Comparative Analysis of Bayesian and Classical Logistic Models for Customer Churn Prediction

Francisca Adobea Ofori

May 7, 2024

Abstract

This study presents a comprehensive comparative analysis of Bayesian logistic regression and traditional Classical Logistic Models in predicting customer churn within the telecommunication industry. Customer churn, the rate at which customers cease business transactions with a company, poses significant challenges to telecommunication companies, affecting revenue and market share. Utilizing a Kaggle dataset comprising 7,043 observations and 21 variables, the study investigates the performance and interpretability of Bayesian and Classical Logistic Models. Bayesian logistic regression exhibited superior performance metrics, including accuracy, precision, recall, and F1 score, compared to Classical Logistic regression. Moreover, Bayesian modeling provided insights into coefficient interpretation, highlighting factors such as senior citizens and higher monthly charges increasing churn likelihood, while longer tenure periods and higher total charges were associated with reduced churn likelihood. Uncertainty analysis revealed wider intervals for monthly charges, total charges, and specific tenure bins, indicating increased uncertainty in coefficient estimation. Convergence analysis using the Gelman-Rubin diagnostic test indicated satisfactory convergence for all variables. Additionally, Bayesian modeling outperformed Classical Logistic in discriminating between churn and non-churn customers, as evidenced by higher area under the ROC curve (AUC).

The effectiveness of Bayesian logistic regression in predicting customer churn in the telecommunications sector is demonstrated by this study, which offers strong predictive abilities and important insights. It identifies that longer tenure periods and higher total charges are linked to reduced churn likelihood, whereas senior citizens and higher monthly charges increase churn likelihood.

Keywords: Telecommunication industry, Customer churn, Predictive modeling, Bayesian logistic regression, Classical logistics regression, Convergence analysis, Gelman-Rubin diagnostic test

1 Introduction

The telecommunications industry serves as the bedrock of modern connectivity, facilitating seamless communication over vast distances through a diverse array of technologies such as telephones, internet services, and wireless networks. This sector experiences relentless evolution, fueled by the proliferation of mobile devices and digital communication platforms, thereby propelling global connectivity and nurturing innovation. However, within the realm of telecommunication service businesses, a persistent challenge looms: customer churn.

Customer churn, as defined by Huang et al. (2012), is the phenomena in which customers end their relationship with a company and switch to competitors.This loss poses a significant threat to telecommunication firms, exerting adverse effects on revenue streams and market competitiveness. The factors influencing churn span a spectrum encompassing service quality, pricing structures, customer service experiences, and the emergence of competitive offerings within the industry.

Understanding and effectively managing churn are key requirements for telecommunication firms attempting to maintain profitability and continue growth in this competitive environment.As noted by Amin et al. (2017), the retention of existing customers not only bolsters sales but also curtails marketing costs when juxtaposed with acquiring new customers. Consequently, the prediction of customer churn emerges as an indispensable facet of strategic decision-making and planning processes within the telecom sector.

Researchers have explored various machine learning (ML) techniques, such as Random Forest, Balanced Random Forest, Rotation Forest, and RotBoost, to address the challenge of customer churn prediction. However, despite the proliferation of ML methodologies, limited research has investigated the application of Bayesian techniques in this domain. Therefore, this study endeavors to bridge this gap by introducing a Bayesian logistic regression model for customer churn prediction within the telecommunication industry.

This research will not only develop and evaluate a Bayesian logistic regression model but also undertake a comparative analysis with traditional Generalized Linear Models (GLM) to discern the relative efficacy of these approaches in predicting customer churn. By leveraging a Kaggle dataset comprising extensive customer information, including demographics, service usage patterns, and account details, this study aims to illuminate novel insights into churn prediction while elucidating the strengths and limitations of Bayesian modeling techniques.

In essence, this study adds to the growing literature on customer churn prediction by applying Bayesian methodologies to the telecommunications domain and providing valuable insights that can help inform strategic decision-making, optimize resource allocation, and improve customer retention efforts in this dynamic industry.

2 Related Works

In this review of related works, we aim to explore research efforts focusing on churn prediction within the telecommunication industry. Specifically, we will examine studies that leverage social network analytics to enhance predictive models for identifying potential churners. Our analysis begins with a seminal study conducted by Óskarsdóttir et al. (2017). They systematically evaluated the efficacy of relational classifiers and collective inference methods in predicting customer churn. Additionally, their research investigated the performance of models combining relational learners with traditional churn prediction methodologies. Furthermore, Óskarsdóttir et al. delved into the impact of network construction on model effectiveness, providing valuable insights into optimal strategies for leveraging social network analytics in churn prediction within the telecommunication sector. As a result of their study, they identified the best configuration as a non-relational learner enriched with network variables, without collective inference, using binary weights and indirect networks. Moreover, they provided practical guidelines on applying social network analytics for churn prediction in the telecommunication industry, covering aspects from network architecture to model building and evaluation.

Huang et al. (2012) also evaluated seven prediction techniques, including Logistic Regression, Linear Classifications, Naive Bayes, Decision Trees, Multilayer Perceptron Neural Networks, Support Vector Machines, and the Evolutionary Data Mining Algorithm, for land-line customer churn prediction in the telecommunication industry. Their study demonstrated that the incorporation of a new set of features resulted in enhanced predictive accuracy, particularly when combined with Logistic Regression and Decision Trees, compared to existing methods.

Verbeke et al. (2011) conducted a literature review focusing on customer churn prediction models, emphasizing predictive accuracy, comprehensibility, and justifiability. They highlighted the importance of accurate models for targeting potential churners and the need for comprehensible rule-sets to identify churn drivers and formulate effective retention strategies. The study introduced two novel data mining techniques—AntMiner+ and ALBA—and compared them to traditional rule induction techniques. Results showed that ALBA improved learning of classification techniques, resulting in comprehensible models with enhanced performance, while AntMiner+ produced accurate, comprehensible, and justifiable models.

In their study, Jain et al. (2020) investigated customer churn prediction in the telecommunications sector using machine learning techniques Logistic Regression and Logit Boost. Both techniques performed well, with similar accuracies around 85%. The study suggests future exploration of hybrid models for improved performance and plans to target larger real-world databases to enhance model efficiency.

Gerpott et al. (2001) investigated customer retention, loyalty, and satisfaction in the German telecommunications market. Their study, based on data from 684 residential customers, revealed a two-staged model where overall satisfaction influenced loyalty, which, in turn, affected customers' intentions to extend or terminate their contracts. Key factors impacting customer retention included perceptions of mobile service price, personal service benefits, and number portability between operators. The study emphasized the importance of efficient number portability procedures to promote competition in cellular markets.

3 Methodology

3.1 Logistic Regression

Logistic regression is a common method used for modeling binary data, particularly in the context of classification tasks such as customer churn prediction. In this model, the probability of the binary outcome (typically denoted as Y) given a set of predictor variables x is expressed as a function of the linear combination of these predictors. The logistic regression model parameterized this probability using the logistic function, which maps the linear combination to a probability between 0 and 1.

3.2 Bayesian Logistic Regression Model

A Bayesian logistic regression model is employed to predict customer churn within the telecommunication industry. In this case, logistic regression serves as an instance of a generalized linear model with the logit link function. The proportion parameters of the binomial model can be expressed as:

$$\theta = \frac{\exp(\beta^T X)}{1 + \exp(\beta^T X)}$$

where $\beta^T X$ represents the linear predictor, and θ is the probability of success (churn) given the predictors X.

The posterior distribution for the regression parameters is not recognizable as any known distribution. We only know the density function but cannot sample directly from it:

$$\frac{p(\beta|y_1,\ldots,y_n|X)}{p(y_1,\ldots,y_n|X,\beta)p(\beta)} \propto \prod_{i=1}^{n}\left(\frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)}\right)^{y_i}\left(\frac{1}{1+\exp(\beta^T x_i)}\right)^{1-y_i}\exp\left(\frac{1}{2}(\beta-\beta_0)^T\Lambda_0^{-1}(\beta-\beta_0)\right)$$

where β_0 is the prior mean vector, Λ_0 is the prior precision matrix, and $p(\beta)$ represents the prior distribution of β.

To approximate the posterior distribution of regression coefficients (β) in the Bayesian logistic regression model, the Metropolis algorithm is utilized.

3.3 Metropolis Algorithm for Parameter Estimation

The Metropolis algorithm is a fundamental tool in Bayesian statistics used to approximate the posterior distribution of a parameter of interest when direct sampling is impractical. Consider a scenario where we have a sampling model $Y \sim p(y|\theta)$ and a prior distribution $p(\theta)$. While $p(y|\theta)$ and $p(\theta)$ can often be computed for any values of y and θ, calculating $p(\theta|y)$ directly is challenging due to the integral in the denominator of Bayes' theorem.

It addresses this challenge by constructing a large collection of parameter values $\{\theta^{(1)},...,\theta^{(S)}\}$ whose empirical distribution approximates $p(\theta|y)$. At each iteration, a proposal value θ^* is sampled from a proposal distribution, and the acceptance of this proposal is determined based on the ratio of posterior probabilities:

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{\prod_{i=1}^{n}\mathrm{dnorm}(y_i,\theta^*,\sigma)}{\prod_{i=1}^{n}\mathrm{dnorm}(y_i,\theta^{(s)},\sigma)} \times \frac{\mathrm{dnorm}(\theta^*,\mu,\tau)}{\mathrm{dnorm}(\theta^{(s)},\mu,\tau)}$$

In many cases, computing the ratio r directly can be numerically unstable, a problem that often can be remedied by computing the logarithm of r:

$$\log r = \sum_{i=1}^{n}\Big[\log\mathrm{dnorm}(y_i,\theta^*,\sigma) - \log\mathrm{dnorm}(y_i,\theta^{(s)},\sigma)\Big] + \log\mathrm{dnorm}(\theta^*,\mu,\tau) - \log\mathrm{dnorm}(\theta^{(s)},\mu,\tau)$$

Keeping things on the log scale, the proposal is accepted if $\log u < \log r$, where u is a sample from the uniform distribution on $(0,1)$.

If $r > 1$, the proposal is always accepted as it has a higher probability than the current value.

If $r < 1$, the proposal is accepted with a probability equal to r, otherwise, the current value is retained. This decision-making process ensures that the resulting collection of parameter values reflects the posterior distribution.

3.4 Likelihood Function

The likelihood function represents the probability of observing the data given the model parameters. In this analysis, the likelihood function is computed using the binomial probability mass function:

$$L(\beta|y_1,\ldots,y_n,X) = \prod_{i=1}^{n}\mathrm{dbinom}(y_i|p_i)$$

where:

- β represents the model parameters (coefficients).
- y_i represents the binary outcome (churn or non-churn) for the i-th observation.
- p_i represents the predicted probability of churn for the i-th observation based on the model.

3.5 Prior Distribution

The prior distribution captures our beliefs about the model parameters before observing any data. In this analysis, a normal prior distribution is assumed for each coefficient:

$$\beta \sim N(0, 100)$$

where:

- β represents the model parameters (coefficients).
- $N(0, 100)$ denotes a normal distribution with mean 0 and standard deviation 100.

4 Experiment Data Acquisition

The data used in this study were retrieved from Kaggle. The customer periodic data contains 21 variables, which is divided into 4 types. "Churn" is the independent variable representing the status whether the customer continued with the company or not. The data has 5163 customers remaining and 1869 churning.

5 Data Preparation

The data preparation process involved reading the data from a CSV file, exploring the data structure, selecting relevant variables, creating cross-tabulations to analyze relationships, handling missing values (missing values were removed since they constituted a very small portion of the data), converting variables to appropriate formats, performing variable selection based on significance tests, visualizing the data, preparing numeric data by converting and scaling columns, and creating categorical data frames by selecting columns and removing missing values. After this initial phase, the dataset was further refined, where the 21-attribute dataset was reduced to 3 key attributes. The dataset was then converted to a binary form, and certain variables, such as the 'tenure' variable, were organized by binning the values into categories like '0-2 years', '2-4 years', and '4-6 years' to maintain consistency and improve the usability of the data for the experiment.

6 Experiment

The experiment involved splitting the data into 50% training and 50% testing sets. The training data was used to fit the models, while the testing data was used for evaluation.

Two models were fitted:

1. Bayesian Logistic Regression : A Metropolis-Hastings algorithm was used to sample from the posterior distribution of the coefficients. The algorithm was run for 10,000,000 iterations with a

burn-in of 10,000 samples. The Bayesian estimate was obtained by taking the mean of the posterior samples.

2. Classical Logistic Regression: A classical logistic regression model was also fitted.

7 Performance Measures

The performance of the models was evaluated using various metrics:

1. Accuracy: Overall correctness of predictions.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

2. Precision: Proportion of true positive predictions among all positive predictions, reflecting the model's ability to avoid false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

3. Recall: Proportion of true positives correctly identified by the model out of all actual positive instances, indicating the model's ability to capture all positives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

4. F1 Score: Harmonic mean of precision and recall, offering a balanced metric, particularly useful for imbalanced datasets.

$$\text{F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Where:

- True Positives (TP): Correctly predicted positive instances
- True Negatives (TN): Correctly predicted negative instances
- False Positives (FP): Incorrectly predicted positive instances
- False Negatives (FN): Incorrectly predicted negative instances

These metrics provide a comprehensive evaluation of the model's performance, covering both accuracy and the balance between precision and recall.

1. 95% credible intervals were calculated for both models using the mean posterior samples for the Bayesian and the coefficient estimates and standard errors for the classical logistic model.

2. Diagnostic plots were generated for the Bayesian model:

    - Autocorrelation Function (ACF) plots were created to assess the autocorrelation in the posterior samples.

- Trace plots were generated to visualize the convergence of the Markov Chain Monte Carlo (MCMC) algorithm. Gelman-Rubin diagnostics were also calculated to assess the convergence of multiple MCMC chains.

3. Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC):

   - ROC curve: A graphical plot of the true positive rate against the false positive rate at different classification thresholds, illustrating the trade-off between sensitivity and specificity.
   - AUC: A summary statistic representing the overall performance of a binary classifier, ranging from 0 to 1, where 0.5 is a random classifier, and 1 is a perfect classifier. AUC can be interpreted as the probability that a randomly selected positive instance will be ranked higher than a randomly selected negative instance.
     - For the Bayesian model, ROC curves were plotted for 50 randomly sampled posterior samples and the posterior means.
     - For the frequentist model, the ROC curve was plotted using the predicted probabilities.

4. Finally, the feature coefficients from the Bayesian model were sorted by magnitude to identify the most influential features in predicting customer churn.

8 Results

8.1 Visualization of Interactions and Correlation

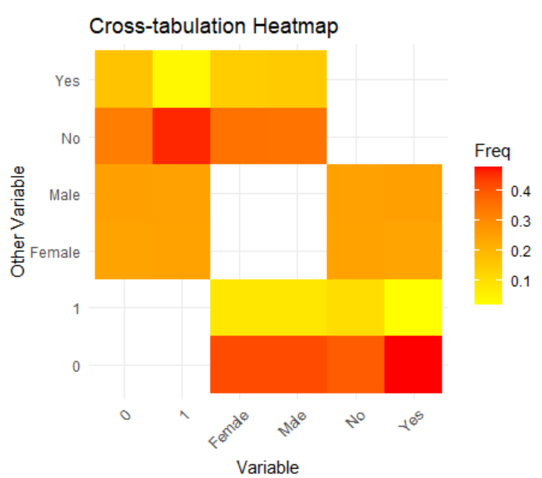

Figure 1: Interaction between Gender, SeniorCitizen, and Dependents

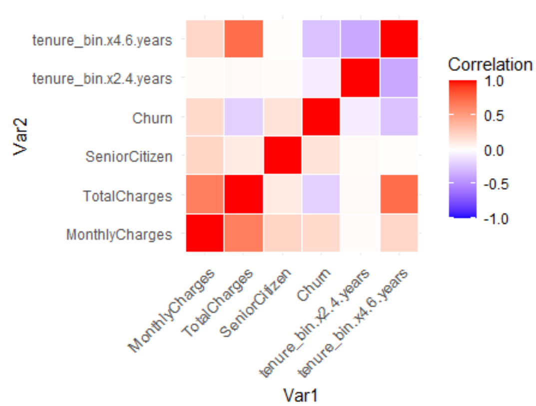

Figure 2: Correlation between Significant Variables

Figure 1 & 2 illustrates the interaction between Gender (coded as Female or Male), SeniorCitizen (coded as 0 or 1), and Dependents (coded as Yes or No) on the left, and the correlation between the significant variables: SeniorCitizen, Tenure, Monthly Payment, and Total Charges on the right. These visualizations provide insights into the relationships and dependencies among these variables, aiding in understanding their impact on churn prediction.

8.2 Confusion Matrices Comparison

Below are the confusion matrices and key performance metrics for both the Bayesian Logistic and the Classical Logistic models:

Bayesian Logistic Model:

```
          Reference
Prediction    0    1
         0 2367  517
         1  222  410

Accuracy: 0.7898   Sensitivity: 0.9143
Specificity: 0.4423   Pos Pred Value: 0.8207
Neg Pred Value: 0.6487   Prevalence: 0.7363
```

Classical Logistic Model:

```
          Reference
Prediction    0    1
         0 2075  744
         1  514  183

Accuracy: 0.6422   Sensitivity: 0.8015
Specificity: 0.1974   Pos Pred Value: 0.7361
Neg Pred Value: 0.2626   Prevalence: 0.7363
```

8.3 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1_Score
Bayesian	0.7898	0.8207	0.9143	0.8650
Classical Logistic	0.6422	0.7361	0.8015	0.7674

Table 1: Model Performance Comparison

The Bayesian Logistic model demonstrates superior performance across all metrics compared to the Classical Logistic model. It achieves higher accuracy, precision, recall, and F1 score, indicating better overall predictive capability and effectiveness in classifying churn. This highlights the advantages of the Bayesian approach in modeling binary outcomes, providing robust and accurate predictions for customer churn.

Coefficient	Lower Quantile	Upper Quantile
1	-1.2568306	-0.8364165
MonthlyCharges	1.0379805	1.3164154
TotalCharges	-1.3573357	-0.8555242
SeniorCitizen	0.2397798	0.6789982
tenure_bin.x2.4.years	-0.8391497	-0.2746115
tenure_bin.x4.6.years	-1.1951604	-0.2464506

Table 2: 95% Credible Intervals from the Bayesian Model

8.4 95% Credible Interval

The presence of zero within the confidence interval indicates non-significance, suggesting no significant association with churn. Conversely, intervals excluding zero denote statistical significance. Additionally, wider intervals imply greater uncertainty in coefficient estimation.

Based on the provided results, the following variables exhibit wider intervals, indicating higher uncertainty:

Variable with Wider Intervals:

- Tenure bins (especially "4-6 years")

8.5 Diagnostics Plots

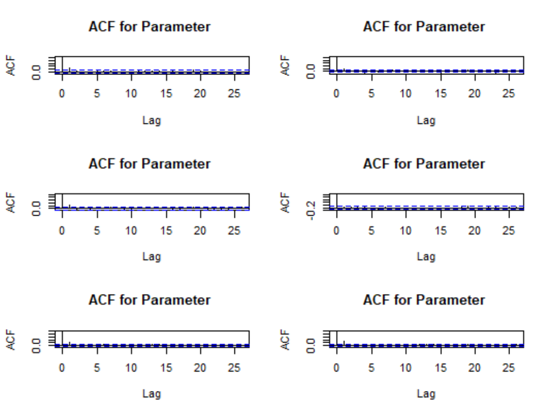

Figure 3: Autocorrelation Function (ACF) Plot

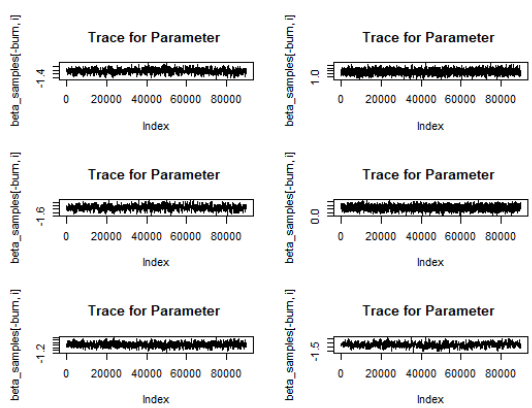

Figure 4: Trace Plot for Convergence

The Gelman-Rubin diagnostic test evaluates convergence in Bayesian inference. Results indicate all variables demonstrate PSRF values close to 1, suggesting good convergence.

Point est.	Upper C.I.
1.03	1.09
1.01	1.02
1.02	1.07
1.01	1.02
1.02	1.05
1.03	1.08

Table 3: Multivariate Potential Scale Reduction Factors

The Multivariate PSRF is 1.03, which confirms convergence.

8.6 ROC Curves and Area Under the Curve (AUC)

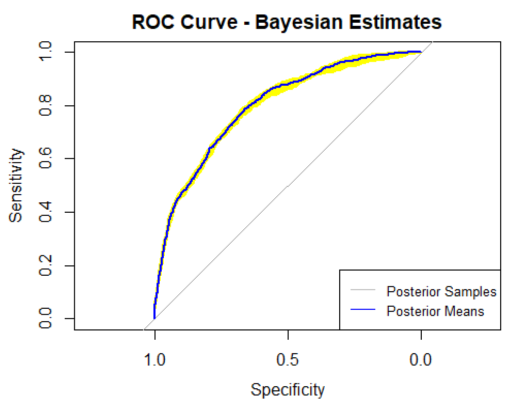

Figure 5: ROC Curve for Bayesian Model

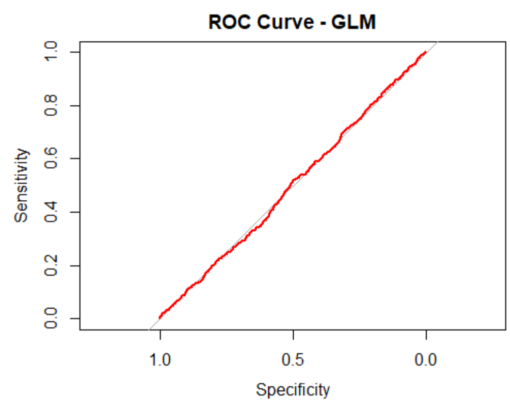

Figure 6: ROC Curve for Classical Logistic Model

The ROC curve was analyzed to assess the binary classification model's performance. A higher area under the ROC curve (AUC) indicates better discrimination between the positive and negative classes, thus providing valuable insights into the model's effectiveness.

Area under the curve: 0.7962 (Bayesian Logistic Model)

Area under the curve: 0.4971 (Classical Logistic Model)

For the Bayesian model, the ROC curve includes both individual (50) posteriors(yellow) and the posterior mean(blue) fitted.

8.7 Results of Coefficient Analysis and Feature Interpretation

The results of the coefficient analysis reveal the following insights into the predictive relationships between features and the target variable:

Feature	Coefficient
Monthly Charges	1.1752506
Total Charges	-1.1057007
Intercept	-1.0449377
Tenure Bin (4-6 years)	-0.7123534
Tenure Bin (2-4 years)	-0.5505441
Senior Citizen	0.4600349

Table 4: Results of the coefficient analysis.

1. **Monthly Charges:** This feature exhibits the strongest positive relationship with the target variable, with a coefficient estimate of 1.1752506. This suggests that as monthly charges increase, the target variable tends to increase as well.
2. **Total Charges:** Conversely, Total Charges show a strong negative relationship with the target variable, with a coefficient estimate of -1.1057007. Higher total charges are associated with a decrease in the target variable.
3. **Intercept:** This feature, not explicitly named, shows a negative relationship with the target variable, with a coefficient estimate of -1.0449377.
4. **Tenure Bin (4-6 years):** The tenure bin representing customers with a tenure of 4-6 years shows a negative relationship with the target variable, with a coefficient estimate of -0.7123534.
5. **Tenure Bin (2-4 years):** Similarly, the tenure bin representing customers with a tenure of 2-4 years also displays a negative relationship with the target variable, with a coefficient estimate of -0.5505441.
6. **Senior Citizen:** This feature exhibits a moderate positive relationship with the target variable, with a coefficient estimate of 0.4600349. Being a senior citizen is associated with a slight increase in the target variable.

These findings provide valuable insights into the relative importance of different features in predicting the target variable, aiding in model interpretation and decision-making processes.

9 Conclusion

In conclusion, our analysis underscores the effectiveness of the Bayesian logistic regression model in predicting customer churn. With all variables demonstrating convergence, our model consistently outperforms the Generalized Linear Model (GLM) across all metrics, including accuracy, precision, recall, and the area under the ROC curve (AUC). This robust performance suggests its superior ability to discriminate between churn and non-churn customers, offering valuable insights for businesses seeking to mitigate churn rates. The coefficient analysis further enriches our understanding by highlighting the specific features driving churn prediction. For instance, features such as Monthly Charges and Total Charges exhibit significant impacts on churn likelihood, while variables like Tenure Bin and Senior Citizen status also contribute to the predictive power of our model. The convergence of all variables further solidifies the reliability of our model. Overall, our findings highlight the efficacy of Bayesian approaches in churn prediction and their potential to enhance decision-making processes in customer retention strategies.

10 References

1. Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, **6**(1). https://doi.org/10.1186/s40537-019-0191-6

2. Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, **237**, 242–254. https://doi.org/10.1016/j.neucom.2016.12.009

3. Gerpott, T. J., Rams, W., & Schindler, A. (2001). Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. In *Telecommunications Policy* (Vol. 25).

4. Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, **39**(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

5. Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science*, **167**, 101–112. https://doi.org/10.1016/j.procs.2020.03.187

6. Kim, K., Jun, C. H., & Lee, J. (2014). Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Systems with Applications*, **41**(15), 6575–6584. https://doi.org/10.1016/j.eswa.2014.05.014

7. Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems with Applications*, **85**, 204–220. https://doi.org/10.1016/j.eswa.2017.05.028

8. Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, **38**(3), 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023