Lung Cancer Survival Prediction After Surgery

Brandon May

Abstract:

Lung cancer is a particularly devastating type of cancer. Even with proper recognition and treatment, many times the cancer can be too advanced to successfully cure and treat. There are two main types of lung cancer: small cell and non-small cell lung cancers. In recent years, mortality from lung cancer has improved likely due to specifically targeted therapies for genes within the tumors themselves (Howlader *et al.*, 2020). Early stage cancers are typically treated surgically with more advanced chemotherapy and radiation added for higher stage disease. The higher the stage of disease, the worse the prognosis.

This dataset was obtained from the UCI Machine Learning Repository and was a patient population studied in Poland from 2007-2011. There were 470 patients in the study with various variables from lung function, presence of vascular risk factors, performance status, and tumor stage with the target variable being survival after lung resection at 1 year. In the dataset, if the patient died, this was coded as true.

There were a mix of categorical and numeric variables in this dataset. There are various predictors of survival in these patients including smoking status, performance status, and how big or advanced the tumor was among others. Cancer may recur even with optimal treatment but with more limited disease, the recurrence rate is lower (Pairolero *et al.*, 1984). The lower the T number of the cancer, the better the prognosis.

This project aimed to determine if machine learning can predict survival after 1 year after lung surgery for lung cancer as well as determining which variables are relatively more important when predicting survival vs. others. If machine learning can identify higher risk patients of death so that closer monitoring and treatment can be performed, machine learning could positively impact treatment of patients with serious disease.

Machine learning algorithms had difficulty predicting both classes of the target variable accurately and were assessed using K Means Clustering, Logistic Regression, Support Vector Machine Classifier, K Nearest Neighbors, and XG Boost methods. The best performing model with the highest F1 score for both predicting survival and death was logistic regression. The limited dataset and the imbalanced target class created difficulties in accurate prediction and further data points on a wide variety of patients and demographics would be required to make the algorithms more accurate.

Introduction:

Lung cancer is a particularly devastating type of cancer. On an annual basis, there are over 135,000 deaths from lung cancer in the United States (Siegel *et al.*, 2019). Even with proper recognition and treatment, many times the cancer can be too advanced to successfully cure and treat. In recent years, mortality from lung cancer has improved likely due to specifically targeted therapies for genes within the tumors themselves (Howlader *et al.*, 2020).

For early stage cancers, the best predictors for survival are functional status as well as pulmonary function tests to determine lung capacity (Midthun, 2020). Some studies have shown that functional status is a particularly important prognostic indicator (Sculier *et al.*, 2008). Cancer may recur even with optimal treatment but with more limited disease, the recurrence rate is lower (Pairolero *et al.*, 1984). Surgery can also be followed with specific chemotherapy or radiation depending upon tumor type and other staging characteristics. Staging for cancers (with some exceptions) usually consists of a TNM classification which is then translated into an overall stage of 1 through 4. The T stands for the size of the tumor and is rated from 1-4. The N stands for any lymph node involvement. The M stands for any presence of metastasis or distant spread of the cancer (National Cancer Institute, 2015). The presence of metastasis or stage 4 disease has a particularly poor prognosis and survival is limited.

This dataset was used in a published paper in Applied Soft Computing and can be found at the following URL: https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

There were 470 study participants in this research study who had lung cancer resections between the years of 2007 and 2011 in Poland (Zieba *et al.*, 2014). Various data was tabulated including any history of heart attack, diabetes, smoking, shortness of breath, etc. Please refer to the appendix below for complete variable descriptions.

The target variable was survival 1 year after the lung resection. If the value was true, then that means the person did not survive. There is a mix of both numeric variables as well as categorical variables in this dataset. The numeric variables include lung volume measurements and age. There are also codes for specific diagnosis codes as well as categories of functional status. The categories of functional status is using the Zubrod scale which is specifically used in cancer patients with a value of 0 being asymptomatic, 1 being symptomatic but able to move, and 2 requiring some assistance and being mobile 50% of the time (West & Jin, 2015).

Interestingly, this dataset is comprised of many of the variables that are medically known to translate to survival in clinical studies. Main questions to answer with this project include determining variable importance that correlates with survival, which vascular risk factors are most important, and to determine which machine learning algorithm is most beneficial for prediction.

Methods:

The data was downloaded from the repository in an arff format. The format was converted to a csv file and then imported into Python using pandas. Two of the columns included identifiers as well as ICD-9 diagnosis codes, which were not useful in analysis, so they were removed. There were no missing values in the dataset. The predictor variables included three numerical values FVC, FEV1, and Age. FVC and FEV1 are measurements of lung capacity and age was the age of the patient. The rest were categorical values indicating performance status, shortness of breath, cough, weakness, diabetes, heart disease, PAD, smoking history, as well as the T stage of the tumor. The columns were renamed to their respective variable names and variables were re-coded in Python to account for the correct data type.

The target variable of 'Risk1Yr' was true if the patient did not survive so that column was changed to 'Died' to account for the correct variable matching.

A descriptive analysis was done of both the numerical and categorical variables. The mean age of the participants in the study population ranged from 21 years old to 87 years old. The mean age was 62. FEV1 mean value was 4.56 and FVC mean value was 3.28. Looking at the max value for FEV1 of 86, that is extraordinarily high and may be an error. When visualizing the categorical variables, very few patients in this study had a history of asthma, diabetes, peripheral arterial disease, or heart attack. Further, when looking at the distribution for the T staging, most people had T1 or 2 tumors which indicate smaller and earlier stage tumors. Few patients had T3 or T4 and since these are more advanced tumors, they would have a worse prognosis and likely not surgical candidates. The majority of the participants were smokers, which is logical in that those who smoke are more likely to get lung cancer.

Died is the target variable and it appears that a large number of participants survived in this population, 400 out of 470 total patients. The majority of the participants did not have pain, hemoptysis, dyspnea, or weakness. Further, the majority did not have diabetes, a heart attack, peripheral arterial disease, or asthma. The majority of the participants did have cough. The most common performance status was Zubrod Scale 1 with 0 being asymptomatic to 2 being more symptomatic. Given these demographics, it bears mentioning that the study population in this dataset have relatively healthy cancer patients without significant co-morbidities. With significant co-morbidities, survival would be expected to be worse after treatment for cancer.

The distribution of these is interesting in that these appear to be relatively low risk lung cancer patients in that they do not have a lot of other co-morbidities, a poor functional status, etc. This could explain why so many participants survived. There were significant outliers in the FVC and FEV1 variables. Most of the values are <5 and there were some that were in the double digits. I believe that these were transposed in error as FEV1 and FVC can be expressed in a numerical volume in liters as well as a percentage of predicted and there were several values that were in the 60-80 range that I suspect were put in as a percentage and not as an absolute value.

There were no highly correlated variables and while FVC and FEV1 variables were negatively skewed, they were not transformed to be as accurate as possible. Further, swarm plots were created to visualize our numerical variables against the target variable, whether the patient died or survived. There were no significant findings from this as these numerical values appeared to have the same distribution between both alive and dead participants.

Given the possible outliers, two CSV files were used with machine learning algorithms. First, the full data was used without redactions using K Means Clustering, Logistic Regression, Support Vector Machine Classifiers, KNN, and XG Boost methods. Then the same methods were used with the outliers removed from the dataset and compared. Similar to previous projects, we would want the algorithms to limit false negatives as much as possible so that high risk patients that may not survive could be identified and cared for to reduce the risk of complications and death.

Results:

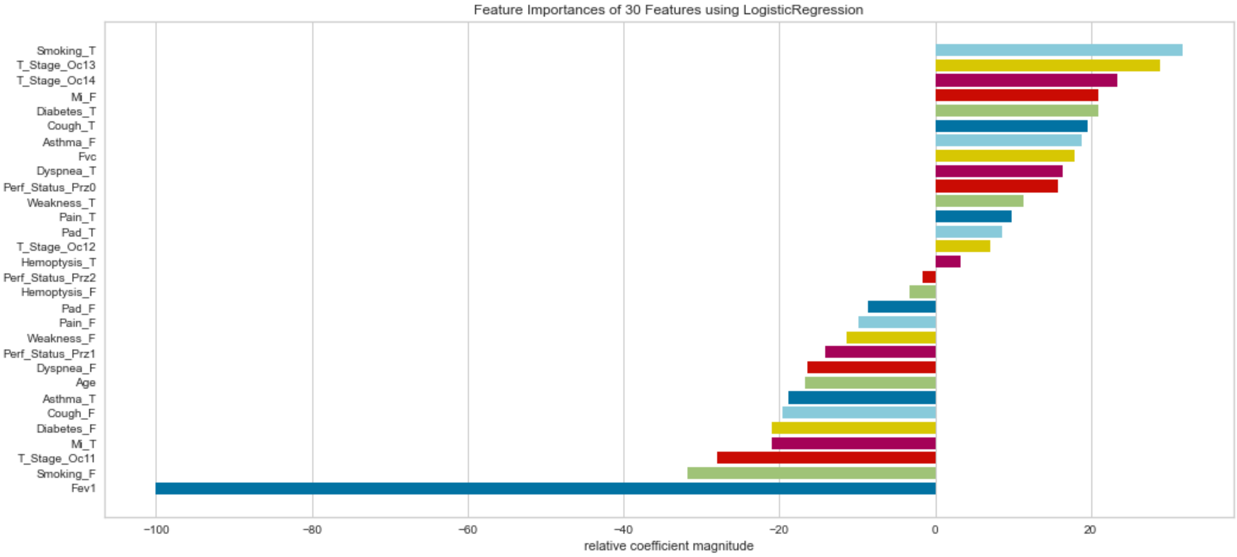
K Means Clustering:

The target variable was dropped from the dataset and using K Means Clustering with two classes (survived vs. died), the algorithm was run and then compared to the true values. The overall accuracy of this method was approximately 62%. When the outliers were removed from the dataset, the accuracy decreased to 38%.

Logistic Regression:

Logistic regression had a high F1 score of 0.789 for predicting survival. It had a comparatively low score of 0.348 when predicting death. ROC was 0.68 for both classes. There were 7 false negatives in that the model predicted survival when the person actually died. After removing the outliers from the dataset, the F1 score as well as the ROC were lower for both predicted classes.

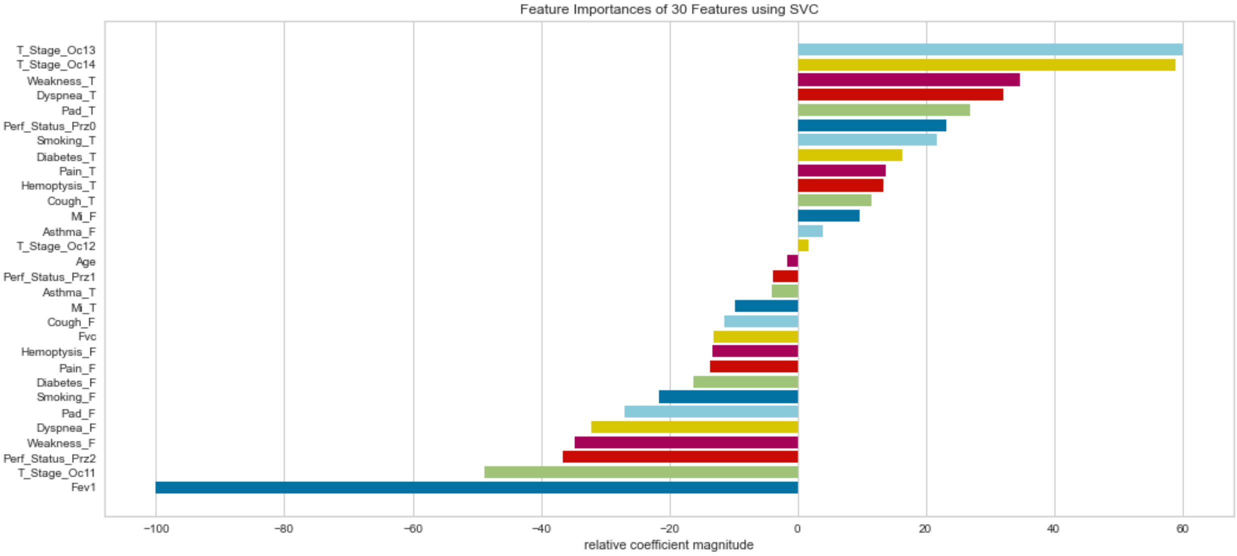
Ranking the feature importance was interesting in that it relatively weighted smokers, T3/4 tumors (higher stage), gender, diabetes, cough, and shortness of breath relatively highly. Interestingly enough, the FEV1 status was considered the least important by relative coefficients.



Support Vector Machine Classifier:

The SVC model performed similarly to the logistic regression model. However, there were 9 false negatives, which was higher than the logistic regression model. The F1 score for predicting survival was higher at 0.821 but the prediction of death was lower at 0.270. The ROC was also lower at 0.66. After removing the outliers, the SVC model also performed worse with lower F1 scores for both classes as well as the ROC.

The SVC model agreed with the logistic regression model in that the T3/4 higher stage tumors were relatively more important in addition to shortness of breath, smoking, and weakness. The FEV1 was considered the least important measure.



KNN:

Using KNN and after standardizing the features, it had a high F1 score at predicting survival at 0.90. Even with tuned hyperparameters with the number of neighbors at 11, p of 1, leaf size of 1, the minkowski metric, and uniform weights, it had a 0 score for predicting death. This would be the worst performing model since the point of the model would be to predict risk of death, not risk of survival. As in the other models, removing the outliers here also resulted in worsened F1 scores for both predicted classes.

XG Boost:

The XG Boost ensemble method classifier had a high F1 score of 0.90 for predicting survival with a 0 score for F1. ROC was 0.62 and performed similarly to the KNN model. Interestingly, when removing the outliers, the F1 score for predicting survival was reduced to 0.883 but the F1 score for predicting death was above 0 minimally at 0.095. The ROC also improved.

XG Boost weighted the worse performance status of 2, T stage, absence of smoking, FVC, the absence of cough, shortness of breath, and pain as relatively more important.

Conclusions:

This dataset was challenging in a multitude of ways. The file format was unusual, and the variables and columns had to be renamed to easily identify which predictors were being identified. Further, for machine learning modelling, the target variable was switched to a numerical Boolean for easier processing.

The main issue with this dataset is that this dataset had a significantly higher survived population than the deceased population. With an imbalanced target class, the algorithms will usually predict the higher proportion class since it is more likely to be correct. With class balancing, this did improve somewhat but all F1 scores for predicting death were lower than 0.3.

The purpose of this data is to develop an algorithm that would predict those at high risk of complications or death. That way, there could be further interventions or intensification of care and monitoring to help prevent that outcome. None of the machine learning models here performed very well with predicting what we want to predict. Overall, the logistic regression did the best at prediction with the highest F1 scores for both predicted survival and death though it was poorly accurate at predicting death. The point of the project is to identify high risk patients so that clinical care could be changed to help improve outcomes more favorably.

It is remarkably interesting that many of the methods seemed to pick out the same variables that are medically well-known to correlate with survival, notably the performance status and T stage of the tumor. This suggests that machine learning can pick up on logical data trends that are logically known, but the algorithm picked this out statistically without outside knowledge of the background of the problem. For a future project, we would require a much larger dataset and a much more varied patient population in order to have highly predictive algorithms.

References Cited:

National Cancer Institute. (2015, March 9). Cancer Staging. Retrieved October 24, 2020, from https://www.cancer.gov/about-cancer/diagnosis-staging/staging

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Howlader N, Forjaz G, Mooradian MJ, Meza R, Kong CY, Cronin KA, Mariotto AB, Lowy DR, Feuer EJ. The Effect of Advances in Lung-Cancer Treatment on Population Mortality. N Engl J Med. 2020 Aug 13;383(7):640-649. doi: 10.1056/NEJMoa1916623. PMID: 32786189.

Midthun, D. E., MD. (2020). Overview of the initial treatment and prognosis of lung cancer. Retrieved October 24, 2020, from https://www.uptodate.com/contents/overview-of-the-initial-treatment-and-prognosis-of-lung-cancer?search=lung+cancer

Pairolero, P. C., Williams, D. E., Bergstralh, E. J., Piehler, J. M., Bernatz, P. E., & Payne, W. S. (1984). Postsurgical Stage I Bronchogenic Carcinoma: Morbid Implications of Recurrent Disease. *The Annals of Thoracic Surgery, 38*(4), 331-338. doi:10.1016/s0003-4975(10)62281-3

Sculier JP, Chansky K, Crowley JJ, Van Meerbeeck J, Goldstraw P; International Staging Committee and Participating Institutions. The impact of additional prognostic factors on survival and their relationship with the anatomical extent of disease expressed by the 6th Edition of the TNM Classification of Malignant Tumors and the proposals for the 7th Edition. J Thorac Oncol. 2008 May;3(5):457-66. doi: 10.1097/JTO.0b013e31816de2b8. PMID: 18448996.

Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020 Jan;70(1):7-30. doi: 10.3322/caac.21590. Epub 2020 Jan 8. PMID: 31912902.

West, H., & Jin, J. O. (2015). Performance Status in Patients With Cancer. *JAMA Oncology, 1*(7), 998. doi:10.1001/jamaoncol.2015.3113

Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (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, 14*, 99-108. doi:10.1016/j.asoc.2013.07.016

Appendix: Variable Description

1. **DGN:** Diagnosis (based on ICD codes)
2. **PRE4:** Forced Vital Capacity -FVC
3. **PRE5:** Volume of Air Exhaled in 1 Second - FEV1
4. **PRE6:** Performance Status (Zubrod Scale) – 3 Values (PRZ0, PRZ1, PRZ2)
5. **PRE7:** Pain Before Surgery (T/F)
6. **PRE8:** Hemoptysis Before Surgery (T/F)
7. **PRE9:** Dyspnea Before Surgery (T/F)
8. **PRE10:** Cough Before Surgery (T/F)
9. **PRE11:** Weakness Before Surgery (T/F)
10. **PRE14:** T in Cancer Stage – Smallest to Largest (OC11, OC12, OC13, OC14)
11. **PRE17:** Type 2 Diabetes – (T/F)
12. **PRE19:** MI w/in 6 Months – (T/F)
13. **PRE25:** Peripheral Arterial Disease – (T/F)
14. **PRE30:** Smoking – (T/F)
15. **PRE32:** Asthma – (T/F)
16. **AGE:** Age at Time of Surgery
17. **Risk1Y:** Survival After 1 Year (T if Died)