Fraudulent Claim Detection

By Anshaj Singh

Executive Summary

This project focuses on developing a predictive system to identify fraudulent insurance claims using historical claim data. The analysis followed a structured data science approach involving data cleaning, exploratory data analysis, feature engineering, and model development.

Two machine learning models—**Logistic Regression** and **Random Forest**—were built and evaluated. While Logistic Regression provided a strong and interpretable baseline, the **tuned Random Forest model demonstrated superior performance** on unseen validation data, particularly in identifying fraudulent claims. The final model achieved higher recall and F1-score, making it more suitable for fraud detection use cases where missing fraudulent claims is costly.

Key Findings

- Fraudulent claims represent a **minority class**, confirming the need for imbalance handling techniques.
- **Claim severity and claim composition** (injury, vehicle, and property claim amounts) are strong indicators of fraud.
- Certain **incident types, collision categories, and severity levels** exhibit significantly higher fraud likelihood.
- Logistic Regression achieved good baseline performance but showed lower fraud recall on validation data.
- The Random Forest model captured **non-linear relationships and feature interactions**, leading to improved fraud detection performance.
- Hyperparameter tuning and feature selection significantly enhanced model stability and generalisation.

Model Performance Summary

Model	Validation Accuracy	Fraud Recall	Precision	F1 Score
Logistic Regression	~82%	~69%	~63%	~66%
Random Forest	~84%	~80%	~64%	~71%

Recommendations

Adopt the Random Forest model for fraud screening
Its higher recall ensures more fraudulent claims are flagged early.

Use probability-based thresholds instead of fixed rules
Adjusting cutoffs allows the business to balance investigation cost and fraud risk dynamically.

Focus investigations on high-severity claims
Claims with unusually high or imbalanced claim components should be prioritised for review.

Periodically retrain the model
Fraud patterns evolve over time; retraining with recent data will maintain effectiveness.

Use feature importance for audit transparency
Feature importance scores can help justify why certain claims are flagged, aiding trust and compliance.

Business Insights

- **Reduced financial losses:** Early detection prevents payouts on fraudulent claims.
- **Operational efficiency:** Automated screening reduces reliance on manual investigations.
- **Improved customer experience:** Genuine claims face fewer delays and unnecessary checks.
- **Scalable fraud detection:** The model can be integrated into real-time claim processing systems.
- **Data-driven decision-making:** The approach provides measurable, explainable insights for risk management teams.

This project aims to answer the following questions

- How can we analyse historical claim data to detect patterns that indicate fraudulent claims?
- Which features are most predictive of fraudulent behaviour?
- Can we predict the likelihood of fraud for an incoming claim, based on past data?
- What insights can be drawn from the model that can help in improving the fraud detection process?

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

Historical claim data can be analysed using a combination of **exploratory data analysis (EDA), bivariate analysis, and statistical modeling**, as demonstrated in the notebook.

In this analysis:

- **Univariate analysis** on training data revealed that claim amount variables (total, injury, vehicle, and property claims) are right-skewed, with higher claim values occurring less frequently.
- **Bivariate analysis** between features and the target variable (fraud_reported) showed that certain claim characteristics have a higher likelihood of fraud. For example:
    - Claims with **major incident severity** had significantly higher fraud rates.
    - **Single-vehicle collisions** and **rear collisions** showed higher fraud likelihood compared to other incident types.
- **Fraud likelihood analysis** using group-by calculations helped quantify how fraud probability changes across different categories (incident type, collision type, authorities contacted, etc.).
- **Correlation analysis** highlighted strong relationships among claim amount variables, suggesting that claim magnitude and composition are important fraud indicators.

Together, these analyses help uncover **recurring patterns and risk signals** in historical claims that differentiate fraudulent claims from legitimate ones.

Which features are most predictive of fraudulent behaviour?

Based on EDA, feature engineering, and model-based feature importance from the notebook, the most predictive features of fraud include:

- **Incident severity** (especially "Major Damage" cases)
- **Claim amount variables**, particularly:
    - total_claim_amount
    - injury_claim
    - vehicle_claim
    - Engineered ratios such as injury_claim_ratio and vehicle_claim_ratio
- **Claim severity score**, created by aggregating different claim components
- **Incident type and collision type**
- **Authorities contacted** (claims involving ambulance or fire services showed higher fraud likelihood)
- **Number of vehicles involved** and **incident hour of the day**

These features consistently appeared as:

- High-impact variables during bivariate analysis
- Important features in Random Forest feature importance
- Significant contributors to model performance in both Logistic Regression and Random Forest models

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

- Yes, the notebook demonstrates that it is possible to predict the probability of fraud for an incoming claim using historical data.

This was achieved by:

- Training supervised classification models (Logistic Regression and Random Forest) on past claim data
- Generating **predicted probabilities** instead of just class labels
- Applying **probability cutoff analysis** to convert probabilities into fraud / non-fraud decisions

The Logistic Regression model provided probability scores that allow flexible cutoff selection, while the Random Forest model achieved **higher validation accuracy and recall**. Validation results confirmed that these models generalise well to unseen data, indicating that they can reliably estimate fraud likelihood for new incoming claims.

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

Several actionable insights were derived from the modeling process:

1. **Fraud detection should prioritise recall over accuracy**
   Missing a fraudulent claim is more costly than investigating a legitimate one. Models with higher fraud recall (such as Random Forest) are therefore more suitable for real-world deployment.
2. **Claim severity is a key fraud indicator**
   Claims with disproportionately high or imbalanced claim components (e.g., high injury claims relative to total amount) are more likely to be fraudulent.
3. **Not all features contribute equally**
   Feature importance analysis showed that a subset of variables carries most of the predictive power. Removing low-importance features improves model stability and reduces overfitting.
4. **Probability cutoff selection matters**
   Adjusting the probability threshold significantly changes fraud detection performance. This allows insurers to tune the system based on business risk tolerance.
5. **Ensemble models outperform linear models**
   While Logistic Regression offers interpretability, Random Forest better captures non-linear relationships and interactions, leading to improved fraud detection performance.

These insights can help insurers:

- Flag high-risk claims early
- Allocate investigation resources more efficiently
- Reduce financial losses due to undetected fraud
- Improve customer experience by avoiding unnecessary scrutiny of low-risk claims

Conclusion

This project demonstrates the successful application of machine learning techniques to insurance fraud detection. By combining exploratory analysis, feature engineering, and robust model evaluation, the final solution delivers meaningful business value. The tuned Random Forest model provides a reliable and scalable framework for proactive fraud detection and can serve as a foundation for future enhancements.