NOVARTIS

NEST

Nurturing Excellence,
Strengthening **Talent.**

TEAM ECORP

IIT KANPUR

Divyansh Mittal
Khush Gandhi
Utkarsh Kumar

Problem Statement – # 2

PREDICTING ACTUAL ENROLLMENT DURATION OF CLINICAL STUDIES WITH EXPLAINABILITY

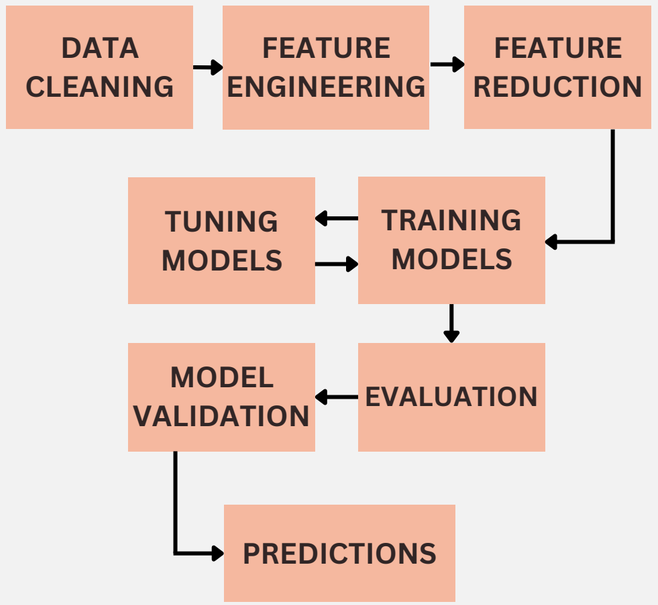
Approach & methodology

Overview

OBJECTIVE

Build an ML-pipeline to predict enrollment duration of clinical trials using structured and unstructured dataset on clinical trials

APPROACH



Methodology

DATA COLLECTION

use_case_2_.csv
withdrawals.txt
eligibility.txt
facilities.txt

NUMERICAL AND CATEGORICAL FEATURES

- Handled erroneous values
- Aggregated statistical & mode values (locations, withdrawals)
- Removed redundant correlated features

TEXTUAL FEATURES

- Handled erroneous values
- Generated word embeddings (12 text columns)
- Enriched embeddings using autoencoders.
- Extracted text insights with TF-IDF.
- Reduced dimensionality of features.

DERIVED FEATURES

-Inclusion/exclusion criteria
-Unique clinical centers: city, state, country
-Frequency bins (locations, periods, reasons, counts of withdrawals)

FEATURE SELECTION

Study summary	Location frequency bin
Outcomes	Study design
Criteria	Enrollment numbers
Condition	Interventions etc

Framework / tools used

NumPy	MatplotLib
Pandas	XgBoost
Pytorch	Shap
ScikitLearn	

NOVARTIS | Reimagining Medicine

Model choice & setup

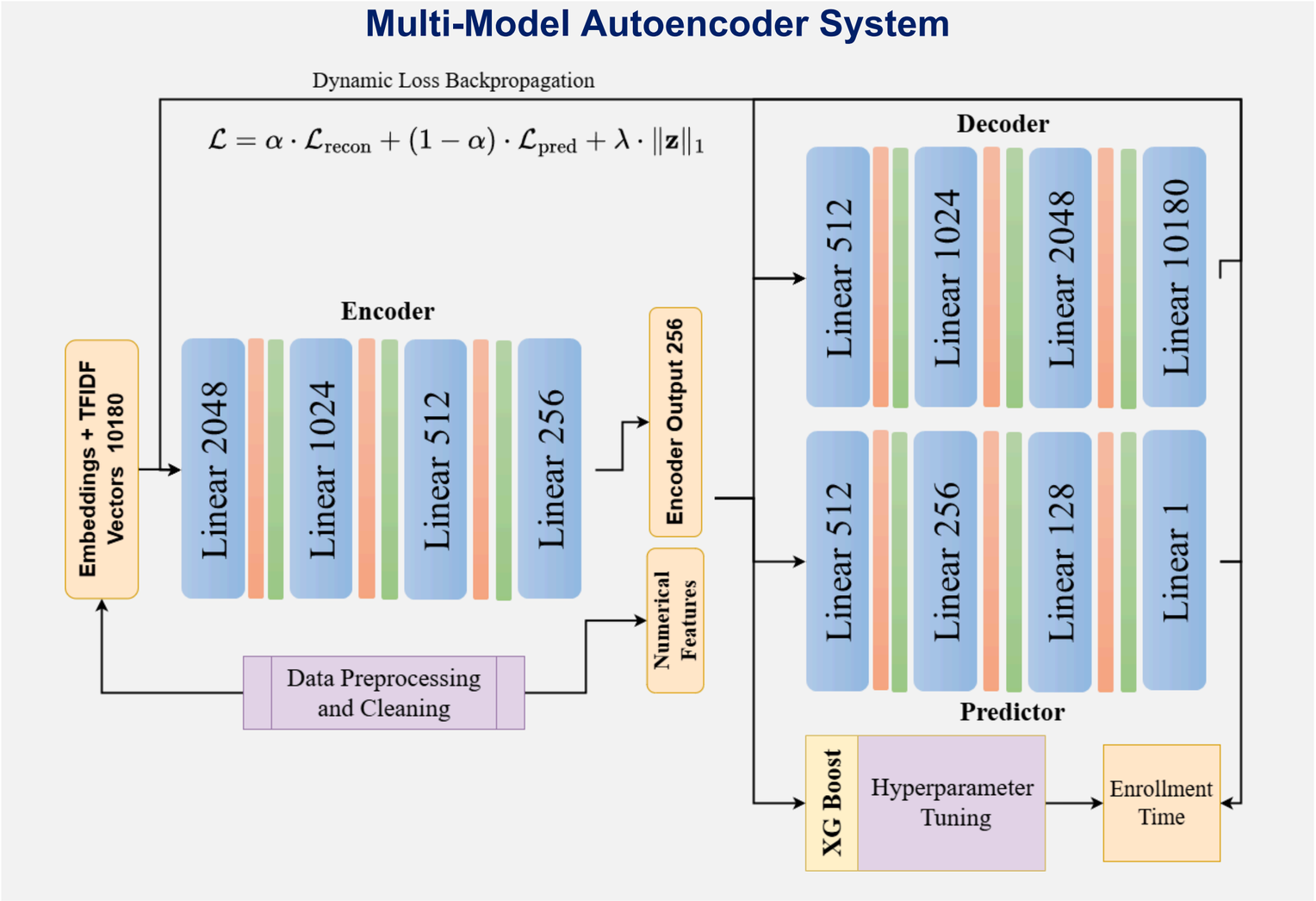
Model Selection

THE NEED

- Handling **High-dimensional text features** (~10,000)
- Handling **Mixed data types** (text and numerical)
- **Preserving predictive information** during feature reduction

KEY ADVANTAGES

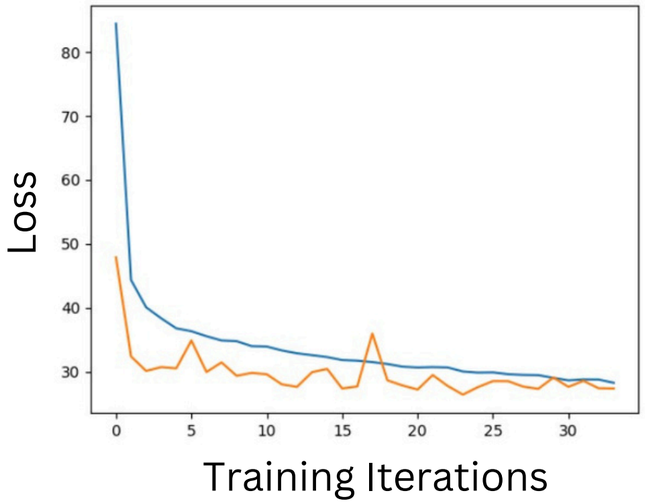
- **Learned compression** vs. statistical methods (e.g., PCA)
- Automated Task-specific **feature selection**
- **Reconstruction loss** ensures meaningful compression
- **L1 regularization for sparse representations**
- Dynamic loss weighting balances objectives
- Better **new data handling with latent space**
- **Transferable Approach**



Model Training & Evaluation

Evaluation Metrics

Models	Train			Testing		
	RMSE	R2 Score	Adj. R2	RMSE	R2 Score	Adj. R2
XGBoost	10.3523	0.5903	0.5880	**12.5457**	**0.4120**	**0.3988**
Neural Network	**5.812**	**0.8901**	**0.8872**	12.6918	0.3990	0.3891
Predictor (Baseline)	11.5962	0.5015	0.4987	12.6559	0.4024	0.3900



ANN Training Process

- Dynamic Loss and Learning Rate using Lr_scheduler
- Early Stopping (min-delta : 1e-4)
- Metrics used for validation : RMSE and R2 Score
- Batch size : 32 and epoch count : 50
- **weight decay** and **gradient clipping** to avoid overfitting

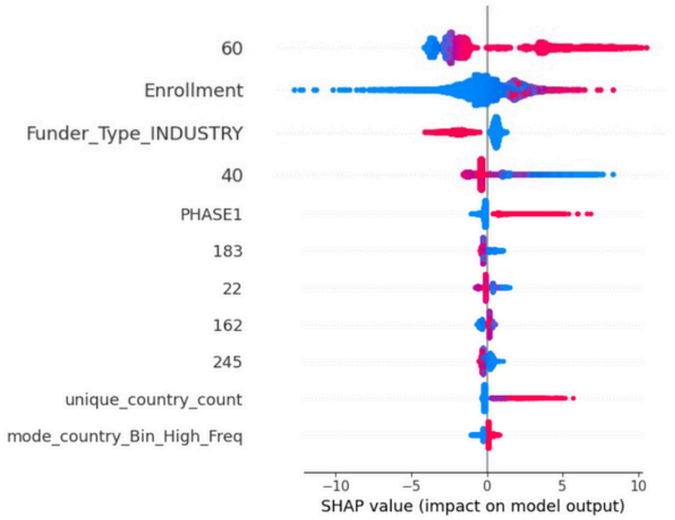
XGBoost Training Process

- Hyperparameter Tuning resulted **28% increase** in R2 Score
- K-fold cross validation and with Optuna
- **max-depth : 6 | learning rate : 0.0082 | gamma: 1.5**
- **n_estimators : 2440 | child_weight : 5 | subsample : 0.87**

Results and visualization

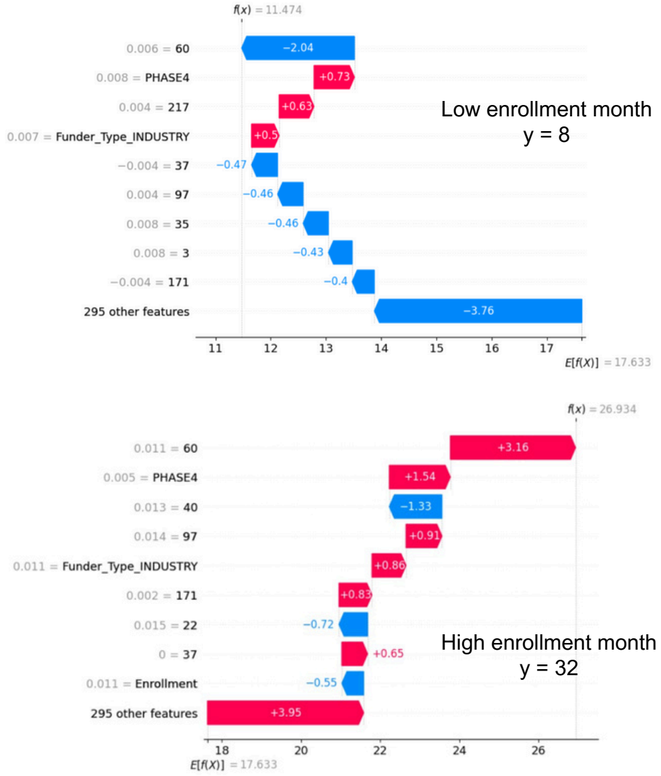
Explainability and Model Outcomes

Feature Beeswarm Plot



Numbers represent the derived 256 textual features

Waterfall Plots



Key Findings

- Predictions are primarily influenced by textual data, along with some numerical and categorical features
- Key interpretations:
  - Higher enrollment numbers correspond to longer enrollment durations
  - Non-industry funders are associated with shorter enrollment durations
  - Phase 1 trials tend to take more months to complete
  - Trials conducted across multiple countries result in longer enrollment periods

Challenges & Next Steps

Practical Applications

Model Deployment:

- Reduced feature space for efficient storage
- Faster inference with compressed representations
- Interpretable latent space

Business Impact:

- Better enrollment time predictions
- Resource allocation optimization
- Trial planning improvements

Limitations

- R2 Score plateaus around 0.42
- Model susceptible to overfitting
- Text embedding vectors demands high computation

Future Enhancement

Feature Engineering:

- Additional text preprocessing features
- Domain-specific text filters to reduce computational cost
- Feature importance analysis

Training:

- Ensemble approaches for multi-model systems
- Active and Curriculum Learning

Thank you!