FRAUDULENT CLAIM DETECTION

Case Study Report

Gokul Narayanan

Saurabh Purohit

May 14 2025

Problem Statement:

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 minimize financial losses and optimize the overall claims handling process.

Following procedure were done to handle the data and create a model out of it

Data Preparation and Cleaning

1) Checked for missing values in each column
2) Handled rows containing null values. "authorities_contacted" had 91 entries of None which was wrongly considered as NAN. It was changed to None contacted
3) Column _c39 was dropped as it was completely empty
4) auto_make and auto_year was dropped as auto_model can give insights around the same data
5) Data types of policy_bind_date and incident_data was fixed and converted to datetime type

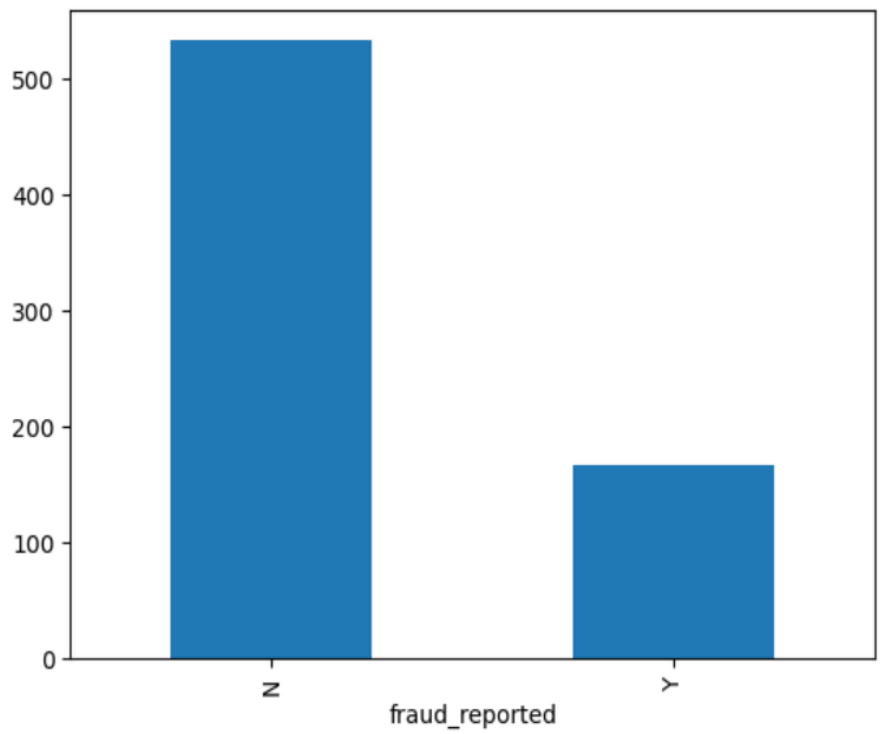
Train and validation Data split

70 percent of data is used for training and 30 percent of data is kept for validation

Exploratory Data Analysis on Training Data

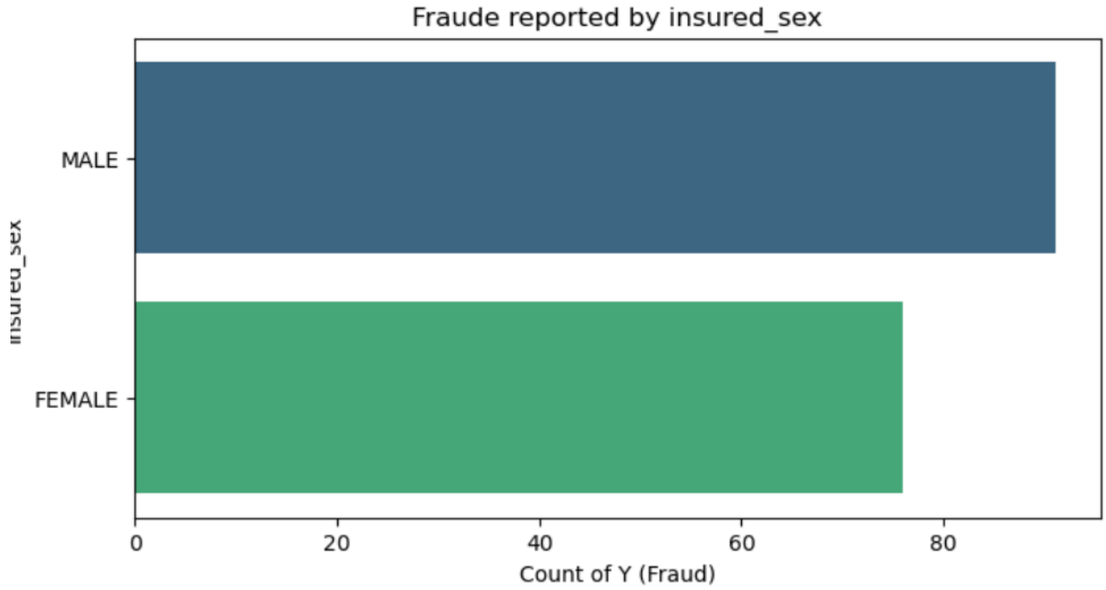
- Univariate analysis on numeric features didn't show up any outliers

- Correlation analysis by plotting heat map on numeric features showed high correlation between vehicle_claim and total_claim_amount. Same is seen for injury_claim and propert_claim. Another highly correlated variables are age and month_as_customer
- Class imbalance data was plotted for target variable "fraud_reported"

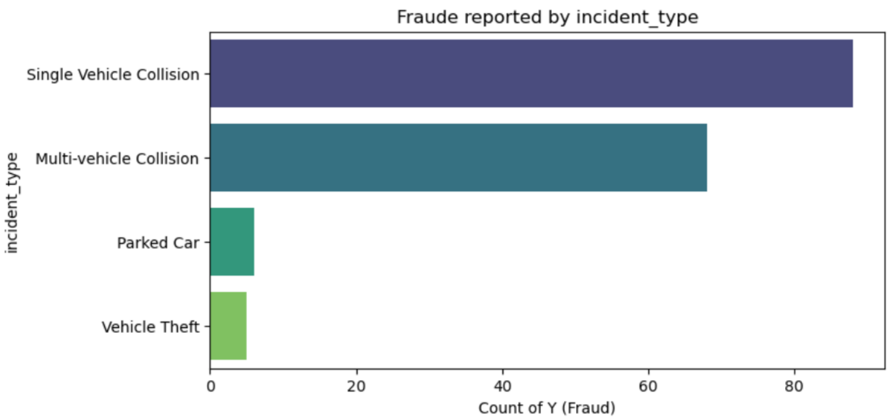


- Bivariate analysis was performed on categorical variables and the target variable

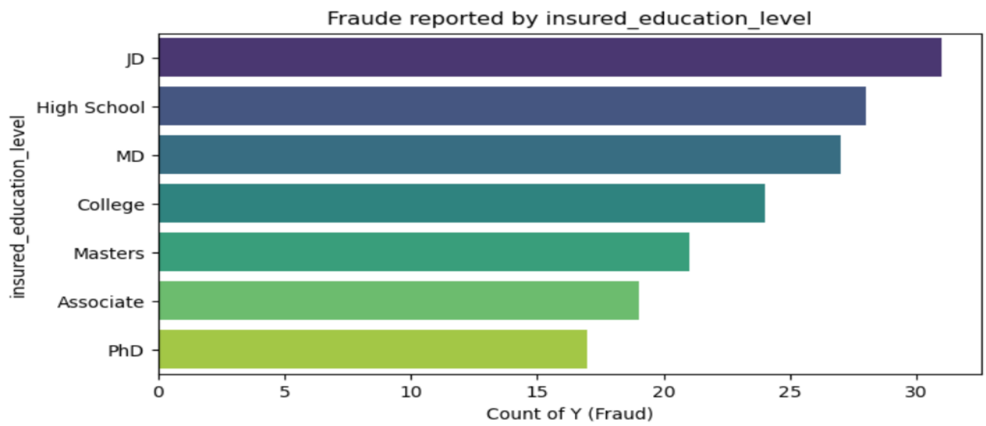
    1) Male seems to report more fraud cases than female



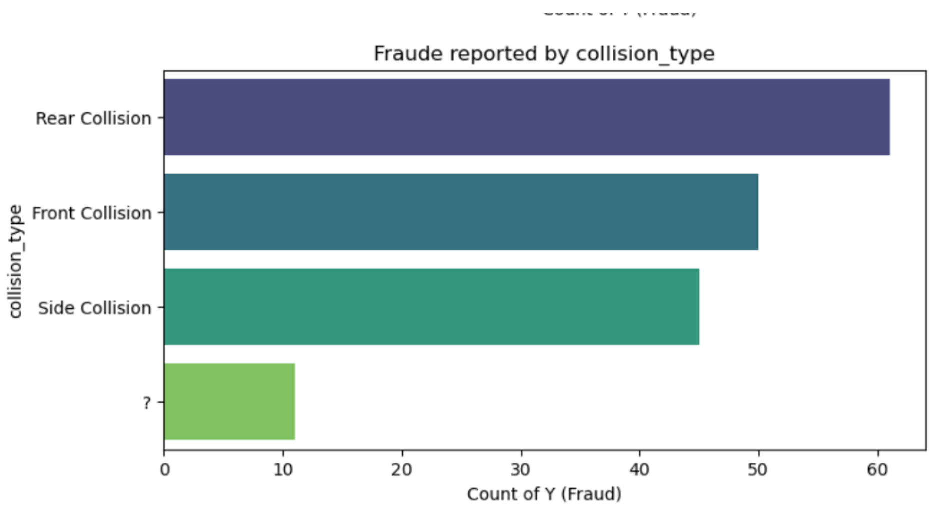
2) In incident_type Single vehicle collision seems to be contributing high for fraud cases



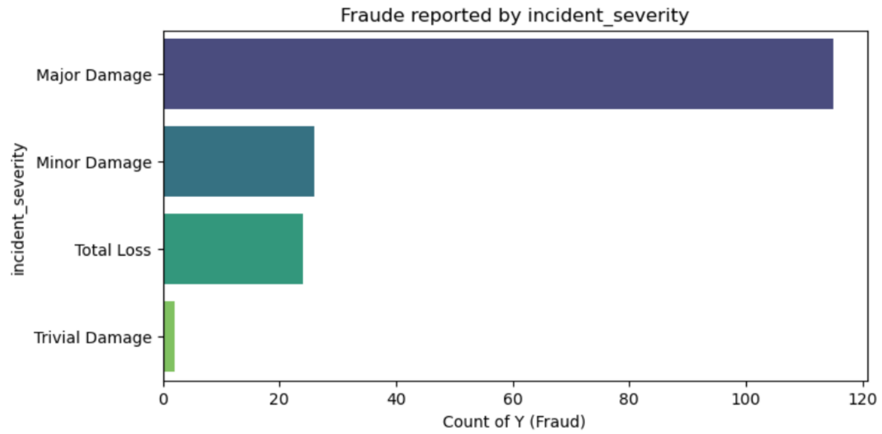
3) In educational level JD seems to be having number of frauds reported compared to other educational qualifications



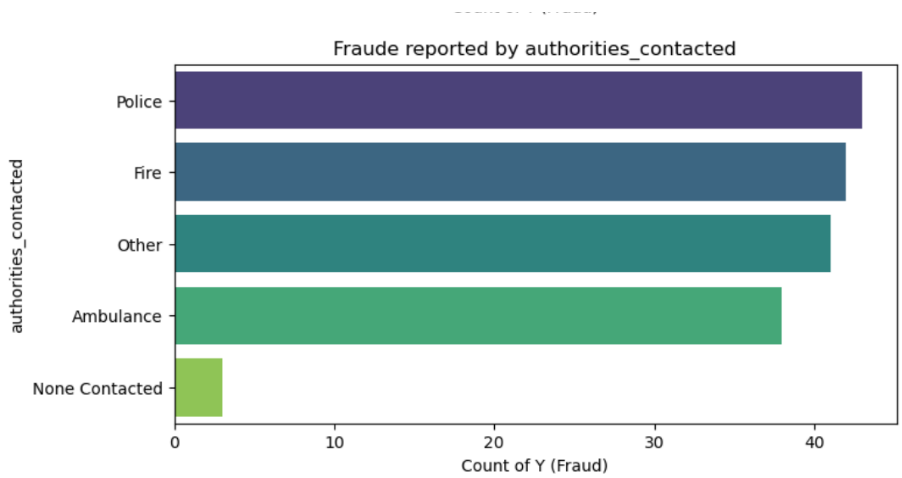
4) Fraud reported by collision_type Rear collision seems to report more to the fraud cases



5) In incident_severity Major Damage seems to contribute more to fraud cases compared to Minor Damage, Total Loss and Trivial damage



6) Police reported seems to make no significant change compared to other authorities contacted like Fire, Ambulance etc in fraudulent claims



Note ** Same EDA techniques is performed on Validation data as well

Feature Engineering

- **RESAMPLING**

Random oversampling technique is used to address the class imbalance issue. This method increases the number of samples in the minority class by randomly duplicating them, creating synthetic data points with similar characteristics

```
Before resampling value
 fraud_reported
N                 533
Y                 167
```

```
After resampling value
 fraud_reported
N                 533
Y                 533
Name: count, dtype: int64
```

- **Feature Creation**

  New features were created from incident_day by extracting the incident_date value . Also, capital_net_gain feature was created from capital-gains-capital-loss

- **Redudant column handling**

  Following features were removed due to non-significance , new feature creation and multi collinearity

  policy_number

  incident_date

  age

  injury_claim

  property_claim

  vehicle_claim

  policy_bind_date

insured_educational_level

insured_occupation

insured_hobbies

insured_relationship

incident_location

incident_state

capital-loss

capital-gain

- **Combining values in categorical columns**
  Refining categorical features by grouping values that have low frequency.

  Incident_severity Trivial Damage and Total loss is respectively mapped to Minor Damage and Major Damage

  Changed the ? values in collision_type and police_report_available to a logical value

  Changed YES or NO in property_damage and police_report_available to values for better dummy variable creation

- **Dummy Variable Creation in both training and validation data**
  Dummy variables are created for both training and validation data for below categorical variables

  policy_state
  policy_csl
  insured_sex
  incident_type
  collision_type
  incident_severity
  authorities_contacted
  incident_city

incident_hour_of_the_day
property_damage
police_report_available
auto_model
incident_day

- **Feature Scaling**

  MinMaxScaler is used to scale the features to a common range to prevent features with larger values from dominating the model

	months_as_customer	policy_deductable	policy_annual_premium
count	300.000000	300.000000	300.000000
mean	0.398636	0.414444	0.509092
std	0.240311	0.400314	0.143318
min	0.002088	0.000000	0.147801
25%	0.210856	0.000000	0.411362
50%	0.386221	0.333333	0.512848
75%	0.551670	1.000000	0.595353
max	1.000000	1.000000	0.930779

	umbrella_limit	insured_zip	number_of_vehicles_involved
count	300.00000	300.000000	300.000000
mean	0.11200	0.385036	0.288889
std	0.23188	0.375836	0.343040
min	-0.10000	0.002710	0.000000
25%	0.00000	0.114735	0.000000
50%	0.00000	0.200047	0.000000
75%	0.00000	0.904157	0.666667
max	1.00000	1.000488	1.000000

	bodily_injuries	witnesses	total_claim_amount	capital_net_gain	...
count	300.000000	300.000000	300.000000	300.000000	...
mean	0.485000	0.530000	0.434998	0.259250	...
std	0.411712	0.351548	0.232911	0.208271	...
min	0.000000	0.000000	0.018725	0.000000	...
25%	0.000000	0.333333	0.303519	0.000000	...
50%	0.500000	0.666667	0.489288	0.253125	...
75%	1.000000	0.666667	0.593102	0.394661	...
max	1.000000	1.000000	0.910207	0.798438	...

	Ultima	Wrangler	X5	X6	Monday	Saturday	\
count	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000	
mean	0.030000	0.030000	0.016667	0.016667	0.126667	0.173333	
std	0.170872	0.170872	0.128233	0.128233	0.333155	0.379168	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

	Sunday	Thursday	Tuesday	Wednesday
count	300.000000	300.000000	300.000000	300.000000
mean	0.116667	0.156667	0.123333	0.153333
std	0.321559	0.364094	0.329369	0.360911
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000

MODEL BUILDING AND EVALUATION

CONCLUSION