.89	97
		1	0.42	0.59	0.49	17
		Avg/total	0.85	0.82	0.83	114

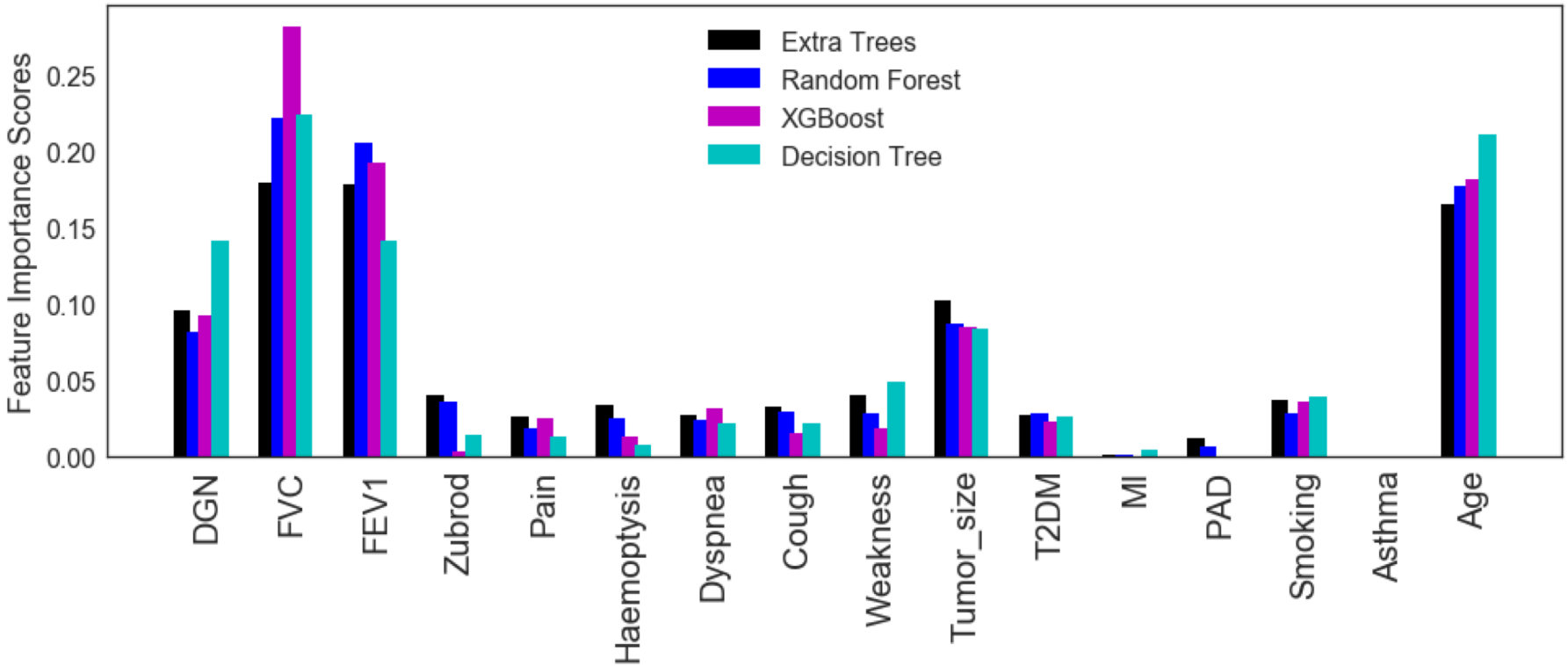
Initial machine learning results:
Test data metrics with sample weighting, stratification, and calibration (calib.).

Feature importance: principal component analysis

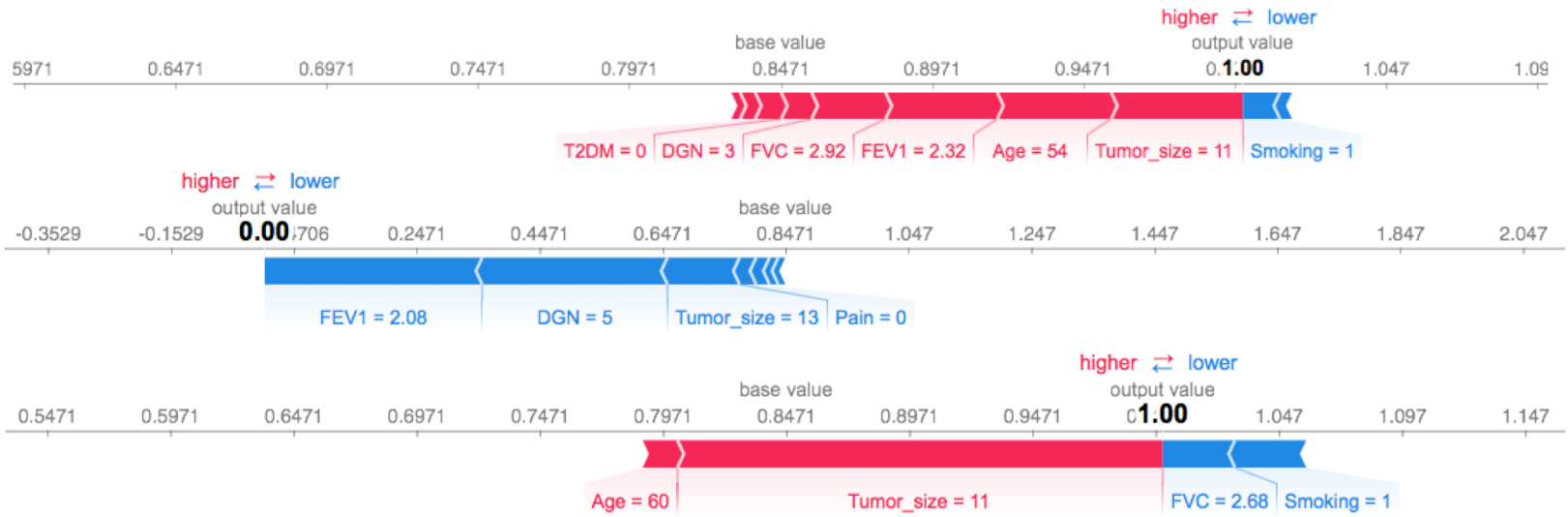


Many components of relatively similar importance in this dataset

Feature importance: by classifier

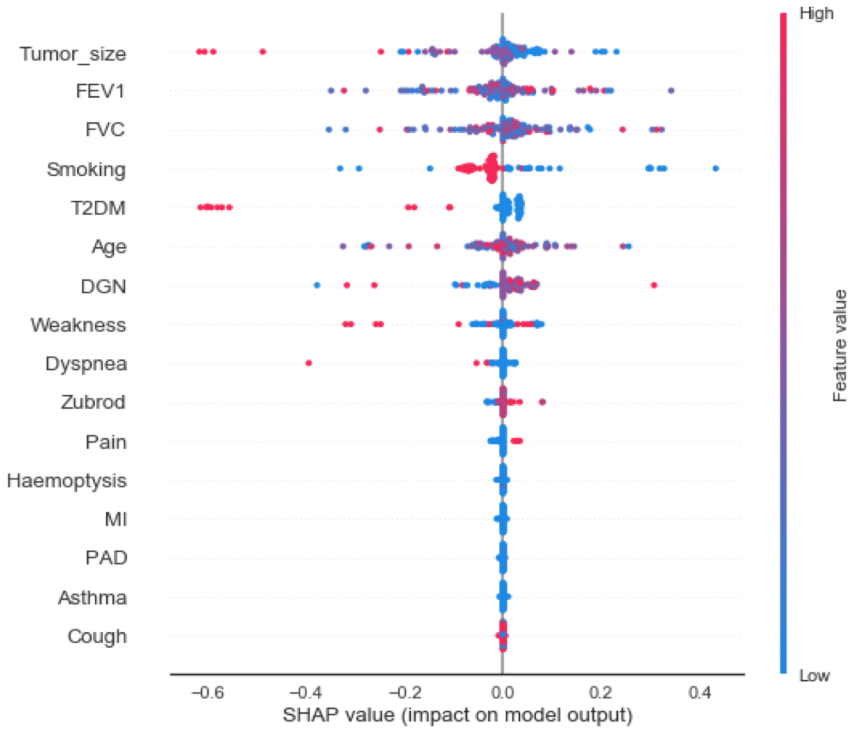


SHAP-derived individual feature contributions: decision tree, all features



In the top panel is descriptive output for a correct prediction of survival in a patient, with values that appear to make sense contributing to this prediction. The middle panel shows a correct prediction of non-survival with values that appear to make sense, apart from the assignment of absence of pain as a contributor to this prediction. The bottom panel shows an incorrect prediction of survival for a patient who did not survive. However, the patient's data are mostly similar to those for patients who do survive, though FVC is a little low.

SHAP summary: decision tree, all features



Model AU-ROC: 0.7462

Correct predictions for 80/97 survivors and 10/17 non-survivors

Tumor size and the respiratory function tests (FEV1, FVC) are treated as most important here, though in some cases feature value contributions appear complicated.

Removed feature(s)	Extra trees AU-ROC	Random forest AU-ROC	XGBoost AU-ROC	Decision tree AU-ROC
Age	0.7884 (IC)	0.7508 (IC)	0.6683 (SC)	0.7859 (UC)
Tumor size	0.6316 (IC)	0.6892 (IC)	0.6331 (UC)	0.6043 (UC)
Diagnosis code	0.7195 (UC)	0.7362 (UC)	0.6771 (IC)	0.6528 (UC)
FVC	0.7295 (IC)	0.6883 (UC)	0.7283 (SC)	0.5992 (UC)
FEV1	0.6898 (IC)	0.6947 (IC)	0.6413 (IC)	0.6874 (UC)
Smoking	0.738 (IC)	0.7623 (SC)	0.7429 (IC)	0.6058 (UC)
Cough	0.6922 (IC)	0.7389 (UC)	0.6689 (UC)	0.722 (UC)
Haemoptysis	0.7878 (IC)	0.782 (SC)	0.6743 (IC)	0.7514 (UC)
Type II diabetes	0.7053 (UC)	0.7347 (UC)	0.6546 (UC)	0.6416 (IC)
Zubrod score	0.7277 (IC)	0.7492 (IC)	0.6998 (IC)	0.722 (UC)
Pain	0.721 (IC)	0.7638 (UC)	0.7029 (SC)	0.6856 (IC)
Dyspnea	0.7044 (IC)	0.732 (UC)	0.6998 (IC)	0.7116 (UC)
Weakness	0.7226 (IC)	0.7216 (IC)	0.6886 (IC)	0.6886 (IC)
MI	0.7814 (IC)	0.7456 (UC)	0.7089 (IC)	0.7359 (UC)
PAD	0.7662 (UC)	0.7195 (SC)	0.7089 (IC)	0.7168 (UC)
Asthma	0.7638 (IC)	0.7338 (UC)	0.7089 (IC)	0.7565 (UC)
Age and haemoptysis	0.7483 (UC)	0.7762 (IC)	0.6692 (SC)	0.7911 (UC)
Age and MI	0.7771 (IC)	0.7671 (UC)	0.6683 (SC)	0.7911 (UC)
Age and PAD	0.7738 (IC)	0.7838 (IC)	0.6683 (SC)	0.7911 (UC)
Age and asthma	0.775 (IC)	0.7632 (UC)	0.6683 (SC)	0.7911 (UC)
Age, haemoptysis, MI	0.7571 (IC)	0.762 (IC)	0.6692 (SC)	0.7962 (UC)
Age, haemoptysis, PAD	0.7435 (SC)	0.7693 (UC)	0.6692 (SC)	0.7911 (UC)
Age, haemoptysis, asthma	0.7632 (IC)	0.7744 (IC)	0.6692 (SC)	0.7911 (UC)
Haemoptysis, asthma, MI, PAD	0.7893 (IC)	0.7544 (UC)	0.6743 (IC)	0.7462 (UC)
Age, haemoptysis, asthma, MI, PAD	0.7859 (IC)	0.7641 (UC)	0.6692 (SC)	0.7962 (UC)

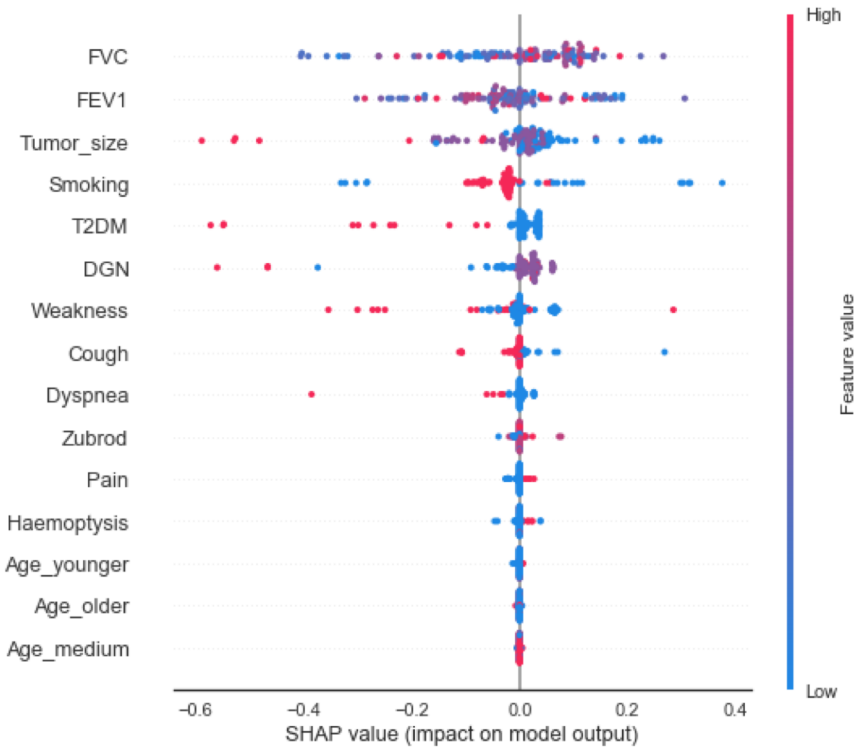
Machine learning results after feature removal: AU-ROC scores with test data for models with one or more features removed, with calibration type noted in parentheses as UC (uncalibrated), IC (isotonic), or SC (sigmoid).

Ultimately contributed to model development with binned age groups, rather than continuous ages, and removal of asthma, MI, and PAD from the model (due to rarity of these conditions).

SHAP-derived individual feature contributions: decision tree, binned ages, removed MI, PAD, asthma

Results from the same patients as shown 3 slides ago. The top two predictions are again correct, with the bottom one still incorrect. Model AU-ROC is slightly improved from before. Most feature contributions here are sensible, with the exception of weakness and pain.

SHAP summary: decision tree, binned ages, removed MI, PAD, asthma



Model AU-ROC: 0.7808

Correct predictions for 84/97 survivors, 12/17 non-survivors

Binned ages play less of a role in model development. MI, PAD, and asthma have very small effects on AU-ROC, but with this dataset they are rare occurrences. FVC, FEV1, tumor size and smoking occupy similar positions as before.

Perspective

- It is challenging to develop a highly accurate predictive model for mortality over the course of a year, but with proper sample weighting and calibration (depending on classifier), it may be possible to generate a moderately useful model.
- Since all the top models presented here are wrong at least 20% of the time, the judgment of the healthcare team is paramount for considering any decisions around surgery for patients under consideration of thoracic surgery for lung tumor resection.
- With the decision tree classifier, it was possible to make a model with an AU-ROC score of 0.78-0.79 and 83% test accuracy, with 84/97 survivors and 12/17 non-survivors correctly identified.
- Feature importance varied some by model, but accurate pulmonary function tests (FVC, FEV1) were key to predictions. Tumor size, diagnosis code, and age also, depending on the model, were important.
- In some models examined here, different features were removed, but with somewhat similar success among most models, a model could be chosen in examination of a patient based on what features are available or of interest.
- Also, while SHAP output and the decision tree classifier are emphasized here, the extra trees classifier also showed consistently high AU-ROC scores for many analyses. The SHAP python package is not currently tuned to the extra trees classifier, but perhaps this will change.
- The model provides a tool for to patient care providers to assess the risks and benefits associated with recommending a patient for lung tumor resection and also can provide a basis to assist patients in considering such risks and benefits for themselves.
- In conjunction with the clinical judgment of a patient's care providers, use of this tool has potential to improve patient care and outcomes.

References

- Falcoz, PE, M Conti, L Brouchet, et al. 2007. The thoracic surgery scoring system (Thorascore): risk model for in-hospital death in 15,183 patients requiring thoracic surgery. *The Journal of Thoracic and Cardiovascular Surgery* 133:325-332.
- Jose, BM. 2017. "Find Fraud in Enron Data." Web. Accessed: 7/12/18. https://bibinmjose.github.io/Enron_find_fraud/ .
- Lundberg, S. "A unified approach to explain the output of any machine learning model." Web. Accessed: 7/12/18. https://github.com/slundberg/shap .
- Miller, MR, J Hankinson, V Brusasco, et al. 2005. Standardisation of spirometry. *European Respiratory Journal* 26:319-338. Also available on the web: https://www.thoracic.org/statements/resources/pfet/PFT2.pdf .
- University of California, Irvine, Machine Learning Repository. Thoracic Surgery Data Set. Web. Accessed: 6/20/18. http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data . Citation associated with dataset: Zieba, M, JM Tomczak, M Lubicz, & J Swiatek. 2014. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing* 14A:99-108. Principal investigator information for component datasets used in this study: M Lubicz (1), K Pawelczyk (2), A Rzechonek (2), and J Kolodziej (2):
  -- (1) Wroclaw University of Technology, wybrzeze Wyspianskiego 27, 50-370, Wroclaw, Poland
  -- (2) Wroclaw Medical University, wybrzeze L. Pasteura 1, 50-367 Wroclaw, Poland