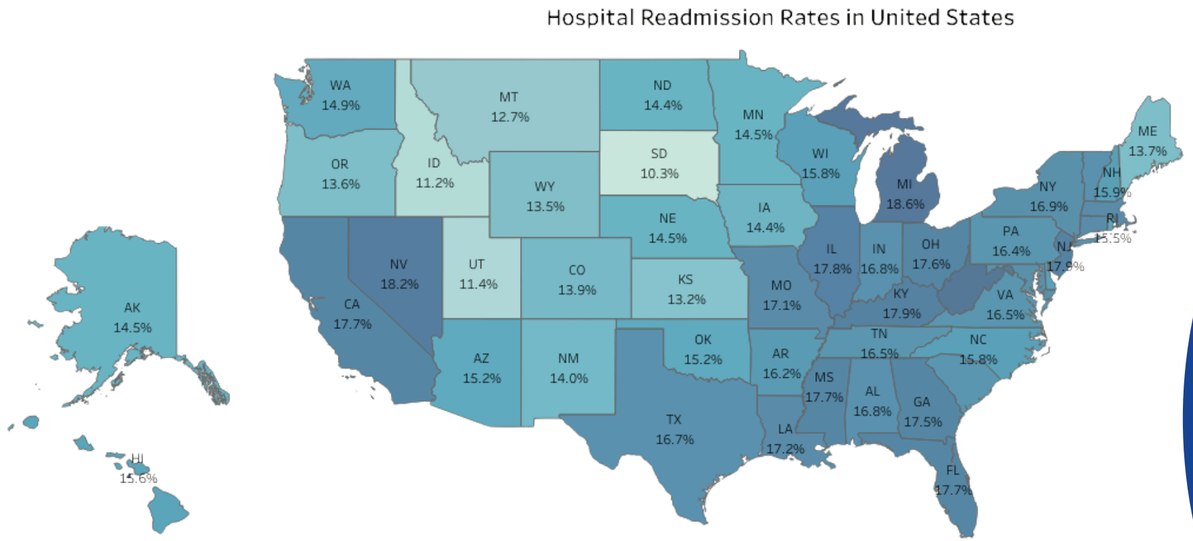
Texas McCombs MSBA
Advanced Machine Learning

Predicting Hospital Readmission for Diabetic Patients

Group 19:
Joseph Bailey • Kristen Lowe • Zan Merrill • Varsha Ramesh • Nisha Sapkota

Fall 2025

Background



Map Source: Pacifica Data (2020-2023)

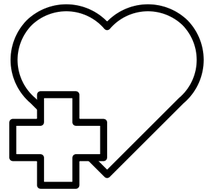
An estimated 3.8 million readmissions occurred in the US in 2018 with an average cost of $15,200 per readmission. That's $57 billion!

Bilicki et al., CDC, 2024

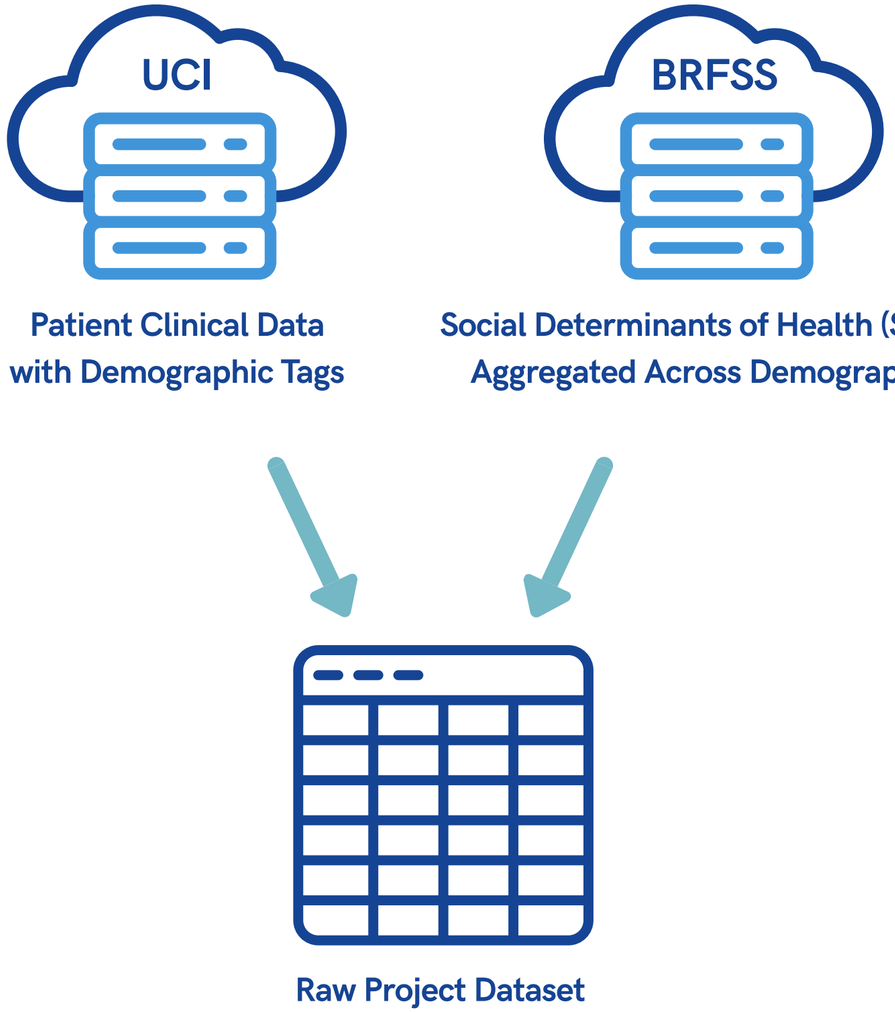
- 30-day readmissions signal gaps in discharge planning, medication management, or follow-up care.
- High readmission rates increase healthcare costs and strain hospital resources.
- Diabetes is one of the leading chronic conditions contributing to avoidable hospitalizations in the U.S.
- Treatment complexity for diabetic patients (multiple medications, comorbidities, varying management adherence) makes predicting risk challenging.
- Current approaches rely heavily on clinical judgment rather than data-driven risk stratification.

Project Goal

- Our goal is to build an interpretable model that helps identify at-risk patients and supports targeted intervention and discharge planning
- 30 day readmissions create major healthcare costs and are financially penalized under the Patient Protection and Affordable Care Act (PPACA) through the Hospital Readmissions Reduction Program (HRRP)
- High-risk patients can cost 5–10× more annually than average patients. (Siekman et al., Cleveland Clinic, 2018)



Stakeholder	Value Provided
Hospitals	Reduce penalties from PPACA, improve quality metrics
Insurance Providers	Reduce total cost of care
Care Management Teams	Identify high-risk patients for intervention programs
Patients	Improve care outcomes



Data Sources

- The University of California Irvine (UCI) Diabetes Dataset supplied clinical and demographic patient features (1999-2008)
- The Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System (BRFSS) provided behavioral and social-determinant survey data (ongoing since 1984)
  - Physical activity, smoking, alcohol use, insurance, employment, etc.
- BRFSS responses were grouped by race, age band, and gender
- Each group's average values were calculated for key health and behavior variables
- These demographic-level BRFSS features were merged onto matching UCI patient records

Combining UCI and BRFSS let us supplement outdated clinical data with population-level social and behavioral factors that meaningfully influence readmission outcomes

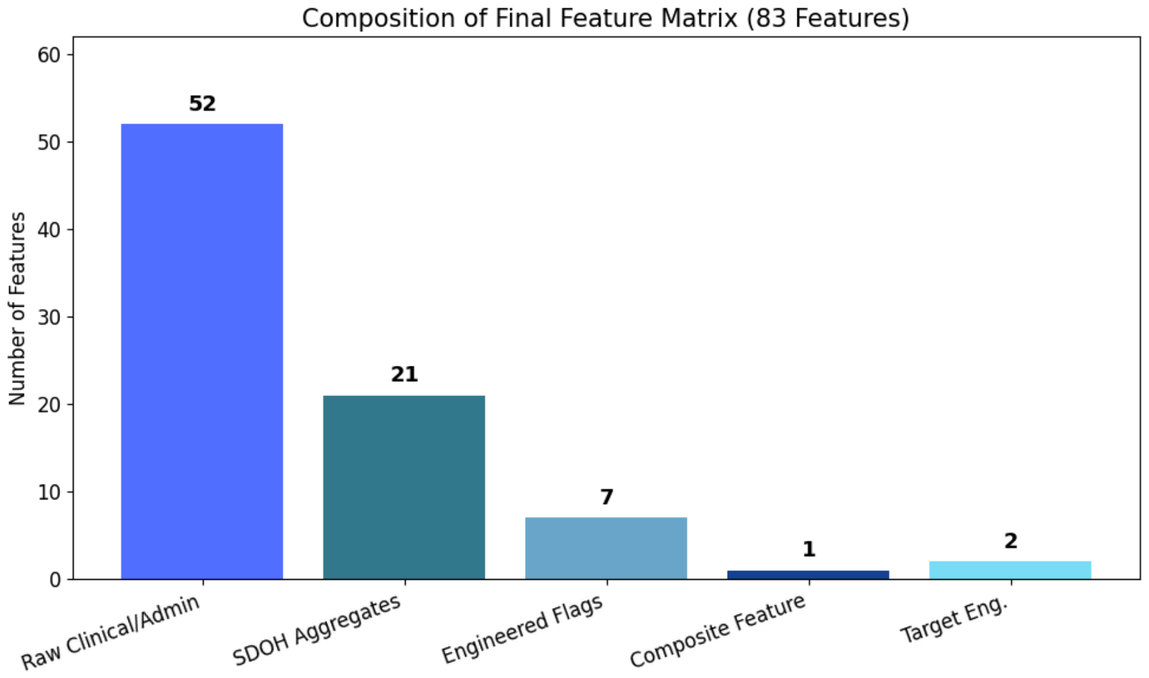
Data Cleaning & Feature Engineering

- Standardized missing values (“?” → NaN), fixed inconsistent text entries, and removed or filled incomplete fields to stabilize the dataset

- Converted structured fields (age buckets, gender, race, yes/no flags, diag codes) into consistent numeric or categorical representations

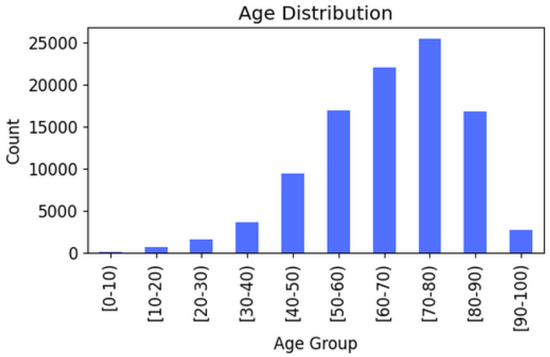
- Created clinically meaningful engineered features such as A1c/Glucose test flags and the discharge-instability score
  - Sum of urgent admit, high procedures, and short stay flags

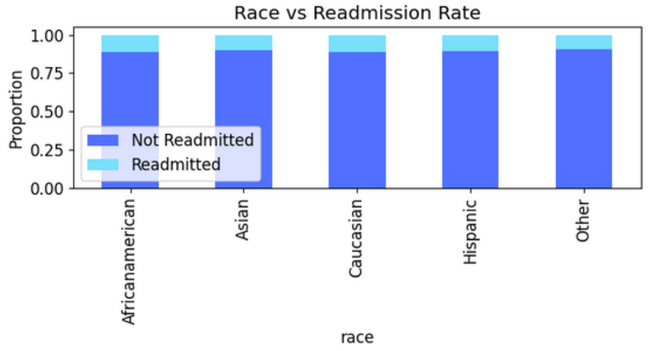
- Filled numeric fields using median imputation and cleaned categorical fields using explicit “Unknown” categories to preserve missing-value information.

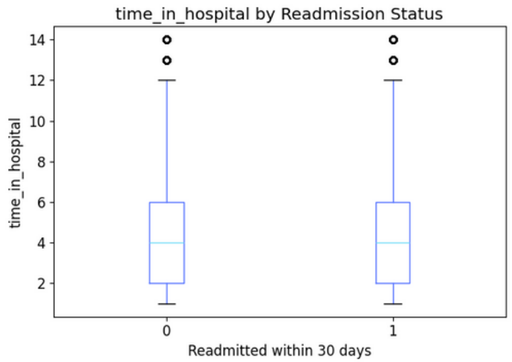


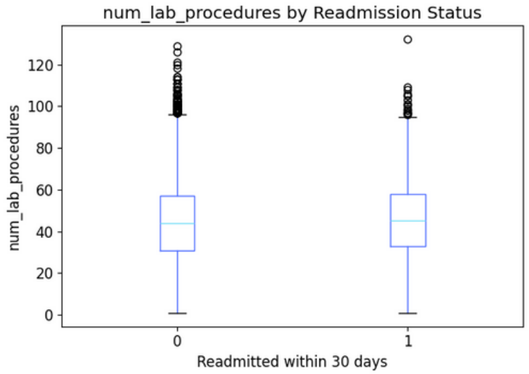
End Result: A unified dataset containing 83 cleaned clinical, behavioral, demographic, and engineered features, ready for downstream modeling.

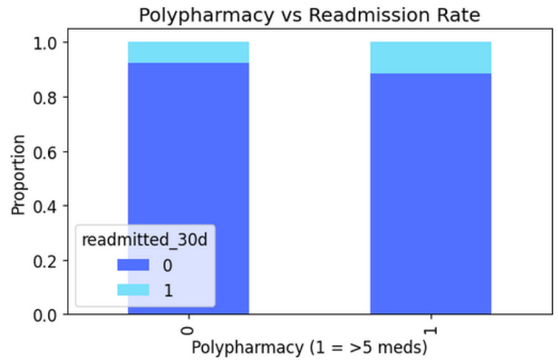
Note: Two target-engineered columns were created during cleaning (readmitted_label and readmitted_30d). Only readmitted_30d will be used as the modeling target; the other is retained only for reference and will be dropped prior to training

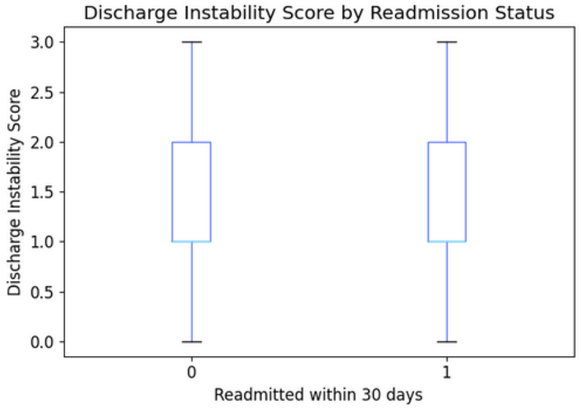












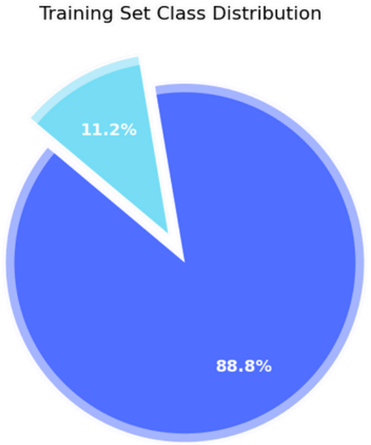
EDA: High-Level Patterns

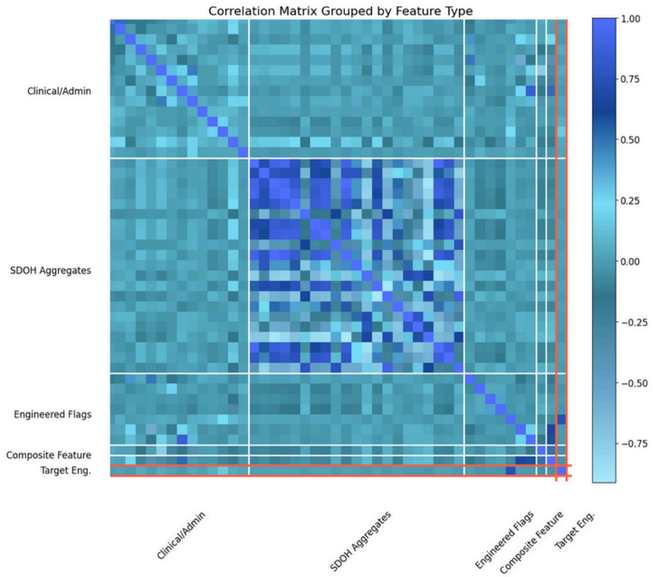
- The 30-day readmission target is strongly imbalanced, with only about 11 percent positive cases
- Correlation analysis shows several internal feature clusters (notably utilization and BRFSS-derived features), but all correlations with the target remain weak
- Core variables such as age group, race, medication count, and time in hospital show only small differences between readmitted and non-readmitted patients
- Several clinical intensity features have long right-tailed distributions, revealing wide patient variability but no clear separation between outcome classes.

These exploratory patterns indicate a diffuse signal and weak univariate structure, underscoring the need for multivariate and nonlinear modeling to uncover meaningful predictive relationships

EDA: Modeling Implications

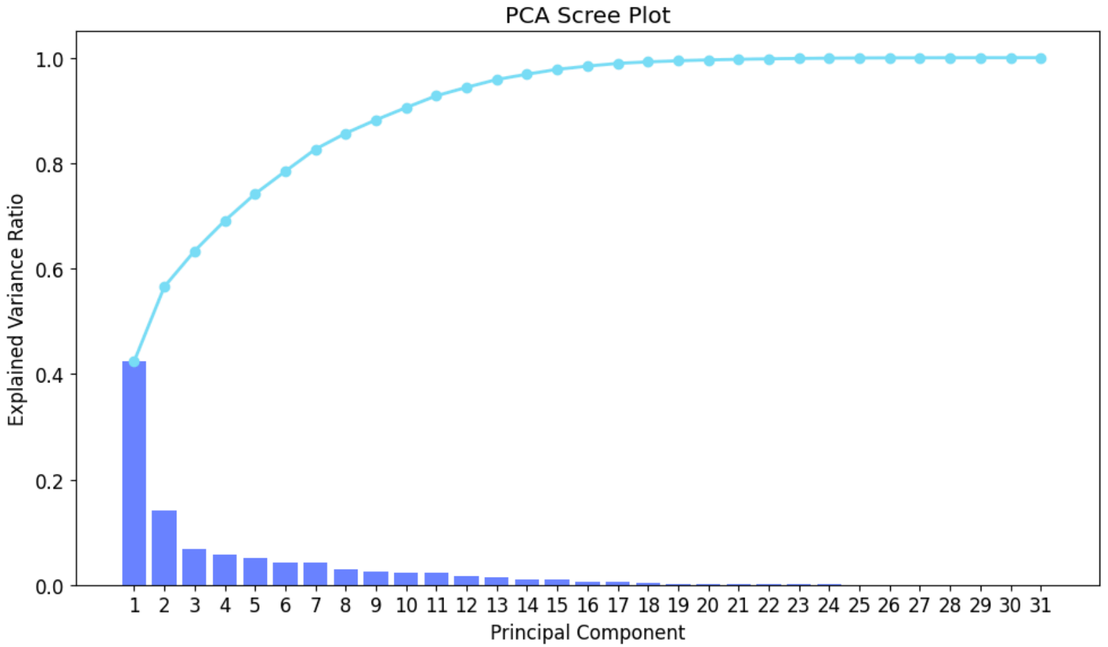
- High-missing and low-variance features suggested the need to remove unstable predictors
- Mixed data types and bucketed age ranges indicated the need for consistent numerical encoding
- Strong class imbalance highlighted the need to modify the prior and loss function
- Weak univariate correlations pointed toward nonlinear, interaction-aware modeling

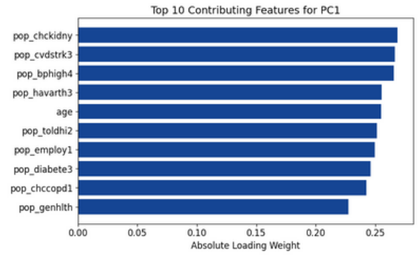


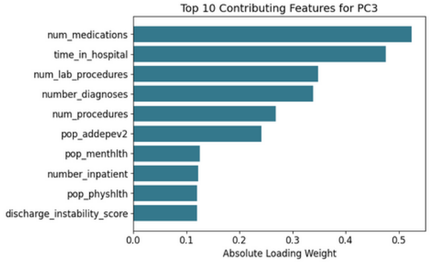


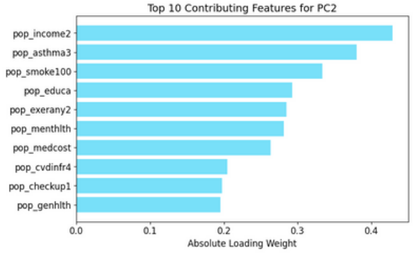
Individual Feature Correlation To Target		
Feature	**Correlation**	**Notes**
readmitted_label	0.739475	dropped later during modeling to avoid leakage
number_inpatient	0.165538	
number_emergency	0.060875	
discharge_disposition_id	0.05063	
number_diagnoses	0.049625	

Even after engineering several features, weak individual correlations highlight the need for models that learn interactions, nonlinearities, and latent structure, such as PCA-enhanced features or tree-based methods









Model Training

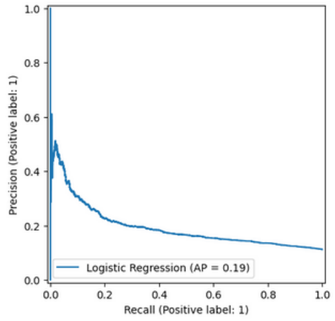
- Removed high-missing and low-variance columns identified during EDA, reducing the feature set from 83 to 44 inputs for modeling
- Converted all features into a unified numeric representation through midpoint mapping, binary recoding, and one-hot encoding
- Scaled continuous features for linear models while preserving raw values for tree-based models
- Applied class weighting and scale_pos_weight to rebalance the loss function during training
- Ran PCA on scaled continuous variables to evaluate redundancy and guide model selection

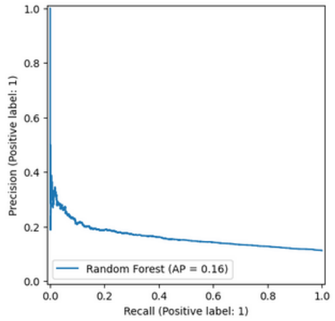
These steps produced a compact, numerically consistent, imbalance-aware feature matrix suitable for training logistic regression, Random Forest, and XGBoost models

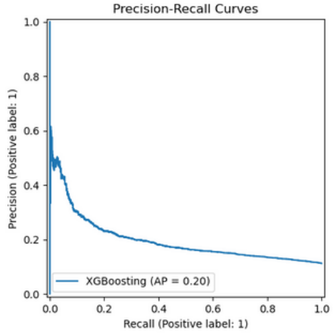
Model Comparison

Model Performance Comparison

Model	Precision	Recall	FPR	AUC	Avg Prec	F1 Score	TP	FP	TN	FN
Log Regression	0.1659	0.4969	0.3158	0.6309	0.19	0.2488	1110	5579	12086	1124
RF	0.1668	0.4866	0.3073	0.6321	0.16	0.2485	1087	5429	12236	1147
XGBoost	0.1633	0.5063	0.3282	0.6333	0.2	0.2469	1131	5797	11868	1103







Model Improvement Over Baseline Prevalence (Average Precision Comparison)

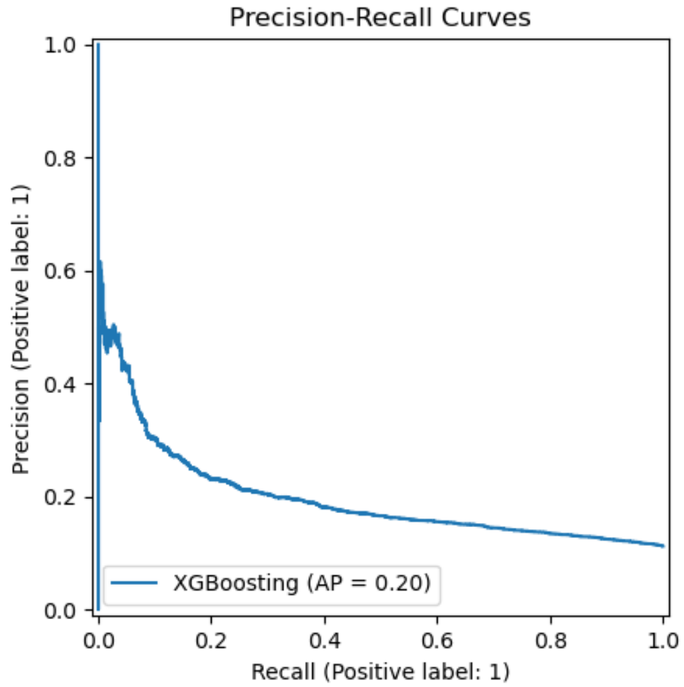
Model	Avg Prec	Baseline AP	Lift (AP/Baseline)	Relative Improvement (%)
Log Regression	0.19	0.11	1.7273	72.7273
RF	0.16	0.11	1.4545	45.4545
XGBoost	0.2	0.11	1.8182	81.8182

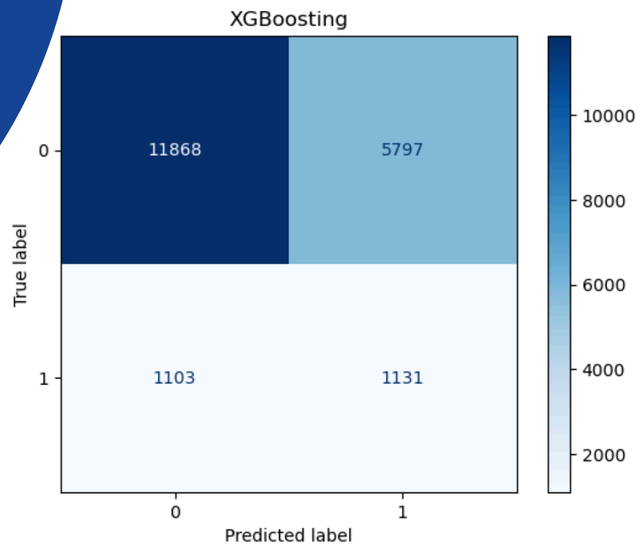
- All three models showed modest predictive strength, consistent with the weak feature relationships identified in EDA
- Logistic regression, Random Forest, and XGBoost achieved similar performance, indicating limited separability in the available features
- Average precision for all models exceeded the baseline prevalence, confirming that each model captured meaningful structure despite the low signal environment
- XGBoost delivered the strongest improvement over the baseline, reflecting its ability to leverage subtle nonlinear patterns

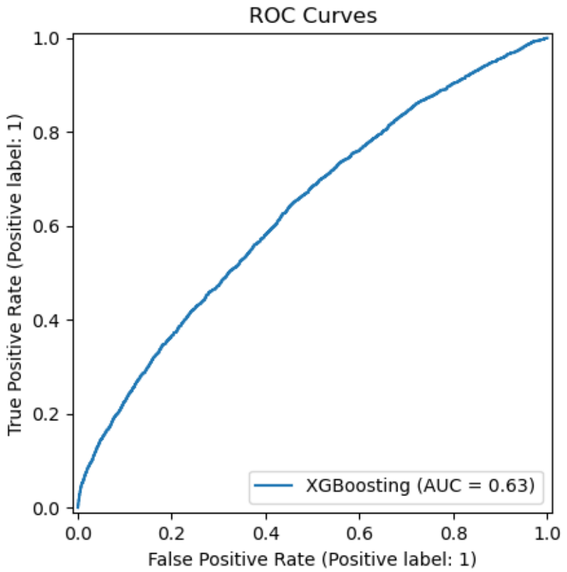
Each model produced measurable improvement over the baseline prevalence, demonstrating that the dataset contains real but limited predictive signal

XGBoost: Focused Evaluation

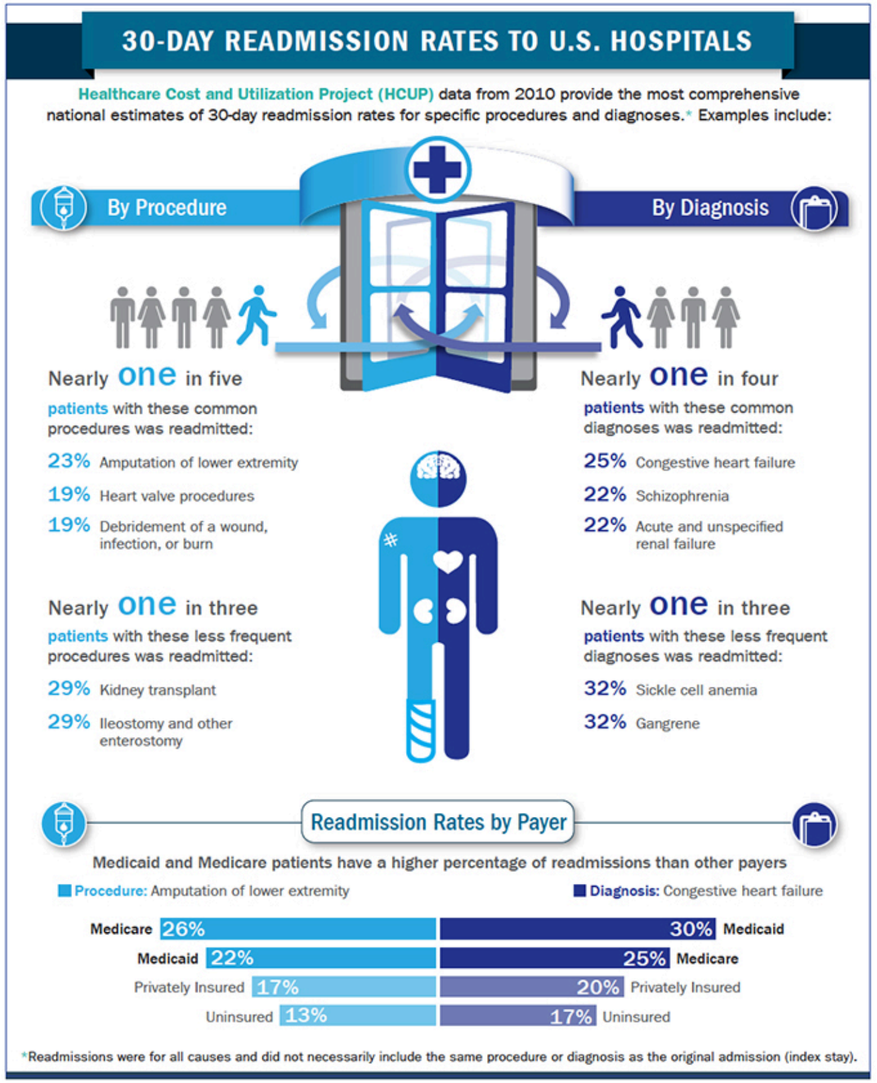
- Tuned a wide range of hyperparameters including learning rate, tree depth, subsampling, and column sampling
- Applied scale_pos_weight to account for class imbalance and emphasize minority class errors
- Achieved the highest AUC and highest average precision of all models tested
- Showed the strongest separation between predicted readmission and non-readmission scores







XGBoost provided the most consistent lift over baseline and delivered the best balance between sensitivity to rare readmissions and overall model stability



Strengthening Hospital Data Inputs

- Many risk drivers like support at home, adherence, or transportation are unobserved; hospitals could collect them through brief discharge or intake screenings
- Clinical deterioration between visits is invisible in single-encounter EHRs; linking encounters or capturing simple longitudinal vitals would help
- Current SDOH inputs come from community averages; hospitals could replace these with patient-level SDOH assessments

Improved prediction depends on capturing key clinical and social risks that are currently unobserved at the individual level

Questions?