Insurance Fraud Detection using Machine Learning

PREDICTIVE MODELING FOR RISK MITIGATION

Sakshi Gupta | Data Science & AI Professional

The Challenge of Insurance Fraud

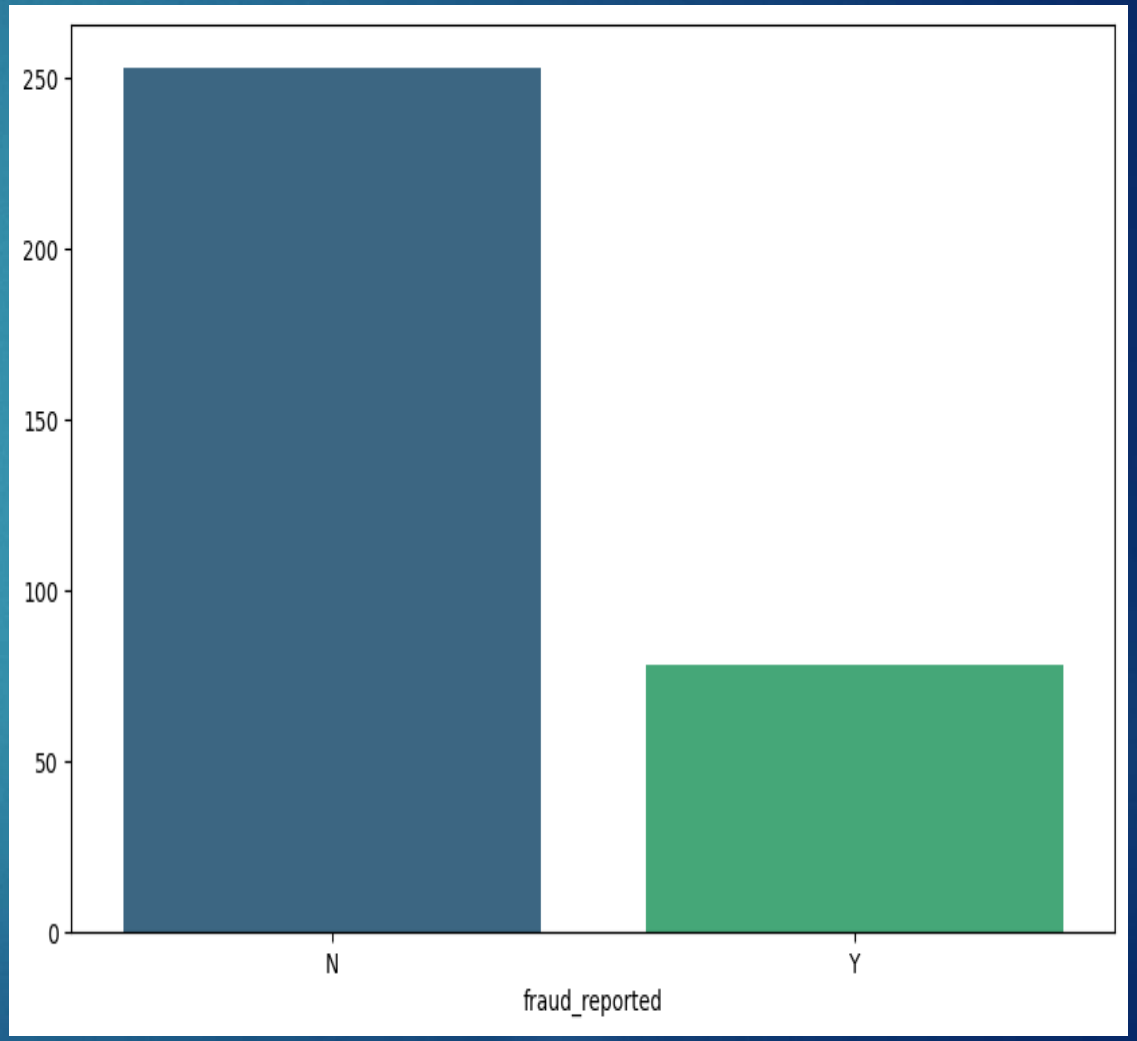
- Fraudulent claims lead to massive annual financial losses.
- Manual verification is slow, expensive, and prone to human error.

Objective:

- To build an automated classification model that flags high-risk claims with high sensitivity.

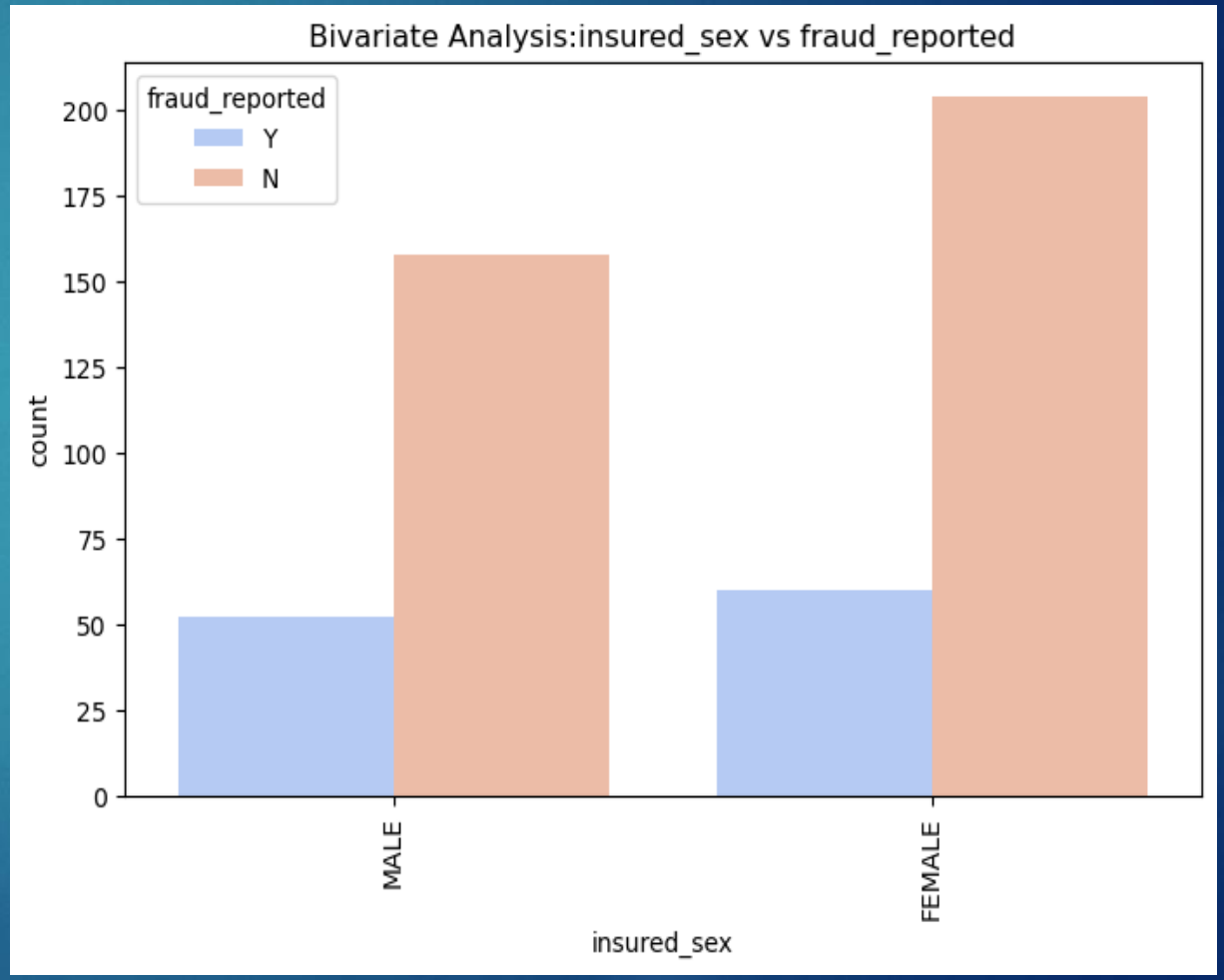
Class Imbalance: Understanding the Target Variable

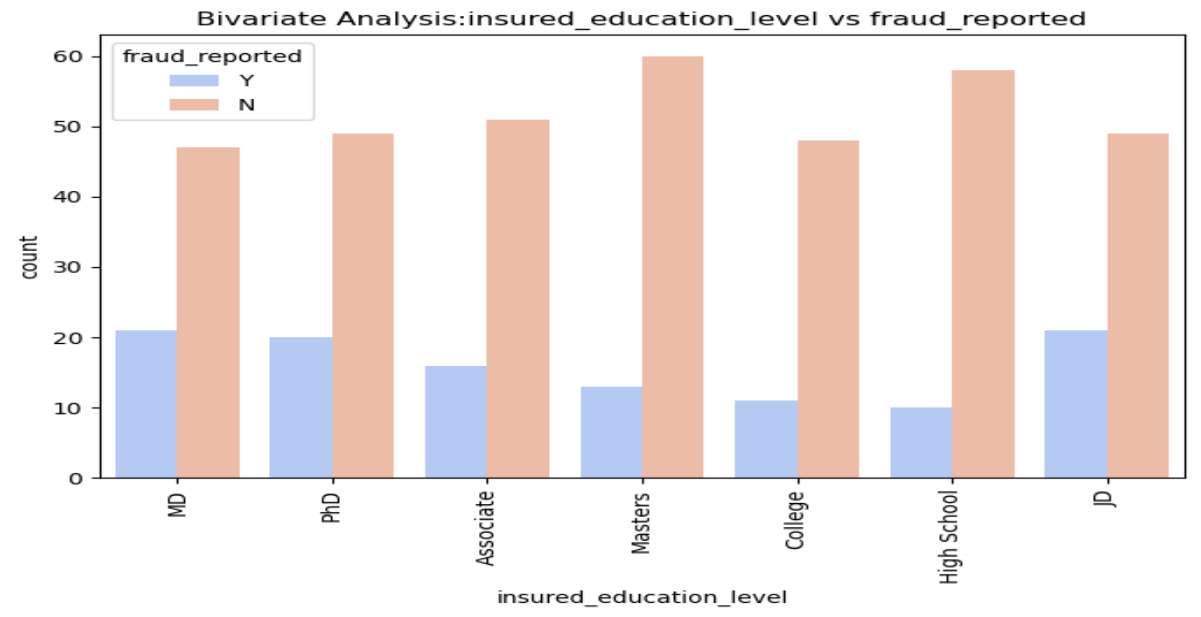
- **Imbalanced Dataset:** Non-fraudulent claims ('N') significantly outnumber fraudulent ones ('Y').
- **The Challenge:** A standard model might achieve high accuracy by simply predicting 'N' every time, but it would fail to catch the actual fraud.
- **Analytical Strategy:** This imbalance necessitates the use of specialized techniques like **Synthetic Minority Over-sampling Technique (SMOTE)** or adjusted class weights during model training.



Analysis of insured sex, education level, and occupation against fraud reports.

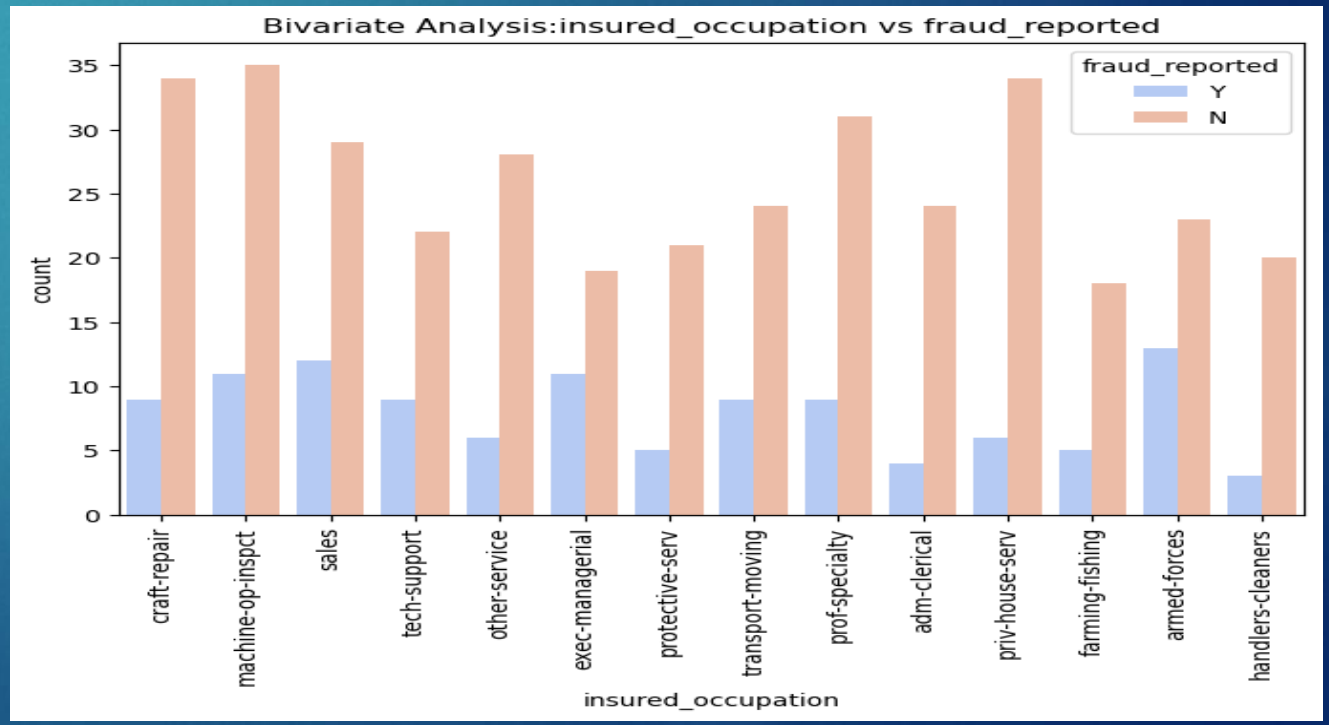
Analysis indicates that fraudulent claims are distributed nearly equally between genders, though **Females** show a slightly higher count of reported fraud in this specific population.





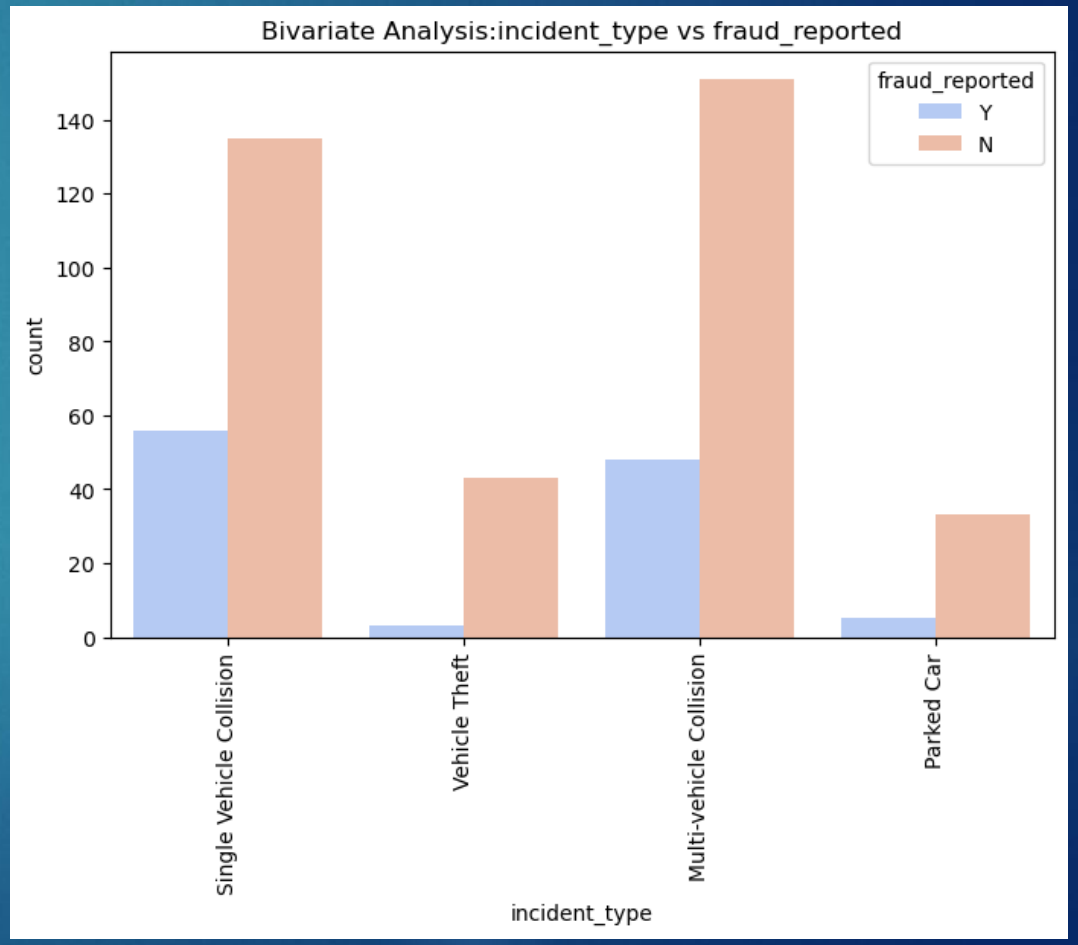
Fraudulent activity is observed across all education levels. Interestingly, **JD and MD** holders show a significant number of fraud reports, suggesting that higher education does not correlate with lower fraud risk.

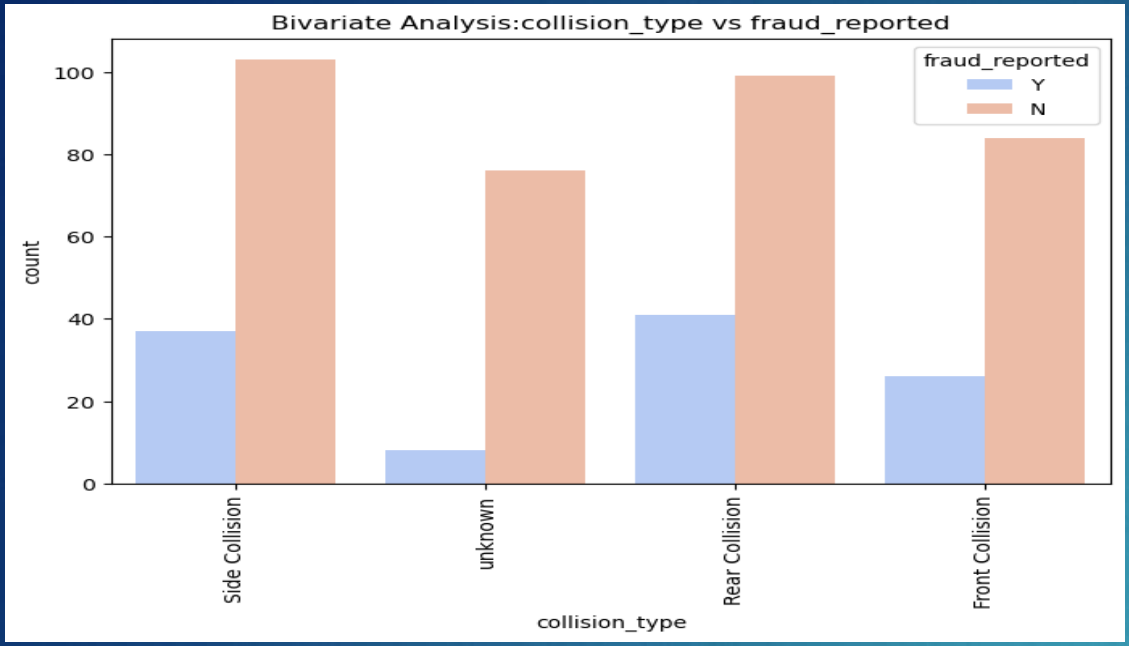
High-risk profiles are more frequent among **Craft-repair, Exec-managerial, and Sales** occupations. These insights allow for better customer segmentation and risk-based pricing.



Visualizing how incident_type, collision_type, and authorities_contacted correlate with fraudulent activity.

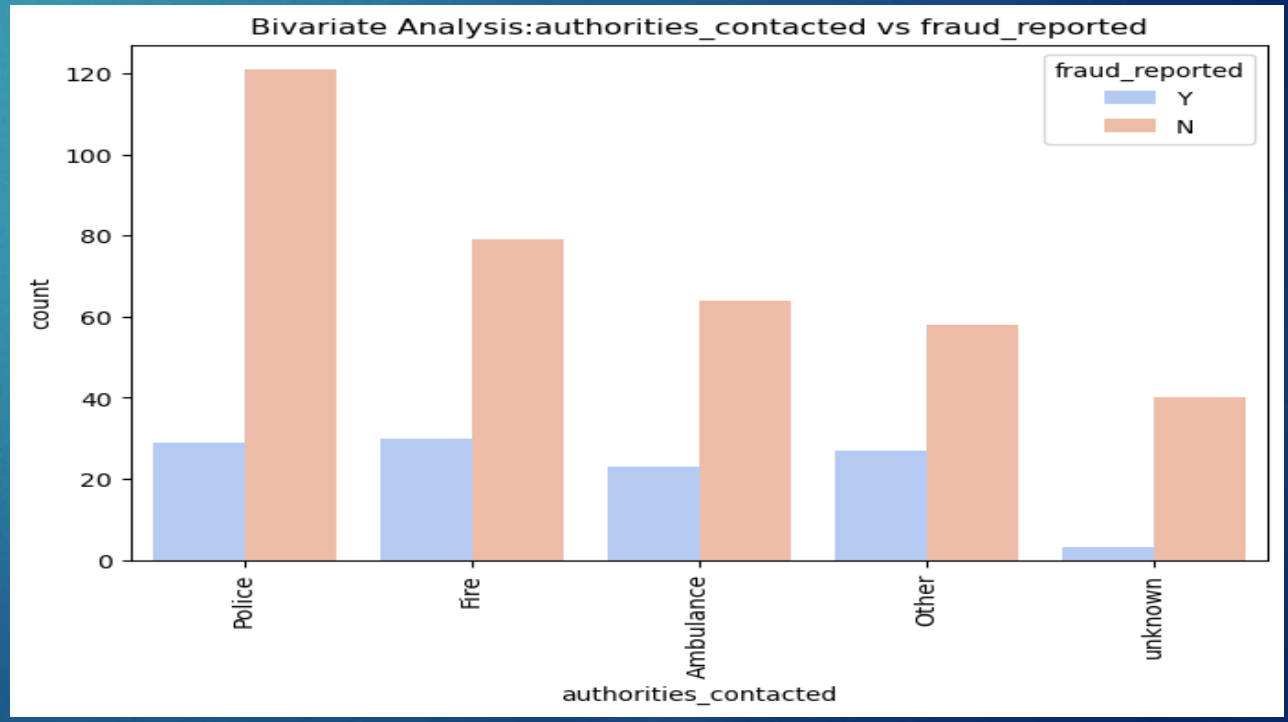
We have noticed that **"Single Vehicle Collision"** and **"Multi-vehicle Collision"** have much higher fraud reporting rates compared to "Parked Car" or "Vehicle Theft."



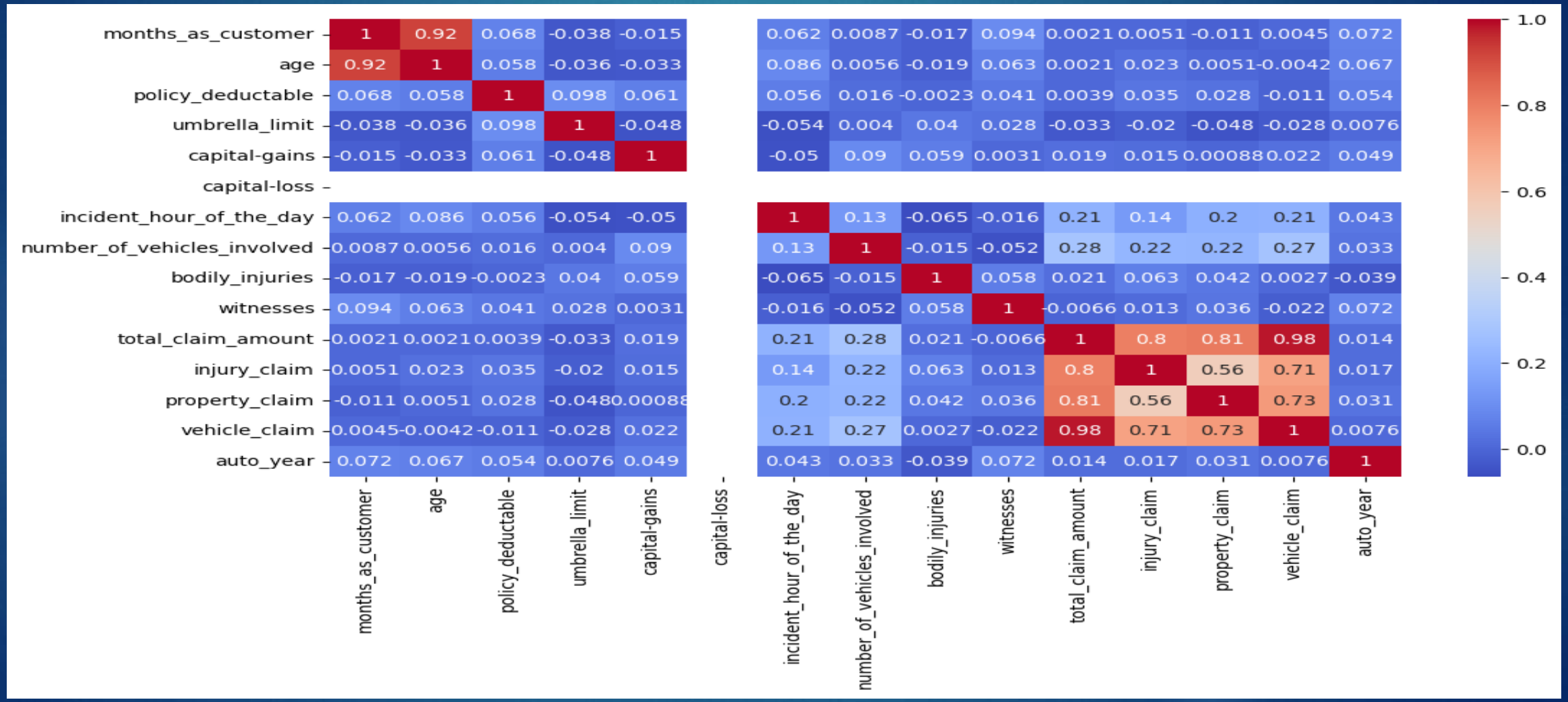


Rear-end and side-impact collisions are statistically higher risk. These are often easier to "stage" or "provoke" in real-world insurance fraud scenarios, and the data reflects this pattern.

The presence of a Police or Fire department report does not guarantee a claim is legitimate; in fact, fraudulent claims often ensure formal authorities are involved to create a paper trail.



Multicollinearity & Feature Inter-relationships



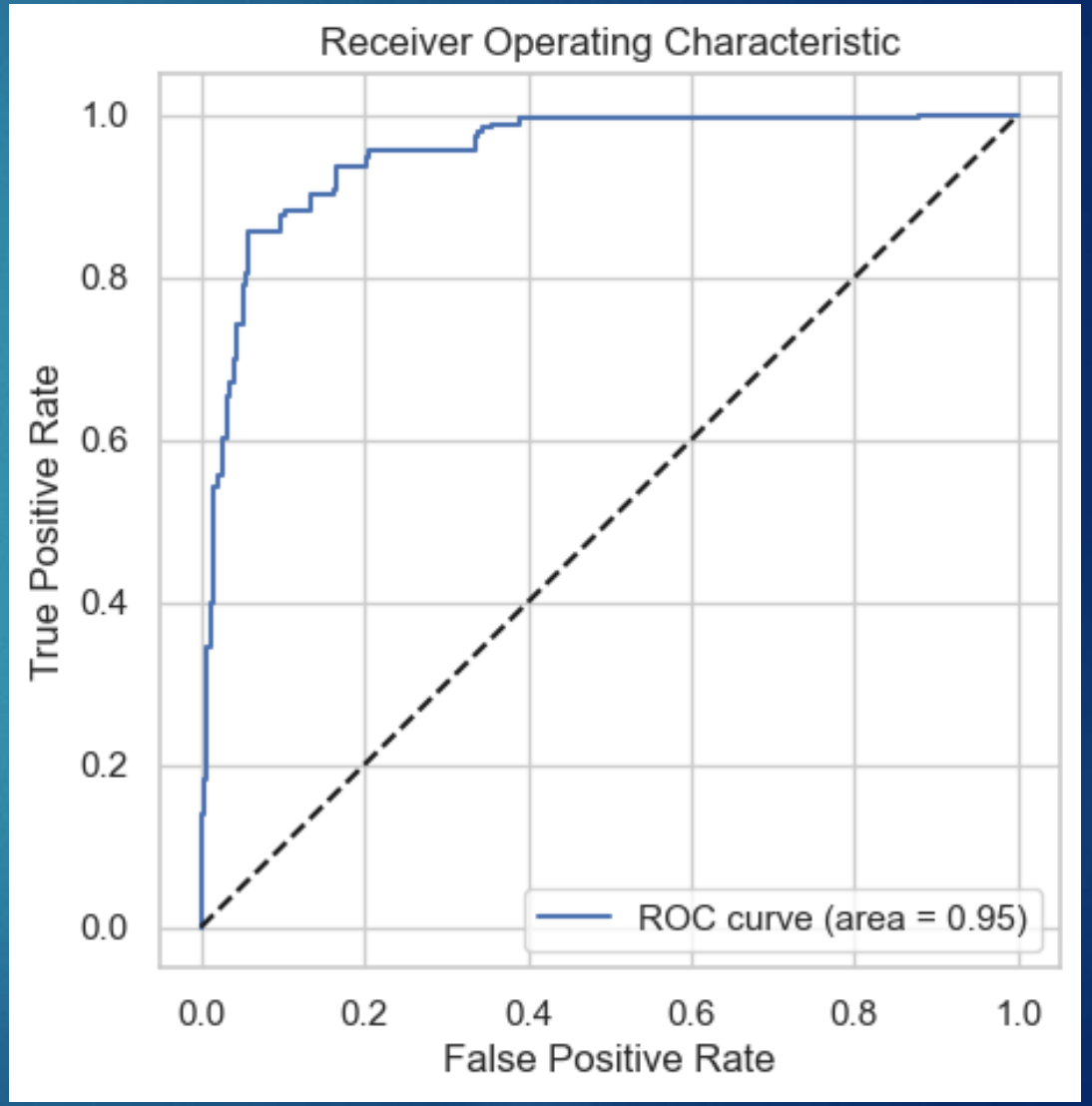
Advanced Feature Engineering & Multicollinearity Analysis

	const	age	policy_deductable	umbrella_limit	capital-gains	number_of_vehicles_involved	total_claim_amount	auto_year	incident_month	is_weekend	...	a
0	1.0	-0.675764	-0.181326	-0.486332	-0.914225	-0.802768	1.172019	-0.487861	2.0	1.427129	...	
1	1.0	-1.444793	-0.181326	-0.486332	1.565718	-0.802768	1.024574	-1.604832	1.0	-0.700707	...	
2	1.0	-1.115209	-1.014923	-0.486332	-0.914225	1.165501	0.750350	-0.009159	2.0	-0.700707	...	
3	1.0	2.729939	-1.014923	-0.486332	-0.914225	-0.802768	0.706634	-0.487861	2.0	1.427129	...	
4	1.0	1.301741	1.485868	-0.486332	-0.914225	-0.802768	-1.829737	-0.647429	2.0	-0.700707	...	

	Features	VIF
75	claim_tier_High	34.687401
6	incident_month	11.490917
76	age_group_Middle-aged	5.067579
4	total_claim_amount	4.912656
28	collision_type_Rear Collision	4.103945
67	auto_model_Other_Model	3.710303
32	incident_severity_Trivial Damage	3.407640
33	authorities_contacted_Fire	3.393391
49	property_damage_unknown	3.264829
35	authorities_contacted_Police	3.255389

Model Diagnostics: ROC Curve & AUC Performance

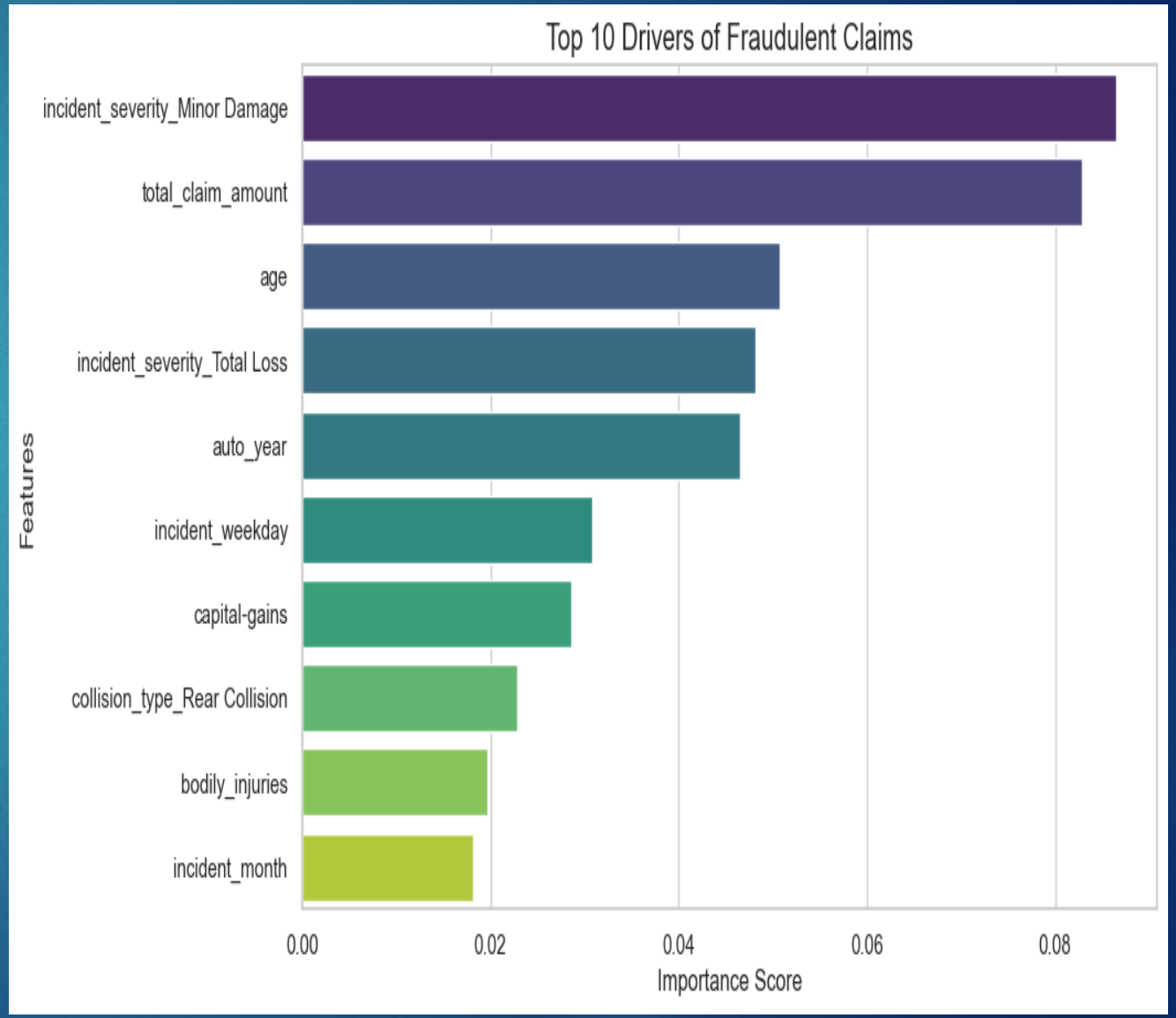
The ROC Curve demonstrates a strong separation between classes, with the **Area Under the Curve (AUC)** indicating that the model has a high probability of correctly distinguishing a fraudulent claim from a legitimate one.



The "Most Wanted" List: Top Indicators of Insurance Fraud

Critical Risk Factors: The model identifies **Incident Severity (Minor Damage/Total Loss)** and **Total Claim Amount** as the most significant predictors. High-impact accidents with high financial stakes are key indicators of potential fraud.

Demographic & Historical Context: Age and **Auto Year** also play a supporting role in detection, suggesting that older vehicles or specific age groups may follow distinct patterns in fraudulent reporting.



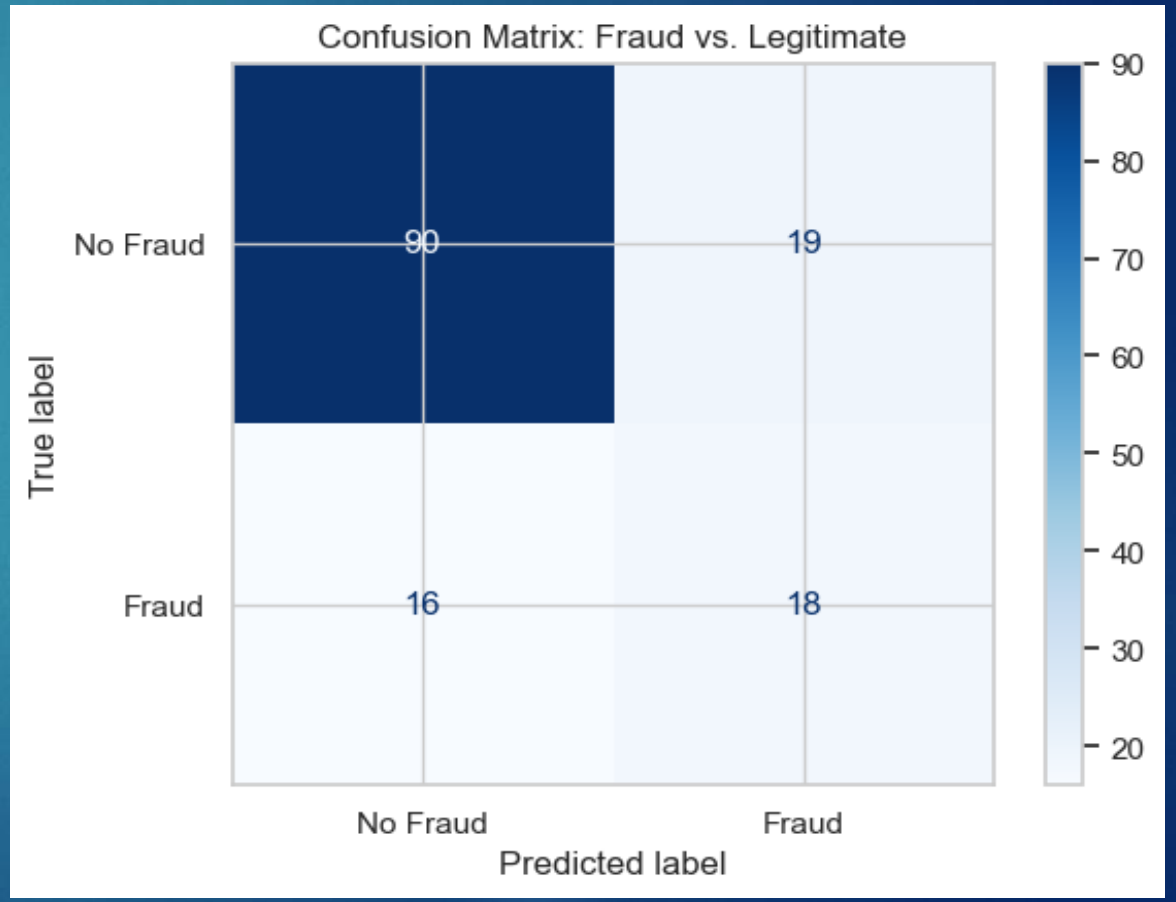
The Model Scoreboard: Assessing Fraud Detection Accuracy

```
Sensitivity: 0.53
Specificity: 0.83
Precision: 0.49
Recall: 0.53
F1-score: 0.51
```

With a **Sensitivity (Recall) of 0.53**, the model successfully identifies 53% of all actual fraudulent claims, providing a proactive shield against financial loss.

A high **Specificity of 0.83** means the model is excellent at recognizing legitimate claims, ensuring 83% of honest customers are not wrongly flagged for investigation.

The final **Validation Accuracy of 75.5%** proves that the model is robust enough to act as a reliable "First Filter" in the insurance approval pipeline.



Final Recommendation: The Winning Fraud Strategy

- **Superior Detection Power:** The **Tuned Random Forest** is our champion model, doubling the fraud detection rate (Sensitivity: **0.53**) compared to traditional Logistic Regression.

- **Operational Efficiency:** By maintaining a **Specificity of 0.83**, the model ensures that the vast majority of legitimate claims are fast-tracked, protecting the experience for honest customers.

- **Strategic Impact:** Transitioning from manual reviews to this AI-driven approach allows Global Insure to automate the audit of **over 80% of claims**, focusing investigative resources only where the risk is highest.