Predicting Stroke Likelihood:
A Bayesian Analysis

Hyun Ko, Griffin McCauley, Eric Tria
DS 6040 - Bayesian Machine Learning
Dr. Don Brown

December 9, 2022
University of Virginia School of Data Science

Problem Description - Background & Motivation

Stroke is the 2nd leading cause of death according to the **WHO** (2021).

Commonly neglected and misdiagnosed as other conditions.

Ongoing studies regarding salient factors contributing to stroke risk.

Problem Description - The Data

~5000 observations with 10 clinical features

- Health metrics: *bmi*, *avg_glucose_level*
- Pre-existing conditions: *hypertension*, *heart_disease*, *smoking_status*
- Demographic information: *gender*, *age*, *ever_married*, *work_type*, *residence_type*

Data originally curated and distributed over Kaggle.

Probability Model - Bayesian Logistic Regression

Binary classification (stroke: 0/1)

Some possible models:

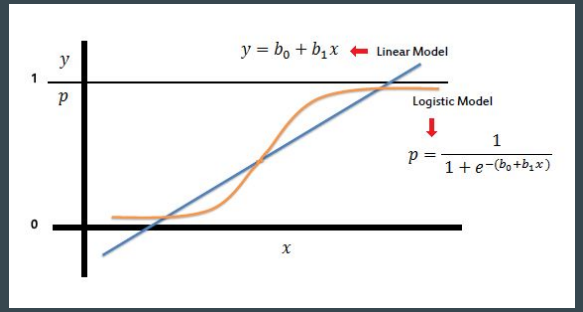
- LDA, QDA, Naïve Bayes, SVM

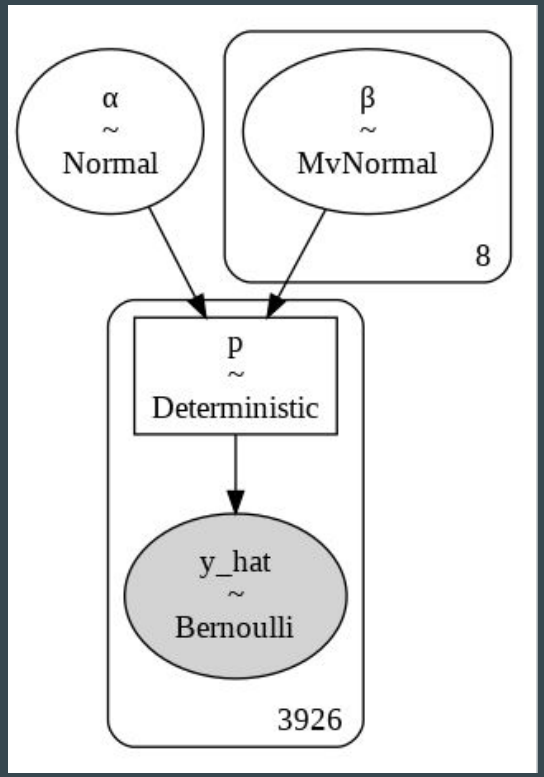
Numeric and categorical features

Bayesian Logistic Regression

- Inference
- Prediction
- Uncertainty

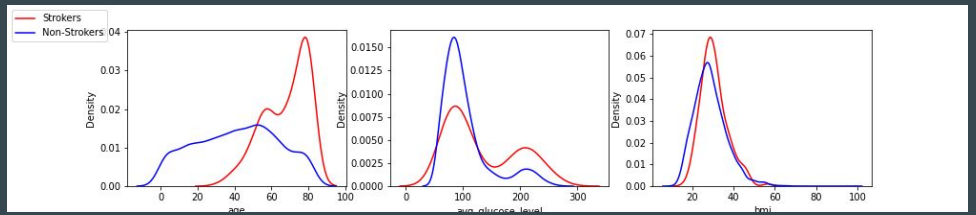
$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

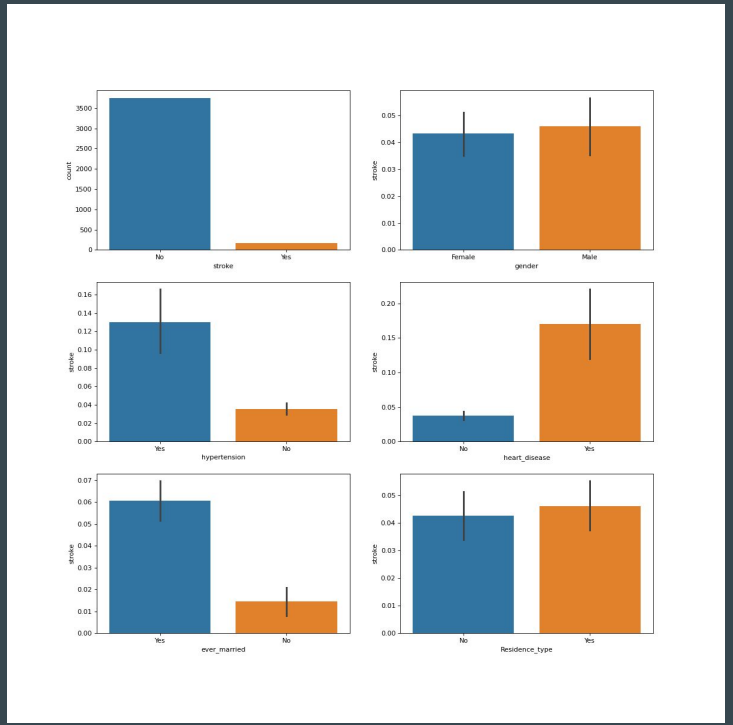




Approach - Exploratory Data Analysis

- Assessed the relationship between each of the predictors and the target variable.
- Potentially notable features:
  - *age*
  - *hypertension*
  - *heart_disease*
  - *ever_married*





Approach - Data Cleaning

Dropped 201 observations containing NAs

Features removed:

- *id* (unique patient identifier)
- *work_type* and *smoking_status* (multi-class categorical predictors)

Features converted into binary variables:

- *gender* (male: 1, female: 0)
- *ever_married* (yes: 1, no: 0)
- *Residence_type* (urban: 1, rural: 0)

Final features:

- 5 binary variables (*gender*, *hypertension*, *heart_disease*, *ever_married*, *Residence_type*)
- 3 numeric variables (*age*, *avg_glucose_level*, *bmi*)

80/20 Train-Test Split

- 3926 training observations
- 982 test observations

Approach - General Trajectory

Full Model

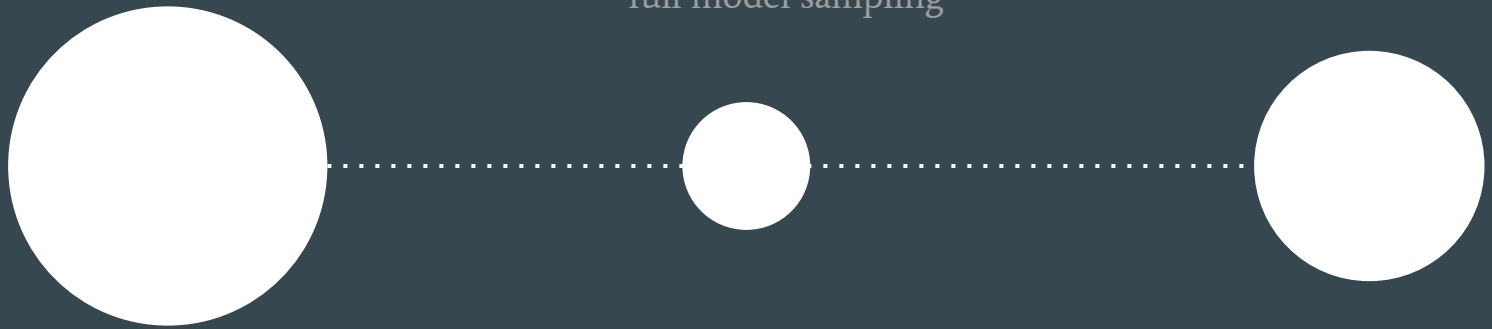
Uses all available features which can be incorporated effectively

Reduced Model

Uses a subset of features determined by the results of the full model sampling

BMA Model

Weights the predictions of the other models based on their WAIC scores



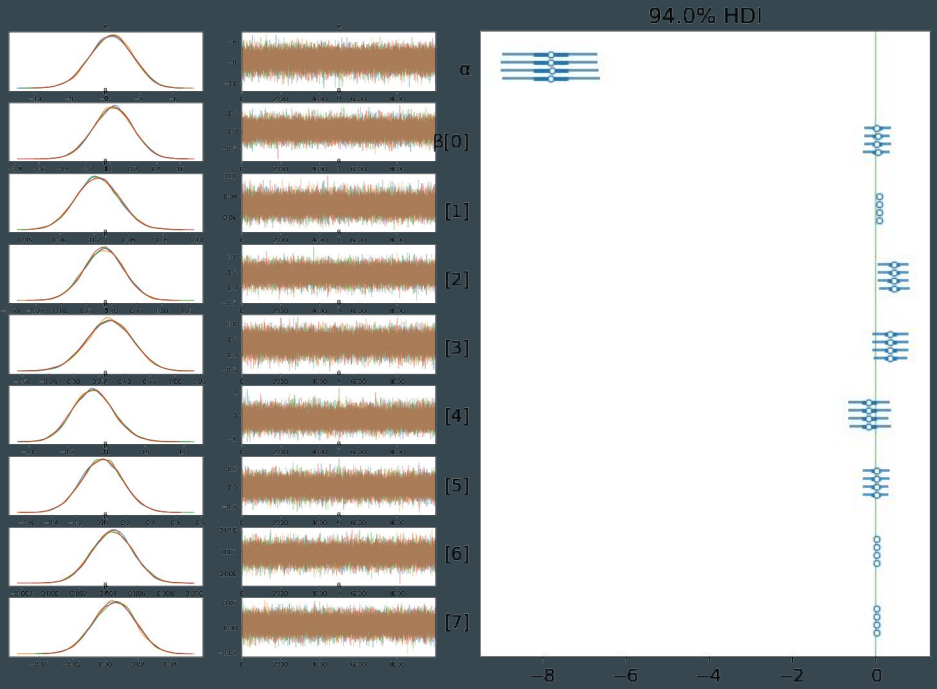
Approach - The Full Model

Predictor variables:

- 5 binary categorical
- 3 numeric

HMC Sampling

- Very slow process
- Good trace plot convergence
- Forest plots and summary table revealed many predictors were not statistically significant



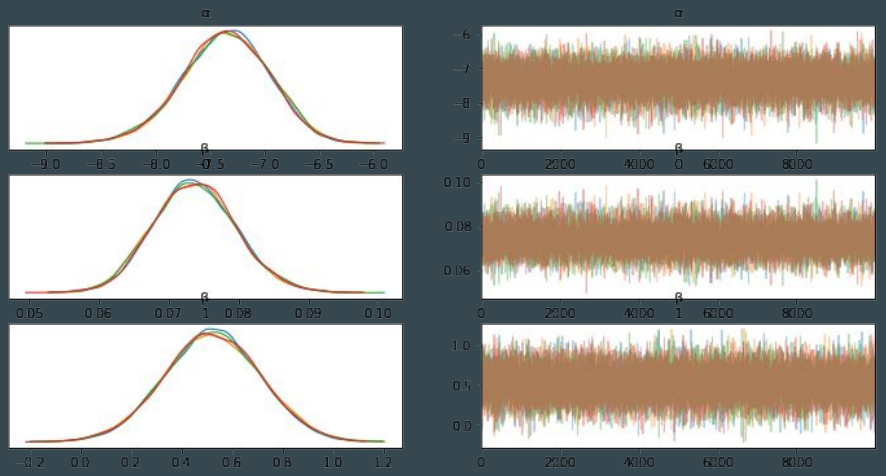
Approach - The Reduced Model and BMA Model

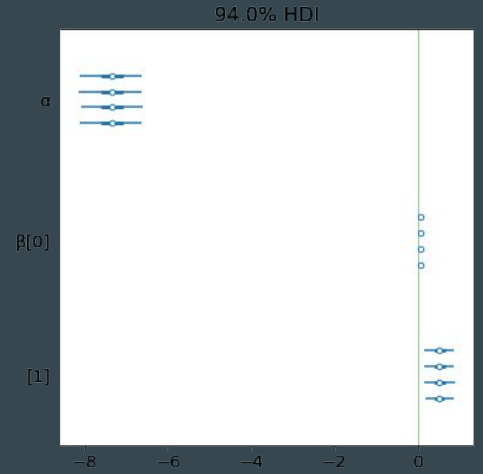
2 predictor variables: *age* & *hypertension*

Strong convergence and statistically significant posterior parameter estimates

WAIC comparison:

- 55.94% full model
- 44.06% reduced model





Modeling Objectives

Minimizing Misclassification Cost

Lower Threshold

- Liberal on diagnosing as ‘positive’ (0.04)
- Impact of misdiagnosing as ‘negative’ (False Negative) is huge
- FP cost = 1, FN cost = 50

Focusing on Recall than Accuracy and Precision

- Huge class imbalance on response variable
- What proportion of actual positives identified correctly?

Inferences on Variability

The Main Goal of Bayesian Statistics

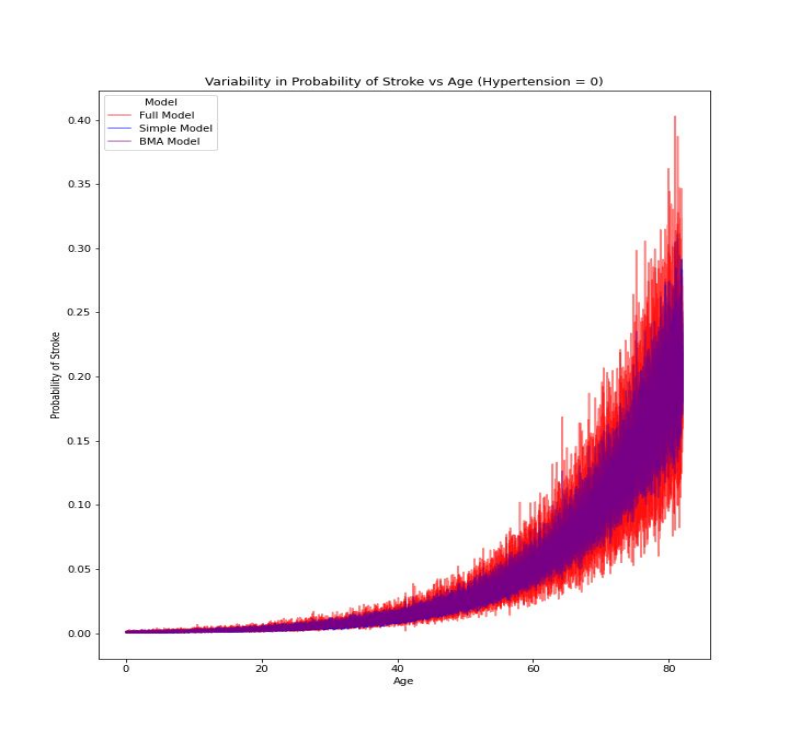
- Identify uncertainty in each model
- Checking mean of posterior graph to get pointwise probability

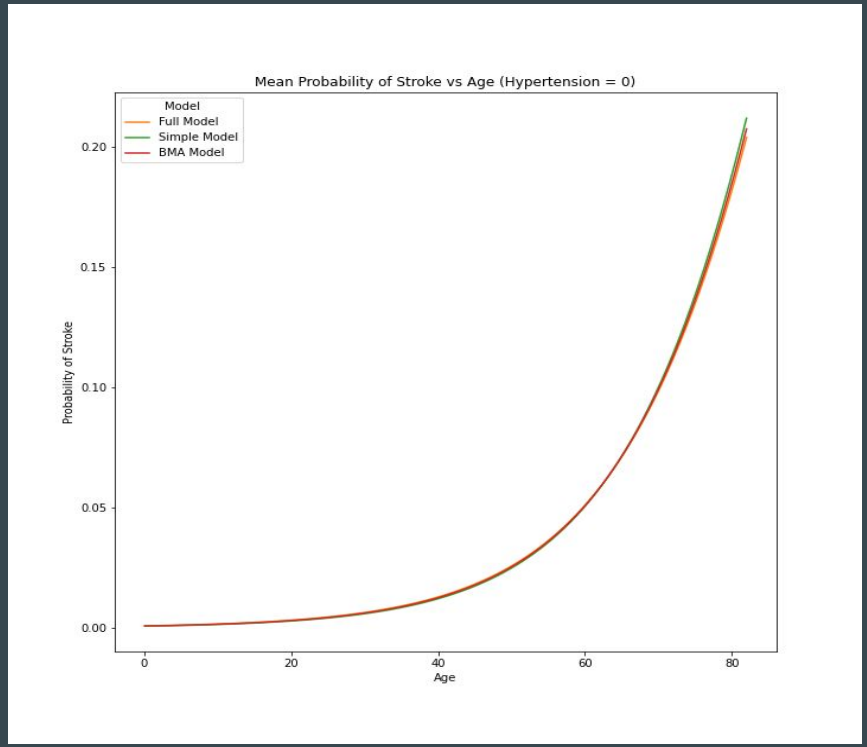
Results

Models / Metrics	AUC	Confusion Matrix	Precision	Accuracy	Recall
Full Sampling	0.854	[[33, 2], [280, 667]]	0.105	0.713	0.943
Reduced Sampling	0.836	[[33, 2], [306, 641]]	0.097	0.686	0.943
BMA	0.847	[[33, 2], [292, 655]]	0.102	0.701	0.943

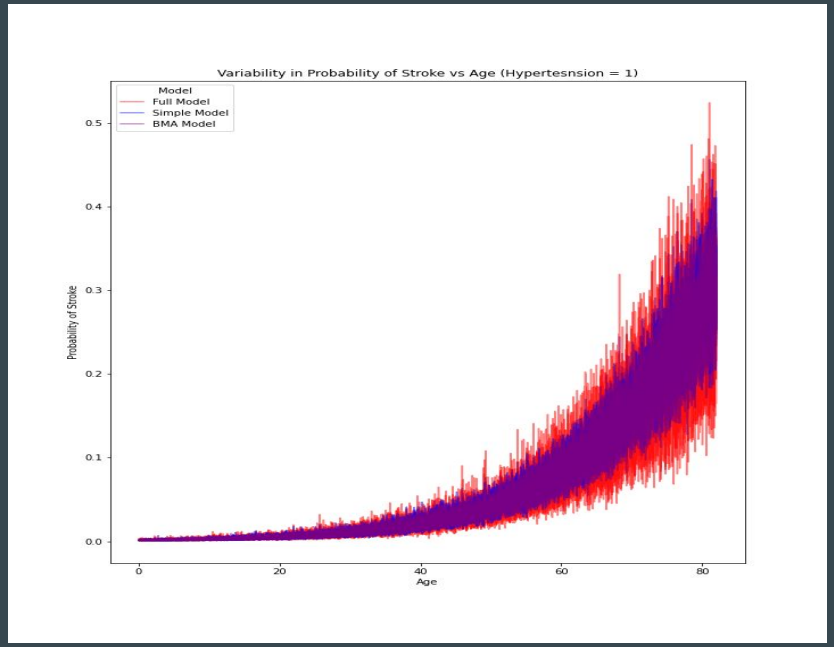
- Precision = TP / (TP+FP) = What proportion of positive identifications are actually correct?
- Recall = TP / (TP + FN) = What proportion of actual positives was identified correctly?

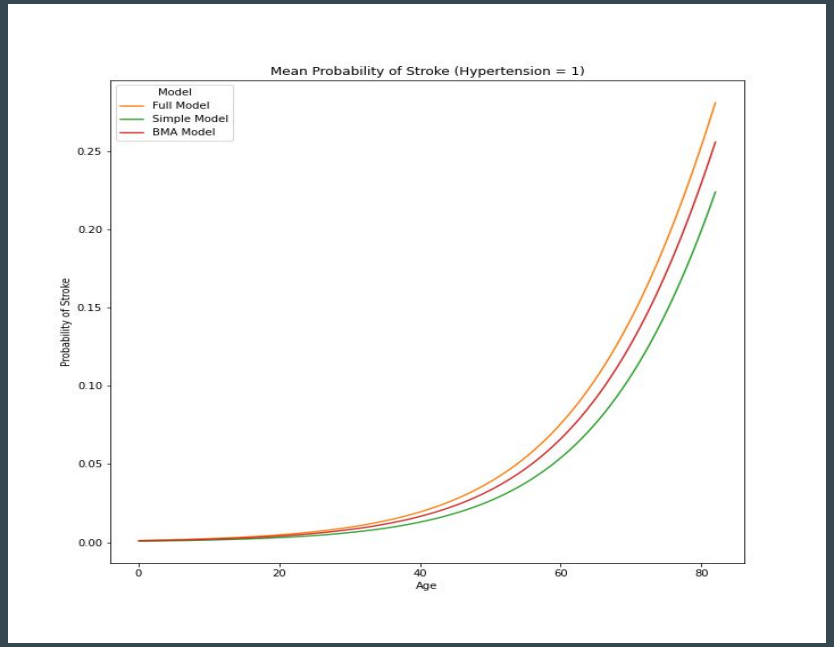
Result: Variability in Age & No Hypertension





Result: Variability in Age & Hypertension





Limitations

Observations	• The dataset only has roughly 5000 observations
Features	• The dataset only has 10 possible predictors
Regional Factors	• Did not take into account possible differences in regions, countries, etc.

Conclusions:

- Bayesian Logistic Regression can provide decent predictions for classifying patients as with or without stroke.
- Bayesian Model Averaging is effective in building a final model that balances prediction variance and overall performance.

The Team

Hyun Ko

M.S. Data Science

B.S. Data Science

Griffin McCauley

M.S. Data Science

Sc.B. Applied Mathematics,
A.B. Economics

Eric Tria

M.S. Data Science

B.S. Computer Science

References

Boehme, A., Esenwa, C., Elkind, M., (2017, February 3). Stroke Risk Factors, Genetics, and Prevention. National Library of Medicine. Retrieved December 6, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5321635/

Fallik, D., (2017, April 6). Nearly 10 Percent of Strokes Are Misdiagnosed (At Least Initially) in the Emergency Department. Neurology Today. Retrieved December 6, 2022, from https://journals.lww.com/neurotodayonline/fulltext/2017/04060/Nearly_10_Percent_of_Strokes_Are_Misdiagnosed__At.8.aspx

Singh, P. (2021, October 28). World Stroke Day. World Health Organization. Retrieved December 6, 2022, from https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day

Soriano, F., (2021, January 26). Stroke Prediction Dataset. Kaggle. Retrieved December 6, 2022, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

The GBD 2016 Lifetime Risk of Stroke Collaborators, (2018, December 20). Global, Regional, and Country-Specific Lifetime Risks of Stroke, 1990 and 2016. The New England Journal of Medicine. Retrieved December 7, 2022, from https://www.nejm.org/doi/full/10.1056/NEJMoa1804492