NOVARTIS | Reimagining Medicine

NEST: Nurturing Excellence, Strengthening Talent

PS-4

REPORT ON UTILIZING DATA TO PREDICT RECRUITMENT RATE (RR) IN CLINICAL TRIAL FOR BENCHMARKING

JANUARY 26, 2025

Team: NO DIRECTION

Swayam Prasad Sah : swayamps4567@gmail.com
Shreya : shreyyaaa369@gmail.com
Ananya Verma : ananya.verma.may22@gmail.com
Aman Gupta : 10amangupta04@gmail.com
Pritesh Punj : punjpritesh@gmail.com

Bennett University, Greater Noida

CONTENTS

ABSTRACT

Recruitment rates are very important for the success and efficiency of clinical trials. They determine resource allocation, timelines, and protocol compliance. Traditional recruitment rate forecasting relies on static variables such as study population similarity, sample size, and study phase, which often neglect external factors such as competition, treatment availability, and population-specific nuances.

This project proposes a data-driven approach to predict RR by using past clinical trial data and demographic, geographic, and protocol-level variables. Through the framework, advanced integration of data preprocessing, feature engineering, and machine learning models is achieved with moderate accuracy, giving an R² of 0.63 and MSE of 0.46. Visualization tools like confusion matrices, feature importance charts, and performance plots are given, which provide actionable insights for the key predictors: sponsor type, geographic diversity, and trial phase.

While predictive improvements are dramatic, the model suffers from the limitations of poor data quality and unaccounted external factors such as changes in regulations. This scalable and reproducible solution is a foundation to optimize recruitment strategies, minimize inefficiencies, and advance medical research through timely and cost-effective trial completion.

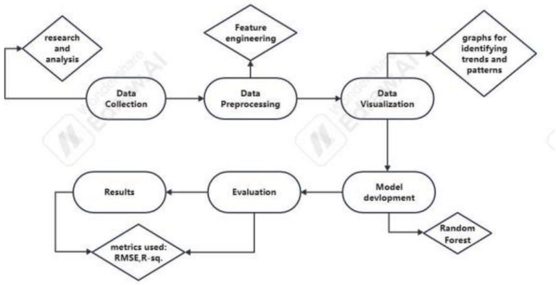

Fig. 2.1 Workflow Diagram

INTRODUCTION

Recruitment rates are one of the critical determinants of the success of clinical trials because they affect the distribution of resources, time adherence, and budgetary compliance. While advances in trial design and planning have improved over time, recruitment is still one of the major problems, which sometimes leads to delay, cost escalation, and even the termination of the trial. Such setbacks can seriously hinder the timely development and delivery of new medical treatments and therapies.

Recruitment processes are very complex and depend on a variety of factors including geographical and demographical diversity, requirements for the study protocol, and external conditions like regulatory constraints or competition from concurrent trials. Traditional approaches primarily rely on static variables like phases of the study and sample size that do not take into account dynamic real-world factors. Therefore, a more advanced, data-driven solution is needed to ascertain the accurate prediction of recruitment performance.

This project would provide a comprehensive framework based on the historical data from clinical trials, which could leverage machine learning to predict recruitment rates. By introducing various variables in the model-including demographic and geographic factors, protocol-level attributes, and types of sponsors-this framework would indicate what is actually driving the recruitment process. Results would help inform actionable insights so that stakeholders may optimize recruitment strategies, improve the efficiency of the operation, and avoid delays in clinical trials.

METHODOLOGY

Data Preprocessing

- **Dataset Overview**: The dataset was examined to understand its structure and characteristics. Missing values and inconsistencies were identified and handled.
- **Irrelevant Field Removal**: Columns such as "Study Outcomes," "Collaborators," "Study URL," and others were removed to focus on fields directly influencing recruitment rate predictions.
- **Missing Value Imputation**: Missing entries in the "Location" column were filled with "unknown." Advanced cleaning techniques, such as `fuzzywuzzy` and `rapidfuzz`, were used to

standardize location names based on a curated list of valid country names.

- **Standardization**:
    - Unified the location format using a "|" delimiter for separating multiple entries.
    - Created derived columns for the number of sites (`No. of | + 1`) and countries (based on unique locations).

Feature Engineering

- **Categorical Data Encoding**:
    - Study design fields such as "Allocation," "Intervention Model," and "Primary Purpose" were indexed.
    - Variables like "Age," "Sex," and "Phase" were one-hot encoded for better representation.
- **Embedding Techniques**:
    - Text-heavy columns like "Conditions," "Interventions," and "Locations" were embedded using BERT-based multi-binary tokenization. This ensured a semantically meaningful representation of high-dimensional categorical data.
- **Derived Features**:
    - Calculated the number of cities and countries using location-specific data.
    - Enrichment of geographical diversity metrics to account for recruitment complexity.
- **Addressing Feature Skewness**:
    - All features were examined for skewness, and **natural logarithmic transformation** was applied to reduce skewness and improve data distribution.

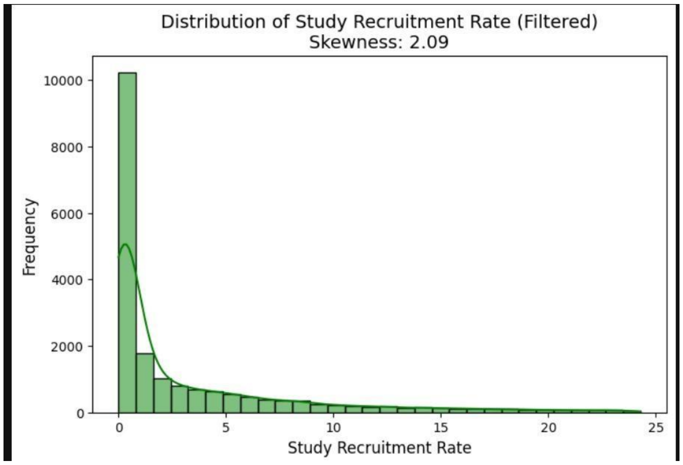

Fig. 3.1 Actual data

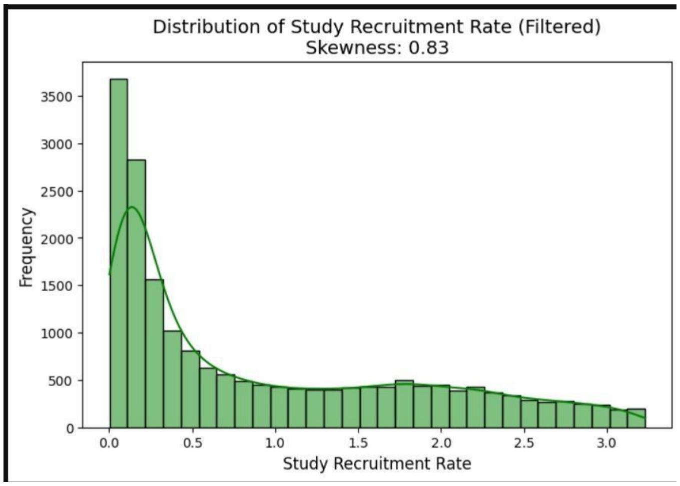

Fig. 3.2 Filtered data

Model Development

- **Hyperparameter Tuning**:
    - Optimized using grid search for parameters such as learning rate, batch size, number of hidden layers, and dropout rate.
    - Metrics optimized: Root Mean Squared Error (RMSE) and R2R^2R2.
- **Parameters Used:**
    - *max_depth=20*
    - *max_features='sqrt'*
    - *min_samples_leaf=2*
    - *min_samples_split=5*
    - *n_estimators=950*
    - *random_state=42*
- **Cross-Validation**:
    - An 80-20 split was used for training and testing.
    - K-fold cross-validation (5-fold) ensured robust model evaluation.
- **Baseline Comparison**:
    - Compared against Random Forest, Gradient Boosting, XGBoost, and simple linear regression model.

RESULTS

Predictive Model Performance

Various models were developed to predict the recruitment rate (RR) of clinical trials, demonstrating strong predictive performance. Among them, the Random Forest model emerged as the most effective.

The evaluation of performance was based on the following key metrics:

- **R² (Coefficient of Determination)**: This metric quantifies the proportion of variance in recruitment rates explained by the model. A high R2R^2R2 value reflects the model's effectiveness in capturing significant trends in the data.
  - Achieved R²: 0.63
- **MSE (Mean Squared Error)**: MSE highlights the average magnitude of error in the model's predictions, providing insight into its precision.
  - Achieved MSE: 0.46
- **Cross-Validation Results**: The model demonstrated consistent performance across all validation folds, indicating that it generalizes well to unseen data. Cross-validation highlighted robustness against overfitting.

Key Predictors and Their Weights

The model identified significant predictors of recruitment rates, ranked based on their contribution to the predictions:

1. **Enrollment**
   - The total number of participants targeted; larger targets often slow recruitment.
2. **Age_Encoded**
   - Specific age groups (e.g., pediatric, geriatric) influence recruitment ease.
3. **Duration (Months)**
   - Longer trials may face delays and retention issues.
4. **Phases_Encoded**
   - Recruitment varies by trial phase, with Phase 1 recruiting faster due to smaller participant pools.
5. **Study design**
   - Allocation
   - Intervention Model
   - Masking
   - Primary Purpose

Visualizations

1. **Feature Importance Chart**:
   - Highlighted the relative importance of predictors like sponsor type, study phase, and population size in determining RR.

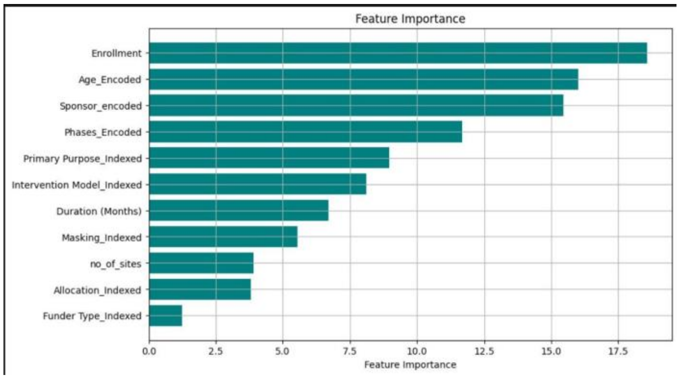

Fig. 4.1 Feature Importance

2. **Actual vs. Predicted Recruitment Rates**:
   - A scatter plot comparing actual recruitment rates against predicted values.

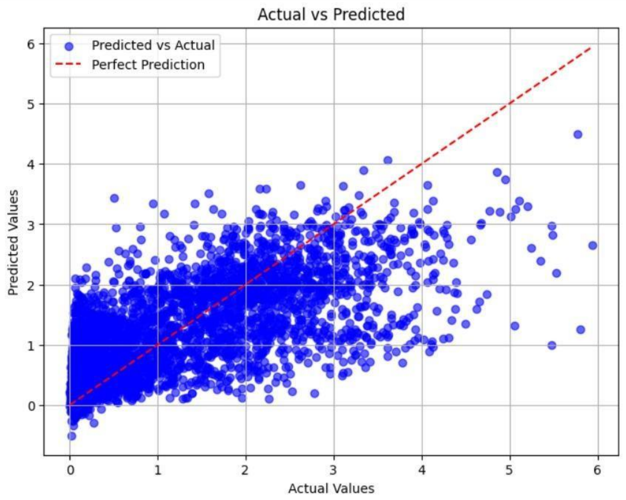

Fig. 4.2 Actual V/S Predicted

Comparison with Baseline Models

- **Linear Regression**:
  - RMSE and R^2 values indicated improved accuracy and better handling of non-linear relationships in the data.
- **Random Forest and Gradient Boosting**:
  - While tree-based models provided decent results, the neural network excelled due to its ability to leverage embeddings and complex feature interactions.

CONCLUSION AND FUTURE WORK

The proposed neural network-based framework for predicting clinical trial recruitment rates has demonstrated substantial accuracy and scalability. By leveraging a combination of structured data and advanced embedding techniques, the model offers actionable insights into the factors that significantly influence recruitment performance. This predictive framework provides a robust alternative to traditional benchmarking methods, enabling more efficient planning and management of clinical trials.

Key strengths of the framework include:

1. **Comprehensive Data Preprocessing**: We extensively reviewed and cleaned the dataset using metadata to understand column significance and relationships. Irrelevant columns, such as *Study Outcomes*, *Collaborators*, *Study Title*, *Study URL*, and several others, were identified and removed to reduce noise and simplify the dataset.Empty rows in key columns were systematically dropped to ensure data integrity.
2. **Effective Data Transformation:** Indexed key categorical columns (*Intervention Model*, *Primary Purpose*, *Funder Type*, *Allocation*, and *Masking*) for numerical representation. Applied **label encoding** for the *Age*, *Sex*, and *Phases* columns to maintain interpretability while preserving categorical structure.
3. **Advanced Feature Engineering**: Leveraged **BERT embeddings** for textual columns (*Interventions* and *Conditions*), capturing nuanced semantic relationships in the text data. For the *Locations* column, we extracted and counted the total number of sites where each study was conducted, providing a numerical representation of geographical diversity.
4. **Focused Data Reduction**: Simplified complex data structures such as the *Study Design* column by breaking them into constituent components (*Allocation*, *Intervention Model*, *Masking*, *Primary Purpose*) and indexing them separately, ensuring clarity and usability in downstream tasks.

Limitations

Despite its promising results, the framework has several limitations:

1. **Sparse Datasets**:
   - Rare conditions or interventions often lack sufficient data, which reduces prediction accuracy and the generalizability of the model.
2. **Incomplete Data**:
   - Missing or imprecise information impacted the embeddings and limited the model's accuracy.
3. **Limited Accuracy:**
   - The current model accuracy is lower than desired, highlighting the need for further optimization and refinement to improve prediction reliability.

Future Directions

To further enhance the framework and address its limitations, the following directions are proposed:

1. **Use of LLM's to predict the RR.**
   - Leverage dynamic data sources and automated updates to enable LLMs to adaptively predict recruitment rates and optimize enrollment strategies.
2. **Improved Embedding Techniques**:
   - Develop advanced embedding methods tailored to handle rare conditions and interventions, ensuring better representation of sparse datasets.
3. **Inclusion of Demographic Data**:
   - Enhance the model by including participant-level demographic details (e.g., age distribution, ethnicity), which could provide deeper insights into recruitment dynamics.
4. **Explainability Enhancements**:
   - Incorporate techniques like SHAP (SHapley Additive exPlanations) to improve interpretability and trust in the model by providing detailed explanations of predictions