TEAM BIOLOGITS

PS-4

Paidi Geetesh Chandra
Raghuveer V
Lalit Karthikeyan M A
Udit Mohan Pradhan
Pratyush Dash

INTRODUCTION

The Problem

Patients are often willing to consent to participation in a clinical trial if they believe that they have an opportunity to receive better treatment or if the results can help others [29,45,89]. Still, failing to enroll a sufficient number of subjects in a trial is a long-standing problem [82,101]. A study of 114 trials in the UK [10] indicated that only 31% met enrollment goals. In addition, Campbell et al. [15] reported that one-third of publicly funded trials required a time extension because they failed to meet initial recruitment goals.

Feller [39] reported that 25% of cancer trials failed to enroll a sufficient number of patients, and 18% of trials closed with less than

Contemporary Clinical Trials Communications 11 (2018) 156-164

Therefore, the biggest opportunity for sponsors to accelerate clinical trials is to increase the speed and improve the efficiency of clinical trial enrollment; however, participant recruitment is increasingly difficult. For example, the rate of clinical trial participants enrolled per site per month in oncology and nononcology Phase 3 trials declined by 14 percent and 54 percent, respectively, in the periods 2012 to 2014 and 2021 to 2023 (Exhibit 2).

Accelerating clinical trials to improve biopharma R&D productivity, McKinsey & Co

Inadequate recruitment remains a persistent challenge. Understanding recruitment rates and identifying appropriate benchmarks are critical for optimising this process. **IS THERE ANY WAY WE CAN SOLVE THIS???**

Our Solution

We present a **data-driven predictive model** to solve this problem. By incorporating real-world variables, **including internal and external data**, our feature engineering goes beyond traditional methods, capturing the multifaceted nature of clinical trial recruitment.

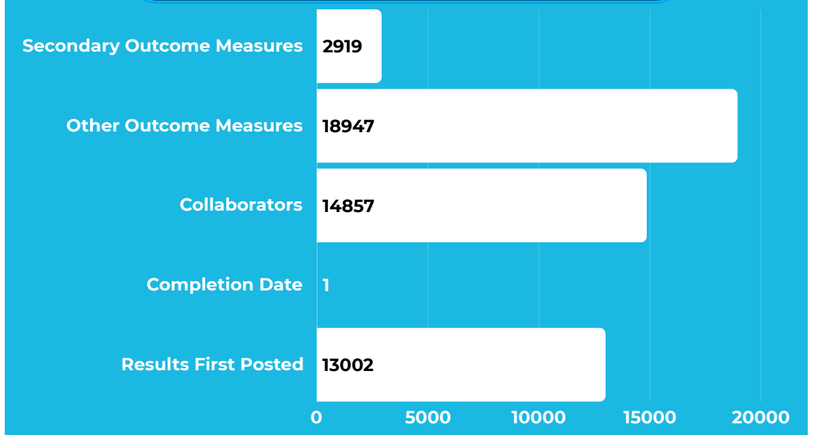
A key advantage of our model is its **lightweight design**, achieved through its efficient boosting algorithm. This architecture is **straightforward to deploy**, **maintain**, and **scale**, making it ideal for production environments. The model's simplicity ensures **quick new data integration** even as resources evolve.

The model features a simple interface where users can **upload new CSV files** (formatted similarly to clinicaltrials.gov) or **enter data manually**. This flexibility empowers people to effortlessly update inputs and generate new predictions.

Our model stands out by delivering accurate, actionable insights that transform trial planning, optimise resource allocation, and drive strategic decision-making, all while offering effortless implementation and low maintenance overhead.

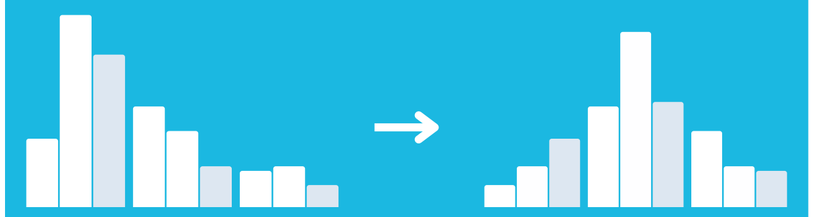
EDA AND PREPROCESSING

Handling NaN Values



Number of missing values in each column

Data Transformations

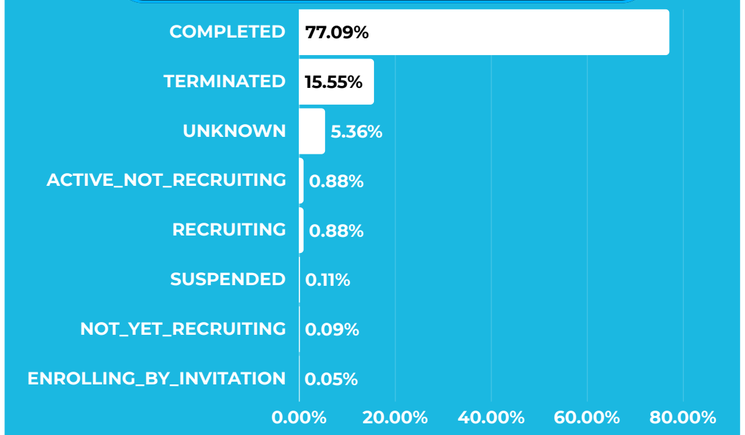


Conducted **Exploratory Data Analysis** using statistical summaries and visualisations to reveal trends and outliers.

Split the study design column into four separate components. Calculated **trial duration (in days)** from start and end dates. Retained the NCT Number as an identifier and **dropped descriptive fields** like titles, URLs, summaries, and dates. Removed interventions and outcome measures to **focus on recruitment-relevant data**.

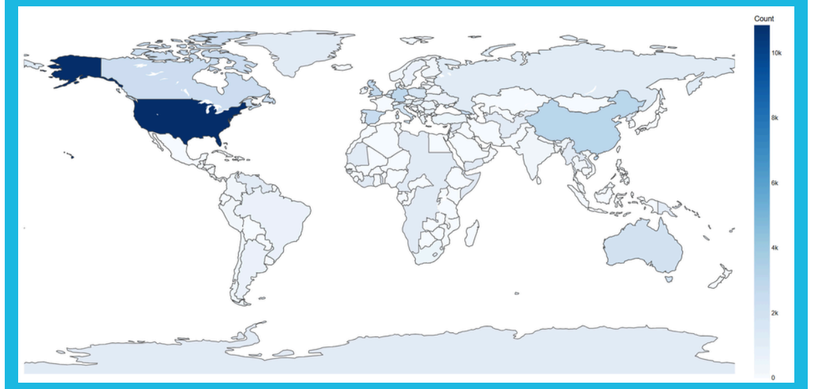
One-hot encoded categorical variables, replacing missing values with 0, and **imputed numerical missing values** with the mean after removing the outliers.

Feature Distribution



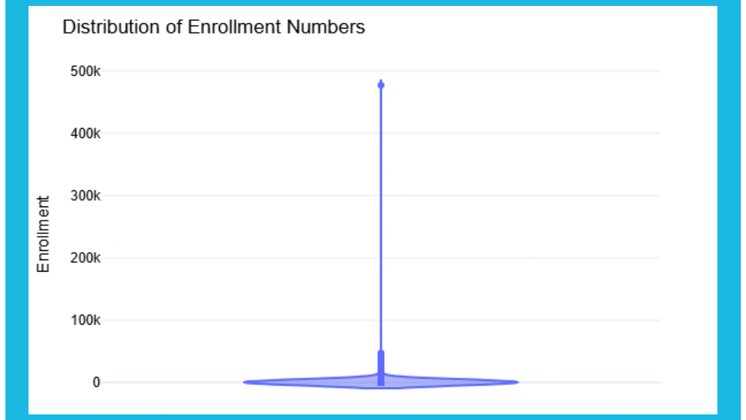
Percentage of trials by overall status

Data Visualisation



Number of Trails by Country

Outlier Detection



Violin plot of enrollment number

FEATURE ENGINEERING

Mesh ID

Mapped conditions to MeSH terms, reducing 7,860 values to 1,328 (106 MeSH IDs).

Used **MedEmbed-small** for embeddings and semantic matching.

Retrieved tree IDs (e.g., C04, D03) via MeSH API for structured representation.

Inclusion and exclusion criteria

Scraped inclusion/exclusion criteria **from study URLs** in the dataset. Measured complexity using **first- & second-order Shannon entropy**.

Entropy quantified eligibility complexity, revealing a clear link between criteria variability and recruitment dynamics.

Geographic Data

Extracted **city and country names** from location data via ClinicalTrials.gov API. Matched city names with **SimpleMaps' World Cities Database** (47,000 cities). to get population data. Included city population as a feature to assess participant availability.

Sponsor Success Rate

Calculated each sponsor's success rate **(successful trials/total trials)**to reflect the influence of experience on outcomes.

Sourced data from **HINT** (Hierarchical Interaction Network for Clinical Trial Outcome Prediction) Paper.

Disease Prevalence

Used **DALYs** from WHO to quantify disease prevalence. Higher prevalence suggests a larger pool of eligible participants.

Aligned conditions with WHO's four-level hierarchy via **semantic matching**. Used **MedEmbed-small** model to ensure accurate condition mapping.

Competition

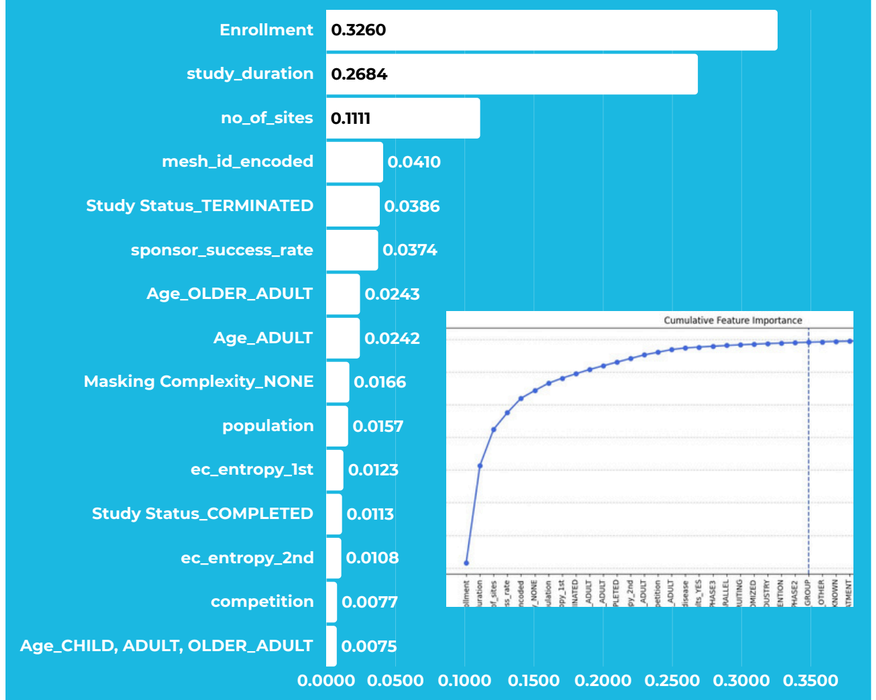
Measured clinical trials recruiting simultaneously in each city.

A higher count indicates more competition, which may limit the number of available participants for any given trial and ultimately impact recruitment success.

FEATURES AND MODEL SELECTION

Feature Selection

We applied **Random Forest feature selection** to rank our 56 features and retained the **top 26**, which account for **99% of the cumulative importance**.



We engineered 9 features, all of which ranked among the top 17, with the remaining features preprocessed from raw data.

Model Selection

Model	R2	MAE
Linear Regression	-0.0796	8.2159
Random Forest	0.6747	1.1382
Gradient Boosting	0.7654	2.3227
XGBoost	0.3841	1.4528
CatBoost	0.5893	1.6472
Decision Tree	0.6475	1.1870
LightGBM	0.3275	2.4417

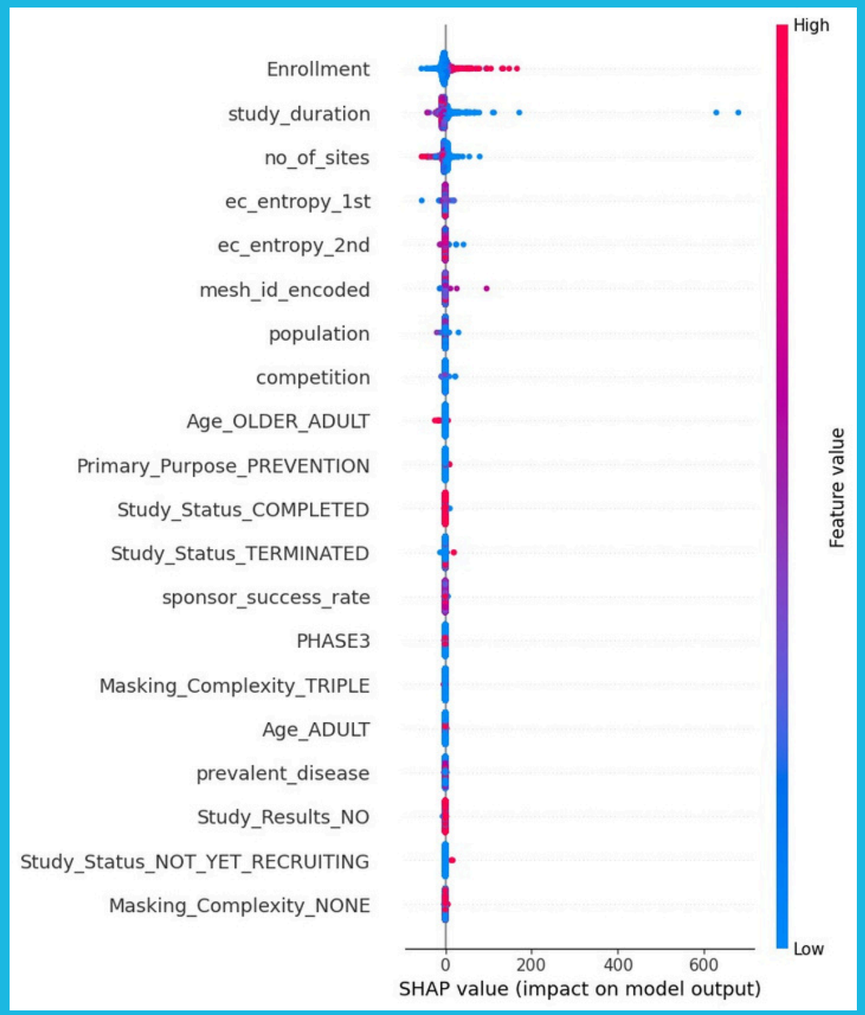
In the baseline testing phase, several Machine learning models were evaluated. The best performance was achieved by **Gradient Boosting**, with an **R2 score** of **0.7654** and an **MAE** of **2.3227**. **Random Forest** also performed well, achieving an **R2 score** of **0.6747** and **an MAE** of **1.1382**. These ensemble methods excel because they combine the strengths of multiple decision trees to capture complex, non-linear patterns while reducing overfitting.

Hyperparameter Tuning: Following the baseline testing, hyperparameter tuning was performed using **Optuna** to fine-tune the models. After 30 trials on GB and RF, Gradient Boosting gave the best results on our test set, with an **R2 score** of **0.8649** and an **MAE** of **1.824**.

HyperParameter	Search Range		Best Value
n_estimators	340	420	407
learning_rate	0.1	0.18	0.145
max_depth	4	7	7
min_samples_split	15	30	22
min_samples_leaf	5	20	11
subsample	0.5	0.8	0.744

EXPLAINABILITY

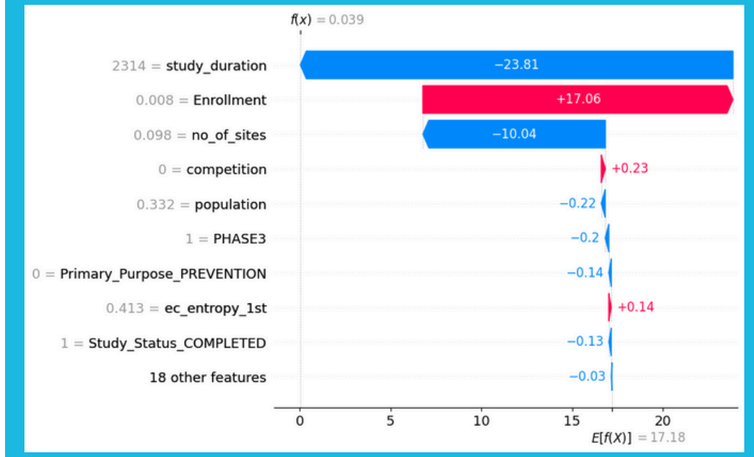
Global SHAP



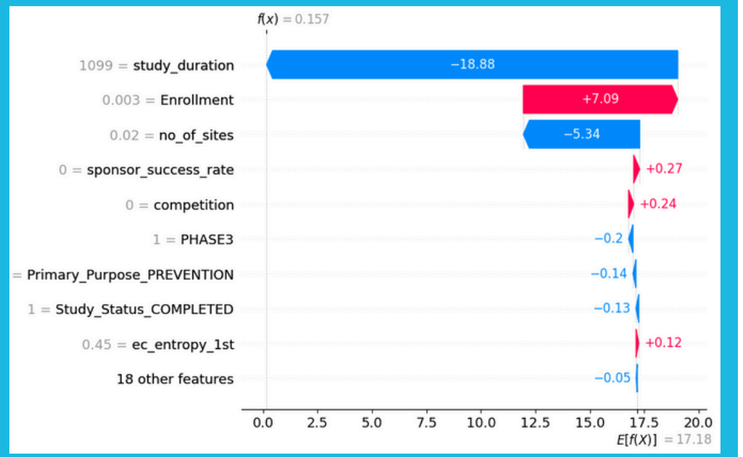
Explanation

SHAP (SHapley Additive exPlanations) helps interpret machine learning models by assigning importance scores to features. **Global SHAP** shows the average impact of each feature across all predictions, helping identify key factors influencing the model. Here, we can see that, on average, **Enrollment number** and **study duration** are the biggest drivers of trial outcomes, followed by **no. of sites**, the **inclusion and exclusion criteria** and so on. **Local SHAP** explains individual predictions by showing how each feature contributed to a specific outcome. This helps understand why a model made a certain decision.

Local SHAP



Actual value : 0.03911
Predicted value : 0.04003



Actual value : 0.15746
Predicted value : 0.15785

INDUSTRY SCOPE AND FUTURE WORK

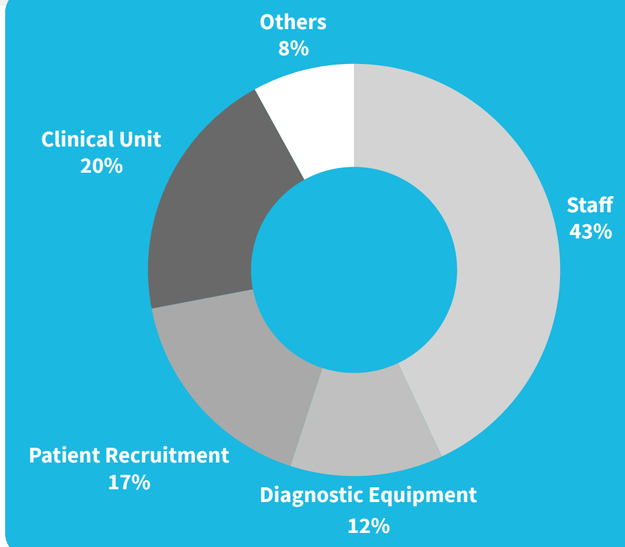
Industrial Value

Clinical trials are crucial for biopharma R&D but **face significant challenges** in **participant recruitment**. The demand for trial participants has increased by nearly **10%** over the past decade, with total target enrollment **growing 18%** from 2019 to 2022. However, participation rates remain low, with only **about 6% of U.S. cancer patients joining trials**. Recruitment is resource-intensive and often uneconomical, with complex protocols requiring specialized equipment and training.

High staff turnover and insufficient financial incentives for physicians further complicate the process. The **uneven distribution** of participants **across geographies** and indications leads to **inefficient resource allocation**, with many trial sites yielding few or no eligible participants. These challenges result in wasted financial and operational resources, undermining the economic sustainability of clinical trials

Accelerating clinical trials to improve biopharma R&D productivity, McKinsey & Co

Financial Effect



Clinical trial delays due to recruitment and retention challenges result in significant financial losses for pharmaceutical companies. It's estimated that each day of delay can cost between **$600,000 and $8million**. The average cost to recruit one patient for a clinical study was over **$6,500** in 2015-2016, while replacing a lost patient cost around **$19,000**. These financial burdens, coupled with the fact that **80% of trials are delayed** by at **least a month** due to recruitment issues, highlight the importance of efficient recruitment.

Future Work

Exploring deep learning models and advanced embedding techniques could enhance feature representation. Dynamically updating the competition factor through **real-time API calls** would improve usability. Sourcing a larger and **more detailed disease prevalence dataset** could help address computational challenges in semantic search with open-source LLMs.

Incorporating external factors such as **political regulations** at clinical sites could be valuable. Some countries have stricter guidelines, which may create potential conflicts and impact trial feasibility.

THANK YOU