Business Analytics

Machine Learning Case study – Group Assignment

Topic:

Fraud Detection using Logistic Regression and Random Forest

Assignment Submitted by

Saloni Sharma

Shiva Shankar Iyer

Amitabh Kumar Biswas

Insurance Claim Fraud Detection: A Machine Learning Approach

Executive Summary

This report outlines the development of a machine learning-based solution for detecting fraudulent insurance claims. Leveraging historical claim data, the project applies Logistic Regression and Random Forest models to uncover patterns indicative of fraudulent behavior. The goal is to improve fraud detection accuracy, reduce manual investigations, and mitigate financial losses for insurance providers.

Project Overview and Business Objectives

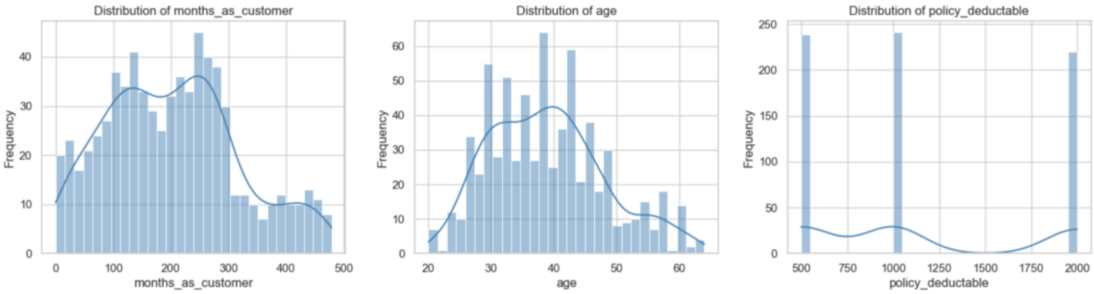
Insurance fraud poses significant challenges, leading to substantial financial losses and increased policy premiums. This project aims to develop a predictive model that classifies incoming claims as either fraudulent or legitimate. By analyzing features such as claim amount, customer profile, and incident specifics, the model enhances decision-making and supports efficient claim validation.

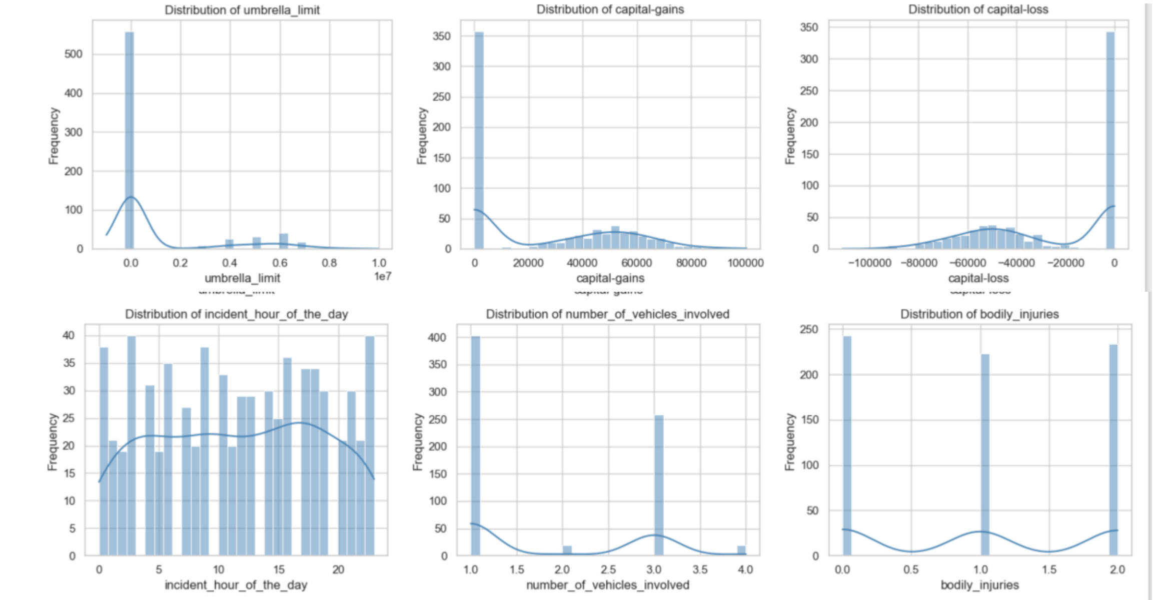
"The model's ability to predict fraudulent claims based on past data is a critical step towards proactive risk management in the insurance industry."

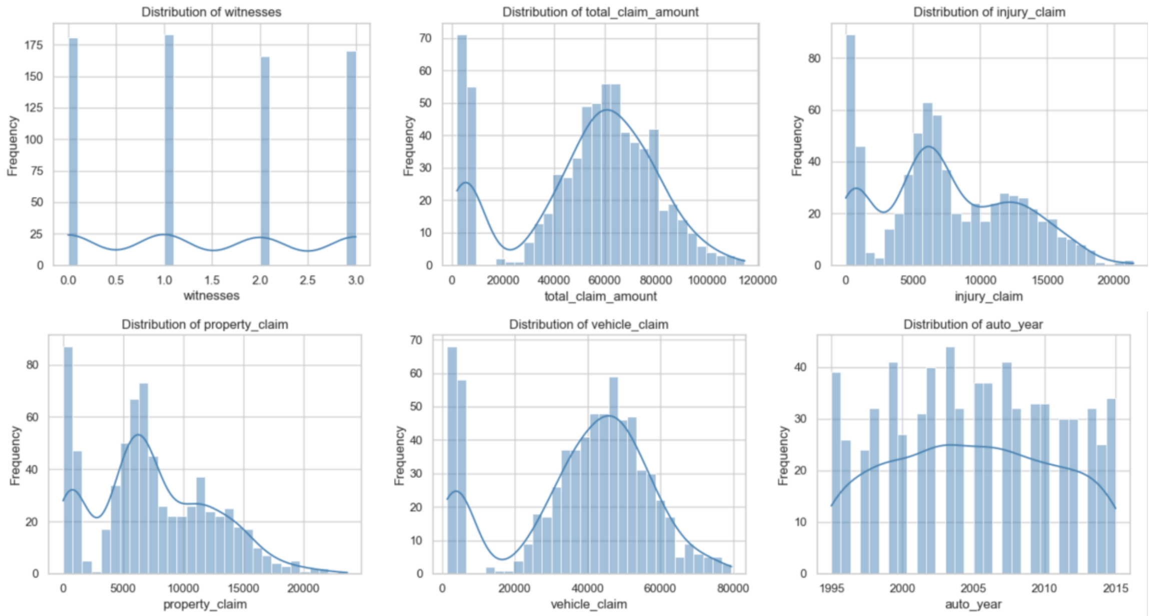
Analyzing Historical Claim Data

To identify fraud indicators, the historical claim data was analyzed using the following techniques:

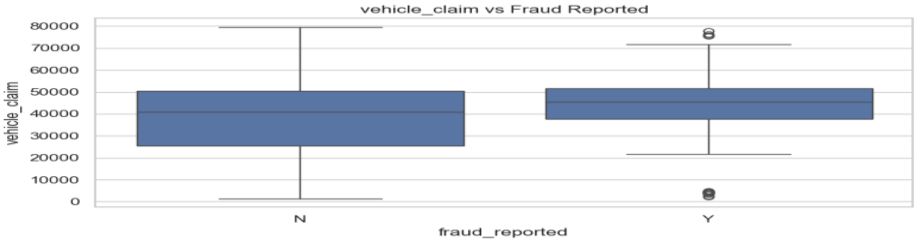
- **Exploratory Data Analysis (EDA):** Distributions and outliers were studied to understand the characteristics of fraud and non-fraud claims.

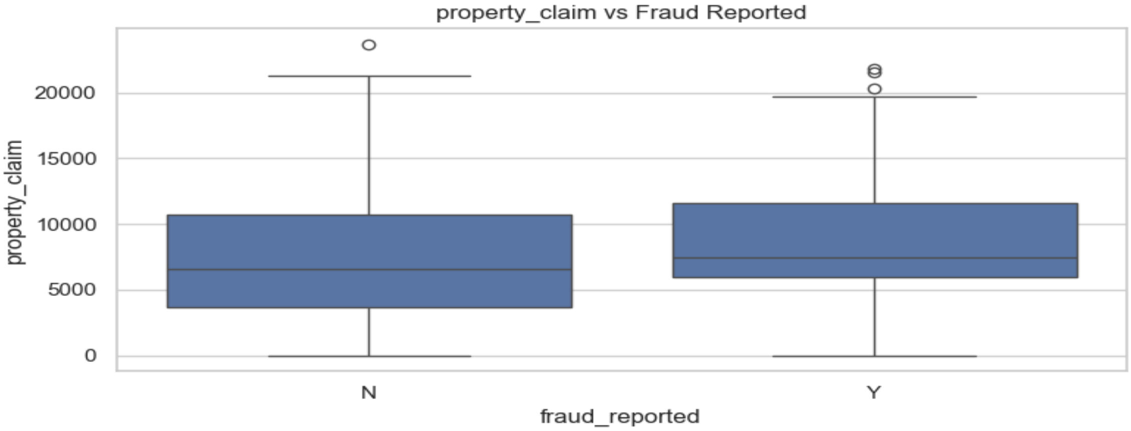


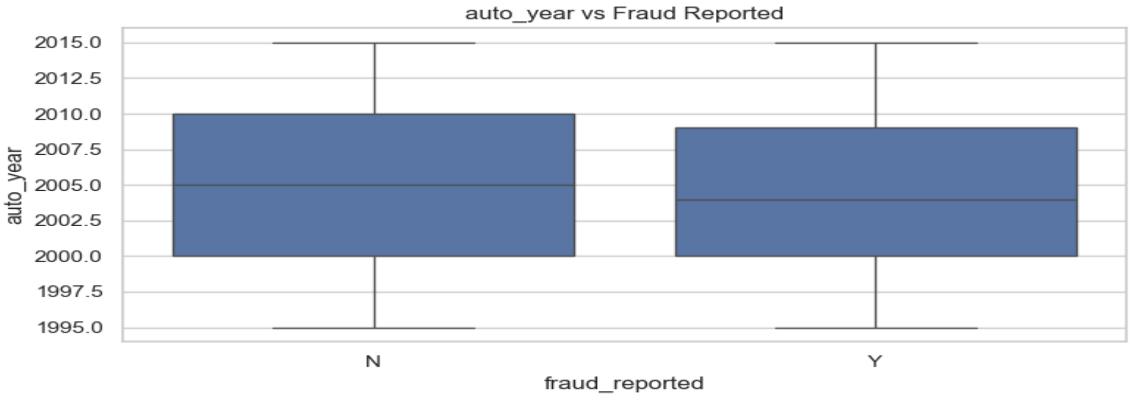




- **Visualization: Boxplots, heatmaps, and count plots helped reveal relationships between variables and fraud labels.**



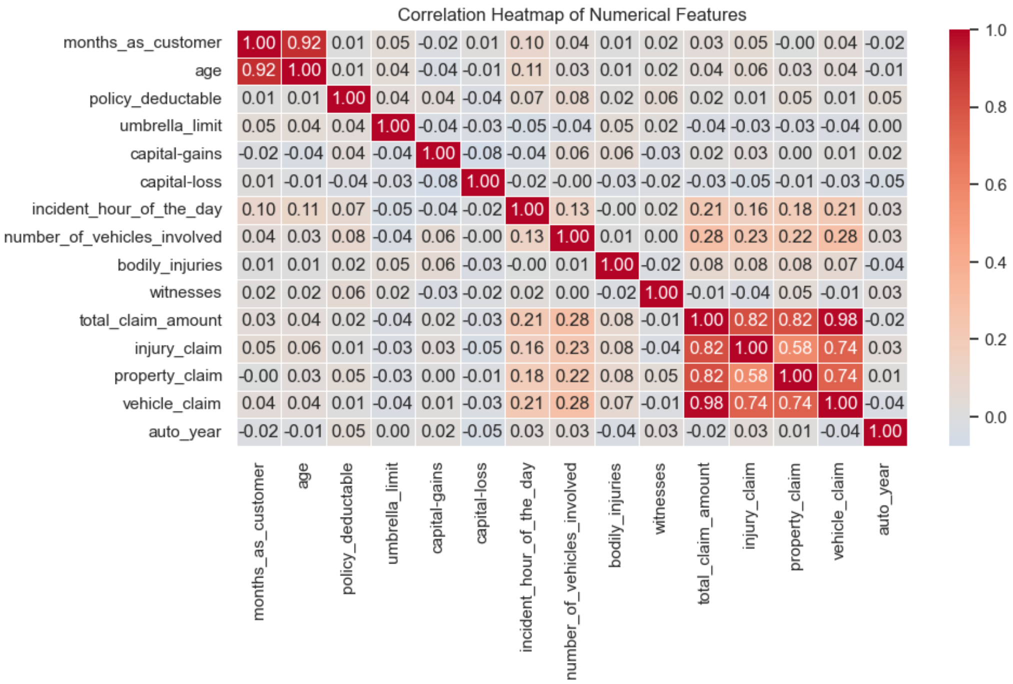




- **Bivariate Analysis: Examined the impact of individual features on fraud classification.**
- **Model-Based Pattern Discovery: Feature importance scores from the Random Forest model guided variable selection.**

Key Predictive Features of Fraudulent Behavior

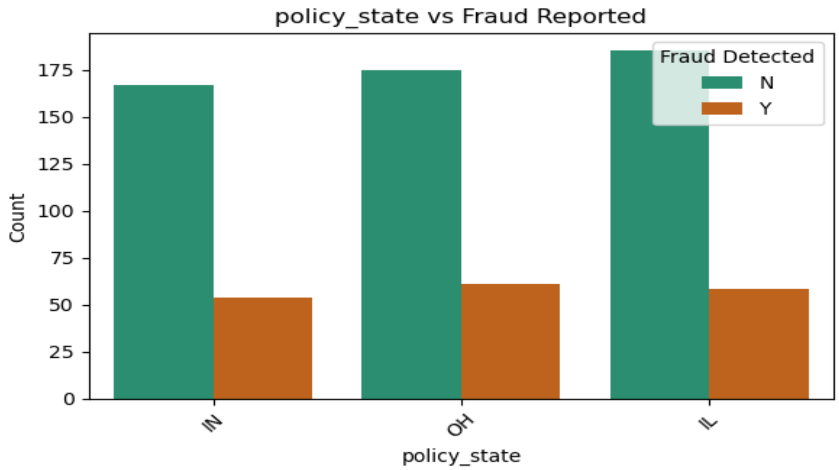
A combination of correlation analysis and model insights identified the following top predictors:

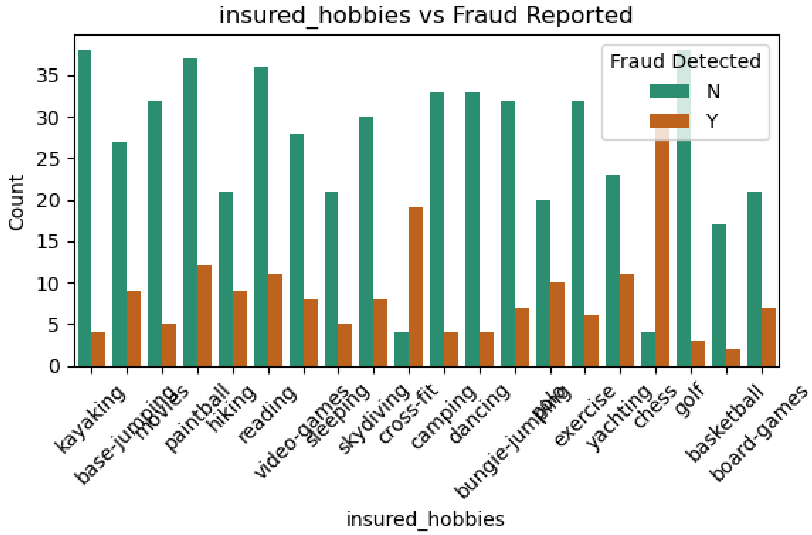


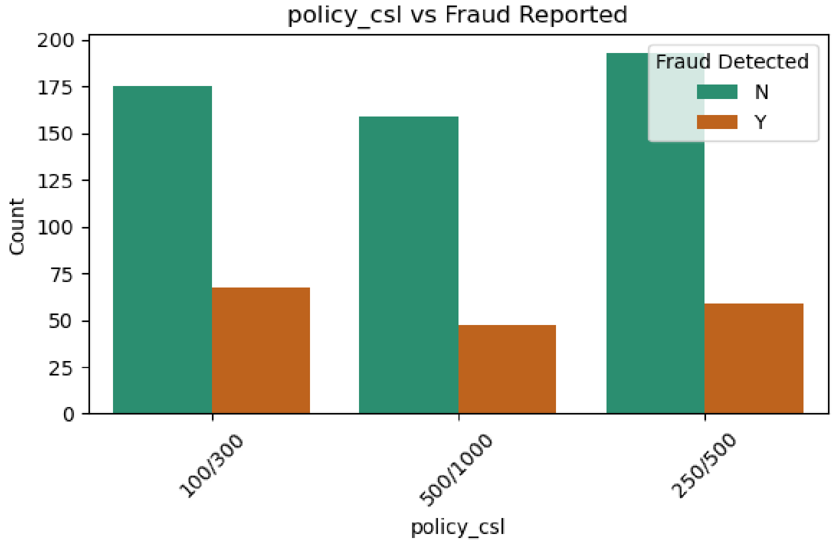
- **total_claim_amount**
- **injury_claim**
- **property_claim**
- **incident_severity**
- **insured_hobbies**
- **months_as_customer**
- **incident_hour_of_the_day**
- **auto_year**
- **witnesses**
- **bodily_injuries**

These features demonstrated strong individual and combined predictive power.

Predicting Fraud Likelihood





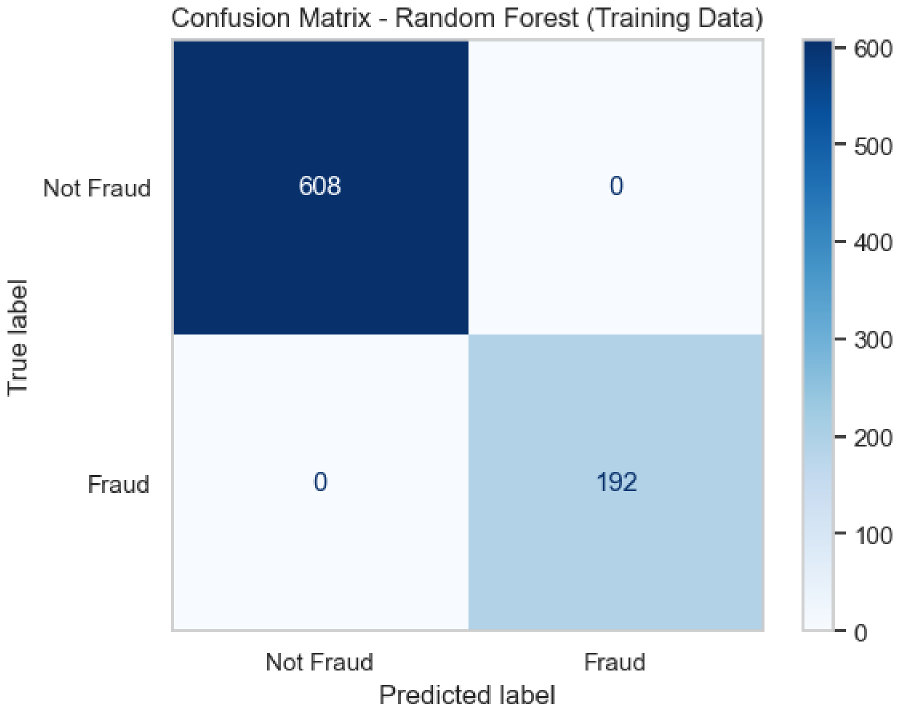


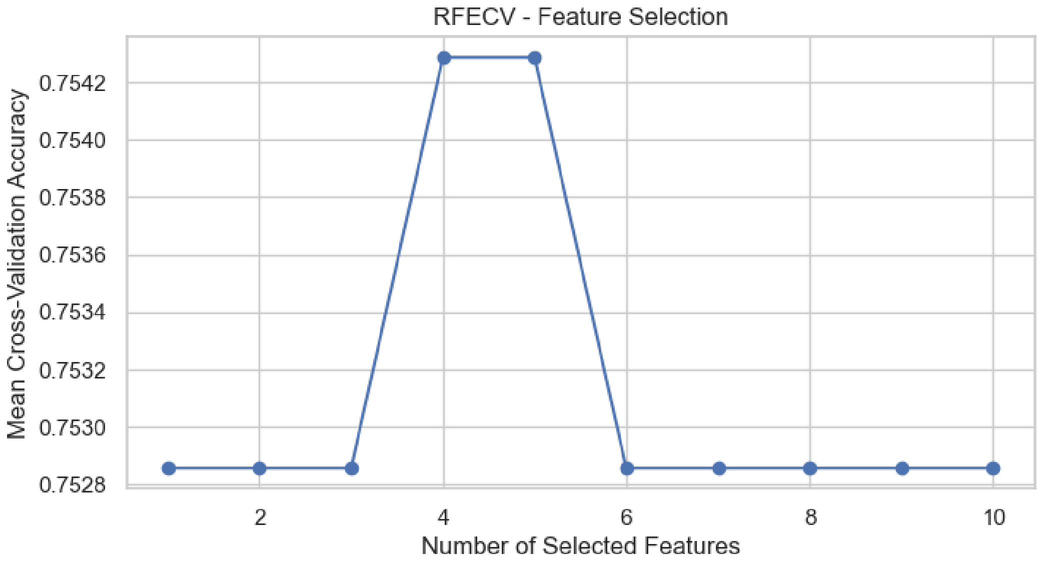
Yes, the likelihood of fraud can be predicted for new claims using trained machine learning models:

- **Logistic Regression: Served as a baseline for performance comparison.**
- **Random Forest Classifier: Tuned using GridSearchCV and demonstrated higher recall and precision, capturing complex interactions in the data.**

Model Evaluation

The models were evaluated using metrics like accuracy, ROC-AUC, F1-score, and confusion matrices. An optimal probability cutoff of 0.10 was selected to ensure high recall, critical for fraud detection.





Final Model Performance (Random Forest):

- **Sensitivity (Recall): 0.9583**
- **Specificity: 0.9951**
- **Precision: 0.9840**
- **F1 Score: 0.9710**

These values indicate a highly effective model that detects most frauds while keeping false positives low.

Insights to Improve Fraud Detection

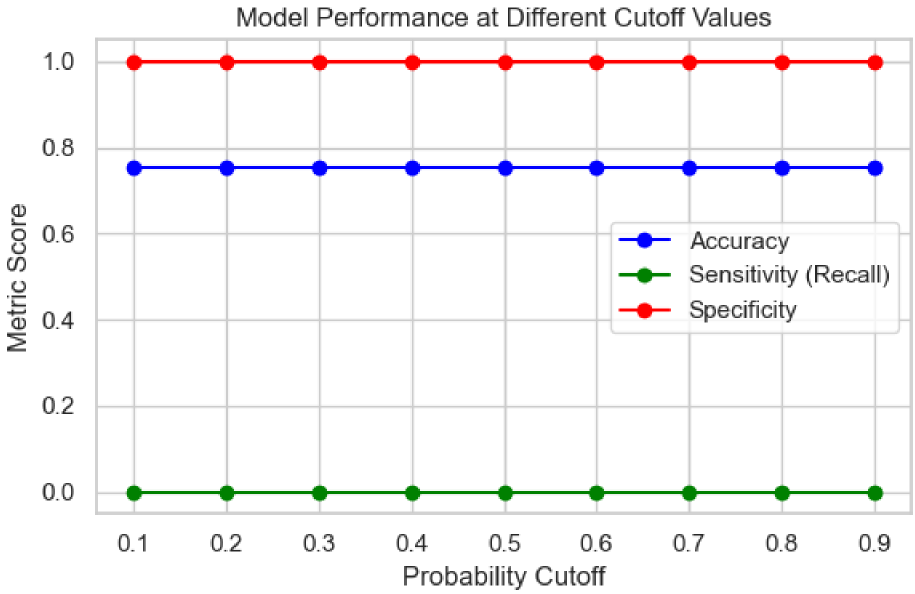
The model offers actionable insights:

- **High-Risk Indicators: Flags claims with suspiciously high values and missing documentation.**
- **Risk Scoring: Assigns fraud probabilities to help prioritize investigations.**
- **Policy Recommendations: Suggests greater scrutiny for claims from new customers.**
- **Audit Transparency: Model explainability enhances decision transparency.**

Observations and Key Considerations

- **Class Imbalance: With ~25% fraud cases, special attention was given to recall.**
- **Variable Relationships: Used correlation matrices and chi-square tests to refine features.**

- **Feature Selection: Recursive Feature Elimination (RFE) improved model focus and reduced overfitting.**
- **Cutoff Optimization: Plots of accuracy, sensitivity, and specificity across different thresholds guided cutoff selection.**



Conclusion

This project successfully developed a robust fraud detection system for Global Insure. With high recall and precision, the Random Forest model can significantly reduce manual intervention and accelerate fraud identification. The insights gained from model features can also shape internal policies and improve claim vetting processes.

The solution is ready for deployment and can be further enhanced through real-time data integration and regular retraining to adapt to evolving fraud tactics.