Increasing the Efficacy of Umbilical Cord Blood Banking Using Machine Learning Algorithms: A Case Study from Royan Cord Blood Bank

Mahdi Haghbayan [a], Behrooz Karimi[b]*, Ashkan Mozdgir [c], Bahareh Abbaspanah [d]

[a] *Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, Tehran, Iran;*

Postal address: Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, 424 Hafez Ave. Tehran, Iran 1591634311

Contact Number: +9386117227

E-mail: m_haghbayan@yahoo.com

[b] *Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, Tehran, Iran;*

Postal address: Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, 424 Hafez Ave. Tehran, Iran 1591634311

Contact Number: +98 9123771542

E-mail: B.Karimi@aut.ac.ir

[c] *Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran;*

Postal address: Kharazmi University No.43 South Mofateh St Tehran, 15719-14911 Iran

Contact Number: +98 (21) 88329220

E-mail: a.mozdgir@khu.ac.ir

[d] *Royan Stem Cell Technology Company, Cord Blood Bank, Tehran, Iran*

Postal address: No. 24, East Hafez Alley, Bani Hashim Square, Bani Hashim Square ST, Sardar Soleimani Highway, Tehran, Iran 1665666311

Contact Number: +98 (21) 27635000

E-mail: b.abbaspanah@gmail.com

***Corresponding author:** Behrooz Karimi

Tel: +98 21 64545374

Mob: +98 9123771542

Fax: +98 21 66954569

Email: B.Karimi@aut.ac.ir

Abstract

Cord blood is the blood that obtains after the birth of a baby. Cord blood is rich in stem cells, which are used to treat a variety of diseases, including cancers and immune disorders. These treatments' effectiveness depends on the quantity of total nucleated cells (TNCs) in cord blood units (CBUs). Both public and private cord blood banks store these CBUs. Public banks rely on government funding for the cost of testing, storing, and maintaining CBUs. In addition, the quantity of TNCs in each CBU remains uncertain until the TNC test is conducted. This study aims to utilize ensemble learning algorithms to aid public banks in identifying and collecting potentially valuable CBUs prior to TNC testing in order to save the cost of TNC testing on CBUs that are not valuable. This study has three main contributions: Firstly, it demonstrates that the XGBoost and LightGBM algorithms can identify CBUs with TNC of more than 0.7×10^9,1×10^9, and $1.5 \times .10^9$; Secondly, the study combines the smote_NC method with Xgboost and LightGBM algorithms and evaluates each algorithm in identifying high TNC samples. Lastly, this article considers the effect of the phlebotomist experience on identifying high TNC samples, a variable overlooked in other studies.

Keywords: Umbilical Cord Blood, Total Nucleated Count, XGBoost, LightGBM, Machine Learning, TNC prediction

1 Introduction

For over 30 years, umbilical cord blood has been used in medicine as a rich source of stem cells to treat 100 indications [1], with about 80 reported in previous studies [2]. However, cord blood has some advantages and also some disadvantages compared to other hematopoietic stem cells (HSCs) sources. The most relevant advantages of UCB as a source of HSCs are readily available once collected and stored, less-precise matching of the donor's human leukocyte antigen (HLA) type to the recipient's HLA type, no risk or pain to the mother or baby, availability, and lower risk of GVHD infection than adult sources [3, 4]. The most crucial disadvantage of umbilical cord blood is the limited number of stem cells in the cord blood unit (CBU), which is typically around ten times lower than bone marrow [5]. Two types of banks have been created to collect and store cord blood units: 1) private banks and 2) public banks. Private banks, also known as family banks, preserve CBU with a link to the identity of the baby. Thus, the family may retake it when they need it. Family banks charge the family to process, test, and cryopreserve the cord blood privately. Public banks preserve donated cord blood for potential use by transplant patients. The CBU is listed in a registry by its tissue type, and the donor remains anonymous. Public banks do not charge parents for donating cord blood. Since the probability of a CBU being appropriate for cord blood unit transplantation is depended on the TNC count, the required tests before cryopreservation and banking the CBU are required. These tests are cost-intensive, and public banks tend to store units that have a high chance of being used [6]. In fact, if public cord blood banks increase the cut-off value of TNC up to 1.5×10^9, the discarded rate of the samples will increase. Therefore, since approximately 75 percent of the samples have TNC below 1.5×10^9 [3], the public cord blood bank incurs a high cost due to the high cost of pre-

storage tests. On the other hand, if the public cord blood bank decreases the cut-off TNC, many stored samples will be inapplicable. Because of these reasons, in 2014, half of the public cord blood banks decided to terminate their activity just two years after they began operating [7]. Therefore, public UCB banks need methods to predict the amount of TNC in the sample before performing the tests to decrease discarded CBUs.

Abbreviations:

UCB	Umbilical cord blood
CBU	Cord blood unit
TNC	Total nucleated cell
GVHD	Graft versus host disease
HLA	human leukocyte antigen
SMOTE_NC	Synthetic Minority Over-sampling TEchnique-Nominal Continuous
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
GOSS	Gradient-Based One Side Sampling
EFB	Exclusive Feature Bundling
ADASYN	Adaptive Synthetic
TP	True positives
FP	False positives
TN	True negatives
FN	False positives
ROC	Receiver operating characteristic
AUC	Area Under the Curve

Machine learning (ML) refers to the usage of computer algorithms that can improve their performance automatically based on experience gained from data. Today, due to the massive amount of data and their complexity, machine learning algorithms have become prevalent, and they are used in various applications to extract important and hidden information. Machine learning algorithms also have many applications in health care to predict diseases such as cancer, heart, lung, and other diseases [8].

Testing the quality of units such as TNC in public cord blood banks can be quite costly. Additionally, the storage of low-quality units can result in further expenses, as most of them are unusable. Therefore, this paper assesses the efficacy of two leading ensemble learning algorithms, XGBoost and LightGBM, in predicting TNC levels of CBUs. By identifying potential samples prior to conducting TNC tests, cord blood banks can avoid unnecessary expenditures and allocate their budget more efficiently. For this purpose, this study compares the performance of two ensemble learning algorithms (XGBoost and LightGBM) in identifying TNC samples that are larger than the cut-off specified cord blood bank. In this article, three cut-offs are examined. The three cut-offs are 0.7×10^9, 1×10^9, 1.5×10^9, which have been used in the literature and international cord blood banks [3, 4, 5, 8, 9].

The two main contributions of this research are as follow: (1) using an ensemble learning algorithm to identify CBU with high TNC in three different cut-offs; (2) Combining the

smote_NC method with Xgboost and LightGBM algorithms and evaluating each algorithm in identifying high TNC samples.

This paper is organized as follows: In Section 2, the literature review is presented. In Section 3, The data and the sets of variables related to UBC are introduced. Then, two algorithms that classify CBUs in terms of TNC are described how these two algorithms' parameters will be tuned are explained. The results of the models and the discussion are presented in Sections 4 and 5. Finally, the conclusion and possible future works are presented in Section 6.

2 Background

In this section, the methods and factors used to predict TNC of umbilical cord blood and bone marrow stem cell, and also techniques that predict the availability of bone marrow donors, are discussed.

Solves et al. [9] identified the cellular dose as the most critical factor limiting the use of stem cells. They used multivariate analysis to identify samples with a cell count of more than 0.8 billion and showed that the variables of sex of newborn, mode of delivery, and weight of placenta would affect cell count. Bouwmeester et al. [10] also introduced the same cell dose as a critical factor for bone marrow stem cell transplantation. They also used multivariable multilevel analysis to predict donor cell dose and considered factors such as age and smoking to affect cell dose prediction. Kristin et al., [11] by using multivariate and univariate analysis methods, examined factors such as gestational age, infant race, infant sex, infant birth weight, maternal age, delivery type, and processing time, and they represented that Among these factors, infant birth weight, race, and sex, as well as processing time, were found to be effective. Manegold-Brauer et al. [12] predicted samples with a cell dose of more than 1.5 billion using a combination of nomogram and multivariate analysis. They first discovered the influencing factors using multivariate analysis, then calculated the sample's probability of at least 1.5 billion TNCs using the nomogram method. Also, factors influencing cell dose were introduced as birth weight and gestational age. Shaoqing Wu et al. [13] introduced factors such as maternal age, birth weight, and vaginal delivery effect on umbilical cord blood stem cell TNC through statistical methods. Reham Al-Qahtani [2] introduced factors such as cord blood volume, birth weight, and method delivery on umbilical cord blood stem cell TNC counts through statistical methods. Lionel Faivre et al. [14] reviewed articles about factors affecting the TNC of umbilical cord blood stem cells. Based on the data collected, they concluded that the fetal weight factor had the greatest effect on cell dose with increasing gestational age. Xinxin Lin et al. [15] showed the effectiveness of placental weight and week of gestation using univariate analysis. But using multivariate analysis, the effect of weeks of gestation was rejected. Since only fetal weight and sex factors affect TNC counts of umbilical cord blood stem cells, they combined a combination of these two factors to define a predictor tool called estimated fetal weight percentile. Using this predictor tool, they improved the banking rate by 30%. Ying Li et al. [16] used machine learning algorithms such as Boosted decision tree, logistic regression, and support vector machine to try to predict access to bone marrow stem cell donors and identify the characteristics of these donors. They showed that one of the main characteristics of donors is the ‘number of days since the last donor contract’. Parmelee Streck et al. [17] used OLS, Lasso, and random forest methods to predict the TNC of

bone marrow stem cells. They demonstrated that the Lasso method has the best performance. Funk et al. [5] predicted samples with a cell dose of more than 1.5×10^9 using the logistic regression method and using maternal and neonatal factors. They indicated that the fetal weight factor has the most significant effect o n the TNC of the samples. In addition, they were able to raise the banking rate from 19.5% to 34.6%.

Since the XGBoost and LightGBM algorithms have not been used in the field of cord blood stem cells before, and on the other hand these two algorithms in other areas of health [18-21], such as prediction of heart disease, have performed effectively; therefore, in this paper, we evaluate these two algorithms.

3 Methodology

In order to predict the TNC of CBUs, Royan cord blood bank data from 1/1/2015 to 1/1/2020 are used as input machine learning data. The feathers of data includes the number of abortions, ethnicity, birth order, gestational age, maternal age, paternal age, and type of delivery. However, infant data such as sex and weight were not accessible for this study.

The methodology used is described below and also shown in Figure. 1:

1. Pre-processing:
   - Convert the response variable from continuous to binary using the specified cut-off
   - Use one hot coding method for categorical variables if using the XGBoost algorithm. For example, if the categorical variable were in 4 categories, then four binary columns would be formed
2. Select 80% of the data as training data and select 20% of the data for test data (in subsets with proportional data quantities for both classes)
3. Data set balancing- SMOTE_NC method
4. Cross-validation of training data-stratified 5-fold
5. Using two algorithms, XGBoost and LightGBM, and evaluating their performance
6. Combine cross-validation and Bayesian optimization algorithm to maximize the AUC value of XGBoost and LightGbm algorithms and find the optimal value of the hyper-parameter of these two algorithms.
7. Evaluation of two algorithms, XGBoost and LightGBM, using test data

INSERT Figure 1 HERE

3.1 Data set balancing

Imbalanced data occurs when the number of instances of the majority class is greater than the number of the minority class, and this phenomenon is widespread in actual samples. This data imbalance is also seen in the TNC data of umbilical cord blood cells in different cut-offs, shown in Fig 2. Data balance is essential because machining learning algorithms may be biased towards the majority class, or in other words, they may have overfitting or underfitting problems. The technique that solves the imbalanced data problem is the Synthetic Minority over-sampling Technique (SMOTE) [22], the modified versions of which include Borderline-SMOTE1, Borderline-SMOTE2, ADASYN [23, 24]. However, the modified version that supports a

combination of continuous and categorical variables is called Synthetic Minority Over-sampling Technique-Nominal Continuous (SNOTE_NC) [25]. In this paper, we evaluated the performance of algorithms in data equilibrium conditions with the SMOTE_NC method and data imbalance conditions.

INSERT Figure 2 HERE

3.2 Machine learning algorithm

In this study, the two algorithms, XGBoost and LightGBM of the boosting algorithms, are evaluated to identify samples whose TNC is above the cut-off. The reasons for choosing these are discussed below.

3.2.1 eXtreme Gradient Boosting (XGBoost)

XGBoost combines Cause Based Decision Tree (CBDT) and Gradient Boosting Machine (GBM) in one effective algorithm [26]. This combination speeds up tree boosting methods and increases the algorithm's accuracy for all data types. Another essential feature of the Xgboost algorithm is that using regularization parameters prevents overfitting[27]. This algorithm has also been used in many Kaggle site competitions [28].

3.2.2 Light Gradient Boosting Machine (LightGBM)

Sergio González et al. [27] evaluated the ensemble learning algorithms and concluded that the LightGBM and XGBoost algorithms performes better than other ensemble learning algorithms. The two Gradient-based One-algorithms Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) are innovative features of the LightGBM algorithm that reduce the size and the number of samples and increase the accuracy and the speed of the LightGBM algorithm. Another advantage of the LightGBM algorithm is supporting categorical features without using feature encoding, such as one-hot encoding [29].

3.3 Hyper-parameter optimization

As mentioned, Xgboost and LightGBM algorithms are composed of many hyper-parameters, which XGBoost algorithm includes learning rate, iteration, max-depth, reg-alpha, reg_lambda, gamma, subsample, colsample_bytree, and LightGBM algorithm includes learning rate, iteration, max-depth, reg-alpha, reg_lambda, num_leaves, bagging_fraction and feature_fraction. Because these two algorithms are composed of several parameters, tuning the parameters is too complicated. Grid search and Bayesian optimization methods can be used to tune the parameters of these two algorithms. Because the Grid search method is used for algorithms that have fewer parameters, the Bayesian optimization method is used to tune the parameters of the two algorithms, XGBoost and LightGBM. The reason for choosing the Bayesian optimization method is that it is suitable for black-box optimization problems [19]. Also, since there is no information available about the derivation of the function of XGBoost and LightGBM algorithms, this method is the most appropriate method for estimating parameters of these two algorithms. This technique uses an approximate objective function instead of the original objective function, called surrogate function. This function uses the Gaussian process. The method also applies another function called acquisition function, which directly samples areas where the probability of improvement is high [30]. In this paper, The upper bound confidence

function is used from three well-known acquisition functions, including Expected improvement, upper bound confidence, and maximum probability of improvement. The Bayesian optimization algorithm is shown below and $D_{1:t-1} = \{(x_1, y_1), ..., (x_{t-1}, y_{t-1})\}$ contains $t-1$ samples taken from the function $f(x)$.

Bayesian optimization algorithm
1: For each t=1, 2, ….,20 2: By maximizing the acquisition function over the surrogate function, find the next sampling point: $x_t = \arg\max_x u\left(x \middle| D_{1:t-1}\right)$ 3: Evaluate the objective function $f(x)$ to obtain the next example: $y_t = f(x_t)$ 4: Add the new sample (x_t , y_t) to the previous samples $D_{1:t} = \left\{D_{1:t-1}, (x_t, y_t)\right\}$ and update the surrogate model.

To obtain the optimal parameters of XGboost and lightGBM algorithms, each time the Bayesian optimization algorithm is executed with a set of parameters, the cross-validation technique is performed on the training data, and the mean AUC score is calculated. After several iterations, the model with the highest mean AUC Score is selected, and this model is used for the test data.

3.4 Performance metrics

Metrics such as accuracy, sensitivity, specificity, F1-score, and area under the curve (AUC) of ROC charts are used to evaluate the proposed methods. Also, another indicator, called banking rate, shows how many cord blood samples remain in the bank after TNC testing. Below are the metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

$$sensivity = \frac{TP}{TP + FN} \quad (2)$$

$$specificty = \frac{TN}{TN + FP} \quad (3)$$

$$precision = \frac{TP}{TP + FP} \quad (4)$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (5)$$

$$banking\ rate = \frac{TP}{TP + TN} \quad (6)$$

Where TP and FP represent the number of samples that are correctly and incorrectly classified TNC above the specified cut-off, respectively. Similarly, TN and FN represent the number of

samples that are correctly and incorrectly classified TNC below the specified cut-off, respectively.

3.5 Software and packages

This paper uses XGBoost, LightGBM, and rBayesianOptimization packages to run XGBoost, LightGBM, and Bayesian optimization algorithms in R software with version 4.1.1 R. In this research, imbalanced-Learn library in Python software version 3.7.3 is used the to implement the SMOTE_NC technique.

4 Experimental results

This study uses XGBoost and LightGBM algorithms to identify CBUs whose TNC exceeds the specified cut-off. The implementation summary of our experiments is as follows:

1- In the first step, the parameters of Xgboost and LightGBM algorithms are calculated by maximizing the AUC function by the Bayesian optimization technique in different cut-offs and in terms of balanced and imbalanced data. The results are shown in Tables 1-4

INSERT Table 1 HERE

INSERT Table 2 HERE

INSERT Table 3 HERE

INSERT Table 4 HERE

2- The TP, TN, FP, FN, and Banking rate values are calculated using test data and optimized models LightGBM, XGBoost, LightGBM_SMOTE_NC, and XGBoost_SMOTE_NC in different cut-offs that are shown in tables 5-7.

INSERT Table 5 HERE

INSERT Table 6 HERE

INSERT Table 7 HERE

3- In order to compare the performance of classifiers, performance metrics are calculated using the values TP, TN, FP, and FN that exist in tables 5-7. The results of these calculations are shown in figures 2-6, and tables 8-10. In addition, to compare the performance of each method for three cut-offs, the AUC curve of each method for different cut-offs is plotted in figures 6-8.

INSERT Table 8 HERE

INSERT Table 9 HERE

INSERT Table 10 HERE

INSERT Figure 3 HERE

INSERT Figure 4 HERE

INSERT Figure 5 HERE

INSERT Figure 6 HERE

INSERT Figure 7 HERE

INSERT Figure 8 HERE

INSERT Figure 9 HERE

INSERT Figure 10 HERE

5 Discussion

In order to discuss the results of the experiment, we first compare the performance of classifiers in each cut-off of samples. Then we compare the performance of models introduced in this article with previous studies from the literature.

5.1 Comparison of the performance of different classifiers in different cut-offs

Based on Table 5, in the cut-off 0.7×10^9 , the XGBoost algorithm has the highest banking rate or precision rate compared to other methods and has improved the banking rate by about 2.6% compared to the case where there is no model. The results indicates that fewer samples are collected by XGBoost algotithem compared to LightGBM and LightGBM_SMOTE_NC methods. Also, the LightGBM method is superior to other methods in terms of AUC, accuracy, sensitivity, and F_score indices (Table 8). In addition, the results demonstrate that combining the XGBoost algorithm with the smote_nc technique improves the detection of samples with TNC less than 0.7×10^9 in exchange for a reduction in the banking rate while combining the smote_nc technique with the LightGBM algorithm reduces the performance of the LightGBM algorithm.

The results from Table 6 and Table 9 show that the most improved banking rate in the cut-off 1×10^9 is related to the XGBoost method. However, the LightGBM method performes better in other indicators such as AUC, accuracy, sensitivity, and F_score. Moreover, the LightGBM method has detected more samples with TNCs above 1×10^9 in exchange for a lower banking rate. Combining the LightGBM algorithm with the SMOTE_NC technique has been able to identify more samples with less than 1×10^9 TNCs. In contrast, combining the XGBoost algorithm with the SMOTE_NC technique has not affected identifying samples with less than 1×10^9 TNCs.

In identifying CBU with a TNC of more than 1.5×10^9according to Tables 7 and 10, the XGBoost banking rate index improves the most compared to other methods, which is 28%. This method is also superior to other methods in other indicators such as AUC, accuracy, and specificity. However, the LightGBM and LightGBM_SMOTE_NC methods are superior in indicators such as sensitivity and F_score and have a more remarkable ability to detect samples

with a TNC greater than1.5×10^9. It also shows that combining the SMOTE_NC technique with the XGBoost and LightGBM algorithms has improved the identification of instances with a TNC higher than 1.5×10^9 and has improved the banking rate relative to the conditions under which no method is used, but this rate has been reduced by combining XGBoost and LightGBM methods with this technique.

According to the XGBoost gain and LightGBM gain obtained from Figures 9 and 10, factors, including maternal age, paternal age, and phlebotomist experience, are the most influential factors on TNC in cord blood samples. In most methods, factors such as ethnicity and gestational age have a moderate effect, and birth order and the number of abrasion factors have a minor effect on TNC.

5.2 Comparison with previous methods

Kristin et al. [31] found that maternal and paternal age are influential factors in predicting the cell dose of cord blood samples. Some studies showed that the birth weight factor has the most significant effect on cell dose prediction. Although the birth weight factor was not analyzed as an essential factor in this study, we were able to improve the banking rate by 28% in the cut-off 1.5×10^9 by analyzing the other factors. Manegold-Brauer et al. [13] and Xinxin Lin et al. [9] investigated the birth weight variable in their research; banking improvement rates were 23% and 30%, respectively. In this study, the experience of phlebotomists is examined, which is not addressed in previous researches. It is found that this factor is influential in the TNC of samples stored in public banks.

6 Conclusion

This paper identifies CBUs whose TNCs are above the cut-off specified by the public Cord Blood Bank before performing TNC quantification tests. We proposed two algorithms, XGBoost and LightGBM, to identify CBUs with a TNC greater than 0.7×10^9,1×10^9, and1.5×10^9. We also evaluated two algorithms using the SMOTE_NC technique in terms of balanced data and imbalanced data. To optimize the hyperparameters of the XGBoost, lightGBM, XGBoost_SMOTE_NC, and LightGBM_SMOTE_NC methods, we proposed the Bayesian optimization method. We found that the most remarkable improvement in the Banking rate is related to the XGBoost method, and the LightGBM method has a better ability to identify samples with TNC greater than 0.7×10^9, 1×10^9 and 1.5×10^9. Determining which method is best for a public cord blood bank depends on the financial situation and the number of samples of the public cord blood bank. When a public cord blood Bank has limited financial resources, it is advisable for them to opt the XGBoost algorithm. This is because it results in a lower number of discarded CBUs following cell dose testing. With a reduced sample selection, it is recommended that public cord blood banks collect samples from individuals of varying races to improve the chance of HLA compliance. However, if the funding for the cord blood bank is sufficient and the cord bank wants to increase the likelihood of HLA compliance, it should use the LightGBM method to collect more samples with the appropriate TNC. At a cut-off 1.5×10^9, the results indicate that the combination of the XGboost or LightGBM algorithm with the SMOTE_NC technique can enable cord blood banks to collect more samples, even though the banking rate may decrease. This effect is particularly significant for the XGboost algorithm. For

future research, it is suggested to evalute these methods with the birth weight factor to determine the effects on improving the banking rate.

References

[1] Dessels, C., Alessandrini, M., and Pepper, M. S. "Factors influencing the umbilical cord blood stem cell industry: An evolving treatment landscape," (in eng), *Stem cells translational medicine,* 7(9), pp. 643-650, (2018).

[2] Al-Qahtani, R., Al-Hedythi, S., Arab, S., Aljuhani, A., et al., "Factor predicting total nucleated cell counts in cord blood units," (in eng), *Transfusion,* 56(9), pp. 2352-2354, Sep 2016.

[3] Kapinos, K. A., Briscombe, B., Gracner, T., et al., "Challenges to the sustainability of the U.S. public cord blood System,". RAND Corporation, (2017).

[4] Bart, T., Boo, M., Balabanova, S., et al., "Impact of selection of cord blood units from the United States and Swiss registries on the cost of banking operations," (in eng), *Transfusion medicine and hemotherapy: offizielles Organ der Deutschen Gesellschaft fur Transfusionsmedizin und Immunhamatologie,* 40(1), pp. 14-20, (2013).

[5] Funk, A., Buechel, J., Huhn, E. A., et al., "Antenatal predictors of stem cell content for successful umbilical cord blood donation," *Archives of Gynecology and Obstetrics,* vol. 304(2), pp. 377-384, (2021).

[6] Park, M., Koh, H., and Lee, Y. H. "Repurposing the public cord blood bank inventory in Korea to enhance cord blood use," (in eng), *Transfus Apher Sci,* 58(3), pp. 332-336, (2019).

[7] Magalon, J., Maiers, M., Kurtzberg, J., et al., "Banking or bankrupting: strategies for sustaining the economic future of public cord blood banks," (in eng), *PLoS One,* vol. 10(12), p. e0143440, (2015).

[8] Abdar, M., Książek, W., Acharya, U. R., et al., " A new machine learning technique for an accurate diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine,* 179, p. 104992, (2019).

[9] Solves, P., Perales, A., Mirabet, V., et al., "Optimizing donor selection in a cord blood bank," (in eng), *Eur J Haematol,* 72(2), pp. 107-112, (2004).

[10] Bouwmeester, W., Fechter, M. M., Heymans, M. W., et al., " Prediction of nucleated cells in bone marrow stem cell products by donor characteristics: a retrospective single centre analysis," (in eng), *Vox Sang,* 98(3), pp. e276-e283, (2010).

[11] Page, K. M., Mendizabal, A., Betz-Stablein, B., et al., "Optimizing donor selection for public cord blood banking: influence of maternal, infant, and collection characteristics on cord blood unit quality," (in eng), *Transfusion,* 54(2), pp. 340-352, (2014).

[12] Manegold-Brauer, G., Borner, B., Bucher, C., et al., "A prenatal prediction model for total nucleated cell count increases the efficacy of umbilical cord blood banking," (in eng), *Transfusion,* 54(11), pp. 2946-2952, (2014).

[13] Wu, S., Xie, G., Wu, J., et al., "Influence of maternal, infant, and collection characteristics on high-quality cord blood units in Guangzhou Cord Blood Bank," (in eng), *Transfusion,* 55 (9), pp. 2158-2167, (2015).

[14] Faivre, L., Chloé, C., Boucher, H., et al., "Associated factors of umbilical cord blood collection quality" *Transfusion,*58, pp. 520–531, (2017).

[15] Lin, X., Torrabadella, M., Amat, L., et al., "Estimated fetal weight percentile as a tool to predict collection of cord blood units with higher cellular content: implications for prenatal selection of cord blood donors," (in eng), *Transfusion,* 58(7), pp. 1732-1738, (2018).
[16] Li, Y., Masiliune, A., Winstone, D., et al., "Predicting the availability of hematopoietic stem cell donors using machine learning," (in eng), *Biol Blood Marrow Transplant,* 26(8), pp. 1406-1413, (2020).
[17] Streck, B.P., Naufal, G., Carrum, G., et al., "Demographic and clinical donor characteristics as predictors of total nucleated cell concentrations in harvested marrow products," *Transplantation and Cellular Therapy,* 27(9), pp. 785.e1-785.e6, (2021).
[18] Vaulet, T., Al-Memar, M., Fourie, H., et al., "Gradient boosted trees with individual explanations: an alternative to logistic regression for viability prediction in the first trimester of pregnancy," *Computer Methods and Programs in Biomedicine,* p. 106520, (2021).
[19] Budholiya, K., Shrivastava, S. K., and Sharma, V. "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences,* (2020).
[20] Gao, H., Ye, Z., Dong, J., et al., "Predicting drug/phospholipid complexation by the lightGBM method," *Chemical Physics Letters,*747, p. 137354, (2020)
[21] Liu, Y., Yu, Z., Chen, C., et al., "Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net," *Analytical Biochemistry,* 609, p. 113903, (2020).
[22] Gök, E. C. and Olgun, M. O. "SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples," (in eng), *Neural Comput Appl,* pp. 1-15, (2021).
[23] Elreedy, D., and Atiya, A. F. "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Information Sciences,* 505, pp. 32-64, (2019).
[24] Haibo, H., Yang, B., Garcia, E. A., et al., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence),* pp. 1322-1328, (2008).
[25] Chawla, N. V., Bowyer, K. W., Hall, L. O., et al., "SMOTE: synthetic minority over-sampling technique," *J. Artif. Int. Res.,* 16(1), pp. 321–357, (2002).
[26] Shehadeh, A., Alshboul, O., Al Mamlook, R. E., et al., "Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, LightGBM, and XGBoost regression," *Automation in Construction,*129, p. 103827, (2021).
[27] González, S., García, S., Del Ser, J., et al., "A practical tutorial on bagging and boosting based ensembles for machine learning: algorithms, software tools, performance study, practical perspectives, and opportunities," *Information Fusion,* 64, pp. 205-237, (2020).
[28] Chen, T. and Guestrin, C. "XGBoost: a scalable tree boosting system," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, (2016). Available: https://doi.org/10.1145/2939672.2939785

[29] Ke, G., Meng, Q., Finley, T., et al., "LightGBM: a highly efficient gradient boosting decision tree," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, (2017).
[30] Wu, J., Chen, X.Y., Zhang, H., et al., "Hyperparameter optimization for machine learning models Based on bayesian optimization," *Journal of Electronic Science and Technology*, 17(1), pp. 26-40, (2019).
[31] Page, K. M., Mendizabal, A., Betz-Stablein, B., et al., "Optimizing donor selection for public cord blood banking: influence of maternal, infant, and collection characteristics on cord blood unit quality," (in eng), *Transfusion*, 54(2), pp. 340-352, (2014).

Mahdi Haghbayan received his BS degree in Eco Insurance Management from Allameh Tabataba'e University and his M.S. degree in Industrial Engineering and Management Systems at Amirkabir University of Technology, in 2018 and 2021, respectively. Currently, he is conducting data analysis on insurance policies using data mining and process mining techniques at an insurance company. His research interests include machine learning, data mining, process mining, simulation, and optimization.

Dr. Behrooz Karimi is a tenured professor at the Faculty of Industrial Engineering and Management Systems at Amirkabir University of Technology. His education includes a bachelor degree in Industrial Engineering from Amirkabir University of Technology, a master degree in Industrial Engineering from Iran University of Science and Technology, and a Ph.D. in Supply Chain Management from Amirkabir University of Technology. The main expertise of Professor Karimi is logistics engineering and supply chain management.

Dr. Ashkan Mozdgir is Assistant Professor of Industrial Engineering at Faculty of Engineering at Kharazmi University of Tehran. He has a PhD in industrial engineering (2018) from K.N.T university of technology, Iran. His research interests include Healthcare management, Supply chain design, and Applications of data science in industry.

Bahareh Abbaspanah is PhD Candidate in Cellular & Molecular Biology. She has over one decade experience in stem cell technology and cell therapy.

Figures Captions

Figure 1. Block diagram of the proposed methods

Figure 2. The number of instances in each class in different cut-offs. The cut-offs are in the billions

Figure 3. Performance evaluation of classifiers for cut-off 0.7×10^9

Figure 4. Performance evaluation of classifiers for cut-off 1×10^9

Figure 5. Performance evaluation of classifiers for cut-off 1.5×10^9

Figure 6. ROC charts in different methods in Cut off 0.7×10^9

Figure 7. ROC charts in different methods in Cut off 1×10^9

Figure 8. ROC charts in different methods in Cut off 1.5×10^9

Figure 9. (A) XGBoost gain at cut-off 0.7×10^9 and in imbalanced data conditions. (B) XGBoost gain at cut-off 0.7×10^9 and in balanced data conditions. (C) XGBoost gain at cut-off $1 \times 10^9$and in imbalanced data conditions. (D) XGBoost gain at cut-off $1 \times 10^9$and in balanced data conditions. (E) XGBoost gain at cut-off 1.5×10^9 and in balanced data conditions. (F) XGBoost gain at cut-off 1.5×10^9 and in imbalanced data conditions.

Figure 10. (A) LightGBM gain at cut-off $0.7 \times 10^9$and in imbalanced data conditions. (B) LightGBM gain at cut-off 0.7×10^9 and in balanced data conditions. (C) LightGBM gain at cut-off 1×10^9 and in imbalanced data conditions. (D) LightGBM gain at cut-off 1×10^9 and in balanced data conditions. (E) LightGBM gain at cut-off $1.5 \times 10^9$and in balanced data conditions. (F) LightGBM gain at cut-off 1.5×10^9 and in imbalanced data conditions.

Tables Captions

Table 1. XGBoost optimized parameters using Bayesian optimization in various cut-offs ($\times 10^9$)

Table 2. XGBoost and Smote_NC optimized parameters using Bayesian optimization in various cut-offs ($\times 10^9$)

Table 3. LightGBM optimized parameters using Bayesian optimization in various cut-offs ($\times 10^9$)

Table 4. LightGBM and Smote_NC optimized parameters using Bayesian optimization in various cut-offs ($\times 10^9$)

Table 5. UCB banking rates after application of ensemble learning algorithm (Cut-off=0.7×10^9)

Table 6. UCB banking rates after application of ensemble learning algorithm (Cut-off=1×10^9)

Table 7. UCB banking rates after application of ensemble learning algorithm (Cut-off=1.5×10^9)

Table 8. Performance evaluation of classifiers for cut-off 0.7×10^9

Table 9. Performance evaluation of classifiers for cut-off 1×10^9

Table 10. Performance evaluation of classifiers for cut-off 1.5×10^9

Figures

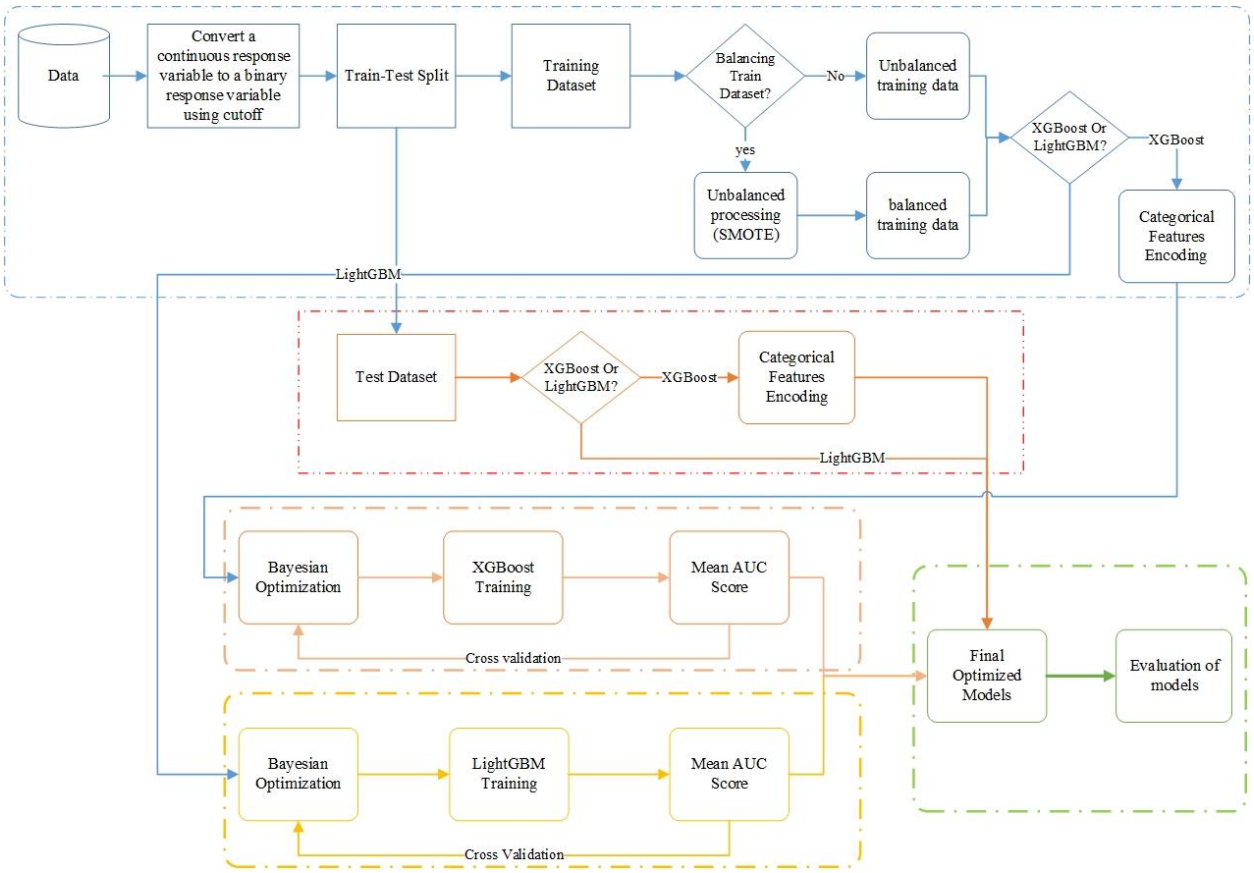

Figure 11.

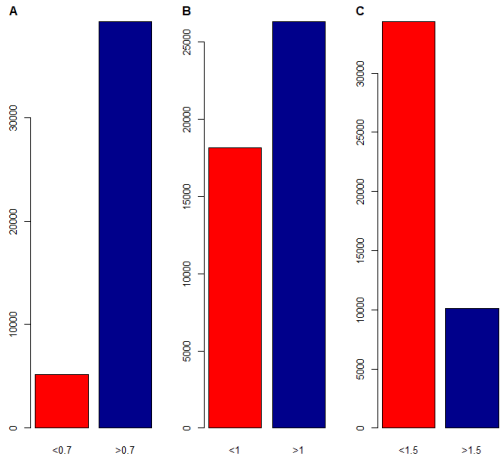

Figure 12.

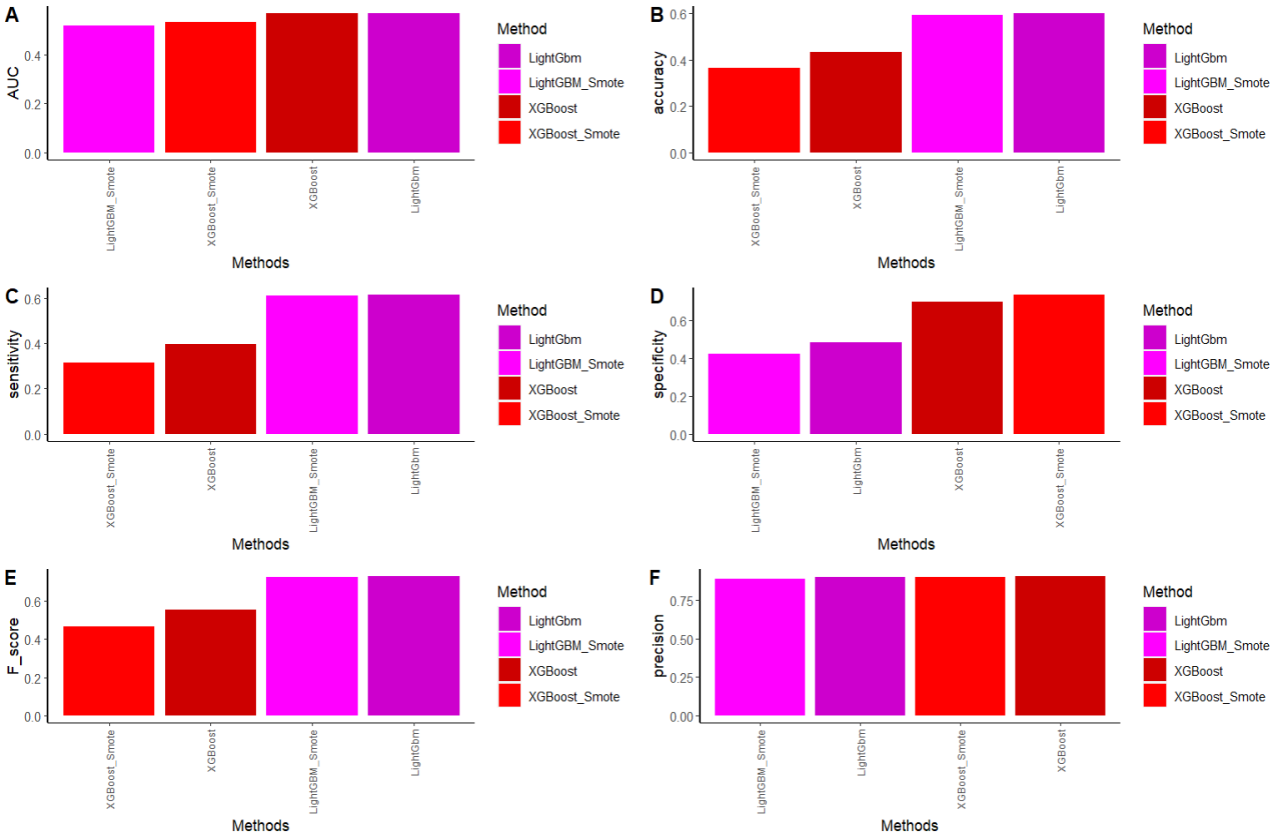
A
AUC
0.0
0.2
0.4
LightGBM_Smote
XGBoost_Smote
XGBoost
LightGbm
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
B
accuracy
0.0
0.2
0.4
0.6
XGBoost_Smote
XGBoost
LightGBM_Smote
LightGbm
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
C
sensitivity
0.0
0.2
0.4
0.6
XGBoost_Smote
XGBoost
LightGBM_Smote
LightGbm
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
D
specificity
0.0
0.2
0.4
0.6
LightGBM_Smote
LightGbm
XGBoost
XGBoost_Smote
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
E
F_score
0.0
0.2
0.4
0.6
XGBoost_Smote
XGBoost
LightGBM_Smote
LightGbm
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
F
precision
0.00
0.25
0.50
0.75
LightGBM_Smote
LightGbm
XGBoost_Smote
XGBoost
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote

Figure 13.

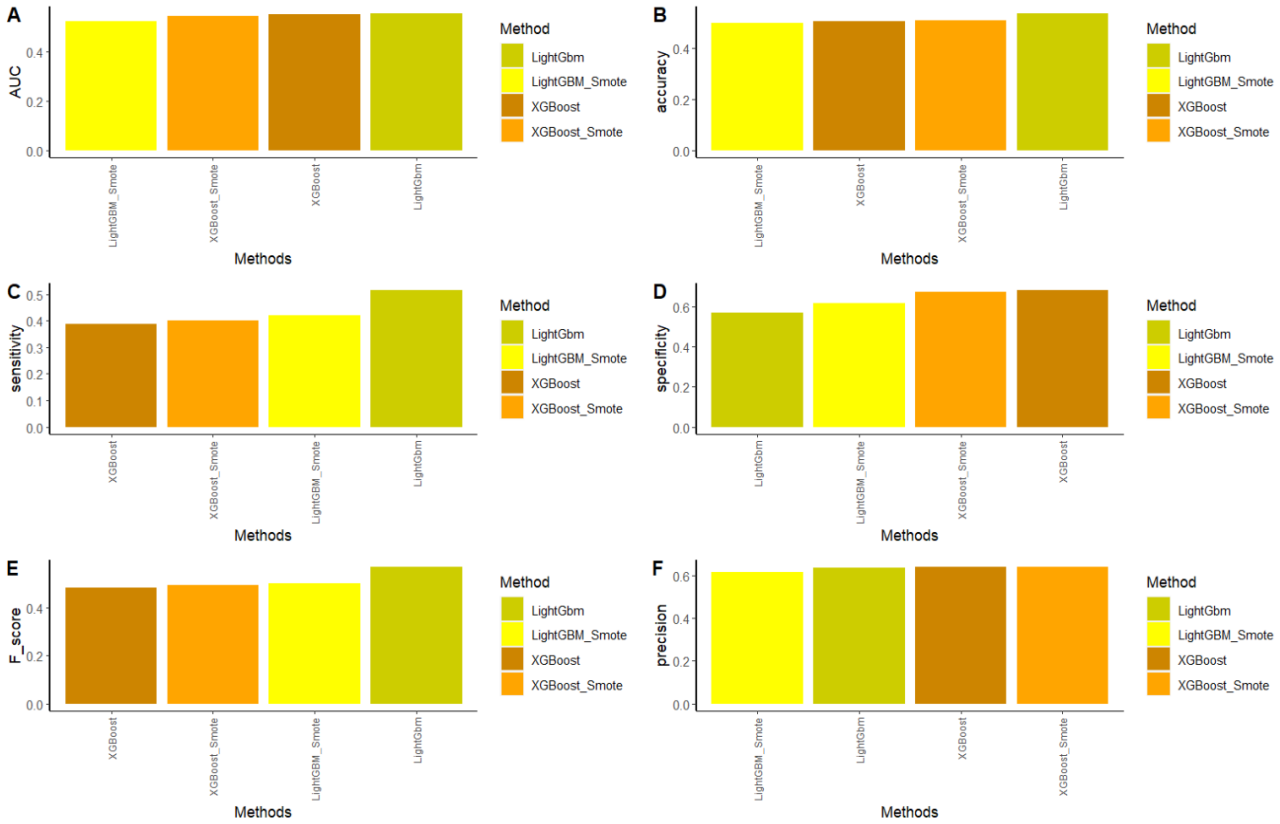
A
AUC
0.0
0.2
0.4
LightGBM_Smote
XGBoost_Smote
XGBoost
LightGbm
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
B
accuracy
0.0
0.2
0.4
LightGBM_Smote
XGBoost
XGBoost_Smote
LightGbm
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
C
sensitivity
0.0
0.1
0.2
0.3
0.4
0.5
XGBoost
XGBoost_Smote
LightGBM_Smote
LightGbm
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
D
specificity
0.0
0.2
0.4
0.6
LightGbm
LightGBM_Smote
XGBoost_Smote
XGBoost
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
E
F_score
0.0
0.2
0.4
XGBoost
XGBoost_Smote
LightGBM_Smote
LightGbm
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
F
precision
0.0
0.2
0.4
0.6
LightGBM_Smote
LightGbm
XGBoost
XGBoost_Smote
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote

Figure 14.

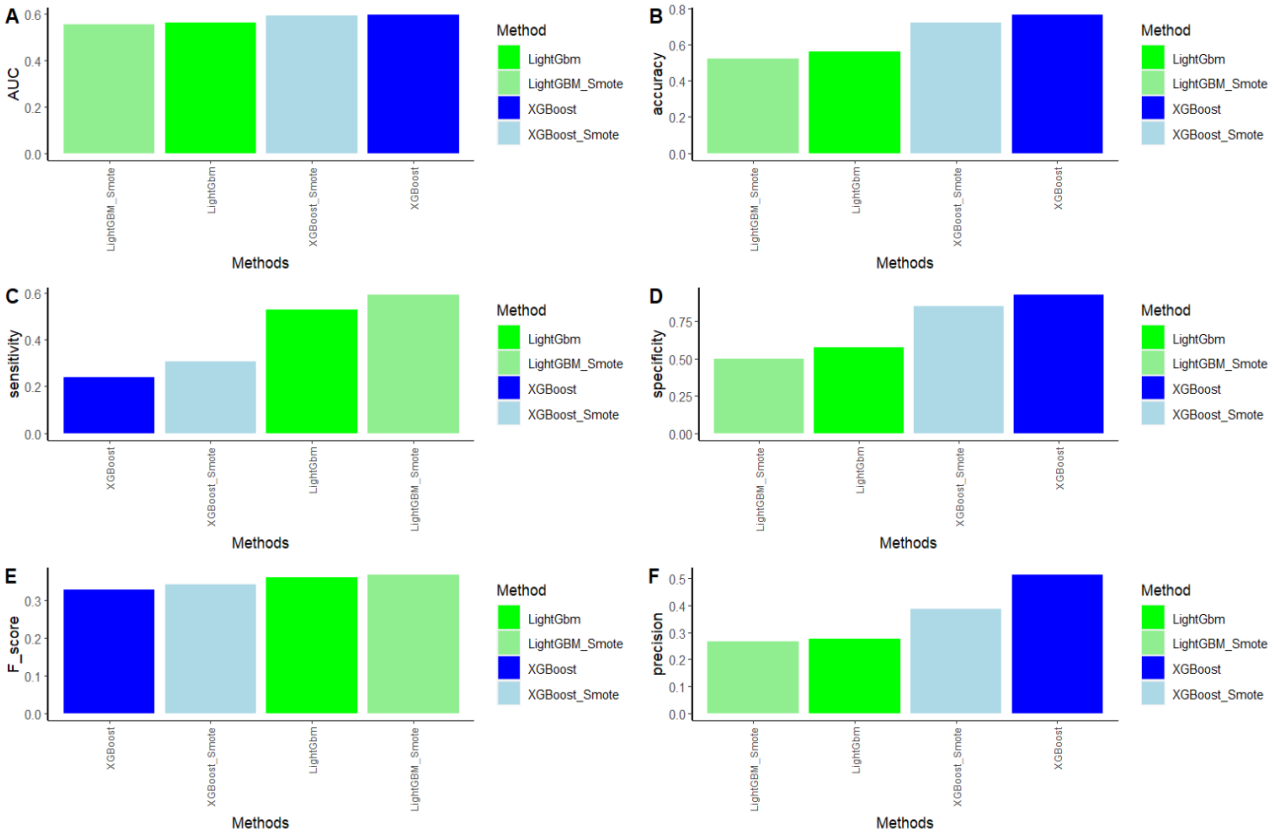
A
AUC
0.6
0.4
0.2
0.0
LightGBM_Smote
LightGbm
XGBoost_Smote
XGBoost
Methods
Method
LightGbm
LightGBM_Smote
XGBoost
XGBoost_Smote
B
accuracy
0.8
0.6
0.4
0.2
0.0
C
sensitivity
D
specificity
0.75
0.50
0.25
0.00
E
F_score
0.3
0.2
0.1
F
precision
0.5
0.4

Figure 15.

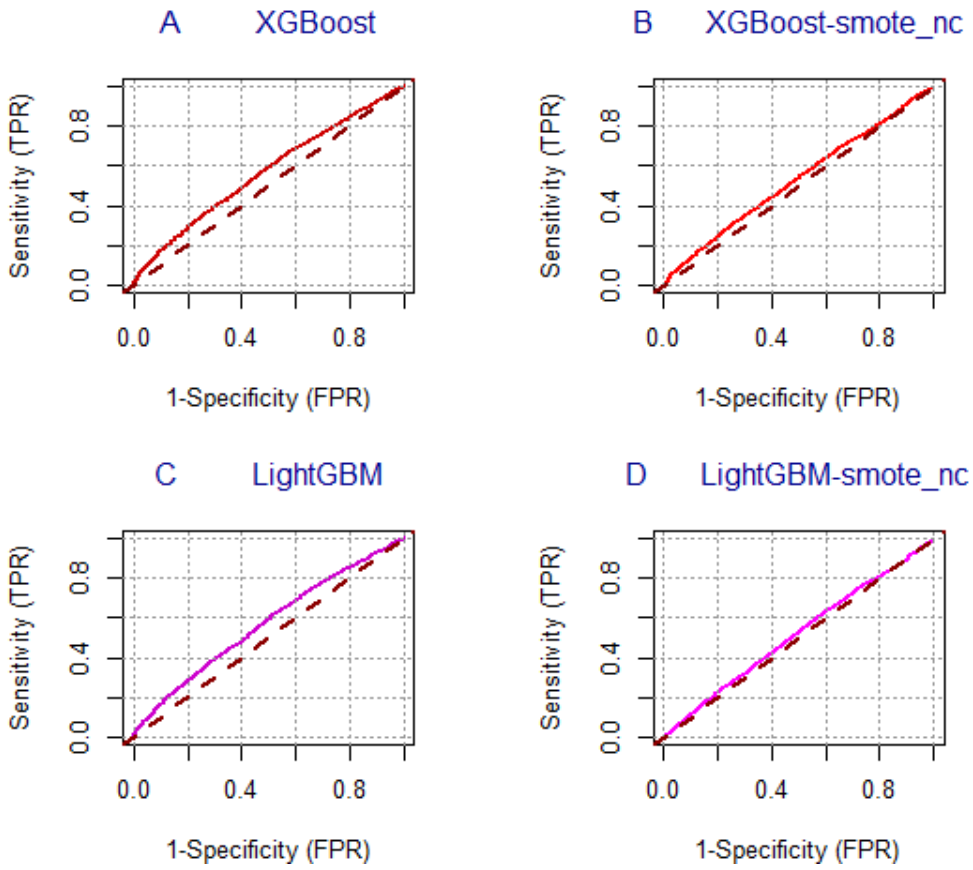
A XGBoost
B XGBoost-smote_nc
C LightGBM
D LightGBM-smote_nc
Sensitivity (TPR)
1-Specificity (FPR)
0.0
0.4
0.8

Figure 16.

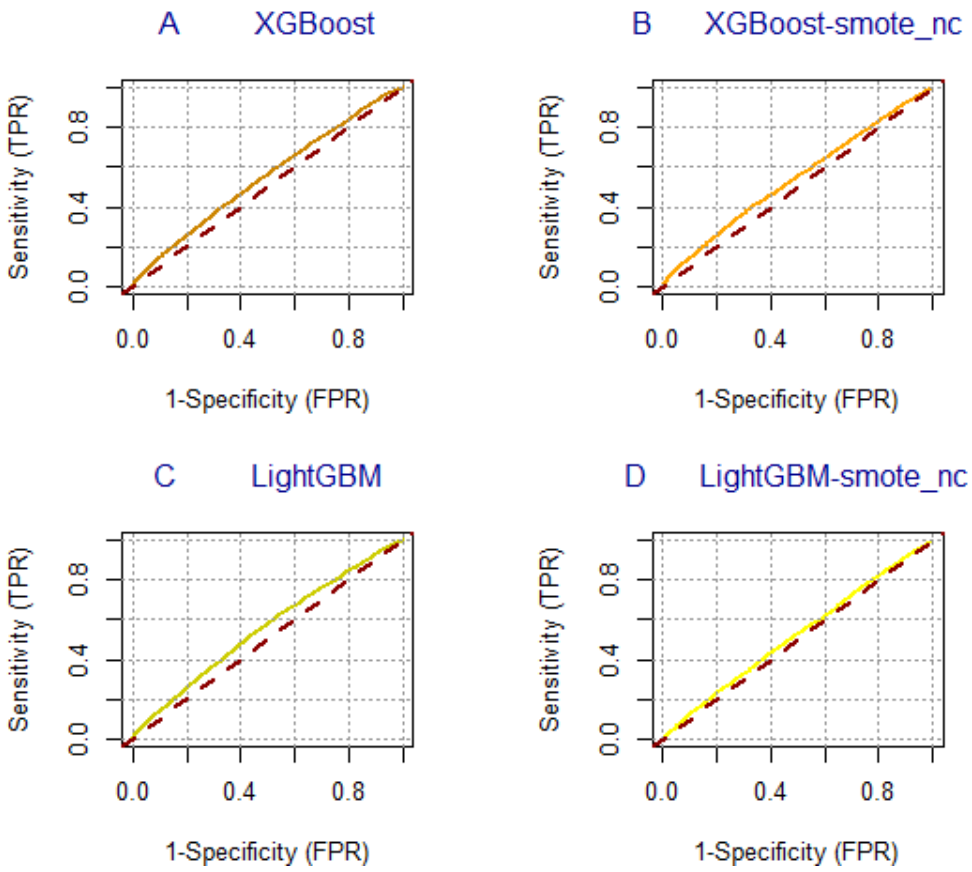
A XGBoost
B XGBoost-smote_nc
C LightGBM
D LightGBM-smote_nc
Sensitivity (TPR)
1-Specificity (FPR)
0.0
0.4
0.8

Figure 17.

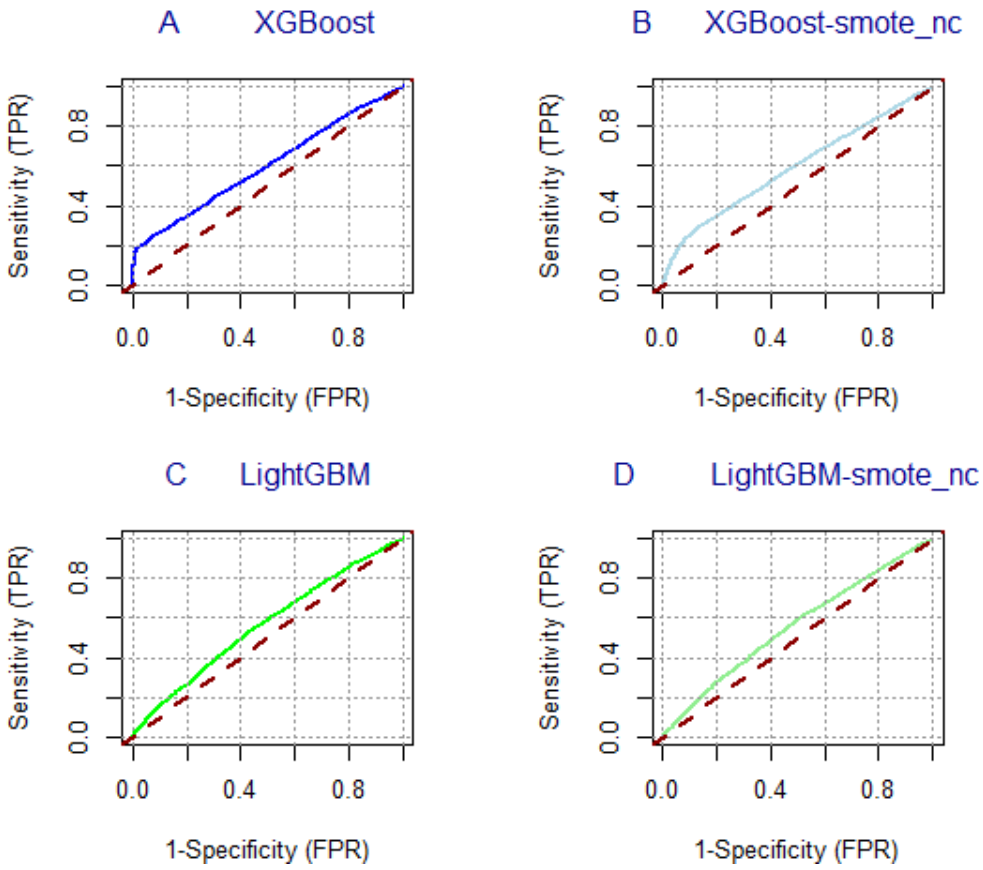
A XGBoost
B XGBoost-smote_nc
C LightGBM
D LightGBM-smote_nc
Sensitivity (TPR)
1-Specificity (FPR)
0.0
0.4
0.8

Figure 18.

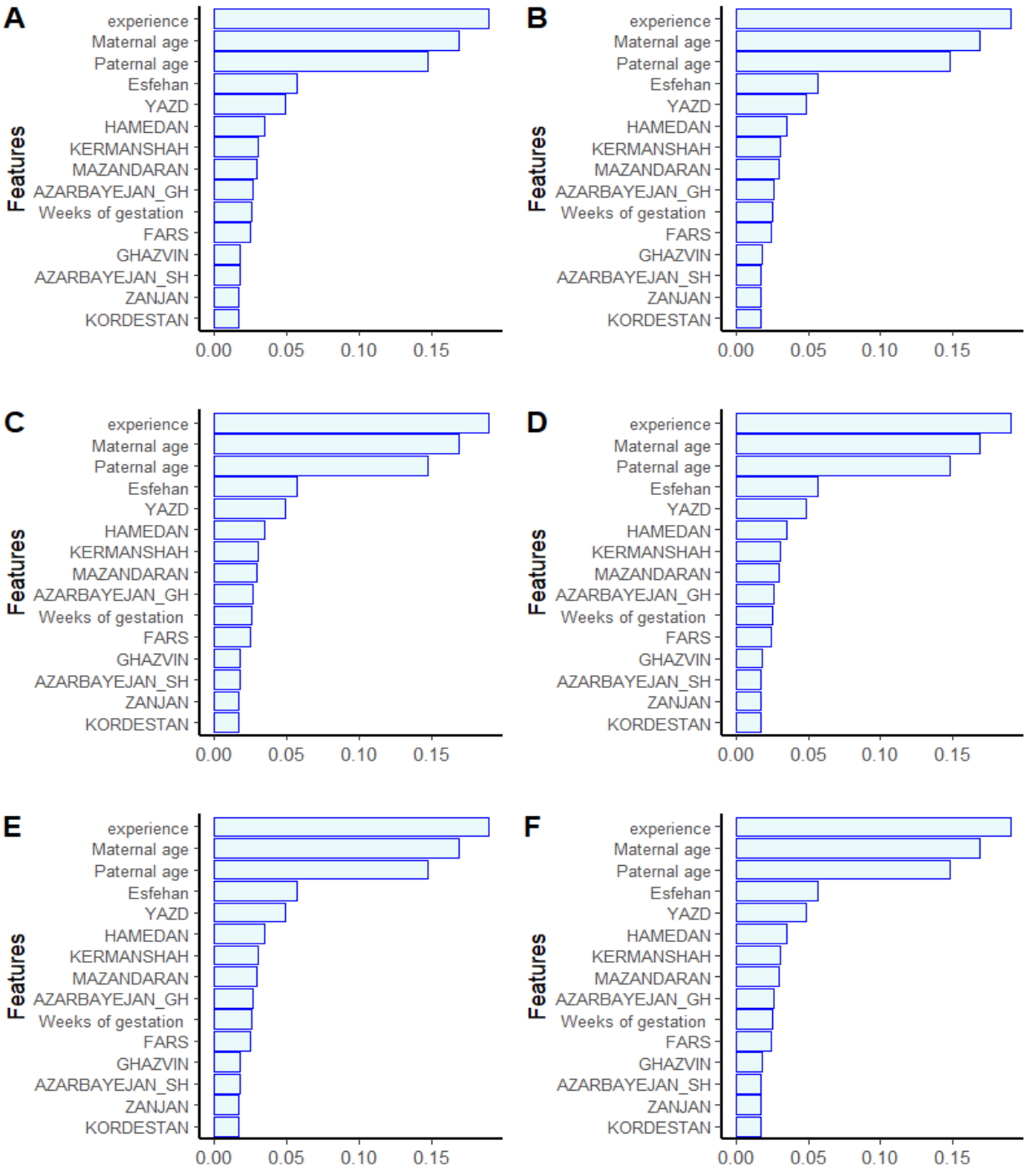
A
B
C
D
E
F
Features
experience
Maternal age
Paternal age
Esfehan
YAZD
HAMEDAN
KERMANSHAH
MAZANDARAN
AZARBAYEJAN_GH
Weeks of gestation
FARS
GHAZVIN
AZARBAYEJAN_SH
ZANJAN
KORDESTAN
0.00
0.05
0.10
0.15

Figure 19.

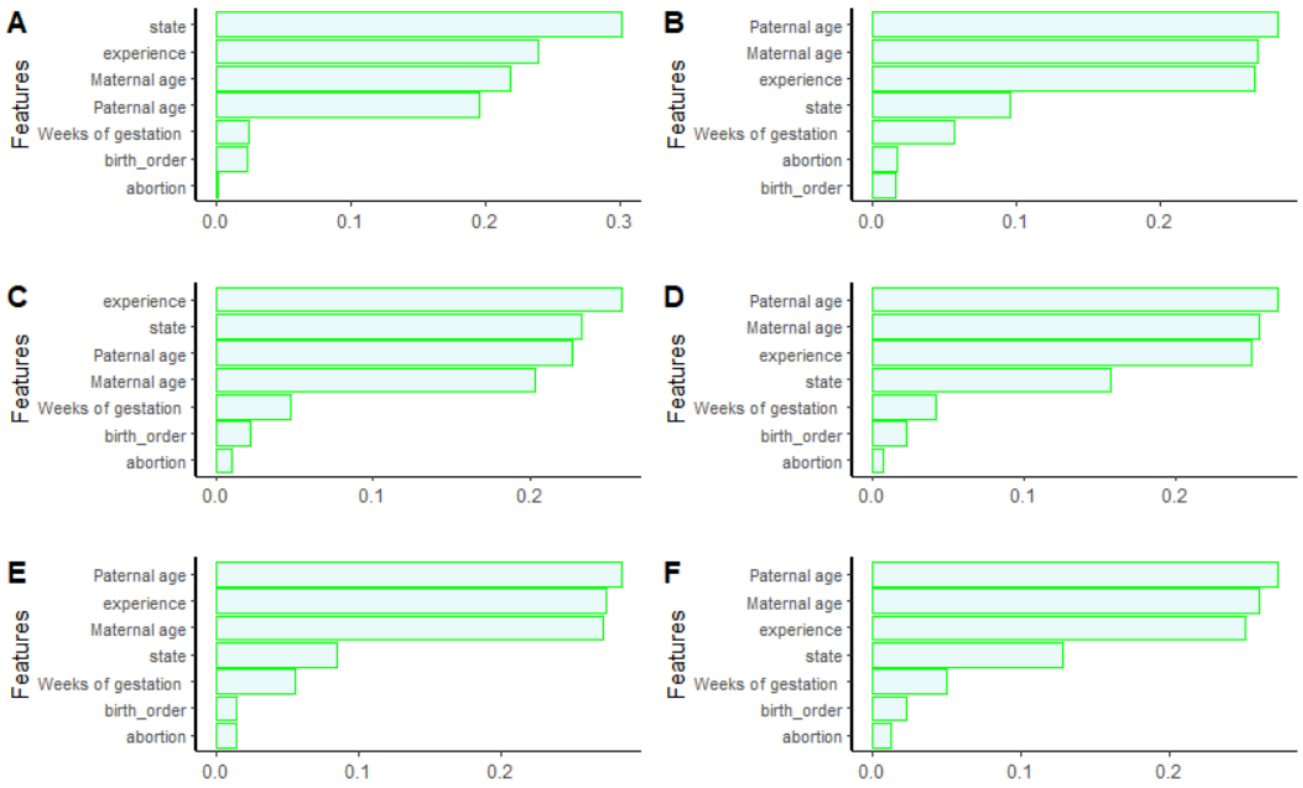
A
state
experience
Maternal age
Paternal age
Weeks of gestation
birth_order
abortion
Features
0.0
0.1
0.2
0.3
B
Paternal age
Maternal age
experience
state
Weeks of gestation
abortion
birth_order
Features
0.0
0.1
0.2
C
experience
state
Paternal age
Maternal age
Weeks of gestation
birth_order
abortion
Features
0.0
0.1
0.2
D
Paternal age
Maternal age
experience
state
Weeks of gestation
birth_order
abortion
Features
0.0
0.1
0.2
E
Paternal age
experience
Maternal age
state
Weeks of gestation
birth_order
abortion
Features
0.0
0.1
0.2
F
Paternal age
Maternal age
experience
state
Weeks of gestation
birth_order
abortion
Features
0.0
0.1
0.2

Figure 20.

Tables

Table 11

Cut-off	Learning rate	iteration	Max_depth	reg_alpha	reg_lambda	gamma	subsample	colsample_bytree	AUC
0.7	0.6	936	10	60	8.86	0.121	0.82	0.761	0.56
1	0.097	308	11	60	45.4	0.021	0.7	0.7	0.54
1.5	0.11	599	19	0	0	0.01	1	1	0.57

Table 12

Cut-off	Learning rate	iteration	Max_depth	reg_alpha	reg_lambda	gamma	subsample	colsample_bytree	AUC
0.7	0.51	928	20	0	60	0.01	0.82	1	0.89
1	0.005	1000	190	0	0	0.255	0.852	0.967	0.63
1.5	0.18	1000	20	3.03	0	0.01	0.809	1	0.82

Table 13

Cut-off	Learning rate	iteration	Max_depth	reg_alpha	reg_lambda	num_leaves	bagging_fraction	feature_fraction	AUC
0.7	0.162	629	7	49.69	49.03	28	0.718	0.71	0.56
1	0.005	300	23	0	60	40	0.889	0.9	0.543
1.5	0.005	900	15	0	0	40	0.9	0.7	0.555

Table 14

Cut-off	Learning rate	iteration	Max_depth	reg_alpha	reg_lambda	num_leaves	bagging_fraction	feature_fraction	AUC
0.7	0.17	900	23	0	0	40	0.9	0.81	0.8
1	0.23	900	30	16.24	12.3	40	0.9	0.9	0.58
1.5	0.21	900	25	0.2	16.44	40	0.7	0.9	0.71

Table 15

Method	Discarded		Test		Collected (n)	Banking rate
	True-negative	False-negative	False-positive	True-positive		
No model	0	0	1276	9844	9844	0.8852
XGBoost	893	5920	383	3924	3924	0.911
XGBoost_Smote_NC	943	6745	333	3099	3099	0.902
LightGBM	620	3761	656	6083	6083	0.902
LightGBM_Smote_NC	540	3802	736	6042	6042	0.891

Table 16

Method	Discarded		Test		Collected (n)	Banking rate
	True-negative	False-negative	False-positive	True-positive		
No model	0	0	4501	6619	6619	0.595
XGBoost	3074	4062	1427	2557	2557	0.641
XGBoost_Smote_NC	3023	3966	1478	2653	2653	0.642
LightGBM	2563	3201	1938	3418	3418	0.638
LightGBM_Smote_NC	2772	3845	1729	2774	2774	0.616

Table 17

Method	Discarded		Test		Collected (n)	Banking rate
	True-negative	False-negative	False-positive	True-positive		
No model	0	0	8507	2613	2613	0.234
XGBoost	7917	1985	590	626	628	0.515
XGBoost_Smote_NC	7246	1812	1261	801	801	0.388
LightGBM	4877	1237	3630	1376	1376	0.274
LightGBM_Smote_NC	4262	1061	4245	1552	1552	0.267

Table 18

Method	AUC	accuracy	sensitivity	specificity	F_score	precision
XGBoost	0.567	0.433	0.398	0.699	0.554	0.911
XGBoost_Smote_NC	0.531	0.363	0.314	0.739	0.466	0.902
LightGbm	0.569	0.602	0.617	0.485	0.733	0.902
LightGBM_Smote_NC	0.517	0.591	0.613	0.423	0.726	0.891

Table 19

Method	AUC	accuracy	sensitivity	specificity	F_score	precision
XGBoost	0.54	0.506	0.386	0.682	0.482	0.641
XGBoost_Smote_NC	0.542	0.51	0.4	0.671	0.493	0.642
LightGbm	0.554	0.537	0.516	0.569	0.570	0.638
LightGBM_Smote_NC	0.521	0.498	0.419	0.615	0.498	0.616

Table 20

Method	AUC	accuracy	sensitivity	specificity	F_score	precision
XGBoost	0.60	0.768	0.240	0.930	0.327	0.515
XGBoost_Smote_NC	0.592	0.723	0.306	0.851	0.342	0.388
LightGbm	0.563	0.562	0.526	0.573	0.361	0.274
LightGBM_Smote_NC	0.557	0.522	0.593	0.500	0.369	0.267