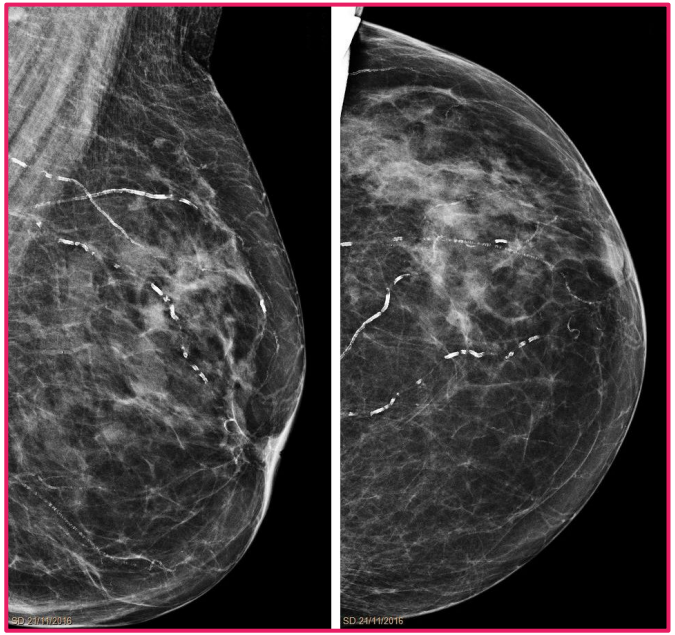
IDS 702 Final Project
Mammography Lesion Mass Analysis

Michael Tang

Motivation

Breast Cancer

- Most commonly diagnosed cancer in women
- Mammography
  - X-ray image of the breast
  - False Positive & False Negative
- Breast Biopsy
  - More accurate
  - Unpleasant



Data

- 961 observations
- 445 +ve and 516 -ve

Severity (Binary)

Benign / Malignant

Shape (Categorical)

- round
- oval
- lobular
- irregular

Density (Categorical)

- high
- iso
- low
- Fat containing

Age (Continuous)

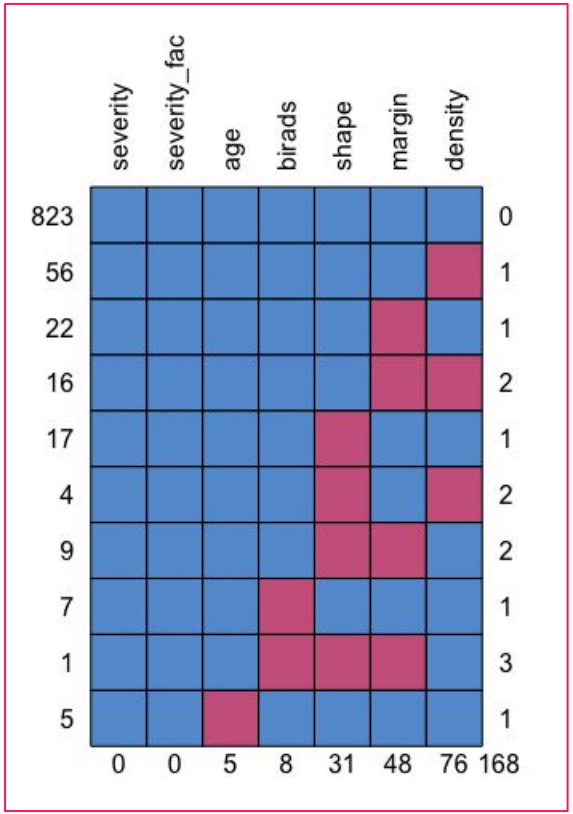
- patient age

Margin (Categorical)

- circumscribed
- microlobulated
- obscured
- Ill-defined
- spiculated

BI-RADS (Categorical)

- 1 - negative
- 2 - benign finding
- 3 - probably benign
- 4 - suspicious abnormality
- 5 - suggestive of malignancy
- 6 - proven malignant



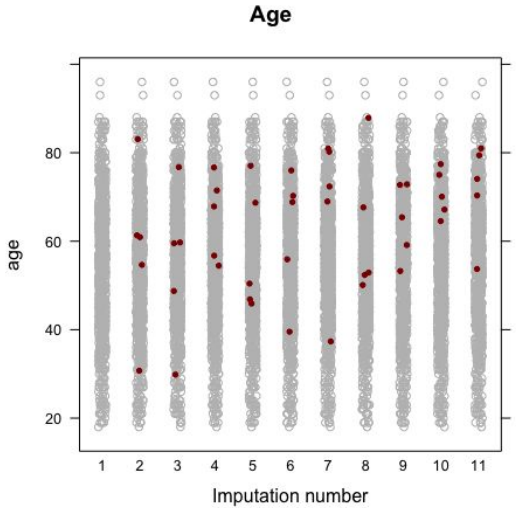
Imputation

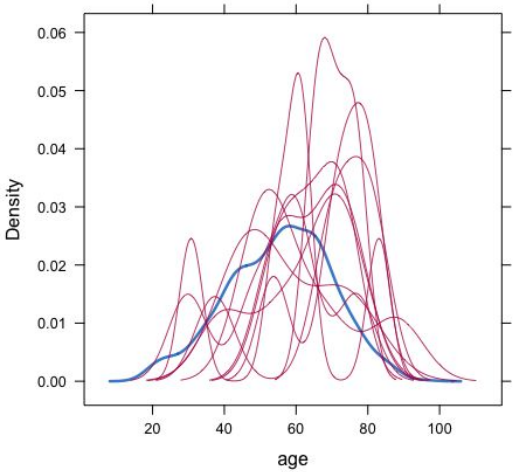
Multiple Imputation

Quality Check on 1 Copy

- 10 copies

- Bayesian Linear Regression
- Logistic Regression
- Multinomial Logistic Regression
- Proportional Odds Model





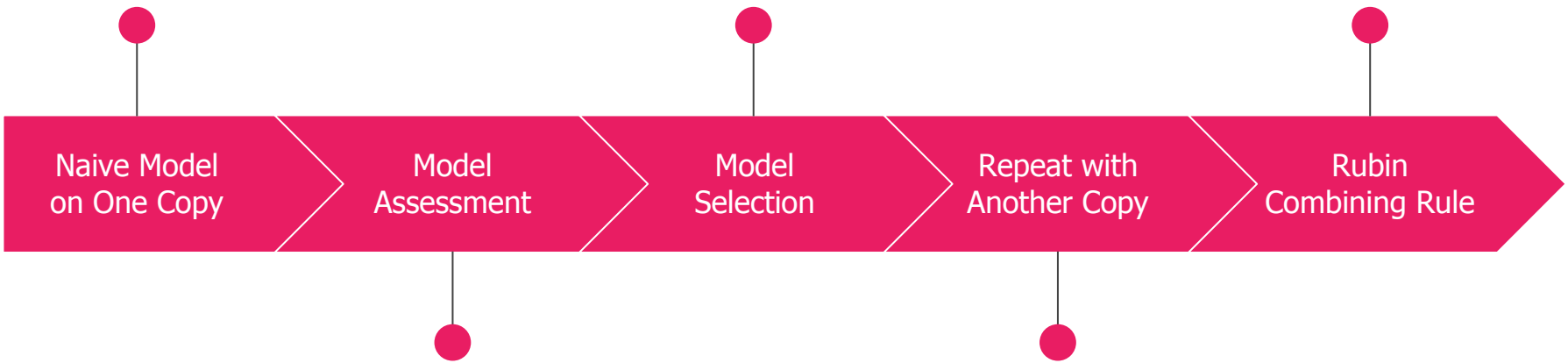
EDA

Objective: Look for potentially significant predictor variables

- Severity vs. Continuous Predictor (Boxplot)
  - Difference in trend -> 'significant'
- Severity vs. Categorical Predictor (Contingency Table)
  - Change in ratio across predictor levels -> 'significant'
- Interaction (Mix)
  - Insufficient sample size -> invalid

	2	3	4	5	6
high	1	1	11	6	0
iso	4	2	50	15	1
low	8	32	485	323	9
fat-containing	1	1	3	6	1

- Findings: All predictors except for Density, No interactions



- Logistic Regression
- All main effects
- Sensitivity: 0.828
- Specificity: 0.869
- AUC: 0.913

- Null: shape & density
- Full: Naive Model
- All 3 selection methods and AIC
- Final: BI-RADS, age, shape, margin
- Sensitivity: 0.831
- Specificity: 0.865
- AUC: 0.911

- Apply to all 10 copies

- Binned residual plots
- Linearity and independence assumptions met

- Sensitivity: 0.800
- Specificity: 0.885
- AUC: 0.910

Final Model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i};$$

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	-5.67	1.18	-4.82	921.42	<0.001
2	shapeoval	-0.27	0.32	-0.85	833.55	0.40
3	shapelobular	0.29	0.40	0.74	648.05	0.46
4	shapeirregular	1.09	0.36	3.06	506.00	<0.001
5	birads3	0.86	1.20	0.72	913.26	0.47
6	birads4	1.17	1.09	1.07	896.97	0.29
7	birads5	3.43	1.11	3.10	891.69	<0.001
8	birads6	2.99	1.35	2.21	823.44	0.03
9	age	0.05	0.01	5.64	935.24	<0.001
10	marginmicrolobulated	0.93	0.60	1.55	916.63	0.12
11	marginobscured	0.56	0.38	1.46	551.83	0.14
12	marginill-defined	0.88	0.33	2.68	370.75	<0.01
13	marginspiculated	1.08	0.41	2.65	479.14	<0.01

Baseline:

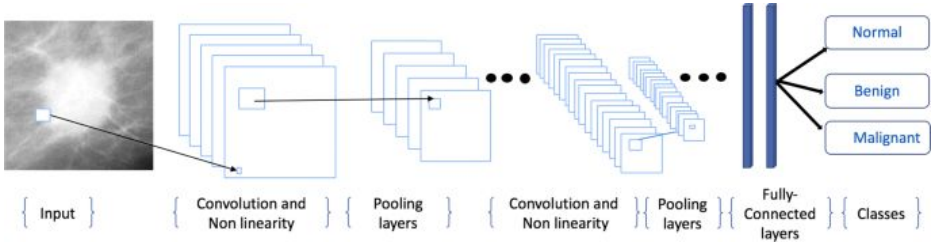
- Age of '0'
- BI-RADS score of 2
- Lesion with a round shape and
- Circumscribed margin

Interesting Findings

- BI-RADS score 🤔
  - Significant but redundant
- Physical attributes
  - Shape and Margin 😎
- Age is important 😉

Future Work

- Image data
- Convolutional Neural Network



Thank You