A Study of Machine Learning in Survival Analysis

Abel Villanueva

University of North Carolina at Charlotte

Fall 2024

Outline

Introduction

Survival Data

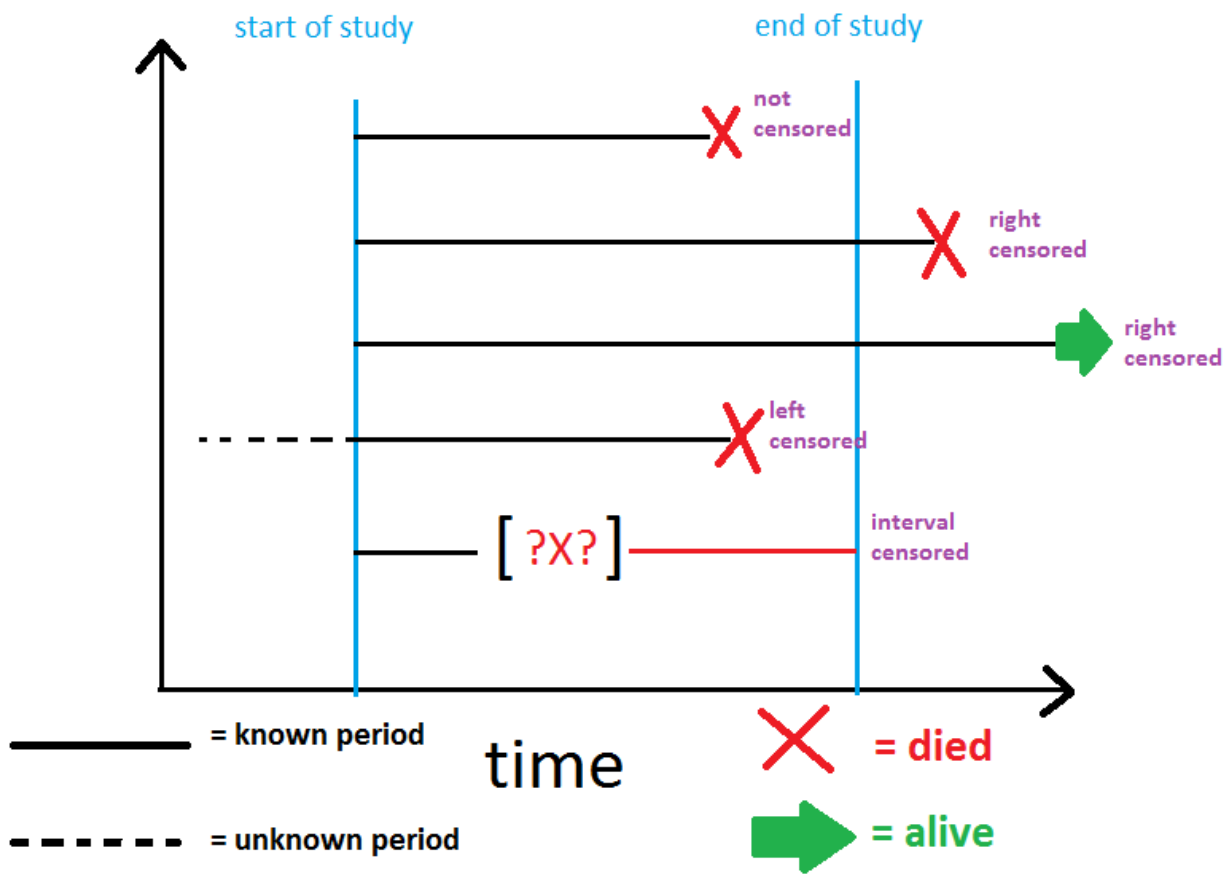
For each instance i, we observe (X_i, y_i, δ_i)

- X_i: Feature Vector
- y_i: Observed Time
- δ_i: Event Indicator

$$y_i = \begin{cases} T_i & \delta_i = 1 \\ C_i & \delta_i = 0 \end{cases} \qquad \begin{array}{l} T_i\text{: True Time} \\ C_i\text{: Censoring Time} \end{array}$$

Applications: Healthcare, Manufacturing, Finance
[Wang et al., 2019]

Censoring


start of study
end of study
not censored
right censored
right censored
left censored
[?X?]
interval censored
= known period
= unknown period
time
= died
= alive

Functions

Survival Function

$$S(t) = P(T \geq t) = \exp[-H(t)]$$

Hazard Function

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}[\ln S(t)] \qquad H(t) = \int_0^t h(u)du$$

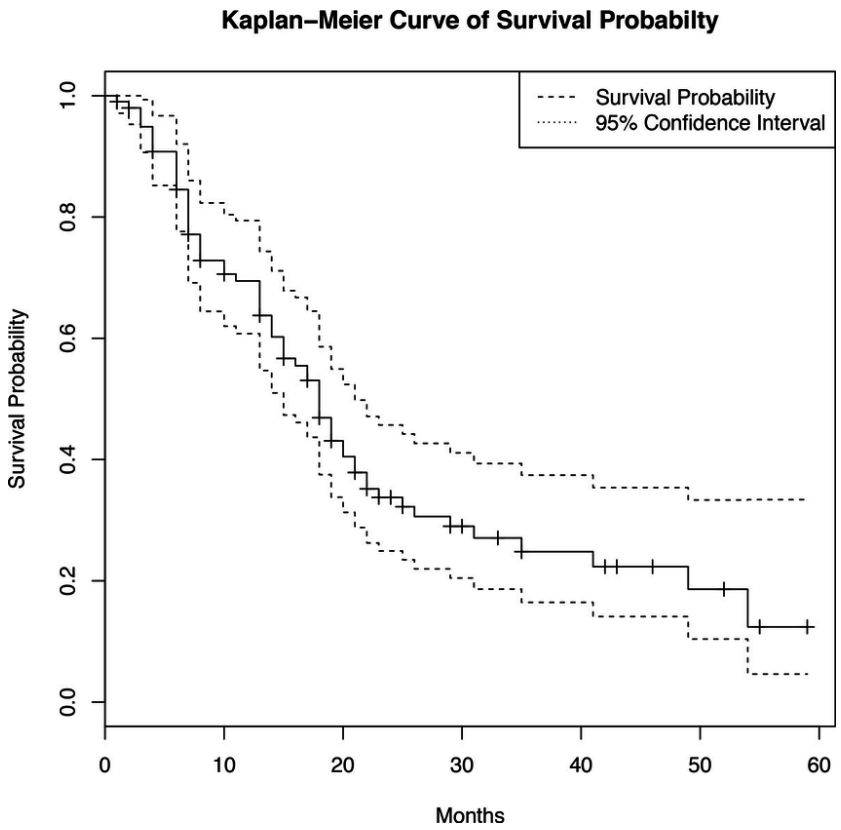
[Reddy and Li, 2015]

Kaplan-Meier Curve

- Non-Parametric Estimate

$$\hat{S}(t) = \prod_{j:T_j<t} \left(1 - \frac{d_j}{r_j}\right)$$

- r_j: number of individuals at risk prior to time T_j
- d_j: number of events at T_j
- Confidence Interval



Logrank Test

$$H_0 : h_0(t) = h_1(t) \qquad H_1 : h_0(t) \neq h_1(t)$$

$$\chi^2_{logrank} = \frac{\left[\sum_{j=1}^{k}\left(d_{0j} - \frac{r_{0j}d_j}{r_j}\right)\right]^2}{\sum_{j=1}^{k}\frac{r_{1j}r_{0j}d_j(r_j - d_j)}{r_j^2(r_j - 1)}}$$

[Fleming and Harrington, 1981]

Cox-Proportional Hazards Model

Cox-Proportional Hazards Model

$$h(t|X_i) = h_0(t)\exp[X_i\beta]$$

$$S(t|X_i) = \exp[-H_0(t)\exp(X_i\beta)]$$

Software

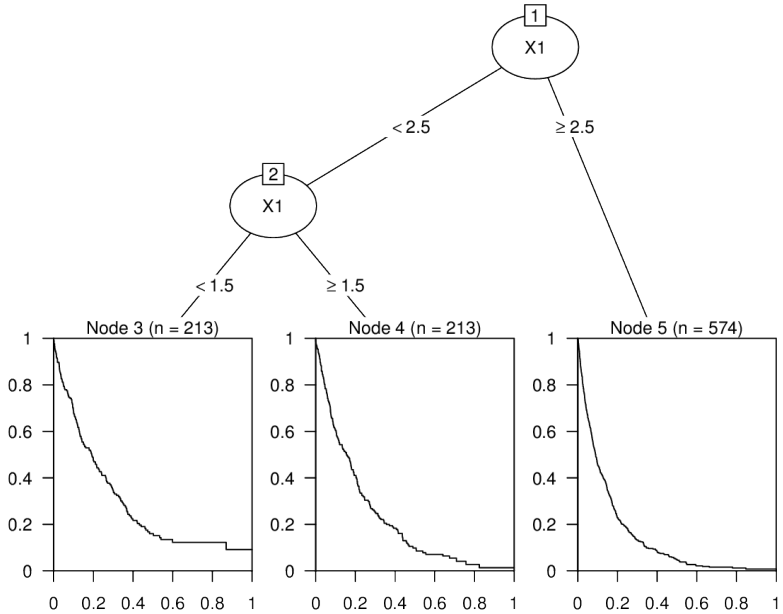
Python

- pandas
- numpy
- scikit-learn
- scikit-survival

Survival Trees

- Splitting Criterion
- Prediction
- Cross Validation
- Pruning

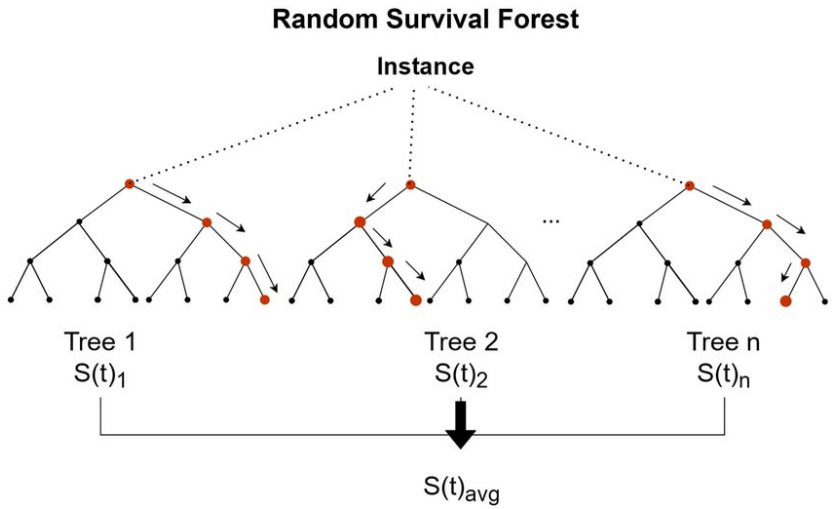
[Safavian and Landgrebe, 1991]



Random Survival Forest

- Ensemble of Survival Trees
- Bagging
- Subspace Sampling
- Feature Importance
- Prediction

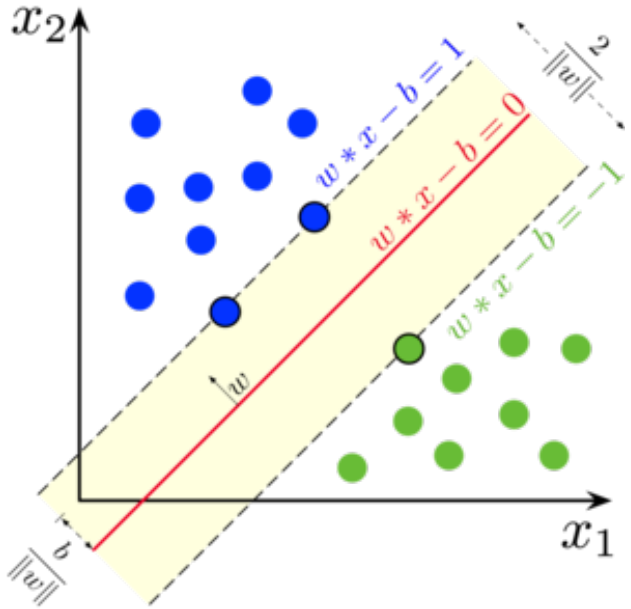
[Ishwaran et al., 2008]

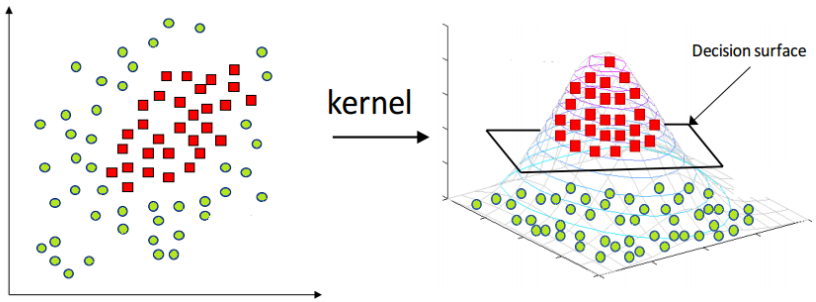


Support Vector Machine (SVM)

- Hyperplane
- Kernels
- Regularization

[Wang et al., 2019]





Evaluation Metrics I

Concordance Index

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I[\hat{S}(y_i|X_i) < \hat{S}(y_j|X_j)]$$

$$= \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I[X_i\hat{\beta} > X_j\hat{\beta}]$$

[Uno et al., 2011]

Evaluation Metrics II

Integrated Brier Score

$$IBS = \int_{t_1}^{t_k} BS(t)dt$$

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{S}(t|X_i) - I[T_i > t]\right]^2$$

[Graf et al., 1999]

Evaluation Metrics III

Cumulative/Dynamic AUC

$$AUC(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I[T_i > t] I[T_j \leq t] I[\hat{S}(t|X_i) \geq \hat{S}(t|X_j)]}{\sum_{i=1}^{n} \sum_{j=1}^{n} I[T_i > t] I[T_j \leq t]}$$

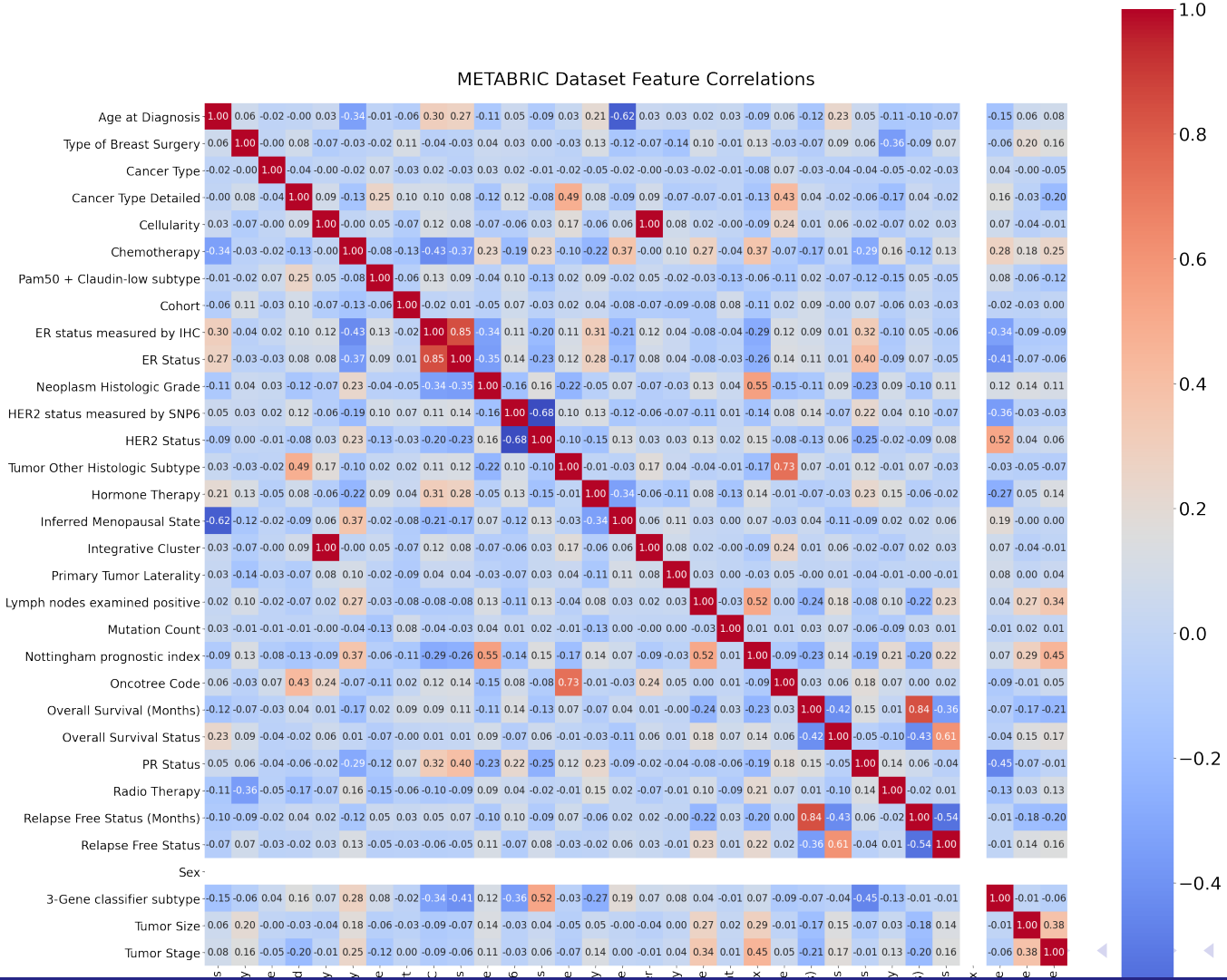
[Uno et al., 2011]

Breast Cancer (METABRIC)

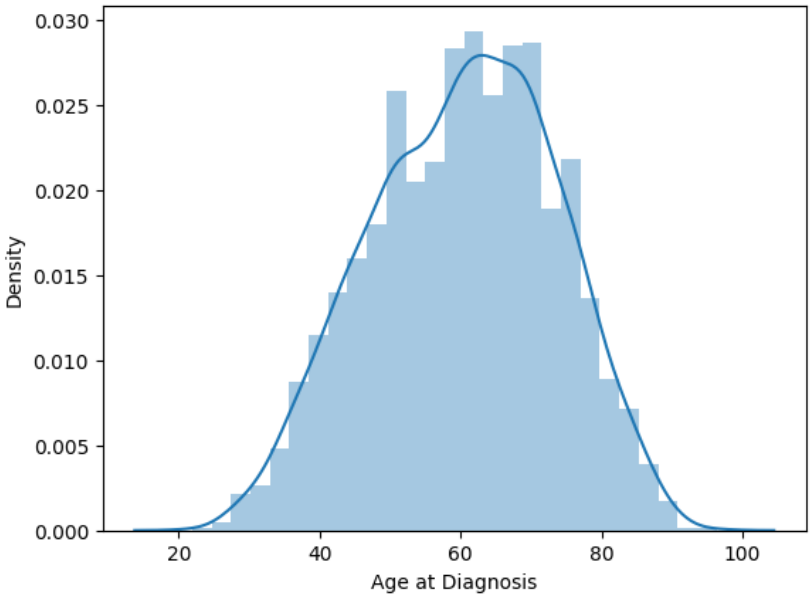
#	Column	Non-Null Count	Dtype
0	Patient ID	2509 non-null	object
1	Age at Diagnosis	2498 non-null	float64
2	Type of Breast Surgery	1955 non-null	object
3	Cancer Type	2509 non-null	object
4	Cancer Type Detailed	2509 non-null	object
5	Cellularity	1917 non-null	object
6	Chemotherapy	1980 non-null	object
7	Pam50 + Claudin-low subtype	1980 non-null	object
8	Cohort	2498 non-null	float64
9	ER status measured by IHC	2426 non-null	object
10	ER Status	2469 non-null	object
11	Neoplasm Histologic Grade	2388 non-null	float64
12	HER2 status measured by SNP6	1980 non-null	object
13	HER2 Status	1980 non-null	object
14	Tumor Other Histologic Subtype	2374 non-null	object
15	Hormone Therapy	1980 non-null	object
16	Inferred Menopausal State	1980 non-null	object
17	Integrative Cluster	1980 non-null	object
18	Primary Tumor Laterality	1870 non-null	object
19	Lymph nodes examined positive	2243 non-null	float64
20	Mutation Count	2357 non-null	float64
21	Nottingham prognostic index	2287 non-null	float64
22	Oncotree Code	2509 non-null	object
23	Overall Survival (Months)	1981 non-null	float64
24	Overall Survival Status	1981 non-null	object
25	PR Status	1980 non-null	object
26	Radio Therapy	1980 non-null	object
27	Relapse Free Status (Months)	2388 non-null	float64
28	Relapse Free Status	2488 non-null	object
29	Sex	2509 non-null	object
30	3-Gene classifier subtype	1764 non-null	object
31	Tumor Size	2360 non-null	float64
32	Tumor Stage	1788 non-null	float64
33	Patient's Vital Status	1980 non-null	object

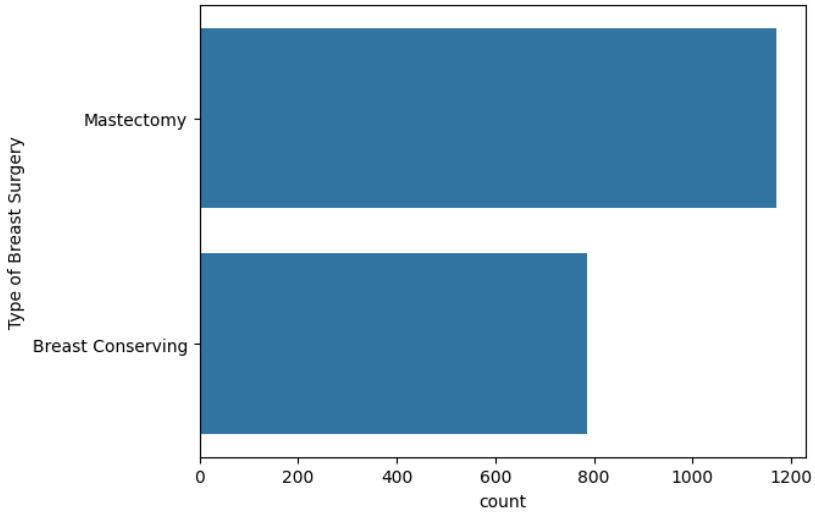
- 2,509 Breast Cancer Patients
- 30 Features
- Overall Survival Status
- Overall Survival (Months)
- Relapse Free Status
- Relapse Free Status (Months)

EDA I

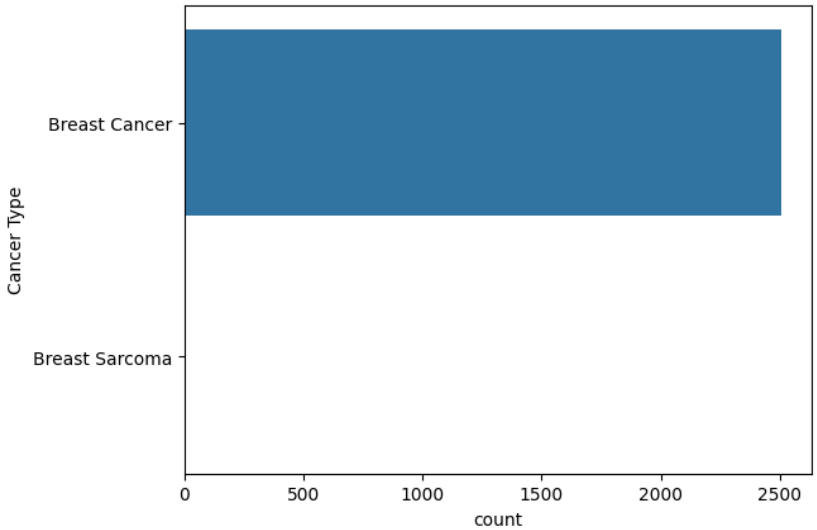

METABRIC Dataset Feature Correlations
Age at Diagnosis
Type of Breast Surgery
Cancer Type
Cancer Type Detailed
Cellularity
Chemotherapy
Pam50 + Claudin-low subtype
Cohort
ER status measured by IHC
ER Status
Neoplasm Histologic Grade
HER2 status measured by SNP6
HER2 Status
Tumor Other Histologic Subtype
Hormone Therapy
Inferred Menopausal State
Integrative Cluster
Primary Tumor Laterality
Lymph nodes examined positive
Mutation Count
Nottingham prognostic index
Oncotree Code
Overall Survival (Months)
Overall Survival Status
PR Status
Radio Therapy
Relapse Free Status (Months)
Relapse Free Status
Sex
3-Gene classifier subtype
Tumor Size
Tumor Stage
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

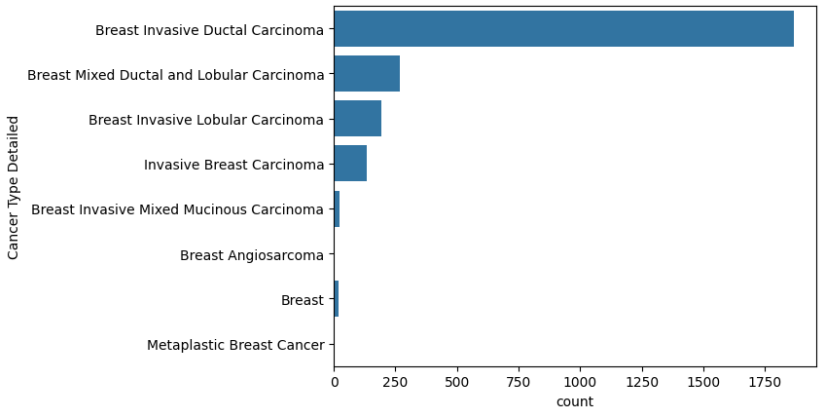
EDA II


0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
20
40
60
80
100
Age at Diagnosis

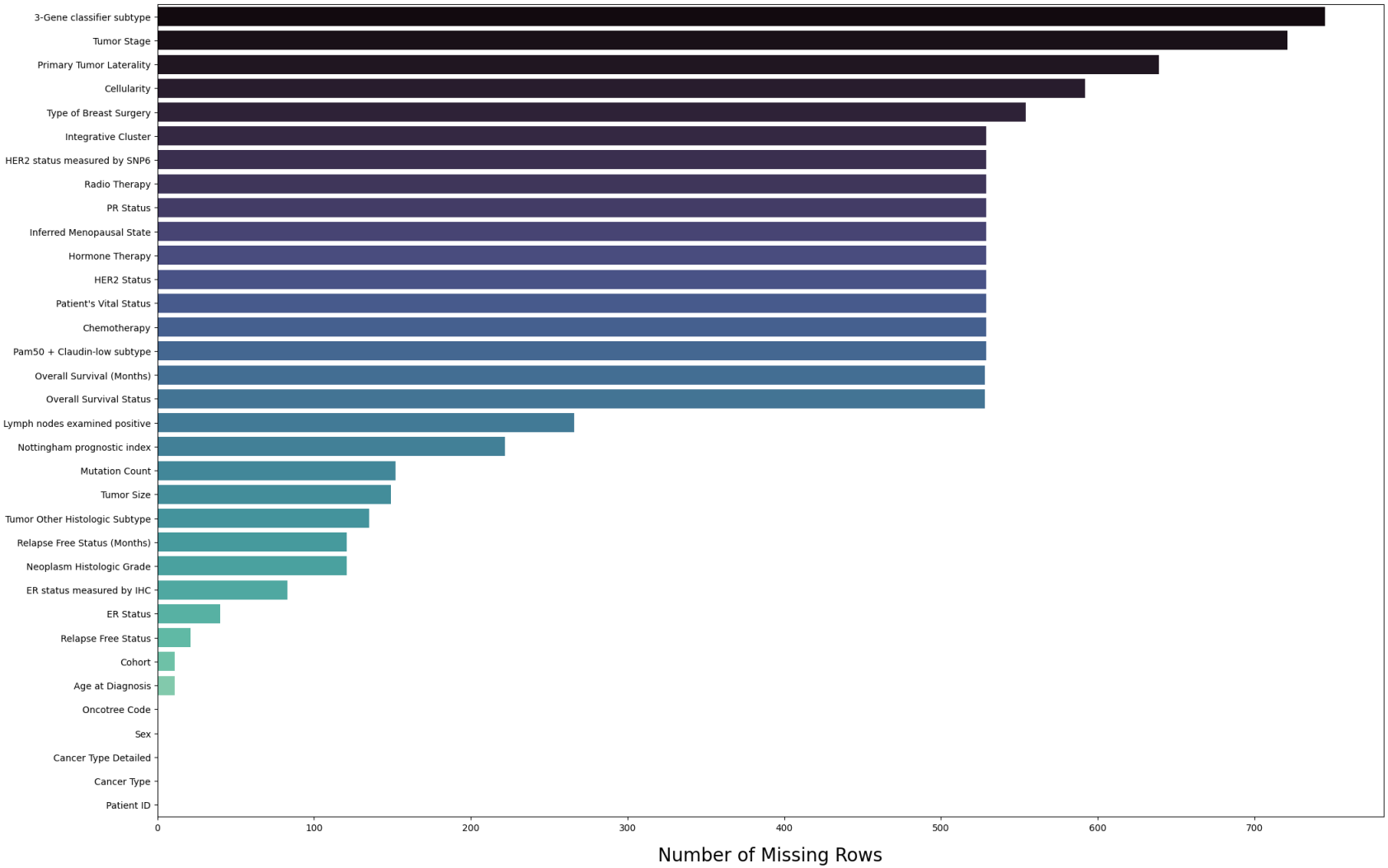

Mastectomy
Breast Conserving
Type of Breast Surgery
0
200
400
600
800
1000
1200
count

EDA III

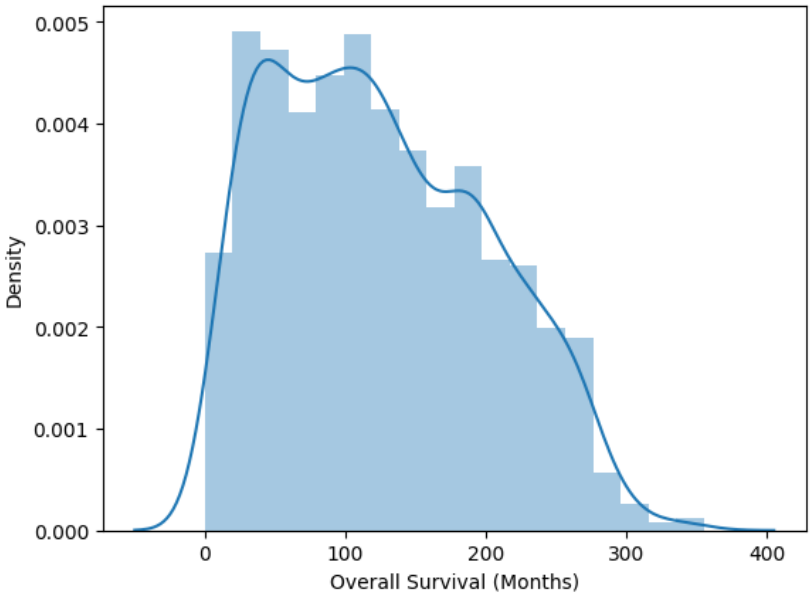

Breast Cancer
Breast Sarcoma
Cancer Type
0
500
1000
1500
2000
2500
count

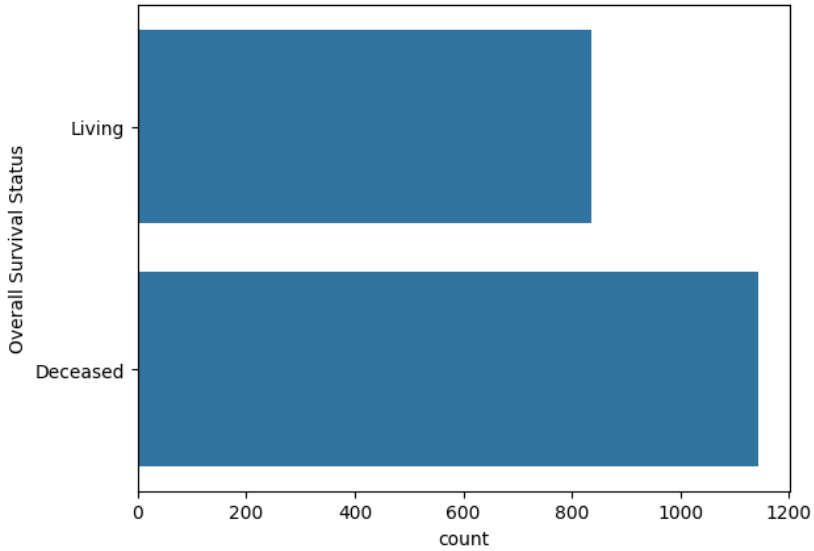

Breast Invasive Ductal Carcinoma
Breast Mixed Ductal and Lobular Carcinoma
Breast Invasive Lobular Carcinoma
Invasive Breast Carcinoma
Breast Invasive Mixed Mucinous Carcinoma
Breast Angiosarcoma
Breast
Metaplastic Breast Cancer
Cancer Type Detailed
0
250
500
750
1000
1250
1500
1750
count

EDA IV

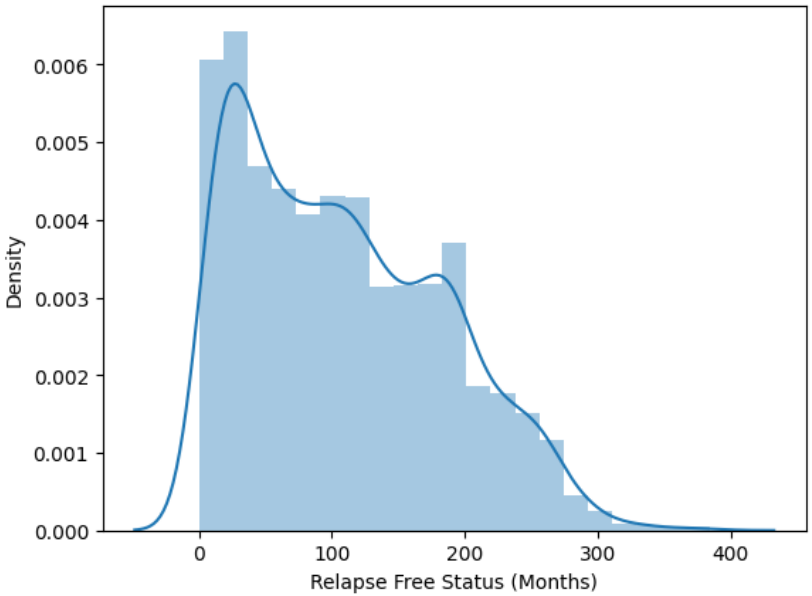

3-Gene classifier subtype
Tumor Stage
Primary Tumor Laterality
Cellularity
Type of Breast Surgery
Integrative Cluster
HER2 status measured by SNP6
Radio Therapy
PR Status
Inferred Menopausal State
Hormone Therapy
HER2 Status
Patient's Vital Status
Chemotherapy
Pam50 + Claudin-low subtype
Overall Survival (Months)
Overall Survival Status
Lymph nodes examined positive
Nottingham prognostic index
Mutation Count
Tumor Size
Tumor Other Histologic Subtype
Relapse Free Status (Months)
Neoplasm Histologic Grade
ER status measured by IHC
ER Status
Relapse Free Status
Cohort
Age at Diagnosis
Oncotree Code
Sex
Cancer Type Detailed
Cancer Type
Patient ID
0
100
200
300
400
500
600
700
Number of Missing Rows

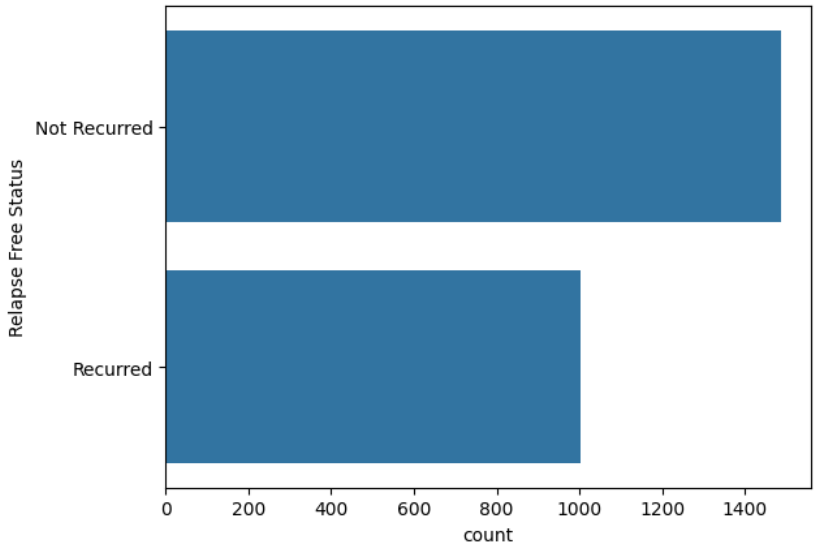
EDA V


0.005
0.004
0.003
0.002
0.001
0.000
Density
0
100
200
300
400
Overall Survival (Months)


Living
Deceased
Overall Survival Status
0
200
400
600
800
1000
1200
count

EDA VI


0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
0
100
200
300
400
Relapse Free Status (Months)


Not Recurred
Recurred
Relapse Free Status
0
200
400
600
800
1000
1200
1400
count

Pre-processing

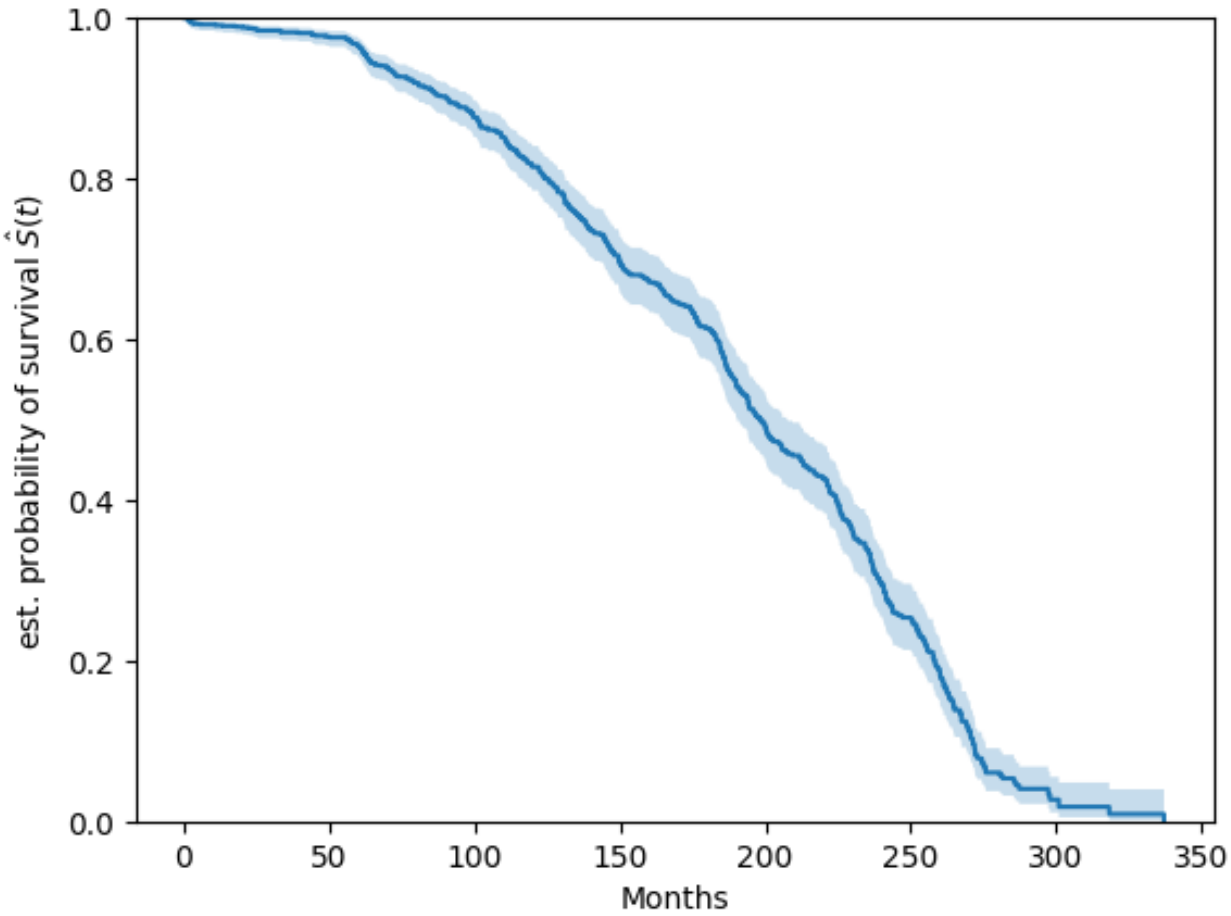
Feature Removal:

- Patient ID
- Cancer Type
- Sex
- Integrative Cluster
- Patient's Vital Status

Complete Dataset

- 1092 Instances
- 8 Numerical
- 16 Nominal
- 1 Ordinal
- 4 Labels
- 45 Features

Kaplan Meier Curve

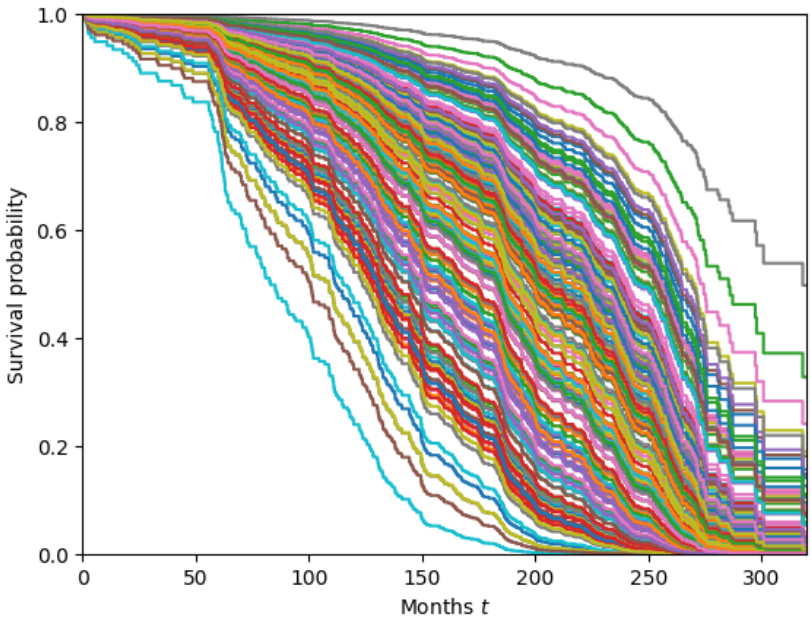


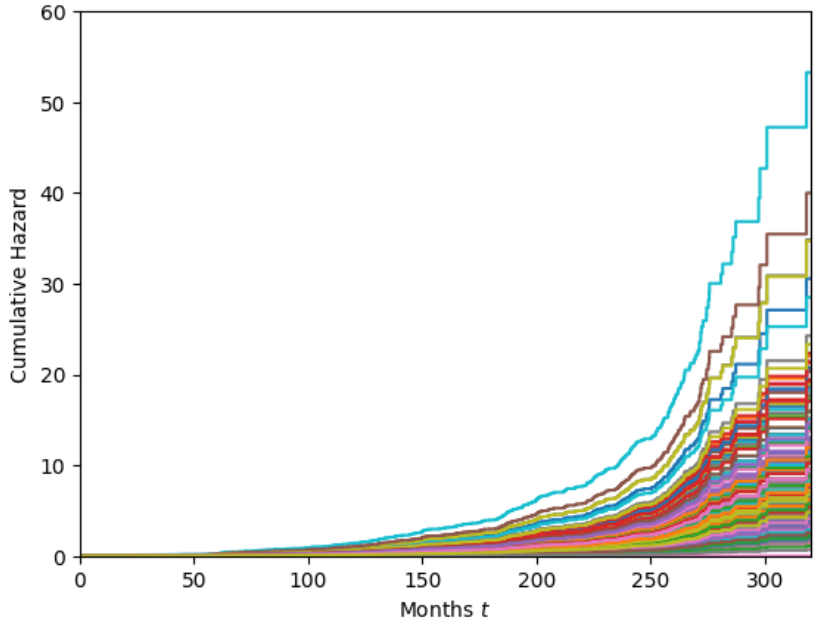
Cox-Proportional Hazards Model I

Cohort	0.677704
Mutation Count	0.612564
Hormone Therapy_Yes	0.578156
Tumor Stage	0.554470
Chemotherapy_Yes	0.552463
Radio Therapy_Yes	0.546204
Primary Tumor Laterality_Right	0.533041
Tumor Size	0.528651
Nottingham prognostic index	0.526868
HER2 status measured by SNP6_Neutral	0.526080

- Cohort: -0.497392
- Mutation Count: -0.141548
- Hormone Therapy_Yes: 0.378709
- Tumor Stage: 0.076158
- Chemotherapy_Yes: 0.493760
- Radio Therapy_Yes: 0.293066
- Primary Tumor Laterality_Right: 0.121881

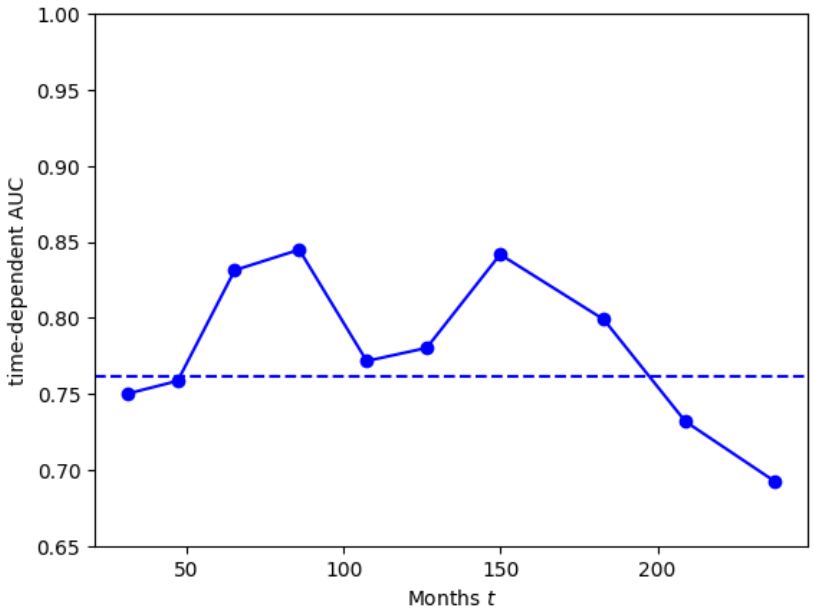
Cox-Proportional Hazards Model II





Cox-Proportional Hazards Model III

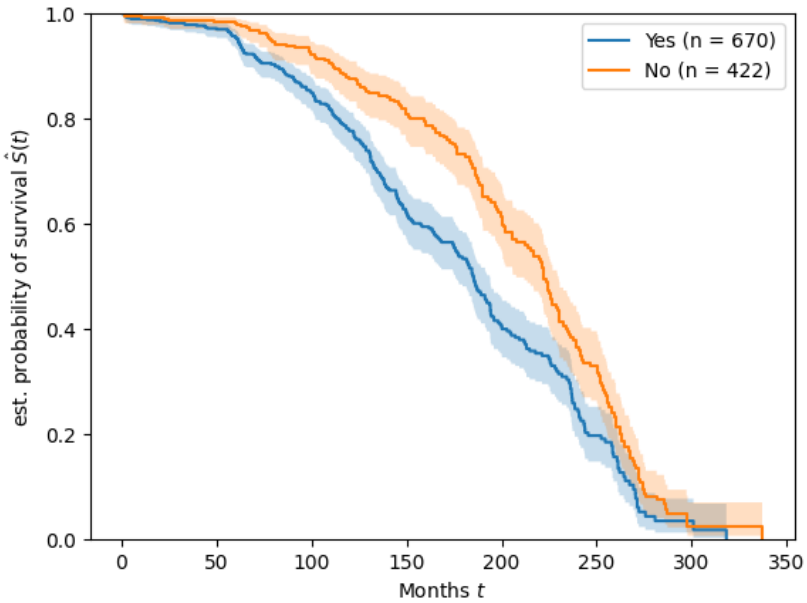
- C-Index: 0.71579
- Integrated Brier Score: 0.12082
- Cumulative AUC Mean: 0.76155



Log Rank Test I

Hormone Therapy

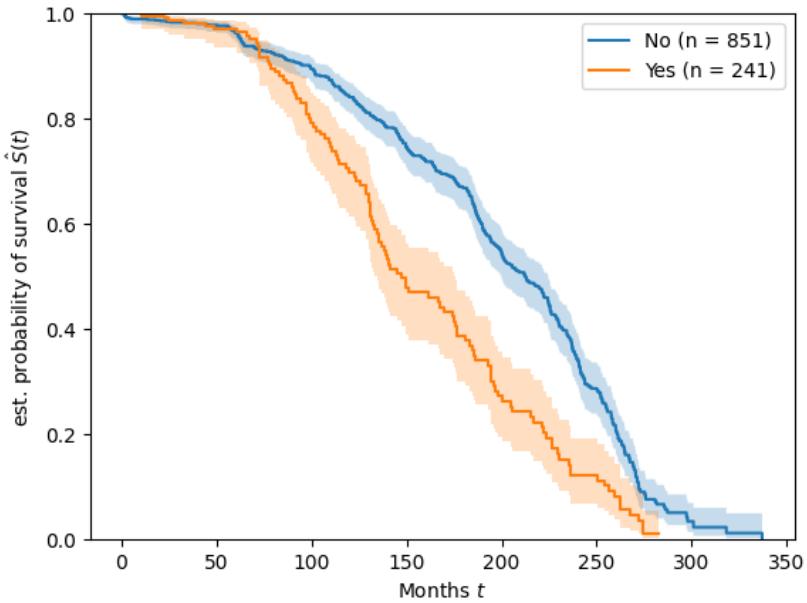
- χ^2: 19.87413
- p-value: 8.27119e-06



Log Rank Test II

Chemotherapy

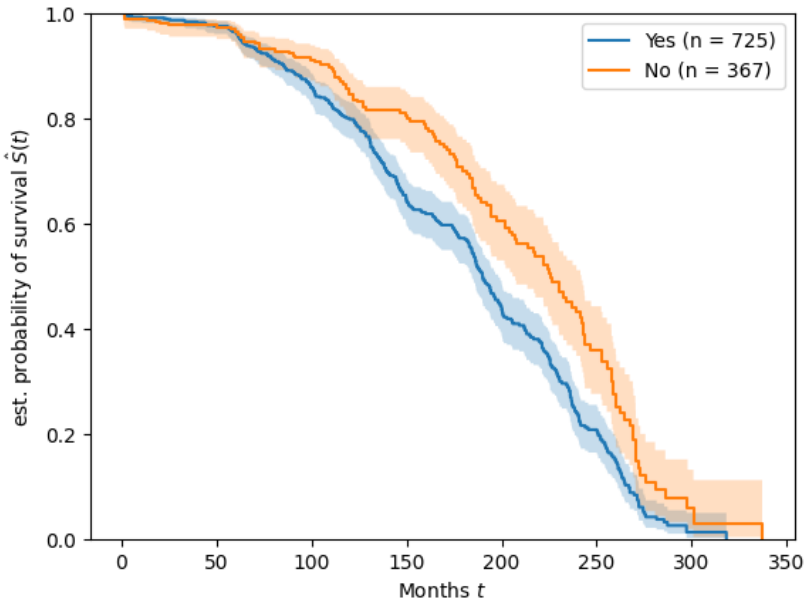
- χ^2: 36.52974
- p-value: 1.50354e-09



Log Rank Test III

Radio Therapy

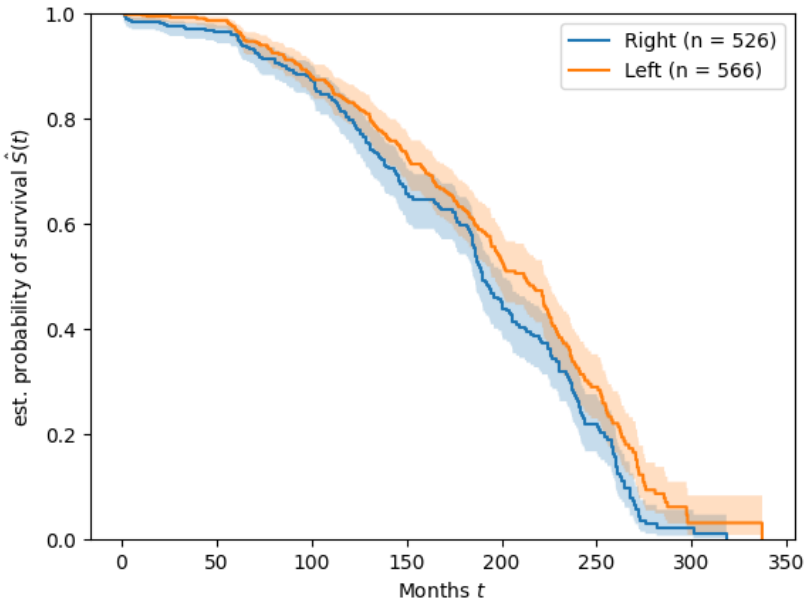
- χ^2: 18.00271
- p-value: 2.20590e-05



Log Rank Test IV

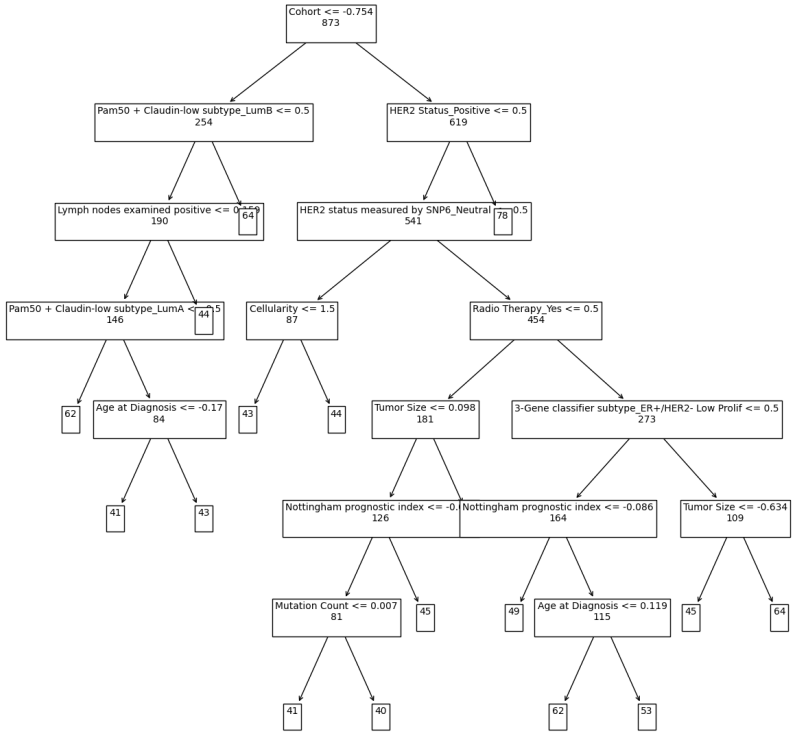
Primary Tumor Laterality

- χ^2: 10.73803
- p-value: 0.01323

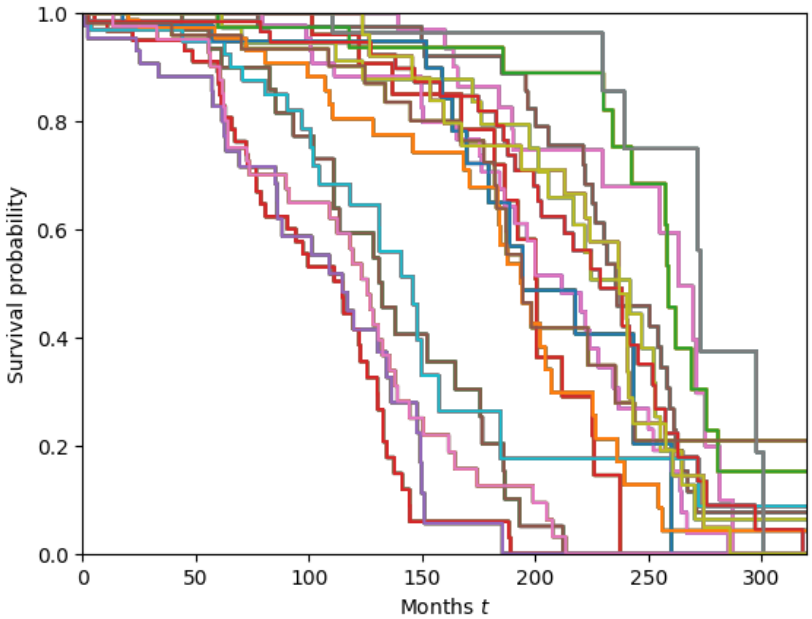


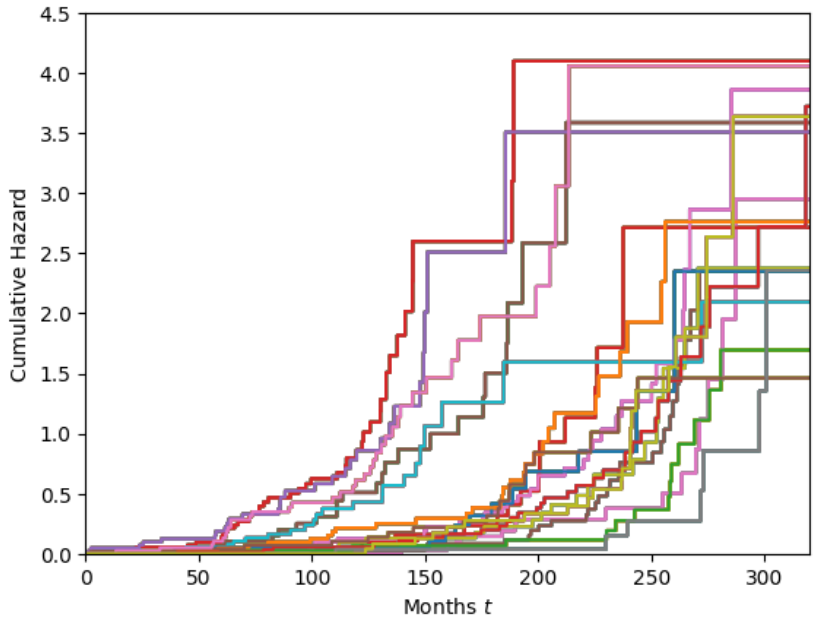
Survival Tree I

- Cross Validaion: 5 Fold
- Max Depth: 8
- Min Samples Leaf: 40
- 17 Leafs



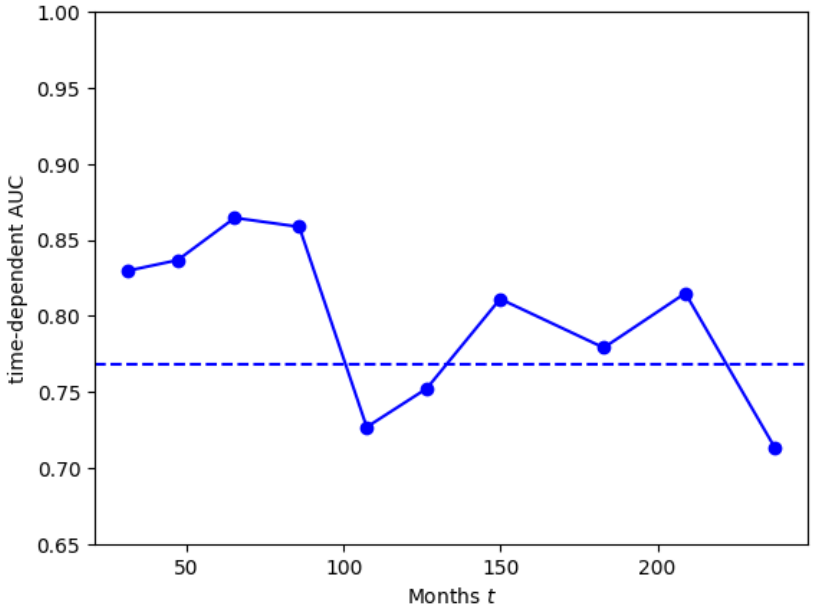
Survival Tree II





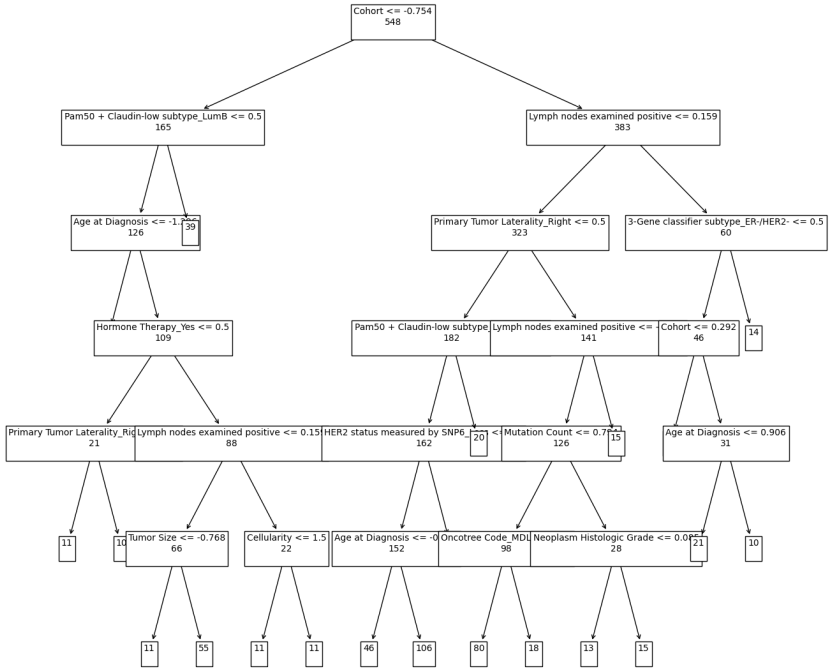
Survival Tree III

- C-Index: 0.71565
- Integrated Brier Score: 0.11133
- Cumulative AUC Mean: 0.76799

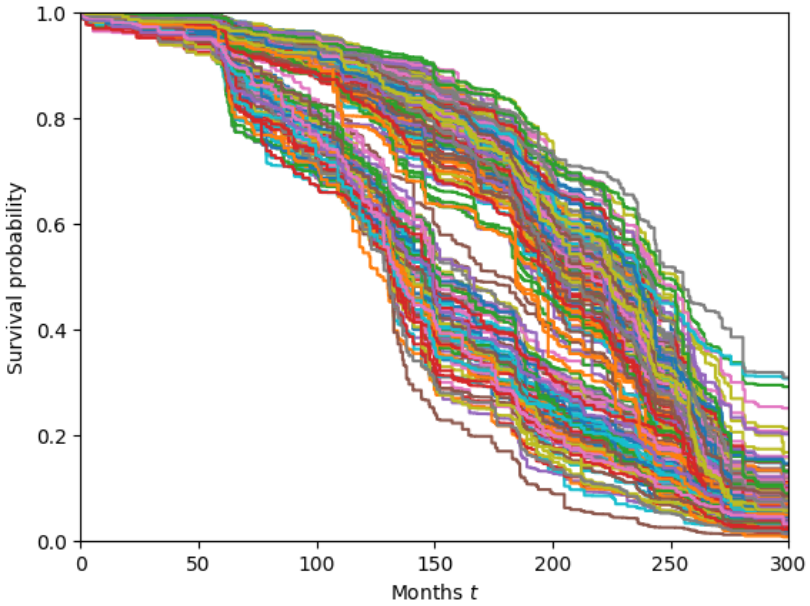


Random Survival Forest I

- Cross-Validation: 5 Fold
- Max Depth: 6
- Min Samples Leaf: 10



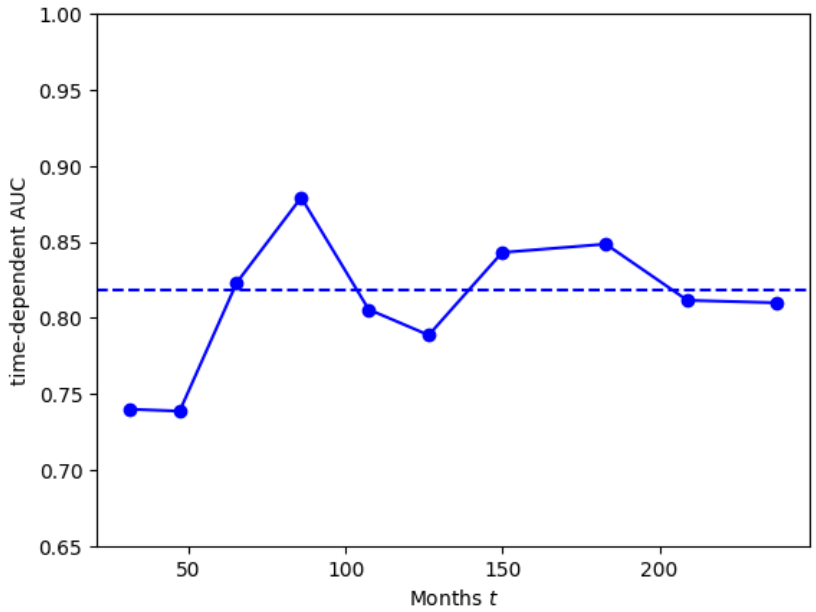
Random Survival Forest II



	importances_mean	importances_std
Cohort	0.173233	0.023901
Hormone Therapy_Yes	0.011999	0.007818
Nottingham prognostic index	0.007982	0.005699
Age at Diagnosis	0.007514	0.003328
Chemotherapy_Yes	0.005528	0.008137
Pam50 + Claudin-low subtype_LumA	0.004899	0.001824
Tumor Stage	0.003955	0.002190
Type of Breast Surgery_Mastectomy	0.003784	0.001758
Tumor Size	0.003568	0.002902
Radio Therapy_Yes	0.002508	0.002876

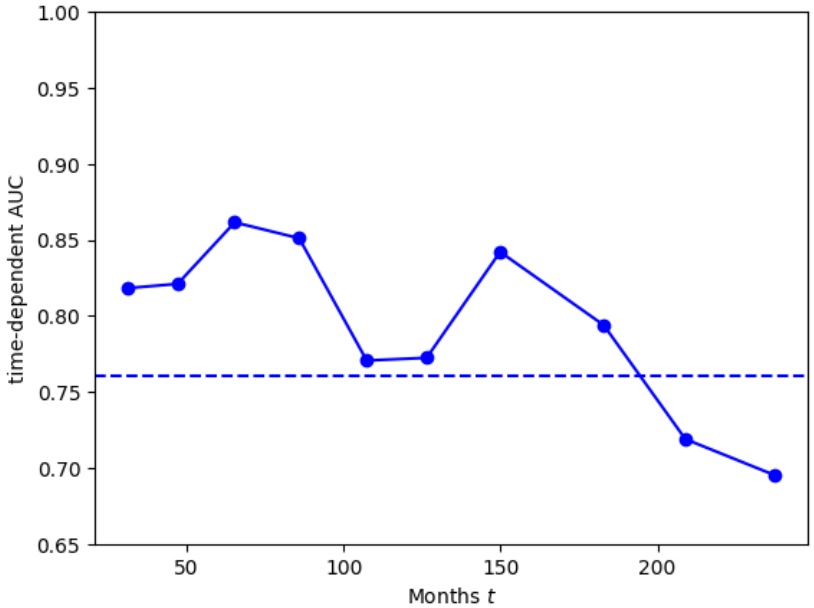
Random Survival Forest III

- C-Index: 0.75570
- Integrated Brier Score: 0.10784
- Cumulative AUC Mean: 0.81813



SVM

- alpha: 0.0625
- C-Index: 0.72064
- Cumulative AUC Mean: 0.76021



Comparison

	C-Index	IBS	AUC Mean
CoxPH	0.7159	0.12082	0.76155
ST	0.71565	0.11133	0.76799
RSF	**0.75570**	**0.10784**	**0.81813**
SVM	0.72064		0.76021

Conclusion

- Survival Analysis
- Machine Learning
- Real Data Analysis

Future Work

- Imputation
- Relapse Free Status
- Other ML Methods [Wang et al., 2019]:
  - Neural Networks
  - Gradient Boosting
  - Bagging Survival Trees

References I

Fleming, T. R. and Harrington, D. P. (1981).
A class of hypothesis tests for one and two sample censored survival data.
Communications in Statistics - Theory and Methods, 10(8):763–794.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999).
Assessment and comparison of prognostic classification schemes for survival data.
Statistics in Medicine, 18(17-18):2529–2545.

References II

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008).
Random survival forests.
The Annals of Applied Statistics, 2(3):841 – 860.

Reddy, C. K. and Li, Y. (2015).
A review of clinical prediction models.
In *Healthcare Data Analytics*.

Safavian, S. and Landgrebe, D. (1991).
A survey of decision tree classifier methodology.
IEEE Transactions on Systems, Man, and Cybernetics, 21(3):660–674.

References III


Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011).
On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.
Statistics in Medicine, 30(10):1105–1117.


Wang, P., Li, Y., and Reddy, C. K. (2019).
Machine learning for survival analysis: A survey.
ACM Comput. Surv., 51(6).