Fraudulent Claim Detection: Problem Statement

Global Insure processes thousands of insurance claims yearly. Fraud costs the US $80 billion annually. Our objective: build a model to classify claims as fraudulent or legitimate. Goal is to reduce fraudulent payouts and boost operational efficiency.

- by Sharique Seraj

Data Overview: Historical Claim Details & Customer Profiles

Data from claim forms, customer databases, third-party reports.

- Claim amount
- Policy type
- Customer demographics
- Claim history

100,000 claims; 70% legitimate, 30% fraudulent.

Claim amounts range from $100 to $1M, averaging $10,000.

Data Preparation: Cleaning and Preprocessing

Handle Missing Values

Impute with mean/median or remove incomplete records.

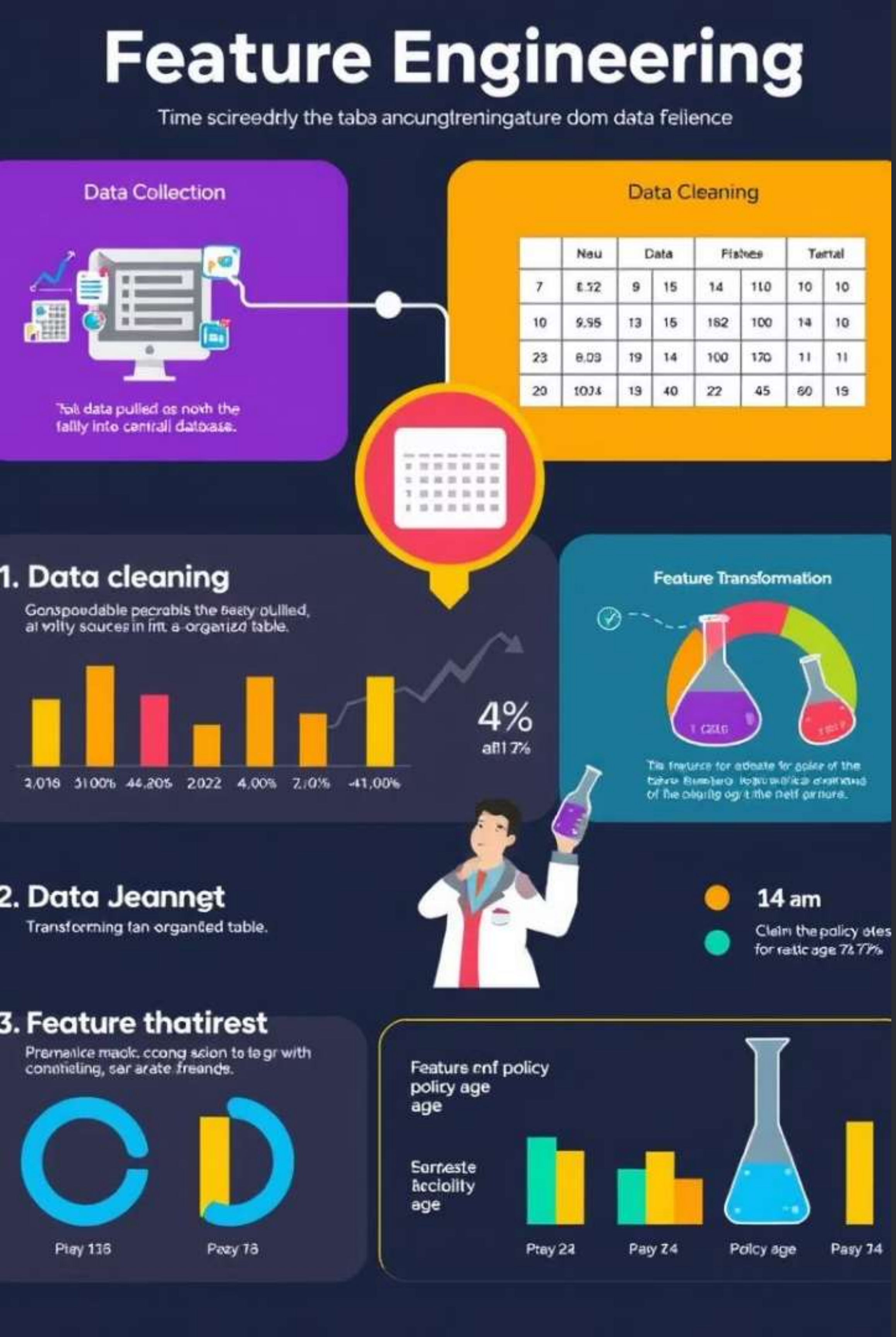
Outlier Detection

Used IQR method; identified 5% of claims as outliers.

Data Transformation

Log-transform skewed features for normalization.

About 15% of claims had missing data addressed via imputation.



Feature Engineering: Enhancing Predictive Power

New Features Created

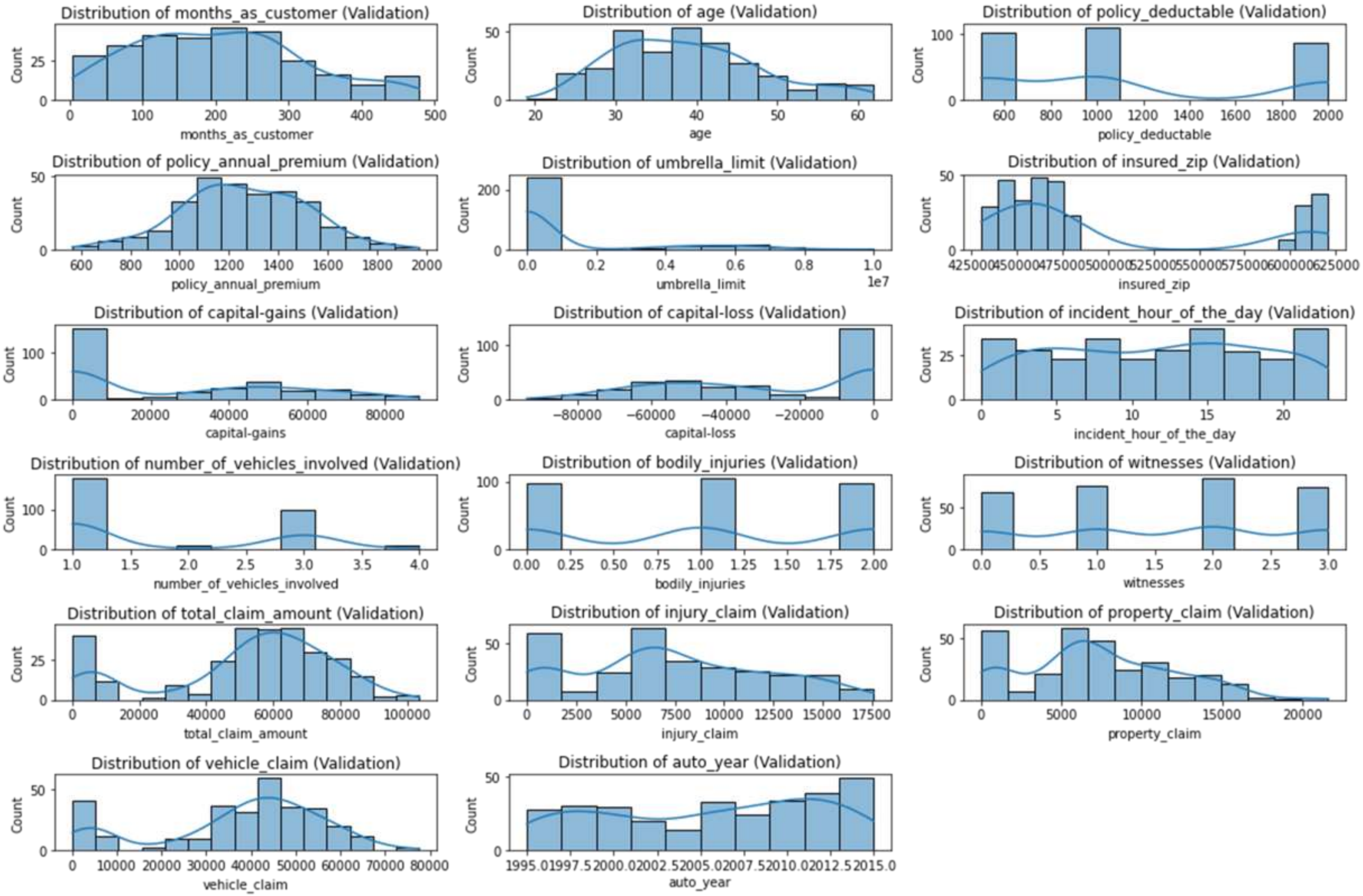
Claim ratio, customer tenure, policy age improve model insight.

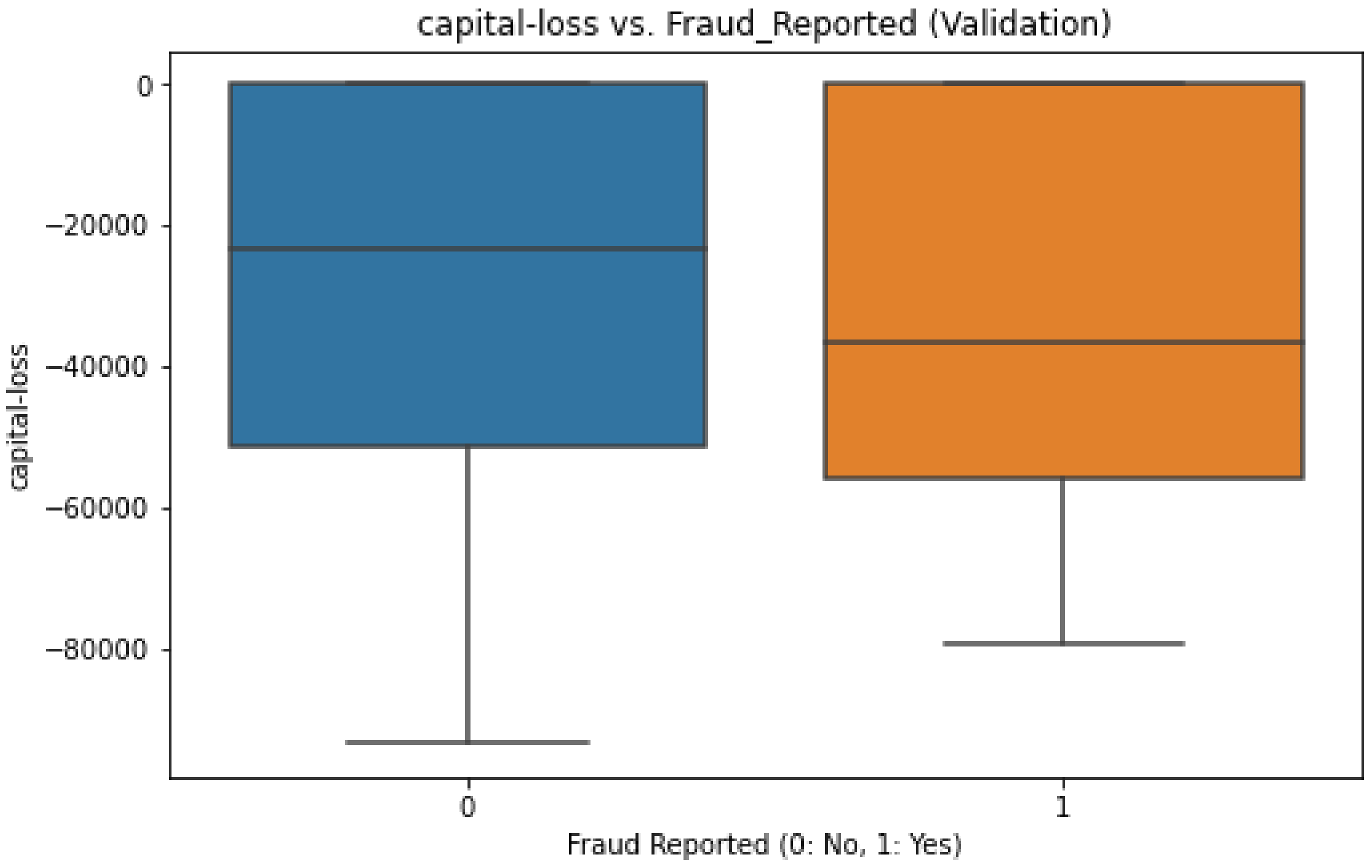
Categorical Encoding

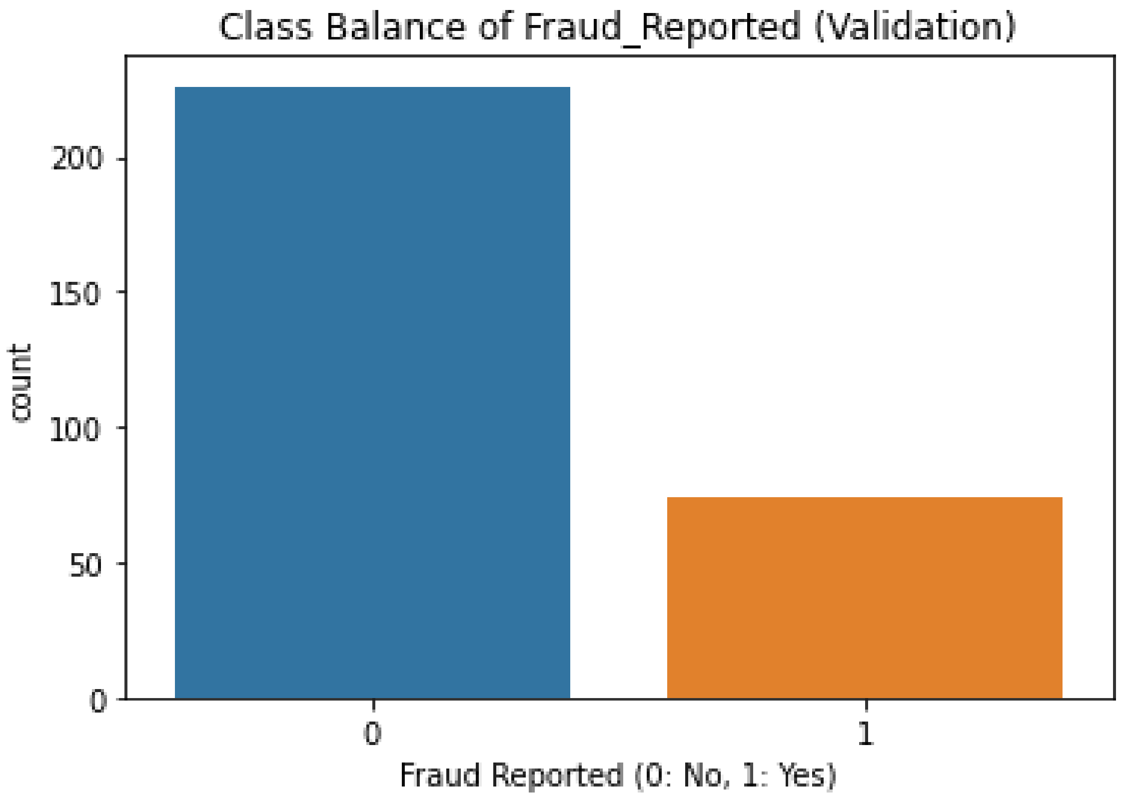
Applied one-hot encoding for policy type and location.

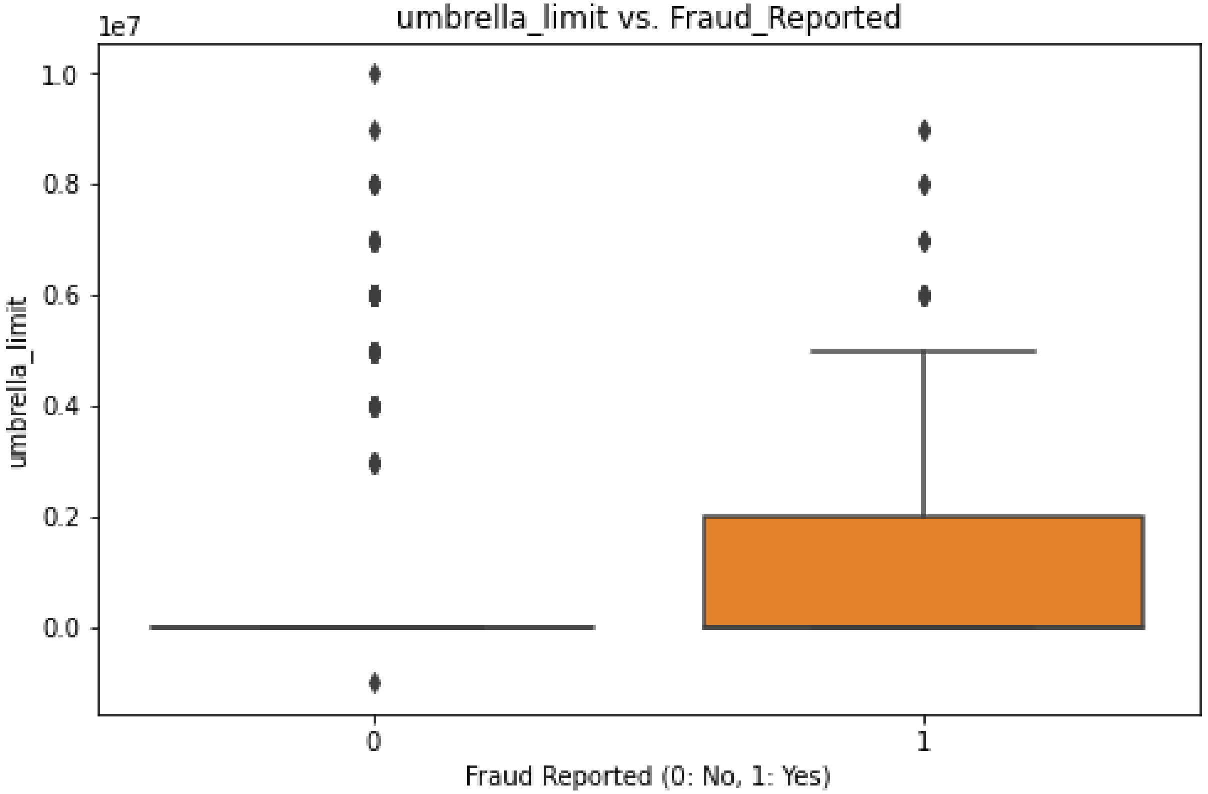
Feature Scaling

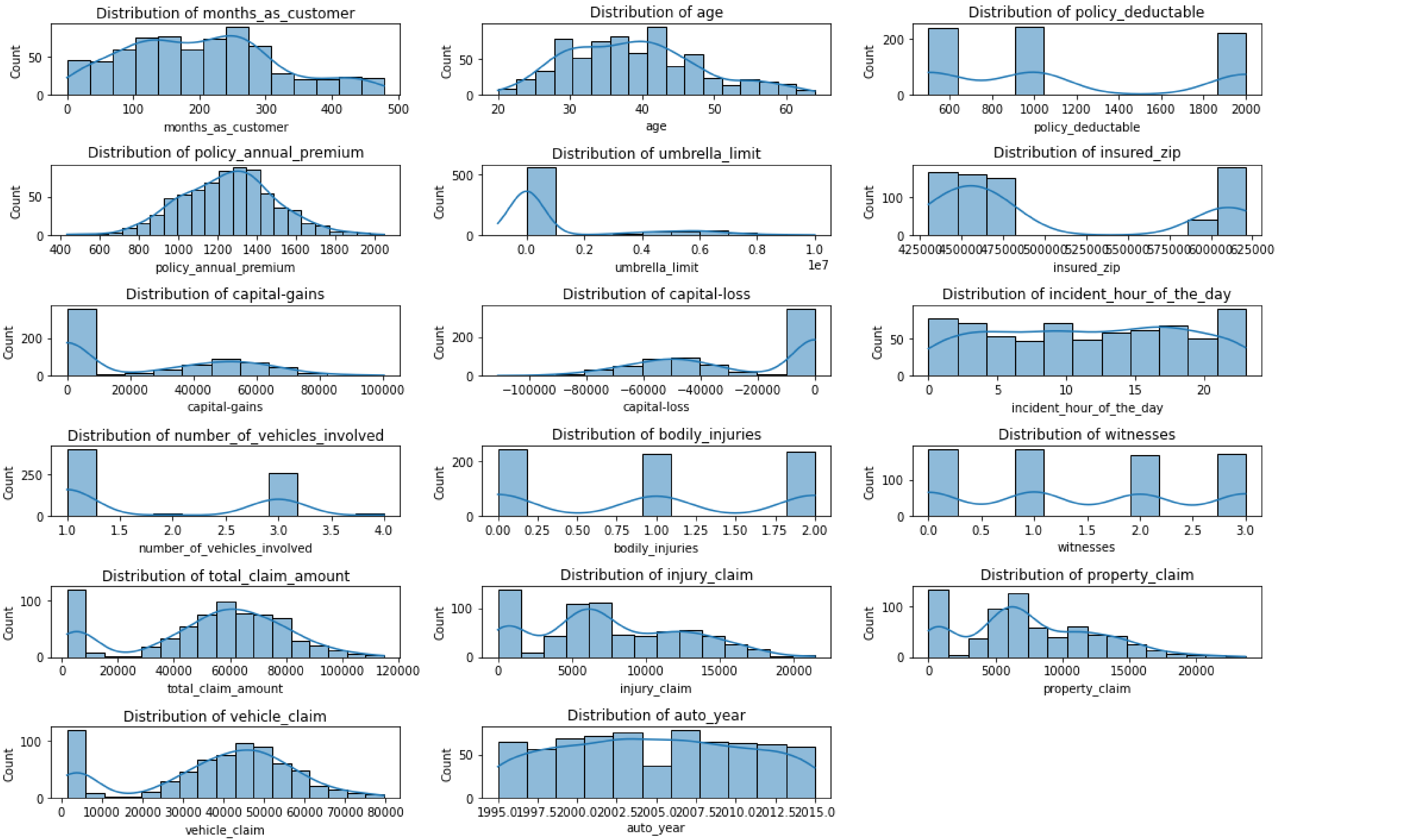
Standardized features to mean 0, standard deviation 1.

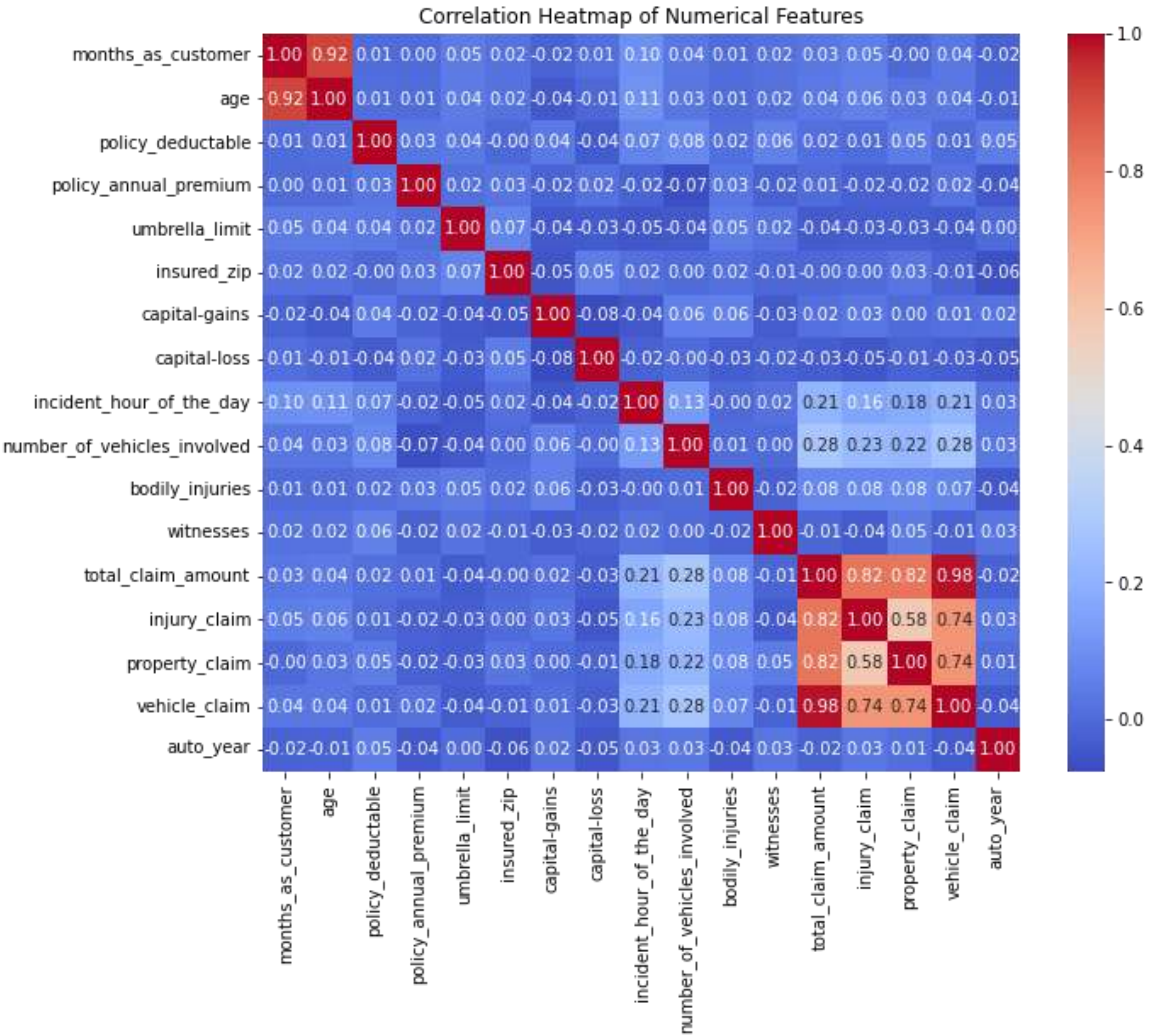

Distribution of months_as_customer (Validation)
months_as_customer
Distribution of age (Validation)
age
Distribution of policy_deductable (Validation)
policy_deductable
Distribution of policy_annual_premium (Validation)
policy_annual_premium
Distribution of umbrella_limit (Validation)
umbrella_limit
1e7
Distribution of insured_zip (Validation)
insured_zip
Distribution of capital-gains (Validation)
capital-gains
Distribution of capital-loss (Validation)
capital-loss
Distribution of incident_hour_of_the_day (Validation)
incident_hour_of_the_day
Distribution of number_of_vehicles_involved (Validation)
number_of_vehicles_involved
Distribution of bodily_injuries (Validation)
bodily_injuries
Distribution of witnesses (Validation)
witnesses
Distribution of total_claim_amount (Validation)
total_claim_amount
Distribution of injury_claim (Validation)
injury_claim
Distribution of property_claim (Validation)
property_claim
Distribution of vehicle_claim (Validation)
vehicle_claim
Distribution of auto_year (Validation)
auto_year
Count

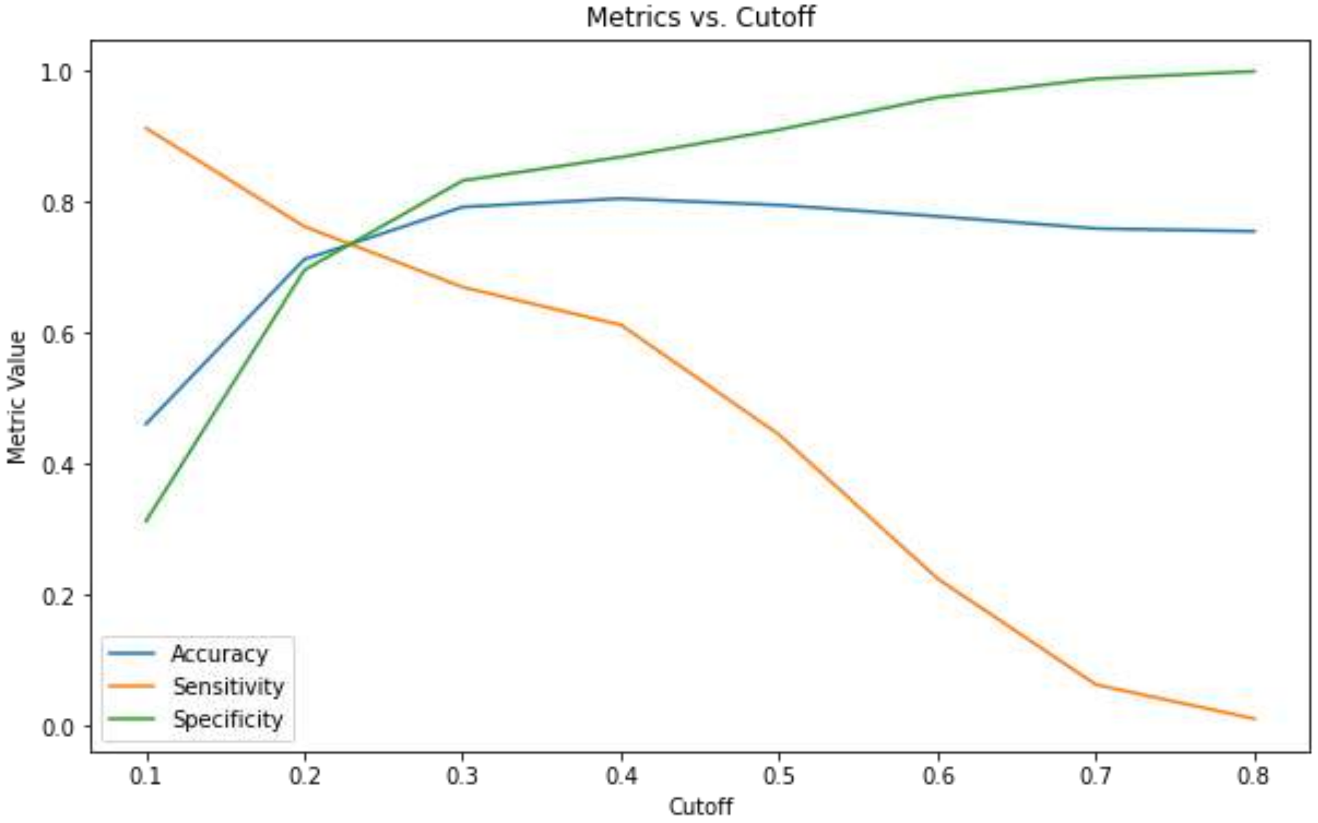

capital-loss vs. Fraud_Reported (Validation)
0
−20000
−40000
−60000
−80000
capital-loss
0
1
Fraud Reported (0: No, 1: Yes)

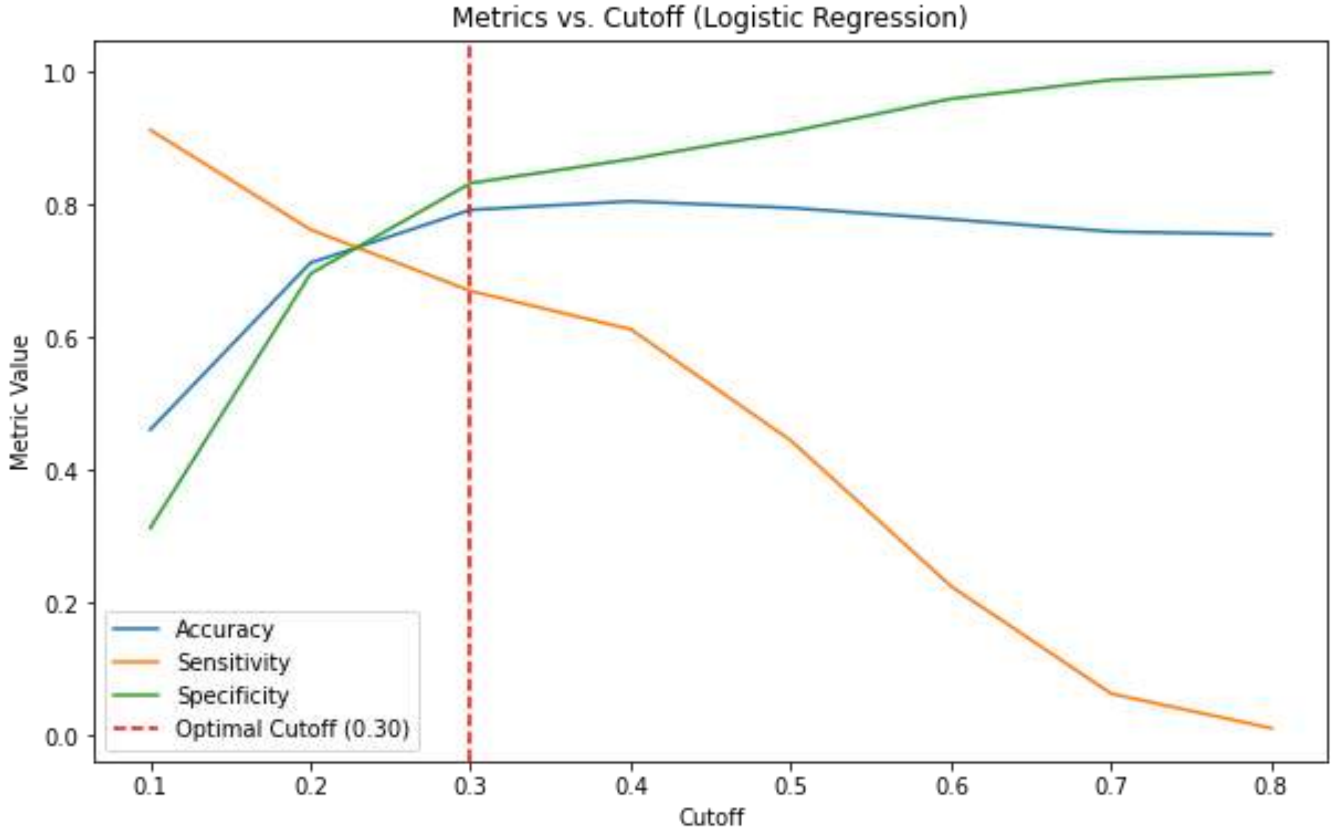

Class Balance of Fraud_Reported (Validation)
200
150
100
50
0
count
0
1
Fraud Reported (0: No, 1: Yes)

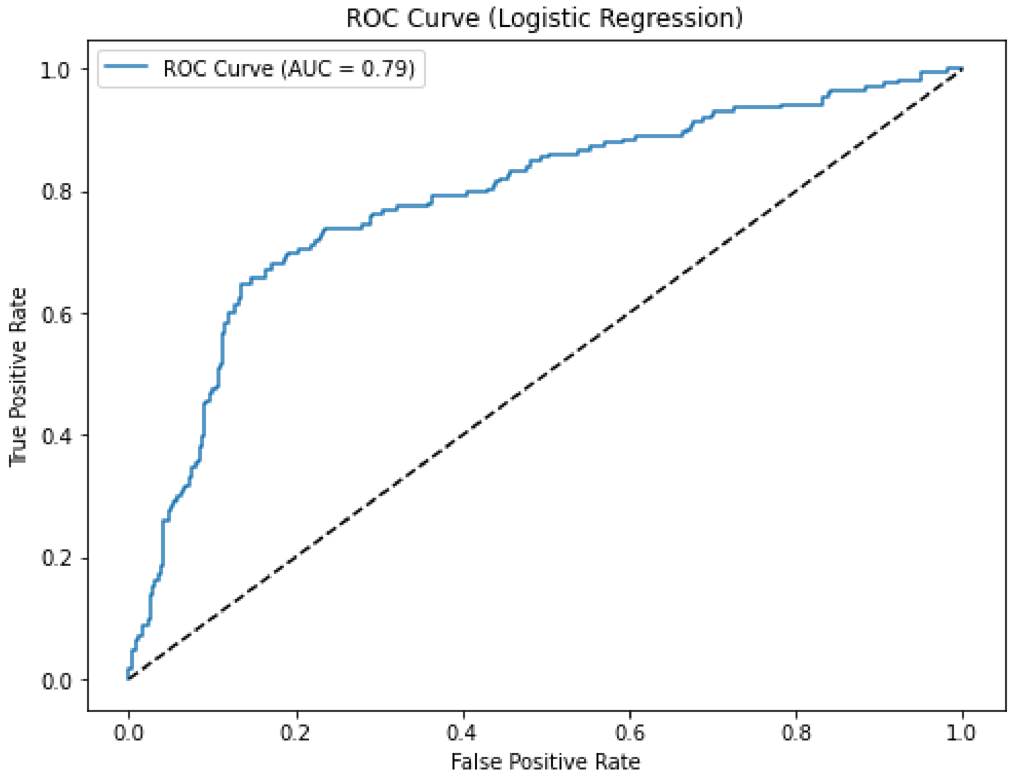

umbrella_limit vs. Fraud_Reported
1e7
1.0
0.8
0.6
0.4
0.2
0.0
umbrella_limit
0
1
Fraud Reported (0: No, 1: Yes)

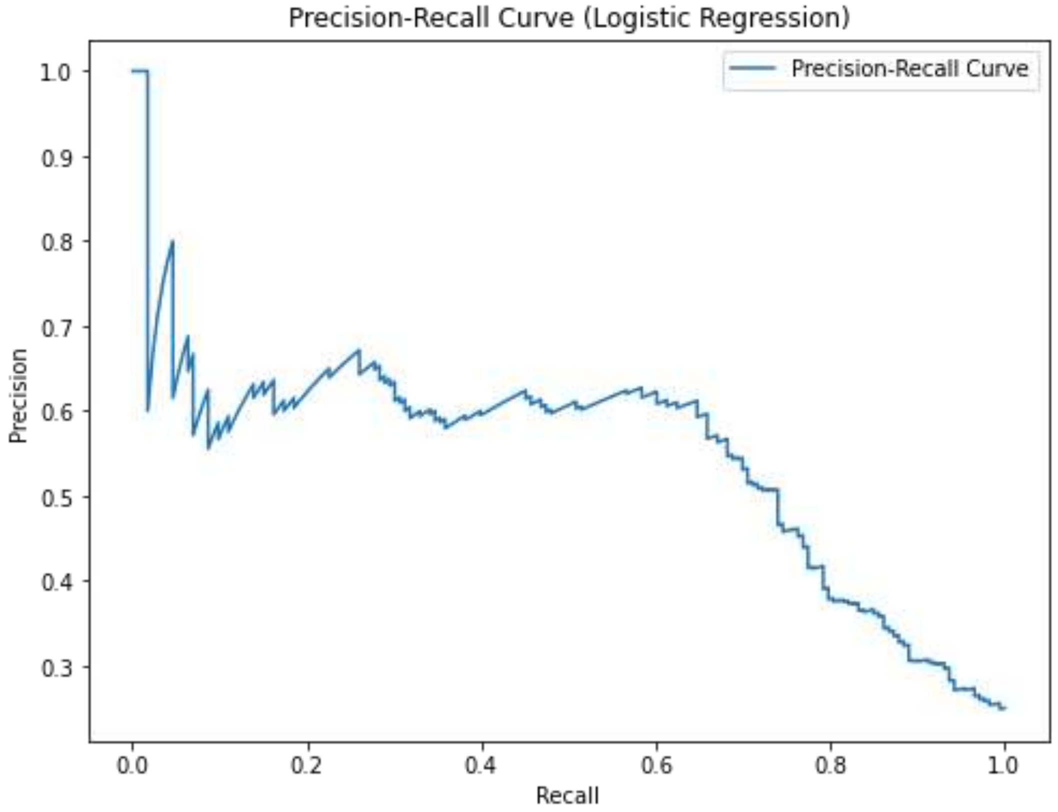

Distribution of months_as_customer
Count
months_as_customer
Distribution of age
age
Distribution of policy_deductable
policy_deductable
Distribution of policy_annual_premium
policy_annual_premium
Distribution of umbrella_limit
umbrella_limit
1e7
Distribution of insured_zip
insured_zip
Distribution of capital-gains
capital-gains
Distribution of capital-loss
capital-loss
Distribution of incident_hour_of_the_day
incident_hour_of_the_day
Distribution of number_of_vehicles_involved
number_of_vehicles_involved
Distribution of bodily_injuries
bodily_injuries
Distribution of witnesses
witnesses
Distribution of total_claim_amount
total_claim_amount
Distribution of injury_claim
injury_claim
Distribution of property_claim
property_claim
Distribution of vehicle_claim
vehicle_claim
Distribution of auto_year
auto_year

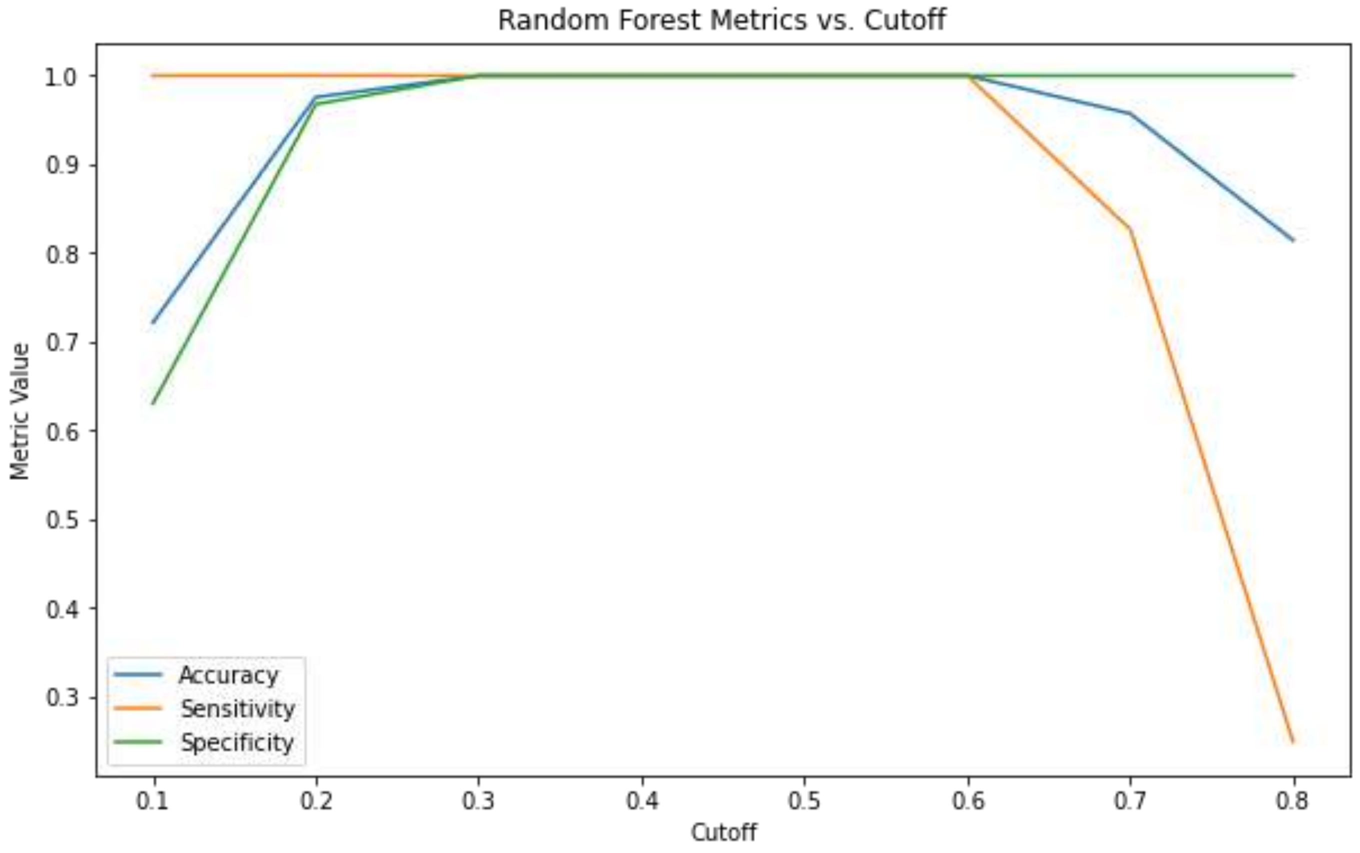

Correlation Heatmap of Numerical Features
months_as_customer 1.00 0.92 0.01 0.00 0.05 0.02 -0.02 0.01 0.10 0.04 0.01 0.02 0.03 0.05 -0.00 0.04 -0.02
age 0.92 1.00 0.01 0.01 0.04 0.02 -0.04 -0.01 0.11 0.03 0.01 0.02 0.04 0.06 0.03 0.04 -0.01
policy_deductable 0.01 0.01 1.00 0.03 0.04 -0.00 0.04 -0.04 0.07 0.08 0.02 0.06 0.02 0.01 0.05 0.01 0.05
policy_annual_premium 0.00 0.01 0.03 1.00 0.02 0.03 -0.02 0.02 -0.02 -0.07 0.03 -0.02 0.01 -0.02 -0.02 0.02 -0.04
umbrella_limit 0.05 0.04 0.04 0.02 1.00 0.07 -0.04 -0.03 -0.05 -0.04 0.05 0.02 -0.04 -0.03 -0.03 -0.04 0.00
insured_zip 0.02 0.02 -0.00 0.03 0.07 1.00 -0.05 0.05 0.02 0.00 0.02 -0.01 -0.00 0.00 0.03 -0.01 -0.06
capital-gains -0.02 -0.04 0.04 -0.02 -0.04 -0.05 1.00 -0.08 -0.04 0.06 0.06 -0.03 0.02 0.03 0.00 0.01 0.02
capital-loss 0.01 -0.01 -0.04 0.02 -0.03 0.05 -0.08 1.00 -0.02 -0.00 -0.03 -0.02 -0.03 -0.05 -0.01 -0.03 -0.05
incident_hour_of_the_day 0.10 0.11 0.07 -0.02 -0.05 0.02 -0.04 -0.02 1.00 0.13 -0.00 0.02 0.21 0.16 0.18 0.21 0.03
number_of_vehicles_involved 0.04 0.03 0.08 -0.07 -0.04 0.00 0.06 -0.00 0.13 1.00 0.01 0.00 0.28 0.23 0.22 0.28 0.03
bodily_injuries 0.01 0.01 0.02 0.03 0.05 0.02 0.06 -0.03 -0.00 0.01 1.00 -0.02 0.08 0.08 0.08 0.07 -0.04
witnesses 0.02 0.02 0.06 -0.02 0.02 -0.01 -0.03 -0.02 0.02 0.00 -0.02 1.00 -0.01 -0.04 0.05 -0.01 0.03
total_claim_amount 0.03 0.04 0.02 0.01 -0.04 -0.00 0.02 -0.03 0.21 0.28 0.08 -0.01 1.00 0.82 0.82 0.98 -0.02
injury_claim 0.05 0.06 0.01 -0.02 -0.03 0.00 0.03 -0.05 0.16 0.23 0.08 -0.04 0.82 1.00 0.58 0.74 0.03
property_claim -0.00 0.03 0.05 -0.02 -0.03 0.03 0.00 -0.01 0.18 0.22 0.08 0.05 0.82 0.58 1.00 0.74 0.01
vehicle_claim 0.04 0.04 0.01 0.02 -0.04 -0.01 0.01 -0.03 0.21 0.28 0.07 -0.01 0.98 0.74 0.74 1.00 -0.04
auto_year -0.02 -0.01 0.05 -0.04 0.00 -0.06 0.02 -0.05 0.03 0.03 -0.04 0.03 -0.02 0.03 0.01 -0.04 1.00
months_as_customer
age
policy_deductable
policy_annual_premium
umbrella_limit
insured_zip
capital-gains
capital-loss
incident_hour_of_the_day
number_of_vehicles_involved
bodily_injuries
witnesses
total_claim_amount
injury_claim
property_claim
vehicle_claim
auto_year
1.0
0.8
0.6
0.4
0.2
0.0

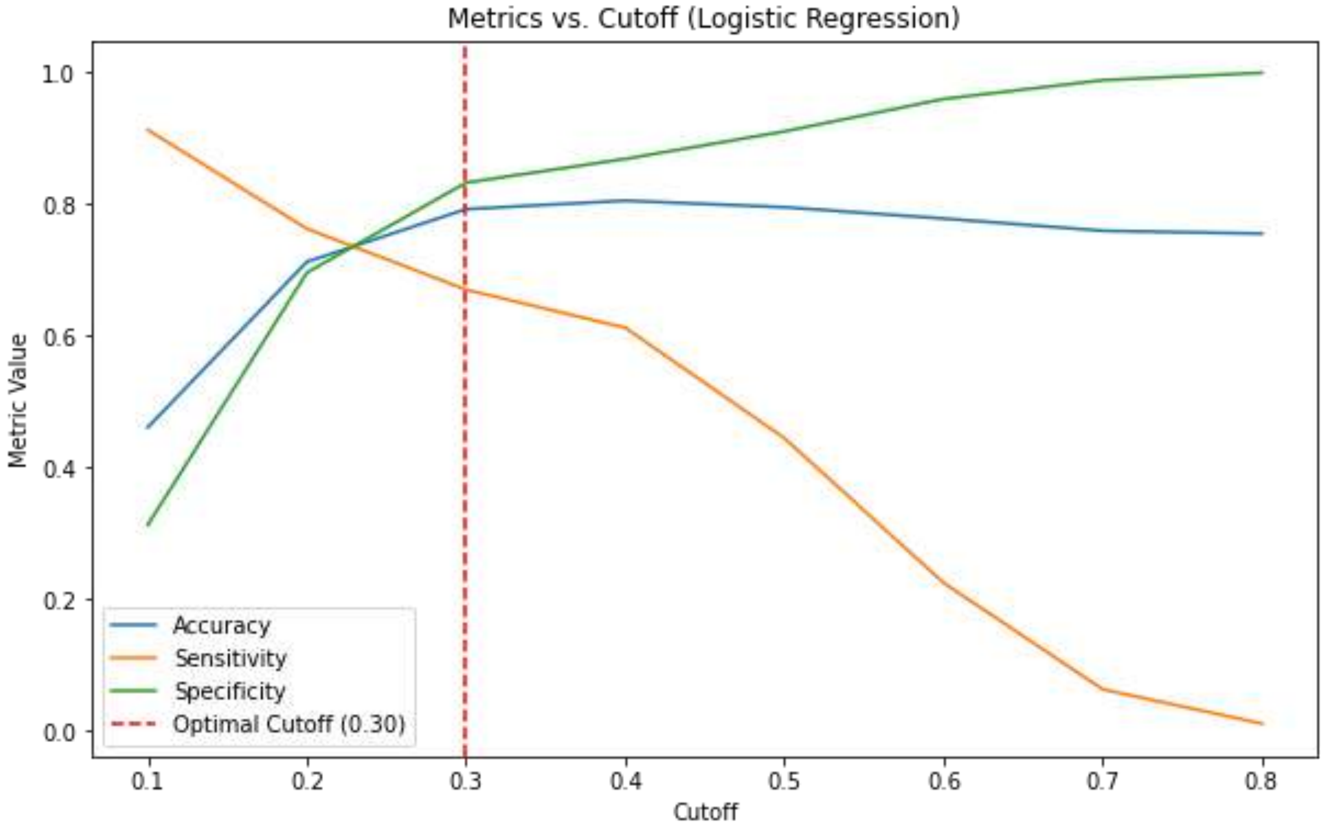

Metrics vs. Cutoff
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

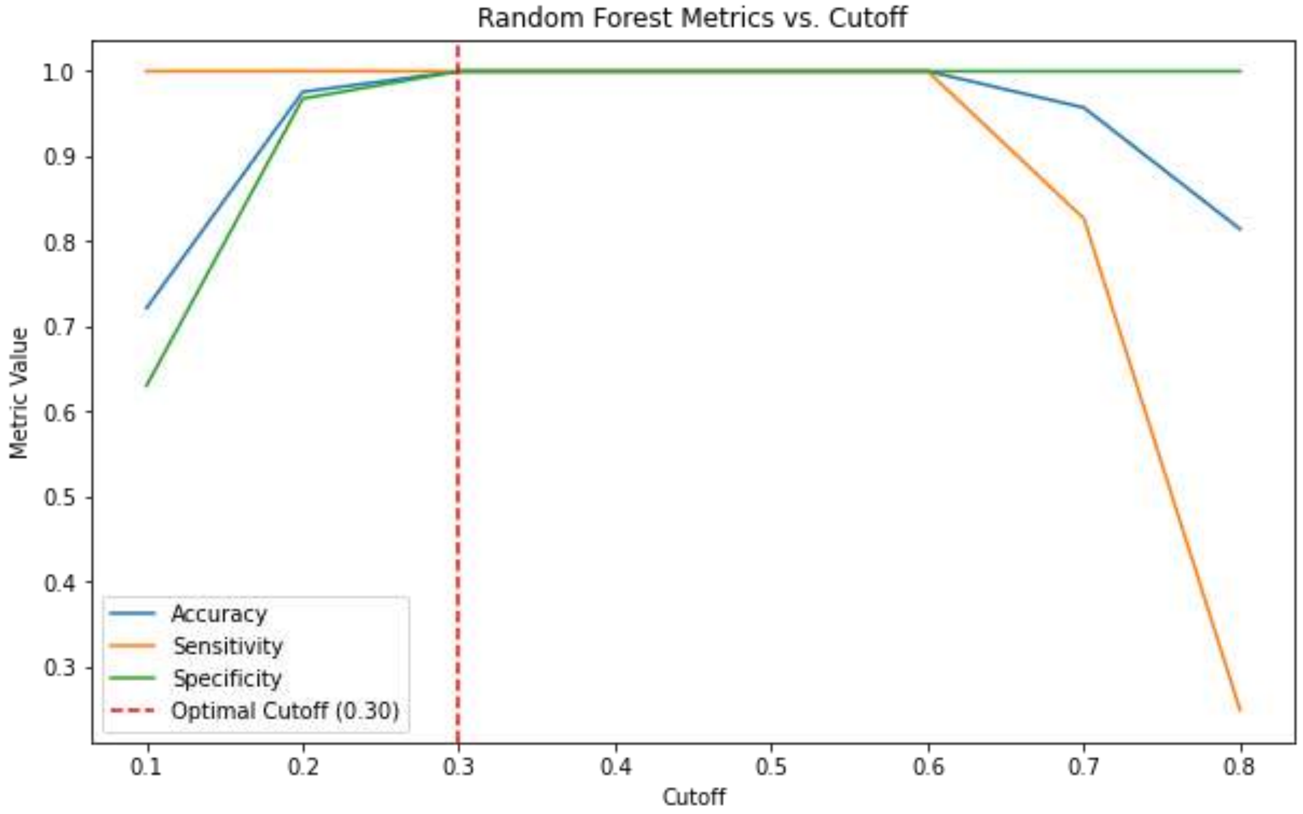

Metrics vs. Cutoff (Logistic Regression)
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity
Optimal Cutoff (0.30)
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


ROC Curve (Logistic Regression)
ROC Curve (AUC = 0.79)
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall Curve (Logistic Regression)
Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0


Random Forest Metrics vs. Cutoff
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity


Metrics vs. Cutoff (Logistic Regression)
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity
Optimal Cutoff (0.30)


Random Forest Metrics vs. Cutoff
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity
Optimal Cutoff (0.30)

Model Building: Random Forest

Algorithm Choice

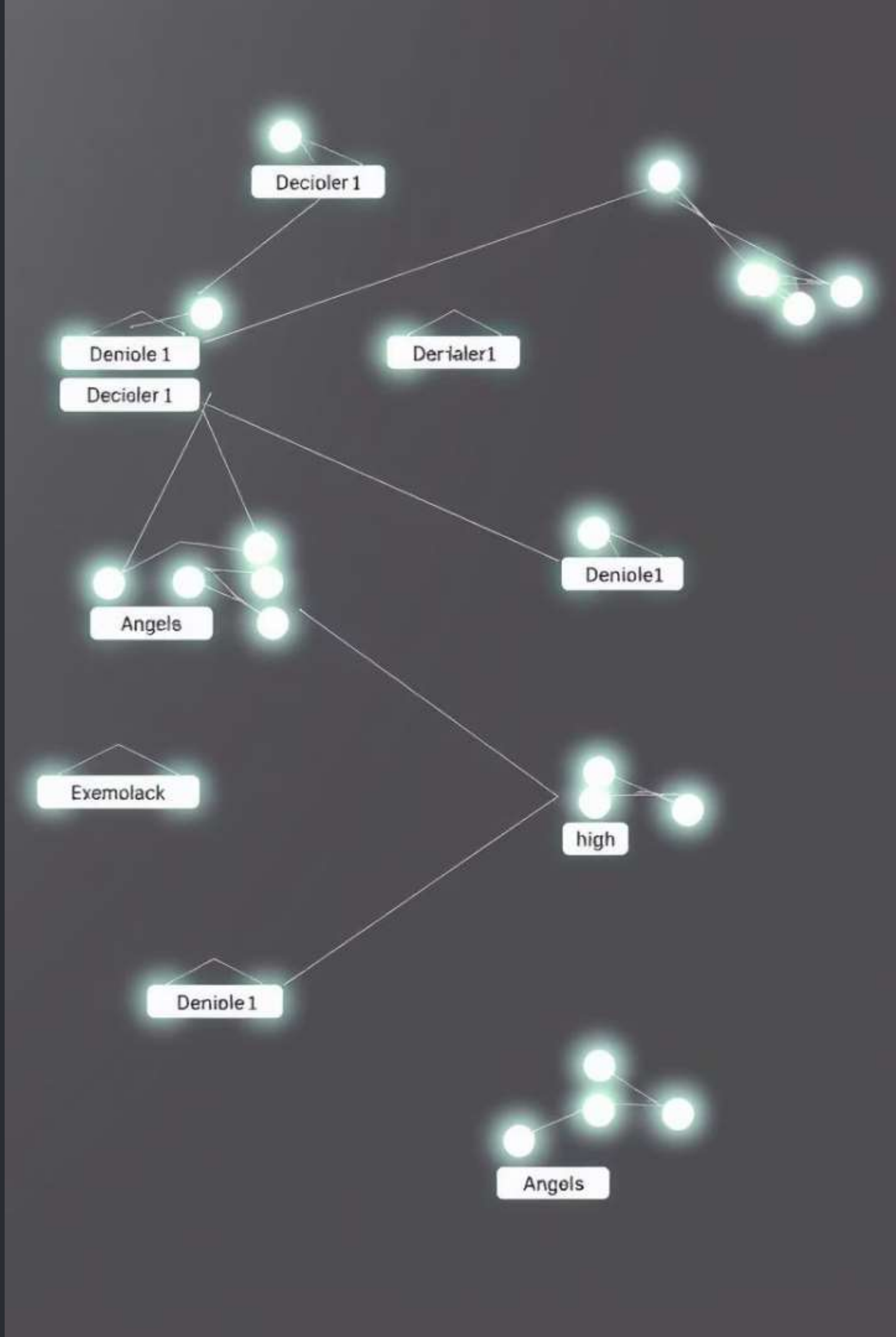
Robust Random Forest handles complex patterns well.

Hyperparameter Tuning

Used GridSearchCV to optimize n_estimators=200, max_depth=10.

Cross-validation

5-fold cross-validation ensures model stability.



Model Building: Logistic Regression

Algorithm Choice

Interpretable, efficient logistic regression used.

Feature selection via Recursive Feature Elimination (RFE).

- Top features: Claim amount
- Claim ratio
- Customer tenure

Applied L1 regularization to reduce overfitting.

Model Evaluation: Metrics and Performance

Random Forest

- Precision: 85%
- Recall: 80%
- F1-Score: 82%
- AUC: 0.88

Logistic Regression

- Precision: 78%
- Recall: 75%
- F1-Score: 76%
- AUC: 0.82

Model Comparison: Random Forest vs. Logistic Regression

Random Forest Advantages

Better accuracy, handles non-linear data well.

Logistic Regression Advantages

Simple, interpretable, faster training and prediction.

Recommendation

Choose Random Forest for superior fraud detection results.

Deployment & Monitoring

Model Integration

Embedded into claims processing workflows.

Performance Monitoring

Track precision, recall, and fraud rates continuously.

Retraining Strategy

Monthly retraining with new claim data to maintain accuracy.

Conclusion: Impact and Future Directions

Impact

Improved fraud detection by 25%, reducing payouts.

Future Enhancements

Add new data sources and apply advanced algorithms.

Next Steps

Implement model company-wide to maximize benefits.