Monica Luo (ML8110)
Data Bootcamp - Fall 2024
Final Project Writeup
December 16, 2024

Cancer is a complex and evolving disease that scientists have been studying for centuries. The number of lung cancer patients has been steadily rising, and the illness itself continually evolves. For my final project, I explored a predictive problem: survival after lung cancer surgery. My goal was to develop a predictive model that could accurately forecast, based on pre-surgery symptoms, whether a patient would pass away within one year post-surgery, effectively resulting in a binary outcome of 1. Information about patient survival within one year is especially crucial, as it offers a clearer understanding of whether undergoing surgery is worthwhile both financially and emotionally. Medical care is costly in any country, whether financed through taxes or personal savings. Thus, from an economic perspective, it's important to allocate resources efficiently. If a patient is unlikely to be saved by the surgery, it may be more beneficial for them to enjoy their remaining life without financial burden, while also ensuring that public funds are used effectively. This project did not find an ideal model due to the imbalance in the dataset. However, the best models for predicting those who would pass away within one year post-surgery were the Random Forest Classifier and XGBoost, based on their F1 scores and Log Loss metrics.

The Dataset:

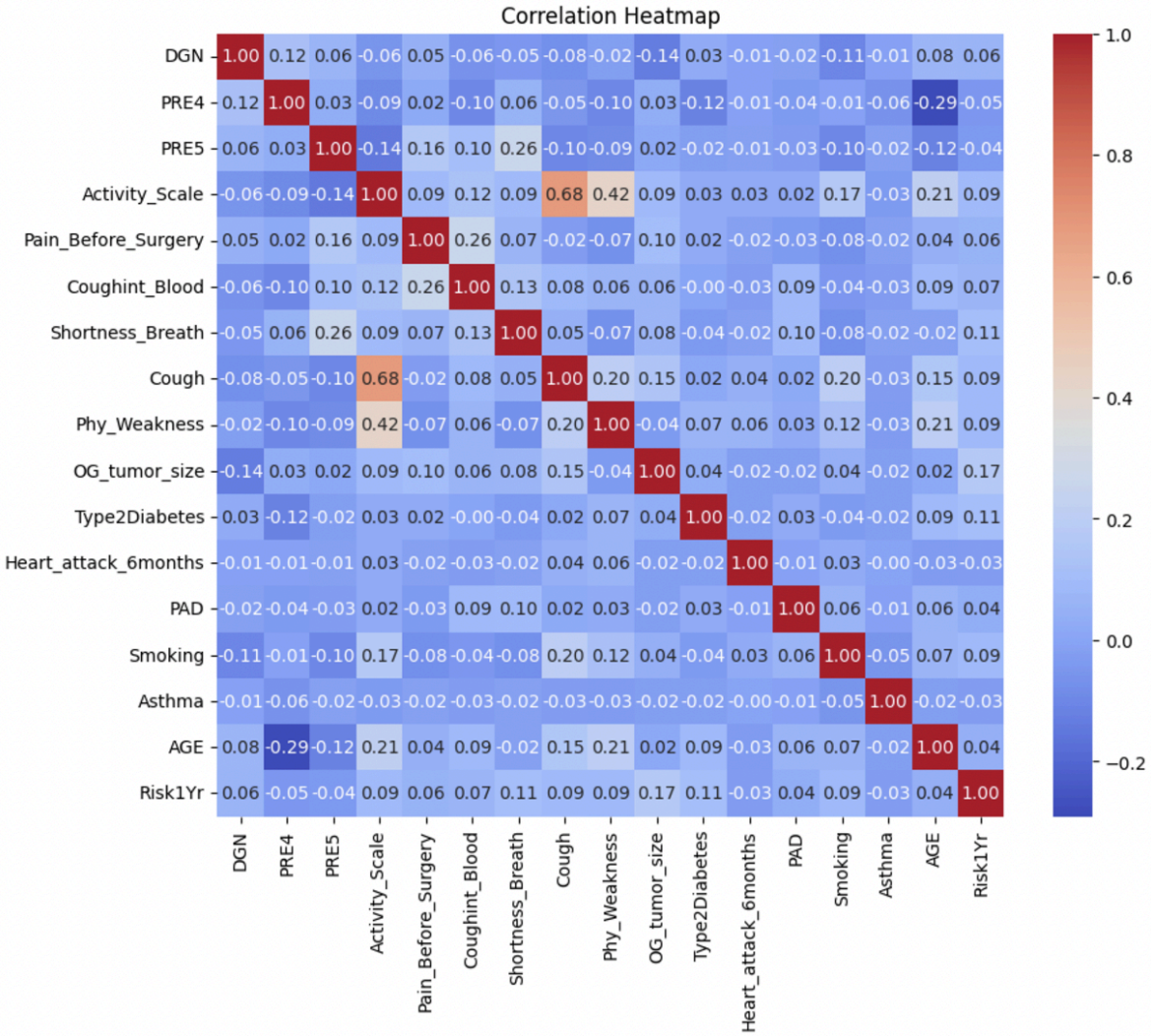
My dataset is provided publically by University of California, Irvine's website, while the data itself was collected at the Wroclaw Thoracic Surgery Center for patients who underwent major lung resections for primary lung cancer from 2007 to 2001. This data is also a part of the National Lung Cancer Registry in Warsaw, Poland. Each datapoint had 16 features, from age to the side of their tumor, and an outcome of living 1 year post surgery or not. The table below describes each factor within a datapoint.

Variable Name	Type	Description	Answers Possible
DGN	Categorical	Diagnosis - type and severity of tumor(s)	1 to 8, higher number means more complicated diagnosis
PRE4	Numerical (continuous)	Forced Vital Capacity	1.44 to 6.3
PRE5	Numerical (continuous)	Volume exhaled at end of first second of forced expiration	0.96 to 86.3
Activity Scale	Categorical	Level of activity they were capable of alone	0 to 2, higher level meant less activity

		before surgery	possible alone
Pain Before Surgery	Binary	Felt pains in their body before surgery	1 = True, 0 = False
Coughing	Binary	Coughed regularly before surgery	1 = True, 0 = False
Shortness Breath	Binary	Often experienced shortness of before surgery	1 = True, 0 = False
Physical Weakness	Binary	Weakness before surgery	1 = True, 0 = False
Original Tumor Size	Categorical	Size of Tumor the surgery was focused on	11 to 14 (no units provided, but larger number meant bigger tumor)
Type 2 Diabetes	Binary	Is the patients diagnosed with Type 2 Diabetes	1 = True, 0 = False
MI	Binary	Has had a heart attack within the last 6 months	1 = True, 0 = False
PAD	Binary	Has Peripheral arterial diseases	1 = True, 0 = False
Haemoptysis	Binary	Coughing blood before surgery	1 = True, 0 = False
Smoking	Binary	Has habit of smoking before surgery	1 = True, 0 = False
Asthma	Binary	Has been diagnosed with Asthma	1 = True, 0 = False
Age	Numerical (Integer)	How long ago was they patient born in the units of years	21 to 87 years old
Risk1Yr	Binary	If the patient was still alive 1 year post surgery	1 = True, 0 = False

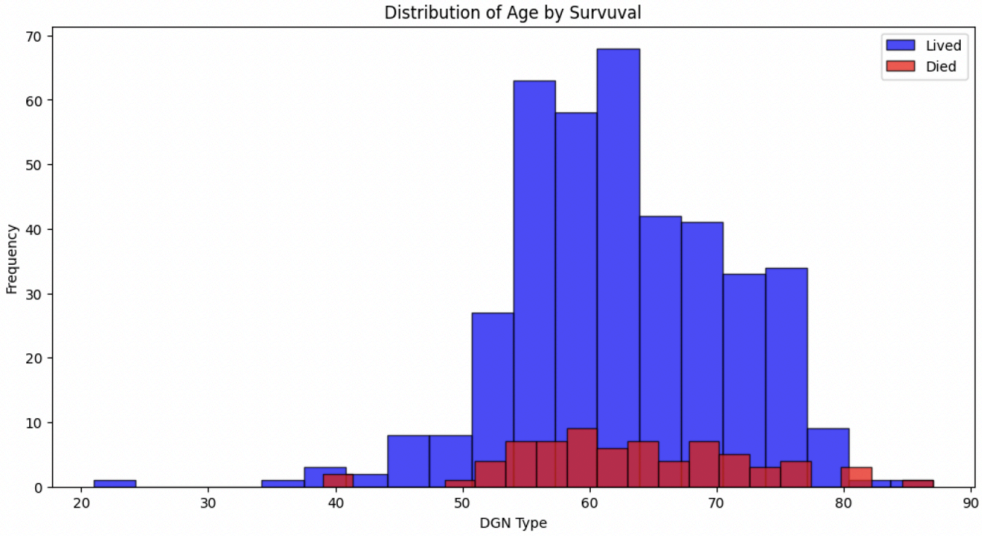
Exploratory Data Analysis:

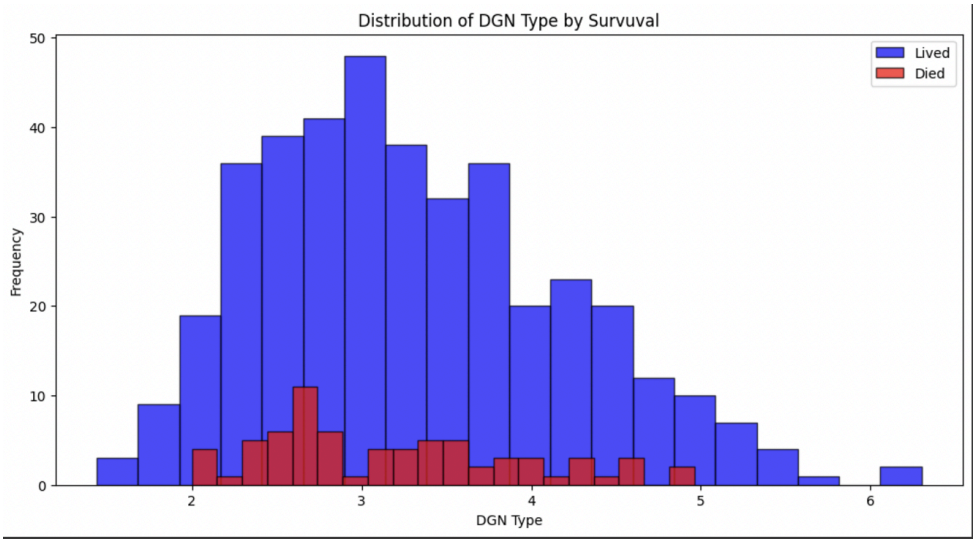
To easily see if there are any correlation within the whole dataset, I first build a correlation heatmap for the data presented below. As the map was mostly filled with varying versions of blue, it seems that the whole dataset did not have many strong correlations. The only correlations we could point out from the map would be between the factors of "Cough" and "Activity Scale" as well as "Physical Weakness" and "Activity Scale". I only proceeded to shallowly look into this correlation as the main correlation I was seeking for was between any of the features and the outcome in "Risk1Yr". From the value counts alone, there seemed to be a trend of greater proportion of people who were on the lower levels of independent activity scale who coughed or experienced physical weakness.



Next, I looked into the split of the binary data and found out that within the dataset, only 14.89% of patients passed away within 1 year post surgery, and the most prevalent symptoms in patients were cough and smoking. As most patients survived surgery, It shows that this dataset is

not as balanced as preferred. As for the other features, the average patient was around 62 years old, with a diagnosis of level 3, activity level of 1, and a tumor size close to the ranking of 12. As for patients who did not live past 1 year after the surgery, the most prevalent symptoms were still coughing and smoking, but none of the patients were diagnosed with Asthma. The average patient who passed also has a diagnosis level around 3, an activity level around 1, a tumor size with a ranking of 12, and at a slightly higher age of 63. Two features that stood out to seem to have a more significant difference was PRE5, or how much air a patient was able to exhale in one go. Patients who did not live past 1 year post surgery had a lower score than the average of the whole dataset and those who survived past 1 year post surgery. Along with Age, I predict that the value of PRE5 may play as a main factor in classifying outcomes by models as these are the only 2 features showing some difference in patients who passed away and survived.





Models:

Chosen Evaluation Metric: F1 Scores and Log Loss

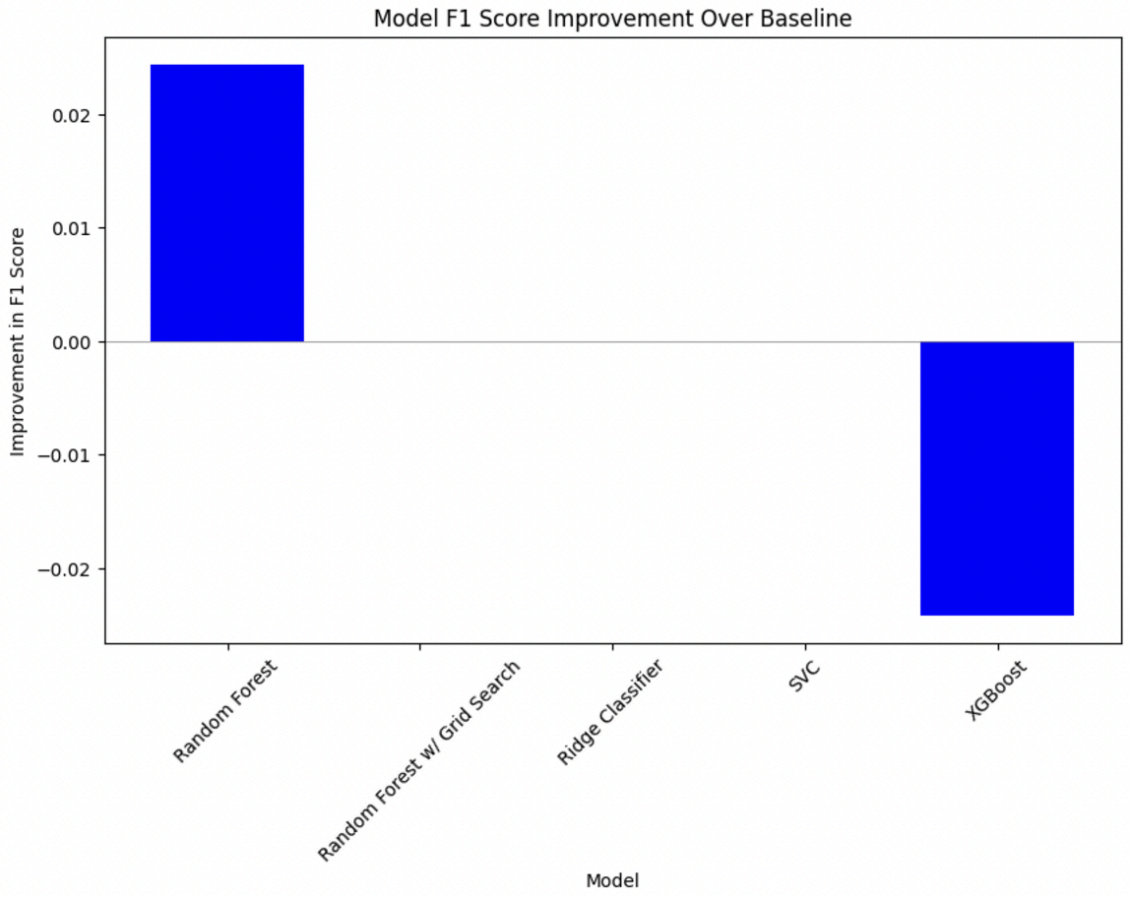
For some models I have chosen the use, the traditional Accuracy and Precision scores were not applicable as the models were complicated and unable to produce a score for the combination of features presented. To analyze all the models on the same scale, we will mainly look at F1 Scores and Log Loss. F1 scores represent a balance in precision and recall, 2 important metrics in classification models, in each model while Log Loss measures how well each model predicted the correct outcome. The goal is to earn higher F1 scores and lower Log Loss scores.

1. Baseline Model: To have a value of accuracy to compare to, I chose the baseline model to be the situation where we always choose 0. In this case, the accuracy percentage will be 85.11%, and with the predictive classifier models we use below, our goal is to get an accuracy score over 85.11%. As for F1 score in my baseline case, we are trying to beat 78.26%
2. Random Forest Classifier: In terms of making the correct diagnosis for if a patient survives 1 year post surgery, the Random Forest Classifier outcomes were not ideal. As my F1 score was .8069, which is relatively high, the F1 score for predicting the minor of those who passed is very low at .13. This model also exhibited a low recall score of 0.07 in identifying 1, or those who passed, meaning the model was not the best for identifying this class of patients.
3. Random Forest Classifier with Grid Search: In an attempt to build upon the Random Forest Classifier model to improve outcomes, I applied Grid Search parameters in order to identify the best parameters and build a better model. Unfortunately the model did not perform better, as the F1 score decreased from 0.8069 to 0.7826, the same F1 score as the baseline model. The model with Grid Searched performed even worse at identifying patients who passed from with an F1 score and recall score of 0 respectively. Although we would expect models to perform better with Grid Search, depending on the dataset it may perform worse, especially for imbalance datasets like the one I have here.
4. Ridge Classifier: In terms of evaluation metrics, the Ridge Classifier performed the same as the Random Forest Classifier with Grid search, with the same overall F1 score of 78.26%, and 0% recall and F1 score in identifying those who passed within 1 year post surgery.
5. SVC Model: Just as the Ridge Classifier did, the SVC model has the exact same statistics as the Baseline model, Random Forest Classifier with Grid Search, and Ridge Classifier models. This model is not ideal in my case of trying to identify patients who passed within a year post surgery, but along with the other classifiers mentioned above, would be great at pointing out patients who will live past a year post surgery.
6. XG Boost Classifier Model: This model returned more interesting stats. Compared to the baseline model, this model had an ever so slightly lower F1 score overall at 75.84%, but the F1 score in predicting the minority of those who passed away within 1 year post surgery saw an improvement to 15% and has a recall score higher at 14%. On the other hand, compared to all the other models, XG Boost was not as great at identifying people

who survived 1 year post surgery as the F1 score in identifying the majority was 86% and a lower recall score of 88%, where in the other models the F1 score for identifying majority was 92% with a 0% recall score.

Overall:

Model	Accuracy	Overall F1 Score	Log Loss
Baseline	85.11%	78.25%	5.37
Random Forest Classifier	86.17%	80.69%	4.99
Random Forest Classifier with Grid Search	85.11%	78.26%	5.37
Ridge Classifier	85.11%	78.26%	5.37
SVC	85.11%	78.26%	5.37
XG Boost Classifier	N/A	75.84	0.67



From all the models, the best 2 outcomes were from Random Forest Classifiers without Grid Search and XG Boost Classifiers. In terms of log loss, which shows how well the predictions align with actual outcomes, XG Boost Classifiers did the best with 0.66 log loss score, as it was the model best able to predict which patients passed away within 1 year after surgery. On the other hand, compared to all the other models, XGBoost Classifier was decent at predicting if patients survived 1 year post surgery, but scored slightly lower on this end of predicting the majority compared to the other models. In terms of accuracy and overall F1 scores, Random Forest Classifiers scored the highest at .8069, even higher than the baseline model, and the second best at identifying those who passed while maintaining a high level of accuracy in predicting those who survived 1 year post surgery, and therefore earned the second lowest log loss score at 4.99

Conclusion and Next Steps:

1. Real World Applications: Cancer remains a complex and evolving disease, making it challenging to predict survival outcomes with absolute accuracy even when we use today's technologies. There are many factors besides symptoms and correlating illnesses that determine if someone does well post surgery, from other major public health events like flu and COVID-19, their everyday habits and diet, to how they view their cancer diagnosis itself, which affects their psychological state. Although this data was less than ideal in a statistical and data analysis way as the outcomes were very unbalanced, it's comforting to know that most patients survived post surgery back in 2007 to 2011. In terms of using this dataset to analyze factors that can contribute to predicting if a patient will survive 1 year post surgery, it may not be the best dataset to examine as we have minimal data about those who do not survive.
2. Next Steps and Suggestions for Further Research: There are many steps and actions we can use to both improve the dataset and model for future research. Some of them include widening the dataset to get more information about patients who pass away within a year post surgery, this will allow the models to have more data to work with in learning patterns within the data. We can also add more features and symptoms to find new patterns. One common pattern I have seen with those around me with lung cancer is that many of them have Chronic obstructive pulmonary disease (COPD). Other factors we can consider would be how much they smoked per day, their diet, height, weight, gender, weight loss since diagnosis, and if this is a recurring condition. It may help to remove unnecessary features, like in the current case none of the patients who passes away within a year post surgery was diagnosed with asthma and therefore could be seen by some as an unnecessary feature. Further data collection may help with a major limitation in the project, data imbalance. Another option can be making the data available about the majority outcome, surviving 1 year post surgery, being a small proportion of the data by randomly selecting datapoints to take away and not evaluating.

As for the models, we can always try with different models, but another option can be to combine models for predictive uses. The main reason I did not attempt to combine models in this project was because I do not have the technical skills to do so yet. Another more feasible option would be to apply grid search to other models, as well as better list hyperparameters for grid search implication. The results of choosing the best features in my use of Grid Search in Random Forest Classifiers was pretty lackluster, as some of them were unable to find the best parameter. This could have been from the imbalance data set to my choice of test parameters.

Overall, lung cancer and survival is a hard topic to predict, shown in both real life and through models. As survival can depend on many factors, and some are not viable to the public, such as psychological state, classifier models may not yet be useful enough to help medical professionals and patients assess if the patient will survive long after surgery and in the end see if the surgery is worth the time, energy, and, depending where you are, money. Although my Random Forest Classifier and XG Boost Classifier models had moderately high F1 scores, they were way better at predicting if patients will survive than will pass, they were still not at a level where many would use to make decisions with. For future research, we can always expand the dataset to collect more information about patients who did pass in an attempt to make a more balanced dataset to define our models with more details by either implementing Grid Search on more models, or combining models to make predictions.

If you would like to interact with my models, feel free to do so using the Streamlit link below: https://data-bootcamp-final-edahphmxdojryjo2wtqmmy.streamlit.app/

Sources:

1. Lubicz, Marek, et al. "Thoracic Surgery Data." UCI Machine Learning Repository, 2014, https://doi.org/10.24432/C5Z60N.