Code Explanation and Best Results for Stillbirth Risk Prediction Model

Overview

This document explains a machine learning pipeline designed to predict stillbirth risk using a Random Forest Classifier. The pipeline includes data preprocessing, stratified sampling, model training with cross-validation, evaluation at multiple thresholds, and feature importance analysis using SHAP. The code processes a large dataset stored in a Parquet file, handles class imbalance with SMOTE, and evaluates model performance with metrics like AUC, F1 score, accuracy, precision, and recall. The best results from the model are summarized based on the test set performance.

Code Explanation

1. Data Loading and Preprocessing

The dataset is loaded from a Parquet file (`telangana_data_with_features_and_targets (1).parquet`) using the `pyarrow` library for efficient batch processing. The `prepare_data_for_targets` function processes the data in chunks to handle large datasets:

- **Required Columns**: Ensures critical columns like `MOTHER_ID` and `GRAVIDA` are present.
- **Numeric Columns**: Converts specified columns (e.g., `AGE`, `HEMOGLOBIN_mean`) to numeric types, handling non-numeric values by coercing them to NaN.
- **Flag Mapping**: Converts categorical flag columns (e.g., `IS_CHILD_DEATH`, `IS_DEFECTIVE_BIRTH`) to binary (0/1) using a predefined `flag_map`.
- **GRAVIDA Cleaning**: Specifically handles the `GRAVIDA` column by converting non-numeric values to NaN and filling missing values with 0.
- **NaN Handling**: Fills missing values in numeric columns with 0 to ensure model compatibility.
- **Output**: A cleaned `pandas` DataFrame (`df`) ready for modeling.

The dataset contains 4,029,571 records, with key target columns like `stillbirth_risk`, `IS_CHILD_DEATH`, and `maternal_mortality_risk`. Since `stillbirth_risk` initially had no positive cases, it was derived from `IS_CHILD_DEATH`, resulting in 48,667 positive cases.

2. Stratified Sampling

The `create_stratified_sample` function creates a balanced sample of 2,000,000 records, prioritizing critical cases (e.g., child deaths, maternal deaths, stillbirths) to ensure sufficient positive cases for modeling:

- **Critical Case Selection**: Includes all records with `IS_CHILD_DEATH=1`, `maternal_mortality_risk=1`, or `stillbirth_risk=1`.
- **Oversampling**: If fewer than 100 positive cases are found, oversamples positive cases to meet the minimum requirement.
- **Remaining Cases**: Fills the remaining sample size with randomly selected non-critical cases.
- **Output**: A sampled DataFrame (`sample_df`) with 48,667 positive and 1,951,333 negative cases for `stillbirth_risk`.

3. Feature Selection

Features are selected based on numeric data types (`float64`, `float32`, `int64`, `int32`, `int8`) and exclusion of leakage-prone columns (e.g., `DEL_COMPLICATIONS`, `CHILD_ID`). The final feature list includes 43 variables like `HEMOGLOBIN_mean`, `AGE`, `GRAVIDA`, and `anemia_mild`.

4. Model Training and Cross-Validation

A Random Forest Classifier is trained using 5-fold stratified cross-validation to handle class imbalance and evaluate performance robustly:

- **Train-Test Split**: Splits the data into 80% training (`X_train_full`, `y_train_full`) and 20% test (`X_test`, `y_test`) sets, maintaining class distribution with stratification.
- **SMOTE**: Applies Synthetic Minority Oversampling Technique (SMOTE) in each fold to balance the training set (1:2 positive-to-negative ratio).
- **Model Parameters**:
  - `n_estimators=200`: Uses 200 trees for robust learning.
  - `max_depth=15`: Limits tree depth to prevent overfitting.
  - `min_samples_split=20`, `min_samples_leaf=10`: Adds flexibility to tree splits.

    - `class_weight={0:1, 1:10}`: Heavily weights the positive class to prioritize recall.
- **Cross-Validation**: Uses `StratifiedKFold` with 5 folds, ensuring each fold maintains the class distribution.

5. Model Evaluation

The model is evaluated at four decision thresholds (0.3, 0.4, 0.5, 0.6) to optimize for high recall, critical for stillbirth prediction. Metrics include:

- **AUC**: Area under the ROC curve for discrimination ability.
- **F1 Score**: Harmonic mean of precision and recall.
- **Accuracy**: Proportion of correct predictions.
- **Precision**: Proportion of positive predictions that are correct.
- **Recall**: Proportion of actual positives correctly identified.
- **Confusion Matrix**: Details true positives, false positives, true negatives, and false negatives.

For each fold, the model computes predictions, calculates metrics, and identifies the optimal threshold maximizing the F1 score using the precision-recall curve.

6. Test Set Evaluation

The best model (highest average F1 score across thresholds) is selected from the cross-validation folds and evaluated on the test set. Metrics are computed at the same thresholds (0.3, 0.4, 0.5, 0.6).

7. SHAP Analysis

SHAP (SHapley Additive exPlanations) is used to interpret feature importance:

- A sample of 1,000 test set records (`X_test_sample`) is used to reduce computation time.
- `shap.TreeExplainer` computes SHAP values for the best model.
- Two plots are generated:
    - **Summary Plot**: Shows the impact of each feature on predictions.
    - **Bar Plot**: Displays mean absolute SHAP values for feature importance.
- A DataFrame (`importance_df`) lists features ranked by SHAP importance.

Best Results

Cross-Validation Mean Metrics

- **AUC**: 0.5859 ± 0.0027
- **Threshold 0.3**:
  - F1 Score: 0.0491 ± 0.0001
  - Accuracy: 0.1087 ± 0.0048
  - Precision: 0.0252 ± 0.0001
  - Recall: 0.9461 ± 0.0051
- **Threshold 0.4**:
  - F1 Score: 0.0507 ± 0.0002
  - Accuracy: 0.1999 ± 0.0047
  - Precision: 0.0261 ± 0.0001
  - Recall: 0.8784 ± 0.0046
- **Threshold 0.5**:
  - F1 Score: 0.0531 ± 0.0005
  - Accuracy: 0.3420 ± 0.0139
  - Precision: 0.0275 ± 0.0003
  - Recall: 0.7582 ± 0.0109
- **Threshold 0.6**:
  - F1 Score: 0.0570 ± 0.0004
  - Accuracy: 0.5257 ± 0.0126
  - Precision: 0.0299 ± 0.0002
  - Recall: 0.5890 ± 0.0145

Test Set Metrics (Best Model from Fold 4)

- **AUC**: 0.5850

- **Threshold 0.3**:

  - F1: 0.0491
  - Accuracy: 0.1051
  - Precision: 0.0252
  - Recall: 0.9488
  - Confusion Matrix: [[32808, 357459], [498, 9235]]
- **Threshold 0.4**:

  - F1: 0.0506
  - Accuracy: 0.1992
  - Precision: 0.0261
  - Recall: 0.8771
  - Confusion Matrix: [[71146, 319121], [1196, 8537]]
- **Threshold 0.5**:

    - F1: 0.0533
    - Accuracy: 0.3433
    - Precision: 0.0276
    - Recall: 0.7600
    - Confusion Matrix: [[129922, 260345], [2336, 7397]]
- **Threshold 0.6**:
    - F1: 0.0568
    - Accuracy: 0.5257
    - Precision: 0.0298
    - Recall: 0.5869
    - Confusion Matrix: [[204580, 185687], [4021, 5712]]
- **Best Threshold**: 0.6 with F1 Score: 0.0568

SHAP Feature Importance

The top 10 features influencing stillbirth risk predictions are:

1. **HEMOGLOBIN_mean**: 0.0395
2. **HEMOGLOBIN_max**: 0.0370
3. **HEMOGLOBIN_min**: 0.0365
4. **anemia_mild**: 0.0328
5. **AGE**: 0.0319
6. **inadequate_weight_gain**: 0.0233
7. **multigravida**: 0.0183
8. **GRAVIDA**: 0.0165
9. **PARITY**: 0.0162
10. **AGE_preg**: 0.0143

Interpretation

- **Model Performance**: The model achieves high recall (0.9488 at threshold 0.3) but low precision (0.0252), indicating it identifies most positive cases but with many false positives. The AUC (0.5850) suggests moderate discriminative ability, likely due to class imbalance and data complexity.
- **Best Threshold**: Threshold 0.6 provides the highest F1 score (0.0568) on the test set, balancing precision and recall.
- **Key Features**: Hemoglobin levels (`HEMOGLOBIN_mean`, `HEMOGLOBIN_max`, `HEMOGLOBIN_min`) and anemia indicators (`anemia_mild`) are the most influential, highlighting the importance of maternal health metrics in predicting stillbirth risk.

- **SHAP Analysis**: The SHAP plots (`shap_summary_plot.png`, `shap_importance_bar.png`) visualize feature impacts, confirming hemoglobin and age-related features as critical drivers.

Conclusion

The pipeline effectively processes a large healthcare dataset, handles class imbalance with SMOTE, and trains a Random Forest model optimized for high recall. While the model successfully identifies most stillbirth cases, the low precision suggests a need for further feature engineering or alternative models (e.g., LightGBM, as partially implemented). The SHAP analysis provides actionable insights into key risk factors, particularly hemoglobin levels and maternal age, which can guide clinical interventions.