NOVARTIS

NEST

Nurturing Excellence,
Strengthening **Talent.**

P.S. 4-Utilizing data to predict recruitment rate (RR) in clinical trial for benchmarking

TEAM MEMBERS:

SATYAM KUMAR
RAUNAK RAJ
DHRUV BANSAL
KRITNANDAN
ANKITA KUMARI

Approach & methodology

Overview

P.S. OVERVIEW:-

- The solution addresses the problem of predicting the **Study Recruitment Rate (RR)** for clinical studies by implementing a structured approach, as this is one of the most critical steps in the **drug circulation process**.

OVERALL APPROACH:-

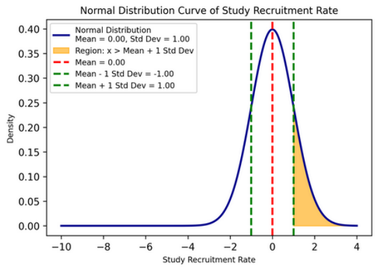
- The dataset is preprocessed by cleaning, handling missing values, and transforming categorical features into numerical representations.
- Textual embeddings are generated using **Large Language Models (LLMs)** via Transformer architectures like **AutoTokenizer** and **AutoModel** and are integrated with numerical features for enhanced predictive modelling.
- The data is then standardized, split into training and testing subsets, and used to train a **Regressor**.
- Performance evaluation is conducted using metrics by different libraries.
- The preprocessed dataset and trained model are saved for future use, ensuring reproducibility.

Methodology

OVERVIEW & PREPROCESSING

- The textual columns were cleaned for alpha-numeric values.
- The selected variables for the model include textual features such as **Study Title**, **Brief Summary**, **Conditions**, **Interventions**, and **Primary Outcome Measures**, as well as a numerical features.
- **Dates** were transformed to **Duration** for capturing its trend.
- Columns containing multiple categorical values like **Study Status**, **Sex**, **Age**, **Phase** were split and **one-hot encoded**.
- We opted to drop irrelevant factors that lacked sensibility for generalizing trends in **Recruitment Rate** predictions.
- Columns such as **NCT Number**, **Study URL**, **Locations**, **Other IDs**, **Sponsor** were dropped due to irrelevance with respect to target variable and too much missing values in the dataset.
- **Result First Posted**, **Collaborators**, **Other Outcome Measures** removed due to missing values along the dataset.
- The columns **Study Design**, **Interventions** contained diverse categorical information with multiple distinct categories within each column which were split to different columns & used in LLM.
- The Numerical columns were standardized using **Standard Scaler**.

ACCURACY METRICS



Frequency Distribution for target variable

MAE handles **skewness** well by providing stable error estimates, while **RMSE** ensures that larger errors are appropriately accounted for, which is crucial for datasets with **extreme values**. **Confidence Intervals** were used to get the trustworthiness of predictions in the presence of skewness.

Framework / tools used

- **Transformers:** To load **BioBERT** for extracting semantic embeddings from textual data.
- **PyTorch:** To utilize GPU-accelerated computations for efficient embedding generation.
- **Scikit-learn:** For training the Gradient Boosting Regressor and evaluating model performance.
- **NumPy:** For handling numerical computations and array manipulations efficiently.
- **Pandas:** For data preprocessing and managing structured datasets.
- **bayes_opt:** Used for efficient hyperparameter tuning via Bayesian Optimization.
- **Gc (Garbage Collection):** To optimize memory usage during batch processing.
- **Pickle**: To save and reload the trained model for future use.
- **Google Colab:** To provide a GPU-enabled environment for computationally intensive tasks.
- **Matplotlib & seaborn**: for visualization of data
- **Regex:** Used for text Cleaning

Model choice & setup

Model Selection

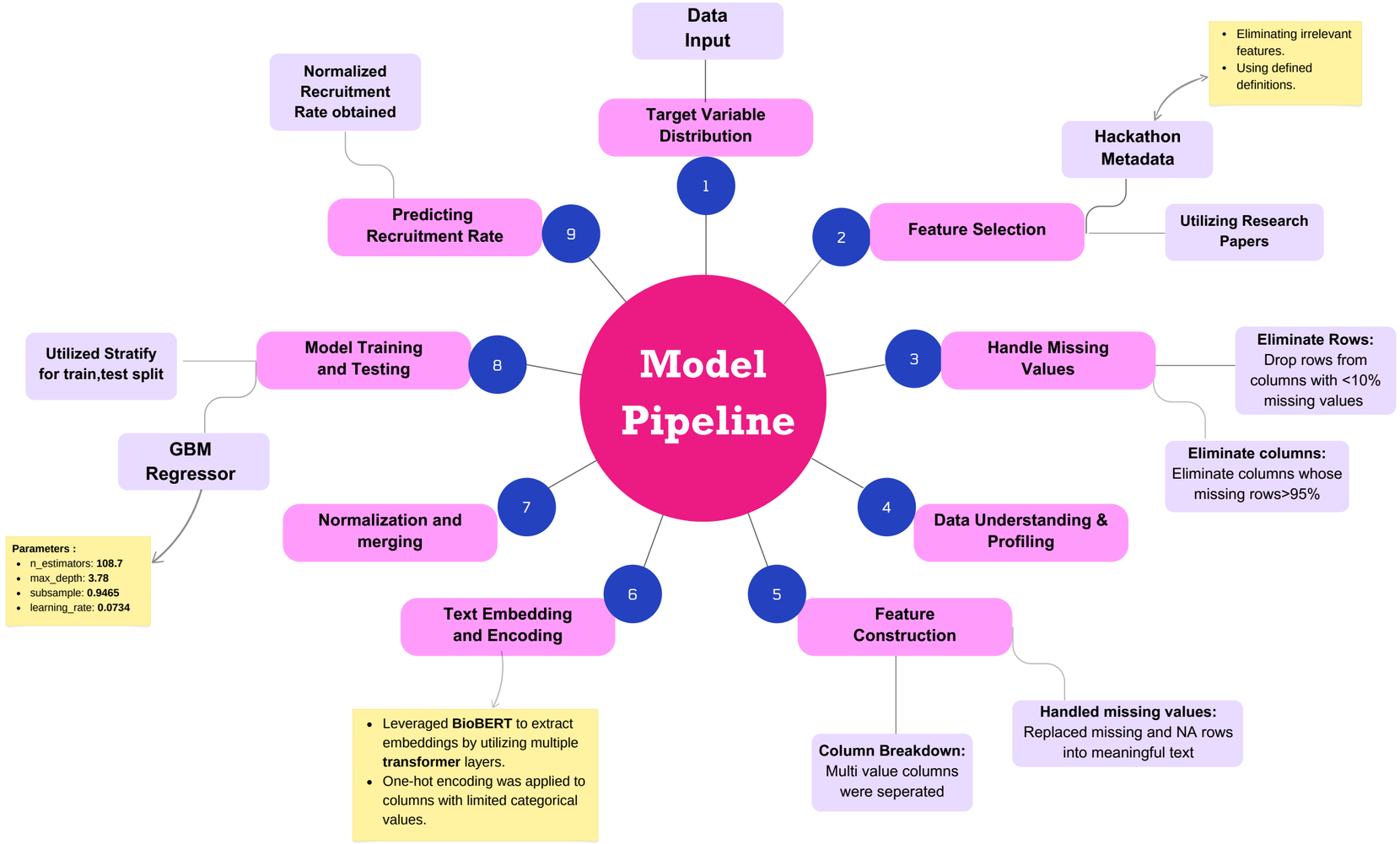
Bio BERT: -

- **Biomedical Specialization: Bio BERT,** trained on **PubMed** and **PMC** data, is tailored for extracting insights from biomedical text.
- **State-of-the-Art Performance:** Its domain-specific training outperforms general-purpose models in understanding biomedical language.
- **Semantic Analysis: Bio BERT** captures nuanced relationships in textual data, essential for analyzing factors influencing **Recruitment Rate**.

GBM Regressor: -

- **Robustness to Outliers:** GBM's boosting mechanism makes it resilient to outliers, ensuring stable predictions.
- **Nonlinear Relationship Handling:** GBM effectively captures complex, nonlinear interactions between features and the target variable.
- **Proven Track Record:** GBM is widely used in research for improving clinical trial recruitment, validating its reliability.

Model Architecture



inhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, Volume 36, Issue 4, February 2020, Pages 1234–1240,

Model Training & Evaluation

Evaluation Metrics

Model Training Process

Data Preparation: -

- **Feature Combination:** Embeddings are combined with numerical features to create the input dataset.
- **Tensor Conversion:** Data is converted into tensors for compatibility with GPU-based computation.

Validation Technique: -

- Implemented **Bayesian Optimization** using the **bayes_opt** library to fine-tune hyperparameters, incorporating a validation approach to evaluate the model's performance iteratively.
- Maintained separate train-test sets to prevent data leakage and ensure unbiased evaluation with **stratified** splitting to preserve class distribution.

Training Workflow: -

- **Gradient Boosting Model** was trained using optimized hyperparameters from Bayesian Optimization.
- **LightGBM Comparison**: Leveraged LightGBM for secondary benchmarking, ensuring robust model selection.

Optimization Details: -

- **Optimized hyperparameters:**
  - **n_estimators, learning rate, max_depth, subsample.**
  - Bayesian Optimization was performed over **20** iterations, with the **negative RMSE** on the validation set serving as the objective function

Evaluation Criteria and Metrics

We used three key metrics: RMSE, MAE, and R^2 Score, to evaluate the model comprehensively.

Root Mean Square Error (RMSE): -

- **Achieved RMSE: 0.34.**
- Indicates the ***model's average prediction*** error in target.
- It reflects how closely the GBM regressor could predict the target variable.

Mean Absolute Error (MAE): -

- **Achieved MAE: 0.083**
- Shows the average magnitude of prediction error, focusing on accuracy.

R-squared (R^2) Score: -

- **Achieved R^2 Score: - 0.45.**
- Demonstrates that ***45% of the variance*** *in the target variable* is explained by the model.

Performance Insights :-

- **High Precision**: **Low RMSE and MAE** indicate accurate and consistent predictions.
- **Explanatory Scope**: **R^2 Score** highlights moderate capability to explain variance in the data.
- **Potential Improvements:** Additional feature engineering or model tuning can further enhance performance.

Results and visualization

Model Outcomes

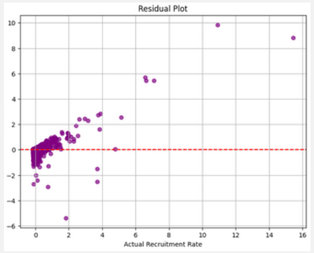
Model Performance and Key outcomes: -

- **RMSE:** 0.338 **MAE:** 0.087 **R^2 score:** 0.452
- These metrics highlight the model's reliability in explaining recruitment rate, variance and supporting actionable insights in clinical trial optimization.
- The model can identify underperforming sites or high-risk trials.

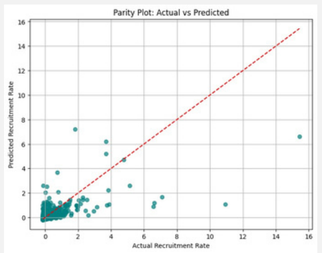
Implications: -

- Clinical trial managers can use these insights to proactively increase site support or optimize recruitment campaigns.
- The model's predictions enable efficient management of recruitment efforts, ensuring that timelines are met and risks are minimized.

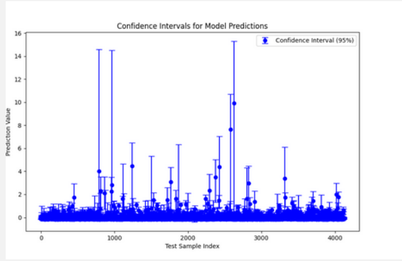
Visualizing Results and Outcomes: -



- The model predicts lower recruitment rates (< 4) accurately, as shown by the residuals being close to zero.
- For higher recruitment rates (>4), the residuals increase, indicating variability in predictions.



- The **correlation** of 0.67 indicates a moderate positive relationship between the predicted and actual recruitment rates, meaning the model captures the general trend of the data well.



- Most predictions have small confidence intervals, indicating high confidence, while a few samples show large intervals, suggesting uncertainty due to noise, outliers, or areas where the model underperforms.

Explainability

Trust in Model Decisions: -

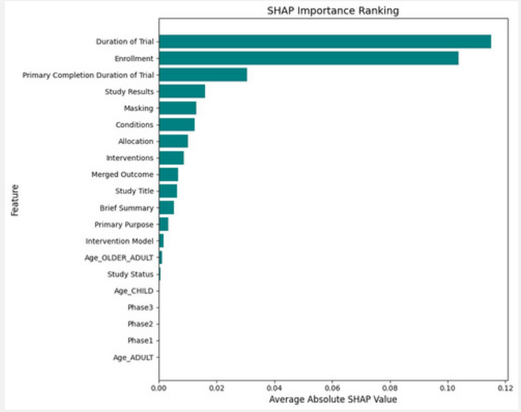
- In critical applications like clinical trials, especially for predicting the **Recruitment Rate (RR)**, explainability is essential to trust a model's decisions.

Explainability Technique - SHAP

- SHAP assigns a value to each feature, representing its contribution to the model's output using a game-theoretic approach.
- It helps to understand how individual features influence predictions for specific instances.

Key Insights from the Graph:

- "Duration of Trial," "Enrollment," and "Primary Completion Duration of Trial" are critical in predicting the target variable, indicating their strong correlation with Recruitment Rate.
- Duration of Trial = (Completion - Start) Date
- Merged outcome = (Primary outcome + secondary outcome) measures



SHAP Importance Ranking Graph

Comparison with Existing Research:

- These primary three features also play a significant role in the **RR** calculation in research papers by using below mentioned formula - [1] [2]

$$Recruitment\ Rate = \frac{Total\ Number\ of\ Participant\ Enrolled}{(Number\ of\ Sites\ *\ Duration\ of\ Recruitment\ Period)}$$

- Positive correlation and metric results validate the model's predictions, ensuring reliability.

Challenges & Next Steps

Limitations

1. **External Factors: Location & Sponsor Influence**
   - **Limitation:** While location and sponsor names do not directly aid generalization, linking location to population data from **external datasets** (specific to the trial period) and considering sponsor-related factors could provide valuable context.
   - **Impact:** Including population-based and sponsor-specific features could improve prediction accuracy by accounting for regional and sponsor-driven variations.
2. **Limitation of BioBERT Embeddings**
   - **Limitation:** **BioBERT's** F1 score of **0.62** for NER indicates moderate performance, falling short of larger LLMs like GPT-4 or Llama-3, which achieve higher accuracy but were **inaccessible** due to limited **GPU** resources
   - **Impact:** Inaccurate embeddings may reduce downstream task efficiency and limit opportunities for leveraging richer representations achievable with larger LLMs.
3. **Skewed Recruitment Rates and Overfitting Potential**
   - **Limitation:** Skewed recruitment rates and the gap between Training RMSE (0.185) and Test RMSE (0.338) suggest potential overfitting.
   - **Impact:** Addressing imbalance through resampling or weighted loss functions and improving regularization can enhance generalization.
4. **Modeling Temporal Dynamics**
   - **Limitation:** Capturing non-linear temporal dynamics is challenging, requiring techniques like **Temporal Fusion Transformers (TFT)** or **RNNs** with attention mechanisms to model time-dependent patterns.
   - **Impact:** Missing temporal trends may lead to inaccurate recruitment rate predictions, affecting decision-making and trial success.

Next Steps

- **Preprocessing Textual Columns Using OpenAI API:** Use OpenAI's **GPT-4o Mini** API to preprocess textual columns by extracting meaningful entities in a structured format. This enhances text representation but requires API access and sufficient budget for large-scale usage.
- **Dynamic Feature Selection**: Apply dynamic feature selection methods using **reinforcement learning** (e.g., Proximal Policy Optimization) to optimize the most relevant predictors for different trial phases.
- **Dynamic Equilibrium Strategy Optimization**: Develop a **dynamic equilibrium model** to optimize stakeholder strategies, adapting to **Recruitment Rates** and resource constraints.[1]
- **Fine-Tuning with Advanced LLMs:** Fine-tune **LLaMA-3.3(70B)** for recruitment rate prediction to utilize its advanced modeling capabilities. This requires extensive GPU resources, which exceed our current infrastructure limits.
- **Hyperparameter Optimization:** Utilize, **Bayesian optimization**, or Hyperband to fine-tune parameters like learning rate and regularization to boost generalization and reduce overfitting.
- **Continuous Learning Framework**: Set up a continuous learning framework where the model can be periodically retrained with new data, ensuring that it adapts to changes in language use or emerging biomedical knowledge.

C. U. O. Kumar, I. Singh and M. Suguna, "Optimizing Patient Recruitment for Clinical Trials: A Hybrid Classification Model and Game-Theoretic Approach for Strategic Interaction," in IEEE Access, vol. 12, pp. 10254-10280, 2024, doi: 10.1109/ACCESS.2024.3351688. keywords: {Recruitment;Clinical trials;Medical services;Game theory;Ensemble learning;Machine learning;Decision making;Ensemble learning;Feature extraction;Clinical trials;decision-making;ensemble learning;feature extraction;game theory;healthcare analytics;stacked ensemble;soft computing},

Thank You!!