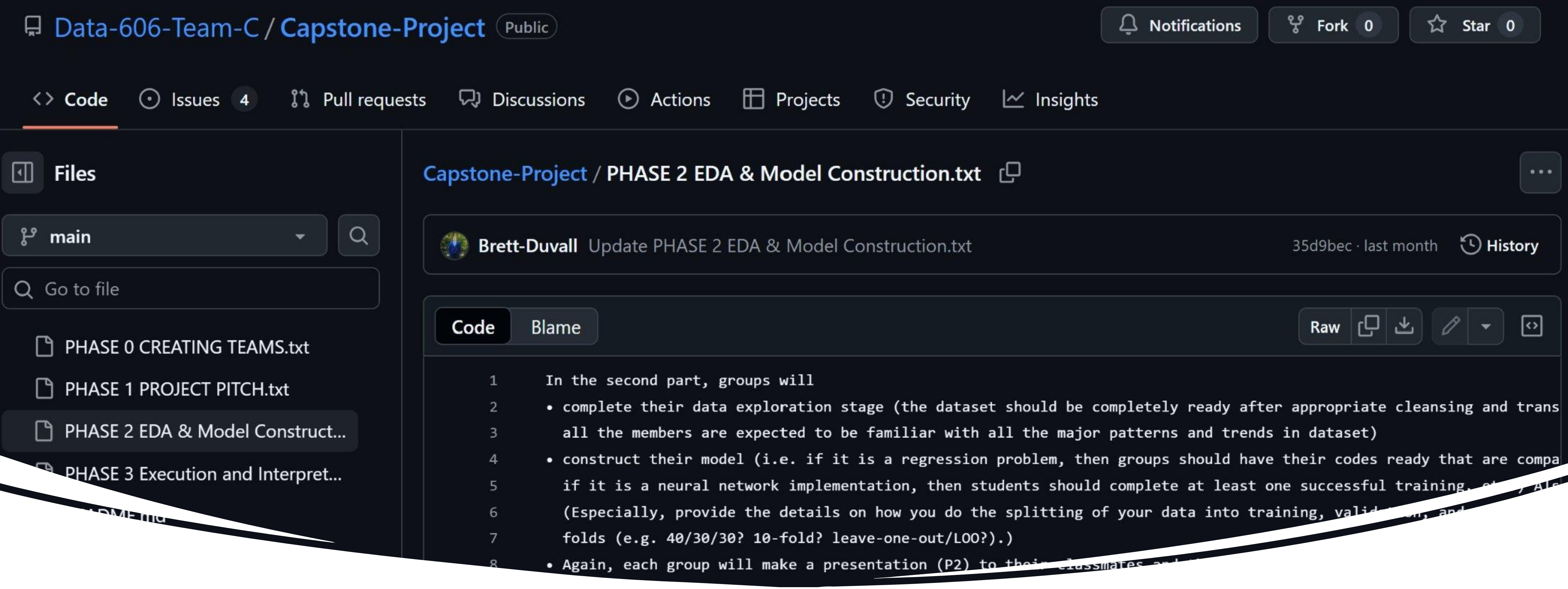
Predicting Hospital Readmissions

Data606 phase 3

Team C:

- Brett Duvall
- Mohammedamin Mussa
- Luis Vargas Ramirez


Data-606-Team-C / Capstone-Project
Public
Notifications
Fork 0
Star 0
Code
Issues 4
Pull requests
Discussions
Actions
Projects
Security
Insights
Files
main
Go to file
PHASE 0 CREATING TEAMS.txt
PHASE 1 PROJECT PITCH.txt
PHASE 2 EDA & Model Construct...
PHASE 3 Execution and Interpret...
README.md
Capstone-Project / PHASE 2 EDA & Model Construction.txt
Brett-Duvall
Update PHASE 2 EDA & Model Construction.txt
35d9bec · last month
History
Code
Blame
Raw

```
In the second part, groups will
• complete their data exploration stage (the dataset should be completely ready after appropriate cleansing and trans
  all the members are expected to be familiar with all the major patterns and trends in dataset)
• construct their model (i.e. if it is a regression problem, then groups should have their codes ready that are compa
  if it is a neural network implementation, then students should complete at least one successful training, etc.) Als
  (Especially, provide the details on how you do the splitting of your data into training, validation, and
  folds (e.g. 40/30/30? 10-fold? leave-one-out/LOO?).)
• Again, each group will make a presentation (P2) to their classmates and
```

https://github.com/Data-606-Team-C

- Size: 101,766 encounters, 50 variables, 18 MB.
- Categories: demographics, diagnoses, labs (A1C, glucose), medication use, utilization.
- Target: 30 day readmission.

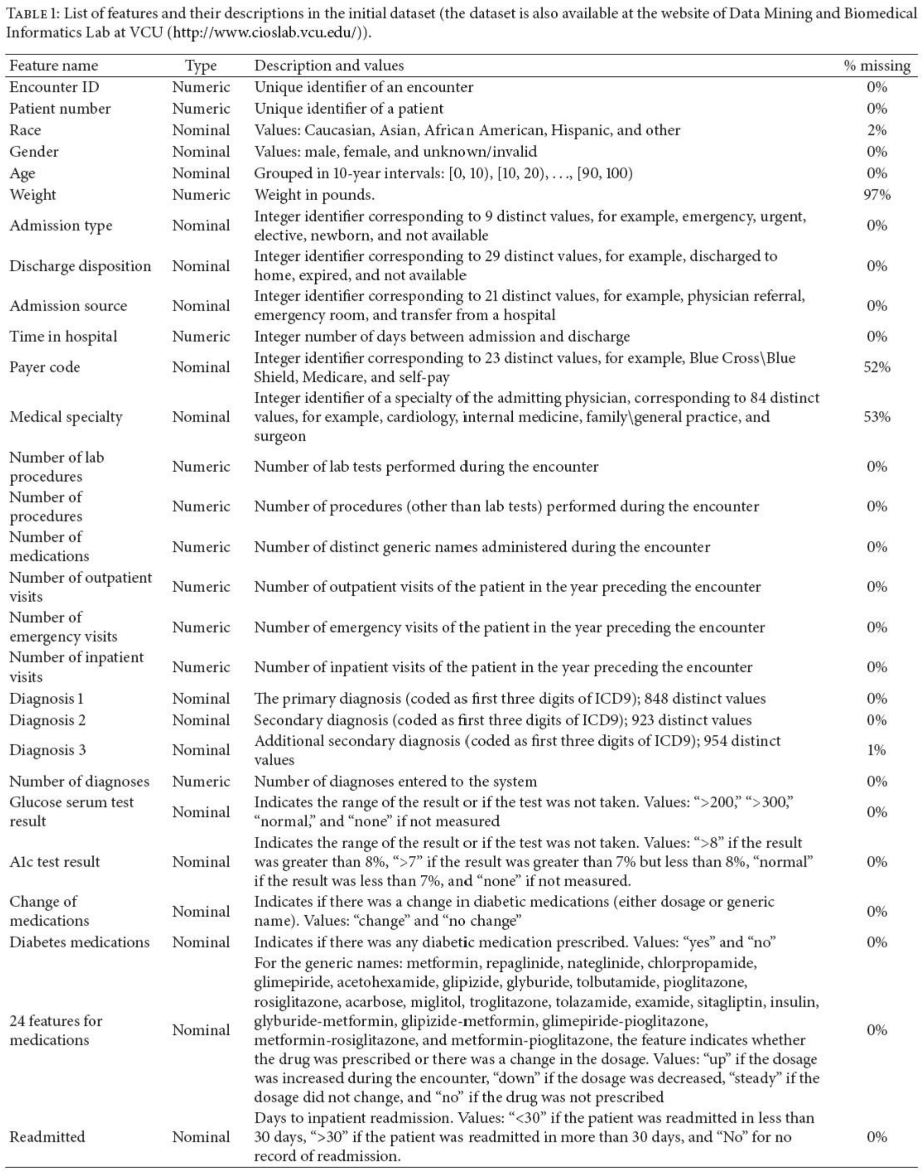


Table 1: List of features and their descriptions in the initial dataset (the dataset is also available at the website of Data Mining and Biomedical Informatics Lab at VCU (http://www.cioslab.vcu.edu/)).

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"	0%
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.	0%

Machine Learning – Classification

Target ('readmitted') has 3 categories:

- <30 (the focus of our classification)
- >30
- NO

Features kept even though missing values:

- A1C test result
- Glucose serum test result

Unnecessary features (administrative or missing values):

- Encounter ID
- patient number
- weight
- admission type
- discharge disposition
- admission source
- payer code
- medical specialty (of admitting physician)

Dataset:
https://drive.google.com/file/d/1fMbjB-5I0Suifh4QrpH1ewtXxvkAtayd/view?usp=sharing

Literature Review

- **Ashfaq et al. (2019)** implemented deep learning models such as convolutional and recurrent neural networks for hospital readmission prediction using electronic health records. Their study demonstrated improved recall and AUC over traditional models, motivating our comparison between ensemble and neural approaches for clinical prediction tasks.
- **Emi-Johnson (2025)** showed the potential of ML algorithms, particularly XGBoost, in enhancing the prediction of 30-day hospital readmissions using structured EHR. By comparing four different models, they demonstrated that ensemble-based methods such as XGBoost not only provide superior predictive performance but also robust interpretability via SHAP analysis.
- **Shukla and Tripathi (2020)** proposed an embedding-based model (EmbPred30) for predicting 30-day readmissions for diabetic patients using the same UCI dataset. Their results showed that categorical embeddings and gradient boosting improved model accuracy compared with traditional one-hot encoding. This supports our focus on feature engineering to enhance model performance.
- **Strack et al. (2014)** analyzed 70,000 clinical records in the UCI Diabetes Readmission dataset and found that conducting HbA1c testing during hospitalization was strongly associated with reduced 30-day readmission rates. Their study introduced the dataset that serves as the foundation for many subsequent projects, including ours.
- **Wang and Zhu (2021)** reviewed challenges and solutions in predictive modeling of hospital readmissions using machine learning. They compared several algorithms, including Random Forest, Support Vector Machines, and Logistic Regression, reporting AUC values typically between 0.62 and 0.70. Their work emphasizes issues such as class imbalance and preprocessing, which our project addresses using class weighting and consistent encoding.

Literature Review Metrics

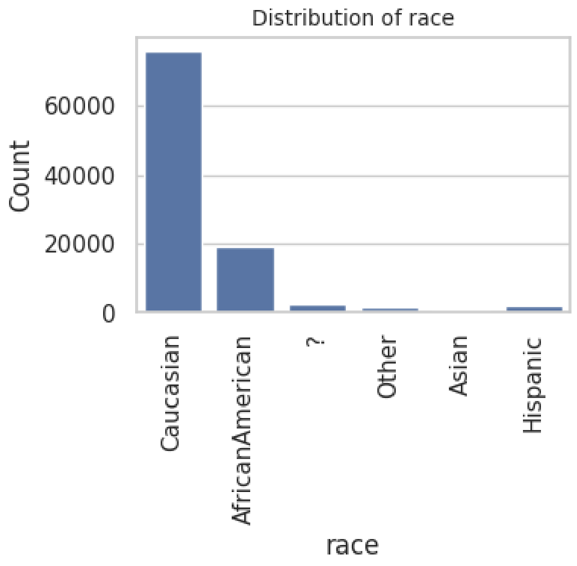
Ashfaq et al. (2019)	• AUC 0.73
Emi-Johnson (2025)	• AUC 0.58 - 0.67
Shukla and Tripathi (2020)	• AUC 0.71
Strack et al. (2014)	• AUC 0.61 – 0.63
Wang and Zhu (2021)	• AUC 0.62 – 0.70

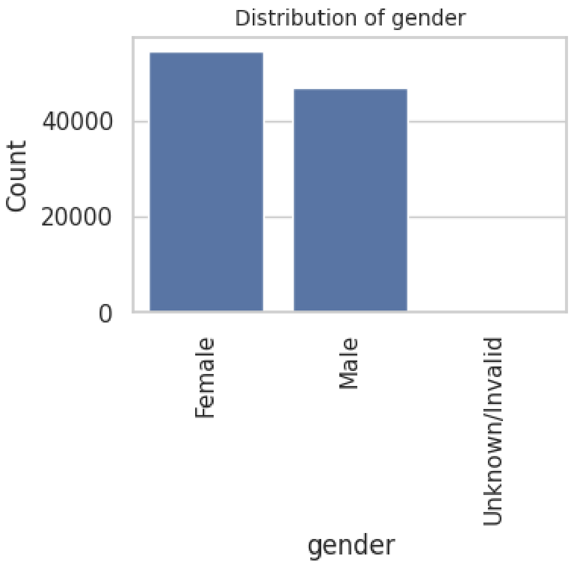
Additional Literature Review

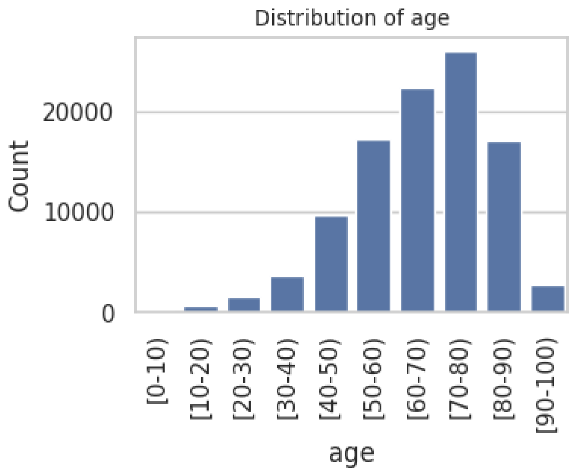
- Combining Patient Visits – Burrill (2025)
  - Adding created features
- A1C Test Reasoning - Strack et al. (2014)
  - 3 main admittance values
    - Diabetes
    - Circulatory
    - Respiratory

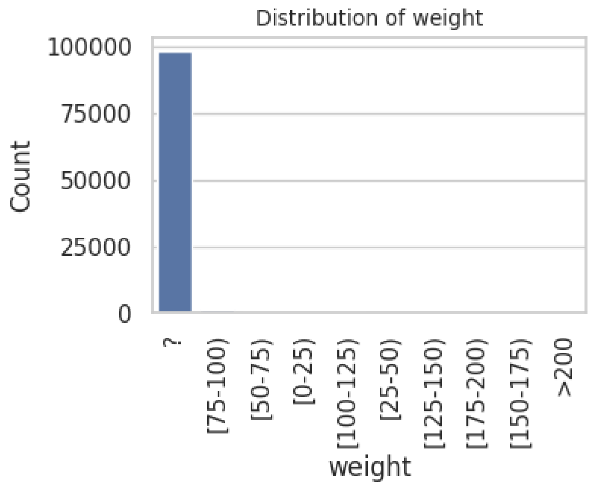
Data Exploration

Data Distribution

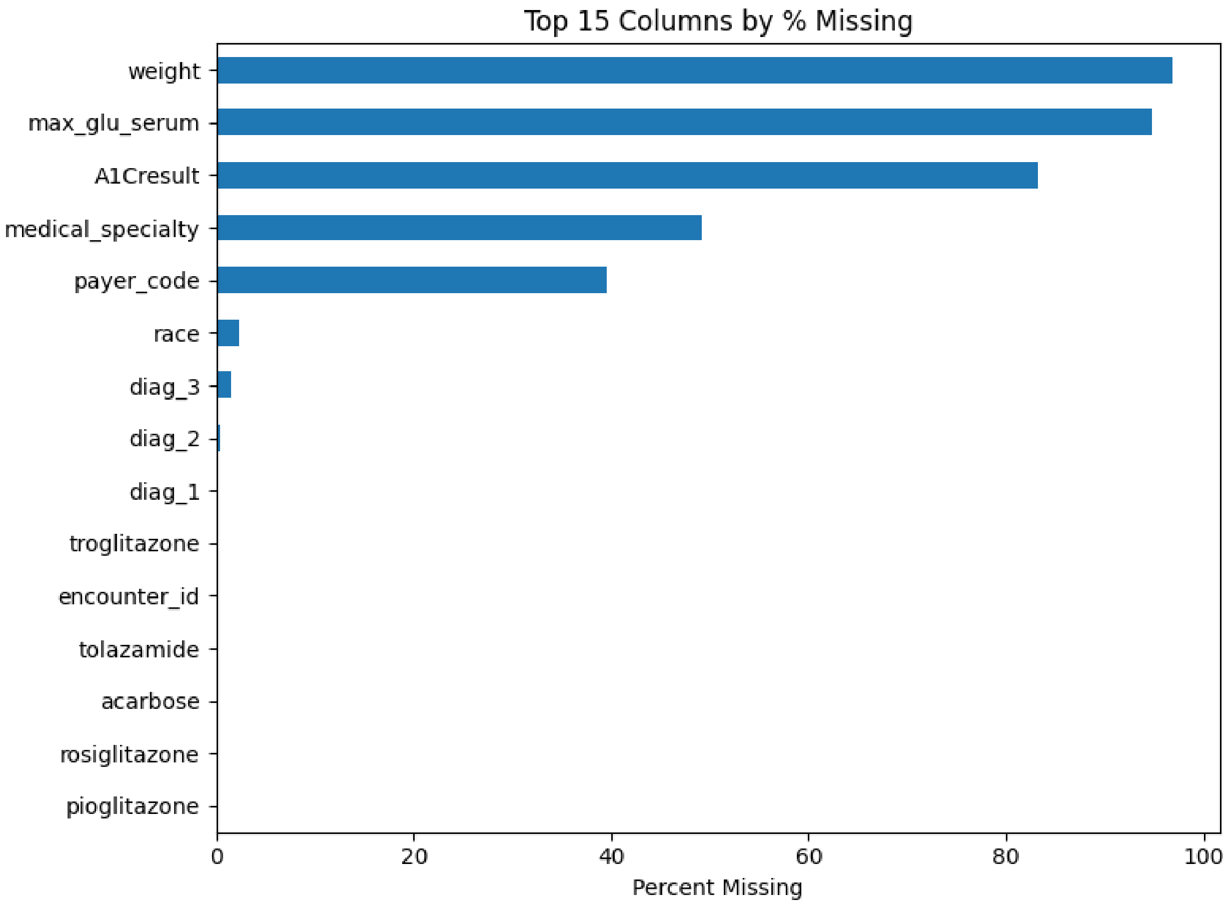

Distribution of race


Distribution of gender


Distribution of age


Distribution of weight

Missing values



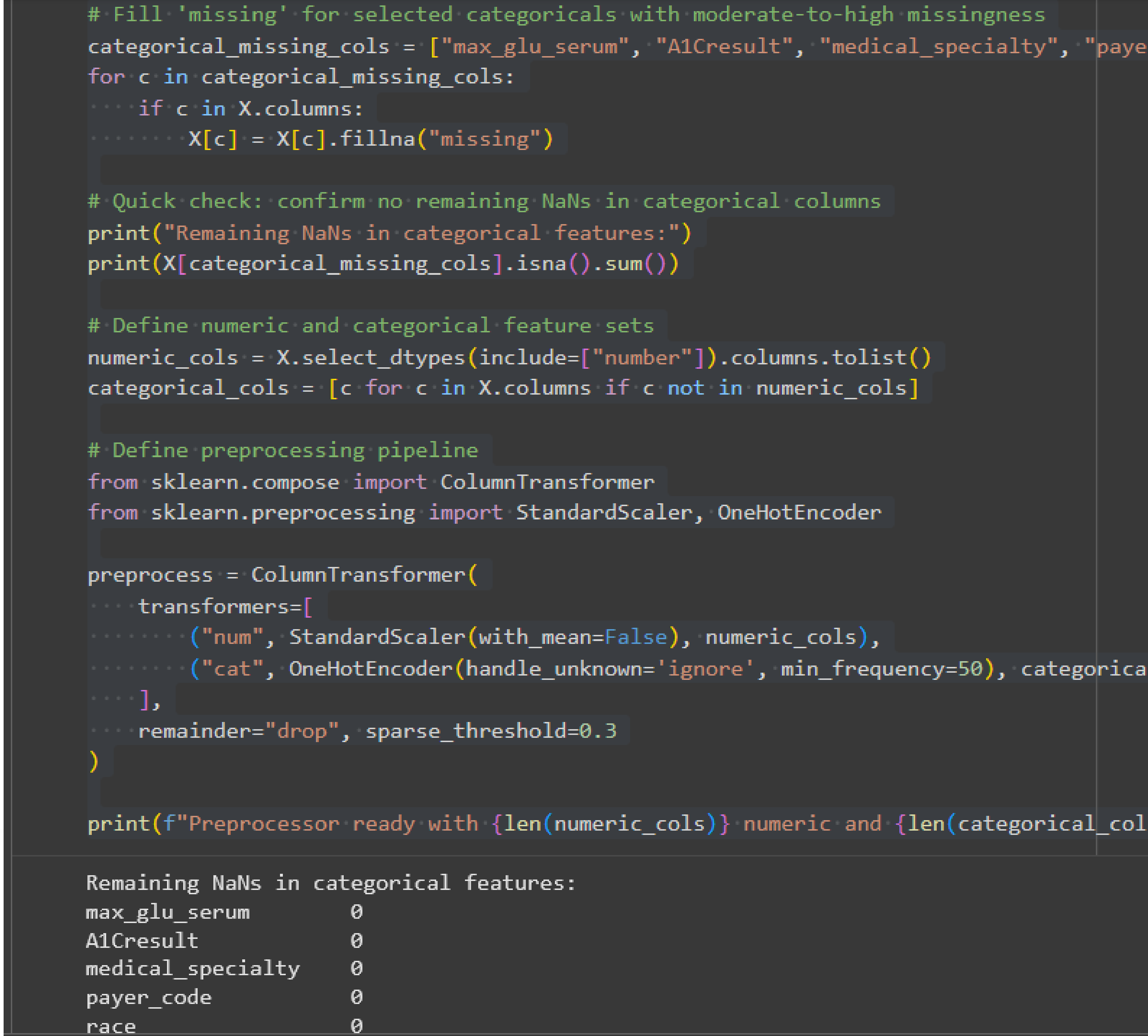
- Serum glucose is a test that measures the amount of glucose (sugar) in a patient's blood.

- A1C test measures the average amount of sugar in your blood over the past few months

	Missing Values	Percentage
max_glu_serum	96420	94.746772
A1Cresult	84748	83.277322

Dealing with missing values

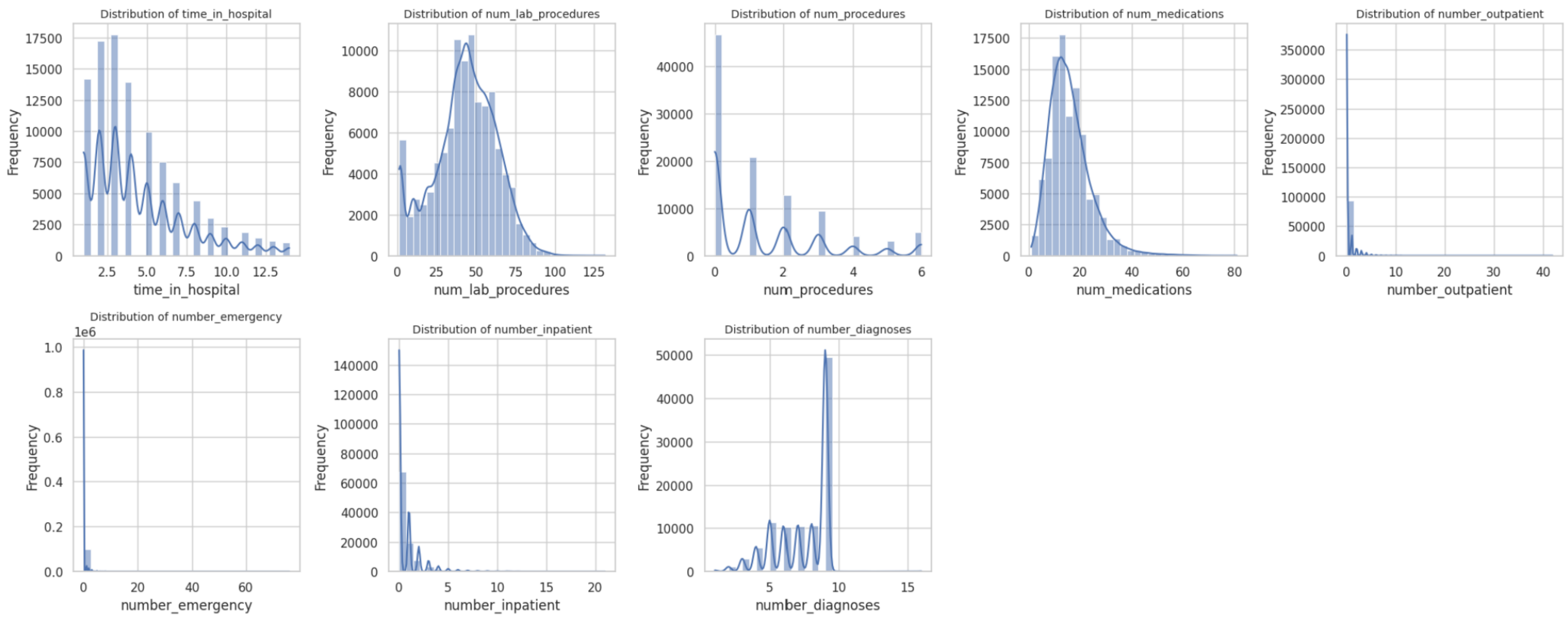
Why keeping two of the highest features with missing values?



```
# Fill 'missing' for selected categoricals with moderate-to-high missingness
categorical_missing_cols = ["max_glu_serum", "A1Cresult", "medical_specialty", "paye
for c in categorical_missing_cols:
    if c in X.columns:
        X[c] = X[c].fillna("missing")

# Quick check: confirm no remaining NaNs in categorical columns
print("Remaining NaNs in categorical features:")
print(X[categorical_missing_cols].isna().sum())

# Define numeric and categorical feature sets
numeric_cols = X.select_dtypes(include=["number"]).columns.tolist()
categorical_cols = [c for c in X.columns if c not in numeric_cols]

# Define preprocessing pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder

preprocess = ColumnTransformer(
    transformers=[
        ("num", StandardScaler(with_mean=False), numeric_cols),
        ("cat", OneHotEncoder(handle_unknown='ignore', min_frequency=50), categorica
    ],
    remainder="drop", sparse_threshold=0.3
)

print(f"Preprocessor ready with {len(numeric_cols)} numeric and {len(categorical_col
```

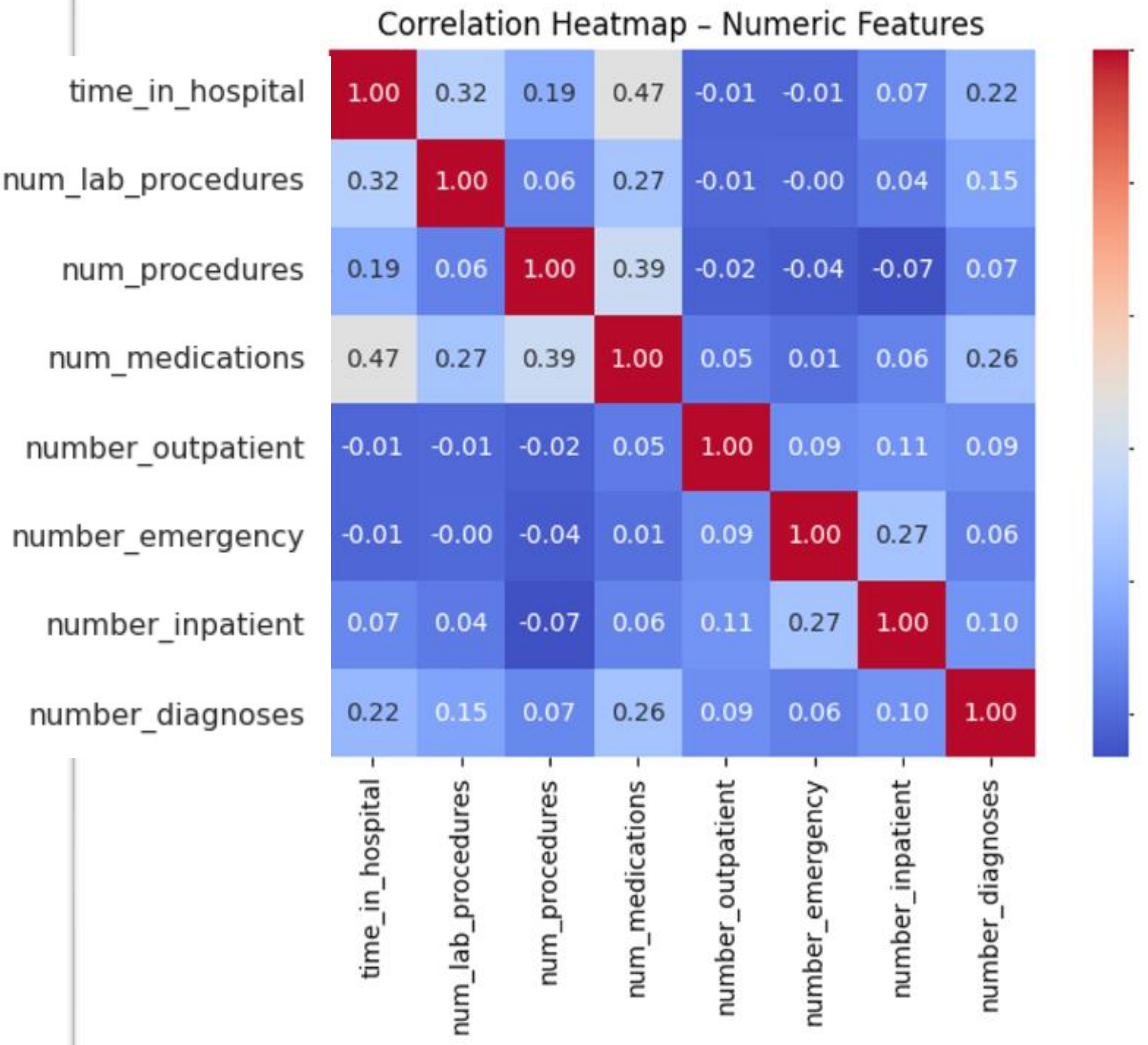
```
Remaining NaNs in categorical features:
max_glu_serum        0
A1Cresult            0
medical_specialty    0
payer_code           0
race                 0
```

Distribution of Numeric Features



Correlation Among Numeric Features

- Moderate positive correlations between longer hospital stays and more medications, also medications and procedures(other than lab tests).
- Most variables have weak relationships.



Correlation Heatmap – Numeric Features

	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	number_diagnoses
time_in_hospital	1.00	0.32	0.19	0.47	-0.01	-0.01	0.07	0.22
num_lab_procedures	0.32	1.00	0.06	0.27	-0.01	-0.00	0.04	0.15
num_procedures	0.19	0.06	1.00	0.39	-0.02	-0.04	-0.07	0.07
num_medications	0.47	0.27	0.39	1.00	0.05	0.01	0.06	0.26
number_outpatient	-0.01	-0.01	-0.02	0.05	1.00	0.09	0.11	0.09
number_emergency	-0.01	-0.00	-0.04	0.01	0.09	1.00	0.27	0.06
number_inpatient	0.07	0.04	-0.07	0.06	0.11	0.27	1.00	0.10
number_diagnoses	0.22	0.15	0.07	0.26	0.09	0.06	0.10	1.00

Model Construction

Model Construction and Validation

Goal: Predict 30-day hospital readmission among diabetic patients using the UCI Diabetes Readmission dataset (101,766 records, 8 numeric, 38 categorical features).

Task Type: Binary classification

Target variable: *readmitted <30 days (1)* vs *≥30 days or no readmission (0)*

Imbalance: ~11% positive (readmitted within 30 days)

Data Split:

- Training: 70%
- Validation: 10%
- Testing: 20%

Cross-Validation:

10-fold Stratified CV to preserve class proportions

Evaluation Metrics:

- ROC-AUC (overall ranking ability)
- PR-AUC (performance on rare positives)
- Precision@Top-10% (clinical targeting usefulness)
- F1 score (balance of precision and recall)

Model Portfolio

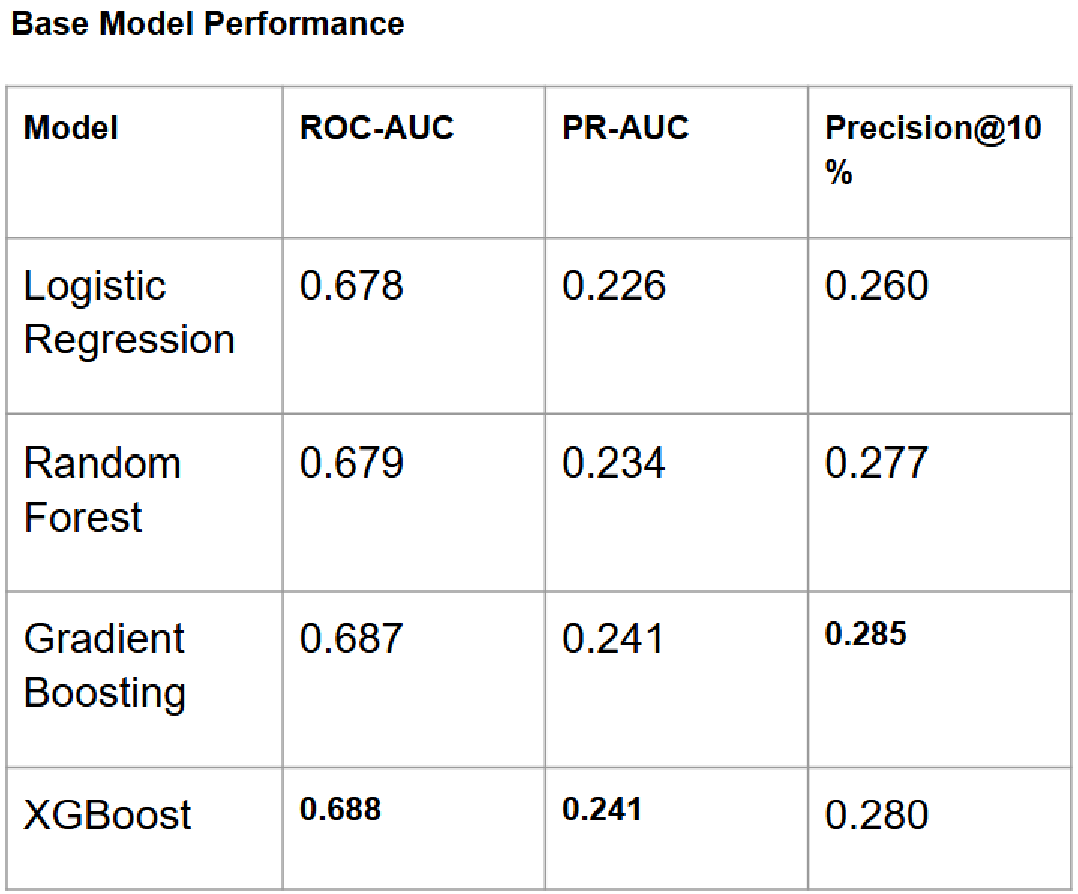
Model Comparison

- Logistic Regression – interpretable linear baseline.
- Random Forest – ensemble of decision trees capturing non-linear patterns.
- Gradient Boosting – sequential tree model emphasizing hard cases.
- XGBoost – optimized gradient boosting; best for structured data.
- SVM – margin-based classifier for comparison.

Design Choices:

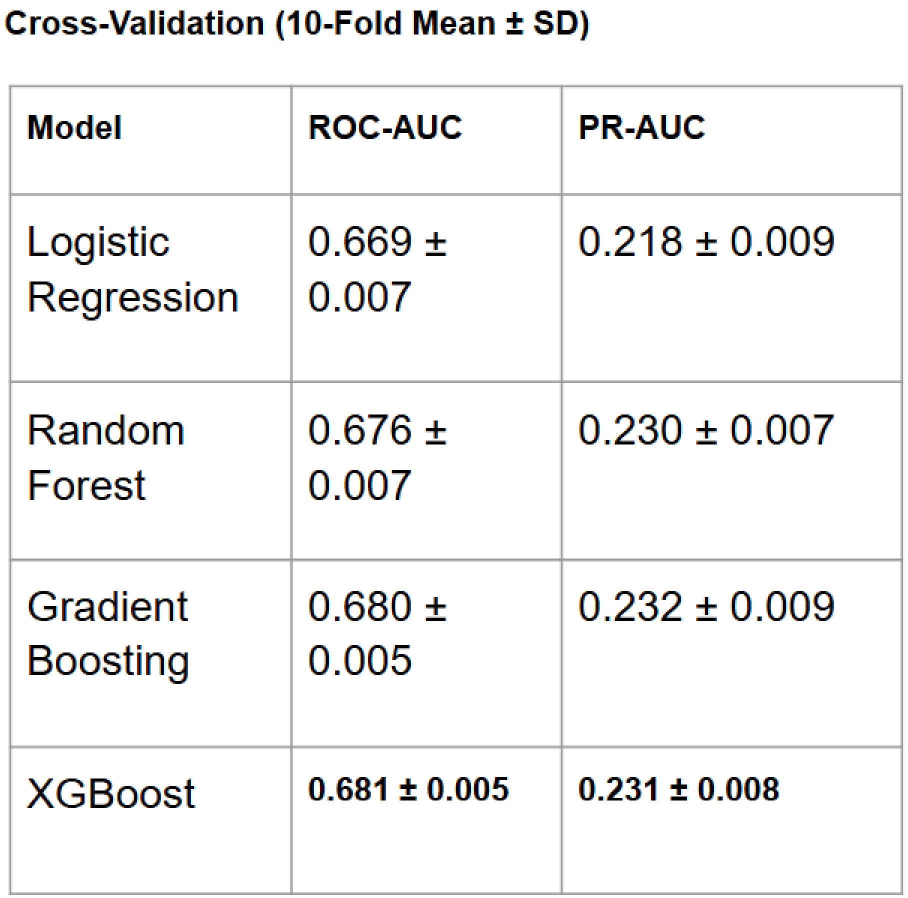
- Missing values handled; categorical variables encoded; numeric features scaled.
- Class imbalance addressed with class_weight=‘balanced’.
- Consistent preprocessing pipeline applied to all models.

Model Performance (Base vs Cross-Validation)



Base Model Performance

Model	ROC-AUC	PR-AUC	Precision@10%
Logistic Regression	0.678	0.226	0.260
Random Forest	0.679	0.234	0.277
Gradient Boosting	0.687	0.241	**0.285**
XGBoost	**0.688**	**0.241**	0.280



Cross-Validation (10-Fold Mean ± SD)

Model	ROC-AUC	PR-AUC
Logistic Regression	0.669 ± 0.007	0.218 ± 0.009
Random Forest	0.676 ± 0.007	0.230 ± 0.007
Gradient Boosting	0.680 ± 0.005	0.232 ± 0.009
XGBoost	**0.681 ± 0.005**	**0.231 ± 0.008**

Choosing the Right Metric for Hospital Readmission

Why Accuracy Is Misleading

Only 11 percent of patients are readmitted, so a model can appear “accurate” just by predicting “no readmission.”

ROC-AUC (Overall Ranking Ability)

Shows how well the model separates readmitted from non-readmitted patients across all thresholds.

Good global indicator but less sensitive to class imbalance.

PR-AUC (Focus on Rare Positives)

Captures the trade-off between precision and recall for the *readmitted* class.

Best single metric for imbalanced clinical data—tells how well the model identifies true readmissions among many non-readmissions.

Precision @ Top 10 Percent (Actionability)

Measures how many of the top 10 percent highest-risk patients truly return within 30 days.

Directly reflects how well hospitals can target limited follow-up resources.

Key Insights

- Baseline PR-AUC ≈ 0.11 (random) → our best models ≈ 0.24 (>2× improvement).
- Precision @ Top 10 % ≈ 0.28 → roughly 28 of 100 flagged patients actually readmit.
- **XGBoost** shows the best overall ranking (ROC-AUC ≈ 0.69).
- **Gradient Boosting** achieves slightly higher precision in the actionable top 10 percent.

Takeaway

Focus on PR-AUC and Precision @ Top 10 % for evaluation—these best represent clinical usefulness in predicting rare readmissions.

Benchmarking Against Prior Studies

How Our Results Compare

- *Strack et al., 2014 (UCI baseline)* – AUC 0.61 – 0.63 → our XGBoost 0.69 = major improvement over the original logistic model.
- *Wang & Zhu, 2021* – AUC 0.62 – 0.70 → our ensemble 0.68 – 0.69 = comparable to modern ML benchmarks.
- *Shukla & Tripathi, 2020* – AUC 0.71 (embedding model) → our XGBoost 0.69 = slightly lower but close.
- *Ashfaq et al., 2019 (deep learning)* – AUC 0.73 → our XGBoost 0.69 = near deep-learning performance with simpler, interpretable methods.

Interpretation

Our ensemble models (XGBoost, Gradient Boosting) perform on par with or better than most published traditional ML approaches.

They achieve strong, stable AUC and PR-AUC scores while remaining easier to interpret and deploy.

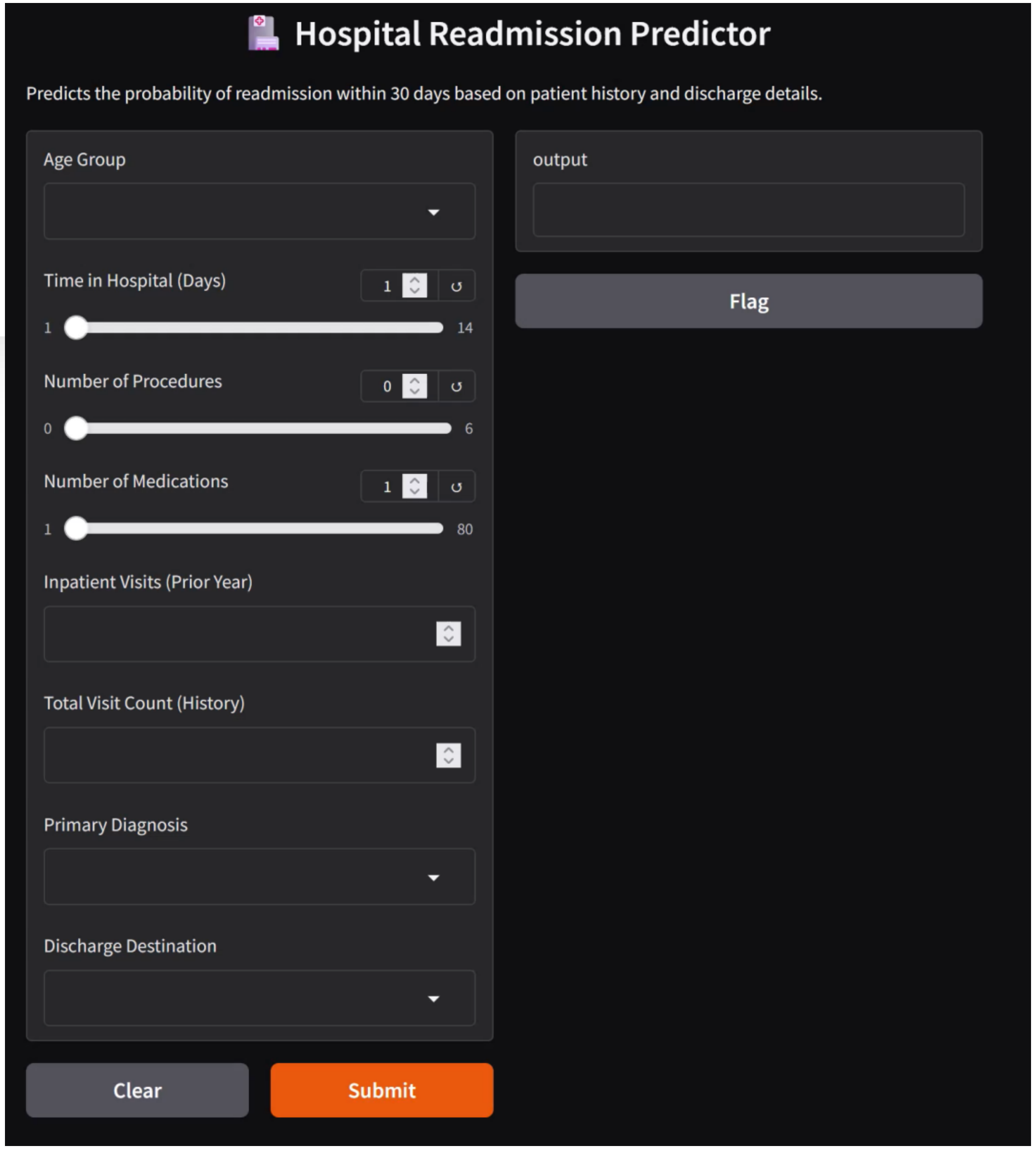
Key Takeaways

- **Best Model:** XGBoost (ROC-AUC ≈ 0.69, PR-AUC ≈ 0.24)
- **Clinical Targeting:** Gradient Boosting slightly higher precision @ 10 %.
- **Consistency:** Results align with the 0.62–0.73 AUC range reported across the literature.

Next Step: Apply SHAP analysis to explain which factors (insulin use, HbA1c level, utilization) drive readmission risk.

Gradio Demonstration

- Model will classify whether patient is at risk for <30 days readmittance
- Client can input top features
    - This is for clean demo purposes
    - Final product can allow for all features
- Client can choose to adjust patient's treatment



Hospital Readmission Predictor

Predicts the probability of readmission within 30 days based on patient history and discharge details.

Age Group

Time in Hospital (Days) 1 (1–14)

Number of Procedures 0 (0–6)

Number of Medications 1 (1–80)

Inpatient Visits (Prior Year)

Total Visit Count (History)

Primary Diagnosis

Discharge Destination

output

Thank You.
Any Questions?

REFERENCES

- Ashfaq, A., Sant'Anna, A., Lingman, M., & Nowaczyk, S. (2019). Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics, 97*, 103271. https://doi.org/10.1016/j.jbi.2019.103271

- Emi-Johnson O., Nkrumah K. "Predicting 30-Day Hospital Readmission in Patients With Diabetes Using Machine Learning on Electronic Health Record Data." (April 17, 2025), Cureus 17(4): e82437. DOI 10.7759/cureus.82437

- Shukla, S., & Tripathi, S. P. (2020). EmbPred30: Assessing 30-days readmission for diabetic patients using categorical embeddings. *arXiv preprint*, arXiv:2002.11215. https://arxiv.org/abs/2002.11215

- Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014, 781670. https://doi.org/10.1155/2014/781670

- Wang, S., & Zhu, X. (2021). Predictive modeling of hospital readmission: Challenges and solutions. *arXiv preprint*, arXiv:2106.08488. https://arxiv.org/abs/2106.08488

- Burrill, L. UC Berkeley (2025). *diabetes_readmission* [Computer software]. https://github.com/lelandburrill/diabetes_readmission