Statistics C116 Final Project: Bayesian Logistic Regression

Maeve Horan-Portelance (305768725)

Professor Handcock

14 June 2024

I. Abstract

This report explores the Bayesian approach to logistic regression, focusing on its theoretical foundations, practical applications, variations in usage, and comparison to frequentist methods. The data used, obtained from the National Institute of Diabetes, predicts the onset of diabetes in Pima Indians using diagnostic measures such as pregnancies, BMI, glucose levels, blood pressure, insulin levels, age, and family history. With this data and the various statistical tools from both Bayesian and frequentist methods, we illustrate the advantages and disadvantages, as well as the workings behind Bayes' approach to logistic regression. We will test four outcomes:

I. Performance of the Bayesian vs. Frequentist logistic models
II. Accuracy of 'MICE' vs. 'BRMS' imputation of missing values
III. Impact of using Horseshoe vs. Laplace priors on Bayesian regression
IV. Impact of adding interaction term

Considering 7 different models, we will analyze which has the best performance by looking at confusion matrices, accuracy measures, and diagnostic plots.

I. Introduction: Bayesian Logistic Regression

A). Key Concepts

1. **Definition of Bayesian Logistic Regression:**
   Bayesian logistic regression is a statistical method used to model binary outcomes (outcomes that can take on one of two possible values), by applying Bayesian principles to the framework of logistic regression. By incorporating prior knowledge or beliefs about the parameters, or predictors, in the form of prior distributions, the method updates these beliefs with observed data to abstain posterior distributions.
   The logistic regression model is used to demonstrate the probability that a given observation belongs to a particular category (for example, either having a disease or not). It uses the logistic, or sigmoid function, which transforms the linear combination of predictor variables into a probability between 0 and 1. The model is given by:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}$$

   where Y is the binary outcome variable, X_1 ,...., X_n are the predictor variables, and β_1 ,...., β_n are the coefficients. In the Bayesian framework, the uncertainty about the model parameters (coefficients) is expressed using probability distributions.

2. **The Role of Prior Distributions**

In Bayesian regression, prior distributions play a critical role in shaping the posterior distribution of the model parameters, as they reflect the initial beliefs or information about the parameters before observing any data. These priors can significantly influence the resulting inferences.

Non-Informative (Flat) Priors: These priors lack prior knowledge about the parameters, and are designed to have minimal influence on the posterior distribution.

Informative Priors: These priors incorporate specific prior knowledge about the parameters, for example, a normal distribution with a specified mean and variance may be used if there is knowledge about the expected value of a coefficient.

Weakly Informative Priors: Priors that are designed to provide regularization and stabilize the estimates without being overly restrictive–for example–a normal distribution with a large variance.

For the sake of this analysis, we will also be examining **sparsity priors**–a type of prior distributions to encourage sparsity in the parameter estimates, shrinking many of the regression coefficients to be exactly zero or very close to zero.

Priors can prevent overfitting, incorporation of external knowledge into the model, and improve the convergence properties of the logistic model. In this case, because we have data with lots of NA values, these priors can also have a significant impact on the posterior distribution, guiding the estimates more strongly.

3. **Posterior Distributions & Inference**
   As the cornerstone of inference, the posterior distribution represents the updated beliefs about the model parameters after incorporating the observed data and prior distributions. It is derived using Bayes' Theorem, and based on the log-likelihood for the logistic model (shown above). While the distribution is often complex and cannot be derived analytically, it is approximated using sampling methods, such as the Markov Chain Monte Carlo (MCMC), which generates samples from the posterior distributions, used to estimate summary statistics, perform hypothesis testing, and make predictions.

B). Comparison with Frequentist Methods

Advantages of Bayesian Methods

1. **Ability to Incorporate Prior Knowledge**
   By incorporating prior knowledge, analysts using the Bayesian framework can incorporate specific, detailed knowledge about the parameters that may be lacking in frequentist methods. Priors can also act as regularizers, preventing overfitting by shrinking unimportant coefficients toward 0 (or another specified value). In addition, whereas sparse data may cause problems for predictive frequentist methods, informative priors can help stabilize the parameter estimates and improve the robustness of the model.

2. **Flexibility in Modeling Complex Relationships**
   In comparison to frequentist methods, Bayesian logistic regression provides a flexible framework for modeling complex relationships and structures within the data, as it can easily incorporate interaction terms, random effects, non-linear relationships and hierarchical models.

3. **Natural Handling of Missing Data**
   Bayesian logistic regression handles missing data naturally by incorporating it into the modeling process, rather than relying on ad hoc imputation methods. Missing values can be treated as additional parameters to be estimated, and the 'mi()' function in 'brms' allows for the inclusion of missing data within the Bayesian framework. By treating missing data probabilistically, Bayesian methods often yield better estimates than traditional imputation methods used in frequentist models, which could underestimate the uncertainty introduced by missing data.

Limitations of Bayesian Methods

1. **Computational Intensity and Convergence Issues**
   Bayesian methods, particularly those involving MCMC samples from posterior distributions, are computationally intensive–making them challenging to implement for complex models or high-volume datasets. Numerous iterations are often required to ensure convergence, and convergence issues may lead to biased or incorrect inferences. Frequentist methods tend to be simpler, less computationally-intensive, and often have effective and reliable convergence.

2. **Dependence on Choice of Prior Distributions**
   The choice of prior distributions in Bayesian analysis can significantly influence the results–if the priors are not well-chosen, they can lead to biased estimates. While non-informative or weakly informative priors are often used to minimize this influence, they may not always be appropriate, and have the opportunity to bias the model. Frequentist models do not have this added risk, meaning that they are immune to the potential harmful effects of choosing inappropriate prior distributions.

3. **Interpretation & Communication of Results**
   Interpreting and communicating the results of Bayesian logistic analysis can be more complex than frequentist methods, especially for those less familiar with Bayesian statistics. For example, instead of p-values, Bayesian analyses provide posterior probabilities, and use WAIC instead of AIC and BIC for the information criterion.

III. Logistic Regression: Diabetes Data Analysis

Background on the Data & Exploratory Analysis

The objective of this dataset is to diagnostically predict whether or not a patient has diabetes for women of Pima Indian heritage over 21 years old. The target variable, outcome, indicates whether or not the individual has diabetes. There are 8 total numerical predictor variables: Age (years), DiabetesPedigreeFunction, BMI, Insulin, SkinThickness, BloodPressure, Glucose, and Pregnancies. The goal of this analysis is to use both frequentist and Bayesian logistic regression to predict diabetes status based on these 8 predictor variables.

Missing Values

Note: While the original dataset contains 0s in the place of missing values, for the sake of this analysis, they were reverted back to NA values.

Both **Figure 1** and **Figure 2** explore the presence of missing values in this dataset. This is relevant to logistic regression, as missing data can significantly impact the results and conclusions drawn from the analysis. Missing values can lead to biased estimates, reduced statistical power, and invalid conclusions if not handled properly. **Figure 1** provides a visualization of the missing data using an aggregation plot–only the Insulin, SkinThickness, Blood Pressure, BMI, and Glucose variables contain missing values, which will be imputed later in the analysis. **Figure 2** shows which percentage of those variables missing values make up. Insulin has the highest percentage of missing values, at a striking 48.7%, followed by skin thickness, at 29.56%. Thus, how the missing values are modeled in logistic regression will clearly have a large impact on the models–mice imputation and Bayesian methods will be explored, and compared for effectiveness.

Figure 1: Aggregation plot of missing values in diabetes dataset

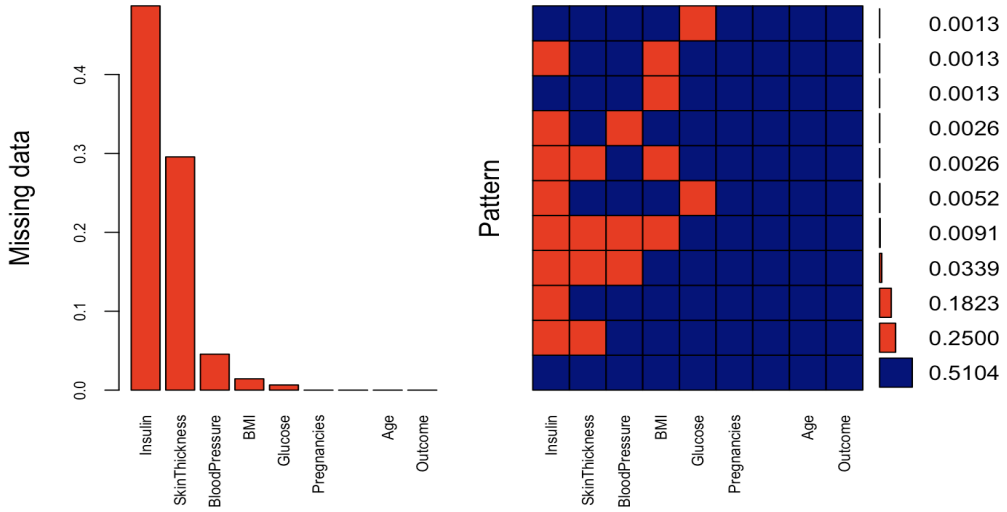


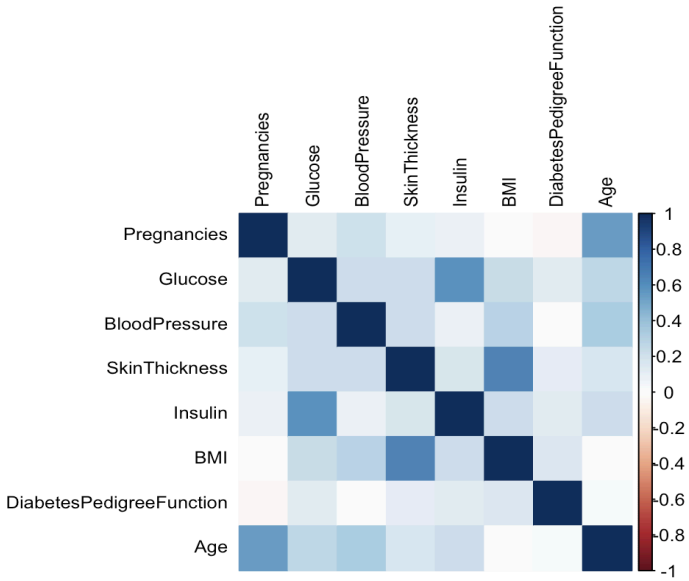
Figure 2: Table of missing data percentages by variable

Variable <chr>	Count <dbl>
Insulin	0.486979167
SkinThickness	0.295572917
BloodPressure	0.045572917
BMI	0.014322917
Glucose	0.006510417
Pregnancies	0.000000000
DiabetesPedigreeFunction	0.000000000
Age	0.000000000
Outcome	0.000000000

Data Preparation

Before analysis began, the data was split into a 70%-30% train-test ratio, so that the models' accuracy could be tested on the test data for more accurate insights. The correlation between the predictors was also examined to make sure multicollinearity would not be an issue in the models. As seen in **Figure 3**, no 2 predictor variables showed an alarming amount of correlation, so all were included in the initial model.

Figure 3: Correlation matrix



For the frequentist methods, the Outcome variable is treated as a factor–for the Bayesian methods, it is converted to a numeric value. In both instances, it is either 0 (no diabetes) or 1 (indicating the presence of diabetes)--the models simply handle the response variable differently.

Part I: Frequentist Analysis

Model I: Frequentist Model, MICE Imputation, All Predictors

The first frequentist model was created using the 'glm' function in R, with all of the predictors included. The MICE package was used to impute the missing values in the data–a common method in frequentist analysis that iteratively creates multiple-imputed datasets, and the results are combined to produce estimates and results.

As shown in **Figure 5**, the model showed an accuracy of 0.7576, a precision score of 0.7798, and an F-1 score of 0.8239. The area under the ROC curve is 0.8468, meaning it performs well in separating diabetes vs. non-diabetes outcomes. In analyzing the diagnostic plots in **Figure 7,** most of the points in the QPP plot fall along the reference line, meaning that the residuals are, for the most part, normally distributed–however, the deviations from the line at the end indicate slight departures from normality (characteristic of logistic regression). The Residuals vs. Fitted plot shows a pattern characteristic of logistic regression, and only a few outliers are identified by the Residuals vs. Leverage plot. The AUC, or area under the ROC curve, is 0.8468, and the ROC curve is shown in **Figure 6.**

As is made clear by the regression coefficients in **Figure 4,** the most significant predictors (with the lowest p-values) include Pregnancies, Glucose, and BMI. DiabetesPedigreeFunction is also fairly significant, with a p-value just above the threshold of 0.05. Thus, we will include these 4 predictors in our best subsets frequentist model.

While this model appears to be a good start for fitting a logistic regression model to the data, we will continue to explore other options in the coming models, and compare their performance and diagnostics.

Figure 4: Regression coefficients for Model I

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.259401   0.970492  -9.541  < 2e-16 ***
Pregnancies               0.105903   0.039240   2.699 0.006958 **
Glucose                   0.042229   0.005280   7.999 1.26e-15 ***
BloodPressure            -0.012326   0.010391  -1.186 0.235535
SkinThickness             0.019172   0.014413   1.330 0.183458
Insulin                  -0.001884   0.001130  -1.668 0.095357 .
BMI                       0.083325   0.023136   3.602 0.000316 ***
DiabetesPedigreeFunction  0.650432   0.345957   1.880 0.060095 .
Age                       0.012389   0.011844   1.046 0.295532
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: Confusion matrix and scores for Model I

Accuracy: 0.7576 Precision: 0.7798 F-1 Score: 0.8239		**Reference**	
		0	**1**
Prediction	**0**	131	38
	1	19	44

Figure 6: ROC curve for Model I

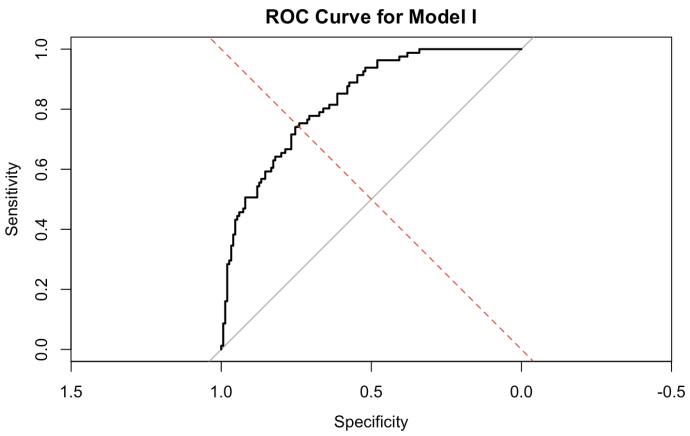
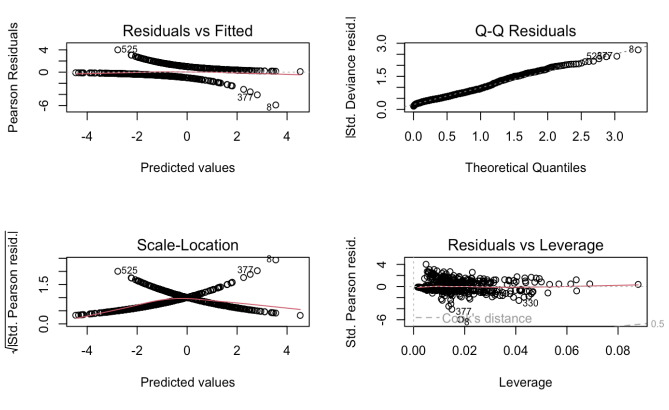


Figure 7: Diagnostic Plots for Model I



Model II: Frequentist Model, MICE Imputation, Best Subset

As the second model in our frequentist exploration, Model II features the same MICE-imputed data, but only the subset of most-significant predictors outlined in the regression coefficients from Model I. Sometimes, reducing the number of predictors is effective because it reduces potential noise in the model created by the less-significant predictors, reduces multicollinearity, and can lead to better generalization on unseen data in certain circumstances. The following predictors were used in the model: Glucose, BMI, DiabetesPedigreeFunction, and Pregnancies. As seen in **Figure 8**, all of these predictors had a statistically significant effect on the model, in comparison to Model I, where only some of the predictors did.

Figure 8: Regression coefficients for Model II

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -9.205268   0.826814 -11.133  < 2e-16 ***
Glucose                  0.037973   0.004311   8.808  < 2e-16 ***
BMI                      0.089339   0.017736   5.037 4.73e-07 ***
DiabetesPedigreeFunction 0.679358   0.339356   2.002 0.045295 *
Pregnancies              0.124425   0.032532   3.825 0.000131 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As shown by the results in **Figure 9**, Model II’s accuracy increased by 1.3%, and the precision and F-1 also show slight increases when the less-significant predictors are removed from the model. While these performance metrics are slightly better, the area under the ROC curve is 0.8379, which is slightly lower than that of Model I, indicating that its measure of separability

between classes may not be as high. The ROC curve is plotted in **Figure 10**–while the area under the curve may be slightly lower, the curves appear almost exactly the same, indicating that this difference is not a major factor in assessing the performance of the two frequentist models.

The diagnostic plots in **Figure 11** also show similar results to Model I–the residuals appear slightly more normally-distributed (fall on the line of the QPP plot), and only one point falls outside of Cook's distance on the Residuals vs. Leverage plot, indicating a lack of outliers.

Figure 9: Confusion matrix and diagnostics for Model II

Accuracy: 0.7706 Precision: 0.7803 F-1 Score: 0.8359		**Reference**	
		0	**1**
Prediction	**0**	135	38
	1	15	45

Figure 10: ROC curve for Model II

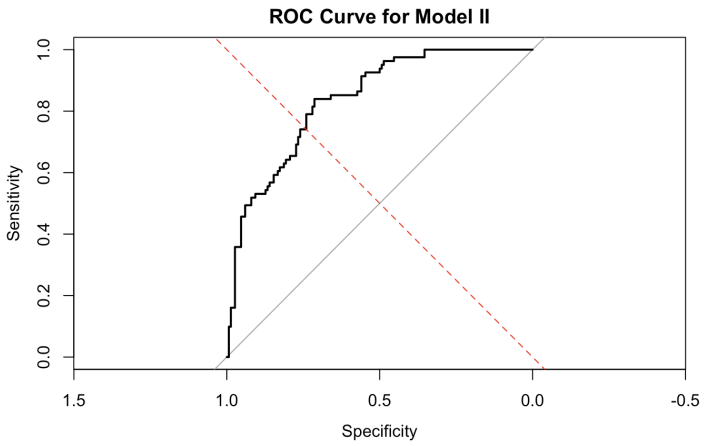
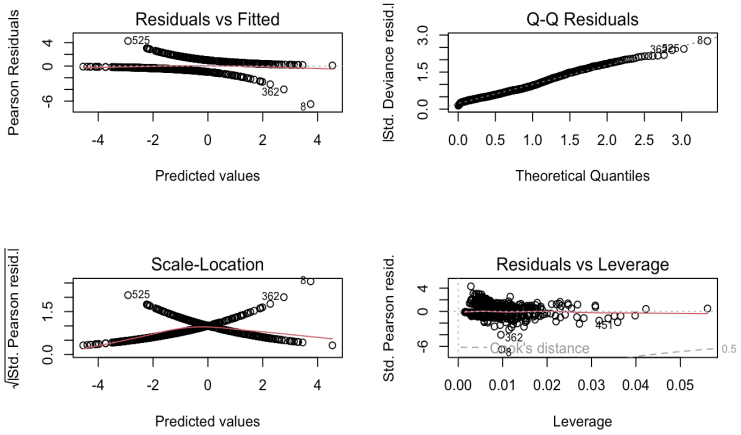


Figure 11: Diagnostic plots for Model II



Conclusion

Taking into consideration the confusion matrix accuracies on the testing data, precision and F-1 scores, the AUC (area under the ROC curve), and residual plots, Model II slightly outperforms Model I. Thus, Model II–frequentist 'glm' with MICE imputation and the best subset of predictors–will be used in comparison with the Bayesian models in the coming analyses.

In addition, this finding illustrates that removing potentially-unimportant predictors has a significant impact on the model's performance. In the coming analysis, we will use Bayesian analysis, which expands on this principle by incorporating prior knowledge about the coefficients and providing a flexible framework for modeling complex relationships within the data.

Part II: Bayesian Analysis

Model III: Bayesian Model, MICE Imputation, Horseshoe Prior

Model III was created using the same MICE-imputed data as the previous two models, however, the package 'brms', particularly the function 'brm,' was used to create a Bayesian model. A horseshoe sparsity prior was also used for this model, which is particularly effective in handling high-dimensional data by shrinking the coefficients of less-significant predictors toward zero, thus preventing overfitting and enhancing model interpretability. 'Brm' differs from the 'glm' function in that it allows for the incorporation of prior distributions, provides a probabilistic interpretation of model parameters, and uses MCMC methods to estimate the posterior distributions of the coefficients.

The output shown in **Figure 12** includes the regression coefficients from the model, which look slightly different from the 'glm' output, as the coefficients in the Bayesian framework are expressed as posterior distributions. This allows for direct quantification of uncertainty around the estimates, offering credible intervals, which provide a range within which the true parameter values are believed to lie with a certain probability.

Figure 12: Regression coefficients for Model II

Regression Coefficients:

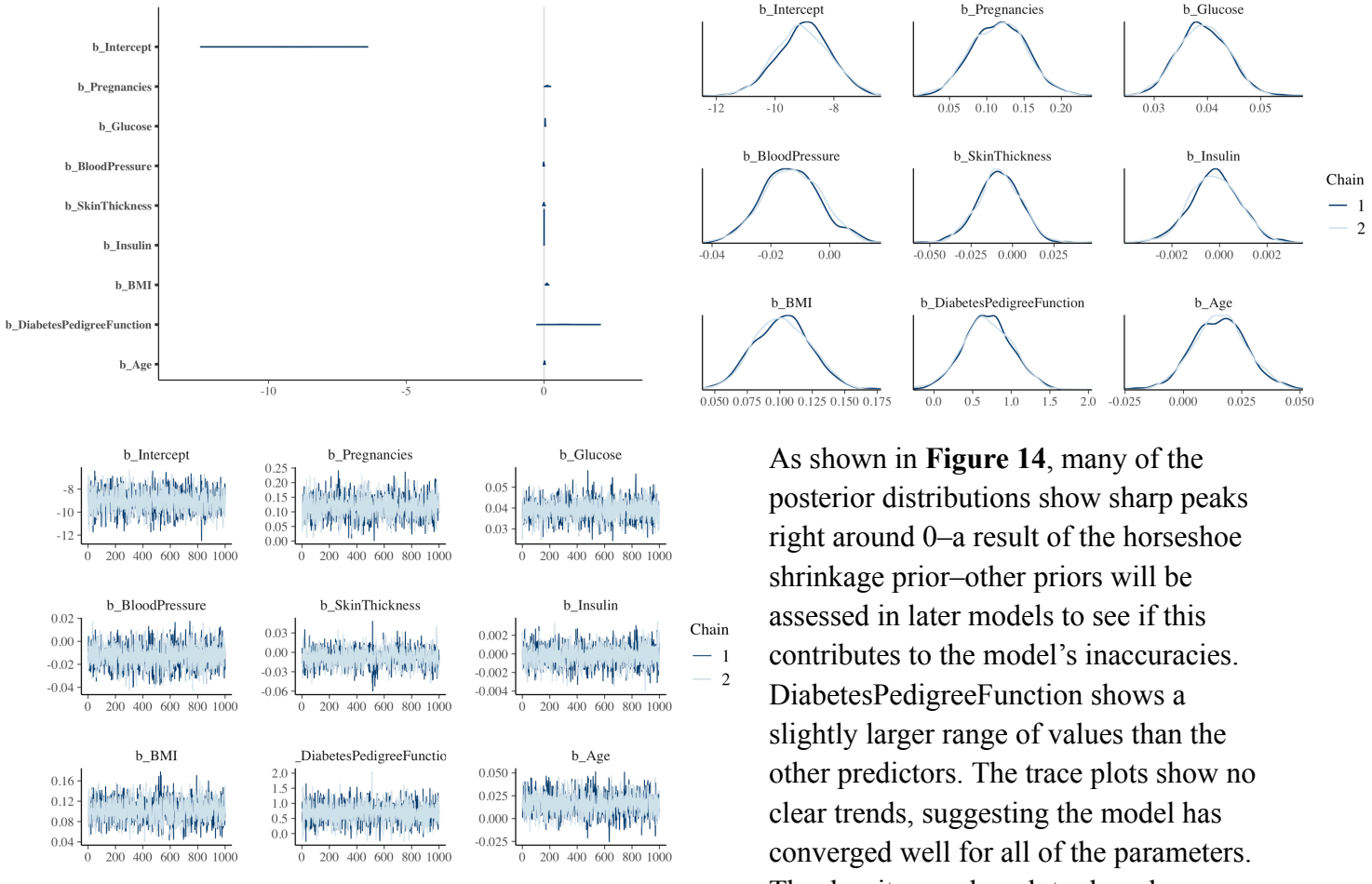
	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-8.94	0.94	-10.81	-7.11	1.01	790	1425
Pregnancies	0.08	0.04	-0.00	0.16	1.00	995	773
Glucose	0.04	0.01	0.03	0.05	1.01	819	1635
BloodPressure	-0.01	0.01	-0.03	0.01	1.00	1524	1584
SkinThickness	0.02	0.02	-0.01	0.05	1.01	147	156
Insulin	-0.00	0.00	-0.00	0.00	1.01	105	30
BMI	0.07	0.02	0.03	0.12	1.01	510	789
DiabetesPedigreeFunction	0.12	0.22	-0.09	0.80	1.00	922	1643
Age	0.01	0.01	-0.01	0.03	1.01	447	1358

Based on the results in **Figure 13**, when Model III was used to predict the Outcome variable on the testing data, it had an accuracy score that is 2.17% lower than Model II's score–the frequentist model using MICE-imputed data. The precision and F-1 scores also decrease from both Model I and Model II, meaning that in general, this model is slightly less effective at predicting the Outcome variable on testing data than the frequentist model. This could be because Bayesian models, while offering flexibility and prior knowledge, are more sensitive to the choice of priors, and may require tuning or other methods of missing value imputation to achieve the same (or higher) level of performance as frequentist models. The MICE imputation may not be the most effective way of modeling the missing values–in the coming models, we will compare these results to Bayesian models with 'BRMS' imputation.

Figure 13: Confusion matrix and diagnostics for Model III

Accuracy: 0.7489 Precision: 0.7706 F-1 Score: 0.8188		**Reference**	
		0	**1**
Prediction	**0**	131	39
	1	9	42

Figure 14: Posterior distributions, trace, and density overlay plots for Model III



As shown in **Figure 14**, many of the posterior distributions show sharp peaks right around 0–a result of the horseshoe shrinkage prior–other priors will be assessed in later models to see if this contributes to the model's inaccuracies. DiabetesPedigreeFunction shows a slightly larger range of values than the other predictors. The trace plots show no clear trends, suggesting the model has converged well for all of the parameters. The density overlay plots show large overlap between the chains for most predictors, indicating consistent sampling and convergence, however, some variables such as BMI and Insulin show some disparity, indicating that this model may not be the best fit for the data.

While Bayesian models offer some advantages, it is clear that Model III is underperforming in comparison to the frequentist models. We will now investigate whether the Bayesian model can be improved upon by other imputation methods and the usage of different priors.

Model IV: Bayesian Model, 'BRMS' Imputation, Horseshoe Prior

Model IV was created using the same 'brm' function from the 'brms' package, all of the predictors, and the horseshoe sparsity prior (same as Model III). The key difference is that it utilizes different data–the original missing values were imputed by the 'brms' package using the 'mi()' function, which allows the model to account for uncertainty in the missing data by treating missing values as additional parameters to be estimated. This method leverages the full Bayesian framework and is much different from MICE, which instead generates multiple datasets, leading to potential inconsistencies between the imputation and the model fitting processes.

Figure 15: Regression coefficients for Model IV

```
Regression Coefficients:
                                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Outcome_Intercept                   -1.08      0.12    -1.32    -0.84 1.00     2223     1111
Glucose_Intercept                  100.72      3.69    93.56   107.78 1.00     2779     1479
Insulin_Intercept                   79.45     21.91    35.12   121.20 1.00     1782     1606
BloodPressure_Intercept             61.55      1.54    58.55    64.70 1.00     2936     1401
SkinThickness_Intercept             24.07      1.64    20.73    27.29 1.00     2087     1698
BMI_Intercept                       32.48      0.91    30.70    34.22 1.00     3398     1610
Outcome_Age                          0.00      0.00    -0.00     0.01 1.00     2220     1423
Outcome_DiabetesPedigreeFunction     0.10      0.05    -0.01     0.20 1.00     3085     1222
Outcome_Pregnancies                  0.02      0.01     0.01     0.03 1.00     3213     1581
Glucose_Age                          0.66      0.10     0.46     0.87 1.00     2787     1594
Insulin_Age                          2.39      0.65     1.13     3.70 1.00     1565     1660
BloodPressure_Age                    0.31      0.04     0.22     0.40 1.00     3132     1263
SkinThickness_Age                    0.17      0.05     0.07     0.27 1.00     1973     1635
BMI_Age                             -0.00      0.03    -0.05     0.05 1.00     3066     1605
Outcome_miGlucose                    0.01      0.00     0.01     0.01 1.00     1808     1325
Outcome_miInsulin                   -0.00      0.00    -0.00     0.00 1.00     1478     1554
Outcome_miBloodPressure             -0.00      0.00    -0.00     0.00 1.00     2143     1925
Outcome_miSkinThickness              0.00      0.00    -0.00     0.00 1.00     2150     1546
Outcome_miBMI                        0.02      0.00     0.01     0.02 1.00     2518     1617
```

In this model, it is clear that, since the "Rhat" values are all 1, all of the chains have converged well. DiabetesPedigreeFunction appears to be the most significant predictor of the outcome variable, with Glucose, BMI, and Pregnancies also have a slight impact. The rest of the predictor variables were shrunk to 0 by the horseshoe sparsity prior.

As seen in **Figure 16**, Model IV predicted the Outcome variable much more accurately than Model III–the accuracy score improved by 2.53%, and the accuracy, precision, and F-1 scores are the best shown thus far. Both Models III and IV used the horseshoe sparsity prior, but because Model IV performed significantly better, it is clear that the imputation of missing values using 'brms' has a positive effect on model outcome.

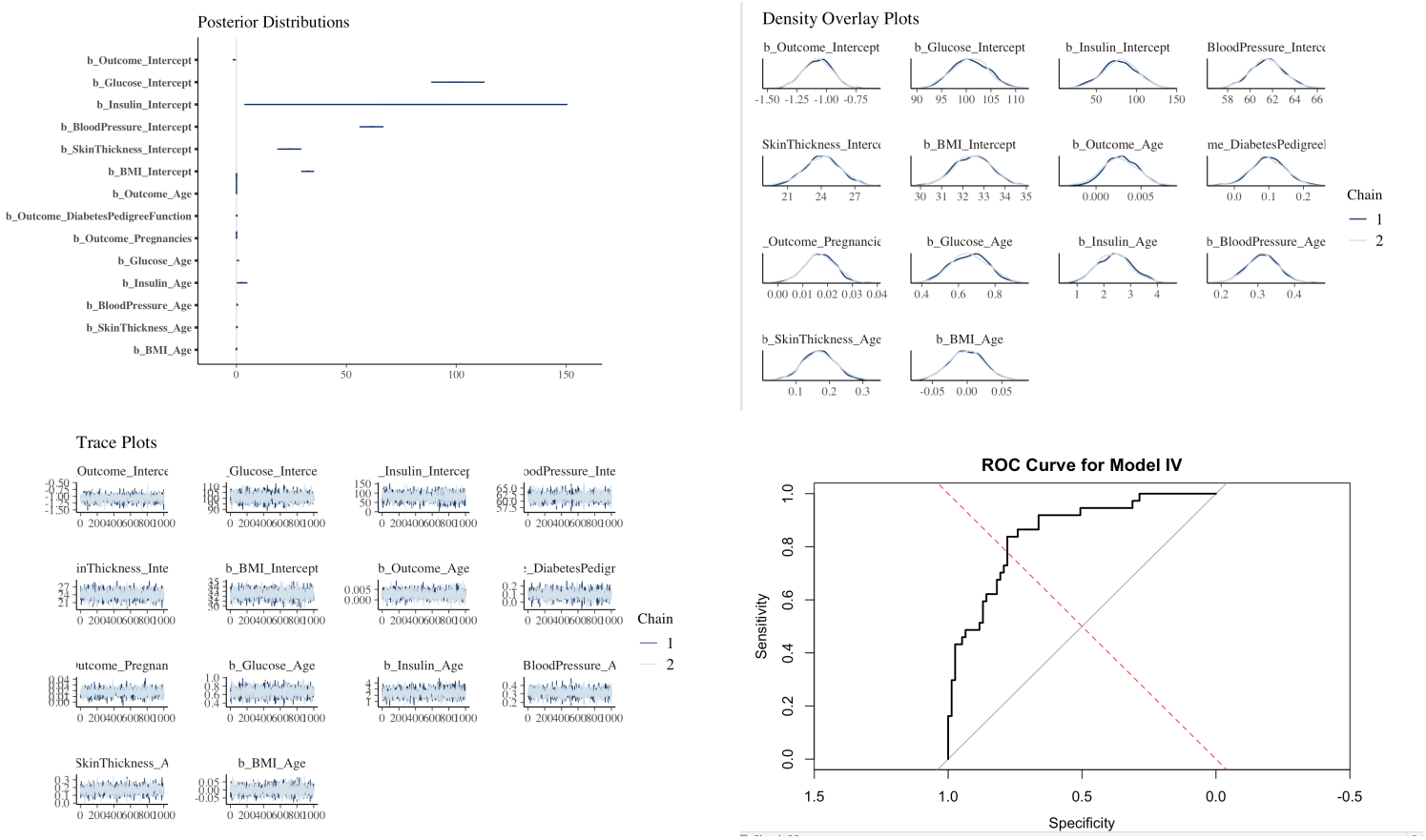
Figure 16: Confusion matrix and diagnostics for Model IV

Accuracy: 0.7742 Precision: 0.7978 F-1 Score: 0.8353		**Reference**	
		0	**1**
Prediction	**0**	71	18
	1	10	25

Looking at **Figure 17,** posterior distributions for the predictors indicate that many of them have been shrunk to 0 (or very close to 0) by the horseshoe prior. DiabetesPedigreeFunction shows a slightly larger range of values than the other predictors. The trace plots show no clear trends, suggesting the model has converged well for all of the parameters. The density overlay plots

show near-perfect overlap between the chains for most predictors, indicating consistent sampling and convergence for the model, which is confirmed by the Rhat values in **Figure 15.** The area under the ROC curve is 0.8529, meaning that the model has an 85.29% chance that it will distinguish between a randomly chosen positive or negative instance from the response variable. This indicates good performance, but certainly leaves room for improvement.

Figure 17: Posterior distributions, trace, density overlay, and ROC plots for Model IV



Model V: Bayesian Model, 'BRMS' Imputation, Laplace Priors

Figure 18: Regression Coefficients for Model V

```
Regression Coefficients:
                                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Outcome_Intercept                  -1.02      0.10     -1.21     -0.82 1.00     3603     1272
Glucose_Intercept                  98.45      3.15     92.24    104.56 1.00     4155     1484
Insulin_Intercept                  74.57     17.59     40.09    108.25 1.00     1462     1800
BloodPressure_Intercept            60.90      1.30     58.38     63.47 1.00     3794     1460
SkinThickness_Intercept            23.99      1.35     21.34     26.63 1.00     2249     1814
BMI_Intercept                      31.94      0.75     30.42     33.43 1.01     2771     1046
Outcome_Age                         0.00      0.00     -0.00      0.01 1.00     2844     1730
Outcome_DiabetesPedigreeFunction    0.13      0.04      0.04      0.22 1.00     4435     1482
Outcome_Pregnancies                 0.02      0.01      0.01      0.03 1.00     4739     1356
Glucose_Age                         0.70      0.09      0.52      0.88 1.00     4453     1436
Insulin_Age                         2.61      0.54      1.58      3.64 1.00     1198     1685
BloodPressure_Age                   0.34      0.04      0.27      0.42 1.00     4275     1400
SkinThickness_Age                   0.16      0.04      0.08      0.24 1.00     1729     1752
BMI_Age                             0.02      0.02     -0.03      0.06 1.00     2740     1302
Outcome_miGlucose                   0.01      0.00      0.01      0.01 1.00     1842     1264
Outcome_miInsulin                  -0.00      0.00     -0.00      0.00 1.00     1598     1584
Outcome_miBloodPressure            -0.00      0.00     -0.00      0.00 1.00     2979     1692
Outcome_miSkinThickness            -0.00      0.00     -0.00      0.00 1.00     2077     1512
Outcome_miBMI                       0.01      0.00      0.01      0.02 1.00     2620     1497
```

Model V involves the use of 'BRMS', or Bayesian Model Selection for handling missing values, with the Laplace prior for regularization. The Laplace prior, or the double-exponential prior, is commonly used for L1 regularization

in Bayesian models, and encourages sparsity in the model by shrinking less important coefficients toward 0. The Laplace distribution for the prior can be expressed as $p(\beta)\ \alpha\ exp(-\ \lambda|\beta|)$, where λ controls the degree of sparsity (larger values of λ result in stronger shrinkage of coefficients). It is different from the horseshoe prior, which performs a more heavy-tailed, adaptive shrinkage.

As seen in **Figure 18**, the regression coefficients look very similar to the ones in Model IV, however, DiabetesPedigreeFunction was given a higher coefficient, while BMI was adjusted just slightly lower–a result of the different methods used in the different sparsity priors. **Figure 19** shows that the Laplace prior made a high impact on the accuracy, precision, and F-1 scores of the model–each score increased by over 2%, showing that the L1 regularization is more effective in the Bayesian logistic regression for this dataset. The diagnostic plots in **Figure 20** show mainly the same results as the prior models–there are no noticeable convergence or sampling issues. The area under the ROC curve is also 0.8529, identical to that of Model V.

Figure 19: Confusion matrix and diagnostics for Model V

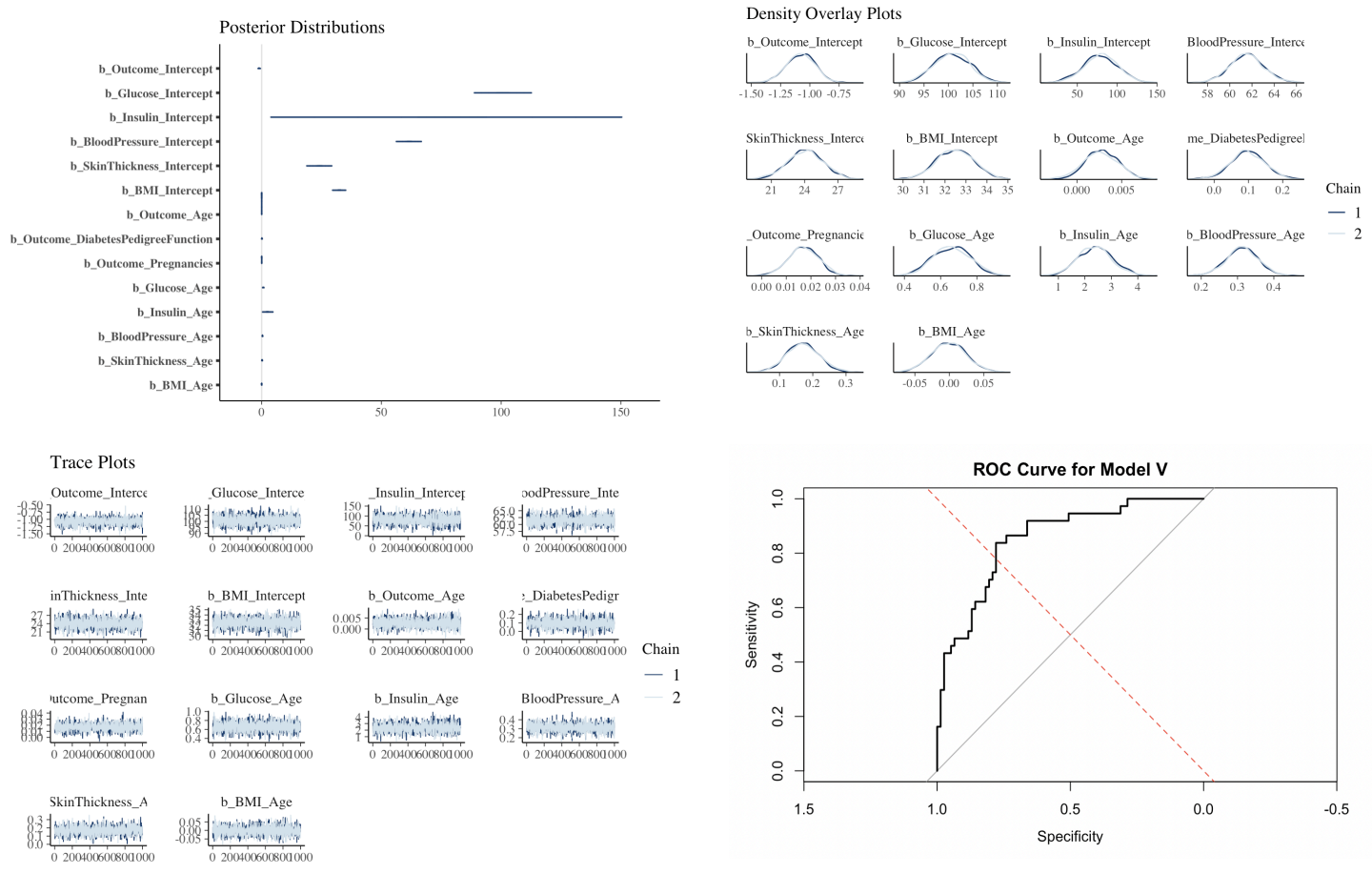
Accuracy: 0.7984 Precision: 0.811 F-1 Score: 0.8521		**Reference**	
		0	**1**
Prediction	**0**	73	17
	1	8	26

Although it is a slight difference, the Laplace prior seems to be a better fit for modeling this data than the horseshoe prior, as it leads to increased accuracy without compromising the model convergence or the consistency of model sampling.

Model VI & VII: Bayesian Model, 'BRMS' Imputation, Laplace Priors, Interaction Terms

Upon analyzing the results above, it is clear that the model that is best at predicting the "Outcome" variable for this dataset is Model V, which uses Bayesian methods, 'BRMS' imputation of missing values, and the Laplace sparsity prior. The last 2 models we will look at take Model V, and add an interaction term between (1) BMI and DiabetesPedigreeFunction and (2) BMI and Glucose. While there are many combinations of variables that could be considered as interaction terms, the purpose of this analysis is to see what happens to the model when an interaction term is added–if it affects convergence, if it has the ability to significantly impact performance, etc. The variable imputation and sparsity priors were kept the same for both models to compare performance with each other, and with Model V.

Figure 23: Posterior distribution, density overlay, trace and ROC plots for Model V



BMI and DiabetesPedigreeFunction were the two most significant predictors in the previous Bayesian analysis–this is why they were chosen for the interaction term in Model VI. The coefficient significance for DiabetesPedigreeFunction increased significantly in this model, while the rest of the predictor coefficient values remain very close to the model without the interaction term. The interaction term itself does not seem to have much of an impact on the model (estimate = -0.01). Based on the confusion matrix in **Figure 25**, the accuracy, precision, and F-1 score decreased by around 0.02 from the previous model, meaning that adding the interaction term had a negative impact on the model's predictive abilities. The Rhat values in **Figure 24** are all 1, and the diagnostic plots in **Figure 27** indicate that there were no issues with convergence, while the area under the ROC curve, 0.854, decreased slightly from Model V.

BMI and Age were chosen for the interaction term in Model VII because, although they are not the two most significant predictors, there is likely to be interaction between those two variables,

as one's BMI will likely vary very heavily with age. In this model, a lot of the predictors were shrunk down to 0, including the interaction term. However, the accuracy, precision, and F-1 scores of the model are almost 2.5% higher than Model VI, and slightly higher than Model VII (not by a significant amount). The area under the ROC curve is 0.878, which is also 2% higher than Models VII.. Looking at **Figure 24,** we can see that the Rhat values in Model VII start to deviate slightly from 1, and in **Figure 27,** there are slightly larger gaps in the overlap of the density overlay plots, indicating that the model may be deviating from perfect convergence. If more interaction terms were added, it is likely that the complexity of the model would increase beyond its capabilities, reducing the accuracy of the results.

Figure 24: Regression Coefficients for Models VI (left) and VII (right)

Regression Coefficients:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Outcome_Intercept	-1.18	0.15	-1.48	-0.87	1.00	3270	1725
Glucose_Intercept	100.76	3.59	93.58	107.75	1.00	4357	1304
Insulin_Intercept	80.06	21.71	38.18	122.05	1.00	1893	1847
BloodPressure_Intercept	61.56	1.57	58.55	64.68	1.00	4167	1270
SkinThickness_Intercept	24.04	1.67	20.76	27.33	1.00	2166	1619
BMI_Intercept	32.49	0.93	30.63	34.34	1.00	5975	1374
Outcome_Age	0.00	0.00	-0.00	0.01	1.00	2960	1790
Outcome_DiabetesPedigreeFunction	0.32	0.20	-0.07	0.72	1.00	2537	1581
Outcome_Pregnancies	0.02	0.01	0.01	0.03	1.00	5607	1506
Glucose_Age	0.66	0.10	0.46	0.87	1.00	4278	1462
Insulin_Age	2.37	0.66	1.12	3.64	1.00	1542	1539
BloodPressure_Age	0.31	0.04	0.23	0.40	1.00	4122	1404
SkinThickness_Age	0.17	0.05	0.07	0.27	1.00	1916	1760
BMI_Age	-0.00	0.03	-0.05	0.05	1.00	6523	1325
Outcome_miInsulin	-0.00	0.00	-0.00	0.00	1.00	2142	1800
Outcome_miBloodPressure	-0.00	0.00	-0.00	0.00	1.00	2746	1794
Outcome_miSkinThickness	0.00	0.00	-0.00	0.01	1.00	2374	1854
Outcome_miBMI	0.02	0.00	0.01	0.03	1.00	2278	1464
Outcome_miGlucose	0.01	0.00	0.01	0.01	1.00	2453	1545
Outcome_miBMI:DiabetesPedigreeFunction	-0.01	0.01	-0.02	0.00	1.00	2545	1459

Regression Coefficients:

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Outcome_Intercept	-0.56	0.27	-1.09	-0.03	1.00	1876	1350
Glucose_Intercept	96.27	3.78	88.80	103.43	1.01	3043	1320
Insulin_Intercept	121.00	21.29	79.91	161.87	1.05	27	127
BloodPressure_Intercept	60.22	1.54	57.10	63.29	1.01	3960	1236
SkinThickness_Intercept	23.92	1.52	20.92	26.88	1.00	1162	1406
BMI_Intercept	31.40	0.88	29.67	33.17	1.00	4913	1407
Outcome_Age	-0.01	0.01	-0.02	0.01	1.00	1768	1286
Outcome_DiabetesPedigreeFunction	0.09	0.05	-0.01	0.20	1.00	5491	1572
Outcome_Pregnancies	0.01	0.01	0.00	0.03	1.01	4454	1079
Glucose_Age	0.80	0.11	0.60	1.02	1.01	3817	1098
Insulin_Age	0.87	0.64	-0.38	2.17	1.11	13	61
BloodPressure_Age	0.37	0.04	0.28	0.46	1.01	5296	1283
SkinThickness_Age	0.16	0.05	0.07	0.25	1.00	1095	1528
BMI_Age	0.03	0.03	-0.02	0.08	1.00	5634	1529
Outcome_miBMI	0.00	0.01	-0.02	0.02	1.00	1771	1283
Outcome_miInsulin	-0.00	0.00	-0.00	0.00	1.01	86	443
Outcome_miBloodPressure	-0.00	0.00	-0.01	0.00	1.00	2536	1662
Outcome_miSkinThickness	-0.00	0.00	-0.01	0.00	1.00	1202	1747
Outcome_miGlucose	0.01	0.00	0.01	0.01	1.00	656	1310
Outcome_miBMI:Age	0.00	0.00	-0.00	0.00	1.00	1846	1402

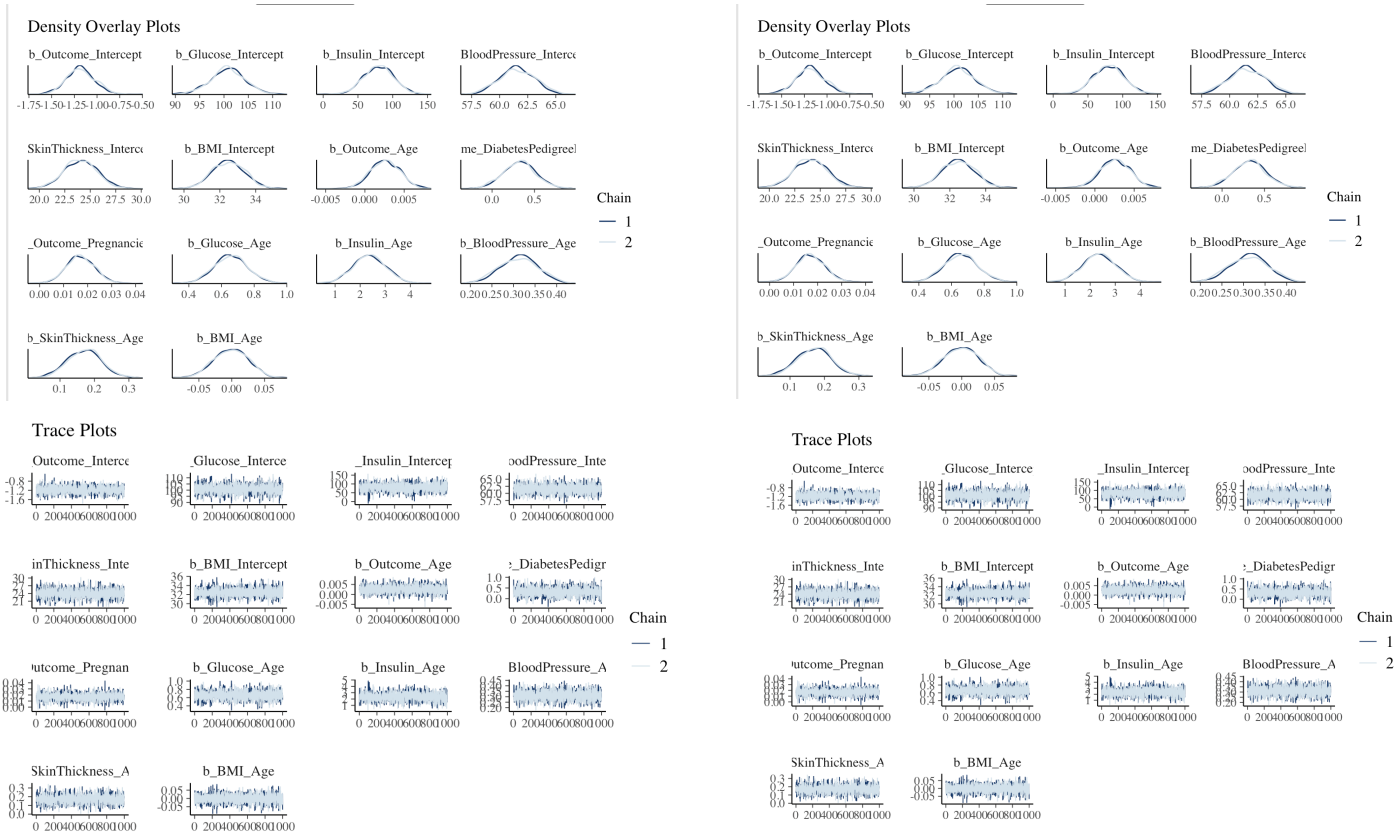
Figure 25: Confusion matrix and diagnostics for Model VI

Accuracy: 0.7719 Precision: 0.7865 F-1 Score: 0.8434		**Reference**	
		0	**1**
Prediction	**0**	70	19
	1	7	18

Figure 26: Confusion matrix and diagnostics for Model VII

Accuracy: 0.8091 Precision: 0.8256 F-1 Score: 0.8712		**Reference**	
		0	**1**
Prediction	**0**	71	15
	1	6	18

Figure 27: Density overlay and trace plots for Models VI (left) and VII (right)



IV. Discussion

Reflection on Results

In this analysis, we focused on 4 different comparisons. The results are outlined below:

I. **Performance of the Bayesian vs. Frequentist logistic models**
Overall, the accuracy of the Bayesian models in predicting the Outcome variable from the other predictors tended to be higher–comparing Models I and II to Models IV-VII show this very clearly. However, this was dependent on the imputation of missing values–use of the BMS imputation increased these models' accuracies significantly. The ability to incorporate prior distributions proved very valuable to the performance of these models, as it allowed for the addition of domain knowledge and the regularization of coefficients, leading to better handling of multicollinearity, and more robust predictors.

II. **Accuracy of 'MICE' vs. 'BRMS' imputation of missing values**
The models that used the BMS imputation for missing values tended to perform better than those using 'MICE' (both in the Frequentist and Bayesian analysis)--comparing Model III to Model IV make this clear, as horseshoe priors and Bayesian models were used in both, but they varied in the NA imputation. This shows that the probabilistic modeling of missing data in BMS imputation provides a more accurate reflection of the underlying data distribution compared to the deterministic approach of MICE, resulting in improved model performance. The Bayesian model's ability to impute missing values in this way is very valuable.

III. **Impact of using Horseshoe vs. Laplace priors on Bayesian regression**
Although both of these priors are sparsity priors, the horseshoe prior is more aggressive in shrinking irrelevant coefficients toward zero while allowing significant predictors to remain large, whereas the laplace prior applies uniform shrinkage to all coefficients. The Bayesian model performed better with the Laplace prior–evident from comparing Model V to Model VI. This indicates that the Laplace prior provided a better balance between bias and variance for this particular dataset, and may be useful for future models using similar data.

IV. **Impact of adding interaction term**
Models VI and VII show that, depending on the variables in the interaction term(s), adding an interaction term can either help or hurt the model (or leave it mostly unchanged). Model VII performed better than Model VI, indicating that choosing interaction terms based on logical suspicion for variable connection is a better option that selecting variables solely based on their significance in the models. Model VII also showed that adding interaction also increases the complexity of the model, which may lead to convergence issues, as the added complexity can cause difficulties in the optimization process and require more computational resources to achieve stable parameter estimates.

Limitations & Potential Improvements

Limitations of the Study Conducted:

I. **Train-Test Split:** Although this experiment used a 70-30% train-test split, cross-validation can offer a more reliable assessment by minimizing the variation in model performance brought on by the particular train-test split. By averaging performance over several splits, cross-validation yields a more accurate assessment of model performance.

II. **Generalizability:** The results of this experiment are specific to the dataset used (and datasets similar in nature), however, the results may not be able to be applied to external datasets, meaning the findings are not generalizable. While the analysis was intended to be specific to this data, the findings may not be as applicable, and other datasets should be considered in the future to see if the same results hold.

III. **Potential for Variability Based on Randomness:** The performance metrics (accuracy, precision, recall, F-1 score) can vary due to randomness in their train-test split. Repeated experiences with different splits or cross-validation provide more stable estimates. While the seed was consistently set throughout this experiment, it is unclear whether the results would change significantly if it was re-ran on a different split, and if the same models would perform better than others. MCMC sampling introduces additional variability in the Bayesian models as well–running the models multiple times can lead to slightly different results due to the stochastic nature of the sampling process.

IV. **Limited Evaluation Metrics:** The analysis primarily focused on accuracy, precision F-1 score and AUC (area under the ROC curve) for model evaluation. However, other metrics such as the Brier score and calibration plots can provide a more comprehensive evaluation of performance, particularly for less-balanced datasets.

Potential Improvements for the Study:

I. **Test for Significance:** Incorporation of formal statistical tests to determine the significance of differences between models would help assess whether the 2%-5% differences in accuracy, F-1, and precision are significant. Techniques like the deLong tests, McNemar's tests, and construction of Bayesian credible intervals would further enhance the understanding of the significance of the findings in this study. While this would require a more nuanced understanding of Bayesian vs. Frequentist analysis, and involve choosing the appropriate analyses, it would nevertheless be a very useful enhancement to these findings.

II. **Cross Validation:** Implementation of cross-validation to evaluate model performance more robustly would reduce dependency on a single train-test split and provide more reliable performance metrics. Using k-fold cross validation, where the data is split into k subsets and the model is trained and tested k times (each time with a different subset as the test set), provides a better estimate of model performance by averaginging results across multiple splits.

III. **Model Complexity vs. Simplification:** Exploring the impact of model complexity on performance could lead to more interpretable models, and potentially better performance by reducing overfitting. Regularization techniques such as lasso and ridge regularization, model selection such as stepwise selection, and other forms of priors (sparsity and non-sparsity) can have different impacts on selecting the most significant predictors, and should be explored further.

IV. **Incorporation of Domain Knowledge:** Integrating domain-specific knowledge into the model development process could lead to the inclusion of relevant predictors and interactions, improving model accuracy and interpretability. Consulting with experts in the field of diabetes research can help identify important variables and potential interactions that may not be apparent from the data alone.

V. **Enhanced Feature Engineering:** Performing more extensive feature engineering to create new variables that may capture underlying patterns in the data more effectively might also help in uncovering relationships that were not evident initially. For example, creating new interaction terms, aggregating variables to capture temporal trends, or incorporating polynomial features could enhance the model's predictive power.

Conclusions: Considerations of the Bayesian vs. Frequentist Models

I. **Influence of Priors**: The choice of priors in Bayesian analysis, as shown above, can significantly influence the results. While we only examined two categories of sparsity priors, the selection of different priors could lead to different conclusions. The dependence on prior distributions introduces a level of subjectivity that is not present in frequentist methods, where parameter estimation is based solely on observed data. Sensitivity to priors can be both a strength and limitation for Bayesian models, depending on the context and availability of prior information. In this case, it proved useful–however, there are several other types of priors that could be considered.

II. **Flexibility in Modeling:** As seen in the models above, Bayesian logistic regression offers greater flexibility in modeling complex relationships and structures within the data–it can easily incorporate non-linear relationships, hierarchical models, interaction terms, and model effects, making it more of a comprehensive framework for analysis, in comparison to frequentist methods, and can integrate itself with most datasets more easily than frequentist methods.

III. **Handling of Missing Data:** Through BMS imputation, the Bayesian models in this analysis were able to model the missing values more effectively than the MICE imputation, which is commonly used in frequentist models. The ability to perform analysis on large datasets without ad-hoc imputation is very useful, as these methods often introduce bias or reduce accuracy.

IV. **Computational Complexity & Convergence**: As seen above, the regression coefficient output for the frequentist models is much more simple and intuitive to interpret than the Bayesian methods, which could make them less practical in certain situations. The Bayesian method also proved much more computationally intensive, and is subject to convergence issues, as seen in Model VII.

In summary, Bayesian logistic regression provides a powerful and flexible framework for analysis, particularly dealing with complex data structures and missing data. Sensitivity to choice

of priors and computational demands are important considerations that have potential to negatively impact the model's accuracy, and must be carefully managed and monitored. In comparison, frequentist methods remain robust and practical, and still exhibit strong performance for most datasets, especially when computational resources are limited, or when prior information isn't available.

References

Alicia A. Johnson, Miles Q. Ott. “A Introduction to Applied Bayesian Modeling.” *Bayes Rules Books*, CRC Press, 1 Dec. 2021, www.bayesrulesbook.com/chapter-13.

Bürkner, Paul. “Handle Missing Values with Brms.” *CRAN-R*, 19 Mar. 2024, cran.r-project.org/web/packages/brms/vignettes/brms_missings.html.

Lukman, P.A. “Bayesian Logistic Regression and Its Application for Hypothyroid Prediction in Post-Radiation Nasopharyngeal Cancer Patients.” *Journal of Physics: Conference Series* , IOP Publishing, 2018, iopscience.iop.org/article/10.1088/1742-6596/1725/1/012010/pdf.

O’Brien, Sean M., and David B. Dunson. “Bayesian Multivariate Logistic Regression.” *OUP Academic*, Oxford University Press, 27 Aug. 2004 academic.oup.com/biometrics/article/60/3/739/7289337.