NOVARTIS

NEST

Nurturing Excellence,
Strengthening **Talent.**

Problem Statement – 3

Predicting Completion of clinical studies with explainability

Approach & methodology

Overview

- **The Challenge We're Solving:**
  - Predicting if a clinical trial will reach completion
  - Think of it like forecasting a journey's success before it begins
  - Real impact: Saving time, resources, and potentially lives
- **Vision:**
  - Create a solution that not only predicts but explains
  - Make complex trial data speak in a language everyone understands
  - Bridge the gap between data science and clinical expertise

Methodology

- **Our Data Journey:**
  - Start with raw data in parquet format
  - Clean and transform data
  - Do Feature engineering (seeing patterns other might miss)
  - Learn from historical patterns(what worked ,what didn't)
- **Features and Insights:**
  - Trial Core Info: → Status→ Phase, Enrollment, Study Type → Timeline metrics (Start to Completion)
  - Location Intelligence: → Geographic distribution → Site-specific patterns
  - Timeline Features: → Study duration calculations → Phase-specific completion times
  - Risk Indicators: → Adverse event severity index → Subject withdrawal patterns
- **Key Metrics That Matter:**
  - Primary Metrics: → Precision → Recall → F1 → AUC-ROC
  - Secondary Analysis: → Class-specific performance metrics → Cross-validation stability scores → Confusion matrix analysis
  - Phase-specific performance ,Geographic success patterns (**ADDITIONAL**)

Framework / tools used

- **Frameworks:**
  - Scikit-learn: Traditional ML pipeline
  - LightGBM/XGBoost: Advanced modeling
  - HuggingFace Transformers: Text processing(BERT)
  - SHAP: Model interpretation
  - Feature-wiz: Automated feature selection
- **Why these choices matter:**
  - Handle large-scale clinical data efficiently
  - Proven reliability in healthcare
  - Strong support for model explainability
  - Makes our predictions trustworthy

Model choice & setup

Model Selection

•**ML Models (Structured Data)**:
•**Logistic Regression**: Baseline for comparision and interpretability.

•**Random Forest**: Captures non-linear patterns and ranks feature importance.

•**LightGBM/XGBoost**: Efficiently handles large-scale structured data with advanced feature interactions.

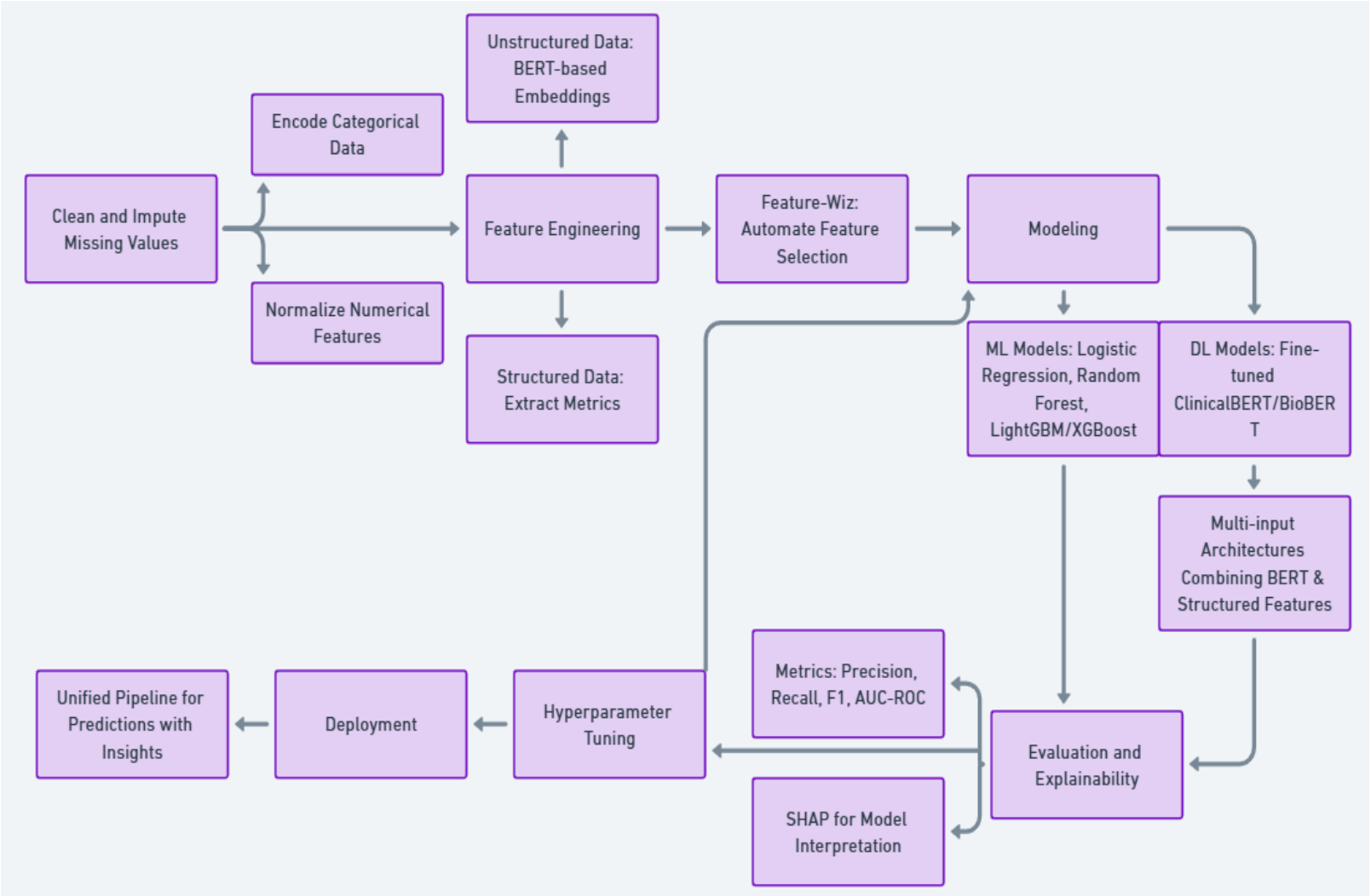
•**GenAI Models (Unstructured Data)**:

•**ClinicalBERT/BioBERT**: Extracts domain-specific insights from clinical text for better contextual understanding.

•**Feature-Wiz**: Automates feature selection to optimize inputs.

This balanced approach addresses the complexity of clinical trial data while ensuring performance, scalability, and interpretability.

Model Architecture



Model Training & Evaluation

Evaluation Metrics

- **Model Training Process**

•**Data Splitting**: Train (70%), Validation (20%), Test (10%).
•**ML Models**: Train Logistic Regression, Random Forest, LightGBM, XGBoost for structured data.
•**DL/GenAI Models**: Fine-tune ClinicalBERT/BioBERT for unstructured data; combine with structured data using multi-input architectures.
•**Optimization**: Hyperparameter tuning for best performance.

- **Evaluation Criteria and Metrics**

•**Primary Metrics**:
•**Precision**: Measures the percentage of true positive predictions among all positive predictions.
•**Recall**: Captures sensitivity by identifying how well the model detects true positives.
•**F1-Score**: Balances precision and recall for overall classification performance.
•**AUC-ROC**: Evaluates the model's ability to distinguish between classes across thresholds.

•**Secondary Metrics**:
•**Class-Specific Analysis**: Precision, Recall, F1 for each class.
•**Cross-Validation Stability**: Ensures consistency across folds.
•**Confusion Matrix**: Highlights misclassifications.

•**Numerical metrics:**
•**Root Mean Square Error (RMSE)**: Quantifies the average prediction error magnitude.
•**Mean Absolute Error (MAE)**: Captures average absolute differences between predicted and true values.
•**R-squared (R^2)**: Indicates the proportion of variance explained by the model.

Results and visualization

Model Outcomes

- **How will we interpret and present the results :**
  - We'll tell the story of each model's journey in predicting trial completion
  - Clear metrics that matter: Focus on what makes a prediction truly useful for clinical researchers
  - Deep dive into understanding why certain predictions work better than others - because numbers alone don't tell the whole story
  - Real-world meaning: Translate complex model outputs into actionable insights

- **Main outcomes**
  - Beyond just numbers: How better predictions could transform trial planning
  - Risk identification: Finding early warning signs
  - Success patterns: Uncovering what really drives trial completion
  - Practical impact: How these insights could help more trials reach completion successfully

- **How we will show results ?**
  - Eye-catching visuals that make complex data easy to grasp
  - Feature importance plots to highlight key predictive factors
  - Model performance comparison charts showing relative strengths across different metrics
  - ROC curves to demonstrate model discrimination capabilities
  - Interactive confusion matrix for detailed performance analysis

Explainability

We plan to implement a multi-layered explainability approach that makes our predictions transparent and trustworthy:

- **Model Explainability Strategy**
  - Start simple - use built-in feature importance from our machine learning models
  - Level up with SHAP (SHapley Additive exPlanations) analysis - this will be our main tool to understand complex model decisions
  - Compare SHAP insights with our initial EDA findings to validate our understanding
  - Feature Importance Analysis: Beyond basic rankings, we'll dive into how different features interact to influence trial outcomes
- **Why this matters ?**
  - Our approach will focus on answering the "why" behind each prediction
  - We'll create intuitive explanations that translate complex model decisions into clear
  - Builds trust in our predictions through transparency
  - Visual interpretation tools will show how different trial characteristics combine to influence the final prediction
  - We'll provide confidence scores alongside explanations, helping stakeholders understand when to rely more heavily on model predictions

Challenges & Next Steps

Limitations

- **Limitations We'll Face (Keeping It Real!):**
  - Our large 3GB dataset means we'll need smart handling (think parquet format!) to work efficiently
  - Clinical trials' diverse nature means some rare conditions might be underrepresented
  - While SHAP helps explain predictions, some complex feature interactions might remain challenging to interpret
  - Sometimes the model might see patterns we need to double-check with medical experts
  - Like learning any new skill, finding the right balance between accuracy and understanding takes time

Next Steps

- **Future Steps and Vision:**
  - Making Our Models Smarter
  - Explore integration of BioBERT and ClinicalBERT for deeper text understanding
  - Investigate advanced feature selection using Feature-Wiz recommendations
  - Create interactive dashboards for stakeholder engagement
  - Turn complex predictions into straightforward recommendations
  - Real-World Impact We Can Create

Thank You !

NOVARTIS | Reimagining Medicine