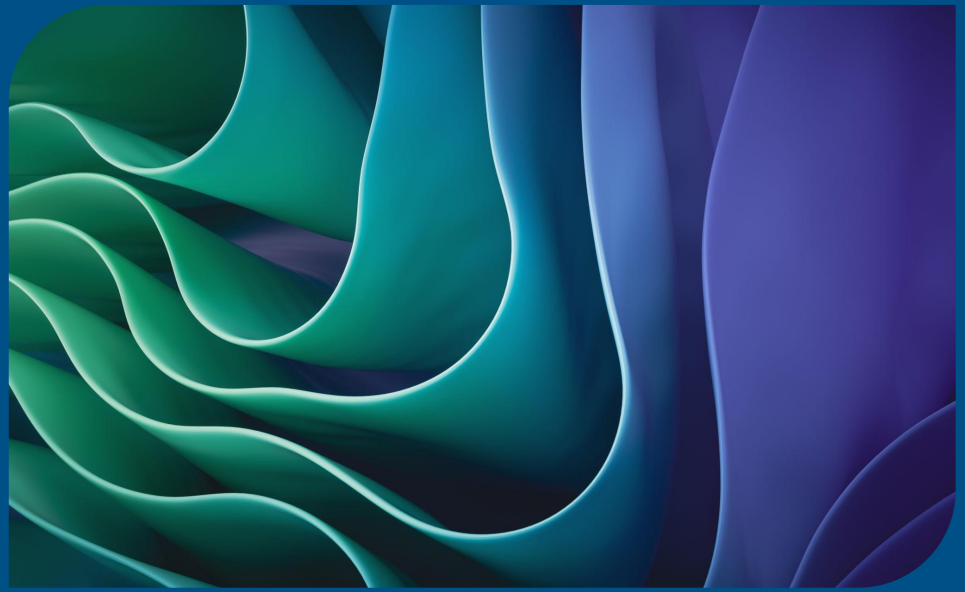
Fraudulent Claim Detection



Lorem Ipsum

Agenda

1. Defining the Problem
2. Data Preparation
3. Data Cleaning
4. EDA
5. Feature Engineering
6. Model Building
7. Predicting and Evaluation

Agenda

1. **Defining the Problem**
2. Data Preparation
3. Data Cleaning
4. EDA
5. Feature Engineering
6. Model Building
7. Predicting and Evaluation

Problem Statement

Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses. The company's current process for identifying fraudulent claims involves manual inspections, which is time-consuming and inefficient. Fraudulent claims are often detected too late in the process, after the company has already paid out significant amounts. Global Insure wants to improve its fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process. This would minimise financial losses and optimise the overall claims handling process.

Business Objective

Global Insure wants to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features like claim amounts, customer profiles and claim types, the company aims to predict which claims are likely to be fraudulent before they are approved

Agenda

1. Defining the Problem
2. **Data Preparation**
3. Data Cleaning
4. EDA
5. Feature Engineering
6. Model Building
7. Predicting and Evaluation

Data Preparation

Provided Data Dictionary

Column Name	Description
months_as_customer	Duration in months that a customer has been with the insurance company.
age	Age of the insured person.
policy_number	Unique identifier for each insurance policy.
policy_bind_date	Date when the insurance policy was initiated.
policy_state	State where the insurance policy is applicable.
policy_csl	Combined single limit for the insurance policy.
policy_deductable	Amount the insured must pay before coverage begins.
policy_annual_premium	Yearly cost of the insurance policy.
umbrella_limit	Extra liability coverage beyond the primary policy.
insured_zip	Zip code of the insured person.
insured_sex	Gender of the insured person.
insured_education_level	Highest educational qualification of the insured person.
insured_occupation	Profession/job of the insured person.
insured_hobbies	Hobbies or leisure activities of the insured person.
insured_relationship	Relationship of the insured to the policyholder.
capital-gains	Profit from selling assets like stocks, bonds, or real estate.
capital-loss	Loss from selling assets like stocks, bonds, or real estate.
incident_date	Date when the incident occurred.
incident_type	Type/category of the incident that led to the claim.

collision_type	Type of collision that occurred.
incident_severity	Extent of damage or injury caused by the incident.
authorities_contacted	Authorities/agencies contacted after the incident.
incident_state	State where the incident occurred.
incident_city	City where the incident occurred.
incident_location	Specific address/location of the incident.
incident_hour_of_the_day	Hour of the day when the incident occurred.
number_of_vehicles_involved	Total number of vehicles involved in the incident.
property_damage	Indicates if there was property damage.
bodily_injuries	Number of bodily injuries resulting from the incident.
witnesses	Number of witnesses present.
police_report_available	Indicates if a police report is available.
total_claim_amount	Total amount claimed for the incident.
injury_claim	Amount claimed for injuries.
property_claim	Amount claimed for property damage.
vehicle_claim	Amount claimed for vehicle damage.
auto_make	Manufacturer of the insured vehicle.
auto_model	Model of the insured vehicle.
auto_year	Year of manufacture of the insured vehicle.
fraud_reported	Indicates whether the claim was reported as fraudulent.
_c39	Unknown or unspecified variable.

- The dimension of the given dataset is 1000 Rows and 40 Columns
- The dataset contains a total of 21 categorical features
  - policy_bind_date
  - policy_state
  - policy_csl
  - insured_sex
  - insured_education_level
  - insured_occupation
  - insured_hobbies
  - insured_relationship
  - incident_date
  - incident_type
  - collision_type
  - incident_severity
  - authorities_contacted
  - incident_state
  - incident_city
  - incident_location
  - property_damage
  - police_report_available
  - auto_make
  - auto_model
  - fraud_reported

- The dataset contains a total of 19 numerical features
  - months_as_customer
  - age
  - policy_number
  - policy_deductable
  - umbrella_limit
  - insured_zip
  - capital-gains
  - capital-loss
  - incident_hour_of_the_day
  - number_of_vehicles_involved
  - bodily_injuries
  - witnesses
  - total_claim_amount
  - injury_claim
  - property_claim
  - vehicle_claim
  - auto_year
  - policy_annual_premium
  - _c39

Agenda

1. Defining the Problem
2. Data Preparation
3. **Data Cleaning**
4. EDA
5. Feature Engineering
6. Model Building
7. Predicting and Evaluation

Data Cleaning

Data Cleaning

Feature Name	Missing Percentage (%)
months_as_customer	0
age	0
policy_number	0
policy_bind_date	0
policy_state	0
policy_csl	0
policy_deductable	0
policy_annual_premium	0
umbrella_limit	0
insured_zip	0
insured_sex	0
insured_education_level	0
insured_occupation	0
insured_hobbies	0
insured_relationship	0
capital-gains	0
capital-loss	0
incident_date	0

Feature Name	Missing Percentage (%)
incident_type	0
collision_type	0
incident_severity	0
authorities_contacted	9.1
incident_state	0
incident_city	0
incident_location	0
incident_hour_of_the_day	0
number_of_vehicles_involved	0
property_damage	0
bodily_injuries	0
witnesses	0
police_report_available	0
total_claim_amount	0
injury_claim	0
property_claim	0
vehicle_claim	0
auto_make	0
auto_model	0
auto_year	0
fraud_reported	0
_c39	100

Unique and Near Unique Columns

Column Name	Unique Values	Proportion (%)	Total Count
months_as_customer	384	42.24	909
age	46	5.06	909
policy_number	909	100	909
policy_bind_date	865	95.16	909
policy_state	3	0.33	909
policy_csl	3	0.33	909
policy_deductable	3	0.33	909
policy_annual_premium	903	99.34	909
umbrella_limit	11	1.21	909
insured_zip	904	99.45	909
insured_sex	2	0.22	909
insured_education_level	7	0.77	909
insured_occupation	14	1.54	909
insured_hobbies	20	2.2	909
insured_relationship	6	0.66	909
capital-gains	312	34.32	909
capital-loss	341	37.51	909
incident_date	60	6.6	909
incident_type	4	0.44	909
collision_type	4	0.44	909
incident_severity	4	0.44	909
incident_state	7	0.77	909
incident_city	7	0.77	909
incident_location	909	100	909

Unique and Near Unique Columns

incident_hour_of_the_day	24	2.64	909
number_of_vehicles_involved	4	0.44	909
property_damage	3	0.33	909
bodily_injuries	3	0.33	909
witnesses	4	0.44	909
police_report_available	3	0.33	909
total_claim_amount	719	79.1	909
injury_claim	616	67.77	909
property_claim	603	66.34	909
vehicle_claim	691	76.02	909
auto_make	14	1.54	909
auto_model	39	4.29	909
auto_year	21	2.31	909
fraud_reported	2	0.22	909

Illogical or Invalid Values

Umbrella Limit

Statistic	Value
Count	909
Mean	1,088,009
Std Dev	2,278,747
Min	-1,000,000
25th Percentile (Q1)	0
Median (Q2)	0
75th Percentile (Q3)	0
Max	10,000,000

Property Damage

Property Damage	Percentage (%)
?	36.23
NO	33
YES	30

Collision Type

Collision Type	Count
Rear Collision	292
Side Collision	275
Front Collision	254
?	8

Police Report Available

Police Report Available	Percentage (%)
?	34.69
NO	34
YES	32

Data Summary

- The dataset contains **909 records** with a mix of **numerical and categorical features**.
- There are **40+ columns**, capturing customer details, insurance policy information, incidents, claims, and whether fraud was reported.
- **No missing values** in the majority of features — 38 out of 40 features have **0% missing data**.
- **authorities_contacted** has **9.1% missing values**, which may need imputation or analysis before modeling.
- **_c39** has **100% missing values**, indicating it can likely be dropped as it contains no useful data.
- Overall, the dataset is mostly complete, with only **2 features needing attention** for missing values.

Key Numeric Insights

- **High Cardinality**:
    - **policy_number, incident_location**: Each has 909 unique values (100% unique).
    - **policy_annual_premium**, **insured_zip** : Show significant variation across records.
- **Claims Data**:
    - Columns like **total_claim_amount, injury_claim, property_claim**, and **vehicle_claim** are **highly skewed** with a **median of 0** but a **maximum as high as 10 million**, indicating a few very large claims.
    - Presence of **negative values** in **umbrella_limit** (e.g., -1,000,000) may need further data cleaning.

Data Quality Observations

- Several columns contain missing or ambiguous values marked as **"?"**, especially in **property_damage** and **police_report_available**.
- Some features have low variance or limited usefulness (e.g., **_c39**, which has a single unique value).
- Columns like **authorities_contacted** have around **9% missing values**.

Categorical Features

- There are **21 categorical features**, including:
    - Customer demographics: **insured_sex**, **insured_education_level**, etc.
    - Incident details: **incident_type**, **collision_type**, **incident_severity**, etc.
    - Vehicle and policy info: **auto_make**, **auto_model**, **policy_state**, etc.
- **Collision Type Distribution**:
    - Rear: 292
    - Side: 275
    - Front: 254
    - Unknown: 8
- **Property Damage**:
    - Unknown (?): 36.23%
    - No: 33.48%
    - Yes: 30.29%
- **Police Report Availability**:
    - Unknown (?): 34.69%
    - No: 33.81%
    - Yes: 31.50%

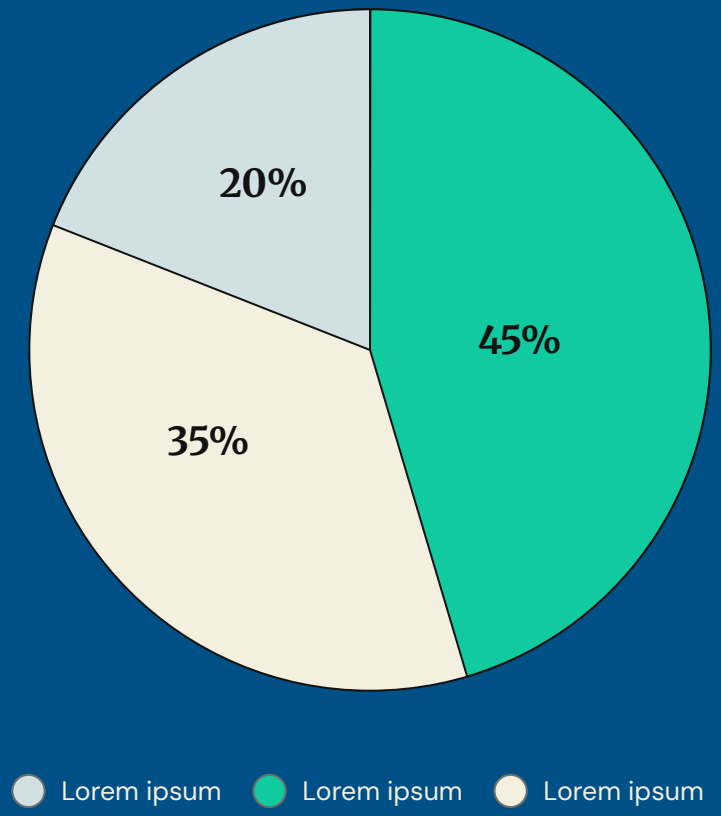
Handling Data

- **Remove rows** where the **authorities_contacted** column contains null values.
- **Drop the column _c39** as it contains only null values.
- **Remove high-cardinality columns** including **policy_number**, **incident_location**, **policy_annual_premium**, and **insured_zip**, since they have unique or near-unique values and provide little predictive power.
- **Delete rows** where the **umbrella_limit** contains negative values, as these are likely erroneous.
- **Replace ambiguous entries (?) with 'UNKNOWN'** in the columns: **police_report_available**, **property_damage**, and **collision_type**.
- **Convert data types** of **policy_bind_date** and **incident_date** from object to datetime for accurate temporal analysis.

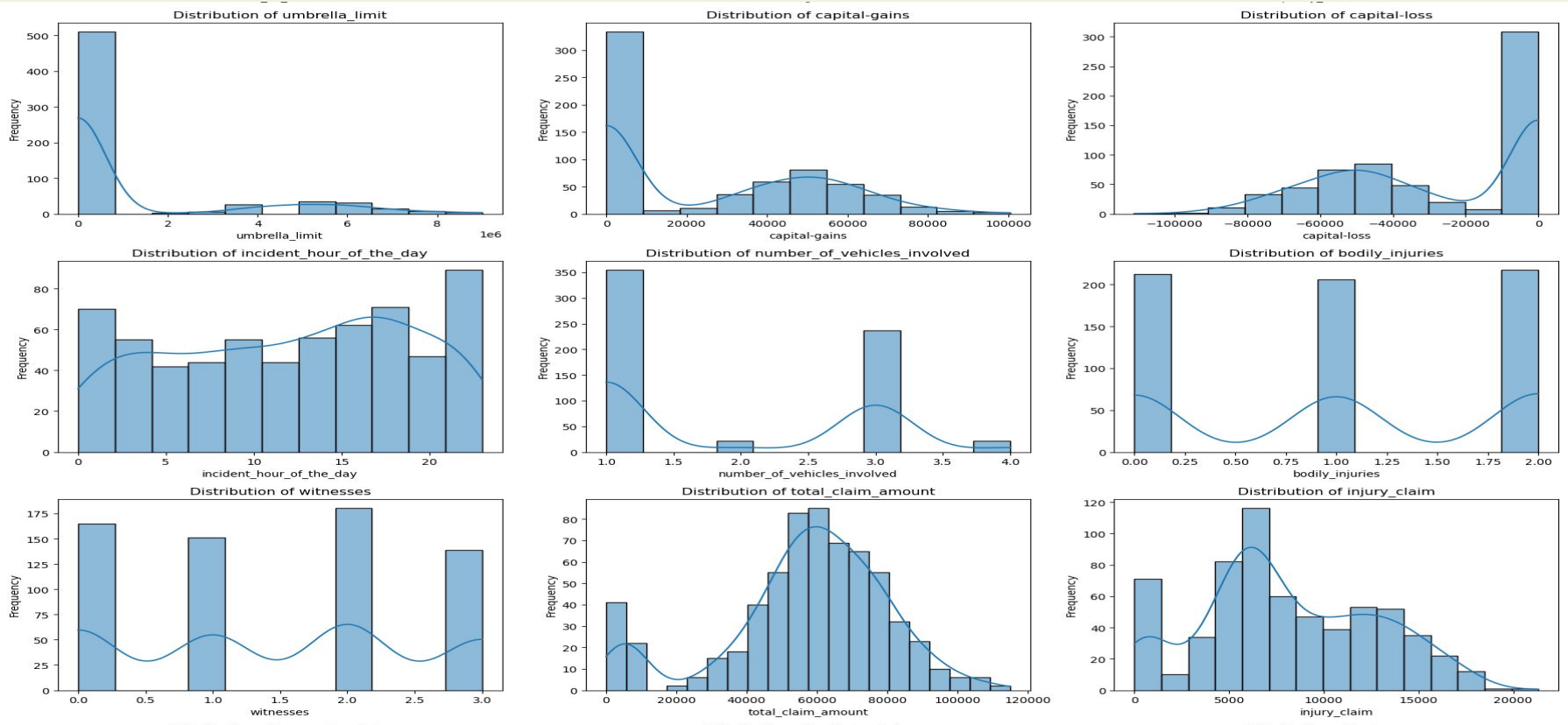
Agenda

1. Defining the Problem
2. Data Preparation
3. Data Cleaning
4. **EDA**
5. Feature Engineering
6. Model Building
7. Predicting and Evaluation

EDA

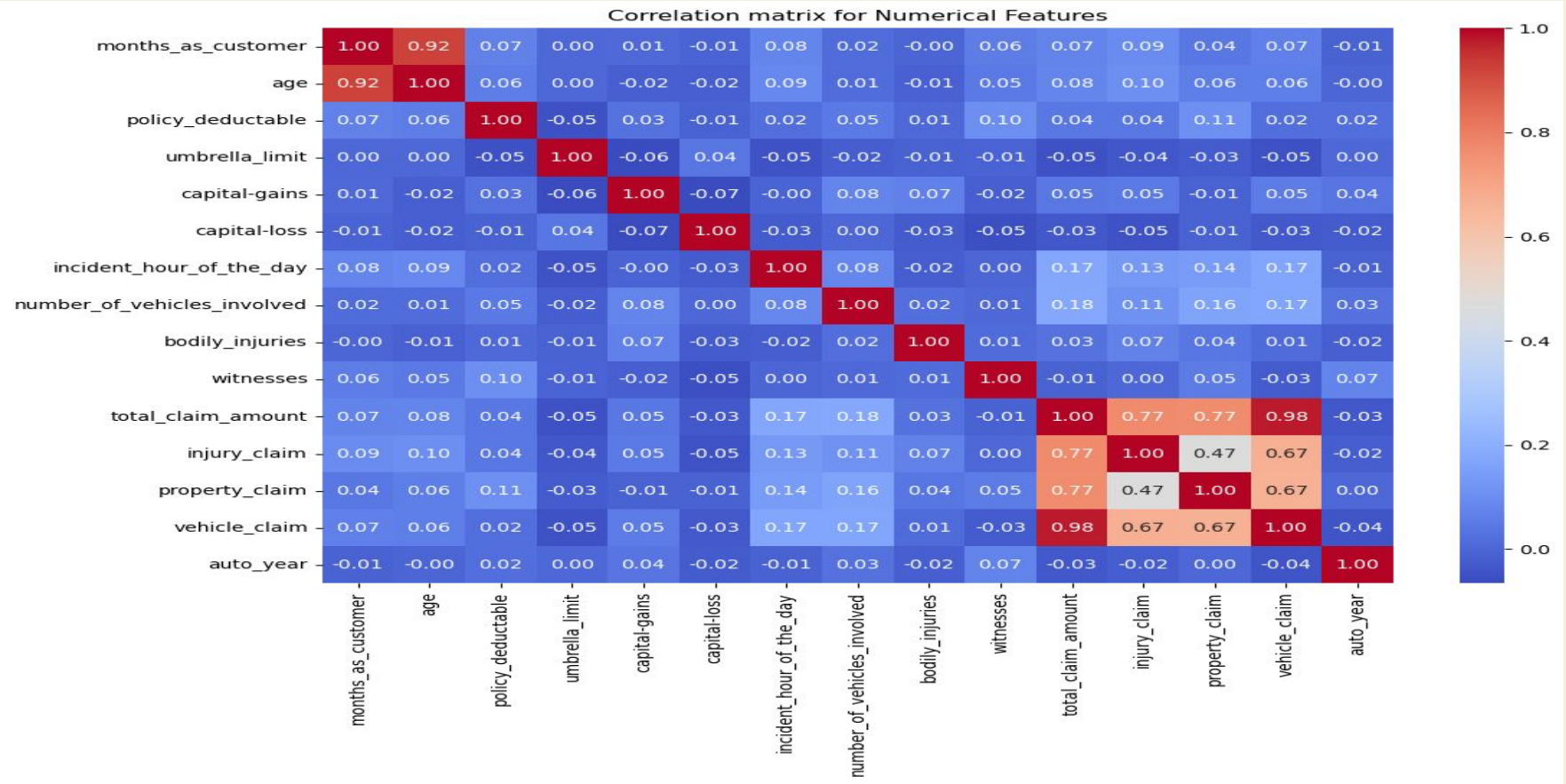


Distribution of Numerical Columns



- **Customer Tenure:** Skewed towards shorter durations with the company.
- **Age:** Roughly normally distributed, concentrated in the 30-40 age range.
- **Policy Deductible:** Most policies have lower deductibles.
- **Umbrella Limit:** Predominantly zero, with a few policies having very high limits.
- **Capital Gains & Losses:** Heavily concentrated at zero.
- **Incident Hour:** Peaks during typical commuting hours.
- **Vehicles Involved:** Most incidents involve one or two vehicles.
- **Bodily Injuries:** Most incidents report zero or one bodily injury.
- **Witnesses:** Typically zero or one witness per incident.
- **Total Claim Amount:** Appears somewhat normally distributed.
- **Injury & Property Claims:** Skewed towards lower amounts.
- **Vehicle Claims:** More normally distributed compared to injury and property claims.
- **Vehicle Model Year:** More recent years appear more frequently in claims.

Correlation Matrix



Strong Positive Correlations:

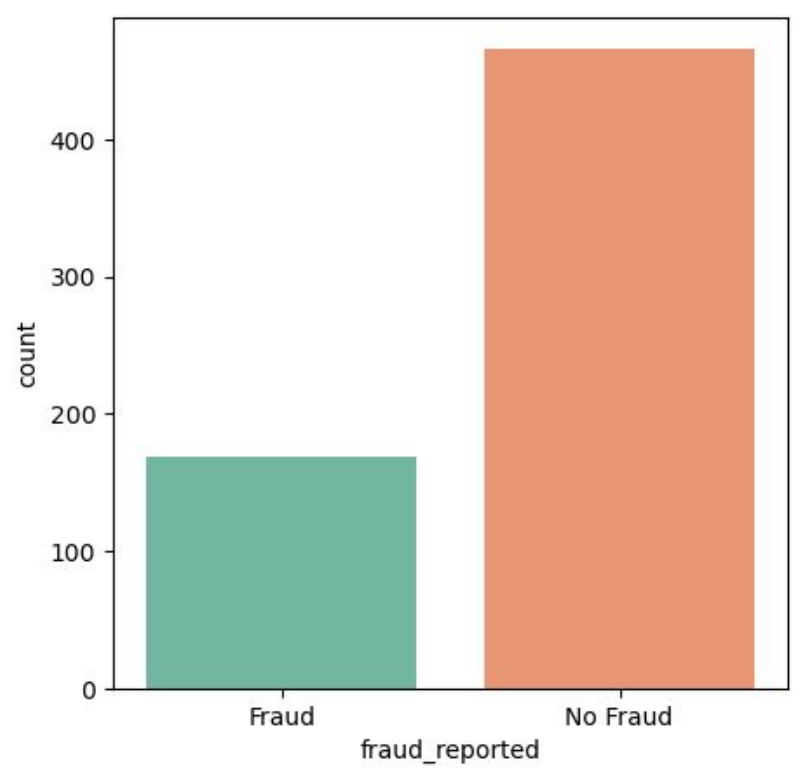
- Total Claim Amount is strongly positively correlated with Injury Claim, Property Claim, and Vehicle Claim.
- Injury Claim and Property Claim show a strong positive correlation.
- Months as Customer and Age have a moderately strong positive correlation.

Weak Correlations: Most other pairs of numerical features exhibit weak or very weak linear correlations (close to zero).

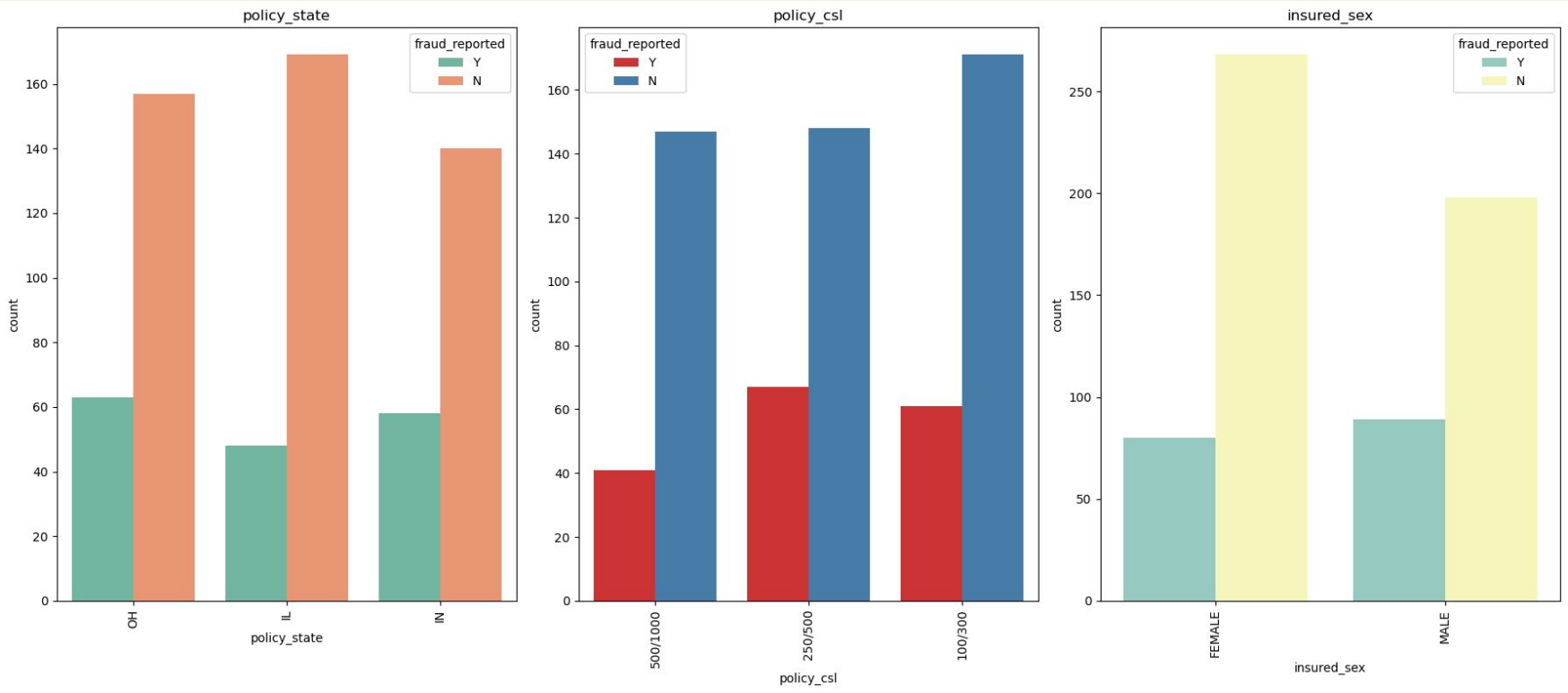
Weak Negative Correlations: A few negligible weak negative correlations exist (e.g., between Capital Loss and Capital Gains, and Auto Year with Months as Customer and Age).

Feature Imbalance

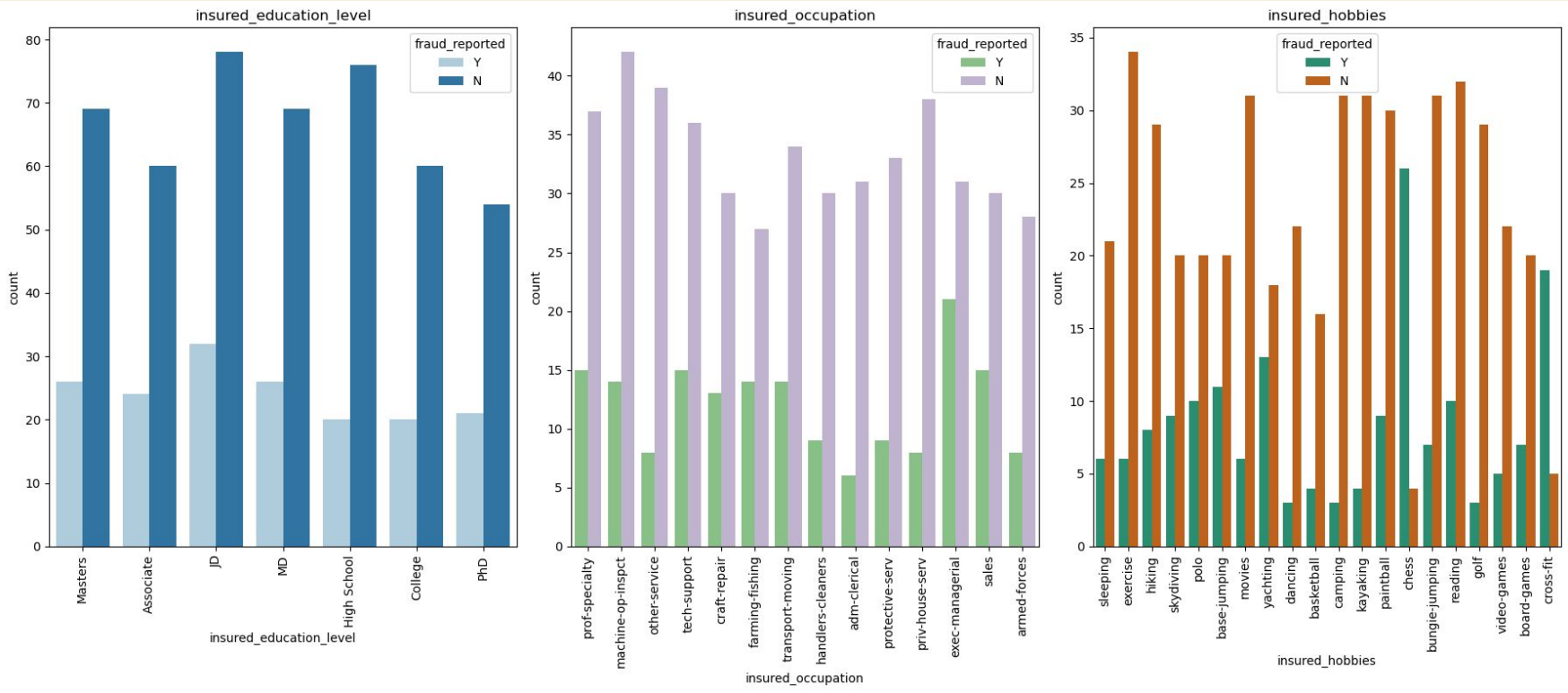
The chart clearly shows that the vast majority of insurance claims in this dataset were *not* reported as fraudulent, while a smaller proportion of claims were reported as fraudulent



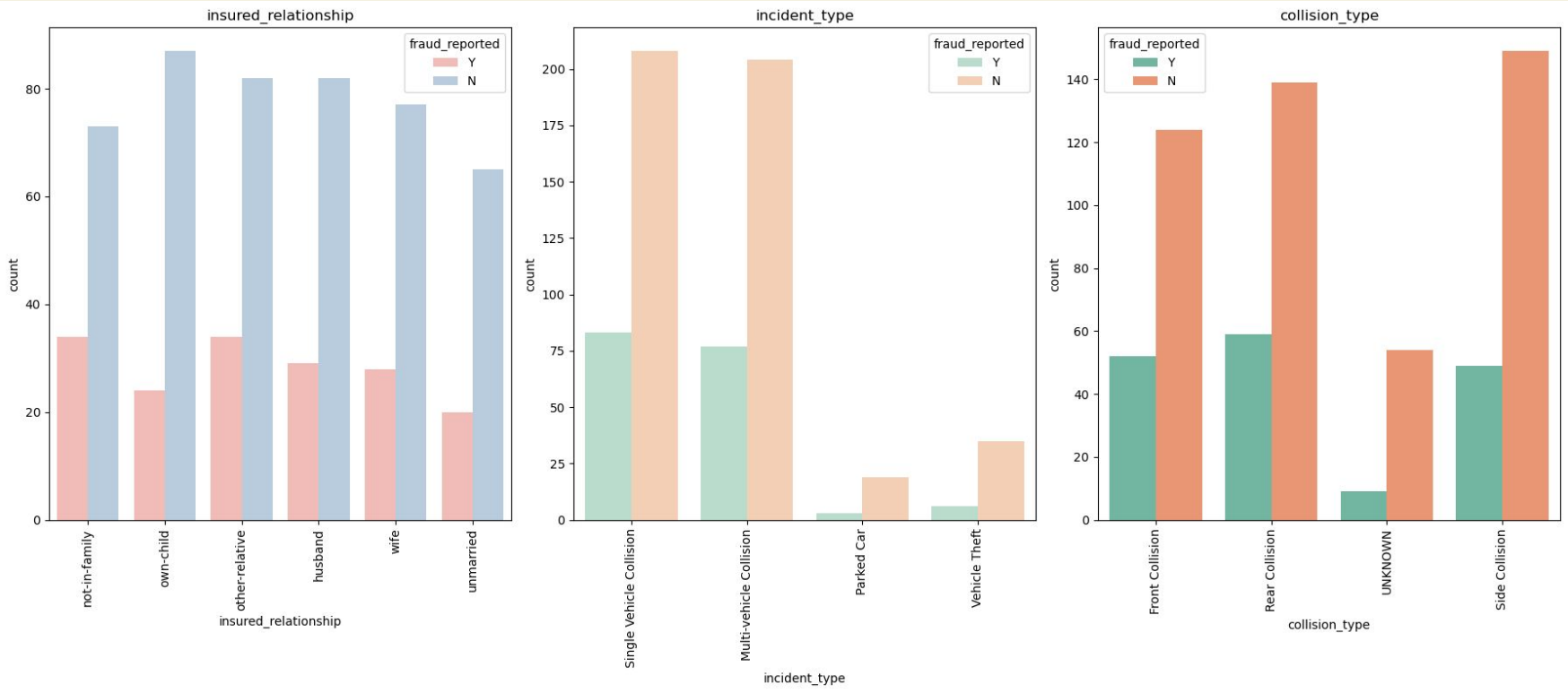
Categorical Feature vs Target Feature



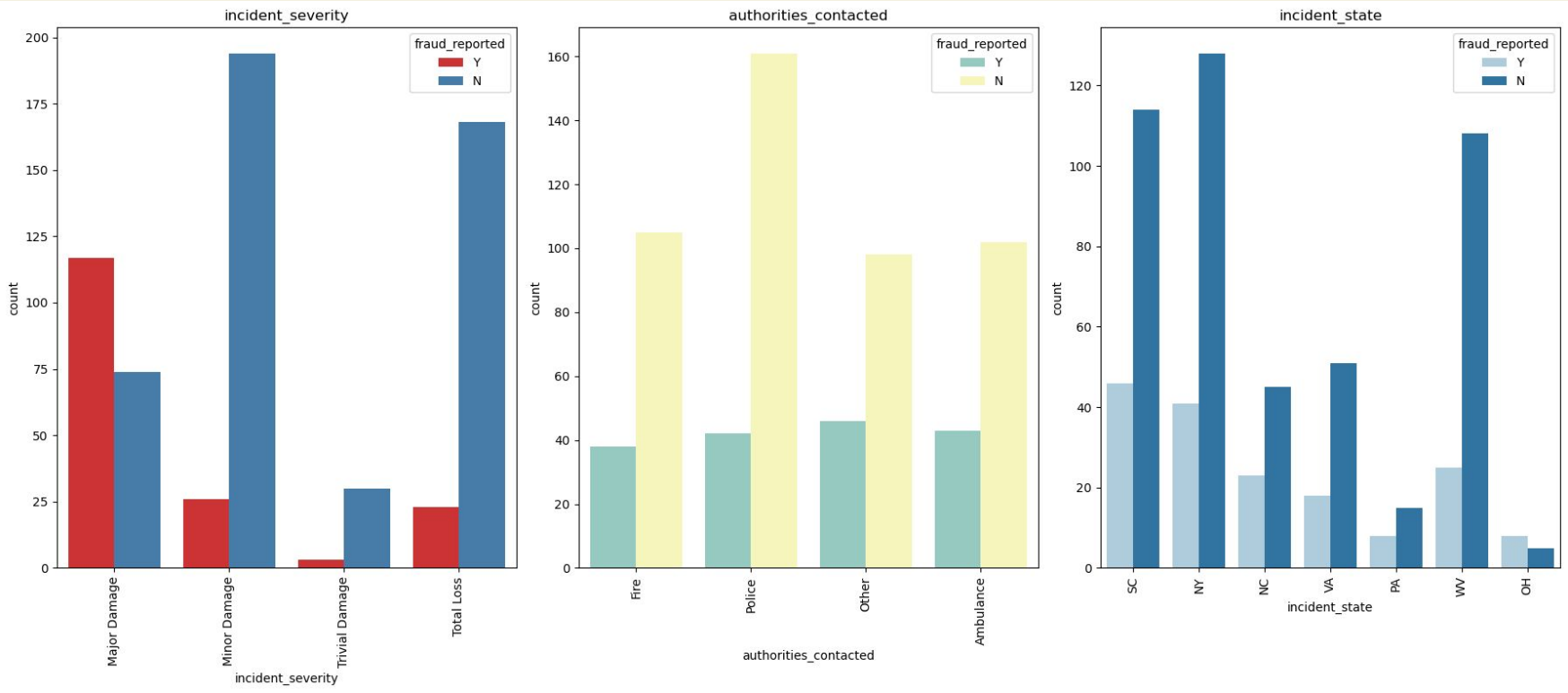
Categorical Feature vs Target Feature



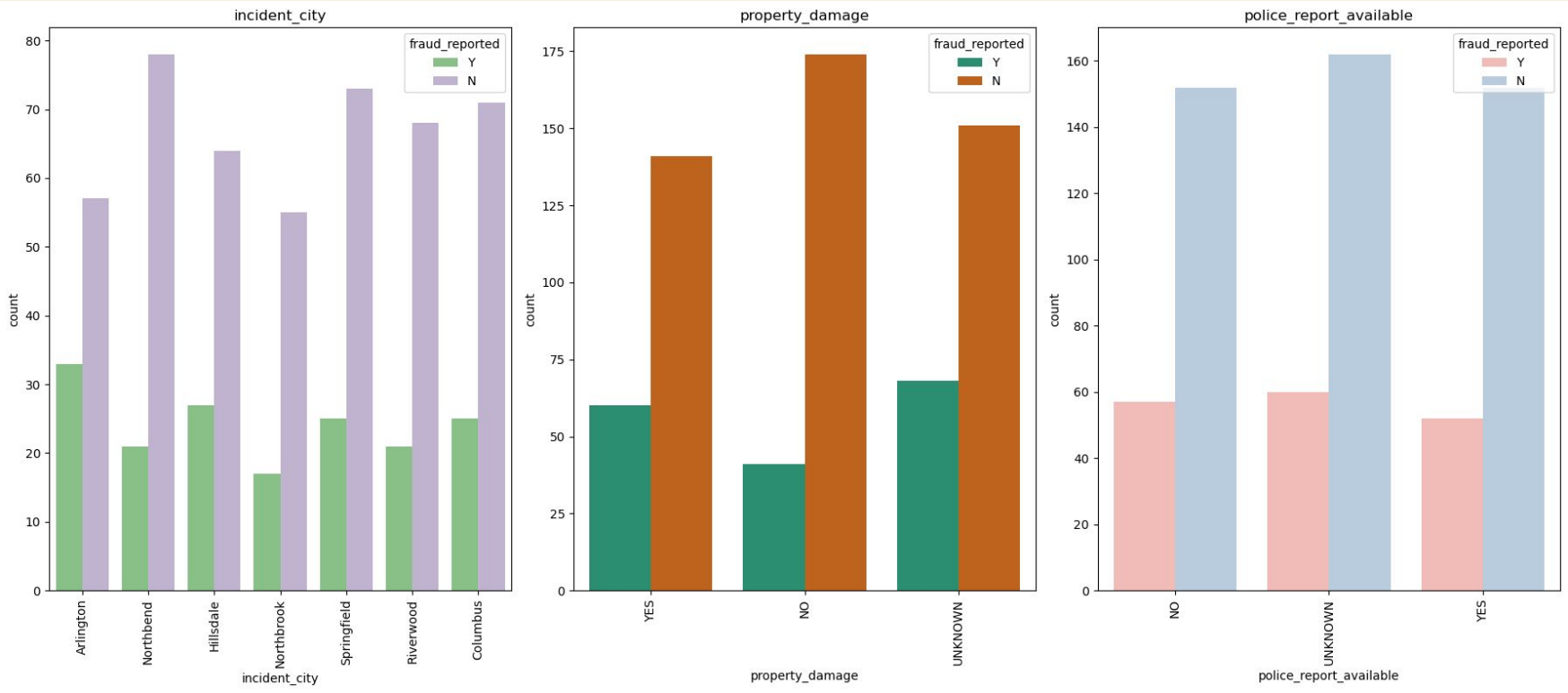
Categorical Feature vs Target Feature



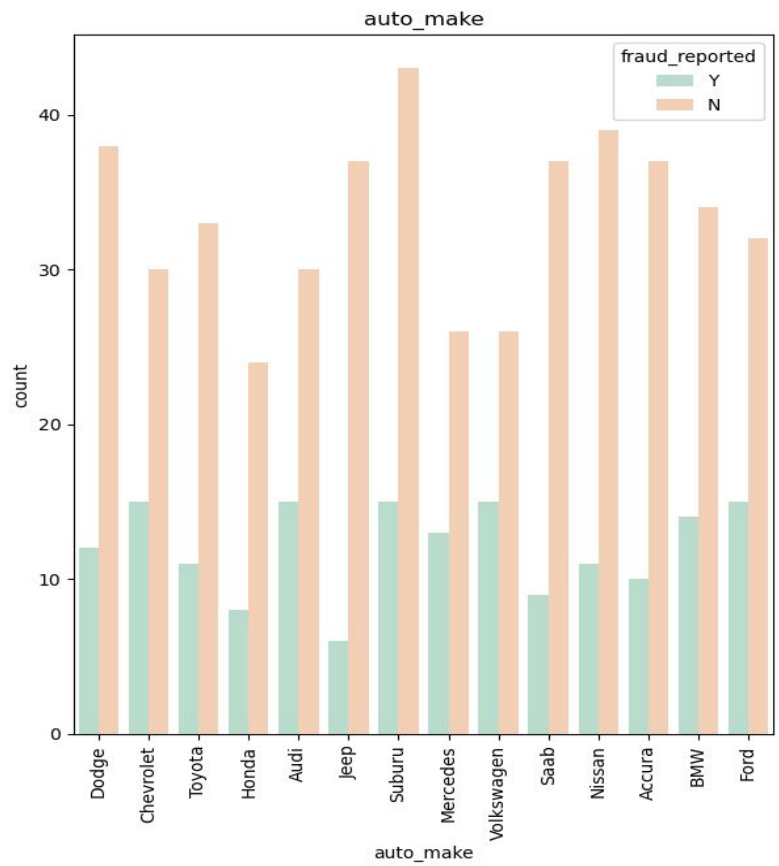
Categorical Feature vs Target Feature

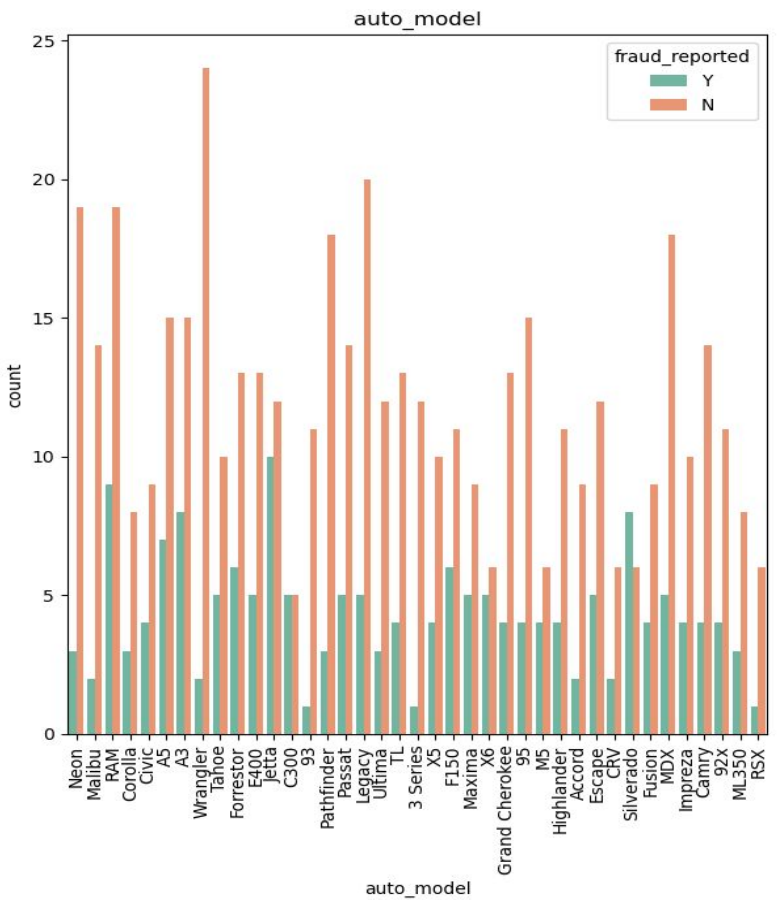


Categorical Feature vs Target Feature



Categorical Feature vs Target Feature





Features Showing Strong Fraud Patterns

- **policy_csl**: Lower coverage (500/1000) has fewer frauds compared to higher ones.
- **incident_severity**: "Minor Damage" and "Major Damage" dominate fraud reports.
- **incident_type**: "Multi-vehicle collision" shows a higher count of frauds.
- **authorities_contacted**: Police and Fire are often contacted in fraudulent claims.
- **insured_hobbies**: Certain hobbies like "chess", "cross-fit", or "skydiving" show slightly different patterns—may be grouped or treated as noise.
- **property_damage**: When damage is marked "YES", fraud reports increase.
- **police_report_available**: Fraud rates are slightly higher when the report is not available ("NO" or "UNKNOWN").

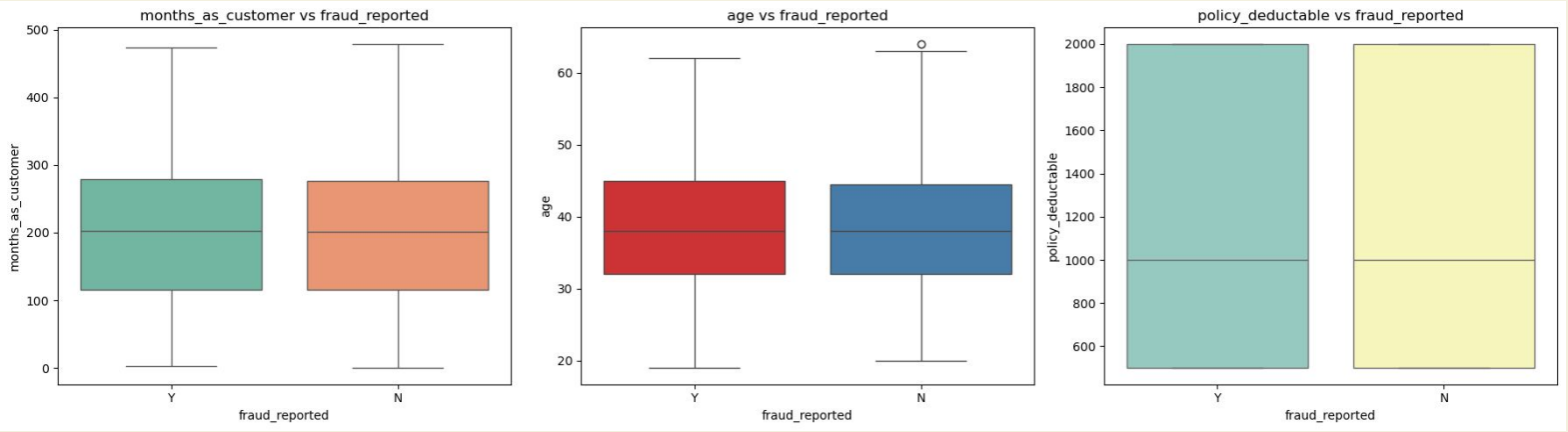
Features with Balanced or Uniform Distribution

- **insured_sex**: Fairly balanced; not strongly indicative of fraud.
- **insured_education_level**: Even distribution; unlikely to affect fraud prediction.
- **policy_state**: No strong fraud pattern based on state alone.

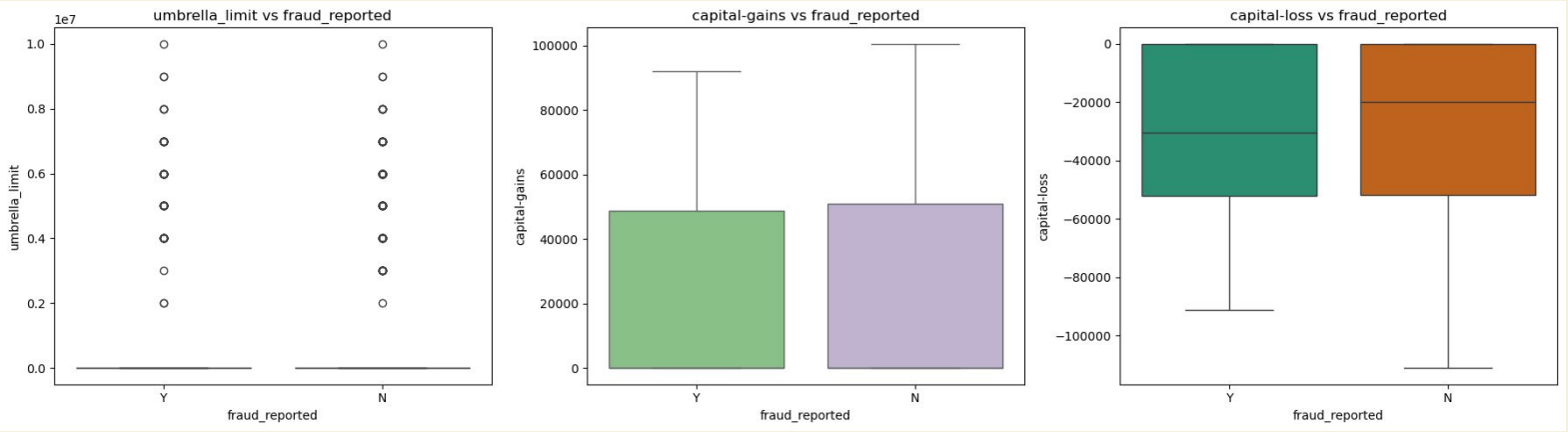
Features with Sparse or Redundant Categories

- **insured_occupation**: Many occupations have very low counts.
- **insured_hobbies**: Many low-frequency hobbies.
- **insured_relationship**: Most common categories like "spouse", "husband", "wife" dominate.
- **auto_model**: Extremely sparse.
- **auto_make**: Similar to auto_model, but less sparse.

Numerical Feature vs Target Feature

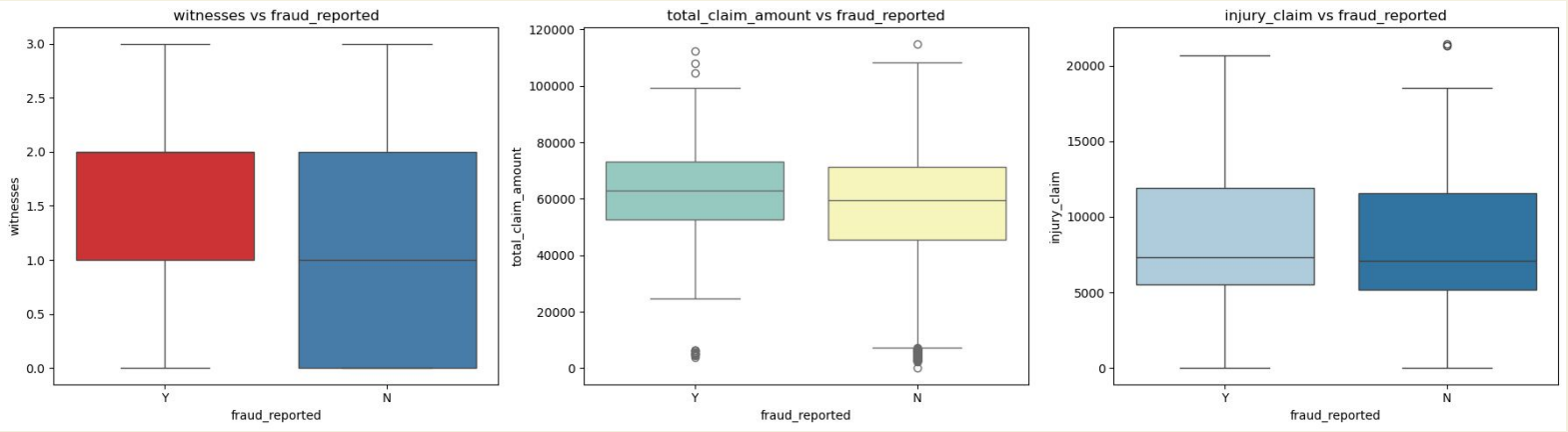


Numerical Feature vs Target Feature

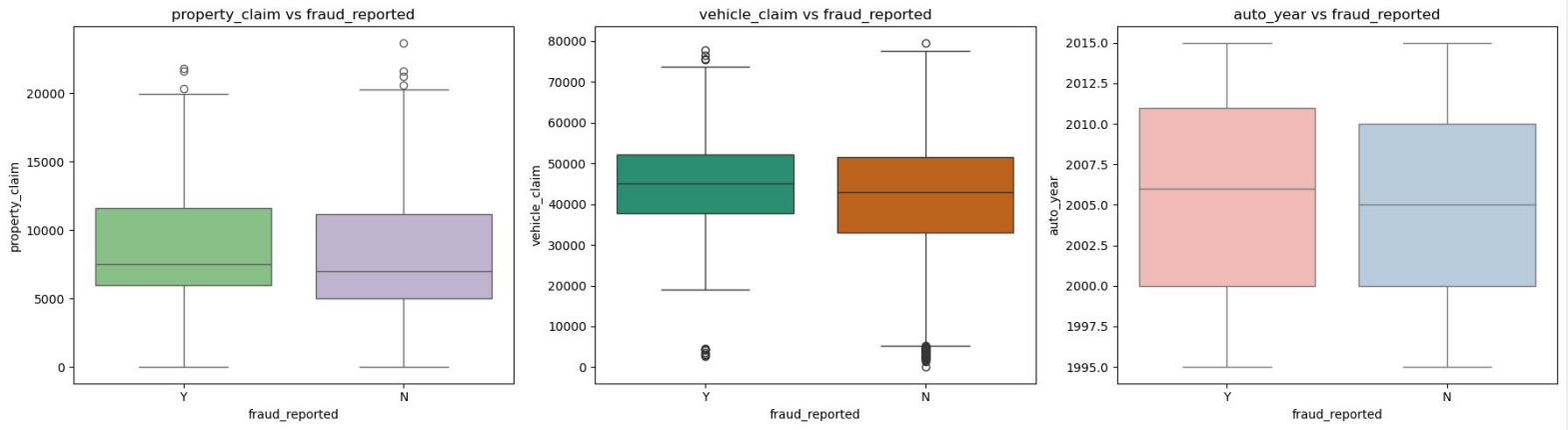


Numerical Feature vs Target Feature

Numerical Feature vs Target Feature



Numerical Feature vs Target Feature



1. Customer & Policy Info

- **months_as_customer** and **age** show similar distributions for fraud and non-fraud cases, indicating little to no discrimination power.
- **policy_deductable** has the same values across both classes, suggesting it might be a constant or categorical-like feature.

2. Financial Limits

- **umbrella_limit** is highly skewed with many outliers but similar for both classes—possible limited utility unless binned.
- **capital-gains** and **capital-loss** show similar spread across fraud and non-fraud, indicating limited predictive power.

3. Incident Details

- **incident_hour_of_the_day** is nearly identical between both classes, offering little insight.
- **number_of_vehicles_involved** and **bodily_injuries** show no significant variation across fraud categories—likely low importance.

4. Witnesses & Claims

- **witnesses** show a slightly higher median for fraud cases, which could suggest relevance.

- **total_claim_amount**, **injury_claim**, **property_claim**, and **vehicle_claim** all have wider ranges and slight distribution shifts, especially in fraud cases—these could be valuable predictors.

5. Vehicle Info

- **auto_year** distributions are nearly identical for both fraud and non-fraud, suggesting minimal impact.

Agenda

1. Defining the Problem
2. Data Preparation
3. Data Cleaning
4. EDA
5. **Feature Engineering**
6. Model Building
7. Predicting and Evaluation

Feature Engineering

incident_month

- Extracts the **month** from incident_date to capture seasonal patterns or monthly trends in fraud.

incident_day_of_week

- Extracts the **day of the week** (0 = Monday, 6 = Sunday) to detect patterns related to weekday/weekend incidents.

policy_age

- Calculates the **number of days between policy_bind_date and incident_date**.
- Indicates how long the policy was active before a claim — useful to spot early fraud attempts.

Feature Engineering

injury_claim_ratio

- Ratio of injury_claim to total_claim_amount.
- Helps normalize injury claims relative to the overall claim size.

property_claim_ratio

- Ratio of property_claim to total_claim_amount.
- Indicates the portion of the total claim attributed to property damage.

vehicle_claim_ratio

- Ratio of vehicle_claim to total_claim_amount.
- Highlights the extent of vehicle-related loss in the total claim.

Feature Engineering

insured_occupation

- Categories with **less than 6%** of total occurrences were replaced with 'Other'.
- Purpose: Prevents the model from assigning high importance to rare job titles.

insured_hobbies

- Categories with **less than 5%** frequency were grouped into 'Other'.
- Purpose: Controls for sparse or overly unique hobbies that may not generalize well.

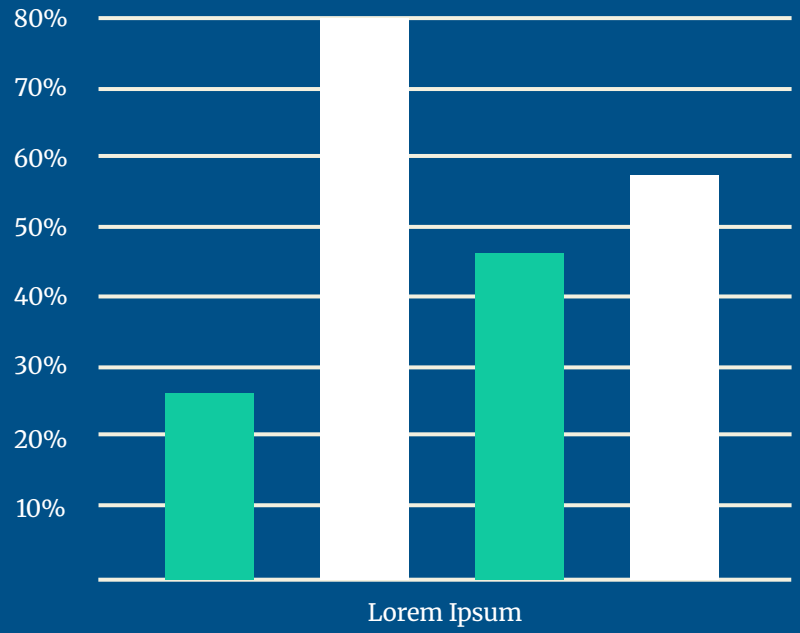
incident_type

- You replaced **incident_type** values using the same low_freq_values intended for insured_hobbies.
- This may be **unintentional**, as incident_type likely has different categories and frequencies than hobbies.

Agenda

1. Defining the Problem
2. Data Preparation
3. Data Cleaning
4. EDA
5. Feature Engineering
6. **Model Building**
7. Predicting and Evaluation

Model Building Logistic Regression



Logistic Regression Model

- Feature Selection using RFECV
- Model Building and Multicollinearity Assessment
- Model Training and Evaluation on Training Data – Fit the model on the training data and assess initial performance.
- Finding the Optimal Cutoff – Determine the best probability threshold by analysing the sensitivity-specificity tradeoff and precision-recall tradeoff.
- FInal Prediction and Evaluation on Training Data using the Optimal Cutoff – Generate final predictions using the selected cutoff and evaluate model performance.

Logistic Regression Default Model

```
Optimization terminated successfully.
         Current function value: 0.270161
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                      Y   No. Observations:                  932
Model:                          Logit   Df Residuals:                      854
Method:                           MLE   Df Model:                           77
Date:                Tue, 13 May 2025   Pseudo R-squ.:                 0.6102
Time:                        00:24:29   Log-Likelihood:                -251.79
converged:                       True   LL-Null:                       -646.01
Covariance Type:            nonrobust   LLR p-value:                1.769e-118
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     -0.6110      0.658     -0.929      0.353      -1.900       0.678
months_as_customer         0.6031      0.370      1.629      0.103      -0.122       1.329
auto_make_Suburu           1.1076        nan        nan        nan         nan         nan
auto_make_Saab            -0.3257        nan        nan        nan         nan         nan
auto_make_Nissan           0.0783      0.878      0.089      0.929      -1.642       1.798
auto_make_Dodge           -0.6696        nan        nan        nan         nan         nan
auto_make_Chevrolet        0.5765      0.942      0.612      0.540      -1.270       2.423
auto_make_BMW              0.2998      0.763      0.393      0.694      -1.195       1.794
auto_make_Audi             0.5834      0.564      1.034      0.301      -0.523       1.689
police_report_available_YES   -0.8008      0.358     -2.236      0.025      -1.503      -0.099
police_report_available_UNKNOWN   -0.7069      0.339     -2.088      0.037      -1.370      -0.043
property_damage_YES        0.8332      0.348      2.394      0.017       0.151       1.515
auto_make_Volkswagen       0.8082      0.722      1.119      0.263      -0.607       2.224
property_damage_UNKNOWN    1.2295      0.332      3.708      0.000       0.580       1.879
incident_city_Riverwood   -0.4586      0.466     -0.983      0.325      -1.373       0.455
incident_city_Northbrook   -0.8855      0.456     -1.943      0.052      -1.779       0.008
incident_city_Columbus     -0.7773      0.417     -1.866      0.062      -1.594       0.039
incident_state_WV          -0.8406      0.403     -2.087      0.037      -1.630      -0.051
incident_state_SC          -0.4457      0.381     -1.169      0.242      -1.193       0.301
incident_state_OH          0.5168      0.905      0.571      0.568      -1.257       2.290
incident_state_NY          -0.2261      0.382     -0.592      0.554      -0.975       0.523
authorities_contacted_Police   0.6803      0.340      1.999      0.046       0.013       1.348
authorities_contacted_Other   0.6341      0.347      1.828      0.068      -0.046       1.314
incident_severity_Trivial Damage   -4.8626      0.758     -6.411      0.000      -6.349      -3.376
incident_city_Springfield  -0.2149      0.392     -0.548      0.584      -0.984       0.554
incident_severity_Total Loss   -4.7433      0.430    -11.038      0.000      -5.586      -3.901
auto_model_92x             -0.9414        nan        nan        nan         nan         nan
auto_model_95              1.8757        nan        nan        nan         nan         nan
auto_model_Wrangler        -1.3482      1.043     -1.292      0.196      -3.393       0.697
auto_model_Ultima          0.8347      1.272      0.656      0.512      -1.659       3.328
```

Logistic Regression Default Model

auto_model_TL	-1.3457	0.933	-1.443	0.149	-3.174	0.482
auto_model_Silverado	1.8857	1.150	1.640	0.101	-0.368	4.140
auto_model_RSX	-1.8183	1.406	-1.293	0.196	-4.574	0.937
auto_model_RAM	0.4663	nan	nan	nan	nan	nan
auto_model_Pathfinder	-2.3652	1.218	-1.942	0.052	-4.753	0.022
auto_model_Passat	1.0581	0.936	1.130	0.259	-0.777	2.894
auto_model_Neon	-1.1359	nan	nan	nan	nan	nan
auto_model_Malibu	-0.6851	1.201	-0.571	0.568	-3.039	1.669
auto_model_93	-1.2600	nan	nan	nan	nan	nan
auto_model_ML350	-1.0836	1.301	-0.833	0.405	-3.634	1.467
auto_model_M5	1.8420	1.140	1.616	0.106	-0.392	4.076
auto_model_Legacy	-0.4221	nan	nan	nan	nan	nan
auto_model_Impreza	0.8959	nan	nan	nan	nan	nan
auto_model_Highlander	1.1884	0.769	1.545	0.122	-0.319	2.696
auto_model_Grand Cherokee	2.7074	0.729	3.712	0.000	1.278	4.137
auto_model_Forrestor	0.6337	nan	nan	nan	nan	nan
auto_model_Civic	1.0213	0.878	1.164	0.244	-0.699	2.741
auto_model_Camry	-0.4555	0.929	-0.490	0.624	-2.276	1.365
auto_model_C300	-0.3092	1.188	-0.260	0.795	-2.638	2.019
auto_model_Accord	-2.3422	1.487	-1.576	0.115	-5.256	0.571
auto_model_MDX	-0.7283	0.815	-0.893	0.372	-2.326	0.870
incident_severity_Minor Damage	-4.8508	0.430	-11.269	0.000	-5.694	-4.007
auto_model_X6	2.8459	1.115	2.553	0.011	0.661	5.031
collision_type_Side Collision	-0.9104	0.323	-2.816	0.005	-1.544	-0.277
insured_hobbies_base-jumping	1.3721	0.665	2.063	0.039	0.068	2.676
insured_occupation_transport-moving	0.2507	0.469	0.534	0.593	-0.669	1.170
insured_occupation_tech-support	0.9745	0.494	1.975	0.048	0.007	1.942
insured_occupation_sales	1.0832	0.597	1.814	0.070	-0.087	2.254
insured_occupation_prof-specialty	0.6602	0.497	1.327	0.184	-0.315	1.635
insured_occupation_exec-managerial	1.5702	0.467	3.364	0.001	0.655	2.485
insured_education_level_PhD	0.6332	0.446	1.419	0.156	-0.241	1.508
insured_education_level_High School	-0.2123	0.415	-0.511	0.609	-1.026	0.602
insured_sex_MALE	0.4663	0.276	1.689	0.091	-0.075	1.007
policy_csl_500/1000	-0.6335	0.344	-1.842	0.065	-1.308	0.041
policy_csl_250/500	0.7889	0.312	2.526	0.012	0.177	1.401
policy_state_OH	0.7797	0.331	2.354	0.019	0.131	1.429
policy_state_IN	1.2836	0.351	3.658	0.000	0.596	1.971
witnesses	0.2474	0.140	1.763	0.078	-0.028	0.522
capital-loss	-0.5572	0.134	-4.163	0.000	-0.820	-0.295
capital-gains	-0.2610	0.137	-1.909	0.056	-0.529	0.007
age	-0.9248	0.370	-2.500	0.012	-1.650	-0.200
insured_hobbies_chess	6.5219	0.707	9.228	0.000	5.137	7.907
insured_hobbies_cross-fit	5.4787	0.678	8.075	0.000	4.149	6.809
insured_education_level_Masters	0.8595	0.407	2.111	0.035	0.061	1.658
insured_hobbies_hiking	0.8521	0.565	1.509	0.131	-0.254	1.959

Logistic Regression Default Model

auto_model_TL	-1.3457	0.933	-1.443	0.149	-3.174	0.482
auto_model_Silverado	1.8857	1.150	1.640	0.101	-0.368	4.140
auto_model_RSX	-1.8183	1.406	-1.293	0.196	-4.574	0.937
auto_model_RAM	0.4663	nan	nan	nan	nan	nan
auto_model_Pathfinder	-2.3652	1.218	-1.942	0.052	-4.753	0.022
auto_model_Passat	1.0581	0.936	1.130	0.259	-0.777	2.894
auto_model_Neon	-1.1359	nan	nan	nan	nan	nan
auto_model_Malibu	-0.6851	1.201	-0.571	0.568	-3.039	1.669
auto_model_93	-1.2600	nan	nan	nan	nan	nan
auto_model_ML350	-1.0836	1.301	-0.833	0.405	-3.634	1.467
auto_model_M5	1.8420	1.140	1.616	0.106	-0.392	4.076
auto_model_Legacy	-0.4221	nan	nan	nan	nan	nan
auto_model_Impreza	0.8959	nan	nan	nan	nan	nan
auto_model_Highlander	1.1884	0.769	1.545	0.122	-0.319	2.696
auto_model_Grand Cherokee	2.7074	0.729	3.712	0.000	1.278	4.137
auto_model_Forrestor	0.6337	nan	nan	nan	nan	nan
auto_model_Civic	1.0213	0.878	1.164	0.244	-0.699	2.741
auto_model_Camry	-0.4555	0.929	-0.490	0.624	-2.276	1.365
auto_model_C300	-0.3092	1.188	-0.260	0.795	-2.638	2.019
auto_model_Accord	-2.3422	1.487	-1.576	0.115	-5.256	0.571
auto_model_MDX	-0.7283	0.815	-0.893	0.372	-2.326	0.870
incident_severity_Minor Damage	-4.8508	0.430	-11.269	0.000	-5.694	-4.007
auto_model_X6	2.8459	1.115	2.553	0.011	0.661	5.031
collision_type_Side Collision	-0.9104	0.323	-2.816	0.005	-1.544	-0.277
insured_hobbies_base-jumping	1.3721	0.665	2.063	0.039	0.068	2.676
insured_occupation_transport-moving	0.2507	0.469	0.534	0.593	-0.669	1.170
insured_occupation_tech-support	0.9745	0.494	1.975	0.048	0.007	1.942
insured_occupation_sales	1.0832	0.597	1.814	0.070	-0.087	2.254
insured_occupation_prof-specialty	0.6602	0.497	1.327	0.184	-0.315	1.635
insured_occupation_exec-managerial	1.5702	0.467	3.364	0.001	0.655	2.485
insured_education_level_PhD	0.6332	0.446	1.419	0.156	-0.241	1.508
insured_education_level_High School	-0.2123	0.415	-0.511	0.609	-1.026	0.602
insured_sex_MALE	0.4663	0.276	1.689	0.091	-0.075	1.007
policy_csl_500/1000	-0.6335	0.344	-1.842	0.065	-1.308	0.041
policy_csl_250/500	0.7889	0.312	2.526	0.012	0.177	1.401
policy_state_OH	0.7797	0.331	2.354	0.019	0.131	1.429
policy_state_IN	1.2836	0.351	3.658	0.000	0.596	1.971
witnesses	0.2474	0.140	1.763	0.078	-0.028	0.522
capital-loss	-0.5572	0.134	-4.163	0.000	-0.820	-0.295
capital-gains	-0.2610	0.137	-1.909	0.056	-0.529	0.007
age	-0.9248	0.370	-2.500	0.012	-1.650	-0.200
insured_hobbies_chess	6.5219	0.707	9.228	0.000	5.137	7.907
insured_hobbies_cross-fit	5.4787	0.678	8.075	0.000	4.149	6.809
insured_education_level_Masters	0.8595	0.407	2.111	0.035	0.061	1.658
insured_hobbies_hiking	0.8521	0.565	1.509	0.131	-0.254	1.959

Logistic Regression Default Model

insured_hobbies_polo	1.4737	0.614	2.401	0.016	0.271	2.677
insured_hobbies_yachting	2.7008	0.575	4.695	0.000	1.573	3.828
incident_type_Other	0.6694	0.785	0.853	0.394	−0.869	2.208
insured_relationship_unmarried	0.6875	0.402	1.710	0.087	−0.101	1.476
insured_hobbies_exercise	−0.3408	0.543	−0.628	0.530	−1.404	0.723
insured_relationship_not-in-family	0.8318	0.355	2.341	0.019	0.135	1.528

Logistic Regression Formula

log_odds=−0.611+0.833 * property_damage_YES+1.229 * property_damage_UNKNOWN−0.801 * police_report_available_YES−0.707 * police_report_available_UNKNOWN+0.680 * authorities_contacted_Police−4.863 · incident_severity_Trivial Damage−4.743 * incident_severity_Total Loss−4.851 * incident_severity_Minor Damage+2.707 * auto_model_Grand Cherokee+2.846 * auto_model_X6−0.910 · collision_type_SideCollision+1.372 * insured_hobbies_base-jumping+0.975 · insured_occupation_tech-support+1.570 · insured_occupation_exec-managerial+0.788 · policy_csl_250/500+0.780 * policy_state_OH+1.284 · policy_state_IN−0.557 · capital-loss−0.925 · age+6.522 * insured_hobbies_chess+5.479 · insured_hobbies_cross-fit+0.860 * insured_education_level_Masters+1.474 · insured_hobbies_polo+2.701 * insured_hobbies_yachting+0.832 · insured_relationship_not-in-familylog left(fracP(Y = 1)1 - P(Y = 1) right) = -0.611 + 0.833 * property_damage_YES + 1.229 * property_damage_UNKNOWN - 0.801 * police_report_available_YES - 0.707 * police_report_available_UNKNOWN + 0.680 * authorities_contacted_Police - 4.863 * incident_severity_Trivial Damage - 4.743 * incident_severity_Total Loss - 4.851 * incident_severity_Minor Damage + 2.707 * auto_model_Grand Cherokee + 2.846 * auto_model_X6 - 0.910 * collision_type_Side Collision + 1.372 * insured_hobbies_base-jumping + 0.975 * insured_occupation_tech-support + 1.570 * insured_occupation_exec-managerial + 0.788 * policy_csl_250/500 + 0.780 * policy_state_OH + 1.284 * policy_state_IN - 0.557 * capital-loss - 0.925 * age + 6.522 * insured_hobbies_chess + 5.479 * insured_hobbies_cross-fit + 0.860 * insured_education_level_Masters + 1.474 * insured_hobbies_polo + 2.701 * insured_hobbies_yachting + 0.832 * insured_relationship_not-in-family

P(Y=1)=1/e^(-log_odds)

VIF Selected Features

auto_make_Chevrolet	incident_state_OH	auto_model_Passat	collision_type_Side Collision	policy_state_IN
auto_make_BMW	incident_state_NY	auto_model_Malibu	insured_hobbies_base-jumping	witnesses
auto_make_Audi	authorities_contacted_Police	auto_model_ML350	insured_occupation_transport-moving	capital-loss
police_report_available_YES	authorities_contacted_Other	auto_model_M5	insured_occupation_tech-support	capital-gains
police_report_available_UNKNOWN	incident_severity_Trivial Damage	auto_model_Highlander	insured_occupation_sales	insured_hobbies_chess
property_damage_YES	incident_city_Springfield	auto_model_Grand Cherokee	insured_occupation_prof-specialty	insured_hobbies_cross-fit
auto_make_Volkswagen	incident_severity_Total Loss	auto_model_Civic	insured_occupation_exec-managerial	insured_education_level_Masters
property_damage_UNKNOWN	auto_model_Wrangler	auto_model_Camry	insured_education_level_PhD	insured_hobbies_hiking
incident_city_Riverwood	auto_model_Ultima	auto_model_C300	insured_education_level_High School	insured_hobbies_polo
incident_city_Northbrook	auto_model_TL	auto_model_Accord	insured_sex_MALE	insured_hobbies_yachting
incident_city_Columbus	auto_model_Silverado	auto_model_MDX	policy_csl_500/1000	incident_type_Other
incident_state_WV	auto_model_RSX	incident_severity_Minor Damage	policy_csl_250/500	insured_relationship_unmarried
incident_state_SC	auto_model_Pathfinder	auto_model_X6	policy_state_OH	insured_hobbies_exercise
				insured_relationship_not-in-family

Trained Logistic Regression Model

Optimization terminated successfully.
Current function value: 0.287491
Iterations 8

Logit Regression Results

Dep. Variable:	Y	No. Observations:	932
Model:	Logit	Df Residuals:	864
Method:	MLE	Df Model:	67
Date:	Tue, 13 May 2025	Pseudo R-squ.:	0.5852
Time:	23:54:13	Log-Likelihood:	-267.94
converged:	True	LL-Null:	-646.01
Covariance Type:	nonrobust	LLR p-value:	2.741e-117

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5749	0.571	-1.008	0.314	-1.693	0.543
auto_make_Nissan	-0.2693	0.809	-0.333	0.739	-1.855	1.317
auto_make_Chevrolet	0.3737	0.867	0.431	0.666	-1.326	2.073
auto_make_BMW	-0.0362	0.722	-0.050	0.960	-1.452	1.380
auto_make_Audi	0.2689	0.464	0.579	0.563	-0.641	1.179
police_report_available_YES	-0.4583	0.331	-1.384	0.166	-1.107	0.191
police_report_available_UNKNOWN	-0.5136	0.317	-1.618	0.106	-1.136	0.108
property_damage_YES	0.9335	0.328	2.850	0.004	0.292	1.576
auto_make_Volkswagen	0.5011	0.654	0.767	0.443	-0.780	1.782
property_damage_UNKNOWN	1.0565	0.315	3.353	0.001	0.439	1.674
incident_city_Riverwood	-0.2802	0.425	-0.659	0.510	-1.114	0.553
incident_city_Northbrook	-0.7705	0.444	-1.737	0.082	-1.640	0.099
incident_city_Columbus	-0.7654	0.396	-1.932	0.053	-1.542	0.011
incident_state_WV	-0.8535	0.385	-2.218	0.027	-1.608	-0.099
incident_state_SC	-0.4035	0.349	-1.158	0.247	-1.087	0.280
incident_state_OH	0.2515	0.903	0.279	0.781	-1.518	2.021
incident_state_NY	-0.3630	0.354	-1.024	0.306	-1.058	0.332
authorities_contacted_Police	0.7534	0.321	2.351	0.019	0.125	1.382
authorities_contacted_Other	0.7153	0.335	2.137	0.033	0.059	1.371
incident_severity_Trivial Damage	-4.4769	0.711	-6.295	0.000	-5.871	-3.083
incident_city_Springfield	-0.0823	0.367	-0.224	0.822	-0.801	0.636
incident_severity_Total Loss	-4.5126	0.401	-11.244	0.000	-5.299	-3.726
auto_model_Wrangler	-1.7608	0.967	-1.820	0.069	-3.657	0.135
auto_model_Ultima	1.2454	1.158	1.075	0.282	-1.025	3.516
auto_model_TL	-1.0636	0.853	-1.247	0.213	-2.736	0.609
auto_model_Silverado	1.6658	1.106	1.506	0.132	-0.502	3.833
auto_model_RSX	-1.8863	1.346	-1.401	0.161	-4.525	0.752
auto_model_Pathfinder	-1.7279	1.133	-1.525	0.127	-3.948	0.493
auto_model_Passat	1.1713	0.914	1.282	0.200	-0.620	2.963
auto_model_Malibu	-0.1979	1.156	-0.171	0.864	-2.463	2.067

Trained Logistic Regression Model

auto_model_ML350	-1.6585	1.210	-1.370	0.171	-4.030	0.713
auto_model_M5	1.9560	1.145	1.709	0.088	-0.288	4.200
auto_model_Highlander	0.9363	0.670	1.398	0.162	-0.376	2.249
auto_model_Grand Cherokee	2.3971	0.646	3.713	0.000	1.132	3.662
auto_model_Civic	0.7200	0.796	0.904	0.366	-0.841	2.281
auto_model_Camry	-0.7115	0.866	-0.821	0.411	-2.409	0.986
auto_model_C300	-0.1853	1.063	-0.174	0.862	-2.269	1.898
auto_model_Accord	-2.4226	1.406	-1.723	0.085	-5.179	0.334
auto_model_MDX	-0.7722	0.728	-1.061	0.289	-2.199	0.654
incident_severity_Minor Damage	-4.6645	0.410	-11.389	0.000	-5.467	-3.862
auto_model_X6	2.7369	1.067	2.564	0.010	0.645	4.829
collision_type_Side Collision	-0.7005	0.299	-2.340	0.019	-1.287	-0.114
insured_hobbies_base-jumping	1.5450	0.611	2.527	0.011	0.347	2.743
insured_occupation_transport-moving	0.7104	0.425	1.673	0.094	-0.122	1.543
insured_occupation_tech-support	1.1560	0.458	2.523	0.012	0.258	2.054
insured_occupation_sales	0.8251	0.554	1.491	0.136	-0.260	1.910
insured_occupation_prof-specialty	0.5461	0.464	1.177	0.239	-0.363	1.455
insured_occupation_exec-managerial	1.5792	0.454	3.478	0.001	0.689	2.469
insured_education_level_PhD	0.5729	0.402	1.424	0.154	-0.215	1.361
insured_education_level_High School	-0.3426	0.401	-0.855	0.392	-1.128	0.443
insured_sex_MALE	0.3665	0.258	1.420	0.156	-0.139	0.873
policy_csl_500/1000	-0.7235	0.329	-2.202	0.028	-1.368	-0.079
policy_csl_250/500	0.6565	0.286	2.295	0.022	0.096	1.217
policy_state_OH	0.9312	0.313	2.976	0.003	0.318	1.545
policy_state_IN	1.3989	0.325	4.302	0.000	0.762	2.036
witnesses	0.1576	0.132	1.198	0.231	-0.100	0.415
capital-loss	-0.5212	0.128	-4.073	0.000	-0.772	-0.270
capital-gains	-0.2332	0.130	-1.793	0.073	-0.488	0.022
insured_hobbies_chess	5.8913	0.665	8.854	0.000	4.587	7.195
insured_hobbies_cross-fit	5.0103	0.630	7.954	0.000	3.776	6.245
insured_education_level_Masters	0.7099	0.387	1.836	0.066	-0.048	1.468
insured_hobbies_hiking	0.9091	0.535	1.701	0.089	-0.139	1.957
insured_hobbies_polo	1.1223	0.583	1.926	0.054	-0.020	2.264
insured_hobbies_yachting	2.5815	0.535	4.829	0.000	1.534	3.629
incident_type_Other	0.3554	0.751	0.473	0.636	-1.117	1.828
insured_relationship_unmarried	0.6511	0.378	1.723	0.085	-0.090	1.392
insured_hobbies_exercise	-0.5949	0.519	-1.147	0.251	-1.611	0.421
insured_relationship_not-in-family	0.6079	0.332	1.832	0.067	-0.042	1.258

Trained Logistic Regression Formula

log_odds = −0.5749 + (−0.2693 * auto_make_Nissan) + 0.3737 * auto_make_Chevrolet + (−0.0362 * auto_make_BMW) + 0.2689 * auto_make_Audi + (−0.4583 * police_report_available_YES) + (−0.5136 * police_report_available_UNKNOWN) + 0.9335 * property_damage_YES + 0.5011 * auto_make_Volkswagen + 1.0565 * property_damage_UNKNOWN + (−0.2802 * incident_city_Riverwood) + (−0.7705 * incident_city_Northbrook) + (−0.7654 * incident_city_Columbus) + (−0.8535 * incident_state_WV) + (−0.4035 * incident_state_SC) + 0.2515 * incident_state_OH + (−0.3630 * incident_state_NY) + 0.7534 * authorities_contacted_Police + 0.7153 * authorities_contacted_Other + (−4.4769 * incident_severity_Trivial Damage) + (−0.0823 * incident_city_Springfield) + (−4.5126 * incident_severity_Total Loss) + (−1.7608 * auto_model_Wrangler) + 1.2454 * auto_model_Ultima + (−1.0636 * auto_model_TL) + 1.6658 * auto_model_Silverado + (−1.8863 * auto_model_RSX) + (−1.7279 * auto_model_Pathfinder) + 1.1713 * auto_model_Passat + (−0.1979 * auto_model_Malibu) + (−1.6585 * auto_model_ML350) + 1.9560 * auto_model_M5 + 0.9363 * auto_model_Highlander + 2.3971 * auto_model_Grand Cherokee + 0.7200 * auto_model_Civic + (−0.7115 * auto_model_Camry) + (−0.1853 * auto_model_C300) + (−2.4226 * auto_model_Accord) + (−0.7722 * auto_model_MDX) + (−4.6645 * incident_severity_Minor Damage) + 2.7369 * auto_model_X6 + (−0.7005 * collision_type_Side Collision) + 1.5450 * insured_hobbies_base-jumping + 0.7104 * insured_occupation_transport-moving + 1.1560 * insured_occupation_tech-support + 0.8251 * insured_occupation_sales + 0.5461 * insured_occupation_prof-specialty + 1.5792 * insured_occupation_exec-managerial + 0.5729 * insured_education_level_PhD + (−0.3426 * insured_education_level_High School) + 0.3665 * insured_sex_MALE + (−0.7235 * policy_csl_500/1000) + 0.6565 * policy_csl_250/500 + 0.9312 * policy_state_OH + 1.3989 * policy_state_IN + 0.1576 * witnesses + (−0.5212 * capital-loss) + (−0.2332 * capital-gains) + 5.8913 * insured_hobbies_chess + 5.0103 * insured_hobbies_cross-fit + 0.7099 * insured_education_level_Masters + 0.9091 * insured_hobbies_hiking + 1.1223 * insured_hobbies_polo + 2.5815 * insured_hobbies_yachting + 0.3554 * incident_type_Other + 0.6511 * insured_relationship_unmarried + (−0.5949 * insured_hobbies_exercise) + 0.6079 * insured_relationship_not-in-family

P(Y=1)=1/e^(-log_odds)

Metrics for the cutoff 0.5

Metrics

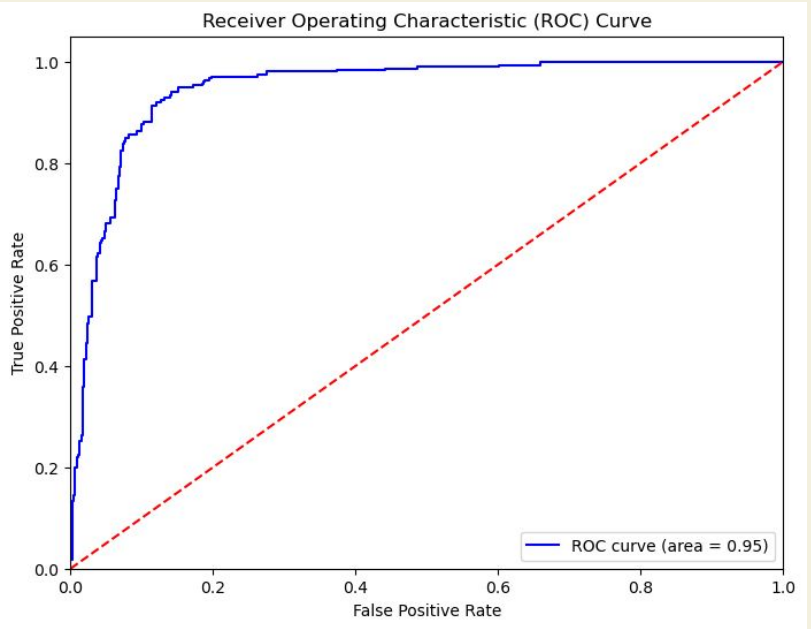
Accuracy	0.83
Sensitivity	0.92
Specificity	0.86
Precision	0.87
Recall	0.92
F1 score	0.90

Confusion Matrix

TP	433
TN	404
FP	62
FN	33

ROC Curve

- **True Positive Rate (TPR)** is plotted against **False Positive Rate (FPR)**.
- The **blue curve** represents the model's performance.
- The **red dashed line** represents a random classifier (baseline).
- The **Area Under the Curve (AUC)** is **0.95**, which indicates **excellent model performance** — the model can distinguish between classes very well.



Model Performance at Different Cutoff

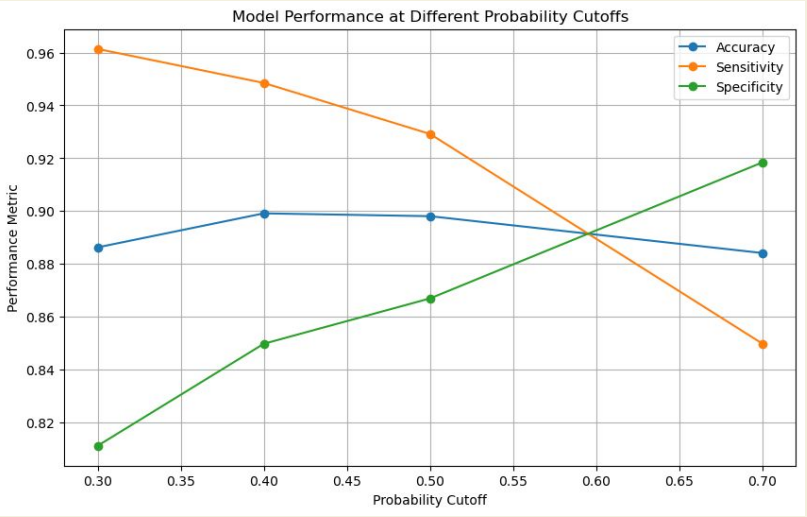
Lower cutoff (like 0.30):

- Catches more positives (**high sensitivity**).
- Makes more mistakes with negatives (**low specificity**).

Higher cutoff (like 0.70):

- Catches fewer positives (**low sensitivity**).
- Does better with negatives (**high specificity**).

Accuracy stays mostly the same around 88–90%.



Metrics for the cutoff 0.3

Metrics

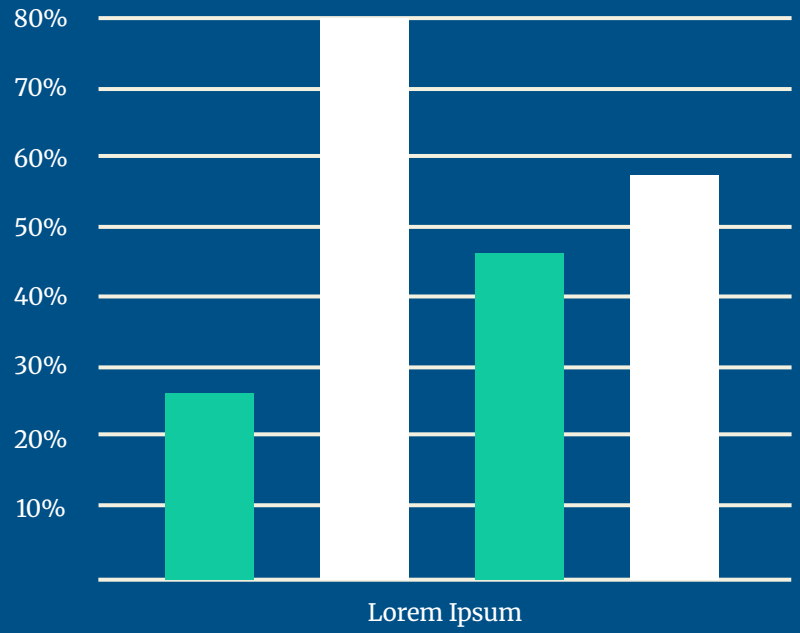
Accuracy	0.88
Sensitivity	0.96
Specificity	0.81
Precision	0.83
Recall	0.96
F1 score	0.89

Confusion Matrix

TP	448
TN	378
FP	88
FN	18

1. Defining the Problem
2. Making Observations
3. Forming a Hypothesis
4. Experiment Results
5. **Drawing a Conclusion**

Model Building Random Forest



Random Forest Model

- - Get Feature Importances - Obtain the importance scores for each feature and select the important features to train the model.
- - Model Evaluation on Training Data – Assess performance metrics on the training data.
- - Check Model Overfitting using Cross-Validation – Evaluate generalisation by performing cross-validation.
- - Hyperparameter Tuning using Grid Search – Optimise model performance by fine-tuning hyperparameters.
- - Final Model and Evaluation on Training Data – Train the final model using the best parameters and assess its performance.

Important Feature for Threshold 0.01

months_as_customer	insured_hobbies_chess
age	insured_hobbies_cross-fit
policy_deductable	incident_severity_Minor Damage
umbrella_limit	incident_severity_Total Loss
capital-gains	insured_hobbies_chess
capital-loss	insured_hobbies_cross-fit
incident_hour_of_the_day	incident_severity_Minor Damage
bodily_injuries	
witnesses	
total_claim_amount	
injury_claim	
property_claim	
vehicle_claim	
auto_year	

Metrics for the Threshold 0.01

Metrics

Accuracy	1.0
Sensitivity	1.0
Specificity	1.0
Precision	1.0
Recall	1.0
F1 score	1.0

Confusion Matrix

TP	466
TN	466
FP	0
FN	0

Hyperparameter Tuning

`max_depth: None`

- The trees in the forest can grow **as deep as they need**.

`min_samples_leaf: 1`

- A leaf (end of a tree branch) can have **just 1 data point**.

`min_samples_split: 5`

- A tree branch will only split if there are **at least 5 data points** in that part of the tree.

`n_estimators: 200`

- You're using **200 decision trees** in the forest.

Metrics for the Hyperparameter Tuning

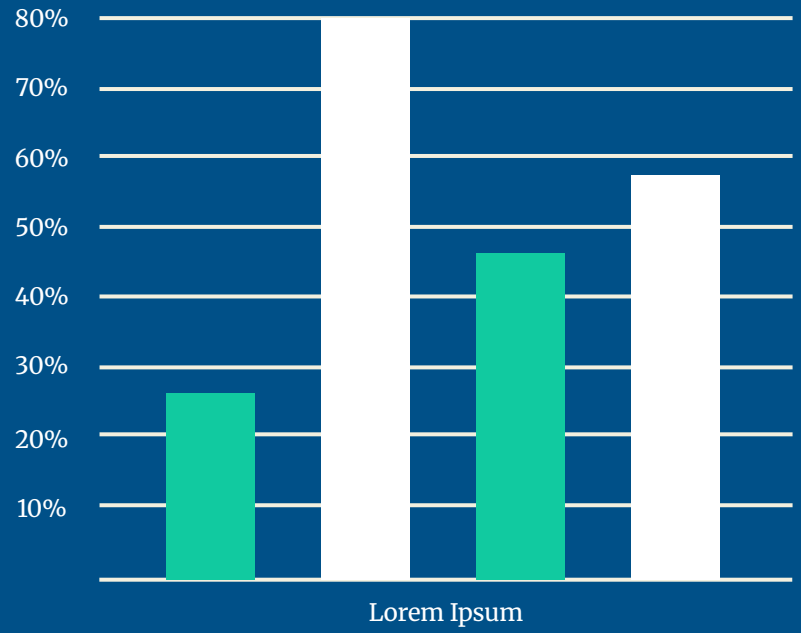
Metrics

Accuracy	0.83
Sensitivity	0.99
Specificity	0.1
Precision	0.1
Recall	0.99
F1 score	0.99

Confusion Matrix

TP	463
TN	466
FP	0
FN	3

Model Evaluation



Metrics for Logistic Regression

Metrics

Accuracy	0.73
Sensitivity	0.75
Specificity	0.72
Precision	0.49
Recall	0.75
F1 score	0.59

Confusion Matrix

TP	54
TN	146
FP	55
FN	18

Metrics for Random Forest

Metrics

Accuracy	0.83
Sensitivity	0.79
Specificity	0.84
Precision	0.64
Recall	0.79
F1 score	0.71

Confusion Matrix

TP	57
TN	170
FP	31
FN	15

Drawing a Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur rhoncus nibh ut odio tempor elementum. Praesent et mattis dolor.

How can we analyse historical claim data to detect patterns that indicate fraudulent claims?

- **Exploratory Data Analysis (EDA)**:
  - Initial patterns are identified by visualizing data. This includes: Comparing the frequency of categorical feature values (e.g., **`incident_type`**, **`collision_type`**, **`incident_severity`**) between fraudulent and non-fraudulent claims. A higher proportion of fraud for a specific category (e.g., 'Major Damage' for **`incident_severity`**) indicates a pattern.

  - Examining distributions of numerical features (e.g., **`total_claim_amount`**, **`age`**, **`number_of_vehicles_involved`**, **`witnesses`**) for fraudulent versus non-fraudulent claims using box plots. Significant differences in medians or ranges (e.g., higher claim amounts or fewer witnesses in fraudulent claims) highlight patterns.

- **Feature Engineering**:
  - New features are created to uncover more complex patterns: Temporal features like **`policy_age`** .
  - A pattern might be that newer policies are more frequently associated with fraud. Ratio features like **`injury_claim_ratio`** (proportion of total claim from injury). An unusually high ratio for a minor incident type could be a pattern.

- **Machine Learning Model Interpretation: Models learn and quantify these patterns**:
  - Logistic Regression: Features selected by RFECV and those with significant p-values and influential coefficients in the model summary (e.g., a high positive coefficient for a dummy variable like **`incident_severity_Major Damage`**) are identified as components of fraudulent patterns.
  - Random Forest: The model provides feature importance scores. Features with high importance (e.g., **`total_claim_amount`**, **`incident_severity`**, **`policy_age`**) are key elements in the patterns the model uses to distinguish fraudulent claims

Which features are most predictive of fraudulent behaviour?

Based on the analysis, the features most predictive of fraudulent behavior are identified by both Logistic Regression (via RFECV and statistical significance) and Random Forest (via feature importance scores).

Considering the Random Forest model generally performed better, its feature importances are particularly insightful. The key predictive features include:

- Incident-Specific Features:
  - **incident_severity**: Different levels (Major Damage, Minor Damage, Total Loss, Trivial Damage) are highly indicative.
  - **witnesses**: The number of witnesses.
  - **incident_hour_of_the_day**: The time of the incident.
  - **number_of_vehicles_involved**.
  - **bodily_injuries**.
- Claim Amount Features:
  - **total_claim_amount**: The overall amount claimed.
  - **injury_claim, property_claim, vehicle_claim**: The breakdown of the claim amount.
  - Engineered ratios like **injury_claim_ratio**, **property_claim_ratio**, **vehicle_claim_ratio**.

Which features are most predictive of fraudulent behaviour?

- Policy and Customer Features:
    i. **policy_age**: An engineered feature representing the policy's age at the time of the incident.
    ii. **months_as_customer**: Duration the customer has been with the insurer.
    iii. **age**: Age of the insured.
    iv. **policy_deductable**.
    v. **umbrella_limit**.
    vi. **capital-gains and capital-loss**.
- Vehicle Features:
    i. **auto_year**: The manufacturing year of the vehicle.

The Logistic Regression model, after RFECV and VIF filtering, also highlighted many of these, such as months_as_customer, policy_deductable, umbrella_limit, capital-gains, various incident_severity levels, witnesses, and property_claim_ratio.

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

1. **High-Impact Features**: The model highlights strong indicators of fraud, such as incident_severity, total_claim_amount, policy_age (engineered), number of witnesses, and capital-gains/loss. This enables investigators to prioritize reviews based on the most critical factors.

2. **High-Risk Scenarios**: By examining certain claim types—like those involving ‘Trivial Damage’ or cases where no police were contacted—that frequently appear in fraudulent activity, the company can identify recurring fraud patterns and flag suspicious claims more quickly.

3. **Implement Automated Prioritization**: The model assigns each claim a fraud score. High-risk claims are escalated to investigators, while low-risk claims are processed faster. This improves efficiency and reduces manual workload.

Conclusion

Both a Logistic Regression model and a Random Forest model were developed to classify insurance claims as fraudulent or legitimate. The performance of each model was evaluated on a validation dataset.

Logistic Regression Model:

- Feature selection was performed using RFECV, and multicollinearity was assessed using VIF.
- The optimal probability cutoff was determined to be **0.3**.
- On the validation data, the Logistic Regression model achieved an **accuracy** of **0.73**, a **sensitivity** of **0.75**, a **specificity** of **0.72**, a **precision** of **0.49**, and an **F1-score** of **0.59**.

Random Forest Model:

- Feature importance scores were used to select the most relevant features.
- Hyperparameter tuning was performed using Grid Search.
- On the validation data, the tuned Random Forest model achieved an **accuracy** of **0.83**, a **sensitivity** of **0.79**, a **specificity** of **0.84**, a **precision** of **0.64**, and an **F1-score** of **0.71**.

The Random Forest model performed slightly better than the Logistic Regression model on the validation data, as indicated by the higher accuracy and F1-score. Both models show reasonable performance in classifying fraudulent claims.