Fraudulent Claim Case Study - Report

Overall Approach of the assignment

1. **Understand the Problem Statement**: The goal was to build a model to detect fraudulent insurance claims using historical claim data.
2. **Data Preparation and Cleaning**: Cleaned the data by handling missing values and encoding categorical features. Checked for class imbalance and prepared the dataset for modeling.
3. **Train-Validation Split (70-30)**: Divided the data into training (70%) and validation (30%) sets.
4. **Exploratory Data Analysis (EDA) on Training Data**: Analyzed feature distributions and visualized patterns. Used correlation analysis to find relationships between features and fraud labels.
5. **Handling Imbalanced Data:** Applied class weight balancing in models to give more importance to the minority (fraudulent) class.
6. **Feature Engineering**: Encoded categorical features (like claim type, customer occupation) into numerical values, created clean and meaningful features by handling missing values and standardizing formats, which helped the model learn better patterns. Focused on important features like claim amount, claim type, and customer profile based on feature importance analysis.
7. **Model Building**: Trained multiple machine learning models - Logistic Regression, Decision Tree, and Random Forest and Tuned the models and selected the best-performing one based on evaluation metrics.
8. **Predicting and Model Evaluation**: Evaluated using accuracy, precision, recall, F1-score, and confusion matrix. Focused more on recall to make sure fraudulent claims are detected effectively.
9. **Provide Insights**: Identified key features that predict fraud (like claim amount, claim type, and customer profile details) and derived business insights to help improve fraud detection processes.
10. **Conclusion and Business Impact:** Summarized how the model can save costs, improve efficiency, and enhance customer trust for the insurance company.

Problem Statement

- Global Insure, a leading insurance company, processes thousands of claims annually.
- However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses.
- The company's current process for identifying fraudulent claims involves manual inspections, which is time-consuming and inefficient.
- Fraudulent claims are often detected too late in the process, after the company has already paid out significant amounts.
- Global Insure wants to improve its fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process.
- This would minimise financial losses and optimise the overall claims handling process.

Methodology used:

1. Data Preparation
2. Data Cleaning
3. Train Validation Split 70-30
4. EDA on Training Data
5. Handling Imbalance
6. Feature Engineering
7. Model Building
8. Predicting and Model Evaluation

The following techniques are used:

- **Univariate analysis** - To understand the characteristics of a single variable in a dataset, offering insights into its distribution, central tendency, and variability
- **Bivariate analysis** – To investigate the relationships between categorical features and the target variable by analysing the target event likelihood (for the 'Y' event) for each level of every relevant categorical feature
- **Correlation analysis** - To investigate the relationships between numerical features to identify potential multicollinearity or dependencies.
- **RandomOverSampler** technique to balance the data and handle class imbalance. This method increases the number of samples in the minority class by randomly duplicating them, creating synthetic data points with similar characteristics. This helps prevent the model from being biased toward the majority class and improves its ability to predict the minority class more accurately.
- **Feature Engineering** technique used to improve the performance and accuracy of machine learning models by transforming raw data into more meaningful and useful features.

Key Insights

- Features like claim amount, claim type, and certain customer profile attributes showed a strong correlation with fraudulent behavior.
- Patterns such as unusually high claim amounts and certain claim types were strong indicators of potential fraud.
- Feature engineering steps, such as creating new interaction terms and handling missing values, contributed to performance improvements.