Fraudulent Claim Detection

Problem statement: - Global Insure (an insurance company) processes thousands of claims annually, with a significant percentage being fraudulent, leading to financial losses. The current manual fraud detection process is inefficient, often detecting fraud too late. The company seeks to improve fraud detection by using data-driven insights to classify claims as fraudulent or legitimate early in the approval process, minimizing financial losses and optimizing claims handling.

Objective: Build a model that will classify claims as either fraudulent or legitimate based on historical claims' and customers' details. This model should help the organization to predict which claims are likely to be fraudulent before they are approved.

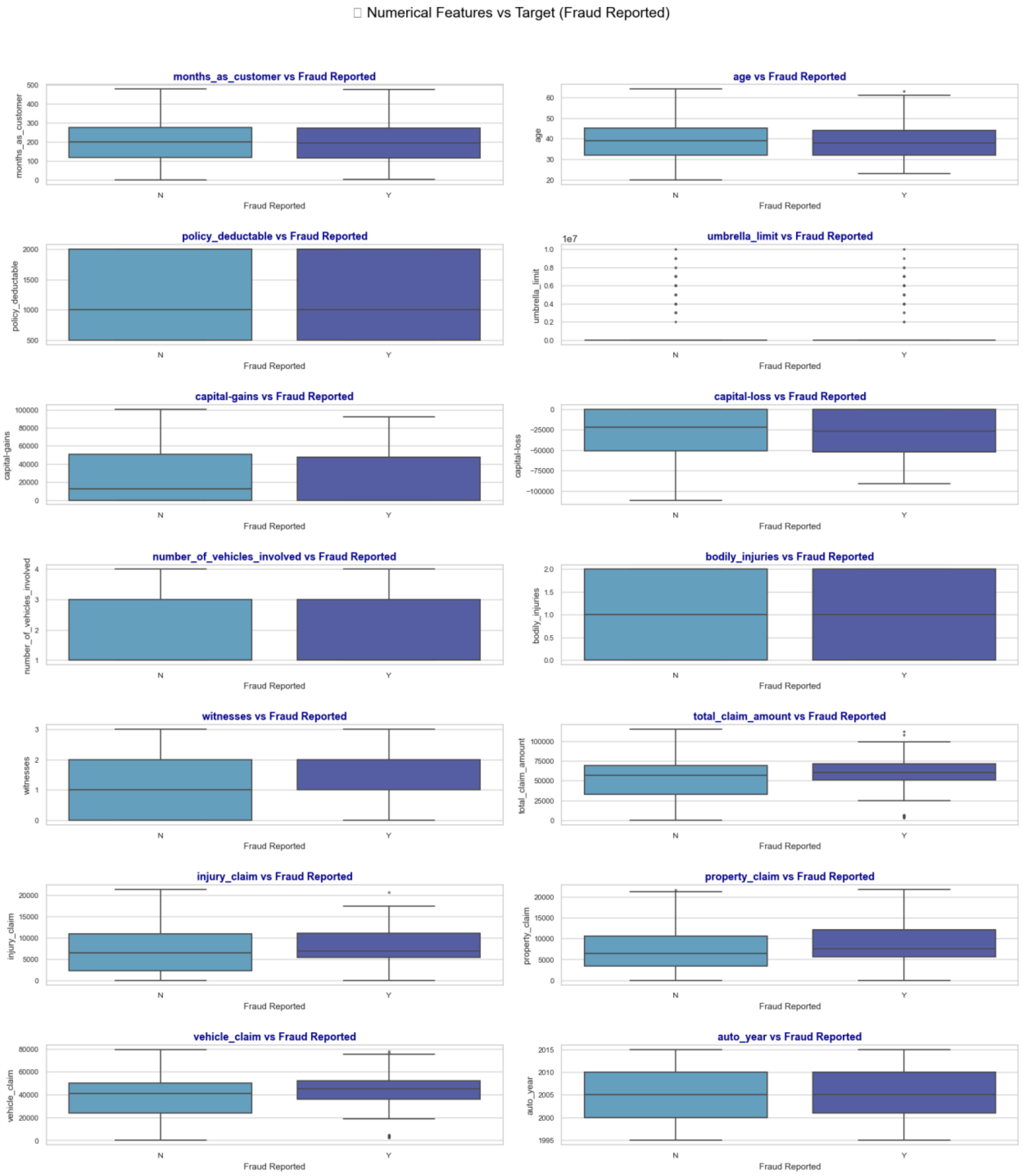
Methodology / Approach: The **CRISP-DM** (Cross-Industry Standard Process for Data Mining) methodology was followed to build the model. The steps included:

1. **Data Preparation**: Processed and prepped the data for analysis.
2. **Data Cleaning**: Handled missing values and fixed datatypes.
3. **Data Splitting**: Divided the data into training and test sets.
4. **Exploratory Data Analysis (EDA)**: Conducted thorough analysis on the training data to understand patterns and distributions.
5. **Feature Engineering**: Created new features to improve model performance.
6. **Model Building**: Developed two models — **Logistic Regression** and **Random Forest**.
7. **Feature Selection**:
    - For Logistic Regression used **RFECV**
    - For Random Forests used **Feature Importance**
8. **Model Fine-Tuning**:
    - Determined the optimal **cutoff** point for Logistic Regression.
    - Tuned hyperparameters using **GridSearchCV** to improve both models.
9. **Model Evaluation**: Assessed model performance using key metrics, including **Accuracy**, **Recall**, and **F1-Score**.

Techniques: Specific tools or methods used at different steps of model building are mentioned below:

1. **Data preprocessing**: One-Hot encoding, Feature scaling through StandardScalar
2. **Model building**: Logistic Regression, Random Forests
3. **Model Optimization**: Accuracy, Specificity & Sensitivity Cutoff graph, GridSearchCV
4. **Evaluation metrics**: Accuracy, Confusion Matrix, Recall

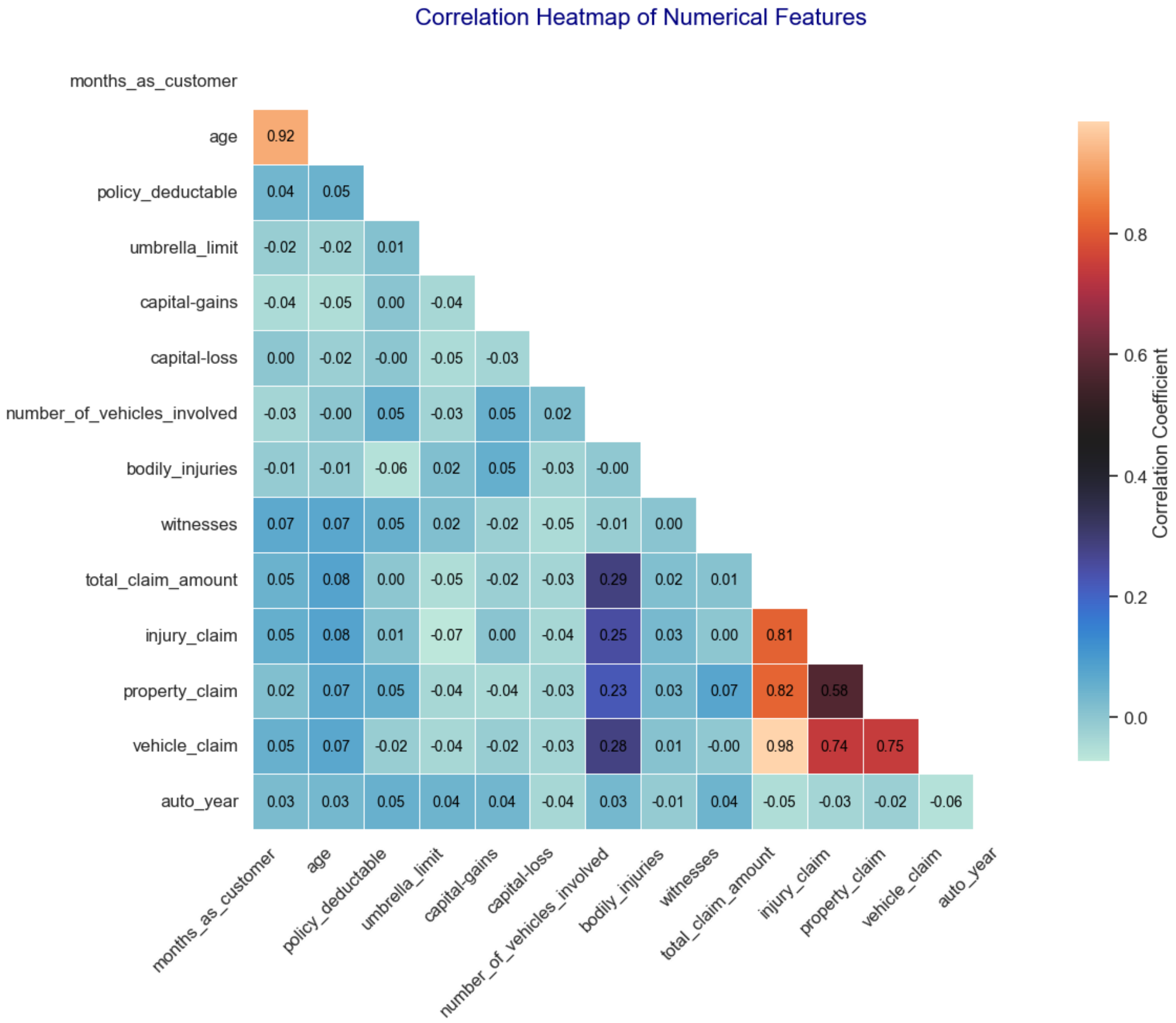
Key Insights: Several significant observations were made during the data analysis:



1. **Months_as_customer**: Fraudulent claims are more frequent among newer customers, suggesting a potential link between shorter customer tenures and higher fraud risk.
2. **Age**: Younger age groups (below 30) are more prone to fraudulent claims.
3. **Policy_deductible**: Fraudulent claims are concentrated within lower deductible brackets, indicating that lower deductibles might incentivize fraudulent behavior.
4. **Umbrella_limit**: No significant differences were observed between fraud and non-fraud cases, suggesting this feature may not be a strong predictor of fraud.
5. **Capital-gains**: Fraudulent claims tend to show slightly higher capital gains, potentially signaling a distinct fraud profile.
6. **Capital-loss**: Similar distributions in fraud and non-fraud cases, indicating low impact for predicting fraud.

7. **Number_of_vehicles_involved**: Multiple vehicles involved in a claim is linked with a higher likelihood of fraud.
8. **Bodily_injuries**: Fraud cases have a slightly higher count of bodily injuries, which may help flag suspicious claims.
9. **Witnesses**: A higher number of witnesses correlates with fraud cases, which could be useful for fraud verification.
10. **Claim components** (Injury_claim, Property_claim, Vehicle_claim): Fraudulent claims show higher amounts in these components.

Features that exhibit High correlation with other variables


Correlation Heatmap of Numerical Features

Months_as_customer, age, total_claim_amount, injury_claim, property_claim, vehicle_claim: multicollinearity was identified among features such as Months_as_customer, Age, Total_claim_amount, Injury_claim, Property_claim, and Vehicle_claim, indicating the need for potential dimensionality reduction

Results: Since the goal is to detect fraudulent claims, Recall becomes the most important evaluation metric for our models. Below is the comparison of both Logistic regression and Random Forests models:

MODEL	ACCURACY (TRAIN)	RECALL (VALIDATION)	ACCURACY (VALIDATION)	RECALL (VALIDATION)
LOGISTIC REGRESSION	86.6%	87.3%	84%	90.5%
RANDOM FORESTS	90.4%	93.3%	77.7%	75.7%

Logistic Regression outperformed **Random Forests** in terms of **Recall**, making it the better model for detecting fraudulent claims, with a **Recall of 90.5%** on the validation data.

Significant Features: The coefficient scores provide insights into the relative importance of each feature in predicting the target variable, which, in this case, is likely related to identifying whether a claim is fraudulent or not. Below is the Regression equation

1.5584 + 4.2153 (insured_hobbies_chess) + 3.5445 (insured_hobbies_cross-fit) – 22.7954 (insured_hobbies_dancing) – 3.4093 (incident_severity_Minor Damage) – 3.2046 (incident_severity_Total Loss)

Insights & Recommendations: Based on the model output, the following insights are drawn:

1. **insured_hobbies_chess:** Claims with the insured person having a hobby of chess are positively associated with fraud. The coefficient (4.2153) suggests that if the insured person plays chess, it increases the log-odds of a fraudulent claim
2. **insured_hobbies_cross-fit:** Similarly, claims with insured individuals who have a hobby of cross-fit are also positively associated with fraud. The coefficient (3.5445) indicates that these claims are more likely to be fraudulent.
3. **incident_severity_Minor Damage:** Claims with minor damage as the incident severity are negatively associated with fraud. The coefficient (-3.4093) suggests that minor damage claims are less likely to be fraudulent compared to more severe damage claims.
4. **incident_severity_Total Loss:** Claims with total loss as the incident severity are also negatively associated with fraud. The negative coefficient (-3.2046) implies that total loss incidents are less likely to be fraudulent compared to minor damage or other less severe claims.

Recommendations:

1. **Flag Claims with Chess and Cross-fit Hobbies**: Claims filed by individuals with chess or cross-fit as hobbies should be flagged for further verification. These claims should undergo additional scrutiny during the review process to minimize the risk of fraudulent payouts.
2. **Faster Processing for Minor Damage Claims**: Since these claims are less likely to be fraudulent, streamline the approval process for minor damage claims. This will help improve turnaround times and operational efficiency.
3. **Lower Fraud Risk for Total Loss Claims**: Given that total loss claims are less likely to be fraudulent, these claims could be processed with fewer checks or flagged as lower priority for manual review.

Additionally, fraud detection efforts should be focused on claims that don’t involve Total Loss and Minor damage as they will have higher likelihood of fraud

Case study team members: - Rakesh Kumar Sahoo, Rohit Vashishth, Saurabh Singh, Sandeep Santhosh