Fraud Detection Model Analysis

1. Introduction

This report summarizes the process and findings from building classification models for insurance fraud detection. We used Logistic Regression and Random Forest models to evaluate performance and derive insights.

2. Data Preparation & Feature Engineering

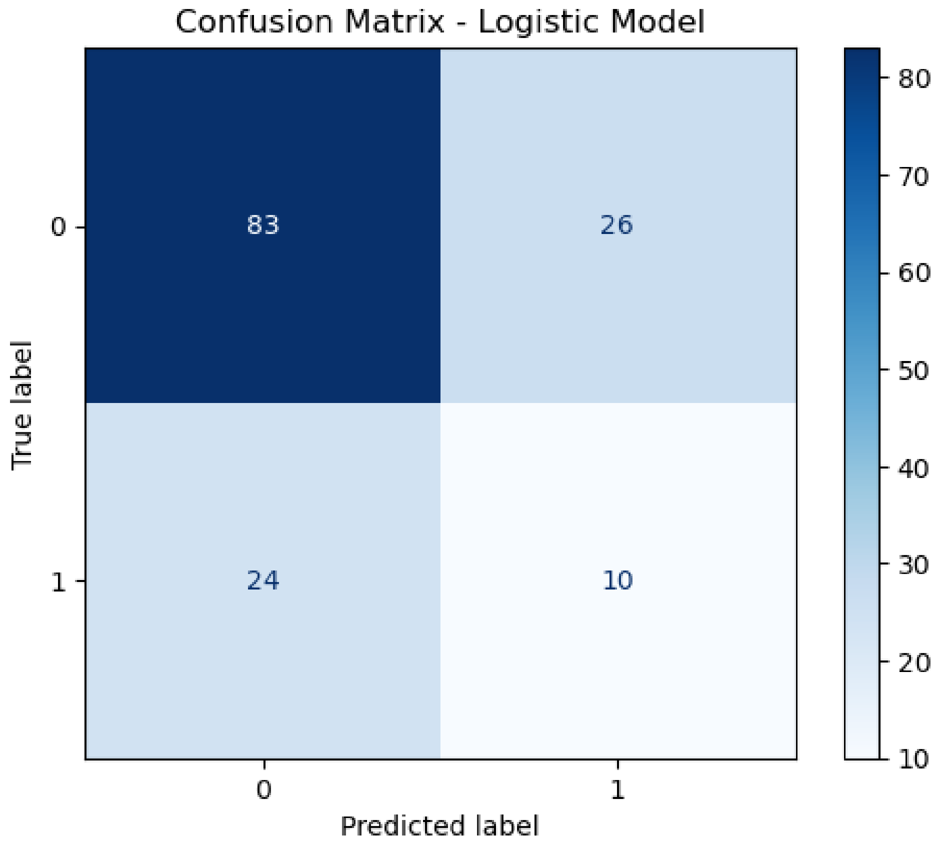
Categorical variables were encoded, and features were engineered such as claim severity ratio. We also checked multicollinearity using VIF for Logistic Regression and feature importance for Random Forest.

feature	VIF	
0	const	1
1	age	1.016279
2	incident_hour_of_the_day	1.078931
3	number_of_vehicles_involved	1.134243
4	claim_severity_ratio	1.184763

3. Logistic Regression Model

A logistic regression model was trained using selected features. VIFs were within acceptable limits. Model evaluation at cutoff 0.2 showed high sensitivity (~79%) and acceptable trade-offs in specificity and precision. It was useful for interpretability and insight generation.

Fraud Detection Model Analysis

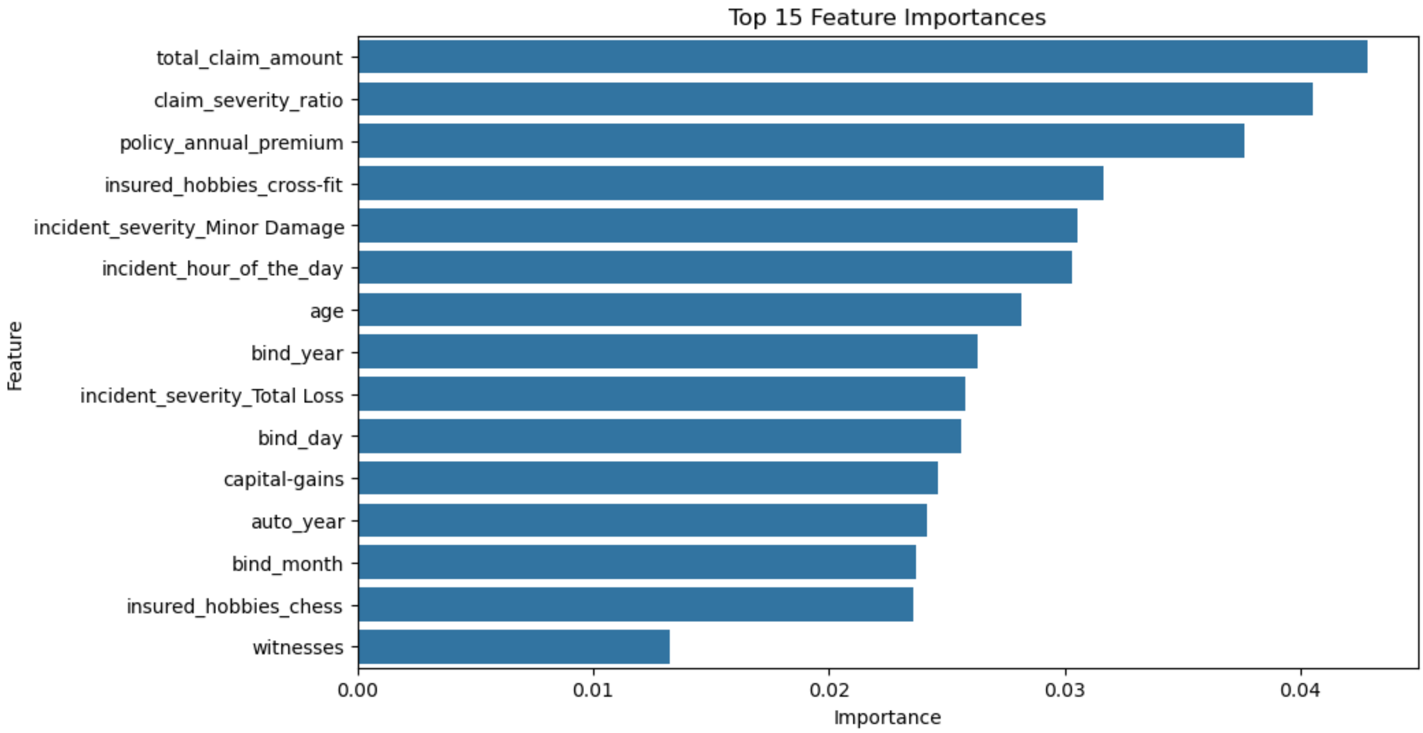


4. Random Forest Model

Random Forest model was trained and hyperparameter tuned. Feature importance was used to select top predictors. This model outperformed logistic regression in recall, accuracy, and F1 score, though it was less interpretable.

RandomForestClassifier(class_weight='balanced', random_state=42)

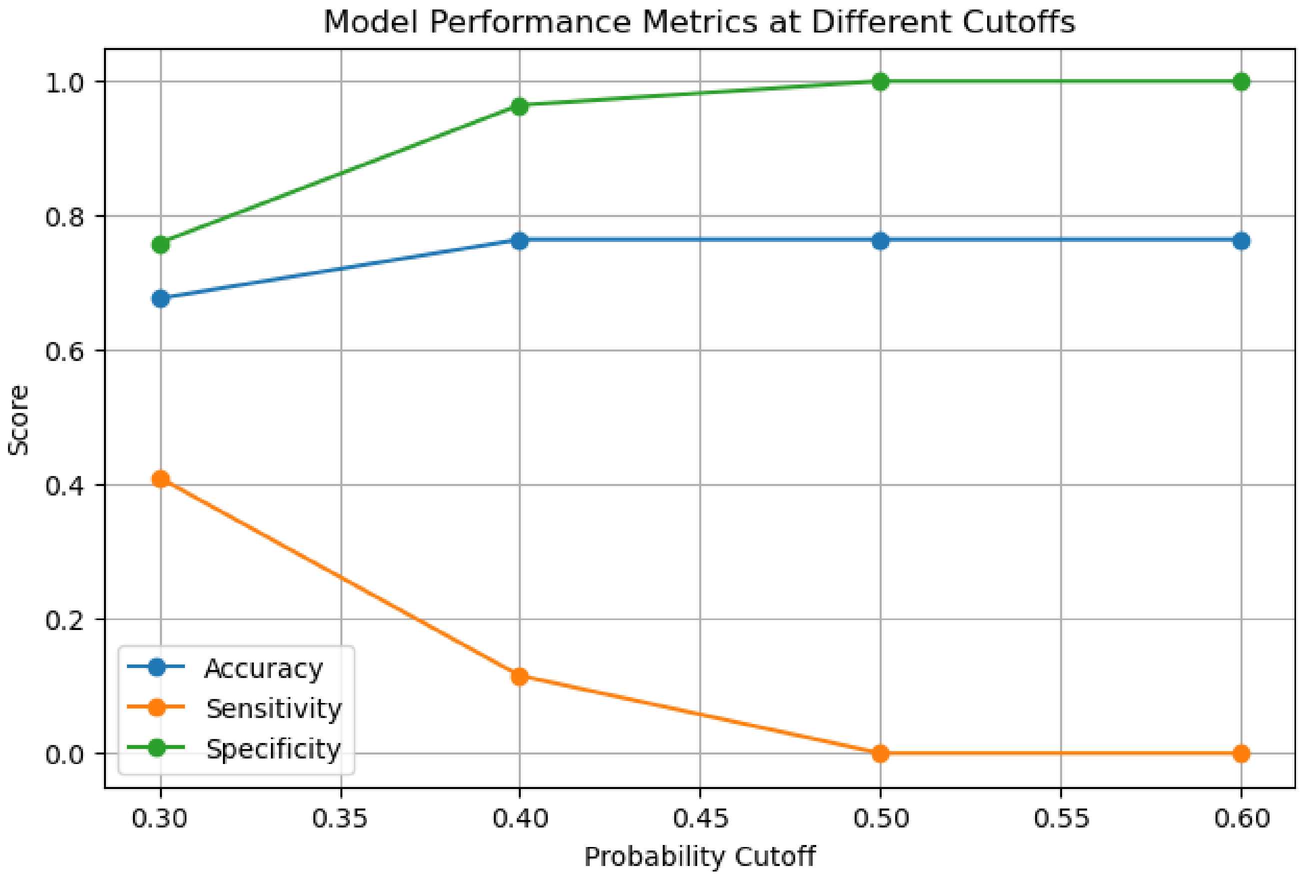
Fraud Detection Model Analysis



5. Cutoff Analysis

Different cutoffs were analyzed for their impact on sensitivity, specificity, and accuracy. A cutoff around 0.2 was ideal for prioritizing fraud detection. Below is the plot showing this tradeoff.

Fraud Detection Model Analysis



6. Conclusion & Recommendation

For operational use where detecting fraudulent cases is critical, the Random Forest model with a 0.2 cutoff is recommended. For cases requiring transparency, Logistic Regression offers interpretability. Model choice should be aligned with business goals.