Fraudulent Claim Detection Report

By Vishal Verma, Vinti Singh

1. Problem Statement & Business Objective

Global Insure aims to detect fraudulent insurance claims before approval. The objectives are to:

- Analyze historical claims to identify fraud patterns
- Determine key predictors of fraudulent behavior
- Develop and evaluate models to flag high-risk claims early.

1. Data Overview

Source: insurance_claims.csv, containing policy details, incident information, customer demographics, claim amounts, and a binary target fraud_reported (Y/N).

Training Set:

- Class Balance: Fraudulent: ~25% (699 × 0.25≈174 samples)
- Non-fraudulent: ~75% (699 × 0.75≈526 samples)
- Imbalance ratio ≈ 3 : 1 (majority: minority)

---

3. Data Preparation & Cleaning

- **Missing Values**
  - Identified and dropped columns with high proportion of missing data.
  - Imputed or removed rows for remaining nulls as appropriate.
  - authorities_contacted has None as one of the categories, but np.nan interprets None as null. Therefore, we will skip all rows where authorities_contacted is np.nan.
  - Empty columns: ['_c39']

- **Redundant & Illogical Entries**
  - Removed duplicate records.
  - Negative values in the dataset:

    umbrella_limit   1
    capital-loss   525
    dtype: int64

    Number of rows with negative values: 526
    Dropping rows with negative values in numeric columns (excluding 'capital-loss')
    Dataset shape after removing rows with negative values: (999, 39)

  - Dropped features with constant or near-constant values.
  - Columns with their percentage of unique values:

    policy_number: 1.0000 (999 / 999)
    incident_location: 1.0000 (999 / 999)
    insured_zip: 0.9950 (994 / 999)
    policy_annual_premium: 0.9910 (990 / 999)
    policy_bind_date: 0.9510 (950 / 999)
    total_claim_amount: 0.7628 (762 / 999)
    vehicle_claim: 0.7257 (725 / 999)
    injury_claim: 0.6386 (638 / 999)
    property_claim: 0.6256 (625 / 999)
    months_as_customer: 0.3914 (391 / 999)
    capital-loss: 0.3544 (354 / 999)
    capital-gains: 0.3383 (338 / 999)
    incident_date: 0.0601 (60 / 999)
    age: 0.0460 (46 / 999)
    auto_model: 0.0390 (39 / 999)
    incident_hour_of_the_day: 0.0240 (24 / 999)
    auto_year: 0.0210 (21 / 999)
    insured_hobbies: 0.0200 (20 / 999)
    insured_occupation: 0.0140 (14 / 999)
    auto_make: 0.0140 (14 / 999)
    umbrella_limit: 0.0100 (10 / 999)
    insured_education_level: 0.0070 (7 / 999)
    incident_state: 0.0070 (7 / 999)
    incident_city: 0.0070 (7 / 999)
    insured_relationship: 0.0060 (6 / 999)
    incident_type: 0.0040 (4 / 999)
    collision_type: 0.0040 (4 / 999)
    incident_severity: 0.0040 (4 / 999)
    authorities_contacted: 0.0040 (4 / 999)
    number_of_vehicles_involved: 0.0040 (4 / 999)
    witnesses: 0.0040 (4 / 999)
    policy_state: 0.0030 (3 / 999)
    policy_csl: 0.0030 (3 / 999)
    policy_deductable: 0.0030 (3 / 999)
    property_damage: 0.0030 (3 / 999)
    bodily_injuries: 0.0030 (3 / 999)
    police_report_available: 0.0030 (3 / 999)

insured_sex: 0.0020 (2 / 999)
fraud_reported: 0.0020 (2 / 999)

Columns with high cardinality (>80% unique values):
['policy_number', 'policy_bind_date', 'policy_annual_premium', 'insured_zip', 'incident_location']

Removing 5 columns with high cardinality
Dataset shape after removing high cardinality columns: (999, 34)

- **Data Types**
  - Converted date fields to datetime objects.
    - Updated data types for date columns: incident_date: datetime64[ns]
  - Cast categorical columns to category dtype.

---

4. Exploratory Data Analysis (EDA)

- **Univariate Analysis**

  Observations from histogram plots:

months_as_customer:
- Mean: 202.57, Median: 199.00
- Skewness: 0.37
- Distribution appears approximately symmetric

age:
- Mean: 38.85, Median: 38.00
- Skewness: 0.51
- Distribution is positively skewed (right-tailed)

policy_deductable:
- Mean: 1150.21, Median: 1000.00
- Skewness: 0.45
- Distribution appears approximately symmetric

umbrella_limit:
- Mean: 1077253.22, Median: 0.00
- Skewness: 1.79
- Distribution is positively skewed (right-tailed)

capital-gains:
- Mean: 25506.01, Median: 0.00
- Skewness: 0.45
- Distribution appears approximately symmetric

capital-loss:
- Mean: -26458.37, Median: -20800.00
- Skewness: -0.41
- Distribution appears approximately symmetric

incident_hour_of_the_day:
- Mean: 11.53, Median: 12.00
- Skewness: -0.01
- Distribution appears approximately symmetric

number_of_vehicles_involved:
- Mean: 1.83, Median: 1.00
- Skewness: 0.49
- Distribution appears approximately symmetric

bodily_injuries:
- Mean: 0.97, Median: 1.00
- Skewness: 0.06
- Distribution appears approximately symmetric

witnesses:
- Mean: 1.46, Median: 1.00
- Skewness: 0.06
- Distribution appears approximately symmetric

total_claim_amount:
- Mean: 52923.61, Median: 58300.00
- Skewness: -0.57
- Distribution is negatively skewed (left-tailed)

injury_claim:
- Mean: 7508.73, Median: 6780.00
- Skewness: 0.27
- Distribution appears approximately symmetric

property_claim:
- Mean: 7399.20, Median: 6780.00
- Skewness: 0.33
- Distribution appears approximately symmetric

vehicle_claim:
- Mean: 38015.68, Median: 42420.00
- Skewness: -0.59
- Distribution is negatively skewed (left-tailed)

auto_year:
- Mean: 2004.96, Median: 2005.00
- Skewness: -0.00
- Distribution appears approximately symmetric - Distribution appears approximately symmetric


months_as_customer
Mean: 202.57, Median: 199.00, Skew: 0.37
Symmetric
age
Mean: 38.85, Median: 38.00, Skew: 0.51
Right-skewed
policy_deductable
Mean: 1150.21, Median: 1000.00, Skew: 0.45
Symmetric
umbrella_limit
Mean: 1077253.22, Median: 0.00, Skew: 1.79
Right-skewed
capital-gains
Mean: 25506.01, Median: 0.00, Skew: 0.45
Symmetric
capital-loss
Mean: -26458.37, Median: -20800.00, Skew: -0.41
Symmetric
incident_hour_of_the_day
Mean: 11.53, Median: 12.00, Skew: -0.01
Symmetric
number_of_vehicles_involved
Mean: 1.83, Median: 1.00, Skew: 0.49
Symmetric
bodily_injuries
Mean: 0.97, Median: 1.00, Skew: 0.06
Symmetric
witnesses
Mean: 1.46, Median: 1.00, Skew: 0.06
Symmetric
total_claim_amount
Mean: 52923.61, Median: 58300.00, Skew: -0.57
Left-skewed
injury_claim
Mean: 7508.73, Median: 6780.00, Skew: 0.27
Symmetric
property_claim
Mean: 7399.20, Median: 6780.00, Skew: 0.33
Symmetric
vehicle_claim
Mean: 38015.68, Median: 42420.00, Skew: -0.59
Left-skewed
auto_year
Mean: 2004.96, Median: 2005.00, Skew: -0.00
Symmetric
Frequency

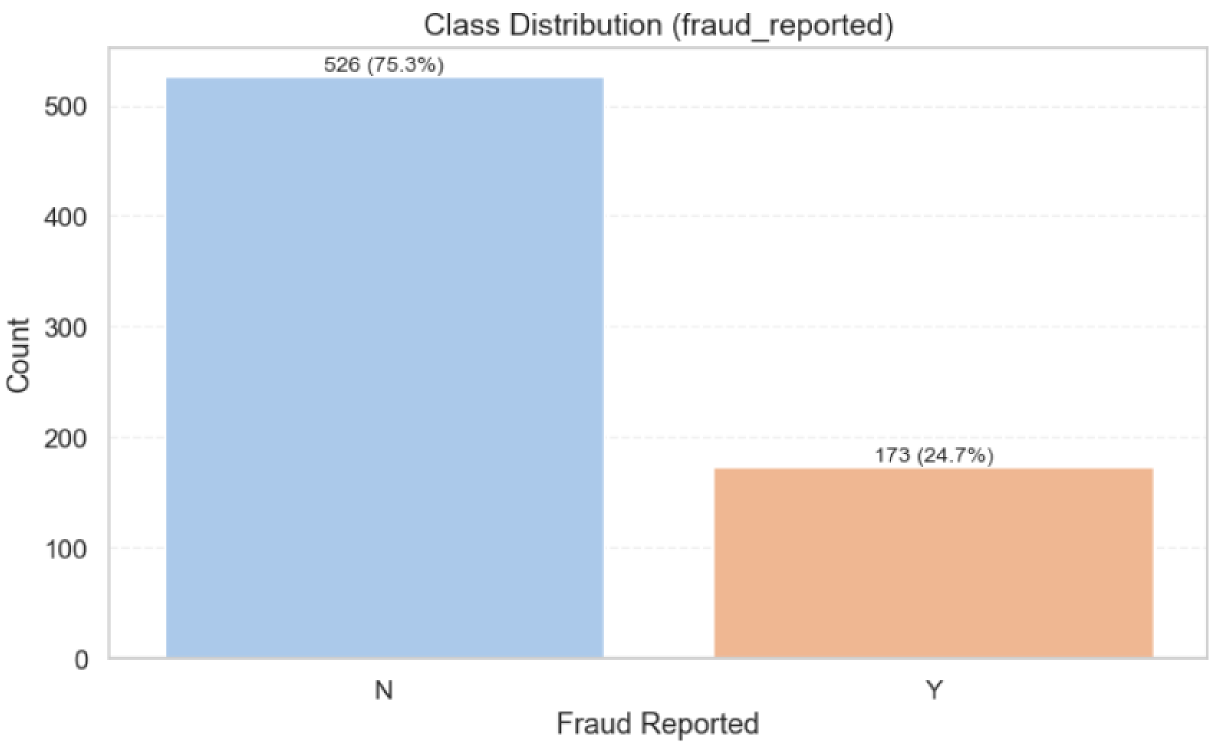
- **Class Balance**

Class imbalance analysis:

Majority class (N): 526 samples (75.25%)
Minority class (Y): 173 samples (24.75%)
Imbalance ratio (majority:minority): 3.04:1

The dataset shows significant class imbalance. This may affect model performance. Consider using techniques such as:

1. Resampling methods (oversampling minority class or undersampling majority class)
2. Using class weights during model training
3. Using algorithms that handle imbalanced data well
4. Using evaluation metrics appropriate for imbalanced datasets (e.g., precision, recall, F1-score, AUC-ROC)



- **Correlation Analysis**
  - Heatmaps indicated moderate correlations between certain numeric features (e.g., injuries_vehicles_interaction and total_claim_amount).
  - Highly correlated feature pairs (|correlation| > 0.7):

    age & months_as_customer: 0.920

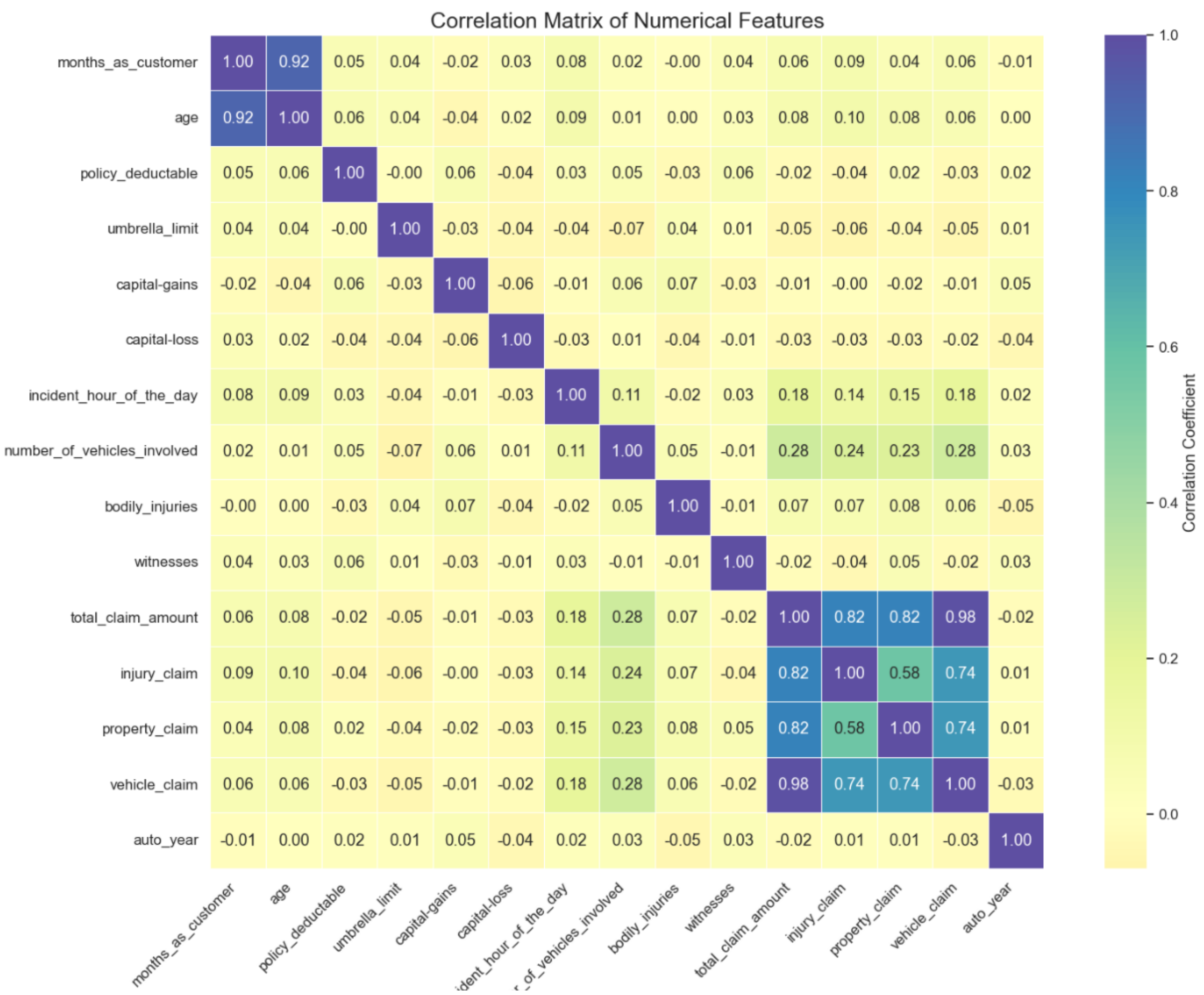
    injury_claim & total_claim_amount: 0.818

    property_claim & total_claim_amount: 0.815

    vehicle_claim & total_claim_amount: 0.984

    vehicle_claim & injury_claim: 0.743

    vehicle_claim & property_claim: 0.742



- **Bivariate Analysis**

  Feature importance based on variance in fraud

  rates: incident_severity: 655.5417

  insured_hobbies: 437.9118

  auto_model: 138.9059

  incident_type: 127.9124

  collision_type: 97.4883

  incident_state: 73.1274

  property_damage: 39.8805

  insured_occupation: 39.3522

  auto_make: 27.8186

  insured_relationship: 24.6759

  authorities_contacted: 23.6709

  incident_city: 14.4581

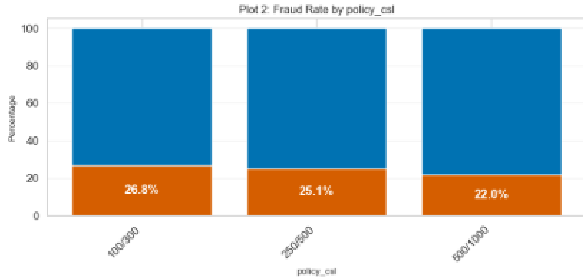
  policy_csl: 6.0253

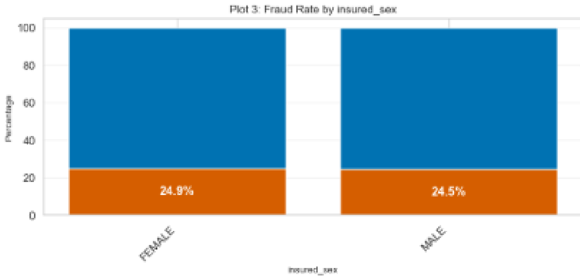
  insured_education_level: 5.3411

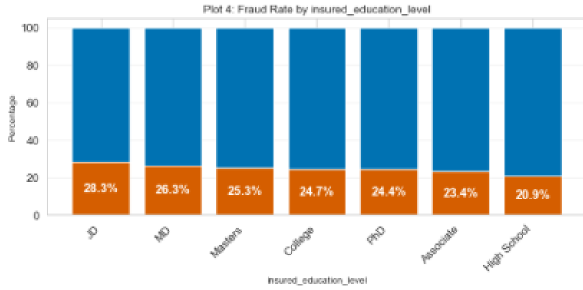
  police_report_available: 2.156

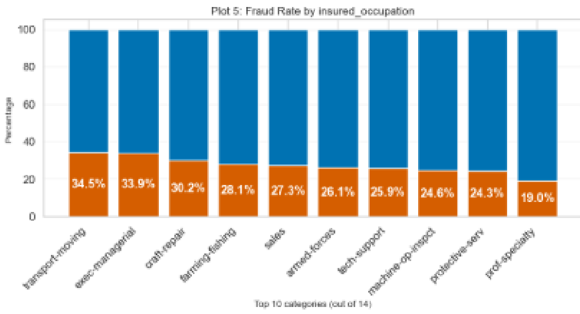
  policy_state: 0.2506

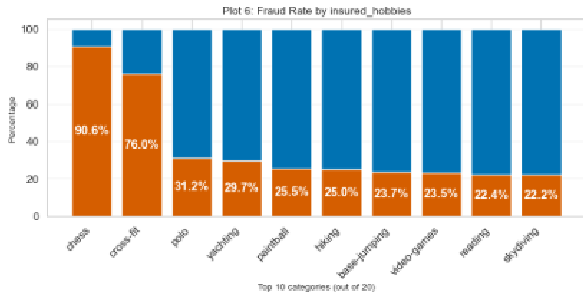
  insured_sex: 0.0773

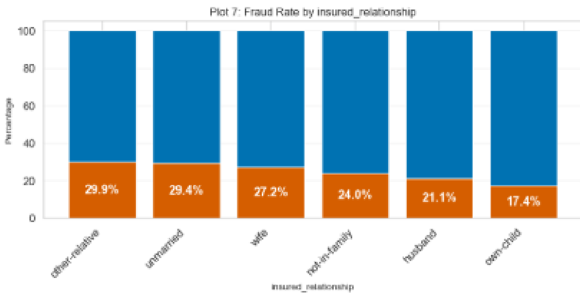


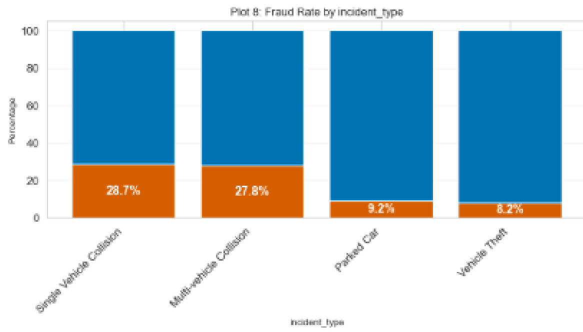


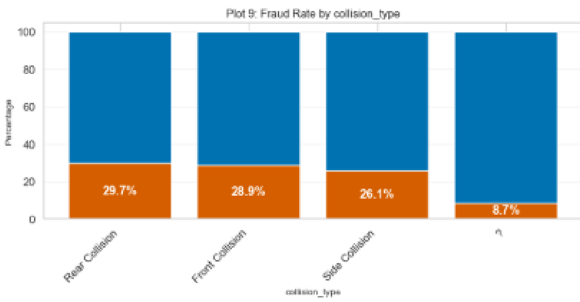


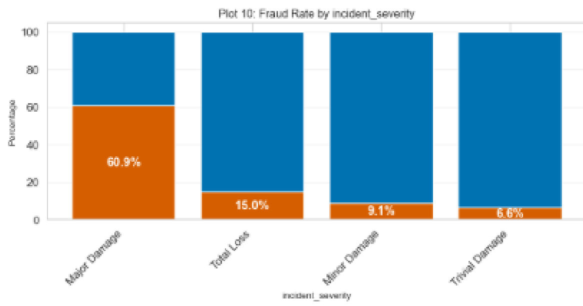












Plot 11: Fraud Rate by authorities_contacted

Plot 12: Fraud Rate by incident_state

5. Feature Engineering

- **Resampling**: explored SMOTE and under sampling to address imbalance.

  Class distribution before resampling:
  fraud_reported
  N 526
  Y 173
  Name: count, dtype: int64

  Class distribution after resampling:
  fraud_reported
  N 526
  Y 526
  Name: count, dtype: int64

  Original training set shape: (699, 33)
  Resampled training set shape: (1052, 33)

- **Feature Creation**:
  - Engineered interaction terms (e.g., time between policy inception and incident).
  - Grouped low-frequency categories into “Other” for stability.
  - Created date-based, claim ratio, time-of-day, interaction, age group, customer tenure features.

    Training set shape after feature creation: (1052, 46)
    Test set shape after feature creation: (300, 46)

    New features:
    - ['incident_day_of_week', 'incident_month', 'is_weekend', 'vehicle_age', 'injury_claim_ratio', 'property_claim_ratio', 'vehicle_claim_ratio', 'vehicles_witnesses_interaction', 'injuries_vehicles_interaction', 'high_claim_amount']

- **Handle Redundant Columns**
  - **Found 3 highly correlated pairs (correlation > 0.85):**
    - age & months_as_customer: 0.9271
    - vehicle_claim & total_claim_amount: 0.9803
    - vehicle_age & auto_year: -1.0000

  Dropping 'incident_date' as we've created derived features from it: incident_day_of_week, incident_month, is_weekend

  Dropping 'auto_year' as we've created derived features from it: vehicle_age

  Dropping 'incident_hour_of_the_day' as we've created derived features from it: incident_time_of_day
  Dropping 'vehicle_claim' due to high correlation (0.9803) with 'total_claim_amount'
  Dropping 'months_as_customer' due to high correlation (0.9271) with 'age'

Removed 5 redundant columns: ['incident_hour_of_the_day', 'incident_date', 'vehicle_claim', 'months_as_customer', 'auto_year']

Training data shape after removing redundant columns: (1052, 41)
Testing data shape after removing redundant columns: (300, 41)

- **Combine values in Categorical Columns**
    - Found 20 categorical features to analyze for combining values and updated below column.
        - Column 'auto_model'
          Total unique values: 39
          Rare categories (< 2% of data): 11
          Reduced categories from 39 to 29
          Top 5 categories after combining: {'Other': 166, 'A5': 50, 'F150': 49, 'RAM': 43, 'A3': 43}
    - Modified 2 categorical columns by combining rare categories
        - incident_severity value distribution:

          Major Damage: 41.3%
          Minor Damage: 28.6%
          Total Loss: 23.9%
          Trivial Damage: 6.2%
        - insured_hobbies value
          distribution: chess: 9.6%
          paintball: 6.9%
          reading: 6.2%
          bungie-jumping: 6.1%
          exercise: 5.3%
          skydiving: 5.2%
          yachting: 5.0%
          base-jumping: 5.0%
          board-games: 4.8%
          hiking: 4.8%
          polo: 4.7%
          cross-fit: 4.7%
          video-games: 4.6%
          movies: 4.5%
          golf: 4.5%
          kayaking: 4.4%
          camping: 4.3%
          sleeping: 3.9%
          dancing: 3.7%
          Other: 1.9%

        - incident_type value distribution:
          Multi-vehicle Collision: 44.9%
          Single Vehicle Collision:
          41.2% Parked Car: 7.7%
          Vehicle Theft: 6.3%

- auto_make value
  distribution: Ford: 9.2%
  Audi: 8.8%
  Chevrolet: 8.7%
  Saab: 8.2%
  Dodge: 8.1%
  Nissan: 7.9%
  Suburu: 7.5%
  BMW: 7.3%
  Mercedes: 6.7%
  Accura: 6.3%
  Toyota: 5.8%
  Jeep: 5.7%
  Volkswagen: 5.4%
  Honda: 4.4%

- insured_relationship value
  distribution: other-relative: 18.3%
  wife: 17.5%
  not-in-family: 17.0%
  unmarried: 16.3%
  husband: 15.9%
  own-child: 15.1%

- **Encoding & Scaling**:
  - One-hot encoded ~20 categorical variables.
    - Cardinality of each categorical column:
      policy_state: 3 unique values policy_csl: 3
      unique values insured_sex: 2 unique
      values insured_education_level: 7 unique
      values insured_occupation: 14 unique
      values insured_hobbies: 20 unique
      values insured_relationship: 6 unique
      values incident_type: 4 unique values
      collision_type: 4 unique values
      incident_severity: 4 unique values
      authorities_contacted: 5 unique values
      incident_state: 7 unique values
      incident_city: 7 unique values
      property_damage: 3 unique values
      police_report_available: 3 unique values
      auto_make: 14 unique values
      auto_model: 29 unique values
      incident_time_of_day: 4 unique values
      age_group: 4 unique values
      customer_tenure_group: 3 unique values

    - Shape of X_train before creating dummy variables: (1052, 41)
    - Shape of X_train after creating dummy variables: (1052, 147)
    - Created dummy variables for dependent feature in training data {'Y': 1, 'N': 0}

- Standardized numeric features via Min–Max scaling.

- **Feature Selection**:
  - **Logistic Regression + RFECV**: Recursive elimination with cross-validation selected the top ~52 predictors.
    - Optimal number of features: 52

[
'policy_csl_250/500', 'insured_education_level_JD',
'insured_education_level_MD', 'insured_education_level_PhD',
'insured_occupation_exec-managerial',
'insured_occupation_farming-fishing',
'insured_occupation_handlers-cleaners',
'insured_occupation_other-service',
'insured_occupation_priv-house-serv', 'insured_hobbies_camping',
'insured_hobbies_chess', 'insured_hobbies_cross-fit',
'insured_hobbies_dancing', 'insured_hobbies_golf',
'insured_hobbies_movies', 'insured_hobbies_sleeping',
'insured_hobbies_video-games', 'insured_relationship_not-in-family',
'insured_relationship_own-child', 'insured_relationship_unmarried',
'incident_type_Vehicle Theft', 'collision_type_Side Collision',
'collision_type_Unknown', 'incident_severity_Minor Damage',
'incident_severity_Total Loss', 'incident_severity_Trivial Damage',
'incident_state_NY', 'incident_state_OH', 'incident_state_PA',
'incident_state_WV', 'incident_city_Northbrook',
'property_damage_Unknown', 'property_damage_YES', 'auto_make_Audi',
'auto_make_BMW', 'auto_make_Chevrolet', 'auto_make_Nissan',
'auto_model_A5', 'auto_model_Camry', 'auto_model_Civic',
'auto_model_F150', 'auto_model_Fusion', 'auto_model_Grand Cherokee',
'auto_model_Legacy', 'auto_model_MDX', 'auto_model_Other',
'auto_model_Pathfinder', 'auto_model_Silverado', 'auto_model_Ultima',
'auto_model_Wrangler', 'auto_model_X5', 'age_group_Young'
]

  - **Random Forest**: Feature importance thresholding (0.01) retained 28 variables. Hyper Parameter Tuning
    - Hyperparameter tuning
      Fitting 5 folds for each of 972 candidates, totalling 4860 fits

      Best Parameters:
      {'bootstrap': True, 'class_weight': 'balanced_subsample', 'criterion': 'entropy', 'max_depth': 12, 'max_features': 0.5, 'min_samples_leaf': 5, 'min_samples_split': 5, 'n_estimators': 200}

      Best ROC-AUC Score: 0.9342

6. Model Building & Evaluation

Model	Validation Accuracy	Sensitivity (TPR)	Specificity (TNR)	Precision	Recall	F1-Score
Logistic Regression Optimized cutoff (0.5282)	0.8000	0.6757	0.8407	0.5814	0.6757	0.6250
Random Forest (Hyperparameter Tuning)	0.7233	0.3378	0.8496	0.4237	0.3378	0.3759

Logistic Regression (Optimized cutoff at 0.5282)

- Achieves 80.00% validation accuracy
- Shows good sensitivity/recall at 67.57% (effectively captures true positives)
- Maintains high specificity at 84.07% (effectively identifies true negatives)
- Delivers precision of 58.14% (moderate confidence in positive predictions)
- Results in F1-Score of 62.50% (balanced performance between precision and recall)

Random Forest

- Reaches 72.33% validation accuracy
- Demonstrates poor sensitivity at only 33.78% (misses many positive cases)
- Maintains high specificity at 84.96% (slightly better than Logistic Regression)
- Shows lower precision at 42.37% (less confidence in positive predictions)
- Results in a substantially lower F1-Score of 37.59%

7. Conclusions & Recommendations

- **Best Model:** Logistic Regression with threshold tuning significantly outperforms Random Forest, showing:
  - Nearly double the sensitivity (67.57% vs 33.78%)
  - Higher precision (58.14% vs 42.37%)
  - Substantially better F1-score (62.50% vs 37.59%)
  - Better overall accuracy (80.00% vs 72.33%)
- **Deployment Suggestion:**
  - Implement the optimized threshold of 0.5282 for flagging potentially fraudulent claims
  - Integrate into claims processing pipeline to trigger manual review for flagged cases
  - Balance false positives (41.86% of flagged claims) against the benefit of catching 67.57% of truly fraudulent claims