Predicting Lung Cancer Survival Time

Pooja Patel, Christopher Wilhite, Eli Parker

Problem Statement

In this study, we attempt to predict the survival time of a patient (remaining days to live) from one three-dimensional CT scan (grayscale image) and a set of pre-extracted quantitative imaging features, as well as clinical data.

The project is based on a supervised survival prediction problem, with the endpoint being a model that can accurately predict survival time, considering censorship and other relevant factors.

Methods

Random Forest Regression – supervised learning model that uses multiple decision trees to output the optimal result.

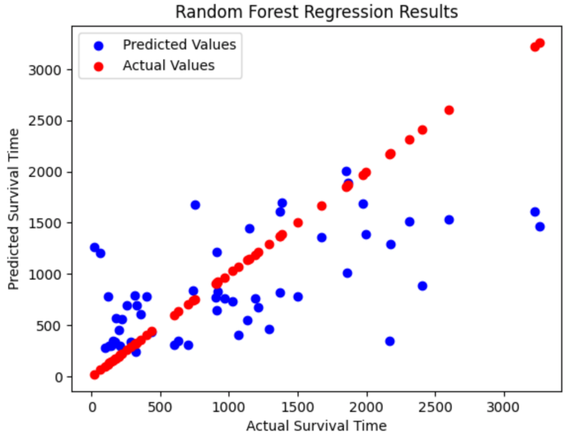
Linear Regression – a supervised learning model that draws a best line of fit of the relationship between the X and y variables.

Kaplan Meier – supervised learning model the shows survival over time. This is the main model for our project.

Libraries include: matplotlib, pandas, seaborn, numpy, lifelines for the Kaplan Meier, sklearn for training test splits, feature scaling, and models

Results

Inside the lungs: A CT scan image provides a detailed look at the structures and tissues of the lungs, allowing for the detection and diagnosis of various respiratory conditions.



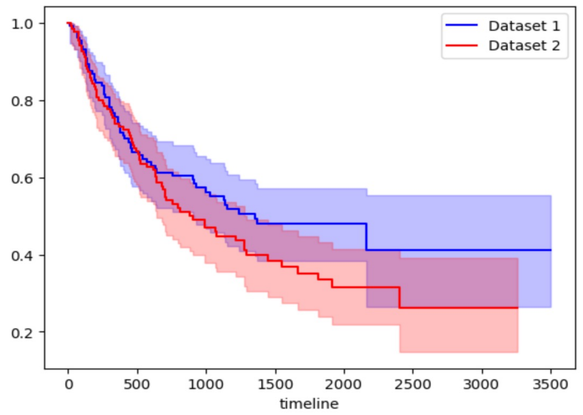

```
Confidence Interval for dataset 1:
       KM_estimate_lower_0.95  KM_estimate_upper_0.95
0.0                  1.000000                1.000000
14.0                 0.947832                0.998937
20.0                 0.947832                0.998937
21.0                 0.947832                0.998937
33.0                 0.940768                0.996189
...                       ...                     ...
2309.0               0.263507                0.554256
2330.0               0.263507                0.554256
2600.0               0.263507                0.554256
3078.0               0.263507                0.554256
3500.0               0.263507                0.554256

[129 rows x 2 columns]
Confidence Interval for dataset 2:
       KM_estimate_lower_0.95  KM_estimate_upper_0.95
0.0                  1.000000                1.000000
25.0                 0.948212                0.998945
31.0                 0.941643                0.996246
43.0                 0.932202                0.992724
77.0                 0.922425                0.988691
...                       ...                     ...
2515.0               0.148192                0.391574
2528.0               0.148192                0.391574
3222.0               0.148192                0.391574
3251.0               0.148192                0.391574
3259.0               0.148192                0.391574

[133 rows x 2 columns]
```

Confidence in survival analysis: Kaplan-Meier tables provide a range of possible outcomes and their confidence intervals, giving researchers a more complete picture of the likelihood of survival over time.

Model	Accuracy
Random Forest Regression	**0.36**
Naive Linear Regression	**0.15**

Random Forest proved to be more accurate than the Linear Regression model.

Survival Curve for the Kapan Meier.



Plotting the course of survival: The Kaplan-Meier curve visually represents the probability of survival over time for a group of subjects with a particular condition or treatment

```
Log-rank test results:
   test_statistic        p  -log2(p)
0        2.077145  0.14952  2.741593
p-value: 0.150
test statistic: 2.077
```

Significance

- This project is beneficial for anyone who is or directly deals with a lung cancer patient.
- This has a large impact because it can be used to determine the best type of treatment through patterns allowing medical personnel to predict treatment reactions.
- Better solutions to the treatment of Lung Cancer could lead to a higher quality of life for the patients as well as a longer life expectancy

Future Plans

- After the first iteration and review of the results of this project, we realized that we had not used proper models to show our results and predict future survival rates.
- When we went back over the information to review our results it became clear to us that we needed to add to and improve upon the models we used in this study.
- Since the endpoint predicts the patient's survival time and the censorship, we need to try more models.
- We plan to try different regression models to get the survival time of patients
- We also plan to try different classification models to get the censorship(Event) of the patients.

College of Engineering
TENNESSEE TECH

Acknowledgments

This project was given to us by our advisor, Dr. Eberle for our Advanced Data Science and Applications course. The project idea itself was proposed by the Owkin company with challenge data from MathA. We would like to thank Dr. Doug Talbert for providing us with guidance on what the next steps for this project would be.

References

- Sklearn.org
- pyradiomics.readthedocs.io/en/2.0.1/features.html
- challengedata.ens.fr/participants/challenges/33/
- www.cancer.org/treatment/understanding-your-diagnosis/staging.html
- https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8
- Deepnote.com for our notebook.