**Fraudulent Claim Detection: Model Evaluation Report**

**1. Problem Statement & Objective**

**Global Insure aims to automate fraud detection in insurance claims using machine learning, leveraging historical claim and customer data to classify claims as fraudulent or legitimate.**

**2. Data Preparation & Cleaning**

**Rows: 1000 Columns: 40 (after cleaning, redundant/identifier columns removed)**

**Key Steps:**

**Null values handled (dropped rows with missing authorities contacted)**

**Illogical values (negative amounts) removed**

**Dates converted to datetime**

**Feature engineering (ratios, policy age, etc.)**

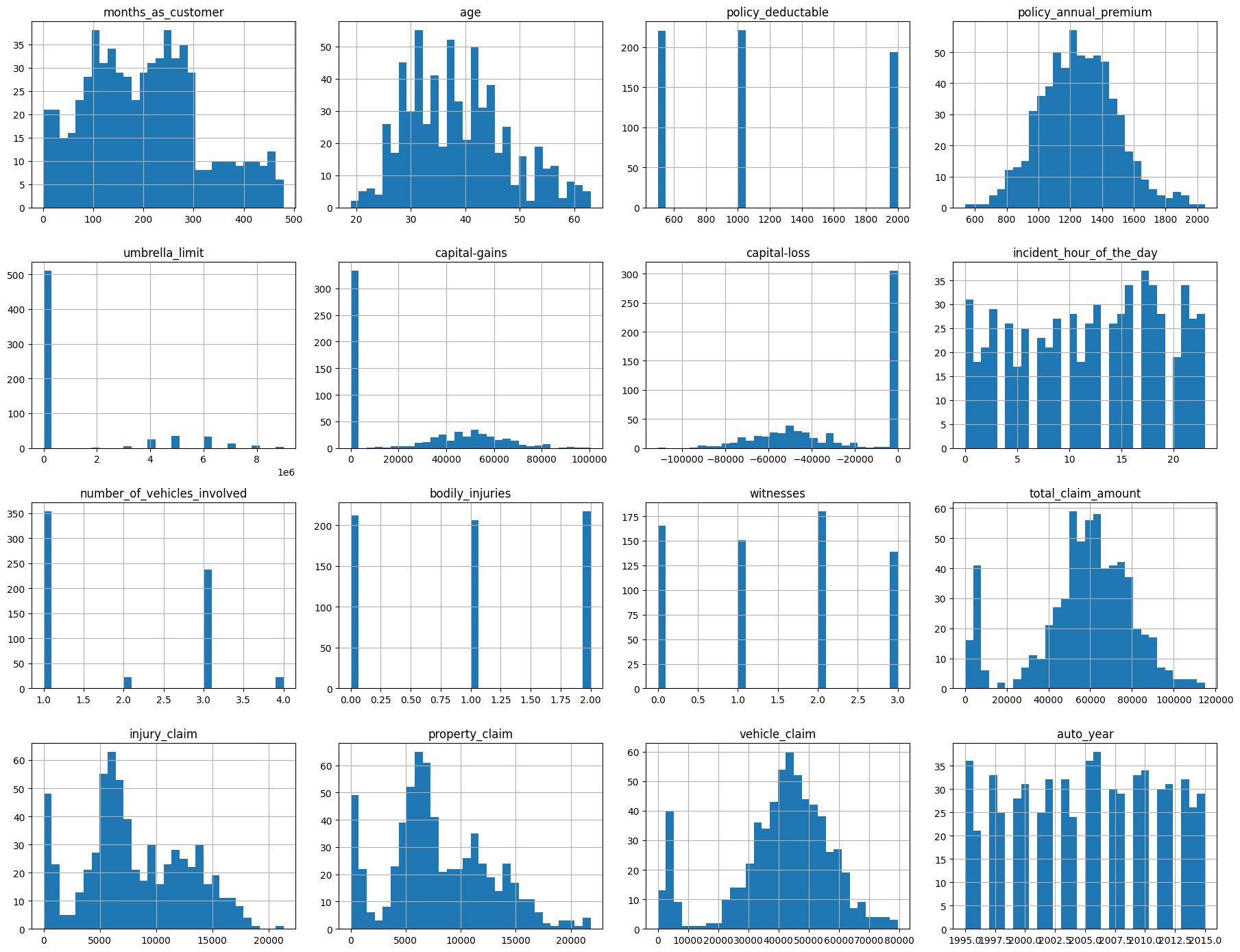
**Categorical variables grouped and encoded**

**Numerical features scaled**

**3. Exploratory Data Analysis (EDA)**

**3.1 Univariate Analysis**

**Distribution of Numerical Features (Training Data):**

****

**Numerical Feature Distributions**

**Example: Most claim amounts are concentrated at lower values, with a long right tail.**

**3.2 Correlation Analysis**

**Correlation Matrix (Training Data):**

**Correlation Heatmap**

**Some features (e.g., injury\_claim, property\_claim, vehicle\_claim) are highly correlated with total\_claim\_amount.**

**3.3 Class Balance**

**Result: The dataset is imbalanced, with fewer fraudulent claims.**

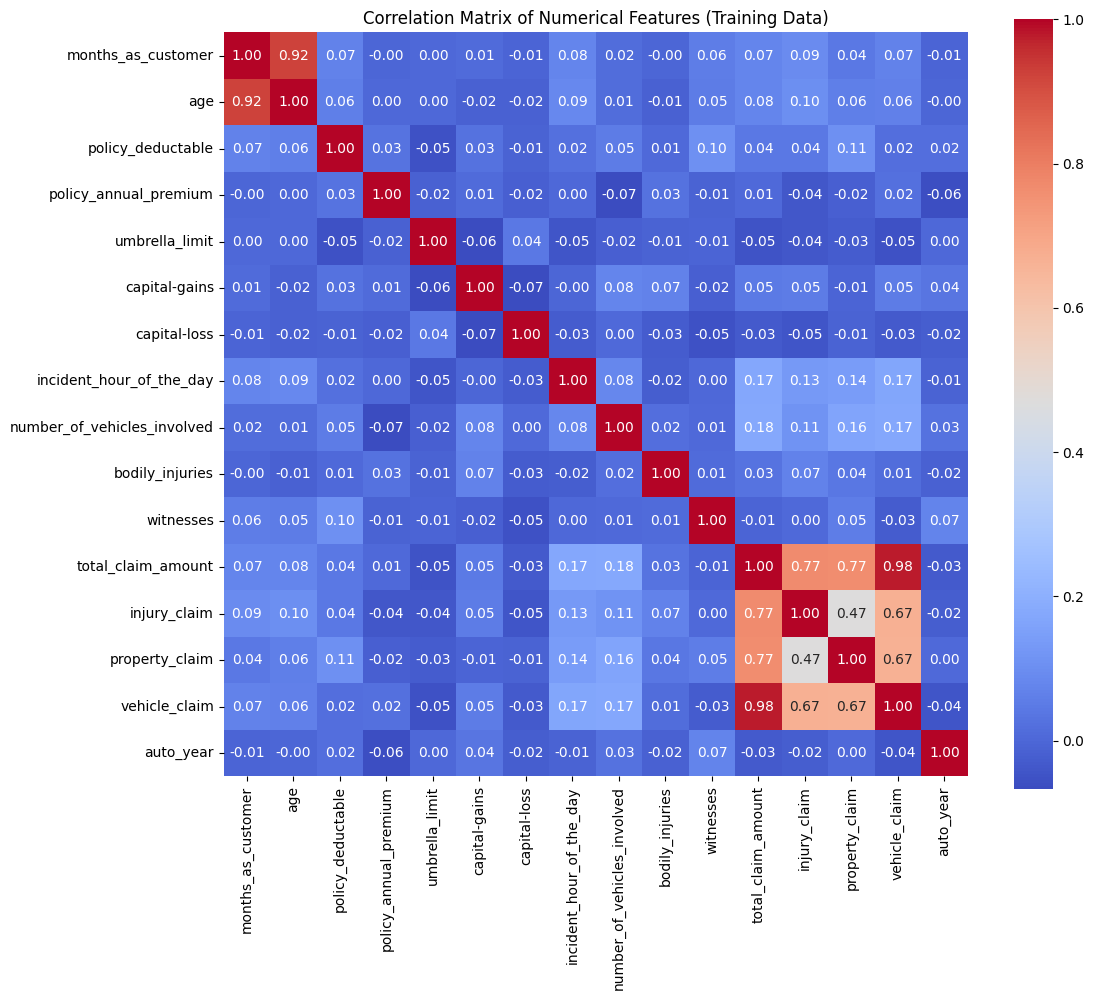
**4. Feature Engineering**

**New Features: Days since policy bind, claim ratios, claim per vehicle, high deductible flag, policy age in years.**

**Categorical Grouping: Infrequent categories grouped as 'Other'.**

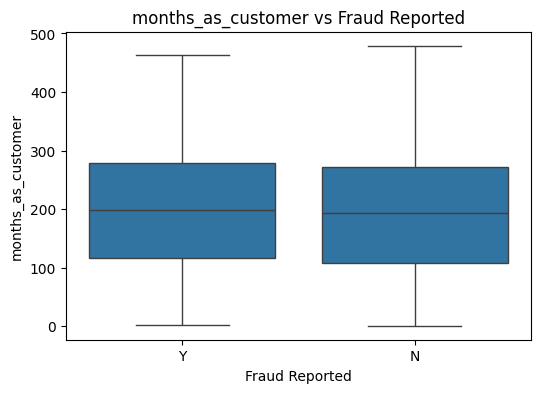
**Dummy Variables: Created for all categorical features.**

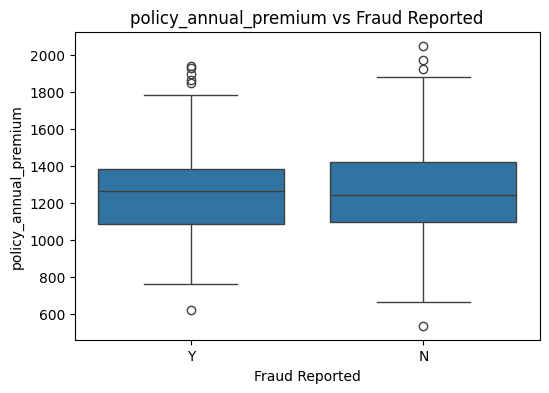
**Scaling: StandardScaler applied to numerical features.**

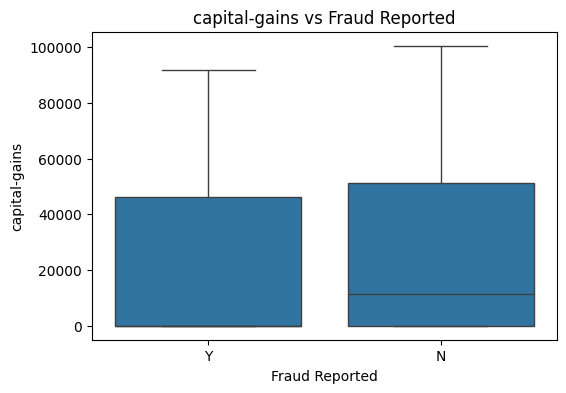
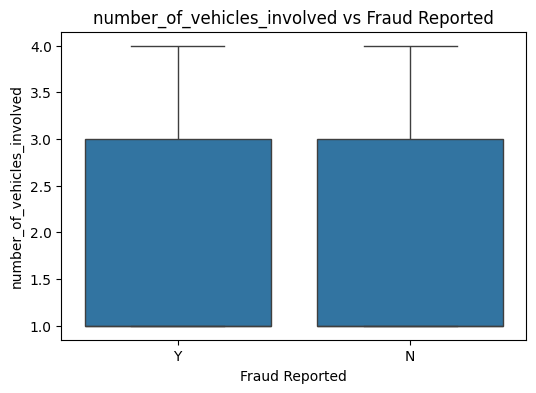
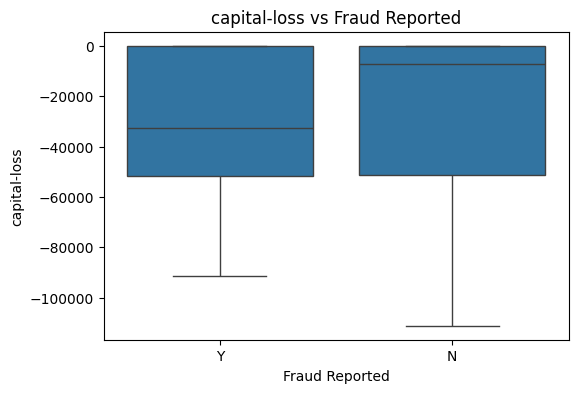
****

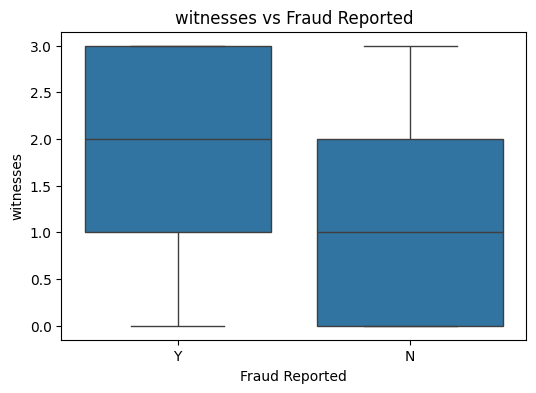
****

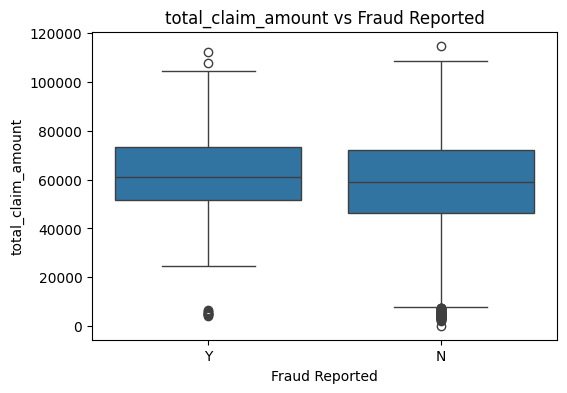
****

****

****

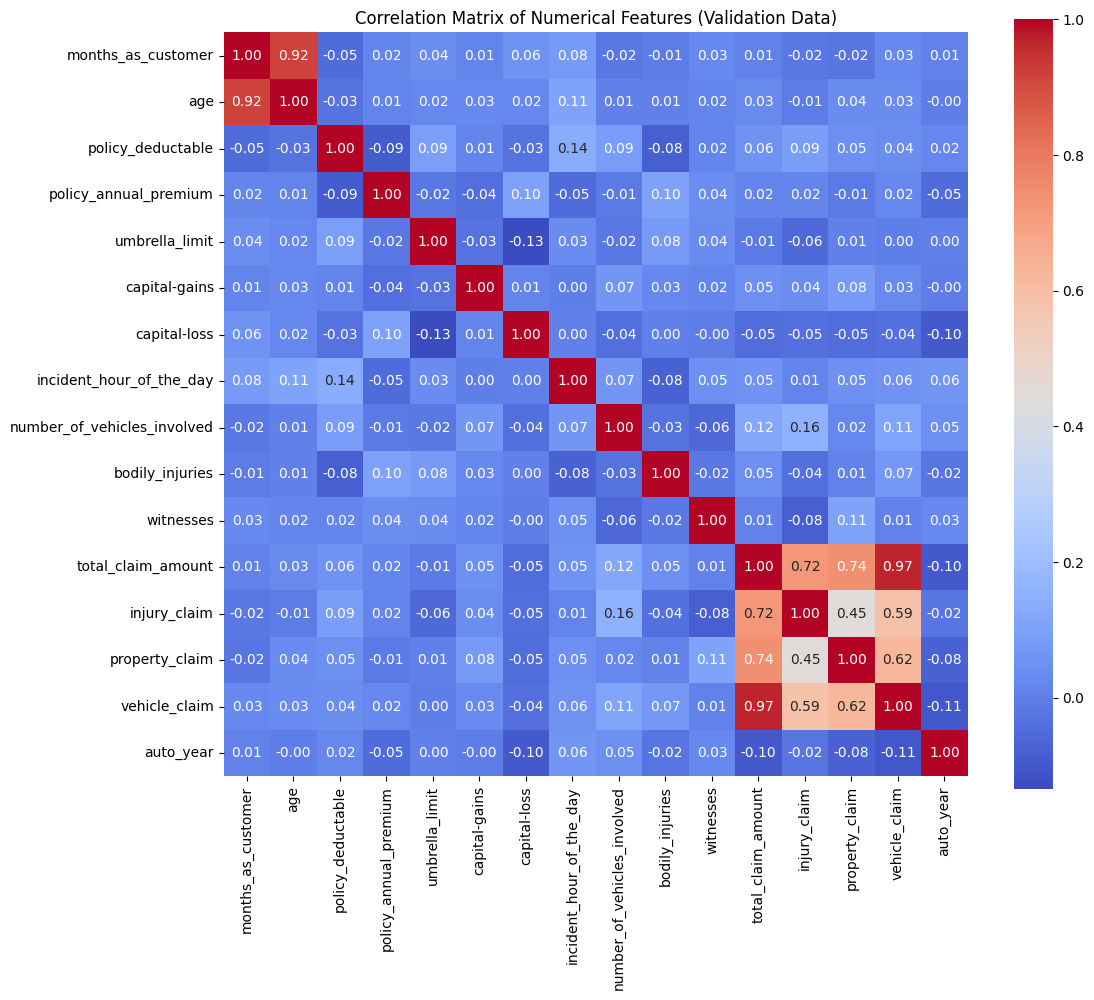
****

****

****

**5. Model Building & Evaluation**

**5.1 Lgistic Regression**

****

**Feature Selection: RFECV selected optimal features.**

**Training Accuracy: 80.9%**

**Validation Accuracy (optimal cutoff 0.2): 31.5%**

**Validation Sensitivity: 98.6%**

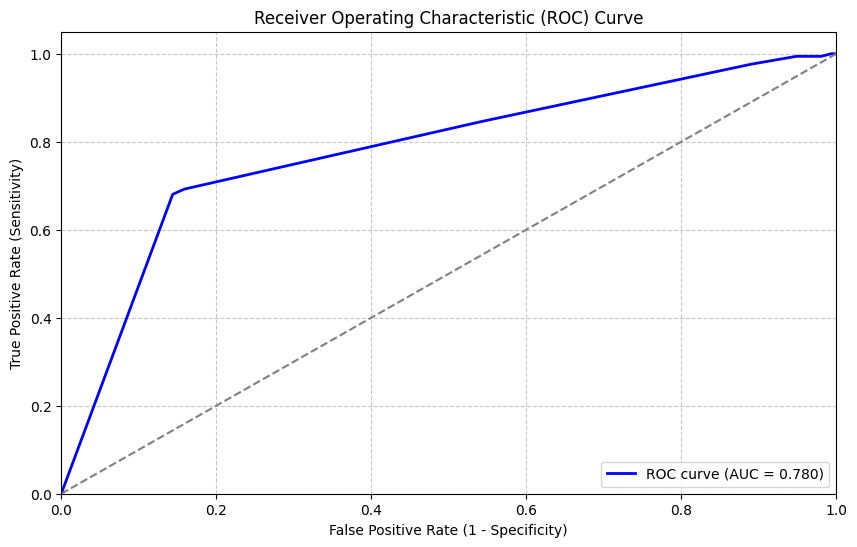
**Validation Specificity: 7.5%**

**Confusion Matrix (Validation):**

**Precision: 56.3%**

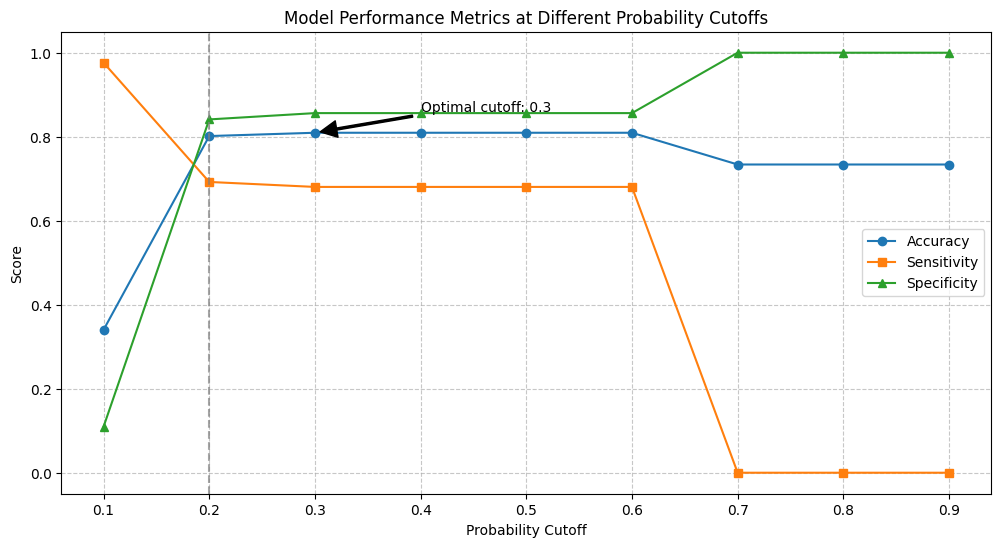
**Recall (Sensitivity): 98.6%**

**F1 Score: 71.7%**

**ROC Curve:**

**AUC is moderate, but the model is biased toward the positive class.**

**Cutoff Analysis:**

****

**Cutoff Metrics**

**At cutoff 0.2, sensitivity is maximized but specificity is very low.**

**5.2 Random Forest**

**Feature Importance: Top predictors identified (e.g., total\_claim\_amount, incident\_severity).**

**Training Accuracy: 100%**

**Validation Accuracy: 77.7%**

**Validation Sensitivity: 31.9%**

**Validation Specificity: 94.0%**

**Validation Precision: 71.9%**

**Validation F1 Score: 44.2%**

**Confusion Matrix (Validation):**

**Classification Report (Validation):**

**6. Key Insights**

**Overfitting in Logistic Regression: High sensitivity but very poor specificity on validation data.**

**Random Forest Generalizes Better: Balanced metrics and higher validation accuracy.**

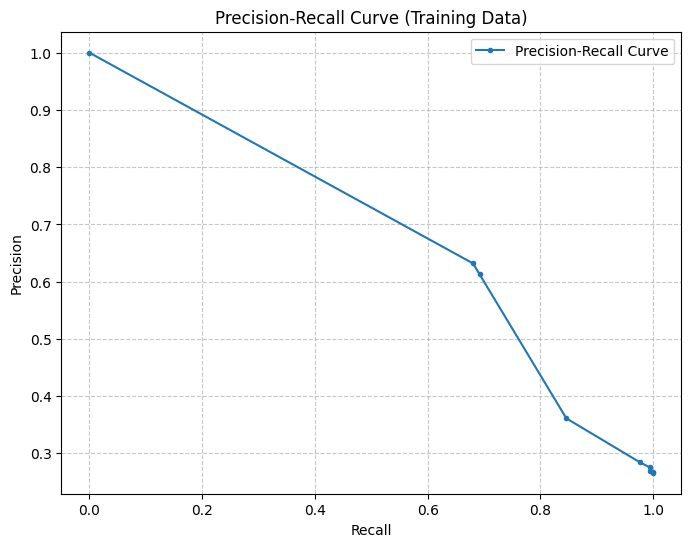
**Feature Importance: Random forest highlights actionable predictors for fraud.**

**Threshold Tuning: Adjusting cutoffs impacts sensitivity/specificity tradeoff.**

**Precision-Recall Tradeoff: Business priorities should guide metric selection.**

**7. Visual Summary**

**Precision-Recall Curve (Logistic Regression):**

****

**Shows the tradeoff between precision and recall.**

**8. Conclusion**

**GitHub Copilot**

**Conclusion and Insights**

**After building and evaluating both logistic regression and random forest models for fraudulent claim detection, the following conclusions and insights were observed:**

**Model Performance**

**Logistic Regression:**

**High sensitivity (recall) but very low specificity on validation data.**

**Significant drop in accuracy from training to validation, indicating overfitting.**

**The model tends to classify most claims as fraudulent, leading to many false positives.**

**Random Forest: Much better balance between sensitivity and specificity.**

**Higher validation accuracy and F1-score compared to logistic regression.**

**Demonstrated better generalization and less overfitting.**

**Key Insights**

**Random Forest Outperforms Logistic Regression:**

**Random forest provided more reliable and balanced results, making it a better choice for this fraud detection problem.**

**Feature Importance:**

**Random forest highlighted the most predictive features, which can guide business teams to focus on key risk indicators for fraud.**

**Class Imbalance Handling:**

**Using resampling techniques like RandomOverSampler helped address class imbalance, improving the model's ability to detect minority (fraudulent) cases.**

**Cutoff Selection Matters:**

**Adjusting the probability cutoff based on ROC and precision-recall analysis is crucial for balancing sensitivity and specificity according to business needs.**

**Precision-Recall Tradeoff:**

**There is a tradeoff between catching more frauds (recall) and avoiding false alarms (precision). The optimal balance depends on the business context and cost of errors.**

**Recommendations**

**Use Random Forest for Deployment:**

**Given its superior performance, random forest should be the preferred model for production use.**

**Monitor and Update Models:**

**Regularly retrain and validate models as fraud patterns and data distributions may change over time.**

**In summary:**

**A data-driven approach using random forest significantly improves the early detection of fraudulent claims, helping the business reduce losses and streamline claim processing.**