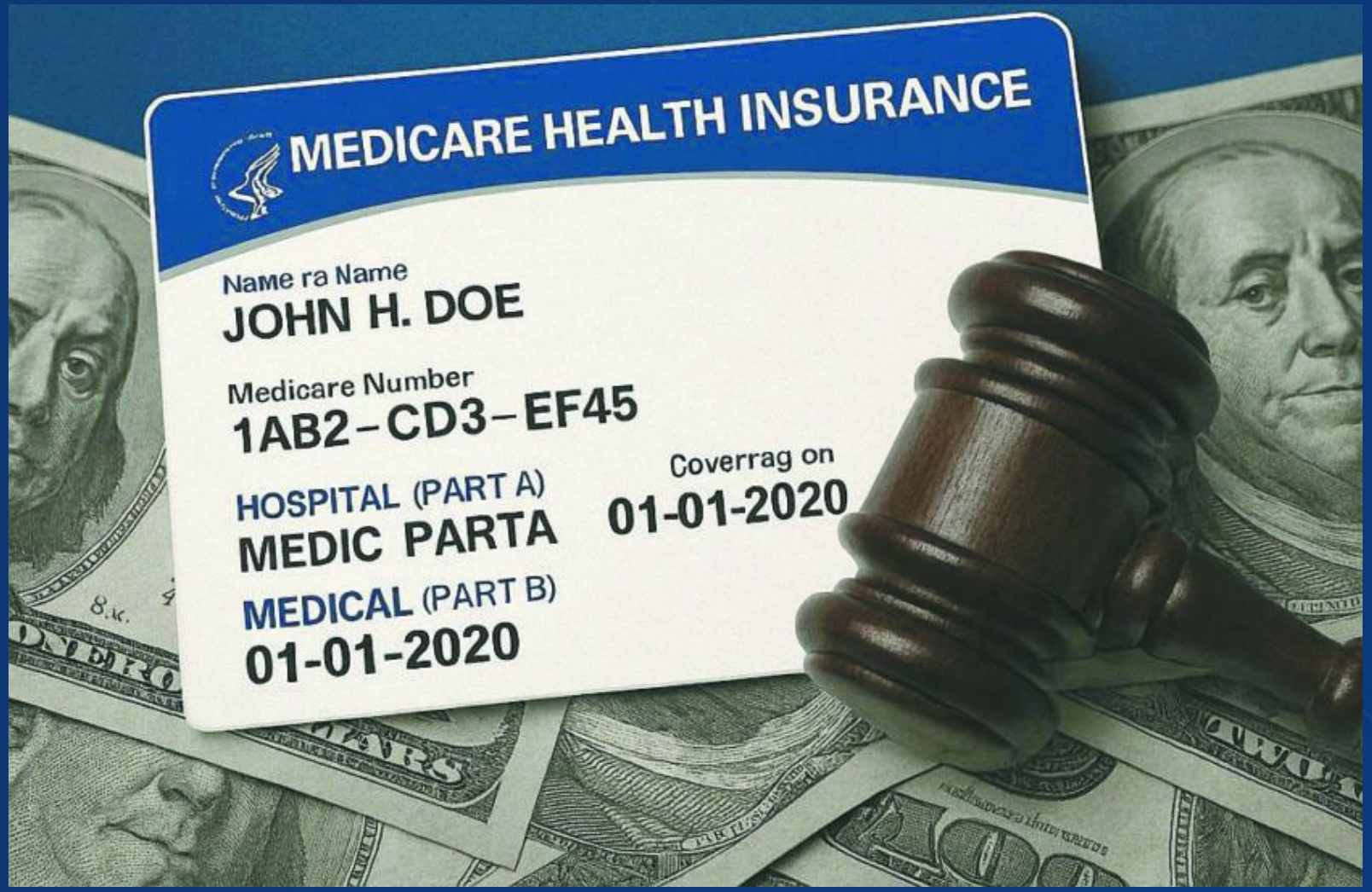
HEALTHCARE PROVIDER FRAUD DETECTION SYSTEM



PRESENTATION ROADMAP

1. BUSINESS CONTEXT

- The $68B Problem
- Fraud Types
- Objective Function

2. DATA ENGINEERING

- Sources & Quality
- Deep Cleaning
- EDA Insights
- Advanced Feature Eng.

3. MODELING

- Handling Imbalance
- Algorithm Selection
- Training Efficiency
- Model Comparison

4. IMPACT

- Error Analysis
- Financial ROI
- Operational Workflow
- Future Roadmap

THE BUSINESS CONTEXT

GOAL: PROACTIVE DETECTION

THE CURRENT STATE: "PAY AND CHASE"

CMS typically pays claims first and investigates later. This reactive model is inefficient.

$68 Billion/Year Estimated lost due to fraud, waste, and abuse.

Resource Constraints
Manual investigators can only review <1% of providers.

DATA INVENTORY

BENEFICIARY DATA

138,556 Records

- Demographics (Age, Race, County)
- Chronic Conditions (11 flags)
- Cost (Deductibles paid)

INPATIENT CLAIMS

40,474 Records

- Admission/Discharge Dates
- Diagnosis Codes (ICD-9)
- DRG Codes
- High Reimbursement Amounts

OUTPATIENT CLAIMS

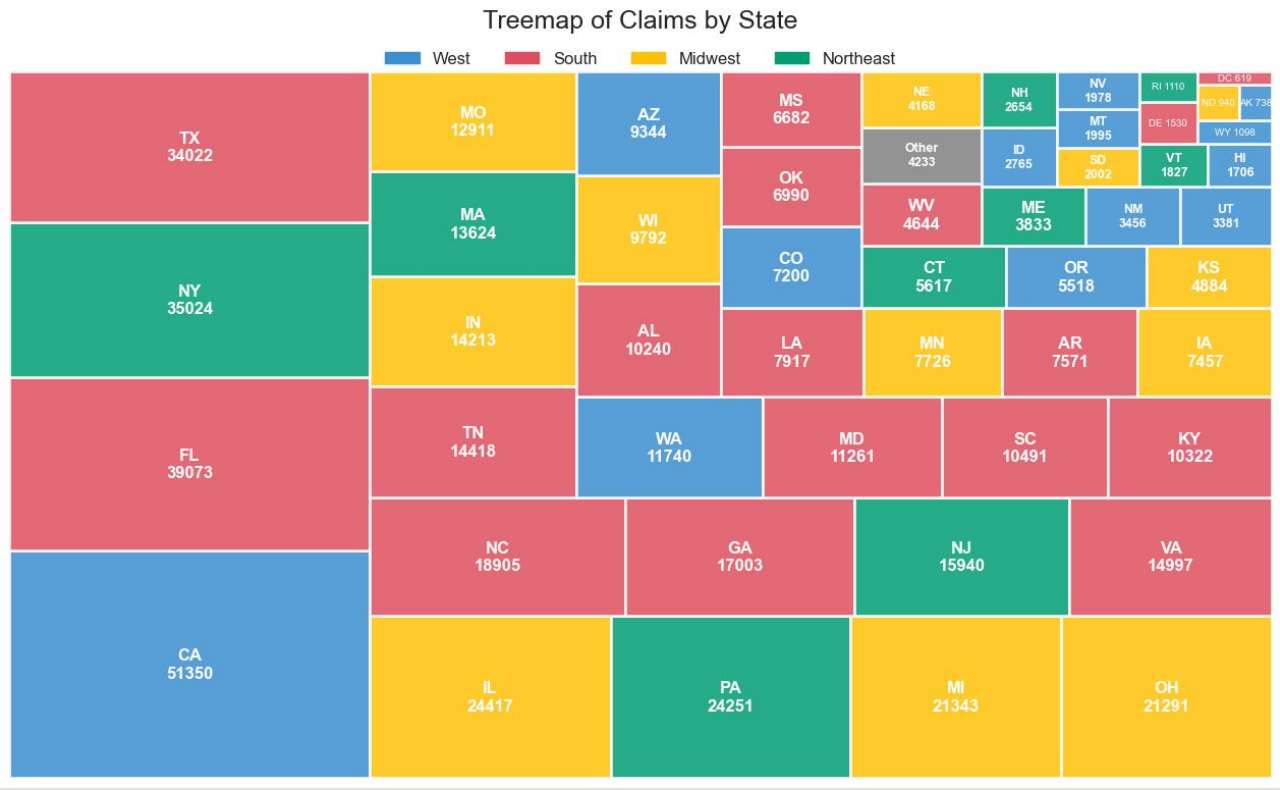
517,737 Records

- Procedure Codes
- Visiting Physicians
- High Volume / Low Cost

DATA QUALITY AUDIT

Feature	Issue Detected	Remediation Strategy
Deductibles	High missingness in Outpatient data.	**Conditional Imputation:** Inpatient defaults to $1,068 (Medicare standard). Outpatient defaults to $0.
Date Logic	32 records where `AdmissionDt` > `DischargeDt`.	**Row Removal:** These illogical records represented <0.01% of data and were dropped to preserve feature integrity.
Categorical	`RenalDiseaseIndicator` coded as 'Y'/'0'.	**Binarization:** Converted to 1/0 boolean flag for model compatibility.
Gender	Coded as 1/2.	**Standardization:** Mapped to 0/1 logic for consistency with other binary flags.

EDA: GEOGRAPHIC DISTRIBUTION



Top States by Claim Volume

CONCENTRATION ANALYSIS

1. California (CA)

2. Florida (FL)

3. NewYork

*Note: While volume is high, fraud *rates* were distributed across states, indicating fraud is not strictly geo-locked.

EDA: THE FINANCIAL FINGERPRINT

KEY INSIGHT: REIMBURSEMENT SKEW

Fraudulent providers have significantly higher average reimbursement claims.

Observation:

The distribution for fraud providers has a "fat tail," meaning they generate extreme outliers (claims > $50k) far more often than legitimate providers.

EDA: PATIENT HEALTH PROFILE

Do fraudulent providers target sicker patients?

RENAL DISEASE
Highest correlation with high-cost claims. Often targeted for expensive dialysis billing.

HEART FAILURE
Common co-morbidity in fraudulent inpatient admissions to justify longer stays.

ALZHEIMER'S
Moderate correlation. Vulnerable population often targeted for phantom billing.

FEATURE ENGINEERING: AGGREGATION STRATEGY

FROM TRANSACTIONS TO BEHAVIOR

Since we predict fraud at the **Provider Level**, we flattened 550k+ claims using aggregation functions.

Raw Feature	Aggregation Function	New Feature Name
InscClaimAmtReimbursed	Mean, Sum, Max	Avg_Reimbursement
ClaimDuration	Mean, Max	Max_LengthOfStay
DiagnosisCode	Count Unique	Diagnosis_Diversity

Many-to-One Transformation

CLINICAL GROUPING LOGIC

Handling the "Curse of Dimensionality" in Diagnosis Codes.

PROBLEM

14,000+ unique ICD-9 Codes.

One-Hot Encoding would create a sparse, massive matrix that overfits easily.

SOLUTION: CCS GROUPING

Mapped codes to **17 Clinical Domains**.

- 401.9 (Hypertension) → **Circulatory**
- 250.0 (Diabetes) → **Metabolic**
- 800.0 (Fracture) → **Trauma**

BEHAVIORAL FEATURES

We engineered 100 features. Here are the top performers:

1. CLAIMDURATION_MAX

The maximum length of stay recorded by a provider. Fraudsters often inflate this to the max allowed limit.

2. DIAGNOSISGROUP_COUNT

The variety of diagnosis codes used. High variance suggests "kitchen sinking" (adding codes to justify cost).

3. INPATIENT_STDDEV

Variability in inpatient claim costs. Fraudsters often have rigid billing patterns (low variance) compared to real life.

THE 9:1 IMBALANCE

WHY THIS MATTERS

Standard algorithms (like Logistic Regression) optimize for Accuracy.

In this dataset, a "dumb" model predicting **No Fraud** for everyone achieves **90.6% Accuracy**.

Result: 0% Fraud Detection.

THE 9:1 IMBALANCE

SOLVING IMBALANCE

We tested 8 different resampling strategies in our pipeline.

1

SMOTE (Oversampling)
Generates synthetic fraud examples by interpolating between existing ones. Good for keeping data volume high.

2

RandomUnderSampler
Deletes non-fraud examples. Fast, but loses information. Worked best for SVM.

3

Class Weights
Mathematical penalization. Telling the model "One fraud error costs as much as 10 non-fraud errors."

4

SMOTE + Tomek
Hybrid approach. Oversamples fraud, then cleans up noisy boundary points.

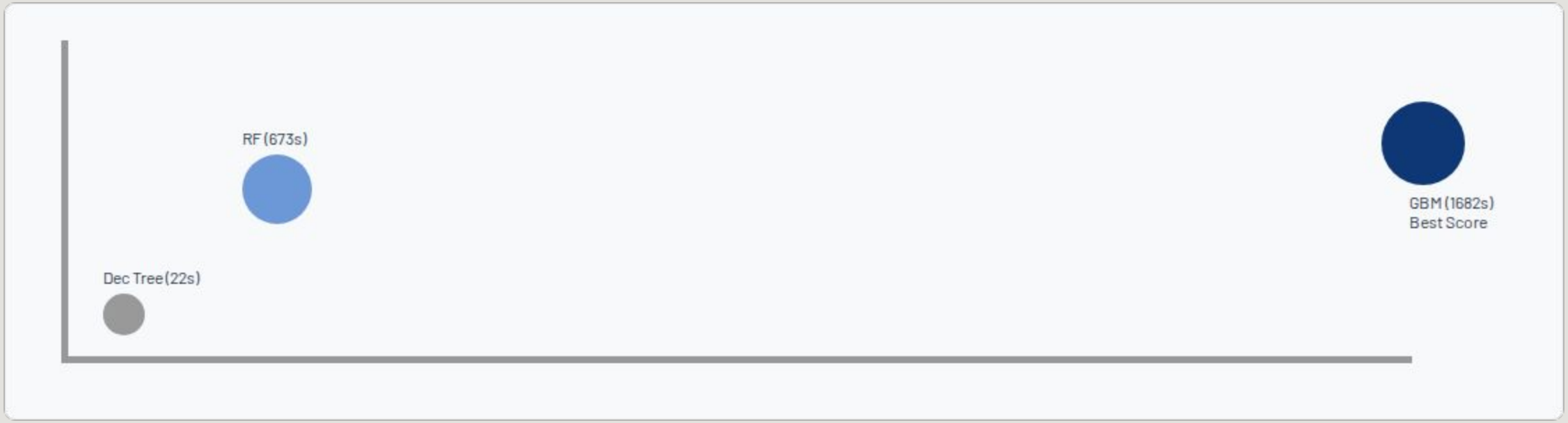
PIPELINE ARCHITECTURE

Crucial Detail: Resampling was applied *inside* the Cross-Validation loop to prevent data leakage.

CANDIDATE ALGORITHMS

Algorithm	Pros	Cons
Logistic Regression	Interpretability, Speed	Fails to capture non-linear fraud patterns.
Support Vector Machine	High Recall (Catch rate)	Computationally expensive, Low Precision.
Random Forest	Robust to outliers, Parallelizable	Struggled with Recall on the minority class here.
Gradient Boosting	**Best performance**, Iterative correction	Slower training time, Requires tuning.

TRAINING EFFICIENCY VS. PERFORMANCE



Tradeoff: Gradient Boosting took ~28 minutes to train (vs 11 mins for RF), but delivered significantly better Recall.

HYPERPARAMETER TUNING

We used GridSearchCV with 5-fold CV to find the optimal settings.

GRADIENT BOOSTING GRID

- **n_estimators:** [100, 200, 300]
- **learning_rate:** [0.01, 0.1, 0.2]
- **max_depth:** [3, 5, 7]
- **subsample:** [0.8, 1.0]
- **min_sample_leaf:** [0.8, 1.0]
- **min_sample_split:** [2, 5]

Winner: 100, 0.01, 3, 0.8, 1.0, 2

ADA Boosting

n_estimators: [50, 100, 200]
learning_rate: [0.01, 0.1, 0.5. 1.0]
algorithm: ["SAMME", "SAMME.R"]
max_depth: [1,3,5]

Winner: 50, 0.01, SAMME, 1

SVM (RandomUnderSampler)

- **C:** [1, 10, 100]
- **kernel:** ['rbf']
- **Gamma:** ["scale" , 0.01, 0.001]
- **Class Weight:** ["Balanced"]

Winner: 1, RBF, Gamma 0.001, Balanced

SVM (NoSampling_ClassWeight)

C: [1, 10, 100]
kernel: ['rbf']
Gamma: ["scale" , 0.01, 0.001]
Class Weight: ["Balanced"]

Winner: 1, RBF, Gamma 0.001, Balanced

EVALUATION METRICS

RECALL (SENSITIVITY)

"Of all the actual fraud, how much did we find?"

Target: MAXIMIZE

PRECISION

"Of all the providers we flagged, how many were actually fraud?"

Target: BALANCE

PR-AUC

Precision-Recall Area Under Curve.

Target: Primary Metric for Imbalance

MODEL RESULTS: TEST SET

Model	Recall	Precision	F1-Score	ROC-AUC
Gradient Boosting (SMOTE)	**82.00%**	**53.00%**	**0.65**	**0.94**
SVM (RandomUnderSampler)	61.00%	55.00%	0.58	0.91
AdaBoost(No Resampling)	69.00%	56.00%	0.61	0.85

DEEP DIVE: SVM RESULTS

THE HIGH RECALL TRAP

SVM with RandomUnderSampler achieved an incredible **92.7%**

Recall.

However, it achieved this by being "trigger happy".

✕**Precision: 32%**

This means for every 3 alerts, 2 are false alarms.

Operally unsustainable for manual review teams.

287

False Positives

141

True Positives

CHAMPION: GRADIENT BOOSTING

THE GOLDILOCKS ZONE

GBM offers the optimal trade-off.

- **82.00% Recall:** Misses very few frauds.
- **53% Precision:** 1 in 2 alerts is real.
- **0.65 F1-Score:** Highest harmonic mean.

0.94

ROC-AUC Score

DRIVERS OF FRAUD (GBM)

Rank	Feature	Importance Score	Interpretation
1	**ClaimDurationDays_max**	0.812	Length of stay is the #1 signal.
2	**DiagnosisGroup_Unique**	0.188	Complexity of cases billed.
3	Inpatient_Std	0.044	Consistency of billing patterns.
4	Coverage_PartB_Std	0.039	Patient insurance variance.

FINAL CONFUSION MATRIX

1360 True Negative (Clean)	**111** False Positive (False Alarm)
26 False Negative (Missed)	**126** True Positive (Caught)

ERROR ANALYSIS: FALSE POSITIVES

THE "TEACHING HOSPITAL" EFFECT

We flagged 111 legitimate providers. Upon review, many were large urban teaching hospitals.

WHY?

- They see the sickest patients (High Diagnosis Count).
- They perform complex surgeries (High Reimbursement).
- They look "anomalous" compared to a small clinic, but are legitimate.

Fix: Normalize features by Bed Count.

ERROR ANALYSIS: FALSE NEGATIVES

THE "FLYING UNDER RADAR" EFFECT

We missed 26 fraud cases. These providers had average billing amounts and normal durations.

WHY?

- They likely commit fraud via **Kickbacks** or **Referral Schemes**.
- These schemes do not show up in transactional metadata (length of stay, cost).

Fix: Network Graph Analysis (Provider-Provider links).

ROI ANALYSIS

COST AVOIDANCE

Detecting 77% of fraud in the test set = Potential savings of millions in improper payments.

EFFICIENCY LIFT

84% reduction in manual review volume. Investigators focus only on high-probability targets.

RISK MITIGATION

Moving from "Pay and Chase" to Pre-payment flags reduces the recovery burden.

THANK YOU

QUESTIONS?