Maternal Mortality Risk Prediction Model Summary

Objective:

To predict rare maternal mortality events from a large-scale Telangana health dataset using a LightGBM classifier with robust handling of class imbalance and interpretability using SHAP.

What We Did

1. Data Preprocessing

- Loaded the dataset in **batches** using `pyarrow` for memory efficiency.
- Ensured type consistency across numeric, flag, and categorical columns.
- Applied **NaN imputation**: median for numerics, mode for categoricals.
- Engineered **interaction features** like `anemia_severe_systolic_bp` and `hypertension_hemoglobin`.
- Cleaned and standardized categorical variables (e.g., top 10 SYS_DISEASE).

2. Stratified Sampling

- Created a **1 million record** balanced sample with all 1,377 positive maternal mortality cases.
- Used **random undersampling** for negatives, preserving class distribution.

3. Feature Selection

- Chose 40+ features across numeric, flag, and encoded categorical variables.
- Used `TargetEncoder` for categorical variables to avoid high dimensionality.

4. Modeling: LightGBM with Cross-Validation

- Applied **5-fold stratified CV** using `StratifiedKFold`.
- Addressed imbalance using **SMOTEENN** (resampling strategy = 0.1).
- Tuned LightGBM with:
    - `max_depth=7`, `n_estimators=500`
    - `scale_pos_weight ≈ 1087`
- Trained using **early stopping** and optimized for **PR-AUC** and **F1 score**.

5. Threshold Evaluation

- Evaluated performance at thresholds: `0.1`, `0.2`, `0.3`, `0.4`.
- Used metrics: **F1 Score**, **Accuracy**, **Precision**, **Recall**, **PR-AUC**, **Confusion Matrix**.
- Best threshold on test: `0.4`, with F1 = `0.0036` and Recall = `86.9%`.

6. Model Interpretation with SHAP

- Used `TreeExplainer` on the best fold's model.
- Generated:
    - **SHAP Summary Plot** (impact of individual features).
    - **SHAP Bar Plot** (mean absolute SHAP values).
- Top influential features:
    - `WEIGHT_child_min`, `inadequate_weight_gain`, `HEMOGLOBIN_mean`, `anemia_mild`, `BLOOD_GRP`.

Best Results Summary

Test Set (Best Fold @ Threshold 0.4)

- **F1 Score:** 0.0036
- **ROC AUC:** 0.6690
- **PR-AUC:** 0.0033
- **Precision:** 0.0018
- **Recall:** 0.8691
- **Confusion Matrix:**
  `[[65655, 134070], [36, 239]]`

Recommendations for Improvement

1. Feature Engineering

- Include **temporal patterns** (ANC visit sequence).
- Derive **risk score composites** (e.g., anemia severity + age).
- Leverage **domain knowledge** to flag critical pregnancies.

2. Advanced Models

- Try **CatBoost** or **XGBoost** with Bayesian optimization.
- Experiment with **tabular neural networks** or **AutoGluon**.

3. Better Sampling Strategies

- Use **cost-sensitive learning** or **focal loss** instead of oversampling.

- Explore **ensemble resampling** (e.g., SMOTE+TomekLinks).

4. Threshold Optimization

- Use **Youden's J index** or **precision-recall tradeoff curves** dynamically per region or hospital.

5. Interpretability

- Combine **LIME** + **SHAP** for patient-level risk explanations.
- Use SHAP values to flag **misclassified positive cases** for manual review