Healthcare Provider Fraud Detection

Done by:

- Sourav Kundu

Coordinated by :
Dr. Kumar Parimal Shrestha
Assistant Professor
IIHMR University, Jaipur

MBA HA-01

About the Dataset

Healthcare provider fraud is one of the biggest problems facing Medicare. It occurs when providers file false or inflated insurance claims to obtain higher reimbursements. The goal is to flag providers who are potentially fraudulent based on their claim behaviour.

Dataset size:

5,410 providers | 138,556 beneficiaries | 40,474 inpatient + 517,737 outpatient claims

Key Features:

Provider claim counts, reimbursement amounts, deductibles, unique patients per provider, claims per patient.

Beneficiary demographics: Age, Gender, Race, State, and 11 chronic disease indicators.

Diagnosis codes (ICD-9) and procedure codes attached to every claim.

Target Variable:

PotentialFraud (binary): whether a provider is flagged as potentially fraudulent (Yes / No).

Objectives:

- To predict potentially fraudulent providers using their claim and patient behaviour.
- To test various supervised ML models and determine the best model for this use case.

Snapshot of the Dataset:

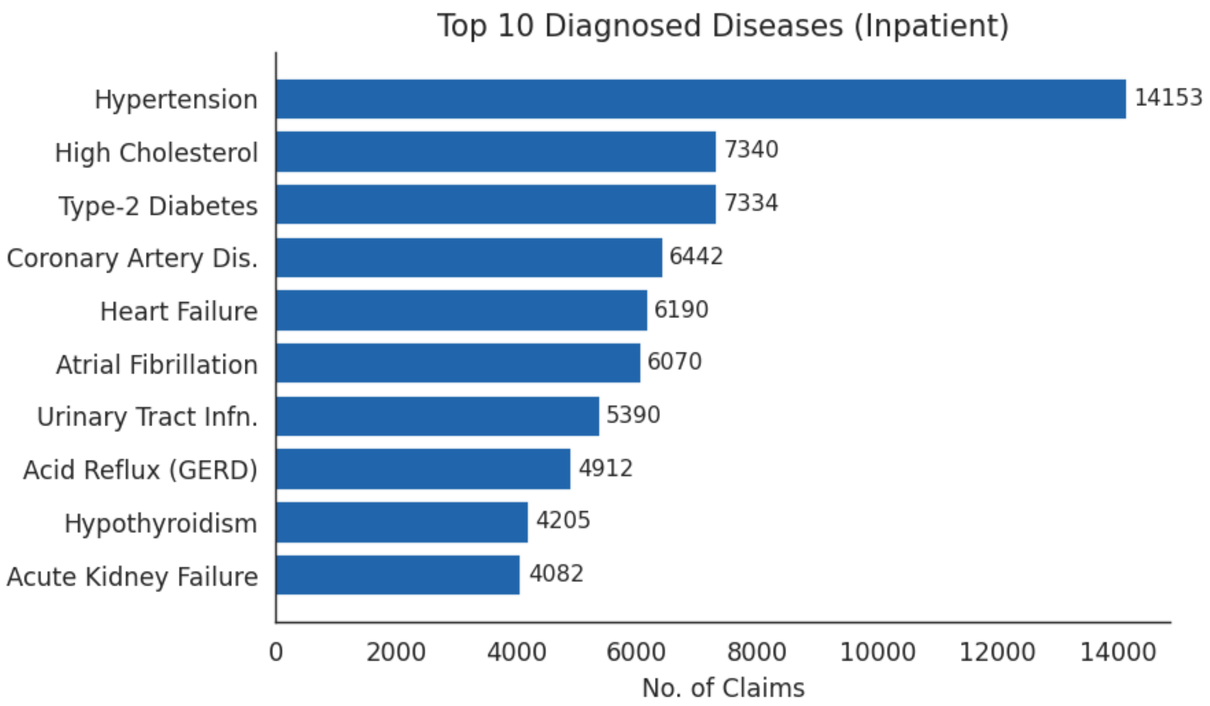
Provider-level features engineered by aggregating all inpatient & outpatient claims:

Provider	TotalClaims	MeanReimb	UniqueBene	TotalReimb	PotentialFraud
PRV51001	25	4186	24	104640	No
PRV51003	132	4588	117	605670	Yes
PRV51004	149	350	138	52170	No
PRV51005	1165	241	495	280910	Yes
PRV51007	72	468	58	33710	No
PRV51008	43	829	36	35630	No

Result (in simple words): We turned millions of raw claims into one simple row per provider, so each doctor/hospital can be compared fairly.

What the Diagnosis Codes Mean

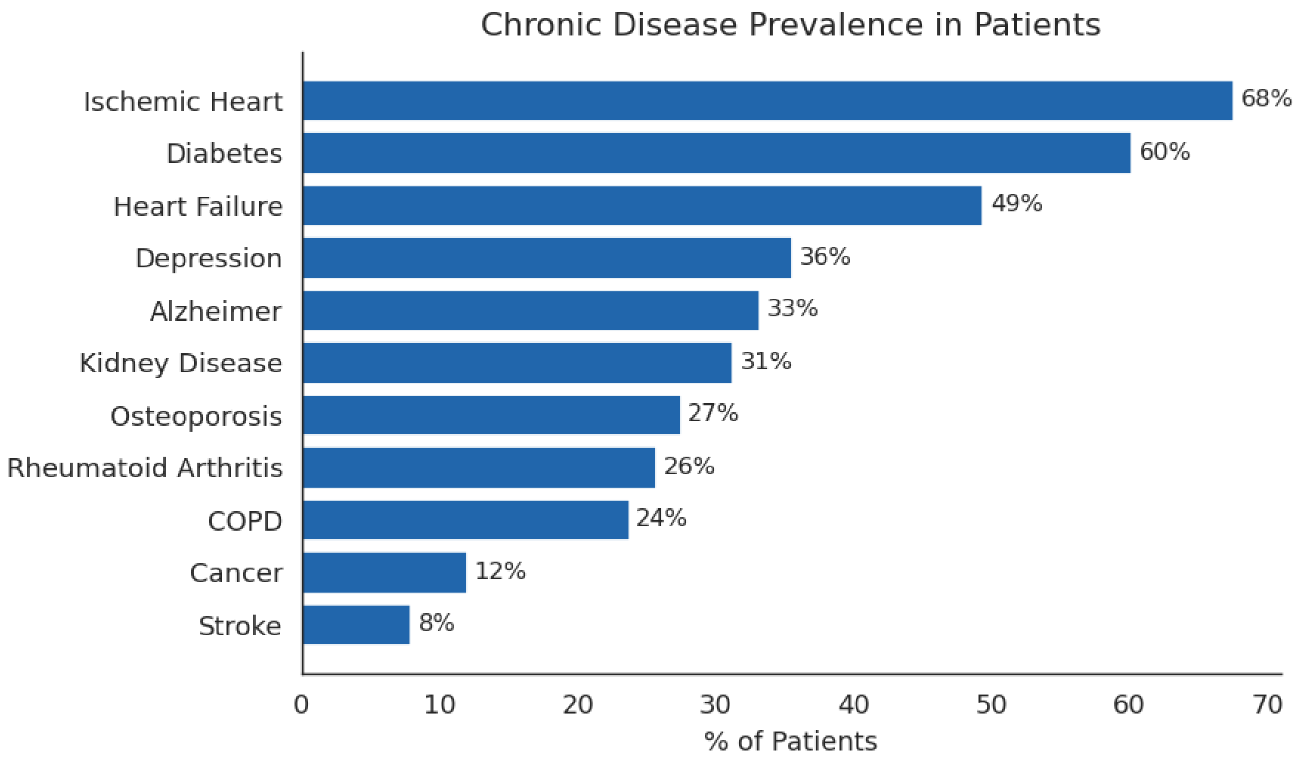
Claims use ICD-9 diagnosis codes. The most frequently diagnosed diseases in this dataset are explained below:



ICD-9	Disease	What it is
4019	Hypertension	Persistently high blood pressure
2724	High Cholesterol	Excess lipids/fats in the blood
25000	Type-2 Diabetes	High blood sugar / insulin resistance
41401	Coronary Artery Disease	Narrowed heart arteries
4280	Heart Failure	Heart can't pump blood efficiently
42731	Atrial Fibrillation	Irregular, rapid heart rhythm
5990	Urinary Tract Infection	Infection in the urinary system
53081	Acid Reflux (GERD)	Stomach acid flowing into food-pipe
2449	Hypothyroidism	Underactive thyroid gland
5849	Acute Kidney Failure	Sudden loss of kidney function

Result (in simple words): Most claims are for common old-age illnesses like high BP, diabetes and heart problems — normal for elderly patients.

Chronic Diseases in Patients



- Most patients are elderly Medicare beneficiaries (avg. age 73 yrs), so chronic disease burden is high.
- Ischemic Heart Disease (68%) and Diabetes (60%) are the most common conditions.
- Heart Failure, Depression and Kidney Disease each affect roughly a third of patients.
- Importantly, the disease mix of fraud vs non-fraud providers is almost identical — fraud cannot be spotted from patient illness alone. It must be found in claim & billing behaviour.

Result (in simple words): Fraud and honest providers treat the same diseases — so illness alone cannot reveal fraud. We must look at billing behaviour.

Data Quality Check

Missing Records

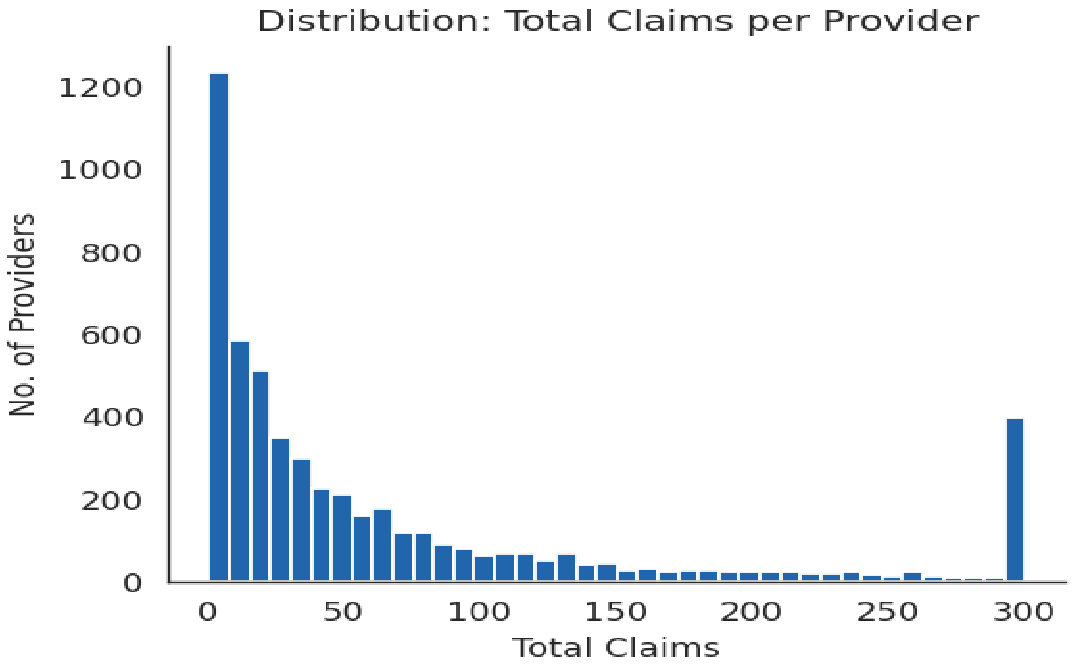
Field	Missing	Why
Date of Death (DOD)	137,135	Patient is still alive
Procedure Codes 2–6	85–100%	Most claims need only 1 procedure
Diagnosis Codes 7–10	Many	Most claims have fewer diagnoses
Operating Physician	16,644	Not every claim needs a surgeon
Other Physician	35,784	Often only one physician involved

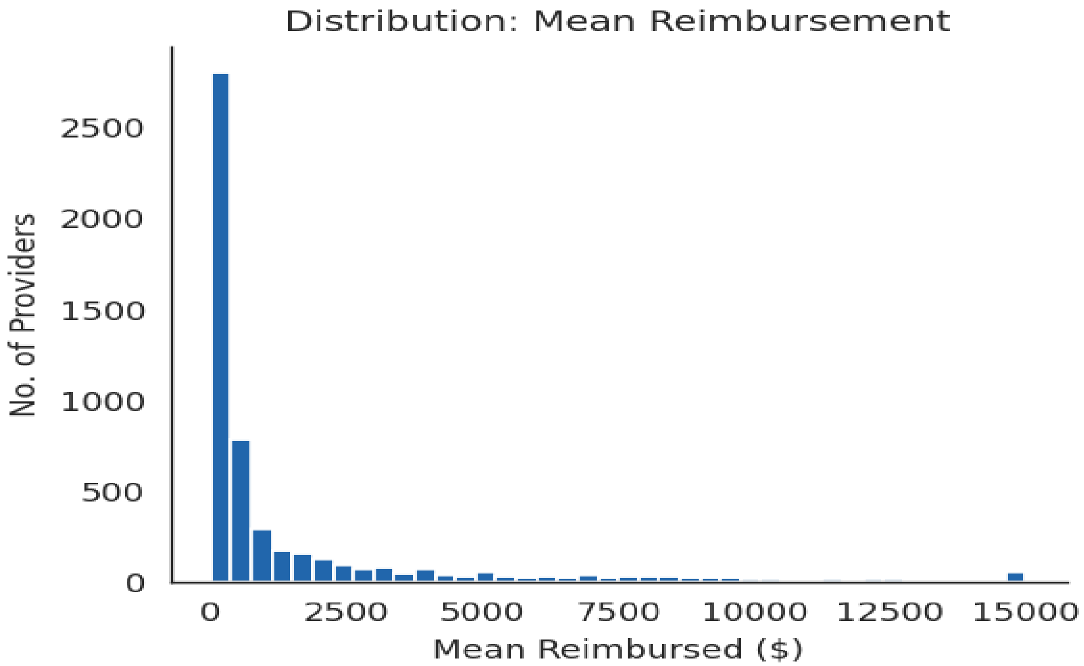
Strategy to deal with missing data:

- Missing values here are structural — they are missing by design, not by error (e.g. DOD is blank because the patient is alive).
- Empty diagnosis / procedure code slots are filled with 0 ("no code present") rather than imputed.
- Missing physician IDs are treated as a separate "none" category.
- Because the data is aggregated to provider level, these structural gaps do not bias the final model.

Result (in simple words): The blank cells are normal (e.g. no death date means patient is alive), so the data is trustworthy and ready to use.

Univariate Analysis

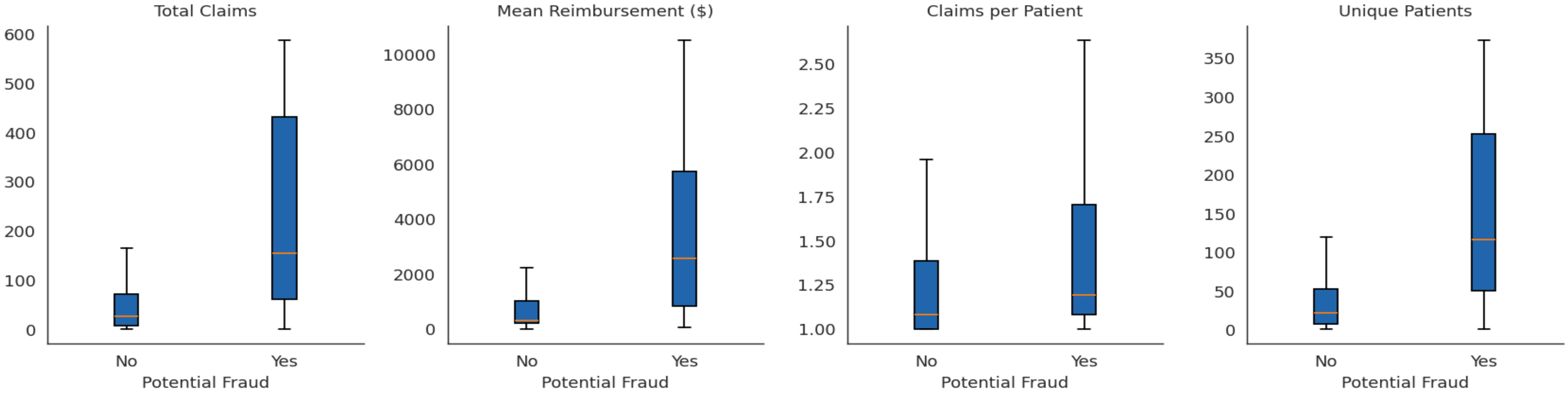




- Both total claims and mean reimbursement per provider are heavily right-skewed — most providers file few low-value claims, but a small tail files very high volumes.
- Only 9.4% of providers are flagged as potential fraud, so the dataset is highly imbalanced — handled using class-weighting in the models.

Result (in simple words): A few providers bill far more than everyone else — these unusual ones are exactly the providers worth checking.

Bivariate Analysis

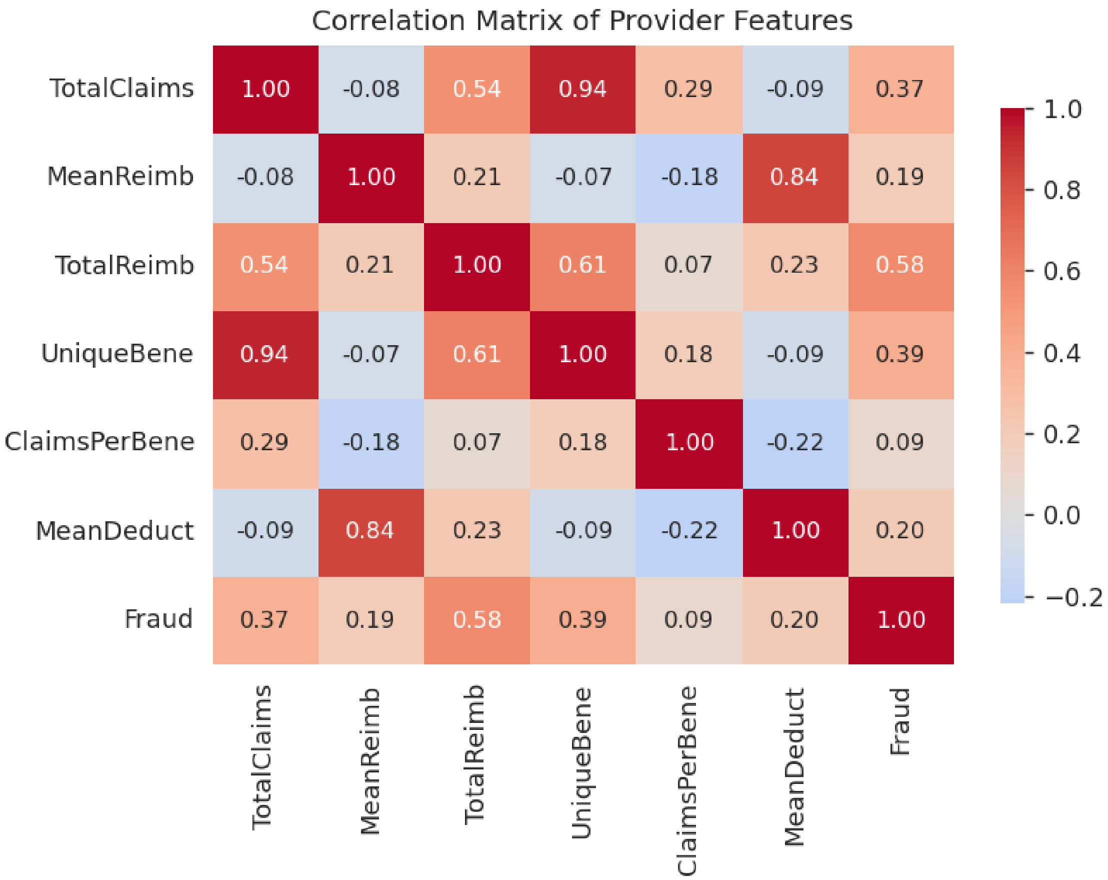


Test Used (Fraud vs feature)	T-Statistic	p-value
T-test for Total Claims	10.8790	0.0000
T-test for Mean Reimbursement	13.1609	0.0000
T-test for Claims per Patient	6.3706	0.0000
T-test for Unique Patients	12.5323	0.0000

- All four claim-behaviour features differ significantly between fraud and non-fraud providers ($p < 0.05$).
- Fraudulent providers file far more claims (avg. 53 vs 10 inpatient claims) and serve more patients.

Result (in simple words): Fraud providers clearly bill more claims and treat more patients than honest ones — a strong warning sign.

Correlation Analysis



- Total Reimbursement has the strongest positive correlation with fraud (r = 0.58).
- Unique Patients (r = 0.39) and Total Claims (r = 0.37) are also moderately correlated with fraud.
- Mean reimbursement per claim is only weakly correlated — fraud is driven by claim volume, not the size of any single claim.

Result (in simple words): The total money claimed by a provider is the biggest hint of fraud — high total payouts deserve a closer look.

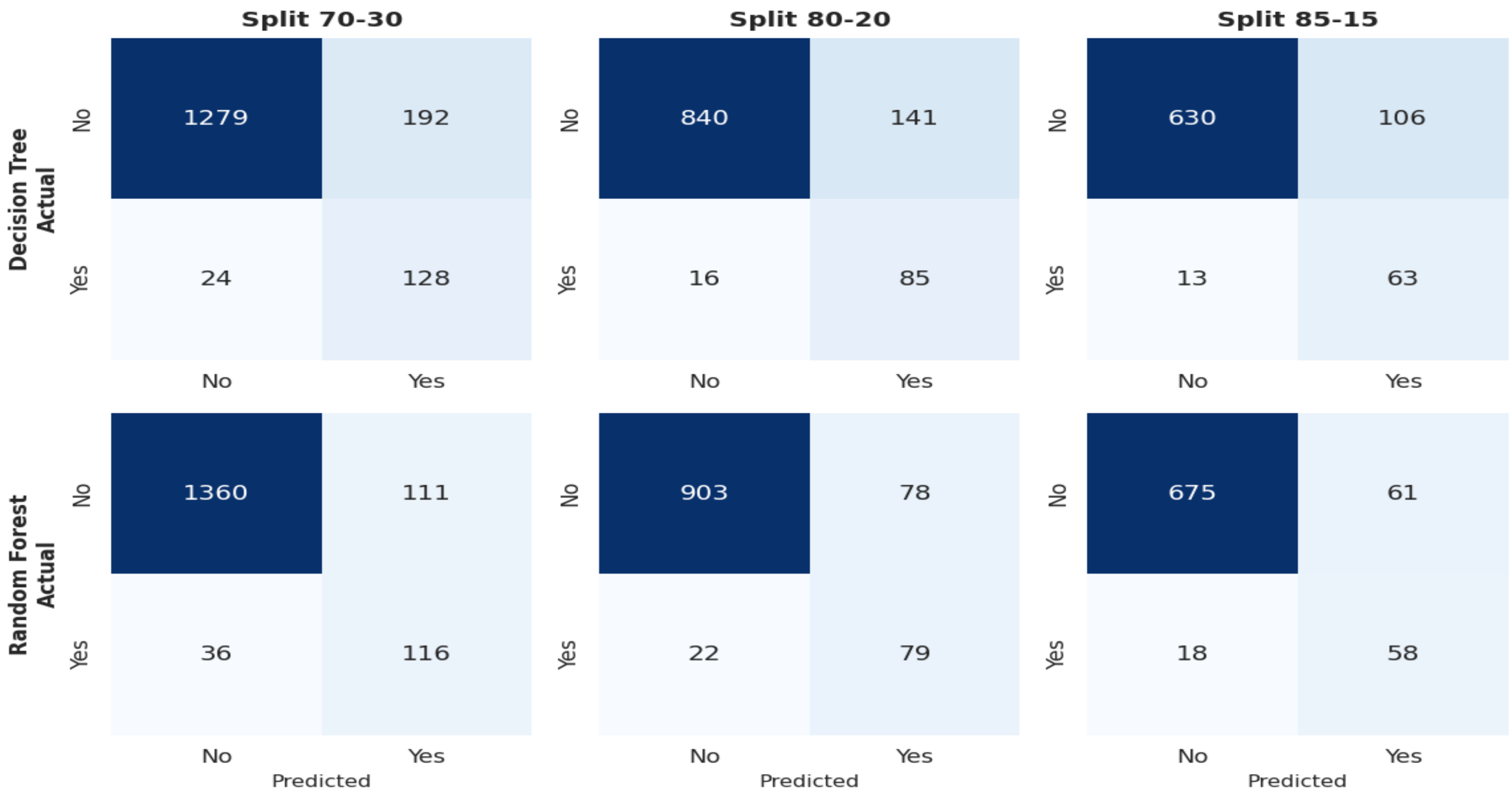
Confusion Matrices

Logistic Regression | SVM | KNN — across three train-test splits

Result (in simple words): These tables show correct vs wrong predictions. KNN catches fraud with the fewest false alarms among these three.

Confusion Matrices

Decision Tree | Random Forest — across three train-test splits



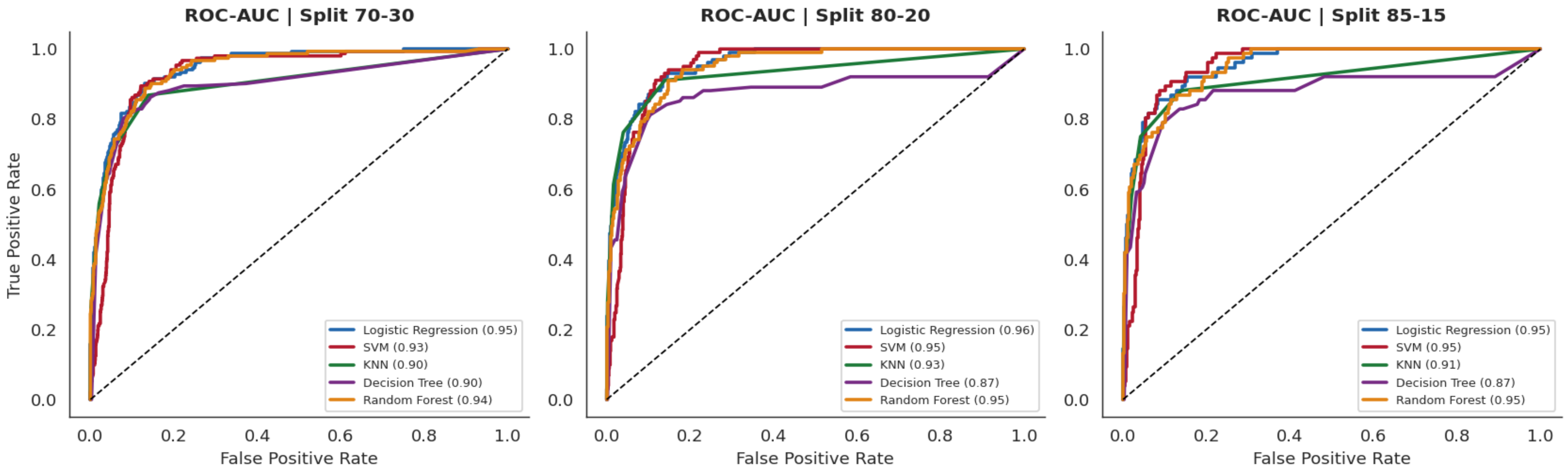
Result (in simple words): Tree-based models also separate fraud well, but they raise more false alarms than KNN does.

Model Performance Comparison

Split	Metrics	Logistic Regression	SVM	KNN	Random Forest	Decision Tree
70-30	Accuracy	0.89	0.87	0.94	0.91	0.87
	Precision	0.46	0.4	0.73	0.51	0.4
	Recall	0.85	0.9	0.56	0.76	0.84
	F1-Score	0.59	0.56	0.63	0.61	0.54
	CV Accuracy	0.89	0.86	0.93	0.91	0.85
80-20	Accuracy	0.89	0.87	0.95	0.91	0.85
	Precision	0.46	0.41	0.79	0.5	0.38
	Recall	0.87	0.92	0.61	0.78	0.84
	F1-Score	0.6	0.56	0.69	0.61	0.52
	CV Accuracy	0.89	0.86	0.93	0.91	0.85
85-15	Accuracy	0.88	0.86	0.94	0.9	0.85
	Precision	0.44	0.39	0.75	0.49	0.37
	Recall	0.86	0.92	0.58	0.76	0.83
	F1-Score	0.58	0.54	0.65	0.59	0.51
	CV Accuracy	0.89	0.86	0.93	0.91	0.85

Result (in simple words): Across every split, KNN gives the best mix of high accuracy and high precision — making it the most reliable choice.

ROC-AUC Characteristics



Result (in simple words): All models score high (0.90–0.96), meaning they tell fraud and honest providers apart well. KNN stays the best all-round pick.

Model of Choice

KNN at 80:20 train-test split

Reasons:

Accuracy = 0.95

Precision = 0.79

Recall = 0.61

F1-Score = 0.69

CV Accuracy = 0.93

Best balance of accuracy & precision

Business takeaway:

The model lets Medicare auditors prioritise the small set of high-risk providers instead of manually reviewing all 5,410 — saving time and recovering fraudulent payouts faster.