STATS205P Bayes Data Analysis Final Project

Chuqi Wang, Shengtong Sun

June 14, 2024

Abstract

Stroke is a significant public health concern and one of the leading causes of death and long-term disability globally[3]. Early prediction and prevention are vital in reducing associated health burdens. This project aims to predict stroke occurrence using a Bayesian Data Analysis approach with logistic regression via the stan_glm function from the R rstanarm package. The analysis incorporates prior knowledge and quantifies uncertainty in predictions, complemented by exploratory data analysis (EDA) to understand underlying patterns and relationships within the data. The dataset includes demographic, medical, and lifestyle information of patients, facilitating insights into factors contributing to stroke events. The Bayesian logistic regression model achieved an overall test accuracy of 95.2% but initially struggled with sensitivity, failing to identify stroke cases. Adjusting the decision threshold to 0.3 improved sensitivity, correctly identifying more stroke cases while maintaining a high overall accuracy of 94.61%. These results highlight the model's potential in predicting stroke risk and guiding healthcare interventions.

1 Introduction

Stroke is a major public health concern, being one of the leading causes of death and long-term disability worldwide. Early prediction and prevention are crucial in reducing the associated health burdens. In this report, we aim to predict the likelihood of stroke occurrence in individuals using a Bayesian Data Analysis approach. We will utilize the stan_glm function from the R rstanarm package to perform logistic regression, allowing us to incorporate prior knowledge and quantify uncertainty in our predictions. This approach will be complemented by exploratory data analysis (EDA) to gain insights into the underlying patterns and relationships within the data.

2 Data Overview

The dataset[1] contains comprehensive information on patients, including demographic, medical, and lifestyle details, to understand factors contributing to stroke events. The dataset comprises 12 columns, each representing different attributes of the patients, and aims to provide insights for predicting stroke occurrences.

2.1 Data Explanation

1.id: This column serves as a unique identifier for each patient, ensuring that each entry in the dataset is distinct. 2. gender: This categorical variable indicates the gender of the patient, with possible values being 'Male' and 'Female'. Gender is an important demographic factor that could influence health outcomes. 3. age: A numerical variable representing the age of the patient. Age is a critical factor in medical studies as it correlates with the risk of various health conditions, including stroke. 4. hypertension [2]: This binary variable indicates whether the patient has hypertension (high blood pressure), with '1' representing the presence of hypertension and '0' representing its absence. Hypertension is a known risk factor for stroke. 5.heart_disease: Another binary variable that denotes whether the patient has any heart disease, with '1' indicating the presence and '0' indicating the absence of heart disease. The presence of heart disease is significant in evaluating stroke risk. 6. ever_married: This categorical variable indicates if the patient has ever been married, with possible values being 'Yes' and 'No'. Marital status can be an indirect indicator of social and emotional support, which may

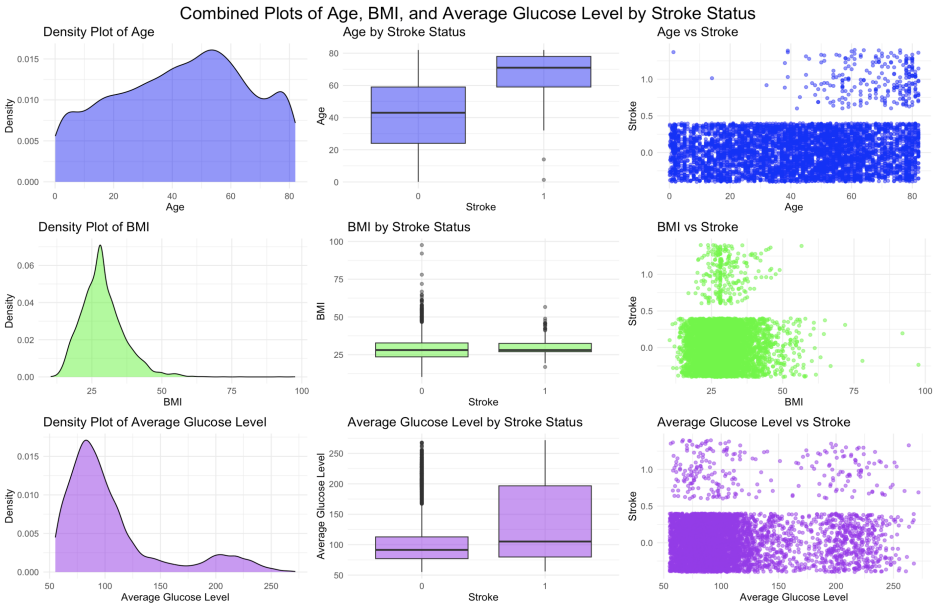

Figure 1: Combined Plots of Age, BMI, and Average Glucose Level by Stroke Status

impact health. 7. work_type: This categorical variable describes the type of work the patient is engaged in, with categories including 'Private', 'Self-employed', 'Govt_job', 'children', and 'Never_worked'. The nature of one's occupation can affect lifestyle and health outcomes. 8. Residence_type: A categorical variable that indicates whether the patient lives in an 'Urban' or 'Rural' area. Residence type may influence access to healthcare and lifestyle choices. 9. avg_glucose_level: This numerical variable records the patient's average glucose level in the blood. Elevated glucose levels are associated with diabetes, which is a risk factor for stroke. 10. bmi: Body Mass Index (BMI) is a numerical variable representing the patient's weight in relation to their height. BMI is a widely used indicator of body fat and an important health metric. 11. smoking_status: This categorical variable describes the patient's smoking habits with categories including 'formerly smoked', 'never smoked', 'smokes', and 'Unknown'. Smoking is a well-known risk factor for many health conditions, including stroke. 12. stroke: The target binary variable indicating whether the patient has had a stroke, with '1' indicating a stroke and '0' indicating no stroke. This variable is the primary focus of the analysis.

2.2 Data Cleaning

Upon initial inspection, it is noted that the 'bmi' column contains 201 missing values. Handling these missing values is crucial for accurate analysis. For this purpose, we will use the median value to fill in the missing values. In addition, smoking_status has unknown values, but we will leave it and treat it as another category in our analysis.

2.3 Data Analysis

In figure 1, it shows a a comprehensive overview of the distributions and relationships between numerical variables: age, BMI, average glucose level, and stroke status. The plots are divided into density plots, box plots, and scatter plots to facilitate a deeper understanding of how these variables interact and influence stroke occurrences.

Age and Stroke Status: The density plot of age shows a relatively even distribution with peaks around 60-70 years. The box plot comparing age by stroke status reveals that individuals who have had a stroke tend to be older, with a median age significantly higher than those who have not experienced a stroke. The scatter plot of age versus stroke status confirms this, showing a higher concentration of stroke cases among older age groups.

BMI and Stroke Status: The density plot of BMI indicates that most individuals have a BMI between 20 and 40, with a noticeable peak around 25-30. When comparing BMI by stroke status, the box plot shows that there is no significant difference in BMI between those who have had a stroke and those who have not, though there are some outliers with extremely high BMI values. The scatter plot of BMI versus stroke status does not reveal a clear pattern, suggesting that BMI alone may not be a strong predictor of stroke in this dataset.

Average Glucose Level and Stroke Status: The density plot for average glucose level shows a skewed distribution with a peak around 75-100 and a long tail extending towards higher glucose levels. The box plot comparing average glucose level by stroke status indicates that individuals who have had a stroke tend to have higher average glucose levels. This is supported by the scatter plot, which shows a higher concentration of stroke cases among individuals with elevated glucose levels.

Overall, the visualizations highlight age and average glucose level as more prominent factors associated with stroke occurrence. Older age and higher average glucose levels are correlated with a higher likelihood of stroke, while BMI does not show a strong correlation in this dataset. These insights can guide further analysis and modeling efforts, such as Bayesian analysis, to quantify the impact of these variables on stroke risk and develop predictive models for better healthcare interventions.

On the other hand, the visualizations in Figures 2 and 3 provide a detailed overview of the distribution of categorical variables in the dataset used for both testing and training. The gender distribution shows a higher count of females (2,994) compared to males (2,115). In terms of hypertension, a significant majority of individuals do not have hypertension (4,612) versus those who do (498). The work type variable is dominated by individuals working in private sectors (2,925), followed by self-employed (819), government jobs (657), children (687), and a minimal number of those who never worked (22). The residence type is nearly evenly split between rural (2,514) and urban (2,596) areas. For heart disease, most individuals do not have heart disease (4,834) compared to a smaller group who do (276). The marital status shows a majority who have been married (3,353) against those who have not (1,757). Smoking status varies widely, with 'never smoked' being the most common (1,892), followed by 'Unknown' (1,544), 'formerly smoked' (885), and 'smokes' (789). Finally, the stroke variable is highly imbalanced, with a vast majority not having experienced a stroke (4,861) compared to those who have (249). These distributions highlight the imbalances and diversity within the dataset, which are crucial considerations for the analysis and modeling processes.

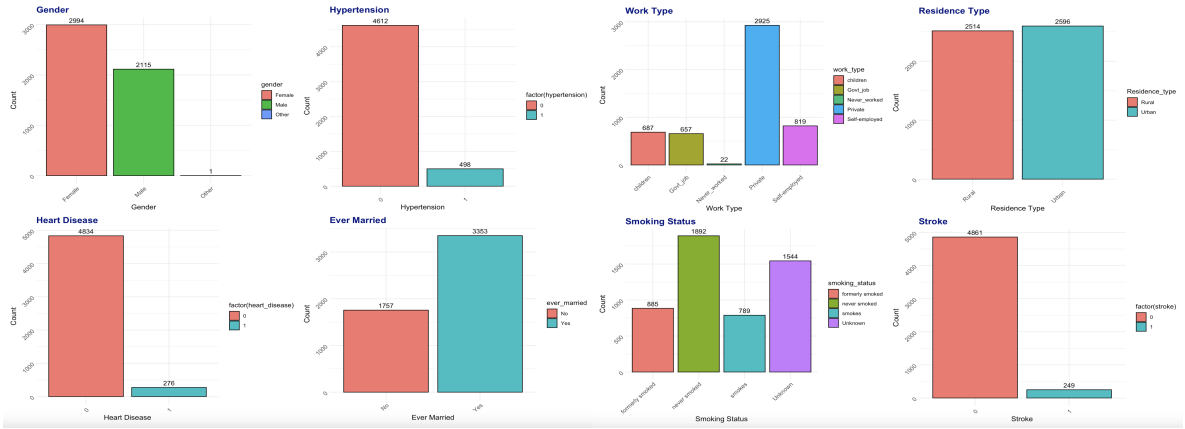

Figure 2: Bar Plot for Stroke Count in Gender, Hypertension, Heart Disease, Ever Married

Figure 3: Bar Plot for Stroke Count in Work Type, Residence Type, Smoking Status

3 Model Analysis

We will use Generalized Linear Model and the 'brms' package is loaded to facilitate Bayesian regression modeling for the binary response. Priors for the model are defined: for coefficients, the prior is set to be a normal distribution with mean 0 and standard deviation 10; for the intercept, the prior is set to

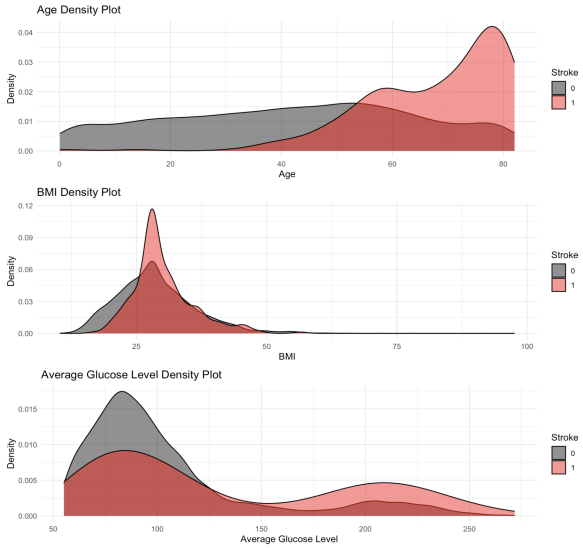

Figure 4: Density Plots for Age, BMI, Average Glucose Level

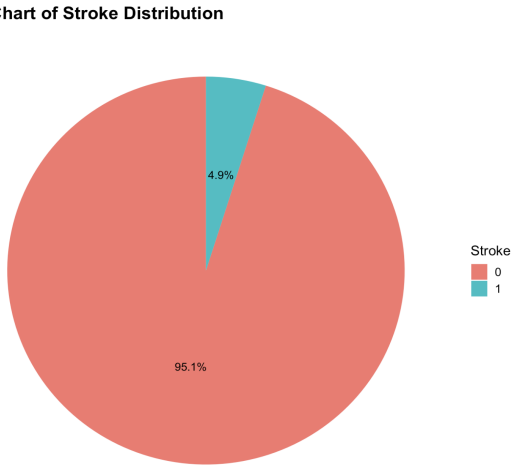

Figure 5: Pie Chart for Counting the Stroke People

be a Student's t-distribution with 3 degrees of freedom, mean 0, and scale 10. The priors for intercept and the coefficients are shown below.

$$\beta_0 \sim t(\nu = 3, \mu = 0, \sigma = 10)$$

$$\beta_i \sim \mathcal{N}(0, 10)$$

The logistic regression model with these priors can be expressed as:

$$\begin{aligned}\text{logit}(P(\text{stroke} = 1)) = \beta_0 &+ \beta_1 \cdot \text{genderMale} + \beta_2 \cdot \text{genderOther} \\ &+ \beta_3 \cdot \text{age} + \beta_4 \cdot \text{hypertension1} + \beta_5 \cdot \text{heart\_disease1} \\ &+ \beta_6 \cdot \text{ever\_marriedYes} + \beta_7 \cdot \text{work\_typeGovt\_job} \\ &+ \beta_8 \cdot \text{work\_typeNever\_worked} + \beta_9 \cdot \text{work\_typePrivate} \\ &+ \beta_{10} \cdot \text{work\_typeSelfEmployed} + \beta_{11} \cdot \text{Residence\_typeUrban} \\ &+ \beta_{12} \cdot \text{avg\_glucose\_level} + \beta_{13} \cdot \text{bmi} \\ &+ \beta_{14} \cdot \text{smoking\_statusneversmoked} + \beta_{15} \cdot \text{smoking\_statussmokes} \\ &+ \beta_{16} \cdot \text{smoking\_statusUnknown}\end{aligned}$$

where

$$\text{logit}(P(\text{stroke} = 1)) = \log\left(\frac{P(\text{stroke} = 1)}{1 - P(\text{stroke} = 1)}\right)$$

The Bayesian logistic regression model is fitted to the training data, and the model uses 2000 iterations for the chain, 4 Markov chains and a random seed of 123 for reproducibility. Roughly 80% of the data are used as training (4088 units) and the rest of 1022 are used for testing. The ten covariates introduced in the data description are all included in the model.

The table presented contains the regression coefficients from the Bayesian logistic regression model predicting the probability of having a stroke. The intercept estimate is -6.94, with a standard error of 0.89. This indicates the log-odds of having a stroke when all predictor variables are at their reference levels or zero. The 95% credible interval for the intercept ranges from -8.86 to -5.43, suggesting that the intercept is significantly different from zero. The coefficient for 'genderMale' is 0.06 with a 95% credible interval of -0.25 to 0.37, suggesting that being male does not significantly change the odds of having a stroke compared to the reference category (female). The 'genderOther' category has an

estimate of -4.97, but with a large standard error (7.03) and a wide credible interval (-21.03 to 5.51), indicating high uncertainty around this estimate. The coefficient for age is 0.07, with a narrow credible interval (0.06 to 0.09). This suggests that for each additional year of age, the log-odds of having a stroke increase by 0.07, indicating a positive and significant association between age and stroke risk. In addition, both hypertension and heart_disease have positive coefficients (0.26), with their credible intervals including zero (-0.13 to 0.64 for hypertension and -0.17 to 0.67 for heart disease). This implies that while there is a positive association, the evidence is not strong enough to conclude a significant effect. The coefficient for ever_marriedYes is -0.18 with a credible interval of -0.66 to 0.35, suggesting no significant difference in stroke risk between those who are married and those who are not. For work type, most of these estimates have credible intervals that include zero, indicating uncertainty in these effects. The estimate for Residence_typeUrban is 0.15, with a credible interval of -0.16 to 0.46, suggesting no significant difference in stroke risk between urban and rural residents. For health metrics, again, the credible intervals for these estimates include zero, indicating uncertainty in the effects.

4 Diagnostics

The Bulk ESS is a diagnostic tool used to assess how well the central part of the posterior distribution is sampled. In Bayesian analysis, we use Markov Chain Monte Carlo (MCMC) methods to draw samples from the posterior distribution. The values range from about 1,345 to 5,858. These values are sufficiently large, indicating good mixing and reliable estimates for the central part of the posterior distributions.

The Tail ESS focuses on the tails of the posterior distribution. Accurate estimation of the tails is crucial for understanding the behavior of extreme values, which can be important for risk assessment and decision-making. The values range from about 1,500 to 3,195. These values suggest that the tails are adequately sampled, providing confidence in the accuracy of extreme quantile estimates. Besides, it is worth noting that all Rhat values are 1.00, indicating that Markov Chains have converged well.

In figure 6, the trace plots for the model parameters provide a visual diagnostic of the convergence and mixing quality of the Markov Chain Monte Carlo (MCMC) sampling process used in the Bayesian logistic regression model. Each subplot represents the sampling trajectory of a specific parameter across four chains, allowing us to assess whether the chains have converged to a common distribution and mixed well. Each trace plot displays the sampled values of a parameter over 1000 iterations for four different chains, represented by different shades of blue. Good convergence is indicated when the chains overlap and exhibit similar distributions. In these plots, all chains for each parameter appear to have converged well, as they mix uniformly without distinct separation, suggesting that the chains are sampling from the same distribution.

The trace plots indicate that the intercept, gender (both genderMale and genderOther), and age parameters show stable and well-mixed chains, suggesting good convergence and reliable estimates. Hypertension (hypertension1) and heart disease (heart_disease1) parameters also demonstrate good mixing. Marital status (ever_marriedYes) and work types (work_typeGovt_job, work_typeNever_worked, work_typePrivate, and work_typeSelfEmployed) show well-mixed chains, although work_typeNever_worked has higher variability. Parameters related to residence type (Residence_typeUrban), average glucose level, BMI, and smoking status (smoking_statusneversmoked, smoking_statussmokes, and smoking_statusUnknown) exhibit good mixing and convergence, ensuring reliable posterior estimates.

The trace plots support the high Bulk ESS and Tail ESS values reported in the coefficient summary. The consistent overlap and lack of distinct trends or patterns across chains suggest that the effective sample sizes are sufficient, providing confidence in the reliability of the parameter estimates. The trace plots for the log prior (lprior) and log posterior (lp) demonstrate good mixing, with the chains overlapping well. This indicates that the overall model fitting process was stable and that the prior and posterior distributions are well-sampled.

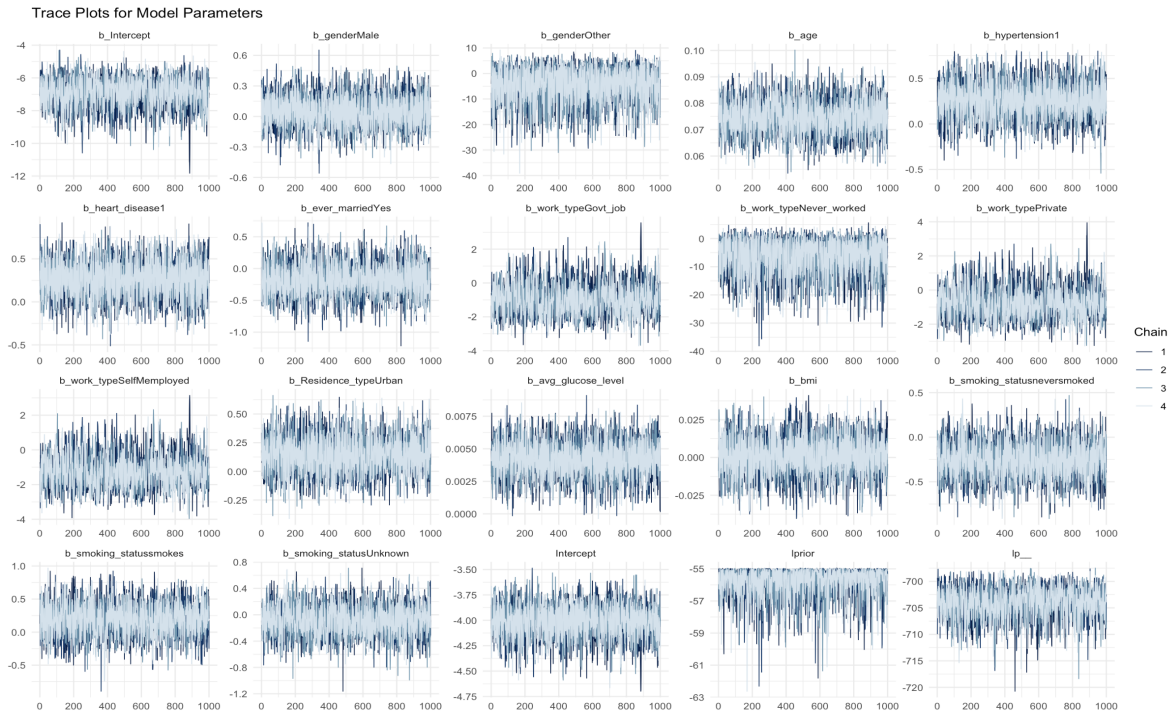

Figure 6: Trace Plots for all parameters

Table 1: Regression Coefficients

	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-6.94	0.89	-8.86	-5.43	1.00	2000	1769
genderMale	0.06	0.16	-0.25	0.37	1.00	4813	2828
genderOther	-4.97	7.03	-21.03	5.51	1.00	4126	2521
age	0.07	0.01	0.06	0.09	1.00	3303	3208
hypertension1	0.26	0.19	-0.13	0.64	1.00	5858	2858
heart_disease1	0.26	0.21	-0.17	0.67	1.00	5849	3051
ever_marriedYes	-0.18	0.26	-0.66	0.35	1.00	4885	2762
work_typeGovt_job	-0.90	0.94	-2.52	1.14	1.00	1366	1586
work_typeNever_worked	-6.60	6.25	-21.41	2.27	1.00	3352	2815
work_typePrivate	-0.70	0.92	-2.31	1.29	1.00	1345	1589
work_typeSelfEmployed	-1.16	0.94	-2.84	0.88	1.00	1359	1500
Residence_typeUrban	0.15	0.16	-0.16	0.46	1.00	5240	2740
avg_glucose_level	0.00	0.00	0.00	0.01	1.00	3981	3084
bmi	0.00	0.01	-0.03	0.03	1.00	4527	2497
smoking_statusneversmoked	-0.26	0.20	-0.64	0.14	1.00	3338	2858
smoking_statussmokes	0.19	0.24	-0.30	0.66	1.00	3182	2829
smoking_statusUnknown	-0.09	0.23	-0.55	0.36	1.00	3342	3195

5 Results

From the summary output of our Bayesian logistic regression, we found that the estimated regression model is given by:

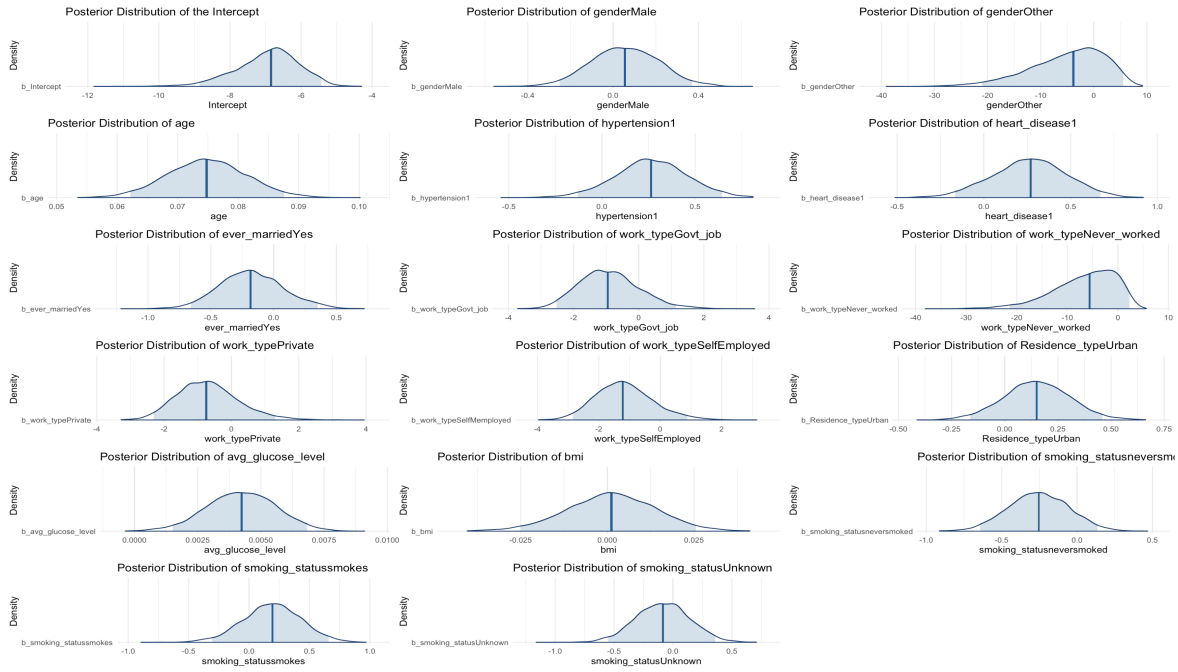

Figure 7: Posterior distributions for all parameters

$$
\begin{aligned}
\text{logit}(P(\text{stroke}=1)) = \; & -6.94 + 0.06 \cdot \text{genderMale} - 4.97 \cdot \text{genderOther} \\
& + 0.07 \cdot \text{age} + 0.26 \cdot \text{hypertension1} + 0.26 \cdot \text{heart\_disease1} \\
& - 0.18 \cdot \text{ever\_marriedYes} - 0.90 \cdot \text{work\_typeGovt\_job} \\
& - 6.60 \cdot \text{work\_typeNever\_worked} - 0.70 \cdot \text{work\_typePrivate} \\
& - 1.16 \cdot \text{work\_typeSelfEmployed} + 0.15 \cdot \text{Residence\_typeUrban} \\
& + 0.00 \cdot \text{avg\_glucose\_level} + 0.00 \cdot \text{bmi} \\
& - 0.26 \cdot \text{smoking\_statusneversmoked} + 0.19 \cdot \text{smoking\_statussmokes} \\
& - 0.09 \cdot \text{smoking\_statusUnknown}
\end{aligned}
$$

To compute the estimated probability of getting stroke, we will use the following equation:

$$
P(\text{stroke}=1) = \frac{1}{1+\exp\left(-\left(\hat{\beta}_0 + \sum_i \hat{\beta}_i X_i\right)\right)}
$$

By utilizing this model to predict stroke status on test data, then we set the decision threshold probability to be 0.5 firstly, the test accuracy is around 95.2% and the confusion matrix is shown in table 2. The confusion matrix indicates that while the model has high overall accuracy (95.2%), but fails to identify any stroke cases (recall = 0), resulting in a significant number of false negatives (49). It has a significant issue with sensitivity (recall) for stroke cases. The specificity is perfect (1), indicating the model is very good at predicting non-stroke cases. Specifically, it failed to correctly identify any stroke cases in the test dataset, resulting in no true positives and a number of false negatives. This suggests that the model is highly conservative and might be overfitting to the majority class (non-stroke) in an imbalanced dataset. Therefore, we changed the threshold probability to 0.3, and we received the test accuracy 94.61%. Even though there is a slight decrease in the accuracy, we successfully predicted more people with actual stroke, but this also increased the wrong results in people without stroke. However, this change is bearable and accuracy is still good enough.

Based on the Figure 8 showing accuracy versus threshold, several conclusions can be drawn. Firstly, the accuracy increases as the threshold increases from 0.1 to around 0.4. Beyond this point, the

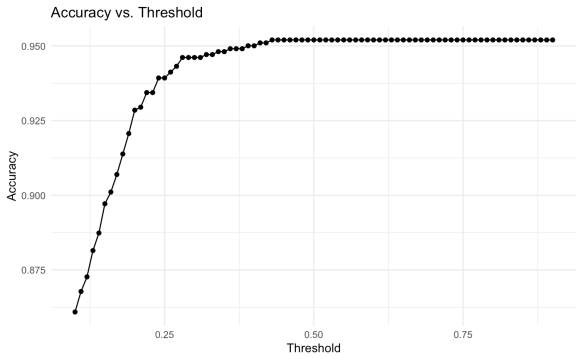

Figure 8: Accuracy vs. Threshold plot

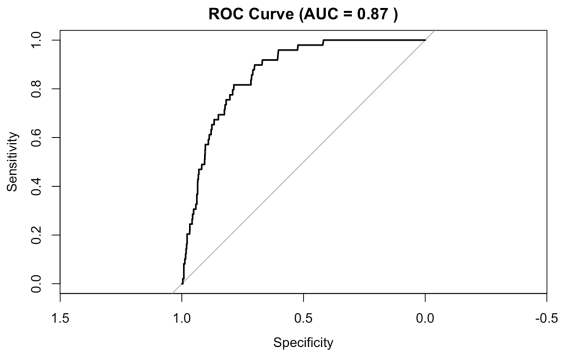

Figure 9: ROC Plot

Table 2: Confusion Matrix using threshold probability=0.5

Actual / Predicted	**Predicted: No Stroke (0)**	**Predicted: Stroke (1)**
Actual: No Stroke (0)	972	0
Actual: Stroke (1)	49	0

Table 3: Confusion Matrix using threshold probability=0.3

Actual / Predicted	**Predicted: No Stroke (0)**	**Predicted: Stroke (1)**
Actual: No Stroke (0)	963	9
Actual: Stroke (1)	45	4

accuracy stabilizes and remains relatively constant at approximately 0.95. This indicates that setting the threshold above 0.4 does not significantly improve the model's performance in terms of accuracy.

The optimal threshold for maximizing accuracy is around 0.4, as this is the point where accuracy stabilizes at its highest level. At lower thresholds (below 0.4), the accuracy is noticeably lower, suggesting that the model may be making more incorrect predictions, likely due to an increase in false positives. Therefore, setting the threshold at around 0.4 is advisable to achieve the best balance between sensitivity and specificity.

6 Conclusion

The Bayesian logistic regression model applied in this study successfully identifies significant predictors of stroke, specifically age and average glucose level, aligning with established medical knowledge. Despite achieving a high overall accuracy of around 95% in the test dataset, the model initially struggled with sensitivity for stroke cases due to the imbalanced nature of the data. Adjusting the decision threshold improved the identification of actual stroke cases, albeit with a minor decrease in overall accuracy. These findings underscore the importance of considering threshold adjustments and data balance in predictive modeling. The model's robustness and reliability, confirmed by convergence diagnostics, suggest its potential utility in clinical settings for early stroke prediction, ultimately aiding in targeted prevention and intervention strategies. Further work could explore integrating additional predictors and refining threshold settings to enhance the model's sensitivity without compromising accuracy.

References

[1] Ahmad Hassan. Stroke prediction dataset, 2023.

[2] Francesca Pistoia, Simona Sacco, Diana Degan, Cindy Tiseo, Raffaele Ornello, and Antonio Carolei. Hypertension and stroke: epidemiological aspects and clinical evaluation. *High Blood Pressure & Cardiovascular Prevention*, 23:9–18, 2016.

[3] Yuexin Qiu, Shiqi Cheng, Yuhang Wu, Wei Yan, Songbo Hu, Yiying Chen, Yan Xu, Xiaona Chen, Junsai Yang, Xiaoyun Chen, et al. Development of rapid and effective risk prediction models for stroke in the chinese population: a cross-sectional study. *BMJ open*, 13(3):e068045, 2023.