Stakeholder Summary
Final Project: Feature Discovery for Predictive Analytics
Diana Larkin

Executive Summary

Fraudulent insurance claims pose a significant risk to operations, driving investigation costs and eroding customer trust. Unsupervised learning was applied to discover hidden patterns in claims data and then tested whether those discoveries improved fraud prediction. The analysis revealed two distinct claim segments: Straightforward Minor Claims and Severe Multi-Vehicle Losses. Integrating these features into predictive models provided modest but meaningful improvements. Logistic Regression remained the most reliable model, while Gradient Boosting benefited slightly from the new features.

The Feature Discovery

Using clustering and PCA, the dataset was segmented into two groups:

- **Straightforward Minor Claims:** Single vehicle incidents, minor or trivial damage, clearer property damage categories, lower fraud rate (23%).
- **Severe Multi-Vehicle Losses:** Multi-vehicle collisions, major damage or total loss, ambiguous property damage, higher fraud rate (27%).

PCA confirmed visual separation between these groups, even though it explained only ~17% of the variance. Together, these methods highlighted that fraud risk is shaped by both incident complexity and policy context.

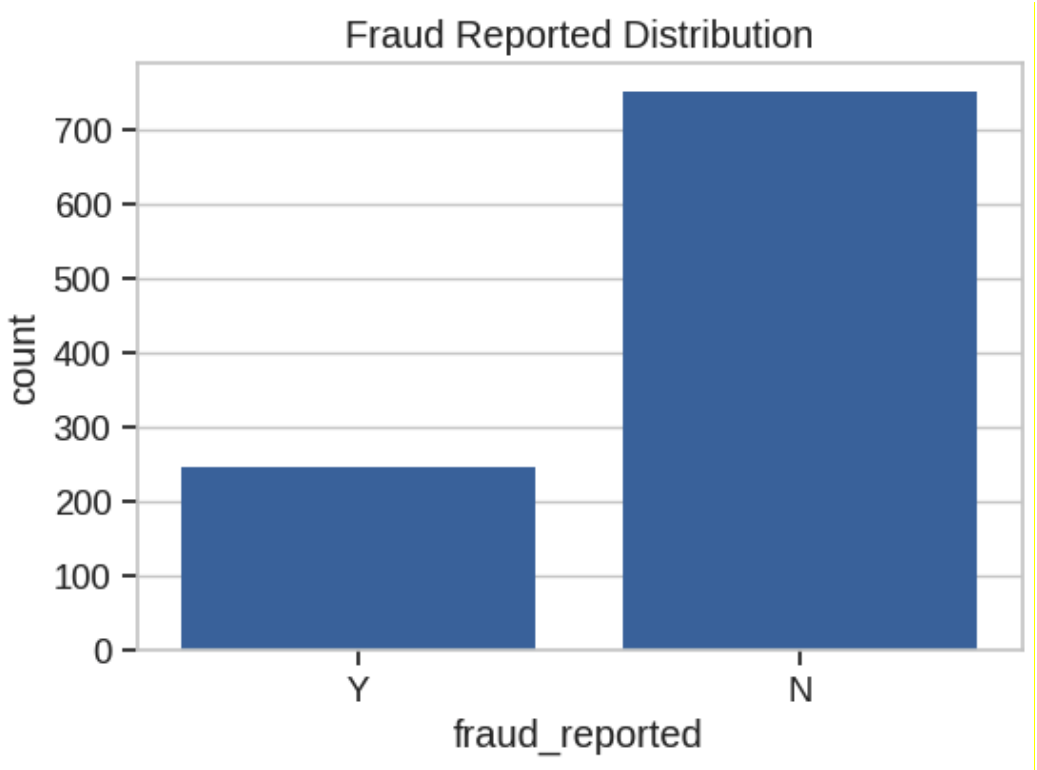

Figure 1: Most claims are legitimate, highlighting class imbalance and the importance of careful fraud detection.

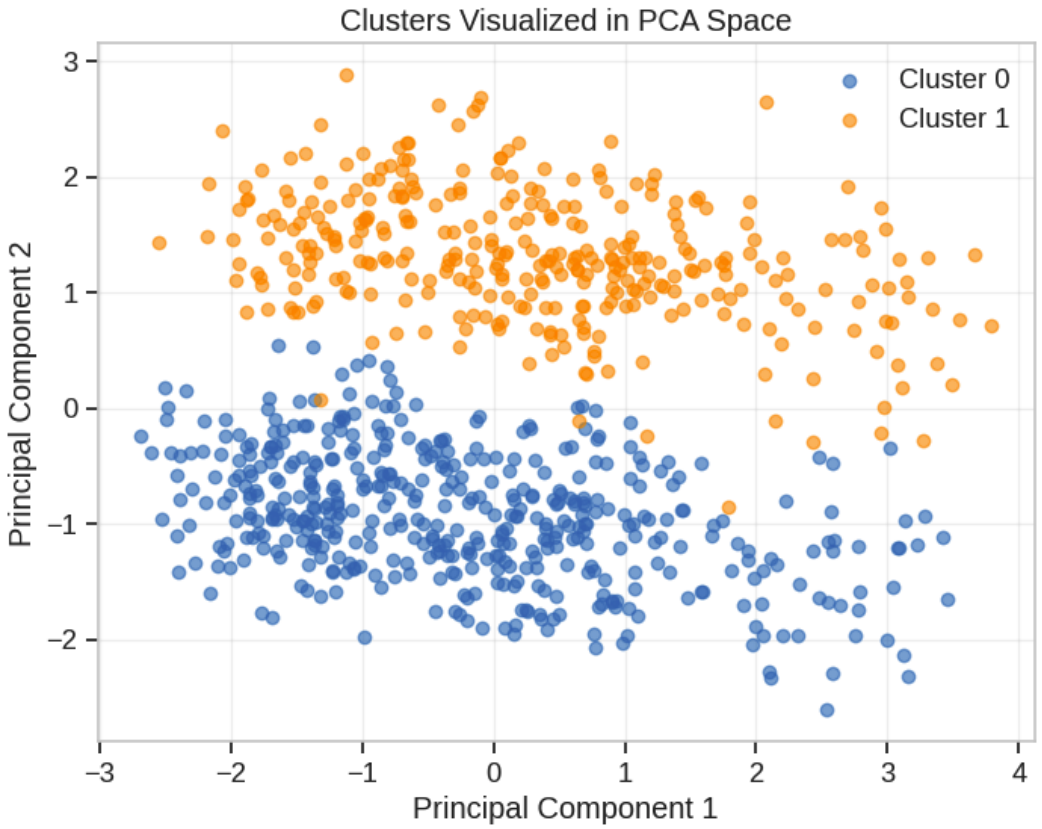

Figure 2: PCA scatterplot shows clear separation between Straightforward Minor Claims and Severe Multi-Vehicle Losses.

The Predictive Models

The target variable was whether a claim was fraudulent (Y/N). Three models were tested:

- Logistic Regression (baseline): Accuracy 0.85, Recall 0.84, F1 0.73 — strongest overall performance.
- Random Forest: Accuracy 0.81, Recall 0.51, F1 0.57 — weaker sensitivity to fraud.
- Gradient Boosting: Accuracy 0.81, Recall 0.55, F1 0.58 — balanced but below Logistic Regression.

When cluster membership and PCA components were added:

- Logistic Regression remained unchanged.
- Random Forest gained slightly in precision (0.64 → 0.66).
- Gradient Boosting improved recall (0.55 → 0.59) and F1 (0.58 → 0.60).

Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression (Baseline)	0.8500	0.6508	0.8367	0.7321
Logistic Regression (With Features)	0.8500	0.6508	0.8367	0.7321
Random Forest (Baseline)	0.8100	0.6410	0.5102	0.5682
Random Forest (With Features)	0.8150	0.6579	0.5102	0.5747
Gradient Boosting (Baseline)	0.8050	0.6136	0.5510	0.5806
Gradient Boosting (With Features)	0.810	0.6170	0.5918	0.6042

Table 1: Comparison of baseline vs. feature-augmented models. Logistic Regression remained strongest overall, while Gradient Boosting showed modest improvement with discovered features.

Business Recommendations

Based on these findings, the following actions are recommended:

1. Segment claims handling: Apply differentiated strategies for Straightforward Minor Claims vs. Severe Multi-Vehicle Losses.
2. Prioritize Logistic Regression: Use it as the operational baseline model due to its strong recall and interpretability.
3. Enhance Gradient Boosting: Explore hyperparameter tuning and resampling (e.g., SMOTE) to further improve recall.
4. Monitor ambiguous categories: Pay closer attention to claims with unclear property damage or missing authority contact, as these align with higher fraud risk.
5. Expand feature engineering: Investigate additional latent features (e.g., interaction terms, distance from cluster center) to capture more complex fraud patterns.

Limitations & Next Steps

- PCA explained limited variance, so its role is primarily visualization, not dimensionality reduction.
- Fraud imbalance remains a challenge; future work should test advanced resampling methods.
- External validation is needed to confirm generalizability beyond this dataset.
- With more time, deeper feature engineering and ensemble tuning could yield stronger improvements.

Closing Reflection

This project demonstrated that unsupervised learning can uncover meaningful fraud risk segments. While predictive gains were modest, the discoveries provide valuable context for claims handling and highlight opportunities for future model refinement. The strongest insight is that fraud risk is not random; it clusters around incident complexity and policy structures.