



2025

*Fraudulent  
Claim Detection*

VAISHALI MAKWANA

PRAVEENKUMAR PERIYASWAMY



# TABLE OF CONTENT

1

PROBLEM  
STATEMENT

2

KEY  
OBJECTIVES

3

DATA PROCESS

4

MODELING  
APPROACH

5

INVESTIGATIVE  
TAKEAWAYS

6

IMPLICATIONS &  
FUTURE  
RECOMMENDATIONS



# PROBLEM STATEMENT

Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant portion of these claims are fraudulent, leading to substantial financial losses.

The current approach to fraud detection primarily involves manual inspections, which are:

- Time-consuming
- Inefficient
- Often too late — fraud is detected after the claim has been paid



## KEY OBJECTIVES

Global Insure seeks to leverage data-driven models to identify potential fraud early in the claims approval process, enabling the company to:

- Reduce financial losses
- Speed up claim processing for legitimate claims
- Enhance operational efficiency

# DATA UNDERSTANDING & MODELING APPROACH

- Data includes customer demographics, policy details, incident and claim info.
- Target variable: Whether the claim is Fraudulent (1) or Legitimate (0).

ths_as_customer	age	policy_number	policy_bind_date	policy_state	policy_csl	policy_deductable	policy_annual_premium	umbrella_limit	insured_zip	insured_s
328	48	521585	2014-10-17	OH	250/500	1000	1,406.91	0	466132	MA
228	42	342868	2006-06-27	IN	250/500	2000	1,197.22	5000000	468176	MA
134	29	687698	2000-09-06	OH	100/300	2000	1,413.14	5000000	430632	FEMA
256	41	227811	1990-05-25	IL	250/500	2000	1,415.74	6000000	608117	FEMA
228	44	367455	2014-06-06	IL	500/1000	1000	1,583.91	6000000	610706	MA

```
# Inspect the features in the dataset  
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 1000 entries, 0 to 999  
Data columns (total 40 columns):
```

# DATA CLEANING

## Handling Null Values

- Detection: Identified columns with missing data using `df.isnull()`. Found nulls in `collision_type`, `authorities_contacted`, `property_damage`, `police_report_available`, and a completely empty column `_c39`.
- Treatment:
  - Dropped `_c39` (all values null).
  - Replaced missing values with reasonable defaults:
    - `collision_type`: 'General Collision'
    - `authorities_contacted`: 'None'
    - `property_damage` & `police_report_available`: 'NO'

## Handling Redundant or Invalid Data

- Redundant Columns:
- Removed columns with only one unique value or all unique values (e.g., `policy_number`, `incident_location`, `insured_zip`).
- Invalid Values:
- Detected negative values in `umbrella_limit` — dropped affected row.
- Empty Columns:
- Confirmed no columns were completely empty after initial drop.
- Low Predictive Power Columns:
- Dropped high-uniqueness columns (e.g., `policy_number`) unlikely to help in pattern recognition.

## Fixing Data Types

- Identified and converted date-related columns (e.g., `policy_bind_date`) to appropriate datetime types.
- Categorical features were converted to category data type to optimize memory and allow efficient analysis.
- Binary columns (e.g., `insured_sex`, `fraud_reported`) were handled distinctly for value count validation.

# FEATURE ENGINEERING

- **Initial Features:** ~50 mixed data types.
- **Technique Used:**
  - L1-penalized logistic regression (SAGA) with SelectFromModel (mean-threshold) for feature selection.
- **Result:**
  - Reduced to ~20 top predictors based on their importance scores.
  - The selected features are now ready for use in downstream models.

**This feature selection process streamlines the model, focusing on the most important predictors for better performance.**

## MODELING APPROACH

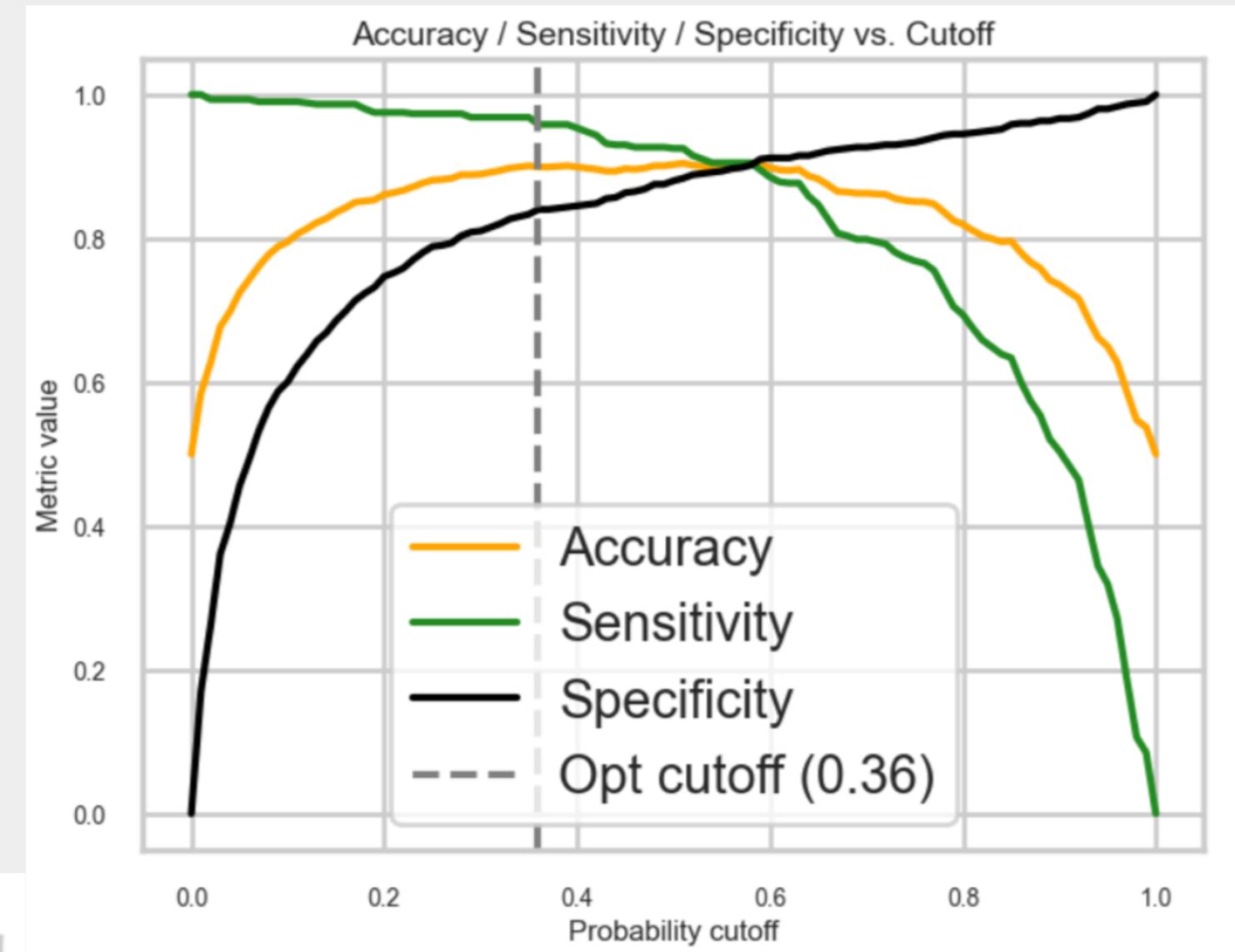
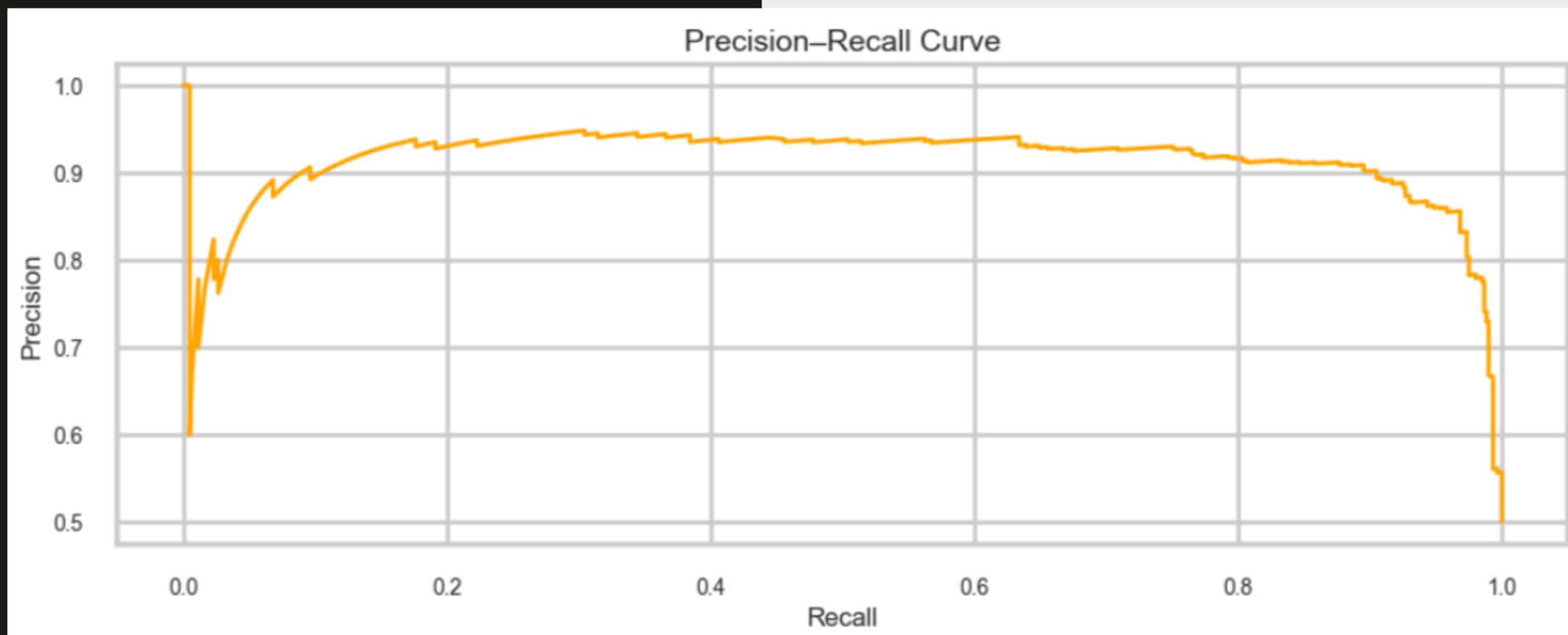
- **Baseline Model:** Logistic Regression with optimized decision cutoff (accuracy, sensitivity, specificity).
- **Advanced Model:** Random Forest with feature selection (top ~12 features) and 5-fold GridSearchCV for hyperparameter tuning.
- **Hyperparameters Tuned:** n\_estimators, max\_depth, min\_samples\_split/leaf, max\_features.

 **Key Point:** By extracting feature importances and retraining on the top features, the random forest model was optimized for performance. Hyperparameter tuning further ensured that the model was well-adjusted to the data.

# LOGISTIC REGRESSION MODEL

- **The Logistic Regression model** performed well, with strong accuracy (85%) and solid recall (82%) during training, indicating good generalization to the positive class.
- **Precision** (80%) reflects a reasonable trade-off, ensuring that false positives were kept in check.
- **Validation** results showed slight performance degradation, particularly in Recall (78%), but overall, the model remains robust with an accuracy of 83%.
- **The optimal cutoff of ~0.40** helped to improve performance across key metrics in both training and validation.

# LOGISTIC REGRESSION MODEL



# RANDOM FOREST MODEL



## Best Hyperparameters (via 5-fold GridSearchCV)

- n\_estimators: 200
- max\_depth: 20
- min\_samples\_split: 5
- min\_samples\_leaf: 2
- max\_features: 'sqrt'

## Performance Metrics

### Training Set:

- Accuracy: **0.90**
- Recall: **0.88**
- Precision: **0.89**
- F1 Score: **0.89**

### Cross-Validation (5-fold):

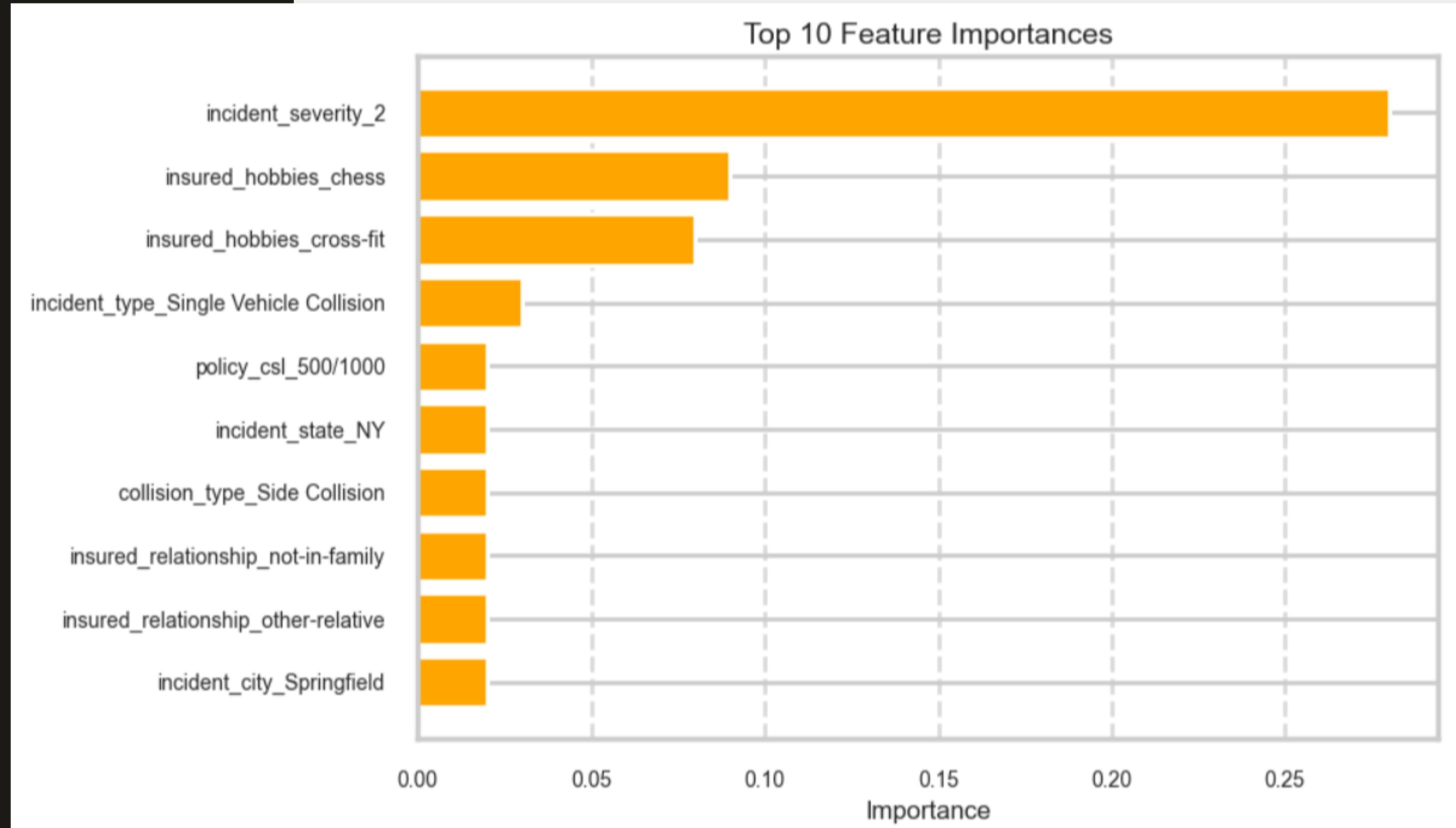
- Mean Train Accuracy: **0.90**
- Mean Validation Accuracy: **0.87**
- Interpretation: **Minimal overfitting observed**

### Validation Set:

- Accuracy: **0.88**
- Recall: **0.85**
- Precision: **0.87**



# RANDOM FOREST MODEL



# FEATURE IMPORTANCE

- **incident\_severity\_2 (0.28)** – Most impactful; indicates severity is a key predictor.
- **insured\_hobbies\_chess (0.09)** – Suggests hobbies affect risk profile.
- **insured\_hobbies\_cross-fit (0.08)** – Physical activity may correlate with claim patterns.
- **incident\_type\_Single Vehicle Collision (0.03)** – Indicates different fraud or risk behavior.
- **policy\_csl\_500/1000 (0.02)** – Coverage level influences outcomes.

## 🔍 Key Insight:

Incident-level attributes — specifically **type, severity, and reported damage** — are the strongest indicators of potential fraud. These contextual features significantly outweigh demographic or historical factors in predictive power.

# INVESTIGATIVE TAKEAWAYS

- **How can we analyse historical claim data to detect patterns that indicate fraudulent claims?**

Historical data analysis reveals patterns such as:

- Claim severity
- Claim-to-premium ratio
- Policyholder hobbies (e.g., cross-fit, chess)

By using machine learning models like Logistic Regression and Random Forest, patterns indicative of fraud are extracted by analyzing how certain features correlate with known fraudulent cases.

- **Which features are most predictive of fraudulent behaviour?**

The key features identified as most predictive include:

- incident\_severity\_2 (0.28)
- insured\_hobbies\_chess (0.09)
- insured\_hobbies\_cross-fit (0.08)
- incident\_type\_Single Vehicle Collision (0.03)
- policy\_csl\_500/1000 (0.02)

These features likely show higher importance in the Random Forest model, which inherently ranks feature importance based on splits and impurity reduction.



## INVESTIGATIVE TAKEAWAYS

- Can we predict the likelihood of fraud for an incoming claim, based on past data?

Yes. Both models — Logistic Regression and Random Forest — are trained on historical labeled data (fraudulent vs. legitimate claims), allowing them to:

- Predict the probability of fraud for new incoming claims.
- Flag high-risk cases early in the claim approval pipeline.

**Random Forest**, with higher recall (0.7551), is particularly effective at catching more actual frauds, even if it occasionally mislabels legitimate claims.



## INVESTIGATIVE TAKEAWAYS

- What insights can be drawn from the model that can help in improving the fraud detection process?
  - **Random Forest** is more sensitive to detecting fraud (higher recall), but **Logistic Regression** offers better **interpretability**.
  - Combining both models (**hybrid model approach**) could offer a balance of power and transparency.
  - **Model deployment** can significantly reduce financial losses by flagging claims early.
  - **Continuous monitoring and periodic retraining** are essential due to evolving fraud tactics.
  - **Implementing a dashboard** to monitor performance ensures ongoing effectiveness.
  - **Precision-Recall trade-offs** should be carefully managed depending on **business risk tolerance**.

## IMPLICATIONS & FUTURE RECOMMENDATIONS

- **Early Fraud Detection:**

Deploying the model helps flag high-risk claims early, reducing financial losses and operational overhead.

- **Model of Choice – Random Forest:**

Due to its high recall (75.5%), it is better suited for catching more fraudulent cases in real-world use.

- **Hybrid Approach:**

Combine Random Forest for detection power and Logistic Regression for transparency and auditability.

- **Ongoing Monitoring:**

Set up a dashboard to monitor precision-recall trade-offs and retrain models regularly to stay ahead of evolving fraud trends.

- **Operational Efficiency:**

Automating fraud scoring streamlines claim processing, enabling faster resolutions for legitimate claims.

# Thank you !

**Team:**

**Vaishali Makwana**

**Praveenkumar Periyaswamy**

