P2 – Milestone 1: Technical Execution

Project: *Predicting 30-Day Readmission for Diabetic Patients & Identifying High-Risk Patients at Discharge*

Course: MSBA 265 – Business Analytics

Student: Vishnu Vaibhav Binde

1: Introduction:

This project mainly focuses on predicting which hospitalized diabetic patients are at high risk with in 30 days

After discharge. The goal is to support hospitals in:

- Identifying **high-risk patients before discharge**
- Reducing early, preventable readmissions
- Improving discharge plans and follow-up care
- Minimizing Medicare penalties related to 30-day readmissions

Milestone P2 covers:

- Data preprocessing and features engineering
- Exploratory analysis
- Train/Test splitting
- Feature selection
- Model development (logistic regression + XG boost)
- Threshold tunning
- Evaluation on hold and test set
- All technical pipeline structures

This milestone gives foundation for the P3 .

2: Data Description

The dataset comes from the UCI diabetes 130 -Us hospitals dataset, containing:

- Total – 101,766 Patient records
- 41 Cleaned input features
- 1 Engineered binary target variable

I have converted the raw text field (readmitted) to a clinically meaningful binary target

Original value (Readmitted)	Binary value (readmitted binary)	Meaning
<30	1	Admitted- high risk
>30	0	Not admitted
NO	0	No readmitted after discharge

3: Data cleaning & Preparation

Data cleaning has been done in the reproducible python pipeline (src/preprocess.py).

Key steps:

3.1 Raw data fixes

- Removed nonclinical identifiers
- Converted “?” to NaN
- Standardized categorical codes (A1c , glucose, insulin categories)

3.2 Variable Engineering

- Encoded categorical variables using one-hot encoding
- Normalized skewed numeric features
- Ensured all features are numeric for ML algorithms

3.3 Preventing the data leakage

- Cleaning has been done before the train/test split
- Encoders fit ONLY. on training data
- Transformations have been applied identically for the test data

3.4 Final shape

- Training: set – 81k rows
- Test set: 20k rows
- Columns: 41 features + target

4 . Exploratory data analysis (EDA)

The following insights were done inside jupyter and python

4.1 Patient utilization patterns :

- High number of inpatient/ER/outpatient visits strongly correlates with readmission risk.
- Length of stay clusters heavily around 3–7 days.

4.2 Medical complexity

- Patients having 6-9 diagnoses have the highest readmission rate

4.3 Medication behavior

- Insulin adjustments (up, down, steady) shows strong predictive patterns
- A1c (medical blood test) levels also plays a significant role in the prediction

4.4 Outcome imbalance

- Only 11 -12% patients have been readmitted in the <30 days
- Imbalanced – needs proper threshold tuning

5: Feature selection

Performed using the SelectKbest implemented in src/feature_sekection.py

5.1 Candidate feature pool (42 total)

Includes

- Demographic
- Admission/ discharge types
- Utilization history
- Lab test results
- Diabetes medication patterns

5.2 Selection results

Model	Feature count	Reasons
Logistic regression	20	avoid overfitting. interpretability
XGBoost	25	Handels nonlinear patterns

This improved the model stability and reduced noise from the other weak variables

6 Model Development

Two models have been developed

6.1 Logistic Regression (baseline)

Purpose?

- Highly interpretable
- Good for explaining risk factors to clinicals

Strengths

- Coefficients easy to interpret
- Captures linear trends

Limitations

- Less in predictive performances
- Cannot capture nonlinear medical patterns

6.2 XGBoost (deployment model)

Purpose?

- Highest predictive power
- Nonlinear relationships
- Better recall

Strengths

- Capture complex interactions between medications, labtets .
- Performs better on imbalanced datasets with the threshold adjustments

This is the final model used for the clinical decision and evaluation

7. Threshold Tuning (critical clinical step)

Traditional ML uses threshold = 0.50

This is not acceptable in a hospital overview

Hospitals require

- High recall – which catches as many as patients as possible
- Willing to tolerate lower prediction – its beter tp flag to many than missing the one true risk patient

Final tuned thresholds values

Model	Threshold	Purposes
Logistic regression	0.450	Max recall baseline
XGBoos	0.100	Strong recall with better F1

Threshold tuning is essential to our safe discharge strategy .

8.Model evaluation (Test set)

Evaluation has been performed in the src/evaluate.py

8.1 Logistic regression

- Recall – 1.0 (excellent - collects all positive)
- Precision – is low to class imbalance
- F1- low-expected
- Serves as interpretability baseline

8.2 XGBoost (Recommended deployment model)

- ROC-AUC: 0.680
- Recall: 0.708
- Precison : low (its because imbalance problems)
- F1 score. : 0.269

Interpretation :

- XGBoost is clearly the stronger model and is chosen for the deployment
- Logistic regression is still valuable for the interpretability and explaining risk factors

9 Clincal interpretation: high risk patients

We will be using tuned XGBoost threshold – 0.100 to categorize predicted risk

- High risk – morely likely to be readmitted within 30days
- Low risk – historically associated with safe discharge

Clinical insights computed from the clinincal_utilis.py: of patients flagged high risk

- % of the toal patients in each risk band
- Observed readmission among low-risk predictions
- Average predicted probability of safety

Why this matters?

- Hospitals face Medicare penalties for high readmission rates
- Identifying high risk diabetic patients at discharge time helps:
- Schedule earlier follow-ups
- Provide additional care instructions
- Consider social work or case wise management
- Reducing costly readmission

10 Challenges and how I have solved them

- Extremely imbalanced target variables – solved by proper threshold tuning + class buildng
- High cardinality medication features - solved with systemic cleaning functions
- Dataset noise?, values, inconsistent formatting – solved by limiting training to sorted features

- Need for clinical friendly outputs – clear threshold based risk labelling, clinical utility summary , streamlit dashboard

11. summary

In this P2 deliverable, we:

- Worked on the UCI diabetes dataset
- Developed a binary readmission target
- Executed feature selection
- Trained two models
- Tuned thresholds to maximize clinical usefulness
- Selected the best model, XGBoost
- Generated a clinically interpretable evaluation summary

This lays the groundwork for the next step, P3: Final Model, Dashboard, and Presentation, where the functional Safe Discharge Command Center dashboard shows how the model can be used by clinicians.