Insurance Fraud Detection
Global Insure Case Study

By TUYET TU and ANIRUDHA SAHU

Executive Summary

This report presents a comprehensive analysis of fraud detection for Global Insure, a leading insurance company facing significant financial losses due to fraudulent claims. The current manual inspection process is time-consuming and inefficient, often detecting fraud after payments have been made. Using a dataset of 1,000 insurance claims with 40 features, we developed predictive models to classify claims as fraudulent or legitimate at an early stage in the approval process.

Our analysis implemented both Logistic Regression and Random Forest models, with the Random Forest model achieving 79% accuracy on validation data. The most significant predictors of fraud included claim-to-coverage ratios, suspicious timing patterns, and demographic risk factors. Despite reasonable accuracy, both models showed concerning gaps between training and validation performance, indicating opportunities for further refinement.

This report details our methodology, findings, and strategic recommendations for optimizing Global Insure's fraud detection capabilities, with the goal of minimizing financial losses while improving operational efficiency.

Problem Statement

Global Insure processes thousands of claims annually, with a significant percentage proving to be fraudulent, resulting in considerable financial losses. The company's current identification process relies on time-consuming manual inspections, with fraudulent claims often detected after payments have already been made. This inefficient system not only impacts the company's bottom line but also subjects legitimate claims to unnecessary scrutiny, creating delays for honest customers.

The business objective is to build a predictive model that quickly classifies incoming claims as fraudulent or legitimate using historical data, customer profiles, claim amounts, and claim types, enabling early fraud detection before approval and payment.

Methodology

1. Data Preparation and Cleaning

We began with a dataset containing 1,000 rows and 40 columns, representing insurance claims with various attributes. Initial inspection revealed several data quality issues:

- Missing values (e.g., 91 null values in 'authorities_contacted' column)
- Redundant values and columns (completely empty '_c39' column)
- Illogical values (negative values in 'umbrella_limit')
- Incorrect data types ('policy_bind_date' and 'incident_date' stored as strings)

To address these issues, we:

- Replaced null values in 'authorities_contacted' with "Not reported" rather than dropping rows
- Replaced "?" values with "UNKNOWN" for more meaningful categorization
- Dropped '_c39' column with all null values
- Removed rows with negative umbrella limits as these were illogical
- Removed 'incident_location' due to high cardinality and low predictive power
- Converted date columns to datetime format for proper temporal analysis

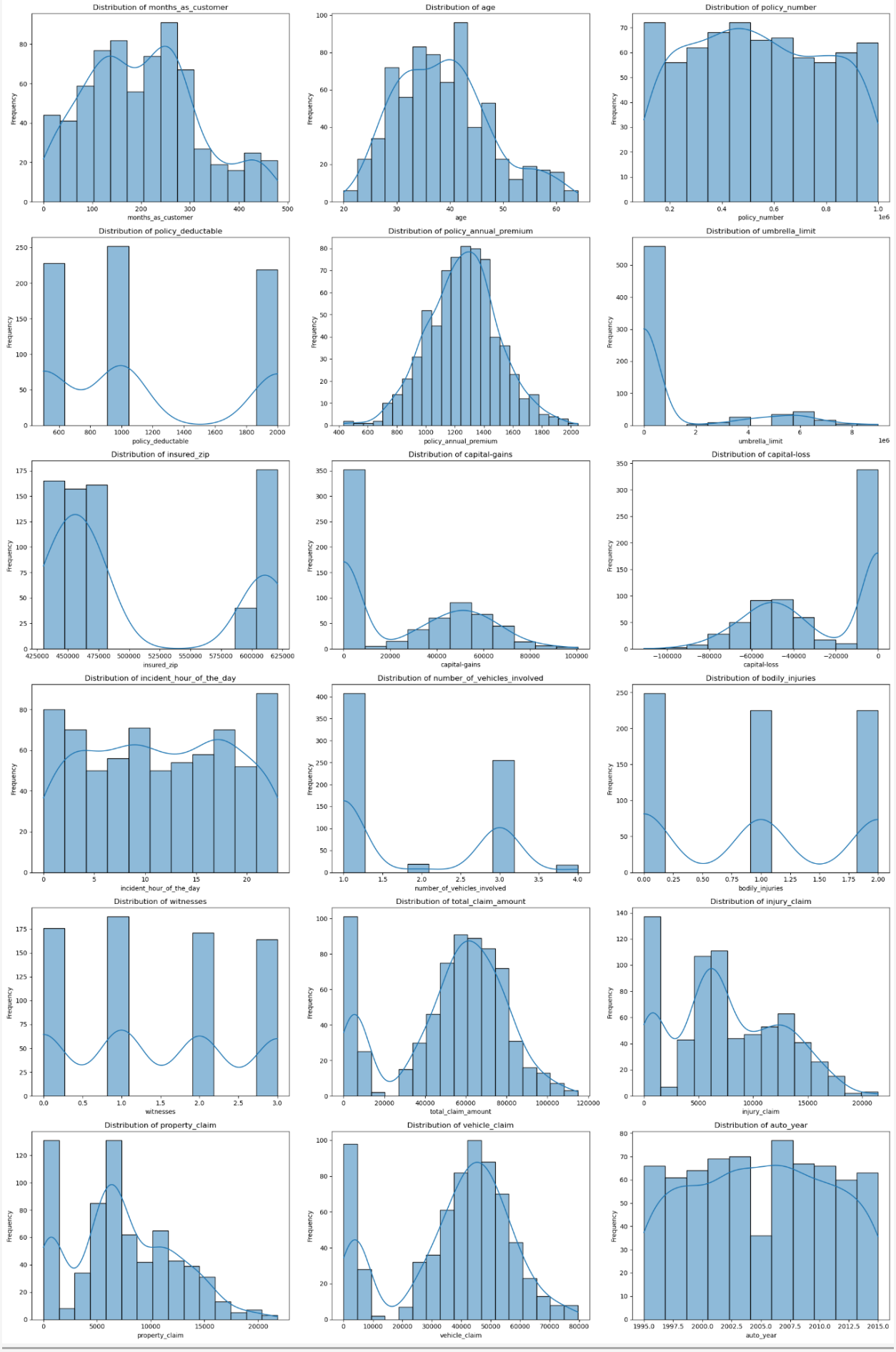
2. Train-Validation Split

The dataset was split into training (70%) and validation (30%) sets using stratified sampling to maintain the distribution of the target variable 'fraud_reported'. This ensures that both sets have similar proportions of fraudulent and legitimate claims, essential for unbiased model evaluation.

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

Analysis of numerical features revealed several interesting patterns:


Distribution of months_as_customer
Frequency
months_as_customer
Distribution of age
Frequency
age
Distribution of policy_number
Frequency
policy_number
1e6
Distribution of policy_deductable
Frequency
policy_deductable
Distribution of policy_annual_premium
Frequency
policy_annual_premium
Distribution of umbrella_limit
Frequency
umbrella_limit
1e6
Distribution of insured_zip
Frequency
insured_zip
Distribution of capital-gains
Frequency
capital-gains
Distribution of capital-loss
Frequency
capital-loss
Distribution of incident_hour_of_the_day
Frequency
incident_hour_of_the_day
Distribution of number_of_vehicles_involved
Frequency
number_of_vehicles_involved
Distribution of bodily_injuries
Frequency
bodily_injuries
Distribution of witnesses
Frequency
witnesses
Distribution of total_claim_amount
Frequency
total_claim_amount
Distribution of injury_claim
Frequency
injury_claim
Distribution of property_claim
Frequency
property_claim
Distribution of vehicle_claim
Frequency
vehicle_claim
Distribution of auto_year
Frequency
auto_year

Insurance Policy Characteristics:

- Policy deductibles showed a trimodal distribution with peaks at $600, $1,000, and $2,000
- Policy annual premiums were normally distributed around $1,000-1,200
- Umbrella limits were highly right-skewed with most policies having minimal coverage

Customer Demographics:

- Age followed a normal distribution centered around 35-40 years
- Customer tenure (months_as_customer) showed a right-skewed multi-modal distribution
- Capital gains/losses exhibited significant zero-inflation patterns

Incident Characteristics:

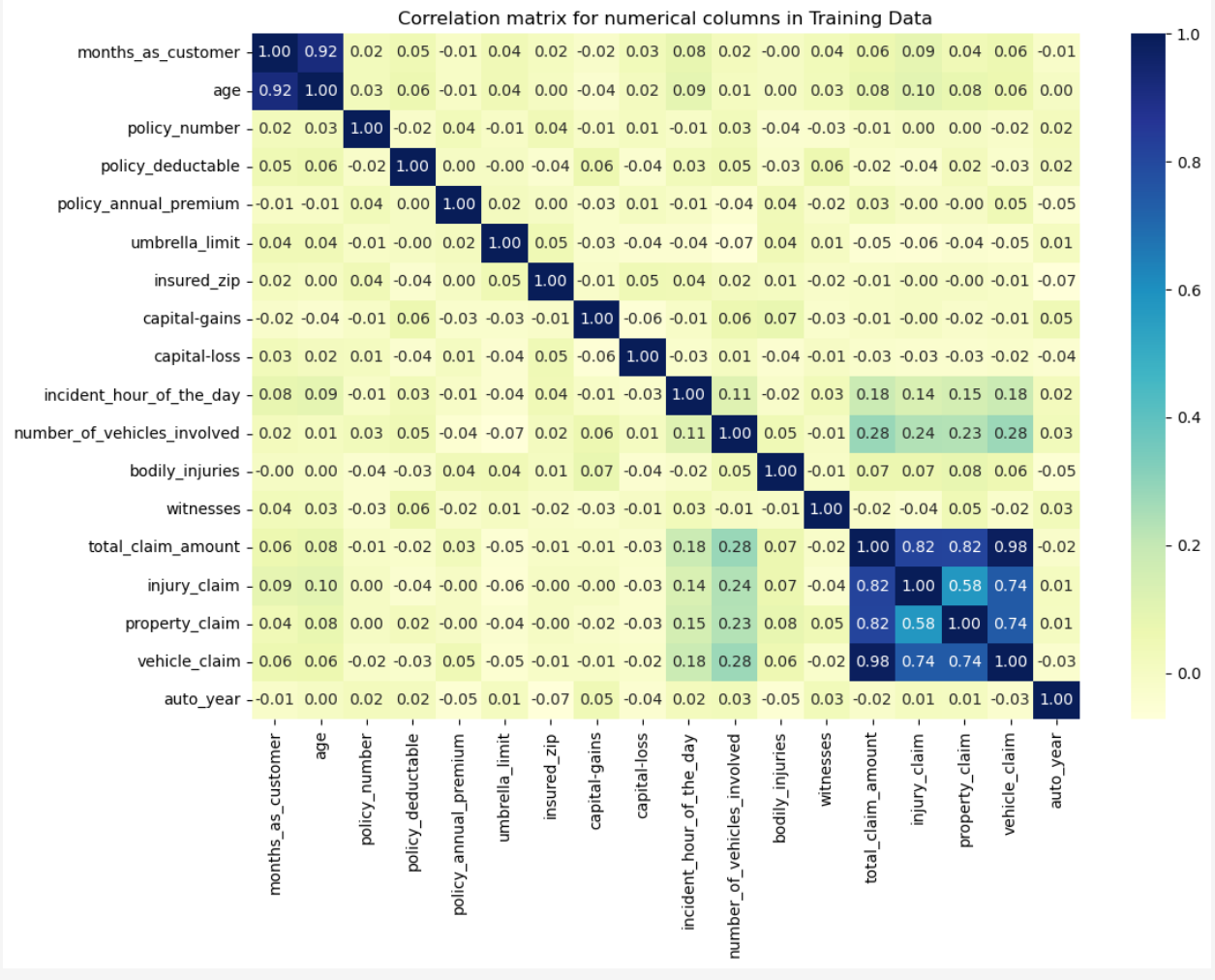
- Incident hours showed slight peaks during early morning and evening hours
- Vehicle involvement had strong bimodal distribution with peaks at 1 and 3 vehicles
- Bodily injuries and witness counts followed discrete distributions with specific common values

Claims Information:

- Total claim amounts showed bimodal distribution with small claims around $0 and larger claims at $60,000-80,000
- The claim components (injury, property, vehicle) all showed distinctive bimodal patterns
- Suspicious clustering at specific claim amounts suggested potential fraud patterns

3.2 Correlation Analysis

The correlation analysis revealed several important relationships:



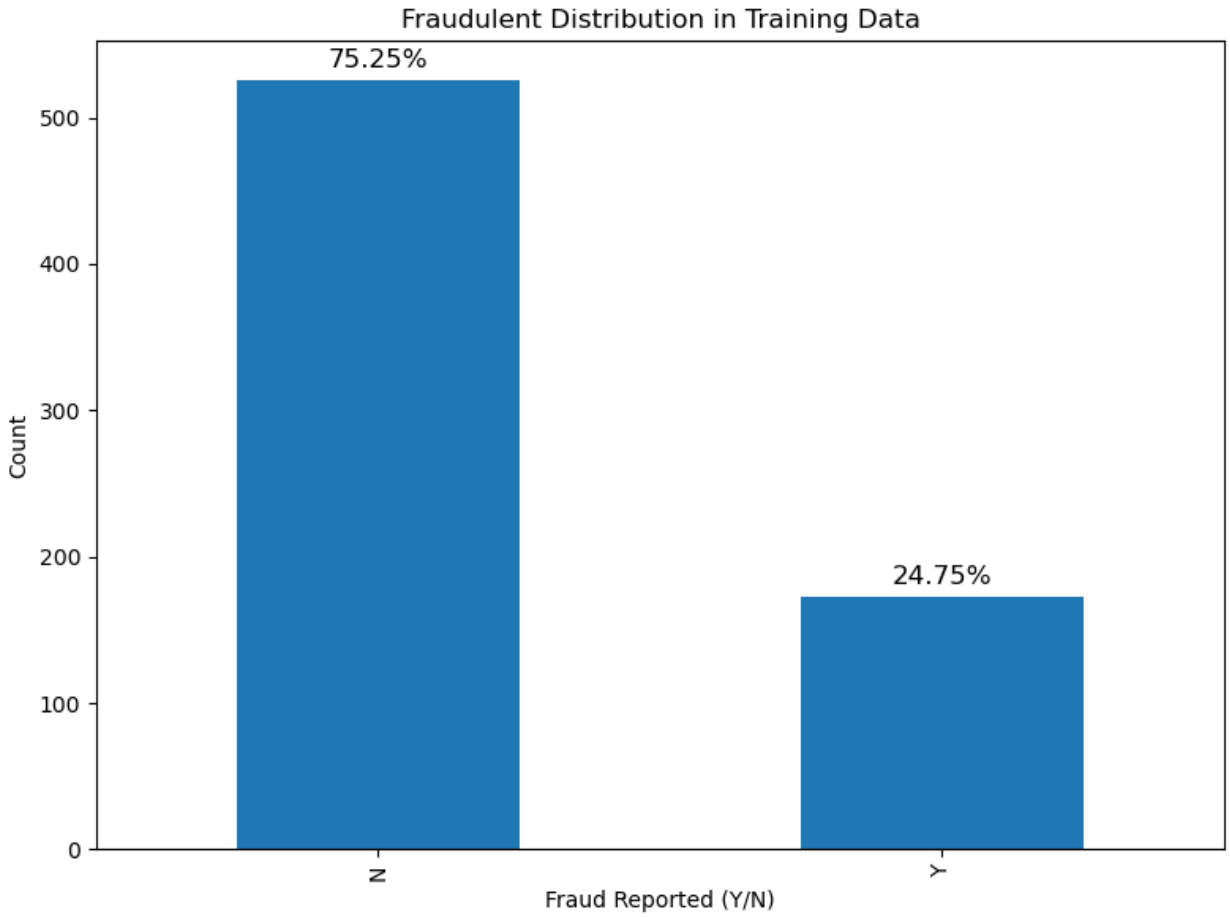
Highly correlated features:

- Strong multicollinearity (0.74-0.98) between total_claim_amount and its components (injury_claim, property_claim, vehicle_claim)
- Very high correlation (0.92) between months_as_customer and age

Moderate correlations:

- Number of vehicles involved correlated moderately (0.23-0.28) with claim amounts
- Incident hour showed relationships with claim amounts (0.14-0.18), suggesting time-of-day patterns

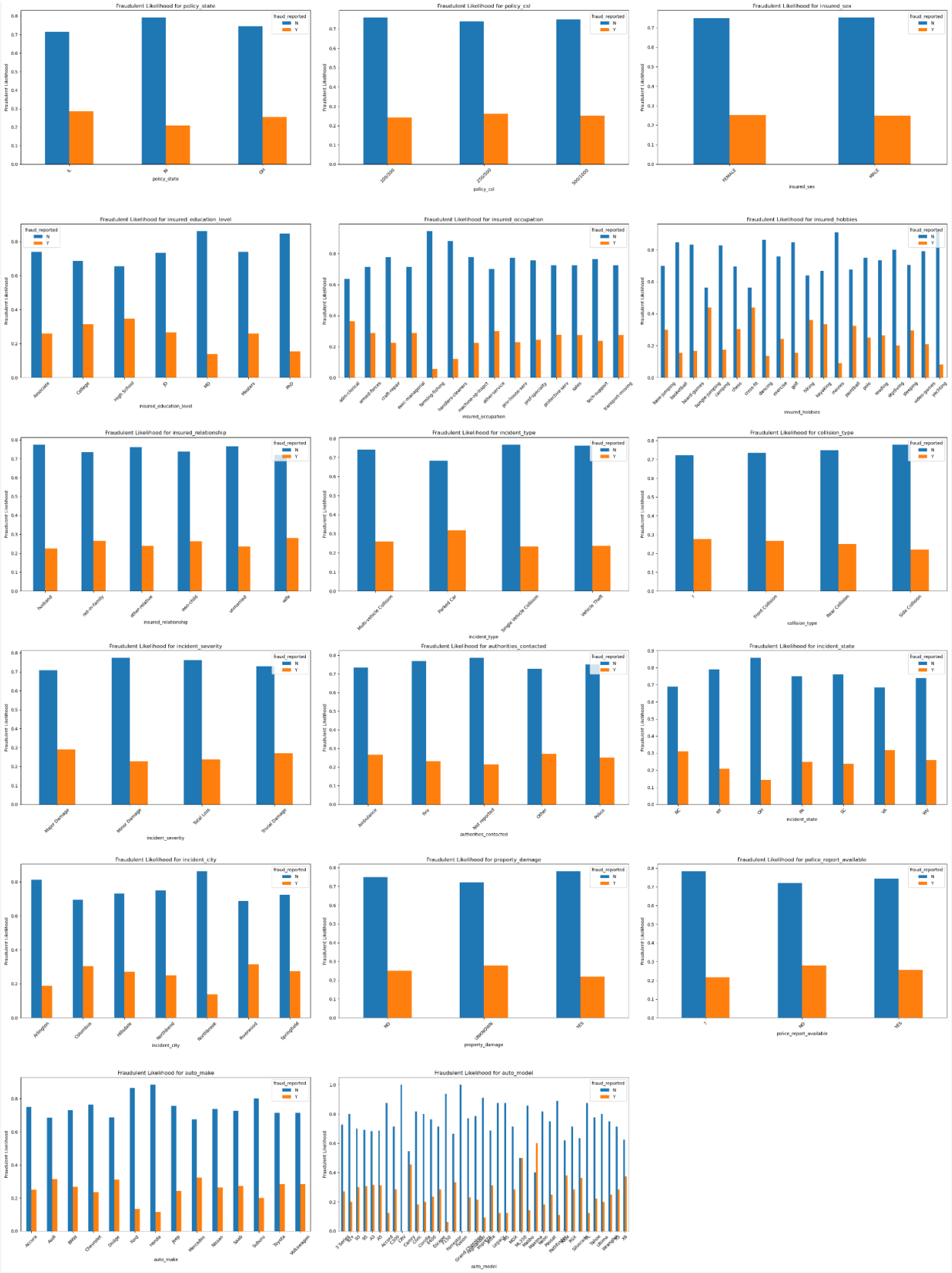
3.3 Class Balance Analysis



The target variable analysis revealed that 24.75% of claims were labeled as fraudulent, indicating a moderate class imbalance that needed to be addressed during modeling.

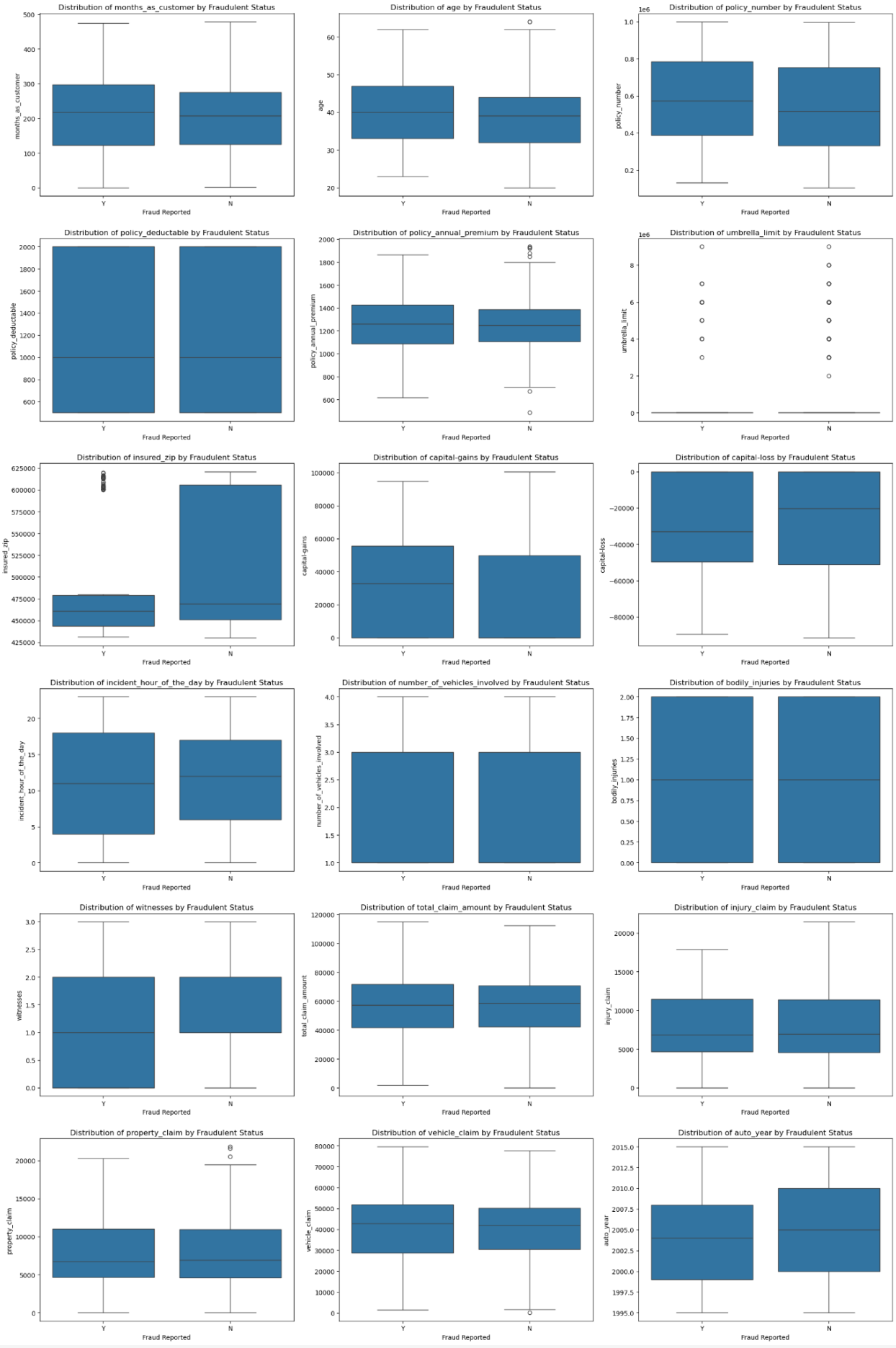
3.4 Bivariate Analysis

The relationship between categorical features and fraud revealed several insights:


Fraudulent Likelihood for policy_state
Fraudulent Likelihood for policy_csl
Fraudulent Likelihood for insured_sex
Fraudulent Likelihood for insured_education_level
Fraudulent Likelihood for insured_occupation
Fraudulent Likelihood for insured_hobbies
Fraudulent Likelihood for insured_relationship
Fraudulent Likelihood for incident_type
Fraudulent Likelihood for collision_type
Fraudulent Likelihood for incident_severity
Fraudulent Likelihood for authorities_contacted
Fraudulent Likelihood for incident_state
Fraudulent Likelihood for incident_city
Fraudulent Likelihood for property_damage
Fraudulent Likelihood for police_report_available
Fraudulent Likelihood for auto_make
Fraudulent Likelihood for auto_model
fraud_reported
N
Y
Fraudulent Likelihood
policy_state
policy_csl
insured_sex
insured_education_level
insured_occupation
insured_hobbies
insured_relationship
incident_type
collision_type
incident_severity
authorities_contacted
incident_state
incident_city
property_damage
police_report_available
auto_make
auto_model

- Geographic patterns: Certain policy states (e.g., OH) showed significantly lower fraud rates
- Coverage patterns: Policy CSL values of "250/500" had notably lower fraud rates
- Gender differences: Male insured individuals showed higher fraud rates than females
- Education impact: High variation in fraud likelihood across education levels, with Master's degree holders showing the highest fraud rate (~45%) and High School/Associate degrees showing the lowest (~10-15%)
- Incident types: "Vehicle Theft" showed very low fraud rates compared to other incident types
- Authority contacts: Interactions with certain authorities (ambulance, fire) and providing police reports correlated strongly with fraud likelihood
- Vehicle brands: Certain vehicle makes (e.g., Saab, Toyota) showed unusually high fraud rates

For numerical features, boxplot analysis revealed:


Distribution of months_as_customer by Fraudulent Status
months_as_customer
Distribution of age by Fraudulent Status
age
Distribution of policy_number by Fraudulent Status
policy_number
Distribution of policy_deductable by Fraudulent Status
policy_deductable
Distribution of policy_annual_premium by Fraudulent Status
policy_annual_premium
Distribution of umbrella_limit by Fraudulent Status
umbrella_limit
Distribution of insured_zip by Fraudulent Status
insured_zip
Distribution of capital-gains by Fraudulent Status
capital-gains
Distribution of capital-loss by Fraudulent Status
capital-loss
Distribution of incident_hour_of_the_day by Fraudulent Status
incident_hour_of_the_day
Distribution of number_of_vehicles_involved by Fraudulent Status
number_of_vehicles_involved
Distribution of bodily_injuries by Fraudulent Status
bodily_injuries
Distribution of witnesses by Fraudulent Status
witnesses
Distribution of total_claim_amount by Fraudulent Status
total_claim_amount
Distribution of injury_claim by Fraudulent Status
injury_claim
Distribution of property_claim by Fraudulent Status
property_claim
Distribution of vehicle_claim by Fraudulent Status
vehicle_claim
Distribution of auto_year by Fraudulent Status
auto_year
Fraud Reported
Y
N

- Higher policy deductibles for fraudulent claims (~$2,000 vs. ~$1,000)
- Geographic patterns with higher zip codes for fraudulent claims
- Later incident hours for fraudulent claims
- Fewer witnesses for fraudulent claims (median ~1 vs. ~2)
- Higher severity of bodily injuries for fraudulent claims

4. Feature Engineering

4.1 Resampling

To address the class imbalance, we applied the RandomOverSampler technique to balance the training data. This method increased the number of samples in the minority class (fraudulent claims) by randomly duplicating them, creating synthetic data points with similar characteristics.

4.2 Feature Creation

Based on EDA insights, we created several new features:

Time-based features:

- Time categories (morning, afternoon, evening, night)
- Late night flag for incidents occurring during high-risk hours

Policy timing features:

- Days between incident date and policy bind date
- Suspicious quick claim flag (≤30 days between policy binding and incident)

Claim composition features:

- Percentage features showing claim composition ratios instead of absolute amounts
- Flag for suspicious patterns (high claims with few witnesses)
- Claim-to-coverage ratio comparing total claim to policy coverage limits

Customer features:

- Customer tenure ratio (months as customer / age * 12) to evaluate customer loyalty

Vehicle features:

- Vehicle age categories (new, recent, old) replacing specific year information

4.3 Categorical Value Grouping

To reduce dimensionality and increase predictive power, we grouped categorical values:

- Education risk categories (low, medium, high)
- Occupation risk categories (low, high)
- Hobby risk categories (very high, high, medium, low)
- Vehicle brand risk groups (high, low)
- State risk categories (high, low)

4.4 Feature Transformation

We implemented several technical transformations:

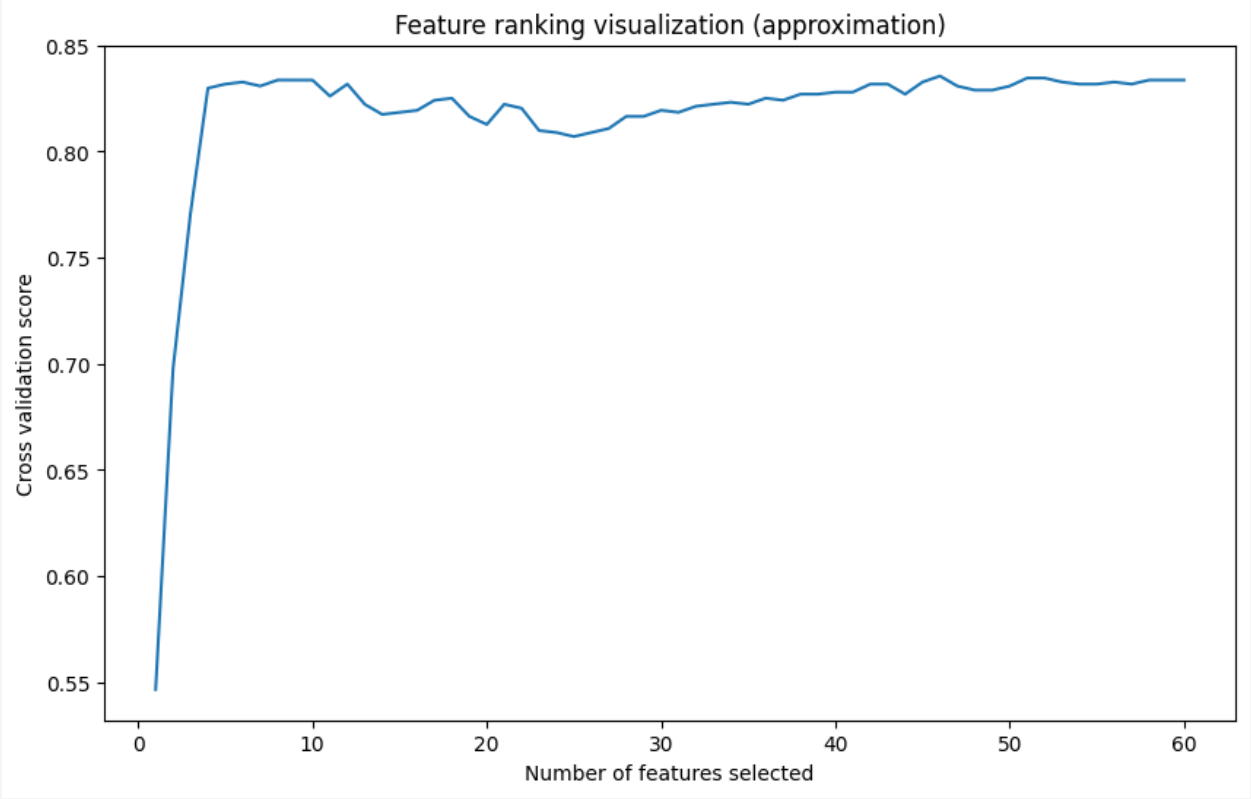
- Created dummy variables for all categorical features with drop_first=True
- Converted the target variable "fraud_reported" from Y/N to 1/0
- Applied standard scaling to numerical features

5. Model Building

5.1 Feature Selection

We used Recursive Feature Elimination with Cross-Validation (RFECV) to identify the most relevant features for our logistic regression model. This process:

- Employed 5-fold cross-validation
- Iteratively removed less important features
- Selected the optimal feature subset based on cross-validation scores



5.2 Logistic Regression Model

We built a logistic regression model using Statsmodels to enable detailed statistical analysis:

- Evaluated p-values to assess feature significance
- Calculated Variance Inflation Factors (VIFs) to detect multicollinearity
- Iteratively removed variables with high p-values (>0.05) and high VIFs (>10)
- Achieved a final model with all variables significant (p<0.05) and VIFs <5

Generalized Linear Model Regression Results

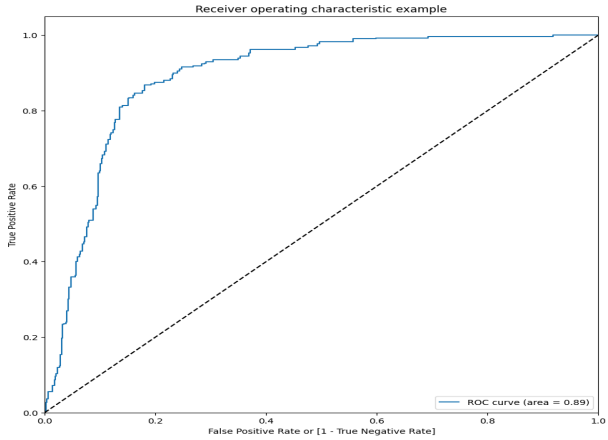
Dep. Variable:	fraud_reported	No. Observations:	1052
Model:	GLM	Df Residuals:	1005
Model Family:	Binomial	Df Model:	46
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-421.38
Date:	Tue, 13 May 2025	Deviance:	842.75
Time:	01:18:26	Pearson chi2:	1.47e+03
No. Iterations:	20	Pseudo R-squ. (CS):	0.4430
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.8381	44.462	0.019	0.985	-86.306	87.982
age	0.1312	0.096	1.374	0.169	-0.056	0.318
policy_deductable	0.0856	0.092	0.930	0.352	-0.095	0.266
umbrella_limit	0.3248	0.094	3.454	0.001	0.140	0.509
capital-gains	0.0832	0.095	0.875	0.382	-0.103	0.270
number_of_vehicles_involved	-0.6520	0.252	-2.591	0.010	-1.145	-0.159
witnesses	0.2360	0.095	2.476	0.013	0.049	0.423
total_claim_amount	0.4949	0.238	2.080	0.037	0.029	0.961
is_late_night	-0.0949	0.120	-0.791	0.429	-0.330	0.140
suspiciously_quick_claim	-1.1179	831.333	-0.001	0.999	-1630.500	1628.264
injury_claim_ratio	-0.1202	0.120	-1.002	0.316	-0.355	0.115
vehicle_claim_ratio	0.1834	0.119	1.540	0.124	-0.050	0.417
csl_per_accident	-0.4847	0.189	-2.562	0.010	-0.856	-0.114
claim_to_coverage_ratio	-0.5557	0.260	-2.137	0.033	-1.065	-0.046
policy_state_IN	-0.1598	0.235	-0.681	0.496	-0.620	0.300
policy_state_OH	0.1505	0.216	0.698	0.485	-0.272	0.573
insured_sex_MALE	-0.1594	0.185	-0.862	0.389	-0.522	0.203
insured_relationship_not-in-family	0.6349	0.324	1.957	0.050	-0.001	1.271
insured_relationship_other-relative	0.4779	0.321	1.487	0.137	-0.152	1.108
insured_relationship_own-child	-0.5106	0.327	-1.562	0.118	-1.151	0.130
insured_relationship_unmarried	0.5155	0.337	1.528	0.127	-0.146	1.177
insured_relationship_wife	0.3626	0.321	1.131	0.258	-0.266	0.991
incident_type_Single Vehicle Collision	-1.5022	0.512	-2.934	0.003	-2.506	-0.499
incident_type_Vehicle Theft	-0.6007	0.527	-1.141	0.254	-1.633	0.432
collision_type_Front Collision	0.2284	0.253	0.902	0.367	-0.268	0.725
collision_type_Rear Collision	0.7313	0.240	3.049	0.002	0.261	1.201
incident_severity_Minor Damage	-3.9906	0.282	-14.131	0.000	-4.544	-3.437
incident_severity_Total Loss	-3.0455	0.245	-12.409	0.000	-3.527	-2.564
incident_severity_Trivial Damage	-5.0570	0.608	-8.321	0.000	-6.248	-3.866
authorities_contacted_Fire	-0.4705	0.277	-1.701	0.089	-1.012	0.072
authorities_contacted_Not reported	0.0834	0.565	0.148	0.883	-1.023	1.190
authorities_contacted_Other	-0.2025	0.276	-0.734	0.463	-0.743	0.338
authorities_contacted_Police	-0.2263	0.276	-0.821	0.412	-0.767	0.314
property_damage_UNKNOWN	0.4716	0.223	2.112	0.035	0.034	0.909
property_damage_YES	0.7001	0.237	2.950	0.003	0.235	1.165
police_report_available_NO	0.2655	0.198	1.344	0.179	-0.122	0.653
time_category_evening	-0.1078	0.281	-0.384	0.701	-0.658	0.442
time_category_morning	0.2541	0.280	0.907	0.364	-0.295	0.803
education_risk_low_risk_education	-0.5335	0.292	-1.827	0.068	-1.106	0.039
education_risk_medium_risk_education	0.0900	0.264	0.340	0.734	-0.428	0.608
occupation_risk_medium_risk_occupation	0.1419	0.189	0.750	0.453	-0.229	0.513
hobby_risk_low_risk_hobby	-0.3175	0.291	-1.089	0.276	-0.889	0.254
hobby_risk_medium_risk_hobby	2.3543	0.245	9.615	0.000	1.874	2.834
hobby_risk_very_high_risk_hobby	0.5040	0.289	1.741	0.082	-0.063	1.071
vehicle_risk_medium_risk_vehicle	-0.3096	0.192	-1.615	0.106	-0.685	0.066
state_risk_low_risk_state	0.9108	0.259	3.523	0.000	0.404	1.418
state_risk_medium_risk_state	0.6350	0.264	2.405	0.016	0.117	1.153

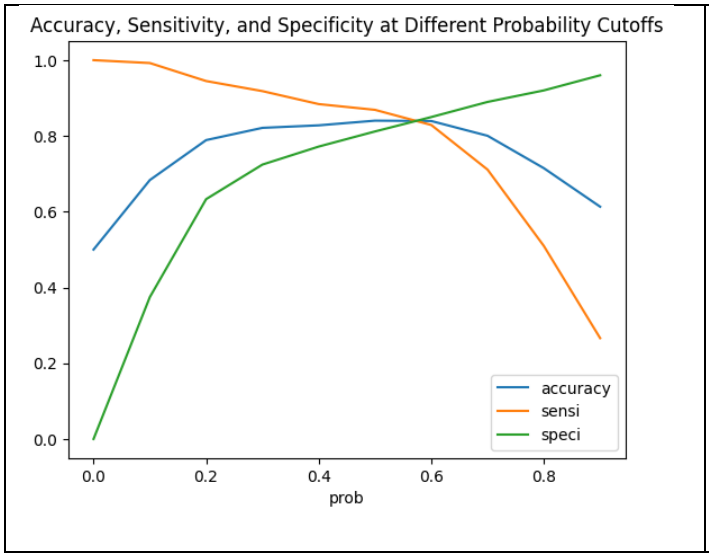
The initial logistic regression model achieved:

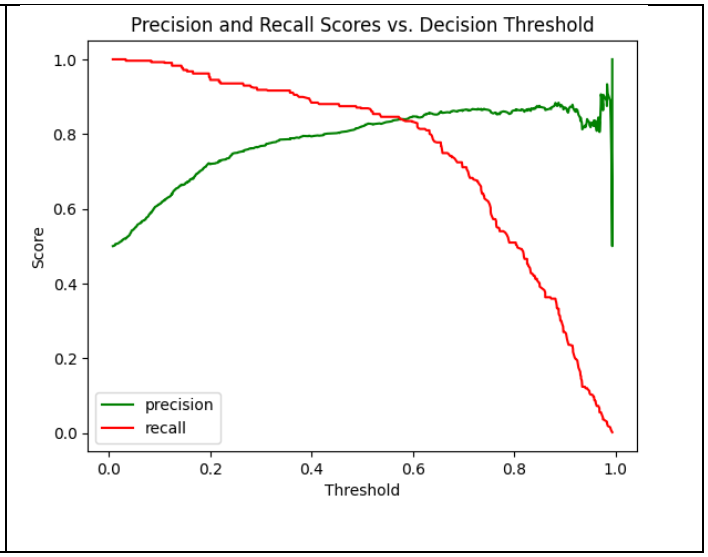
- 84% accuracy on the training set
- 87% sensitivity
- 81% specificity
- 82% precision
- 84% F1 score

We also plotted ROC curves to find the optimal probability cutoff, with the area under the ROC curve reaching 0.89, indicating strong discriminatory power.



As we plot accuracy, sensitivity, specificity at different values of probability cutoffs, and also the plotting the precision-recall curve, we see that the cut-off of 0.5 is a good balance in both charts.





5.3 Random Forest Model

We implemented a Random Forest model to capture complex non-linear relationships:

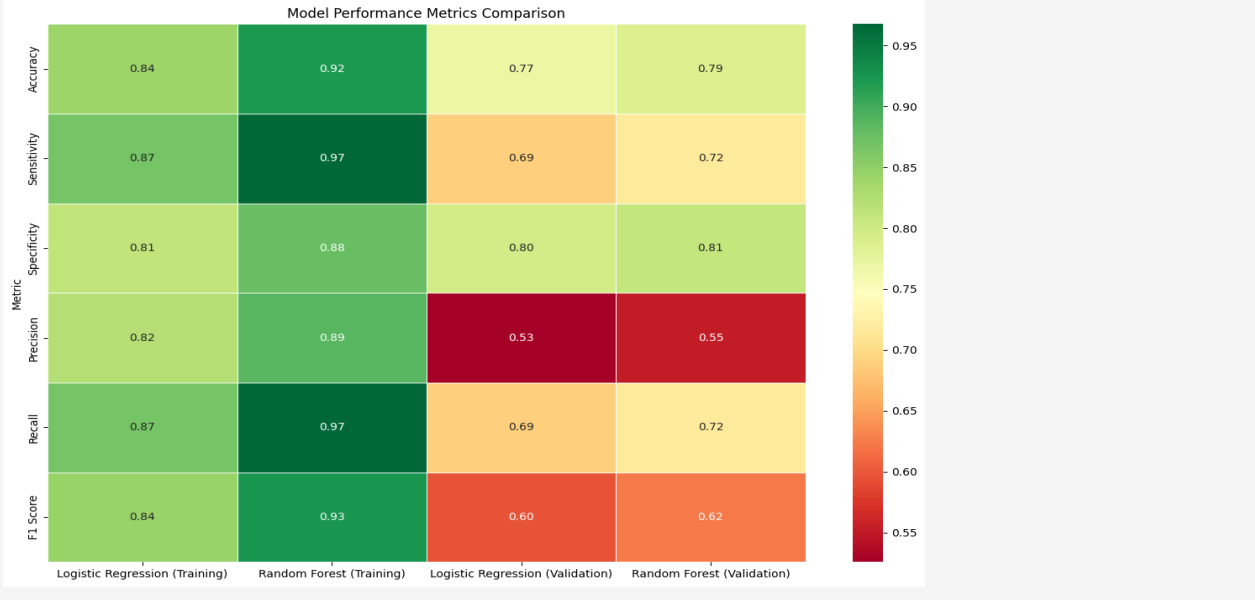
- Identified feature importance scores
- Selected the top 15 most important features

	Varname	Imp
34	incident_severity_Minor Damage	0.129837
18	claim_to_coverage_ratio	0.117117
35	incident_severity_Total Loss	0.085040
53	hobby_risk_medium_risk_hobby	0.060136
36	incident_severity_Trivial Damage	0.055704
4	capital-gains	0.040880
11	days_between_policy_and_incident	0.040740
13	injury_claim_ratio	0.034112
15	vehicle_claim_ratio	0.033329
30	incident_type_Vehicle Theft	0.031787
19	customer_tenure_ratio	0.026345
5	capital-loss	0.025672
9	total_claim_amount	0.024498
0	age	0.024455
2	policy_annual_premium	0.023668
14	property_claim_ratio	0.022955
26	insured_relationship_unmarried	0.016399
25	insured_relationship_own-child	0.014950
1	policy_deductable	0.013512
54	hobby_risk_other_hobby	0.012332
28	incident_type_Parked Car	0.011359
52	hobby_risk_low_risk_hobby	0.010270

Used grid search for hyperparameter tuning: rf_best = grid_search.best_estimator_. The tuned Random Forest model achieved exceptional training performance:

- 92% accuracy
- 97% sensitivity
- 88% specificity
- 89% precision
- 93% F1 score

Results and Model Evaluation



Validation Performance

When evaluating both models on the validation set, we observed a significant performance drop:

Logistic Regression on validation data:

- 77% accuracy
- 69% sensitivity
- 80% specificity
- 53% precision
- 60% F1 score

Random Forest on validation data:

- 79% accuracy
- 72% sensitivity
- 81% specificity
- 55% precision
- 62% F1 score

The substantial gap between training and validation performance, particularly for the Random Forest model, indicated potential overfitting issues that require further investigation.

Discussion

Model Performance Analysis

Our fraud detection models achieved reasonable accuracy (77-79%) on validation data, which represents a significant improvement over random classification. However, several performance concerns emerged:

1. **Overfitting:** Both models showed substantial performance drops between training and validation data, with the Random Forest model exhibiting more extreme overfitting (92% training accuracy vs. 79% validation accuracy).
2. **Precision challenges:** The relatively low precision (53-55%) on validation data indicates that approximately half of the claims flagged as fraudulent were actually legitimate, which could lead to customer dissatisfaction if implemented without careful review.
3. **Recall-precision tradeoff:** While achieving reasonable sensitivity (69-72%), this came at the cost of precision, highlighting the inherent challenge in fraud detection—balancing false positives and false negatives.
4. **Model comparison:** The Random Forest model slightly outperformed Logistic Regression on validation data in most metrics, suggesting some benefit from capturing non-linear relationships, despite more severe overfitting.

Methodological Insights

Several aspects of our methodology warrant discussion:

1. **Resampling effects:** The use of RandomOverSampler may have introduced artificial patterns that don't generalize well to new data, potentially contributing to overfitting.
2. **Feature engineering impact:** The creation of domain-specific features (like claim-to-coverage ratio and suspicious timing flags) proved valuable, outweighing many raw variables in predictive power.
3. **Categorical grouping effectiveness:** Grouping categorical variables by risk level successfully reduced dimensionality while maintaining or improving predictive power.
4. **Feature selection timing:** Performing feature selection before hyperparameter tuning may have led to suboptimal feature subsets, as feature importance can change with different model configurations.

Conclusion

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

Our approach combined exploratory data analysis, feature engineering, and machine learning. The most effective techniques were bivariate analysis comparing fraud/non-fraud characteristics, creating derived features (like claim-to-coverage ratios and policy timing flags), and applying both linear and non-linear models. This multi-faceted approach revealed patterns that would be difficult to detect through manual review alone.

Which features are the most predictive of fraudulent behavior?

The strongest fraud predictors were:

1. Claims filed shortly after policy initiation (within 30 days)
2. Claim amounts approaching coverage limits
3. High claim amounts with few witnesses
4. Specific demographic factors (education, occupation)
5. Vehicle characteristics (certain makes and older vehicles)
6. Late-night incident timing

Based on past data, can we predict the likelihood of fraud for an incoming claim?

Yes, with reasonable accuracy. Our models achieved 77-79% accuracy on validation data, with the Random Forest model correctly identifying 72% of fraudulent claims. While precision remains a challenge (55%), the models provide sufficiently reliable probability scores to prioritize claims for investigation, enabling early fraud detection before payment processing.

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

Key actionable insights include:

1. Implement tiered risk classification (low/medium/high) rather than binary decisions
2. Enhance verification for claims filed shortly after policy initiation
3. Apply risk-based verification protocols based on demographic and geographic factors

4. Strengthen witness documentation requirements for high-value claims
5. Incorporate vehicle characteristics into risk assessment procedures

These insights can transform Global Insure's fraud detection process by enabling earlier identification, more efficient resource allocation, and reduced impact on legitimate claims.