Clinical Trial Enrollment Prediction

Data Science Institute
COLUMBIA UNIVERSITY

Authors: Naiqi Chen, Yixuan Liu, Yuhan Jin, Zeyu Jin, Xingyu Wei
Industry mentor: Lars Hulstaert
Faculty mentor: Adam S. Kelleher

Data Science Capstone Project with JnJ

Background and Objective

Our project is a quantitative study of how socioeconomic and demographic factors affect patient enrollment rates in clinical trials.

The first objective of our project is to predict the enrollment rate of clinical trials. The ability to predict enrollment rates can help assess which proposed trials will meet enrollment targets and provide a data-driven decision support system to determine the locations of institutions conducting clinical trials. Thus, an accurate prediction model can help reduce the cost incurred by initiating under-enrolled trials.

The second objective of our project is to accelerate enrollment rates. By systematically understanding how socioeconomic and demographic factors influence patient recruitment in clinical trials, we can identify actionable levers that can be impacted to enhance the diversity of clinical trial populations and thereby reduce health inequity.

In our project, we combined internal datasets containing patient enrollment information provided by Johnson & Johnson with external datasets from SDOH and CDC. We experimented with various approaches, including regression and classification, and obtained a decent result.

System Design

This project carries through the entire process of Data Science. Initially, we have gathered features from three external sources joined by zip codes to our internal data. After cleaning the dataset, we standardized and transformed features to feed into our models. We have trained and tuned both regression and classification models. These models considers different levels of attributions which study-level features have onto the model prediction. Finally, we assessed our model performance with metrics and identified key features with SHAP.

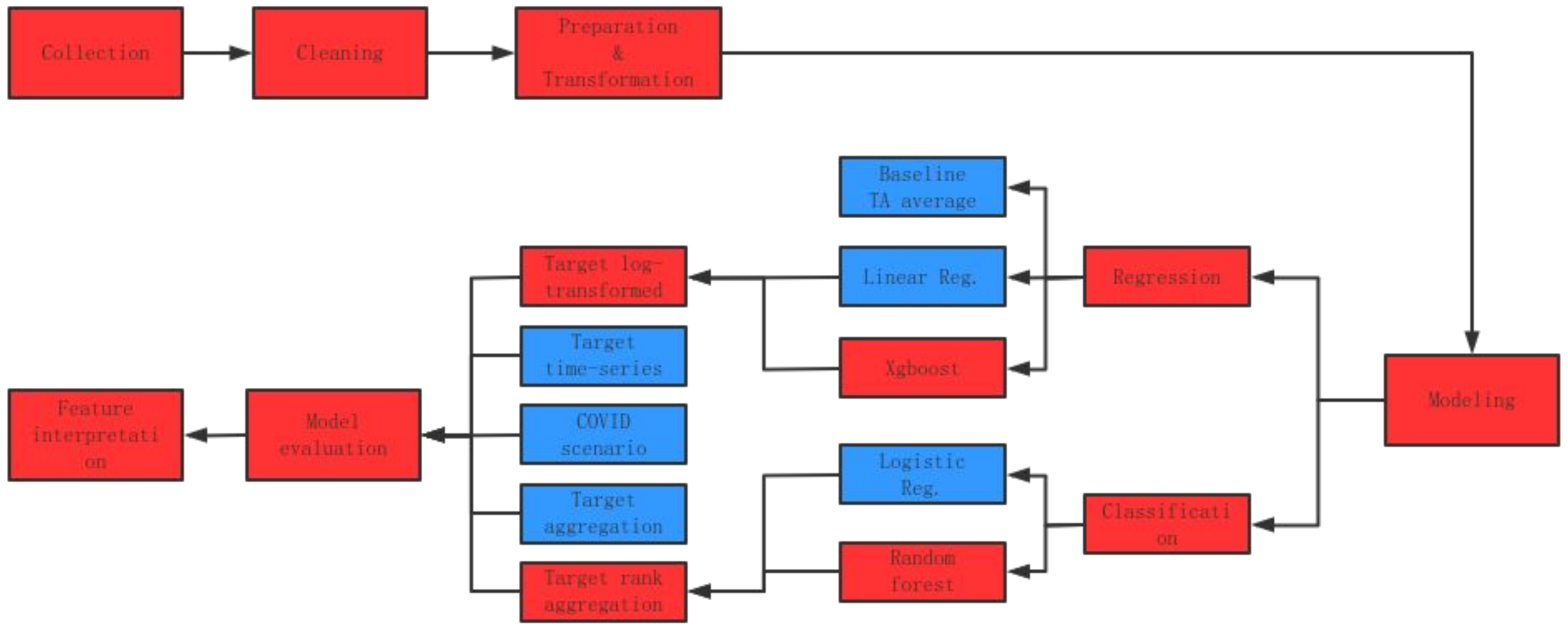

Figure 1. System Design & Implementation Flow Chart

Result

The two following graphs are the SHAP visualizations of our final models. The one on the left corresponds to the XGBoost Regressor and the right one corresponds to the Random Forest Classifier. In terms of performance, the regression model with 0.63 R2 score on the test set is better than the classification model with 0.55 accuracy on the test set. However, the SHAP interpretation indicates that the regression model is dominated by study related features, such as therapeutic area and primary indication of the study, instead of the socio-economic and demographic factors that are associated with site selection. With the transformation of ranking within each study, our classification model focuses more on these socio-economic and demographic factors that could facilitate patient recruitment in clinical trials.

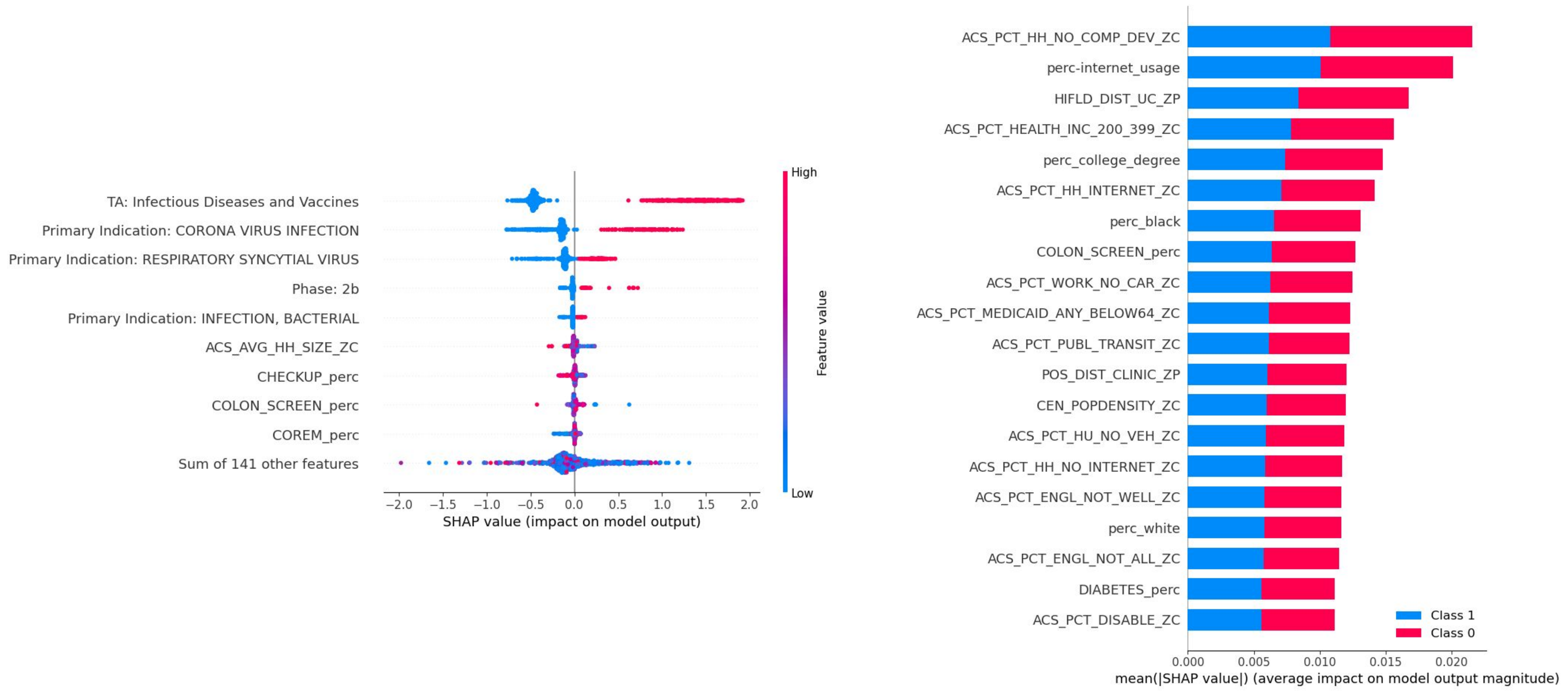

Figure 2. SHAP Interpretation of Final Models (left: XGBoost Regressor; right: Random Forest Classifier).

Conclusion

Through regression and classification machine learning models such as xgboost and random forest, this study distinguishes key factors that affect clinical trial enrollment rates. Building on this result, we would like to perform the same analysis on larger datasets which would hopefully increase the importance of socio-economic features.

Acknowledgments

We would like to express our gratitude to Lars and Professor Kelleher for their advice and supports during the course of this project.

References

Bieganek, C., Aliferis, C. & Ma, S. Prediction of clinical trial enrollment rates. *PLoS One* 17, e0263193 (2022).

Desai, M. Recruitment and retention of participants in clinical studies: Critical issues and challenges. *Perspect Clin Res* 11, 51–53 (2020).