Predicting True Sepsis and Culture-positive Sepsis in Intensive Care Unit with Machine Learning Techniques

Zeyuan Wu

Master's thesis
2024:E11

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

Predicting True Sepsis and Culture-positive Sepsis in Intensive Care Unit with Machine Learning Techniques

Zeyuan Wu

Supervisors:
Prof. Andreas Jakobsson, Prof. Attila Frigyesi

LUND UNIVERSITY

Centre for Mathematical Sciences
Lund University

Abstract

Sepsis, a serious medical condition often leading to patients requiring intensive care, has prompted numerous scientists to employ mathematical techniques to aid in its diagnosis. This thesis uses logistic regression and a machine learning technique, XGBoost, to predict true sepsis (as opposed to sepsis mimics) and culture-positive sepsis (among true sepsis) in critical care using blood test results, physiological measurements and other patient characteristics.

In this study, the dataset employed for constructing the prediction models comprises the information of 2,667 patients across 105 variables. Notably, a considerable portion of these variables exhibits missing values. To address this issue, imputation techniques are systematically applied to rectify the gaps within the dataset.

The predictive models acquired in this study are evaluated with the area under the operating characteristic curve (AUC) and using cross-validation. To address the imputed missing values within the dataset, a modified cross-validation technique is employed. This methodology ensures that imputed values are exclusively utilized during the training phase, while the testing phase exclusively involves the use of the original, unaltered data. Variable selection and analysis have been conducted employing forest plots for regression, while for XGBoost models, significance is determined through the utilization of importance plots and SHAP value plots.

The result of this study shows that XGBoost performs better than the regression models. In predicting true sepsis, the XGBoost model achieves an AUC of 0.74, while the regression model yields an AUC of 0.72. In predicting culture positivity, the XGBoost model attains an AUC of 0.77, whereas the regression model yields an AUC of 0.74. Both the XGBoost algorithm and regression models demonstrated efficacy in predicting true sepsis and culture-positive sepsis. The performance of these prediction models exhibits potential for enhancement with the utilization of a more extensive dataset. Consequently, mathematical models serve as valuable and effective aids in supporting medical professionals' clinical judgement.

Keywords: Machine Learning, Diagnosis of Sepsis, XGBoost, Logistic Regression

Contents

1 Introduction

1.1 Backgrounds

Sepsis is a severe condition resulting from the body's hyperactive response to an infection, leading to damage to its own tissues and organs. It poses a risk to individuals of all ages but particularly affects older adults, infants, pregnant women, and those with underlying health issues, as suggested in [1, 2]. Typical indicators of sepsis encompass fever, elevated heart rate, rapid breathing, cognitive disorientation, and bodily discomfort. If untreated, from [3], it can progress to septic shock, multiple organ failure, and even mortality. While bacterial infections commonly trigger sepsis, it can also arise from viral, parasitic, or fungal sources. Timely identification and appropriate clinical intervention, including optimal antimicrobial administration and fluid replenishment, will improve the chances of survival as shown in [4, 5].

To treat sepsis effectively, healthcare practitioners usually first observe concerning symptoms and conduct tests for accurate diagnosis. Then it will come to identifying the primary infection source which often involves conducting blood culture tests as discussed in [4]. However, from [4], diagnosing sepsis presents challenges due to overlapping symptoms with other ailments. Conditions such as pneumonia, urinary tract infections, and influenza may present symptoms that resemble those of sepsis. Therefore, it becomes imperative to explore alternative methodologies to augment the diagnosis of sepsis.

Apart from healthcare professionals' diagnoses, applying mathematical and data science approaches has demonstrated its potential in addressing this problem. In particular, "whether or not" problems can be effectively treated as binary classification problems, leveraging statistical techniques. By leveraging medical data obtained from previous patients, predictive models can be developed to aid in diagnosing new patients. Investigating and assessing these models aims to advance the field of medical diagnostics and provide healthcare practitioners with valuable insights. Leveraging statistical predictive models in diagnostic processes presents several advantages. Firstly, it operates independent of a physician's experience, relying solely on data and model analysis for diagnoses. Moreover, diagnoses are rendered post the patient's completion of all requisite tests, offering rapidity crucial in time-sensitive scenarios, expediting diagnoses and potentially hastening treatment onset. Additionally, mathematical models, inherently impartial, yield relative accuracy, particularly when supported by extensive data sets, as shown in [6]. With proper training and validation, these models consistently exhibit high accuracy rates in diagnosing illnesses.

Numerous instances and research endeavors have showcased the application of mathematical methodologies. Hazem Koozi and collaborators, for instance, developed a statistical model aimed at predicting patient mortality, yielding commendable outcomes as evidenced by their work in [7]. Notably, their model demonstrated comparable performance to the widely utilized SAPS III system, a well-established tool in forecasting the mortality of patients in intensive care units illustrated in [8]. There are many studies in predicting sepsis occurrence among patients in ICUs and most of them acquired favourable outcomes, such as [9].

In conclusion, sepsis continues to pose a significant medical challenge, particularly in its diagnosis. Conventional diagnostic techniques, which rely predominantly on clinical observation and laboratory tests, may not always suffice for accurate sepsis identification. Therefore, the integration of mathematical and data science methodologies presents promising opportunities to enhance diagnostic precision and efficiency in addressing this critical healthcare concern.

1.2 Previous Studies

Numerous investigations have employed mathematical models to assess the suitability of biomarkers or medical tests as indicators for predicting sepsis. Tan et al. determined that, with respect to the diagnostic accuracy of C-reactive protein (CRP) for sepsis, the comprehensive area under the summary receiver operator characteristic (SROC) curve was 0.73. Conversely, in evaluating the diagnostic accuracy of procalcitonin (PCT) for sepsis, the overall area under the SROC curve was calculated to be 0.85. This suggests that PCT holds greater promise as an indicator for predicting sepsis when compared to CRP [10]. Wacker and colleagues also reported procalcitonin (PCT) as a valuable biomarker for distinguishing between sepsis and other non-inflammatory syndromes, achieving an area under the receiver operating characteristic (AUROC) of 0.85 [11]. Their investigations have revealed that specific variables exhibit significant relevance to the prediction outcome of "having sepsis." Consequently, increased emphasis will be placed on these variables in future model building within this study.

Moreover, additional investigations have been conducted, constructing models by leveraging biomarker and medical test data to ascertain the likelihood of a patient having sepsis. Barton and colleagues developed a predictive model utilizing the XGBoost machine learning algorithm in [12] for sepsis prognosis. Their model demonstrated a superior performance with the highest achieved area under the receiver operating characteristic curve (AUC) being 0.88, notably surpassing the AUC of the Systemic Inflammatory Response Syndrome (SIRS), which stood at 0.66. Their research demonstrated the efficacy of employing the XGBoost algorithm as a viable approach for addressing this particular challenge. Consequently, the models within this study will also incorporate the XGBoost algorithm.

Calvert and collaborators introduced a novel machine learning system named InSight in [13]. The InSight system exhibits noteworthy predictive capabilities for sepsis, demonstrating a comparatively high area under the receiver operating characteristic curve (AUC), reaching a peak value of 0.965. Hye Jin Kam and Ha Young Kim employed both the InSight system and deep neural networks to construct models for the early detection of sepsis in [14]. Their results yielded an area under the receiver operating characteristic curve (AUC) approximately at 0.9. They also showed that the efficacy of neural networks in outperforming alternative methodologies in specific instances.

The predictive models established in these studies demonstrated favourable outcomes; however, they are not without certain limitations to the patients they studied. Notably, these models are founded on data from all patients admitted to the ICU. The ICU admits patients for a spectrum of reasons, making it intricate to distinguish sepsis cases from other conditions. This differentiation will offer limited insights into sepsis diagnosis. In the clinical setting, physicians can readily distinguish that a patient admitted to the Intensive Care Unit (ICU) due to cerebral hemorrhage is unlikely to be suffering from sepsis, obviating the necessity for mathematical models in such cases. However, the challenge lies in accurately identifying ambiguous cases where patients exhibit symptoms that overlap with those of sepsis. Hence, it is advisable for studies to focus on patients manifesting at least a subset of symptoms associated with sepsis, rather than conducting predictions for all ICU patients indiscriminately.

1.3 Thesis Objective

As elucidated previously, several factors necessitate alterations in the construction of predictive models for sepsis diagnosis, thereby ensuring their alignment with real-world complexities and enhancing their practical significance. In this study, predictive models are exclusively developed within the subset of patients who undergo blood culture testing, indicating suspicion of sepsis. By focusing solely on this cohort, the predictive model aligns more closely with real-world scenarios. It aims to offer practical assistance and valuable insights to healthcare practitioners encountering difficulty in determining whether a patient presents with sepsis or not. Moreover, the novel predictive models developed in this study for anticipating culture blood test outcomes have not been previously introduced. These models serve as a valuable resource for healthcare practitioners, aiding in the early identification of the specific type of infection present in patients.

Predictive models will be constructed employing logistic regression alongside various machine learning methodologies such as XGBoost, neural networks, and Random Forests. The findings will primarily showcase the outcomes derived from the XGBoost method and logistic regression. The presentation will solely include the results from these two methodologies as other machine learning methods, upon evaluation, demonstrated inferior performance compared to XGBoost. Also, predictive models in this study not only aid in diagnosing sepsis but also forecast the likelihood of positive culture results in patients, contributing to comprehensive care strategies.

2 Data Pre-processing

2.1 General Pre-processing

The data set employed in this research study, denoted as **D**, are medical data of 2667 patients provided by the local hospital in Lund and Malmö. The data set **D** encompasses an extensive array of variables, numbering over one hundred in total. These variables are categorised into distinct domains, including basic demographic information, laboratory test results, medical history regarding diseases, past medication records, and doctor's notes. To enhance the data set's coherence and utility, an initial step involves the consolidation of certain variables. This consolidation is necessary due to variations arising from the use of different testing instruments, even though they pertain to the same target variables.

Subsequently, a rigorous data pre-processing procedure is enacted to ensure the data set's quality. Outliers, extraneous variables, and those with an excess of 35% missing values are meticulously identified and removed. As a result of this meticulous curation, a refined dataset is obtained, characterised by a total of 66 variables pertaining to a cohort of 2,667 patients.

However, it is worth noting that even after this curation process, the data set is not yet amenable to rigorous analytical procedures. A significant portion of the remaining variables still contain missing values. Therefore, the crucial preparatory step before embarking on model-building activities is the implementation of a data imputation strategy. This data imputation process is essential to address the gaps in the data set and render it suitable for subsequent analyses and modelling endeavours.

2.2 Data Imputation

The presence of missing values in dataset **D** may arise due to instances where certain patients either did not undergo specific medical tests or the recorded test results were not documented within the system. Readers can gain a general overview of the extent of missing values within **D** by referring to the histogram in figure 1. In this graphical representation, the x-axis denotes the percentage of missing values associated with each variable. For instance, variables with less than 5% missing values, represented by the first bar in 1 exhibit a frequency of approximately 60, indicating that these variables constitute around 60% of the total variables.

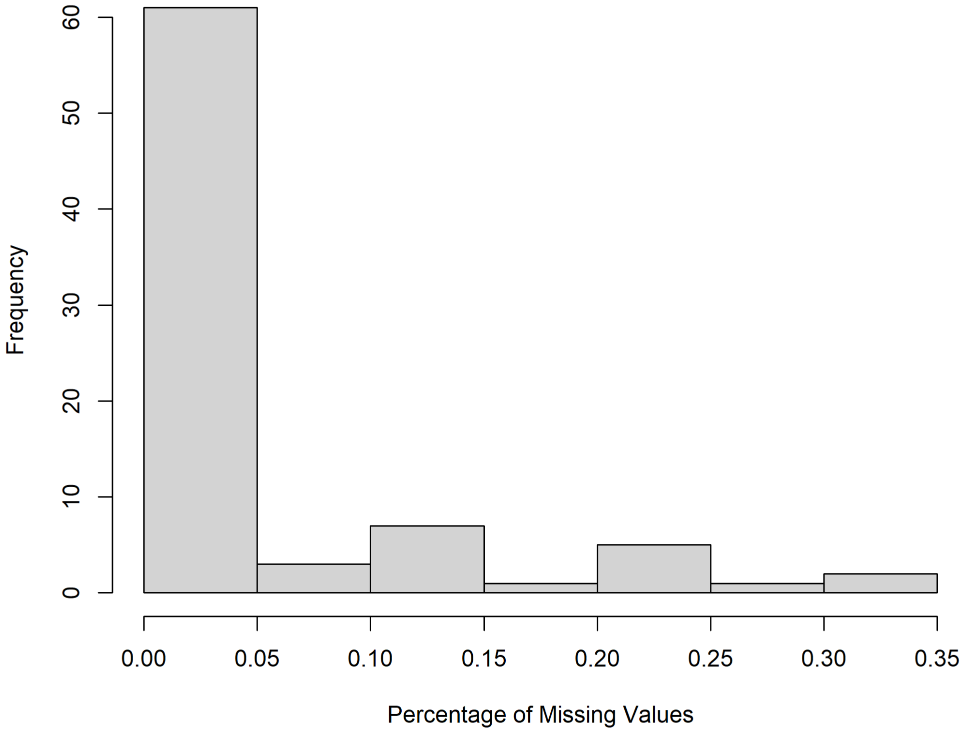


Figure 1: Histogram of the percentage of missing values of variables in data set **D**.

As depicted in the histogram, following the exclusion of variables with in excess of 35% missing values, a noteworthy observation is that the majority of the remaining variables exhibit a relatively modest presence of missing values, typically falling below the 5% threshold. This finding suggests that these missing data points are amenable to resolution through an appropriate imputation method, facilitating a more comprehensive and analytically robust data set for subsequent analyses.

In the context of data imputation, there typically exist two prominent choices, namely MissForest, illustrated in [15] and the K-nearest neighbours (kNN) imputation techniques. MissForest stands as a machine learning algorithm designed specifically for the imputation of missing values within datasets. Leveraging the random forest methodology, it predicts missing values by drawing insights from the available non-missing data across various columns. Notably, its efficacy remains robust even when dealing with a substantial proportion of missing data, outperforming other conventional imputation techniques.

K-Nearest Neighbours (KNN) imputation stands as a method employed for completing missing values within a data set by approximating them through neighbouring data points. Numeric missing values are imputed by assigning them the mean value derived from their k nearest neighbours. Categorical missing values are filled by employing a majority voting approach, wherein the most frequently occurring category among their k nearest neighbours determines the imputed value. While considered straightforward and user-friendly, its efficacy can diminish when dealing with data sets characterized by high dimensionality or sparsity.

To ascertain the optimal method for imputation and rigorously evaluate its performance, a simulation study is conducted using a subset of the dataset characterised by the absence of missing values across all variables. Fortunately, this subset comprises data entries from a cohort of 246 patients who exhibit complete data records for all relevant variables. A new data set, denoted as $\mathbf{D}_c$, is constructed from the information contained within these 246 patients. Next, a certain percentage of data is randomly removed from $\mathbf{D}_c$. In this process, the percentage of missing values corresponding to each variable in the original data set is mirrored within $\mathbf{D}_c$, resulting in the creation of a new data set, $\mathbf{D}_{cm}$, which incorporates missing values. By implementing imputation on $\mathbf{D}_{cm}$ and comparing the imputed data set with the original data set $\mathbf{D}_c$, one can infer how different imputation techniques will perform for this data set.

Subsequently, both MissForest and kNN imputation methods are applied to $\mathbf{D}_{cm}$, yielding an imputed data set, $\mathbf{D}_{ci}$. To evaluate the efficacy of these imputation techniques, a comparative analysis is performed by contrasting $\mathbf{D}_c$ against $\mathbf{D}_{ci}$ using two distinct criteria: the normalised root-mean-square error (NRMSE) for numeric variables and the proportion of false classification (PFC) for categorical variables. NRMSE is defined as:

$$\text{NRMSE} = \sqrt{\frac{\text{mean}\ \left((X_{\text{true}} - X_{\text{imp}})^2\right)}{\text{var}\left(X_{\text{true}}\right)}},$$

where X_{true} is the real data set or data matrix and X_{imp} is the imputed data set or matrix ,following the definition in [16]. NRMSE quantifies the disparity between predicted and actual data values in the context of real-valued data. A lower NRMSE signifies superior imputation accuracy. PFC is defined as the ratio between the number of falsely classified instances and the total number of instances in the data set. It serves as a measure for evaluating the efficacy of imputation methods applied to categorical data. A reduced PFC value indicates enhanced performance of the imputation model.

The outcomes of this comparative assessment, elucidating the performance of the two imputation methods, are presented in table 1 and 2 for reference.

Table 1: NRMSE of two imputation methods.

	NRMSE	95% CI
MissForest	0.569	(0.525, 0.624)
kNN, k = 3	0.696	(0.616, 0.715)
kNN, k = 5	0.661	(0.556, 0.767)
kNN, k = 7	0.659	(0.568, 0.750)

Table 2: PFC of two imputation methods.

	PFC	95% CI
MissForest	0.425	(0.403, 0.446)
kNN, k = 3	0.499	(0.480, 0.518)
kNN, k = 5	0.481	(0.453, 0.508)
kNN, k = 7	0.467	(0.449, 0.485)

As delineated in the table, it is discernible that MissForest imputation method demonstrates superior performance. Nevertheless, it is imperative to underscore that this simulation serves primarily as a reference point. This stems from the fact that the data set $\mathbf{D}$ at large is considerably more extensive, tenfold in magnitude, compared to the reduced dataset $\mathbf{D}_c$ utilised in this simulation.

In forthcoming modelling endeavours, it is envisaged that both of these imputation techniques, MissForest and kNN, will be employed. The ultimate selection between the two will hinge upon their respective performance in enhancing the modelling outcomes, with the area under the receiver operating characteristic curve (AUC) serving as the primary metric of evaluation within this study. The approach is to prioritise the imputation method that contributes to superior modelling performance, as gauged by the AUC metric.

The MissForest method is ultimately selected for imputation in this study. The inclusion of both methods serves as a prudent decision, as assessing imputation performance is impossible without knowing the actual values of the missing data.

3 Predicting "TrueSepsis"

The complex and elusive nature of sepsis makes accurate diagnosis a formidable task for healthcare professionals. Consequently, the need to develop robust methods for distinguishing sepsis in patients is of paramount significance. To address this imperative, this section aims to apply mathematical methodologies, specifically logistic regression and XGBoost, as tools to tackle the challenge of sepsis identification. The target variable $\mathbf{Y}$ under consideration is "TrueSepsis" within the data set denoted as $\mathbf{D}$, other independent variables will be denoted as $\mathbf{x}$. When the binary variable $\mathbf{Y}$ is assigned a value of true, it signifies the presence of sepsis in the patient, while a false value indicates the absence of this condition. These analytical techniques will be leveraged to provide a reliable means of discerning the presence or absence of sepsis in patients, thereby contributing to enhanced clinical decision-making and patient care.

3.1 Logistic Regression

Given that the desired outcome $\mathbf{Y}$ is a binary variable, employing logistic regression represents a dependable methodology for addressing this issue. Logistic regression is a statistical technique applied in binary classification, aiming to forecast the likelihood of an event or outcome occurrence, often represented as either 1 (positive class) or 0 (negative class), as described in [17]. Its versatility extends across diverse domains including machine learning, where it serves as a foundational method for predictive modelling. Hence, it will serve as a valuable benchmark for the ostensibly more potent machine learning methods under consideration in this study.

3.1.1 Full Model

First of all, a full model using logistic regression including all variables is introduced.

When $Y_i = 1$ for the i-th patient, it signifies the presence of sepsis, whereas $Y_i = 0$ indicates the absence of sepsis in that individual. Let $\mathbf{x}_i = (x_{i1}, \cdots, x_{ik})$ denote the independent variables or medical test results for the i-th patient, where k is the total number of independent variables.

To perform the logistic regression, assume that the probability of $Y_i = 1$ is q_i,

$$\Pr(Y_i = 0) = 1 - \Pr(Y_i = 1) = p_i.$$

Y_i is a random variable with Bernoulli distribution,

$$Y_i \sim \text{Bernoulli}(q_i).$$

The logistic regression model will be:

$$\mathbb{E}(Y_i) = \text{logit}\,(q_i) = \log \frac{q_i}{1-q_i} = \mathbf{x}_i \boldsymbol{\beta},$$

where $\boldsymbol{\beta}^\top = (\beta_0, \cdots, \beta_k)$ is the vector of unknown parameters for each variable. $\boldsymbol{\beta}$ will be estimated through maximum likelihood method and after that, the regression model will be complete. If $\beta_k = 0$, it means that the k-th independent variable should not be included in the regression model.

In assessing the performance of the developed model, the metric utilized will be the Area Under the Receiver Operating Characteristic (ROC) Curve, commonly referred to as AUC. This metric serves as an evaluation tool specifically designed for assessing statistical and machine learning models, particularly in the context of binary classification problems.

The ROC curve graphically represents a classifier system's performance while varying its discrimination threshold. It plots the true positive rate (TPR) or sensitivity against the false positive rate (FPR) or specificity at different threshold settings. Sensitivity and specificity are defined as follows,

$$\begin{aligned} \text{Sensitivity} &= \frac{\text{True Positives}}{\text{True Positives } + \text{ False Negatives}}, \\ \text{Specificity} &= \frac{\text{True Negatives}}{\text{True Negatives } + \text{ False Positives}}, \end{aligned}$$

where terms like "True Negatives" are defined in the table 3,

Table 3: Definitions of TP, FP, FN, TN.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

The harmonic mean of specificity and sensitivity is defined as F1-score, which is:

$$\text{F1-score} = 2 \times \frac{\text{specificity } \times \text{ sensitivity}}{\text{specificity } + \text{ sensitivity}}.$$

The AUC quantifies the entire two-dimensional area under this curve, spanning from the coordinates (0,0) to (1,1). A sample ROC plot is given in figure 2, where the x-axis is specificity and the y-axis is sensitivity.

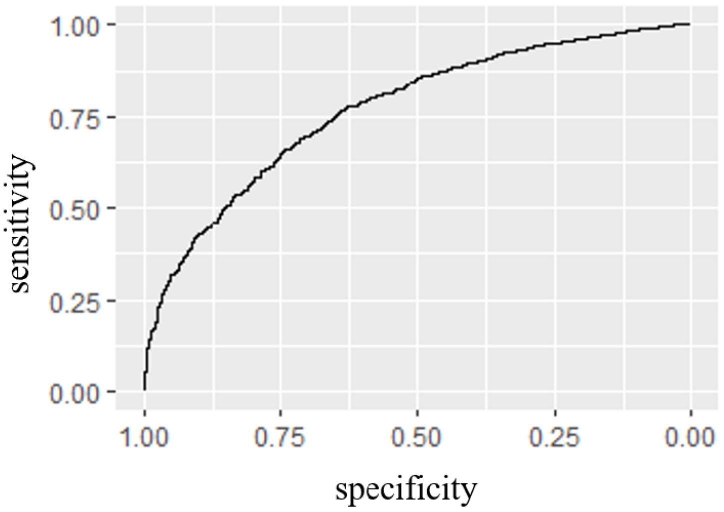


Figure 2: Illustration of the ROC curve.

The prediction outcomes derived from a logistic regression model yield probabilities ranging from 0 to 1. The predicted probability q_i, denoted as $\hat{q}_i$, signifies the probability of $Y_i = 1$, indicating the probability of the i-th patient exhibiting sepsis. The primary objective of the regression model is to classify patients into respective classes. Consequently, a threshold is typically designated for this purpose. If q_i falls below the chosen threshold, the patient is classified into the class of not having sepsis; conversely, if q_i surpasses the threshold, the patient is classified into the opposite class.

Therefore, varying the threshold value will result in different predicted outcomes, thereby influencing specificity and sensitivity measures. The ROC curve illustrates the relationship between specificity and sensitivity across the entire range of threshold values from 0 to 1. For instance, if the threshold is set to 0, all patients will be classified into class "1," leading to a sensitivity of 1 and specificity of 0. Conversely, if the threshold is chosen as 1, all patients will be predicted as "0", and the sensitivity will be 0 and the specificity will be 1.

The AUC is defined to be the area under the ROC curve. The AUC's value is confined within the range of 0.5 to 1. A higher AUC value signifies superior discrimination capability within the model. An AUC of 1 denotes flawless discrimination, while an AUC of 0.5 indicates performance equivalent to random chance.

This metric's robustness lies in its insensitivity to the classification threshold employed. Consequently, AUC serves as a prevalent tool for comparing distinct models or refining models throughout the developmental stages.

Additionally, for model evaluation purposes, cross-validation will be employed. Cross-validation serves as a technique utilized in statistical analysis and machine learning to appraise the performance of predictive models. Its primary objective is to gauge the model's ability, trained on a specific data set, to generalize effectively to an independent dataset.

In the process of building and testing models for regression analyses or machine learning, it is customary to partition the dataset randomly into two subsets: the training set and the test set. The training set is utilized to train the model, while the test set serves to evaluate the model's performance. Notably, the test set comprises data that the model has not encountered during its training phase, enabling an assessment of the model's generalization capabilities to unseen data. Typically, the majority of the dataset, ranging from 70% to 80%, is allocated to the training set, with the remainder reserved for the test set. However, allocating only 20% of the data to the test set may not adequately reflect the model's true performance, prompting the adoption of cross-validation techniques. Through cross-validation, each data point in the dataset serves as the test set at least once, facilitating a more comprehensive evaluation of the model's performance.

The fundamental concept underlying cross-validation involves the partitioning of the dataset into multiple subsets, commonly referred to as folds. Subsequently, the model is trained on a subset of the data, known as the training set, and subsequently evaluated on the remaining data, designated as the test set. This iterative process is repeated multiple times, each iteration involving a distinct partitioning of the data into training and test sets.

The prevalent approach in cross-validation is k-fold cross-validation shown in [18], involving the division of the dataset into k subsets or folds of comparable sizes. Through this method, the model undergoes training k times, utilizing k-1 folds for training in each iteration and reserving the remaining fold for validation or testing. This process ensures that each data point contributes to the validation process precisely once. The resulting performance metrics from each fold are amalgamated to derive an overall assessment of the model's performance.

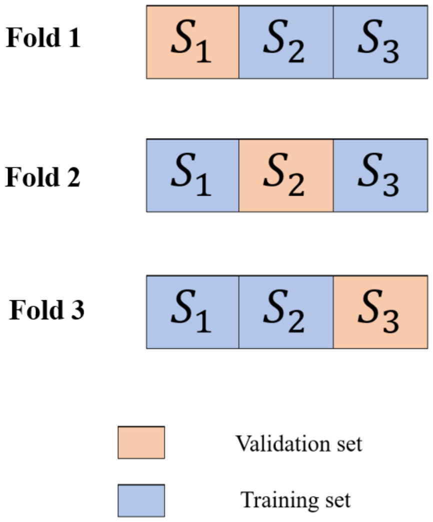


Figure 3: Example of 3-fold cross validation

Figure 3 illustrates an instance of k-fold cross-validation with k set to 3. Initially, the dataset SS is divided into three equal parts. Subsequently, in each fold, one of these parts (S_1, S_2, or S_3) is designated as the validation or test set, while the remaining two parts are utilized as the training set. The process iterates such that each part serves as the test set once, while the others collectively form the training set. Following the training and testing of all three folds, the model's performance is evaluated by aggregating the outcomes from each fold. Specifically, in this thesis, the result of a 3-fold cross-validation is represented by the mean of the three AUC values obtained from the models of the respective folds.

Cross-validation plays a pivotal role in more accurately estimating the model's performance compared to a single split of the dataset into training and testing sets. By leveraging multiple data splits, it effectively mitigates variance in performance estimation, contributing to a more reliable evaluation methodology.

The full regression model's performance is evaluated by 5-fold cross-validation, where 20 per cent of the data was used as a validation set, detailed results shown in table 4. In the train test split, no individual will appear in both sets. The mean of test AUC is 0.719, with the 95% confidence interval (0.689, 0.750).

Table 4: Results of cross validation for the full regression model predicting "TrueSepsis."

Resample	**AUC**	**Spec**	**Sens**	**F1-score**
Fold1	0.702	0.898	0.229	0.365
Fold2	0.717	0.912	0.213	0.345
Fold3	0.691	0.937	0.206	0.338
Fold4	0.753	0.935	0.227	0.365
Fold5	0.733	0.935	0.244	0.387

In this context, cross-validation can only serve as a point of reference due to the potential presence of imputed data within the validation set. It is impractical to utilise the entire set of real data as a validation set in the full model, as only 246 patients out of the total number 2667 have complete medical data without any missing values. A test set comprising less than 10% of the total dataset may not provide sufficient evaluative capacity. However, in subsequently reduced models, validation will exclusively be performed using real data. A modified cross-validation approach tailored for data sets containing missing values will be introduced to address the challenge of missing values. This adapted method ensures appropriate validation and assessment of the model's performance.

3.1.2 Stepwise Selection Based on P-value

In regression analysis, there is a general inclination towards favouring a smaller reduced model over the full model. This preference is rooted in the reduced model's tendency to offer increased generality and a measure of mitigation against overfitting. Additionally, within this study, numerous variables exhibit missing values. These missing values, even post-imputation, might adversely impact the model's performance. The strategic removal of unnecessary variables can mitigate the potential influence of these missing values on the model.

In this section, a backwards stepwise selection based on the P-value is implemented to reduce the full model. In regression analysis, a small p-value, typically below a chosen significance level such as 0.05, associated with a predictor variable indicates substantial evidence against the null hypothesis. This suggests that the predictor variable is statistically significant in its contribution to predicting the response variable. Conversely, a large p-value indicates limited evidence against the null hypothesis, implying that the predictor variable may not be statistically significant in predicting the response variable.

Stepwise selection based on p-values serves as a technique for identifying the significant variables within a regression model. This iterative process involves the continual addition or removal of variables contingent upon their statistical significance, commonly assessed through p-values. This method offers an expedient and automated approach to variable selection, particularly advantageous when handling a substantial pool of potential predictors. The intricacies of this stepwise methodology are delineated in the accompanying flowchart in figure 4.

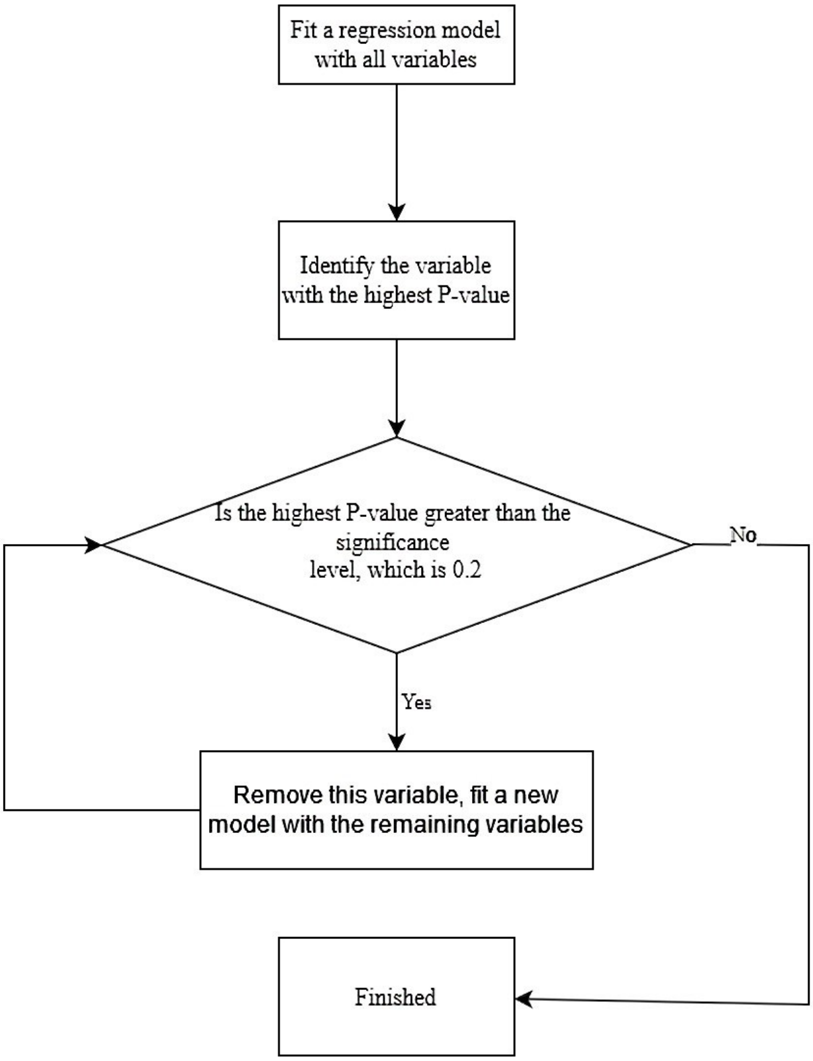

Fit a regression model with all variables
Identify the variable with the highest P-value
Is the highest P-value greater than the significance level, which is 0.2
No
Yes
Remove this variable, fit a new model with the remaining variables
Finished

Figure 4: Procedure of the stepwise selection method based on P-value.

Following the stepwise selection process, a reduced model of 12 features is obtained. This model's training area under the curve (AUC) is 0.757 with a 95% confidence interval (0.749, 0.765) , while the test AUC is measured at 0.714 with a 95% confidence interval (0.709, 0.719).

3.1.2.1 Forest Plot

To further analyse this model, the forest plot raised in [19] is introduced. A forest plot within regression analysis serves as a graphical depiction presenting estimated coefficients alongside their associated confidence intervals for distinct predictor variables within a regression model. Frequently utilized in meta-analyses and multiple regression analyses, this visualization facilitates the comparative examination of effect sizes among diverse predictors. Its utility lies in enabling researchers to expeditiously evaluate the influence of varied predictors within the model. Furthermore, by visually contrasting the lengths and positions of confidence intervals, this plot offers valuable insights into the relative significance and precision of these estimates.

The forest plot depicting the key findings of the reduced model is presented as figure 5.

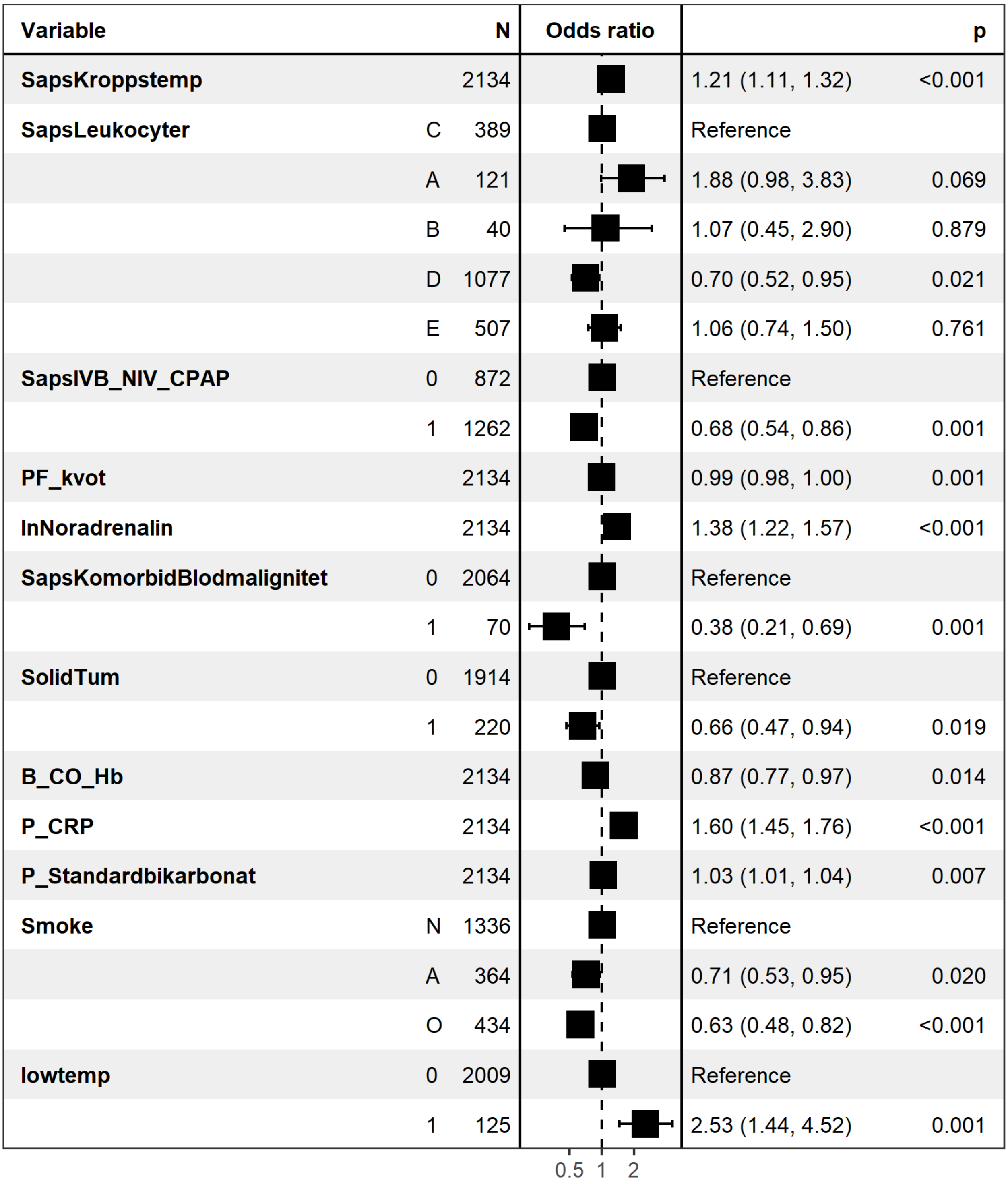
The plot encompasses several key components:

Variables: Each predictor variable within the regression model is delineated along the vertical axis of the plot.

Effect Size: Represented by squares, the estimated effect sizes or coefficients for each predictor are graphically depicted as points.

Confidence Intervals: Horizontal lines extending from these points signify the confidence intervals encircling the estimated coefficients. The length of these lines serves as an indicator of estimate precision, with shorter lines reflecting higher precision.

Vertical Line: A vertical line, positioned at the null value, aids in discerning the statistical significance of a predictor's effect. If the confidence interval does not intersect or cross this line, it implies statistical significance.



Variable		N	Odds ratio		p
SapsKroppstemp		2134		1.21 (1.11, 1.32)	<0.001
SapsLeukocyter	C	389		Reference	
	A	121		1.88 (0.98, 3.83)	0.069
	B	40		1.07 (0.45, 2.90)	0.879
	D	1077		0.70 (0.52, 0.95)	0.021
	E	507		1.06 (0.74, 1.50)	0.761
SapsIVB_NIV_CPAP	0	872		Reference	
	1	1262		0.68 (0.54, 0.86)	0.001
PF_kvot		2134		0.99 (0.98, 1.00)	0.001
lnNoradrenalin		2134		1.38 (1.22, 1.57)	<0.001
SapsKomorbidBlodmalignitet	0	2064		Reference	
	1	70		0.38 (0.21, 0.69)	0.001
SolidTum	0	1914		Reference	
	1	220		0.66 (0.47, 0.94)	0.019
B_CO_Hb		2134		0.87 (0.77, 0.97)	0.014
P_CRP		2134		1.60 (1.45, 1.76)	<0.001
P_Standardbikarbonat		2134		1.03 (1.01, 1.04)	0.007
Smoke	N	1336		Reference	
	A	364		0.71 (0.53, 0.95)	0.020
	O	434		0.63 (0.48, 0.82)	<0.001
lowtemp	0	2009		Reference	
	1	125		2.53 (1.44, 4.52)	0.001

Figure 5: Forest plot of the stepwise selection model for predicting "TrueSepsis."

From the forest plot, a compilation of 12 variables was considered in this model, with three variables exhibiting confidence intervals that do not intersect the vertical line: "SapsKroppstemp," "InNoradrenalin," and "P-CRP." It is noteworthy that the variable "P-CRP" demonstrates a notable odds ratio of 1.6, accompanied by a relatively narrow confidence interval. This finding underscores a robust correlation between "P-CRP" and the outcome variable "TrueSepesis."

"P-CRP" denotes the measurement of C-reactive protein, a standard medical test utilized to assess CRP levels in the bloodstream. C-reactive protein is synthesized by the liver as a response to inflammation. Elevated CRP levels in the blood typically signify ongoing inflammation within the body. Within the context of sepsis, which is triggered by the body's exaggerated reaction to an infection, notable increases in CRP levels often occur. Therefore, the considerable significance of "P-CRP" within the regression model is rational.

Regarding "SapsKroppstemp," it represents the measurement of body temperature, also indicative of inflammation within the body and it has been included in Systemic inflammatory response syndrome(SIRS) criteria from [20] for defining sepsis. Studies conducted by P. Póvoa et al. in [21] proved that CRP serves as a superior marker for sepsis compared to temperature. This observation aligns with the findings derived from the linear model in this study.

"InNoradrenalin" represents the measurement of noradrenaline, pivotal in regulating blood pressure and modulating the body's stress response mechanisms. During sepsis, the body often enters a state of shock prompted by an overwhelming immune reaction. This can result in a decline in blood pressure. To sustain blood pressure levels and ensure adequate blood circulation to essential organs, the body releases noradrenaline as a component of its natural stress response. Therefore, the presence of noradrenaline can serve as an indicator for the presence of sepsis, justifying its inclusion within the model.

Additionally, several other variables exhibit statistically significant odds ratios, although it is imperative to acknowledge that these findings may be influenced by the limited sample size of certain categories. Consequently, the confidence intervals associated with these odds ratios tend to be wider.

For instance, consider the category "1" within the variable "lowtemp," which constitutes only 6% of the total training set. Despite this relatively small representation, it yields a substantial odds ratio of 2.53. However, due to the limited sample size in this category, the corresponding confidence interval is noticeably wider, suggesting the need for a cautious interpretation of this particular result.

In the regression models of this study, which involve multiple hypothesis tests, the application of the Holm-Bonferroni method for p-value adjustment is deemed necessary. However, it is important to note that the p-values of the variables are not the primary consideration for determining their importance in this study. Following the completion of the stepwise selection model, variable importance analyses are conducted by examining the odds ratio component in the forest plot.

Throughout the stepwise selection process, the p-value serves as a criterion. However, only the variable with the largest p-value is removed at each step. For the largest p-value, the significance level remains unadjusted in the Holm-Bonferroni method and remains at 0.05. The threshold chosen for p-value stepwise selection in this study is 0.2. Consequently, the absence of p-value adjustments is not anticipated to impact the model's outcome.

3.1.2.2 Modified Cross Validation

Considering the reduced variables within the stepwise selected model, there has been a notable increase in the size of the non-imputed real data. To mitigate this change, a modified cross-validation method is introduced in this study by the author. This newly developed modified cross-validation technique ensures that exclusively authentic data are utilized as test data, thereby enhancing the alignment of the test results with real-world scenarios. Initially, the patients' medical data is partitioned into two subsets: P_i (containing imputed data) and P_r (comprising non-imputed real data).

For the training phase, only P_i is utilised. Subsequently, P_r is randomly partitioned into K segments denoted as $\{P_r^1, P_r^2, ..., P_r^k\}$, where each segment represents approximately 20 per cent of the entire set. In each iteration of the training process, the ith fold utilises P_r^i as the validation set, while the remaining P_r segments are employed as the training set alongside P_i. After conducting K training folds, the average test-AUC across these folds is calculated. Readers may refer to the figure below to understand how this model operates. Figure 6 illustrates the functioning of the modified cross-validation method when the parameter k is set to 3.

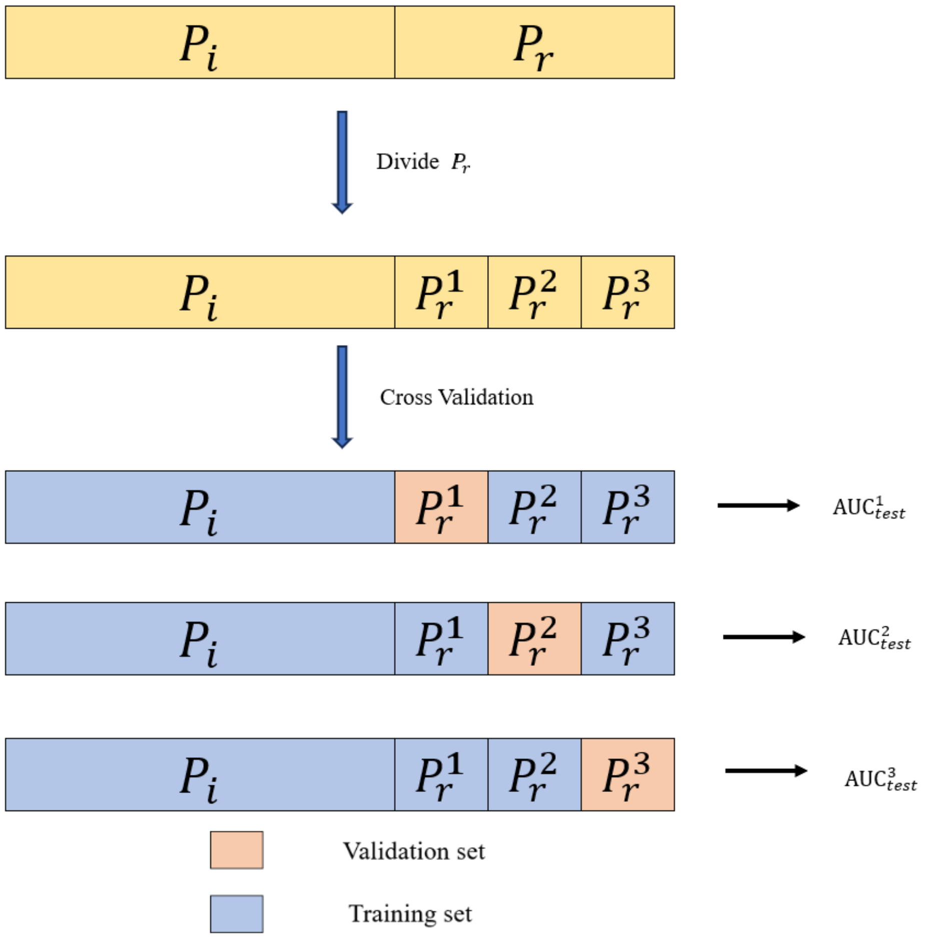


Figure 6: Procedure of modified cross validation method.

The process is repeated ten times to account for variability and ensure robustness, effectively removing uncertainty and generating more reliable results.

When employing this method, it is crucial to exercise caution regarding the proportion of outcomes within the prediction results of P_r. Specifically, the ratio between "True" and "False" outcomes in P_r should align with that of the entire data set. Failure to ensure this consistency may lead to intriguing yet inaccurate results.

The results from the modified cross-validation method are given as follows. P_r here contains 1111 patients' medical data, and 356 patients' variable "TrueSepsis" is "False", which is relatively consistent with the ratio 0.25 from the entire set. Detailed results of the ten tests is shown in table 5. The mean test-AUC is 0.716 with a 95% confidence interval (0.711, 0.721).

Table 5: Results of the modified cross validation for the stepwise selection model for predicting "TrueSepsis."

Training-AUC	**Test-AUC**
0.755	0.720
0.754	0.718
0.753	0.720
0.753	0.713
0.755	0.722
0.753	0.726
0.752	0.714
0.756	0.712
0.754	0.703
0.755	0.710

The findings derived from modified cross-validation demonstrate a slight improvement over the ordinary cross-validation method. Given the substantial difference in sizes between the training and test sets, it is worth noting that imputed data could exert a more pronounced influence on the test set compared to the training set.

Interpreting the results suggests that the regression model holds a 71% probability of accurately discerning whether a patient suspected of having sepsis indeed exhibits septic conditions. This statistical accuracy, when juxtaposed with clinical judgment, appears relatively acceptable. Within the dataset **D**, 75% of patients receive a correct diagnosis. Notably, the false instances of the variable "TrueSepsis" denote misdiagnoses as sepsis, excluding cases where patients wrongly classified as having sepsis are not encompassed within **D**. Consequently, the actual rate of correct diagnoses by physicians might be lower, rendering the performance of the regression model relatively acceptable.

In comparing the outcomes with those of the full model, the AUC (Area Under the Curve) remains nearly identical. This similarity underscores the success of the stepwise selection method based on p-values, as it effectively reduces model complexity while preserving performance. Furthermore, it highlights that certain variables bear minimal relevance to the target variable "TrueSepsis."

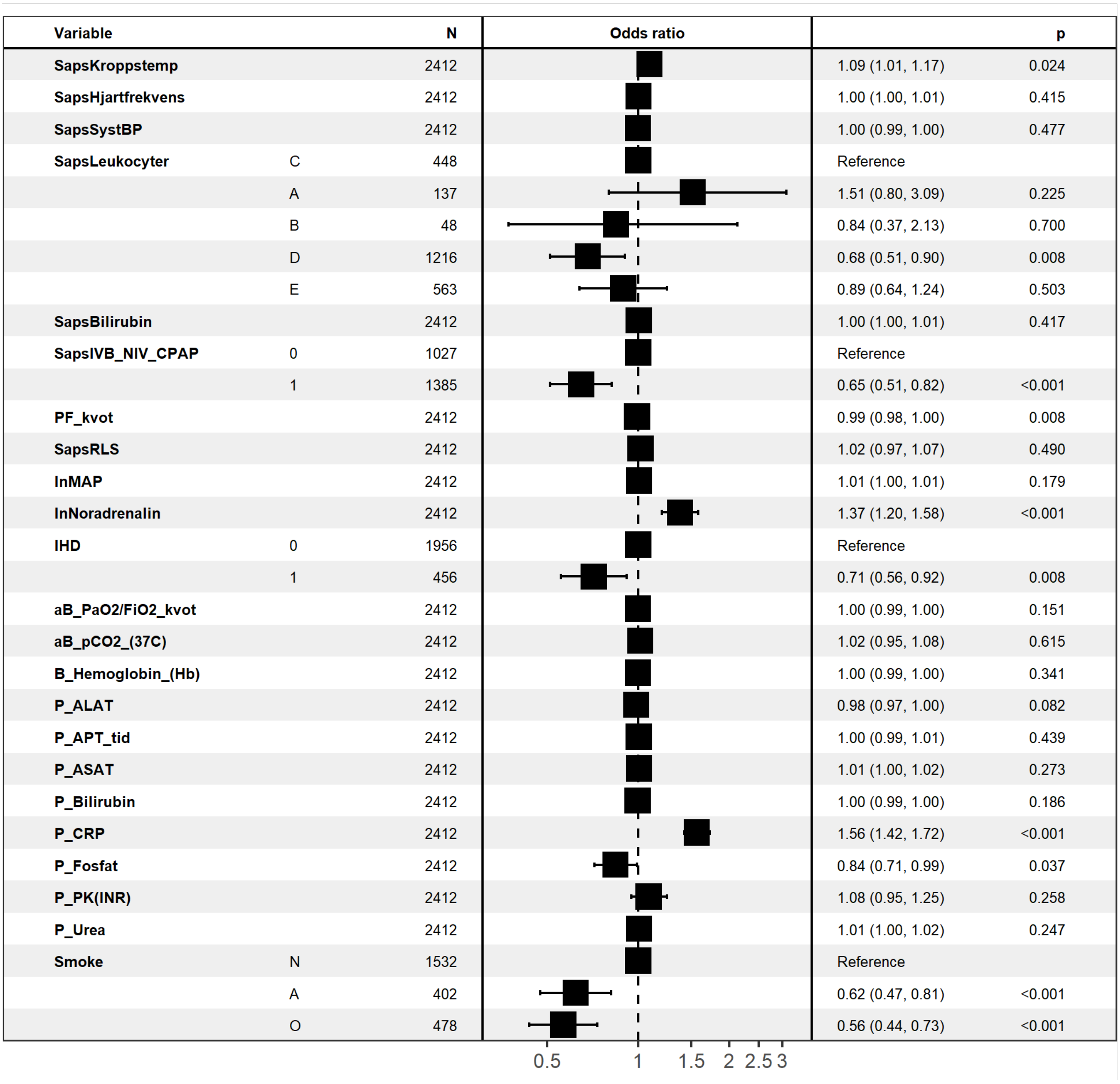
3.1.3 Uni-variate AUC

While the outcomes obtained through the stepwise selection method are promising, it is essential to acknowledge the inherent limitations of this approach. The variables selected may exhibit variance when applied to diverse training sets. To mitigate this potential variability, a strategic approach involves integrating alternative models developed through distinct criteria for variable reduction. This technique aims to temper the adverse impact stemming from the potential instability of variable selection across varying training data sets.

In this section, a uni-variate method will be incorporated for model reduction. Initial models will be constructed for each variable in conjunction with the outcome variable. Subsequently, variables whose models achieve an AUC exceeding a predefined threshold will be identified and chosen for inclusion. A novel model will be constructed using these selected variables and the outcome variable.

Thus, the selection of variables hinges upon the targeted evaluation metric, prioritizing optimal performance. Nonetheless, this methodology overlooks potential interrelationships among variables. Certain variables might exhibit predictive efficacy when in conjunction with others, despite lacking robust individual associations.

As before, this model's forest plot is presented below as figure 7.



Variable		N	Odds ratio		p
SapsKroppstemp		2412		1.09 (1.01, 1.17)	0.024
SapsHjartfrekvens		2412		1.00 (1.00, 1.01)	0.415
SapsSystBP		2412		1.00 (0.99, 1.00)	0.477
SapsLeukocyter	C	448		Reference	
	A	137		1.51 (0.80, 3.09)	0.225
	B	48		0.84 (0.37, 2.13)	0.700
	D	1216		0.68 (0.51, 0.90)	0.008
	E	563		0.89 (0.64, 1.24)	0.503
SapsBilirubin		2412		1.00 (1.00, 1.01)	0.417
SapsIVB_NIV_CPAP	0	1027		Reference	
	1	1385		0.65 (0.51, 0.82)	<0.001
PF_kvot		2412		0.99 (0.98, 1.00)	0.008
SapsRLS		2412		1.02 (0.97, 1.07)	0.490
lnMAP		2412		1.01 (1.00, 1.01)	0.179
lnNoradrenalin		2412		1.37 (1.20, 1.58)	<0.001
IHD	0	1956		Reference	
	1	456		0.71 (0.56, 0.92)	0.008
aB_PaO2/FiO2_kvot		2412		1.00 (0.99, 1.00)	0.151
aB_pCO2_(37C)		2412		1.02 (0.95, 1.08)	0.615
B_Hemoglobin_(Hb)		2412		1.00 (0.99, 1.00)	0.341
P_ALAT		2412		0.98 (0.97, 1.00)	0.082
P_APT_tid		2412		1.00 (0.99, 1.01)	0.439
P_ASAT		2412		1.01 (1.00, 1.02)	0.273
P_Bilirubin		2412		1.00 (0.99, 1.00)	0.186
P_CRP		2412		1.56 (1.42, 1.72)	<0.001
P_Fosfat		2412		0.84 (0.71, 0.99)	0.037
P_PK(INR)		2412		1.08 (0.95, 1.25)	0.258
P_Urea		2412		1.01 (1.00, 1.02)	0.247
Smoke	N	1532		Reference	
	A	402		0.62 (0.47, 0.81)	<0.001
	O	478		0.56 (0.44, 0.73)	<0.001

Figure 7: Forest plot for the uni-variate model for predicting "TrueSepsis."

The uni-variate model in total has 23 variables, which is about double the size of the stepwise selection method. When building the uni-variate model, all variables whose regression model using "TrueSepsis" as the outcome have an AUC larger than 0.53 were included. This criterion is meant to include all variables that may contribute to the prediction.

Upon scrutinizing the odds ratios, akin to the stepwise selection model, certain variables—specifically, "SapsKroppstemp," "InNoradrenalin," and notably, "P-CRP" —demonstrate relative significance. Particularly, "P-CRP" exhibits the highest odds ratio value. Besides, two variables, "P-Fosfat" and "P-PK(INR)," absent in the stepwise selection model, exhibit significance in this univariate model. However, "P-PK(INR)" displays an odds ratio only marginally above 1, while the confidence interval for "P-Fosfat" is considerably wide and nearly intersects with the vertical line. The prominence of "P-CRP" remains unaffected. On comparative analysis, it could be argued that these two models bear similarities concerning significant variables.

Regarding its performance, the Modified CV results are shown in table 6. The mean test-AUC is 0.709, with a 95% confidence interval (0.706, 0.711).

Table 6: Results of the modified cross validation for the uni-variate model for predicting "TrueSepsis."

Training-AUC	**Test-AUC**
0.751	0.712
0.751	0.712
0.752	0.705
0.751	0.707
0.751	0.711
0.751	0.706
0.751	0.711
0.752	0.706
0.751	0.702
0.751	0.714

Table 7: Performance results of two regression models for predicting "TrueSepsis."

	test-AUC	**95% CI**
Stepwise selection	0.716	(0.711, 0.721)
Univariate AUC	0.709	(0.706, 0.711)

From table 12, the mean test-AUC observed in the univariate model is marginally lower than that of the stepwise selection model. However, this discrepancy appears rational when considering the variance inherent in certain test datasets. It is plausible that with an alternative test set, the univariate model could potentially outperform the stepwise selection model. Hence, one might contend that the AUC performance of these models is notably similar. Consequently, it suggests a scenario where only a limited number of variables exhibit association with the outcome. Furthermore, the notion that augmenting the model with additional variables—despite their individual AUC test validation—might not significantly enhance predictive capability. Such a scenario raises concerns about model efficacy, as relying solely on three variables might not yield a robust prediction. This inherent issue might persist as the dataset remains static. At present, it is plausible to assert that the predictive performance of regression models is limited, yielding an AUC of 0.716 with a 95% confidence interval (0.711, 0.721).

To choose between the two regression models, the stepwise selection model boasts a more concise set of variables and demonstrates superior performance in terms of AUC. Consequently, the stepwise selection model is the more suitable choice within the regression domain.

3.2 Machine Learning: XGBoost

XGBoost, an abbreviation for eXtreme Gradient Boosting illustrated in [22], is a prominent and formidable machine learning algorithm predominantly applied to regression and classification tasks. It is classified within the domain of gradient boosting algorithms, which strategically construct an ensemble of decision trees to facilitate predictive modelling. Renowned for its exceptional computational efficiency, remarkable predictive accuracy, and high processing speed, XGBoost has consistently outperformed its peers in numerous machine-learning competitions. This exceptional performance has solidified XGBoost as a favoured choice in academic and industrial circles. Subsequently, a model utilising XGBoost will be presented below. Its importance plot and Shap value plot will be given.

3.2.1 The Theory

Before implementing the XGBoost algorithm, an explanation of it mathematical theory will be first given. XGBoost is based on gradient boosted trees, which originated from gradient boosting developed in [23] and detailed explanations of the gradient boosted trees will be shown in the following sections.

3.2.1.1 Elements of the learning algorithm

XGBoost is a supervised machine learning algorithm, which uses the training data $\boldsymbol{x}_i$ to predict the outcome y_i. The main aim of the learning algorithm is to find function $\boldsymbol{f}$ such that

$$\forall i, \quad \boldsymbol{f}(\boldsymbol{x}_i) = y_i.$$

Parameters The function $\boldsymbol{f}$ typically lacks a closed-form expression. To address this challenge, existing functions are often employed to approximate $\boldsymbol{f}$, with these approximations commonly referred to as models. An illustrative example of such models in supervised machine learning is the employment of linear models, denoted as $\hat{y}_i = \hat{\boldsymbol{f}}(\boldsymbol{x_i}) = \sum_j \theta_j x_{ij}$, as discussed in the preceding section.

Within the model function, certain components remain undetermined until derived from the provided data. These components, commonly known as parameters, are typically denoted as θ, analogous to its usage in the linear model. It is noteworthy that various models may entail multiple parameters. For instance, within the XGBoost algorithm, several commonly utilized parameters include η, denoting the learning step size, λ, representing the L2 regularization term, and the maximum number of leaves.

Objective Function In seeking the appropriate values for θ, a fundamental objective entails the minimization of $|\hat{y}_i - y_i|$, often achieved through the minimization of the loss function $L(\theta)$.A prevalent selection for the loss function L is the mean squared error, defined as:

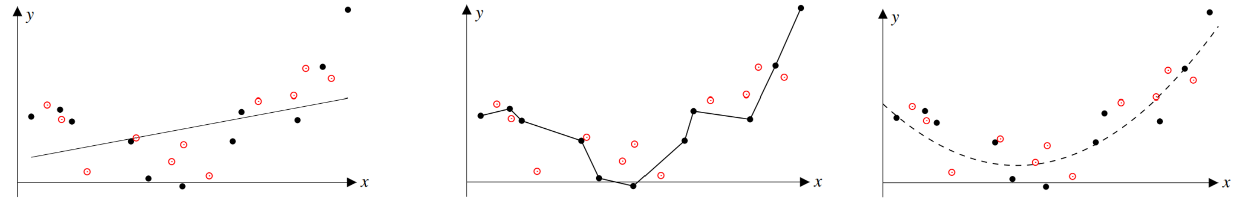
$$L(\theta) = \sum_i l\left(y_i, \hat{y}_i\right) = \sum_i \left(y_i - \hat{y}_i\right)^2.$$

For classification problems in this study, the logistic loss function is chosen as:

$$L(\theta) = \sum_i l\left(y_i, \hat{y}_i\right) = \sum_i \left[y_i \ln\left(1 + e^{-\hat{y}_i}\right) + \left(1 - y_i\right) \ln\left(1 + e^{\hat{y}_i}\right)\right].$$

Nevertheless, the minimization of the loss function primarily ensures that the parameter θ adequately fits the training data $\{\boldsymbol{x}_i, y_i\}$. Subsequently, the trained model is applied to new data i.e. test data, to forecast the outcomes. It is noteworthy that, on occasions, despite achieving a smaller loss function value, the performance of the model may deteriorate. Figure 8

[24], depicted below, serves as an illustrative example, demonstrating that although the training data is fitted well, the model may fail to generalize effectively to the test data.



(a) underfitting (b) overfitting (c) balanced model

Figure 8: An example of different training strategies affect the test results. Black dots are the training set, while red dots are the test sets.

Figure 8 illustrates three distinct levels of model performance. In Figure 8(a), the depicted straight line inadequately fits both the training data and test data. Conversely, Figure 8(b) portrays a scenario where the model perfectly fits the training data but fails to provide accurate predictions for the test data. Finally, Figure 8(c) exemplifies a balanced model, exhibiting proficient fitting capabilities for both the training and test datasets.

The exclusive minimization of $L(\theta)$ may lead to overfitting issues. To mitigate this challenge, a regularization term is introduced to regulate the model's complexity. Together, the regularization term denoted as $R(\theta)$ and loss function $L(\theta)$ forms the objective function F_{obj},

$$F_{obj} = L(\theta) + R(\theta).$$

The objective function is a representation of the bias-variance trade-off concept in machine learning as shown in [25]. Generally, as the number of tunable parameters in a model is increased, its flexibility increases, leading to improved adaptation to a given training dataset. This enhanced flexibility results in reduced bias. Nonetheless, with heightened model flexibility, there is a propensity for increased variance in the model's fit when applying different samples to generate new training datasets. Consequently, with variations in the training dataset, a substantial divergence in parameters will be observed. This divergence implies that the trained model may exhibit suboptimal performance when applied to test data, leading to diminished predictive accuracy. Thus, the primary objective of the XGBoost model is to minimize the objective function, with detailed elucidation provided in subsequent sections.

3.2.1.2 Classification and Regression Tree Ensembles

To find θ that minimise the objective function, ensemble tree methods are applied in the learning process of XGBoost algorithm. In machine learning, ensemble methods combines multiple simple individual models to formulate a more potent and robust predictive model. The fundamental concept underlying ensemble methods posits that through the amalgamation of predictions from numerous simple models, enhanced performance can be attained compared to any individual model in isolation, as proved in [26, 27]. These methodologies find extensive application across diverse machine learning tasks and are renowned for augmenting both the overall accuracy and generalization capabilities of the model.

The ensemble model employed in XGBoost comprises a collection of classification and regression tree (CART). CART initiates by meticulously selecting a specific feature to partition the dataset. Upon the selection of the splitting feature, the dataset undergoes division into two subsets contingent upon the values of that feature. For instance, if the chosen splitting feature pertains to "age," the dataset may be bifurcated into subsets comprising individuals either younger or older than a certain age threshold. Subsequent to the split, each terminal node, also referred to as a leaf, is assigned a predictive value. The predictive value assigned to each terminal node serves as a crucial component within the algorithm, facilitating the derivation of prediction outcomes in subsequent stages of analysis. Figure 9 is an example of CART for predicting whether or not one likes Kpop music.

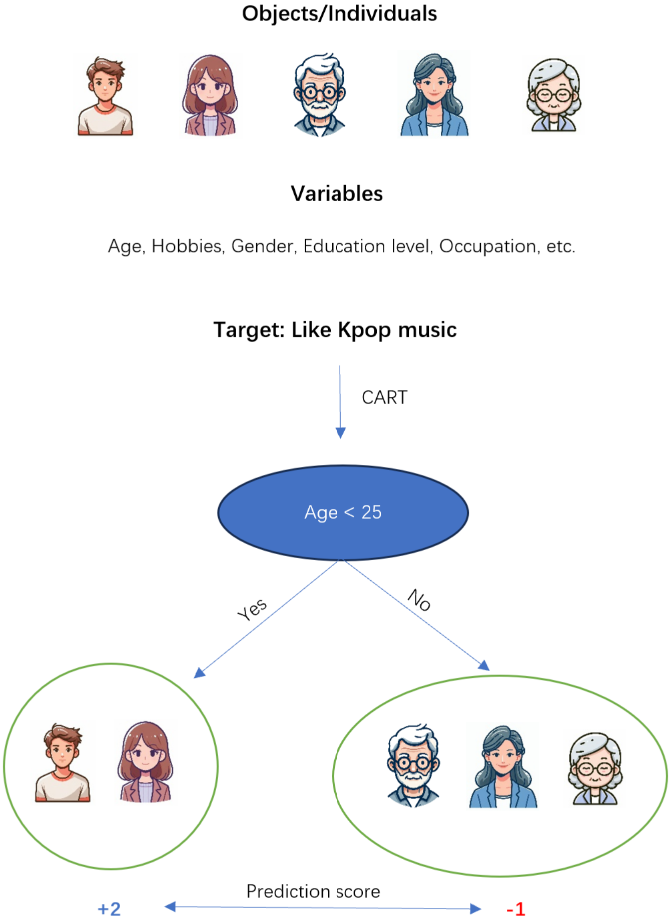


Figure 9: Illustration of CART based on five individuals.

The prediction in the aforementioned example is constructed using data from 5 individuals and several input variables. The CART, based on the variable "age," is illustrated in figure 9. Each leaf node within the tree is assigned a prediction score. The final prediction is determined by aggregating the prediction scores from the leaf nodes across multiple trees, reflecting the underlying principle of ensemble methods.

Figure 10 is an example of an ensemble of two trees. The final prediction scores are given for all the five individuals.

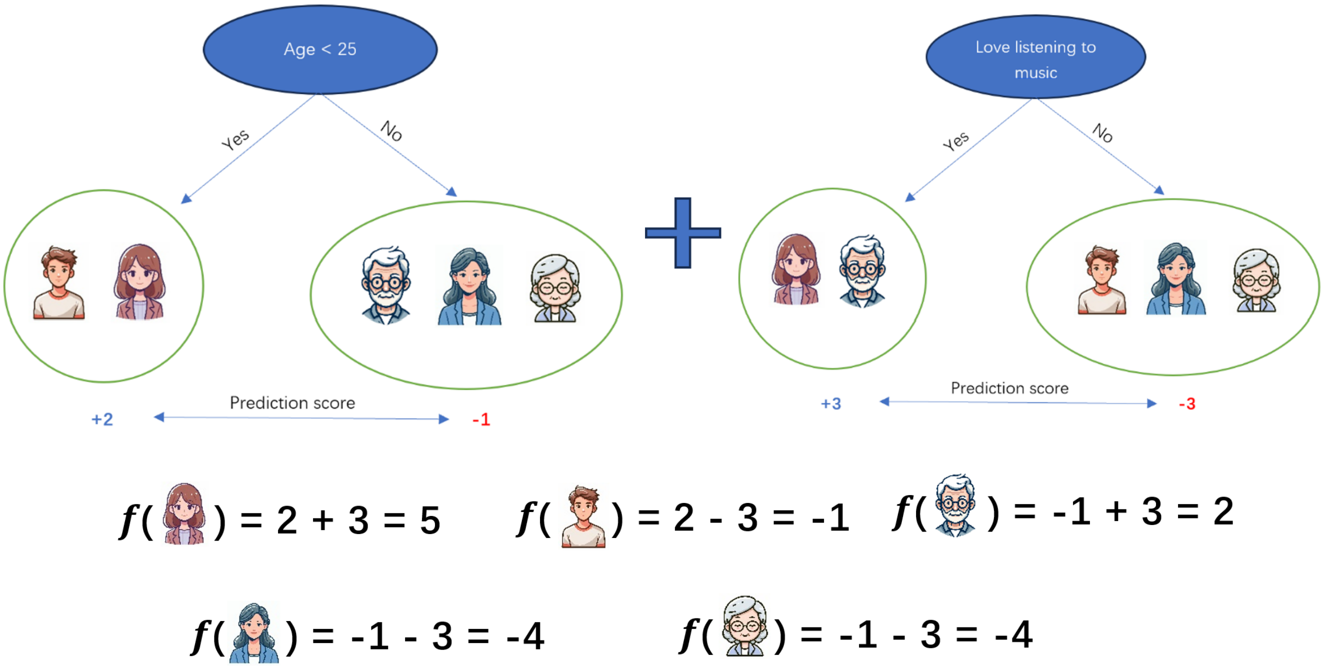


Figure 10: Ensemble of two CARTs and calculation of the prediction score for each individual.

Following this example, the mathematical expression of the ensemble trees are given as below:

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right), f_k \in \mathcal{F},$$

where K is the number of CARTs, and f_k is a function in the functional space $\mathcal{F}$, which is the space of all possible CARTs. The objective function to be minimised is given by

$$F_{obj}(\theta) = \sum_{i}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \omega\left(f_k\right),$$

where $l\left(y_i, \hat{y}_i\right)$ is the loss function and $\omega(f_k)$ is the complexity of the tree f_k, and it is meant to be the regularization term. Detailed definition of $\omega(f_k)$ will be given in the later sections.

3.2.1.3 Model Training

As observed in the aforementioned example, the prediction scores assigned to each leaf node, such as "+2", are not arbitrary but are determined through the training or learning process, as are the structures of the trees themselves. From a mathematical perspective, it is imperative to identify a suitable function f_k, that encapsulates both of these aspects. The learning process of f_k is done additively, whereby a new tree is sequentially incorporated into the model based on the previous training outcomes. This is necessary as training all the trees simultaneously by merely considering the gradient is not feasible. Denote the prediction value $\hat{y}_i$ at step s as $\hat{y}_i^{(s)}$, which can be written as:

$$\begin{aligned}
&\hat{y}_i^{(0)} = 0, \\
&\hat{y}_i^{(1)} = f_1\left(x_i\right) = \hat{y}_i^{(0)} + f_1\left(x_i\right), \\
&\hat{y}_i^{(2)} = f_1\left(x_i\right) + f_2\left(x_i\right) = \hat{y}_i^{(1)} + f_2\left(x_i\right), \\
&\quad \cdots \\
&\hat{y}_i^{(s)} = \sum_{k=1}^{s} f_k\left(x_i\right) = \hat{y}_i^{(s-1)} + f_s\left(x_i\right).
\end{aligned}$$

The objective function can be updated as:

$$\begin{aligned}
{F_{obj}}^{(s)} &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(s)}\right) + \sum_{i=1}^{s} \omega\left(f_i\right) \\
&= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(s-1)} + f_s\left(x_i\right)\right) + \sum_{i=1}^{s} \omega\left(f_i\right).
\end{aligned}$$

Since trees before f_s have been learned and established, the complexity term can be simplified, whereby the complexity of the prior trees can be reduced to a constant value:

$${F_{obj}}^{(s)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(s-1)} + f_s\left(x_i\right)\right) + \omega\left(f_s\right) + \text{constant.}$$

In the application of XGBoost models to various problem domains, including regression and classification, the choice of loss function varies. Certain loss functions may not yield a straightforward closed-form expression for the objective function. Consequently, a second-order

Taylor expansion is employed to approximate the loss function:

$$
\begin{aligned}
{F_{obj}}^{(s)} = & \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(s-1)}\right) + \partial_{\hat{y}_i^{(s-1)}} l\left(y_i, \hat{y}_i^{(s-1)}\right) f_s\left(x_i\right) + \frac{1}{2} \partial^2_{\hat{y}_i^{(s-1)}} l\left(y_i, \hat{y}_i^{(s-1)}\right) f_s^2\left(x_i\right) \right] \\
& + \omega\left(f_s\right) + \text{constant}.
\end{aligned}
$$

To simplify the expression, define the two partial derivatives as g_i and h_i respectively:

$$
\begin{aligned}
g_i &= \partial_{\hat{y}_i^{(s-1)}} l\left(y_i, \hat{y}_i^{(s-1)}\right), \\
h_i &= \partial^2_{\hat{y}_i^{(s-1)}} l\left(y_i, \hat{y}_i^{(s-1)}\right).
\end{aligned}
$$

As mentioned before, the training process is done additively so f_s and $\hat{y}_i^{(s-1)}$ has already be decided, which means they are constant. Since constant terms do not affect the minimisation, they can be removed to simplify the expression:

$$
{F_{obj}}^{(s)} = \sum_{i=1}^{n} \left[g_i f_s\left(x_i\right) + \frac{1}{2} h_i f_s^2\left(x_i\right) \right] + \omega\left(f_s\right).
$$

Following the treatment of the loss function, it is pertinent to provide the definition of the regularization term. First, define f_s as:

$$
f_s(x) = v_{q(x)}, v \in R^T, q : R^d \rightarrow \{1, 2, \cdots, T\},
$$

where v is a vector recording the prediction scores on leaves, and q is a function whose input is the data or object and the output is the leaf the data belongs, and T is the total number of leaves. Reader can refer to figure 11 as an example.

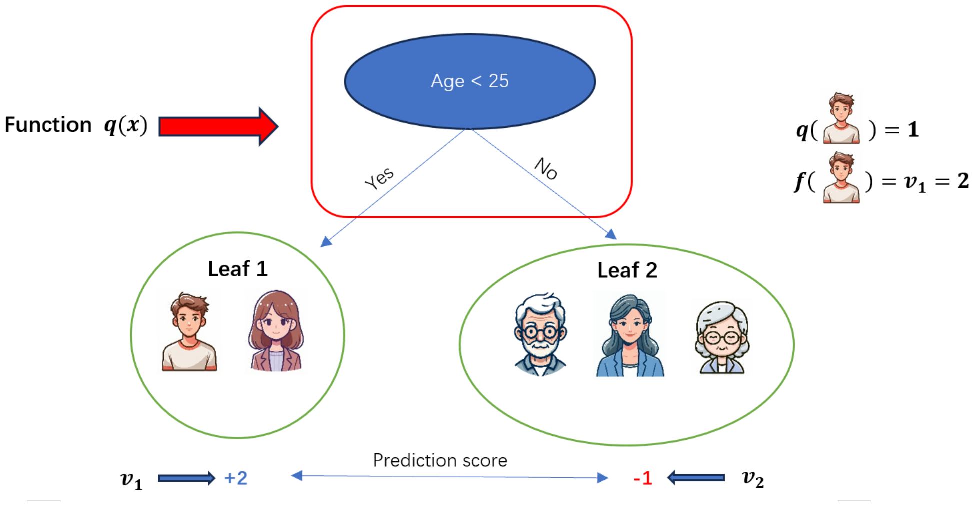


Figure 11: Explanation of the tree function f_s.

In this example, the tree has two leaves, and function q allocate the young boy in to leaf 1. Hence, the prediction score of the young boy in this tree is v_1 which is 2.

In XGBoost models, the model complexity $\omega(f)$ is usually defined as:

$$\omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} v_j^2.$$

This definition comes out of experience as shown in [22].

Finally, combining the loss and regularization term, the objective function can be expressed as:

$$\begin{aligned} {F_{obj}}^{(s)} &\approx \sum_{i=1}^{n} \left[g_i v_{q(x_i)} + \frac{1}{2} h_i v_{q(x_i)}^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} v_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) v_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) v_j^2 \right] + \gamma T, \end{aligned}$$

where $I_j = \{i \mid q(x_i) = j\}$ is a set of indices, indicating data points that are assigned to the j-th leaf. The second equality comes from that the prediction scores of data points belong to the same leaf are the same. To further simplify this equation, define

$$G_j = \sum_{i \in I_j} g_i,$$
$$H_j = \sum_{i \in I_j} h_i.$$

The expression can be reduced to:

$$F_{obj}{}^{(s)} = \sum_{j=1}^{T} \left[G_j v_j + \frac{1}{2} \left(H_j + \lambda \right) v_j^2 \right] + \gamma T,$$

which is a quadratic function of v_j.

Since the prediction scores v_j in different leaves are independent, the quadratic form $G_j v_j + \frac{1}{2} \left(H_j + \lambda \right) v_j^2$ has a minimum, which is achieved at v_j^*:

$$v_j^* = -\frac{G_j}{H_j + \lambda}.$$

The minimum at v_j^* given the structure function q is:

$$F_{obj}^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

For varying tree structures, denoted by different q, the minimum of F_{obj} will vary accordingly. The minimum, denoted as F_{obj}^*, is commonly termed as the structure score. A lower value of the minimum structure score indicates that the tree structure provides a better fit for the particular problem.

Ultimately, the remaining question pertains to the method for learning the true structure function q. One potential approach involves employing the exhaustive method, entailing the calculation of F_{obj}^* for all conceivable combinations. However, this approach may prove impractical due to the substantial computational time required.

To resolve this, a common strategy involves splitting a leaf into two leaves. The new structure score gained from this split is defined as:

$$\text{Gain } = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma.$$

From left to right, each term in this expression means, the score of the new leaf in the left, the score of the new leaf in the right, the score of the original leaf, the regularization term of the additional leaf. If the Gain is less than 0, meaning that it is less than the regularization term, this new branch should not be added since the gain in the loss function does make up the loss in model complexity.

Now the whole tree boosting theory is complete. In the next sections, specification on how to evaluate the performance of XGBoost models will be given.

3.2.2 Evaluation Methods

In XGBoost, importance plots serve as visual aids illustrating the relative significance of various features within a predictive model. These graphical representations facilitate the comprehension of the most impactful variables or features on the model's predictions. The importance of features is evaluated primarily through the metric known as "Gain" or "Importance score." This metric gauges the enhancement in AUC attributable to a specific feature within the decision branches it influences. A higher gain value denotes a more pivotal feature in the decision-making process of the model.

Another way to measure the feature's importance is using SHAP (SHapley Additive exPlanations) in [28]. SHAP values serve as a method to comprehensively assess the contribution of individual features towards predictions for specific instances within a model. SHAP value plots provide a visual depiction of the influence exerted by various features on model predictions for individual data points. In the context of classification models, SHAP value plots illustrate the significance of each feature through horizontal bars, delineating their positive or negative contributions. The length of each bar within the plot signifies the magnitude of a feature's impact on the prediction. This visualization technique proves instrumental in not only gauging the overall importance of features within the model but also in delineating the nuanced impact of each feature on individual predictions.

3.2.3 Model Analyses

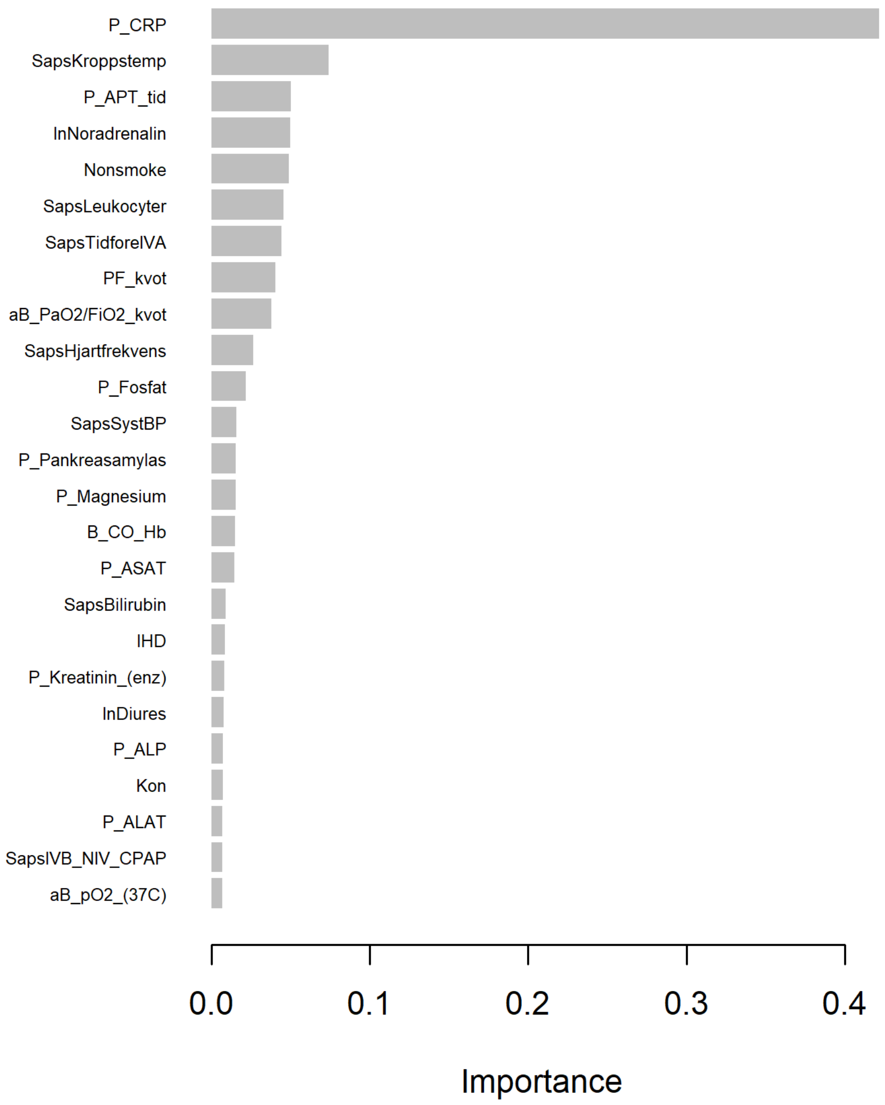


Figure 12: Feature importance of the top 20 variables used in the XGBoost model for predicting "TrueSepsis".

From the importance plot in figure 12, the XGBoost model encompassed 25 variables, some of which were not previously presented in the regression models. Notably, "P-APT-tid," absent in the regression models, ranked third in importance within the XGBoost model. This discrepancy might stem from XGBoost's capacity to discern nonlinear relationships during training. However, the identification of new variables potentially contributing to prediction might not markedly enhance model performance. This observation arises from the substantial importance attributed to "P-CRP," significantly outweighing the importance of all other variables within this model.

Alternatively, it can be inferred that the remaining variables do not contribute substantial or valuable information to the model. This observation aligns with our earlier findings in the context of regression models, where "P-CRP" consistently emerged as the dominant factor. In light of these findings, it is conceivable that augmenting the data set with additional data may improve the model's performance.

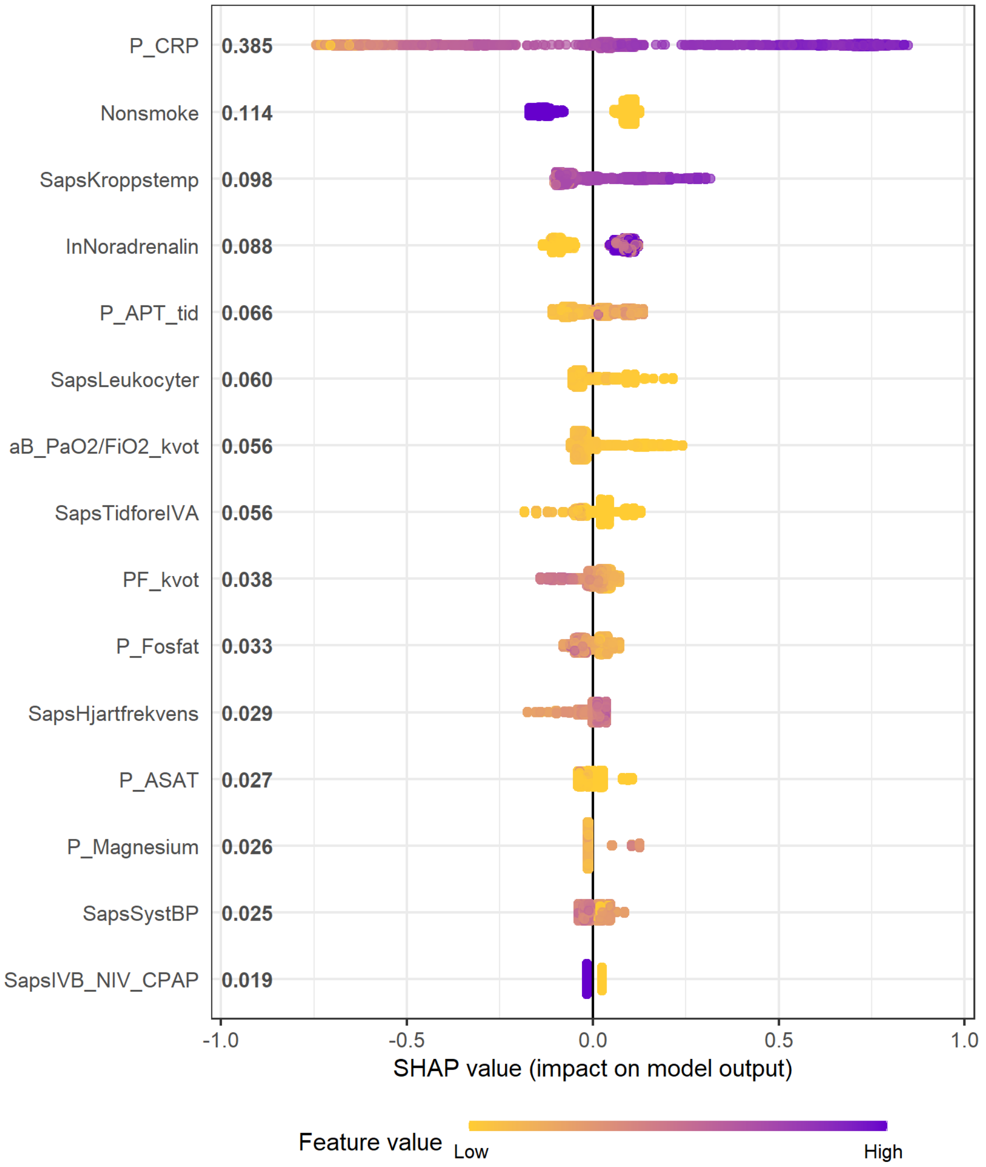


Figure 13: SHAP summary plot of top 5 most important variables the XGBoost model for predicting "TrueSepsis."

Within the SHAP summary plot in figure 13, several components are discernible:

Variables: Each predictor variable in the XGBoost model is displayed along the vertical axis of the plot. Typically, features are arranged in descending order of importance, highlighting the most influential ones at the top.

Feature Importance: The significance of each feature in the model's prediction is depicted through its SHAP value. A higher SHAP value indicates a more substantial impact on the model prediction. The average absolute SHAP value for each variable is indicated adjacent to the variable name on the y-axis.

Feature Effects: This section illustrates how each feature contributes to either elevating or reducing the model's prediction. Positive SHAP values signify contributions towards "True" of the outcome, while negative values indicate contributions towards "False" of the outcome.

Color Gradients: The SHAP values are visualized using varied colors corresponding to the magnitude of each feature value, enabling the visualization of each feature's impact.

Baseline: The baseline value selected for prediction serves as a reference point. In this context, the baseline is established at 0, with two categories set to 1 and -1, respectively, for this particular study. This choice provides context for interpreting the effects and contributions of the features within the model.

In reviewing the SHAP summary plot, the prominence of "P-CRP" remains evident, exhibiting both the widest range and highest magnitude of SHAP values. "SapsKroppstemp" and "InNoradrenalin" continue to hold top positions in terms of SHAP values, albeit with notably lower importance compared to "P-CRP." These findings align with the earlier importance plot and regression results.

Interestingly, "Nonsmoke" occupies a significant position in SHAP values despite its relatively lower importance in the regression model. However, it is pertinent to note that this variable contains over 30% missing data, casting uncertainty on its actual impact within the model.

As a result, the findings from the SHAP value plot correspond harmoniously with the importance values and the logistic regression models, reaffirming the dominance of "P-CRP" within this model. This suggests that the existing data might not suffice for the development of an improved model.

In addition to the SHAP summary plot, the SHAP contribution dependency plot of the XGBoost model will be presented. This visualization illustrates the correlation between a feature's value and its influence on the model's output. These plots serve to elucidate how alterations in a specific feature's value throughout the data set impact predictions, facilitating the validation of model reliability. They offer precise insights into how the magnitude of variables directly influences predictions.

In this study, the dependence plot encompasses solely the top 5 most crucial variables. Given the relatively smaller SHAP values and importance of other variables, their inclusion might not yield substantial analytical significance.

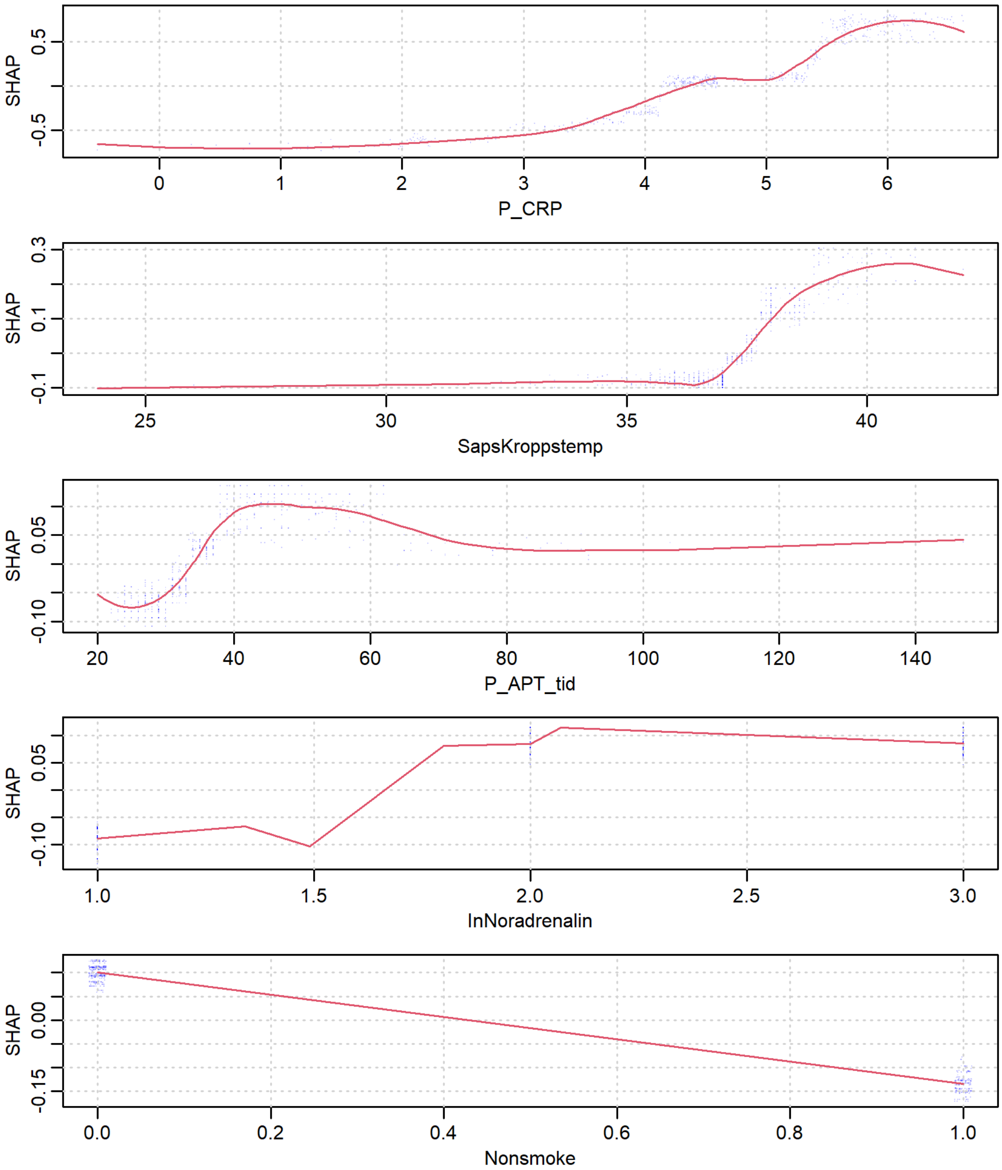


Figure 14: Shap value plot of top 5 important variables used in the XGBoost model for predicting "TrueSepsis."

The dependency plots as in figure 14 for the variables "P-CRP," "SapsKroppstemp," "In-

Noradrenalin," and "P-APT-tid" reveal an overall increasing trend concerning variable values. Although occasional instances of decreasing trends exist, these exhibit relatively modest slopes. It can be inferred that as the values of these four variables increase, the prediction tends toward "True" for "TrueSepsis." These outcomes align with real-world scenarios. For instance, when the value of "SapsKroppstemp" hovers around 37, indicating normal body temperature, its corresponding SHAP value approximates 0. However, an increase beyond 37 leads to a positive SHAP value, shifting the prediction toward "True," signifying the presence of sepsis. This observation corresponds to the SIRS criteria, where a patient displaying a body temperature exceeding 38°Cindicates a potential septic condition.

Conversely, "Nonsmoke," being a categorical variable, is represented as a straight line in the plot. Here, "True" for "Nonsmoke" denotes a patient who is a non-smoker, resulting in a prediction of "False," indicating the absence of sepsis.

As to the performance of the XGBoost model, Modified CV will again be implemented for evaluation. The outcomes are provided in table 8. The mean test-AUC is 0.742 with a 95% confidence interval (0.738, 0.745). The omission of Training-AUC in the table is deliberate due to its variability, spanning from 0.85 to 0.97 with different parameters. Notably, this metric demonstrates significant fluctuation, whereas the test-AUC remains relatively stable despite parameter variations.

Table 8: Results of the modified cross validation for the XGBoost model for predicting "TrueSepsis."

Test-AUC
0.744
0.742
0.734
0.746
0.749
0.740
0.747
0.743
0.736
0.736

When contrasting the logistic regression model with XGBoost models, the latter prove more straightforward to construct due to the incorporated regularization term and automated variable selection. In logistic regression, the process involves manual reduction of models based on diverse criteria, followed by comparisons among the reduced models. While adjusted regression methods such as Lasso regression in [29] and Ridge regression in [30] exist, their application to the data set **D** in this study unfortunately yields suboptimal performance.

Table 9: Performance results of regression and XGBoost models for predicting "TrueSepsis."

	test-AUC	**95% CI**
Stepwise selection	0.716	(0.711, 0.721)
Univariate AUC	0.709	(0.706, 0.711)
XGBoost	0.742	(0.738, 0.745)

As to performance, from table 9, XGBoost exhibits a modest improvement relative to the regression models; nevertheless, it falls short of achieving a significantly favourable outcome. It is essential to acknowledge that, given the existing data set, there are limited avenues for enhancing the model's performance. The prospect of conducting further investigations and achieving improved results would be contingent upon acquiring a more extensive data set.

Even though same as logistic regression, the main contributing variable of the XGBoost is "P-CRP" only, XGBoost's superior performance underscores the advantages of ensemble methods in mitigating variance. However, the enhancement in performance is marginal, primarily due to the dominance of "P-CRP," rendering the model close to a simple logistic regression.

The promising training performance exhibited by XGBoost suggests potential for improved models with larger datasets, both in terms of size and variable count or fewer missing values.

Considering the current findings, the model's diagnostic accuracy is only 1% lower than that of a medical professional, an accuracy that could potentially diminish, as previously mentioned. Nevertheless, the process of training a healthcare professional necessitates a considerable duration. For example, in Sweden, acquiring a medical license necessitates completion of extensive 5 to 6 years of degree studies. Remarkably, a model capable of providing near-comparable diagnostic efficacy within seconds is quite remarkable and effective. However, it is essential to note that while the current results offer valuable insights, they should not be relied upon exclusively. They could possibly serve as a valuable reference aiding diagnostic decision-making processes.

4 Predicting "CultureBloodPositive"

Blood culture tests serve as vital diagnostic tools for identifying the presence of microorganisms, such as bacteria or fungi, within a patient's bloodstream, particularly in the context of sepsis.

Their significance lies in pinpointing the specific pathogen causing the infection, a critical step in sepsis diagnosis. This identification is pivotal as it enables the prescription of precise and effective antibiotic or anti-fungal treatments. Upon the identification of the causative microorganism via blood cultures, healthcare providers possess the capability to customize antibiotic therapy to effectively target the specific pathogen. Timely administration of appropriate medications is imperative in managing sepsis and preventing its escalation to severe stages.

These tests not only facilitate the identification of the responsible microorganism but also aid in determining the most suitable medications, especially in instances where the infection might exhibit resistance to certain antibiotics. As a result, culture blood tests play an integral role in the diagnostic process for sepsis, guiding healthcare providers in administering timely and tailored treatments to patients.

Statistical methods' predictions concerning the culture blood test results aim to further stratify patients and expedite doctors' diagnostic processes.

This section focuses on predicting the blood culture test outcomes, specifically the outcome variable "CultureBloodPositive" within dataset **D**. "CultureBloodPositive" represents a categorical variable, where "True" indicates a positive test result and "False" denotes a negative test result. For the prediction task, logistic regression and XGBoost methodologies will be employed.

It is important to note that the blood culture test results are exclusively documented for septic patients within dataset **D**, comprising a total of 2011 patients, constituting approximately 75% of the entire dataset. The independent variables utilized for prediction will mirror those employed in predicting "TrueSepsis."

4.1 Logistic Regression

4.1.1 Full Model

Consistent with the previous approach, logistic regression is utilised as a benchmark. To commence, the full model encompassing all variables will be presented, followed by subsequent iterations featuring reduced models.

The results of cross validation for the full model are shown in table 10.

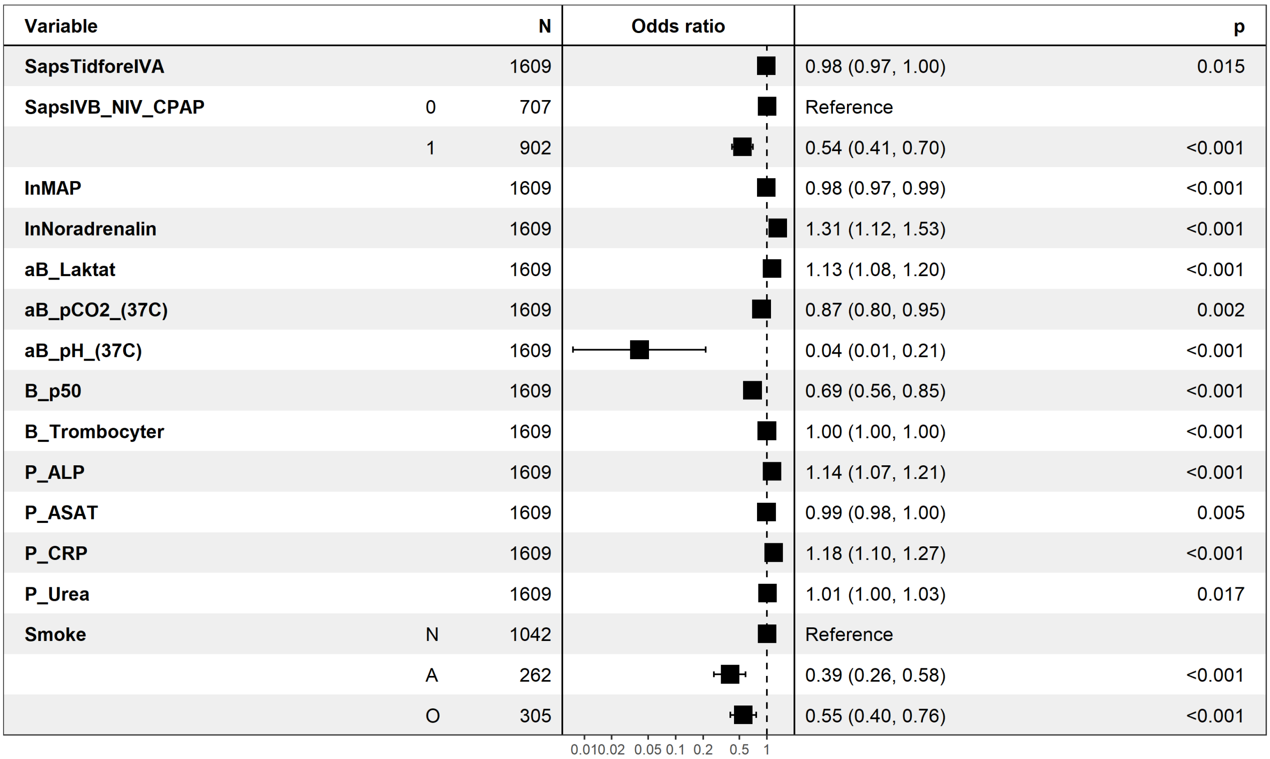
Table 10: 5 fold cross-validation results for the full regression model for predicting "Culture-BloodPositive."

Resample	Test-AUC	Spec	Sens	F1-score
Fold1	0.777	0.485	0.875	0.624
Fold2	0.781	0.522	0.860	0.650
Fold3	0.767	0.492	0.842	0.621
Fold4	0.728	0.507	0.804	0.622
Fold5	0.726	0.507	0.812	0.624

The mean test-AUC stands at 0.756 with a 95% confidence interval (0.723, 0.789). Interestingly, compared to the full regression model for predicting "TrueSepsis," this AUC is marginally higher, despite the smaller dataset size. This slight elevation can be attributed to the greater number of contributing variables in this specific model, a point that will be elucidated further in the subsequent analyses of the reduced models.

4.1.2 Stepwise Selection Based on P-value

In the context of the stepwise selection method, the analysis unfolds as follows: The forest plots in figure 15 reveal the presence of seven variables with notably significant odds ratio values. The model has 14 variables in total and half of them are of high significance. This notable discrepancy compared to the "TrueSepsis" prediction, in which "P-CRP" singularly dominates, suggests the existence of multiple variables holding valuable information. Having more contributing variables signifies a positive development in model prediction, notably reflected in the elevated AUC value, despite the reduction in dataset size.



Variable		N	Odds ratio	p
SapsTidforeIVA		1609	0.98 (0.97, 1.00)	0.015
SapsIVB_NIV_CPAP	0	707	Reference	
	1	902	0.54 (0.41, 0.70)	<0.001
lnMAP		1609	0.98 (0.97, 0.99)	<0.001
lnNoradrenalin		1609	1.31 (1.12, 1.53)	<0.001
aB_Laktat		1609	1.13 (1.08, 1.20)	<0.001
aB_pCO2_(37C)		1609	0.87 (0.80, 0.95)	0.002
aB_pH_(37C)		1609	0.04 (0.01, 0.21)	<0.001
B_p50		1609	0.69 (0.56, 0.85)	<0.001
B_Trombocyter		1609	1.00 (1.00, 1.00)	<0.001
P_ALP		1609	1.14 (1.07, 1.21)	<0.001
P_ASAT		1609	0.99 (0.98, 1.00)	0.005
P_CRP		1609	1.18 (1.10, 1.27)	<0.001
P_Urea		1609	1.01 (1.00, 1.03)	0.017
Smoke	N	1042	Reference	
	A	262	0.39 (0.26, 0.58)	<0.001
	O	305	0.55 (0.40, 0.76)	<0.001

Figure 15: Forest plot of the stepwise selection model for predicting "CultureBloodPositive."

Regarding the Modified Cross Validation result, the mean test-AUC is observed at 0.743, with a 95% confidence interval (0.739, 0.747). Results of the ten tests are shown in table 11. Although marginally lower than the full model, this difference is reasonably accounted for by the reduction in the number of variables utilized in this analysis.

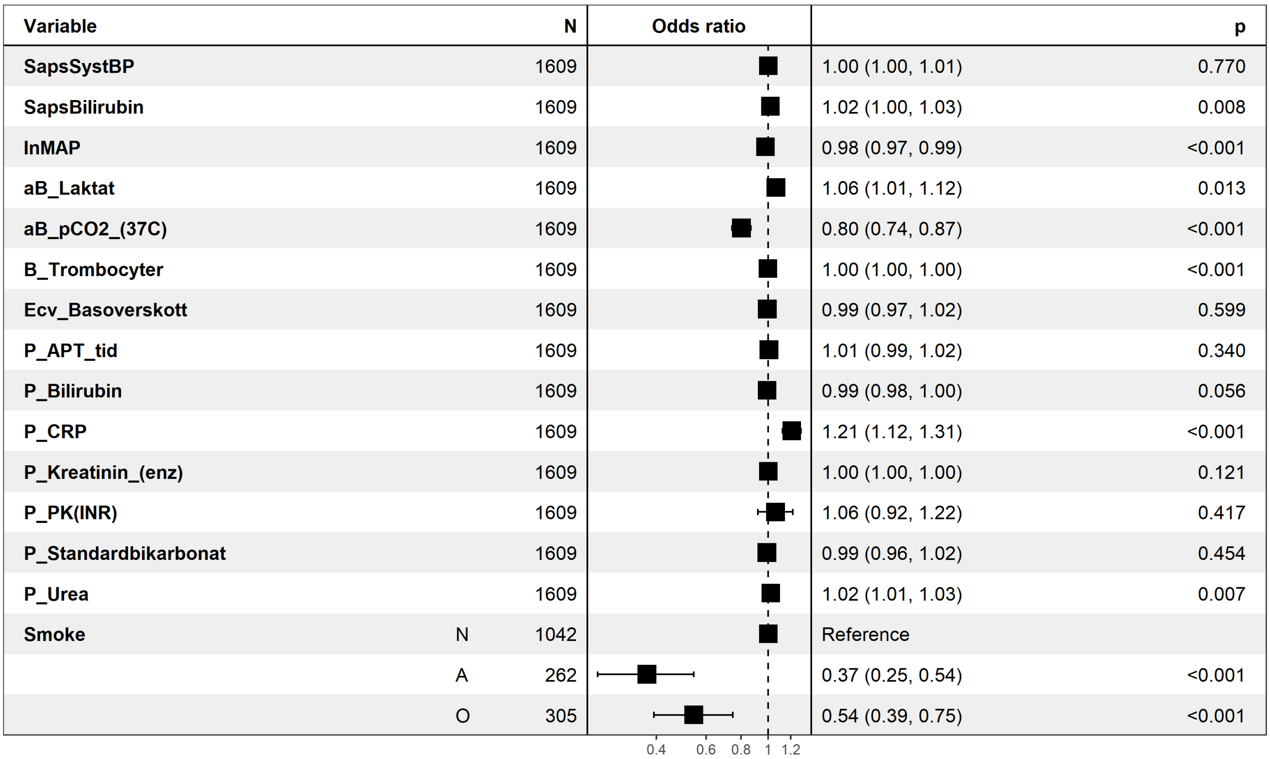
Table 11: Modified cross validation results of the stepwise selection model for predicting "CultureBloodPositive."

Training-AUC	**Test-AUC**
0.798	0.737
0.797	0.746
0.796	0.741
0.796	0.741
0.797	0.747
0.798	0.746
0.797	0.742
0.798	0.752
0.797	0.746
0.798	0.734

4.1.3 Uni-variate AUC

Concerning the uni-variate model, it is noteworthy that its performance and model size closely mirror those of the stepwise selection method.

Nevertheless, in contrast, only three variables within the uni-variate model exhibit substantial odds ratio values. Variables highly significant in the stepwise selection model, such as "InNoradrenalin," "aB-pH-(37C)," and "B-p50," are notably absent from the uni-variate model. This absence may suggest their reliance on combination with other variables to contribute effectively to the prediction.



Variable		N	Odds ratio		p
SapsSystBP		1609		1.00 (1.00, 1.01)	0.770
SapsBilirubin		1609		1.02 (1.00, 1.03)	0.008
lnMAP		1609		0.98 (0.97, 0.99)	<0.001
aB_Laktat		1609		1.06 (1.01, 1.12)	0.013
aB_pCO2_(37C)		1609		0.80 (0.74, 0.87)	<0.001
B_Trombocyter		1609		1.00 (1.00, 1.00)	<0.001
Ecv_Basoverskott		1609		0.99 (0.97, 1.02)	0.599
P_APT_tid		1609		1.01 (0.99, 1.02)	0.340
P_Bilirubin		1609		0.99 (0.98, 1.00)	0.056
P_CRP		1609		1.21 (1.12, 1.31)	<0.001
P_Kreatinin_(enz)		1609		1.00 (1.00, 1.00)	0.121
P_PK(INR)		1609		1.06 (0.92, 1.22)	0.417
P_Standardbikarbonat		1609		0.99 (0.96, 1.02)	0.454
P_Urea		1609		1.02 (1.01, 1.03)	0.007
Smoke	N	1042		Reference	
	A	262		0.37 (0.25, 0.54)	<0.001
	O	305		0.54 (0.39, 0.75)	<0.001

Figure 16: Forest plot of the Uni-variate model for predicting "CultureBloodPositive."

Table 12: Performance results of two regression models for predicting "CultureBloodPositive."

	test-AUC	**95% CI**
Stepwise selection	0.756	(0.723, 0.789)
Univariate AUC	0.748	(0.745, 0.751)

The mean test-AUC from the modified cross-validation is 0.748 with a 95% confidence interval (0.745, 0.751), which is almost the same as the stepwise selection method. Results of the ten tests are given in table 13.

Table 13: Modified cross validation results of the Uni-variate model for predicting "Culture-BloodPositive."

Training-AUC	**Test-AUC**
0.768	0.748
0.767	0.755
0.768	0.747
0.768	0.748
0.768	0.740
0.767	0.751
0.767	0.750
0.767	0.745
0.767	0.748
0.768	0.749

4.2 Machine Learning: XGBoost

In contrast to the predictive model for "TrueSepsis," the XGBoost model for forecasting "CultureBloodPositive" encompasses a notably more extensive set of variables. Observed from the importance plot, the variable "B-Trombocyter" wields the highest influence; however, it is imperative to underscore that numerous other variables exhibit high importance values, coinside with the observation from the logistic regression models. Additionally, it is noteworthy that no variable demonstrates extreme dominance over others, a departure from the scenario observed in models predicting "TrueSepsis."

This phenomenon underscores the presence of valuable information distributed across multiple variables, which, in turn, augments the potential for a more robust model performance.

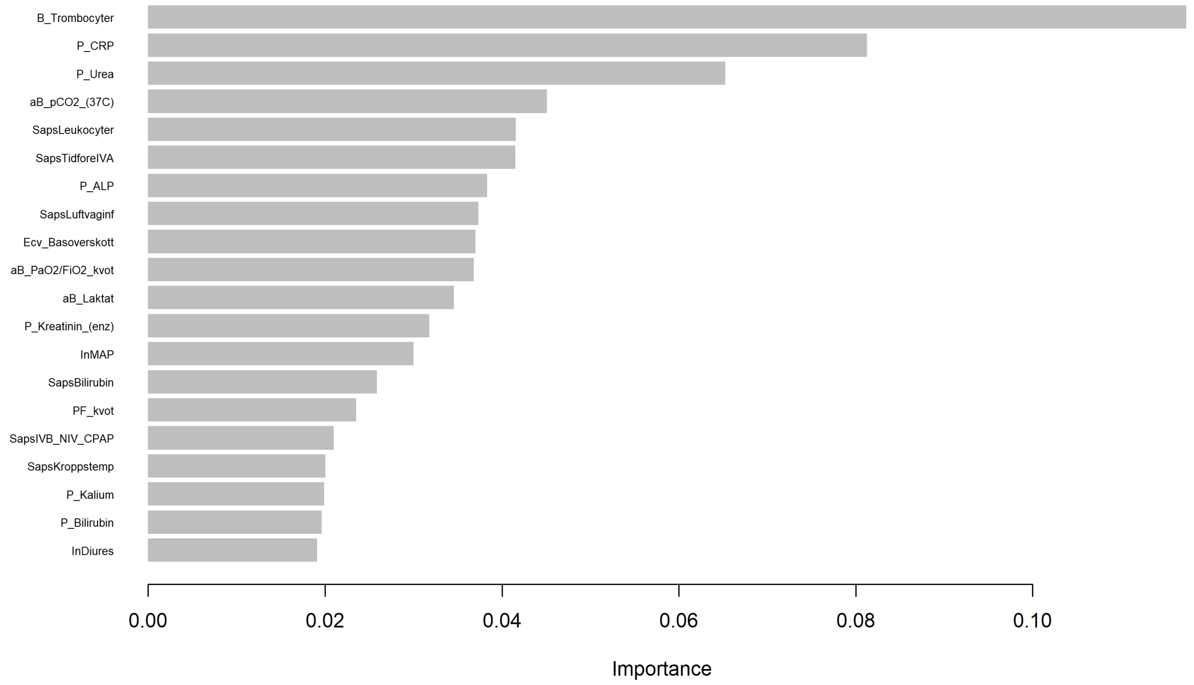


Figure 17: Importance plot of the XGBoost model for predicting "CultureBloodPositive."

The variable "B-Trombocyter" did not exhibit significance in either of the two reduced regression models despite having the highest importance. This discrepancy might stem from its nonlinear relationship with the outcome variable which logistic regression models cound not catch.

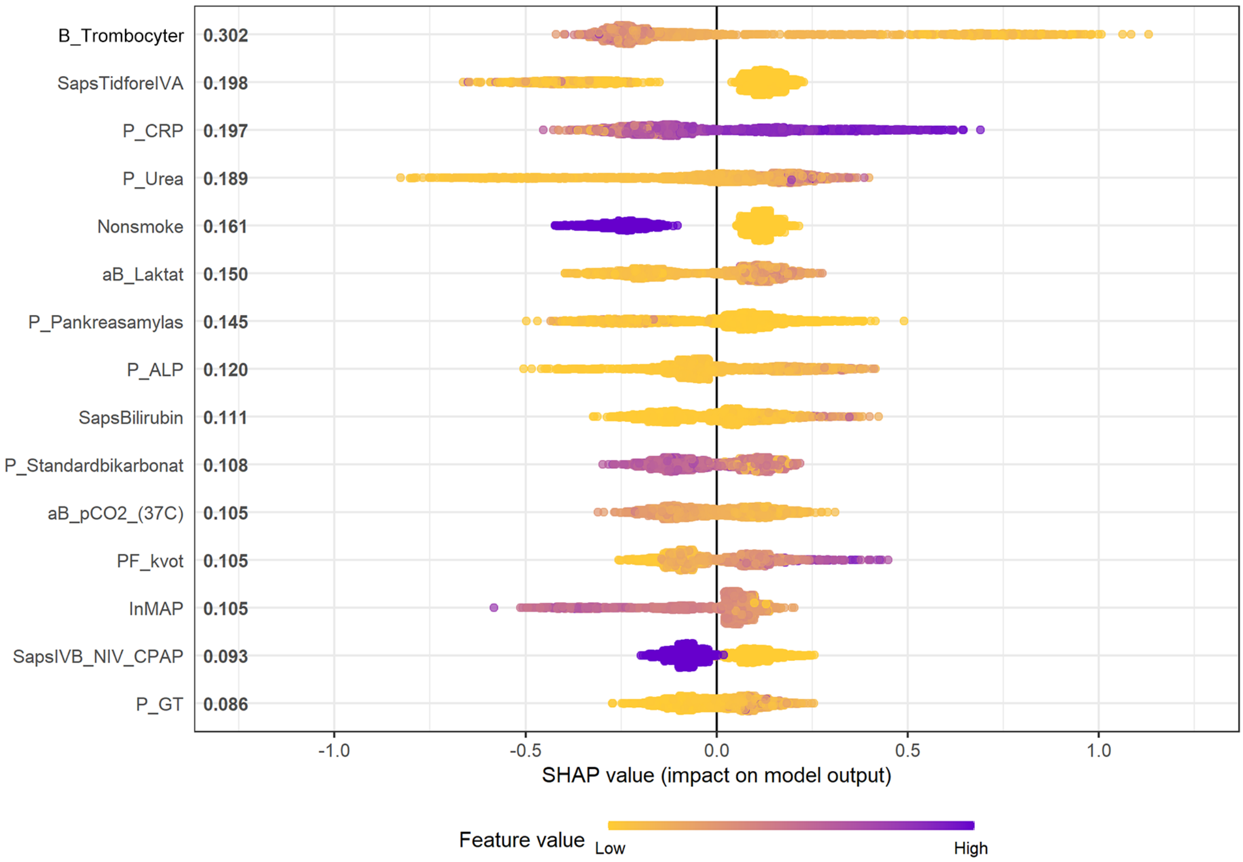


Figure 18: SHAP value summary plot of the XGBoost model for predicting "CultureBloodPositive."

Much like the importance plot, the SHAP value summary plot also indicates the absence of the singular dominant variable, highlighting the significance of numerous variables. The disparity in the magnitude of SHAP values across different variables isn't substantial. These observations, along with insights from regression models, suggest that multiple variables carry valuable information for predicting "CultureBloodPositive." This insight could potentially guide improvements in predicting "TrueSepsis." In addition to expanding the data set by recording data from more patients, augmenting the significance tests could serve to enhance the model's performance.

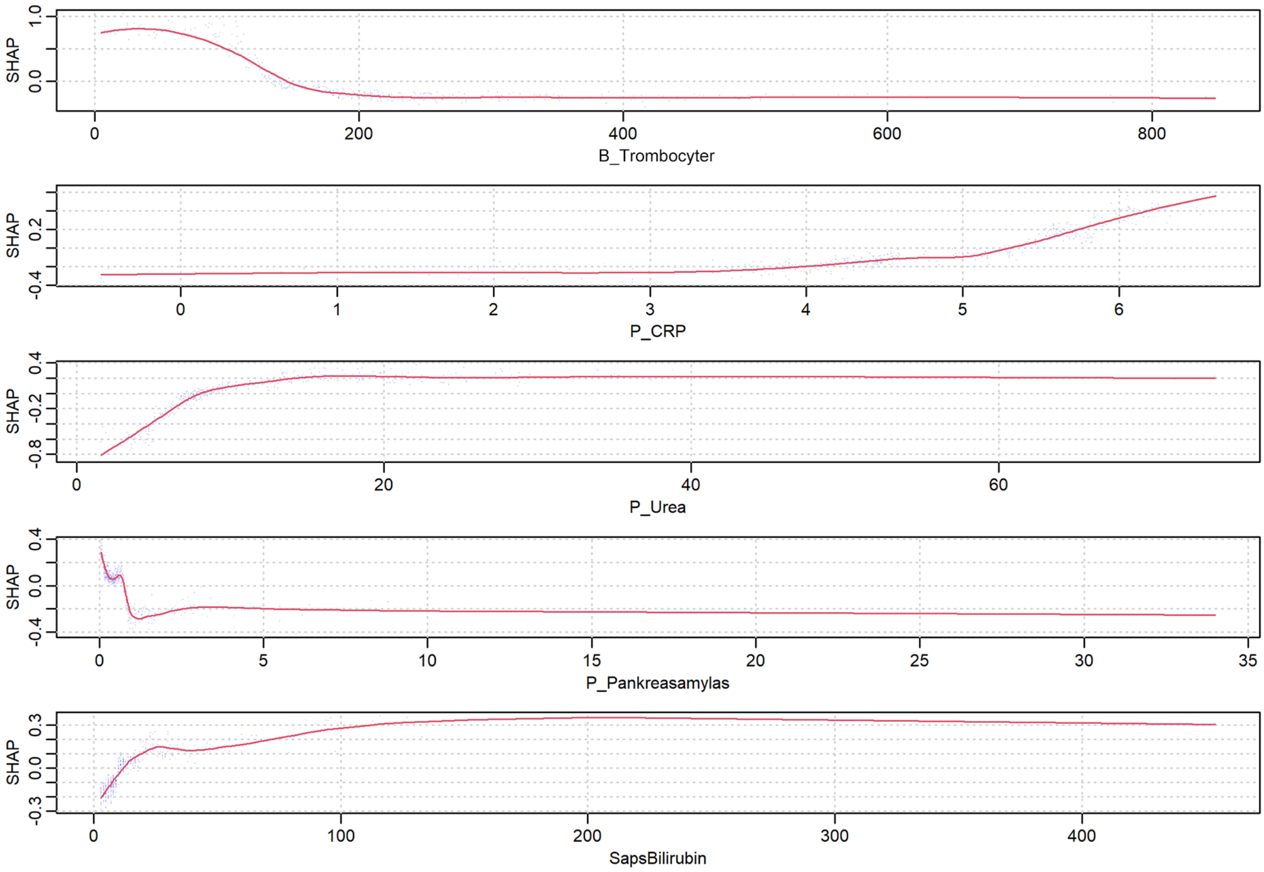


Figure 19: SHAP contribution dependency plot of the top 5 most important variables used in the XGBoost model for predicting "CultureBloodPositive."

The SHAP contribution dependency plot reveals a pattern where the curve levels off before or after a specific point. This observation suggests that the blood culture test result might be influenced when the test outcome deviates from the normal range or exceeds certain thresholds, which aligns with reasonable expectations.

As to the model performance, the average test-AUC obtained through the modified cross-validation is 0.774 with a 95% confidence interval (0.770, 0.777). Detailed results are shown in table 14

Table 14: Modified cross validation results of XGBoost model for predicting "CultureBloodPositive."

Test-AUC
0.773
0.780
0.780
0.779
0.767
0.777
0.772
0.773
0.769
0.766

Table 15: Performance of two XGBoost models

	test-AUC	**95% CI**
Predicting "TrueSepsis"	0.742	(0.738, 0.745)
Predicting "CultureBloodPositive"	0.774	(0.770, 0.777)

From table 15, the performance of the "CultureBloodPositive" prediction slightly surpass those of the preceding XGBoost model. It is crucial to recognize that this prediction pertains exclusively to patients with a "True" value for "TrueSepsis." Consequently, this improved performance is achieved with a smaller the data set size. The observed improvement is evident in both the regression and XGBoost models. It could be inferred that incorporating additional variables with pertinent information would likely elevate the performance of the classification model.

Once again, the performance of the XGBoost model surpassed that of the regression models. As observed previously, adjusting the parameters can elevate the training-AUC of the XGBoost model to approximately 0.85 to 0.97, yet the test-AUC remains consistent. This pattern implies the potential for further enhancement within the XGBoost model. Strategies for improvement will be deliberated upon in the subsequent sections.

5 Model Comparison

Numerous studies have explored the subject matter, and their findings consistently demonstrate superior performance in the prediction models' AUC values. Some investigations have even achieved an impressive area under the curve (AUC) of approximately 0.9 as shown in [3, 9, 12, 13, 14, 31]. This outcome variance can likely be attributed to the disparity in dataset sizes. Our dataset comprises only 2667 patients, whereas most other studies have amassed over 10,000 patient samples. Augmenting current dataset with additional data can enhance our results significantly. Also, in this study, the samples exclusively comprise individuals suspected of "TrueSepsis." In contrast, other research endeavours have utilised datasets encompassing all patients within the ICU, a subset of whom were admitted for reasons unrelated to sepsis, such as severe injuries. Consequently, the inclusion of these non-"TrueSepsis" patients in the dataset has the potential to yield an elevated AUC due to the relatively easier task of distinguishing them from "TrueSepsis" cases.

Of particular concern is the predictive accuracy of "TrueSepsis," as indicated by the importance plots, which reveal that only the "P-CRP" variable exerts a dominant influence on our model. Conversely, previous research, such as [9, 32, 33, 34] highlights the significance of other variables, such as neutrophil to lymphocyte ratio(NLR) and procalcitonin(PCT), in contributing positively to the model's performance. Also, "PCT" and "NLR" exhibit a greater contribution to the predictive modeling of sepsis compared to "P-CRP", as shown in [34, 35]. Regrettably, these variables could not be consider due to their substantial missing data. Only 367 patients in dataset **D** have complete records for all three tests.

Furthermore, it is important to note that most of the variables in data set **D** contain varying degrees of missing values. While imputation techniques have been applied to address these gaps, there will inevitably be disparities between the imputed values and the actual data, potentially impacting the model's overall performance.

Lastly, it is worth acknowledging the inherent complexity in identifying cases of "TrueSepsis" for medical professionals. In dataset **D**, misidentified patients exist, which could adversely affect the model's accuracy and reliability.

In light of these considerations, further research and data augmentation efforts are warranted to improve the robustness and generalisability of models in this study.

6 Conclusions

Within this investigation, predictive models for sepsis diagnosis are developed utilizing both logistic regression and the XGBoost algorithm. The outcomes yielded by these models shown in section 3.2.3 is only 1% lower compared to the accuracy of trained healthcare professionals' judgement. Hence, it is appropriate to assert that the mathematical models demonstrate considerable efficacy, thereby offering valuable assistance to medical professionals in their clinical assessments. It is evident that mathematical methodologies hold promise in aiding the prediction and identification of certain medical conditions. Nevertheless, from the discussion in section 5, other studies produce better results with larger data sets, so the efficacy of these mathematical approaches may be contingent upon the extent and quality of the available data set.

In comparative analyses of prediction models, XGBoost machine learning models have demonstrated a slight performance advantage, which is about 4% higher as shown in section 3.2.3, over logistic regression models. This observed difference aligns with expectations. However, it is important to note that from section 5, the exclusion of two significant variables from this study due to a considerable amount of missing data may limit the predictive capacity of these models, constraining their ability to yield exceptional results.

For the issue of predicting "TrueSepsis," it is advisable for medical practitioners to consider incorporating the Neutrophil-to-Lymphocyte Ratio (NLR) and Procalcitonin (PCT) tests for patients exhibiting suspicious symptoms. This observation arises from the significant correlation between these tests and "TrueSepsis," and they have been proved to be better predictors in modelling of sepsis in comparison to "P-CRP" as shown in [34, 35]. These diagnostic tests are anticipated to make substantial contributions to future investigations on this subject. However, in data set **D**, more than 50% data of these two variables are missing and they were excluded in the data pre-processing step in section 2. Hence, it is recommended for medical practitioners to take these two tests into consideration.

Enhancing both prediction models necessitates the accumulation of more patient data. A refined model will emerge as the dataset accumulates and expands. For future studies, the construction of models on an expanded dataset is advisable. If the research framework necessitates reliance on data from local patients, a requisite duration is essential for the systematic accrual of pertinent data. Conversely, if the research scope allows for a departure from local patient data, alternative open-source datasets such as MIMIC III may be employed. In leveraging external datasets, the strategy could involve utilizing them exclusively to enhance model performance and subsequently applying these refined models to the local patient data for predictive analyses, which means using the richer data sets as training set only and the target local data set as the test set. An alternative approach involves amalgamating the open-source dataset with the local dataset, followed by the development of models through a train-test split methodology.

7 Acknowledgements

I extend my gratitude to Prof. Attila Frigyesi for initiating and creating the foundational project that formed the basis of this thesis. His visionary insights and unwavering support were instrumental in shaping this research.

I am immensely grateful for the patient guidance, invaluable instructions, and scholarly mentorship provided by both Prof. Andreas Jakobsson and Attila Frigyesi. Their collective expertise and encouragement significantly enriched the development and execution of this work. Besides, their patient and dedicated guidance, along with a keen eye for detail, played a crucial role in shaping the research. Both professors' willingness to assist and their thoughtful insights significantly contributed to the quality and depth of this work.

Special appreciation goes to Dr. Maria Lengquist for her invaluable assistance in processing the dataset. Her dedication were pivotal in navigating the complexities of the data, enabling a deeper analysis and understanding.

I would also like to acknowledge the local hospital for generously providing access to the dataset essential for this research.

Last but not the least, I would like to express my sincere appreciation to Lund University for creating a nurturing and inspiring academic environment. The school's commitment to excellence and the provision of a welcoming atmosphere significantly contributed to the success of this study.

References

[1] John Joseph et al. “Sepsis in pregnancy and early goal-directed therapy”. In: *Obstetric Medicine* 2.3 (Sept. 2009), pp. 93–99. DOI: 10.1258/om.2009.090024.

[2] Mercedes Ibarz et al. “The critically ill older patient with sepsis: A narrative review”. In: *Annals of Intensive Care* 14.1 (Jan. 2024), p. 6. DOI: 10.1186/s13613-023-01233-7.

[3] S. M. Brown, J. Jones, K. G. Kuttler, et al. “Prospective evaluation of an automated method to identify patients with severe sepsis or septic shock in the emergency department”. In: *BMC Emerg Med* 16 (2016), p. 31. DOI: 10.1186/s12873-016-0095-0.

[4] *Sepsis*. World Health Organization Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/sepsis.

[5] R. P. Dellinger et al. “Surviving Sepsis Campaign: International Guidelines for Management of Severe Sepsis and Septic Shock: 2012”. In: *Critical Care Medicine* 41.2 (Feb. 2013), pp. 580–637. DOI: 10.1097/CCM.0b013e31827e83af.

[6] Yikai Liu, Ruozheng Wu, and Aimin Yang. “Research on Medical Problems Based on Mathematical Models”. In: *Mathematics* 11.13 (2023). DOI: 10.3390/math11132842.

[7] Hazem Koozi et al. “A simple mortality prediction model for sepsis patients in intensive care”. In: *Journal of the Intensive Care Society* 24.4 (2023), pp. 372–378. DOI: 10.1177/17511437221149572.

[8] DH Lee and BK Lee. “Performance of the simplified acute physiology score III in acute organophosphate poisoning: A retrospective observational study”. In: *Human & Experimental Toxicology* 37.3 (2018), pp. 221–228. DOI: 10.1177/0960327117698541.

[9] C. F. Duncan, T. Youngstein, M. D. Kirrane, et al. “Diagnostic Challenges in Sepsis”. In: *Curr Infect Dis Rep* 23 (2021), p. 22. DOI: 10.1007/s11908-021-00765-y.

[10] Mengmeng Tan et al. “The diagnostic accuracy of procalcitonin and C-reactive protein for sepsis: A systematic review and meta-analysis”. In: *Journal of Cellular Biochemistry* 120.4 (Apr. 2019), pp. 5852–5859. DOI: 10.1002/jcb.27870.

[11] Charlotte Wacker et al. "Procalcitonin as a diagnostic marker for sepsis: A systematic review and meta-analysis". In: *The Lancet Infectious Diseases* 13.5 (May 2013), pp. 426–435. DOI: 10.1016/S1473-3099(12)70323-7.

[12] C. Barton, U. Chettipally, Y. Zhou, et al. "Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs". In: *Comput Biol Med* 109 (2019), pp. 79–84. DOI: 10.1016/j.compbiomed.2019.04.027.

[13] J. Calvert, T. Desautels, U. Chettipally, et al. "High-performance detection and early prediction of septic shock for alcohol-use disorder patients". In: *Ann Med Surg* 8 (2016), pp. 50–55. DOI: 10.1016/j.amsu.2016.04.023.

[14] H.J. Kam and H.Y. Kim. "Learning representations for the early detection of sepsis with deep neural networks". In: *Comput Biol Med* 89 (2017), pp. 248–255. DOI: 10.1016/j.compbiomed.2017.08.015.

[15] Daniel J. Stekhoven and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1 (2012), pp. 112–118. DOI: 10.1093/bioinformatics/btr597.

[16] *missForest Package Documentation*. https://www.rdocumentation.org/packages/missForest/versions/1.5/topics/missForest.

[17] Moo K. Chung. *Introduction to logistic regression*. 2020. DOI: 10.48550/arXiv.2008.13567.

[18] M. Stone. "Cross-Validatory Choice and Assessment of Statistical Prediction (with Discussion)". In: *Journal of the Royal Statistical Society (Series B)* 36 (1974), pp. 111–147. DOI: 10.1111/j.2517-6161.1974.tb00994.x.

[19] Michael Borenstein et al. *Introduction to Meta-Analysis*. John Wiley & Sons, 2009.

[20] R. C. Bone et al. "Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis". In: *Chest* 101.6 (June 1992), pp. 1644–1655. DOI: 10.1378/chest.101.6.1644.

[21] P. Póvoa et al. “C-reactive protein as a marker of infection in critically ill patients”. In: *Clinical Microbiology and Infection* 11.2 (2005), pp. 101–108. ISSN: 1198-743X. DOI: 10.1111/j.1469-0691.2004.01044.x.

[22] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

[23] Jerome H. Friedman. “Greedy function approximation: A gradient boosting machine.” In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451.

[24] Patrik Edén and Mattias Ohlsson. *Introduction to Artificial Neural Networks and Deep Learning*. 2022.

[25] Ulrike von Luxburg and Bernhard Schoelkopf. *Statistical Learning Theory: Models, Concepts, and Results*. 2008. arXiv: 0810.4752 [stat.ML].

[26] Lior Rokach. “Ensemble-based Classifiers”. In: *Artificial Intelligence Review* 33 (2010), pp. 1–39. DOI: 10.1007/s10462-009-9124-7.

[27] D. Opitz and R. Maclin. “Popular Ensemble Methods: An Empirical Study”. In: *Journal of Artificial Intelligence Research* 11 (Aug. 1999), pp. 169–198. ISSN: 1076-9757. DOI: 10.1613/jair.614.

[28] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017). DOI: 10.48550/arXiv.1705.07874.

[29] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178.

[30] Arthur E. Hoerl and Robert W. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. In: *Technometrics* 42.1 (2000), pp. 80–86. ISSN: 00401706. URL: http://www.jstor.org/stable/1271436.

[31] Md. Mohaimenul Islam et al. “Prediction of sepsis patients using machine learning approach: A meta-analysis”. In: *Computer Methods and Programs in Biomedicine* 170 (2019), pp. 1–9. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2018.12.027.

[32] Belal Azab, Marlene Camacho-Rivera, and Emanuela Taioli. “Average values and racial differences of neutrophil lymphocyte ratio among a nationally representative sample of United States subjects”. In: *PLoS One* 9.11 (Nov. 2014), e112361. DOI: 10.1371/journal.pone.0112361.

[33] C. P. de Jager, P. T. van Wijk, R. B. Mathoera, et al. “Lymphocytopenia and neutrophil-lymphocyte count ratio predict bacteremia better than conventional infection markers in an emergency care unit”. In: *Crit Care* 14 (2010), R192. DOI: 10.1186/cc9309.

[34] Lars Ljungström et al. “Diagnostic accuracy of procalcitonin, neutrophil-lymphocyte count ratio, C-reactive protein, and lactate in patients with suspected bacterial sepsis”. In: *PLOS ONE* 12.7 (July 2017), pp. 1–17. DOI: 10.1371/journal.pone.0181704.

[35] Tobias Schupp et al. “C-reactive protein and procalcitonin during course of sepsis and septic shock”. In: *Irish Journal of Medical Science (1971 -)* (May 2023). DOI: 10.1007/s11845-023-03385-8.

Master's Theses in Mathematical Sciences 2024:E11
ISSN 1404-6342

LUNFMS-3124-2024

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/