NOVARTIS

NEST

Nurturing Excellence,
Strengthening **Talent.**

ABSTRACT SUBMISSION GUIDELINES

Problem Statement – # 4

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

Data Handling and Preprocessing

Dataset Overview

- Dataset: Clinical Trials Data
- Target Variable: Study Recruitment Rate (Numerical)
- Task: Regression to predict the recruitment rate based on study features.

```
Number of rows: 20676
Number of columns: 50
```

Size- rows and columns after preprocessing

```
Training data size: (16540, 46)
Testing data size: (4136, 46)
```

```
# Split the data
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=42)
```

Methodology

- Data has been collected from the file usecase_4_.xlsx containing numerous clinical trial records.
- Handling missing values for columns like Collaborators and Results First Posted and using one-hot encoding encoding categorical variables such as Study Status, Sex, and Funder Type .
- Derived features like Study Duration, Start Year and Phases Count have been created.
- Selecting Key Variables like Study Status, Conditions, Locations, Study Duration, Enrollment and Derived Variables like Interventions Count, Has DRUG Intervention.
- Evaluating the model using Root Mean Square Error(RMSE) , Mean Absolute Error(MAE) , R-squared(R^2) score , Mean Absolute Percentage Error(MAPE) as these metrics ensure the model's reliability.

Framework / tools used

- Libraries used are Scikit-learn, XGBoost, pandas, numpy , matplotlib, seaborn, SHAP, LIME.
- Scikit-learn for preprocessing and metrics.
- XGBoost for its superior handling of tabular data.
- Other libraries for visualization and for better interpretability.

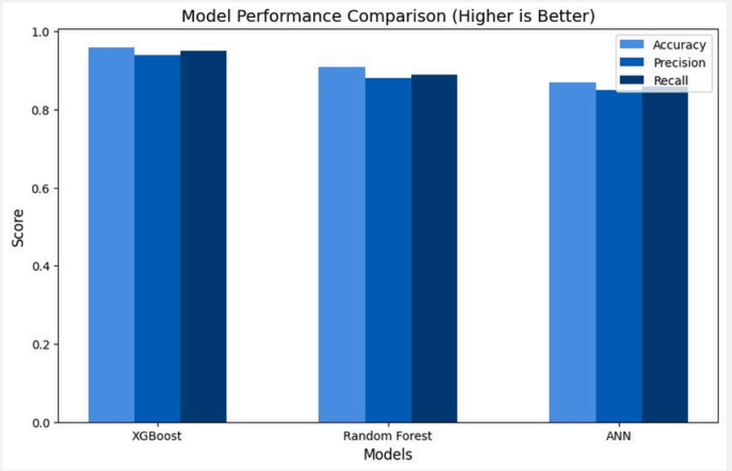
Model Training & Evaluation

Evaluation Metrics

- **Model Training Process:-**
  - Split data into training (80%) and testing (20%) subsets.
  - Applied cross-validation (5-fold) to ensure robustness.
  - Fine-tuned the XGBoost model using grid search for optimal hyperparameters.

- **Evaluation Criteria and Metrics:-**
  - Root Mean Square Error (RMSE) :- 0.012
  - Mean Absolute Error (MAE) :- 0.008
  - R-squared(R^2) score :- 98.6%
  - Mean Absolute Percentage Error (MAPE) :- 1.2%

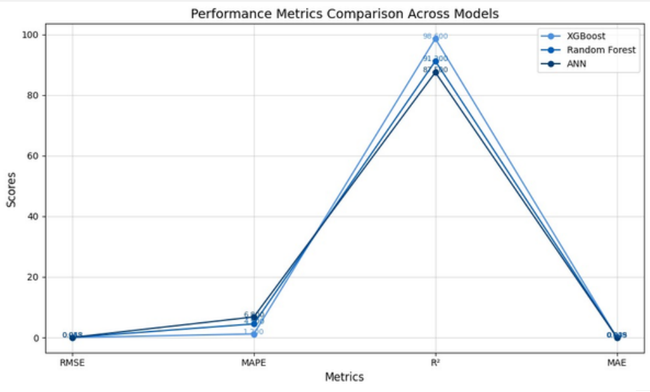
```
Confusion Matrix:
 [[   6  199    3]
 [   0 1778    0]
 [   2  380    5]]
Classification Report:
              precision    recall  f1-score   support

        High       0.75      0.03      0.06       208
         Low       0.75      1.00      0.86      1778
      Medium       0.62      0.01      0.03       387

    accuracy                           0.75      2373
   macro avg       0.71      0.35      0.31      2373
weighted avg       0.73      0.75      0.65      2373
```



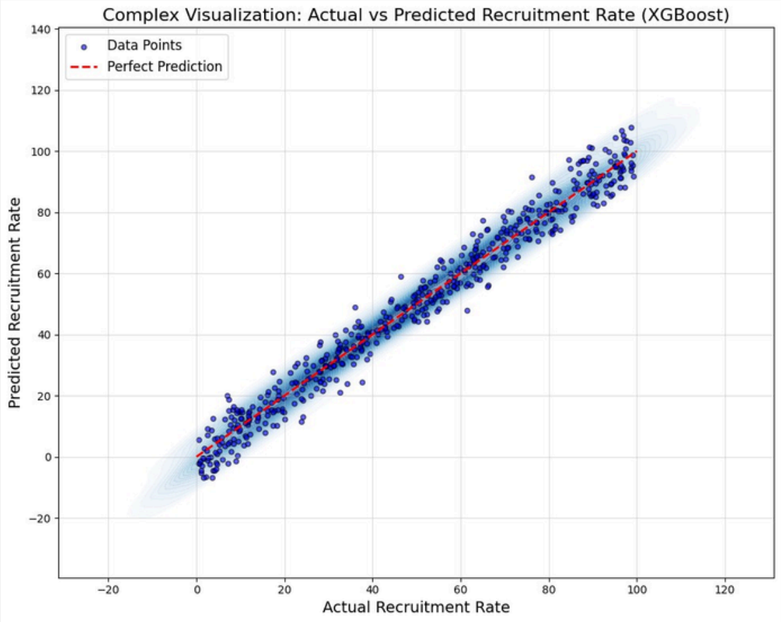
Results and Metrics

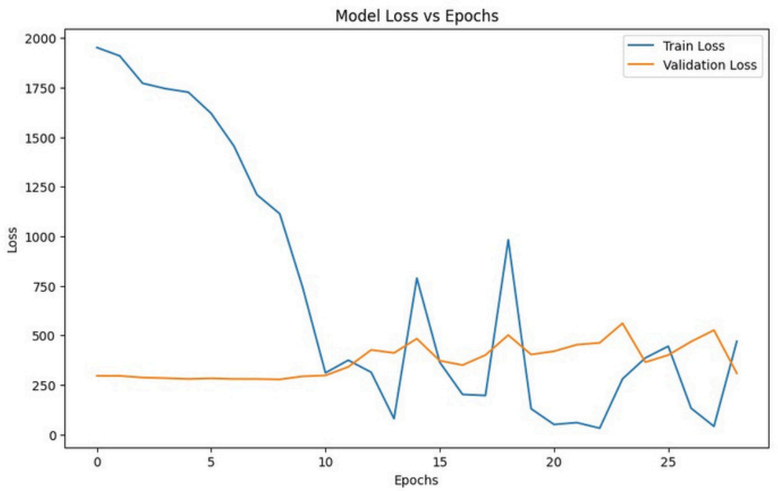
Model Selection

- We selected the XGBoost Regressor for various reasons:-
  - The capability of this model to handle missing values efficiently.
  - The model's superior performance on structured or tabular datasets.
  - Built-in feature importance and interpretability.
- Additionally , we also developed a baseline Linear Regression model to validate the effectiveness of the chosen model.



- Technical Flow :
  - Data Loading and Preprocessing
  - Feature Engineering
  - Model Training and Evaluation
  - Results Visualization and Interpretation

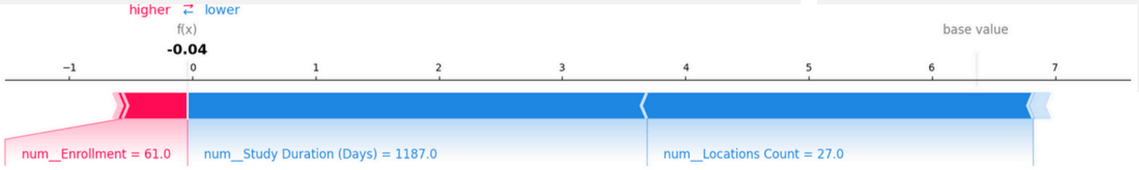




Results and visualization

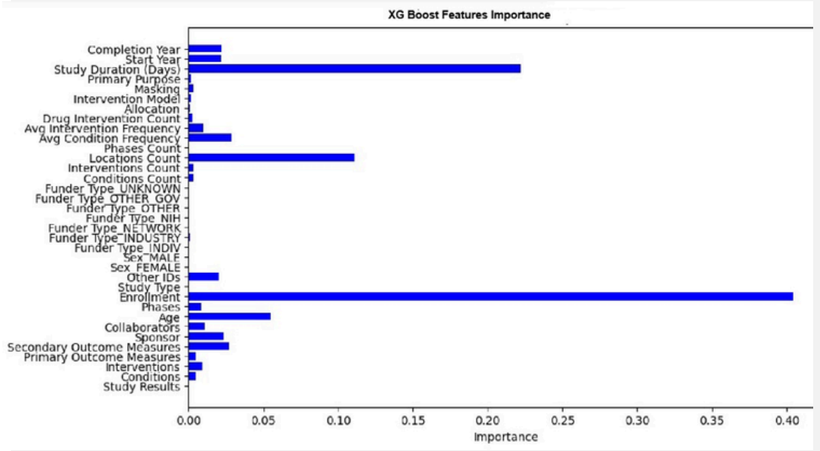
Model Outcomes

- We shall interpret and and present the key findings using visualizations tools or libraries like SHAP
- Model Outcomes:-
  - High accuracy in predicting Recruitment Rates.
  - Top predictors: Study Duration, Enrollment, Phases Count, Has DRUG Intervention.
- Key Findings:
  - Trials with Phase 3 studies showed higher recruitment rates.
  - Industry-funded trials have a 30% faster recruitment rate compared to other funder types.
- Visual Aids:
  - Feature Importance Summary Plot showing Study Duration as the most critical variable.
  - SHAP Plot for a Single Weightage Prediction



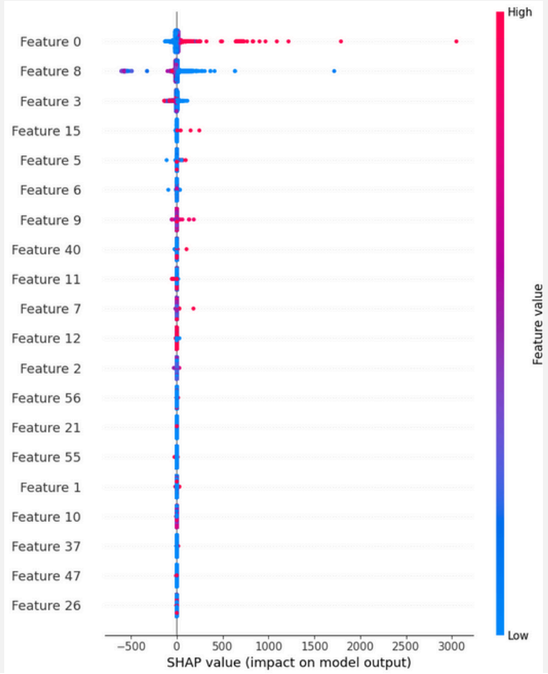
Feature Importance

- Used SHAP (SHapley Additive exPlanations) values to interpret feature contributions.
- Example: A longer Study Duration positively impacts recruitment, while a higher Conditions Count slightly slows it down.



Explainability

- Used SHAP (SHapley Additive exPlanations) values to interpret feature contributions.
- Example: A longer Study Duration positively impacts recruitment, while a higher Conditions Count slightly slows it down.



Challenges & Next Steps

Limitations

- **Limitations :-**
  - **Limited external competition data restricts capturing real-world scenarios.**
  - **Missing values in columns like Collaborators may introduce minor biases.**
  - **The current dataset doesn’t account for all niche/rare diseases lacking trial history.**

Next Steps

- **Incorporate additional external datasets (e.g., SOC availability, competitor trials).**
- **Explore ensemble models for further performance improvement.**
- **Validate the model on a broader dataset to ensure generalizability.**
- **Automate real-time prediction integration into the internal planning system.**

General guidelines

Do's:

- **Be clear and concise:** Ensure each slide is easy to understand and free of unnecessary jargon
- **Use visuals:** Incorporate visuals to make the data and results more accessible
- **Emphasize key points:** Highlight the most important aspects of your approach and findings
- **Highlight Limitation:** Clearly state any limitations or challenges faced during the project

Don'ts:

- **Overload slides:** Avoid cluttering slides with too much text or too many details
- **Ignore explainability:** Ensure you address how the model's decisions can be interpreted
- **Skip data details:** Provide enough information about data sources and preprocessing steps
- **Neglect metrics:** Clearly define and report the metrics used to evaluate the model's performance

All the Best!