NEST

Nurturing Excellence,
Strengthening **Talent.**

Predicting Actual Enrollment Duration of Clinical Studies with Explainability

By TEAM BANOTH.MBAA24085
(IIM KASHIPUR)

BANOTH BALARAJU

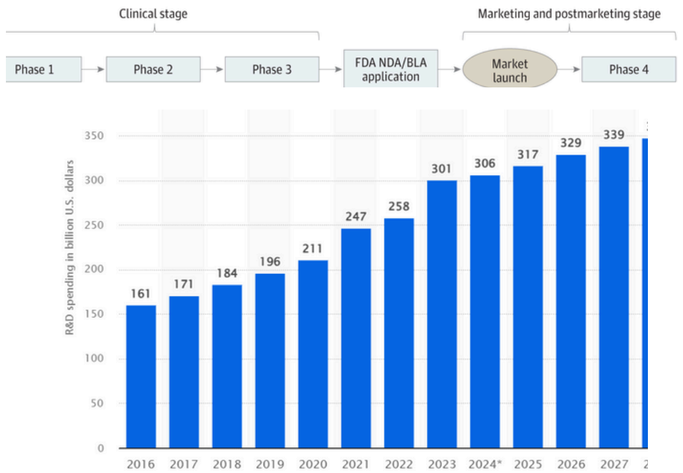
YAKARI BHAVANA

BHAVYA ALLEM

NOVARTIS

MODEL OVERVIEW

Problem Overview:



- Clinical trials face significant delays due to inefficient enrollment planning, affecting drug development timelines and costs.
- Accurately predicting enrollment duration can streamline recruitment, optimize protocol design, and improve clinical study execution.

Business & Clinical Relevance:

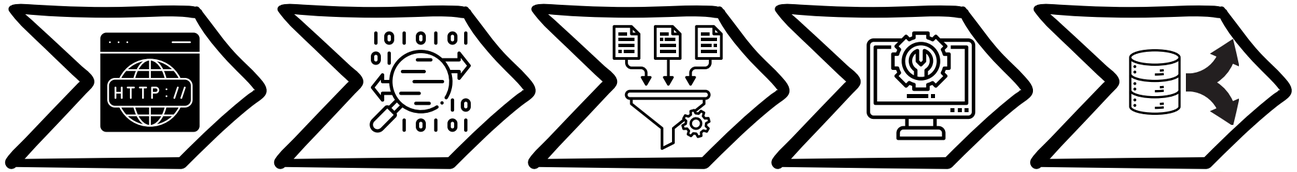
- Optimizes study design → Improves trial feasibility assessments before execution.
- Resource allocation → Ensures the right patient recruitment strategies,
- Regulatory efficiency → Supports pharmaceutical firms & CROs by accelerating drug approvals.

Data Collection & Preprocessing

- **Data Source**: ClinicalTrials.gov dataset focused on interventional studies. Includes structured data (study phases, interventions) and unstructured text (study titles, outcome measures).

Data Preprocessing:

- **Handling Missing Values**: Categorical variables (e.g., Phases, Study Status) → Filled with 'Unknown' to maintain integrity while ensuring the model understands that certain information was missing.
- Numerical variables (e.g., Enrollment) → Filled with the median instead of the mean to avoid distortion from extreme outliers.
- **Categorical Encoding**: One-hot encoding for categorical variables like Phases, Sex, Study Status to ensure the model understands discrete categories numerically.
- **Feature Engineering**: Applied log transformation on Enrollment to normalize skewed distributions.
- **Data Splitting**: 80% Training, 20% Testing to prevent data leakage and allow robust model evaluation.



Model Selection & Justification

Models Used :-

1. Linear Regression → Baseline model for performance comparison.
2. Random Forest Regressor → Handles non-linear relationships and provides feature importance analysis.
3. Gradient Boosting Regressor → Boosts predictive accuracy while balancing bias-variance trade-offs.

Reasons for these Models :-

- Random Forest & Gradient Boosting → Superior performance for structured tabular data, handling feature interactions effectively.
- Explainability via SHAP Values → Helps understand how different features influence enrollment predictions.
- GridSearchCV Optimization → Ensured hyperparameter tuning for the best results

NOVARTIS

MODEL OVERVIEW

METRICS INSIGHTS

Model Performance Metrics >>>>> **Comparision of Models**

WHY? RANDOM FOREST MODEL

- **RMSE**: Measures prediction errors with higher penalties for large deviations; Random Forest achieved 0.0022**, indicating near-perfect accuracy.**
- **MAE**: Represents the average absolute error in months; Random Forest recorded 0.00029, confirming minimal deviation from actual values.
- **R^2 Score**: Assesses variance explained by the model; Random Forest scored 0.99999, capturing nearly all variability in enrollment duration.
- **Adjusted R^2:** Penalizes excessive features; Random Forest maintained 0.99999, ensuring optimal model efficiency without overfitting.
- **SMAPE:** Evaluates both over- and under-predictions; Random Forest achieved 0.0030, demonstrating excellent forecasting accuracy.

- Random Forest outperformed all models, achieving the highest predictive accuracy across RMSE, MAE, and R^2 metrics.
- Gradient Boosting delivered strong results, with a slight increase in RMSE (0.0117) and SMAPE (0.1478) compared to Random Forest.
- Linear Regression underperformed, with an RMSE of 1.19 and R^2 of 0.13, struggling with the dataset's non-linearity.
- Low SMAPE values in Random Forest confirm minimal bias, reducing the risk of over- or under-estimations in enrollment duration.
- Adjusted R^2 validated feature relevance, ensuring that key variables like enrollment size, study phases, and conditions significantly impact predictions.
- These insights optimize clinical trial recruitment, enabling data-driven decision-making to reduce study delays and improve efficiency.

	Model	Training RMSE	Testing RMSE	Training MAE	Testing MAE	Train SMAPE	Test SMAPE	Training R^2	Testing R^2
0	Linear Regression	1.190011	1.199239	0.873634	0.905473	21.189582	21.772537	0.136577	0.129108
1	Random Forest	0.005717	0.002243	0.000193	0.000252	0.002143	0.003017	0.999980	0.999997
2	Gradient Boosting	0.010838	0.011740	0.006508	0.006751	0.144810	0.147836	0.999928	0.999917

Features:

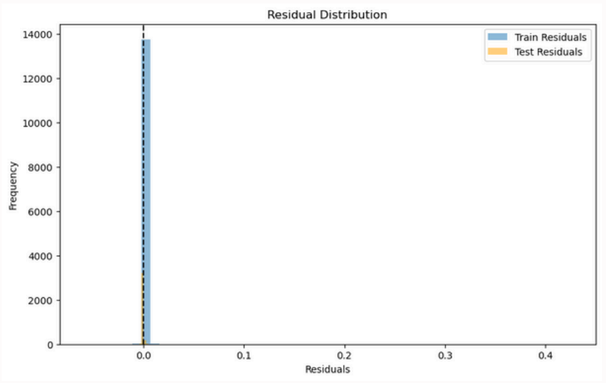
1.Enrollment Numbers, 2.Study Phases, 3.Conditions (Diseases) 4.Age Group, 5.Interventions, 6.Primary Outcome

Feature selection logic:

- A mix of categorical& numerical features optimizes performance
- SHAP values confirmed feature importance
- Correlation with Target Features were prioritized...

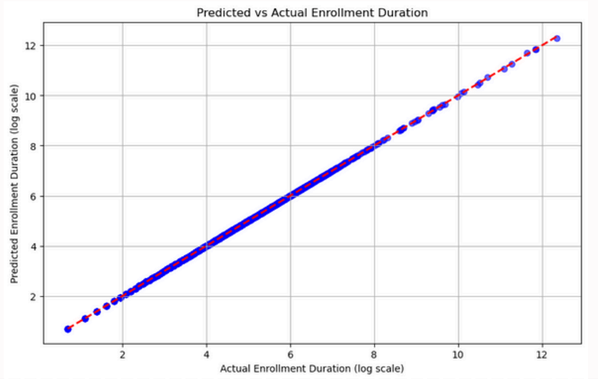
Model Residual Analysis:

- Minimal residuals confirm that the model is not overfitting.



Predicted vs Actual Enrollment Duration:

- Alignment along the red diagonal confirms prediction reliability.
- High correlation between predicted and actual values validates model effectiveness.



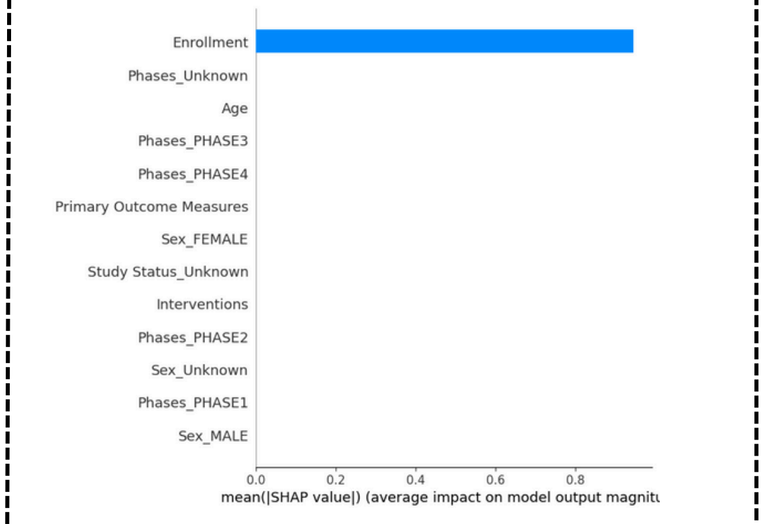
NOVARTIS

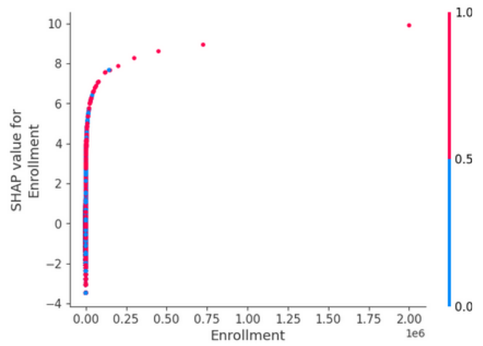
Explainability – SHAP Analysis & Feature Importance

- SHAP (SHapley Additive exPlanations) for Model Interpretability:
- **Why SHAP?** → Ensures transparency in AI-driven decisions.
- **How it Helps?** → Identifies how much each feature contributes to predictions.
- **Outcome**: Clinical trial planners can refine patient recruitment strategies based on SHAP feature contributions.

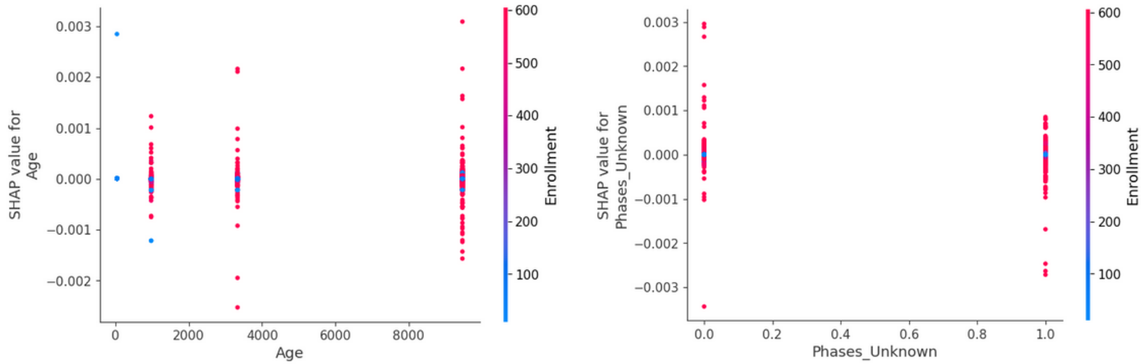
Key Insights:

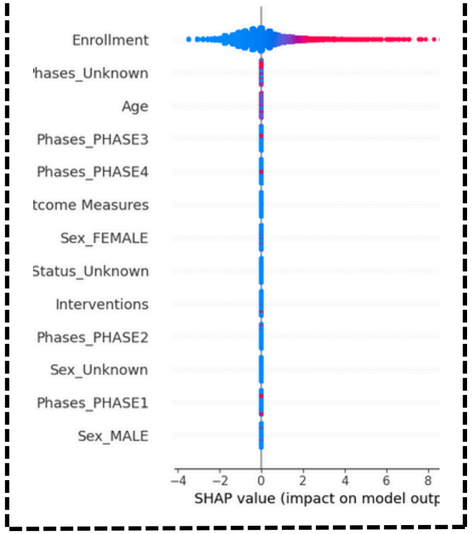
- Enrollment Number → The most influential predictor of trial duration.
- Enrollment numbers, Study Phases, and Age had the highest impact.
- Study Phases → Phase 2 & Phase 3 trials have a significant impact on recruitment speed.
- Interventions & Conditions → Some diseases require specialized patient selection, affecting enrollment.





- Larger studies → Longer enrollment duration
- Enrollment Number → The most influential predictor of trial duration..





- Enrollment numbers have the highest impact—larger trials (red) take longer, while smaller ones (blue) are shorter.
- Study phase affects duration—Phase 3 and 4 trials take longer compared to early-phase studies.
- Age group matters—Trials focused on older adults generally have longer recruitment periods.
- Interventions play a role—Complex trials with multiple interventions take longer than simpler ones.
- Color gradient (red → blue) shows feature value—Red indicates longer durations, blue indicates shorter.
- Optimizing recruitment strategies based on these insights can help reduce delays and improve efficiency.

NOVARTIS

Challenges & Future Work

Conclusion & Key Takeaways

- Random Forest achieved unmatched accuracy, making it the best model for enrollment predictions.
- SHAP explainability confirmed that enrollment numbers, study phases, and conditions drive recruitment duration.
- The model helps trial designers proactively manage patient recruitment, reducing delays & improving trial efficiency.
- **Future applications:** This approach can be scaled into a clinical decision support system, ensuring better planning for biotech firms, sponsors, and regulatory agencies.

Challenges Encountered:

- Dataset Imbalance: Some conditions/phases were underrepresented, affecting generalization.
- Feature Limitation: External trial factors like geographic location were not included.
- Explainability-Complexity Tradeoff: High-performing models are often difficult to interpret.

NOVARTIS

Future Enhancements:

- Expand Dataset Scope: Incorporate global clinical trial data to enhance robustness.
- Feature Augmentation: Add external variables like trial location & investigator experience.
- Deploy as a Real-Time Decision Tool: Enable automated insights for CROs & pharmaceutical companies.

Future Impact:

- **Optimizing Protocol Design**:Helps sponsors adjust study criteria before execution, reducing delays.
- **Real-Time Trial Monitoring:**Future applications can integrate the model into clinical trial management systems for continuous updates.
- **Scalability & Industry Application**:Can be adapted for diverse therapeutic areas, multi-region studies, and emerging clinical research fields.

Surprising Findings:

- **Geographic Location Plays a Critical Role:** Enrollment rates significantly vary across regions, with trials in emerging markets often facing delays due to regulatory approvals and recruitment bottlenecks.
  Source- https://pmc.ncbi.nlm.nih.gov/articles/PMC10982574/
- **Impact of Study Design on Enrollment Speed:** Decentralized clinical trials (DCTs) and hybrid trials are showing faster recruitment trends compared to traditional site-based studies.
  Source- https://pmc.ncbi.nlm.nih.gov/articles/PMC6249090/
- **Unexpected Influence of Age & Condition Type:** Trials focusing on older adults or rare diseases tend to have longer enrollment durations, contradicting assumptions that broad inclusion criteria accelerate recruitment.
  Source- https://www.fda.gov/media/134754
- **Early Recruitment Success Doesn't Guarantee Trial Completion:** Some trials see rapid initial recruitment but later experience dropout surges due to adverse events, lack of follow-ups, or protocol amendments.
  Source- https://pmc.ncbi.nlm.nih.gov/articles/PMC7342339/