STAT-I 529
Customer Churn Analysis Using Bayesian and Frequentist Approaches

Fahad Mehfooz
fmehfooz@iu.edu

Problem Statement

Objective:

To predict customer churn and analyze the factors contributing to it using both Bayesian and frequentist methods, providing insights into uncertainty quantification and the impact of various features.

Data Description

- **Dataset used: Telco Customer Churn**
- Dataset Source- Kaggle
- 7,032 customers, 20 features
- **Customer Demographics-** Includes details like gender, senior citizen status, and whether the customer has a partner or dependents.
- **Account Information -** Covers the tenure with the company, contract type, billing method (paperless or not), and payment method.
- **Services Subscribed-** Details whether the customer has phone service, multiple lines, internet service, and additional services like online security, backup, device protection, tech support, streaming TV, and movies.
- **Charges-** Includes monthly charges and total charges incurred by the customer.
- **Churn-** Indicates whether the customer has churned (binary).
- Dataset is imbalanced. We have 73% non-churned samples and 27 % churned samples.

More About The Data

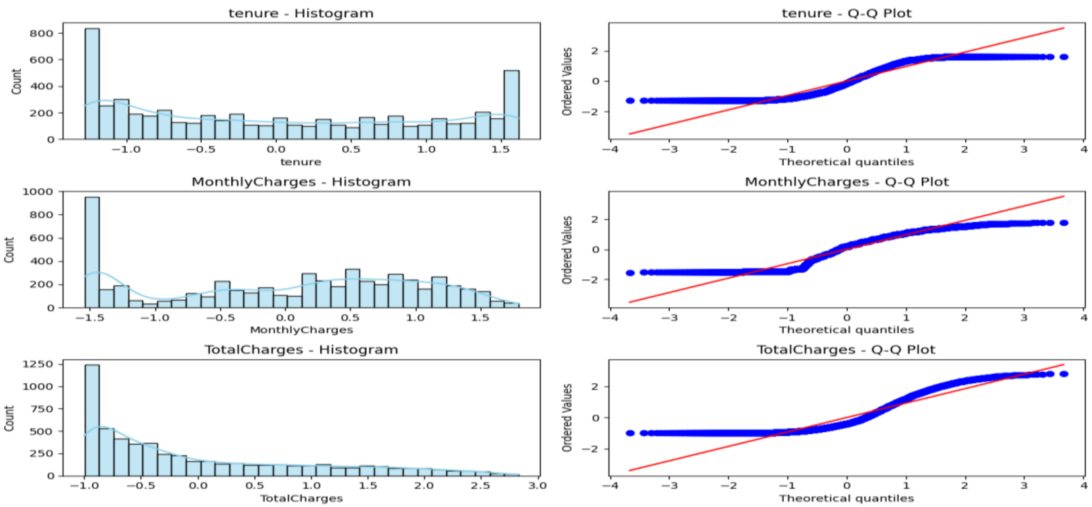
- customerID: Unique identifier for each customer (not used in modeling).
- gender: Gender of the customer (categorical).
- SeniorCitizen: Indicates if the customer is a senior citizen (binary).
- Partner: Indicates if the customer has a partner (binary).
- Dependents: Indicates if the customer has dependents (binary).
- tenure: Number of months the customer has been with the company (continuous).
- PhoneService: Indicates if the customer has phone service (binary).
- MultipleLines: Indicates if the customer has multiple lines (binary).
- InternetService: Type of internet service (categorical).
- OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV,
- StreamingMovies: Indicates various services the customer subscribes to (binary).
- Contract: Type of contract the customer is under (categorical).
- PaperlessBilling: Indicates if the customer has paperless billing (binary).
- PaymentMethod: Payment method used by the customer (categorical).
- MonthlyCharges: Monthly charges for the customer (continuous).
- TotalCharges: Total charges incurred by the customer (continuous).
- Churn: Target variable indicating whether the customer has churned (binary).

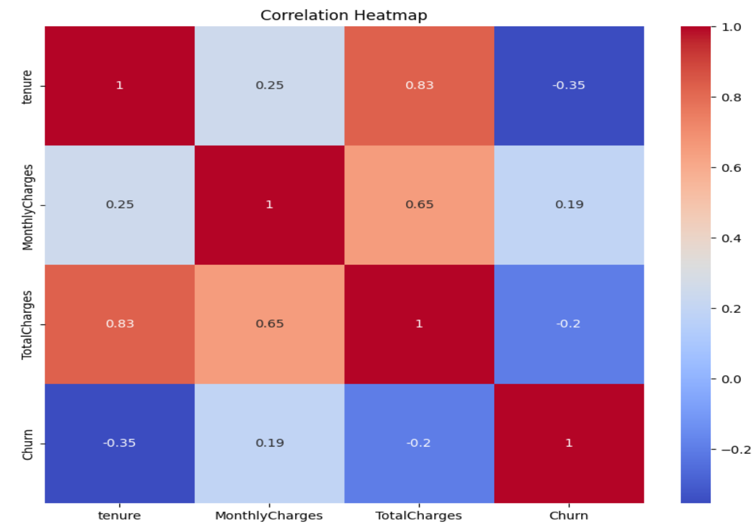
Data Preprocessing

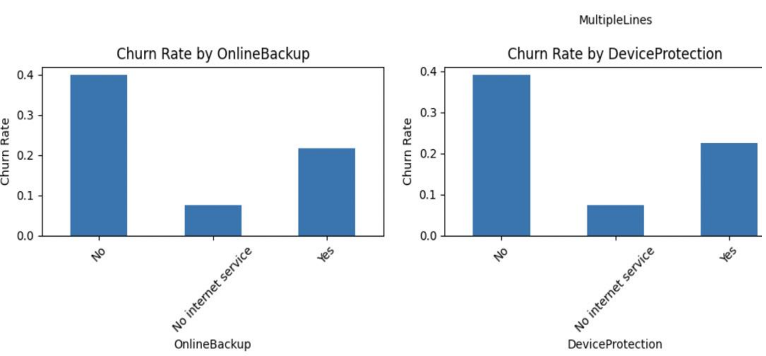
- Handling Missing Values: Check for missing values and impute or remove them.
- Encoding Categorical Variables: Used label encoding for categorical/binary variables.
- Feature Scaling: Standardize continuous variables.

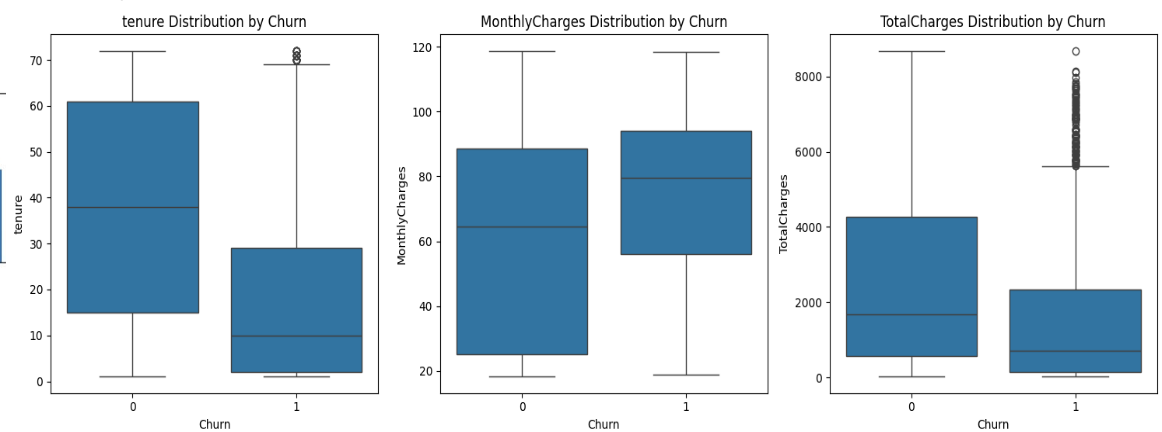
Exploratory Data Analysis (EDA)

- Strong correlations between: Tenure and total charges Monthly. Charges and total charges
- High churn rates observed for: Fiber optic internet service, customers without online security/backup, customers without device protection/tech support









Statistical Testing Results

- Significant association between contract type and churn- Used chi-square test.
- Monthly charges differ significantly between churned and non-churned customers- Independent Samples t-Test (Welch’s)
- Proportion of customers with online security differs between churned and non-churned- Proportion Z-Test
- Strong correlation between total charges and tenure- Pearson correlation test.

Variable Selection

- Used all predictors for Bayesian modelling.
- We can later comment on if we need to drop some features as part of backward feature selection.

Bayesian Framework

Choosing Priors

- **Priors** represent initial beliefs about parameters before oberving data.
- **Normal priors** were chosen for all predictors because they are:
  - Symmetric and unbounded, making them versatile for regression coefficients.
  - Non-restrictive, allowing data to shape the posterior effectively.
- A **mean of 0** and a moderate **standard deviation** provide a reasonable starting point given the lack of strong prior knowledge.
- Priors can be adjusted based on posterior analysis as needed.
- Sampling: Metropolis-Hastings algorithm
- 2,000 tuning steps, 2,000 draws, 4 chains

Logistic Regression Likelihood Calculation

- **Likelihood of logistic regression is calculated as:**

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\pi_i^{y_i}(1-\pi_i)^{1-y_i}\right]$$

Where:

- $\pi_i = P(y_i = 1 | \mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{1+e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}}$: The logistic function.
- $\mathbf{x}_i$: The feature vector for the i-th observation.
- $\boldsymbol{\beta}$: The parameter vector (coefficients of the model).
- y_i: The observed binary outcome for the i-th observation.

Posterior Calculation

Uses the formula to compute the posterior.

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})}$$

Posterior Summary Analysis For Normal Priors

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
beta_dependents	-0.298	0.103	-0.498	-0.127	0.008	0.005	187.0	304.0	1.01
beta_device_protection	-0.175	0.093	-0.347	0.006	0.023	0.017	17.0	106.0	1.10
beta_gender	-0.032	0.070	-0.156	0.101	0.014	0.010	24.0	46.0	1.09
beta_is_dsl	0.026	0.229	-0.363	0.390	0.137	0.109	3.0	87.0	1.74
beta_is_fiber_optic	0.634	0.377	0.058	1.149	0.247	0.203	3.0	16.0	1.83
beta_monthly_charges	0.421	0.210	0.074	0.715	0.135	0.110	3.0	10.0	1.86
beta_multiple_lines	0.233	0.098	0.049	0.389	0.018	0.015	27.0	35.0	1.09
beta_online_backup	-0.197	0.097	-0.382	-0.055	0.015	0.011	41.0	134.0	1.04
beta_online_security	-0.623	0.097	-0.783	-0.448	0.024	0.018	18.0	175.0	1.11
beta_paperless_billing	0.348	0.077	0.225	0.504	0.011	0.008	53.0	67.0	1.05
beta_partner	0.069	0.084	-0.097	0.190	0.008	0.006	102.0	108.0	1.01
beta_phone_service	-0.986	0.186	-1.355	-0.663	0.096	0.074	4.0	53.0	1.55
beta_senior_citizen	0.354	0.094	0.190	0.537	0.008	0.005	156.0	175.0	1.01
beta_streaming_movies	0.107	0.114	-0.119	0.303	0.045	0.033	7.0	82.0	1.27
beta_streaming_tv	0.085	0.112	-0.130	0.278	0.042	0.031	7.0	53.0	1.25
beta_tech_support	-0.642	0.102	-0.884	-0.491	0.010	0.007	99.0	219.0	1.04
beta_tenure	-1.704	0.154	-1.976	-1.449	0.073	0.055	5.0	19.0	1.39
beta_total_charges	0.649	0.184	0.370	1.005	0.079	0.059	6.0	25.0	1.33
intercept	-0.758	0.453	-1.392	-0.120	0.309	0.258	3.0	21.0	1.96

1. Convergence Issues (R-hat > 1.1)- Poor convergence for **beta_is_dsl**, **beta_is_fiber_optic**, **beta_monthly_charges**, **beta_tenure**, and intercept. Likely due to insufficient iterations or poorly chosen priors.

2. Low Effective Sample Size (ESS)- High autocorrelation for **beta_phone_service**, **beta_streaming_movies**, **beta_streaming_tv**, and intercept. Chains are mixing poorly; more iterations needed.

3. Wide HDI Intervals- High uncertainty for **beta_is_fiber_optic** ([0.058, 1.149]) and **beta_tenure** ([-1.976, -1.449]). Indicates vague priors or insufficient data.

4. High Monte Carlo Standard Error (MCSE)- Requires additional sampling for **beta_is_dsl**, **beta_fiber_optic**, and **beta_monthly_charges**.

5. Positive Coefficients (Strong Influence)

- **beta_is_fiber_optic**: Strong (+0.680), but wide HDI and poor convergence.
- **beta_paperless_billing**: Reliable (+0.365, low R-hat).
- **beta_senior_citizen**: Consistent positive effect.

6. Negative Coefficients (Strong Influence)

- **beta_tenure**: Strong negative (-1.692, well-defined posterior).
- **beta_online_security**: Substantial negative (-0.619, reliable).
- **beta_phone_service**: Negative (-0.871), but poor convergence.

7. Unclear Influence (HDI Includes Zero)

- **beta_device_protection**, **beta_gender**, **beta_streaming_movies**: No significant effect.

8. Key Actions

- **Adjust Priors:** Tighten priors for parameters with wide HDIs.
- **Increase Sampling:** Resolve convergence issues (high R-hat, low ESS).

Proposing New Priors

- **Updated Priors**: Cauchy prior selected for parameters like beta_is_dsl, beta_is_fiber_optic, beta_monthly_charges, and beta_total_charges.
- Remaining Priors: Unchanged.

- **Reason**: Cauchy has heavy tails, making it more **robust to outliers** and **extreme values**.**Context**: Logistic regression often involves uncertain predictors, and the **Normal prior** might be too restrictive.

- **Effectiveness**: The Cauchy prior allows the model to explore a wider range of parameter values, ensuring flexibility for features with potential heavy-tailed distributions (like total_charges or monthly_charges).

Posterior Summary Analysis For Updated Priors

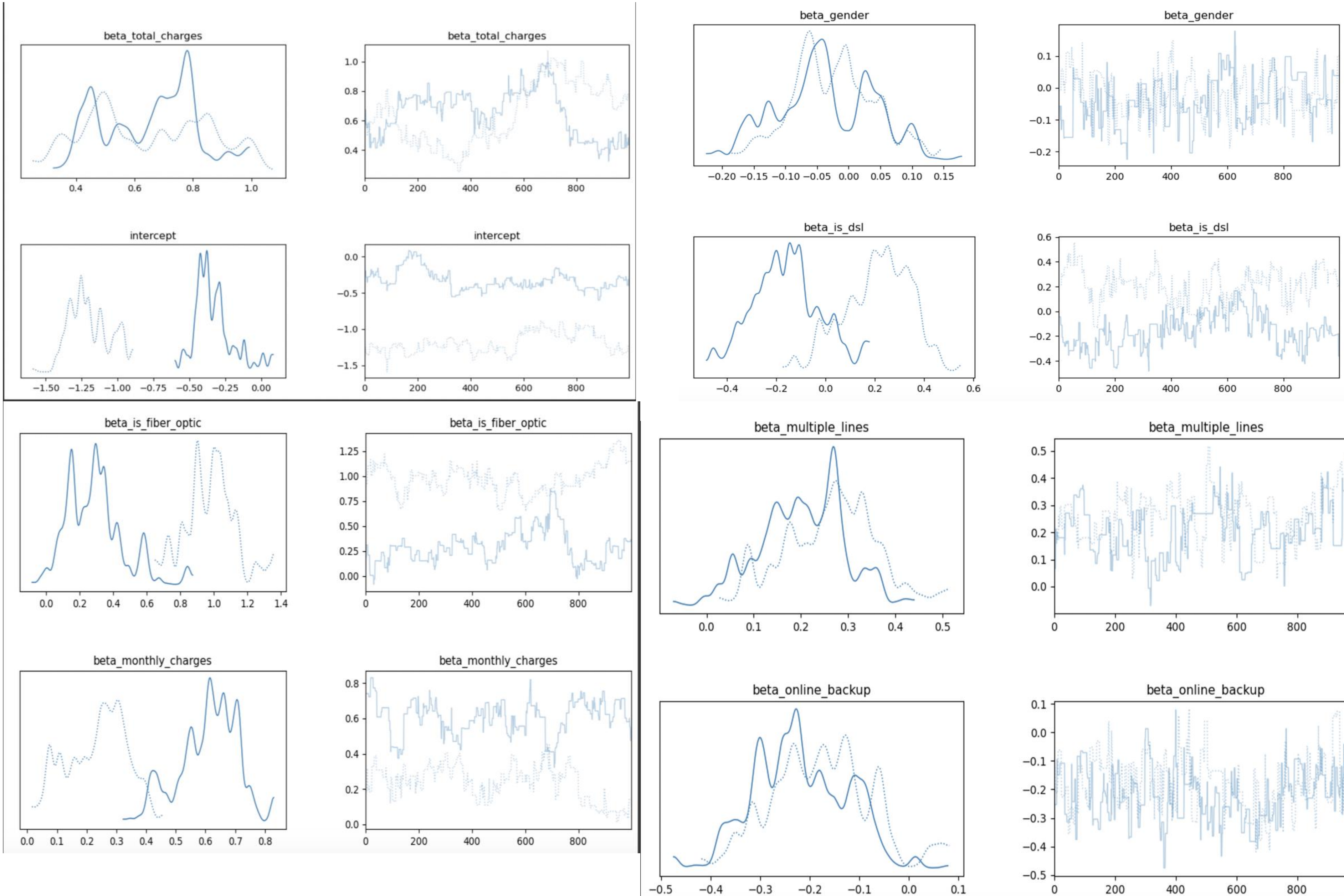
	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
beta_dependents	-0.298	0.101	-0.492	-0.122	0.004	0.003	528.0	794.0	1.01
beta_device_protection	-0.115	0.116	-0.333	0.100	0.037	0.027	10.0	44.0	1.30
beta_gender	-0.031	0.075	-0.182	0.102	0.003	0.002	509.0	714.0	1.01
beta_is_dsl	0.350	0.483	-0.666	1.019	0.229	0.174	5.0	13.0	2.70
beta_is_fiber_optic	1.205	0.876	-0.650	2.336	0.422	0.321	4.0	12.0	3.15
beta_monthly_charges	0.282	0.966	-1.055	2.296	0.463	0.352	5.0	12.0	2.94
beta_multiple_lines	0.262	0.119	0.044	0.500	0.039	0.029	9.0	59.0	1.36
beta_online_backup	-0.147	0.111	-0.369	0.053	0.032	0.023	12.0	70.0	1.25
beta_online_security	-0.571	0.114	-0.772	-0.355	0.035	0.026	11.0	88.0	1.29
beta_paperless_billing	0.342	0.082	0.185	0.490	0.004	0.003	341.0	660.0	1.02
beta_partner	0.061	0.086	-0.091	0.231	0.004	0.003	441.0	820.0	1.01
beta_phone_service	-0.707	0.326	-1.324	-0.214	0.149	0.113	5.0	18.0	2.13
beta_senior_citizen	0.357	0.094	0.185	0.533	0.004	0.003	691.0	1215.0	1.00
beta_streaming_movies	0.206	0.186	-0.141	0.509	0.079	0.059	6.0	21.0	1.77
beta_streaming_tv	0.192	0.178	-0.124	0.512	0.074	0.056	6.0	20.0	1.74
beta_tech_support	-0.602	0.124	-0.860	-0.396	0.038	0.028	11.0	92.0	1.31
beta_tenure	-3.068	0.260	-3.544	-2.566	0.027	0.019	92.0	150.0	1.04
beta_total_charges	0.969	0.285	0.475	1.551	0.031	0.022	84.0	157.0	1.04
intercept	-1.499	1.073	-2.901	0.681	0.521	0.397	4.0	12.0	3.38

Priors Comparison: New vs. Earlier Priors

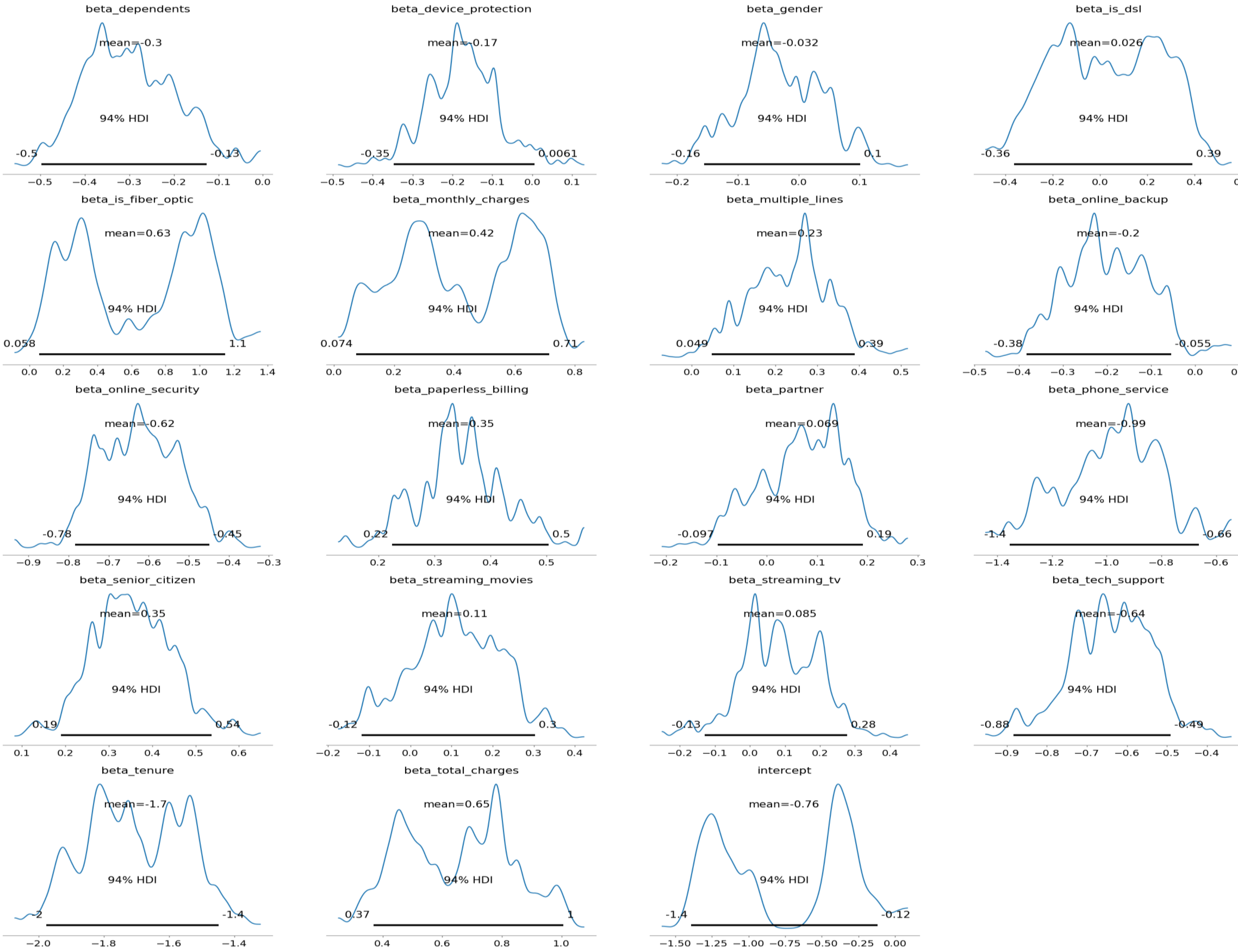
1. **Convergence (R-hat)**
   - **New Priors**: Most parameters have R-hat close to 1.00, but the intercept (3.38) indicates potential convergence issues.
   - **Earlier Priors**: All R-hat values are close to 1.00, indicating stable convergence.
2. **Effective Sample Size (ESS)**
   - **New Priors**: High ESS for most parameters, though some like beta_is_dsl and beta_is_fiber_optic have lower ESS, indicating less effective sampling.
   - **Earlier Priors**: Lower ESS values across parameters, especially for beta_is_fiber_optic and beta_is_dsl.
3. **Parameter Means**
   - **New Priors**: Means are reasonable; e.g., beta_monthly_charges (0.282), beta_is_fiber_optic (1.205).
   - **Earlier Priors**: Means are similar but slightly higher for some parameters like beta_monthly_charges (0.421).
4. **HDI (Highest Density Interval)**
   - **New Priors**: Narrow HDIs, indicating good precision (e.g., beta_senior_citizen 0.185–0.533).
   - **Earlier Priors**: Similar narrow HDIs.
5. **MCSE (Monte Carlo Standard Error)**
   - **New Priors**: Most MCSE values are small, but higher for beta_is_fiber_optic and beta_phone_service, indicating some variability.
   - **Earlier Priors**: Similar MCSE values, with some larger values.
6. **Implications**
   - **New Priors**: Provide more stable estimates and better sampling, though some parameters need further optimization.
   - **Earlier Priors**: Stable convergence but lower ESS, suggesting less effective exploration of the parameter space.

- Conclusion: Choosing normal priors as the difference is insignificant.

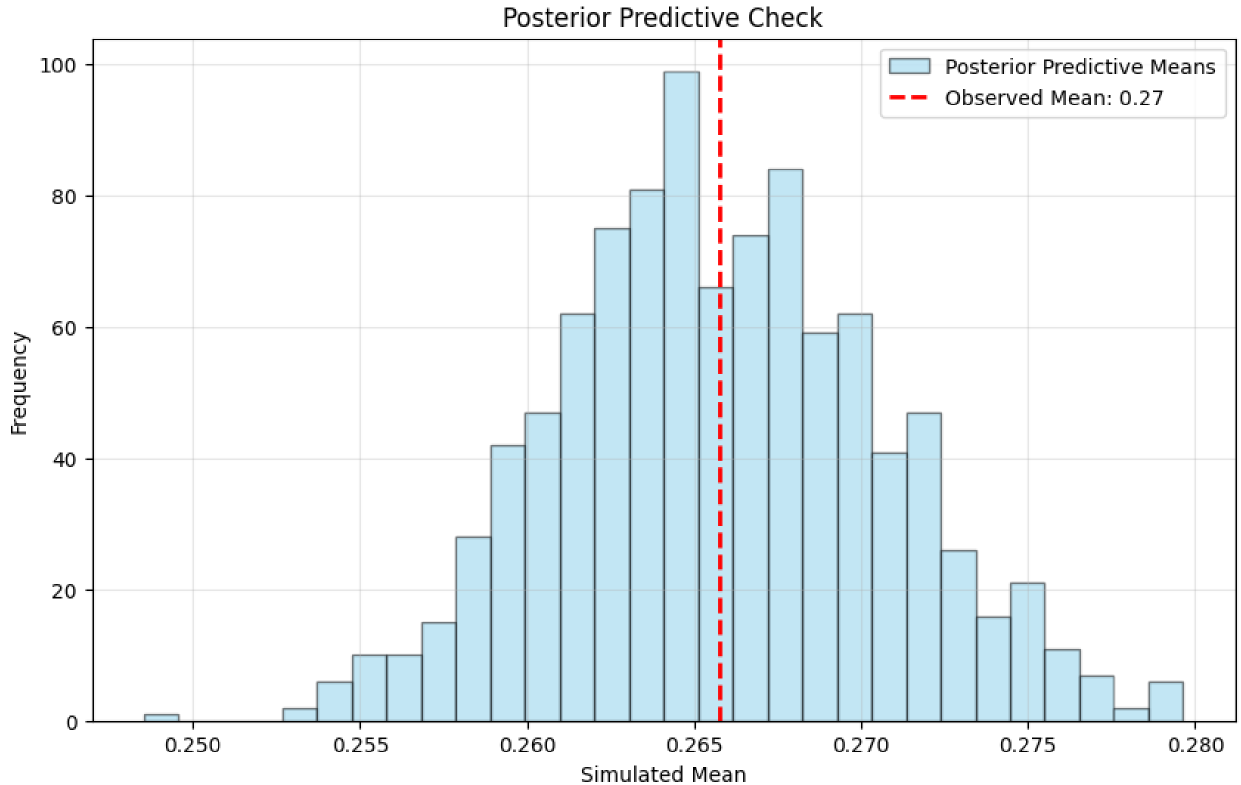
Further Posterior Analysis:



- **Convergence Issues**:
- The trace plots for parameters like the intercept and beta_total_charges show irregular patterns, indicating potential convergence issues. These parameters might require more iterations or refined priors to ensure better mixing and more reliable estimates.
- **Well-Defined Parameters**:
- Parameters such as beta_is_fiber_optic and beta_monthly_charges exhibit clear, unimodal posterior distributions with good mixing in the trace plots. These parameters show strong posterior certainty, indicating they are reliable predictors in the model.
- **Multimodal Distributions**:
- The density plots for parameters like beta_dependents and beta_device_protection suggest multimodal distributions, which could be caused by insufficient sampling or model misspecification. These parameters should be closely examined to ensure proper model specification and sampling.
- **Autocorrelation**:
- Some trace plots (e.g., for beta_total_charges) show high autocorrelation, which implies that the sampler may not be efficiently exploring the parameter space. This could be mitigated by increasing the number of iterations or adjusting the priors.
- **Action Items**:
- **Increase Iterations**: For parameters like intercept and beta_total_charges, run the sampler for more iterations to improve convergence and ensure stable mixing.
- **Refine Priors**: Adjust priors for parameters with high uncertainty or poor convergence (e.g., intercept), potentially using more robust priors like Cauchy or adjusting scale parameters.
- **Posterior Predictive Checks**: Validate the model fit by comparing posterior predictions with observed data to ensure the model generalizes well.
- **Use Diagnostics**: Employ R-hat and Effective Sample Size (ESS) to quantitatively assess convergence and identify potential sampling inefficiencies.


beta_dependents
mean=-0.3
94% HDI
-0.5
-0.13
beta_device_protection
mean=-0.17
94% HDI
-0.35
0.0061
beta_gender
mean=-0.032
94% HDI
-0.16
0.1
beta_is_dsl
mean=0.026
94% HDI
-0.36
0.39
beta_is_fiber_optic
mean=0.63
94% HDI
0.058
1.1
beta_monthly_charges
mean=0.42
94% HDI
0.074
0.71
beta_multiple_lines
mean=0.23
94% HDI
0.049
0.39
beta_online_backup
mean=-0.2
94% HDI
-0.38
-0.055
beta_online_security
mean=-0.62
94% HDI
-0.78
-0.45
beta_paperless_billing
mean=0.35
94% HDI
0.22
0.5
beta_partner
mean=0.069
94% HDI
-0.097
0.19
beta_phone_service
mean=-0.99
94% HDI
-1.4
-0.66
beta_senior_citizen
mean=0.35
94% HDI
0.19
0.54
beta_streaming_movies
mean=0.11
94% HDI
-0.12
0.3
beta_streaming_tv
mean=0.085
94% HDI
-0.13
0.28
beta_tech_support
mean=-0.64
94% HDI
-0.88
-0.49
beta_tenure
mean=-1.7
94% HDI
-2
-1.4
beta_total_charges
mean=0.65
94% HDI
0.37
1
intercept
mean=-0.76
94% HDI
-1.4
-0.12

1. **Strongly Influential Parameters**:
    1. beta_tenure (mean ≈ -1.7): Strong negative effect with high certainty (narrow HDI).
    2. beta_total_charges (mean ≈ 0.65): Strong positive effect with clear posterior definition.
2. **High Certainty Parameters**:
    1. beta_gender, beta_multiple_lines, and beta_online_backup have narrow HDI ranges, reflecting strong certainty.
3. **Multimodality in Posteriors**:
    1. Parameters like beta_is_fiber_optic and beta_paperless_billing exhibit multimodal distributions, indicating potential interactions or insufficient data.
4. **Wide Uncertainty**:
    1. beta_tech_support and beta_senior_citizen show broader HDI ranges, reflecting greater uncertainty.
5. **Intercept**:
    1. Mean ≈ -0.76 with moderate uncertainty, indicating its effect is not as well-defined.
6. **Recommendations**:
    1. Investigate multimodal parameters further to address potential issues.
    2. Leverage high-certainty coefficients for actionable insights.



- The observed mean is consistent with the predictions from the model's posterior distribution.
- This suggests the model is capable of capturing the central tendency of the data.

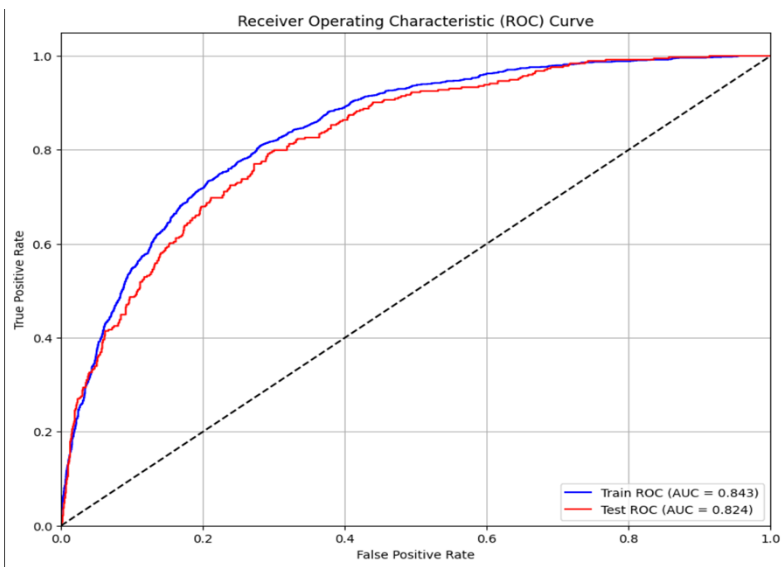
Odds Ratio

1. **Strong Predictors**:
    1. **beta_is_fiber_optic** and **beta_total_charges** are the strongest predictors with high mean values and large effect sizes.
    2. Policies or actions focusing on these variables could have the most substantial impact on the outcome.
2. **Moderate Predictors**:
    1. Variables like **beta_monthly_charges**, **beta_senior_citizen**, and **beta_multiple_lines** show moderate influence on the outcome.
    2. These should also be considered in decision-making as they provide significant contributions.
3. **Weak Predictors**:
    1. Predictors such as **beta_tenure** and **beta_phone_service** have relatively minor effects and may not be critical to focus on for optimization.
4. **Uncertainty in Estimates**:
    1. Variables with wider 95% HDI intervals, such as **beta_is_fiber_optic**, indicate higher uncertainty in their effect sizes.
    2. This warrants further analysis or additional data collection to improve the reliability of these estimates.
5. **Positive Effects**:
    1. All variables exhibit positive effects, as their 95% HDI intervals do not cross zero.
    2. This suggests that all predictors positively contribute to the outcome to varying degrees.

	mean	median	2.5%	97.5%
beta_dependents	0.746321	0.735960	0.618600	0.937789
beta_device_protection	0.843440	0.837170	0.706459	1.024223
beta_gender	0.971120	0.964913	0.846217	1.108469
beta_is_dsl	1.053784	1.027664	0.681319	1.486079
beta_is_fiber_optic	2.020480	1.985811	1.073966	3.484501
beta_monthly_charges	1.557498	1.503962	1.076758	2.097223
beta_multiple_lines	1.268699	1.273198	1.053791	1.531558
beta_online_backup	0.824668	0.815193	0.687010	1.012538
beta_online_security	0.538980	0.534543	0.454742	0.641938
beta_paperless_billing	1.419787	1.409279	1.221447	1.647958
beta_partner	1.075621	1.080851	0.909142	1.238848
beta_phone_service	0.379629	0.379375	0.256068	0.516975
beta_senior_citizen	1.431373	1.419643	1.187269	1.737385
beta_streaming_movies	1.120043	1.111795	0.895691	1.388494
beta_streaming_tv	1.095130	1.083552	0.846072	1.331336
beta_tech_support	0.529023	0.526585	0.416743	0.641940
beta_tenure	0.184116	0.179577	0.138545	0.239933
beta_total_charges	1.945797	1.962414	1.403960	2.698041

Evaluation: Bayesian Model

- Evaluation is done on an 80:20 train-test split.

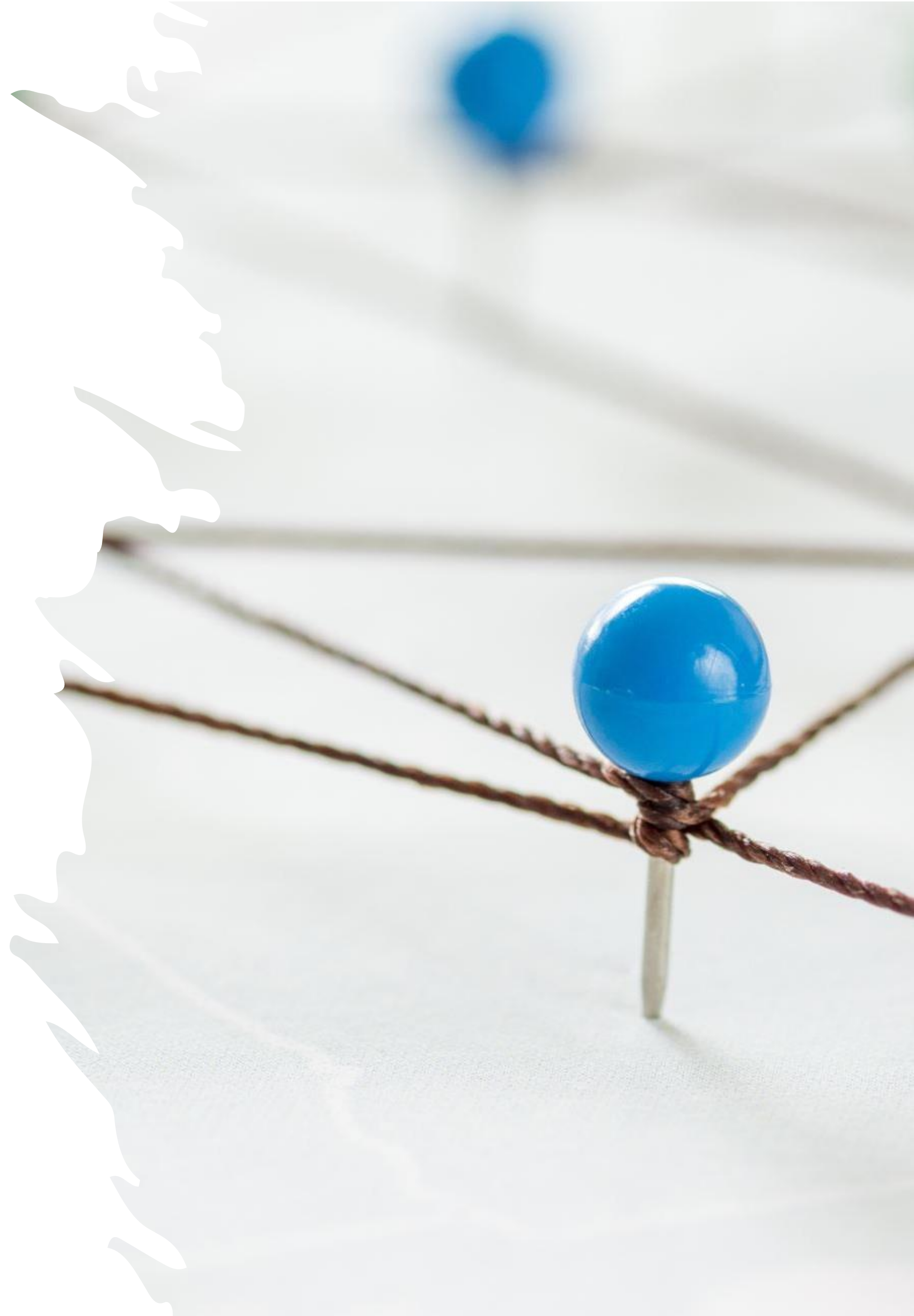


Model Evaluation Metrics:

	Accuracy	Precision	Recall	F1 Score	ROC AUC
Train	0.806044	0.667774	0.537793	0.595776	0.842860
Test	0.786780	0.626712	0.489305	0.549550	0.824007

Frequentist Approach

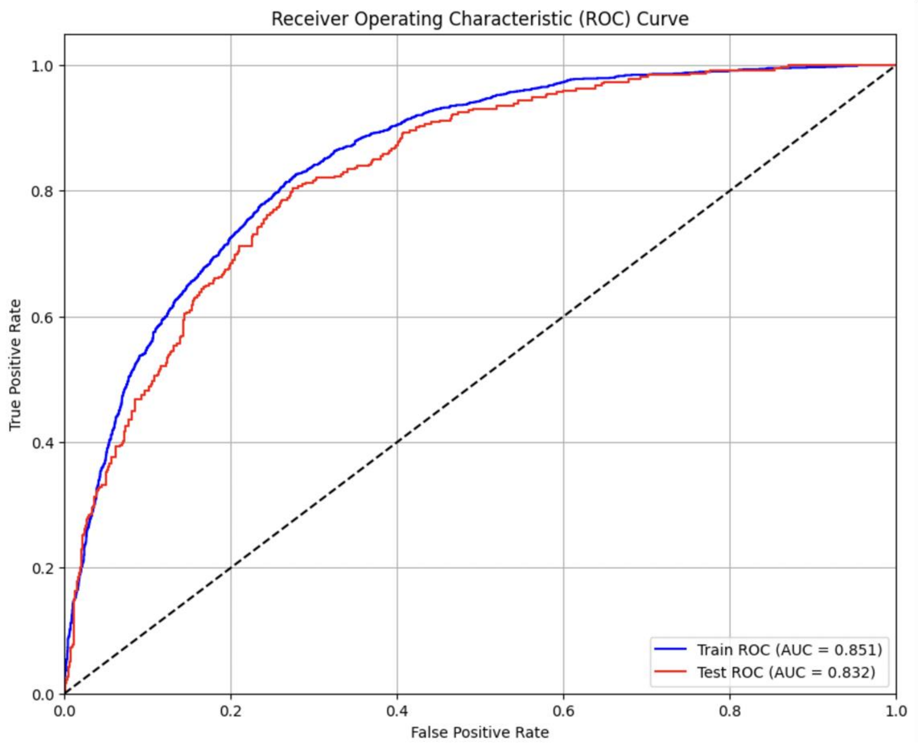
- Train a logistic regression model for churn prediction.
- Weight initialization- Fixed single point estimates.



Evaluation Frequentist Model

Model Evaluation Metrics:

	Accuracy	Precision	Recall	F1 Score	ROC AUC
Train	0.806222	0.659574	0.559866	0.605644	0.851260
Test	0.786780	0.619355	0.513369	0.561404	0.831828



Conclusions and Next Steps

- Bayesian approach provides insights into uncertainty
- Model identifies key factors influencing churn
- Improvements needed:Refine priors for problematic parameters
- Increase sampling iterations
- Consider alternative model structures