Bayesian Modeling of Heart Disease

Cory Clayton, Luke Moles, Siddharth Surapaneni

Problem

- Heart Disease (HD) represents significant health risk
  - According to the CDC, around 655,000 people die from Heart Disease in the US every year
- Early diagnosis can help improve outcome
- Understanding and addressing uncertainty can lead to an improved understanding to the diagnosis of heart disease

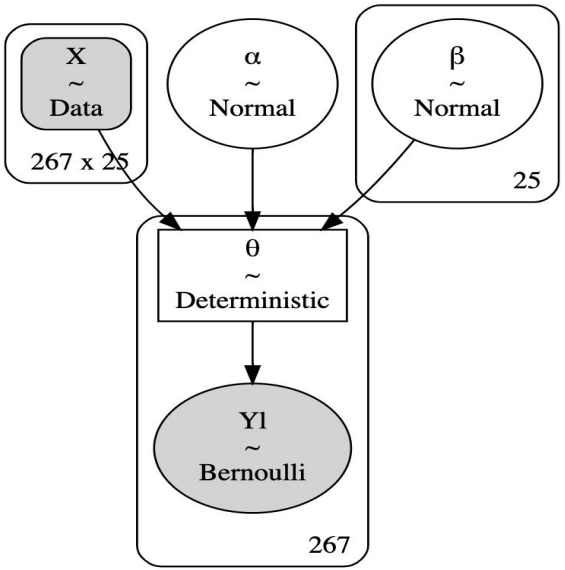
Dataset

- Sourced from UCI Machine Learning Repository
- Patients from four worldwide hospitals
  - Focus on 297 complete observations from the Cleveland Clinic Foundation
- 14 attributes including HD diagnosis level
  - Quantitative Attributes: Age, Resting Blood Pressure(trestbps), Serum Cholesterol Levels(chol),Maximum Heart Rate(thalach), ST depression(oldpeak)
  - Categorical Attributes: Sex, Chest Pain Type(cp), Fasting Blood Sugar(>120 mg/dl), Slope, Excercise Induced Angina(exang), Vessels Colored By Flouroscopy(ca), Thal, Rest ECG

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	hd
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	1
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0

Initial Modeling

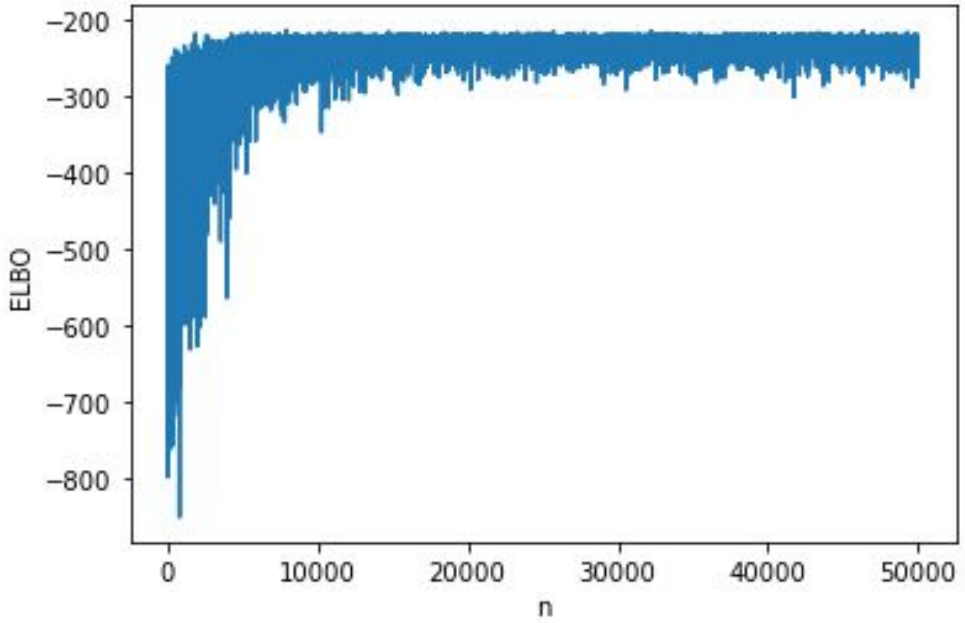
- 13 Predictors
- Scaled our quantitative variables
- Transformed response variable to a binary variable
- One Hot Encoded Categorical Predictors with more than 2 levels
    - e.g. Chest Pain Type, Slope
- Held out 30 observations for evaluation purposes
- Created a Bayesian Logistic Regression Model with Uninformed Gaussian Priors



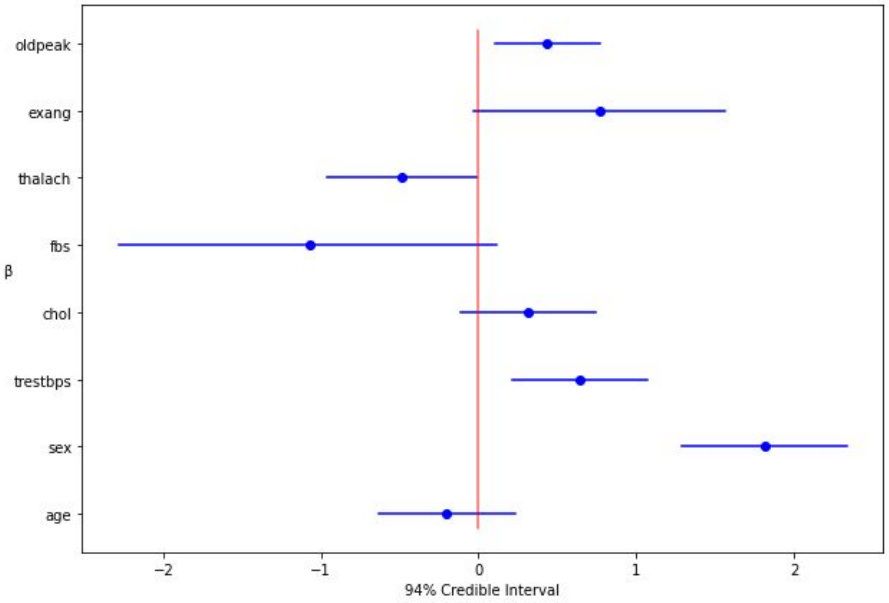
$$\ln \frac{p}{1-p} = \alpha + \beta_i X_i$$

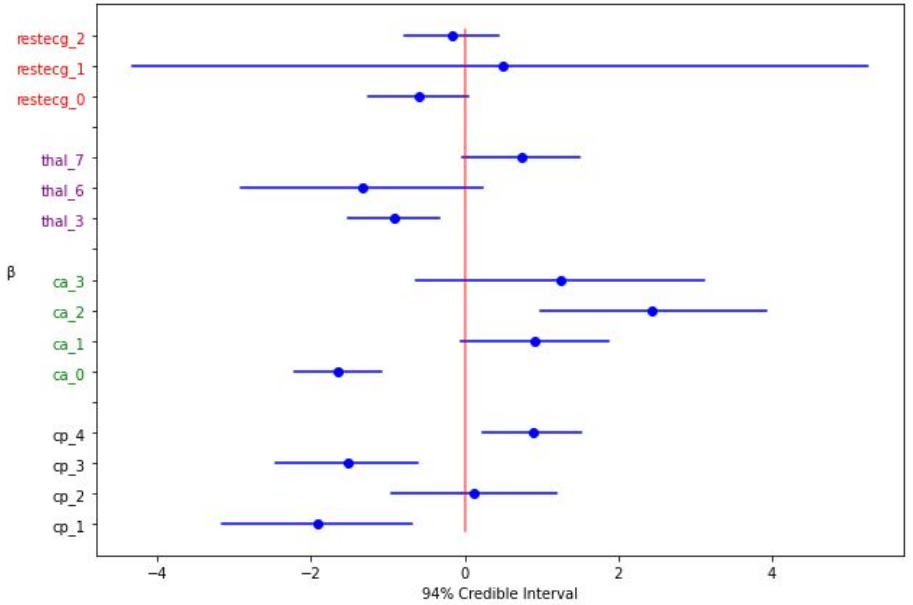
ELBO Plot

- Initially tried Sampling
- Sampling wasn't converging
- Switched to ADVI



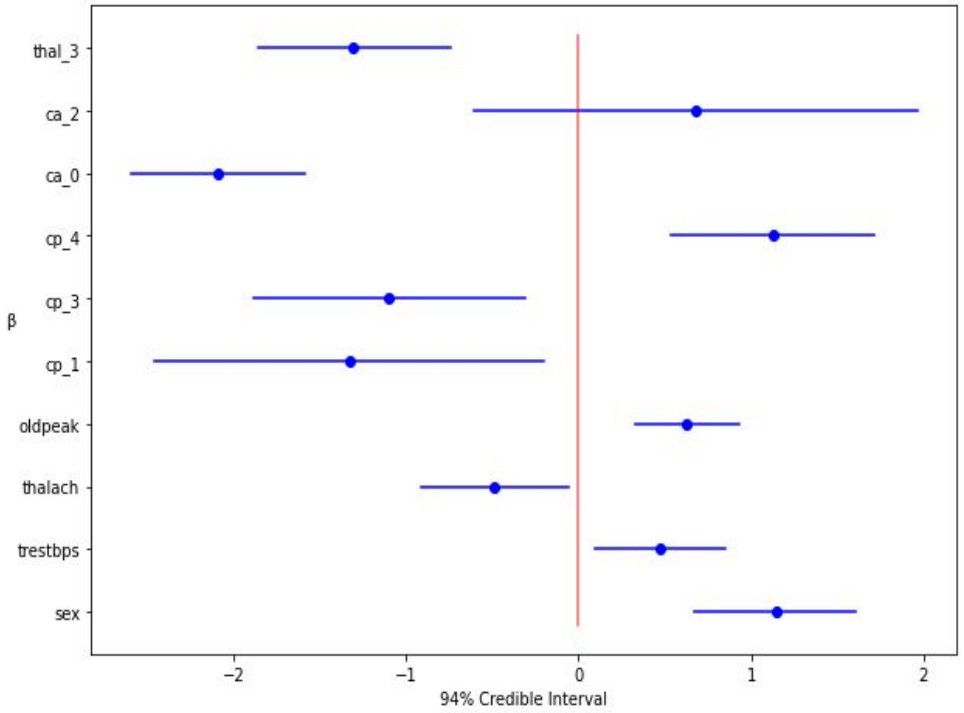
Forest Plots





Reduced Model

- Used the 10 predictors with significant credible intervals
- Used a logistic model



Model Selection

- Compared with a reduced model using 10 predictors
- Full Model WAIC of -123.7
- Reduced Model WAIC -107.7
- BMA 0.3% Full Model and 99.7% Reduced Model

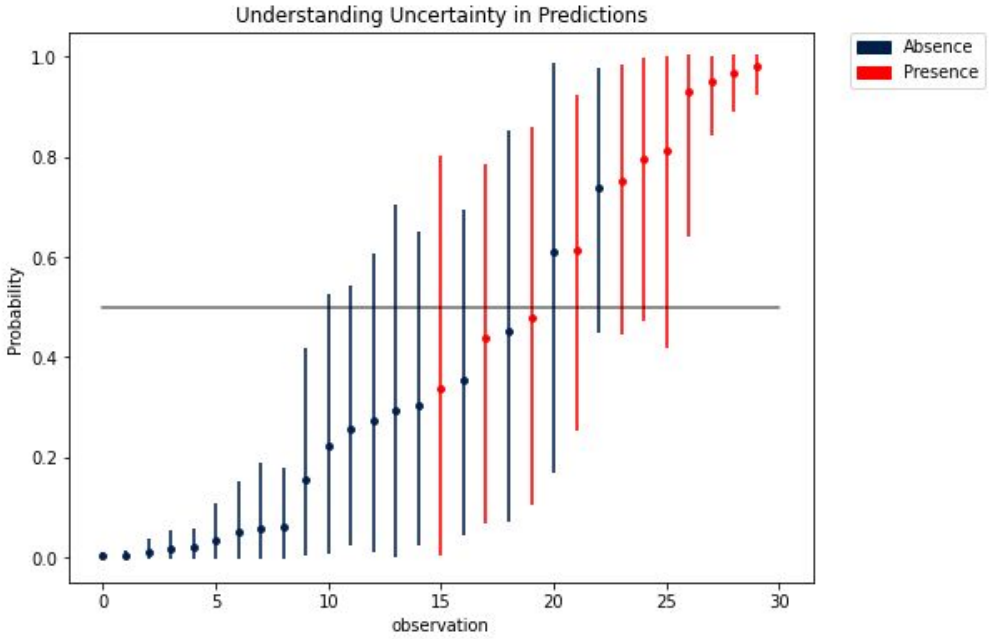
$$w_k = \frac{exp(-\frac{1}{2}\Delta_k)}{\sum_{i=1}^{K} exp(-\frac{1}{2}\Delta_i)}$$

$$\Delta_k = \text{WAIC}_k - \text{WAIC}^*$$

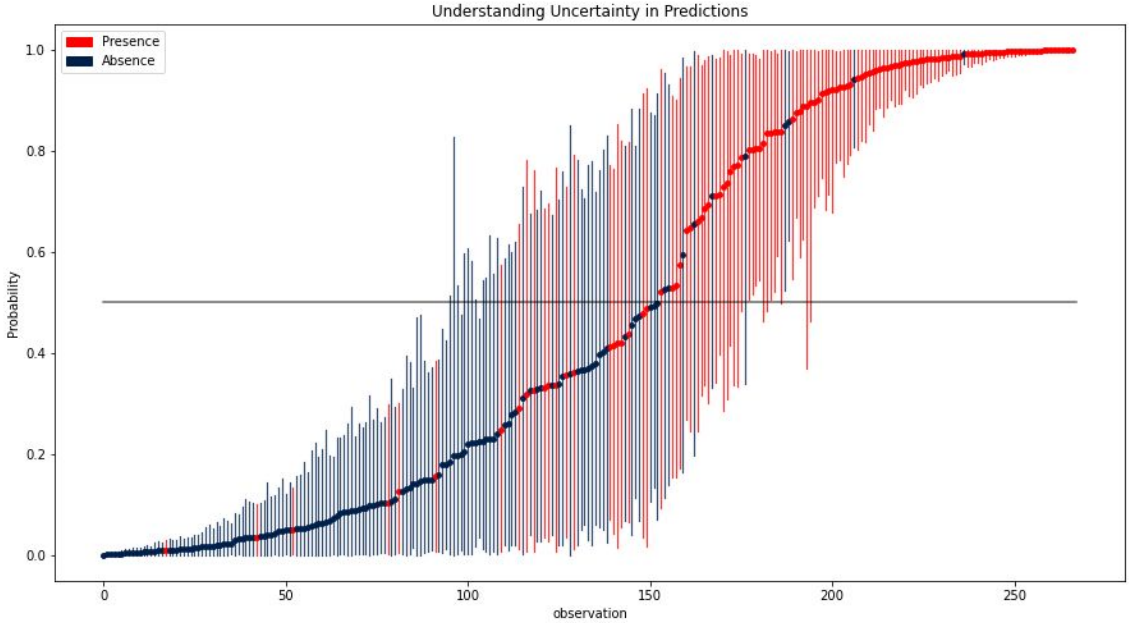
Results

- 83% accuracy on evaluation set
- 20% false positive
- 15% false negative

Confusion Matrix		Predicted Diagnosis	
		Absence	Presence
Actual Diagnosis	Absence	17	2
	Presence	3	8



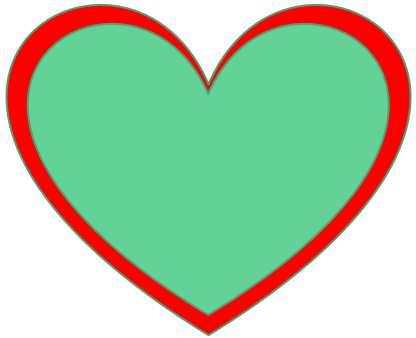
Analysis of Uncertainty



- Less accuracy among in-sample predictions with high uncertainty

Conclusions

- Able to identify uncertainty of observations to call for further investigation of Heart Disease
- Model still showed substantial uncertainty in predictors
- Future modeling could focus on hierarchical approach to certain variables
  - Sex
  - Chest pain
  - Resting ECG



References

https://archive.ics.uci.edu/ml/datasets/heart+disease

https://www.cdc.gov/heartdisease/facts.htm