Fraudulent Claim Detection Case Study

Advanced ML Case Study, Data Science C71

Group Members

Parul Chopra
Pavithra Sri S
Sushil Santoshrao Muli

1. Problem Statement

The objective is to build a model to classify insurance claims as either fraudulent or legitimate using historical data. By analyzing features like claim amounts, customer profiles, claim types, and approval times, the goal is to identify potentially fraudulent claims before they are approved.

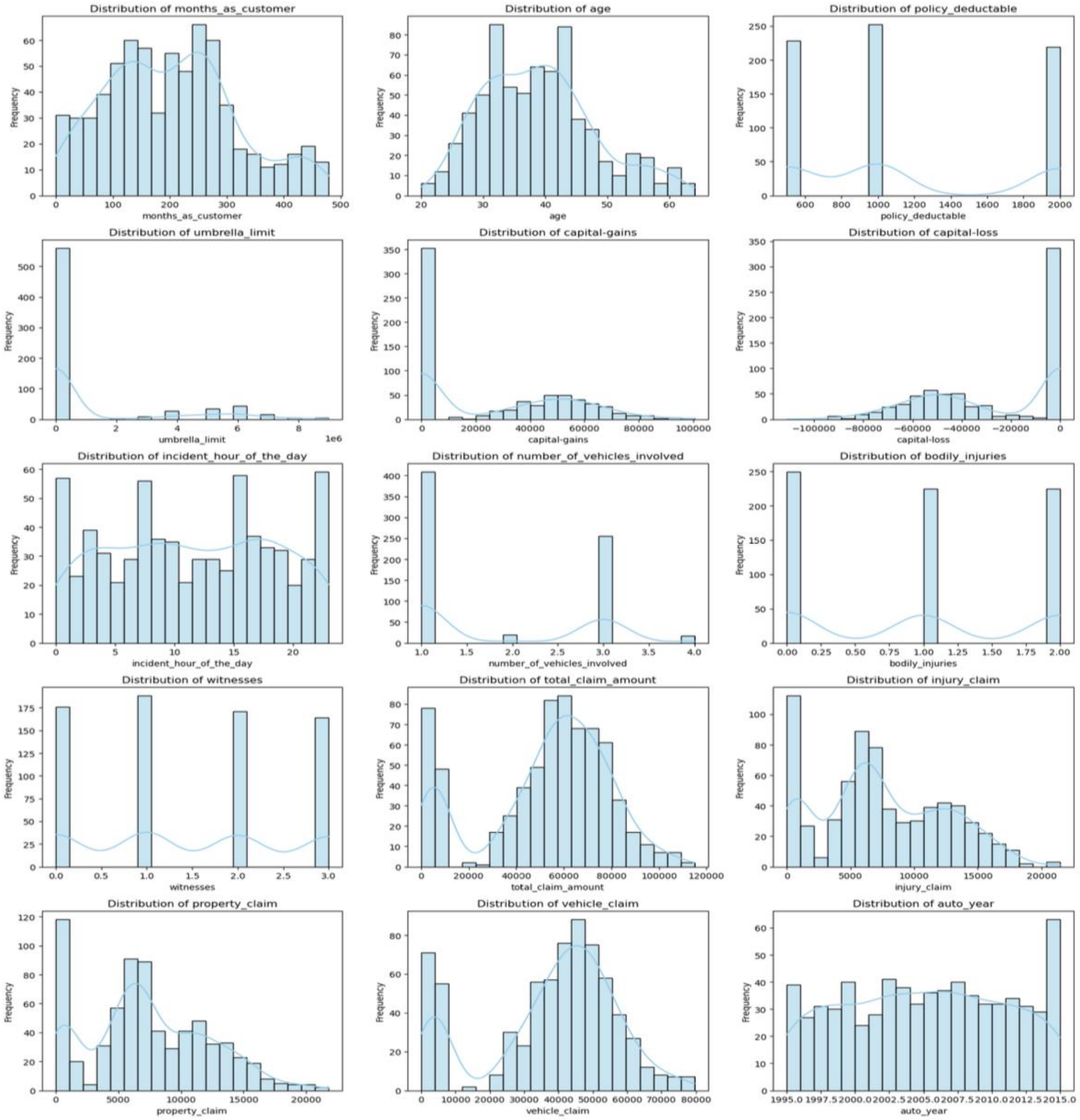
2. Goals

1. Enhance Global Insure's fraud detection capabilities by leveraging historical data.
2. Identify key indicators that differentiate fraudulent claims.
3. Predict the likelihood of fraud in new claims to enable early intervention and reduce financial loss.

3. Methodology

- Imported libraries and loaded dataset.
- Cleaned data: handled missing values, dropped redundant or ID columns, corrected datatypes.
- Converted date fields and created new date-based features.
- Target variable: `fraud_reported`.
- Performed train-validation split (70:30).
- Conducted EDA: univariate and bivariate analysis.

EDA

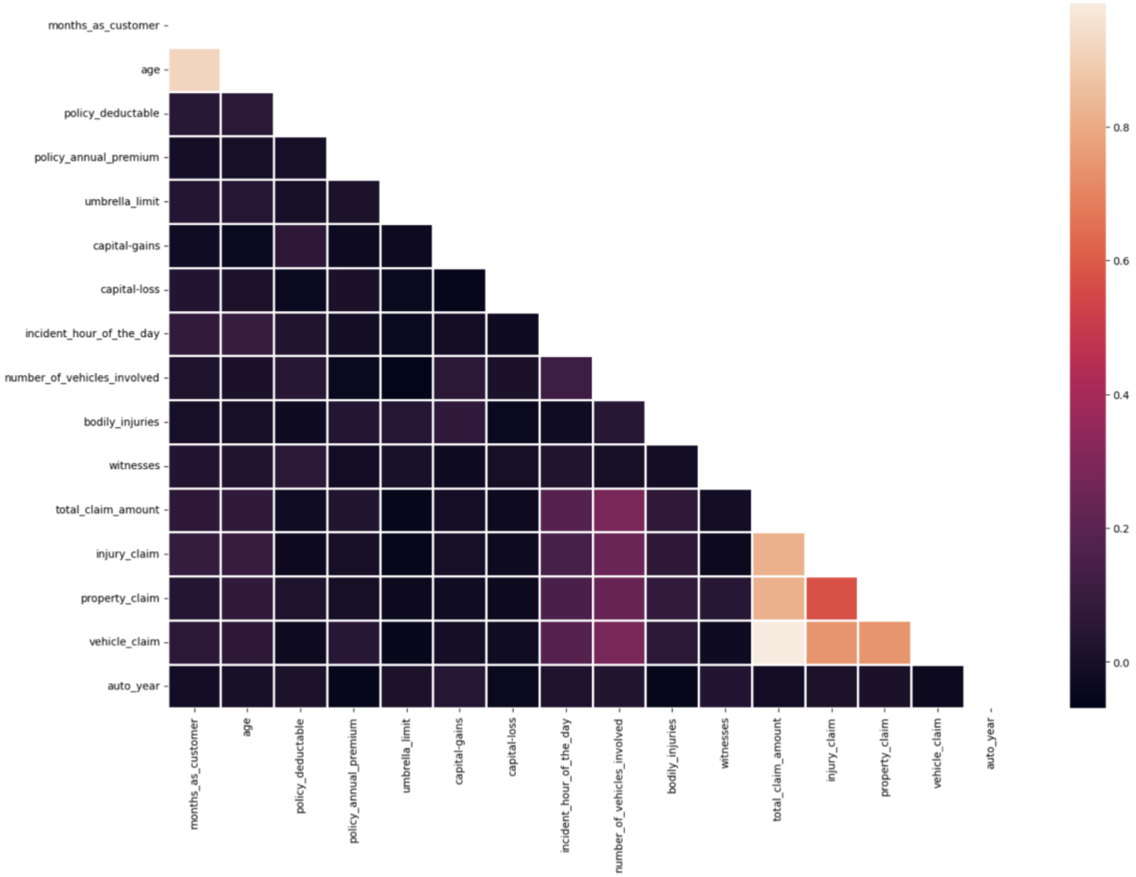


- months_as_customer: Slightly right-skewed, with most customers in the 100–300 month range.
- age: Bell-shaped distribution, centered around 35–45 years. Most customers are aged between 30 and 50 years.
- policy_deductable: Discrete values with strong peaks at 500, 1000, and 2000—likely predefined policy options.
- policy_annual_premium: Normally distributed, centered around 1200–1400.
- number_of_vehicles_involved: Strong peak at 1 vehicle, indicating most incidents are single-vehicle.
- bodily_injuries: Typically ranging between 0 to 2 injuries per incident.
- witnesses: Common values being 0 and 3 witnesses.
- property_claim: Right-skewed distribution with most values clustered below 10,000. Few high-value claims exist.
- vehicle_claim: Appears approximately normally distributed, centered around 40,000–50,000.

Correlation analysis:

Inference: We observed a high correlation between age and months_as_customer, indicating redundancy. Therefore, age column has to be dropped to avoid multicollinearity.

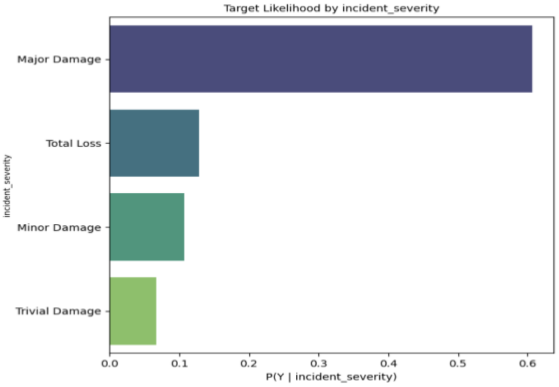
Additionally, there is a strong correlation among total_claim_amount, injury_claim, property_claim, and vehicle_claim, which is expected since total_claim_amount is the sum of the individual claim components.

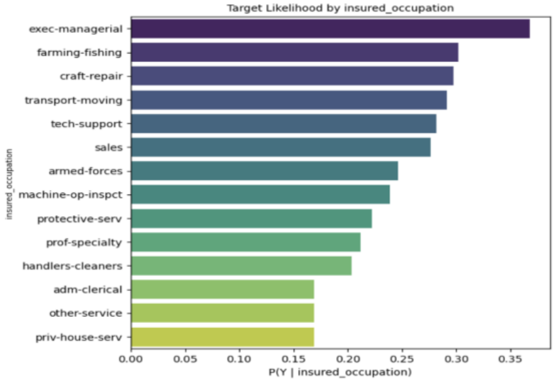


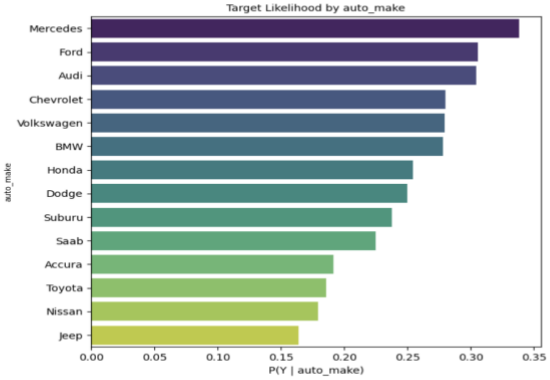
Target Likelihoods by various features:

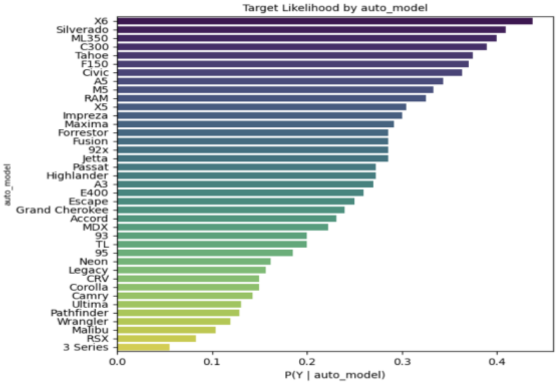
Features that contribute more to the prediction

- **incident_severity:** Huge variation (e.g., Major Damage: 60.7% vs Trivial Damage: 6.7%) → very predictive.
- **insured_occupation:** Fraud likelihood varies significantly (e.g., exec-managerial: 36.8%, other-service: 16.9%).
- **auto_make and auto_model:** Certain models/makes are fraud-prone (e.g., Mercedes X6: 43.7%, 3 Series: 5.5%)



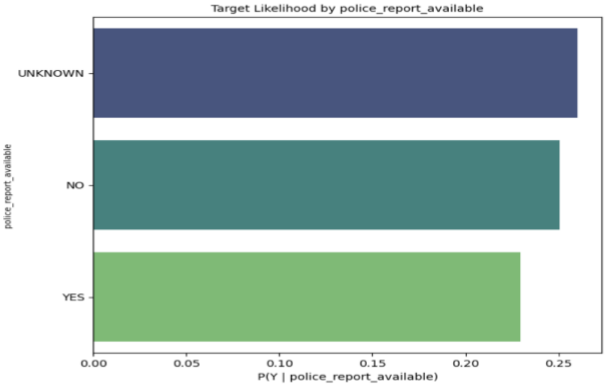


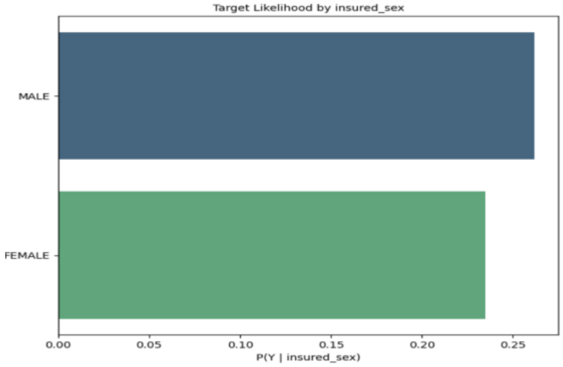


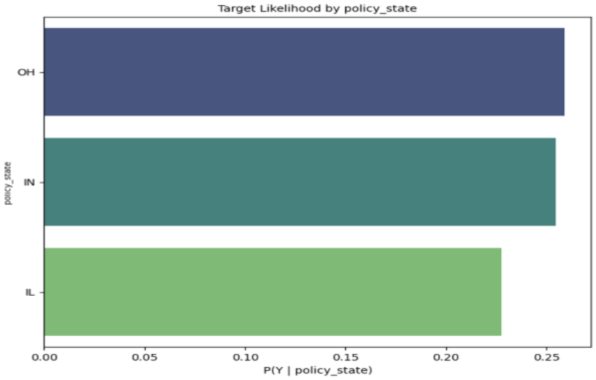


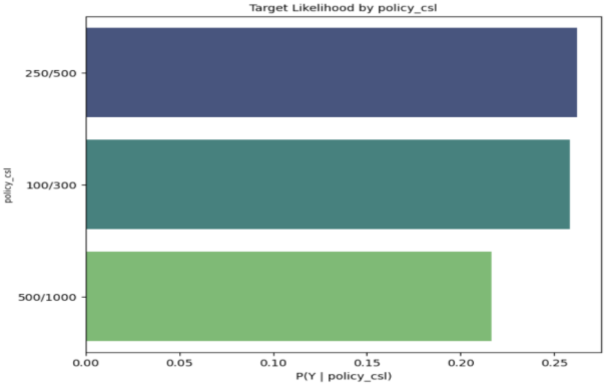
Features with less importance: These may not significantly contribute to the model's performance and can be considered for removal.

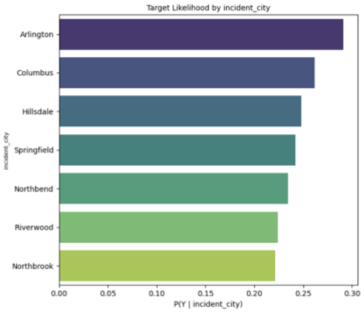
- **policy_state:** Very similar fraud rates across OH, IN, IL (22.8%–25.9%).
- **policy_csl:** Minor variation (21.6%–26.2%).
- **insured_sex:** Small difference between Male (26.1%) and Female (23.4%).
- **police_report_available:** All values are ~23%–26%.

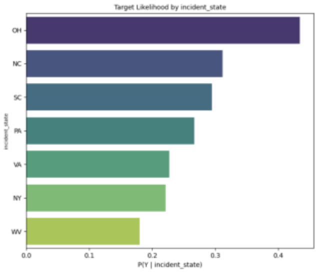


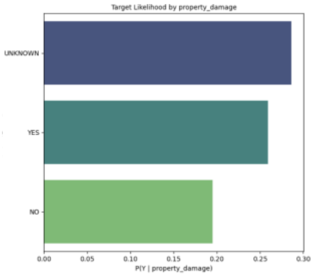


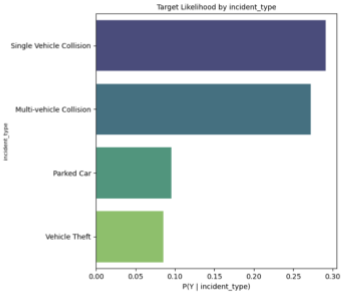


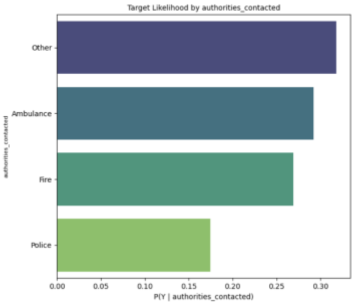


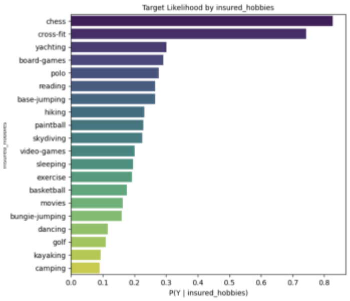


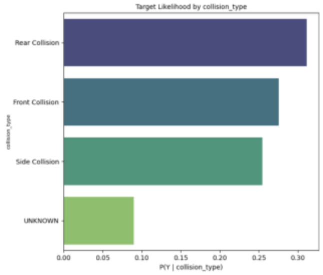


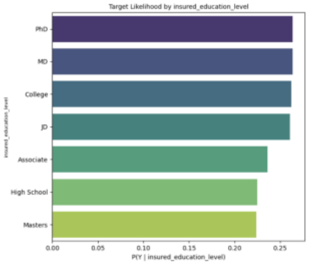


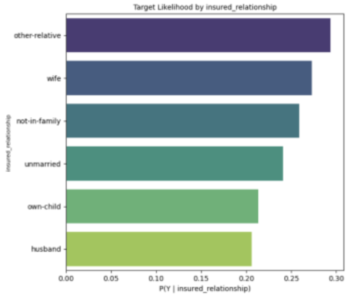






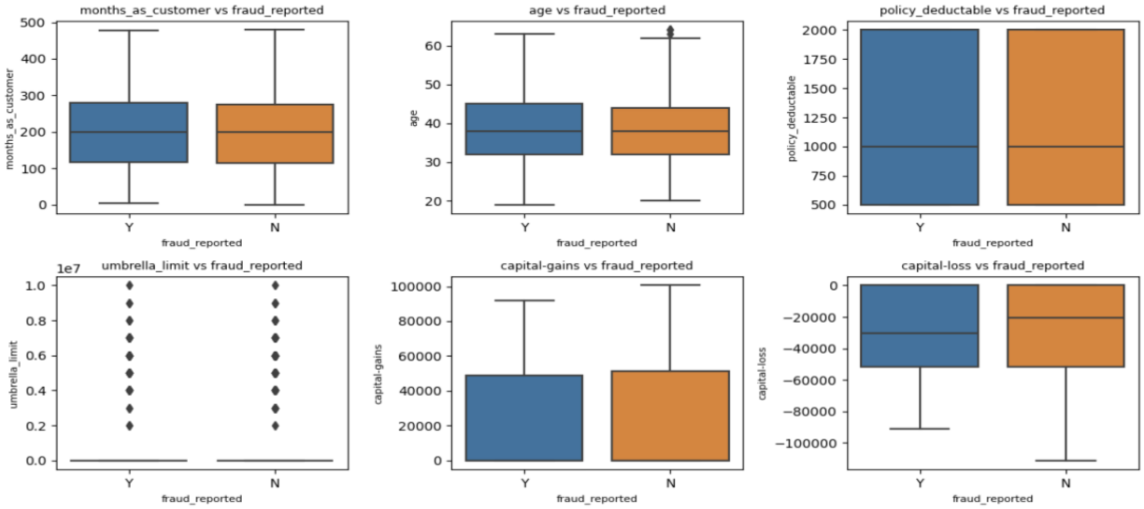


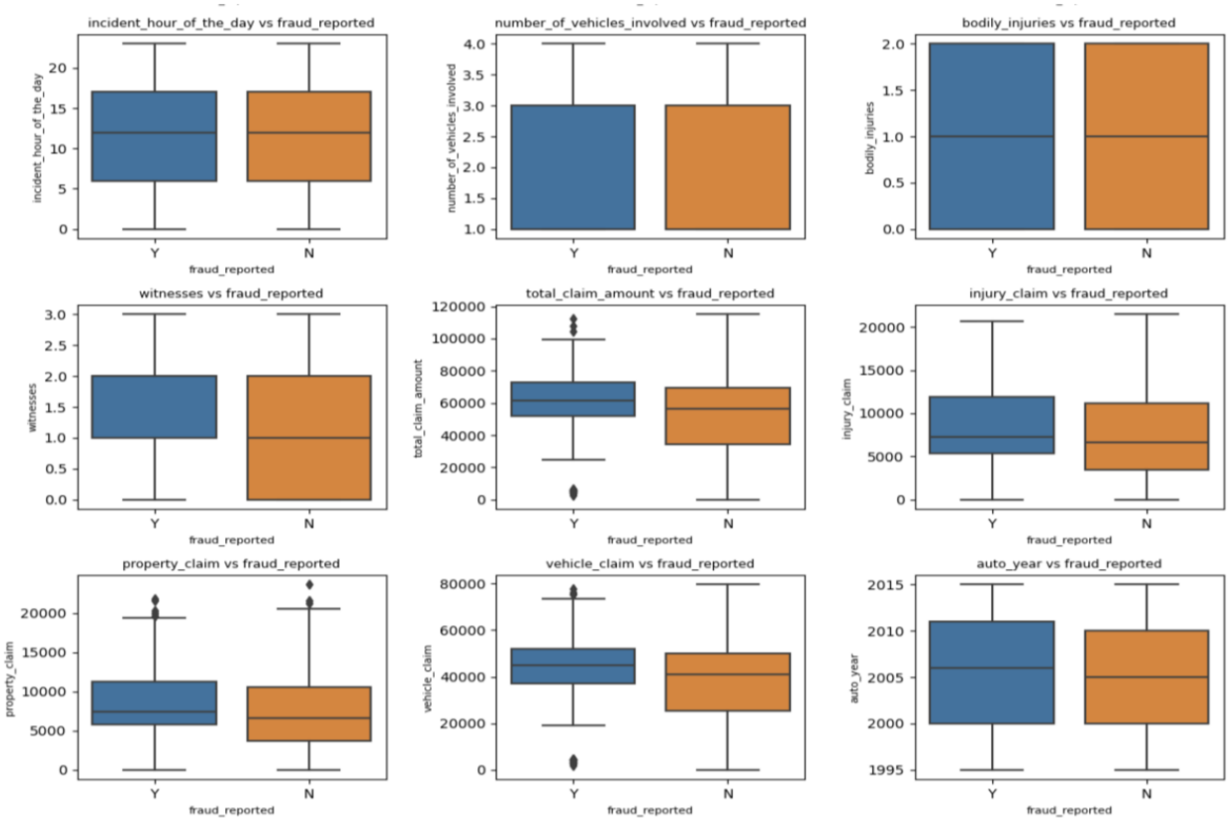


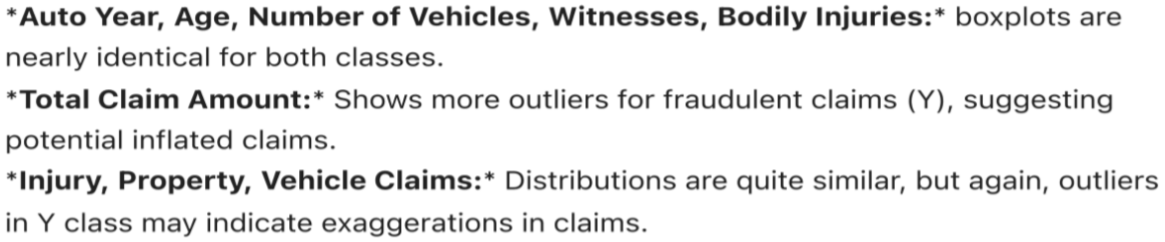


Other features

Relation between numerical variables and Fraud Reported

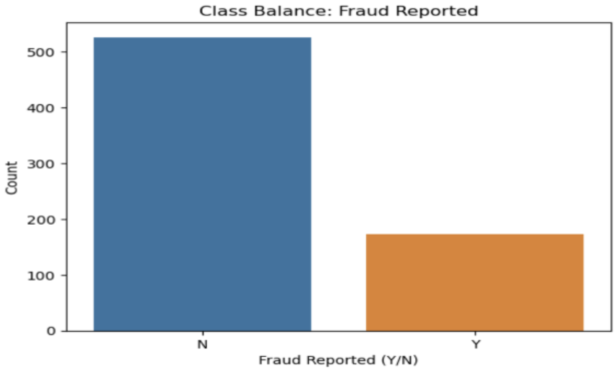






Auto Year, Age, Number of Vehicles, Witnesses, Bodily Injuries: boxplots are nearly identical for both classes.
Total Claim Amount: Shows more outliers for fraudulent claims (Y), suggesting potential inflated claims.
Injury, Property, Vehicle Claims: Distributions are quite similar, but again, outliers in Y class may indicate exaggerations in claims.

- Handled class imbalance using Random OverSampler. (75% non-fraud, 25% fraud initially).



Inference: The graph above clearly indicates a significant imbalance between reported and non-reported fraud cases. This imbalance needs to be addressed to ensure reliable model performance.

- Created new features:
  - Month and year extraction from date columns for better analysis.
  - Binning of the column incident_hour_of_the_day to get useful insight.
  - Log transformation on the column Umbrella_limit as it is highly skewed.
  - Binary mapping on the column insured_sex.
- Applied one-hot encoding and standardized numerical features.
- Built Logistic Regression and Random Forest models.
- Feature selection using RFECV and VIF to remove multicollinearity.
- Evaluated models using various metrics.

4. Techniques Used

- Missing Value Imputation: Replaced nulls and '?' with "UNKNOWN".
- Feature Elimination: Removed high-cardinality or redundant columns.
- Encoding: One-hot encoding of categorical features.
- Scaling: StandardScaler for numerical data.
- Multicollinearity Check: VIF.
- Feature Selection: RFECV with Logistic Regression.
- Modeling: Logistic Regression and Random Forest.
- Cutoff Tuning**:** Used Precision-Recall curves to set optimal thresholds.
- Evaluation**:** Accuracy, Precision, Recall, F1-score, AUC.
- Visualization**:** Boxplots, barplots, PR curves, confusion matrices.

5. Model Performance Comparison

Metric	**Logistic Regression (Train)**	**Logistic Regression (Validation)**	**Random Forest (Train)**	**Random Forest (Validation)**
Accuracy	0.80	0.72	1.00	0.80
Recall (Sensitivity)	0.72	0.59	1.00	0.65
Precision	0.62	0.45	1.00	0.59
Specificity	0.85	0.76	1.00	0.85
F1 Score	0.66	0.51	1.00	0.62

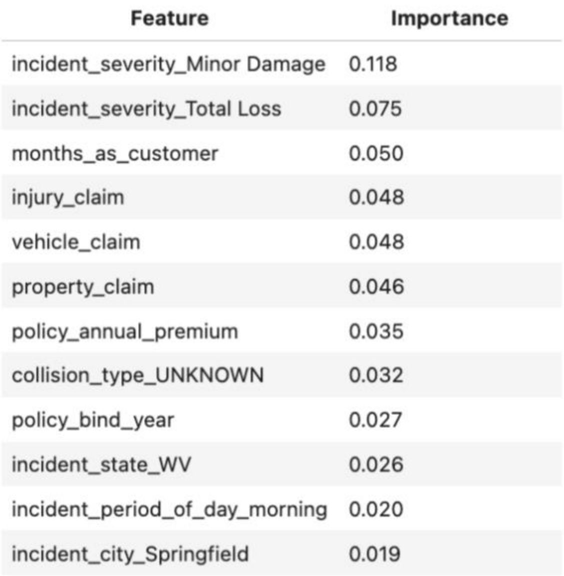
Insights

- Random Forest (Tuned) performs best on validation data, offering a balanced F1 score and higher recall.
- Logistic Regression shows signs of overfitting (train vs validation gap).
- Random Forest without tuning shows perfect train performance → overfitting, but hyperparameter tuning fixed this.
- Tuned Random Forest clearly outperforms Logistic Regression across all performance metrics.

- Recall and F1 Score are especially crucial for fraud detection. A high recall ensures that more fraudulent claims are identified early. The F1 Score balances the trade-off between detecting fraud and minimizing false positives.

6. Top Features by Random Forest

- Claim-related variables (like `injury_claim`, `vehicle_claim`, `incident_severity`) are strong indicators of fraud.
- Customer behavior and policy attributes (e.g., `months_as_customer`, `policy_premium`) also significantly contribute.
- These high-importance features can be leveraged to develop targeted fraud detection strategies for early and efficient intervention.



Feature	Importance
incident_severity_Minor Damage	0.118
incident_severity_Total Loss	0.075
months_as_customer	0.050
injury_claim	0.048
vehicle_claim	0.048
property_claim	0.046
policy_annual_premium	0.035
collision_type_UNKNOWN	0.032
policy_bind_year	0.027
incident_state_WV	0.026
incident_period_of_day_morning	0.020
incident_city_Springfield	0.019

7. Recommendation

Recommended Model: Random Forest (with hyperparameter tuning)

    - Strong generalization to new data
    - Balanced performance: high recall and precision
    - Better fraud detection capability with fewer false negatives
- Business Impact:
    - Reduces financial loss
    - Automates fraud scoring in claim pipeline

8. Business Impact

1. Enables early and proactive detection of fraudulent insurance claims.
2. Helps in minimizing financial losses by preventing payouts on suspicious claims.

3. Facilitates efficient, automated, and data-driven claim triaging, reducing manual workload. Early flagging for investigation
4. Empowers the organization to focus investigations on high-risk cases, using the most predictive features identified by the model.

9. Deployment Suggestions: These may be done further to increase to make it more impactful

1. Risk Scoring System:
   - Use RF model to assign a probability score (0–1) to each claim.
2. Threshold-based Triage for dearly identification of fraudulent claims
   - Above 0.75: Immediate investigation
   - 0.44–0.75: Manual review queue
   - Below 0.44: Normal processing
3. Alert Mechanism can be integrated for high-risk claims
4. Model Re-training every month, retraining with new fraud labels

9. Conclusion

A robust machine learning pipeline was created for insurance fraud detection. Through careful feature engineering, balanced data preparation, and comparative modeling, the Random Forest model with hyperparameter tuning emerged as the most suitable solution. With strategic deployment and threshold tuning, this model can drive impactful real-time fraud detection for Global Insure.