Fraudulent Claim Detection Report

1. Problem Statement

Insurance fraud represents a substantial challenge for insurance companies, leading to considerable financial losses and undermining trust in the claims process. The objective of this project is to develop a predictive model that can identify fraudulent insurance claims using historical data. Specifically, the task is to classify claims as either fraudulent (`'Y'`) or not (`'N'`) based on various features associated with each claim.

2. Methodology

2.1 Data Collection and Preparation

- The dataset includes multiple features related to claims, such as the type of accident, policy details, claim amounts, and more.
- The target variable is `fraud_reported`, which is highly imbalanced, with significantly fewer fraudulent (`'Y'`) claims compared to non-fraudulent (`'N'`) claims.

2.2 Preprocessing Steps

- Categorical variables were encoded using label encoding and one-hot encoding.
- Missing values were handled appropriately through imputation or removal.
- The dataset was split into training and testing sets to evaluate model performance.

2.3 Model Selection

- A **Random Forest Classifier** was selected for its ability to handle complex, nonlinear relationships and its robustness against overfitting.
- The model was trained on the prepared dataset and evaluated using several performance metrics.

3. Techniques Used

- **Random Forest Algorithm:** For classification due to its ensemble nature and ability to rank feature importance.
- **Evaluation Metrics:** AUC-ROC, Accuracy, Precision, Recall, and F1-Score.
- **Confusion Matrix:** To analyze prediction distribution across true positives, false positives, true negatives, and false negatives.
- **Feature Importance Analysis:** To identify which features contribute most to predictions.

4. Data Analysis Key Findings

4.1 Class Imbalance

- The target variable `fraud_reported` shows a **major class imbalance**, with a large proportion of non-fraudulent claims.
- This imbalance negatively impacts the model's ability to accurately detect the minority class (`'Y'`).

4.2 Model Performance

- **AUC-ROC Score:** 0.772 – indicating moderate discriminative power.
- **Accuracy:** 0.75 – reflects good overall prediction accuracy.
- **Precision for 'Y' class:** 0.40 – among claims predicted as fraudulent, only 40% were correct.
- **Recall for 'Y' class:** 0.03 – only 3% of actual fraudulent claims were correctly identified.
- **F1-Score:** 0.05 – due to low recall, the F1-score remains low.

4.3 Confusion Matrix Analysis

- High number of **True Negatives** – the model is effective in identifying legitimate claims.
- Very low number of **True Positives** – it often misses fraudulent claims.

4.4 Feature Importance

- Top 10 features identified by the model can be used for future **feature engineering** and **model simplification**.
- These features are critical for understanding what drives fraudulent behavior.

5. Visualizations

- **Confusion Matrix Heatmap**: Showed stark contrast between correct predictions for non-fraud vs. fraud.
- **ROC Curve**: Demonstrated the trade-off between sensitivity and specificity.
- **Feature Importance Plot**: Ranked the most influential features in the model’s decision-making.

6. Insights

- The model is biased towards the majority class (`'N'`) and performs poorly on the minority class (`'Y'`).
- Low recall indicates **most fraudulent claims go undetected**, which undermines the core purpose of the model.
- Despite moderate AUC-ROC and accuracy, the **business value is limited** due to poor performance on identifying frauds.

7. Actionable Outcomes

1. **Address Class Imbalance:**
    - Apply **SMOTE** (Synthetic Minority Over-sampling Technique) or **under sampling** to improve detection of fraudulent claims.
    - Experiment with **class weighting** in model training.
2. **Model Optimization:**
    - Test alternate algorithms like **XG Boost** or **Light GBM** which are known to handle imbalance well.
    - Perform **hyperparameter tuning** to improve generalization.
3. **Threshold Adjustment:**
    - Lowering the probability threshold for predicting fraud might increase recall at the cost of precision.
4. **Enhanced Feature Engineering:**
    - Use insights from feature importance to create new variables or interactions that might better capture fraud signals.
5. **Use in Practice:**
    - Instead of automated classification, use the model to **flag high-risk claims** for **manual review**.

8. Conclusion

The current model provides a foundational approach for fraud detection in insurance claims. However, it requires significant improvements, especially in handling class imbalance and boosting recall for fraudulent cases. By integrating advanced techniques and tuning strategies, the model can be enhanced to deliver better accuracy and actionable predictions that add tangible value to fraud prevention efforts.