Model Performance Summary for Provider Fraud Prediction

This document summarizes the performance of the Logistic Regression and XGBoost models trained to predict potentially fraudulent healthcare providers. The models were evaluated on a validation set, and class imbalance was addressed using SMOTE during training.

Evaluation Metrics Overview

Given the highly imbalanced nature of the target variable (fraudulent vs. non-fraudulent providers), the following metrics are crucial for evaluation:

- **Precision (Positive Predictive Value):** The proportion of correctly predicted positive cases (fraud) out of all cases predicted as positive. High precision means fewer false alarms.
- **Recall (Sensitivity/True Positive Rate):** The proportion of actual positive cases (fraud) that were correctly identified. High recall means fewer missed fraud cases.
- **F1-Score:** The harmonic mean of Precision and Recall, providing a balance between the two.
- **ROC-AUC (Receiver Operating Characteristic - Area Under the Curve):** Measures the model's ability to distinguish between the positive and negative classes across various classification thresholds. A higher AUC indicates better discriminatory power.
- **Confusion Matrix:** A table showing the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Feature Engineering

The initial raw datasets (Train_Beneficiarydata, Train_Inpatientdata, Train_Outpatientdata, and Train) were transformed and combined to create a rich set of provider-level features. This process involved:

1. **Date Conversions and Age Calculation:** Dates of birth (DOB) and death (DOD) in the beneficiary data were converted to datetime objects. Beneficiary age was calculated as of a reference date (2009-12-31) or at the time of death.
2. **Categorical to Numerical Mapping:** RenalDiseaseIndicator ('Y'/'0') and chronic condition flags (1/2) were mapped to binary (1/0) numerical representations.
3. **Claim-Level Feature Creation:**
    - **Inpatient Claims:** LengthOfStay (Discharge Date - Admission Date) and ClaimDuration (Claim End Date - Claim Start Date) were calculated. Counts of diagnosis codes (NumDiagnosisCodes) and procedure codes (NumProcedureCodes) were derived. Binary flags (HasAttendingPhysician,

HasOperatingPhysician, HasOtherPhysician) were created to indicate the presence of different physician roles.

    - **Outpatient Claims:** ClaimDuration, NumDiagnosisCodes, NumProcedureCodes, and physician presence flags were similarly created.

4. **Merging Beneficiary and Claims Data:** Beneficiary demographic and health condition features were joined to their respective inpatient and outpatient claims.
5. **Provider-Level Aggregation:** Claims and beneficiary data were aggregated to the Provider level. This involved calculating various statistics for each provider, such as:
    - Total, average, min, max, and standard deviation of InscClaimAmtReimbursed and DeductibleAmtPaid.
    - Average and maximum LengthOfStay (for inpatient) and ClaimDuration.
    - Number of unique beneficiaries and unique physicians associated with the provider.
    - Proportion of claims involving specific physician roles.
    - Average age, most common gender, most common race, and average prevalence of chronic conditions among the provider's beneficiaries.
6. **Combined Provider Features:** New features were created by combining inpatient and outpatient aggregations, including TotalClaims, TotalReimbursement, AvgReimbursementPerClaim, TotalUniqueBeneficiaries, ClaimsPerBeneficiary, TotalUniqueAttendingPhysicians, TotalUniqueOperatingPhysicians, TotalUniqueOtherPhysicians, InpatientClaimsRatio, InpatientReimbursementRatio, AvgHasOperatingPhysician, and AvgHasOtherPhysician.
7. **Target Variable Preparation:** The PotentialFraud column from the Train dataset was merged with the aggregated provider features and converted from 'Yes'/'No' to numerical 1/0.

Missing Values Summary Across Datasets

During the data management and feature engineering phases, missing values were identified and handled. Here's a summary of the key observations regarding missing data:

- **Train_Beneficiarydata-1542865627584.csv:**
    - DOD (Date of Death) had a significant number of missing values (137,135 out of 138,556), which were handled by calculating age as of a reference date for living beneficiaries.
    - All other columns were complete.

- **Train_Inpatientdata-1542865627584.csv:**
    - OtherPhysician (35,784 missing) and OperatingPhysician (16,644 missing) had substantial missingness, which was addressed by creating binary "Has" flags and counting unique physicians.
    - ClmDiagnosisCode_2 to ClmDiagnosisCode_10 and ClmProcedureCode_1 to ClmProcedureCode_6 showed increasing missing values for higher-numbered codes, indicating fewer diagnoses/procedures per claim. These were addressed by counting non-null codes.
    - DeductibleAmtPaid had 899 missing values, which were imputed with 0.
    - AttendingPhysician had a minor number of missing values (112).
- **Train_Outpatientdata-1542865627584.csv:**
    - ClmProcedureCode_1 to ClmProcedureCode_6 had very high missingness (with ClmProcedureCode_5 and ClmProcedureCode_6 being entirely empty), suggesting procedures are rarely recorded for outpatient claims. These were handled by counting non-null codes.
    - OperatingPhysician (427,120 missing), ClmAdmitDiagnosisCode (412,312 missing), and OtherPhysician (322,691 missing) also showed high missingness. These were handled similarly to inpatient data.
    - ClmDiagnosisCode_2 to ClmDiagnosisCode_10 had increasing missing values for higher-numbered codes.
    - AttendingPhysician had minor missing values (1,396).
- **Train-1542865627584.csv:**
    - This dataset was complete with no missing values.

After the entire feature engineering pipeline, including aggregation and specific fillna strategies, the final final_df used for modeling contains **zero missing values**, ensuring a complete dataset for training.

1. Logistic Regression Model Performance

The Logistic Regression model was tuned using Randomized Search CV and trained on SMOTE-resampled data.

Classification Report (on Validation Set):

Class	Precision	Recall	F1-Score	Support
0 (No Fraud)	0.99	0.90	0.94	981

1 (Fraud)	0.49	0.89	0.63	101
Accuracy			0.90	1082
Macro Avg	0.74	0.90	0.79	1082
Weighted Avg	0.94	0.90	0.91	1082

Confusion Matrix (on Validation Set):

	Predicted No Fraud	**Predicted Fraud**
Actual No Fraud	887	94
Actual Fraud	11	90

ROC-AUC Score: 0.9686**Error! Filename not specified.**

Interpretation: The Logistic Regression model demonstrates strong performance, particularly in identifying actual fraudulent providers with a high Recall of 0.89. This means it correctly identifies 89% of all fraudulent cases. Its Precision for fraud is 0.49, indicating that roughly half of the providers flagged as fraudulent by this model are indeed fraudulent, while the other half are false positives. The high ROC-AUC score of 0.9686 suggests excellent overall discriminatory power.

2. XGBoost Model Performance

The XGBoost model was also tuned using Randomized Search CV and trained on SMOTE-resampled data.

Classification Report (on Validation Set):

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0 (No Fraud)	0.97	0.97	0.97	981
1 (Fraud)	0.70	0.75	0.73	101
Accuracy			0.95	1082
Macro Avg	0.84	0.86	0.85	1082
Weighted Avg	0.95	0.95	0.95	1082

Confusion Matrix (on Validation Set):

	Predicted No Fraud	Predicted Fraud
Actual No Fraud	949	32
Actual Fraud	25	76

ROC-AUC Score: 0.9683**Error! Filename not specified.**

Interpretation: The XGBoost model exhibits robust performance, with a high Precision of 0.70 for the fraud class, meaning 70% of providers flagged as fraudulent are indeed fraudulent, which is a significant improvement in reducing false alarms compared to Logistic Regression. Its Recall for fraud is 0.75, indicating it correctly identifies 75% of actual fraudulent providers. The F1-Score of 0.73 shows a strong balance between precision and recall for the fraud class. The ROC-AUC score of 0.9683 is very high, indicating excellent discriminatory power, comparable to Logistic Regression in this regard.

Comparative Analysis

Both Logistic Regression and XGBoost models demonstrate very strong discriminatory power with high ROC-AUC scores (around 0.968). However, XGBoost significantly outperforms Logistic Regression in terms of **Precision** for the fraud class (0.70 vs. 0.49). This means XGBoost is much better at minimizing false positives, which is often a critical factor in fraud detection to reduce the burden of investigating non-fraudulent cases. While Logistic Regression achieved a slightly higher Recall (0.89 vs. 0.75), XGBoost offers a better balance between Precision and Recall (higher F1-Score of 0.73 vs. 0.63 for fraud). Depending on the business objective (e.g., minimizing false alarms vs. catching every possible fraud), the preferred model might vary, but XGBoost generally provides a more balanced and efficient solution for identifying fraudulent providers with fewer false positives.