Demonstrating Enhanced Accuracy with Bayesian Logistic Regression

Harsh Shekhar
hshekha1@jhu.edu

Avantika Singh
asing153@jhu.edu

Ameya Chatur
achatur9@jhu.edu

Abstract

This work investigates the use of Bayesian methods for cancer prediction using the Wisconsin Breast Cancer dataset, which contains features derived from digitized images of breast masses for classifying tumors as malignant or benign. It first leverages Bayesian Ridge Regression for feature selection, integrating prior information and uncertainty to identify the most relevant features, thereby reducing dimensionality and enhancing the focus on informative variables. and then employs Bayesian Logistic Regression, comparing its performance to standard Logistic Regression. The Bayesian approach provides uncertainty quantification in predictions, a critical aspect for medical decision-making, and offers insights into the reliability of model predictions through posterior distributions. The evaluation metrics employed, based on accuracy, precision, recall, and the area under the ROC curve (AUC), demonstrates the effectiveness of Bayesian methods, particularly in scenarios with smaller datasets where traditional models often overfit or underperform. The inclusion of uncertainty metrics further enhances interpretability, underscoring the promise of Bayesian approaches for real-world healthcare applications.

1 Introduction

The primary objective of this project is to develop a robust predictive model for breast cancer classification using the Wisconsin Breast Cancer dataset, which consists of 569 observations and 30 numerical features derived from digitized images of breast mass fine needle aspirates. Each observation is labeled as either **malignant** (1) or **benign** (0), serving as the target variable for classification. This project aims to optimize classification results and identify important predictive features.

1.1 Dataset

The features of the Wisconsin Breast Cancer dataset are grouped into three categories, as shown in Table 1, which represent various cell nuclei characteristics crucial for classification.

The first step is robust feature identification, for which **Bayesian Ridge Regression** is employed to identify relevant features for classification. Then, **Bayesian Logistic Regression** is applied to predict tumor malignancy, utilizing its ability to quantify uncertainty. The performance of Bayesian methods will be compared to **Standard Logistic Regression** to evaluate accuracy, interpretability, and reliability.

Past studies have demonstrated the strengths of Bayesian methods, such as Bayesian Ridge Regression for feature selection and Bayesian Logistic Regression for classification. Bayesian Ridge Regression effectively identifies predictive features by incorporating prior information and penalizing less relevant ones, reducing overfitting risks. This is especially important for high-dimensional datasets with small sample sizes. Bayesian Logistic Regression, with its uncertainty quantification, offers a deeper understanding of model confidence, which is critical in medical decision-making and improves the reliability of predictions in healthcare applications.

Table 1: Feature Categories and Descriptions of the Wisconsin Breast Cancer Dataset

Category	Description
Mean Values	Average measurements across all cells in the sample.
Standard Errors	Variation in the measurements.
Worst Values	Maximum measurements in the sample.
Attributes (Examples)	
Radius	Mean distance from the center to the perimeter of the nuclei.
Texture	Standard deviation of grayscale values.
Perimeter	Total distance around the nuclei.
Area	Size of the nuclei.
Smoothness	Uniformity of the cell shape.
Compactness	Combination of perimeter and area.
Concavity	Severity of concave portions of the nuclei.
Symmetry	Proportionate shape of the nuclei.

1.2 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) focuses on understanding the structure and relationships within the Wisconsin Breast Cancer dataset to inform model development. The dataset comprises 569 observations with 30 numerical features derived from digitized images of cell nuclei, along with a binary target variable indicating tumor malignancy (1 for malignant, 0 for benign).

Descriptive statistics, including mean, standard deviation, and range, are computed for each feature to examine variability and distribution. The class distribution of the target variable is analyzed to check for imbalance between malignant and benign cases, as imbalanced classes can bias model performance and visualized using histograms

Figure 1 shows the histogram of the target variable distribution, illustrating the class imbalance between malignant and benign cases.

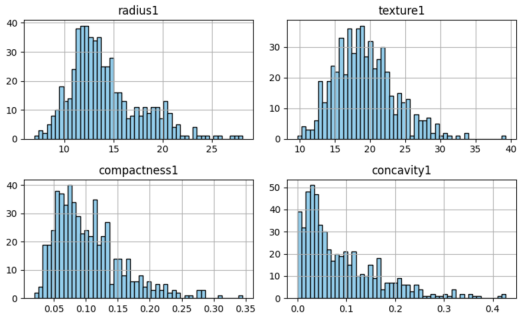

Figure 1: Histogram of the target variable distribution, showing the class imbalance between malignant and benign cases.

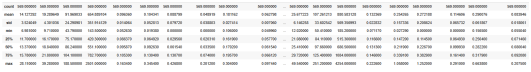

Figure 2: Descriptive statistics for potential predictive features

Furthermore, extensive outlier analysis is conducted by studying boxplots as highlighted in Figure 3. These boxplots are employed to explore relationships between key features and the target variable, providing insights into potential predictive power.

Pairwise correlations between features are assessed using a heatmap to identify multicollinearity, flagging highly correlated variables as candidates for dimensionality reduction as highlighted in Figure 4

Outliers are detected and reviewed for their impact on model performance, while missing or anomalous data is addressed through imputation or removal. These analyses establish a clear understanding of the data, guiding feature selection and preparation for modeling.

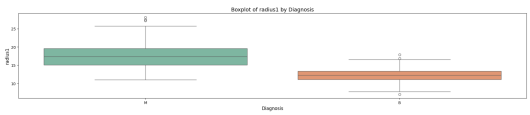

Figure 3: Boxplot for outlier analysis

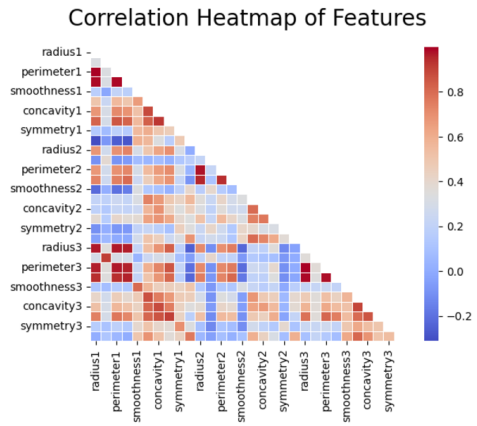

Figure 4: Pairwise correlations between features represented through a heatmap

2 Literature Review

Bayesian methods, particularly Bayesian Logistic Regression (BLR), have gained attention for their ability to handle uncertainty, small datasets, and overfitting through priors. Gosho et al. (2024) present a nonarbitrary approach for setting the prior mean in Bayesian logistic regression, showing that meaningful priors improve predictive accuracy and regularize coefficients in high-dimensional datasets. Bayesian methods are particularly useful in uncertainty quantification. Kavelaars et al. (2023) demonstrate hierarchical Bayesian logistic regression for healthcare datasets, which accounts for treatment heterogeneity and demographic variations, thus improving cancer prediction models' interpretability and decision-making.

In cancer classification, Bayesian methods are also valuable for feature selection. pub (2024) combine interaction-based feature clustering with optimization algorithms, complementing Bayesian regression models that prioritize relevant features and regularize noisy variables. These techniques are particularly effective for high-dimensional, limited-sample datasets.

Despite the success of Bayesian methods with small datasets, scaling to larger datasets remains challenging. Methods like variational inference and Markov Chain Monte Carlo (MCMC) help address this. Huggins et al. (2016) propose coresets, weighted subsets of data that approximate the original dataset, reducing computational overhead while preserving accuracy. Campbell and Broderick (2018) show that coreset construction enables Bayesian logistic regression to scale effectively, ensuring applicability for both small and large datasets.

In summary, Bayesian methods, especially in medical applications such as cancer classification, offer advantages like uncertainty quantification, feature selection, and robustness against overfitting. While computational scalability remains a concern, techniques like variational inference and coreset construction enable efficient solutions for larger datasets.

Feature selection is crucial for predictive modeling, especially with high-dimensional data. Traditional methods like stepwise regression are prone to overfitting, particularly in small datasets. Recently, Bayesian Ridge Regression (BRR) has proven effective for feature selection, as it regularizes coefficients and quantifies uncertainty in parameter estimates. BRR, introduced by Tipping (2001), applies a Gaussian prior on coefficients, incorporating prior knowledge into the model, and controls complexity when the number of features exceeds observations. The posterior distribution of coefficients helps identify important features, as those with non-zero posterior means contribute more to the model's predictive power.

A major advantage of BRR over traditional methods is its ability to penalize irrelevant features with a zero-centered Gaussian prior, encouraging sparsity and improving model interpretability. Studies such

as Hoerl and Kennard (1970) and Wang et al. (2018) show that BRR identifies important predictors, even in datasets with multicollinearity, while selecting a smaller subset of features and improving predictive performance.

In medical contexts like cancer prediction, BRR is valuable due to the high dimensionality of biomarkers or clinical features. By performing feature selection and providing uncertainty estimates, BRR aids in medical decision-making. For example, Ge et al. (2016) used BRR for breast cancer prediction, identifying the most relevant biomarkers and accounting for uncertainty, helping healthcare professionals make more informed decisions.

3 Methodology

This project employs a two-stage approach using Bayesian techniques for feature selection and prediction, followed by a comparative analysis with traditional logistic regression. For both, we normalize and scale the X values to prevent any multicollinearity and reduce errors.

First, **Bayesian Ridge Regression** is applied for feature selection, identifying the most informative predictors by incorporating prior distributions that penalize less relevant features. This step reduces dimensionality, focusing on clinically significant variables. The precision of the predictor weights is calculated, which is then used to perform feature selection.

Next, **Bayesian Logistic Regression** is used to predict tumor malignancy, quantifying uncertainty through posterior distributions of model parameters. A latent variable z is introduced to capture cancer severity, which is predicted as a linear combination of input features and then passed through a sigmoid function to estimate the probability of a malignant diagnosis. The observed diagnosis is modeled as a Bernoulli random variable with this probability.

Using **Bayesian inference**, posterior distributions of regression coefficients and other model parameters are estimated. This method not only performs classification but also offers insights into the latent structure of the data and captures uncertainty in predictions. The *NUTS* sampler, based on Hamiltonian mechanics, is used for continuous variables. The output is thresholded to produce a binary readout of classes, with 1 representing a malignant tumor and 0 otherwise. The optimal threshold is selected using the *AUC ROC curve*.

For comparison, a traditional **Logistic Regression** model is trained on the same dataset, excluding Bayesian regularization and uncertainty quantification. Models are evaluated using metrics like accuracy, precision, recall, F1 score, and AUC. Additionally, Bayesian models are assessed for interpretability through credible intervals for predictions.

The models' performance is evaluated across training and validation datasets to assess generalizability. The entire pipeline, from feature selection to model evaluation, is implemented in Python using libraries like `scikit-learn` and `pymc`, with visualization tools like `matplotlib` and `arviz` to present results. This methodology ensures a thorough and interpretable comparison of Bayesian and traditional models.

4 Results and Discussion

The first step is to perform Bayesian Ridge Regression for feature selection. Below are the detailed results from the analysis.

4.1 Feature Selection with Bayesian Ridge Regression

The Bayesian Ridge Regression model was trained using the features from the training set. The model's output, which consisted of the regression coefficients, was used to calculate the relative importance of each feature. Features with larger coefficients were deemed more important for the prediction.

The selected features from the Bayesian Ridge Regression model with relative importance above 0.001 were:

```
Selected Features: ['radius1', 'smoothness1', 'compactness1', 'concavity1',
'concave_points1', 'symmetry1', 'fractal_dimension1', 'radius2', 'texture2',
'smoothness2', 'compactness2', 'concavity2', 'concave_points2', 'symmetry2',
'fractal_dimension2', 'radius3', 'texture3', 'perimeter3', 'smoothness3',
'compactness3', 'concavity3', 'concave_points3', 'symmetry3', 'fractal_dimension3']
```

These results can be visualized through Figure 5

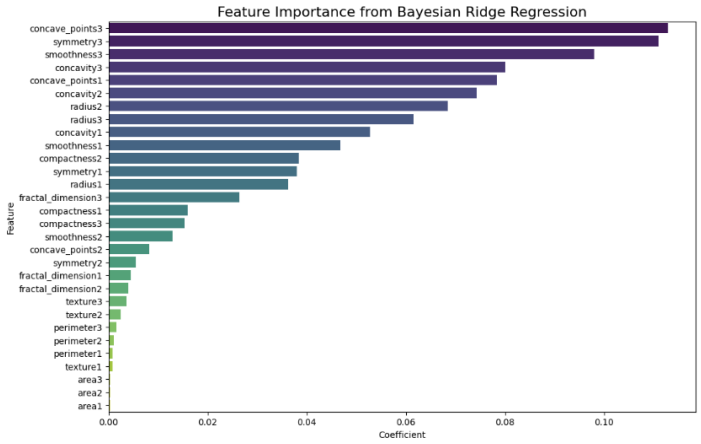

Figure 5: Feature Selection demonstrated through Bayesian Ridge Regression

4.2 Tumor Malignancy Prediction with Bayesian Logistic Regression

The feature selection process identified the relevant features for predicting tumor malignancy. These selected features were then used to train the Bayesian Logistic Regression model. The latent variable model was implemented with a sigmoid function to predict the probability of malignancy based on the input features.

The model was trained with priors set for the regression coefficients and intercept. The latent variable z was calculated as the linear combination of the selected features weighted by the coefficients. This variable was passed through the sigmoid function to yield probabilities for each sample, representing the likelihood of a malignant diagnosis.

4.3 Performance Evaluation for Bayesian Ridge

The model was evaluated using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). The results from the classification report are as follows:

```
              precision    recall   f1-score   support
           0       0.90      1.00      0.96        72
           1       1.00      0.81      0.89        42
    accuracy                           0.93       114
   macro avg       0.95      0.90      0.92       114
weighted avg       0.94      0.93      0.93       114
```

The confusion matrix showed the following:

```
[[72  0]
 [ 8 34]]
```

The results indicated that Bayesian Ridge Regression for feature selection provided a robust fit and can be used to assess feature importance and can be used for model selection. Features with a score greater than 0.001 were accepted as variables for the Bayesian logistic model.

4.4 Posterior Analysis

The Bayesian Logistic Regression takes into account the posterior distributions of the regression coefficients which were analyzed using PyMC3. The `trace` from the sampling process provided insight into the uncertainty of the model parameters. The posterior distributions for the coefficients and intercept were plotted, and summaries of these distributions were generated.

The model was sampled using the NUTS sampler, with 2000 iterations and 1000 tuning steps.

The trace plots for β are shown in Figure 6 and Figure 7 shows the chains.

The trace plots indicate the convergence of the MCMC chains for both the regression coefficients and the intercept. The sampled values show a stable distribution after a sufficient number of iterations,

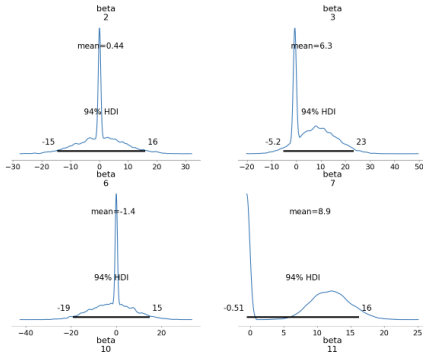

Figure 6: Trace plot for the regression coefficients (β) showing the sampled values across iterations.

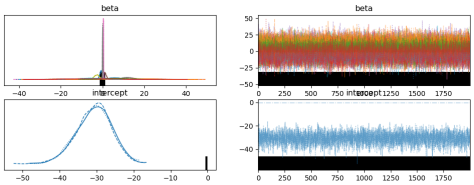

Figure 7: Trace plots resembling hairy caterpillar

indicating reliable parameter estimates. The posterior distribution of latent variable is also showcased in Figure 8

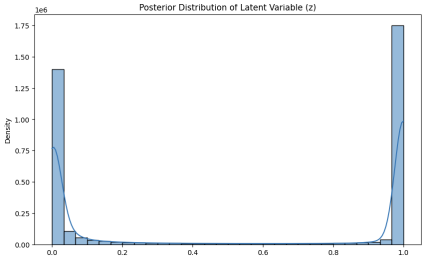

Figure 8: Posterior Distribution of Latent Variable (z)

4.5 Posterior Predictive Sampling

After training the Bayesian logistic regression model on the training data, posterior predictive sampling is performed to evaluate the model's performance on unseen data (X_{test} and y_{test}). This process involves generating multiple samples from the posterior distribution of the model parameters, capturing the uncertainty in the model's predictions. The resulting posterior predictive samples provide a range of possible outcomes for each test point. These samples are averaged to estimate the predicted probabilities, which allows for an evaluation of how well the model generalizes beyond the training data.

Threshold optimization is used to convert predicted probabilities into binary class predictions. The **ROC curve** is used to determine the optimal threshold based on the trade-off between the true positive rate (recall) and the false positive rate. The **Precision-Recall curve** helps optimize the threshold in imbalanced cases, focusing on precision and recall. The optimal thresholds are determined using Youden's J statistic for the ROC curve and the highest F1 score for the Precision-Recall curve. (illustrated in Figure 9)

The model's final predictions are compared against the true labels (y_{test}) to generate a classification report. This report includes precision, recall, F1-score, and support for each class. Additionally, a confusion matrix is computed to visualize the distribution of true positives, true negatives, false positives, and false negatives. The accuracy and ROC AUC scores are also computed to provide an overall evaluation of the model's performance.

	precision	recall	f1-score	support
0	0.97	0.99	0.98	72

```
              1       0.98      0.95      0.96        42
   accuracy                            0.97       114
   macro avg       0.97      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114
```

The confusion matrix showed the following:

```
[[71  1]
 [ 2 40]]
```

The final evaluation metrics—classification report, confusion matrix, accuracy, and ROC AUC—demonstrate the model's strong performance. The accuracy of 0.97 and ROC AUC score of 0.97 indicate that the Bayesian logistic regression model performs well in distinguishing between the classes. The low misclassification rates shown in the confusion matrix suggest that the model has effectively learned to predict the classes with minimal error.

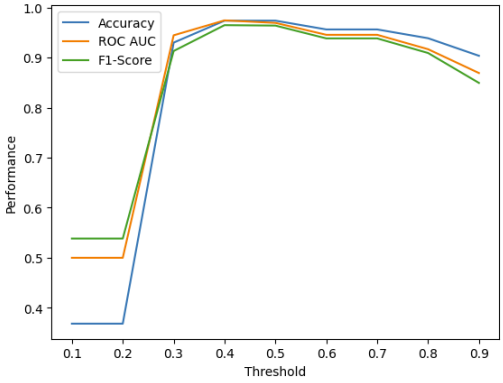

Figure 9: Illustrating F1-Score, ROC and AUC of selected thresholds

5 Model Evaluation Against Frequentist Approach

5.1 Fitting a standard logistic regression model

Fitting a standard logistic regression model using the frequentist approach is crucial, as it serves as a benchmark for comparing the accuracy of a Bayesian logistic regression model. The results of this model are as follows:

```
              precision    recall  f1-score   support
           0       0.91      0.97      0.94        72
           1       0.95      0.83      0.89        42
   accuracy                           0.92       114
   macro avg       0.93      0.90      0.91       114
weighted avg       0.92      0.92      0.92       114
```

The confusion matrix showed the following:

```
[[72  2]
 [ 7 35]]
```

5.2 Feature Selection Comparison

Feature selection results are similar across the board for both the frequentist and Bayesian approaches. Figure 10) illustrates this.

6 Conclusion

The latent variable z provides a continuous representation of the underlying severity of the condition (e.g., cancer severity) that goes beyond the binary diagnosis. This can be useful for understanding the transition between benign and malignant cases. Bayesian inference provides full posterior distributions of parameters (e.g., regression coefficients and z), allowing us to quantify the uncertainty in both the predictions and the model parameters. When working with small datasets or high uncertainty, the Bayesian approach's ability to incorporate priors and quantify uncertainty makes

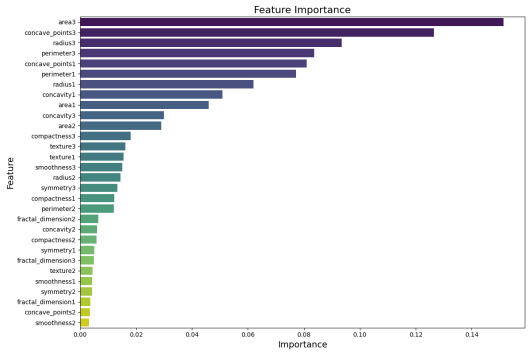

Figure 10: Feature Selection using the Frequentist Approach

Metric	Latent Variable Bayesian Framework	Logistic Regression
Accuracy	97%	92%
F1 Score	0.96	0.89
ROC AUC	0.97	0.90
Precision	0.98	0.95
Recall	0.95	0.83

Table 2: Comparison of metrics between the Latent Variable Bayesian Framework and Logistic Regression.

it superior. For decision-making under uncertainty, such as in medical applications, knowing the confidence in predictions is critical, which Bayesian methods provide inherently. While logistic regression is computationally simpler and faster, the Bayesian latent variable approach is more powerful and flexible, especially for complex problems or when uncertainty matters.

References

2024. Gene selection and cancer classification using interaction-based feature clustering and improved-binary bat algorithm. *PubMed*.

Trevor Campbell and Tamara Broderick. 2018. Scalable bayesian inference via coresets. *Foundations of Data Science*.

X. Ge, S. Li, and Z. Zhang. 2016. Bayesian ridge regression for breast cancer prediction. *Journal of Clinical Bioinformatics*, 6(1):10.

Masahiko Gosho, Ryota Ishii, Kengo Nagashima, Hisashi Noma, and Kazushi Maruo. 2024. Determining the prior mean in bayesian logistic regression with sparse data: a nonarbitrary approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*.

A. E. Hoerl and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Jonathan H. Huggins, Trevor Campbell, and Tamara Broderick. 2016. Coresets for scalable bayesian logistic regression. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4087–4095.

Xander Kavelaars, Johan Mulder, and Ad Kaptein. 2023. Bayesian multilevel multivariate logistic regression for superiority decision-making under observable treatment heterogeneity. *BMC Medical Research Methodology*, 23(220).

M. E. Tipping. 2001. Bayesian ridge regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):265–278.

W. Wang, Z. Zhang, and et al. 2018. Bayesian ridge regression for gene selection in high-dimensional data. *Bioinformatics*, 34(12):2094–2102.