BTU Cottbus-Senftenberg

Brandenburgische Technische Universität Cottbus - Senftenberg

Patient Readmission Prediction using H2O AutoML

Natasha Shereen Benita
Guide: Prof. Dr. Ireneusz Jablonski

Introduction

Problem Statement:

- Hospital readmission prediction is critical for resource allocation, patient care, and cost reduction.

Objective:

- Use H2O AutoML to automate model selection and improve prediction accuracy.

Key Result:

- Best-performing model (GBM) achieved **99% accuracy, AUC = 1.0**.

Literature Review

Traditional Approaches:

- Logistic Regression, Decision Trees – struggled with complex medical data.

Advancements:

- ML models like **XGBoost, Random Forests, GBM** improve predictions.

Why AutoML?

- Automates model selection & tuning, making healthcare predictions more accessible.

Dataset & Preprocessing

Dataset: eICU Collaborative Research Database (PhysioNet).

Key Tables Used: Patient, AdmissionDX, Diagnosis, Lab, Treatment.

Preprocessing Steps:

- Removed duplicates, optimized data types, merged tables.
- Dropped **medication table** due to high memory usage.
- Final dataset: **1000 rows, 14 columns**.

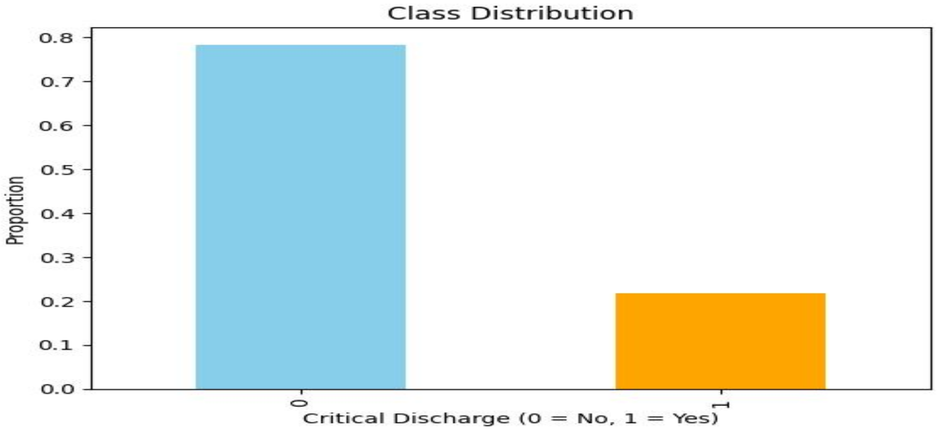
Feature Engineering & EDA

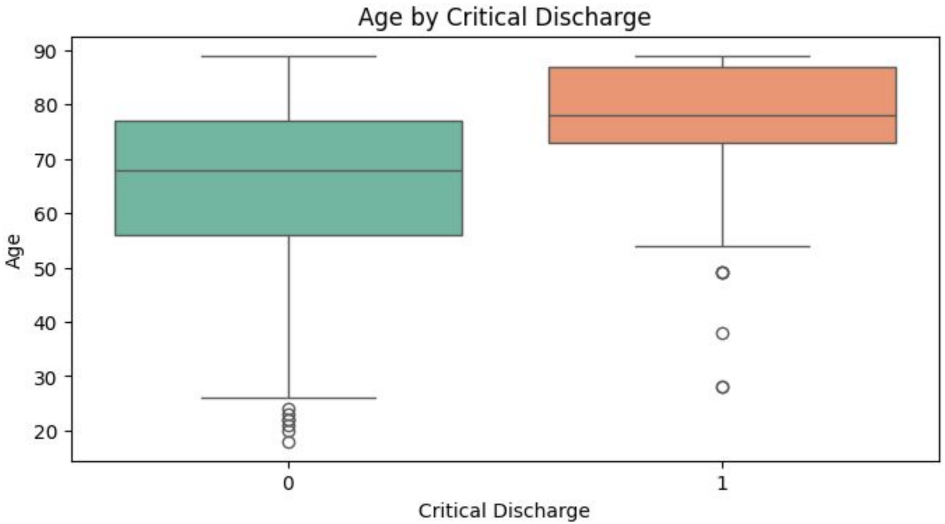
New Features Created:

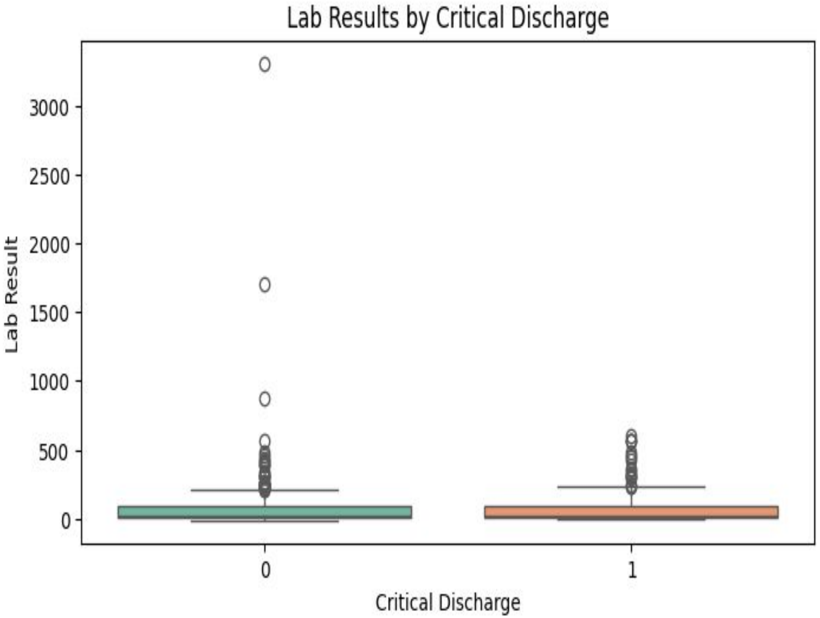
- **Age Groups, Diagnosis Length, Gender Encoding, Critical Discharge Indicator.**

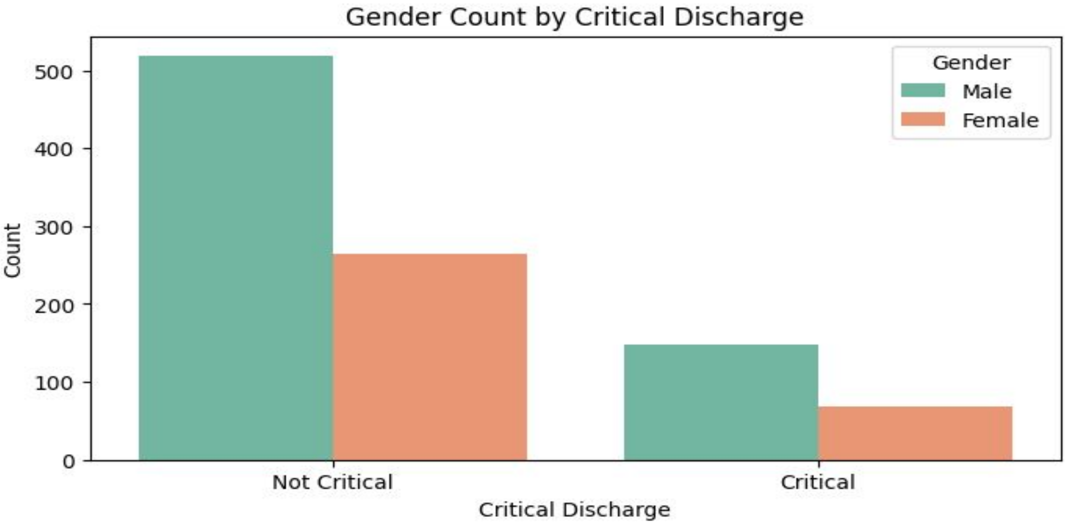
Target variable

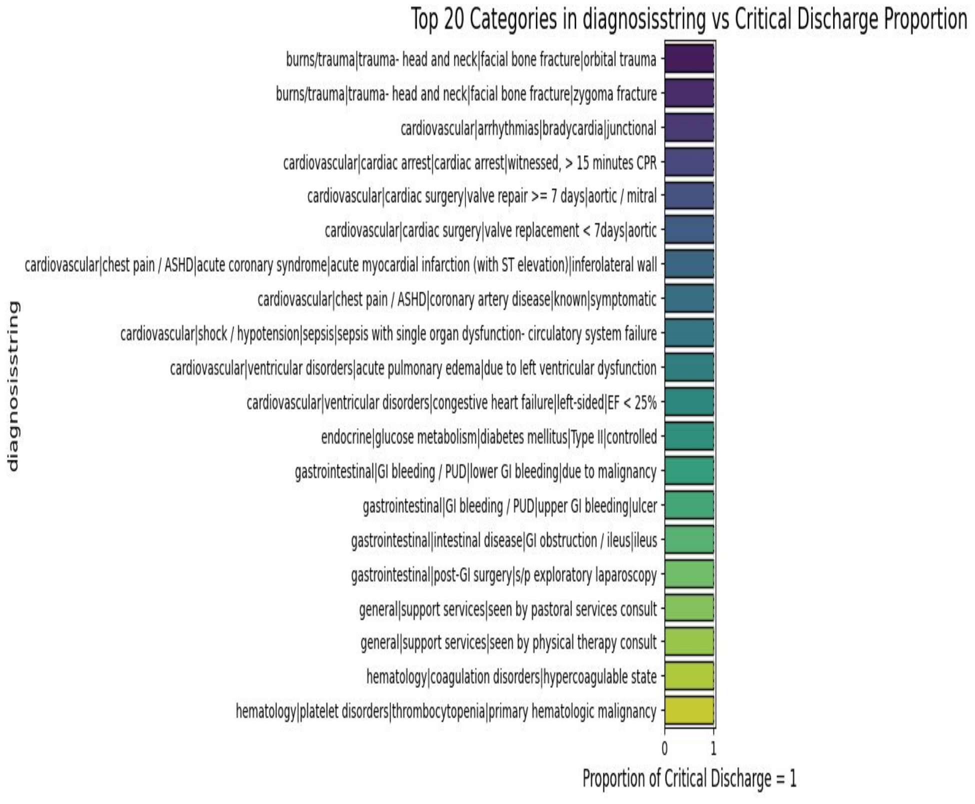
- **Research has also shown that discharge outcomes are strong indicators of readmission risk**
- **Binary feature**
- **Early identification of high-risk patients to enhance post-discharge care and reduce readmission rates.**
- **Imbalanced distribution (78.3% non-critical, 21.7% critical)**

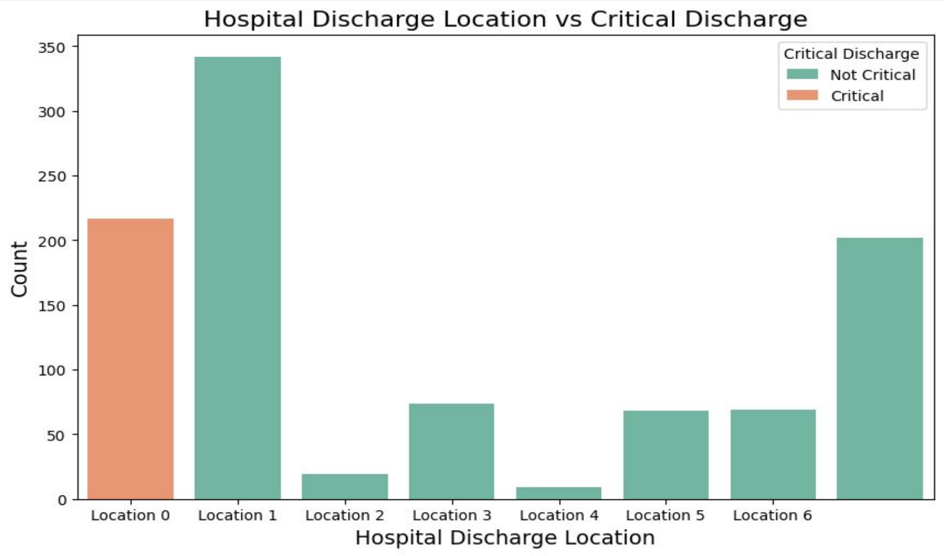

Class Distribution
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
Critical Discharge (0 = No, 1 = Yes)


Age by Critical Discharge
Age
20
30
40
50
60
70
80
90
0
1
Critical Discharge


Lab Results by Critical Discharge
Lab Result
0
500
1000
1500
2000
2500
3000
0
1
Critical Discharge


Gender Count by Critical Discharge
Count
0
100
200
300
400
500
Gender
Male
Female
Not Critical
Critical
Critical Discharge





burns/trauma-related diagnoses) have the highest association with readmission

Model Training & Results

- **Best Model: Gradient Boosting Machine (GBM).**

Performance Metrics:

- **AUC:** 1.0
- **Log Loss:** 3.459
- **Optimal Threshold:** 0.993 (for best F1-score).

model_id	rmse	mse	mae	rmsle	mean_residual_deviance
GBM_grid_1_AutoML_1_20250120_172333_model_17	1.39721e-07	1.9522e-14	7.27513e-08	7.05875e-08	1.9522e-14
GBM_grid_1_AutoML_1_20250120_172333_model_19	1.39721e-07	1.9522e-14	7.27513e-08	7.05875e-08	1.9522e-14
GBM_grid_1_AutoML_1_20250120_172333_model_2	1.39721e-07	1.9522e-14	7.27513e-08	7.05875e-08	1.9522e-14
GBM_grid_1_AutoML_1_20250120_172333_model_13	1.39721e-07	1.9522e-14	7.27513e-08	7.05875e-08	1.9522e-14
GBM_grid_1_AutoML_1_20250120_172333_model_1	3.30671e-07	1.09343e-13	8.27802e-08	3.11002e-07	1.09343e-13
GBM_grid_1_AutoML_1_20250120_172333_model_16	0.000634837	4.03018e-07	0.000277034	0.000408196	4.03018e-07
StackedEnsemble_AllModels_1_AutoML_1_20250120_172333	0.00273976	7.50627e-06	0.00156764	0.00172258	7.50627e-06
GLM_1_AutoML_1_20250120_172333	0.0034585	1.19612e-05	0.00285516	0.00282503	1.19612e-05
StackedEnsemble_BestOfFamily_1_AutoML_1_20250120_172333	0.00374225	1.40044e-05	0.00224006	0.00249188	1.40044e-05
GBM_2_AutoML_1_20250120_172333	0.0110038	0.000121084	0.00605146	0.00810294	0.000121084

[30 rows x 6 columns]

```
AUC: 1.0
Logloss: 3.459647956244093e-05
Confusion Matrix:
Confusion Matrix (Act/Pred) for max f1 @ threshold = 0.9931157914848616
       0    1    Error    Rate
-----  ---  ---  -------  ----------
0      157  0    0        (0.0/157.0)
1      0    43   0        (0.0/43.0)
Total  157  43   0        (0.0/200.0)
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       157
           1       1.00      0.98      0.99        43

    accuracy                           0.99       200
   macro avg       1.00      0.99      0.99       200
weighted avg       1.00      0.99      0.99       200
```

Variable Importances:

variable	relative_importance	scaled_importance	percentage
hospitaldischargestatus	577.2741089	1.0	0.4031135
hospitaldischargelocation	329.3840332	0.5705851	0.2300106
diagnosisstring	116.7477036	0.2022396	0.0815255
treatmentstring	67.1490936	0.1163210	0.0468906
admitdxtext_length	58.7481613	0.1017682	0.0410241
age_group	53.1897964	0.0921396	0.0371427
hospitaladmitsource	47.4351959	0.0821710	0.0331242
patientunitstayid	46.5215569	0.0805883	0.0324862
ethnicity	38.9848671	0.0675327	0.0272233
labresult	33.5315590	0.0580860	0.0234153
age	28.8300037	0.0499416	0.0201321
icd9code	23.4633026	0.0406450	0.0163845
is_female	10.7791386	0.0186725	0.0075271

Why GBM is the best?

Boosting Reduces Errors Iteratively:
GBM builds trees sequentially, correcting mistakes from previous iterations, making it more robust for structured medical data.

The dataset has **78.3% non-critical vs. 21.7% critical discharges**.
GBM can handle imbalance effectively by assigning higher weights to minority-class samples.

Performs Well with Mixed Data Types:

H2O AutoML **tunes hyperparameters** automatically, finding the best GBM settings for your data.

Future Work

- XGBoost can be included: XGBoost **requires more memory** than H2O’s native GBM. If AutoML detects potential memory issues, it **disables XGBoost automatically**.
- Check for overfitting
- Data related to expired people can be removed when working with the whole dataset as it will form a significant group. However, for this study it was not excluded since the data suggests otherwise.

(From the data we can see that some patients are alive and are not critically discharged even though they are categorized under deadly diseases while some others die from minor infections. If removed from this subset valuable information related to other variables and features will be lost. Overall, the numbers correspond to the class imbalance in the target column.)