**Capstone 2 Report 2**

Team 6

Duyen Do

Lakshmi Kambathanahally Lakshminarasap

Jessica Nguyen

Shrinidhi Sudhir

BZAN 6361: Capstone II

April 19th, 2024

Table of Contents

[Executive Summary and Recommendations 3](#_Toc164457202)

[Reflections 4](#_Toc164457203)

[Strengths from previous findings: 4](#_Toc164457204)

[Opportunities for improvement: 4](#_Toc164457205)

[Analysis 5](#_Toc164457206)

[Understand the target variable “HasCLABSI” 5](#_Toc164457207)

[Addressing missing values in the dataset based on data type and correlation with the target variable 5](#_Toc164457208)

[Reduce dimensionality with correlation matrix and pattern matching 6](#_Toc164457209)

[Refine and enhance the target variable to improve the prediction of CLABSI within 3 days 6](#_Toc164457210)

[Modeling Design 7](#_Toc164457211)

[Neural Networks (NN) 7](#_Toc164457212)

[k-Nearest Neighbors (kNN) 7](#_Toc164457213)

[Naïve Bayes 8](#_Toc164457214)

[Decision Tree 9](#_Toc164457215)

[Evaluation Plan 10](#_Toc164457216)

[Overcome overfitting and data leakage. 10](#_Toc164457217)

[Overcome imbalanced data 11](#_Toc164457218)

[Metrics to evaluate models’ performance: 11](#_Toc164457219)

[Performance adjustment within algorithms based on models’ inherent biases: 12](#_Toc164457220)

[Further on Recommendations 14](#_Toc164457221)

[The trade-off between type 1 and type 2 errors 14](#_Toc164457222)

[The practicality of a cost matrix 14](#_Toc164457223)

# Executive Summary and Recommendations

**Challenge:**

Central Line-Associated Bloodstream Infections (CLABSI) remain a prominent life-threatening threat to PICU patients. While timely and early detection is critical, vast amounts of data and limited adaptation of machine learning methods in healthcare have revealed shortcomings in this area.

**Recommendations:**

* Utilizing a cost-matrix, a cost-sensitive model, and weighted accuracy can help decision makers optimize the results to fit the clinical setting.
* Healthcare providers should consider models that minimize type 2 errors to save lives while preserving the institutions’ quality ratings and reputation.

**Key findings:**

* Advanced and tailored approaches are necessary to develop predictive machine learning models. This tailored approach is implemented throughout the process, from data-preprocessing to model evaluation.
* Four models were examined, each with its own strength and inherent biases that require further adjustments.
* Decision tree models exhibit the most balanced performance across crucial metrics.

Some additional solutions that are worth considering to incorporate and complement the current approach:

* Transferred learning: leverage pretrained models for effectiveness
* Multiple models: employ several different models at distinct stages of the process.

# Reflections

*Summary: Advanced data analysis and preprocessing significantly improve model performance and outcomes.*

## **Strengths from previous findings**:

### ***Thorough data preprocessing and strategic planning for later stages:***

**A meticulous approach to data preparation allowed us to address multiple challenges within the dataset, such as sparse data and feature engineering. The team recognizes the importance of this foundational step, as it strengthens the impact and effectiveness of our models in later stages.**

### ****Varied** *data analysis techniques, including advanced statistical modeling, show proficiency***:

The team effectively utilized a variety of models from Python's libraries (Pandas, NumPy, Seaborn, Matplotlib, Statsmodels, and linearmodels) to handle a broad spectrum of data analysis tasks. Additionally, we implemented advanced statistical models like PanelOLS to uncover data relationships. This broad array of advanced models allows us to confidently draw data-driven decisions and recommendations based on in-depth analysis.

### ***Effective data visualization to communicate findings in a business context:***

**Analysts successfully implemented visualization tools like Seaborn and Matplotlib to communicate technical findings in a business context. These visualizations effectively conveyed the direction and scope of the complex analysis to clients.**

## ****Opportunities for improvement:****

### ***Explore and employ explicit data validation and dynamic data handling measures:***

**The team will investigate and implement additional methods for data validation to address data quality issues encountered. Additionally, considering the dynamic nature of healthcare factors and patient data, we aim to transition and broaden our model from descriptive analysis to more adaptable models that can handle dynamic data attributes, thereby improving scalability.**

### ***Optimize queries and documentation to improve clarity:***

**There are opportunities to streamline our queries with additional comments. This will allow easier model fine-tuning and will help collaborators grasp the scope of our script.**

# Analysis

*Summary: Further data preparation and a deeper understanding of the target variable led to improved prediction of CLABSI within a 3-day period.*

The analysis thoroughly examines the project's challenges across key domains after dynamically handling sparse data based on the target variable. It focuses on understanding the target variable, addressing missing values, developing new methods to reduce dimensions, and refining the target variable to enhance the prediction of CLABSI risk within a three-day window.

## Understand the target variable “HasCLABSI”

The target variable, “HasCLABSI” attribute in the data frame, is a binary categorical variable indicating whether a patient has contracted CLABSI (True) or not (False).

The distribution of these values is as follows:

* True: The patient has contracted CLABSI, in 52 instances
* False: The patient has not contracted CLABSI, over 14,000 instances

This reveals a significant class imbalance, with more instances of patients not contracting CLABSI compared to those who did. This imbalance presents challenges in predicting infection rates. If the majority class (no CLABSI) is not handled appropriately, a model can learn to focus on and predict the majority class while ignoring the minority class, which is our target group. Addressing class imbalance is crucial when building predictive models to ensure accurate and meaningful results.

## Addressing missing values in the dataset based on data type and correlation with the target variable

The dataset contains numerous missing values, which in clinical datasets can introduce significant bias for models. Hence, handling missing values is a critical part of data preprocessing. A targeted approach for addressing missing values is implemented, considering both the data type and its correlation with the target variable.

The process for managing missing values is as follows:

* *With EDA*, understand the sparsity of the dataset and the types of columns affected. Then, categorize missing values by data type (object, Boolean, numeric).
* *Object data type:* For object data type, missing values are filled with the string 'NA', which serves as a placeholder for unknown or missing categorical values.
* *Numeric data type*: Missing values are handled based on the target variable “HasCLABSI” due to the aforementioned class imbalance. Hence, a tailored approach to handle missing value is necessary.
  + For instances where “HasCLABSI” is True, missing values are filled with the median of non-missing values within the corresponding data subset. The median is preferred due to its central tendency and resilience to skewed data.
  + Similarly, for rows where “HasCLABSI” is False, missing values are replaced with the median of non-missing values within the corresponding data subset.

This nuanced approach not only improves the quality of data feeding into the model but also aligns with clinical understandings that certain measurements may vary significantly between infected and non-infected patients.

## Reduce dimensionality with correlation matrix and pattern matching

The dataset initially contains 278 attributes. However, a high number of attributes can hinder performance, especially since not all dimensions carry the same effect on the target variable. Hence, new methods for dimensionality reduction were developed, resulting in a new dataset with 168 attributes.

The process to reduce dimensionality is as follows:

* *Correlation matrix:* Utilize the correlation matrix and set a threshold to understand each attribute's impact on the target variable. This information is used in an automated procedure to identify and remove highly correlated columns that introduce redundancy and inefficiency. Eliminating columns with strong correlations exceeding a predefined threshold minimizes redundancy, thereby enhancing overall model performance and computational efficiency.
* *Pattern matching:* Numeric details are extracted from column names to facilitate dimensionality reduction. For example, when column names include numeric identifiers such as "MedicationsInjectedLast30," "15," "5," "3," and "2" days, numeric suffixes are extracted based on a specified pattern, retaining only the column with the highest numeric suffix for each prefix. This pattern-matching technique ensures the retention of the most relevant features and helps preventing overfitting, improving the model's generalizability.

A 40% reduction in dimensionality simplifies the dataset while preserving crucial information.

## Refine and enhance the target variable to improve the prediction of CLABSI within 3 days

Early detection of CLABSI allows healthcare providers to employ timely intervention strategies. To refine the prediction, two new target variables are created to capture the risk of CLABSI occurrence over time. These two variables are: “HasCLABSI\_NextDay” and “HasCLABSI\_Next3Days.” These new target variables provide more granular and actionable insights into the risk of CLABSI development over time.

The process for determining and employing this approach is as follows:

* *Sorting:* First, the data frame is sorted by 'PatientKey' and 'Date' columns. This is essential for subsequent operations because later queries involve shifting values based on patient identifiers in chronological order.
* *“HasCLABSI\_NextDay” is created*: This variable indicates whether a patient is at risk of developing a CLABSI infection the following day. The value of "HasCLABSI" is shifted by -1 for each patient group to generate this new variable.
* *“HasCLABSI\_Next3Days” is created:* This variable is built on the previous step and captures the risk of developing CLABSI within the next three days. This variable is constructed by combining the shifted "HasCLABSI" values for the next three days using conditional logic operations. As a result, if a patient is at risk at any time within the next three days, it is reflected in the target variable.

# Modeling Design

*Summary: An evaluation of 4 different models, each demonstrating unique strengths. All can capture complex data patterns, effectively handle high-dimensional data structures containing missing values and class imbalance.*

Choosing the appropriate classification models for predicting CLABSI in PICU settings depends on several factors such as specific characteristics of the medical dataset, the complexity of the prediction problem, and the interpretability of the machine learning models. All of which are carefully examined for each contending model: Neural Networks (NN), k-Nearest Neighbors (kNN), Naïve Bayes, and Decisions Trees. These models go through the same three-step preparation process:

* Data preprocessing phase includes splitting the data into train-test sets and handling missing values and class imbalance
* Training phase
* Model evaluation phase

## Neural Networks (NN)

Given the current dataset’s characteristics - high-dimensional data structure, missing values, and complex data patterns - NN is a model worth considering. NN’s interconnected nodes across multiple layers allow it to process data effectively and exhibit flexibility in detecting patterns within complex datasets.

### Advantages of NN for this dataset:

* *Sequential dependency:* With this data set, panel data (observations of multiple entities over time) is a concern. However, NN is adept at capturing sequential dependencies, making it well-suited for time series forecasting and other time-dependent duties.
* *Nonlinear relationships*: The model is suitable for tasks where the relationships between predictors and the target variable are nonlinear and hard to interpret, which are present in our dataset.
* *Uninformative data*: It also works well with uninformative and sparse data and can handle class imbalance with ease.

### Challenges of NN for this dataset:

However, neural networks lack interpretability due to intricate architecture. Additionally, because of its ability to learn complicated patterns, it tends to be overfitting. To resolve these challenges, analysts can:

* *Adjust hidden layers:* Adjust the number of nodes in the hidden layer because the number of nodes heavily impact how intricate and complicate the model becomes.
* *Adaptive learning rate:* Monitor the learning rate 𝜂 to change values as the training moves forward is also an effective way to prevent overfitting.

## k-Nearest Neighbors (kNN)

KNN is an instance-based learning algorithm that does not make strong assumptions about the underlying data distribution and is mostly used for classification tasks.

### Advantages of kNN for this dataset:

* *Simplicity and efficiency:* kNN is notable for its ease of implementation and the ability to handle large and complex data. It offers rapid exploration of medical data which enables quick analysis and generation of initial insights.
* *Flexibility:* It also provides flexibility in handling missing data, crucial for such datasets that have various missing values.
* *Adaptability:* kNN is known to belong to a group of “lazy learning” methods because kNN does not explicitly "train" on data like other models. Instead, it stores a training dataset including all available input data points and calculates distances between new data points to make predictions. For that reason, new data can be added seamlessly, making it advantageous when dealing with panel data where the data distribution may change over time.

### Challenges of kNN for this dataset:

* *Computational cost:* Because of its adaptability, it requires a large memory capacity and expensive computational cost as the size of dataset keeps increasing.
* *Conditional effectiveness:* Its effectiveness may depend on the specific characteristics of the panel data and the chosen distance metric. This method also requires standardizing or scaling the data before performing calculations.

## Naïve Bayes

It is a probabilistic classifier that applies Bayes' theorem. When provided with a labeled dataset, it calculates the probability of each data point belonging to each class. This calculation is based on the observed records and the model assumes strong independence between the data points or features. The class with the highest probability will be predicted as the outcome.

### Advantages of Naïve Bayes for this dataset:

* *Efficient analysis:* This model is adept at providing timely insights in medical settings, especially for CLABSI prediction. This allows healthcare providers to take prompt intervention measures.
* *High dimensional data:* Naïve Bayes can handle high-dimensional data, especially sparse data. Additionally, it exhibits a high resistance to overfitting.
* *Impactful attributes:* It controls the impact of redundant features that might be present in the medical dataset and focuses only on the most informative variables for CLABSI detection. This makes the model easy to train and interpret.

### Challenges of Naïve Bayes for this dataset:

* *Feature independence assumption:* Naïve Bayes’ assumption of feature independence may not always be true for all datasets, especially real-world scenarios. It oversimplifies relationships which reduces flexibility in catching complex data patterns.
* *Loss of information:* This happens because Naïve Bayes tends to ignore potential correlation and dependence of features.

Overall, while Naive Bayes does not directly affect Type I and Type II errors, they can be influenced by the choice of classification threshold, which simulates the trade-off between Type I and Type II errors.

## Decision Tree

Decision Tree model splits the data into subsets using feature rules, resulting in a tree with a collection of decision nodes and leaf nodes.

### Advantages of Decision Tree for this dataset:

* *Simplicity and interpretability:* Easy to understand and interpret, Decision Tree model is valuable in areas like healthcare where interpretability is crucial for gaining insights of CLABSI risk factors.
* *Effective handling of complex data:* It can handle panel data well and provides understanding of the variables that contribute to CLABSI risk over time. Decision Tree can also capture non-linear relationships between predictors and the target variable and handle missing data and class imbalance well. Together, these ensure unbiased model performance in medical datasets.
* *Discrete target variable:* Because the interest of outcome is binary (HasCLABSI: True/False), Decision Tree is a suitable model because it requires the target class to be discrete. This means the target variable must be clearly classified as either belonging to a particular class (CLABSI or non-CLABSI).
* *Feature selection:* Decision Tree utilizes a distinct method of selecting features that measures how well it splits the data. It determines the most relevant risk factors that best separate the data into CLABSI vs. non-CLABSI classes. Then the algorithms make decisions on which features to prioritize when building the tree structure, leading to more accurate predictions or classifications.

### Challenges of Decision Tree for this dataset:

* *Overfitting:* Decision Tree is prone to overfitting. It occurs when the tree is too finely tuned and fully grown and becomes too complex. This means the model can perform exceptionally on the training data but poorly on newly introduced data.
* *Solution*: One way to prevent overfitting in Decision Tree is to prune nodes and branches to increase the generalizability of the results.

# Evaluation Plan

*Summary: Address data leakage, imbalanced data, and models’ inherent biases. Among the evaluated models, Decision Tree model emerges as the most balanced method.*

We plan to evaluate the model’s performance on several key aspects: overfitting, imbalanced data, classification errors, and their business implications. This plan outlines how the models can be tailored to meet the high standards required in clinical settings, ensuring that they are not only technically sound but also relevant and easy to interpret.

## Overcome overfitting and data leakage.

Overfitting is a critical concern in machine learning tasks, especially in medical data sets with a high feature-to-sample ratio. To address this, we consider several aspects that can contribute to overfitting and underfitting:

### Address data leakage with cross validation and withholding a small data set:

In machine learning, analysts build predictive models with currently available data so that the model can predict on new data. Considering that new data is not available, the practice of splitting a dataset into train and test sets is commonly used.

However, data leakage occurs when the model learns from the entire dataset, invalidating the estimated performance. Data leakage typically happens when analysts normalize the whole data set to estimate the performance of the model. This skews the model’s performance and typically results in overly optimistic results.

Our team addressed data leakage with cross validation and withholding a small data set. First, cross validation is employed to evaluate the model’s robustness and ensure it performs well on unseen data. This involves partitioning the data into multiple subsets based on certain parameters and training the model on several subsets while evaluating the remaining. Any data preprocessing is done separately for each cross-validation step and any normalization is performed based on the data in the split, not the entire set. This process helps in assessing the model's ability to generalize. Second, as we gather more data, a small data set can be reserved for the last step of evaluating the performance of our model. If the model yields abnormally high accuracy or “too good to be true” performance, data leakage has occurred.

### Address overtly complicated models and overfitting with regularization:

We aim for a simple, as-complicated-as-needed model so that it can scale and adapt to new data. Models that are finely tuned to match the current data set can perform exceptionally well with the current data. However, this model will then be less effective and useful when new data is introduced.

To prevent overtly complicated models, we utilize regularization techniques. These techniques are incorporated in predictive models like Neural Networks and Decision Tree to discourage overtly complex models. Regularization adds a penalty term to the loss function used to train the model, which can control the magnitude of model coefficients and help in reducing overfitting.

Additionally, we can reduce model complexity by pruning decision tree or reducing the layers or number of units in neural networks to ensure that the model does not learn the noise in the training data. Finally, certain selection techniques can be applied to keep the most relevant and impactful variables, reducing the dimensionality, and improving model interpretability as it scales.

## Overcome imbalanced data

The significant class imbalance noted with “HasCLABSI” with most instances being “False” can lead the model to be biased towards predicting the majority class. To address this, two balancing techniques are implemented: SMOTE and a tailored sampling approach.

### Synthetic Minority Over-sampling Technique (SMOTE):

This is utilized to balance the dataset by creating synthetic examples rather than over-sampling with replacement. This helps in providing a more balanced dataset, which improves the training process and enables the models to learn more about the minority class.

The process of utilizing SMOTE is as follows: SMOTE is implemented after the train-test split on training set to avoid data leakage. After applying SMOTE, analysts now have synthetic samples on the minority class to train our models. If SMOTE is adopted before train-test split, the synthetic data is created based on the whole data set which will invalidate the estimated performance of the model. The model’s performance is measured after it being applied to the test set.

### Tailored sampling technique:

A tailored sampling technique can be utilized to complement balancing techniques such as SMOTE. Analysts can implement different sampling techniques for different models to ensure that each model type is provided with the best representation of both data classes.

## Metrics to evaluate models’ performance:

Several metrics will be used to evaluate the models:

* Precision and Recall: Important for this dataset where the cost of a false negative (failing to predict CLABSI) could be higher than a false positive. While false positives can be costly in other contexts, false negatives cost patients’ lives in the medical settings.
* F1-Score: Harmonic mean of precision and recall, useful when we need a balance between precision and recall.
* ROC-AUC Score: Represents a model's ability to discriminate between the classes at various threshold settings.
* Confusion Matrix: Helps visualize the performance of the algorithm.
* Accuracy: This is not the golden standard in this context due to the class imbalance. Almost 90% of instances are without CLABSI. Hence, a model can accurately predict patients without CLABSI and yield a high accuracy rate. However, the objective is to determine patients with CLABSI for prompt treatment and intervention. A model with a lower accuracy rate but higher precision, recall, and lower type 2 errors will be preferred.

## Performance adjustment within algorithms based on models’ inherent biases:

After conducting some model evaluations, we notice that some algorithms inherently reduce type I (false positive) or type II (false negative) errors. Therefore, we explored some strategies to leverage these biases towards achieving more clinically useful predictions. More specifically, different threshold adjustments such as sensitivity and specificity according to the clinical needs.

An overview of models’ performance in predicting the incidence of CLABSI, with the aforementioned metrics as benchmarks:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Best Performing** | **Contending** | **Reason for Fit** |
| **Precision** | Decision Tree | Naive Bayes | Decision Tree shows the highest precision for classifying class 1 instances correctly (0.59) |
| **Recall** | Decision Tree | Naive Bayes | Decision Tree also leads in recall for class 1, managing to identify 62% correctly |
| **False Positives (Type 1)** | K-Nearest Neighbours | Decision Tree | K-Nearest Neighbours has more Type 1 errors compared to Decision Tree in relative terms |
| **False Negatives (Type 2)** | Decision Tree | Naive Bayes | Decision Tree has fewer Type 2 errors compared to Naive Bayes, aligning with higher recall |
| **Overall Model Accuracy** | Decision Tree | Neural Network | Decision Tree and Neural Network both achieve high accuracy, but Decision Tree performs slightly better in other metrics |
| **Inference** | Decision Tree | Naive Bayes | Decision Tree exhibits better capability at identifying CLABSI, at the cost of precision compared to other models. It suggests a more balanced approach than the Neural Network |

### Presicion and recall: Decision Tree models shows its effectiveness:

Notably, the Decision Tree model outperforms others in precision and recall, indicating its effectiveness in correctly identifying actual CLABSI cases while minimizing false alarms. With a precision of 0.59, the Decision Tree model correctly predicts CLABSI occurrences almost 60% of the time when it detects an infection is present. This level of precision is particularly important in clinical settings to ensure that patients are not subjected to unnecessary treatments. Moreover, the Decision Tree leads in recall with 62%, implying it successfully identifies 62% of all true CLABSI cases, which is crucial for a condition where early detection can significantly influence patient outcomes and reduce hospital stays.

### Type 1 and type 2 errors: kNN tends to over-diagnose type 1 errors while Decision Tree has lower type 2 errors

K-Nearest Neighbours model is noted to have more instances of type 1 errors in comparison to the Decision Tree, indicating a tendency to over-diagnose. While over-diagnosis is less dangerous than under-diagnosis in certain medical contexts, it can still lead to unnecessary treatments and higher healthcare costs. Conversely, the Decision Tree model demonstrates a better balance by having fewer Type 2 errors, meaning it is less likely to miss actual cases of infection. This balance is critical because Type 2 errors can be particularly perilous in a healthcare setting, as failing to detect a true case of CLABSI could result in worsening patient conditions and fatal outcomes.

### Overall: Decision Tree model appears to be the most balanced:

Decision Tree model emerges as the most reliable, exhibiting the most balanced performance across all metrics, making it a compelling choice going forward. This shows the model's superior capability to discern CLABSI instances without excessively compromising either precision or recall.

# Further on Recommendations

## The trade-off between type 1 and type 2 errors

Type 1 errors (false positive) and type 2 errors (false negative) exist in a trade-off relationship where a decrease in one result in an increase in another.

Within the context of business and healthcare, the complexity of clinical decision-making demands consideration of both the statistical performance of models and their practical implications. It is true that both type 1 and type 2 errors are detrimental and should be minimized. The nuanced trade-offs between detecting true positives and avoiding false alarms are pivotal in choosing the right model.

On one hand with false positives, healthcare providers waste resources in treating a non-infected patients who will, as a result, face high healthcare charges. Considering that resources, both personnel and materials, are limited in a hospital setting, we cannot simply ignore the cost of type 1 errors.

However, on the other hand, a high type 2 errors costs lives because a patient’s health is endangered due to undetected and untreated CLABSI. Additionally, hospital quality ratings, patients’ satisfaction, and reimbursement rates are all negatively affected when type 2 errors are prominent within an institution.

## The practicality of a cost matrix

A cost matrix assigns different costs and benefits to each type of error, depending on the decision-makers’ discretions. Thus, it can be useful to help decision makers determine the optimal approach depending on the context of the problem.

With a cost matrix, we can evaluate the cost and benefits of each combination of possibilities. It is true that in different context, type 1 errors will be costlier than type 2 errors. However, within the scope of medicine, type 2 errors are more detrimental. Hence, can assign a higher cost to type 2 errors to align it with clinical outcomes. This approach can be extended to models for a cost-sensitive learning. With cost-sensitive learning, models can factor in the cost associated with misclassification during its training process. Overall, adopting a cost matrix and cost-sensitive learning can help decision makers decide the optimal approach to address both business and clinical’s needs. Below is a simplified example of a cost matrix

|  |  |
| --- | --- |
| True positive (correctly identify an infected patient) | Low cost |
| True negative (correctly identify a non-infected patient) | Low cost |
| False positive (Type 1 error: incorrectly identify infection) | Medium cost (cost resources) |
| False negative (Type 2 error: incorrectly identify no infection) | High cost (patient’s health and life at risk) |