NOVARTIS

NEST

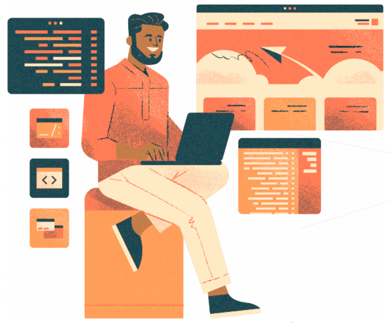
Nurturing Excellence,
Strengthening **Talent.**

Problem Statement – 4
Prediction of Study Recruitment Rate in Clinical Trials

NOVARTIS

Unique Value Proposition

1 Static Benchmarking Fails to adapt Real-time Changes



- Utilizing Active Learning to acquire data points for the model to adapt to real-time recruitment trends
- Allows the model to dynamically adjust to evolving trial conditions and minimize failures caused by traditional static benchmarks.

2 Traditional Methods Struggle with Niche & Rare Diseases

- BioBERT embeddings capture medical context, improving recruitment predictions for rare diseases with limited trial data.
- Enhances recruitment accuracy for niche diseases, overcoming data scarcity and historical biases.

3 Lack of Confidence & Explainability in AI-Driven Estimates

- Confidence intervals, SHAP analysis, and Bayesian uncertainty estimates ensure transparent and interpretable predictions.
- Builds trust in AI-driven recruitment forecasts, enabling data-driven decision-making with clear model insights.

4 Adaptive Data Preprocessing & Feature Engineering

- Identifies and prioritizes the most relevant factors influencing recruitment rates for better prediction accuracy..
- Corrects skewed distributions to ensure consistent and reliable forecasting, even in cases of extreme variations.

5 Ignoring External Factors Leads to Unrealistic Projections

- Integrating external datasets (funding, geography, locations) to provide context-aware and realistic recruitment predictions.
- This ensures recruitment forecasts reflect real-world conditions by incorporating external factors.

EXPLORATION AND RESEARCH

Selection of Features

- Research highlights - Duration of Trial, Enrollment Trends, Study Design, Patient Eligibility, and Primary Completion Date as major factors affecting recruitment rates.
- We used primary and secondary outcome measures as they assess trial success, effectiveness, side effects, and long-term impact of a drug.

NLP on Biomedical Text

- Research paper used: NLP-based techniques including **LLaMA**, **BERT**, **ClinicalBERT**, and **BioBERT**, etc.
- Our choice is BioBERT because It is trained on Wiki, Books, **PubMed**, and **PMC datasets**.
- Why not LLaMA or ClinicalBERT? Llama lacks biomedical specialization, while ClinicalBERT is limited to the **MIMIC-III database**.

Selection of Model

- Explored Models: **LightGBM**, **Neural Network**, **XGBoost**, **CatBoost**, and **GBM Regressor**.
- Best Performance: GBM Regressor achieved the best prediction results.
- Research Support: Prior studies also validate the effectiveness of GBM Regressor

Model Explainibility

- We used **SHAP** analysis to assign scores based on each feature's contribution to model predictions.
- The top three influential features - Duration of Trial, Enrollment, and Primary Completion Duration were identified during analysis.
- These findings strongly align with established research and widely accepted recruitment rate formulas.

NOVARTIS

TECHNICAL WORKFLOW

01 Data Input & Analysing

Took dataset and analyze the target variable—the recruitment rate—to understand its distribution and influencing factors

02 Feature Selection

Applied research-backed feature selection, focusing only on the most relevant columns, eliminating noise for better accuracy.

03 Data Preprocessing

Dropped highly incomplete columns.
Removed Potential Outliers.
Normalized numerical data (for consistency.

04 Textual Data Handling

We used **BioBERT**, a domain specific language model for medical data, to convert **text into embeddings**, making it machine-readable while retaining meaning

05 Feature Scaling

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

Model Training 06

Splitted into training and validation set using **Stratified sampling** .
We chose the **GBM Regressor**—a robust algorithm for efficiently handling structured data. for training.

Optimization 07

To maximize the performance, we fine-tuned it using **Bayesian Optimization**, which systematically adjusts hyperparameters for better accuracy

Predicting Recruitment Rate 08

After training the model it is used for predicting the **RR** for validation set

Model Evaluation 09

Evaluated our model's performance using **RMSE, MAE, and R^2 Score**, ensuring its predictions align closely with actual values

Deployment & Performance Monitoring 10

The trained model is deployed. To ensure it remains effective, we implement performance monitoring, tracking predictions over time.

Adaptability Under Constraints

Robust Data Handling

Our solution ensures reliable predictions even if there are fluctuations in our data or the input data is incomplete or noisy

Adaptive Learning

Instead of retraining on the entire dataset, the model selectively learns from the most valuable new data points, improving efficiency.

Time & Cost Efficiency

Our model optimizes computational resources by leveraging automation and parallelization, reducing processing time while minimizing infrastructure costs.

Scalability

The model scales accordingly whether deployed in high-performance cloud environments or constrained on-premise systems. It can run efficiently on lower-end GPUs or CPUs,

Future-Proof Design

The framework allows easy integration of new features, datasets without major change and can also be diversified to predict other factors also

KEY LIMITATIONS

Lack of Location & Sponsor Data biased Recruitment Rate Predictions

- **Prediction Bias:** Missing location and sponsor external data skews recruitment rate accuracy.
- **Biological Impact:** Just as genetics and environment shape disease prevalence, regional and financial factors affect trial participation.

How Early Terminations and Skewed Data Mislead Recruitment Predictions

- **Bias from Early Terminations:** Unexplained trial dropouts skewed recruitment rate predictions, leading to misleading insights. Excluding or down-weighting these trials could help reduce bias.
- **Addressing skewed data :** While log transformation helped smooth out imbalanced recruitment rates, additional techniques like Box-Cox transformation and synthetic data generation could improve accuracy even further.

Enhancing Recruitment Predictions with Advanced Embeddings

- **Sharper Insights with Advanced NER:** BioBERT captures basic text patterns, but LLMs with 30M+ parameters offer superior Named Entity Recognition (NER) for textual data. However, their deployment requires significant infrastructure.
- **Cloud-Powered, Yet Resource-Intensive:** Cloud GPUs enable advanced model use, but the high computational demands of LLMs remain a limitation in production.

Improving Phase-Wise Recruitment Rate Modelling

- **Phase-Specific Precision:** Our model already captures recruitment data separately for each phase, enhancing granularity.
- **Unlocking Deeper Trends:** Using advanced temporal models like TFT can reveal hidden phase-to-phase recruitment patterns for smarter predictions.

NOVARTIS

NEXT STEPS

Preprocessing Textual Columns Using OpenAI API

- Transforming Text into Data: Extracting trial outcomes and timelines with OpenAI API converts unstructured text into meaningful insights, enhancing recruitment predictions.
- Addressing Key Challenges: Scaling to large datasets requires budget planning, while privacy concerns, infrastructure reliance, and compliance expertise add complexity to implementation.

Optimising Recruitment Rate Predictions with Adaptive Feature Selection

- Smarter Feature Selection: Just as doctors prioritise key symptoms for diagnosis, RFE and reinforcement learning identify the most important features for better recruitment predictions.
- Adapting to Trial Phases: Adaptive feature selection improves accuracy, but resource constraints and missing phase-specific data limit full reinforcement learning implementation.

Implementing Dynamic Optimisation for Recruitment Strategies

- Optimised Resource Allocation: Use MDPs and dynamic programming to allocate budgets and trial sites efficiently, maximising recruitment success.
- Adaptive Strategy Refinement: Integrate real-time data for dynamic recruitment adjustments, but challenges like data limitations and complex modelling must be addressed.

Fine-Tuning LLMs for Recruitment Prediction

- Leveraging LLMs for Better Predictions: Fine-tuning LLaMA-3.3 (70B) can extract deeper insights from trial outcome measures, enriching structured features like Enrollment and Study Design.
- Addressing Computational Limits: Implementing this requires multi-GPU setups, but resource constraints and time limitations pose challenges to full adoption.

Game-Theoretic Optimisation: Enhancing Recruitment Strategies

- Strategic Decision Modelling: Using game theory, we can model how patients, doctors, and research firms interact, optimising recruitment strategies with utility functions and payoff matrices.
- Implementation Challenges: Applying Nash equilibrium and backward induction requires high computational resources and expert knowledge, while limited stakeholder data makes defining realistic payoffs difficult.

NOVARTIS

Execution Workflow for Production

01 Developing Modularized Code

Modularization and use of **Object Oriented Approach** (Custom Classes) will enable modular updates, scalability, reusability, and easier debugging, ensuring easy integration & deployment.

02 Pipelines-Based Workflow

Use an MLOps framework like **ZenML** to orchestrate and manage ML workflows for training and deployment, ensuring seamless integration with tools like **AWS**, **MLflow**, and Kubernetes.

03 Model Tracking & Versioning

Utilization of **MLflow Model Registry** will enable version control, staged deployments (Staging, Production), and audit logging, ensuring efficient model tracking, comparison, and reproducibility.

04 Model Deployment & API Hosting

The model will be deployed using **Flask**, exposing REST API endpoints for real-time predictions & health monitoring. It will be **containerized** with **Docker** for portability & deployed on **AWS** to for scalability and efficient model serving.

05 Automating Model Retraining & Deployment (CI/CD)

GitHub Actions to automate model training and deployment. **Apache Airflow** will be utilized for scheduled retraining, while **EvidentlyAI** will monitor data drift & model degradation, ensuring performance improvement.

06 PERFORMANCE MONITORING

AWS CloudWatch to track performance metrics, latency, and resource utilization. Implement model performance evaluation using **A/B testing** and feedback loops. Detect model drift and trigger retraining when necessary.

Thank You!!

Team-Satyamkmr22