

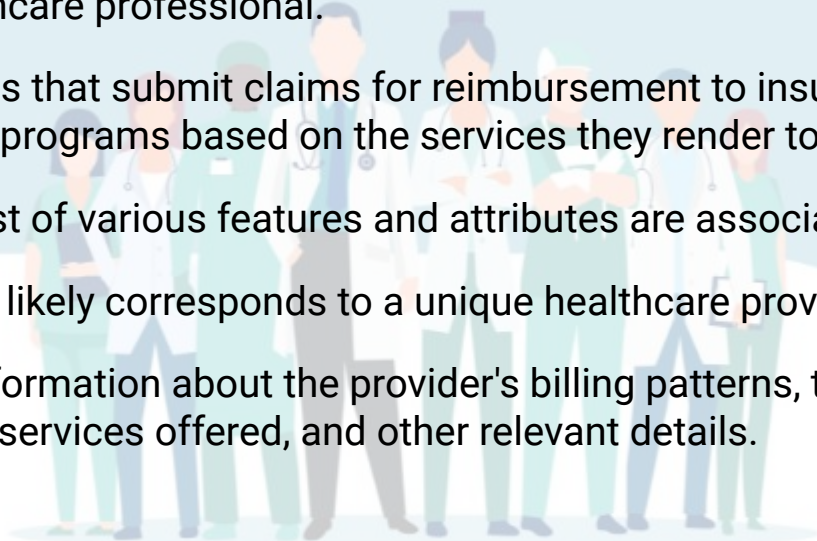
Healthcare Provider Fraud Detection Analysis And Prediction

Shabna Nasser, M.Tech



Who is a Provider?

- A "provider" typically refers to an entity or individual that offers healthcare services, such as a hospital, clinic, or healthcare professional.
- Providers are the entities that submit claims for reimbursement to insurance companies or government healthcare programs based on the services they render to patients.
- Used the dataset consist of various features and attributes are associated with providers.
- Each row in the dataset likely corresponds to a unique healthcare provider,.
- The features provide information about the provider's billing patterns, the number of claims submitted, the types of services offered, and other relevant details.



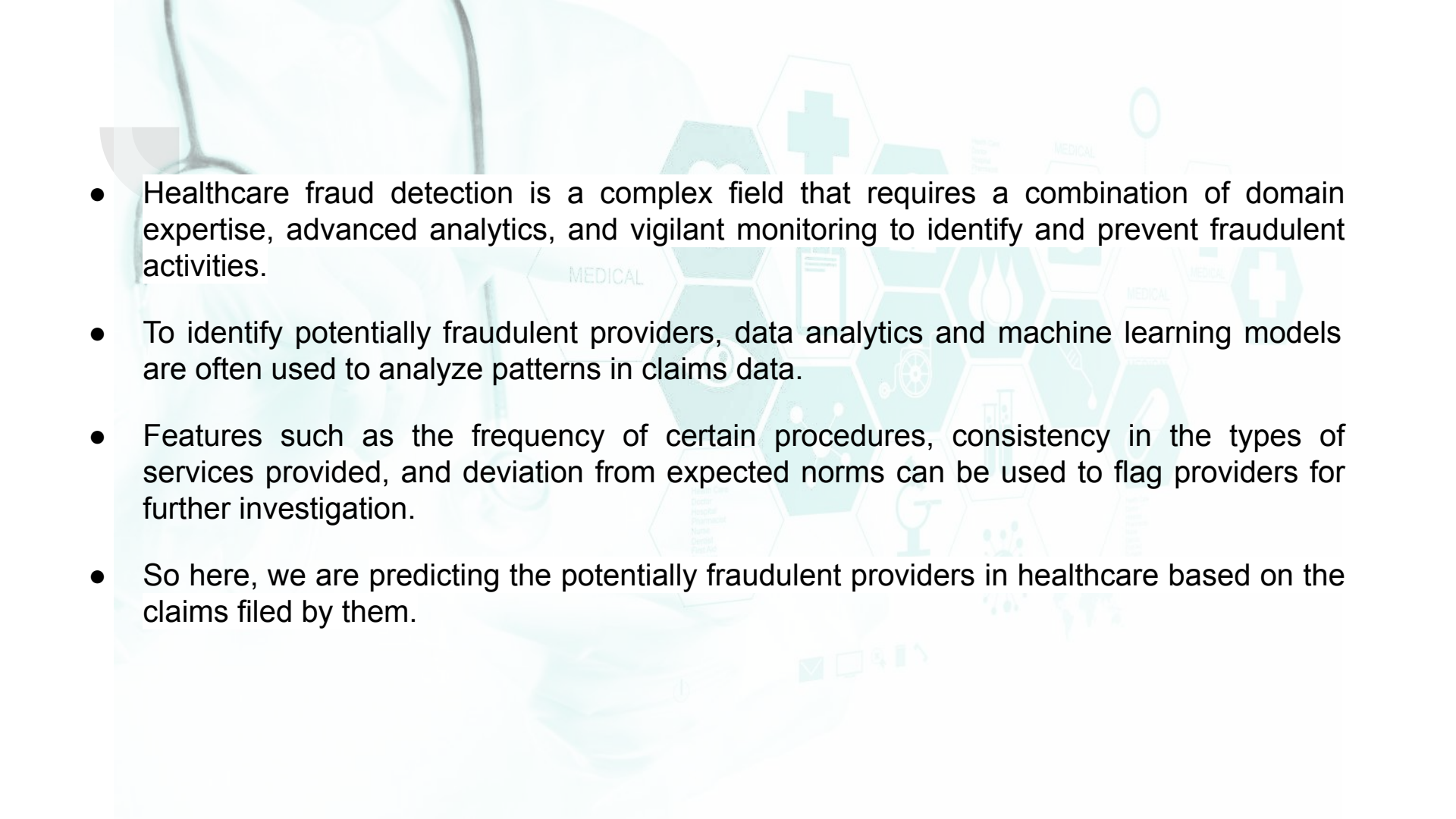
Problem

Providers commonly engaging in deceptive practices such as:

- Submitting claims for services that were never rendered.
- Repetitively submitting claims for the same service.
- Falsifying the description of the service provided.
- Charging for a service of higher complexity or cost than the one actually delivered.
- Billing for a service covered by insurance, when the service provided was not covered.

Goal: Predict potentially fraudulent providers in healthcare based on the claims filed by them.



- 
- The background of the slide features a light teal color with a pattern of hexagons. Inside some hexagons are medical icons such as a cross, a stethoscope, and a pill. The word "MEDICAL" is repeated in a light teal font across the background. In the top left corner, there is a faint image of a stethoscope.
- Healthcare fraud detection is a complex field that requires a combination of domain expertise, advanced analytics, and vigilant monitoring to identify and prevent fraudulent activities.
 - To identify potentially fraudulent providers, data analytics and machine learning models are often used to analyze patterns in claims data.
 - Features such as the frequency of certain procedures, consistency in the types of services provided, and deviation from expected norms can be used to flag providers for further investigation.
 - So here, we are predicting the potentially fraudulent providers in healthcare based on the claims filed by them.



Data Overview

For this project, we are considering Inpatient claims, Outpatient claims and Beneficiary details and provider details (kaggle dataset):

- **Inpatient Data:** Insights of claims filed for patients who are admitted in the hospitals. Also, details like admission and discharge dates and admit diagnosis code.
- **Outpatient Data:** Details of claims filed for those patients who visit hospitals and not admitted in it.
- **Beneficiary Details Data:** Beneficiary KYC details like DOB, DOD, Gender, Race, health conditions, State, Country they belong to, etc.
- **Provider Data:** Provider numbers and corresponding whether this provider is potentially fraud or not.



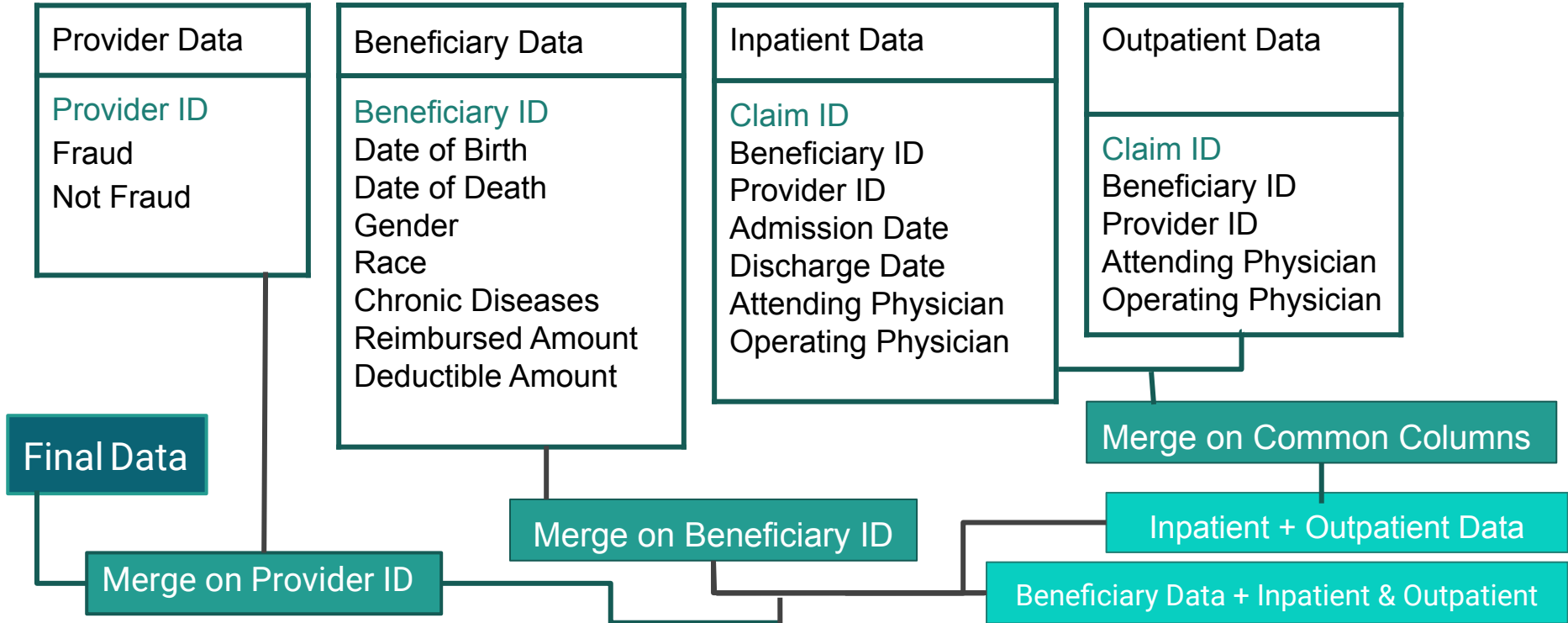
Data Wrangling

- Cleaned each dataset and made the data ready for analysis.
- For this we first changed the target labels of provider data into 0 and 1.
- Checked for null values, duplicates, dropped unwanted columns, filled null. Likewise, explored each dataset.
- Merged them into a single dataframe.
- Added aggregated features.
- Finally, got a dataset with 558211 rows and 325 columns.

| Data | | Rows | Cols. |
|------------|-------------|--------|-------|
| Train Data | Provider | 5410 | 2 |
| | Beneficiary | 138556 | 25 |
| | Outpatient | 517737 | 27 |
| | Inpatient | 40474 | 30 |
| Test Data | Provider | 1353 | 1 |
| | Beneficiary | 63968 | 25 |
| | Outpatient | 125841 | 27 |
| | Inpatient | 9551 | 30 |

Overall Representation of The Dataset

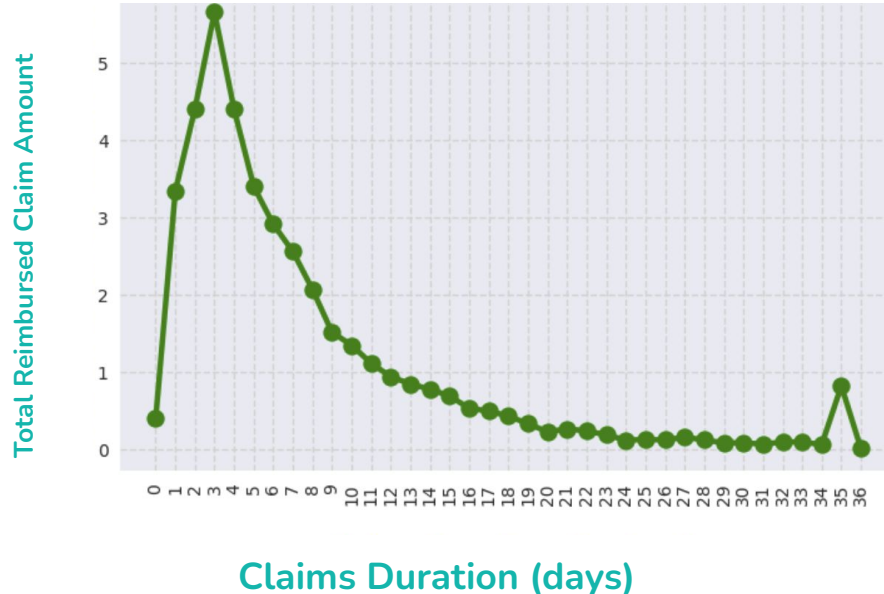
Healthcare Dataset



Data Exploration

Inpatient Claim Amount Vs Duration

Trend of Total Reimbursed Claim Amount for each filed duration(in days)

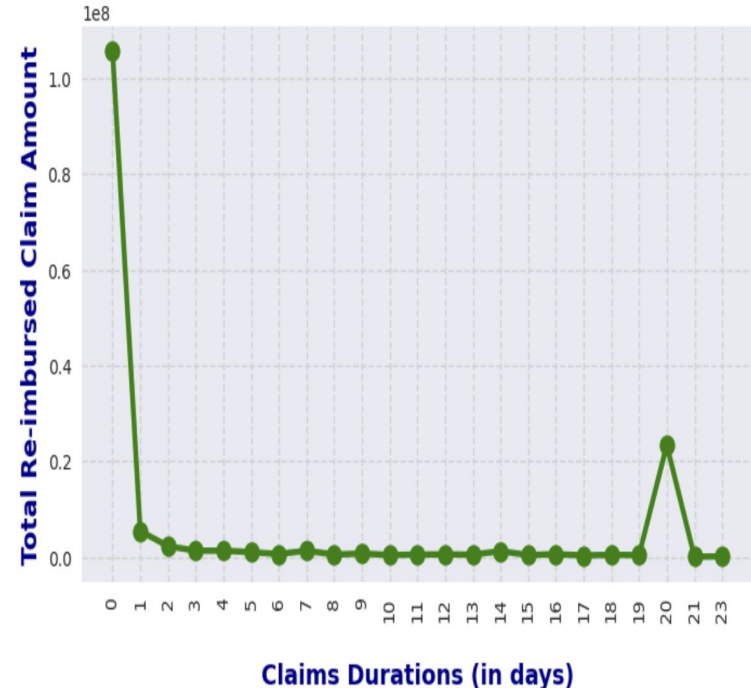


- Relationship between features by plotting them into graphical visualizations.
- Total Reimbursed Amount is the highest for 3 days claims for inpatients.
- For claims with durations from 12 to 34 the total reimbursed amount is very less, and for 35 days duration a clear spike, that can be a potential sign of fraud.

Outpatient Claim Amount Vs Duration

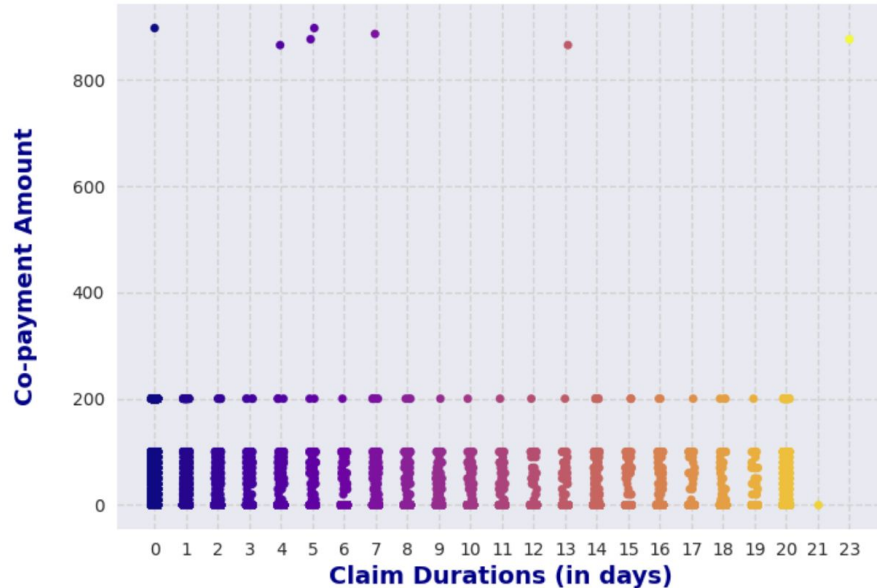
- Total Reimbursed Amount is the highest for 0 days claims.
- Total Reimbursed Claim Amount for majority of the claims filed for each filed duration for outpatients is less than or equal to 2 days.
- For claims with durations from 2 to 19 the total reimbursed amount is very less or similar.
- For a 20 days duration a clear spike, that can be a potential sign of fraud.

Trend of Total Re-imbursed Claim Amount for each filed duration(in days)



Co-payment vs Claim Duration

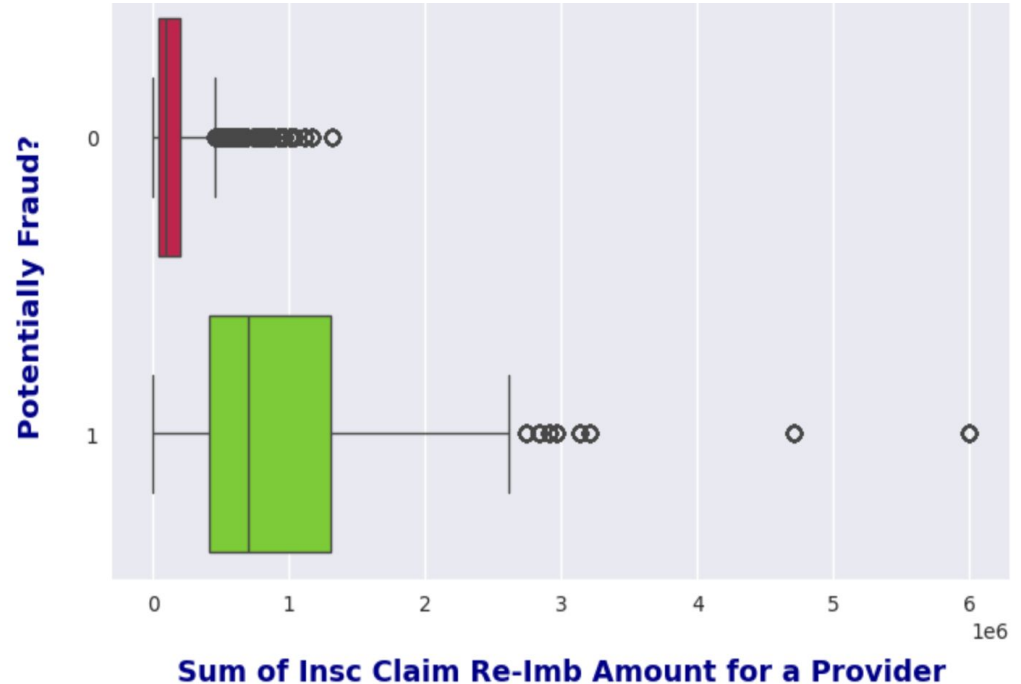
Various co-payments paid for different claim durations



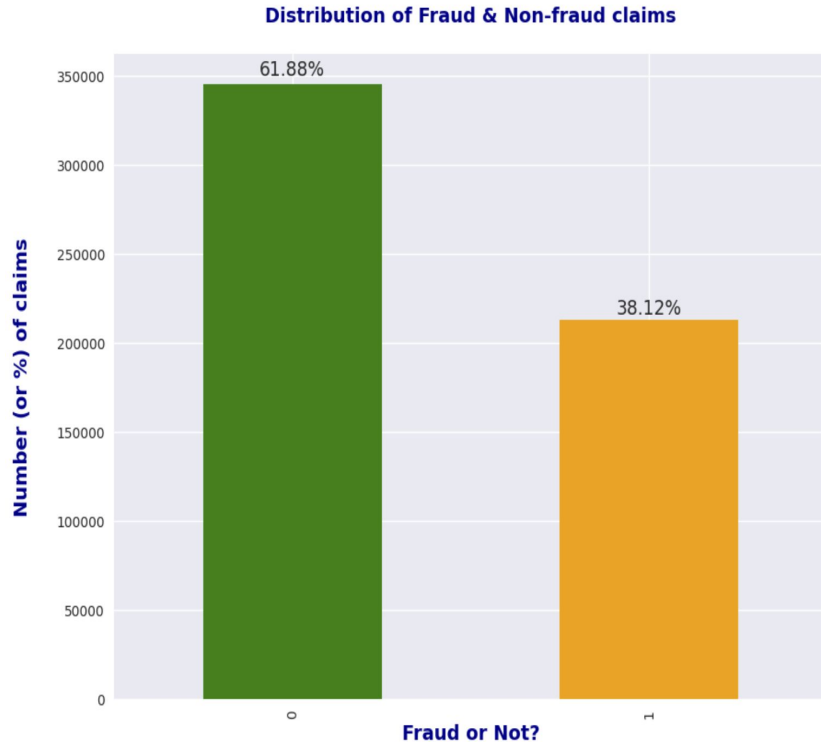
- Co-payment is similar across the various durations.
- Most of the co-payments with 200.
- There are few co-payment which are very high or more than 800.

Distribution of Sum of Insurance Claim Reimburse Amount for a Provider

- If Provider Insurance Claim Reimburse Amount is high then it increases the chances of fraud.



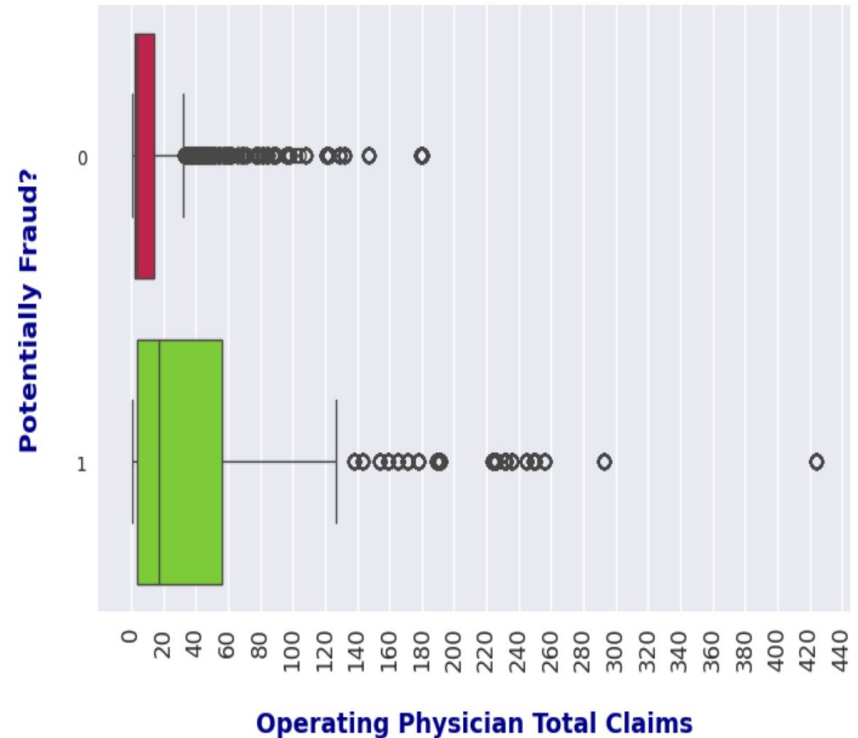
Fraud/Not Fraud vs Number of Claims



- After merging the datasets, checked the Distribution of Fraud & Non-fraud claims.
- Found that 62% of claims are Non-Fraud and 38% of them are Fraud.

Total Number of claims filed by Operating Physician

- If total claims filed by a Operating Physician are greater than 100 then chances of being fraudulent are high.





Modeling

Data Preprocessing Steps:

- 1) One-hot encode categorical features using pandas `get_dummies`.
- 2) Split the dataset into training and testing sets (80-20 split).
- 3) Standardize the magnitude of numeric features using a scaler (RobustScaler).

Cross Validation (CV) for hyperparameter tuning:

- 5 Fold CV.
- Using sklearn `cross_val_score` method.

- Training classifiers using 80% of the whole data.
- Evaluation metric: Accuracy, Precision, Recall, F1-score.

Testing

- Make predictions using holdout dataset (20% of the whole data).
- Performance comparison & feature importance checks.



Classification Algorithms Used

1) Logistic Regression

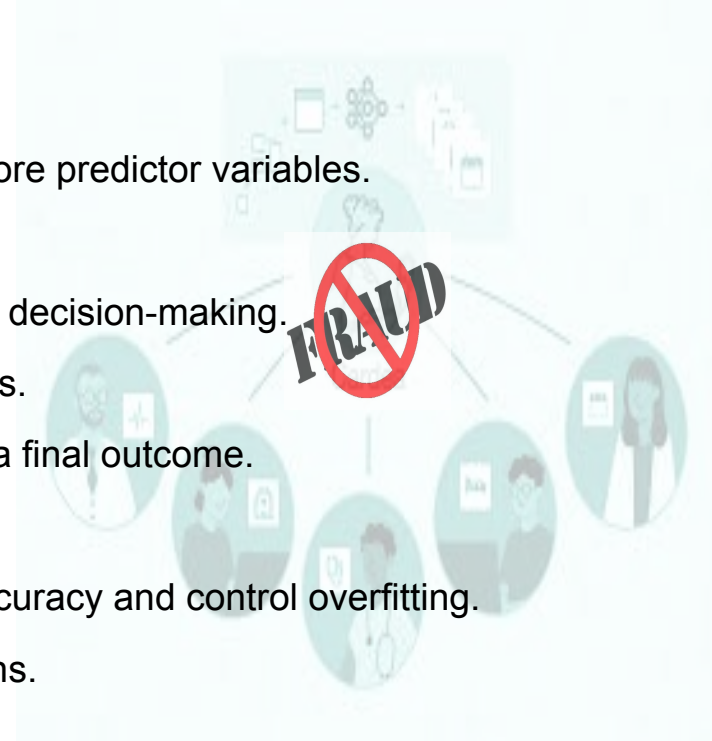
- It is well-suited for binary classification problems.
- Models the probability of a binary outcome based on one or more predictor variables.

2) Decision Tree

- Decision Trees are selected for their ability to capture complex decision-making.
- Partition the feature space into regions based on feature values.
- The tree structure represents a series of decisions, leading to a final outcome.

3) Random Forest

- This combines multiple decision trees to improve predictive accuracy and control overfitting.
- Builds a multitude of decision trees and merges their predictions.





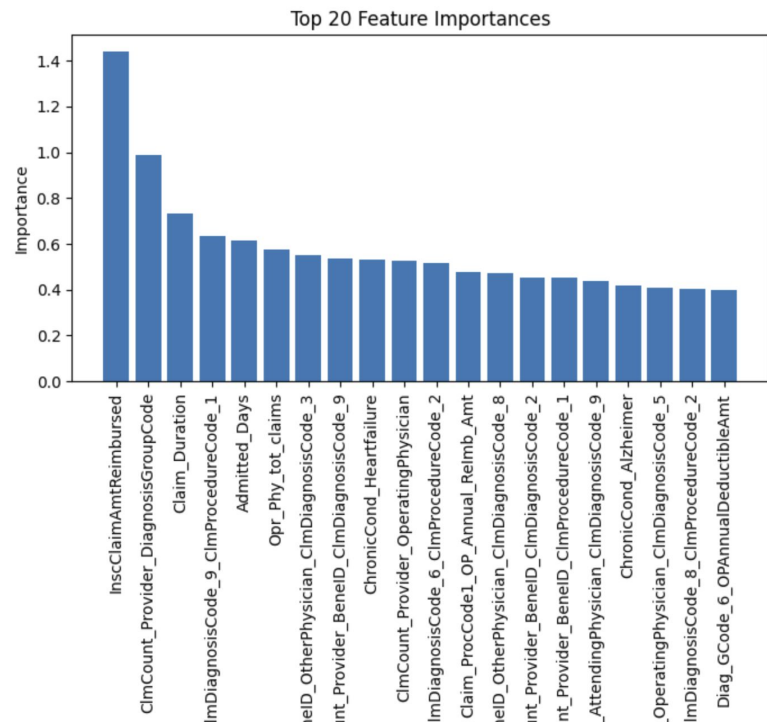
Model Metrics

| Models | Accuracy | Precision | Recall | F1-Score |
|---------------------|----------|-----------|--------|----------|
| Logistic Regression | 0.927 | 0.701 | 0.447 | 0.546 |
| Decision Tree | 0.892 | 0.453 | 0.514 | 0.482 |
| Random Forest | 0.922 | 0.677 | 0.380 | 0.487 |

Best performing model among these models is **Logistic Regression** with an **Accuracy of 0.927** and **F1-Score: 0.546**.



Feature Importance



- The feature importances were calculated.
- Bar chart to provide insights into the relative importance of each feature.
- The top 20 features that exert the most influence on the model's predictions.
- The features are InscClaimAmtReimbursed, ClmCount_Provider_DiagnosisGroupCode, Claim_Duration, Admitted_Days, ClmCount_Provider_BeneID_ClmDiagnosisCode_9_ClmProcedureCode_1, Opr_Phy_tot_claims, ClmCount_Provider_BeneID_OtherPhysician_ClmDiagnosisCode_3, ChronicCond_Heartfailure, ClmCount_Provider_BeneID_ClmDiagnosisCode_9, ClmCount_Provider_OperatingPhysician.



Top Features

- **InscClaimAmtReimbursed:** This feature represents the total amount reimbursed by insurance for a claim. Unusually high reimbursement amounts could be a potential indicator of fraud, as providers might attempt to maximize payouts through inflated claims.
- **ClmCount_Provider_DiagnosisGroupCode:** This indicates the count of claims made by a provider for a specific Diagnosis Group Code. Providers engaging in fraud might show irregular patterns in the distribution of claims across different diagnosis groups.
- **Claim_Duration:** Measure of the duration of a claim, which could be relevant for identifying anomalies. Extremely short or long claim durations may raise suspicion and warrant further investigation.
- **ClmCount_Provider_BeneID_ClmDiagnosisCode_9_ClmProcedureCode_1:** This involves the count of claims made by a provider for a specific beneficiary with the ninth diagnosis code and the first procedure code. It could capture specific patterns that deviate from normal billing practices.



Contd..

- **Admitted_Days:** The number of days a patient was admitted, providing insight into the length of hospital stays. Unusually long or short stays might be indicative of fraudulent billing practices.
- **Opr_Phy_tot_claims:** This denotes the total number of claims associated with operating physicians. Fraud detection might involve identifying providers with disproportionately high or low numbers of claims for specific types of procedures.
- **CImCount_Provider_BeneID_OtherPhysician_CImDiagnosisCode_3:** This feature involves the count of claims made by a provider for a specific beneficiary with the third diagnosis code and when a different physician is involved. It can capture collaborative patterns that might be indicative of fraud.
- **ChronicCond_Heartfailure:** This feature indicates whether the beneficiary has chronic heart failure. Chronic conditions might be exploited for fraudulent claims, and this information helps identify potential vulnerabilities.



CONCLUSION

- This project has a comprehensive modeling and evaluation process, aiming to develop a robust predictive model for identifying fraudulent activities within healthcare claims.
- Employed machine learning algorithms, Logistic Regression, Decision Tree, and Random Forest.
- Evaluated performance based on key metrics such as accuracy, precision, recall, and F1-score.
- Logistic Regression model performed best, demonstrating superior predictive capabilities.
- Feature importance analysis highlighted factors contributing to the model's decision-making process, providing valuable insights into the intricate patterns within the dataset.