pHealth 2020
B. Blobel et al. (Eds.)

doi:10.3233/SHTI200622

Prediction of a Due Date Based on the Pregnancy History Data Using Machine Learning

Oleg METSKER[a,1], Georgy KOPANITSA[b], Eduard KOMLICHENKO[a], Maria YANUSHANETS[a] and Ekaterina BOLGOVA[b]
[a] *Almazov National Medical Research Centre, Saint-Petersburg, Russia*
[b] *ITMO University, Saint-Petersburg, Russia*

Abstract. Prediction of a labor due date is important especially for the pregnancies with high risk of complications where a special treatment is needed. This is especially valid in the countries with multilevel health care institutions like Russia. In Russia medical organizations are distributed into national, regional and municipal levels. Organizations of each level can provide treatment of different types and quality. For example, pregnancies with low risk of complications are routed to the municipal hospitals, moderate risk pregnancies are routed to the reginal and high risk of complications are routed to the hospitals of the national level. In the situation of resource deficiency especially on the national level it is necessary to plan admission date and a treatment team in advance to provide the best possible care. When pregnancy data is not standardized and semantically interoperable, data driven models. We have retrospectively analyzed electronic health records from the perinatal Center of the Almazov perinatal medical center in Saint-Petersburg, Russia. The dataset was exported from the medical information system. It consisted of structured and semi structured data with the total of 73115 lines for 12989 female patients. The proposed due date prediction data-driven model allows a high accuracy prediction to allow proper resource planning. The models are based on the real-world evidence and can be applied with limited amount of predictors.

Keywords. Pregnancy, Prediction, Due Date, Machine learning, Random Forest

Introduction

Prediction of a labor due date is important especially for the pregnancies with high risk of complications where a special treatment is needed [1]. This is especially valid in the countries with multilevel health care institutions like Russia. In Russia medical organizations are distributed into national, regional and municipal levels. Organizations of each level can provide treatment of different types and quality. For example, pregnancies with low risk of complications are routed to the municipal hospitals, moderate risk pregnancies are routed to the reginal and high risk of complications are routed to the hospitals of the national level. In the situation of resource deficiency especially on the national level it is necessary to plan admission date and a treatment team in advance to provide the best possible care.

[1] Corresponding Author. Oleg Metsker, Almazov National Medical Research Centre, Saint-Petersburg, Russia; E-mail: olegmetsker@gmail.com

Experience of implementing machine learning methods shows that efficient due date prediction can be made based on the ultrasound data [2]. Artificial neuron networks showed high accuracy in the due date prediction [3].

The development of perinatal episodes is characterized by a significant number of heterogeneous interconnected factors with different contributions in etiological and pathological terms at different stages. This significantly complicates the development of decision support models. In this situation intellectual data analysis and data-driven models [4] can become a good basis for clinical decision support systems that can support doctors and healthcare organizers to plan necessary resources on the different healthcare delivery levels [5]. Latest systematic reviews [6–11] have demonstrated the inability to accurately predict the due date [12] due to the lack of available data.

Existing perinatal decision support models consider only the specific risk factors associated with the development of pregnancy. However, there are no comprehensive models that can predict a due date based on the various parameters available in the electronic health record. This is largely due to the lack of patient data, which makes it impossible to build sufficiently accurate mathematical models of pregnancy development.

Thus, despite the experience gained in developing decision-making models and forecasting of maternal risks, there is still room for improvement of the models. The development of such models will to predict a due date with a high accuracy to allow resources planning and efficient patient routing.

The goal of this study is to develop a model for prediction of a labor due date based on the pregnancy history data using machine learning methods and to identify the most important predictors.

1. Methods

We have retrospectively analyzed electronic health records from the perinatal Center of the Almazov perinatal medical center in Saint-Petersburg, Russia. The dataset was exported from the medical information system. It consisted of structured and semi structured data with the total of 73115 lines for 12989 female patients (Dataset A) for the period between 1st of January 2015 and 31st of December 2019**.**

1.1. Data Preprocessing

- We have extracted 73115 lines with 97 structured features with a mother's anamnesis. Each line presents a female patient that underwent treatment or observation in the perinatal center.
- All the lines that did not contain labor date were removed from the dataset. This resulted in the 62734 lines representing c corresponding amount of female patients
- A target column was a length of a gestation in days.

1.2. Correlation and Features Importance

Correlation analysis was performed using F-score based method of scikit-learn library to select the most relevant predictors.

1.3. Sensitivity Analysis and Grid Search

Each experiment ran in the setting of stratified 5-fold cross-validation i.e., random 80% of training dataset was used for training and random 20% for testing of training dataset (70% random selection from the study dataset) for testing. Target class ratios in the folds were preserved. The gradient search parameters were: params = {'min_child_weight':[4,5], 'gamma':[i/10.0 for i in range(3,6)], 'subsample':[i/10.0 for i in range(6,11)], 'colsample_bytree':[i/10.0 for i in range(6,11)], 'max_depth': [2,3,4]}. We compared Gradient Boosting regression, Random forest regression, Linear regression and Voting regression. Root mean square error was used as a performance metric. After determining the optimal dataset and model parameters, we performed a validation with the testing dataset (30% random selection from the study dataset). Scikit-learn library was used for the experiment. Mean Absolute Error (MAE) was used as a performance metric. The best performing regressor was evaluated on the test dataset (20% random selection from the study dataset). For this study we used Python 3.6.3 and scikit-learn 0.19.1[2] as the basic framework for machine learning models.

2. Results

2.1. Features Importance

This section describes predictors that are either well-known (for example, the age of the mother, etc.), as well as previously weakly described predictors (such as the sex of the child, RH factor, gastrointestinal diseases). The results of the features importance analysis are presented in the Figure 1.

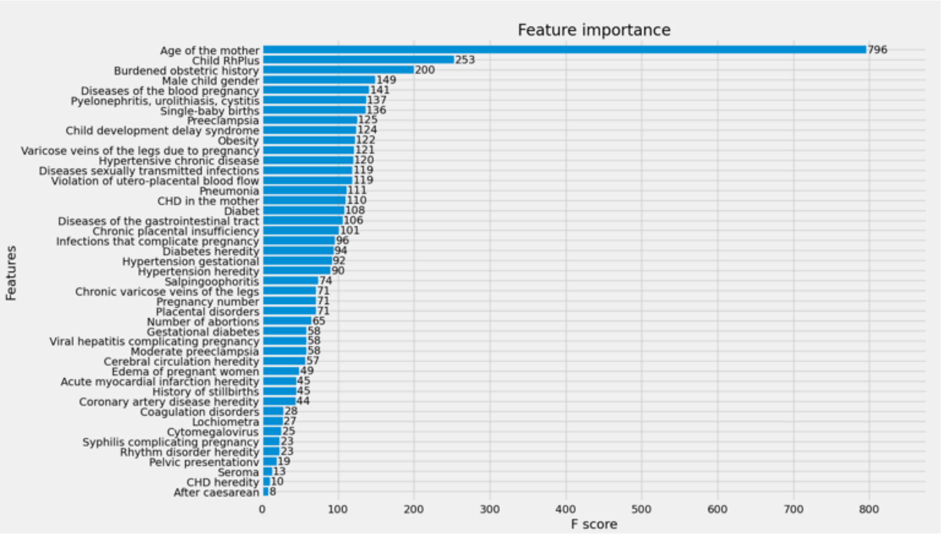

Figure 1. Features importance for the due date prediction

Figure 2 and Table 1 presents the results of the grid search for the optimal regression model for a due date prediction.

[2] https://scikit-learn.org/stable/

Table 1. Prediction efficiency for different regressors

Regressor	MAE
Random Forrest	3.72
Gradient Boosting	8.02
Linear regression	7.12
Voting regression	6.58

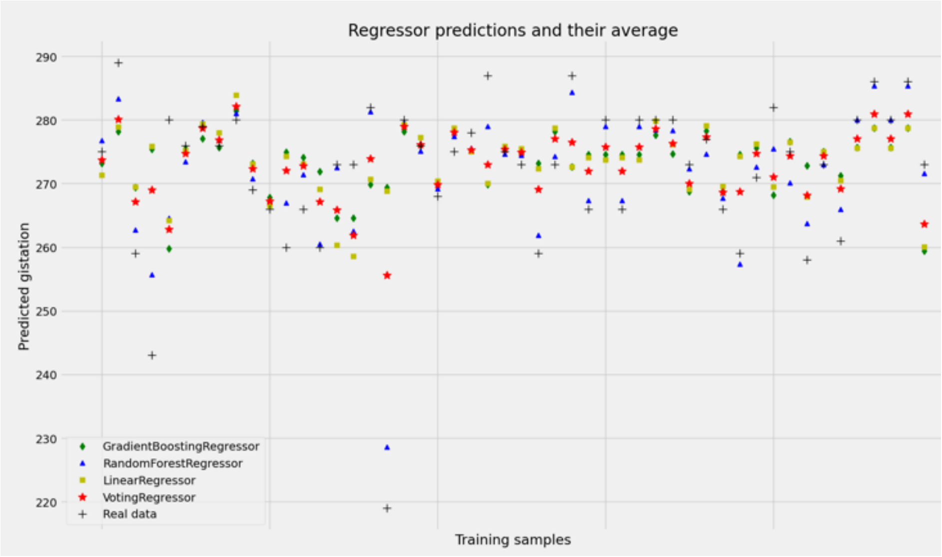

Figure 2. Due date Regression prediction

The grid search resulted in the optimal grid parameters: {'colsample_bytree': 0.9, 'gamma': 0.3, 'max_depth': 2, 'min_child_weight': 4, 'subsample': 1.0}. We used a MAE for the delivery due date accuracy assessment. The random forest regression gave the best value of MAE of 3.85 on the test dataset.

3. Discussion

The proposed due date prediction data-driven model allows a high accuracy prediction to allow proper resource planning. The models are based on the real-world evidence and can be applied with limited number of predictors. We have also identified the most important features to predict the labor due date. This will help policy makers to establish proper data collection channels to have the most important information in the electronic health records.

Acknowledgment

The reported study was funded by RFBR, project number 20-37-70047.
The work of Georgy Kopanitsa was financially supported by the Government of the Russian Federation through the ITMO fellowship and professorship program.

References

[1] Olesen AW, Thomsen SG. Prediction of delivery date by sonography in the first and second trimesters. Ultrasound Obstet Gynecol. 2006 Sep; 28,3: 292-7. doi:10.1002/uog.2793.

[2] Naimi AI, Platt RW, Larkin JC. Machine Learning for Fetal Growth Prediction. Epidemiology. 2018 Mar; 29,2: 290-298. doi:10.1097/EDE.0000000000000788.

[3] Podda M, Bacciu D, Micheli A, Bellù R, Placidi G, Gagliardi L. A machine learning approach to estimating preterm infants survival: development of the Preterm Infants Survival Assessment (PISA) predictor. Sci. Rep. 2018; 8: 13743. doi:10.1038/s41598-018-31920-6.

[4] Tsui KL, Chen N, Zhou Q, Hai Y, Wang W. Prognostics and health management: A review on data driven approaches. Math. Probl. Eng. 2015; 2015,6: 1-17. doi:10.1155/2015/793161.

[5] Krikunov AV, Bolgova EV, Krotov E, Abuhay TM, Yakovlev AN, Kovalchuk SV. Complex data-driven predictive modeling in personalized clinical decision support for Acute Coronary Syndrome episodes. Procedia Comput. Sci. 2016; 80: 518-529. doi:10.1016/j.procs.2016.05.332.

[6] Aoyama K, D'Souza R, Pinto R, Ray JG, Hill A, et al. Risk prediction models for maternal mortality: A systematic review and meta-analysis. PLoS One. 2018; 13,12: e0208563. doi:10.1371/journal.pone.0208563.

[7] Verstraete EH, Blot K, Mahieu L, Vogelaers D, Blot S. Prediction models for neonatal health care-associated sepsis: A meta-analysis. Pediatrics April; 2015; 135,4: e1002-e1014. doi:10.1542/peds.2014-3226.

[8] Ukah UV, De Silva DA, Payne B, Magee LA, et al. Prediction of adverse maternal outcomes from pre-eclampsia and other hypertensive disorders of pregnancy: A systematic review. Pregnancy Hypertens. 2018 Jan; 11: 115-123. doi:10.1016/j.preghy.2017.11.006.

[9] Verhagen TEM, Hendriks DJ, Bancsi LFJMM, Mol BWJ, Broekmans FJM. The accuracy of multivariate models predicting ovarian reserve and pregnancy after in vitro fertilization: A meta-analysis. Hum Reprod Update. Mar-Apr 2008; 14,2: 95-100. doi:10.1093/humupd/dmn001.

[10] Lamain – de Ruiter M, Kwee A, Naaktgeboren CA, Franx A, Moons KGM, Koster MPH. Prediction models for the risk of gestational diabetes: a systematic review. Diagn Progn Res. 2017 Feb 8; 1: 3. doi:10.1186/s41512-016-0005-7.

[11] Sananès N, Langer B, Gaudineau A, Kutnahorsky R, et al. Prediction of spontaneous preterm delivery in singleton pregnancies: Where are we and where are we going? A review of literature. J Obstet Gynaecol. 2014 Aug; 34,6: 457-61. doi:10.3109/01443615.2014.896325.

[12] Draper ES, Manktelow B, Field DJ, James D. Prediction of survival for preterm births by weight and gestational age: retrospective population based study. BMJ. 1999 Oct 23; 319,7217: 1093–1097. doi:10.1136/bmj.319.7217.1093.