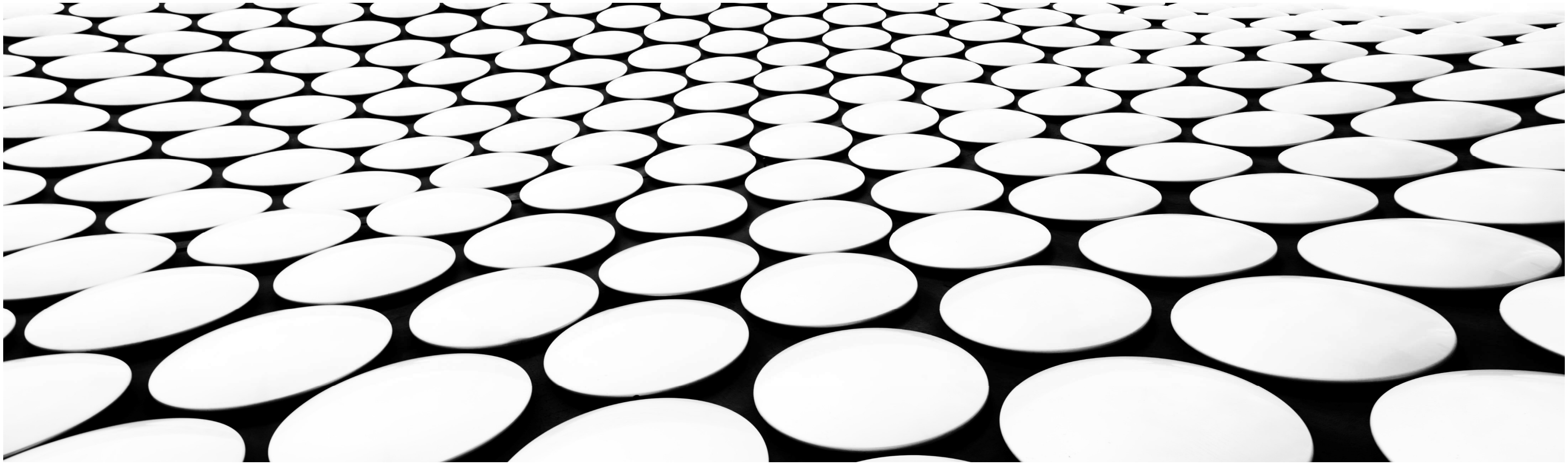
FRAUDULENT CLAIM DETECTION (CASE STUDY)

CHINMAYAJEET OJHA

CHAITANYA BANDARU



OBJECTIVE

This document will outline the summary of the findings from the data analysis of a leading insurance company, Global Insure. As part of this initiative, we have built a predictive model using the historical claim details to find patterns which can be used by the company to prevent fraudulent claims and eliminate financial risks associated with it.

SOLUTION APPROACH

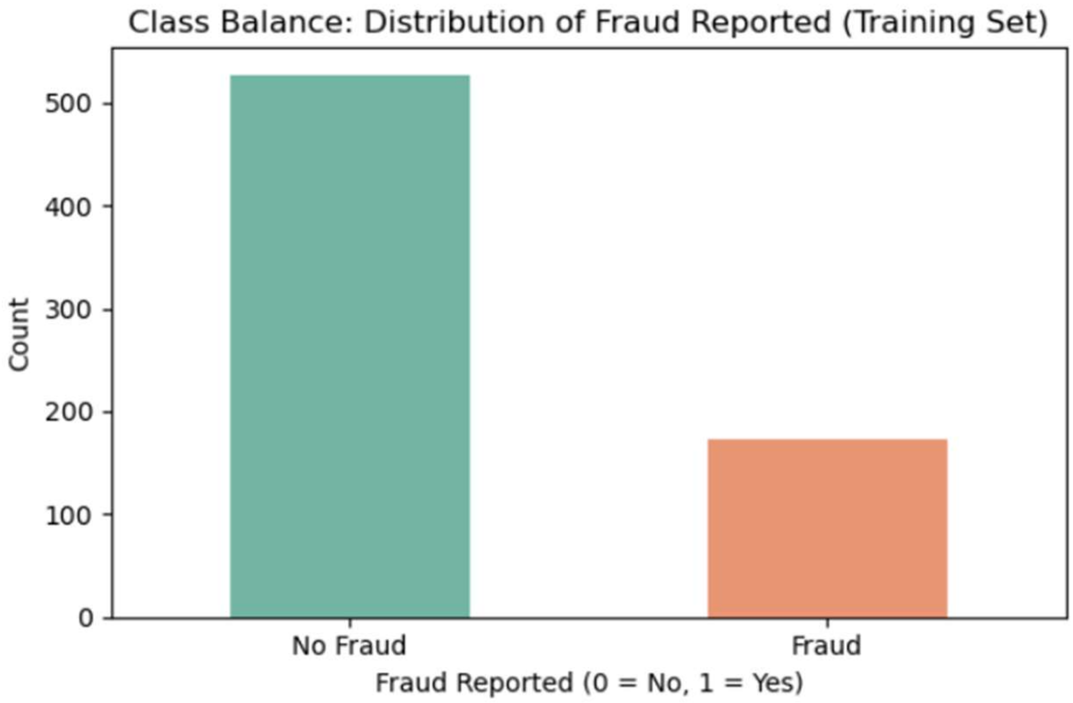
The analysis of fraudulent insurance claims followed a comprehensive data science methodology. The process began with data acquisition and initial setup, followed by thorough exploratory data analysis (EDA) and visualization to understand underlying patterns and relationships. The dataset was then systematically split into training and test sets, with careful attention to data preparation and feature scaling to ensure optimal model performance. Correlation analysis was performed to understand variable relationships, and feature selection was optimized using Recursive Feature Elimination with Cross-Validation (RFECV) to identify the most relevant predictors. The final stages involved building and training the predictive model, followed by rigorous evaluation and validation to assess its effectiveness in identifying fraudulent claims. This structured approach ensured a robust and reliable fraud detection system

ANALYSIS RESULTS

CLASS BALANCE

Used plots to examine the distribution of the target variable to identify potential class imbalances using visualization for better understanding.

Based on the data analysis we can see that the distribution of fraud is around ~35% which is quite high



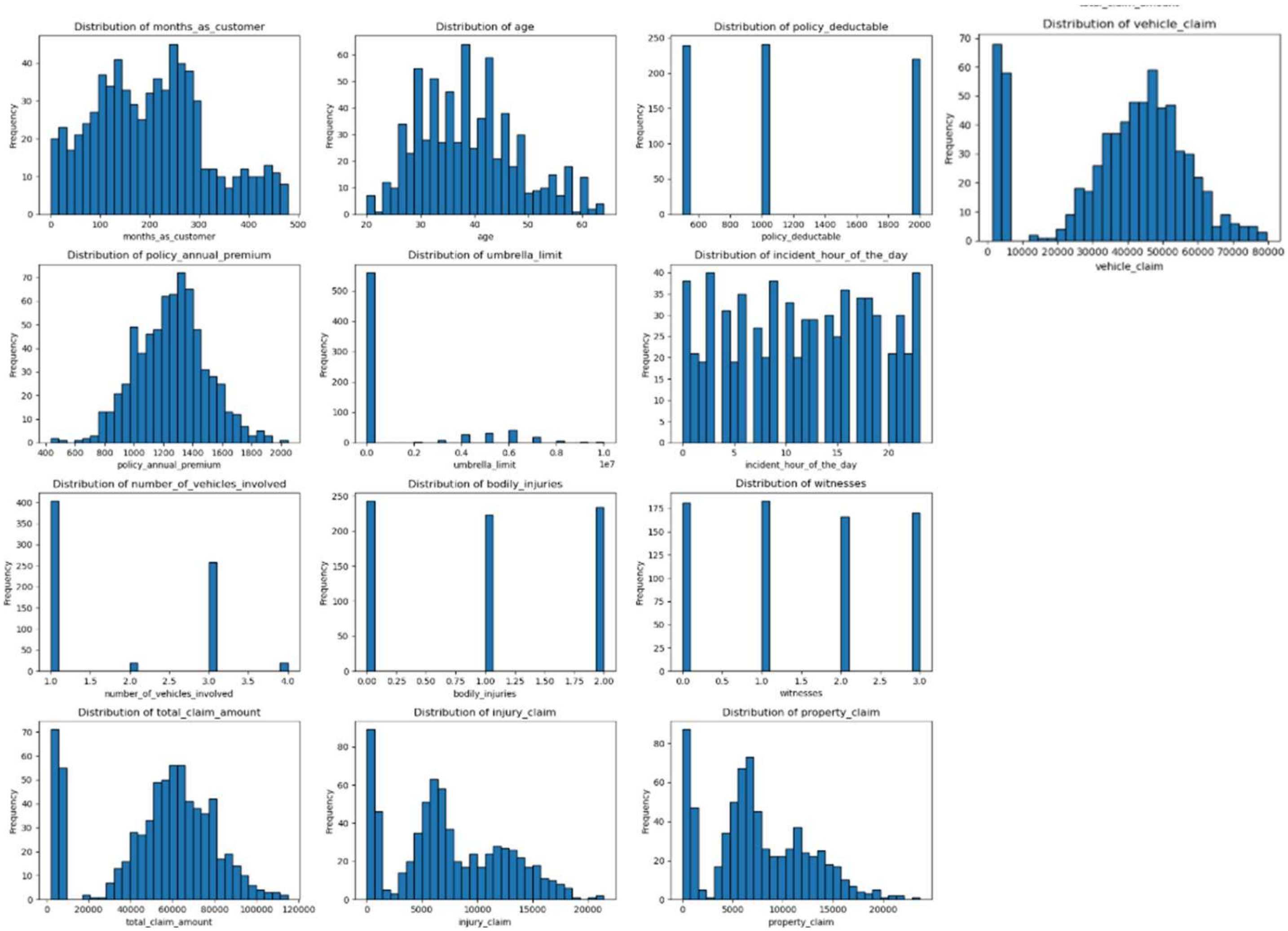
ANALYSIS RESULTS

VISUAL DISTRIBUTION OF NUMERICAL VARIABLES

Used histplot to visualize the distribution of selected numerical variables in dataset.

This analysis gives a visuals of the following key aspects in the dataset:

1. Distribution – Normal or not
2. Where bulk of data is
3. Range of values and Variability
4. Outliers
5. Helps in data cleaning decisions

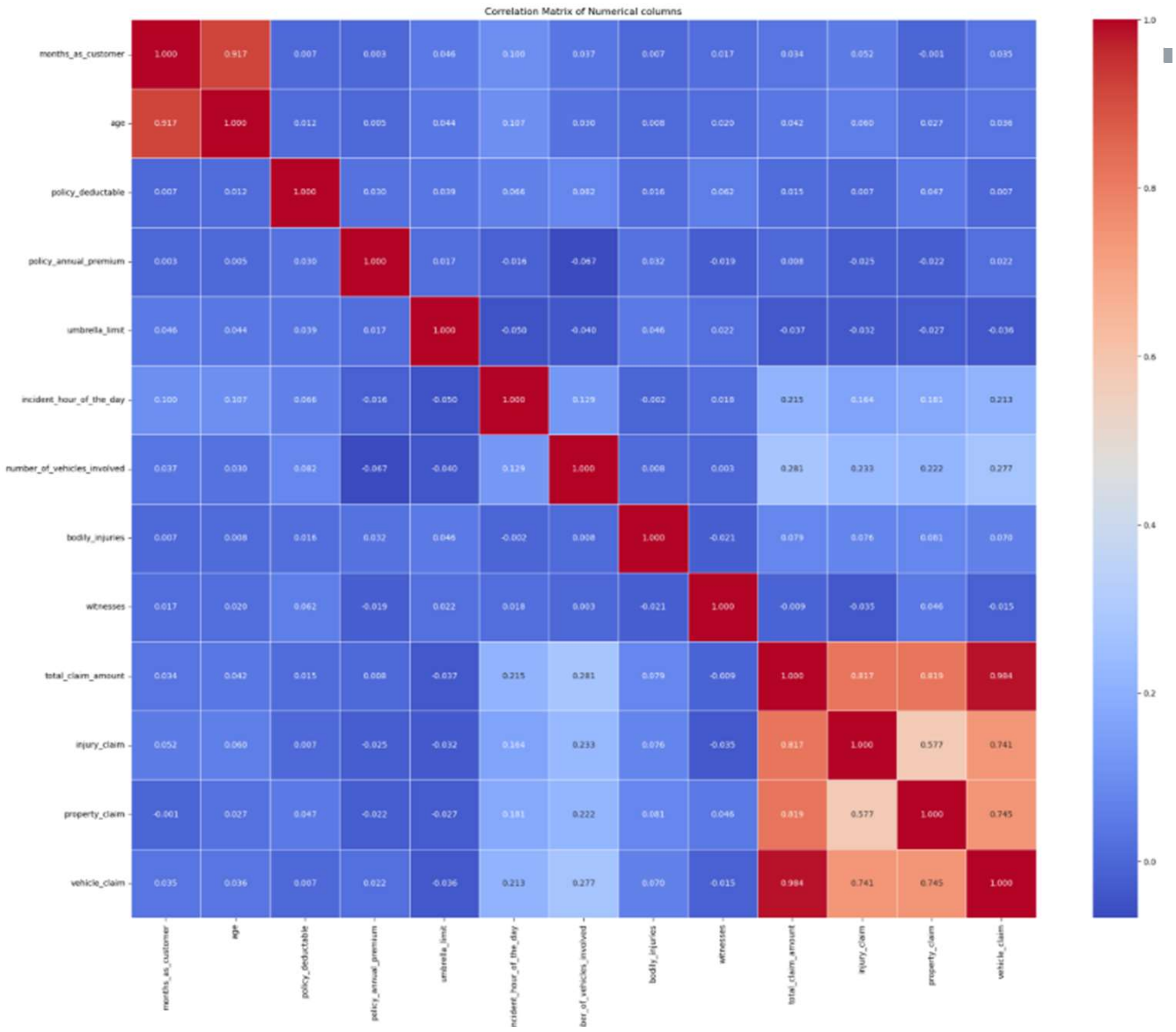


ANALYSIS RESULTS

CORRELATION ANALYSIS

Used heatmap for the correlation matrix. This helps to understand the relation between numeric variables to identify potential multicollinearity or dependencies.

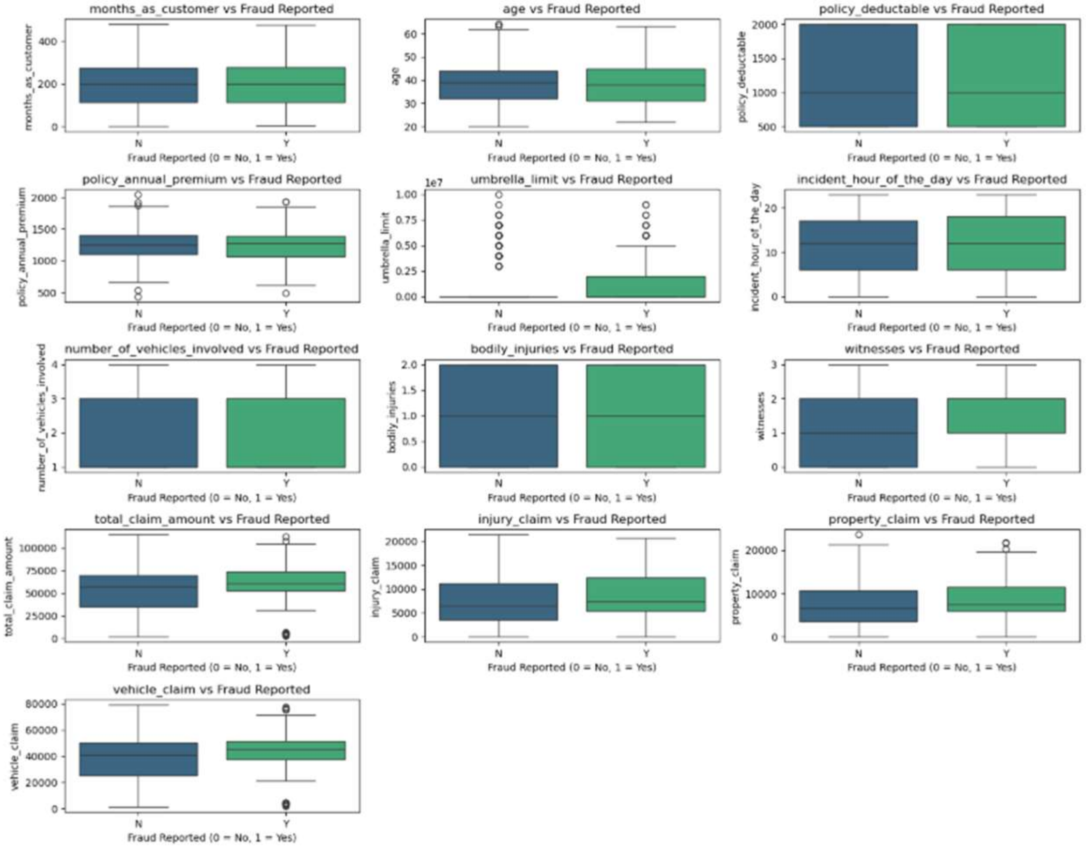
This analysis shows that there is higher positive collinearity of total_claim_amount with injury_claim, property_claim and vehicle_claim.

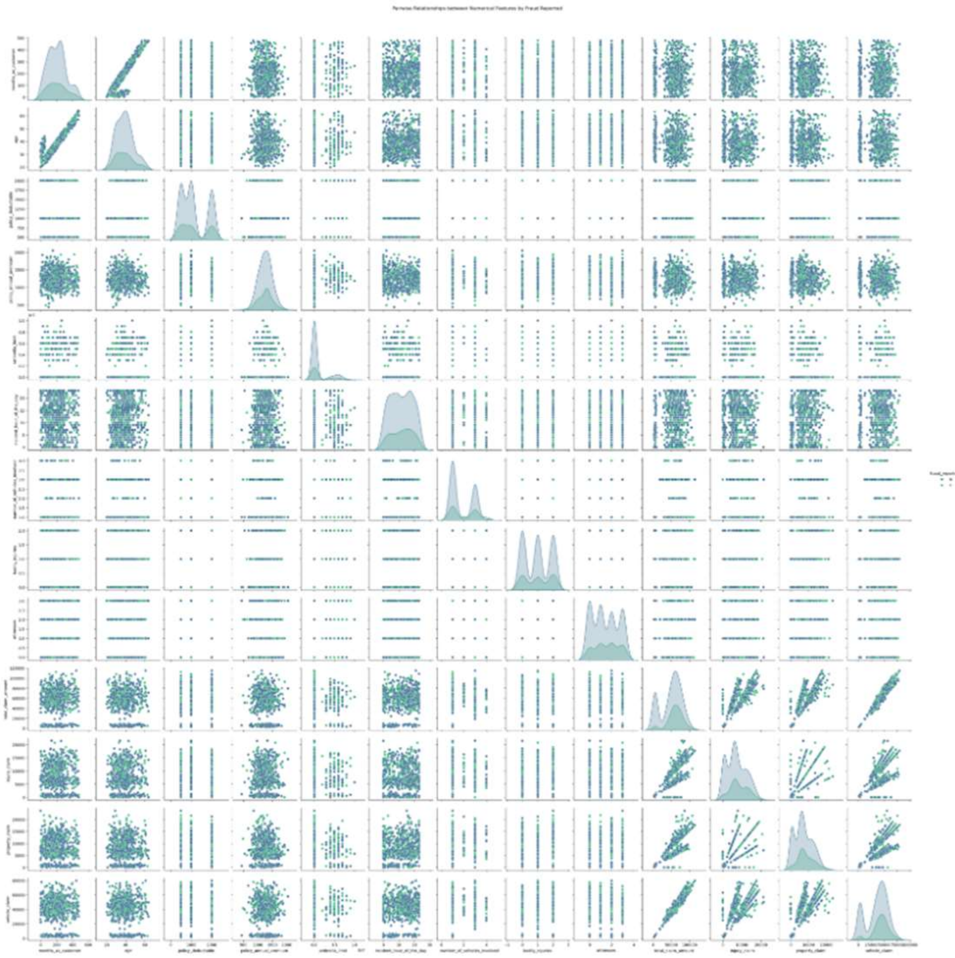

Correlation Matrix of Numerical columns

	months_as_customer	age	policy_deductable	policy_annual_premium	umbrella_limit	ncident_hour_of_the_day	ber_of_vehicles_involved	bodily_injuries	witnesses	total_claim_amount	injury_claim	property_claim	vehicle_claim
months_as_customer	1.000	0.917	0.007	0.003	0.046	0.100	0.037	0.007	0.017	0.034	0.052	-0.001	0.035
age	0.917	1.000	0.012	0.005	0.044	0.107	0.030	0.008	0.020	0.042	0.060	0.027	0.036
policy_deductable	0.007	0.012	1.000	0.030	0.039	0.066	0.082	0.016	0.062	0.015	0.007	0.047	0.007
policy_annual_premium	0.003	0.005	0.030	1.000	0.017	-0.016	-0.067	0.032	-0.019	0.008	-0.025	-0.022	0.022
umbrella_limit	0.046	0.044	0.039	0.017	1.000	-0.050	-0.040	0.046	0.022	-0.037	-0.032	-0.027	-0.036
incident_hour_of_the_day	0.100	0.107	0.066	-0.016	-0.050	1.000	0.129	-0.002	0.018	0.215	0.164	0.181	0.213
number_of_vehicles_involved	0.037	0.030	0.082	-0.067	-0.040	0.129	1.000	0.008	0.003	0.281	0.233	0.222	0.277
bodily_injuries	0.007	0.008	0.016	0.032	0.046	-0.002	0.008	1.000	-0.021	0.079	0.076	0.081	0.070
witnesses	0.017	0.020	0.062	-0.019	0.022	0.018	0.003	-0.021	1.000	-0.009	-0.035	0.046	-0.015
total_claim_amount	0.034	0.042	0.015	0.008	-0.037	0.215	0.281	0.079	-0.009	1.000	0.817	0.819	0.984
injury_claim	0.052	0.060	0.007	-0.025	-0.032	0.164	0.233	0.076	-0.035	0.817	1.000	0.577	0.741
property_claim	-0.001	0.027	0.047	-0.022	-0.027	0.181	0.222	0.081	0.046	0.819	0.577	1.000	0.745
vehicle_claim	0.035	0.036	0.007	0.022	-0.036	0.213	0.277	0.070	-0.015	0.984	0.741	0.745	1.000

ANALYSIS RESULTS

BI-VARIATE ANALYSIS





SUMMARY OF ANALYSIS

- Class Imbalance identified in the distribution analysis of target variable ‘Fraud reported’.In this case, with ~65% No (non-fraud) and ~35% Yes (fraud), this class imbalance can impact model performance in the following ways:
  - Model may be biased towards predicting "No" (majority class)
  - Risk of overlooking fraudulent cases
  - May underperform in detecting actual fraud cases
  - High overall accuracy can be mis-leading.
  - May have business impact resulting in missing fraud cases and hence increased financial risk.
- Derived the following insight from data analysis
  - Incident_severity and total_claim_amount highly predictive of the fraud.
  - Features like policy_annual_premium, days_between_policy_and_incident, and incident_hour_of_the_day also show strong importance in the Random Forest model.
- Model performance is as follows
  - Logistic Regression and Random Forest showed decent performance.
  - After hyperparameter tuning, Random Forest improved Recall from 0.27 to 0.55 and F1 Score from 0.30 to 0.56, balancing between capturing frauds and avoiding false alarms

SUMMARY OF ANALYSIS

- Optimal cutoff
  - Fixed threshold (0.5) was suboptimal.
  - Precision-recall tradeoff analysis suggests lowering threshold may be beneficial for better recall (detecting more frauds).

ANSWERING THE SUMMARY QUESTIONS

1. How can we analyze historical claim data to detect patterns that indicate fraudulent claims?

It is quite critical to get the right set of data or more specifically right set of variable which can help to derive insights into the potential fraudulent claims. In this case study, we followed the below steps to make sure the model meets the requirements of the underlying business case.

- Data collection and Preparation: We cleaned the data by removing the least important variables (low-variance, which may not help in prediction), redundant columns, empty columns etc.
- Exploratory Data Analysis: Analyzed the distribution of numeric variables. Analyzed the distribution of fraud cased in the data. There was a class imbalance, hence we used the sampling technique (*RandomOverSampler*) to address the class imbalance.
- Feature Engineering: We created relevant features from raw-date such as *days_between_policy_and_incident* and *age_group*, to get relevant information pertinent to fraud possibility.
- Categorical Grouping: Key categorical variables in the dataset were grouped to improve model predictability and accuracy.
- Statistical Analysis: performed statistical tests such as Univariate and Bi-variate analysis to understand the distribution and measure relationship strength respectively.
- Model Development: Built model by splitting the dataset into 70% train and 30% validation sets.
- Model Evaluation: Used machine learning models line Logistic regression and Random Forest to assess model accuracy, review confusion matrix and detect non-linear patterns in dataset.

ANSWERING THE SUMMARY QUESTIONS

2. Which features are the most predictive of fraudulent behavior?

Based on the analysis using techniques such as RFECV and Random Forest Importance scores. The following features found to be higher predictive than others:

- incident_severity
- total_claim_amount
- policy_annual_premium
- days_between_policy_and_incident

These features either capture the scale or suspicious timing of an incident, or represent contextual indicators that correlate with fraud.

ANSWERING THE SUMMARY QUESTIONS

3. Based on past data, can we predict the likelihood of fraud for an incoming claim?

Below are the performance metrics:

- ~67% Sensitivity: Catches 2/3 of actual fraud cases
- ~62% Precision: 62% of fraud predictions are correct
- ~64% F1 Score: Good balance of precision and recall
- ~87% Specificity: Strong in identifying legitimate claims

Based on these, Yes, we can predict the likelihood of fraud for incoming claims based on these results. The model shows a good balanced performance with reasonable capacity to detect fraud. However, considering 1 out of 3 actual fraud cases potentially can be missed to be predicted correctly, there can be business impact. So the better implementation strategy will be to use this model as part of a broader fraud detection system combined with expert reviews, establishing clear threshold policies and regular model monitoring and updates to model.

ANSWERING THE SUMMARY QUESTIONS

4. What insights can be drawn from the model that can help in improving the fraud detection process?

Here are the key insights that can be drawn from the model to improve the fraud detection process:

- Model performance is relatively good with ~87% accuracy in identifying legitimate claims
- Top predictors like *incident_severity* and *total_claim_amount* can be indicating factors for triggering manual reviews.
- Implement strategic actions such as update of fraud detection policies and improving data collection methods.
- Threshold tuning can allow better control over false positives and false negatives.
- Model should be monitored regularly to understand the changes in pattern and accordingly update the model.
- Train staff on fraud indicators, share identified patterns and improve documentation practices.

These insights can help create a more effective and efficient fraud detection system while optimizing resource utilization.

BUSINESS IMPACT

- Fraud Loss Reduction: Early detection of fraud helps in preventing financial losses, especially by flagging high-value suspicious claims (total_claim_amount).
- Operational Efficiency: Automating the detection using models will help reduce the burden on human investigators by prioritizing high-risk claims.
- Better Resource Allocation: High-risk cases can be fast-tracked to special investigation teams, while low-risk claims can be approved faster—improving customer satisfaction.
- Regulatory Compliance: Having a well-documented and explainable model helps demonstrate due diligence to insurance regulators and audit teams

RECOMMENDATIONS

- During model building:
  - Address class imbalance using re-sampling techniques.
  - Feature engineering should be used to select relevant features for the model.
  - Focus on precision, F1-scores and monitor false positives and negatives (Hyper-parameter tuning). Also use ROC-curves for evaluation metrics.
  - Threshold tuning can allow better control over false positives and false negatives.
  - Optimize the model trying different algorithms and ensemble methods.
  - Regular re-training of the model for model optimization.

- Operational recommendations to the insurance company:
  - The analysis was done with the top predictors but the pattern can change over time. Hence, the patterns should monitored and required updates should be done to the model to get accurate predictions.
  - Due to class imbalance, there is still chance of false positives or false negatives. Hence, due audit processes need to be added to avoid any financial risk due to false prediction.
  - Automated alerts for threshold breach and manual review process for high-risk cases.
  - Strategic steps should be taken by the organization to update fraud detection policies and improving data collection methods.
  - Train staff on fraud indicators, share identified patterns and improve documentation practices.

THANK YOU