ClinicalTrials.gov

STATS 418: Final Project

Isaac Schwarz

Exploratory Data Analysis

Data Source:

- **ClinicalTrials.gov** API
- Initially retrieved ~10,000 study records
- Final dataset included **~100 high-quality records** after cleaning
  - Due to time and technical constraints, a larger sample was not processed

General Observations:

- Most columns were **fairly clean**
- Some columns contained **missing or ambiguous values** (e.g., “Unknown”)
- A subset of **irrelevant variables** was dropped from the analysis

Variables

Feature Variables (10)

- **Categorical (9):**
  - study_type, allocation, intervention_model, masking,
  - primary_purpose, gender, healthy_volunteers,
  - lead_sponsor_class, intervention_types
- **Numerical (1):**
  - minimum_age (parsed from string format)

Target Variables (3):

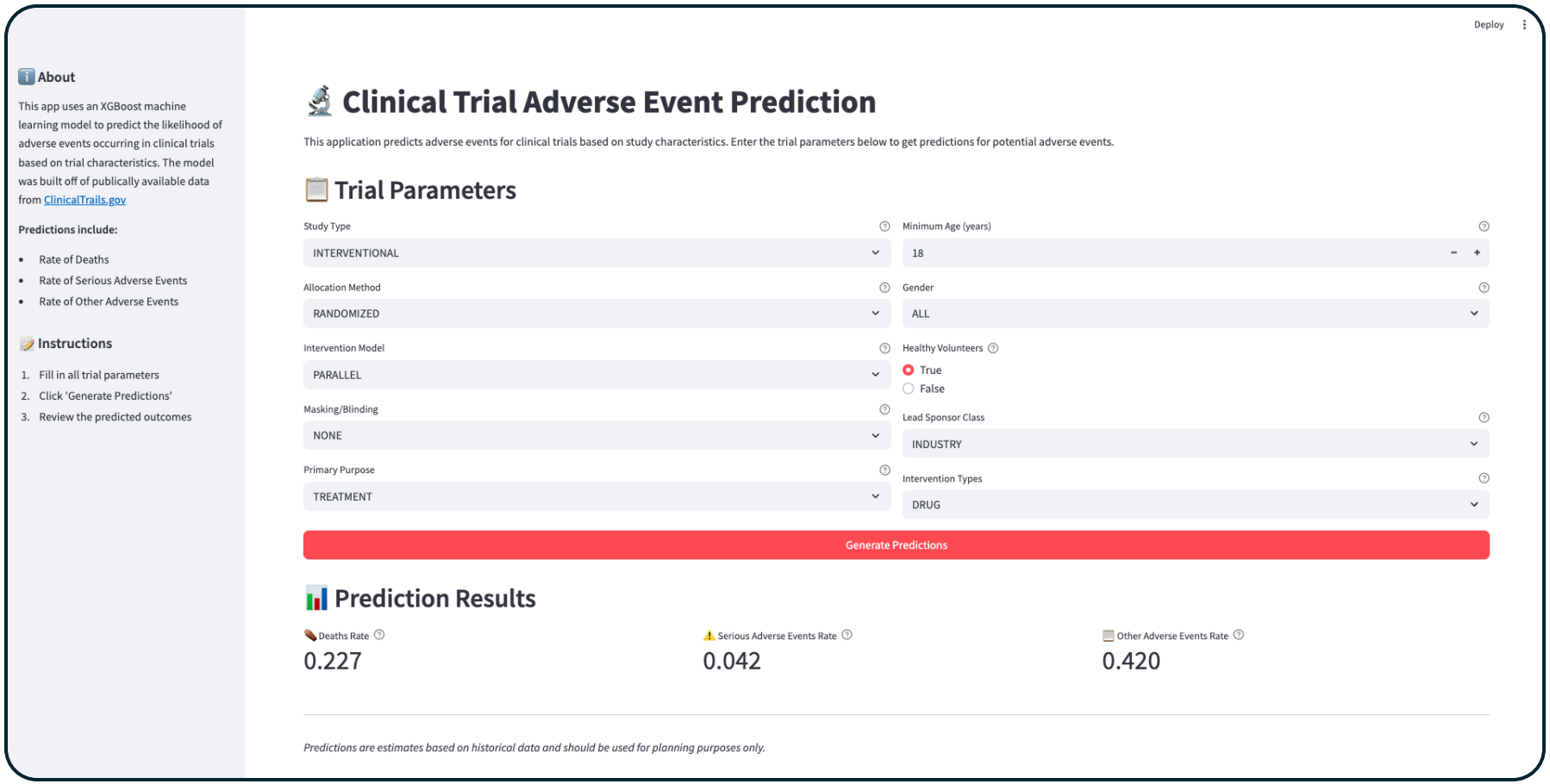
- death_rate
- serious_adverse_event_rate
- other_adverse_event_rate

Methodology

Model Choice: XGBoost

- Selected for its ability to handle **missing data**, **nonlinear relationships**
- Robust performance with **mixed-type input data** (categorical + numerical)
  - Can handle **categorical variables** natively **without encoding**
- Suitable for **multi-output regression** tasks

Web App



Clinical Trial Adverse Event Prediction