NOVARTIS

NEST

Nurturing Excellence, Strengthening Talent.

PS(04)-Utilizing data to predict recruitment rate (RR) in clinical trial for benchmarking

Satyam Kumar - satyamkmr22@iitk.ac.in
Raunak Raj - raunakr22@iitk.ac.in
Dhruv Bansal - dhrubb22@iitk.ac.in
Kritnandan - kritnandan22@iitk.ac.in
Ankita Kumari - 22je0130@iitism.ac.in

Approach & Methodology

Overview of P.S.

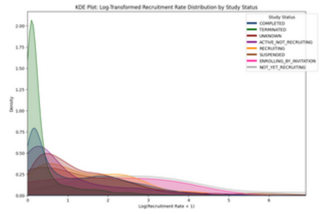
- The solution addresses the problem of predicting the **Study Recruitment Rate** (RR) for clinical studies by implementing a structured approach, as this is one of the most critical steps in the **drug circulation process**.
- Accurate RR prediction ensures that clinical trials meet their enrollment targets within the desired timeframe, minimizing delays in bringing new treatments to market.
- By leveraging AI/ML techniques, developing a solution that considers both internal and external factors would help us predict precise recruitment rates, enabling more effective planning and optimal resource allocation for clinical trials.

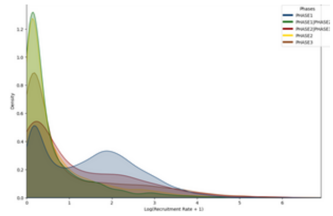
Overall Approach

- Our approach to predicting the **Study Recruitment Rate (RR)** began with analyzing the relationship between features and their impact on clinical trials.
- We processed date columns by transforming them into formats suitable for modeling and conducted **Exploratory Data Analysis (EDA)** to evaluate categorical, textual, and numerical features. Complex categorical features with multiple subcategories were refined and categorized, while textual data with potential relevance to the target variable was cleaned to enhance its significance.
- We leveraged **AutoTokenizer** and **BioBERT** to generate embeddings for textual features, capturing their semantic meaning. Binary and label encoding were applied to categorical features where appropriate, while less relevant features were dropped based on evidence and intuition. **Active Learning** was also considered as a strategy to intelligently acquire data points that could improve model training efficiency.
- Finally, numerical features and the target variable were standardized, and **Regressors** were trained to make predictions, resulting in a comprehensive and data-driven modeling pipeline.

METHODOLOGY

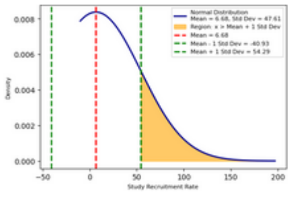
We start with analyzing our data along with the definitions of metadata file. The **NCT Number**, a unique identifier for each drug trial, was dropped as it does not contribute to trend identification. Similarly, the **Study URL**, containing unique web links, and **Collaborators**, listing contributors' names, were excluded as they provide no actionable insights or generalization value. While the **Sponsor** and **Funder Type** columns have potential utility, incorporating them meaningfully into the model would require additional **external datasets** to account for factors such as funding amounts. Lastly, the **Other Outcome Measures** column was dropped due to excessive missing values, as including it would have introduced bias into the model training, potentially compromising the reliability and accuracy of the predictions. These exclusions ensure the dataset remains focused and optimized for features that can be effectively leveraged with the current data.

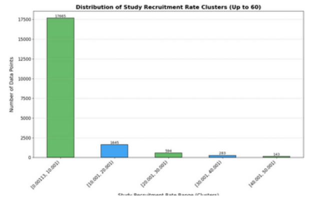




We then proceeded to analyze the **distribution of the target variable and its relationship** with various categorical columns. To achieve this, we exploded columns such as **Phases** and **Age** to better understand their contribution and performed **One-Hot Encoding** on the resulting exploded columns. The **Study Status** column, which included 7 categories across the dataset, was label encoded without being exploded. Additionally, we generated **KDE plots** for Phases, considering that clinical trials typically progress in a phase-wise manner, to visualize and interpret their distribution in relation to the **Log-Transformed** target variable.

We included columns like **Phases**, **Sex,** and **Study Status**, as their significant biomedical relevance, observed through the *Metadata Definitions* and cited *Research articles.*





We plot the target variable to analyze its **skewness**. The **Study Recruitment Rate** is **positively skewed**, with a mean of **6.68** and **a High Standard Deviation of 47.61**, indicating significant variability. Most values fall within one standard deviation **(−40.93 to 54.29)**, but negative values suggest the need for adjustments. The skewness highlights the potential benefit of normalization, such as Log Transformation, to improve symmetry and model performance.

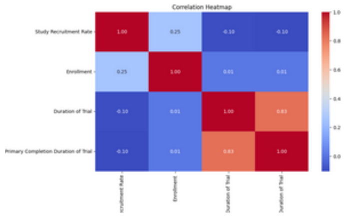
Next, we focus on analyzing numerical columns, starting with an assessment of their scope.

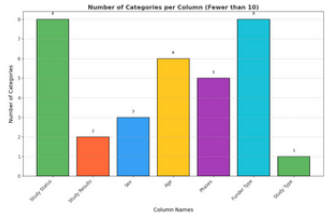
Date columns such as **Start Date**, **Primary Completion Date**, **Completion Date**, **First Posted**, **Results First Posted**, and **Last Update Posted** were available. However, the **Posted Dates** were **excluded** due to **high missing values** and their limited relevance to the recruitment process. Specifically, Results Posted Dates, which indicate when trial information became available on clinicaltrials.gov, are unlikely to have a significant impact on patient recruitment.

Instead, we focused on **Start Date**, **Primary Completion Date**, and **Completion Date**, as they provide the duration of the trial, a critical factor influencing recruitment rates. The relevance of these columns is **supported by citations** [1] in previously published research articles, reinforcing their importance in the context of recruitment analysis.

We now proceed to textual columns, crucial for extracting semantic insights and hidden patterns.

Key textual features selected for the model include **Study Title**, **Brief Summary**, **Conditions**, **Interventions**, and **Primary & Secondary Outcome Measures**, alongside numerical features. The **Study Design** column was further exploded into distinct features based on its values to capture more granular information. Since **BioBERT** is pretrained on medical text, we only needed to clean the textual columns by removing non-alphanumeric characters and lowercasing them. BioBERT's efficiency eliminates the need for additional preprocessing.





The numerical features—Enrollment, Duration of Trial, and Primary Completion Duration of Trial—were included due to their significant relevance to predicting the Study Recruitment Rate, as shown in the heatmap. Enrollment has a *positive correlation* **(r=0.25)** with the target variable, highlighting its importance in understanding how participant numbers influence recruitment efficiency. **Duration of Trial** and **Primary Completion Duration of Trial** are strongly correlated (r=0.83), emphasizing their shared role in capturing trial timelines, which are critical to recruitment dynamics. Although Duration of Trial and Primary Completion Duration of Trial show a weaker direct correlation with the Study Recruitment Rate (r=−0.10), their temporal significance in capturing recruitment efficiency and timelines justifies their inclusion. This is particularly relevant because other features in the dataset exhibit even lower correlations, making these features comparatively more valuable.

The graph on the right illustrates the number of categories with fewer than 10 unique values across the entire dataset. We then addressed **missing values** by intelligently removing rows after assessing their overall percentage and evaluating the *impact of data loss*, particularly for data points belonging to rare clusters.

LIBRARIES UTLIZED AND METRICS USED

Core libraries like **numpy**, **pandas**, and **os** manage computations and data manipulation, while **matplotlib.pyplot** and **seaborn** handle visualization. Machine learning tools include **sklearn**, **lightgbm**, **xgboost**, **catboost**, and **torch**. For Bayesian optimization, **gpytorch** and **botorch** are utilized, and NLP tasks are supported by the transformers library using **AutoTokenizer** and **AutoModel**. Additional tools like **joblib** and **shap** support model saving and explainability.

We used several metrics to evaluate model performance and ensure reliable predictions for the **Study Recruitment Rate**. **Confidence intervals** quantified prediction uncertainty, while **MAE** and **RMSE** measured error magnitude, with **RMSE** penalizing larger deviations. The **R² Score** assessed how well the model explained target variance, and residual plots ensured errors were randomly distributed, confirming model robustness and accuracy.

NOVARTIS

Model choice & setup

03

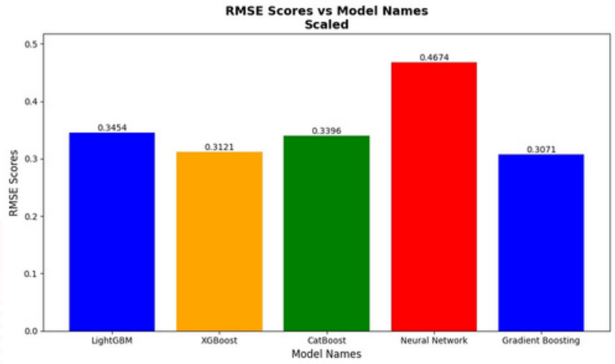
Why BioBERT for Biomedical NLP?

- Specialized for Biomedical NLP Tasks:
  - **BioBERT** is a language model specifically pre-trained for **biomedical natural language processing (NLP)** tasks, making it highly suitable for **embedding creation** in biomedical datasets.
- Domain-Specific Pre-training:
  - Trained on domain-specific corpora such as **PubMed and PMC**, enabling a deep understanding of biomedical terminologies.
  - The updated **BioBERT v1.1** was pre-trained for **1 Million steps**, enhancing its capability to capture context-rich representations.
- Performance on NLP Tasks:
  - BioBERT consistently outperformed **state-of-the-art models** on six out of nine biomedical NLP datasets.
  - BioBERT v1.1 achieved a **0.62 increase** in the **micro-averaged F1 score**, demonstrating its improvement over earlier versions.
- Excellence in Relation Extraction:
  - Achieved a **2.80 F1 score** improvement in biomedical **relation extraction** tasks.
  - Demonstrated effectiveness in identifying complex relationships such as gene-disease and protein-chemical associations.
- Highly Effective for Biomedical Embedding Creation:
  - BioBERT's domain-specific training and superior performance metrics make it an ideal tool for embedding creation in datasets involving medical terminologies.

Why Handling Complex Medical Text with BioBERT?

- Handling Complex Medical Terminologies:
  - BioBERT's pre-training on biomedical texts enables it to effectively understand and process intricate medical terminologies, such as **drug codes and trial descriptions**.
- Effective Tokenization with WordPiece:
  - BioBERT employs **WordPiece tokenization** and uses the **[CLS] token** to represent the entire input sequence, ensuring robust performance even with out-of-vocabulary words and domain-specific terms.
- Rich Context-Aware Embeddings:
  - BioBERT generates embeddings that are rich in context and tailored specifically for biomedical text.
  - These embeddings enhance feature extraction, improving **downstream machine learning** models' ability to identify patterns and trends.

Why Use GBM Regressor and not Others?



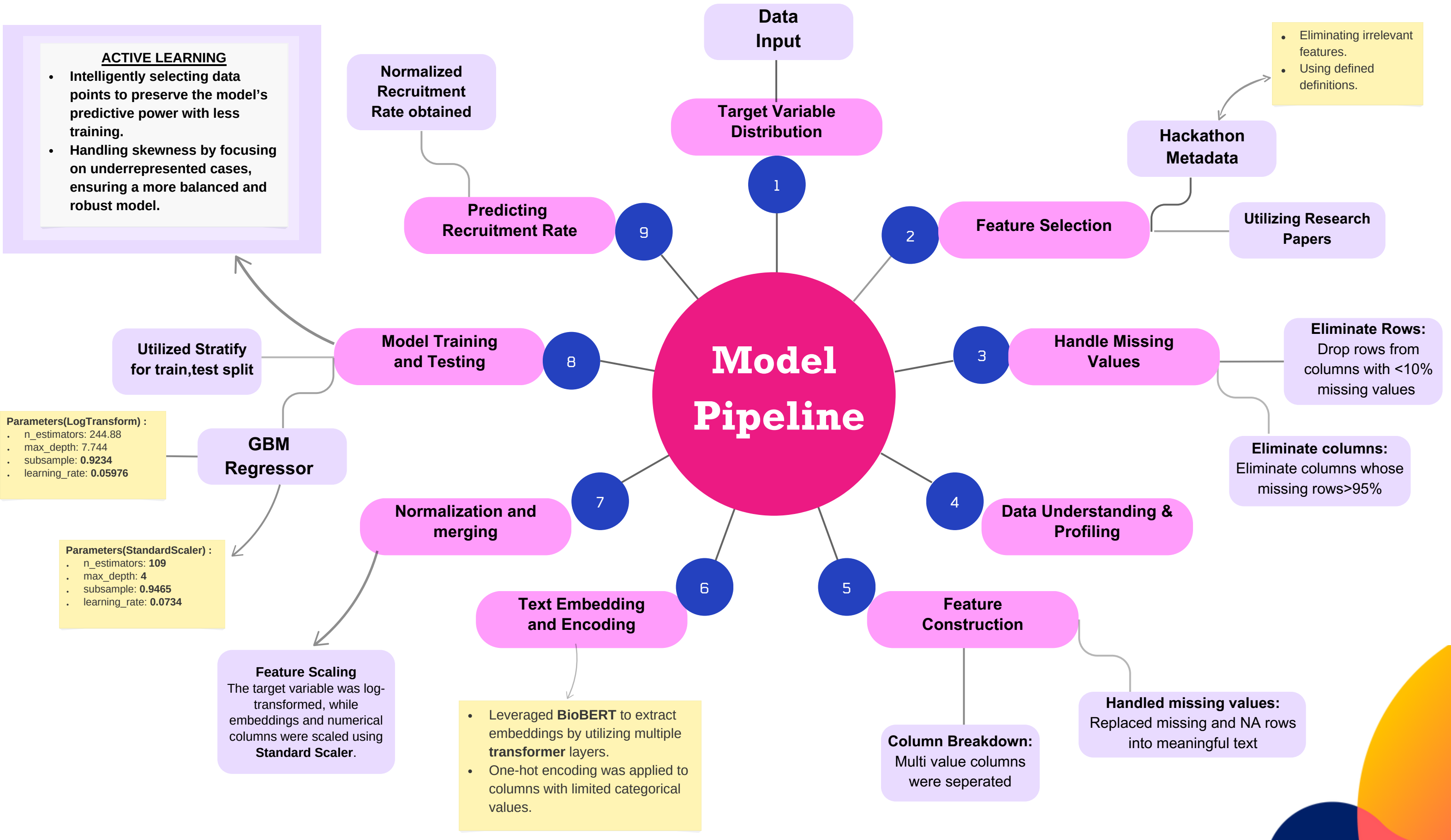
Gradient Boosting Regressor (GBM) is chosen for the following key reasons:

- Robustness to Outliers:
  - GBM's boosting mechanism reduces the impact of outliers, ensuring stable and reliable predictions.
- Nonlinear Relationship Handling:
  - GBM effectively captures complex, nonlinear interactions between features and the target variable, making it highly suitable for datasets with intricate patterns.
- Proven Track Record:
  - GBM is widely used in research and practical applications, such as improving clinical trial recruitment, validating its reliability and effectiveness in solving real-world problems.

Significance in Clinical Trial Use-Case

- Clinical trials demand accurate predictions for recruitment rates to allocate resources effectively and ensure timely execution. The use of GBM ensures that **subtle trends in the dataset**, which may be overlooked by simpler models, are captured with precision.
- Given the interplay of medical, logistical, and demographic factors, GBM's iterative boosting mechanism provides an edge by focusing on the **hardest-to-predict cases**, which often hold the most critical insights for trial success.

Model Architecture



Model Pipeline

Data Input

1 **Target Variable Distribution**

2 **Feature Selection**

Hackathon Metadata

- Eliminating irrelevant features.
- Using defined definitions.

Utilizing Research Papers

3 **Handle Missing Values**

Eliminate Rows: Drop rows from columns with <10% missing values

Eliminate columns: Eliminate columns whose missing rows>95%

4 **Data Understanding & Profiling**

5 **Feature Construction**

Column Breakdown: Multi value columns were seperated

Handled missing values: Replaced missing and NA rows into meaningful text

6 **Text Embedding and Encoding**

- Leveraged **BioBERT** to extract embeddings by utilizing multiple **transformer** layers.
- One-hot encoding was applied to columns with limited categorical values.

7 **Normalization and merging**

Feature Scaling
The target variable was log-transformed, while embeddings and numerical columns were scaled using **Standard Scaler**.

8 **Model Training and Testing**

Utilized Stratify for train,test split

GBM Regressor

Parameters(LogTransform) :
- n_estimators: 244.88
- max_depth: 7.744
- subsample: **0.9234**
- learning_rate: **0.05976**

Parameters(StandardScaler) :
- n_estimators: **109**
- max_depth: **4**
- subsample: **0.9465**
- learning_rate: **0.0734**

ACTIVE LEARNING
- **Intelligently selecting data points to preserve the model's predictive power with less training.**
- **Handling skewness by focusing on underrepresented cases, ensuring a more balanced and robust model.**

9 **Predicting Recruitment Rate**

Normalized Recruitment Rate obtained

NOVARTIS

Model Training & Evaluation

04

Model Training Process

Data Preparation:

Handling Missing Values

- Missing values in numerical columns were replaced with the median of the respective columns. This approach ensures that the central tendency of the data remains preserved.
- Textual columns with large proportions of missing values were dropped from the dataset to minimize noise.

Categorical Data Encoding

- One-Hot Encoding: Applied to categorical variables with a relatively small number of unique categories to create binary indicator variables.
- Label Encoding: Used for categorical features with a larger number of unique values to transform them into integer representations.

Exploratory Data Analysis (EDA)

- EDA techniques were employed to understand the underlying data distributions, detect outliers, and identify correlations between features. This step also involved feature selection to retain the most informative attributes and reduce redundancy.

Embedding Generation

- BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Texts) was employed to generate embeddings to handle textual data. BioBERT is specifically trained to process biomedical data, ensuring a contextual understanding of domain-specific text.
- These embeddings were combined with other numerical and categorical features to create the final training dataset.

Training Process:

Stratified Splitting

- The dataset is split into **training and Validation subsets** using **stratified sampling** to prevent data leakage and ensure an unbiased evaluation. This method preserves the class distribution across subsets, enhancing the reliability of model performance evaluation.

Data Transformation

- **Standardization** was applied to scale the combined input features, including numerical features, **one-hot-encoded** categorical variables, and textual embeddings, to same scale. This ensured that no feature **disproportionately influenced** the model training due to varying scales.
- For the target variable, a **Log transformation** was used to address its high skewness, compressing large values and making the regression task more balanced.
- After predictions, the log transformation was inverted to return predictions to the original scale. These preprocessing steps enhanced the model's ability to generalize, **reduced the impact of outliers**.

Validation Technique

Bayesian Optimization

- Bayesian Optimization is implemented using the **bayes opt** library to optimize model performance. This approach iteratively **fine-tunes** hyperparameters, balancing **exploration and exploitation** to achieve the best results on a validation set.
- The following **hyperparameters** were optimized:
  - Number of estimators, Learning rate, Maximum depth of trees, Subsampling rate.

Active Learning

- We utilized active learning leveraging acquisition functions like Expected Improvement and **Upper Confidence Bound** for selecting informative samples iteratively.
- This method is integrated into the training process by iteratively selecting and acquiring new data points based on the **acquisition function's** recommendations.
- **The most informative samples are identified** and **added to the training** set and then the model is retrained using both the initial data and the newly acquired samples.
- The resampling ensures balanced representation across the target variable. **Gaussian Process** models guided sample selection by predicting uncertainty.

Evaluation Criteria and Metrics

Three key metrics and following visuals are employed to evaluate model performance:

Root Mean Square Error (RMSE) $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$

- Achieved **RMSE: 0.3071**
- Interpretation: RMSE quantifies the average magnitude of prediction errors in the same units as the target variable. A lower RMSE reflects better predictive accuracy. In this case, the RMSE indicates that the model achieves a relatively high degree of precision, with predictions being close to the observed values.

Mean Absolute Error (MAE) = $\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$

- Achieved **MAE: 0.079**
- Interpretation: MAE provides an average measure of prediction error magnitude, focusing on accuracy without penalizing large errors disproportionately.

R-squared (R^2) Score $1-\frac{\sum_i(y_i-\hat{y}_i)^2}{\sum_i(y_i-\overline{y})^2}$

- Achieved **R^2 Score: 0.458**
- Interpretation: The R^2 score measures the proportion of variance in the target variable that is explained by the model. While an R^2 of 0.45 indicates that the model captures some patterns in the data, it also highlights the potential for improvement.

Results and Discussion

Performance Insights

- **High Precision:** The low RMSE and MAE scores reflect the model's capability to provide accurate and consistent predictions.
- **Explanatory Scope:** With an R^2 score of 0.45, the model explains a moderate portion of the variance in the data. However the **Log Transformation of the target variable yielded a significant boost in R^2 to 0.8**, revealing the model's capability to better capture relationships under certain transformations. However, this gain came with a slight increase in RMSE, suggesting a trade-off between explanatory power and predictive precision.

Comparative Analysis

- Utilized various models, including XGBoost and CatBoost, to compare their performance and identify the model that delivers the best prediction of Recruitment rates.
- This approach ensures a comprehensive evaluation and serves as a benchmark for selecting the most effective model.

Visual Evaluations

- **Residual Plot:** Residual plots were used to examine the distribution of errors. A randomly scattered pattern around zero confirmed the absence of systematic bias, while any clear trends highlighted potential areas for model refinement.
- **Prediction vs. Actual Plot**: These scatter plots compared the predicted values to the actual values, helping to assess the model's ability to approximate the true distribution. A higher density along the diagonal line indicated better predictive performance.
- **Confidence Interval Plot**: This plot visualizes the range within which predictions are expected to fall, considering the model's uncertainty. The confidence intervals highlighted the reliability of the predictions, with narrower intervals indicating higher certainty in model outputs. The inclusion of this plot enhanced our understanding of prediction confidence

Conclusion

- This study demonstrates the successful application of **Gradient Boosting Models** for predictive analytics, integrating Bayesian Optimization for hyperparameter tuning.
- While the model exhibits high predictive accuracy, there is scope for improvement in explanatory power and generalizability. Future efforts will focus on feature engineering and advanced modeling techniques to overcome these challenges.

Results and visualization

Model Performance and Key Outcomes: -

Performance Metrics:

Model Name	RMSE	RMSE (SCALED)	MAE	R2
LightGBM	0.3454	17.7731	0.0829	0.3150
GBM Regressor	0.3071	15.9883	0.0791	0.4584
XGBoost	0.3121	16.3760	0.0681	0.4407
CatBoost	0.3396	17.5442	0.0938	0.3378
Neural Network	0.4674	23.5620	0.1949	-0.2545

- When we used the **StandardScaler** for both features and the target variable then we got the above-mentioned result for different models
- Here, the GBM Regressor does better work where RMSE indicates low prediction errors on the scaled target, while the **R²** of **0.45** suggests the model explains only 45% of the target's variance. This suboptimal fit may stem from scaling both the features and target, limiting the model's ability to capture the full variability in the original scale.

Model Name	RMSE	RMSE (SCALED)	R2
GBM Regressor	0.4713	17.4784	0.8035
XGBoost	0.5068	18.1580	0.7910
CatBoost	0.5012	17.6090	0.7955

- When we used **log transformation** over target variables and Standard Scaler over features the R² score value significantly increased, this suggests that the log transformation helped to stabilize the variance and made the target variable's distribution more suitable for modeling, improving the model's ability to capture the variability.

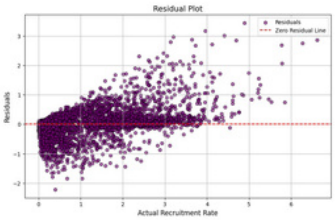
Implications:

- Clinical trial managers can leverage these insights to enhance site support or refine recruitment strategies.
- Predictions enable efficient management of recruitment efforts, ensuring timelines are adhered to and risks are minimized.

Visualizing Results and Outcome

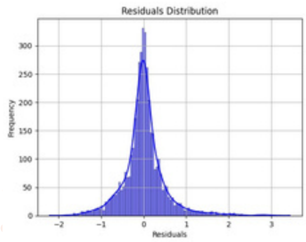
These visualization plots for the Log Transformation on target variable because for skewed data, log transformation is better as it reduces skewness, dampens outliers, improves interpretability, and allows the model to better utilize features. In contrast, the StandardScaler only normalizes the range without addressing the core issues of skewness and outliers, potentially leading to suboptimal model performance.

- ***Residual Plot:***



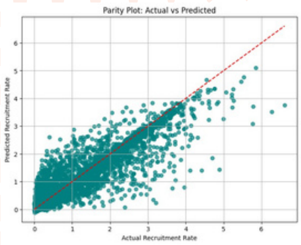
- Observation: The plot shows that most of the predicted and actual recruitment rates (RR) align closely for smaller values (RR < 4). However, for higher RR values (RR > 4), there is increased variability in the predictions.
- Reason: The limited number of rows with higher RR values, treated as outliers, may lead to insufficient training data for those cases, causing the model to struggle with accurately predicting higher RR values.

- ***Residual Distribution:***



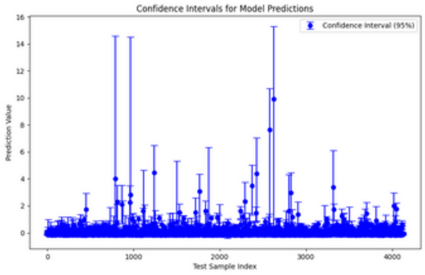
The residual distribution is centered around 0, forming a bell-shaped curve, indicating unbiased predictions and symmetrical error distribution. Most residuals are close to 0, reflecting accurate predictions with low variance. However, the long tails suggest a few outliers where the model struggled, potentially due to the higher value of the Recruitment rate. Overall, the model performs well for the log transformation over target variable.

- ***Parity Plot:***



- Observation: The **correlation** of **0.68** indicates a moderate positive relationship between the predicted and actual recruitment rates, this suggests that the model predicts RR accurately for log transformation over target variable. Points near the parity line highlight the model's ability to capture trends effectively
- Reason: Outliers in the data can disproportionately impact the correlation, lowering its value by introducing larger deviations between predicted and actual RR rates.

Confidence Intervals Graph:



- Observation: The plot shows predicted values with **95% confidence intervals**, with some predictions having wide intervals (greater uncertainty) and others having narrow intervals (higher confidence).
- Reason: Wide intervals may arise from less informative embeddings being generated, while narrow intervals, on the other hand, suggest that the embeddings effectively capture consistent and meaningful patterns in the data.

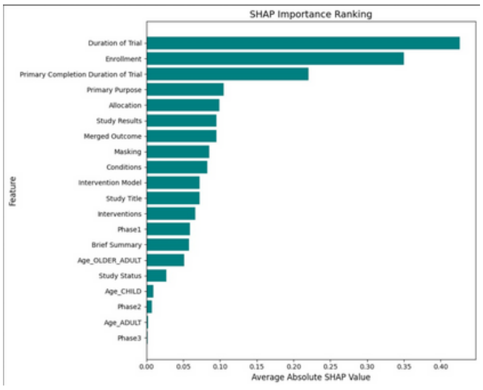
Explainability:

- SHAP assigns a contribution value to each feature, quantifying its impact on the model's output.
- This approach helps interpret how specific features influence predictions, enabling better decision-making.

SHAP (SHapley Additive explanations):

- It enhances trust by making the model's decision-making process transparent, allowing us to align model predictions with human reasoning.
- In critical applications like clinical trials, especially for predicting the Recruitment Rate (RR), trusting a model's decisions is crucial for ensuring accurate predictions, optimizing recruitment strategies, and maintaining ethical standards.

Key Insights from the SHAP Graph:



- **Duration of Trial**, the most impactful feature, indicates that longer trials provide more time for participant recruitment, directly enhancing recruitment rates. **Enrollment**, the second most impactful feature, highlights the importance of larger participant pools in validating trial effectiveness. The third feature, **Primary Completion Duration**, reflects the operational pacing of trials, linking it to recruitment efficiency and timely primary outcome data collection.
- While moderate-impact features such as **Primary Purpose**, **Allocation**, and **Study Results** influence the recruitment rate based on trial objectives and participant allocation methods, transparency in results can positively attract participants, thereby enhancing recruitment.

Comparison with SHAP analysis with Research papers:

- The top three features—"Duration of Trial", "Enrollment", and "Primary Completion Duration of Trial"—strongly align with established research findings and widely accepted recruitment rate formulas, validating the model's relevance and reliability in capturing critical factors influencing clinical trial success.
- Formula validation:[1] [2]

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

- https://www.corpuspublishers.com/assets/articles/cjct-v2-21-1006.pdf
- https://trialhub.com/resources/articles/clinical-trial-recruitment-rate-4-things-to-know

Challenges & Next Steps

Limitations

External Factors: Location & Sponsor Influence

- **Limitation:** Location and sponsor names lack external data enrichment (e.g., demographics, sponsor performance), reducing their contextual value. This limits the model's ability to capture regional and sponsor-driven variations, affecting prediction accuracy.
- **Enhancing Contextual Features with External Data:** Integrate demographic data like population density and healthcare infrastructure to enrich location features.
- Add sponsor metrics (e.g., funding capacity, trial history) and categorize them into tiers or performance scores for better representation.
- **Impact:** Enriched features improve prediction accuracy and generalization by capturing regional and sponsor-specific patterns.
- Provides actionable insights for optimizing trial strategies based on location and sponsor characteristics.

Skewness in Recruitment Rate

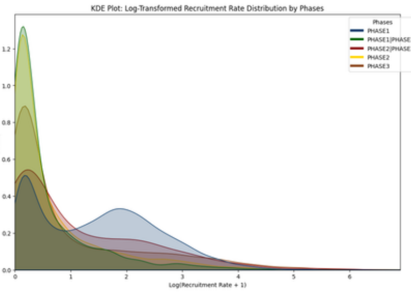
- **Limitation**: Skewed recruitment rate data causes issues with StandardScaler, leading to negative values that distort the distribution and affect model performance. This limits generalization, particularly for low recruitment rates.
- **Addressing Skewness with Alternative Transformations:** Use transformations like log, square root, or Box-Cox to normalize the distribution and prevent negative scaling.
- Apply resampling or GAN-based synthetic data generation to balance underrepresented ranges, if feasible.
- **Impact:** Improved data representation enhances model performance, particularly for skewed data, and ensures better generalization to low recruitment rates.

Impact of Early Termination on Recruitment Rates

- Many low recruitment rates correspond to trials terminated early due to unknown medical or ethical reasons.
- This could be a potential reason for the recruitment rates data being skewed, as our model is primarily trained on this data.
- The lack of contextual information biases the model, reducing its accuracy for terminated trials.
- These unknown factors make it challenging to differentiate between genuine low recruitment rates and early termination cases.
- **Controlled Selection for Bias Mitigation**
- **Selective Data Inclusion:** Implement controlled selection of data points during model training, focusing on excluding or down-weighting terminated trials with low recruitment rates.
- **Bias Reduction**: This approach ensures the model is less biased toward terminated trials, improving its ability to generalize to active and completed trials with meaningful recruitment rates.

Lack of Phase-Wise Modelling

- **Limitation:** The model uses overall trial duration, ignoring distinct recruitment dynamics in Phase1, Phase2, and Phase3. This misses critical phase-specific trends, reducing predictive accuracy. **Incorporate Phase-Specific Features**
- Add **Phase1, Phase2,** and **Phase3** durations as independent features to capture unique recruitment patterns.
- Model interactions between phases and variables like **Enrollment** and **Funder Type** for phase-specific impacts.
- Use temporal models like **Temporal Fusion Transformer** to capture dependencies across phases.
- **Impact:** Phase-specific features improve predictive accuracy by capturing nuanced recruitment patterns.
- Enables better resource allocation and strategy optimization, enhancing trial efficiency and success rates.



Suboptimal Embeddings

- **Limitation:** While BioBERT embeddings offer moderate performance (F1 score 0.62) and effectively represent textual features, they fall short of larger LLMs like GPT-4 or LLaMA-3, which capture more nuanced patterns. Resource constraints make these advanced models inaccessible.
- **Advanced Embeddings:** Use advanced LLMs like GPT-4 or LLaMA-3 for richer embeddings that capture nuanced patterns.
- Leverage cloud-based GPUs to overcome resource limitations for efficient implementation.
- **Impact:** Advanced embeddings improve generalization and capture complex patterns, enhancing recruitment rate predictions and reliability.

Next Steps

Preprocessing Textual Columns Using OpenAI API

- **Using OpenAI API:**
- Extract structured entities like outcomes and timelines from the Primary and Secondary Outcome Measures column.
- Encode these entities as numerical or categorical features.
- Improve recruitment rate predictions by incorporating trial-specific characteristics.
- The cost of using it is very minimal, making it an ideal choice for our use case.
- **Challenges:**
- Moderate costs require budget planning for large datasets.
- Privacy concerns and reliance on external infrastructure add complexity.
- Expertise is needed for fine-tuning and data compliance.

Dynamic Feature Selection

- **Implementation:**
- Use methods like **Recursive Feature Elimination (RFE)** or mutual information to rank impactful features for recruitment rate predictions.
- Apply **reinforcement learning** (e.g., **PPO**) to adaptively select features based on trial phase or context.
- Enhance accuracy by prioritizing relevant features, reducing overfitting, and improving computational efficiency.
- **Challenges:**
- Resource constraints and complexity in setting up reinforcement learning hindered implementation.
- Lack of phase-specific labels and reliance on static methods limited feasibility.

Dynamic Equilibrium Strategy Optimization

- **Implementation:**
- Use **Markov Decision Processes (MDPs)** or **dynamic programming** to allocate resources like budget and trial sites efficiently.
- Integrate real-time data and define a reward function to balance recruitment success with resource efficiency.
- Adapt strategies dynamically during the trial lifecycle to improve recruitment rate predictions.
- **Challenges:**
- Unavailability of high-quality real-time data limited feasibility.
- Complexity of dynamic modeling and limited expertise in advanced optimization techniques posed barriers.

Fine-Tuning with Advanced LLMs

- **Implementation:** Fine-tune a 70B LLM like LLaMA-3.3 to extract contextual embeddings from Primary and Secondary Outcome Measures.
- Combine embeddings with structured features like Enrollment, Phase, and Study Design for a unified feature set.
- Use the enriched feature set with LightGBM to improve recruitment rate predictions.
- **Challenges:** Fine-tuning a 70B LLM requires extensive computational resources, including multi-GPU setups, which were unavailable.
- The column's size and diversity may not fully leverage the model, and time constraints further hindered adoption.

Game-Theoretic Framework [1]

- **Implementation:** Develop a game-theoretic framework using utility functions and payoff matrices to model interactions between patients, doctors, and research firms.
- Use BioBERT embeddings and LightGBM outputs as inputs, refining strategies with Nash equilibrium analysis.
- Apply backward induction in iterative simulations to optimize stakeholder strategies and align predictions with real-world behaviour.
- **Challenges:** Integrating machine learning with game theory required advanced expertise, beyond the project's scope.
- Iterative simulations were computationally intensive, exceeding infrastructure capabilities.
- Limited data on stakeholder interactions hindered realistic payoff model development.

REFERENCES

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!!