FINTAP

PRESENTATION

AI-Powered Fraud Detection in Auto Insurance:
Predictive Modeling for Smarter Claims Management

PREPARED BY:

TEAM QUANTACORES

TEAM ID - TEAM(MB6)4_4

PROBLEM STATEMENT

Auto insurance fraud is a persistent and costly issue for insurance companies, leading to significant financial losses and impacting honest policyholders. Traditional methods of detecting fraud are manual, time-consuming, and often miss subtle patterns. With the increase in claim volume and complexity, there is a critical need for intelligent systems that can accurately identify suspicious claims. This project aims to leverage machine learning models to analyze claim data, predict the likelihood of fraud, and support faster, smarter decision-making in claims management.

Objectives:

- Increasing claim volume makes pattern detection difficult
- Need for AI-powered models to predict and flag suspicious claims

Financial management is crucial for the stability and growth of any organization.

DATASET BREAKDOWN

The dataset includes four CSV files and one data dictionary text file. CSV1 is used for training, CSV2 for testing (with fraud labels removed), and CSV3 for final predictions. Each file contains detailed insurance claim information across 50+ features related to policy, vehicle, customer, and accident details.

CSV 1 consist of model data with the fraud index

CSV 2 consist of user data with the fraud index for testing

CSV 3 consist of user data but the fraud index is not there we have to generate that using models

CSV 4 consist of claim id and fraud index of csv 3 and we have added a coloumn called fraud probablity (likelihoood)

OUR UNIQUE APPROACH

Our approach involves training machine learning models using CSV1 with known fraud labels, then testing them on CSV2 after removing the Fraud_Ind column to simulate real-world unseen data. We predict fraud likelihood as a percentage rather than just a Yes/No output. This enables better prioritization of suspicious claims. Finally, we apply the trained models on CSV3 and compare results for similarity, risk, and confidence across models.

Model Diversity

Train models using XGBoost, Random Forest, Decision Tree, and AdaBoost on labeled data (CSV1) to capture different patterns of fraud

Realistic Evaluation

Drop Fraud_Ind from CSV2, predict fraud using trained models, and evaluate predicted likelihood for each claim.

Likelihood Scoring

Generate a fraud probability score (0 to 100%), allowing flexible risk-based claim assessment rather than binary outputs

SELECTED PARAMETERS

✅ Selected Parameters (and Why They Were Chosen):

We carefully selected features that are most relevant to fraud detection based on domain knowledge, data quality, and correlation with the target:

Parameter	Reason for Selection
`Accident_Severity`	Indicates claim seriousness; often inflated in fraud.
`Total_Claim`	High claim amounts can signal potential fraud.
`Accident_Hour`	Odd-time accidents (late night) are more suspicious.
`Police_Report`	Fraud cases often skip police involvement.
`Age_Insured`	Younger or older profiles may reflect risk patterns.
`Injury_Claim`	Frequently exaggerated in fraudulent cases.
`Vehicle_Claim`	Manipulated repair claims are a common fraud area.
`Witnesses`	Lack of witnesses may indicate staged events.
`Property_Damage`	Used to detect inflated or fake damage reports.

DROPPED PARAMETERS

❌ Dropped Parameters (and Why They Were Excluded):

We excluded several parameters due to missing values, irrelevance to fraud behavior, or redundanc

Parameter	Reason for Dropping
`Vehicle_Registration`	Unique for each car, no pattern or predictive value.
`DL_Expiry_Date`	Mostly missing or inconsistent across datasets.
`Garage_Location`	High missing rate and low correlation to fraud.
`Customer_Life_Value1`	More useful for business profitability than fraud risk.
`Check_Point`	Internal audit field, not useful for prediction.
`Hobbies`	Not relevant for detecting fraud in claims.
`Occupation`	Too diverse and inconsistent to model clearly.

MODELS USED

Random Forest (RF)

- Ensemble of decision trees with bagging
- Handles overfitting and missing values well
- Robust and interpretable

XGBoost (Extreme Gradient Boosting)

- Boosted tree model that minimizes errors iteratively
- Highly accurate, handles complex relationships
- Efficient with large datasets and missing values

Decision Tree (DT)

- Simple tree-based model
- Easy to interpret with visual flows
- Baseline for understanding feature importance

AdaBoost (Adaptive Boosting)

- Combines weak learners to create a strong model
- Focuses on previously misclassified samples
- Useful for improving simpler models like DTs

XG BOOSTING

XGBoost is used because it delivers high accuracy, handles missing data efficiently, reduces overfitting through regularization, and is optimized for speed, making it ideal for complex fraud detection tasks

Confusion Matrix

Predicted: Y Predicted: N

Actual: Y (Fraud) 5233 (True Pos.) 87 (False Neg.)

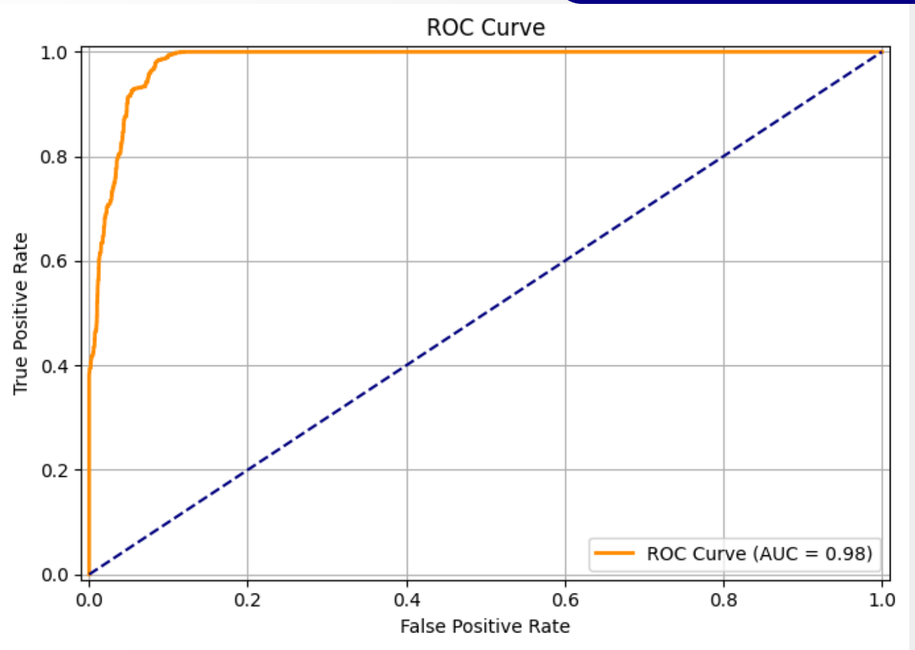
Actual: N (Not Fraud) 1293 (False Pos.) 13387 (True Neg.)

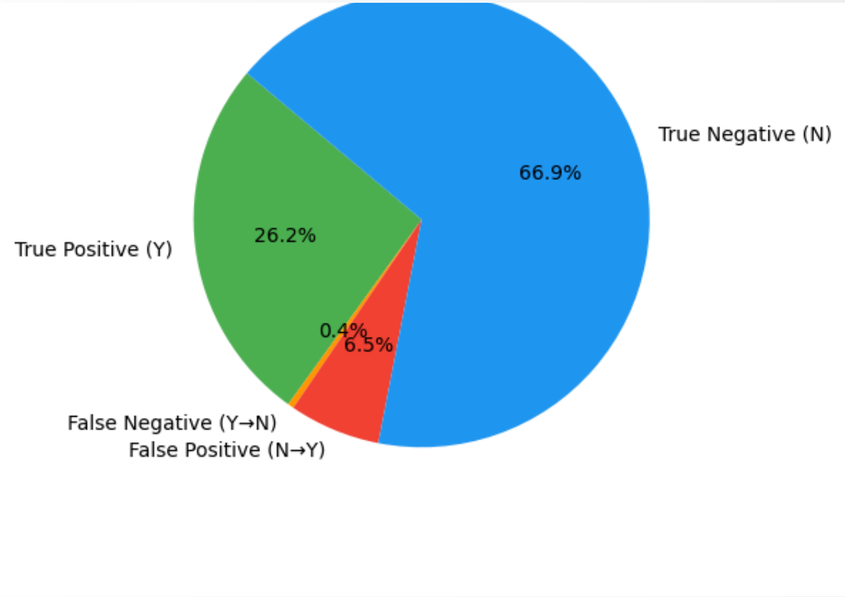
📊 Key Metrics

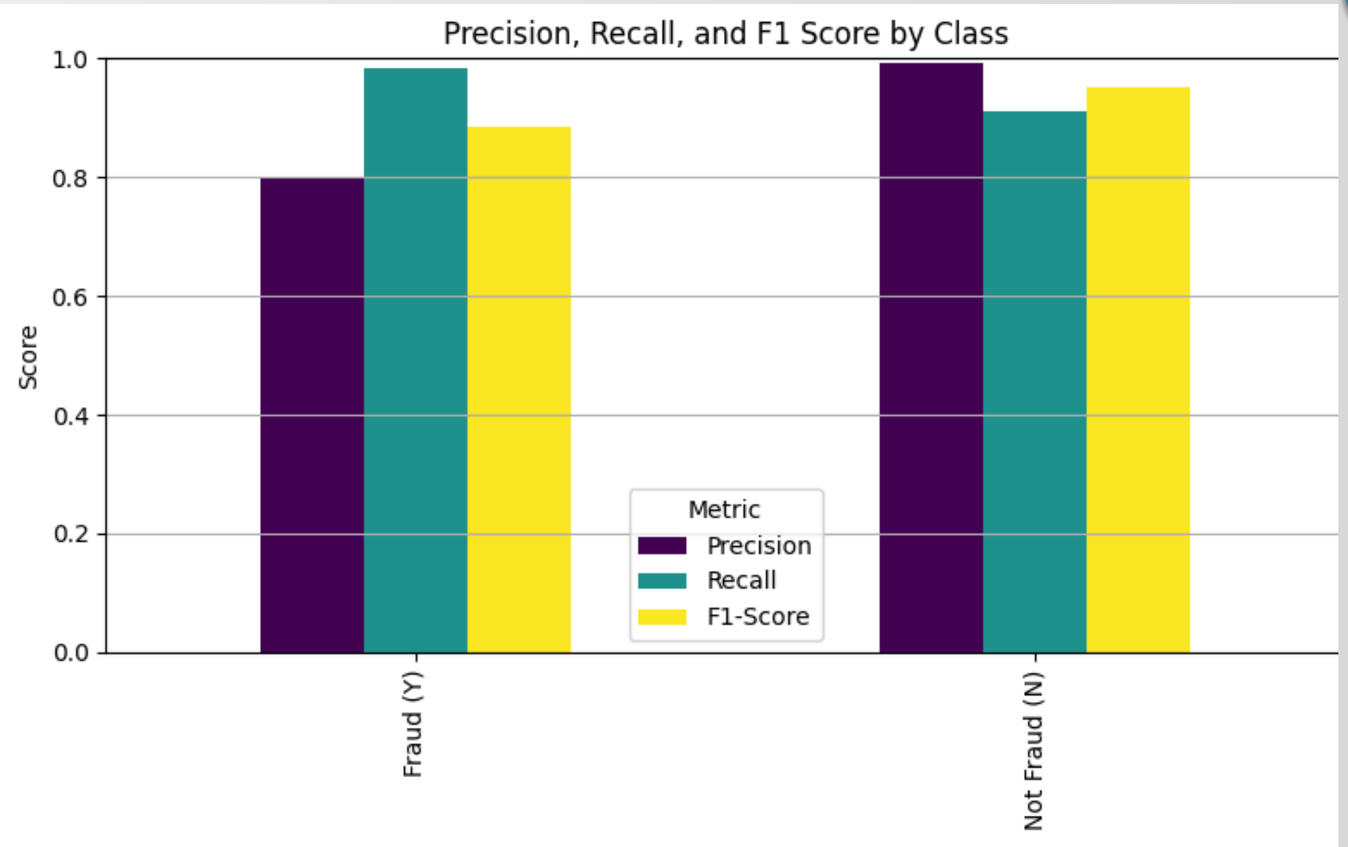
Metric	Fraud (Y)	Not Fraud (N)
Precision	80.19%	99.35%
Recall	98.36%	91.19%
F1 Score	88.35%	95.10%

✅ Overall Accuracy: 93.10%

METRICES


ROC Curve
True Positive Rate
False Positive Rate
ROC Curve (AUC = 0.98)
0.0
0.2
0.4
0.6
0.8
1.0


True Negative (N)
66.9%
True Positive (Y)
26.2%
0.4%
6.5%
False Negative (Y→N)
False Positive (N→Y)


Precision, Recall, and F1 Score by Class
Score
Metric
Precision
Recall
F1-Score
Fraud (Y)
Not Fraud (N)
0.0
0.2
0.4
0.6
0.8
1.0

RANDOM FOREST & DECISION TREE

Random Forest (RF) is used for its robustness, accuracy, and ability to handle noisy data, while Decision Tree (DT) is chosen for its simplicity, interpretability, and fast computation.

🔍 Confusion Matrix

	Predicted: Y	Predicted: N
Actual: Y (Fraud)	5320 (True Pos.)	0 (False Neg.)
Actual: N (Not Fraud)	0 (False Pos.)	14680 (True Neg.)

📊 Key Metrics

Metric	Fraud (Y)	Not Fraud (N)
Precision	100.00%	100.00%
Recall	100.00%	100.00%
F1 Score	100.00%	100.00%

✅ Overall Accuracy: 100.00%

🔍 Confusion Matrix

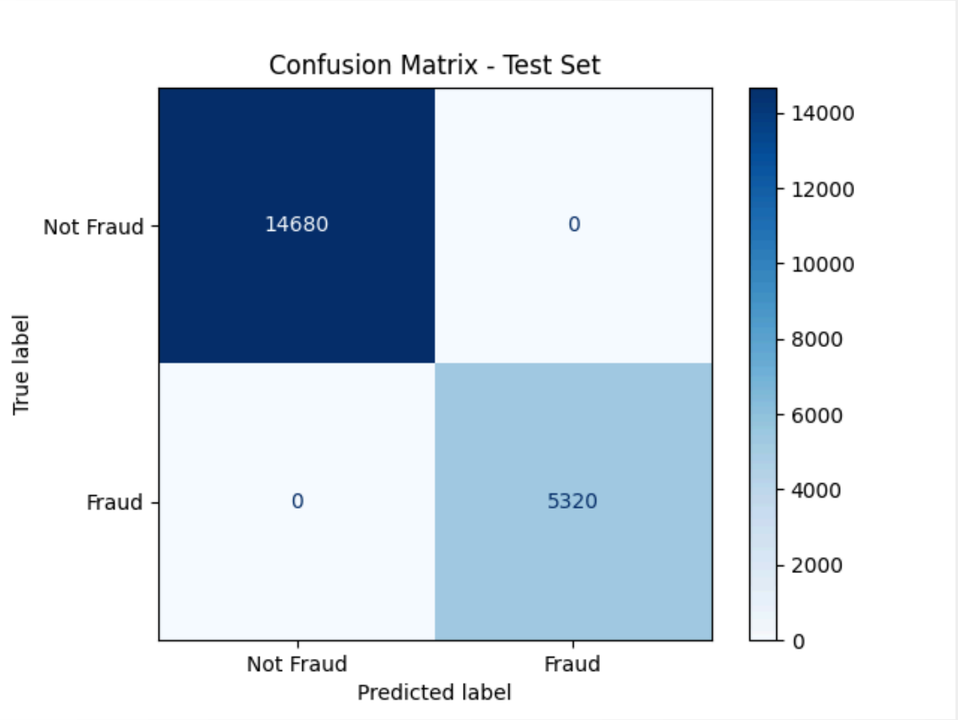
	Predicted: Y	Predicted: N
Actual: Y (Fraud)	5320 (True Pos.)	0 (False Neg.)
Actual: N (Not Fraud)	0 (False Pos.)	14680 (True Neg.)

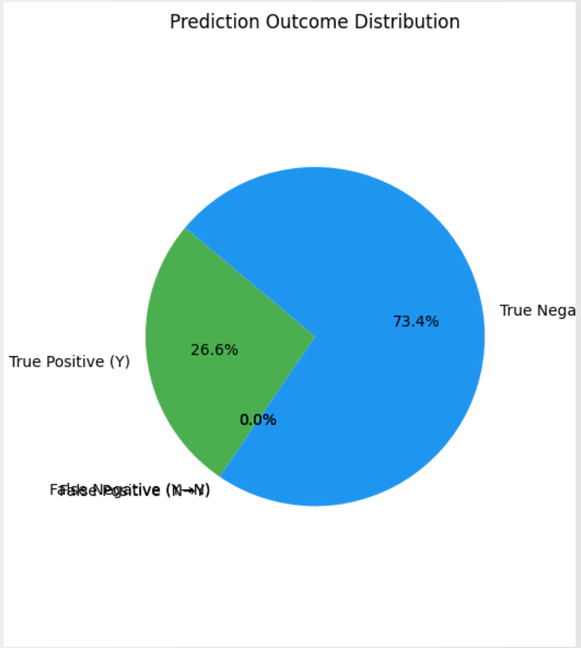
📊 Key Metrics

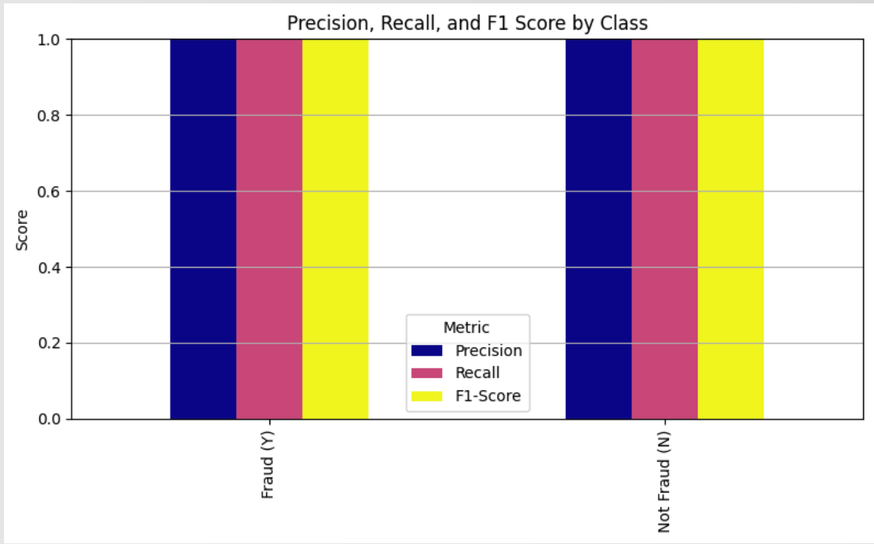
Metric	Fraud (Y)	Not Fraud (N)
Precision	100.00%	100.00%
Recall	100.00%	100.00%
F1 Score	100.00%	100.00%

✅ Overall Accuracy: 100.00%

METRICES


Confusion Matrix - Test Set
Not Fraud
Fraud
True label
14680
0
0
5320
Not Fraud
Fraud
Predicted label
14000
12000
10000
8000
6000
4000
2000
0


Prediction Outcome Distribution
True Positive (Y)
26.6%
73.4%
True Nega
0.0%


Precision, Recall, and F1 Score by Class
Score
1.0
0.8
0.6
0.4
0.2
0.0
Metric
Precision
Recall
F1-Score
Fraud (Y)
Not Fraud (N)

ADAPTIVE BOOSTING

- **Focuses on misclassified samples to improve model accuracy over iterations**
- **Combines weak learners into a strong, more accurate predictive model**

🔍 Confusion Matrix

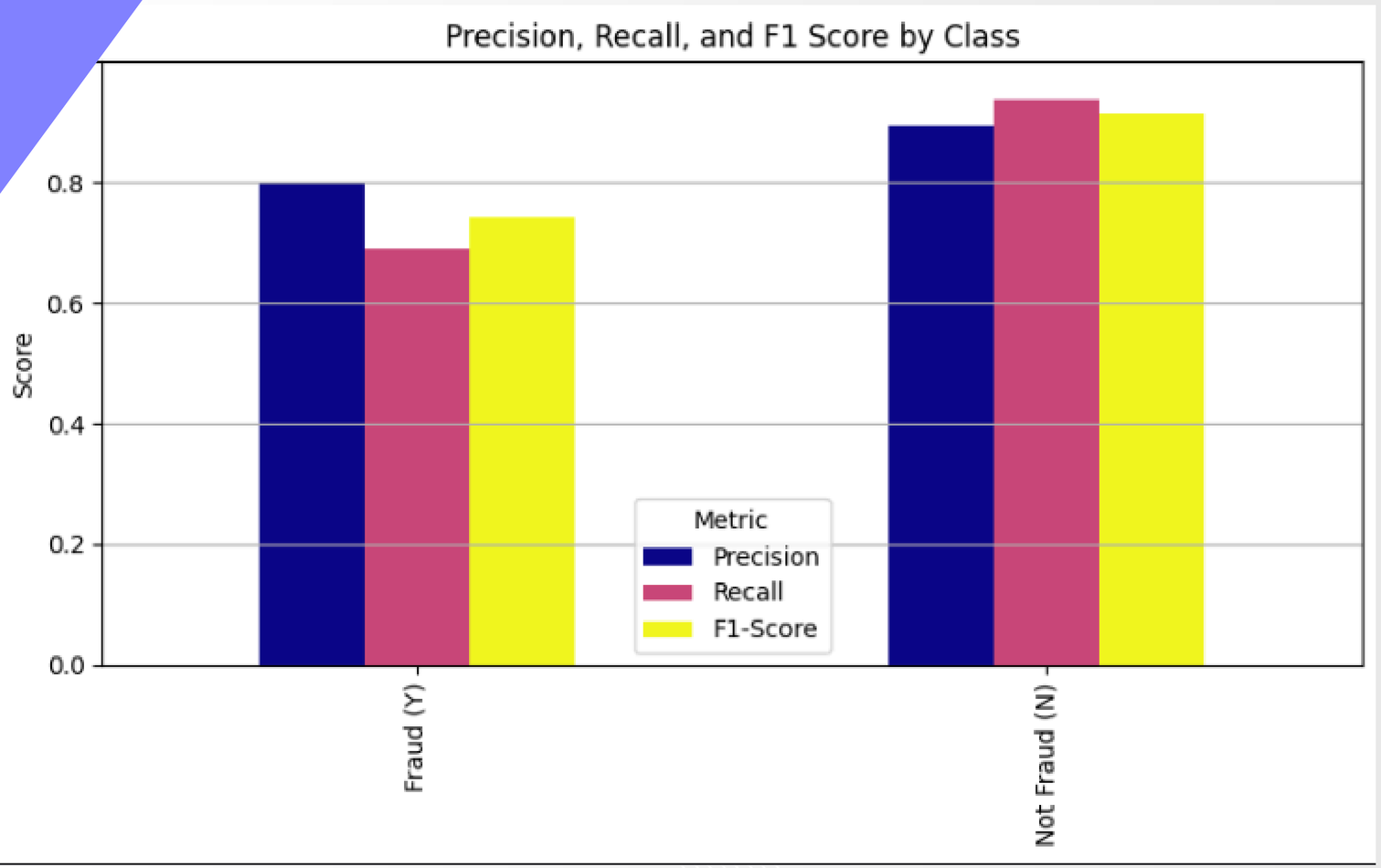
	Predicted: Y	Predicted: N
Actual: Y (Fraud)	3678 (True Pos.)	1642 (False Neg.)
Actual: N (Not Fraud)	910 (False Pos.)	13770 (True Neg.)

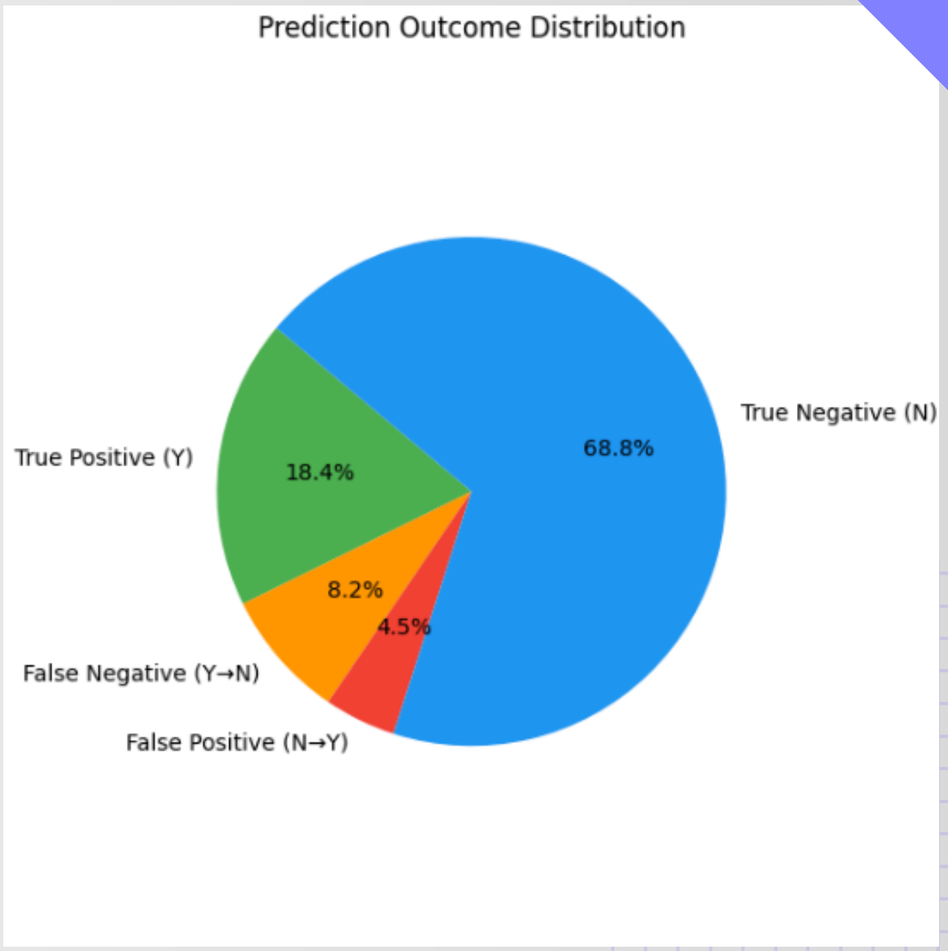
📊 Key Metrics

Metric	Fraud (Y)	Not Fraud (N)
Precision	80.17%	89.35%
Recall	69.14%	93.80%
F1 Score	74.24%	91.52%

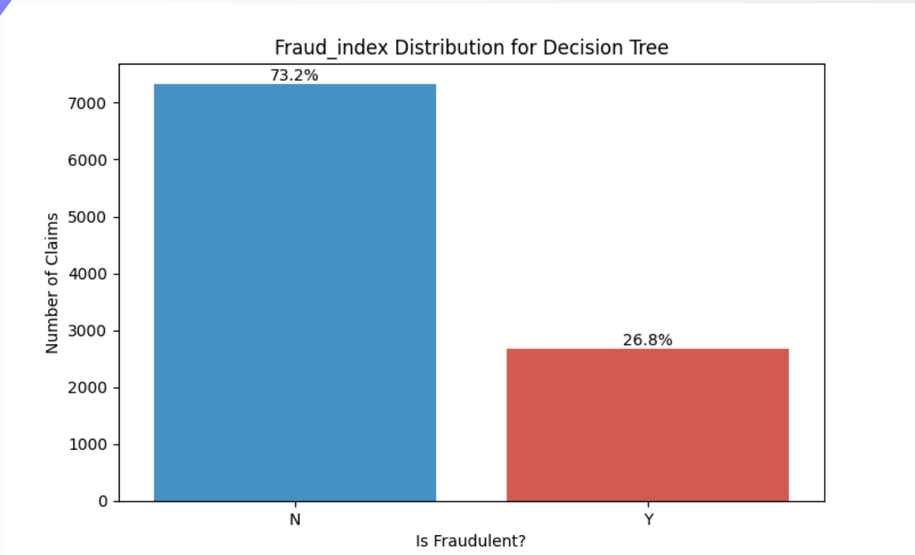
✅ Overall Accuracy: 87.24%

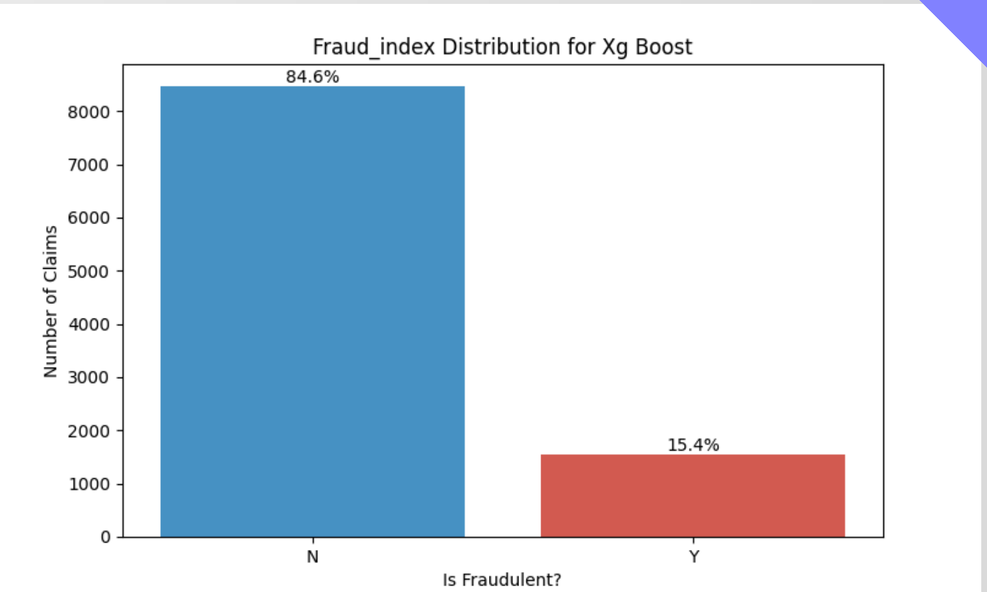
ADAPTIVE BOOSTING

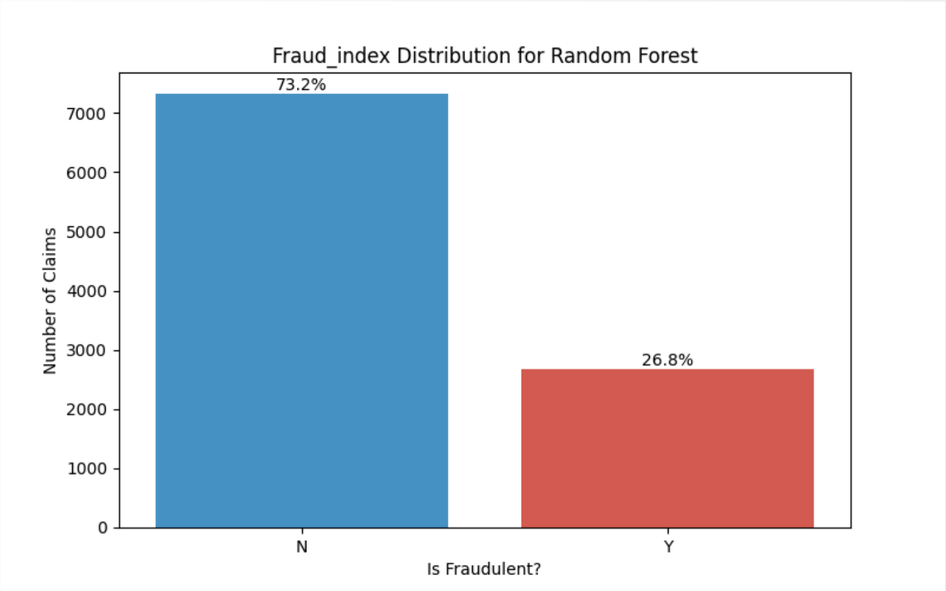


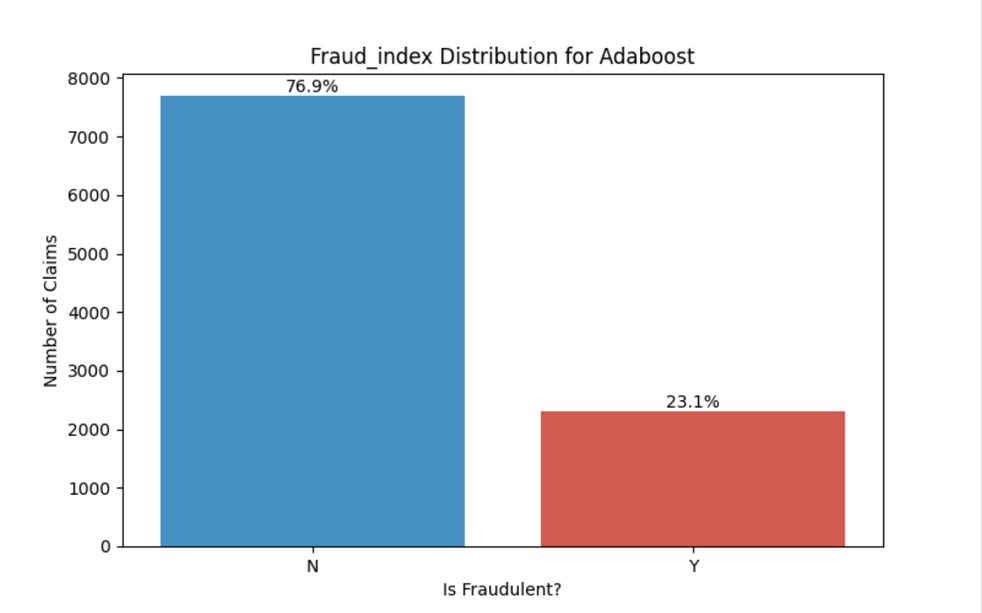


EVALUATION OF UNSEEN DATA





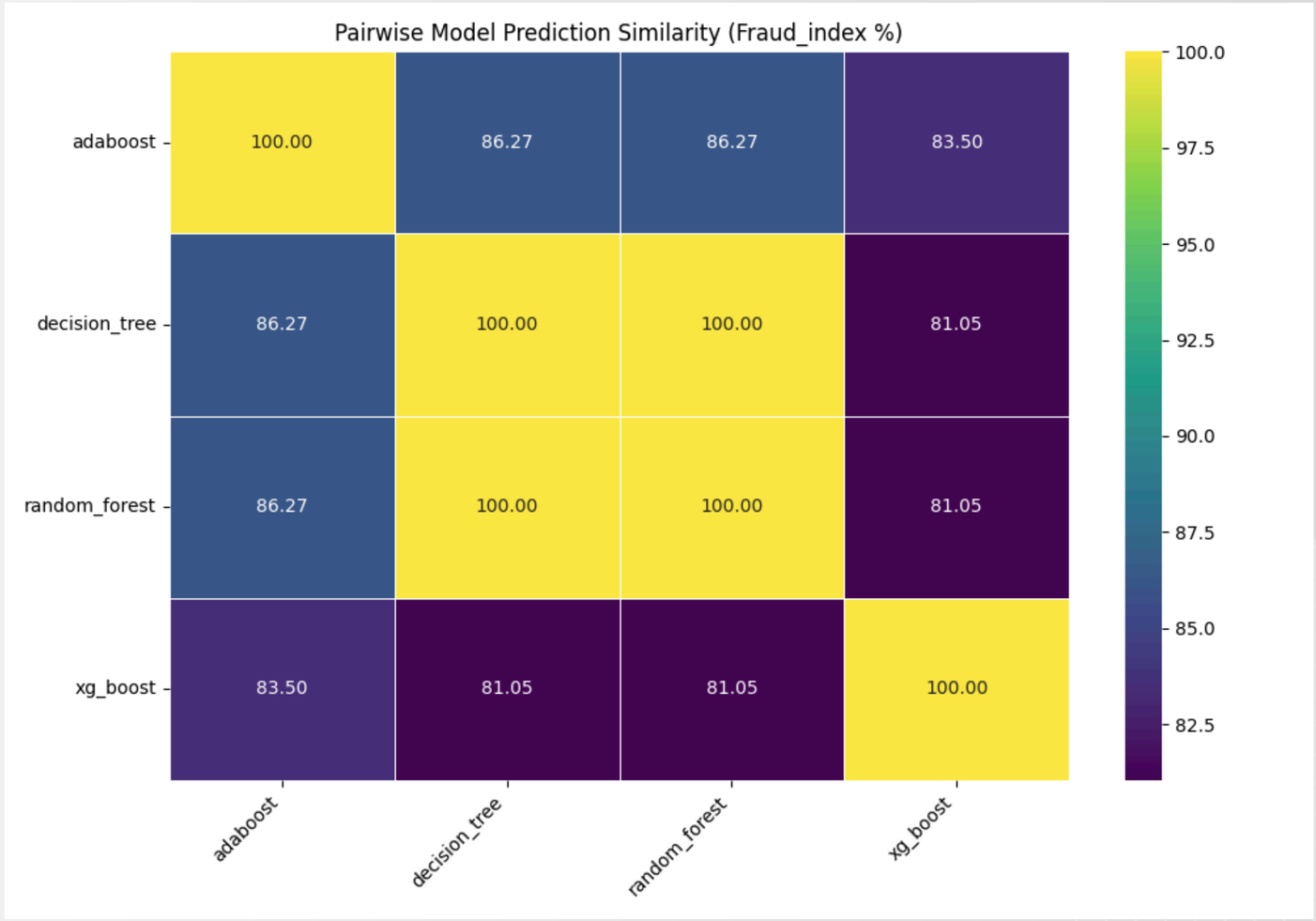




SIMILARITY REPORT

Each model's predictions on CSV3 were compared for similarity in fraud likelihood scores. This helped identify consistent high-risk claims and showed agreement levels across models, improving trust in final fraud detection outcomes.

```
--- Analysis Within Each File ---

--- Adaboost ---
Distribution of Fraud_index:
Fraud_index
N    76.87%
Y    23.13%

--- Decision Tree ---
Distribution of Fraud_index:
Fraud_index
N    73.20%
Y    26.80%

--- Random Forest ---
Distribution of Fraud_index:
Fraud_index
N    73.20%
Y    26.80%

--- Xg Boost ---
Distribution of Fraud_index:
Fraud_index
N    84.57%
Y    15.43%

--- Pairwise Similarity Analysis Between Files ---
Comparing the 'Fraud_index' predictions for each pair of models.

Similarity Matrix (%):
               adaboost  decision_tree  random_forest  xg_boost
adaboost         100.00          86.27          86.27     83.50
decision_tree     86.27         100.00         100.00     81.05
random_forest     86.27         100.00         100.00     81.05
xg_boost          83.50          81.05          81.05    100.00
```



WHY LIKELIHOOD MATTERS

- Prioritizes Investigations: Helps insurers focus on high-risk claims first, improving fraud detection efficiency and resource allocation
- Reduces False Alarms: Avoids wrongly flagging honest claims by using confidence scores instead of binary outputs
- Supports Smart Decisions: Enables customizable risk thresholds for auto-approval, review, or rejection of claims

FUTURE SCOPES

The system can evolve by integrating real-time data, NLP for claim descriptions, and deploying interactive dashboards for insurers.

- **NLP Integration: Analyze claim narratives using Natural Language Processing to detect fraud-related language patterns**
- **Real-Time Deployment: Enable instant fraud scoring during claim submission via APIs or mobile apps**
- **Dashboard Interface: Build visual dashboards for adjusters to monitor, compare, and manage fraud predictions effectively**

CONCLUSION

01 **Effective Prediction: Accurately detects fraud using trained ML models.**

02 **Risk-Based Output: Provides fraud likelihood scores for better decisions.**

03 **Scalable Approach: Easily applies to real-world insurance claim systems.**

Our model accurately detects fraudulent claims using ML, enhances decision-making, and provides scaOur model effectively detects fraudulent auto insurance claims using machine learning algorithms. By providing fraud likelihood scores instead of binary outputs, it enhances decision-making, prioritizes high-risk cases, and offers a scalable, intelligent solution for modern claims management systems.

Ask ChatGPT

lable, risk-based fra

THANK YOU

FOR YOUR ATTENTION

LEARNATHON 4.0