Model Selection – Detect Insurance Fraud Claim

- Nisarg Shah
- Jagdsh Chand

Problem Statement & Business Objective

- Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses. The company's current process for identifying fraudulent claims involves manual inspections, which is time-consuming and inefficient. Fraudulent claims are often detected too late in the process, after the company has already paid out significant amounts. Global Insure wants to improve its fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process. This would minimise financial losses and optimise the overall claims handling process.

Business Objective

Global Insure wants to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features like claim amounts, customer profiles and claim types, the company aims to predict which claims are likely to be fraudulent before they are approved.

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

1. Descriptive Pattern Analysis (EDA)

- Visualization and aggregation on the raw and engineered data to reveal characteristic differences between fraud and non-fraud claims.
- Visualizing the distribution of numerical and categorical features using appropriate plots revealed that as claim increases fraud reported also increased with concentration in the middle range
- Analyzing features like the **Claim Consistency Ratio** (Policy Annual Premium / Total Claim Amount) reveals if fraudulent claims statistically involve a much higher payoff relative to the customer's premium investment
- Applying correlations showed that claims are always correlated and incident_of_the_hour is correlated with claim amount.

2. Feature Engineering:

- By applying dummy variables and scaler we ensures that all features contribute equally to the pattern detection process, preventing mathematical bias.
- We can use StandardScaler or MinMax scaler according to the data.

3. Predictive Modeling and Feature Importance

- **Feature Importance Ranking (RFECV):** The final model's built-in **Feature Importance** output reveals the variables that were most instrumental in separating fraud from non-fraud. The highest ranked features
  (e.g., Claim Consistency Ratio, Vehicle Claim) are the strongest indicators of fraud

Which features are most predictive of fraudulent behaviour?

1. Claim Components:

- **property_claim (0.155):** This is the single most predictive feature, suggesting that the amount claimed for property damage is a primary indicator used by the model. Fraudsters may inflate these values or target scenarios where property damage is hard to verify.
- **Claim_component_balance (0.123):** This engineered feature (likely the sum of injury_claim and property_claim) is the second most predictive. Its high rank shows that the model relies heavily on the combined non-vehicle loss amount, often associated with exaggerated soft-tissue injury claims.

2. Financial Stress and Inconsistency (The Motive):

- **Claim_consistency_ratio (0.079):** This is a key fraud indicator. Its high score confirms that claims that are disproportionately large compared to the low Policy Annual Premium are highly predictive of fraud.
- **Capital_activity (0.069):** This feature (likely derived from capital-gains and capital-loss) measures customer financial dist**ress**. Its rank confirms that the customer's financial motive is a strong predictor of fraud.

3. Time, Tenure, and Stability (The Context):

- **months_as_customer (0.082) and Policy_Tenure (0.058):** The presence of both original and engineered tenure features confirms that claims filed early in the policy's life (low tenure) or by newer customers are strongly indicative of fraud risk.
- **incident_hour_of_the_day (0.055):** This temporal feature confirms that claims clustering at suspicious times (e.g., late night/early morning) are predictive, as these times minimize witnesses.

4. Categorical and Geographic Indicators

- **valid_single_collision_damage (0.068):** This binary feature (likely flagging a specific, easy-to-stage incident type) shows that the mechanics of the incident are important.
- **incident_state_WV (0.041):** The fact that a specific state WV is a top predictor suggests that fraud is geographically concentrated, potentially due to organized rings, local laws, or economic distress unique to that region.

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

Yes, absolutely. Predicting the likelihood of fraud for an incoming claim is the primary objective of the machine learning process. The Machine Learning model acts as a highly specialized filter, using the statistical patterns from the past to identify suspicious claims in the present.

1. Training Phase (Learning Patterns):

- The model was trained on historical data where the outcome was known (labeled as 'Fraud' or 'Not Fraud'). The model will learn the complex, non-linear relationships that distinguish a fraudulent claim by observing the top predictive features.
- High-Risk Signature: We learned that if a claim has a high Claim Consistency Ratio, low Policy Tenure, and involves a high Property Claim amount, the likelihood of fraud is statistically high.

2. Prediction Phase (Generating Likelihood):

- Feature Extraction: The new claim's details are preprocessed and transformed using categorical encoding, numerical scaling, and creation of engineered features like Claim Consistency Ratio for the training data.
- Probability Output: The model uses its learned rules (the "forest" of decision trees) to score the new claim. Instead of outputting a simple 'Yes' or 'No', the model outputs a probability score between 0 and 1 (e.g., 0.85)

3. Decision Phase (Threshold Application):

- Standard Prediction: If Likelihood>0.50, the claim is flagged as "Fraud.", The default which we applied in the assignment.
- Optimized Prediction: If Likelihood>T (where T is your business-optimized threshold, often 0.65 or 0.70 to reduce false alarms), the claim is routed for immediate investigation.

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

1. **Investigative Focus (Prioritizing Features)**
   - **Claim_consistency_ratio:** show that the model is primarily exploiting financial inconsistencies.
     - **Action:** Investigate claims where the ratio of **Total Claim Amount to Annual Premium** is abnormally high, as this is the most common financial red flag.
   - **Tenure Risk:** The high importance of months_as_customer and incident_hour_of_the_day suggests that simple filters should be applied to new claims.
     - **Action:** Claims filed by customers with low policy tenure (e.g., <6 months) and those occurring during unusual hours (e.g., 2 AM to 5 AM) should automatically receive an initial high-risk score.
2. **Operational Improvement**
   - **Geographic Risk Allocation:** The importance of features like incident_state_WV suggests that fraud risk is not evenly distributed.
     - **Action:** Allocate more investigative resources (e.g., assign more experienced adjusters or implement mandatory on-site inspections) to regions identified as statistically high-risk by the model.
3. **Data Integrity and Future Modeling**
   - **Data Quality Feedback:** The model highlighted the need to clean problematic placeholder values (like '?') and handle outliers (inf VIF). This process to ensure cleaner data for future models.
   - **Feature Engineering Value:** The high predictive power of engineered features (Claim_consistency_ratio, Claim_component_balance) proves that simple raw data is insufficient.
     - **Action:** The claims database should be modified to pre-calculate and store these derivative risk ratios to allow for real-time risk assessment and improve future model training speed and accuracy.