Problem Statement – 3

PREDICTING COMPLETION OF CLINICAL STUDIES WITH EXPLAINABILITY

TEAM NAME:
ACHAUHAN_BE22
TEAM MEMBERS:
ANANYA CHAUHAN
MEDHANSH JAIN

PROBLEM : Statement

Objective:

- *Predict the clinical study completion status ("**Completed**" or "**Not Completed**," including "**Suspended**," "**Terminated**," and "**Withdrawn**").*

Significance:

- *Helps in improving **resource allocation**, **trial predictability**, and **overall success rates** in the clinical trial process.*
- *Provides insights for **better decision-making** in clinical trial designs.*

Challenges

Class Imbalance in Trial Outcomes

Minority outcomes like "Suspended" and "Withdrawn" are harder to predict, leading to biased models.

Handling Unstructured Data

Transforming text features like study titles and summaries into meaningful embeddings is complex.

Ensuring Explainability

Providing clear, interpretable insights for predictions is crucial to gain stakeholder trust.

Integration of Diverse Features

Combining structured and unstructured data effectively for accurate predictions is challenging.

KEY QUESTIONS ADDRESSED

1. ***What factors contribute most to the completion status of clinical trials?***
2. ***How can we manage high-dimensional data to ensure meaningful predictions?***
3. ***How can explainability improve trust and understanding of predictions?***

DATASETS USED

- ***Source***: Publicly available clinical trial data from clinicaltrials.gov.
- ***Features:***
  - **Structured:** Study design, criteria, etc.
  - **Unstructured:** Study title, brief summary, etc.
- ***Additional Files to Join:*** drop_withdrawals.txt, eligibilities.txt, facilities.txt, and reported_events.txt using **nct_id** as the key column.

NOVARTIS | Reimagining Medicine

Modeling and Optimization

BASELINE MODELS:

Logistic Regression and Random Forest for initial assessments.

DEEP LEARNING:

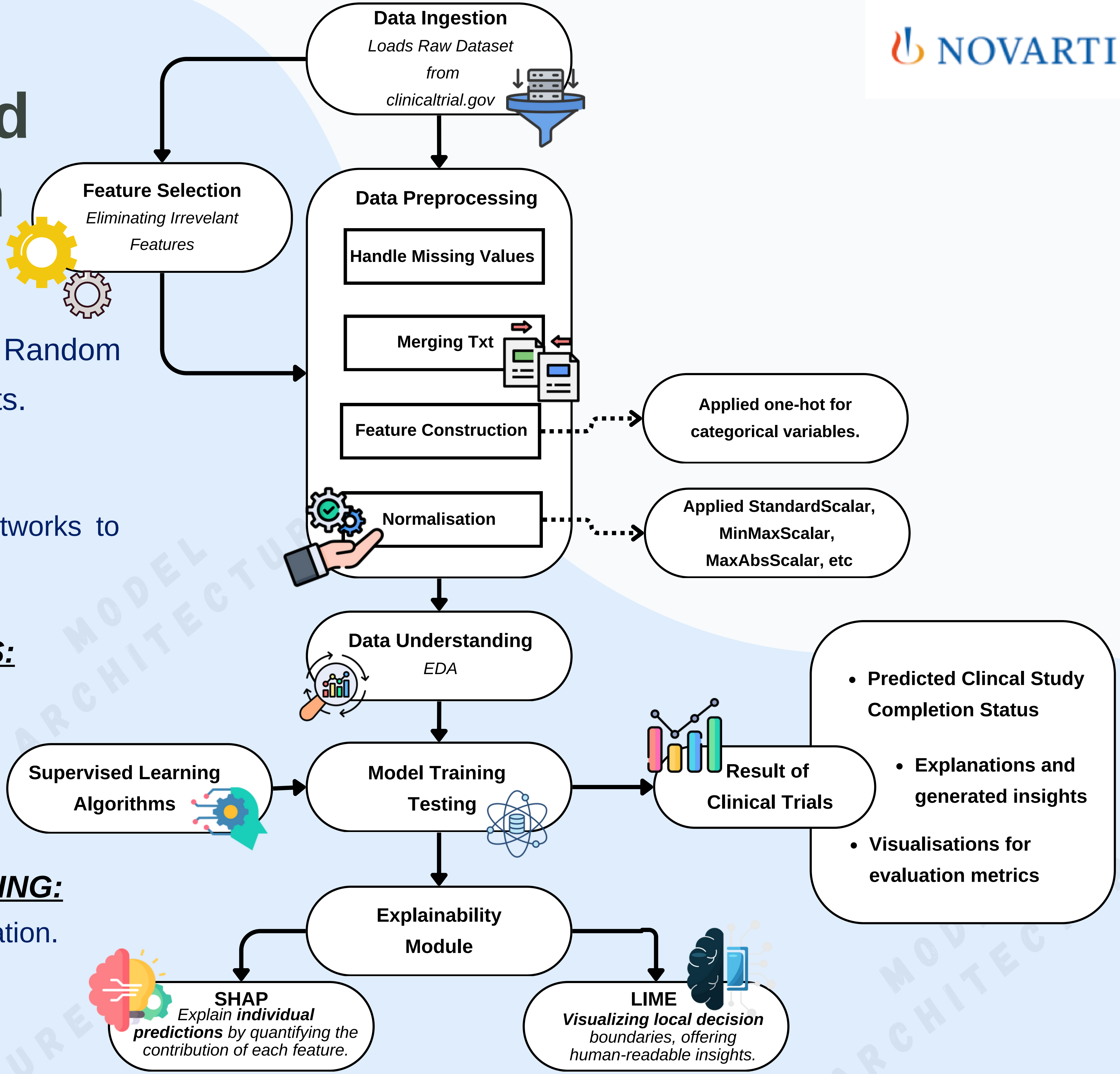
TensorFlow-based Neural Networks to capture complex relationships.

ADVANCED TECHNIQUES:

Gradient Boosting models (***XGBoost, LightGBM, CatBoost***) for enhanced accuracy.

HYPER PARAMETER TUNING:

Leveraged ***optuna*** for optimization.



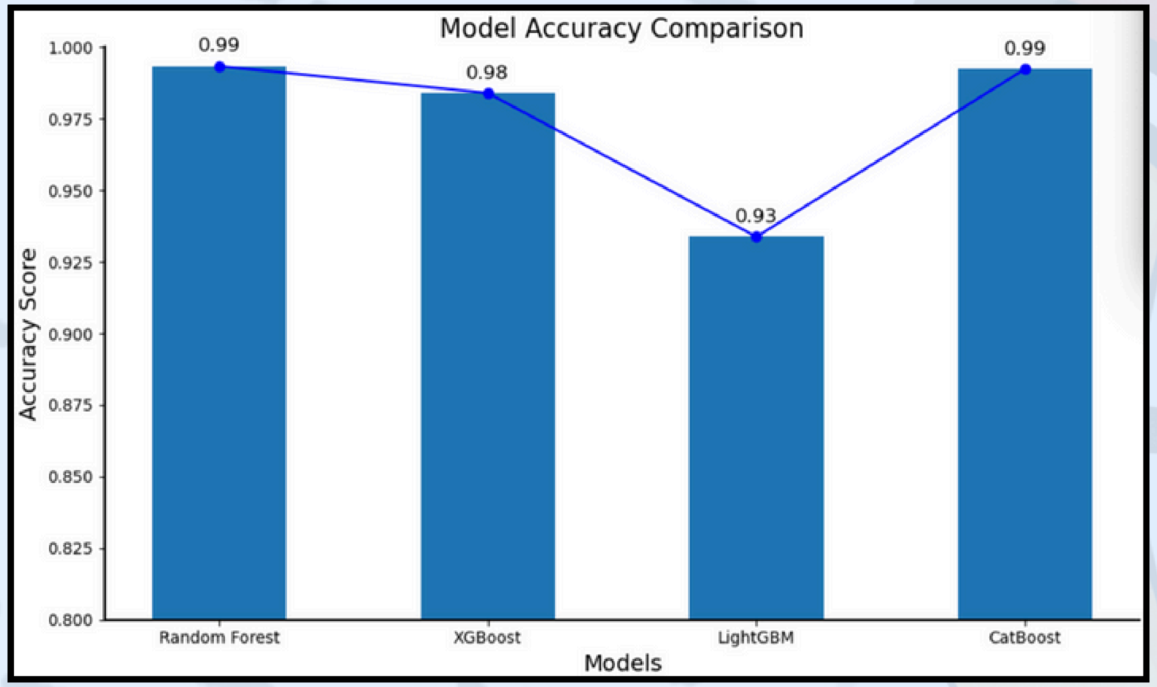
FrameWorks and Libraries Used

Results and Visualisations

Model	MAE	MSE	RMSE	R2 Score	Variance Score
CatBoost	0.0153	0.0306	0.1751	0.9408	0.9408
LightGBM	0.1325	0.2650	0.5148	0.4881	0.4881
XgBoost	0.0321	0.0642	0.2534	0.8759	0.8763

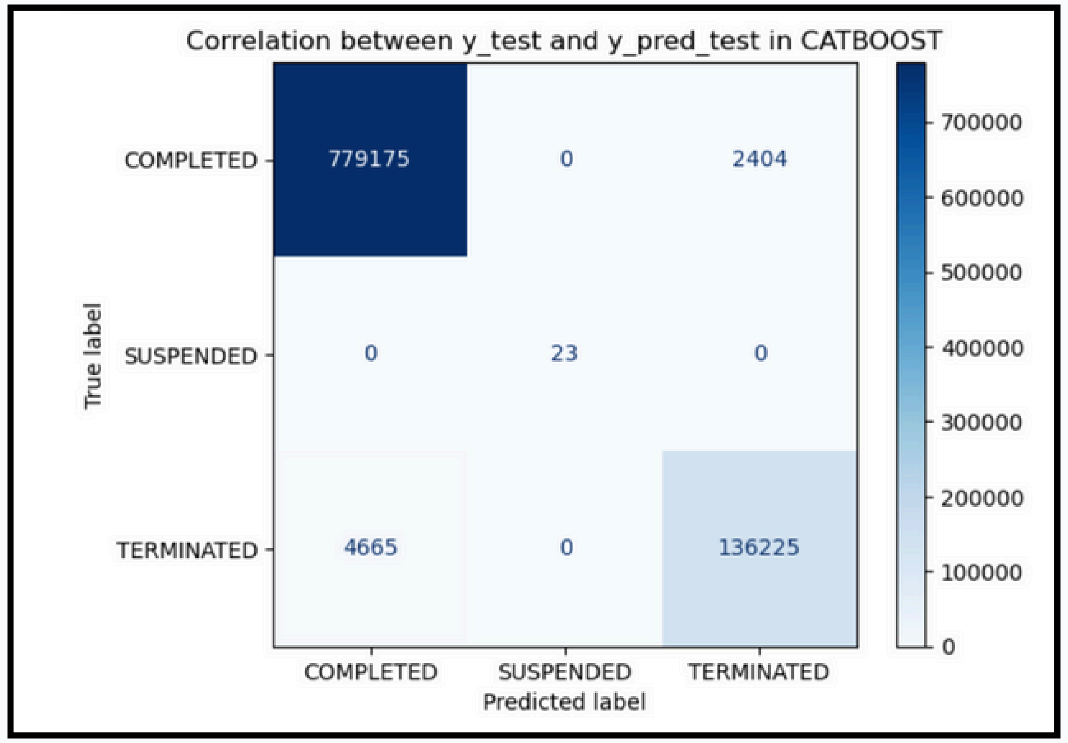
Model	Accuracy (%)
CatBoost	99.23
LightGBM	93.37
XgBoost	98.39
Random Forest	99.33

- CatBoost achieved the ***lowest error rates (MAE, MSE, and RMSE)***, showcasing its ability to make ***precise predictions***.
- ***High R^2 score*** indicates that the model effectively explains the variance in the dataset, making it a ***strong contender for regression tasks***.



- ***Random Forest and CatBoost*** demonstrated the ***highest accuracy*** scores, both achieving ***0.99***, indicating their ***robustness*** in handling the dataset.
- ***XGBoost*** also performed exceptionally well, with an ***accuracy of 0.98***, showcasing its efficiency in ***structured data modeling***.
- ***LightGBM*** had the ***lowest accuracy of 0.93***, although still ***respectable***, suggesting room for ***improvement*** in handling the given dataset's characteristics.

Evaluation Metrics



Compared to other models, ***CatBoost*** excels with ***superior*** precision, recall, and F1-scores across all classes.

- ***Precision***: ***High*** precision for "***Completed***" and "***Terminated***" classes, as the ***false positives are minimal***.
- ***Recall***:
  - ***Excellent recall*** for "***Completed***" (most instances correctly classified).
  - ***Slight confusion*** with "***Terminated***," as a small portion is misclassified as "Completed."
  - ***Almost perfect*** for "***Suspended***" with negligible errors.
- ***F1-Score***: ***High*** F1-score due to ***balanced precision and recall*** across major classes.
- ***Overall Accuracy***: ***Strong*** accuracy due to ***accurate predictions*** across all major classes.

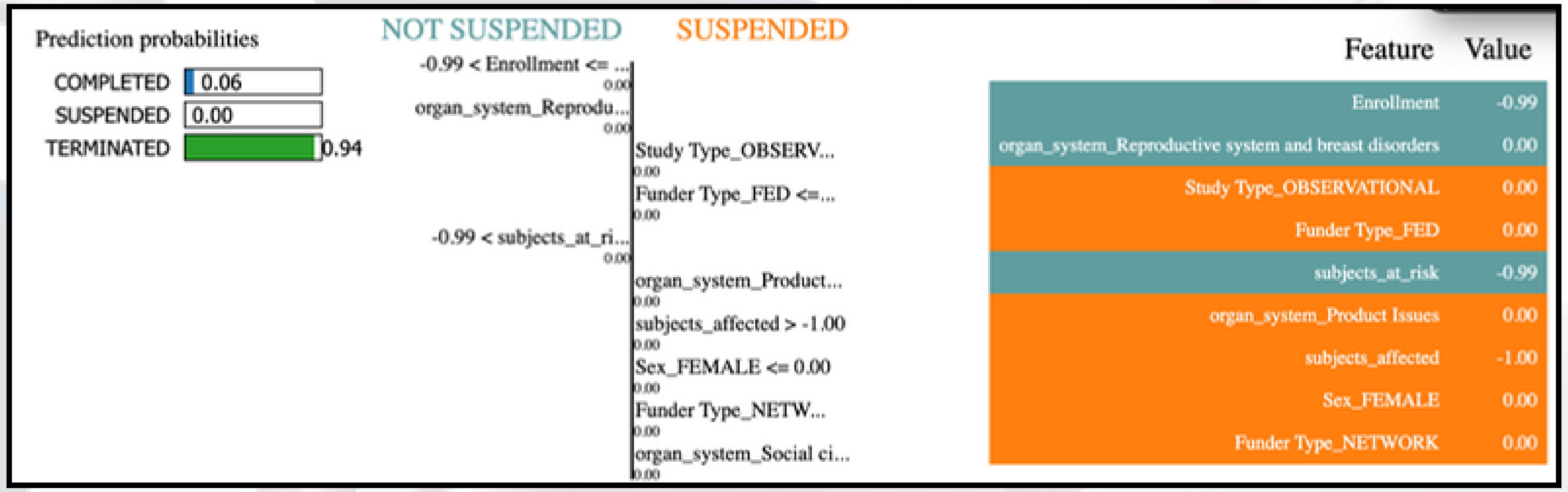
Explainability

Trust in Model Decisions:

- In critical applications like clinical trials, especially for predicting the ***Completion status***, explainability is essential to trust a model's decisions.

LIME:

- ***Local Interpretability***: LIME provides a clear, instance-specific explanation of how a model behaves for a single data point.
- ***Feature Attribution for Predictions***: LIME explains a model's prediction by showing the ***contribution of individual features*** towards the outcome.



- The visualization reveals that the model's decision to classify the instance as "***TERMINATED"*** is driven primarily by specific feature values (***Enrollment, subjects_at_risk, subjects_affected***)
- Other features, such as ***organ_system_Reproductive system*** and ***breast disorders***, contribute slightly but are ***less significant***.
- The ***absence of contribution*** from features like ***Funder Type*** and ***Sex_FEMALE*** indicates that they play a ***negligible role*** in this specific prediction.
- ***LIME*** helps make the model's decision process transparent and builds trust in its outputs.

CONCLUSION

Best Model

CatBoost outperformed other models with the ***highest precision, recall, and F1-scores***, handling edge cases with ***minimal misclassifications***, making it ideal for deployment.

Alternate Model

- ***Random Forest*** provided comparable results with high accuracy and recall,
- ***XGBoost*** performed well but struggled with minority classes.

LightGBM Challenges

LightGBM showed higher misclassifications, lowering overall accuracy. ***Tuning or feature engineering*** may improve its performance.

Key Takeway

LightGBM showed higher misclassifications, lowering overall accuracy. Tuning or feature engineering may improve its performance.