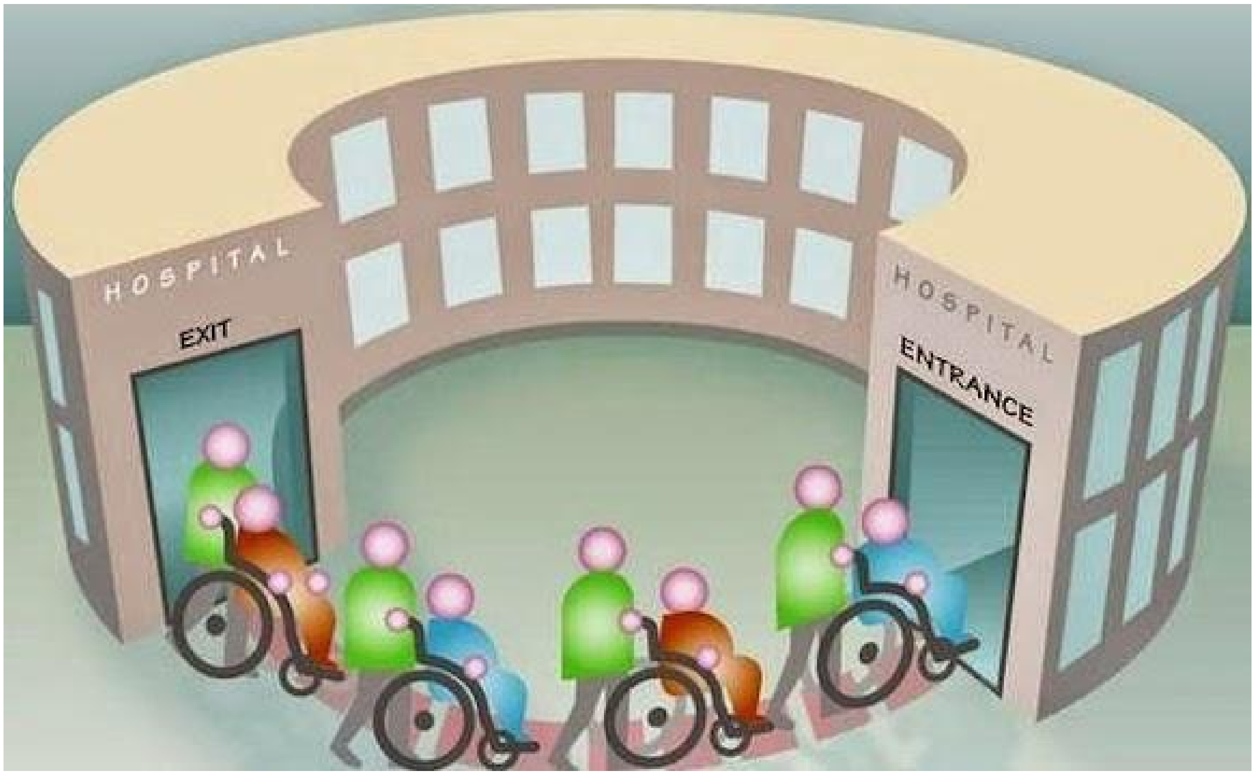
HOSPITAL READMISSION PREDICTION



JULY 23, 2025

SHENG MIAO

Table of Contents

1. Introduction

1.1 Opening Sentence

Hospital readmissions represent a significant challenge in modern healthcare, contributing to increased costs, resource strain, and poorer patient outcomes. This report presents a comprehensive predictive model developed using the MIMIC-IV dataset to forecast 30-day readmission risks, leveraging advanced machine learning techniques on clinical notes and structured data. By identifying key risk factors and achieving an average precision of 0.16 with logistic regression, my approach offers actionable insights to enhance preventive care and reduce readmission rates.

1.2 Context for the Problem

Hospital readmissions represent a significant burden on healthcare systems, contributing to escalated costs, strained resources, and adverse patient outcomes. In the United States alone, unplanned readmissions account for billions in annual expenses, with many cases potentially preventable through early identification of at-risk patients. Leveraging data science and predictive modeling can enable proactive interventions, such as tailored discharge planning and follow-up care, to mitigate these risks and improve overall healthcare efficiency.

This project utilizes the MIMIC-IV database (version 3.1), a comprehensive, de-identified resource from PhysioNet containing critical care data from over 65,000 ICU patients and 200,000 emergency department admissions at Beth Israel Deaconess Medical Center between 2008 and 2022 (available at https://physionet.org/content/mimiciv/3.1/).

By analyzing its rich contents—including demographics, hospital stays, laboratory results, and clinical notes, I aim to develop robust models for 30-day readmission prediction, ultimately supporting better clinical decision-making and resource allocation.

2. Dataset

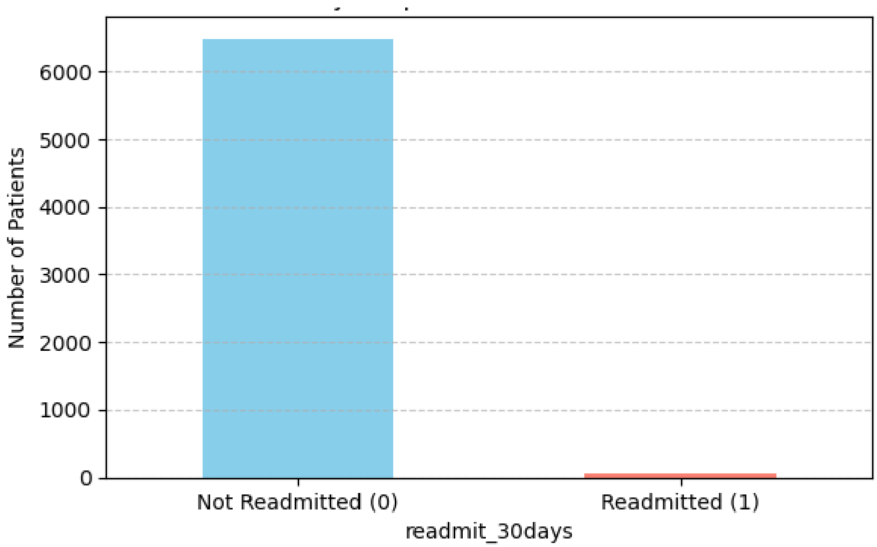
2.1 Data Source

MIMIC-IV database (version 3.1) from PhysioNet

2.2 Dataset Description

MIMIC-IV (Medical Information Mart for Intensive Care, version IV) is a large, freely available, de-identified database containing detailed information on patients admitted either to the emergency department or to an intensive-care unit at Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA. Spanning more than 65 000 ICU admissions and over 200 000 emergency-department visits, MIMIC-IV captures four broad categories of data—patient demographics, hospital/ICU stays, diagnoses & treatments, and rich unstructured clinical notes—making it one of the most comprehensive public resources for critical-care research.

For this study I selected 10 000 hospital admissions representing 6 544 unique adult patients. Because my SQL join was performed through the icustays table, the resulting cohort consists exclusively of ICU encounters; admissions without an ICU stay do not appear in the analysis (a limitation revisited in the discussion). Restricting the sample further to encounters with complete discharge notes yielded 63 readmissions within 30 days—a rate of 0.96 %. This is markedly lower than the 13.9 % all-cause 30-day readmission rate reported in the HCUP Nationwide Readmissions Database for 2020 (available at https://hcup-us.ahrq.gov/reports/statbriefs/sb304-readmissions-2016-2020.jsp), highlighting both the narrow scope of an ICU-only cohort and the potential under-capture of readmissions that occur at other hospitals or among patients who do not survive to discharge. Each patient (subject_id) may have multiple hospital admissions (hadm_id), and each admission can contain several ICU episodes (icustay_id). From the structured tables I engineered stay-duration features, while demographic, diagnostic, and treatment variables are slated for future integration. Clinical notes were retained for text-based feature extraction to enrich the predictive modeling pipeline.


6000
5000
4000
3000
2000
1000
0
Number of Patients
Not Readmitted (0)
Readmitted (1)
readmit_30days

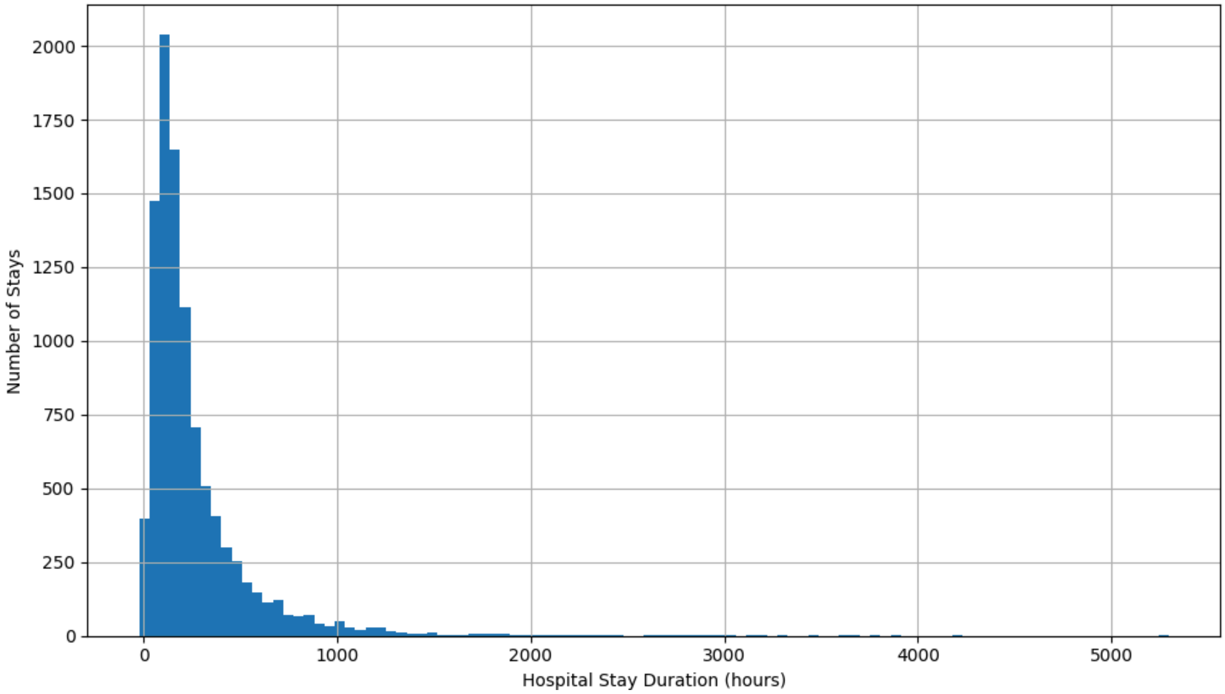
3. Exploratory Data Analysis (EDA)

3.1 Key Decisions

- Patient-level deduplication: A single patient (subject_id) may have multiple hospital admissions (hadm_id). For patients with more than one admission, we retained only the earliest admission and created a binary target variable indicating whether a 30-day readmission occurred (calculated as the interval between the first discharge and the next admission).
- ICU-stay consolidation: An admission can include several ICU episodes (icustay_id). Because non-ICU features remain identical across these episodes, we treated multiple ICU stays within the same admission as duplicates and kept a single record.
- Data-quality check: Ten admissions were dropped because their recorded discharge time preceded the admission time.

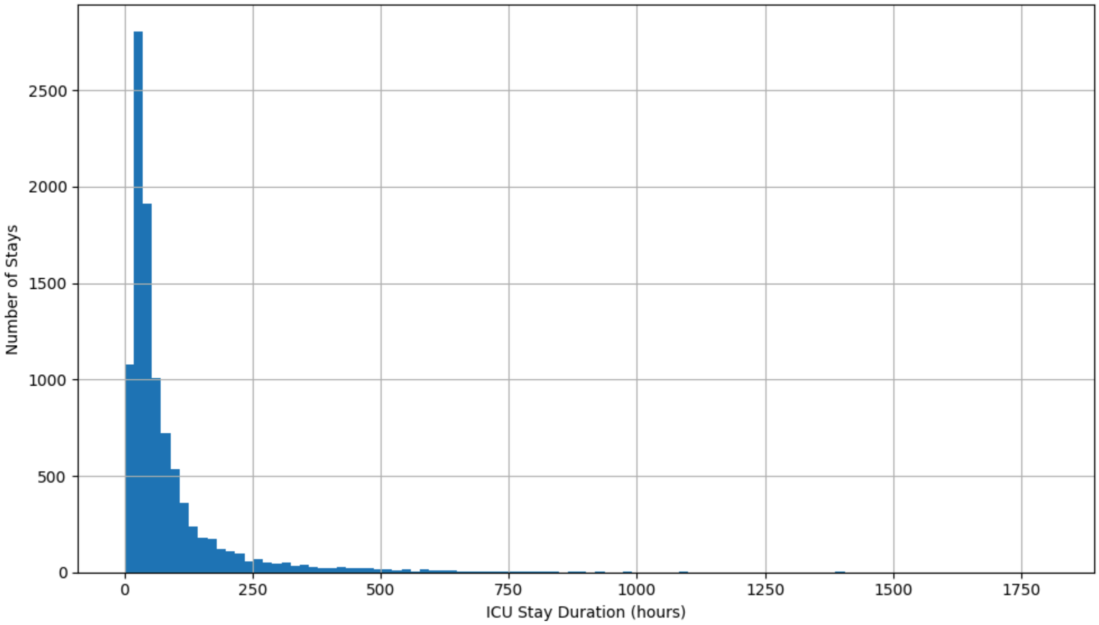
3.2 Univariate Analysis

- *Distribution of Hospital Stay Duration*



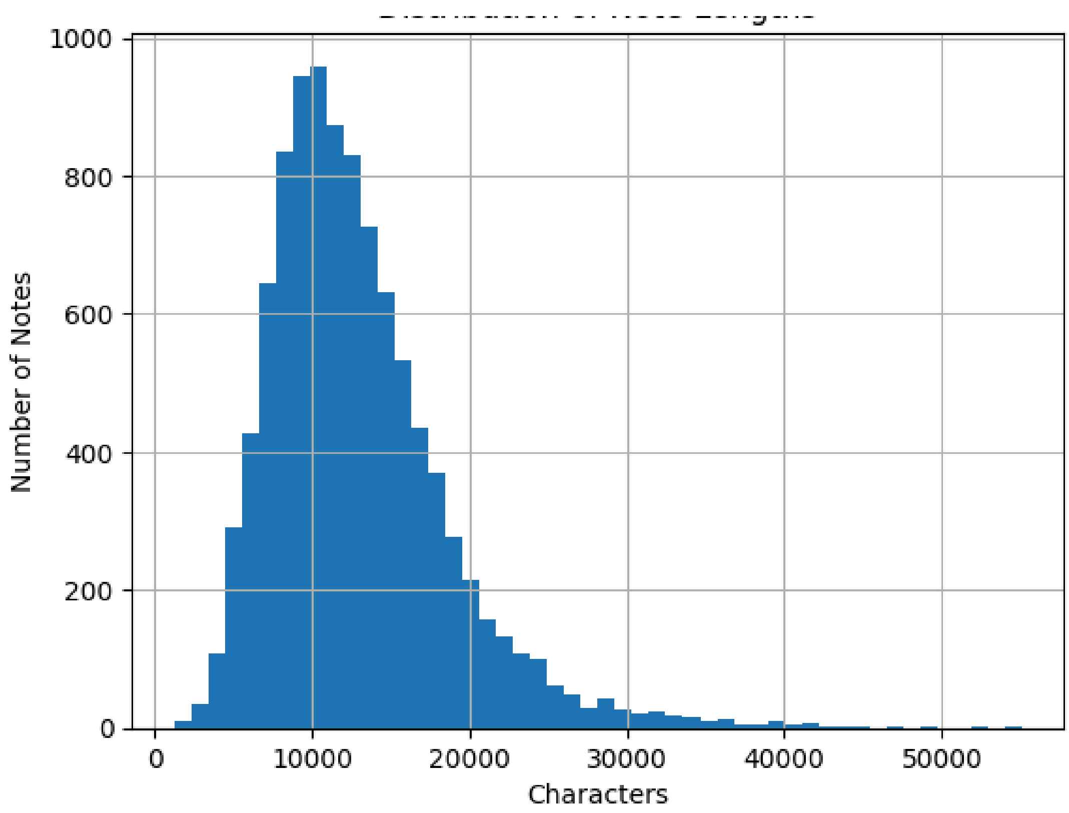
This histogram illustrates the distribution of hospital stay durations (in hours) among patients in the dataset. The data is heavily right-skewed, indicating that the majority of hospital stays are relatively short, with a large number concentrated below 500 hours. A sharp peak is visible around the lower durations, after which the frequency of stays decreases rapidly. A long tail extends beyond 1,000 hours, with a few outliers showing hospital stays exceeding 5,000 hours, suggesting occasional cases of extended hospitalization.

- *Distribution of ICU Stay Duration*



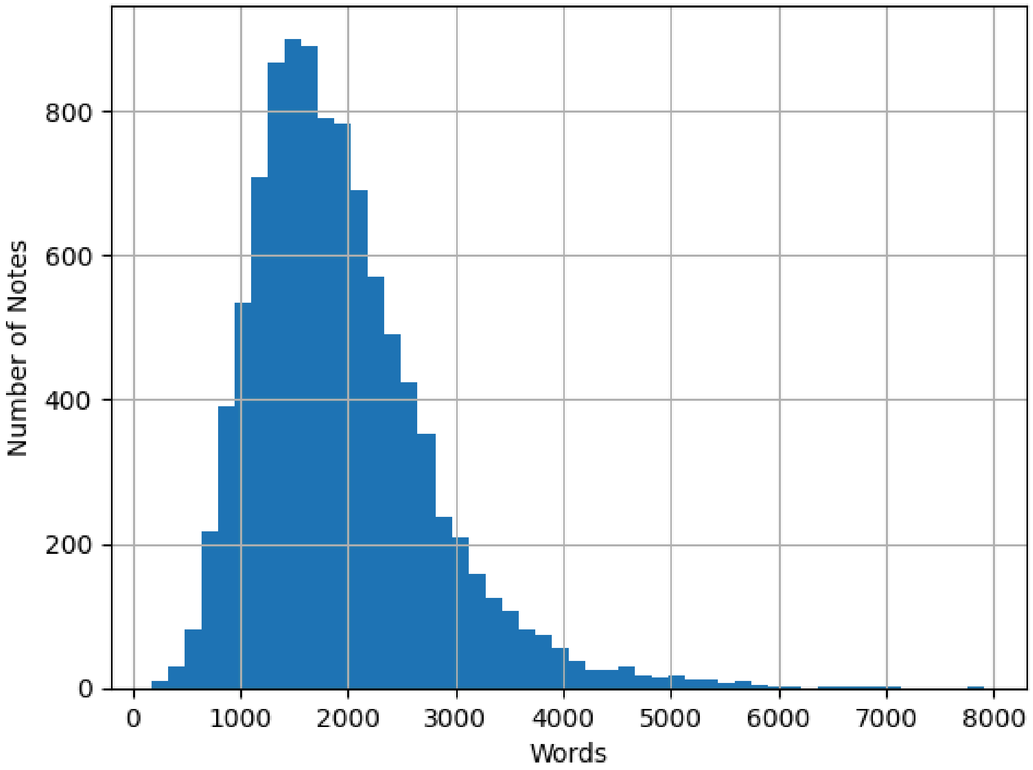
This histogram displays the distribution of ICU stay durations (in hours) for patients in the dataset. The distribution is highly right-skewed, with the majority of ICU stays being relatively short—most occurring under 100 hours. A pronounced peak is observed at the lower duration range, indicating that short-term ICU admissions are common. The frequency rapidly declines as the duration increases, with very few stays extending beyond 500 hours, and only rare outliers exceeding 1,000 hours. This pattern reflects typical ICU utilization where most patients require intensive care for brief, acute episodes.

- *Distribution of Note Lengths*



This histogram shows the distribution of clinical note lengths based on character count. The data is approximately right-skewed and unimodal, with most notes falling between 5,000 and 20,000 characters, and a clear peak around 10,000 characters. As note length increases beyond this range, the frequency declines steadily, though a long tail extends beyond 50,000 characters, indicating the presence of unusually long notes. Overall, this suggests that while most clinical documentation is of moderate length, a small subset of records are exceptionally verbose.

- *Distribution of Word Count*

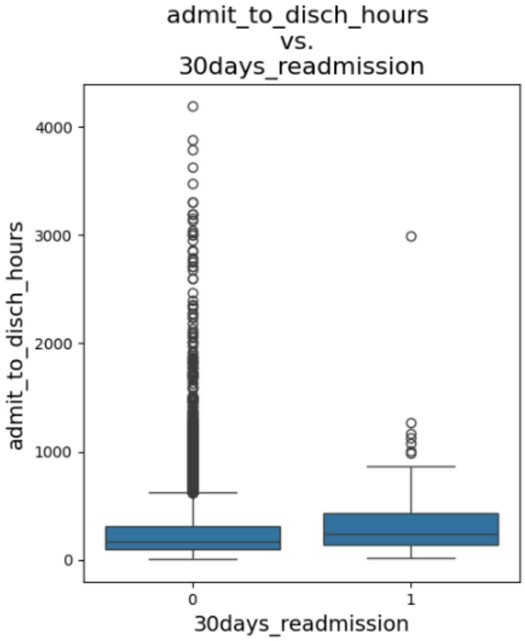


This histogram depicts the distribution of clinical note lengths based on word count. The data is unimodal and moderately right-skewed, with the majority of notes falling between 1,000 and 3,000 words, and the most frequent length occurring around 1,700–1,800 words. The number of notes decreases steadily as word count increases, with relatively few notes exceeding 5,000 words. This pattern indicates that most clinical documentation is moderately lengthy, while a small number of notes are exceptionally long and may reflect complex or extended hospitalizations.

3.3 Bivariate Analysis - Feature Relationships with readmission

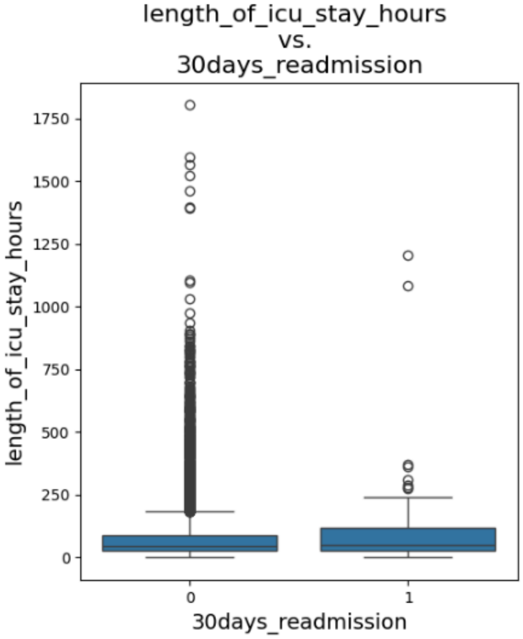
The box plots below illustrate the relationship between each feature and the target variable, 30-day readmission.

3.3.1 hospital duration vs. readmission



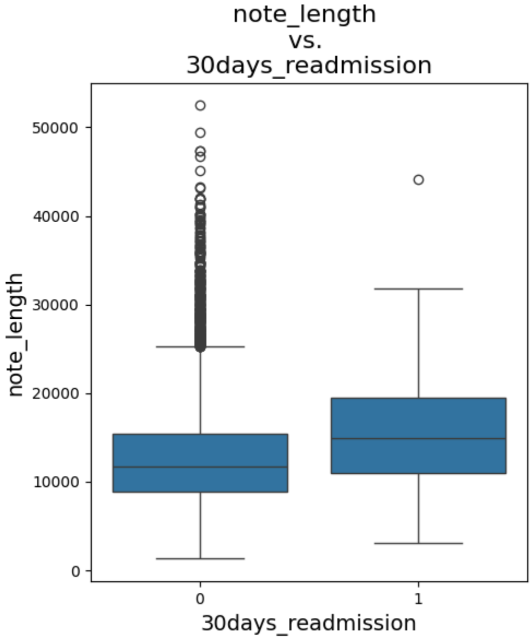
This boxplot compares hospital stay durations (in hours from admission to discharge) between patients who were not readmitted within 30 days (label 0) and those who were (label 1). Both groups have similar median stay durations, but the non-readmitted group shows a wider interquartile range and significantly more extreme outliers, with some stays exceeding 4,000 hours.

3.3.2 ICU stay vs. readmission



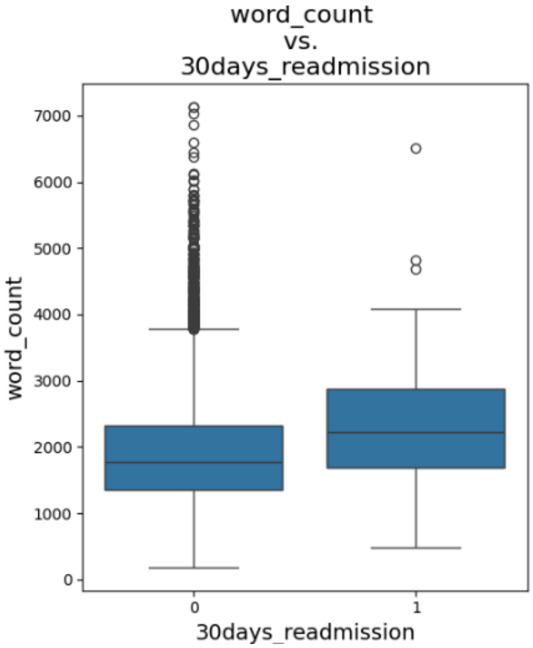
This boxplot compares ICU stay durations (in hours) between patients who were not readmitted within 30 days (category 0) and those who were (category 1). Both groups share a similar median ICU stay duration and interquartile range, indicating that the central tendency of ICU length is comparable regardless of readmission outcome. However, the non-readmitted group shows a larger number of extreme outliers, with several ICU stays extending beyond 1,500 hours, while the readmitted group has fewer high-duration outliers. Overall, there is no clear distinction in ICU stay length between the two groups, suggesting that ICU duration alone may not strongly differentiate readmission risk.

3.3.3 Clinical Note length vs. readmission



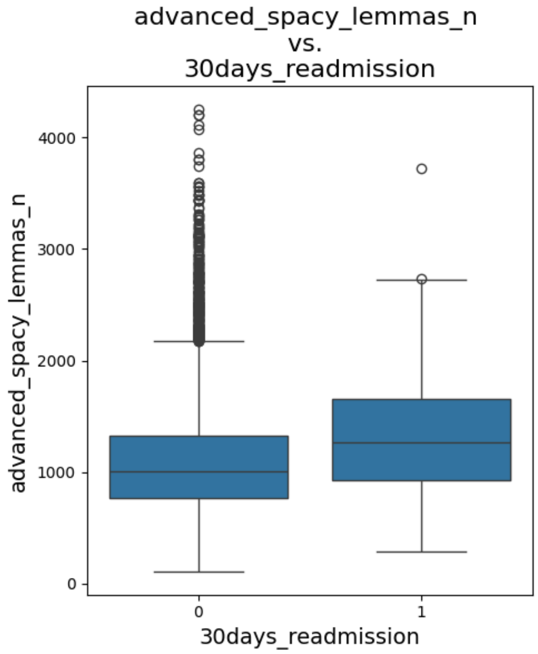
This boxplot shows the distribution of clinical note lengths (measured in characters) for patients grouped by 30-day readmission status. Patients who were readmitted (label 1) tend to have longer clinical notes on average, with a higher median and upper quartile compared to those who were not readmitted (label 0). The spread of note lengths is wider among readmitted patients, and while both groups contain outliers, extreme values are more densely concentrated in the non-readmitted group. This suggests that longer clinical documentation may be modestly associated with higher readmission risk, possibly reflecting greater clinical complexity or more severe conditions.

3.3.4 Clinical Notes word count vs. readmission



This boxplot compares the word count of clinical notes between patients who were not readmitted within 30 days (label 0) and those who were (label 1). The median word count is noticeably higher in the readmitted group, and the interquartile range also shifts upward, indicating that clinical notes tend to be longer for patients who experience readmission. While both groups show the presence of outliers, the non-readmitted group has a higher concentration of extremely long notes. These findings suggest that longer clinical notes—potentially reflecting more complex or severe cases—may be modestly associated with a greater likelihood of 30-day readmission.

3.3.5 Lemmas count of clinical Notes vs. readimission

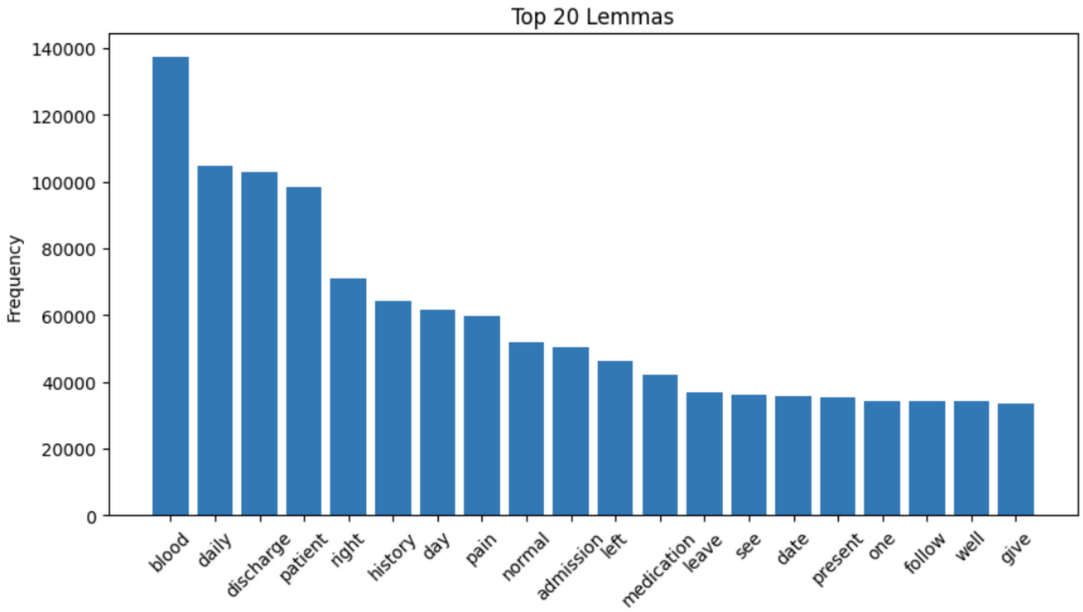


This boxplot compares the number of lemmatized tokens (using advanced spaCy processing) between patients who were not readmitted within 30 days (label 0) and those who were (label 1). The readmitted group shows a slightly higher median and wider interquartile range, suggesting that their clinical notes tend to contain more meaningful content words after linguistic preprocessing. Although both groups contain outliers, extreme values are more concentrated in the non-readmitted group. Overall, the distribution suggests a modest association between longer, content-rich documentation and the likelihood of 30-day readmission, possibly reflecting higher clinical complexity.

3.4 Text Preprocessing (Cleaning, Tokenization and Lemmatization)

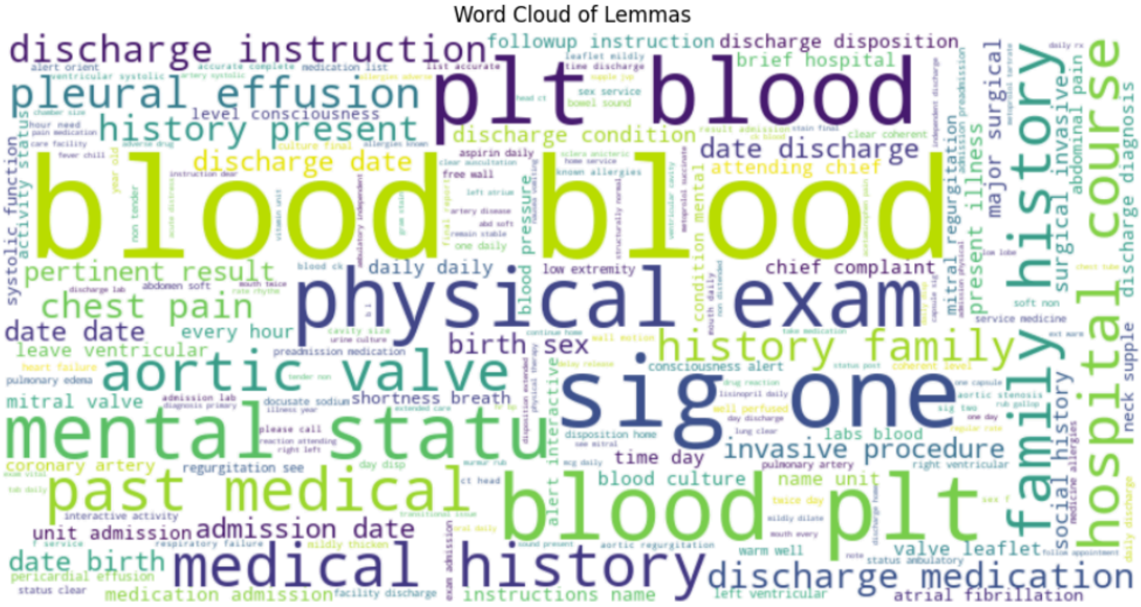
To prepare clinical notes for text analysis, a text preprocessing pipeline was implemented using spaCy's efficient batch processing capabilities. This pipeline included three key steps: cleaning, where irrelevant characters, punctuation, and common stopwords were removed to reduce noise; tokenization, which split the text into individual linguistic units (tokens); and lemmatization, which reduced each word to its base form to consolidate variations of the same concept (e.g., “running” → “run”). By applying spaCy’s NLP model in batches, the preprocessing was optimized for speed and scalability, enabling consistent and linguistically aware transformation of thousands of clinical documents into structured, analyzable formats.

3.4.1 Most Frequent Lemmas



This bar chart displays the top 20 most frequent lemmas extracted from clinical notes in the dataset. The lemma “blood” appears most often, followed by other common clinical terms such as “daily,” “discharge,” “patient,” and “right.” These frequently occurring lemmas reflect key aspects of clinical documentation, including procedures (e.g., discharge, medication), symptoms or conditions (e.g., pain, normal), and temporal or positional references (e.g., day, left, date). The consistent presence of such terms highlights the routine focus of clinical notes on patient monitoring, treatment, and physical status.

3.4.2 Word Cloud

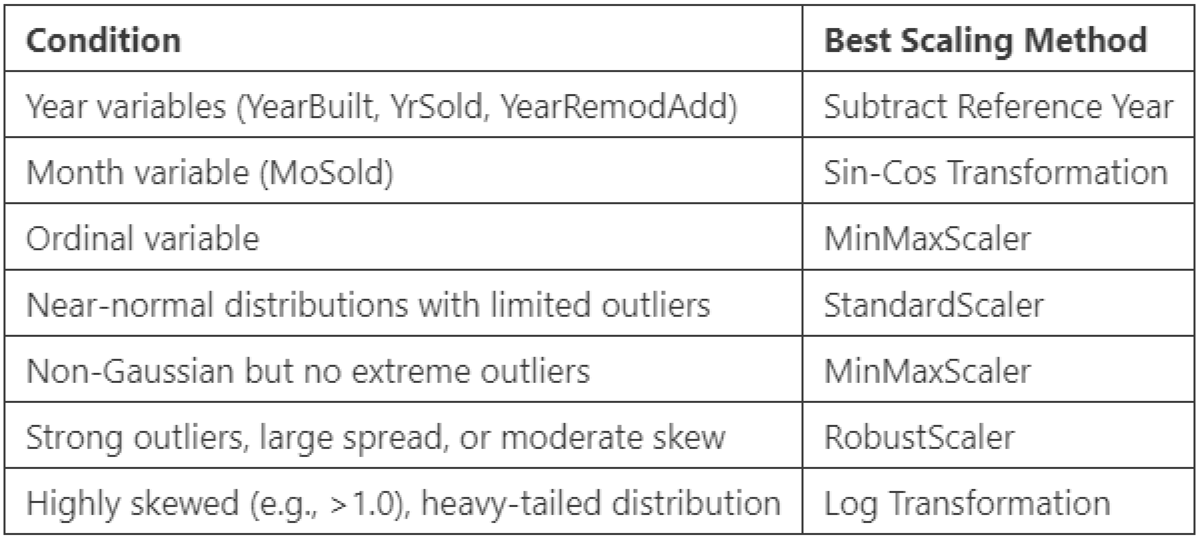


This word cloud visualizes the most frequent lemmas found in the clinical notes, where word size indicates relative frequency. Prominent terms such as “blood,” “plt,” “physical exam,” “mental status,” “medical history,” and “discharge” suggest common themes across documentation, including lab results, physical and mental assessments, and patient history. Other frequently observed phrases like “aortic valve,” “chest pain,” “medication,” and “hospital course” reflect typical medical concerns, procedures, and care routines. Overall, the word cloud highlights the central vocabulary used in patient evaluations and clinical decision-making processes.

4. Feature engineering

To represent the clinical notes in numerical form, two distinct text feature extraction techniques were applied. The first approach was TF-IDF, which generated a sparse representation of word importance across documents, resulting in a 2,000-dimensional feature vector. The second approach used ClinicalBERT, a contextualized language model pre-trained on medical corpora, to produce dense, semantically rich embeddings with a dimensionality of 768. These complementary methods capture different aspects of textual information—TF-IDF focuses on term frequency, while ClinicalBERT emphasizes contextual meaning.

In addition to text, five structured features were included to represent core clinical and textual metadata: length_of_icu_stay_hours, admit_to_disch_hours, note_length, word_count, and advanced_spacy_lemmas_n. These features capture key aspects of hospital stays and documentation volume. The appropriate scaling method for each feature was selected based on its distribution and statistical properties. The logic behind the scaling method selection is outlined below.



Condition	Best Scaling Method
Year variables (YearBuilt, YrSold, YearRemodAdd)	Subtract Reference Year
Month variable (MoSold)	Sin-Cos Transformation
Ordinal variable	MinMaxScaler
Near-normal distributions with limited outliers	StandardScaler
Non-Gaussian but no extreme outliers	MinMaxScaler
Strong outliers, large spread, or moderate skew	RobustScaler
Highly skewed (e.g., >1.0), heavy-tailed distribution	Log Transformation

All features were automatically assigned to appropriate scaling methods using a custom function that categorized them based on statistical thresholds and predefined ordinal classifications. As a result, all five features were grouped under log transformation due to their highly skewed, heavy-tailed distributions—an essential preprocessing step for managing the wide-ranging numeric values commonly observed in clinical datasets. After inspecting the distribution of each feature, the automatic classification appeared reasonable.

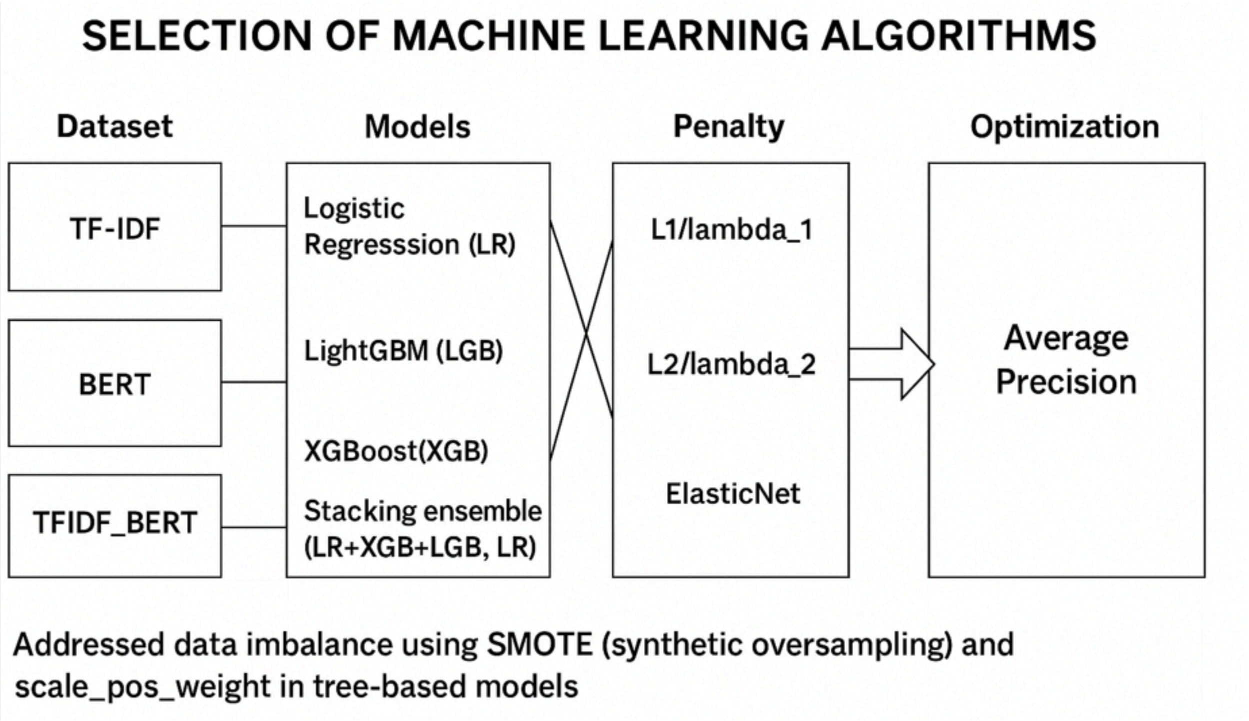
To explore the impact of combining structured and unstructured data, three hybrid datasets were constructed. The Tfidf dataset combined TF-IDF vectors with structured features, forming a 2,005-dimensional input space. The BERT dataset merged ClinicalBERT embeddings with structured features, yielding 773 dimensions. Finally, the most comprehensive dataset, Tfidf_BERT, integrated both TF-IDF and ClinicalBERT features alongside the structured data, resulting in a 2,773-dimensional representation. These combinations allowed for comparative evaluation of different feature fusion strategies in downstream modeling.

5. Modeling

5.1 Modeling and optimization

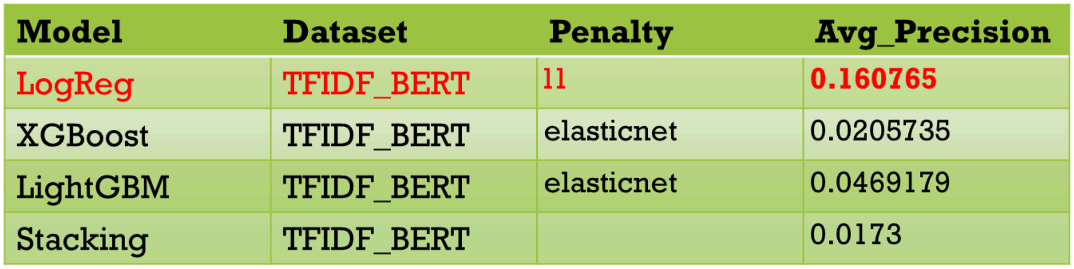
To identify the most effective model for predicting 30-day hospital readmission, I adopted a modular strategy that systematically explored combinations of feature sets, algorithms, penalties, and optimization criteria. Three types of input datasets were prepared: TF-IDF, ClinicalBERT embeddings, and their combination (TFIDF_BERT), each integrated with structured clinical features. For each dataset, I aimed to maximize average precision by evaluating a range of machine learning models, including Logistic Regression (LR) and tree-based algorithms such as LightGBM (LGB) and XGBoost (XGB), each configured with one of three regularization penalties: L1, L2, or ElasticNet. To further enhance performance, I implemented a stacking ensemble that combined LR, XGB, and LGB as base learners, with LR serving as the meta-learner—designed to leverage the complementary strengths of these individual models.

Model performance was optimized using Average Precision (AP) as the primary evaluation metric, aligning with the project's focus on imbalanced classification. To address the extreme class imbalance in the dataset, I applied SMOTE (Synthetic Minority Oversampling Technique) and scale_pos_weight adjustment. This comprehensive strategy enabled fair comparisons across model families and dataset variants while ensuring robustness in handling rare positive cases.


SELECTION OF MACHINE LEARNING ALGORITHMS
Dataset
Models
Penalty
Optimization
TF-IDF
BERT
TFIDF_BERT
Logistic Regresssion (LR)
LightGBM (LGB)
XGBoost(XGB)
Stacking ensemble (LR+XGB+LGB, LR)
L1/lambda_1
L2/lambda_2
ElasticNet
Average Precision
Addressed data imbalance using SMOTE (synthetic oversampling) and scale_pos_weight in tree-based models

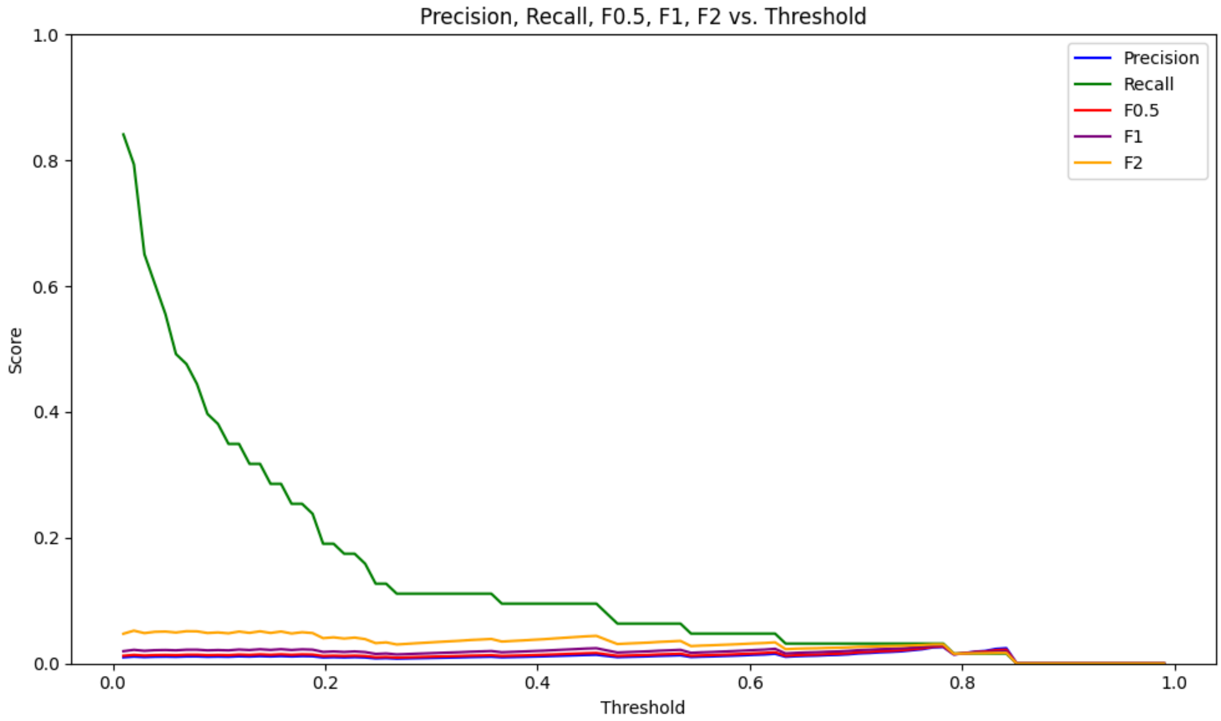
5.2 Model performance Summary

The performance and predictive accuracy of each model were evaluated. The results summarized in this table show that Logistic Regression with L1 regularization on the TFIDF_BERT dataset achieved the highest performance, yielding an Average Precision (AP) of 0.1608, significantly outperforming all other models. In comparison, XGBoost and LightGBM with ElasticNet penalty produced much lower AP scores of 0.0206 and 0.0469, respectively. Surprisingly, the stacking ensemble, despite integrating LR, XGB, and LGB as base learners, achieved the lowest AP at 0.0173. These findings highlight the effectiveness of sparse linear modeling (L1) in high-dimensional settings and suggest that more complex models may not generalize well under extreme class imbalance.



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

To better understand the trade-off between false positives and false negatives in our highly imbalanced hospital readmission prediction task, a precision-recall threshold analysis was conducted for the best-performing model (Logistic Regression with L1 penalty on TFIDF_BERT). Rather than using the default threshold of 0.5, this approach evaluates performance across a range of thresholds from 0 to 1. By calculating and plotting precision, recall, F0.5, F1, and F2 scores at each threshold, we can identify an optimal decision point aligned with clinical or operational goals—whether it be minimizing missed readmissions (recall-focused), reducing unnecessary interventions (precision-focused), or balancing both.

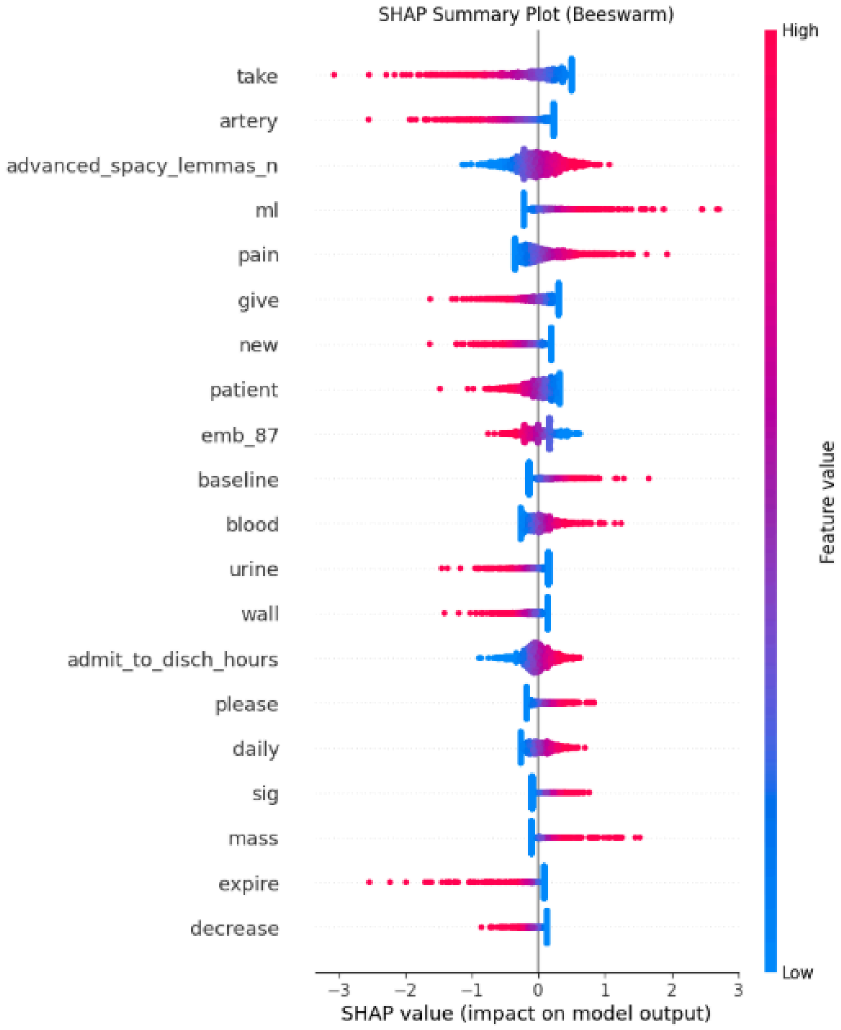


The precision-recall threshold analysis reveals hallmark patterns of extreme class imbalance and weak model calibration. The precision curve remains consistently low—below 0.05 across most thresholds—meaning that even at the best points, over 95% of predicted positives are false alarms. Recall, while initially perfect at a threshold of 0.0, drops sharply as the threshold increases, indicating that the model can only catch true positives when it is extremely lenient. Meanwhile, the F-scores (F0.5, F1, F2) remain near zero across the board, with F2 performing slightly better due to its emphasis on recall, but still offering no viable operating point above 0.1. This suggests the model is unable to maintain a useful balance between false positives and false negatives, no matter the threshold.

Overall, the pattern reflects a model struggling under the weight of a highly imbalanced dataset—likely with a ~1% positive rate—where rare true signals are drowned out by abundant noise. While the model does exhibit some ability to rank examples (as indicated by an average precision of ~0.16), its predicted probabilities are poorly calibrated: most true positives receive low scores, indistinguishable from negatives. This may stem from limitations in the feature space (e.g., TF-IDF + BERT + structured features may not capture meaningful discriminative signals) or from the imbalance itself overpowering the learning signal. Additionally, this threshold tuning process has limitations. It was based on predictions from cross_val_predict() on the training set, which may not fully reflect the model’s generalization ability. To move forward, strategies like cost-sensitive learning, refining feature selection, reweighting, calibration tuning, domain-specific feature engineering or leveraging temporal or sequence-based features from EHR data are needed before the model can be considered reliable for deployment. Finally, external validation on a truly unseen test set and a deeper error analysis will help assess the real-world applicability of the chosen threshold.

5.4 Feature Importance under the best model

To interpret the predictions of the selected best model—Logistic Regression with L1 regularization on the TFIDF_BERT dataset—and to understand the influence of each input feature, I employed SHAP (SHapley Additive exPlanations) values. Rooted in cooperative game theory, SHAP assigns each feature a contribution value for individual predictions, providing a consistent and theoretically grounded measure of feature importance. Unlike traditional methods that often rely on model-specific heuristics, SHAP offers a unified, model-agnostic framework that ensures local accuracy and interpretability. In this analysis, SHAP was used to quantify and visualize the impact of each feature on model output, delivering clearer insights into how the model makes predictions and which features drive its behavior.



This SHAP summary plot (beeswarm) visualizes the impact of the top 20 features on the predictions made by the best-performing model—Logistic Regression with L1 penalty on TFIDF_BERT features. Each dot represents a SHAP value for a feature in one sample, with color indicating the feature's actual value (red = high, blue = low). Features toward the top have the greatest overall influence. Among the most impactful and reasonable features are clinical concepts such as “pain,” “artery,” and “blood,” which align with the model’s goal of predicting adverse outcomes like hospital readmission. Structured features like “admit_to_disch_hours” (hospital stay duration) and “advanced_spacy_lemmas_n” (note complexity or verbosity) also appear prominently, reinforcing their interpretability. Additionally, “emb_87”, a ClinicalBERT embedding dimension, likely encodes high-level semantic signals learned from clinical text and may capture underlying patterns relevant to patient status or clinical events.

On the other hand, the presence of less meaningful features such as “take,” “ml,” “give,” “new,” “patient,” “please,” and “baseline” suggests some overfitting or noise sensitivity in the model. These terms are generic and frequently found in routine clinical documentation (e.g., "take medication," "give 5 ml," "new patient admitted"). Their high SHAP influence may not reflect true clinical causality but rather correlation artifacts—for example, boilerplate phrases disproportionately present in positive cases. Such spurious patterns are common in imbalanced datasets, where neutral or procedural language can become associated with the minority class simply by co-occurrence. While SHAP reveals strong associations, it does not guarantee semantic relevance; some of these features may be proxies for systemic documentation habits rather than meaningful medical indicators. Future improvements may include removing boilerplate text, adding domain-specific filters, or using techniques like causal inference or attention maps to better isolate clinically relevant signals.

6. Limitations, Challenges and Future Work

This study faced several key limitations stemming from cohort definition and data filtering choices. First, by focusing solely on ICU admissions and requiring complete clinical notes—compounded by the mistaken use of INNER JOIN in SQL filtering—the cohort likely excluded many valid hospitalizations lacking associated notes or ICU stays, substantially shrinking the dataset and limiting the model’s generalizability. Furthermore, the pipeline did not exclude patients who died post-discharge, neonates, or cases with missing timestamps—all of which may have biased the readmission label assignment and skewed cohort composition. These combined limitations likely contributed to the low prevalence of 30-day readmissions in the final dataset and weakened the model’s discriminative power.

To address these issues, future work should begin with refining cohort construction—broadening inclusion criteria to cover all hospitalization events, not just ICU encounters. Better filtering logic (e.g., LEFT JOINs, exclusion flags for death, age filters) can help ensure clinical relevance and cleaner outcome labeling. Incorporating richer predictive features such as the Charlson Comorbidity Index, diagnostic codes, lab values, and vital signs could significantly boost model performance by capturing the underlying health status and treatment trajectory of each patient—factors directly linked to readmission risk.

On the modeling side, additional gains could come from transforming and normalizing structured features to better expose relationships during learning. Further, using temporal patterns, sequence-aware models, or ensemble methods with calibrated thresholds may offer stronger performance in the face of class imbalance and heterogeneous documentation. Overall, addressing both the data quality and feature representation gaps will be critical for improving predictive accuracy and translating this work into a robust tool for clinical decision support.

7. Conclusion

This project established a baseline machine learning framework to predict 30-day hospital readmissions using an adult ICU cohort from the MIMIC-IV dataset. The analysis was challenged by an extremely imbalanced target variable, with a readmission prevalence of only 0.96%. Despite applying class-imbalance strategies such as SMOTE and class weighting, and testing multiple models—including stacked ensembles—the signal remained weak.

The best-performing model, a Logistic Regression with L1 regularization on combined TF-IDF, ClinicalBERT, and structured features, achieved an average precision of 0.16. Precision-recall analysis highlighted persistent challenges with high false positive rates and low F-scores under severe imbalance. SHAP analysis revealed that key predictive drivers included terms from clinical notes (e.g., “pain,” “artery,” “blood”) and length-of-stay metrics, which may inform early discharge risk assessment.

However, several data limitations constrained model performance, including the restriction to ICU admissions, reliance on complete clinical notes, and the exclusion of post-discharge deaths and external readmissions—all of which significantly reduced the number of positive cases. Future work should focus on expanding the cohort to include all hospitalizations, integrating more comprehensive clinical features such as diagnoses, laboratory results, and comorbidity indices, and enhancing modeling strategies to improve recall and maximize real-world clinical utility.