Gradient Boosting Predictive Model of Ovarian Response for Hormonal Therapy in Infertility Treatment

Kseniia Tikhaeva [1,2,3], Natalia Nesterova [1], Evgeny Tomilov [1], Stanislav Sotkin [1], Anna Muhina [2], Pavel Zavyalov [4], Elena Rosyuk [4]

[1] *Analytical Laboratory Sodas LLC, Volgograd, Russia*
[2] *Volgograd State Medical University, Volgograd, Russia*
[3] *LLC «Clinic Family», Volgograd, Russia*
[4] *«Center for Family Medicine», Ekaterinbug, Russia*

Abstract

To develop models for predicting the response to hormonal stimulation in the treatment of infertility, assisted reproductive technologies, data from 700 cases of infertility treatment with IVF were used. The forecasting system was built using the R 4.0.3 language. Modeling was carried out based on linear regression, regression trees, regression trees, k-nearest neighbors, gradient boosting.

Keywords

Gradient boosting, regression trees, linear regression, k-nearest neighbors, predictive models

1. Introduction

The number of oocytes obtained during the infertility treatment using methods of assisted reproductive technologies is a quantitative variable that is a significant predictor of the pregnancy probability. By the number of oocytes obtained, the ovarian response can be identified as optimal - 10-16 oocytes, suboptimal - 4-9 oocytes, poor - less than 4 oocytes, as well as a hyper response of more than 16 oocytes [1,2].

The problem of a poor ("weak") response is a pressing challenge in reproductive medicine. An important fact is that a poor and suboptimal response occurs not only in the group of women with the reduced ovarian reserve, but also in the group with normal and high ovarian reserve. While in the first case the unsatisfactory result of ovarian stimulation is predictable, in the second case it is unexpected. At the same time, it is known that in the group of patients with an unexpectedly poor ovarian response to stimulation with gonadotropins, the pregnancy rate is significantly lower in comparison with the group of patients with an expected weak ovarian response - 6–7% and 17–26%, respectively [3,4,5].

Therefore, it is important to compare the forecast of the oocytes number with the actual number of oocytes obtained.

In addition, it is important that most of the systems for predicting the number of oocytes are aimed at calculating the total number of oocytes, including immature and poor quality, or calculating the number of preovulatory follicles. A complex system is needed to predict the number of mature and high-quality oocytes suitable for fertilization. This prognosis system should include not only indicators of ovarian reserve, but also the sensitivity of follicles to gonadotropins. The basis of such a forecasting system can be a predictive model that includes all the variety of influencing factors.

Thus, the aim of the study is to develop a basic model that would be able to restore the functional relationship between predictor variables and the number of mature oocytes.

Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 10th International Conference «Integrated Models and Soft Computing in Artificial Intelligence» (IMSC-2021), May 17–20, 2021, Kolomna, Russian Federation
EMAIL:
ORCID:

CEUR Workshop Proceedings (CEUR-WS.org)

2. Materials and methods

Retrospective data on cases of ovarian stimulation and the number of oocytes obtained have been used to develop the model. A total of 658 cases of ovarian stimulation with gonadotropins have been included.

SMAPE (symmetric mean absolute percentage error) has been chosen as a metric for assessing the effectiveness of the model [6]. This error formula is:

$$sMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2},$$

where:

- n – sample size
- A_t – true response value
- F_t – value response, returned by the model

This error has been chosen since:

- The target variable in the available data is strictly positive.
- The target variable is measured on an absolute scale.
- sMAPE is an intuitively interpreted metric that improves the understanding of model quality by experts.
- sMAPE is a symmetric metric, it adjusts the model equally both in the direction of too high predictions and too low.

Thus, the target accuracy of the final model is determined as 85%, which is expressed in terms of 1 - sMAPE.

To increase the accuracy of the estimate, the mean absolute error (MAE) is also used [7]. The number of mature oocytes in the data ranges from 1 to 20 and at this stage the target MAE is 3 mature oocytes.

The mean absolute error formula is defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n},$$

where:

- n – sample size
- y_i – true response value
- x_i – value response, returned by the model

Thus, the formal development goals are as follows:

- Model with MAE <= 3
- Model with sMAPE <= 0.15.

The R 4.0.3 language has been chosen for the technical implementation of the simulation.

After the formation of a subset of the most informative variables using the Boruta method, the sample has included 31 predictors and 1 target variable.

The target variable was the number of mature oocytes.

Data on medical history, reproductive history, presence of somatic and reproductive diseases, objective anthropometry data, laboratory and instrumental indicators of ovarian reserve, blood levels of steroid and non-steroid hormones, clinical blood analysis indicators, biochemical blood test reflecting carbohydrate, fat and protein homeostasis have been used as potential predictors.

The variables that have been used to develop the model were categorical or quantitative.

All categorical variables for linear regression models, decision tree, random forest, k-nearest neighbours have been converted to binary using the one-hot method. The built-in target encoding method has been used for catboost.

All missing values in the variables have been restored by the method of imputation with auxiliary models. It should be noted that variables with the percentage of missing values more than 15 have been removed from the set of variables. Thus, the likelihood of improbable restoration of missing values has been reduced.

All quantitative variables have been normalized by Z-scaling.

$$z = \frac{x - \mu}{\sigma},$$

where:

- x – value of the variable
- μ – average value of this variable
- σ – standard deviation of the variable

In total, after clearing invalid objects, the selection currently contains 658 data rows.

The basic relationship diagram can be represented as follows:

$$number\ of\ mature\ oocytes \sim a + b_1 * x_1 + b_2 * x_2 + \cdots + b_n * x_n\ + \varepsilon,$$

where:

- *number of mature oocytes* is the target variable
- a is a constant
- b_n – coefficients
- x_n – the corresponding values of the variables
- ε – “noise”, a random error that is inevitably present in the data.

The dataset has been divided into training and testing samples in a ratio of 80% / 20% (529/129). Hyperparameters in the first subset have been selected by the cross-validation method, the second subset has been left for testing the model.

where “number of mature oocytes” is the target variable, a is a constant, bn – coefficients, xn – the corresponding values of the variables, ε – “noise”, a random error that is inevitably present in the data.

The dataset has been divided into training and testing samples in a ratio of 80% / 20% (529/129). Hyperparameters in the first subset have been selected by the cross-validation method, the second subset has been left for testing the model.

The models have been compared according to the MAE and sMAPE metrics, upon which the best one has been selected.

3. Modeling

3.1. Linear regression model

For the basic model, the stepwise regression method has been used according to the Akaike criterion:

$$AIC\ =\ 2k - 2ln\ (L),$$

where:

- k is the number of parameters in the statistical model,
- L is the maximized value of the likelihood function of the model, after which the one model with the smallest AIC value is chosen [8].

Table 1

Metrics of the linear regression model

Metric	Value
MAE	2.40
sMAPE	38.5%
1 - sMAPE	61.5%

The MAE of the baseline model is less than 3. The alternative models were expected to have lower error, than the baseline naive model, in order to be considered more efficient.

3.2. Regression tree model

The regression tree model does not differ significantly from the baseline model, having an even larger sMAPE [9].

Table 2
Regression tree model metrics

Metric	Value
MAE	2.40
sMAPE	38.7%
1 - sMAPE	61.3%

3.3. Random forest model

A model based on a random forest has low interpretability, but it allows one to assess the significance of individual features in the model, and also combines all the advantages of decision trees, compensating for their shortcomings [11].

Table 3
Metrics of the random forest model

Metric	Value
MAE	1.7
sMAPE	29.4%
1 - sMAPE	70.6%

As can be seen from the table of results, the resulting model has an average error of 1.7 oocytes, which is a significant improvement compared to the regression tree.

sMAPE dropped to 29.4%, or by 9.3%.

3.4. K-nearest neighbours model

The standard Euclidean metric has been chosen as the metric [11].

The number of neighbors K has been chosen to be 13.

Table 4
Metrics of the k-nearest neighbours model

Metric	Value
MAE	2.3
sMAPE	37.4%
1 - sMAPE	62.6%

The metrics of this model are closer to linear regression and decision tree models, which gives grounds to class them as poorly usable.

3.5. Catboost model

The catboost model is a gradient boosting algorithm developed by Yandex based on decision trees [12].

Table 5
Metrics of the catboost model

Metric	Value
MAE	1.09
sMAPE	17.7%
1 - sMAPE	82.3%

This model is of the highest quality of all previously investigated. MAE is only 1.08 oocytes, while SMAPE is 17.7%. Thus, we accept catboost as a working model for the following reasons:

- The lowest error closely approaching the target level.
- The optimized library allows you to create the fastest implementation of the model for use in real practice.

4. Conclusion

In order to develop a model for predicting the response of the ovaries to hormonal stimulation in the treatment of infertility using assisted reproductive technologies, data from 700 cases of infertility treatment with ART have been used, for each of the cases, anonymized and impersonal data regarding anamnesis, patient status, ovarian reserve, treatment metrics, and the result of stimulation have been collected. The forecasting system has been built using the R 4.0.3 language. Modeling has been carried out on the basis of linear regression algorithms, regression trees, random forest, k random neighbours method, gradient boosting.

The predictive model developed on the basis of random forest methods and the “catboost” gradient boosting variant predicts the number of mature oocytes with an average error (MAE) of 1.09 of a mature oocyte and with an accuracy of 82.3% (sMAPE).

Predictive modeling can solve the problem of predicting ovarian response in the treatment of infertility using ART methods. The most effective for this task is a variant of the gradient boosting method “catboost” due to the mechanism of its operation, which consists in constructing a set of successively correcting decision trees.

5. References

[1] Li Y, Li X, Yang X, Cai S, Lu G, Lin G, Humaidan P, Gong F. Cumulative Live Birth Rates in Low Prognosis Patients According to the POSEIDON Criteria: An Analysis of 26,697 Cycles of in vitro Fertilization/Intracytoplasmic Sperm Injection. Front Endocrinol (Lausanne). 2019 Sep 19;10:642. doi: 10.3389/fendo.2019.00642. PMID: 31608011; PMCID: PMC6761219.

[2] Loutradis, Dimitris et al. “FSH receptor gene polymorphisms have a role for different ovarian response to stimulation in patients entering IVF/ICSI-ET programs.” Journal of assisted reproduction and genetics vol. 23,4 (2006): 177-84. doi:10.1007/s10815-005-9015-z

[3] van der Gaast MH, Eijkemans MJ, van der Net JB, de Boer EJ, Burger CW, van Leeuwen FE, Fauser BC, Macklon NS. Optimum number of oocytes for a successful first IVF treatment cycle. Reprod Biomed Online. 2006 Oct;13(4):476-80. doi: 10.1016/s1472-6483(10)60633-5. PMID: 17007663.

[4] Wu CX, Zhang T, Shu L, Huang J, Diao FY, Ding W, Gao Y, Wang W, Mao YD, Cui YG, Liu JY. [Cumulative live birth rates per oocytes retrieved cycle: evaluation of clinical outcomes of IVF/ICSI]. Zhonghua Fu Chan Ke Za Zhi. 2018 Mar 25;53(3):160-166. Chinese. doi: 10.3760/cma.j.issn.0529-567X.2018.03.004. PMID: 29609229.

[5] Drakopoulos P, Blockeel C, Stoop D, Camus M, de Vos M, Tournaye H, Polyzos NP. Conventional ovarian stimulation and single embryo transfer for IVF/ICSI. How many oocytes do we need to maximize cumulative live birth rates after utilization of all fresh and frozen embryos? Hum Reprod. 2016 Feb;31(2):370-6. doi: 10.1093/humrep/dev316. Epub 2016 Jan 2. PMID: 26724797.

[6] Flores, B. E. (1986) "A pragmatic view of accuracy measurement in forecasting", Omega (Oxford), 14(2), 93–98. doi:10.1016/0305-0483(86)90013-7

[7] Willmott, Cort J.; Matsuura, Kenji (December 19, 2005). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". Climate Research. 30: 79–82. doi:10.3354/cr030079

[8] Rencher, Alvin C.; Christensen, William F. (2012), "Chapter 10, Multivariate regression – Section 10.1, Introduction", Methods of Multivariate Analysis, Wiley Series in Probability and Statistics, 709 (3rd ed.), John Wiley & Sons, p. 19, ISBN 9781118391679

[9] Quinlan, J. R. (1986). "Induction of decision trees" (PDF). Machine Learning. 1: 81–106. doi:10.1007/BF00116251. S2CID 189902138

[10] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5

[11] Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression" (PDF). The American Statistician. 46 (3): 175–185. doi:10.1080/00031305.1992.10475879. hdl:1813/31637

[12] Dorogush, Anna Veronika; Ershov, Vasily; Gulin, Andrey (2018-10-24). "CatBoost: gradient boosting with categorical features support". arXiv:1810.11363 [cs.LG]