**Model Description**

This model aims to detect fraudulent healthcare claims using machine learning on the ‘Healthcare Provider Fraud Detection Analysis’ dataset from Kaggle. The dataset contains information about healthcare providers, their transactions, and whether or not they were flagged as fraudulent. To enhance the model’s performance, Bayesian optimization is employed for hyperparameter tuning, improving its predictive power.

Model Type: Supervised Classification

Input:

* Training data: 32958 records which was split into training, validation and test set
* Features: 54 columns (e.g., total\_claims, unique\_benefeciaries, inpatient\_days)
* Target variable: ‘PotentialFraud’ (binary: ’Yes’ or ‘No’)

The dataset includes following key features:

* TotalClaims: Number of claims submitted by the provider.
* InpatientDays: Total number of inpatient days claimed.
* UniqueBeneficiaries: Number of unique patients for the provider
* OutpatientDays: Total number of outpatient days claimed.
* TotalReimbursetmentAmount: Total money reimbursed by insurance.

Output:

* Target variable: ‘PotentialFraud’ (binary: ’Yes’ or ‘No’)

**Model Architecture:**

Model Type: Supervised Classification

Preprocessing Steps

1. Missing value Handling
2. Feature encoding
3. Scaling

Baseline Model

* Model: Random Forest Classifier
* Metrics:

Accuracy: 0.85

Precision: 0.88

Recall: 0.82

F1-Score: 0.85

Bayesian Optimization for Hyperparameter Tuning

To improve the model’s performance, we applied Bayesian Optimization for hyperparameter tuning in order to find the optimal values.

* Hyperparameters Tuned:

N\_estimators: Number of trees in Random Forest

Max\_depth: Maximum depth of each tree

Min\_samples\_split: Minimum number of samples required to split a node.

Min\_samples\_leaf: Minimum number of samples required at each leaf node.

* Bayesian Optimization Process:

Acquisition Function: Expected Improvement (EI)

Number of iterations: 50 iterations to maximize performance

Libraries: Scikit-optimize (skopt)

**Performance**

Metrics

After applying Bayesian optimization, the tuned Random Forest model demonstrated improved performance:

* Accuracy: 0.88
* Precision: 0.90
* Recall: 0.87
* F1-Score: 0.89
* ROC AUC Score: 0.91

Feature importance

The following features were found to be the most important in determining healthcare fraud:

1. TotalClaims
2. UniqueBeneficiaries
3. TotalReimbursementAmount
4. InpatientDays
5. OutpatientDays

Ethical Considerations

1. Bias and Fairness: Ensuring that the model dies not unfairly penalize specific healthcare providers based on incomplete or biased data. Continuous monitoring for fairness is essential.
2. Transparency: The model is designed to provide explanations for predictions, using feature importance and SHAP (Shapley Additive exPlanations) values to ensure decisions are interpretable.

**Limitations**

* Data Quality: Fraud detection models heavily rely on quality of data, and inaccuracies in the dataset may lead to false positives or negatives.
* Class imbalance: Fraudulent claims may be underrepresented in the dataset, which could affect model performance.
* Scalability: Bayesian optimization can be computationally expensive for larger datasets or more complex models.
* Due to large size of the dataset the time taken for execution is quite high.