

Summary and recommendation

Based on model performance, we recommend deploying the Random Forest model, as it offers strong recall and specificity, effectively identifying fraudulent claims. Integrate the model into the claims workflow to automatically score and flag high-risk claims for manual review. Adjust the classification threshold based on the business cost of false positives vs. missed fraud. Regularly retrain the model to adapt to evolving fraud patterns. Combine model predictions with business rules for improved accuracy. Use SHAP or feature importance for explainability to stakeholders. Finally, train fraud analysts to interpret and act on model outputs, enhancing decision-making and fraud investigation efficiency.

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

To detect fraud, analyze historical claim data using exploratory data analysis to identify unusual patterns like frequent claims, high amounts, or early claims after policy start. Engineer features that capture customer behavior and claim timing. Apply machine learning models like Random Forest or XGBoost, using oversampling to handle imbalance. Use evaluation metrics like recall and precision, and interpret results with SHAP to uncover meaningful fraud indicators.

Which features are most predictive of fraudulent behavior?

features	Imp
incident_severity_Minor Damage	0.088142
incident_severity_Total Loss	0.057772
authorities_contacted_Unknown	0.036399
incident_severity_Trivial Damage	0.035396
policy_annual_premium	0.033319
collision_type_Unknown	0.024426
vehicle_claim_range_40k-60k	0.023208
insured_relationship_own-child	0.021713
insured_relationship_other-relative	0.020784
combined_limit_500	0.019508
insured_sex_MALE	0.019384
incident_type_Parked Car	0.019108
incident_type_Vehicle Theft	0.017087
insured_occupation_priv-house-serv	0.015043
property_claim_range_5k-10k	0.014860

These are the Important features to predict the fraudulent behavior

Can we predict the likelihood of fraud for an incoming claim, based on past data?

Yes, we can predict the likelihood of fraud for an incoming claim using historical claim data. By training machine learning models like Logistics regression Random Forest on past labeled data (fraudulent vs. non-fraudulent), the model learns patterns associated with fraud. Once trained, it can assign a probability score to new claims, helping prioritize high-risk cases for further investigation or automatic flagging.

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

The Random Forest model highlights key fraud indicators such as claim timing, amount, and frequency. High recall and specificity suggest it effectively detects fraud while minimizing false positives. Feature importance helps identify strong fraud predictors, enabling targeted rule creation. These insights support better claim triaging, enhance investigator focus on high-risk claims, and allow continuous improvement of fraud strategies through regular model monitoring and feedback.

Answers
