**Identify Healthcare Insurance Fraud**

**What is Healthcare Insurance Fraud**

Healthcare fraud is an organized crime which involves peers of providers, physicians, beneficiaries acting together to make fraud claims.

Some of the most common types of frauds by providers are:

* Billing for services that were not provided.
* Duplicate submission of a claim for the same service.
* Misrepresenting the service provided.
* Charging for a more complex or expensive service than was actually provided.
* Billing for a covered service when the service actually provided was not covered.

**Business Problem**

Statistics shows that 15% of the total Medicare expense are caused due to fraud claims. Insurance companies are the most vulnerable institutions impacted due to these bad practices. Insurance premium is also increasing day by day due to this bad practice.

**Objective**

Our objective is to predict whether a provider is potentially fraudulent or probability score of that provider's fradulent activity and also find the reasons behind it as well to prevent the financial loss.

Find out the important features which are the reasons behind the potentially fraudulent providers. Such as if claim amount is high for a patient whose risk score is low, then it is suspicious.

**Architecture Diagram of Machine Learning**

A blue and white diagram with blue text

Description automatically generated

A diagram of a machine learning project

Description automatically generated

**ML Formulation**

Build a binary classification model based on the claims filled by the provider along with Inpatient data, Outpatient data, Beneficiary details to predict whether the provider is potentially fraudulent or not.

**Business constraints**

a). The cost of misclassification is very high. False Negative and False Positive should be as low as possible. If fraudulent providers are predicted as non-fraudulent (False Negative) it is a huge financial loss to the insurer and if legitimate providers are predicted as fraudulent (False Positive) it will cost for investigation and also, it’s a matter of reputation of the agency.

b) Model interpretability is very important because the agency or insurer should justify that fraudulent activity and may need to set up a manual investigation. It should not be a black box type prediction.

c)  The insurer should pay the claim amount to the provider for legitimate claims within 30 days. So, there are no such strict latency constraints, but it should not take more than a day because depending on the output of the model the agency may need to set up an investigation.

**Performance metric**

As the dataset in healthcare fraud is highly imbalanced (very few fraud cases), ‘accuracy’ won’t be the proper metric. An important initial step will be to plot the confusion matrix. Then we need to check the misclassification i.e., FP and FN. FN means the cases predicted by the model are legitimate, but it is fraudulent. FP means the cases detected by the model are fraudulent, but it is legitimate.

So, the performance metrics are:  
**a) Confusion Matrix:** It is the table where TP, FP, TN, FN counts will be plotted. From this table, we can visualize the performance of the model.

**b) F1 Score:**It is the harmonic mean of precision and recall.  
*F1 Score = 2(Precision \* Recall)/ (Precision + Recall)*  
where **Precision = TP/(TP+FP)** and **Recall = TP/(TP+FN)**. As F1 score consists of both Precision and Recall it will be correct metric for this problem.

**c) AUC Score:** AUC stands for Area Under ROC (Receiver Operating Characteristics) Curve. ROC plots TPR concerning FPR for different thresholds. The area under the curve depends on the ranking of the predicted probability score, not on absolute values.

**d) FPR and FNR:** As the cost of misclassification is very high, we need to check the FPR and FNR separately, it should be as low as possible.

**Approach**

a. Split the data into Train and Validation (80:20)  
b. Oversample the data using SMOTE (majority: minority) to make 80:20, 75:25, 65:35, and 50:50.

c. Feature importance will be done  
d. Use Logistic Regression, Decision Tree, Support Vector Classifier, and Naive Bayes for all these 4 oversampled datasets. Pick the best model based on the performance score.

**Tools and Technology**

**Softwares**

* Google Chrome for web browser
* Python
* Anaconda, Jupyter Notebook
* Frontend – HTML, React, Flask
* Database – LiteSQL

**Deployment/Hosting**

* Frontend + Backend --> Heroku

**Demo**

For demo purposes, we will be able to showcase the following: -

* We will display a web interface where we can input the attributes of the claim, following which a model will execute to predict whether the claim if fraudulent or not.
* Furthermore, we can explain the working of the ML model as well as the tuning done to optimize the model with respect to the dataset at hand.

**Next steps (Future Improvements)**

As future steps, we can look forward to integration of Deep Learning for capturing intricate patterns and relationships within vast and heterogenous healthcare datasets. By trying to leverage Deep Learning, we will be able to further refine our ability to detect subtle anomalies for fraud detection.

**Meet the Team**

* Kuldeep Jolly
* Aman Mishra
* Abhishek Shah