HOSPITAL READMISSION PREDICTION USING MIMIC DATASET

Sheng Miao. Ph.D

AGENDA ITEMS

1. Introduction to Hospital Readmission Prediction
2. Exploring the Mimic Dataset
3. Data Analysis and Feature Engineering
4. Building the Predictive Model
5. Model Interpretation and Insights
6. Limitations, Challenges and Future Work

1. INTRODUCTION TO HOSPITAL READMISSION PREDICTION

IMPORTANCE OF PREDICTING HOSPITAL READMISSIONS

Saves Lives: Helps doctors spot patients at risk of getting worse and returning to the hospital.

Prevents Problems Early: Allows care teams to step in before issues become emergencies.

Improves Patient Care: Keeps people healthier at home, not in hospitals.

Reduces Healthcare Costs: Cuts down on expensive repeat hospital stays.

Supports Smarter Healthcare: Helps hospitals plan better and use resources wisely.

OVERVIEW OF THE MIMIC DATASET

MIMIC-IV

Medical Information Mart for Intensive Care (MIMIC)-IV, a large deidentified dataset of patients admitted to the emergency department or an intensive care unit at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA.

Size

over 65,000 patients admitted to an ICU and over 200,000 patients admitted to the emergency department.

Contents

Four main types of information: **patient demographics**, **hospital/ICU stays**, **diagnoses and treatments**, and **clinical notes**, all collected from real ICU and hospital visits.

DATASETS CREATED FROM MIMIC-IV

- Patient cohort included:
  - 10,000 hospital admissions from 6544 unique patients (Only adult, age > 18)
  - ICU-only cohort (Unintentionally—this is discussed further in the limitations).
  - Restrict to patients with complete notes only (Unintentionally—this is discussed further in the limitations)
  - 63 unique patients from 110 admissions were readmitted within 30 days of discharge,
  - Readmission Rate: 63/6544 = 0.96% (Extremely imbalance!!!)

  Note:

  ✔ Each subject_ID could have multiple admission_ID (multi-admission),

  ✔ Each admission_ID could have multiple icu_ID (multip-ICU),

- Structured data:

  ✔ Features Included: duration for each admission, duration for each ICU stay,

  ✔ Features Not Included Yet: patient demograpic (such as age), Dignosis and Treatment Information,

- **Clinical notes**

OBJECTIVES OF THE PROJECT

Develop Predictive Model

Forecast 30-day readmission probability for every discharge.

Identify Key Predictors

Reveal the clinical and operational factors that most influence readmission risk.

Actionable Insights

Package insights into practical checklists and decision aids that frontline teams can deploy.

Who might care?

Patients and Families

Clinicians (Doctors, Nurses)

Hospital Administrators

Data Scientists & Analysts

Health IT Teams

Insurance Providers / Payers

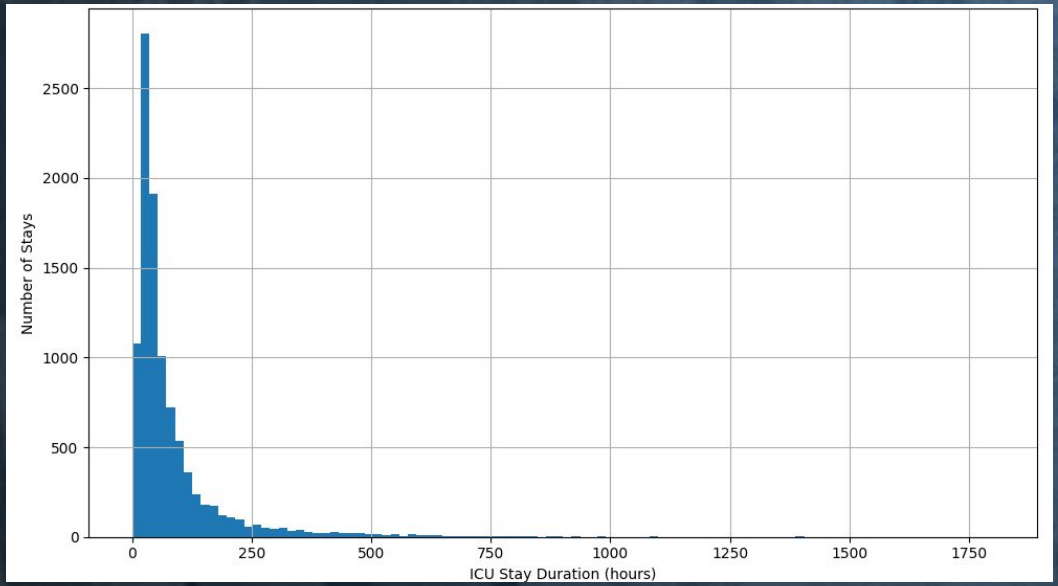
Regulatory Bodies (e.g., CMS, Medicare)

Researchers & Academics

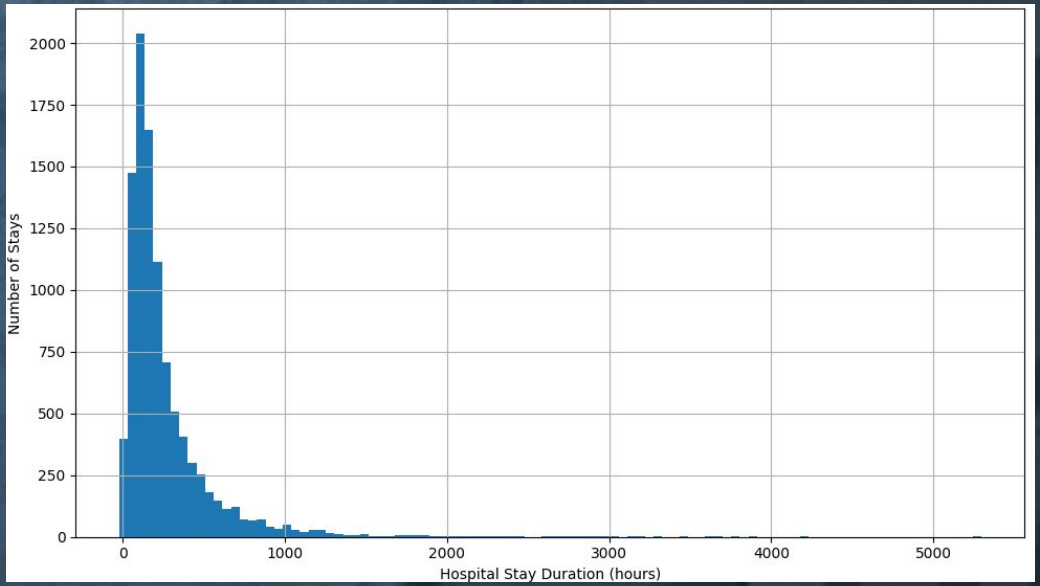
2. EXPLORING THE MIMIC DATASET

TWO FEATURES INCLUDED IN STRUCTURED DATA

Distribution of ICU Stay Duration

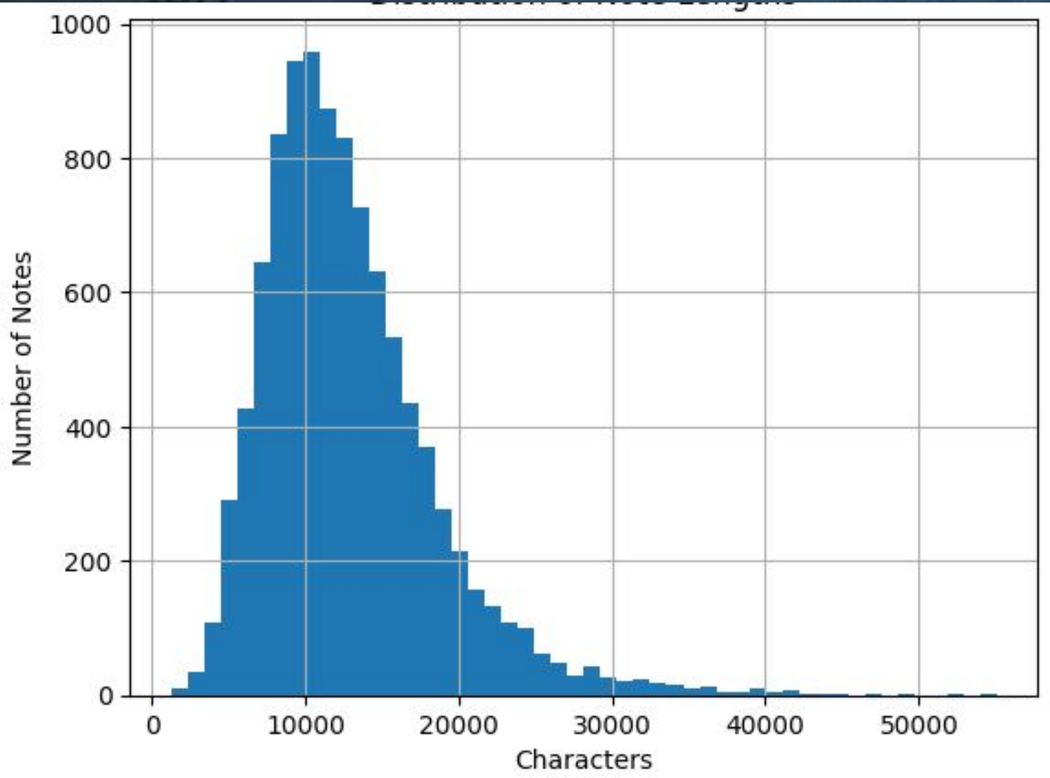


Distribution of Hospital Stay Duration

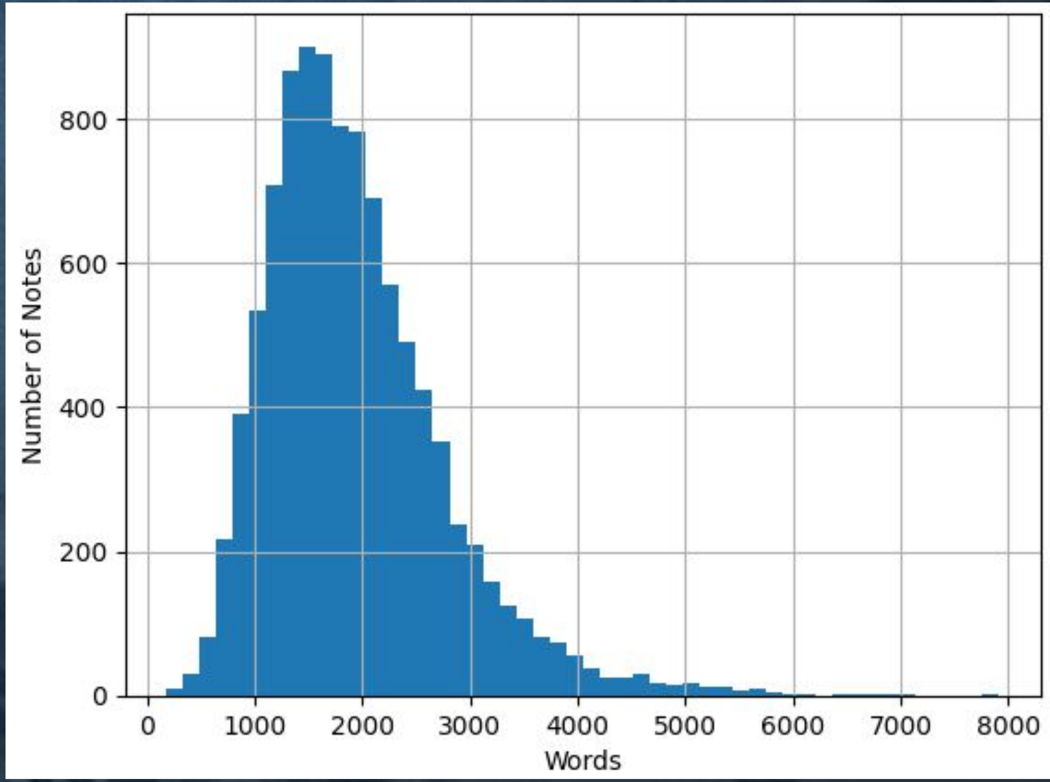


BASIC STATISTIC OF CLINICAL NOTES

Distribution of Note Lengths

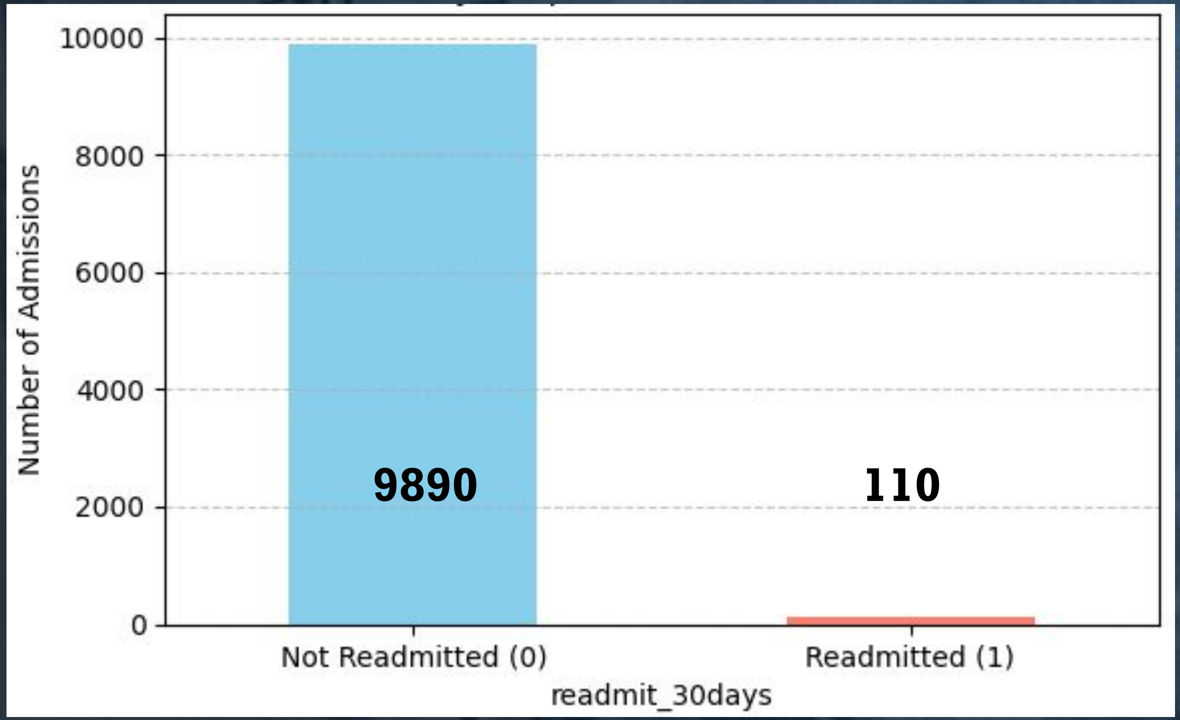


Distribution of Word Count

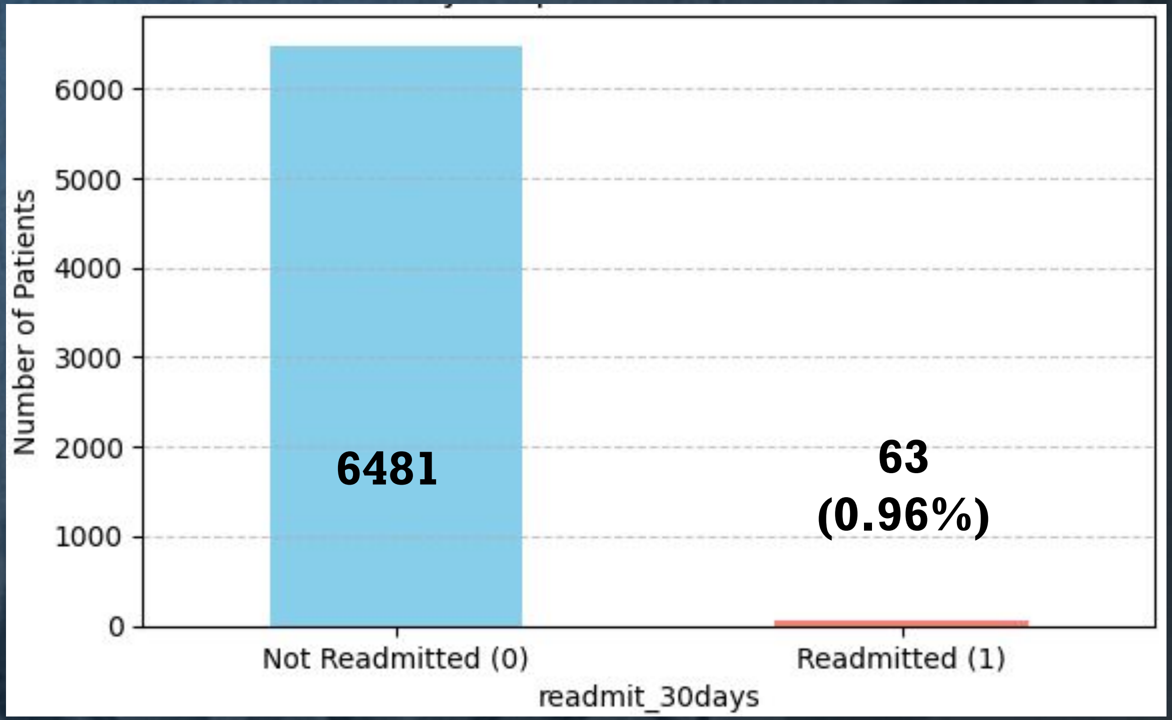


CREATE A BINARY TARGET VARIABLE FOR 30-DAY-READMISSION

30-day Hospital Readmission
Count of Admissions



30-day Hospital Readmission
Count of Patients

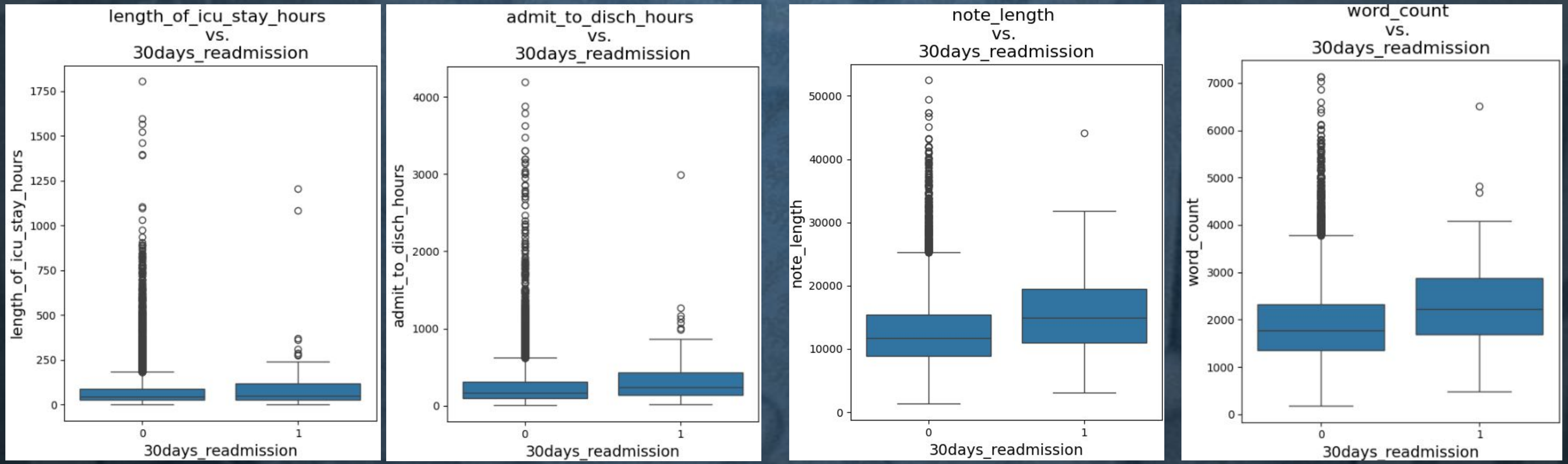


DATA WRANGLING AND CLEANING

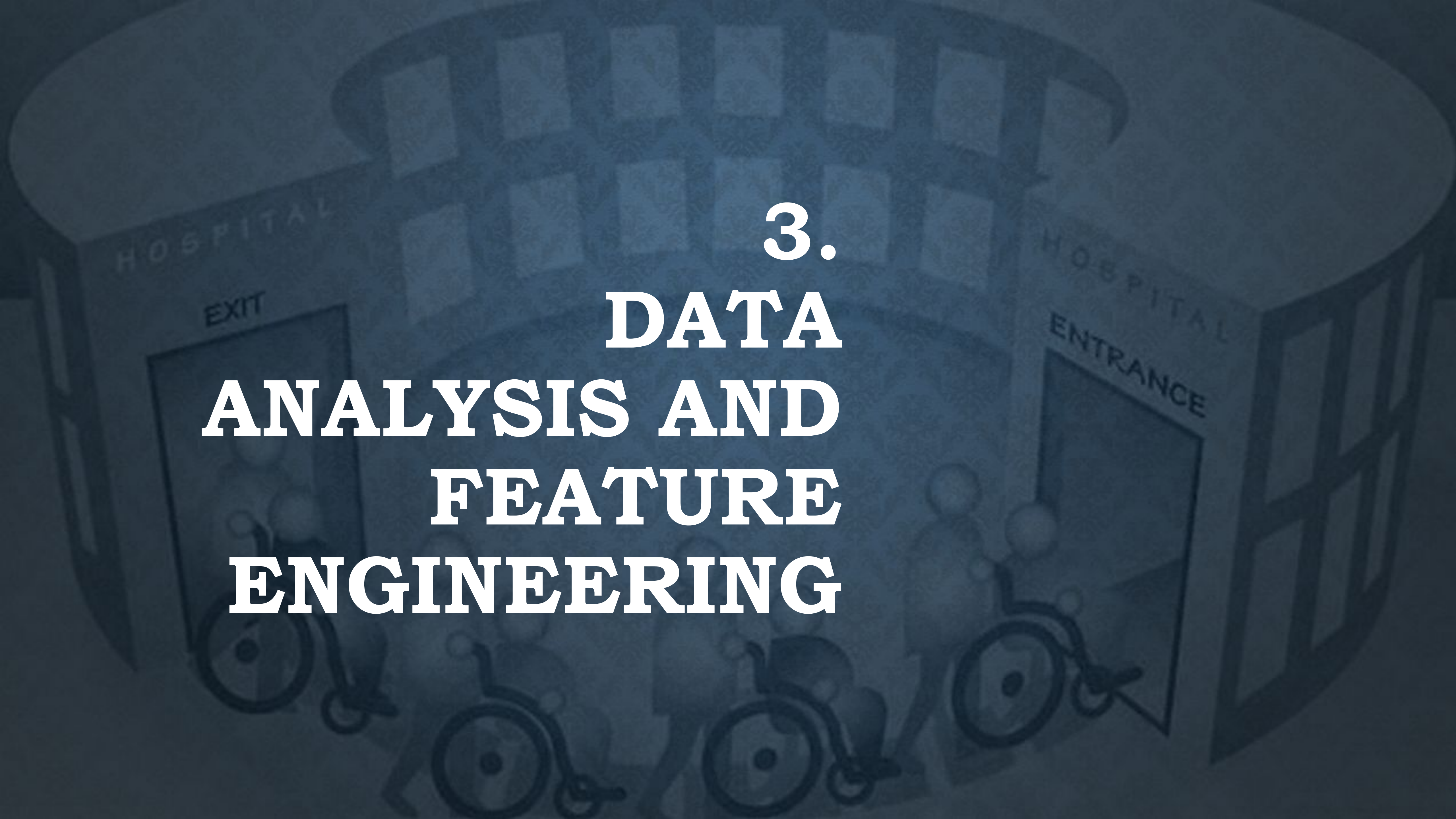
1. Removed readmission records for patients with more than one admissions.

2. Removed 10 records with data errors where dischtime is earlier than admittime.

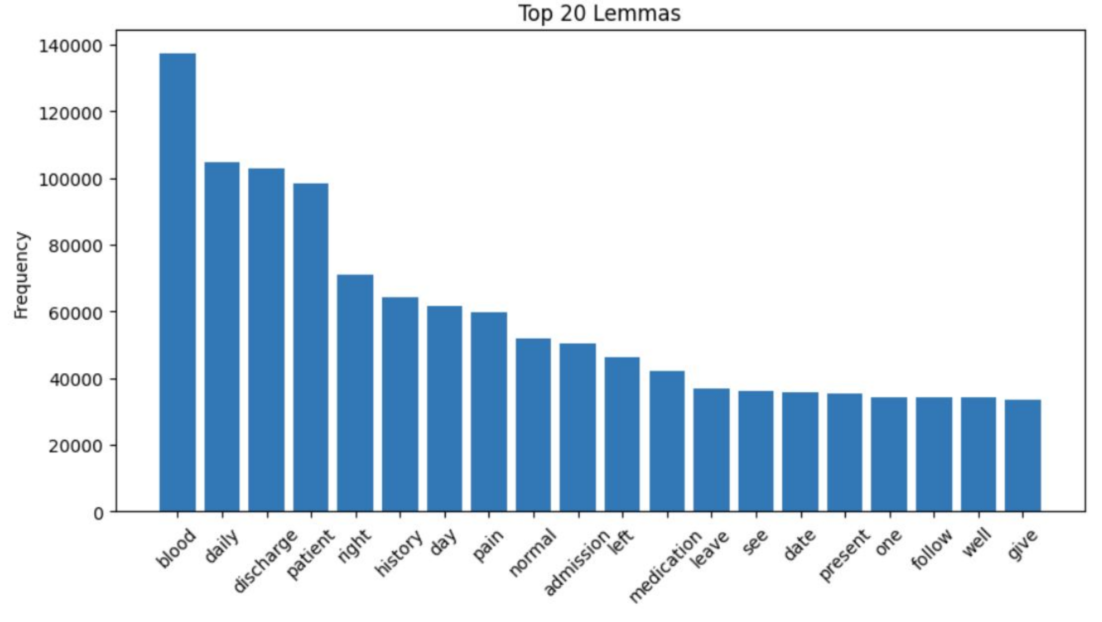
3. Elimianted Non-Predictive or Redundant Columns.

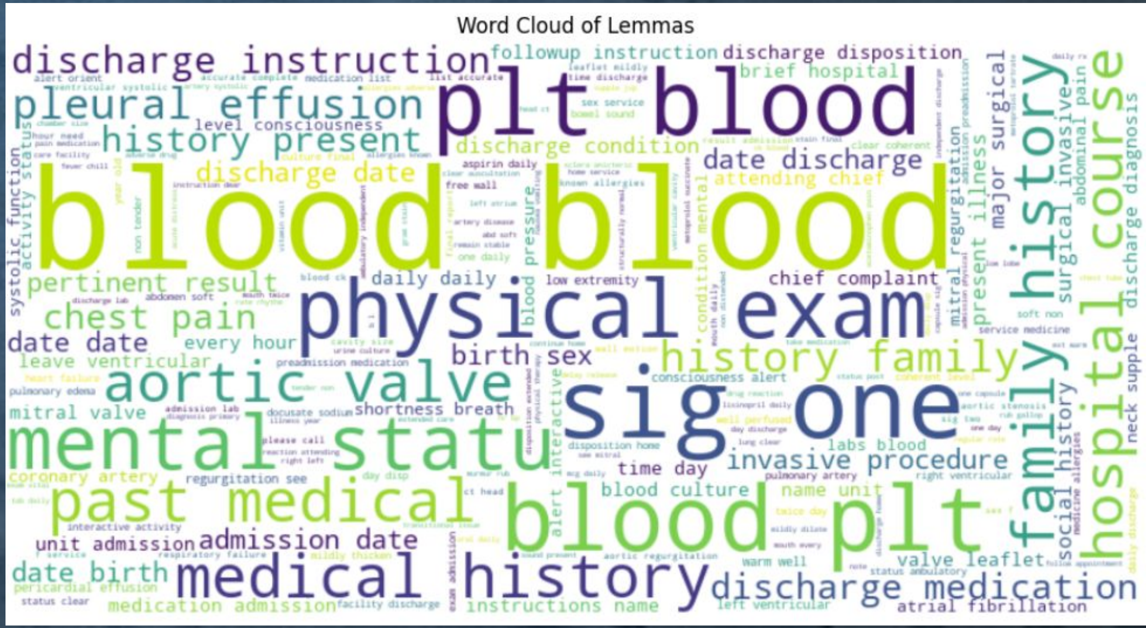

length_of_icu_stay_hours vs. 30days_readmission
length_of_icu_stay_hours
30days_readmission
0
1
0
250
500
750
1000
1250
1500
1750
admit_to_disch_hours vs. 30days_readmission
admit_to_disch_hours
30days_readmission
0
1000
2000
3000
4000
note_length vs. 30days_readmission
note_length
30days_readmission
0
10000
20000
30000
40000
50000
word_count vs. 30days_readmission
word_count
30days_readmission
0
1000
2000
3000
4000
5000
6000
7000

3. DATA ANALYSIS AND FEATURE ENGINEERING



TEXT PREPROCESSING: CLEANING, TOKENIZATION, LEMMATIZATION





spaCy Batch Processing Pipeline

FEATURE EXTRACTION, ENGINEERING AND COMBINATION

Two ways for text feature extraction:

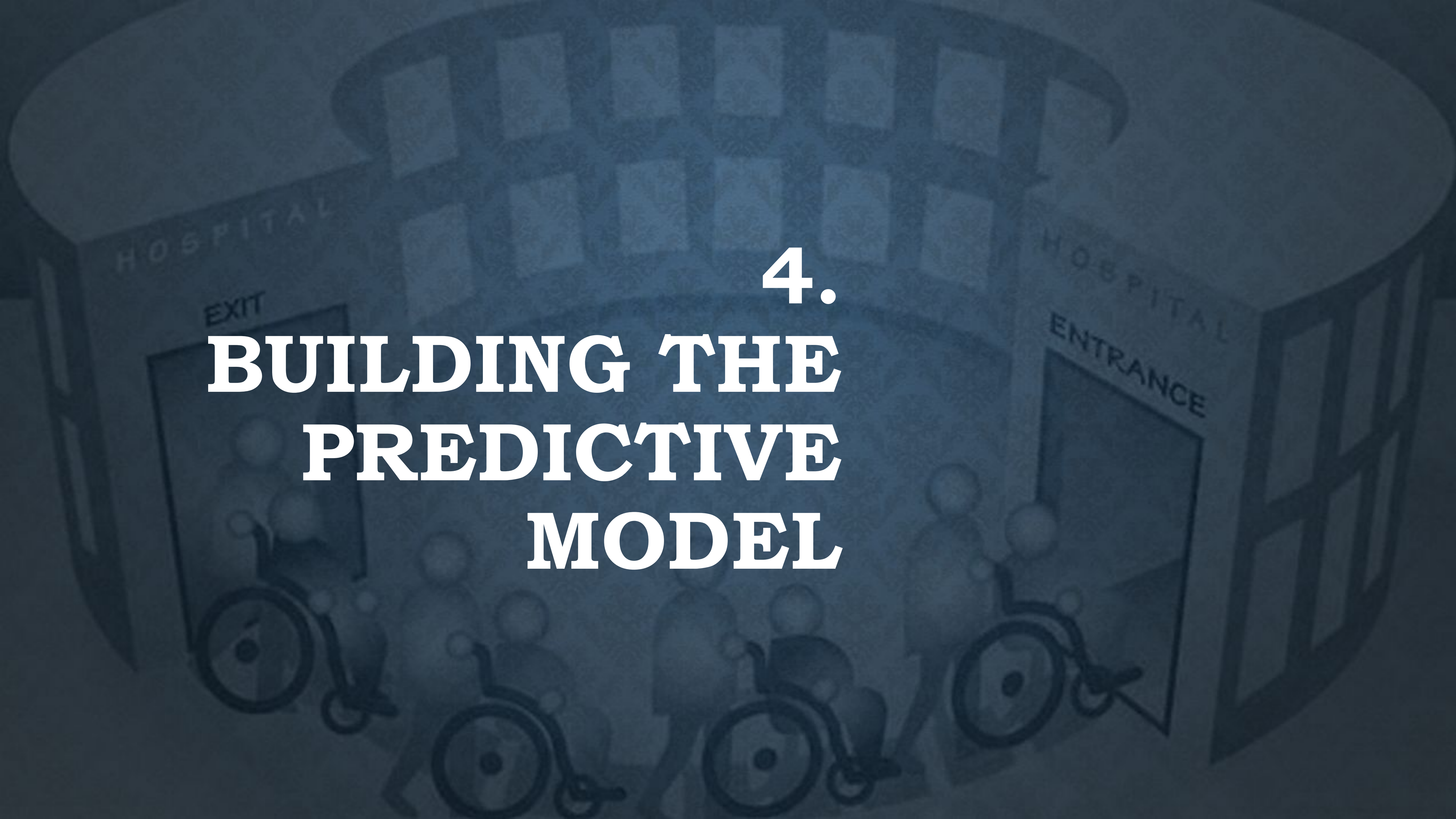
- IF-IDF -- (, 2000)
- ClinicalBERT -- (, 768)

Structured data – (, 5)

- ['length_of_icu_stay_hours', 'admit_to_disch_hours', 'note_length', 'word_count', 'advanced_spacy_lemmas_n'],
- Feature engineering: log transformation

Three combined datasets:

- Tfidf== IF-IDF vectors + structured data -- (, 2005)
- BERT==ClinicalBERT embedding + structured data – (, 773)
- Tfidf_BERT==IF-IDF vectors + ClinicalBERT embedding + structured data – (, 2773)



4. BUILDING THE PREDICTIVE MODEL

SELECTION OF MACHINE LEARNING ALGORITHMS

Dataset

- TF-IDF
- BERT
- TFIDF_BERT

Models

- Logistic Regression(LR)
- LightGBM(LGB)
- XGBoost(XGB)
- Stakcing ensemble (LR+XGB+LGB, LR)

Penalty

- L1/lambda_l1
- L2/lambda_l2
- ElasticNet

Optimization

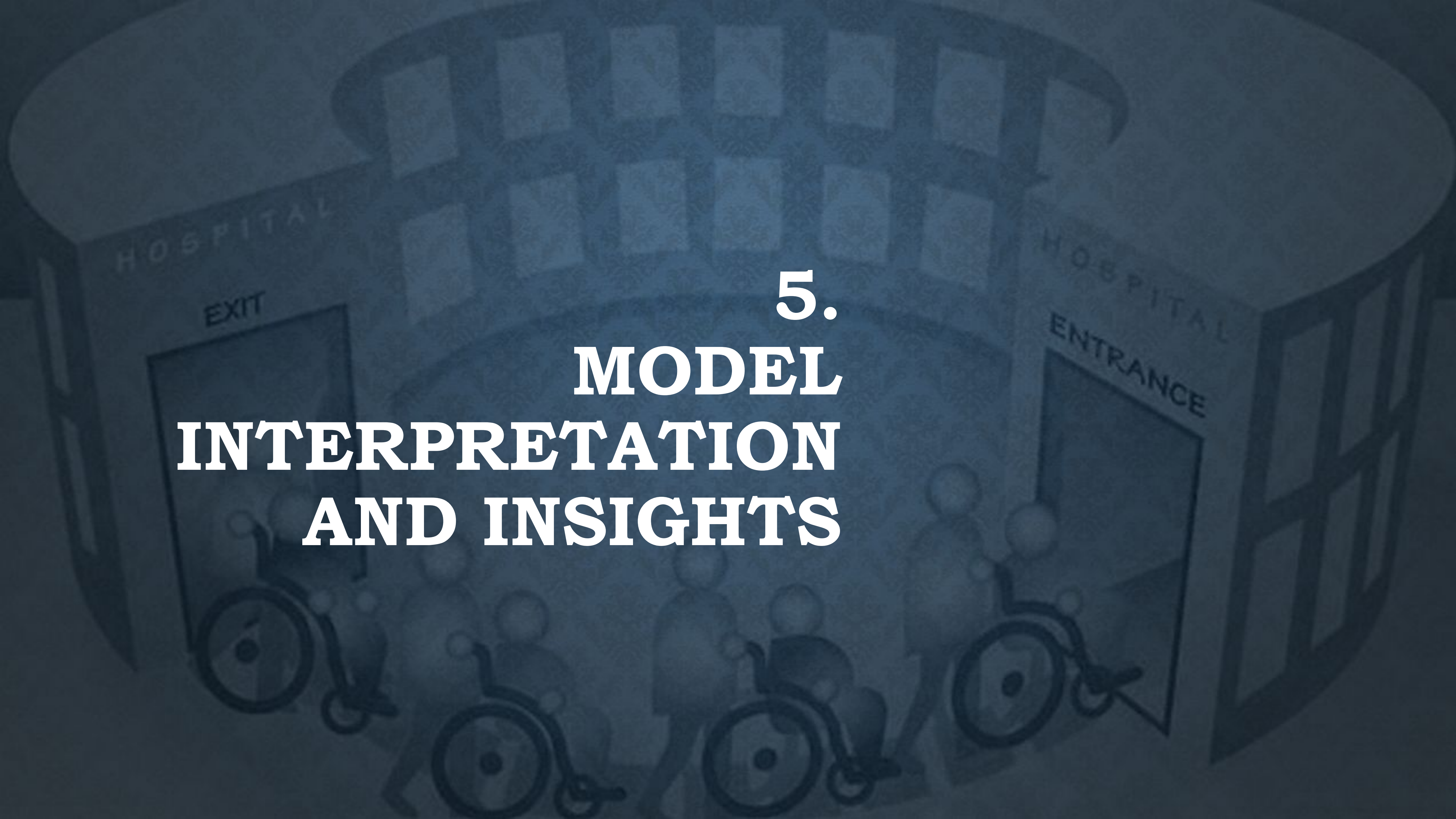
- Average Precision

Addressed data imbalance using SMOTE (synthetic oversampling) and scale_pos_weight in tree-based models

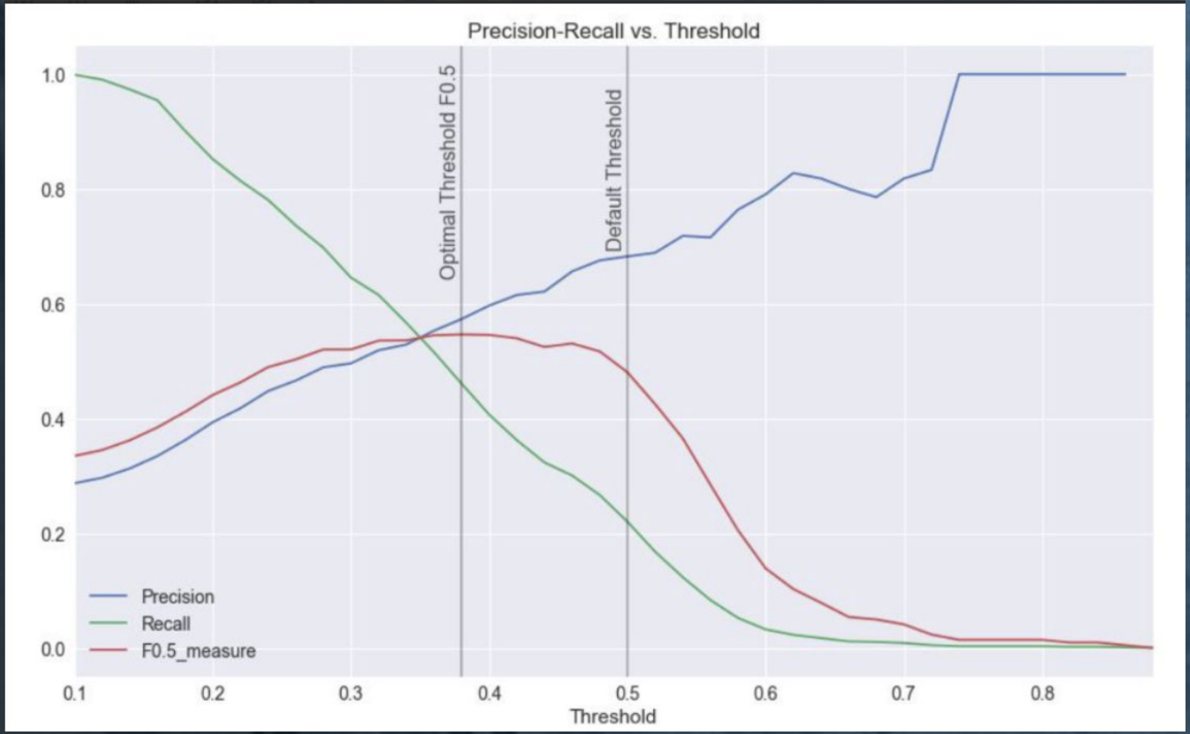
AVERAGE PRECISION AT BEST MODEL TUNING

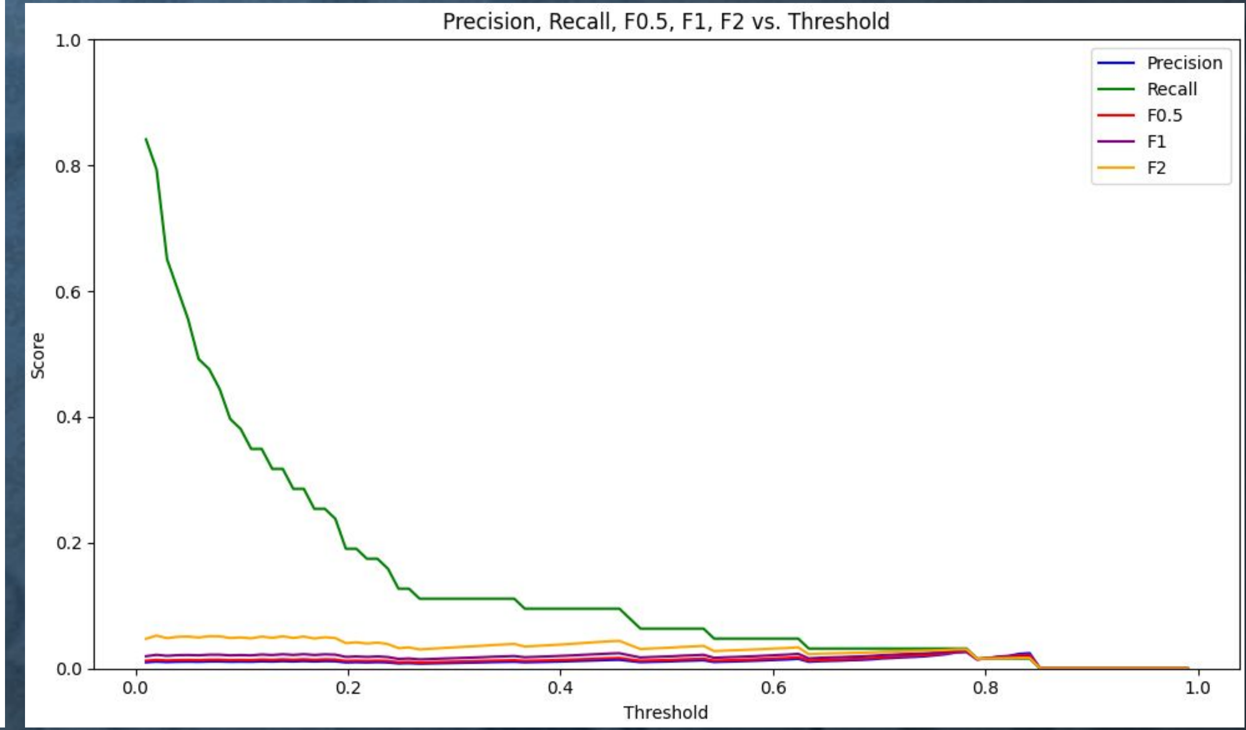
Model	Dataset	Penalty	Avg_Precision
LogReg	TFIDF_BERT	l1	**0.160765**
XGBoost	TFIDF_BERT	elasticnet	0.0205735
LightGBM	TFIDF_BERT	elasticnet	0.0469179
Stacking	TFIDF_BERT		0.0173

5. MODEL INTERPRETATION AND INSIGHTS



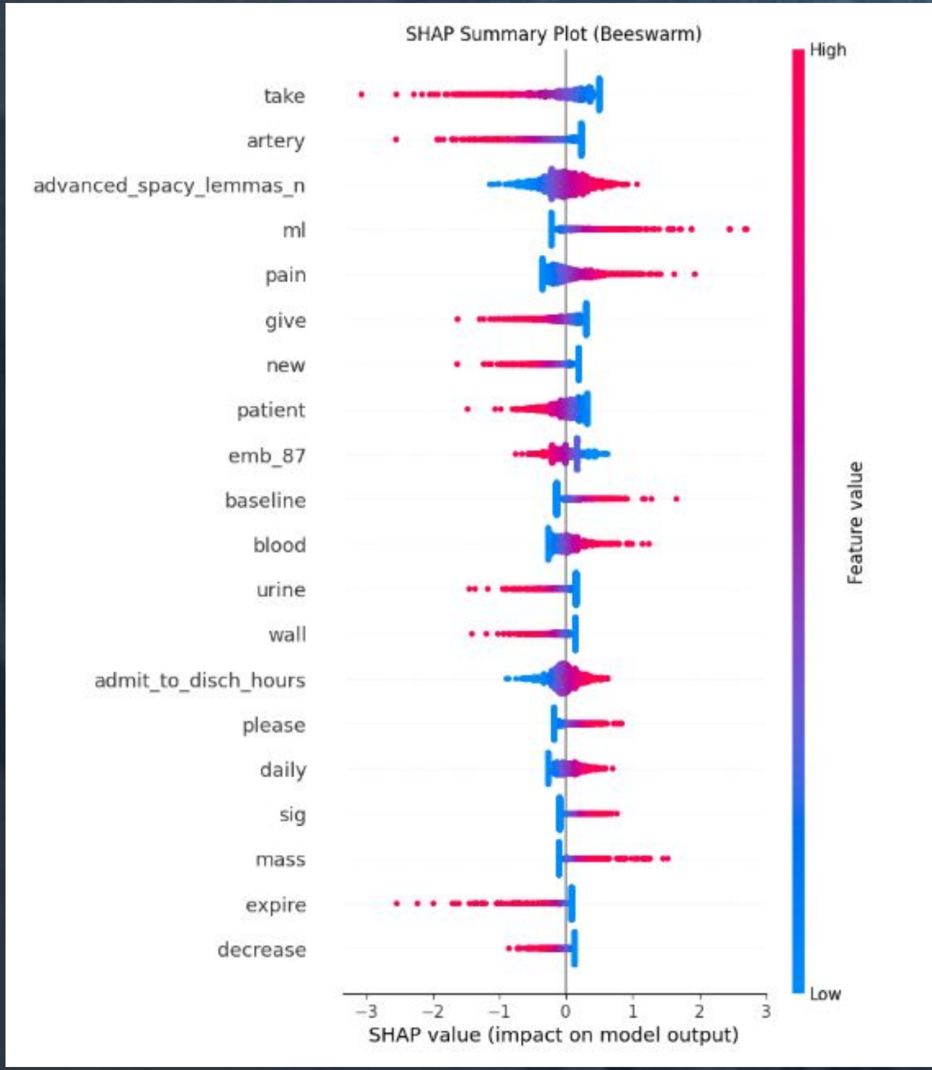
PRECISION RECALL THRESHOLDING CURVE FOR THE BEST MODEL





This result screams "highly imbalanced dataset with weak signal."

KEY PREDICTORS OF HOSPITAL READMISSIONS



- Reasonable Features:
  "Pain", "artery", "blood“,

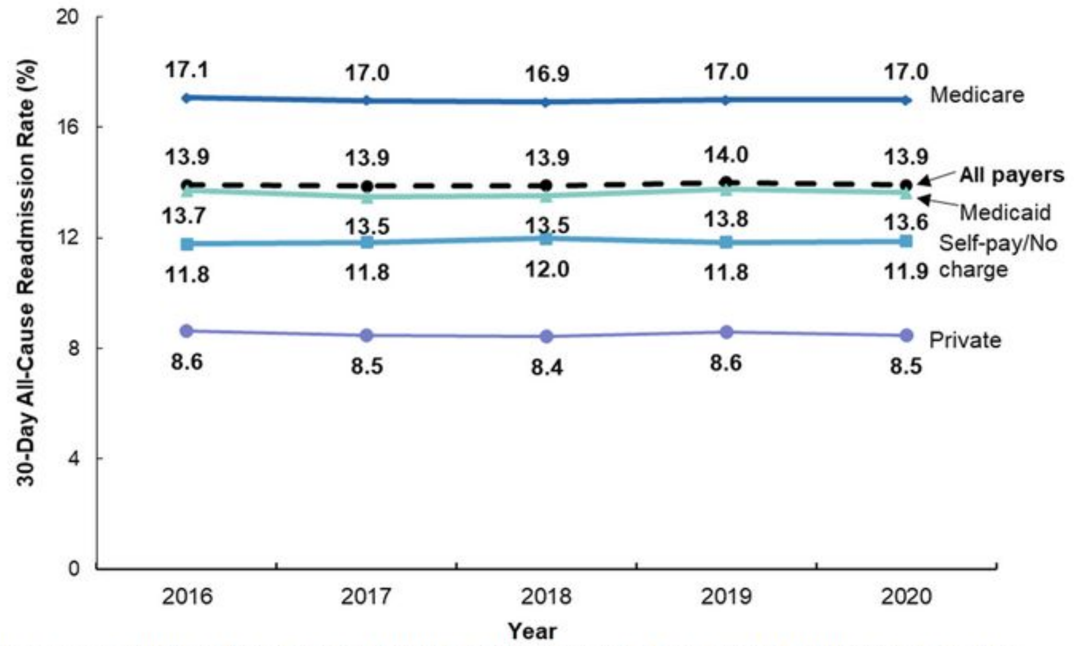
- Unreasonable Features:
  "Take", "ml", "give", "new", "patient", "baseline", "please" seem generic/stopword-like

LIMITATIONS, CHALLENGES AND FUTURE WORK

0.96% 30-DAY HOISPITAL READMISSION RATE IS MUCH LOWER THAN PUBLIC BENCHMARKS REPORTED FOR GENERAL HOSPITALIZED POPULATIONS

Healthcare Cost and Utilization Project (HCUP) Statistical Brief

Figure 1. Rates of 30-day all-cause readmissions by expected primary payer, 2016-2020



Source: Agency for Healthcare Research and Quality (AHRQ), Healthcare Cost and Utilization Project (HCUP), Nationwide Readmissions Database (NRD), 2016-2020.

Table 3. Rate of readmission for all causes within 30 days by principal diagnosis category at index admission, 2020

Rank	Principal diagnosis at index admission[a]	Readmission rate[b]	Number of all-cause readmissions[c]
1	Blood diseases	23.8	79,720
2	Neoplasms	19.0	212,954
3	Endocrine, nutritional, and metabolic diseases	17.3	223,149
4	Genitourinary system diseases	17.3	238,130
5	Respiratory system diseases	17.0	304,627
6	Mental, behavioral, and neurodevelopmental disorders	16.2	303,313
7	Digestive system diseases	16.0	447,677
8	Infectious and parasitic diseases	15.6	478,007
9	Circulatory system diseases	15.3	647,861
10	Skin diseases	13.4	61,403
11	Injury, poisoning, and other external causes	13.4	331,496
12	Nervous system diseases	13.3	101,948
13	Eye and adnexa diseases	8.8	2,234
14	Congenital malformations, deformations, and abnormalities	8.7	5,173
15	Conditions of newborn originating in the perinatal period	8.6	41
16	Musculoskeletal system diseases	7.4	112,376
17	Ear and mastoid process diseases	6.5	1,886
18	Pregnancy, childbirth, and the puerperium	3.6	134,260
N/A	**Overall (any diagnosis)**	**13.9**	**3,850,413**

https://hcup-us.ahrq.gov/reports/statbriefs/sb304-readmissions-2016-2020.jsp

LIMITATION ON DATA FILTERING, COHORT DEFINITION

- only capturing ICU admissions (not all hospitalizations),
- Restrict to patients with complete notes only

I mistakenly used inner join

```
  n.note_id, n.text
FROM `physionet-data.mimiciv_3_1_hosp.admissions` AS h
JOIN `physionet-data.mimiciv_3_1_icu.icustays` AS i
  ON h.hadm_id = i.hadm_id
JOIN `physionet-data.mimiciv_note.discharge` AS n
  ON h.hadm_id = n.hadm_id
```

- Not exclude Patients died post-discharge.
- Not exclude neonates or missing timestamps

These filtering limitation combined may shrink the positive pool and reduce significantly 30-day hospital readmission rate in my cohort.

POTENTIAL IMPROVEMENTS AND FUTURE DIRECTIONS

- **Improve Data filtering to include all hospitalizations admissions**

- **Include more predictive features on diagnosis and treatment**

Such as Charlson Comorbidity Index, diagnostic categories, lab values, vitals

- **Refining Predictive Models**

Transforming data points can help in normalizing data and making relationships more apparent for better analysis.

CONCLUSION

- **Built a baseline ML framework** to predict 30-day readmission using an adult ICU cohort from MIMIC-IV (10 k admissions, 6.5 k patients). Readmission prevalence was **0.96 %—extremely imbalanced.**
- **Class-imbalance strategies** (SMOTE + class weighting) and a stacked LR / XGB / LightGBM ensemble achieved modest uplift, but the signal remains **weak .**
- Developed a predictive model for 30-day hospital readmissions using MIMIC-IV data, achieving an Average Precision of 0.16 with Logistic Regression on combined TF-IDF + ClinicalBERT + structured features.
- Precision-Recall analysis revealed challenges with severe imbalance (0.96% readmission rate), leading to high false positives and low F-scores.
- **Top drivers** include clinical-note terms ("pain", "artery", "blood") and length-of-stay features, offering actionable cues for discharge planning.
- **Key limitations:** ICU-only admissions, notes-only records, exclusion of external readmissions, and an inner-join cohort filter—all suppress true-positive counts and push the rate far below public 10-20 % benchmarks .
- **Future work:** broaden to all hospitalizations, add diagnoses/labs/comorbidity indices, and refine models to boost recall and clinical utility

ACKNOWLEDG EMENT

Benjamin Bell