FRAUDULENT CLAIM DETECTION CASE STUDY

SUBMITTED BY:

- SHIVANSHU KUMAR SINGH (shivanshu.singh2102@gmail.com)
- ANKIT SRIVASTAVA (toankit1987@gmail.com)

INTRODUCTION & PROBLEM STATEMENT

- Global Insure processes thousands of claims annually, leading to financial losses due to fraudulent claims.
- Traditional manual fraud detection is time-consuming and inefficient.
- Goal: Develop a data-driven fraud detection system to flag suspicious claims before approval, minimizing losses.

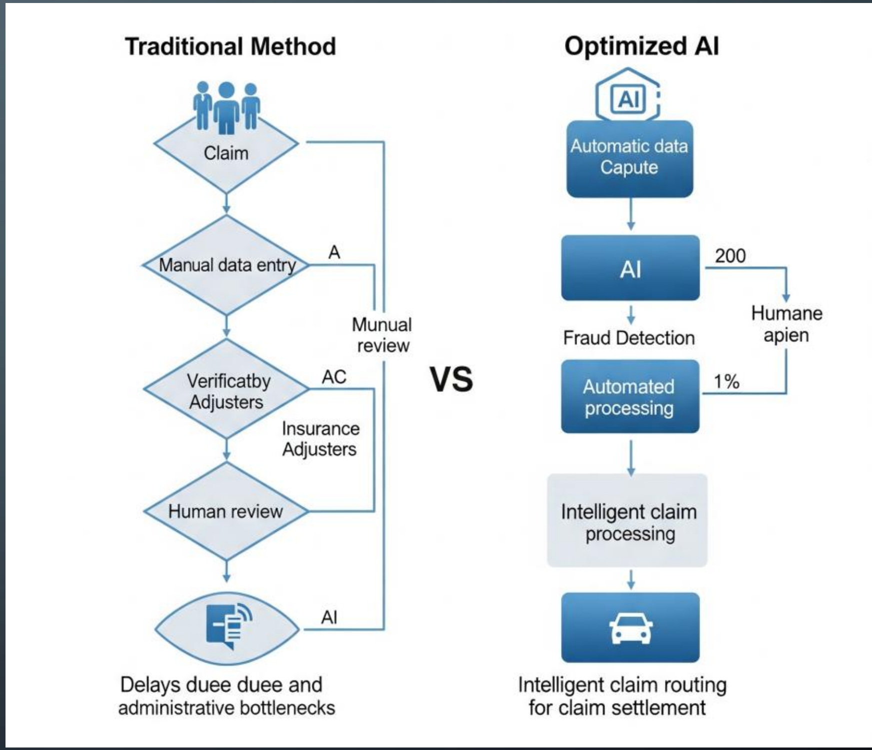

Figure 1: A flowchart illustrating the traditional claims process vs. an optimized AI-based approach. Generated by AI

APPROACH & METHODOLOGY

- Data Preparation & Cleaning:
  - Handling missing values, redundant columns, and unique identifiers.
- Exploratory Data Analysis (EDA):
  - Univariate and bivariate analysis to detect fraud patterns.
- Feature Engineering:
  - Customer tenure, claim ratios, incident severity.
- Model Building & Evaluation:
  - Logistic Regression & Random Forest models.
  - Hyperparameter tuning & validation.

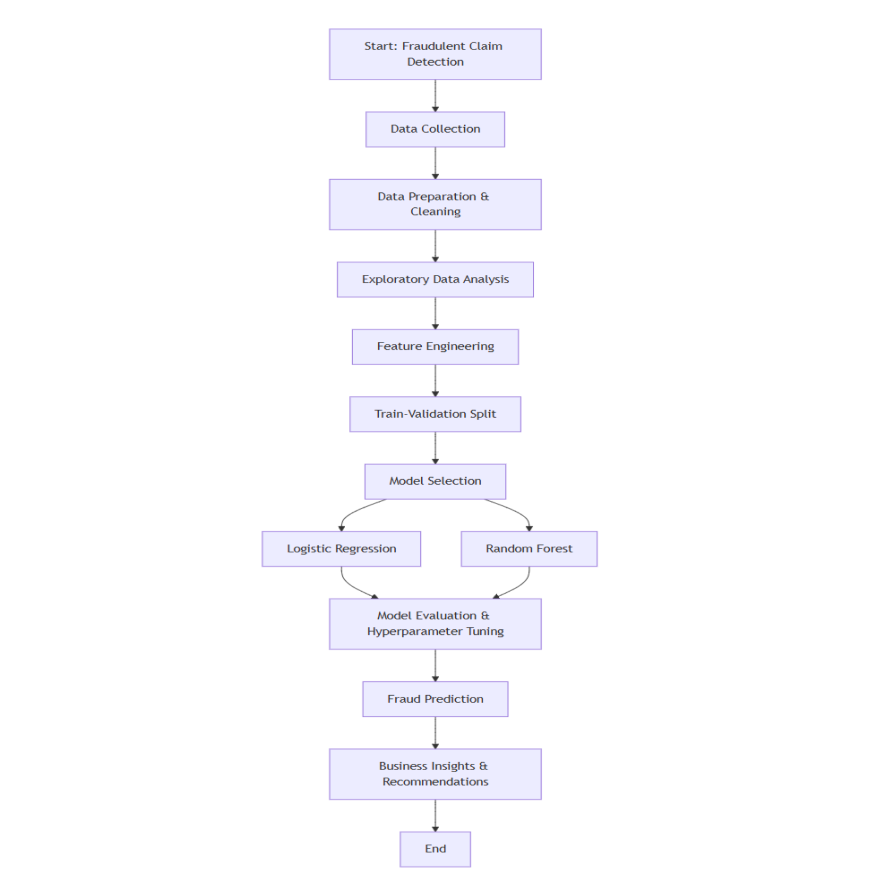

Figure 2: A high-level workflow diagram of the methodology.

HOW CAN WE ANALYZE HISTORICAL DATA TO DETECT FRAUD?

- **Pattern Recognition:** Analyzing fraudulent vs. legitimate claims.
- **Feature Importance:** Identifying the most relevant factors.
- **Correlations:** Examining how attributes (like claim amounts, occupation, policy deductible) influence fraud.

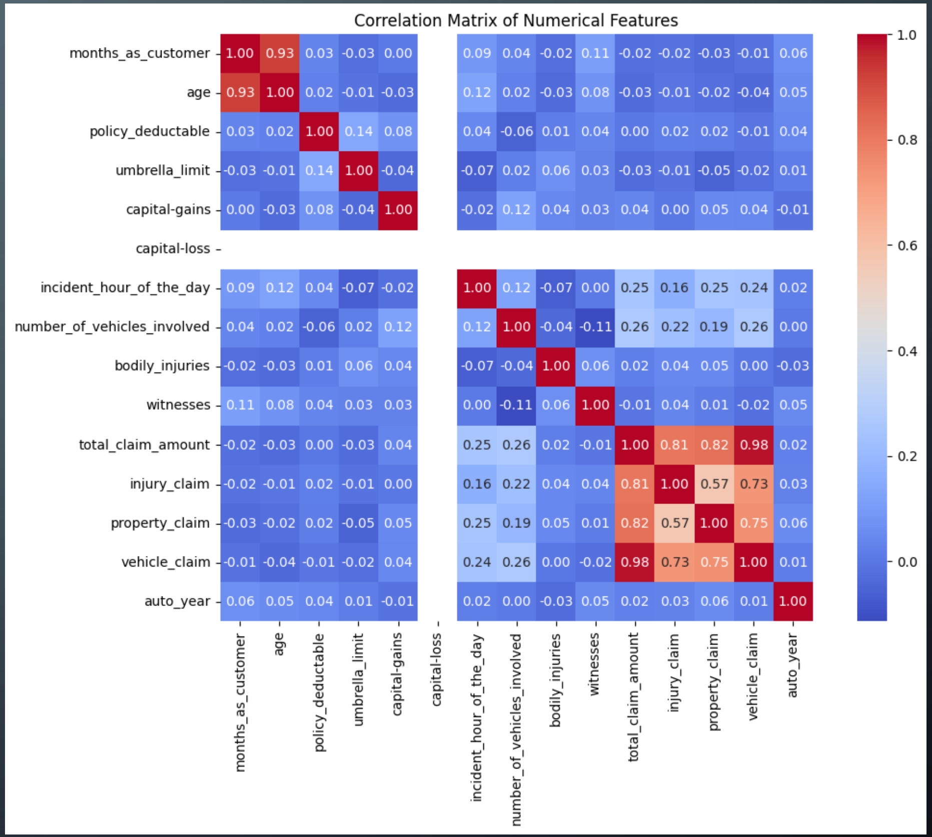

Figure 3: A heatmap showing feature correlations.

WHICH FEATURES ARE MOST PREDICTIVE OF FRAUDULENT BEHAVIOR?

- High Claim Amounts
- Policy Deductibles – Higher deductibles correlate with fraud.
- Incident Severity – Major damage & single-vehicle collisions show higher fraud likelihood.
- Absence of Police Reports & Property Damage Labeled "Unknown".
- Luxury Auto Brands (Mercedes, Ford).

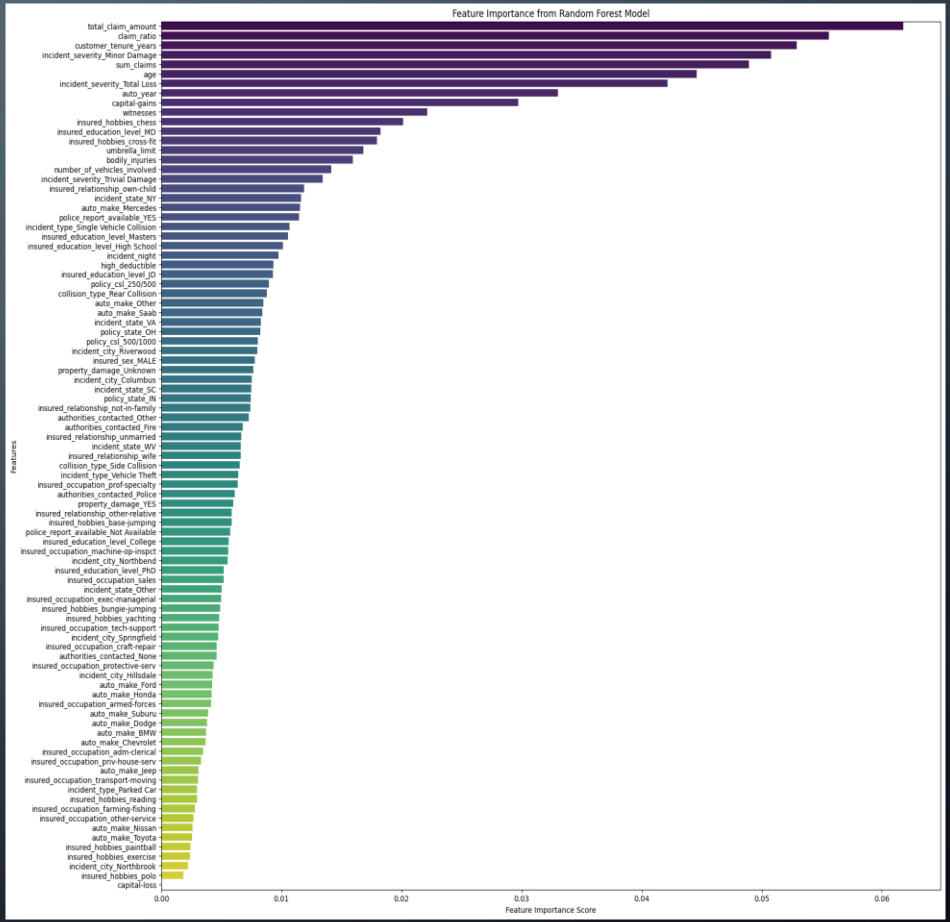

Figure 4: Feature importance graph from Random Forest.

CAN WE PREDICT THE LIKELIHOOD OF FRAUD FOR AN INCOMING CLAIM?

- Using **historical fraud data**, our model predicts fraud probability.
- Logistic Regression offers **probability estimates**, while Random Forest gives **classification strength**.
- A fraud detection threshold ensures **claims above 70% probability are flagged for review.**

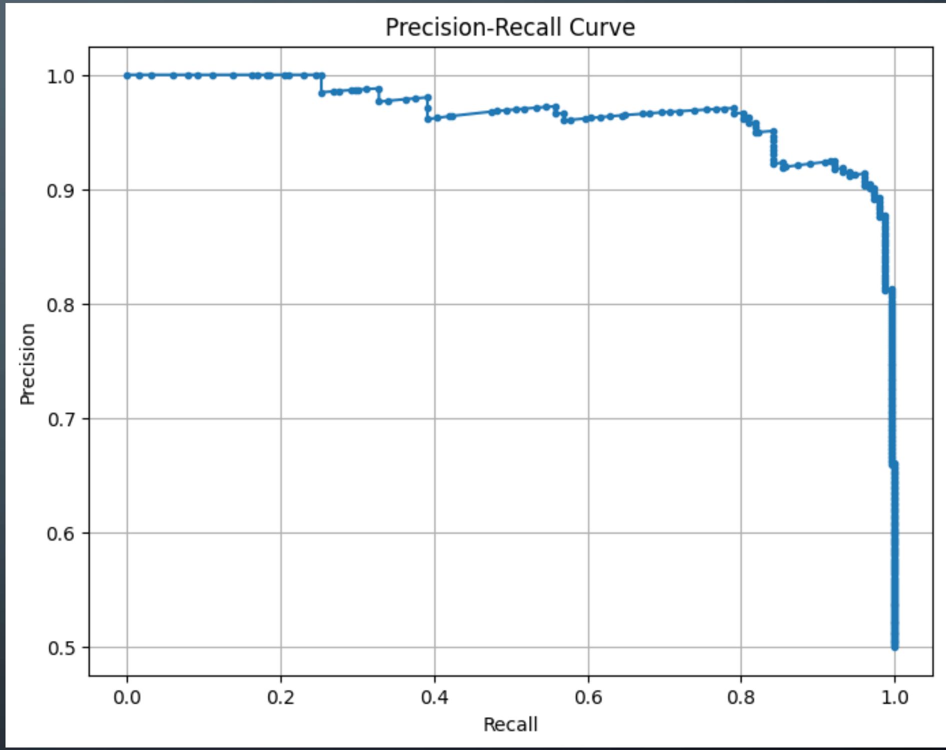

Figure 5: Precision-recall curve graph.

INSIGHTS TO IMPROVE THE FRAUD DETECTION PROCESS

- **Major claims need deeper scrutiny –** High-value claims show high fraud rates.
- **Customers with less tenure are riskier –** New policyholders file more fraudulent claims.
- **Investigate cases without police reports or with unknown property damage.**
- **Automate fraud detection using ML models,** reducing manual intervention.

- Model Performance & Evaluation
  - Logistic Regression (Validation Set)
    - Accuracy: 76.92%
    - Sensitivity (Recall): 64.71%
    - Specificity: 80.73%
    - Precision: 51.16%
    - F1 Score: 57.14%
  - Random Forest (Validation Set)
    - Accuracy: 76.92%
    - Sensitivity (Recall): 23.53%
    - Specificity: 93.57%
    - Precision: 53.33%
    - F1 Score: 32.65%

BUSINESS IMPLICATIONS & RECOMMENDATIONS

- **Financial Savings –** Prevent fraudulent payouts and reduce claim processing costs.
- **Operational Efficiency** – Automate fraud detection, freeing manual resources.
- **Customer Trust & Compliance** – Lower false rejections, ensuring fair claims

- **Deploy Random Forest Model in production** for real-time fraud detection.
- **Improve data collection mechanisms –** Ensure **police reports & property damage descriptions** are mandatory.
- **Regularly update the fraud detection system** with new patterns.
- **Implement flagging thresholds** for suspicious claims based on detected fraud probability.

CONCLUSION

- Machine Learning enables **early fraud detection**, minimizing insurance fraud risks.
- Continuous **retraining and updates** will improve fraud detection accuracy.