# 

# 

# **Fraudulent Claims Detection**

## Analysis Report

**Presented by:**

Rohit Gupta & Venkata Raghunadh Illinda

Batch: ML C72

# **Table of Contents**

[1. Executive Summary 3](#_83deccvpq4dj)

[2. Key Problems 3](#_eteqwfybo77c)

[3. Tasks Undertaken 3](#_ib74bi3i27ak)

[4. Recommendations 4](#_m3sr8yhokwaq)

[5. Business Implications 4](#_yhicroil5t1m)

[6. Key Questions Asked 4](#_lpug9h34tliq)

[7. Conclusion 5](#_c10dykoo99ay)

# **1. Executive Summary**

Global Insure, an insurance company, faces financial losses due to fraudulent claims. The assignment focuses on building a data driven fraud detection system using historical claim and customer data. To tackle this, we built two machine learning models, Logistic Regression and Random Forest, using past insurance claim data. The goal is to help the company identify frauds early in the claim process and improve efficiency and reduce the financial risks.

# **2. Key Problems**

* High number of fraudulent insurance claims
* Manual fraud detection is time-consuming and inefficient
* Late identification of fraud after claim approval
* Presence of missing, inconsistent, and illogical data

# **3. Tasks Undertaken**

* Data Cleaning and Preprocessing
* Exploratory Data Analysis
* Handling Missing and Redundant Data
* Feature Engineering and Transformation
* Class Balancing using Resampling
* Model Building: Logistic Regression & Random Forest
* Model Evaluation using multiple metrics
* Hyperparameter Tuning & Cross-Validation
* Business Insight Generation and Final Recommendations

# **4. Recommendations**

* Use Random Forest for Quick Detection
* Use Logistic Regression for Clarity
* Work on Improving Recall
* Combine model predictions with human expertise for a hybrid decision-making process.
* Keep monitoring and retraining the model with fresh claim data every 3–6 months.

# **5. Business Implications**

* The models can automatically spot suspicious claims, so that manual checks can be reduced.
* This will help in less time wastage, faster claim processing, and saving money by avoiding fake payouts.
* Your fraud investigation team can focus on the riskiest cases first, making their work more efficient and impactful.

# **6. Key Questions Asked**

Q1: How can we analyze historical claim data to detect patterns that indicate fraudulent claims?

* We analyzed the data using Exploratory Data Analysis (EDA) and correlation studies.
* Patterns were identified in features like incident\_severity, insured\_hobbies, and authorities\_contacted.
* Fraudulent claims often had different distributions in severity, hobbies, and reporting behavior.

Q2: Which features are most predictive of fraudulent behavior?

* From both models and feature selection steps, the following features stood out:
* Incident severity: Minor or trivial damages were often marked as fraud.
* Insured hobbies: Certain hobbies, like cross-fit and chess, appeared more in fraud cases.
* Authorities contacted: Fraud cases often did not involve official reporting.
* Claim amounts: Extremely high claim amounts were suspicious.

Q3: Can we predict the likelihood of fraud for an incoming claim based on past data?

* Yes. Our models, especially Random Forest, predicted fraud with a validation accuracy of ~83%, with 60% precision and 49% recall, meaning it is capable of flagging suspicious claims early.

Q4: What insights can be drawn from the model that can help in improving the fraud detection process?

* Not all fraud cases are obvious; some appear under trivial incidents.
* Behavioral patterns (like hobbies and non-reporting to authorities) can be strong indicators.
* Early detection is possible using these patterns, helping the claims team prioritize review.

# **7. Conclusion**

* We built two models, Logistic Regression and Random Forest to help detect fraudulent insurance claims.
* The Logistic Regression model gave us clear insights into which factors matter most like incident severity and customer hobbies.
* The Random Forest model was more powerful overall and achieved better accuracy. However, it did miss few fraud cases.
* The Random Forest model correctly identified about 49% of fraud cases from the recall.
* As per the Specificity it spotted nearly 90% of genuine claims
* As per the precision 60% of flagged frauds were correct
* This means that the model is better at avoiding false alarms by not calling the genuine claims as frauds.