

Predicting Patients' Attrition Percentage in Clinical Trials

Leveraging Machine Learning and Geospatial Analysis
to Understand Clinical Trial Retention

Prepared By:

Fadwa Elfeituri, Maxwell Lewis, Tsegie Kassahun
Masters of Health Informatics and Data Science
Massive Health Data Fundamentals - HIDS6001
Biomedical Graduate Education
Georgetown University

Introduction

Predicting patient attrition in clinical trials is a significant challenge for the medical and research community. High dropout rates can compromise the integrity of trial outcomes, delay regulatory approval, and increase overall costs. This project aimed to address this issue by leveraging machine learning models to predict attrition percentages based on features extracted from ClinicalTrials.gov. In addition to prediction, the project explored geographic trends by mapping trial locations and their associated dropout rates.

Data for the project were sourced from ClinicalTrials.gov via its API and an attrition dataset containing dropout percentages for 1325 clinical trials. Features relevant to the study design, participant demographics, and trial logistics were selected to build predictive models. The project's dual objectives were to develop a machine learning model to predict attrition and visualize trial facility locations to explore geographic influences on retention.

Methodology

The dataset integrated information from multiple sources, including trial-specific data from ClinicalTrials.gov, dropout percentages from the attrition dataset, and Rural Urban Commuting Area (RUCA) codes to categorize trial locations. The features selected for modeling encompassed participant characteristics, study design, geographic location, and trial outcomes such as adverse events. These features were chosen because they have a significant impact on understanding the attrition rate, as they capture both individual-level factors and external influences that contribute to participant dropout or retention in clinical trials. A total of 23 features were included in the analysis, with key variables such as trial duration, the number of trial locations, and participant masking receiving particular attention.

The dataset underwent preprocessing to address missing values and encode categorical features. Missing values were imputed using the median, and one-hot encoding was applied to categorical features, including study phases, participant sex, and intervention types. After preprocessing, the dataset was split into training and testing sets in an 80/20 ratio.

Two machine learning models were employed for this task: Random Forest Regressor and XGBoost Regressor. Random Forest was selected for its ability to handle high-dimensional data and provide insights into feature importance, while XGBoost was chosen for its optimization capabilities and ability to handle feature interactions effectively. Both models were evaluated using metrics such as R² score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Model Results

The Random Forest model achieved a R^2 score of 0.31 and an RMSE of 8.79, while the XGBoost model slightly outperformed Random Forest with an R^2 score of 0.33 and an RMSE of 8.67. These results highlight the utility of machine learning models in predicting attrition percentages, although the modest R^2 scores suggest that additional features or alternative modeling approaches may further improve predictions. The model configuration specifies a regression task using squared error as the loss function, with parameters such as maximum tree depth to control complexity, a learning rate to adjust updates, and subsampling and column sampling to reduce overfitting. Root Mean Squared Error (RMSE) is used as the evaluation metric during training. The model trains for up to 100 rounds, with early stopping implemented to halt training if no improvement is observed for 10 rounds.

Feature importance analysis revealed that trial duration was the most significant predictor of attrition, emphasizing the challenges of retaining participants in longer studies. The number of trial locations also played a critical role, suggesting that accessibility to trial sites influences retention. Other important features included adverse events and enrollment size, which reflect the operational and experiential aspects of clinical trials.

Visualizations provided additional insights into the data. Feature importance rankings were also visualized using bar plots, aiding in the interpretation of model results. Geographic mapping of trial facilities illustrated the spatial distribution of trials and their associated attrition rates, further emphasizing the importance of location in participant retention.

Table 1. Comparison of Results Between the Two Models

Comparison Table:

	Model	R^2 Score	MSE	RMSE
0	Random Forest	0.312719	77.348382	8.794793
1	XGBoost	0.331875	75.192476	8.671360

Figure 1. Bar Chart Illustrating Feature Importance in the Random Forest Regressor Model

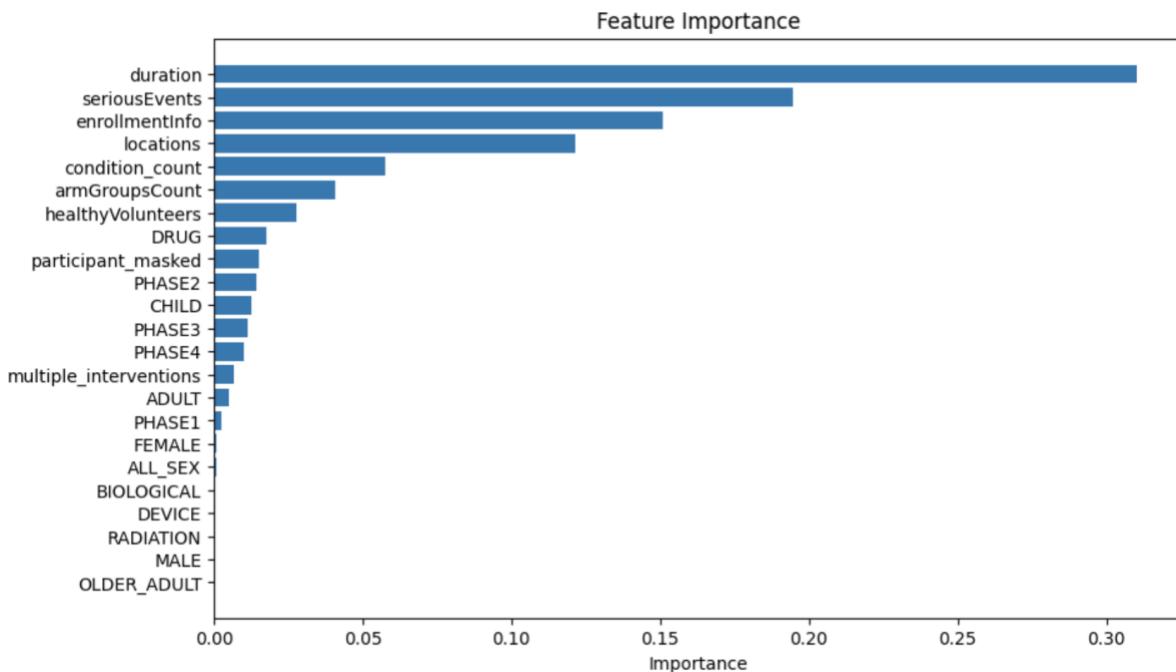
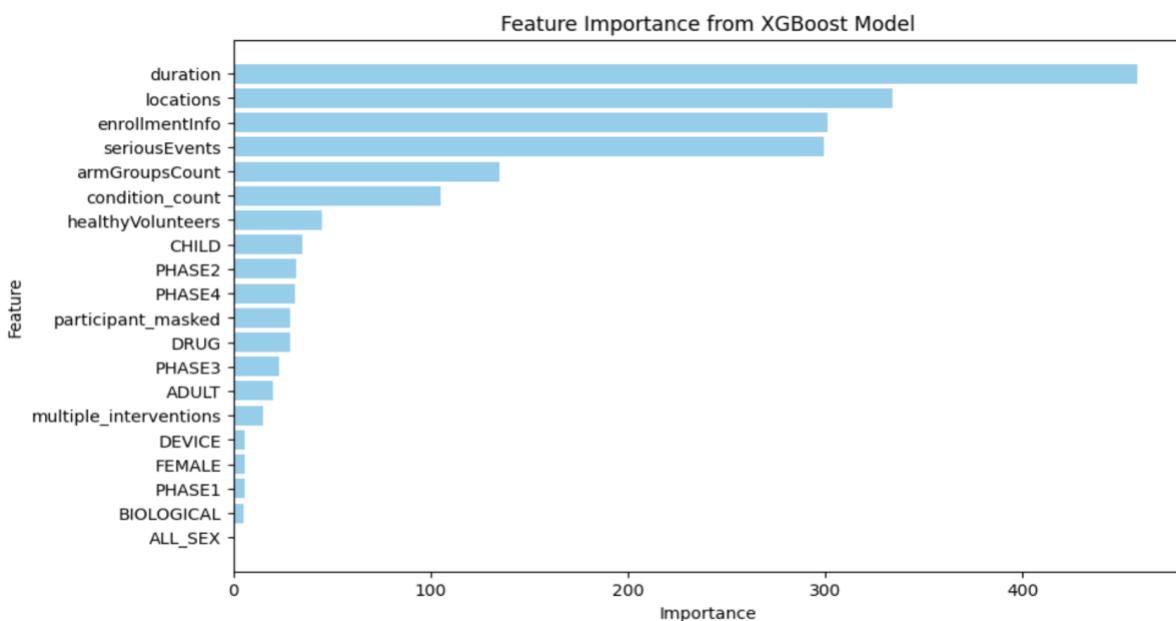


Figure 2. Bar Chart Illustrating Feature Importance in the XGBoost Model



Geospatial Mapping

Geospatial mapping in this project involved integrating data from the ClinicalTrials.gov API and the RUCA dataset to visualize geographic trends in clinical trial dropout rates. Locations were categorized into urban and rural types based on RUCA codes, and dropout percentages were color-coded on an interactive map. The map displayed clinical trial facilities with information such as dropout percentages, facility names, and location types. This approach helped identify regional patterns of attrition, providing valuable insights for improving trial planning and participant retention across different geographic areas.

The interactive map is available through the following [link](#).

Figure 3. Map displaying clinical facility locations categorized by dropout percentage



Drop Out Percentage

- [Green square] Less than 20%
- [Yellow square] Between 20% - 50%
- [Orange square] Between 50% - 80%
- [Red square] More than 80%

Conclusion

This project successfully developed machine learning models to predict attrition percentages in clinical trials and identified key factors influencing participant retention. Trial duration, the number of locations, and adverse events emerged as the most critical predictors of dropout rates. These findings underscore the importance of designing trials with participant accessibility and experience in mind to minimize attrition.

The visualization of trial locations added a geographic dimension to the analysis, revealing trends in urban versus rural dropout rates and providing actionable insights for trial planning. While the models performed adequately, future work could focus on incorporating additional features, such as socioeconomic and environmental factors, to enhance prediction accuracy. Exploring patient-level data and leveraging ensemble modeling techniques could further improve model performance.

By addressing the challenges of patient attrition, this project contributes to the optimization of clinical trial designs and outcomes, ultimately supporting the advancement of medical research and patient care.