The Silent Emergency - Predicting Preterm Birth

The Erdős Institute - Data Science Boot Camp Project Fall 2023

Members: Katherine Grillaert, Divya Joshi, Noah Rahman, Alex Sutherland, and Kristina Zvolanek

Overview: Preterm birth is a primary cause of infant mortality and morbidity in the United States, affecting approximately 1 in 10 births.[1] This rate is notably higher among Black women (14.6%), compared to White (9.4%) and Hispanic women (10.1%).[2] Despite its prevalence, predicting preterm birth remains challenging due to a complex etiology rooted in environmental, biological, genetic, and behavioral interactions.[3] Our project harnesses machine learning to predict preterm birth using electronic health records. This data intersects with social determinants of health, reflecting some of the interactions contributing to preterm birth.[4] Recognizing that under-representation in healthcare research perpetuates racial and ethnic health disparities, we take care to use diverse data to ensure equitable model performance across underrepresented populations.[5]

Project Stakeholders: Pregnant individuals, prospective parents, medical professionals involved with maternal care and births, hospital systems, insurance companies.

Approach: We constructed two models to predict preterm birth, one with demographic features and one with health and lifestyle features. Our data source was the National Institute of Health's All of Us Research Program controlled tier of de-identified medical data. The Demographics model was trained on a dataset of 13690 births between the years 2009 and 2022. Most demographic information for each individual was available only as summary statistics based on their zip code. Race and ethnicity were available on the individual level. The Lifestyle model was trained on a dataset of 8771 births between the years 2011 and 2022. Features included drinking, smoking, drug use, body mass index, diabetes, and mental health.

Methods: The Demographics model used a support vector classifier and the Health and Lifestyle model a logistic regression. Both employed class weights, optimized using grid search cross-validation, to prioritize prediction of preterm birth. Our baseline model was a weighted coin flip that reflected the ratio of preterm births in our data. We trained and evaluated the models using 10-fold stratified cross-validation. The package AI Fairness 360 tested that our model predictions performed equally well across race and ethnicity.

Key Performance Indicators (KPIs): We prioritized minimizing costly false negatives, accepting the possibility of increased false positives.

Demographic Model		Baseline
Recall	0.413	0.145
F1	0.242	0.137
PR-AUC	0.172	0.192
SPD	*	
Equalized Odds[†]	0.0	

Health and Lifestyle Model		Baseline
Recall	0.473	0.137
F1	0.247	0.136
PR-AUC	0.197	0.196
SPD	*	
Equalized Odds[†]	0.0	

Conclusion and Future Work: Our models performed only as well as the baseline model, highlighting the challenges of predicting preterm birth with only electronic health records. Predictive models may need to incorporate features from more than one domain, including environmental, behavioral, biological, and genetic factors.[4] Future work should consider the collection of thorough, individual-level data, observed during the pregnancy, in order to provide a high-quality data source for machine learning predictions.

Acknowledgements

Thank you to Evelyn Huszar for her enthusiastic mentorship during this project. Our gratitude also to Roman Holowinsky, Matt Osborne, Alec Clott, and The Erdős Institute for their support during this boot camp. Finally, thank you to the NIH and the All of Us program for collecting, anonymizing, and allowing us to use the data for this project.

More Information

Our project documentation is hosted on our GitHub. You'll also find our five-minute presentation, exploratory analysis visualizations, code, and a synthetic data set for hands-on exploration. Check it out and contact us at: https://www.github.com/kgrillaert/preterm_birth

[*] *The full table of SPD metrics is available on our GitHub.*

[†] *An important caveat to the SPD and Equalized Odds metrics: because our model is not reliably predicting preterm birth, these fairness metrics might not provide accurate insights or hold substantial value.*

> "The world is facing a silent emergency... of preterm births." - UNICEF[6]

References

1. Csaba Siffel, Andrew K Hirst, Sujata P Sarda, Michael W Kuzniewicz, and De-Kun Li. The clinical burden of extremely preterm birth in a large medical records database in the united states: Mortality and survival associated with selected complications. *Early Human Development*, 171:105613, 2022.

2. Preterm birth. `https://www.cdc.gov/reproductivehealth/maternalinfanthealth/pretermbirth.htm`, October 2023. Accessed: 2023-11-18.

3. Tracy A Manuck. Racial and ethnic differences in preterm birth: a complex, multifactorial problem. In *Seminars in perinatology*, volume 41, pages 511–518. Elsevier, 2017.

4. Dhelia M Williamson, Karon Abe, Christopher Bean, Cynthia Ferré, Zsakeba Henderson, and Eve Lackritz. Current research in preterm birth. *Journal of Women's Health*, 17(10):1545–1549, 2008.

5. CM Vajdic, A Kricker, M Giblin, J McKenzie, J Aitken, GG Giles, and BK Armstrong. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. 2013.

6. World Health Organization et al. *Born too soon: decade of action on preterm birth.* World Health Organization, 2023.