Summary Report: Fraudulent Claim Detection

1. Introduction

Global Insure, a leading insurance company, faces significant financial losses due to fraudulent claims. The current manual inspection process is inefficient, leading to late detection of fraud. This project aims to implement a data-driven approach to classify claims as fraudulent or legitimate early in the approval process, minimizing losses and optimizing claim handling.

2. Business Objective

To develop a predictive model that classifies insurance claims based on historical data, utilizing features such as claim amounts, customer profiles, and claim types.

3. Data Preparation

- **Data Source**: The dataset contains 1,000 rows and 40 columns, encompassing various features related to customer profiles and claim details.
- **Data Cleaning**:
  - Handled missing values in the authorities_contacted column (9.1% missing) by replacing with 'Other'.
  - Dropped the _c39 column due to 100% missing values.
  - Removed rows with illogical values, such as negative values in columns that should only contain positive values.

4. Exploratory Data Analysis (EDA)

- **Univariate Analysis**: Analyzed individual feature distributions using histograms and bar plots.
- **Bivariate Analysis**: Investigated relationships between categorical features and the target variable (fraud_reported) using count plots.
- **Correlation Analysis**: Identified multicollinearity among numerical features using correlation matrices and scatter plots.

5. Feature Engineering

- **Resampling**: Used RandomOverSampler to address class imbalance in the training set, achieving an equal distribution of fraudulent and legitimate claims.

- **New Features**: Created features from existing ones, such as extracting year, month, and day from dates, and interaction terms for better model performance.

6. Model Building

- **Logistic Regression**:
  - Utilized Recursive Feature Elimination with Cross-Validation (RFECV) to select significant features.
  - Evaluated model performance using metrics like accuracy, sensitivity, specificity, and F1-score.
  - Achieved an accuracy of approximately 86.3% on the training data.
- **Random Forest Classifier**:
  - Conducted hyperparameter tuning using Grid Search to optimize model performance.
  - Obtained an accuracy of approximately 83.3% on the validation data.

7. Evaluation Metrics

- **Logistic Regression**:
  - Accuracy: 86%
  - Sensitivity: 88.4%
  - Specificity: 85.4%
  - Precision: 66.8%
  - F1 Score: 76.1%
- **Random Forest**:
  - Accuracy:84%
  - Sensitivity: 79.7%
  - Specificity: 86.2%
  - Precision: 65.5%
  - F1 Score: 71.1%

8. Key Insights

- Higher fraud rates were associated with certain demographics, including:
  - Males
  - Customers with specific educational backgrounds (College students)
  - Certain occupations (e.g., exec-managerial, armed-forces)
  - Certain hobbies (e.g., chess, cross-fit)
- Incident types such as Multi-vehicle and Single Vehicle Collisions showed higher fraud likelihood.
- The number of vehicles involved and specific auto makes/models also correlated with fraudulent claims.

9. Conclusion

Both models (Logistic Regression and Random Forest) demonstrated strong predictive capabilities in identifying fraudulent claims. The logistic regression model provided slightly better performance in terms of accuracy and sensitivity. Insights from the analysis can significantly enhance the fraud detection process, allowing for better resource allocation and quicker responses to fraudulent activities.

10. Recommendations

- Implement the logistic regression model for initial fraud detection due to its interpretability and performance.
- Consider using the Random Forest model for complex scenarios where non-linear relationships between features may provide additional insights.
- Continuously monitor model performance and retrain with new data to adapt to evolving fraud patterns.
- Further research into high-risk demographics and claim types to fine-tune detection strategies.