



# Predicting Congestive Heart Failure

Brandon Ryu, Douglas Hilton, Stephen Kita

# Why Healthcare Analytics

- Common interest in helping patients
- Career Opportunities
  - \$11.5 Billion market in 2019
  - \$40.8 Billion market projected in 2025
  - Compound Annual Growth Rate: 23.55%
  - Growing opportunities for data scientists
- Meaningful Use
  - Increase in volume
  - Increase in interoperability



# Why Congestive Heart Failure?

## Prevalence

- 6.2 Million adults in US
  - 10% of patient population
- 13.4% of deaths in 2018 mentioned heart failure
- \$30.7 Billion to treat CHF in 2012

## Survival Rate

- ~50% 5 year survival rate
- 5 year survival rates per stage (2007 study, 2000 subjects)
  - Stage A: 97%
  - Stage B: 95.7%
  - Stage C: 74.6%
  - Stage D: 20%

## Difficulty

- Hard to clearly diagnose with few labs
  - Acute Kidney Failure: diagnosed with few labs
- Gold Standard: Electrocardiogram (EKG)

# Data Source

## MIMIC IV v0.4

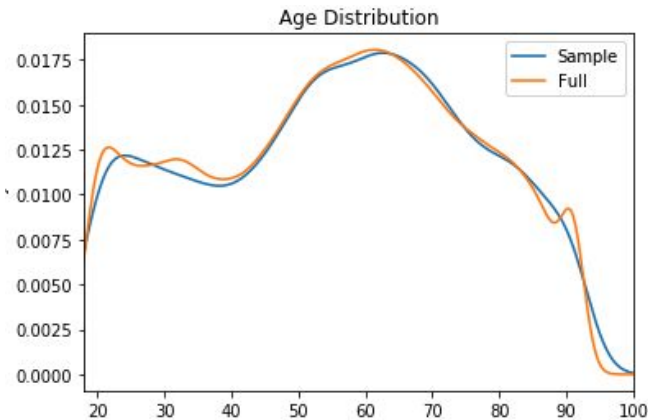
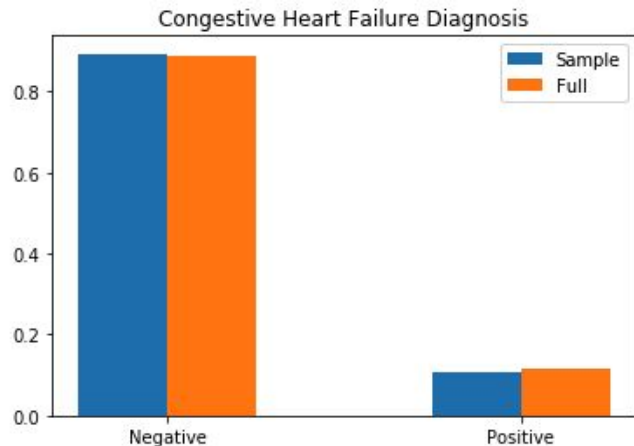
## Tables

## Rows

- |   |   |   |
|---|---|---|
| <ul style="list-style-type: none"><li>- Medical Information Mart for Intensive Care (MIMIC)</li><li>- Beth Israel Deaconess Medical Center (Boston, MA)</li><li>- Intensive Care Unit (ICU) and Emergency Department (ED) patients from 2008~2019</li><li>- Ethics course requirement</li></ul> | <p>Used 8 of 34 total tables</p> <ul style="list-style-type: none"><li>- Patients (sub_id)</li><li>- Admissions (sub_id, adm_id)</li><li>- Diagnoses (adm_id, diag_id)</li><li>- Labevents (adm_id, lab_id)</li><li>- Chartevents (adm_id, event_id)</li><li>- Dictionaries (lab_id, event_id, diag_id)</li></ul> | <ul style="list-style-type: none"><li>- 260,000 total patients<ul style="list-style-type: none"><li>- 200,000 adult patients</li></ul></li><li>- 530,000 hospital admissions<ul style="list-style-type: none"><li>- 460,000 adult admissions</li></ul></li><li>- 500,000,000 events</li></ul> |
|---|---|---|

# Data Processing

- Aggregate
  - First hospital admission with CHF diagnosis for patients with CHF
  - Random hospital admission for patients without CHF
  - Loss due to age
  - Loss due to lack of labs
- Full Data vs Current, Future Samples
  - Positive Diagnosis
    - 10.7% vs 11.4%, 11.9%
  - Patients
    - 257.4k vs 11.5k, 8k



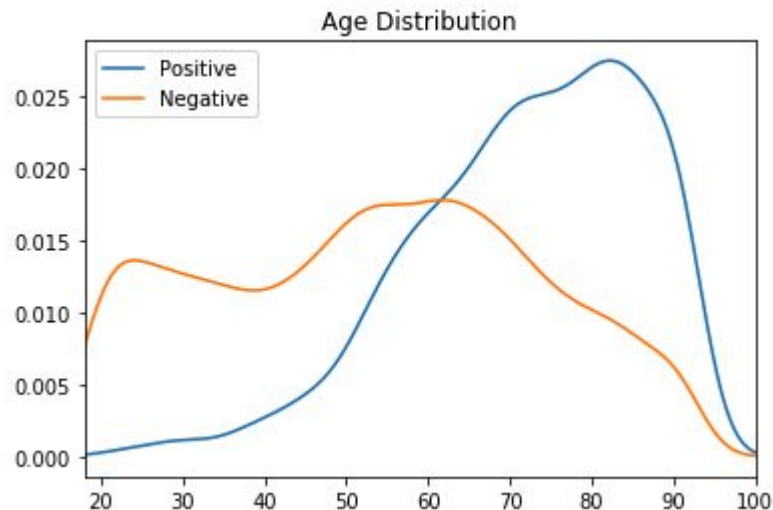
# Feature Engineering

- Labs Events
  - > 120m records
  - 1625 different tests
  - Lots of missingness
- Grab most relevant labs
  - 20 most common
  - 5 known to be related
- Aggregate
  - Min
  - Mean
  - Max
  - Abnormal %
  - Above Max
  - Below Min
- 150 Total Lab Event Features

	itemid	label	fluid	category	loinc_code	
	115	50868	Anion Gap	Blood	Chemistry	1863-0
	212	50882	Bicarbonate	Blood	Chemistry	1963-8
	222	51464	Bilirubin	Urine	Hematology	5770-3
	282	50893	Calcium, Total	Blood	Chemistry	2000-8
	442	50902	Chloride	Blood	Chemistry	2075-0
	511	50911	Creatine Kinase, MB Isoenzyme	Blood	Chemistry	6773-6
	512	50912	Creatinine	Blood	Chemistry	2160-0
	634	50920	Estimated GFR (MDRD equation)	Blood	Chemistry	NaN
	723	50931	Glucose	Blood	Chemistry	6777-7
	761	51221	Hematocrit	Blood	Hematology	4544-3
	771	51222	Hemoglobin	Blood	Hematology	718-7
	1008	50960	Magnesium	Blood	Chemistry	2601-3
	1013	51248	MCH	Blood	Hematology	785-6
	1014	51249	MCHC	Blood	Hematology	786-4
	1016	51250	MCV	Blood	Hematology	787-2
	1189	50970	Phosphate	Blood	Chemistry	2777-1
	1211	51265	Platelet Count	Blood	Hematology	777-3
	1233	50971	Potassium	Blood	Chemistry	2823-3
	1279	51274	PT	Blood	Hematology	5902-2
	1307	51277	RDW	Blood	Hematology	788-0
	1312	51279	Red Blood Cells	Blood	Hematology	789-8
	1373	50983	Sodium	Blood	Chemistry	2951-2
	1494	51003	Troponin T	Blood	Chemistry	6598-7
	1506	51006	Urea Nitrogen	Blood	Chemistry	3094-0
	1598	51301	White Blood Cells	Blood	Hematology	804-5

# Feature Engineering

- Additional Features
  - Comorbidities
  - Time in Emergency Department
  - Age
  - Gender
  - Ethnicity
  - Insurance
- Lots of data cleaning
  - Text parsing
  - Bad data entry
  - Differing mapping systems
- Standardize the data
- Check for multicollinearity
  - 173 total features
  - Variance Inflation Factor checks
  - Reduce features to 110-120 depending on the sample



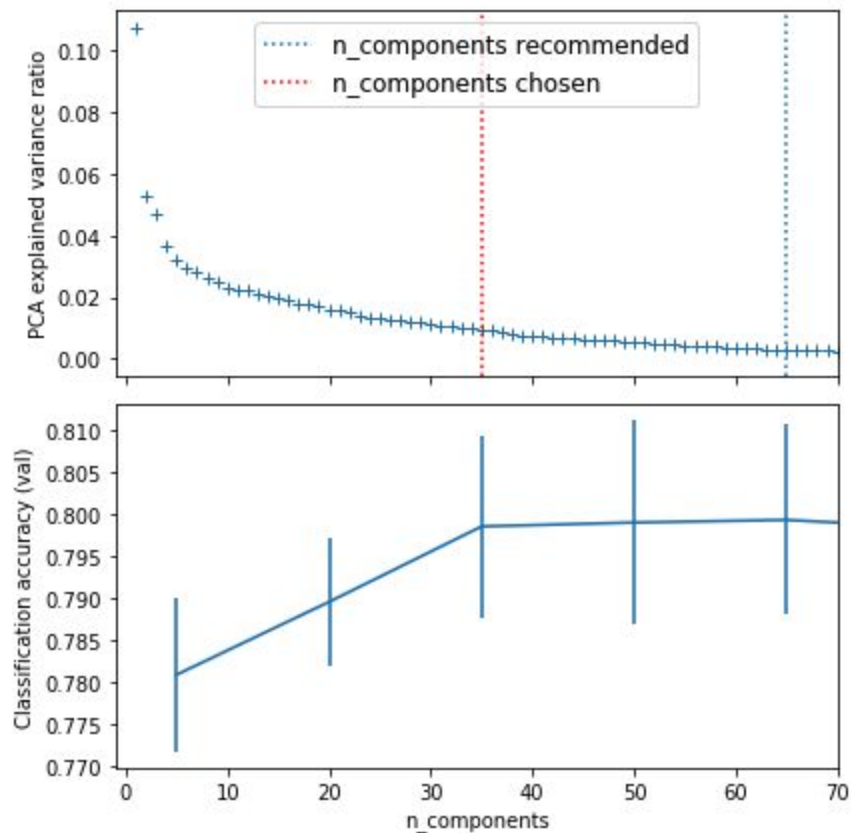
# Imputation Methods/Missing Data

- Data is Missing At Random
  - The reason the lab data is missing is because doctors don't need it
  - They either assume it's normal or they can infer based on other lab results
- Normal value imputation
  - Generate 50 normally distributed values around the middle of the normal range
  - Gives a very small chance of an abnormal value
- KNN imputation
  - $\sqrt{N}$  Nearest Neighbors
  - More likely to return abnormal values



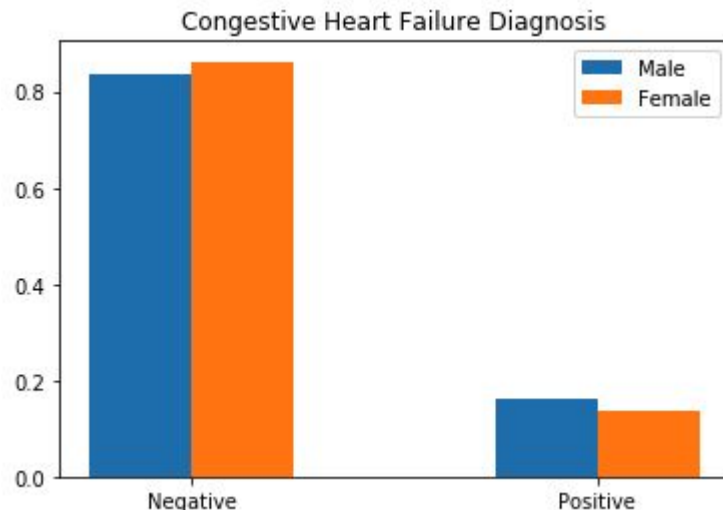
# Feature Selection

- Principal Component Analysis
  - Reduced feature set from 110 to 35
  - Harder to interpret
  - No better than other models
- Recursive Feature Elimination
  - Reduced feature set from 110 to 36, 60, 101
  - Made model less complex, for the most part



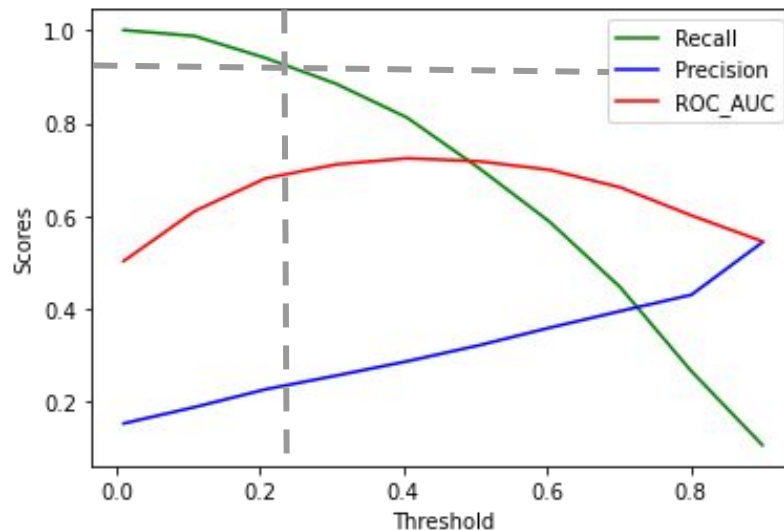
# Final Samples for Models

- Current Predictions
  - 11.5k Patients (1.3k Positive Diagnosis)
  - Logistic: 36 Features
  - Random Forest: 101 Features
  - XGBoost: 60 Features
- Gender Based Models
  - Male
    - 8.6k Patients (1.4k Positive Diagnosis)
    - Logistic: 63 Features
  - Female
    - 8.7k Patients (1.2k Positive Diagnosis)
    - Logistic: 77 Features
- Future Predictions
  - 8k Patients (0.9k Positive Diagnosis)
  - All Models: 78 Features



# Modeling Objective

- Predict Chronic Heart Failure:
  1. Current visit (present)
  2. Within 1 yr (future)
  3. Differences b/t Male/Female
- Interpretable Models:
  - Logistic Regression
  - Random Forest
- Accurate Models:
  - Gradient Boosting
- Tuning:
  - Area Under Curve (AUC)
  - Adjust threshold for 95% Recall



# Modeling Results

	<u>Present Model</u> (11.5K patients, 11.4% with disease)			<u>Future Model</u> (8K patients, 11.9% with disease)		
Score	Logistic (36 features)	Random Forest (101 features)	XG Boost (60 features)	Logistic (78 features)	Random Forest (78 features)	XG Boost (78 features)
<b>Recall</b>	95%	95%	95%	95%	95%	95%
<b>ROC/AUC</b>	68%	70%	71%	65%	71%	66%
<b>Accuracy</b>	49%	52%	55%	44%	55%	46%
<b>Precision</b>	22%	23%	24%	21%	25%	21%
<b>F1 Score</b>	36%	37%	38%	34%	39%	35%

# Feature Importance Comparison

	Present			Future		
Top Features	Logistic	Random Forest	XGBoost	Logistic	Random Forest	XGBoost
#1	Atherosclerosis	Insurance Medicare	Below Min Bicarbonate	Arrhythmia	Atherosclerosis	Atherosclerosis
#2	Age	Above Max Urea Nitrogen	Abnormal % Chloride	Ethnicity (Other)	Diabetes	Age
#3	Max Bicarbonate	Age	Abnormal % Calcium	Atherosclerosis	Age	Mean Urea Nitrogen
#4	Min Urea Nitrogen	Abnormal % Urea Nitrogen	Below Min Phosphate	Above Max Red Blood DW	Hypertension	Above max Urea Nitrogen
#5	Min Red Blood DW	Abnormal % Troponin T	Below Min Creatine Kinase	Age	Above Max Urea Nitrogen	Hypertension

# Demographic Comparisons (Present)

	Logistic Regression	
Score	Female (77 features)	Male (63 features)
<b>Recall</b>	95%	95%
<b>ROC/AUC</b>	72%	71%
<b>Accuracy</b>	55%	54%
<b>Precision</b>	22%	26%
<b>F1 Score</b>	36%	41%

Top Features	Female	Male
<b>#1</b>	Age	<b>Atherosclerosis</b>
<b>#2</b>	<b>Diabetes</b>	Age
<b>#3</b>	Hypertension	Max Bicarbonate
<b>#4</b>	Max Anion Gap	Max Anion Gap
<b>#5</b>	Max Bicarbonate	Ethnicity Asian (-)

# Conclusion

## Challenges

Dataset was:

- Sparse
- Messy
- Complicated

## Model Outcomes

Predicting present

- Similar to doctors
  - max/min labs

Gender

- Female: diabetes
- Male: atherosclerosis

Predicting Future

- Similar accuracy
- More comorbidities

## Applications

- Decision support tool for doctors
  - Additional tests
- Early warning tool for patients
  - Lifestyle changes

# Future Work

- Compare Ethnicities
- Ensemble different models
  - Sklearn voting classifier
- Incorporate different tables
  - Prescriptions
  - Chartevents
- Predict expire from CHF (severity)