# Fraudulent Claim Detection

Summary & Recommendations

Submitted by:
Ketan Chincholikar, Milan Banura, Lakhan Varshney

# Problem Statement

- Global Insure faces considerable financial losses dues to fraudulent claims
- Current process is time-consuming and inefficient
- Fraudulent claims are detected after the claim payment
- Global Insure wants to improve the fraud detection process using data-driven insights to classify claims as fraudulent or legitimate and minimize financial losses and optimize the overall claims handling process

# Business Objective

- Build a predictive model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles
- Use Logistic Regression and Random Forest models to identify the key features associated with fraudulent claims

# Data Handling

- Loaded and inspected data
- Handled null values
- Removed missing and illogical values (e.g., negative losses)
- Dropped high-cardinality identifiers
- Categorical encoding and feature scaling applied using StandardScaler()
- Class imbalance resolved using RandomOverSampler

# Exploratory Data Analysis – Key Insights

- Performed univariate and bivariate analysis
  - People with hobbies like chess and cross-fit show high fraud risk
  - Insured occupation like transport-moving and exec-managerial have more fraud percentage
  - Single Vehicle Collision  and Multi-vehicle Collision have more chance of fraud
  - MD, PhD , JD (people with very high education level) are likely to fraud
  - Incidents occurring in Ohio has highest fraud likelihood
  - Auto make – Audi, Mercedes, Ford lead in fraud
- Correlation analysis
  - Total claim amount and vehicle claim, property claim, injury claim
  - Age and month as customer
- Class Imbalance
  - Fraud reported in ~25% of cases

# Feature Engineering

- Created binary indicators for risky segments
- Combined rare categories to reduce sparsity
- Applied one-hot encoding for categorical features
- Used RandomOverSampler to balance classes

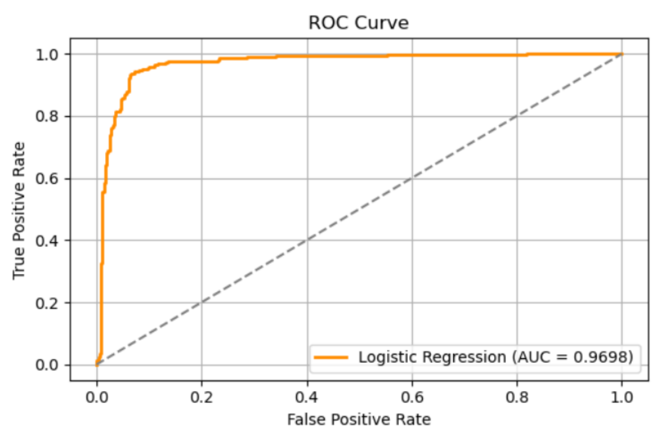
# Model Building

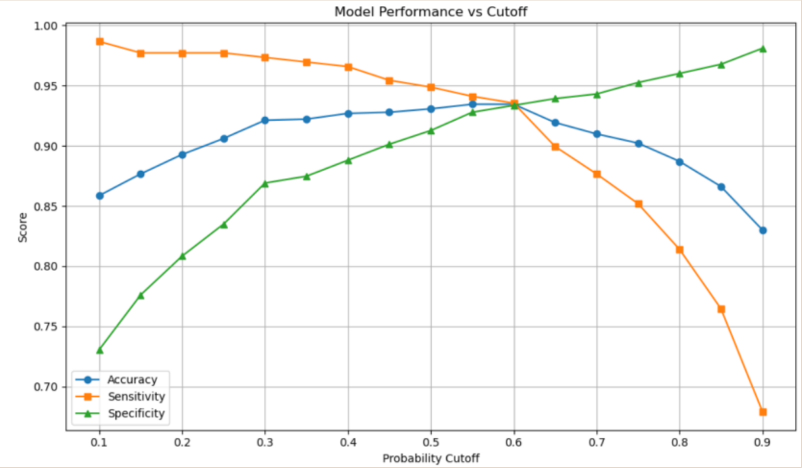
## Logistic Regression

| Cutoff | Accuracy | Sensitivity | Specificity | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| 0.6 | 93.5% | 0.94 | 0.93 | 0.93 | 0.94 | 0.93 |

## Random Forest

| Hyperparameter Tuning | Accuracy | Sensitivity | Specificity | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| Yes | 94.87% | 0.99 | 0.90 | 0.91 | 0.99 | 0.95 |



# Model Evaluation

## Logistic Regression Model

| Cutoff | Accuracy | Sensitivity | Specificity | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| 0.6 | 79% | 0.54 | 0.87 | 0.58 | 0.54 | 0.56 |

## Random Forest Model

| Hyperparameter Tuning | Accuracy | Sensitivity | Specificity | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| Yes | 83.67% | 0.77 | 0.86 | 0.64 | 0.77 | 0.70 |

# Recommendations & Business Implications

- Fraudulent claims exhibit identifiable patterns linked to incident severity, customer hobbies, customer occupation types, and education making predictive modelling highly viable
- Random Forest model is recommended for operational deployment due to its superior accuracy, balanced precision and recall, and robustness after hyperparameter tuning
- Logistic Regression model showed slightly lower performance; it remains valuable for its interpretability
- Prioritize investigation for high-risk segments (e.g., major damage, unusual hobbies)
- Enhance claims triage using predictive scores

# Thank you

Ketan Chincholikar

Milan Banura

Lakhan Varshney