Analyzing Hospital Readmission Using SDOH Factors

David Deng, Josiah Domercant, Xiaotian Dai

Department of Mathematics
Illinois State University

May 12, 2025

Introduction

- In this project, our objective is to find out whether SDOH variables have influence on readmission condition within 30 days.
- The following modeling strategies are applied hospital data we received:
  - **XGBoost**
  - **Bayesian Regression** with **Spike-and-Slab Prior**
- Our goal is to identify the most relevant predictors and the significancy of SDOH variables based on metrics such as:
  - **ROC Curve** for overall classification effect for each model
  - **Variable Importance** for **XGBoost**
  - **95% Credible Interval** of the coefficient for Bayesian Regression

Bayesian Regression Overview

- Bayesian regression combines **prior beliefs** with observed data to generate **posterior distributions** over model parameters.
- Unlike traditional regression, it provides **uncertainty quantification** for each coefficient.
- Be capable of:
  - Variable selection by checking credible interval
  - Good interpretability
- Offers natural regularization and incorporates prior domain knowledge.

Spike-and-Slab Prior

- The spike-and-slab prior is a commonly used Bayesian prior for variable selection.
- Each coefficient β_j follows a mixture prior:

$$\beta_j \sim (1 - \pi) \cdot \delta_0 + \pi \cdot \mathcal{N}(0, \tau^2)$$

- Interpretation:
  - **Spike**: A point mass at 0 which indicates the variable is likely irrelevant.
  - **Slab**: A wide normal distribution which allows for significant effect sizes.
- Advantages:
  - Automatically selects important variables
  - Reduces model complexity and improves generalizability

Motivation for using Bayesian Regression

- Our project aims to **identify meaningful variables**, not just achieve high predictive accuracy.
- The spike-and-slab prior enables:
  - Reliable variable selection via posterior distributions
  - Shrinkage of unimportant variables toward zero
- Complements machine learning models (e.g., XGBoost):
  - XGBoost focuses on prediction performance
  - Bayesian regression emphasizes interpretability and variable selection

Hospital Readmission Data Overview

- **Number of observations:** 74,994
- **Number of variables:** 101
- **Variable Type:** Numerical(1), Categorical(100)
- **Missing values:**
  - `VAR_SDOHAlcoholUseDomianRisk_CAT`: 87.5% missing
  - `VAR_PhysicalActivityDomian_CAT`: 90.0% missing
  - `VAR_SDOHScoialConnectionDomian_CAT`: 91.8% missing
  - `VAR_SDOHStressDomian_CAT`: 90.4% missing

SDOH Variable Explanation

- `VAR_SDOHAlcoholUseDomianRisk_CAT`

  SDOH risk assessment for the patient, within Alcohol Use domain
- `VAR_FinancialResourceStrainDomainCollected_FLG`

  Availability of SDOH risk assessment for Financial Strain domain
- `VAR_SDOHFoodInsecurityDomainCollected_FLG`

  Availability of SDOH risk assessment for Food Insecurity domain
- `VAR_SDOHHousingStabilityDomainCollected_FLG`

  Availability of SDOH risk assessment for Housing Stability domain
- `VAR_SDOHPhysicalActivityDomain_CAT`

  SDOH risk assessment for the patient, within Physical Activity domain

SDOH Variable Explanation

- `VAR_SDOHSafetyandDomesticViolenceDomainCollected_FLG`

  Availability of SDOH risk assessment for Safety and Domestic Violence domain
- `VAR_SDOHSocialConnectionDomain_CAT`

  SDOH risk assessment for the patient, within Social Connection domain
- `VAR_SDOHStressDomain_CAT`

  SDOH risk assessment for the patient, within Stress domain
- `VAR_SDOHTransportationDomainCollected_FLG`

  Availability of SDOH risk assessment for Transportation domain
- `VAR_SDOHUtilitiesDomain_FLG`

  Availability of SDOH risk assessment for Utilities domain

Independency Check

In our dataset, there is only one numerical variable, we choose to use Chi-Square Test to see the independency.

	Alcohol	**Financial**	**Food**	**Housing**	**Physical**
p-value	0.01	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$	10^{-10}

	Safety	**Social**	**Stress**	**Transportation**	**Utilities**
p-value	$< 10^{-16}$	0.05	0.96	$< 10^{-16}$	$< 10^{-16}$

95%: Not Independent if p-value < 0.05

Data Visulization

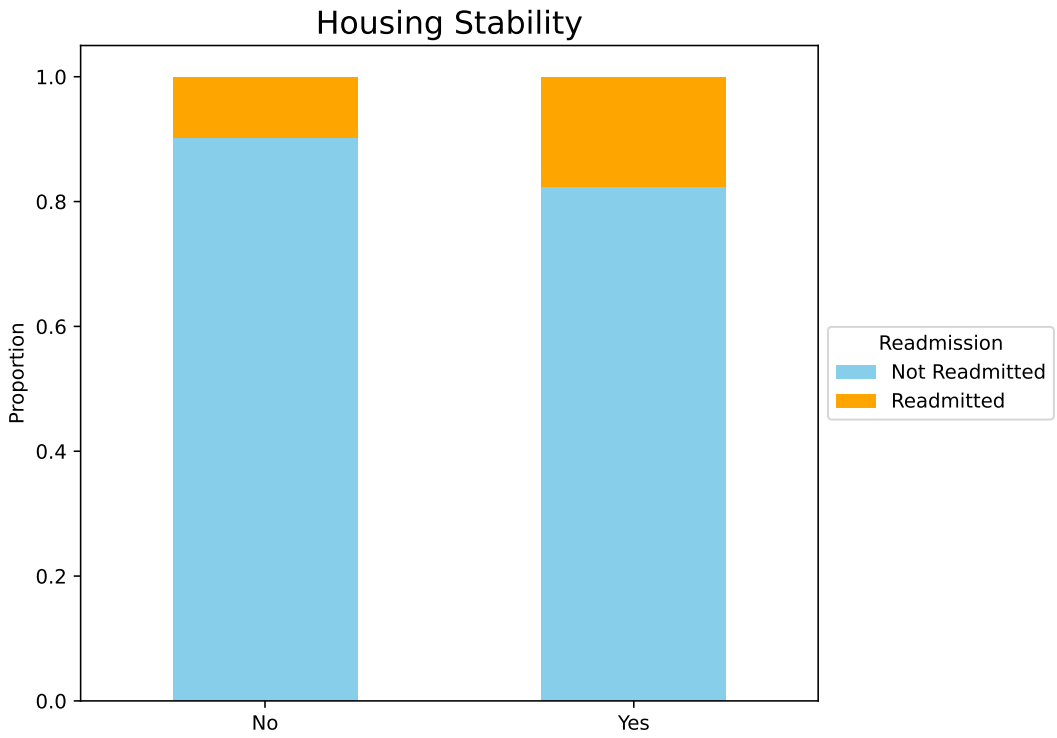


Figure: Readmission Condition of different category for Housing Stability.

Data Visulization

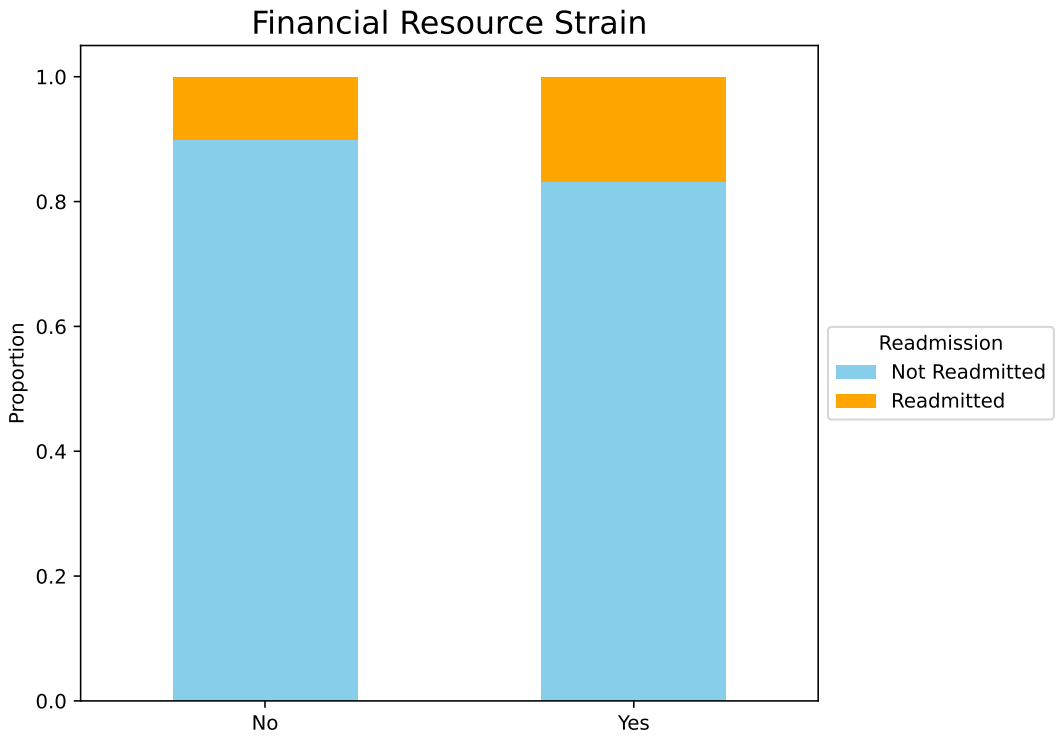


Figure: Readmission Condition of different category for Financial Resource Strain.

Data Visulization

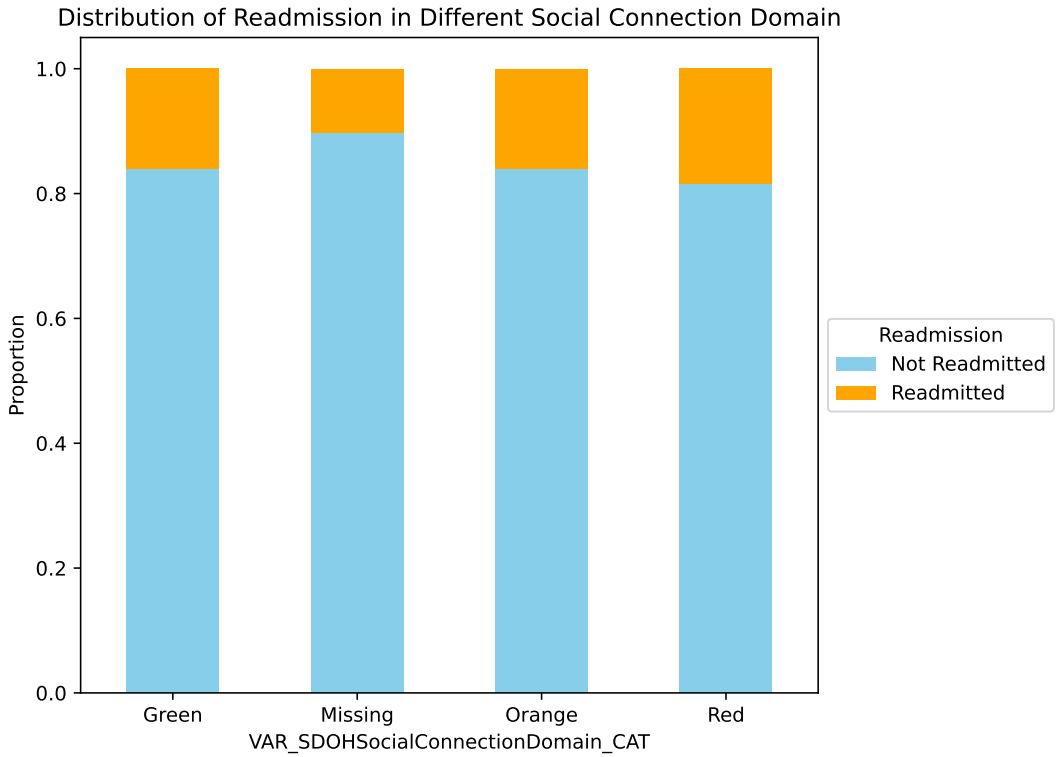


Figure: Readmission Condition of different category for Food Insecurity.

Data Visulization

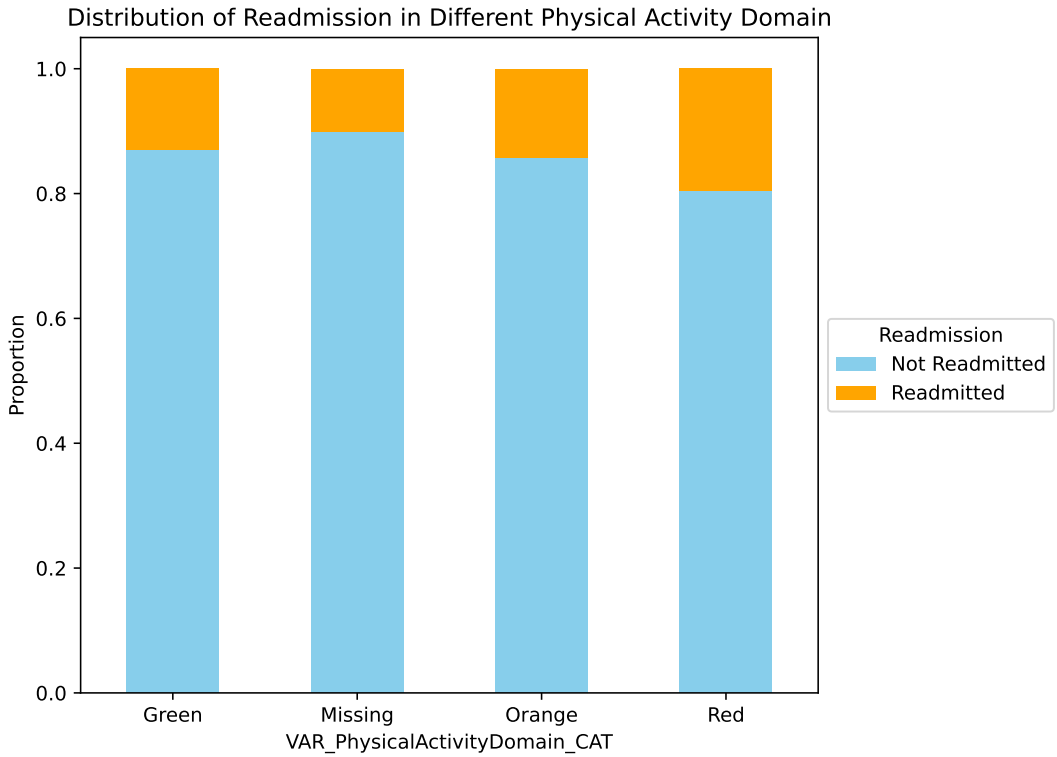


Figure: Readmission Condition of different category for Transportation.

Data Preprocessing

For XGBoost:

- **Handling Missing Values:** For categorical variables, `NA` was treated as a separate level.
- **Training/Testing Splitting:** 50% Training, 50% Testing

For Bayesian Regression:

- **Handling Missing Values:** For categorical variables, `NA` was treated as a separate level.
- **Training/Testing Splitting:** 50% Training, 50% Testing
- **Encoding Categorical Variables:** Create new variable represent different level for each categorical predictor.

XGBoost: Hyperparameter Tuning

- **Key Parameters Tuned:**
  - `n_estimators`: number of trees
  - `learning_rate`: step size shrinkage
  - `max_depth`: maximum depth of a tree
  - `subsample`: fraction of samples used per tree
  - `scale_pos_weight`: control the weight of positive sample
- **Tuning Strategy:** Grid search with 5-fold cross-validation on training set

Model Evaluation

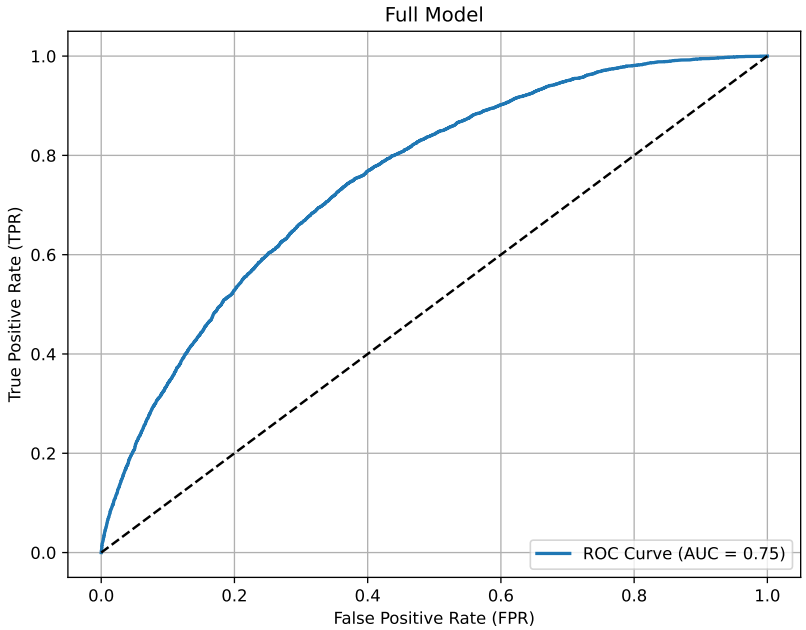


Figure: ROC Curve for XGBoost, AUC = 0.75

Feature Importance

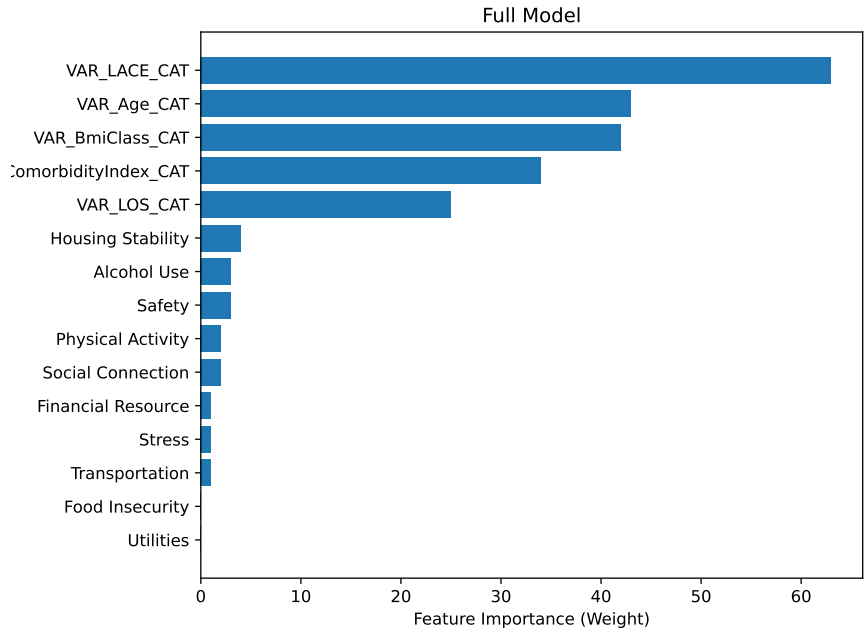


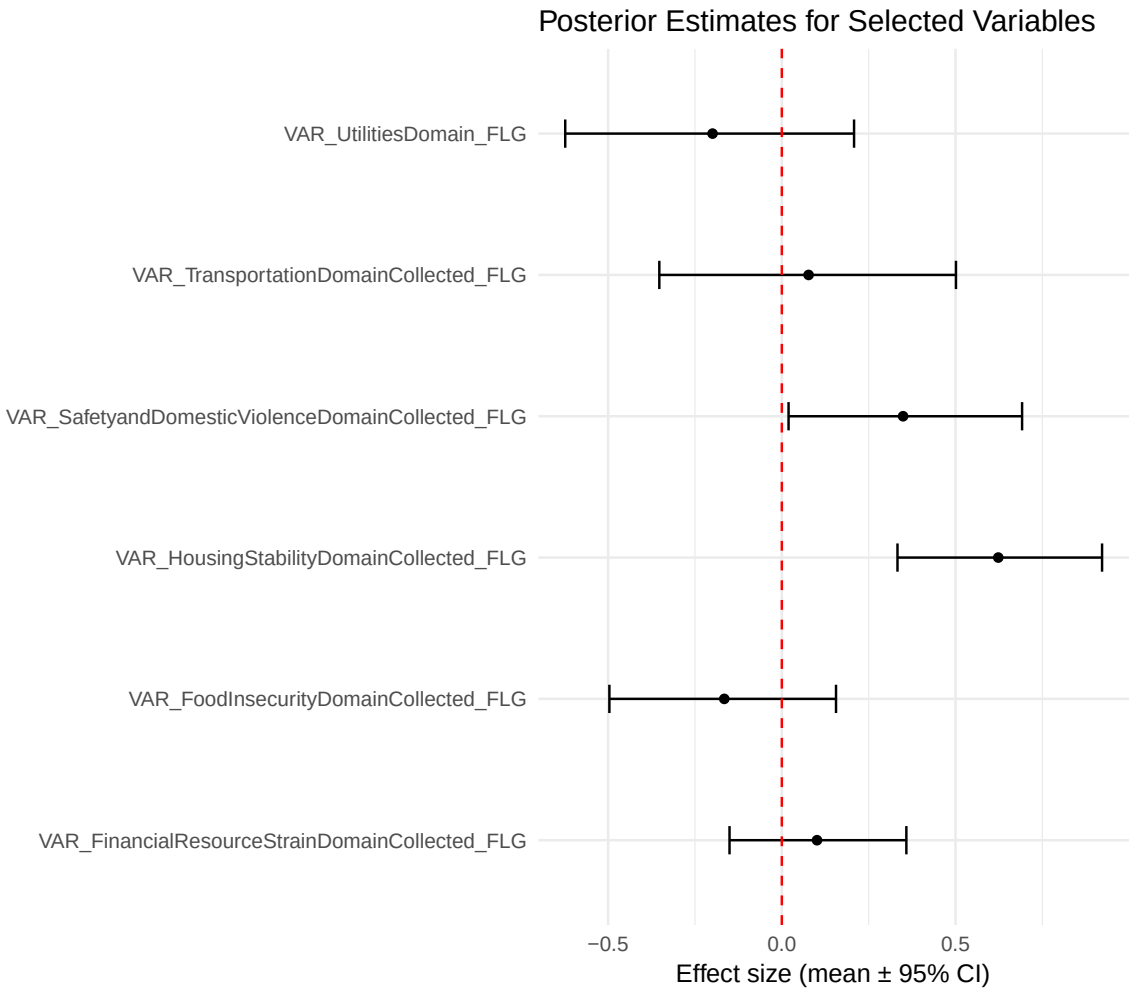
Figure: Readmission Condition of different category for Financial Resource Strain.

- The model shows that SDOH variables do have some importance, but there are several dominant variables.

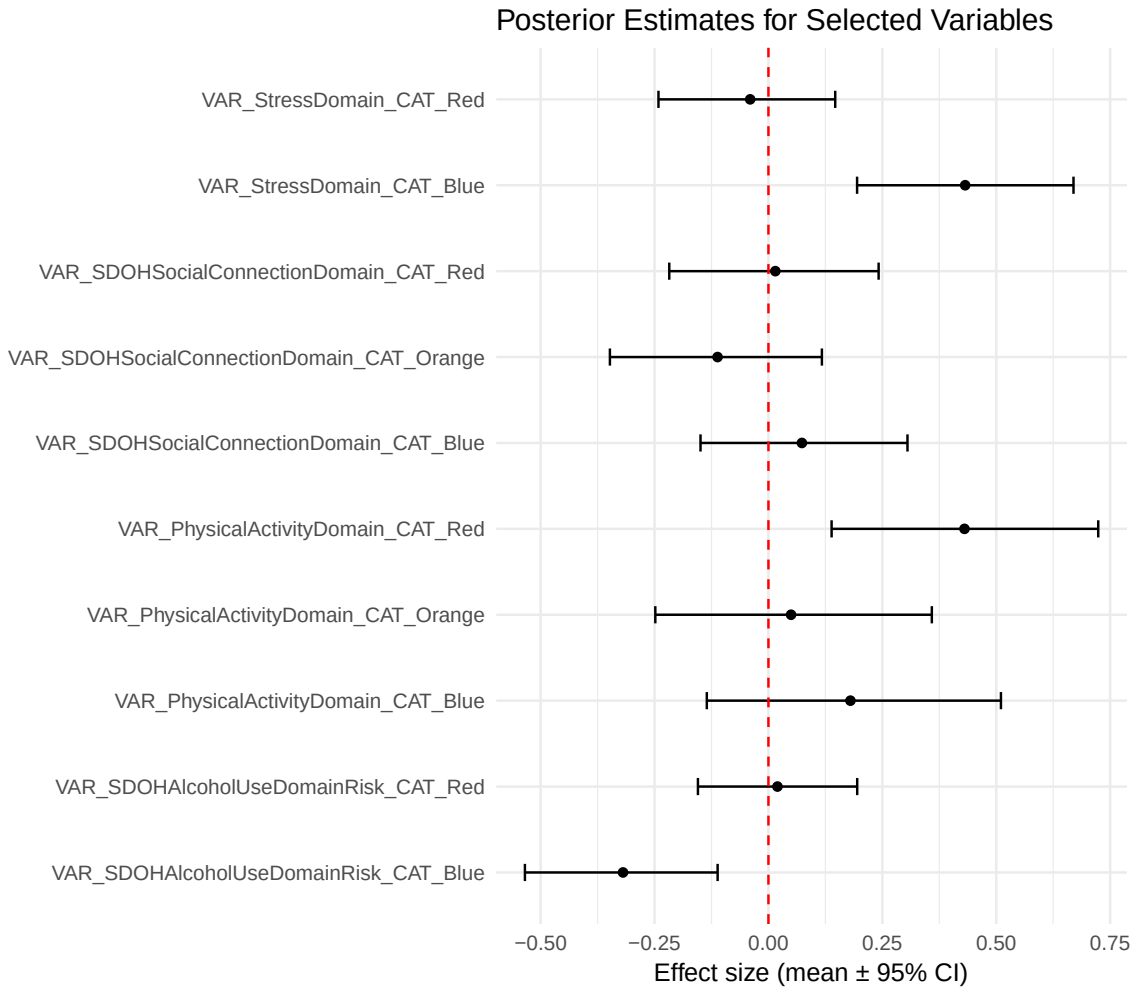
XGBoost Summary

- The model achieves quite strong predictive performance.
- However, XGBoost does not provide a result that can indicate positive or negative relationship between the predictor and the response.
- Next, we apply **Bayesian Regression** to obtain uncertainty-aware variable selection.

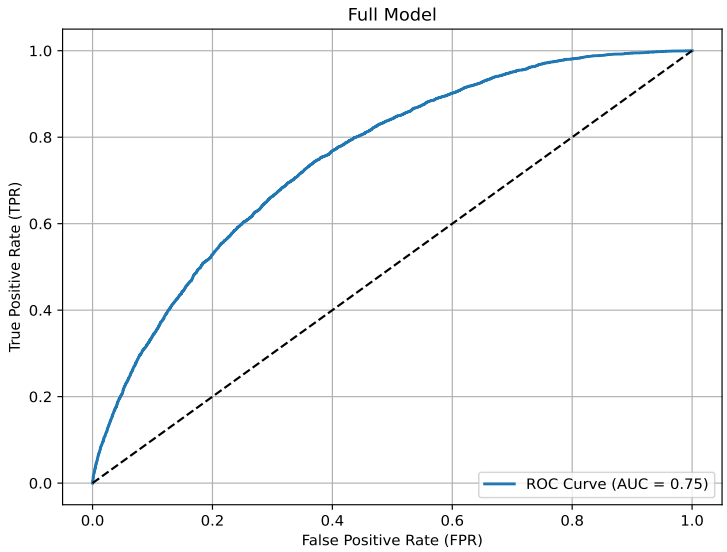
Posterior Inference and Variable Importance



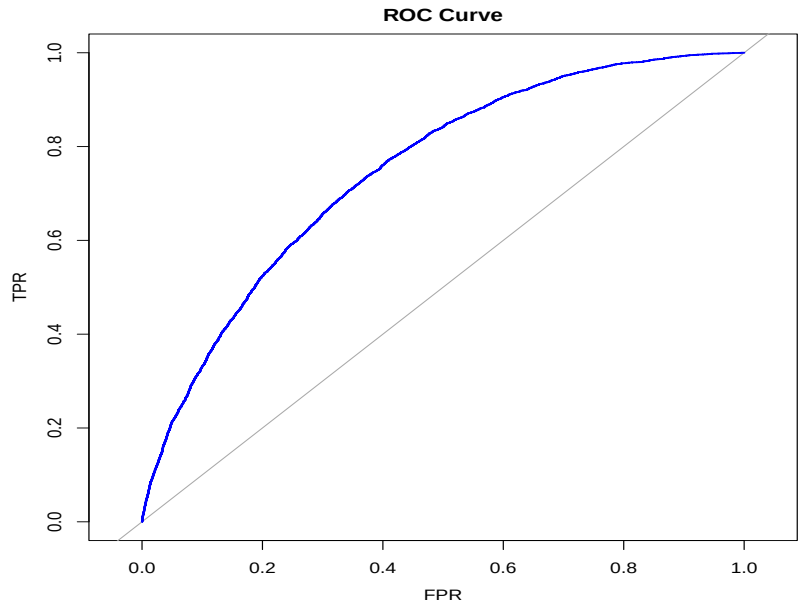
Posterior Inference and Variable Importance



Bayesian Model Evaluation



(a) XGBoost $AUC \approx 0.75$



(b) Bayesian Regression $AUC = 0.7473$

Figure: ROC Curve for XGBoost and Bayesian Regression.

Conclusion

- Both models achieved moderate performance (AUC = 0.75), indicating a moderate predictive power overall.
- SDOH variables do have some importance in the result of XGBoost, but the dominant variable may make their importance be underestimated.
- Safety and Housing Stability have positive relationship with readmission condition.
- If the result of the Physical Activity assessment is red, it will make the readmission condition increase.
- The patient does not have assessment on Alcohol Use and Stress have some potential reason to make their readmission condition increase.

Thanks!