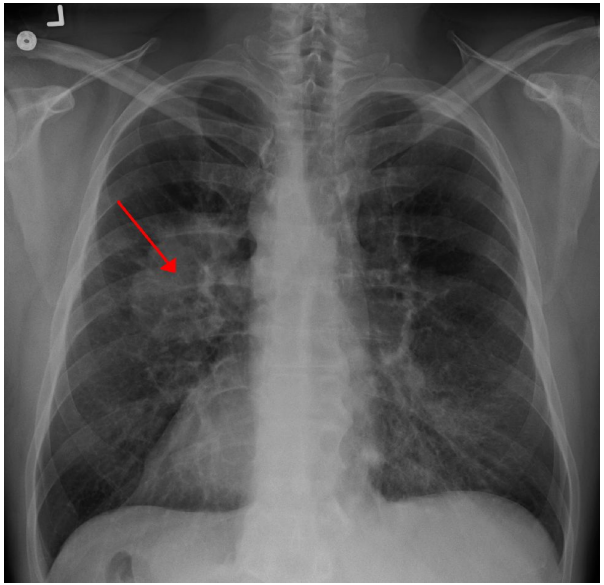
Survival Analysis of Lung Cancer Data for Non-Invasive Cancer Detection

Alan Wu
Faculty Mentor: Dr. Subhajyoti De

Problem Statement

Non-invasively Detect Cancer

- Cancer is a leading cause of death in the United States
- Detecting cancer can be invasive
  - CT Scan
  - MRI
- Modern Techniques: Liquid biopsy and DNA sequencing



Goal: Predict Survival Times of Lung Cancer Patients

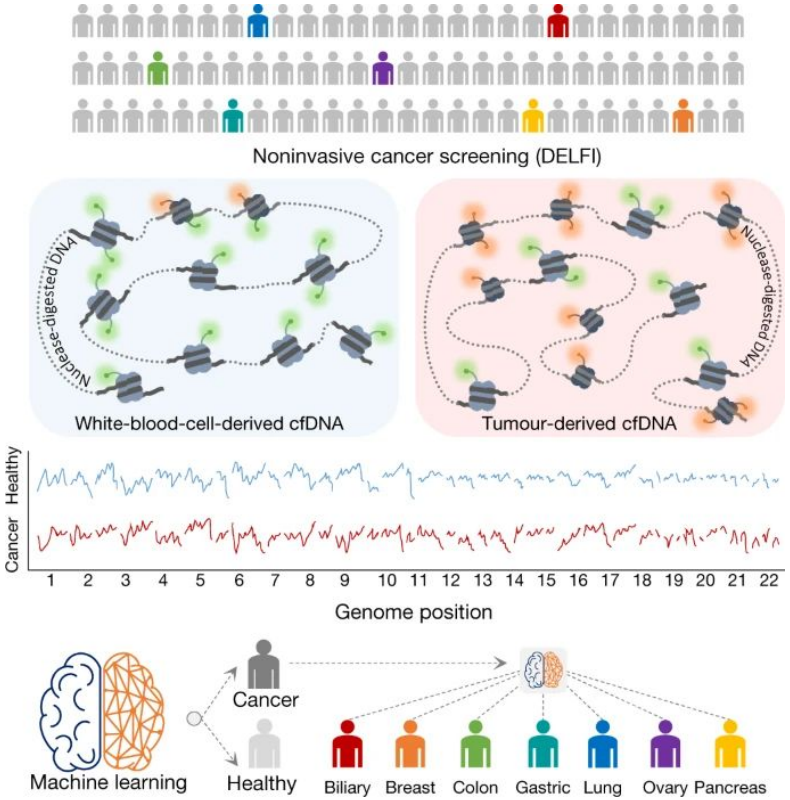
1. What are the significant features that attribute the most to predicting the survival time of a patient who already has cancer?
2. Can we find a model that predicts the survival time of patients well based on the data?

Background Work

1. Data: LUCAS Lung Cancer Data
    a. 95 patients with lung cancer
    b. Days alive
    c. Survival status
    d. Stages i-iv
    e. Delfi score
    f. Treatments: no treatment, surgery, palliative chemotherapy/surgery, surgery + adjuvant treatment, chemotherapy+surgery, etc.
2. Survival Analysis Techniques
    a. Some samples are censored (the event that has been designated the ‘target event’ has not occurred for a certain subset of samples)
    b. Cannot generalize traditional ML and statistical algorithms to this data
    c. Methods: survival ML algorithms, survival curves, hazard function

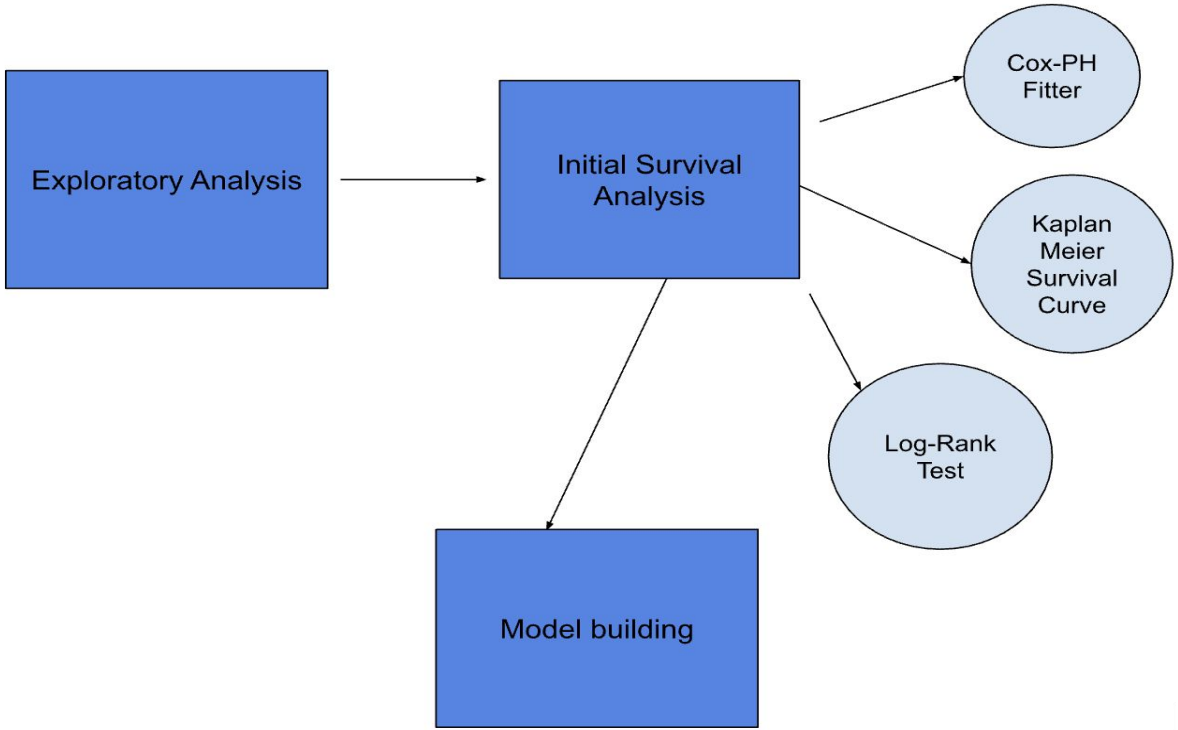
DELFI Score

- DELFI (DNA evaluation of fragments for early interception)
  - Greater alteration patterns = greater sensitivity to changes in alterations
  - If the tumor is very aggressive, there are lots of molecular features in the liquid biopsy
  - Maximize probability of detecting changes in the fragmentation profile of cfDNA (cell-free DNA) by detecting larger numbers of alterations in genome-wide analysis.



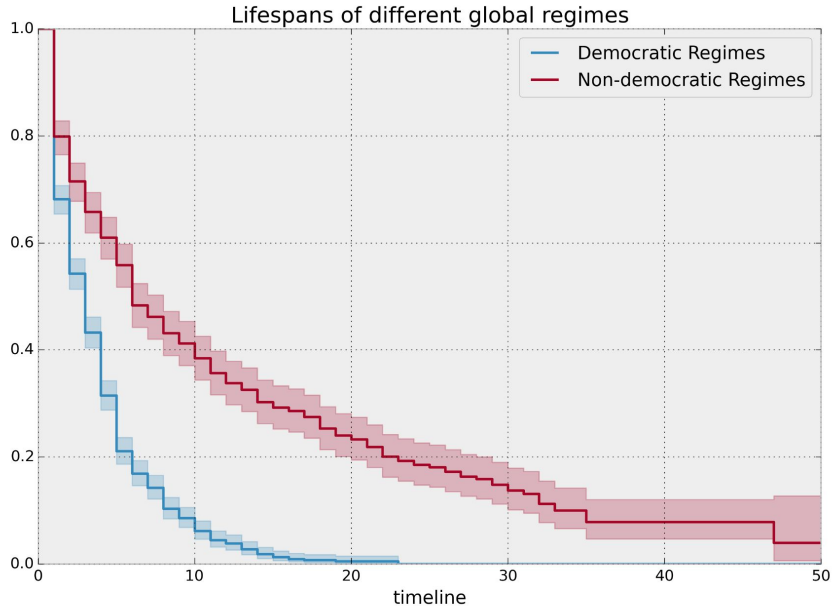
Methodology

Workflow


Exploratory Analysis
Initial Survival Analysis
Cox-PH Fitter
Kaplan Meier Survival Curve
Log-Rank Test
Model building

Significance Testing

- Kaplan Meier Fitter
  - Non parametric method (fit survival curve of population)
- Log-Rank Test
  - Hypothesis test difference between two or more groups
- Cox-Proportional Hazards Model
  - Multivariate analysis



Model Building Workflow

- Data preprocessing
  - One-hot encoding
- Split data using equal partitions of censored and uncensored data
- Fit an initial model
- Tune hyperparameters (regularization, etc)
- Cross-validate model using StratifiedKFold CV to guarantee proportional maintenance
  - For models that do not directly predict survival time, predict the survival function and take the median survival time as the given survival time
- Fit model on test set

Models + Metrics

Worked with the following survival-based algorithms:

1. Regularized Cox-PH Models (LASSO, Ridge, ElasticNet)
2. Linear Accelerated Failure Time Model
3. Survival SVM
4. RSF (Random Survival Forests)

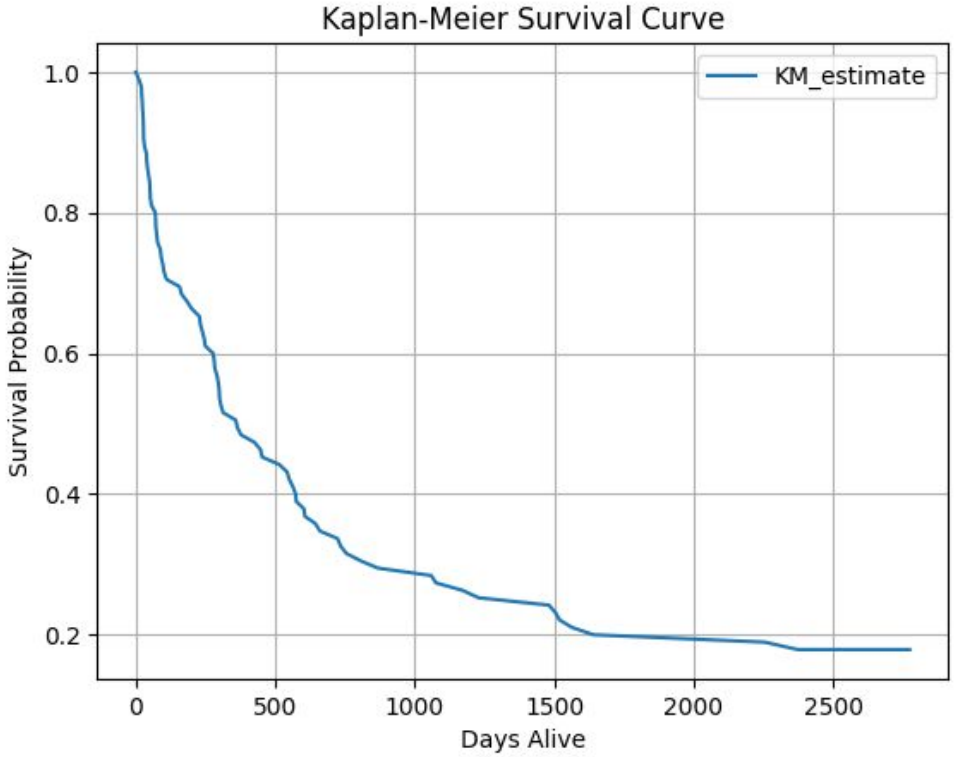
Metrics:

1. Adjusted weighted and log MAE (Mean absolute error)
2. C-Index (Concordance Index)

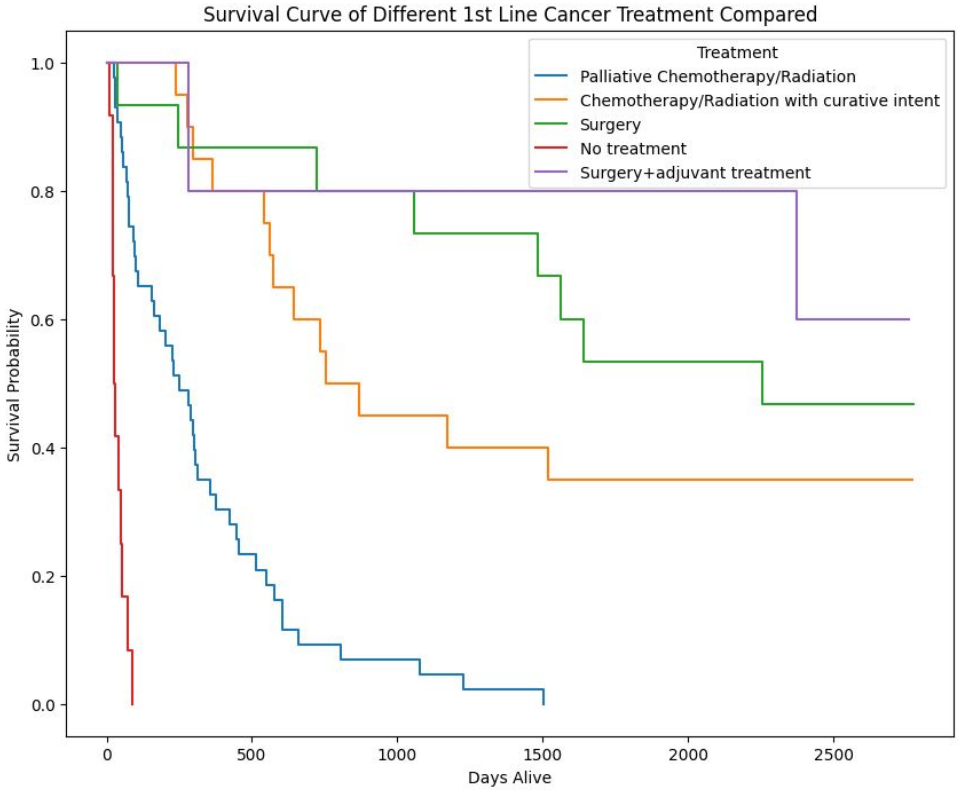
Results

Survival Function

- Survival function gives information about the survival time of the entire population in data.
- Median survival time: 364 days



Treatment Significance

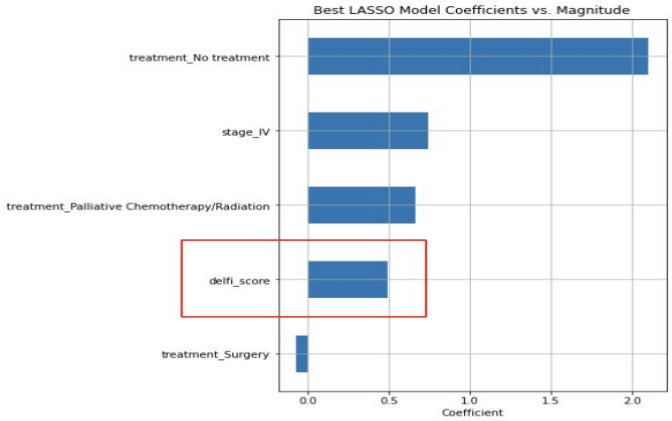


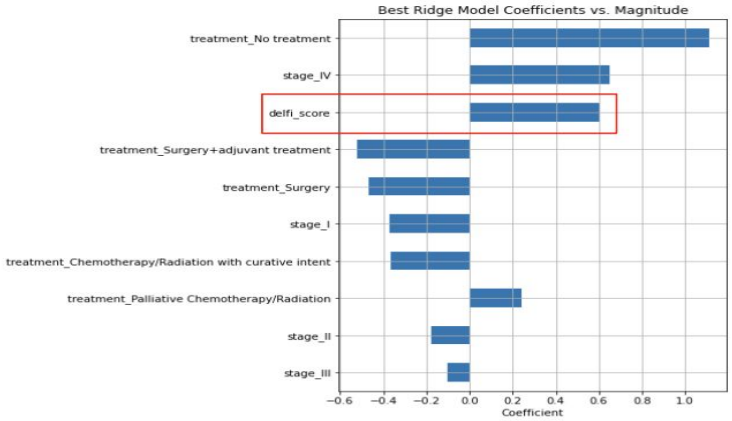
Simpler is Better

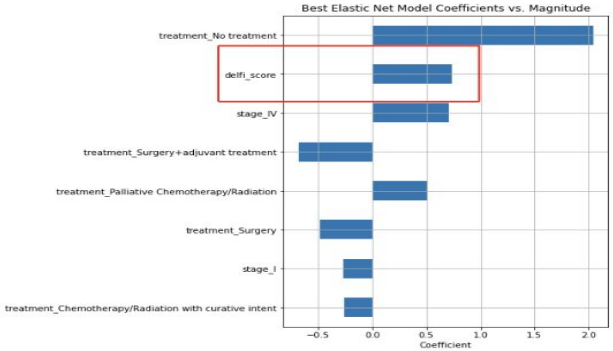
- Linear models (Cox PH) and Accelerated Failure Time on average performed with better MAE and C-Index than more complex models
- Could be attributed to low sample size and significance of the features

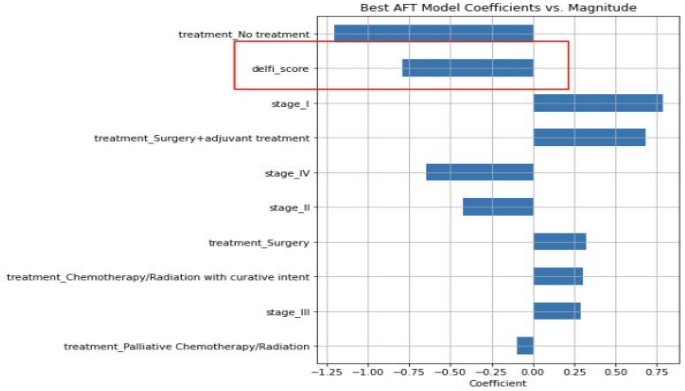
Model	Weighted MAE	C-Index	Log MAE
Kaplan Meier Fitter	769.704	0.500	1.378
Ridge Regularized Cox-PH	557.213	0.813	0.899
Lasso Regularized Cox-PH	581.493	0.825	0.910
ElasticNet Regularized Cox-PH	**557.951**	**0.828**	**0.875**
AFT	554.073	0.803	0.921
Survival SVM	**483.828**	**0.831**	**0.838**
RSF (Random Survival Forest)	592.314	0.819	0.902

Feature Inspection









Biological and Computational Conclusions

What do our results mean?

Biologically

1. Stage is consistent with what we already know about cancer
2. Treatments vs. No Treatment expected
3. DELFI is a way to measure cancer related signals and can also indicate survival

Computationally

1. Complex survival models may not be the best solution when working with small datasets
2. Regularization makes a massive difference in performance of algorithms

References

- Cristiano, S., Leal, A., Phallen, J. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385–389 (2019). https://doi.org/10.1038/s41586-019-1272-6
- Clark, T., Bradburn, M., Love, S. et al. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer 89, 232–238 (2003). https://doi.org/10.1038/sj.bjc.6601118