Predicting Breast Cancer Survival and Identifying Biomarkers for Survival

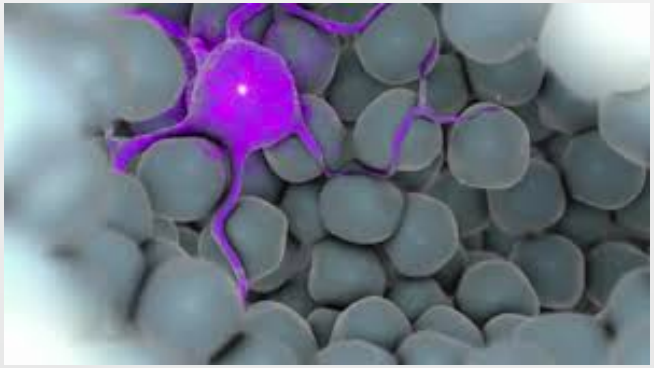
Shannon Ballard, Ph.D.
Springboard Data Science Career Track

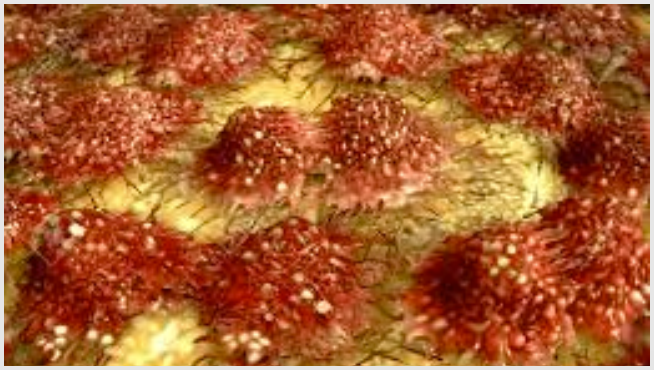
The Big Picture Question:

Can we predict how a disease will affect an individual?

Breast Cancer

- Increase in the proliferation of cells within the breast tissue, forming a tumor
- As of Jan. 2020, over 3.5 million women will have had or currently have breast cancer in the United States
- The average 10-year survival rate for women with invasive breast cancer is 84%
- men have a 1 in 883 chance of developing breast cancer





The Specific Questions:

Can a patients' outcome from breast cancer be predicted from measured diseased-state features?

Can specific genes be identified as potential breast cancer biomarkers and therapeutic targets?

Description of Data Collected:

- Netherlands Cancer Institute (NKI)
- 272 breast cancer patients
- 1,569 features

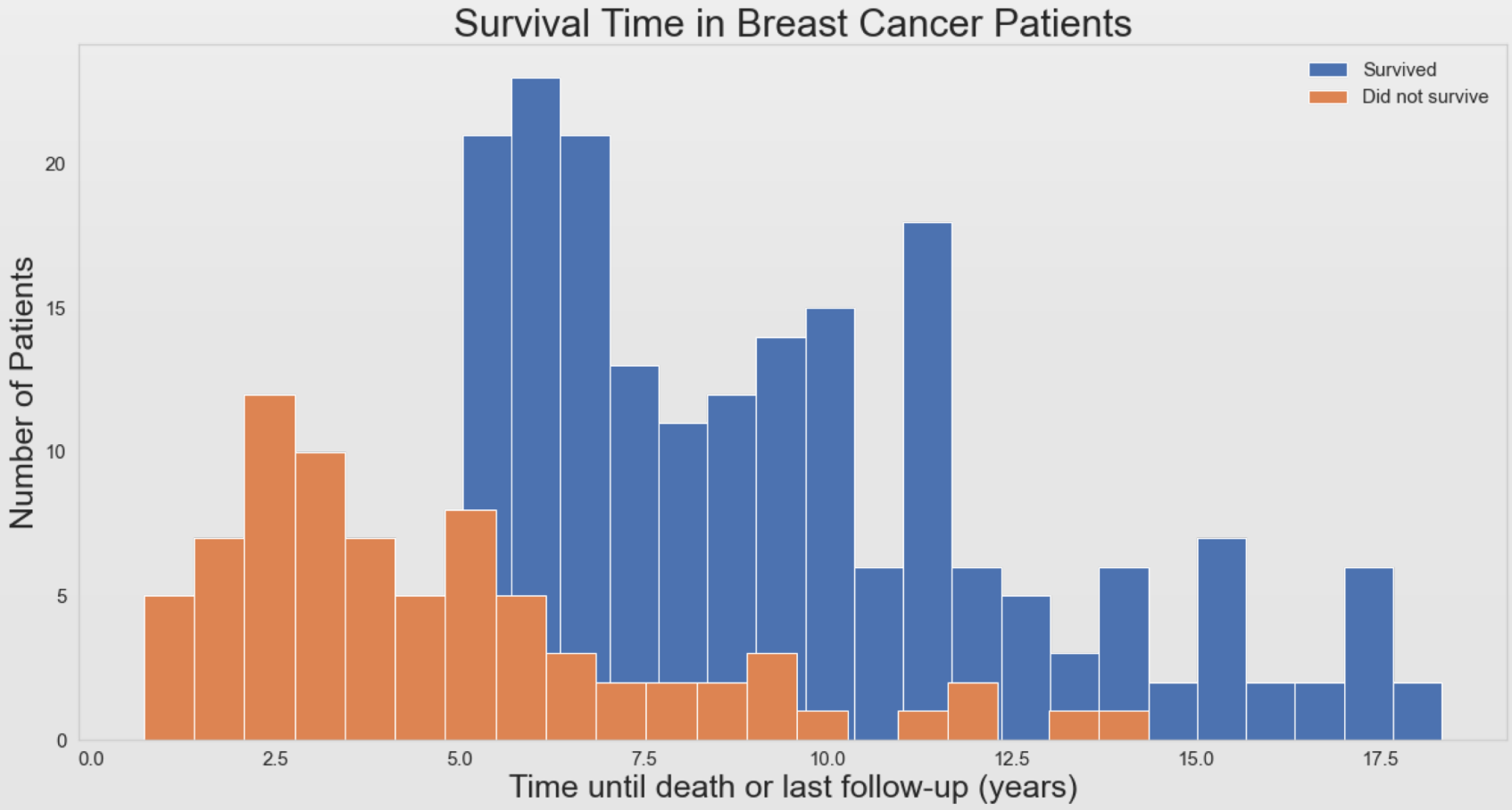
Description of Data Collected:

Variable	Details	Type
Patient	Patient sample number	Continuous
ID	Patient ID	Continuous
age	Age at which patient was diagnosed with breast cancer	Continuous
eventdeath	0 = alive, 1 = death	Categorical
survival	Time (in years) until death or last follow-up	Continuous
timerecurrence	Time (in years) until cancer recurrence or last follow-up	Continuous
chemo	chemotherapy used (yes=1/no=0)	Categorical
hormonal	Hormonal therapy used (yes=1/no=0)	Categorical
amputation	Mastectomy (yes = 1/no = 0)	Categorical
histtype	Histological grade based on 3 morphological features	Categorical
diam	Diameter of primary tumor	Continuous
posnodes	number of lymph nodes that contained cancerous cells	Continuous
grade	Pathological grade based on cell differentiation & growth rate (1=low, 2=intermediate, 3=high)	Categorical
angioinv	Vascular invasion 1= absent, 2= minor, 3 = major	Categorical
lymphinfil	level of lymphocytic infiltration	Categorical
barcode	sample barcode	Continuous
1,554 Genes	each gene is provided as an individual variable; given as an intensity ratio to that of reference pool	Continuous

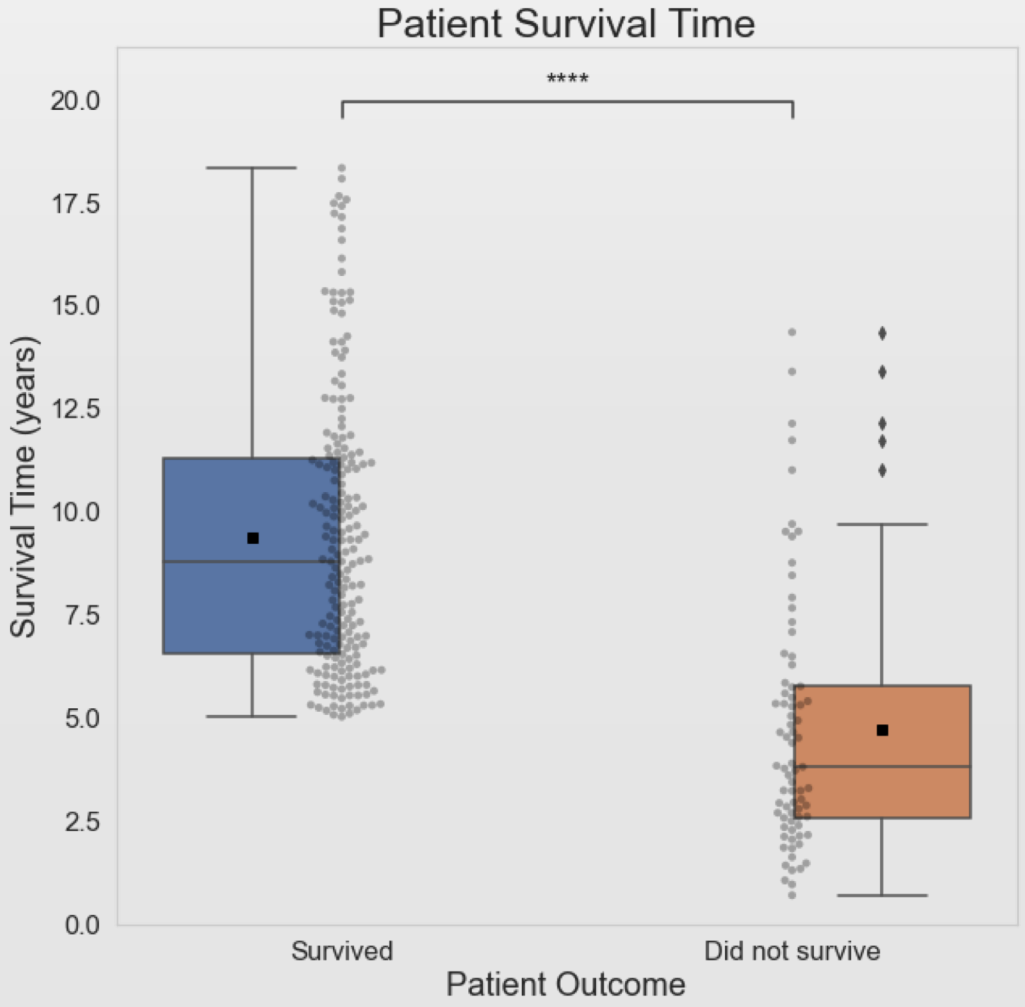
Data Analysis: Data Wrangling

Variable	Details	Type
~~Patient~~	~~Patient sample number~~	~~Continuous~~
~~ID~~	~~Patient ID~~	~~Continuous~~
age	Age at which patient was diagnosed with breast cancer	Continuous
eventdeath	0 = alive, 1 = death	Categorical
survival	Time (in years) until death or last follow-up	Continuous
timerecurrence	Time (in years) until cancer recurrence or last follow-up	Continuous
chemo	chemotherapy used (yes=1/no=0)	Categorical
hormonal	Hormonal therapy used (yes=1/no=0)	Categorical
amputation	Mastectomy (yes = 1/no = 0)	Categorical
histtype	Histological grade based on 3 morphological features	Categorical
diam	Diameter of primary tumor	Continuous
posnodes	number of lymph nodes that contained cancerous cells	Continuous
grade	Pathological grade based on cell differentiation & growth rate (1=low, 2=intermediate, 3=high)	Categorical
angioinv	Vascular invasion 1= absent, 2= minor, 3 = major	Categorical
lymphinfil	level of lymphocytic infiltration	Categorical
~~barcode~~	~~sample barcode~~	~~Continuous~~
1,554 Genes	each gene is provided as an individual variable; given as an intensity ratio to that of reference pool	Continuous

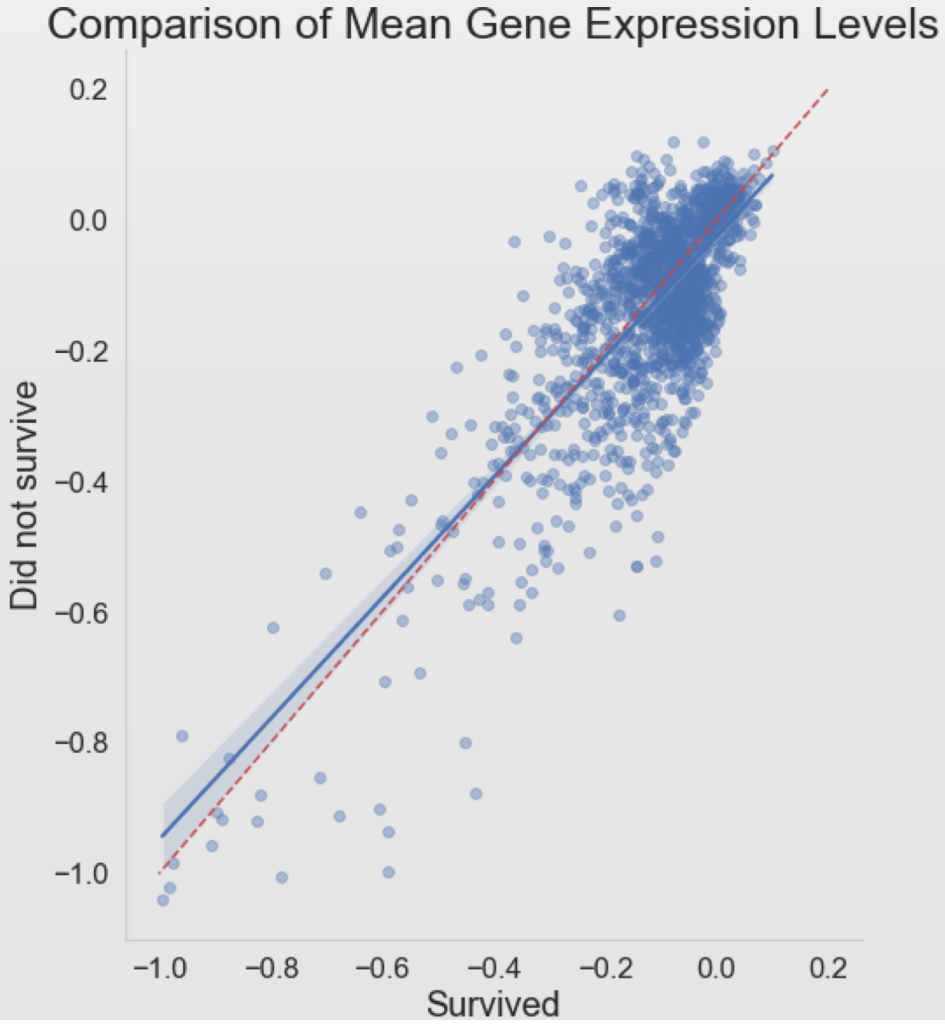
Data Analysis: Initial Findings



Data Analysis: Initial Findings



Data Analysis: Initial Findings



Data Analysis: Generating Models to Predict Patient Survival

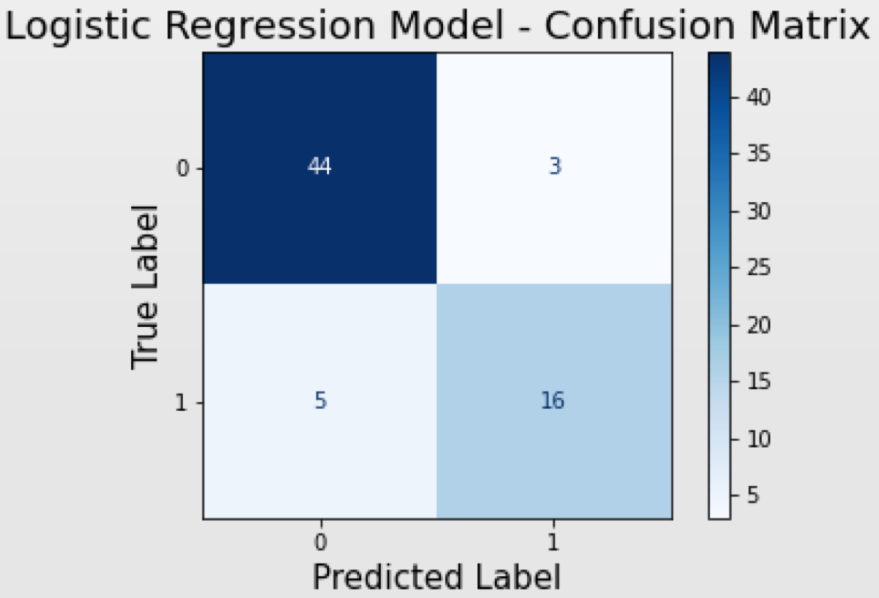
- Assume that features other than 'eventdeath' will account for the patients' outcome (survived or did not survive)

- 3 Models Tested:
  - **Logistic Regression**
  - **Random Forest**

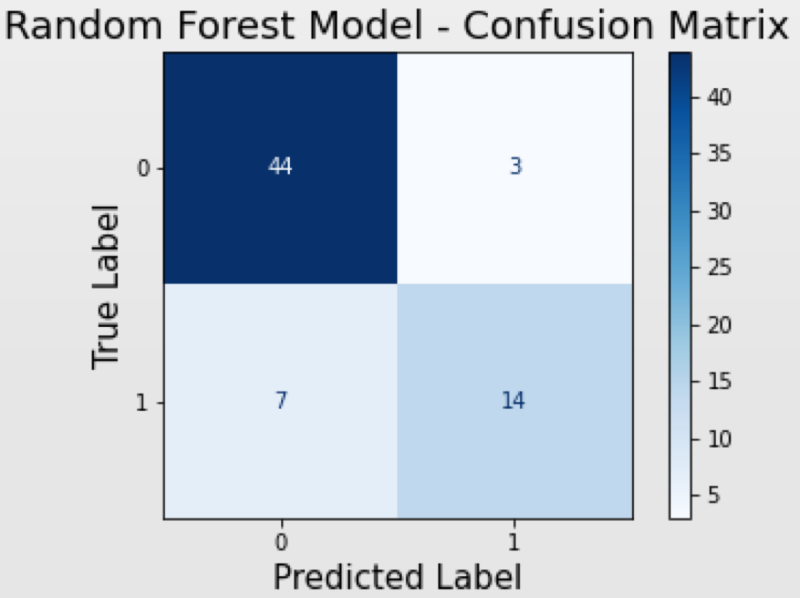
  } calculate the probability of an event and classify an event given the other features

  - **Cox Proportional Hazard** → explores the relationship between patient outcome and the other features, where 'eventdeath' is the hazard function at a given time

Data Analysis: Logistic Regression vs. Random Forest

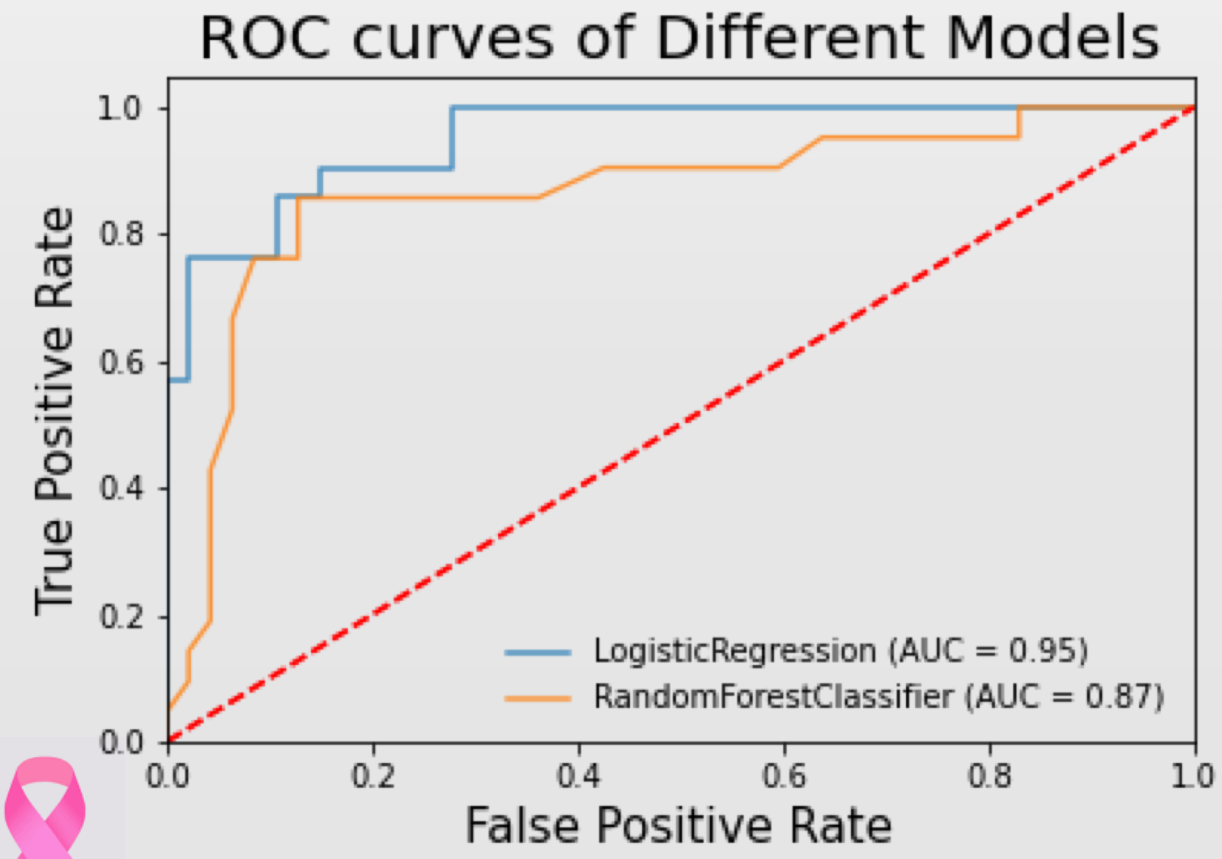


- Out of the 19 (16 + 3) predicted to be positive, 16 were correctly identified.
- Out of the 21 (5 + 16) positive cases, 16 were correctly identified.



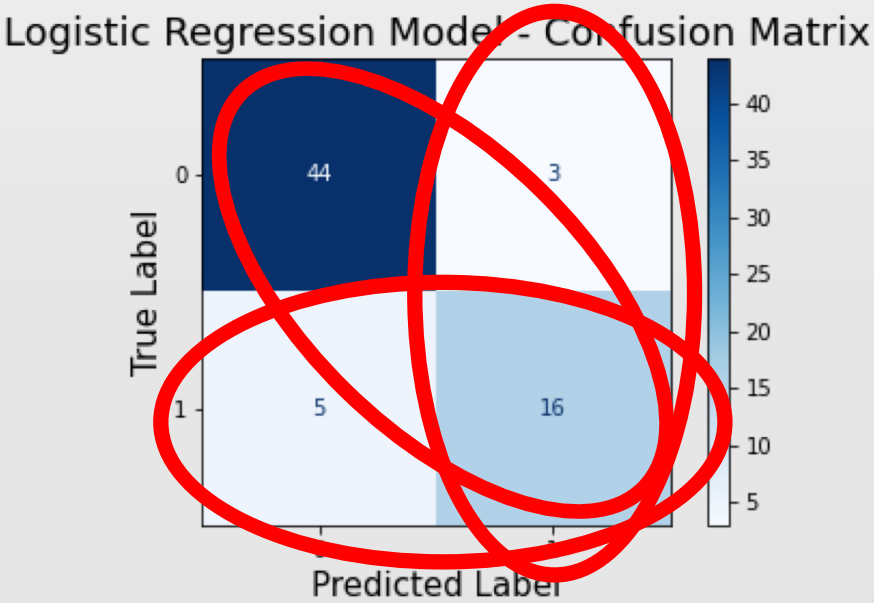
- Out of the 17 (14 + 3) predicted to be positive, 14 were correctly identified.
- Out of the 21 (7 + 14) positive cases, 14 were correctly identified.

Data Analysis: Logistic Regression vs. Random Forest



- **Receiver Operating Characteristic (ROC) curve** –
  - shows the tradeoff between a true positive (Do not survive) and a false positive (Survive).
- **Area under the ROC curve (AUC)** -
  - helps determine how well the model classifies positive and negative outcomes
  - The closer the AUC value is to 1, the better the model

Data Analysis: Choosing the Model



- **Higher Precision** - What proportion of positive identifications was actually correct?
- **Higher Recall** - What proportion of actual positives was identified correctly?
- **Higher F1 Score** - Combines Precision and recall to measure of a model's accuracy
- **Higher Accuracy** - What percentage of the predictions were correct?
- **Higher AUC** - What is the the probability that the model ranks a random positive example more highly than a random negative example?

Data Analysis: Logistic Regression vs. Random Forest

Model	Precision	Recall	F1 score	Accuracy	AUC
Logistic Regression (C:1000)	0.84	0.76	0.8	0.88	0.95
Random Forest (n_estimators = 67)	0.82	0.67	0.74	0.85	0.87

Data Analysis: Cox Proportional Hazard Summary

Concordance index (ci) – evaluates predictions made by the model

- 0.5 → no better than random
- 0.7-0.8 → Good model
- 0.8 -1.0 → Strong model

ci = 0.72

Data Analysis: Identifying Potential Biomarkers for Survival

1, 565 Features

↓ Independent t-test between 'Survived' and 'Did not survive' ($p <= 0.05$)

618 Features

↓ Determine fold change between groups (>=2 or <=-2)

217 Features

Top 10 Features Based on p-value

Feature	Description
timerecurrence	Time (in years) until cancer recurrence or last follow-up
survival	Time (in years) until death or last follow-up
NM_007019	ubiquitin conjugating enzyme E2 C (UBE2C)
NM_004456	enhancer of zeste 2 polycomb repressive complex 2 subunit (EZH2)
AL049265	mRNA; cDNA DKFZp564F053 - putative ER target gene
NM_020974	CUB domain and EGF like domain containing 2 (SCUBE2)
AL137566	mRNA; cDNA DKFZp586G0321
NM_000849	glutathione S-transferase mu 3 (GSTM3)
NM_000926	progesterone receptor (PGR)
Contig56390_RC	unknown function

Data Analysis: Identifying Potential Biomarkers for Survival

Logistic Regression Model

↓

Top 10 Positive and negative coefficients

↓

Compare to 217 features

↓

Determine whether a change in a feature makes the death of a patient more likely (positive) or less likely (negative)

Data Analysis: Identifying Potential Biomarkers for Survival

Feature	Description
survival	Time (in years) until death or last follow-up
timerecurrence	Time (in years) until cancer recurrence or last follow-up
diam	Diameter of primary tumor
NM_000853	glutathione S-transferase theta 1

Glutathione *S*-Transferase Mu and Theta Polymorphisms and Breast Cancer Susceptibility

Montserrat García-Closas, Karl T. Kelsey, Susan E. Hankinson, Donna Spiegelman, Kathryn Springer, Walter C. Willett, Frank E. Speizer, David J. Hunter

Association Between Glutathione *S*-Transferase M1, P1, and T1 Genetic Polymorphisms and Development of Breast Cancer

*Kathy J. Helzlsouer, Ornella Selmin, Han-Yao Huang, Paul T. Strickland, Sandra Hoffman, Anthony J. Alberg, Mary Watson, George W. Comstock, Douglas Bell**

Future Research and Recommendations

- **Recommended Model** – Logistic Regression Model
- Test other models
- Gather more data for testing the model
- Include control patients in gene expression analyses
- Can be used by physicians , researchers, and patients to:
  - Help predict patient outcome
  - Find novel biomarkers for breast cancer
  - Test novel therapeutics to treat breast cancer