Applying the XGBoost Model to Processing Patient Data

M.A. Zaynullina[1], V.V. Mokshin[1]

Introduction

The aim of the study is to study and apply ML methods to predict the risk of diabetes among patients

Tasks:
- Perform primary data processing
- Identify significant signs affecting whether a patient has diabetes using correlation analysis
- Provide SMOTE data balancing
- Learn and apply the XGBoost model
- Get and analyze results

The objective funtion of the XGBoost model

The objective function of the model:

$$\boxed{\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t),}$$

where l is the loss function,

$y_i, \hat{y}_i^t$ — s the value of the i-th element of the training sample and the sum of the predictions of the first t trees,

x_i – is a set of features of the i–th element of the training sample,

f_t - is a function (in our case, a tree) that we want to train at step t,

$f_t(x_i)$ – is a prediction on the i-th element of the training sample,

$\Omega(f)$ is the regularization of the function f.

In the next step, using the Taylor expansion to the second term, we can approximate the optimized function $\mathcal{L}^{(t)}$ with the following expression:

$$\boxed{\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + 0.5 h_i f_t^2(x_i)\right) + \Omega(f_t)}$$

In turn:

$$\boxed{g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}},\ h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial^2 \hat{y}_i^{(t-1)}}}$$

Initial data

Parameter	*Parameter description*	*Parameter type*
x_1	Number of pregnancies	int64
x_2	Plasma gluco concentration 2 hours in ar oral glucose tolerance test	int64
x_3	Diastolic blood pressu (mm Hg)	int64
x_4	Triceps skin fold thickne (mm)	int64
x_5	2-Hour serum insulin (n U/ml)	int64
x_6	Body mass index (kg/m^2)	float64
x_7	Age	int64
x_8	Diabetes pedigree function	float64
y	Having diadetes	int64

Fig. 1. Parameters description

r-Pearson correlation coefficients

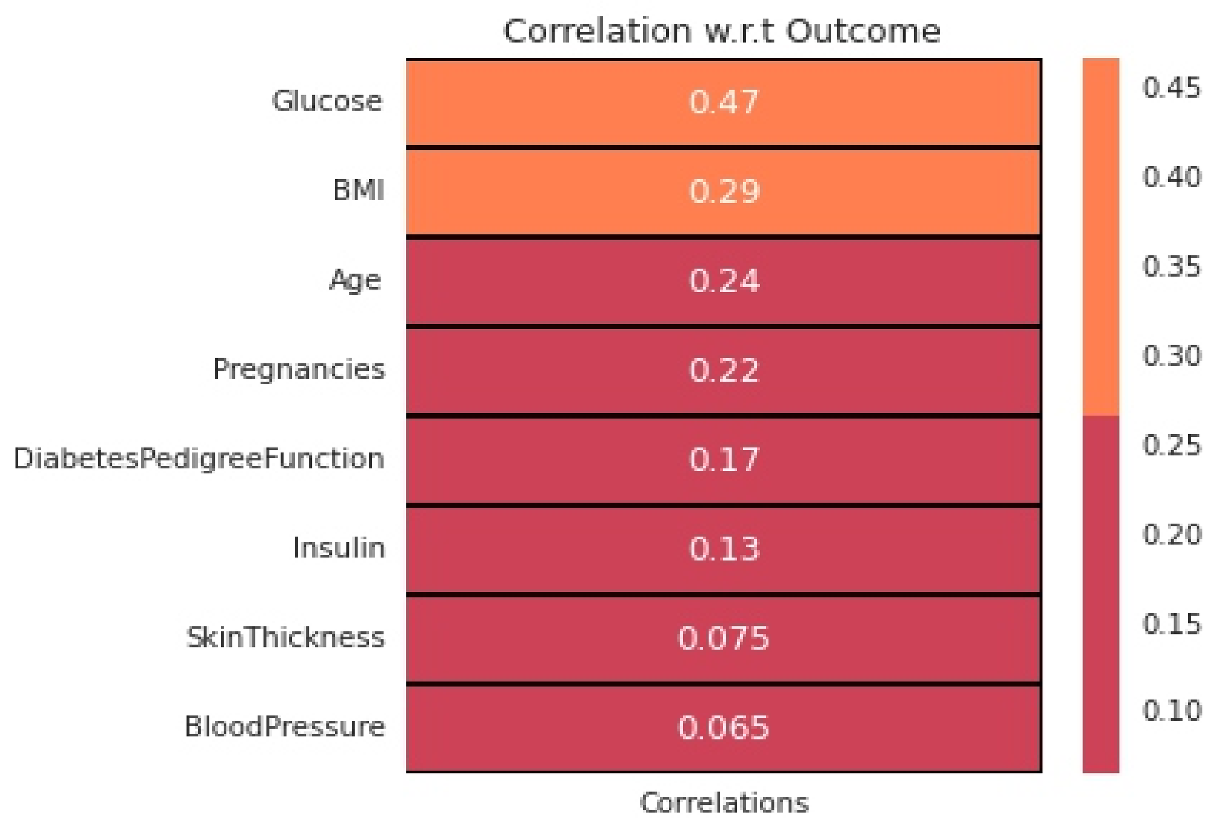

Fig. 2. r-Pearson correlation coefficients

Minimization the model error

Since we want to minimize the model error on the training sample, we need to find the minimum $\mathcal{L}^{(t)}$ for each t.

The minimum of this expression with respect to $f_t(x_i)$ is at the point:

$$\boxed{f_t(x_i) = \frac{-g_i}{h_i}}$$

Each individual tree of the ensemble $f_t(x_i)$ is trained by a standard algorithm.

Results of training

Since the binary classification problem is being solved (does the patient have diabetes), a reasonable AUC value should be greater than 0.5, and a good classification model has an AUC index greater than 0.9 (the value varies depending on the scope of application)

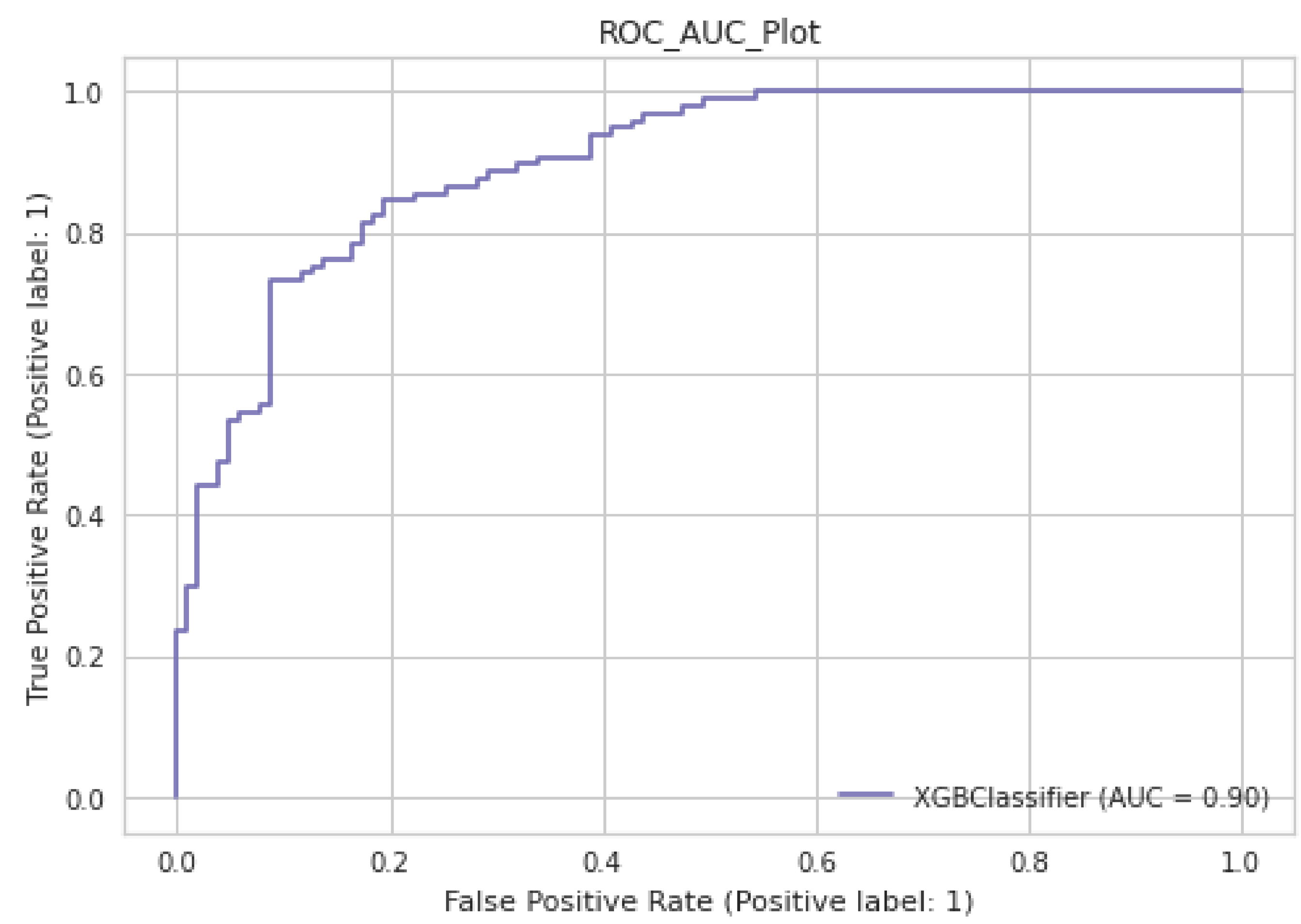

Fig. 3. The ROC curve for the weighted set of the Xboost method

Conclusion

The paper reviewed the XGBoost machine learning model for the purpose of subsequent prediction. The analysis of ways to separate significant features using correlation analysis, rebalancing of SMOTE classes is carried out. The quality of this solution for this particular task is estimated at 0.9 AUC, which is an excellent result.

[1] **KAZAN NATIONAL RESEARCH TECHNICAL UNIVERSITY NAMED AFTER A. N. TUPOLEV - KAI**
10 K.Marx St., Kazan,
Tatarstan 420111, Russia (www.kai.ru)

Крылья для жизни!