MIMIC-III ICU Admission ML Models Report

Introduction

Using the MIMIC-III Clinical Database,

> “a collection of deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (including post-hospital discharge)”,

I tested a variety of 6 different machine learning predictive models to identity whether or not (yes/no) each hospital visit resulted in the patient being admitted into an intensive care unit (ICU).

> ⚠️ Note: the MIMIC-III data in **intentionally** heavily imbalanced towards ICU patient admission (98% of visits resulted in admission, 2% did not)

Input parameters included:

- Medical Diagnosis
- Gender
- Age
- Marital Status
- Ethnicity
- Vitals from the first 24 hours of hospital admission:
    - Heart Rate

    - Systolic Blood Pressure
    - Diastolic Blood Pressure
    - Body Temperature
    - Respiratory Rate
    - Blood Oxygen Level (SpO2)

Output prediction:

- ICU Admission (yes/no)

Methods

After engineering my input parameter vector from various CSV files from the MIMIC-III database, I used a 80/20 training/test data split to evaluate the performance of 6 different machine learning models:

1. Logistic Regression
2. Random Forest
3. Gradient Boosting
4. Support Vector Machine (SVM)
5. Naive Bayes
6. K-Nearest Neighbors (KNN)

Each model's performance was evaluated using 9 various metrics:

1. Accuracy
2. Precision
3. Recall
4. F1 Score
5. ROC-AUC
6. PR-AUC
7. Specificity

8. Training Time
9. Prediction Time

Link to the full process Python Notebook

Results

Overall best model: Gradient Boosting

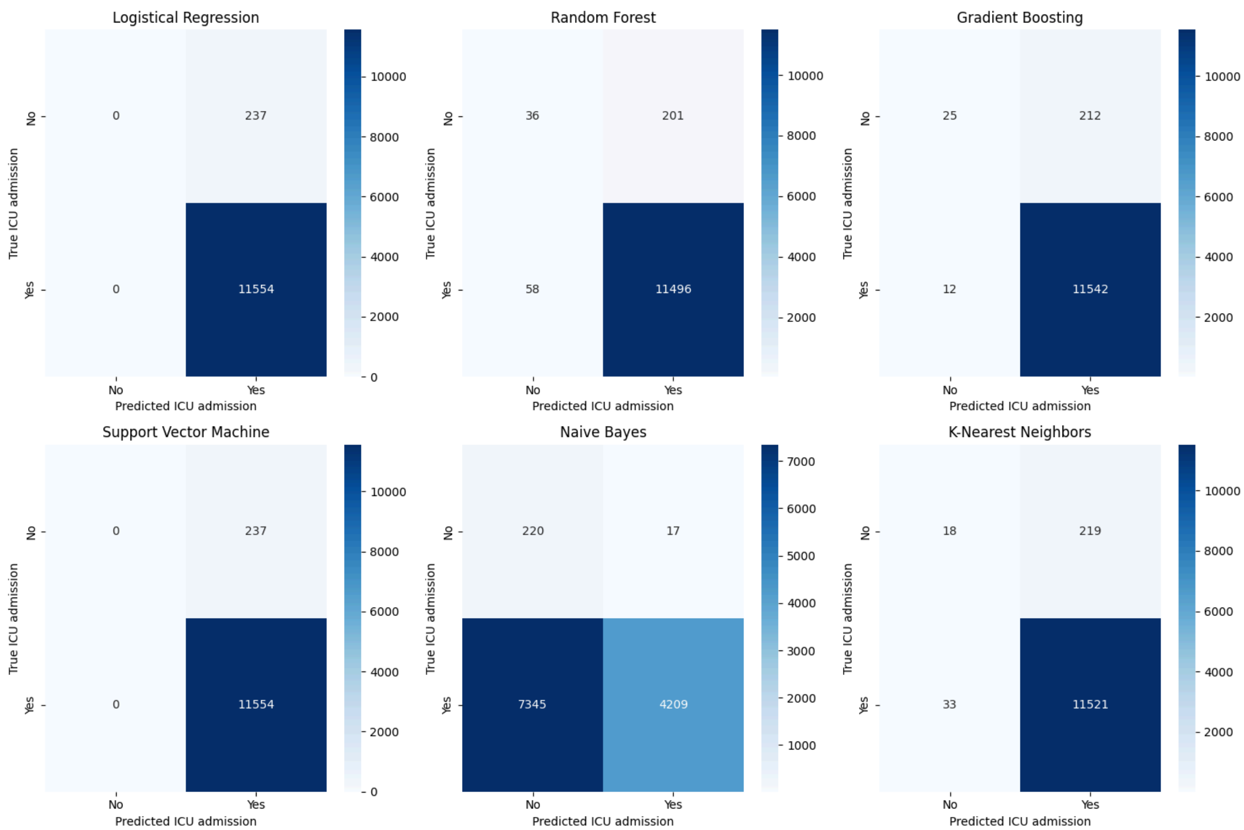
Gradient boosting proved to be the most accurate model. Furthermore, it posed the greatest ability to distinguish ICU admission for each patient by capturing non-linear patterns in diagnosis and vitals despite taking into account the severely imbalanced data (Highest PR-AUC score). PR-AUC is a key metric in this data study because ROC-AUC can be misleading in data as imbalanced as this set is. The 26 second training time indicates slower, but not detrimental time needed to train.

	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	PR-AUC	Specificity	Training Time (s)	Prediction Time (s)
0	Logistical Regression	0.979900	0.979900	1.000000	0.989848	0.870311	0.996443	0.000000	0.877320	0.017896
1	Random Forest	0.978034	0.982816	0.994980	0.988861	0.869137	0.995357	0.151899	5.948558	0.232902
2	Gradient Boosting	0.981002	0.981964	0.998961	0.990390	0.927164	0.997738	0.105485	25.785991	0.028031
3	Support Vector Machine	0.979900	0.979900	1.000000	0.989848	0.842691	0.995746	0.000000	193.959173	9.005790
4	Naive Bayes	0.375625	0.995977	0.364289	0.533460	0.853739	0.995025	0.928270	0.058352	0.024841
5	K-Nearest Neighbors	0.978628	0.981346	0.997144	0.989182	0.691984	0.987583	0.075949	0.004389	3.099962

Key insights on each of the other models:

- **Random Forest**: Best specificity among competitive models, good for identifying true non-ICU cases, though recall remains the priority in this setting
- **Logistic Regression**: Fast baseline but limited by linear assumptions
- **SVM**: Too slow (194s training) for marginal performance gain
- **KNN**: Poor generalization (ROC-AUC: 0.692), doesn't scale well to high-dimensional sparse features
- **Naive Bayes**: High specificity but misses 64% of ICU patients, unacceptable for clinical use

Additionally, accuracy by model can be visualized via a confusion matrix with Predicted Value on the X axis, and True Target Value on the Y axis. It is worth noting that the confusion matrix shows that Logistic Regression and SVM did not make a “no” prediction a single time. This indicates that the models choose to exploit their knowledge of the imbalanced data rather than making their prediction based on discriminating the data.



Conclusion

In the specific clinical setting where an ICU transfer decision is needed and the diagnosis + first 24 hours of vitals data is available, a Gradient Boosting machine learning model can be used to augment this assessment. Hospitals with high volumes of critically ill patients could implement a similar model to improve the determination of needing a Rapid Response Team. Future work on more balanced datasets, such as the forthcoming Children's Health study, is expected to yield more generalized results, and I look forward to presenting those findings.

References

Data access from PhysioNet. Data engineering, machine learning modeling, and analysis done by Griffin Kuchar (TCU Computer Science, 2027) under Dr. Robin Chataut.

Sources:

- Johnson, Alistair, et al. "MIMIC-III Clinical Database" (version 1.4). PhysioNet (2016). RRID:SCR_007345.
- Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825–2830 (2011).
- Harris, C.R., et al. "Array programming with NumPy." *Nature*, 585, 357–362 (2020).
- McKinney, Wes. "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference* (2010).
- Hunter, J.D. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, 9(3), 90–95 (2007).
- Waskom, M.L. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software*, 6(60), 3021 (2021).
- Van Rossum, G. & Drake, F.L. *Python 3 Reference Manual*. Python Software Foundation (2009). *(ast — Python standard library)*
- Virtanen, P., et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods*, 17, 261–272 (2020).