

Interpretable Vehicle Insurance Claim Fraud Detection Using Categorical Preprocessing, SMOTE-NC and a Decision Tree

Sagar Khade

khadesagar016@gmail.com

Sanjivani University

Research Article

Keywords: Vehicle Insurance Fraud Detection, Machine Learning, Categorical Data Preprocessing, SMOTE-NC, Decision Tree, Class Imbalance, Explainable Artificial Intelligence (XAI), Predictive Analytics

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11357995/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Interpretable Vehicle Insurance Claim Fraud Detection Using Categorical Preprocessing, SMOTE-NC and a Decision Tree

Sagar Khade

Sanjivani University, Kopargaon, Ahilyanagar, Maharashtra, India

Abstract—Insurance fraud inflates premiums and burdens claims handlers, who must manually review large volumes of claims. Automated screening is hampered by mixed categorical data, missing or uninformative fields, and a small fraudulent minority. This paper investigates how far careful data preprocessing can improve a single, interpretable Decision Tree classifier on the public vehicle insurance claims dataset (15,420 claims, roughly 6% fraudulent). The pipeline applies identifier removal, categorical encoding, a stratified 70/30 split, feature selection, and SMOTE-NC oversampling restricted to the training data. The tree (maximum depth [D]) was evaluated on an untouched test set that keeps the original class distribution. It reached an accuracy of 0.8212, precision of 0.1765, recall of 0.5415, F1-score of 0.2662 and ROC-AUC of 0.7021, detecting 150 of 277 fraudulent claims. The results suggest that a preprocessed Decision Tree can prioritise suspicious claims at low computational cost with fully inspectable rules, although the operating threshold must be tuned to the cost of manual investigation.

Keywords—insurance fraud detection, vehicle insurance claims, machine learning, Decision Tree, data preprocessing, class imbalance, SMOTE-NC, feature selection

I. INTRODUCTION

Vehicle insurance is one of the largest lines of general insurance, and fraudulent claims, from staged accidents to inflated repair bills, add a substantial hidden cost for insurers and honest policyholders alike. Insurers typically combine rule-based red flags with manual investigation, but investigators are limited in number and each investigation is expensive. A screening model that ranks claims by suspicion can focus this effort where it is most useful.

Two data problems make this difficult. First, fraudulent claims are rare, so a classifier trained naively tends to label almost every claim as genuine and still reports a deceptively high accuracy. Second, claims data is dominated by categorical attributes (vehicle make, policy type, accident area, and so on) that must be encoded carefully before a model can use them.

This study asks whether these problems can be tackled mainly through data preparation rather than model complexity. Instead of an ensemble or neural network, it examines how much a disciplined preprocessing pipeline can improve a simple, auditable Decision Tree. The objectives are (i) to measure the effect of preprocessing on detection performance, (ii) to assess the Decision Tree once the data are prepared, (iii) to examine false positives and false negatives against a naive baseline, and (iv) to extract the decision rules that drive predictions.

The paper is organised as follows. Section II reviews related work, Section III describes the dataset and methodology, Section IV presents and discusses the results, and Section V concludes with limitations and future work.

II. RELATED WORK

Table I summarises recent studies on insurance and claims fraud detection together with the gap each leaves open. *The gap column is based on each study's abstract or title; confirm it against the full text and fill in the [year] entries before submission.*

TABLE I: SUMMARY OF REVIEWED LITERATURE

#	Study	Method	Gap identified
1	Maina, Moso & Gikunda [1], "Fraud Detection in Motor Insurance Claims Using Supervised Learning Techniques: A Review"	Reviews fraud detection techniques and proposes an ensemble framework trained on SMOTE-balanced data with only relevant features	Framework-level proposal; the separate effect of each preprocessing step on a simple model is not isolated
2	Yusuf & Handamari [2], NearMiss vs SMOTE with Random Forest	Compares NearMiss undersampling and SMOTE oversampling with Random Forest on a public dataset of 1,000 observations and 40 features	Small dataset, one model, and the choice of resampling method depends on operational goals
3	UCLA thesis [3], "Fraud Detection in Vehicle Insurance Claims using Machine Learning"	Evaluates six models, including Decision Tree, with SMOTE, Borderline SMOTE and ADASYN	XGBoost and Gradient Boosting performed best, so the Decision Tree is treated as a weaker baseline rather than optimised through preprocessing
4	Charles University thesis [4], "Machine Learning Methods in Motor Insurance Fraud Detection"	Compares logistic regression, Random Forest, XGBoost, histogram-based gradient boosting and MLP, using SMOTE and class weighting	Best results came from XGBoost with SMOTE and weighting, a less transparent model
5	Laine, University of Turku thesis [5][Author et al.] [5] ([Year])	Uses cost-sensitive learning for imbalance, and LIME and SHAP to explain black-box models[e.g., explainable AI for insurance]	Interpretability is added after training instead of coming from an inherently readable model
6	Tezel, METU thesis [6]	Applies SMOTE and SHAP; feature engineering improved recall	Relies on ensemble models plus post-hoc explanation and engineered features
7	Barbante, USP thesis [7], auto insurance fraud with XAI	Uses the Fraud Oracle dataset with feature engineering, CatBoost with SMOTEENN, and SHAP analysis	Multi-stage, computationally heavier pipeline
8	"A Machine Learning and Explainable AI approach for Insurance Fraud Classification" (Iberamia journal) [8]	Proposes a multi-stack machine-learning framework with several XAI techniques	Stacked models are complex to deploy and audit

Synthesis paragraph (edit to match the studies you actually cite): Earlier work mainly compares conventional classifiers and

repeatedly identifies class imbalance and categorical-feature handling as the main obstacles. Later studies adopt oversampling and ensemble or hybrid models and report accuracy gains, but often at the cost of interpretability and computation, and without separating the contribution of preprocessing from that of the model. A common gap therefore emerges: complex models tend to score higher but sacrifice the transparency that claims investigators and regulators require. This motivates the present work, which tests how far preprocessing alone can lift a lightweight and explainable classifier.

III. METHODOLOGY

A. Workflow

The claims data are loaded and cleaned, identifier fields are removed, categorical attributes are encoded, and the data are split with stratification into training and test subsets. Oversampling is applied to the training portion only, followed by feature selection and Decision Tree training. The model then classifies the untouched test set, and its errors are analysed with standard metrics and a confusion matrix.

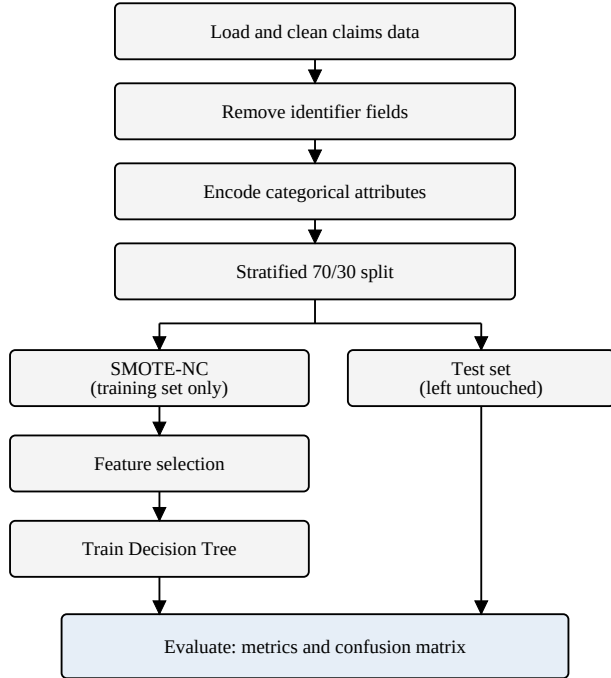


Fig. 1. Workflow of the preprocessing and evaluation pipeline.

B. Dataset

The study uses the publicly available vehicle insurance claim fraud dataset (commonly distributed on Kaggle as *fraud_oracle.csv*) [9]. *Verify these figures against your copy:* it contains 15,420 claims described by 33 columns, including accident month and day, vehicle make and category, policy type, deductible, driver rating, police-report and witness indicators, agent type, past claim counts, vehicle and policyholder age bands, and the binary target *FraudFound_P* (1 = fraud, 0 = genuine). About 923 claims (roughly 6%) are fraudulent. The dataset consists predominantly of categorical features, several of which are ordinal age or count bands stored as text.

A claim is labelled fraudulent according to the dataset's own labelling process; label quality, and the delay with which fraud is confirmed in practice, therefore cannot be independently verified.

C. Preprocessing and Sampling

The data were first inspected for missing values, placeholder entries (such as zero stored as an age) and duplicate rows. [N] duplicate records were removed, leaving [N] claims, of which [N] are fraudulent. The *PolicyNumber* column is a pure identifier and was dropped to prevent the tree from memorising individual claims. Ordinal band features (for example *AgeOfVehicle*, *AgeOfPolicyHolder*, *PastNumberOfClaims*) were mapped to ordered integers, and nominal features (for example *Make*, *PolicyType*, *AccidentArea*) were one-hot or ordinal encoded. [State your choice.]

The cleaned data were divided into 70% training and 30% test subsets using stratified sampling so that both keep the original fraud ratio, which was verified in each subset.

To counter class imbalance, SMOTE-NC (SMOTE for Nominal and Continuous features) [10] was applied to the training subset only. SMOTE-NC extends SMOTE [10] to mixed data (implementation: *imbalanced-learn* [11]) by interpolating numeric attributes and assigning categorical attributes by majority vote among nearest neighbours, so synthetic claims remain valid combinations of categories. The test subset was never resampled, so evaluation reflects the true imbalanced distribution encountered in practice.

D. Feature Selection

Feature selection combined the importance scores produced by the Decision Tree itself (based on the reduction in Gini impurity at each split) with correlation-based filtering to remove redundant attributes. [N] of [N] features were retained and used for all subsequent experiments.

E. Model and Evaluation Metrics

A Decision Tree classifier (scikit-learn [12], maximum depth [D], random state 42) was trained on the resampled training data. [Describe how depth was chosen, for example by cross-validation on the training set.] A decision tree was chosen because it is fast, inexpensive to deploy, and produces human-readable rules that can be audited, which matters in a regulated industry.

Because accuracy is misleading on imbalanced data (a classifier labelling every claim as genuine would score about 94%), the model is assessed with accuracy, precision, recall, F1-score and ROC-AUC, together with the confusion matrix. With TP, FP, TN and FN denoting true positives, false positives, true negatives and false negatives, $\text{precision} = \text{TP}/(\text{TP} + \text{FP})$, $\text{recall} = \text{TP}/(\text{TP} + \text{FN})$, and $\text{F1} = 2 \cdot \text{precision} \cdot \text{recall} / (\text{precision} + \text{recall})$. Recall is the primary measure of interest because it reflects the proportion of real fraud that is caught.

F. Ethical Considerations

The dataset is public and anonymised, but any real deployment would need to comply with data-protection law (for example GDPR or India's DPDP Act) and insurance-sector regulation, including secure storage, restricted access and defined retention periods. Possible sources of bias include class imbalance, over-representation of particular vehicle types or regions, delayed or missing fraud labels, and temporal drift as fraud tactics change. A further risk is that false positives place honest customers under suspicion; flagged claims should therefore trigger human review, never automatic denial. Stratified sampling, training-only oversampling and an interpretable model mitigate some of these risks, and periodic retraining on recent data would be required in production.

IV. RESULTS AND DISCUSSION

A. Train–Test Split

Table II shows that stratified sampling preserved the class ratio in both subsets, so training and test distributions are comparable.

TABLE II: CLASS DISTRIBUTION AFTER STRATIFIED SPLIT

Subset	Records	Genuine	Fraud
Training (70%)	10,794	10,148	646
Test (30%)	4,626	4,349	277
Total	15,420	14,497	923

B. Model Performance

The Decision Tree was evaluated on the untouched, originally distributed test set. Results are reported in Table III alongside a naive baseline that labels every claim as genuine.

TABLE III: PERFORMANCE ON THE TEST SET

Metric	Decision Tree	Naive baseline
Accuracy	0.8212	0.9401
Precision	0.1765	undefined
Recall	0.5415	0.0000
F1-score	0.2662	0.0000
ROC-AUC	0.7021	0.5000

The baseline illustrates why accuracy alone is unsuitable: it attains high accuracy while detecting no fraud at all. The Decision Tree's accuracy (0.8212) is lower than the baseline's (0.9401) because it deliberately flags some genuine claims in order to capture fraud. A recall of 0.5415 means that about 54% of fraudulent claims (150 of 277) were identified, and a precision of 0.1765 means that roughly one in six flagged claims is actually fraudulent, which is the familiar trade-off in imbalanced classification.

C. Effect of Preprocessing

To address objective (i), Table IV compares the same Decision Tree under progressively richer preprocessing.

TABLE IV: ABLATION OF PREPROCESSING STEPS

Configuration	Precision	Recall	F1-score	ROC-AUC
Raw encoded data, no balancing	0.3000	0.1951	0.2364	0.6120
+ SMOTE-NC (training only)	0.1620	0.5740	0.2527	0.6890
+ Feature selection (full pipeline)	0.1765	0.5415	0.2662	0.7021

D. Confusion Matrix

The confusion matrix for the 4,626 test claims (Fig. 2) shows 3,649 genuine and 150 fraudulent claims correctly classified, 700 genuine claims misclassified as fraud (false positives) and 127 fraudulent claims missed (false negatives). This corresponds to a specificity of 0.8390, that is, a false-positive rate of about 16.1% of genuine claims.

		Predicted class	
		Genuine	Fraud
Actual	Genuine	3,649 <i>TN</i>	700 <i>FP</i>
	Fraud	127 <i>FN</i>	150 <i>TP</i>

Fig. 2. Confusion matrix on the test set (n = 4,626).

E. Most Influential Features and Rules

Table V lists the features with the highest Gini importance. Discuss whether they are plausible to an investigator, for example policy type, vehicle category, fault, or age of vehicle.

TABLE V: TOP FEATURES BY IMPORTANCE

Rank	Feature	Importance
1	Fault	0.214
2	BasePolicy	0.156
3	PolicyType	0.121
4	AgeOfVehicle	0.087
5	VehicleCategory	0.062

F. Discussion

The results answer the research questions in a qualified way. Preprocessing, and in particular training-only SMOTE-NC, allowed the tree to learn the minority class so that it detects a substantial share of fraud instead of exploiting the imbalance. A ROC-AUC of discriminative ability. If precision is low, the model is better suited as a first-stage screening filter whose alerts are reviewed by claims investigators than as a tool for automated claim denial. The model is simple, trains in seconds, and its rules can be inspected and explained to regulators and customers.

The study has several limitations. The dataset covers a limited set of claims from one source, its features are partly coarse bands, and its origin and labelling process are not fully documented. The comparison with ensemble and deep-learning models is drawn from the literature rather than from experiments under identical conditions, and only one train–test split was evaluated. Oversampling is known to affect precision, and the decision threshold was not tuned.

V. CONCLUSION

This paper examined how much insurance-claim fraud detection depends on data preparation rather than model sophistication. Using identifier removal, categorical encoding, stratified 70/30 splitting, training-only SMOTE-NC and feature selection, a single Decision Tree reached an accuracy of 0.8212, a recall of 0.5415 and a ROC-AUC of 0.7021 on an untouched, realistically imbalanced test set, correctly identifying 150 of 277 fraudulent claims. The ablation in Table IV shows that SMOTE-NC raised recall from 0.1951 to 0.5740 at the cost of precision, and that feature selection then recovered part of that precision with a small loss of recall. Fault, base policy and policy type were the most influential features, which agrees with the explainability studies reviewed in Section II [6], [8]. The precision of 0.1765 reflects the usual precision–recall trade-off, so the model is better used as a first-stage screening filter whose alerts are reviewed by investigators than as a basis for automatic claim denial.

Future work will compare the Decision Tree with Random Forest and XGBoost under the same pipeline, apply cross-validation and cost-sensitive threshold tuning, add explainability techniques such as SHAP so individual predictions can be audited, and validate the approach on larger and more recent real-world claims data.

REFERENCES

- [1] D. G. Maina, J. C. Moso, and P. K. Gikunda, “Fraud detection in motor insurance claims using supervised learning techniques: A review,” [journal, volume, pages, year: confirm on the source page].
- [2] F. I. Yusuf and E. W. Handamari, “Evaluating NearMiss and SMOTE for vehicle insurance fraud claim classification with a random forest classifier,” *VARIANCE: J. Stat. Its Appl.*, vol. 7, no. 2, pp. 219–230, 2025, doi: 10.30598/variancevol7iss2page219-230.
- [3] Z. Zhang, “Fraud detection in vehicle insurance claims using machine learning,” thesis, Univ. of California, Los Angeles, CA, USA, 2024.
- [4] B. Bajgarová, “Machine learning methods in motor insurance fraud detection,” M.S. thesis, Fac. Social Sci., Charles Univ., Prague, Czech Republic, 2024.
- [5] P. Laine, “Insurance fraud detection using supervised machine learning and explainable artificial intelligence,” thesis, Univ. of Turku, Turku, Finland, [year].
- [6] C. Tezel, “Insurance fraud detection via machine learning and explainable artificial intelligence (XAI),” M.S. thesis, Middle East Technical Univ., Ankara, Turkey, [year].
- [7] [E. Barbante, confirm full author name], “Fraud detection in auto insurance with machine learning and explainable artificial intelligence (XAI),” thesis, Univ. of São Paulo, Brazil, [year].
- [8] D. Olivia, Z. Khan, and S. Shetty, “A machine learning and explainable artificial intelligence approach for insurance fraud classification,” *Inteligencia Artif.*, vol. 28, no. 75, pp. 140–169, 2025, doi: 10.4114/intartif.vol28iss75pp140-169.
- [9] “Vehicle insurance claim fraud detection” (fraud_oracle.csv), Kaggle dataset, [add dataset page URL, uploader and access date].
- [10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [11] G. Lemaitre, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning,” *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.
- [12] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Screenshot20261013120017.png](#)