Fraud Claim Detection

Problem Statement

The problem

Insurance fraud detection currently depends on manual time-consuming processes that delay investigation, miss fraudulent activity and subject genuine claims to unnecessary scrutiny.

These inefficiencies result in financial losses, poor customer experience and suboptimal use of resources

Business goals

- Enhance fraud detection using historical insurance claim data
- Identify patterns and indicators that distinguish fraudulent and genuine claims
- Develop a predictive model to proactively assess fraud risk and reduce financial losses

Data Overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 40 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   months_as_customer           1000 non-null   int64
 1   age                          1000 non-null   int64
 2   policy_number                1000 non-null   int64
 3   policy_bind_date             1000 non-null   object
 4   policy_state                 1000 non-null   object
 5   policy_csl                   1000 non-null   object
 6   policy_deductable            1000 non-null   int64
 7   policy_annual_premium        1000 non-null   float64
 8   umbrella_limit               1000 non-null   int64
 9   insured_zip                  1000 non-null   int64
 10  insured_sex                  1000 non-null   object
 11  insured_education_level      1000 non-null   object
 12  insured_occupation           1000 non-null   object
 13  insured_hobbies              1000 non-null   object
 14  insured_relationship         1000 non-null   object
 15  capital-gains                1000 non-null   int64
 16  capital-loss                 1000 non-null   int64
 17  incident_date                1000 non-null   object
```

Total Records: 1000 claims

Features: 40 columns

Target Variable: fraud_reported (Y = fraud, N = no fraud)

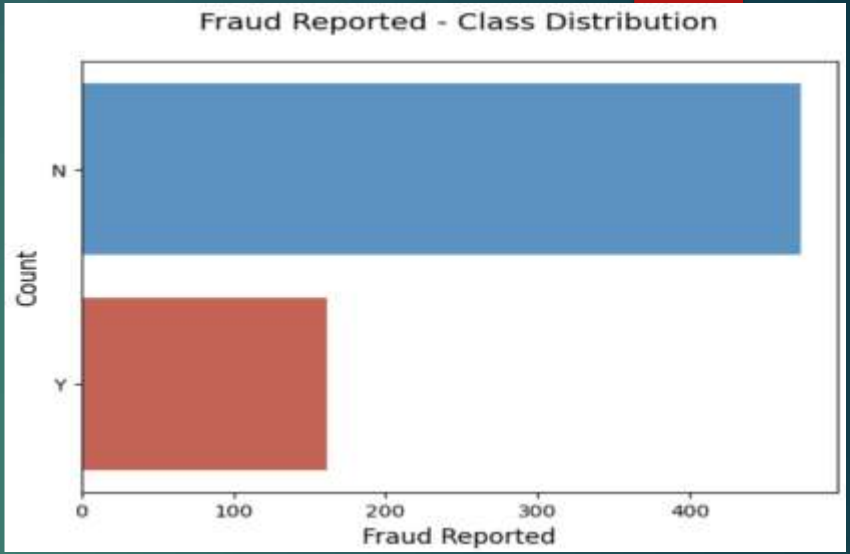
Data Challenges

Highly correlation amongst features

Imbalanced distribution between fraud and genuine claims

Presence of redundant and uninformative feature

Missing values in multiple columns



Data Preparation

Preparing Data

Handling of missing values

Dropping identifier columns

Cleaning and validating numerical data

Converting data to correct types

Feature Engineering

Oversampled fraud claim to balance the distribution of the existing minority class

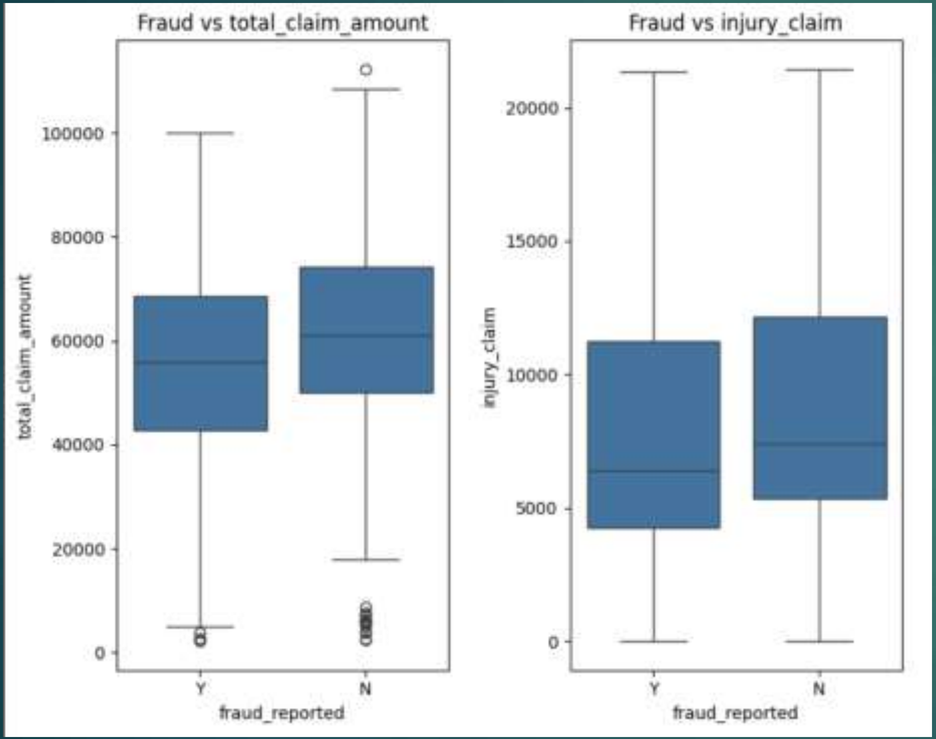
Created new features from existing features

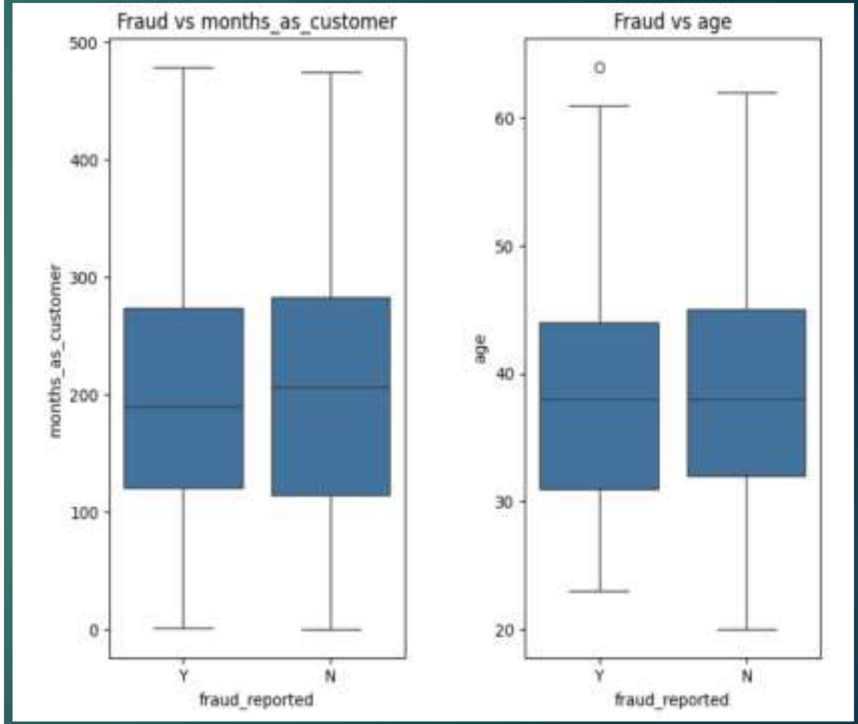
Created dummy variables for categorical features

Standardized numerical features

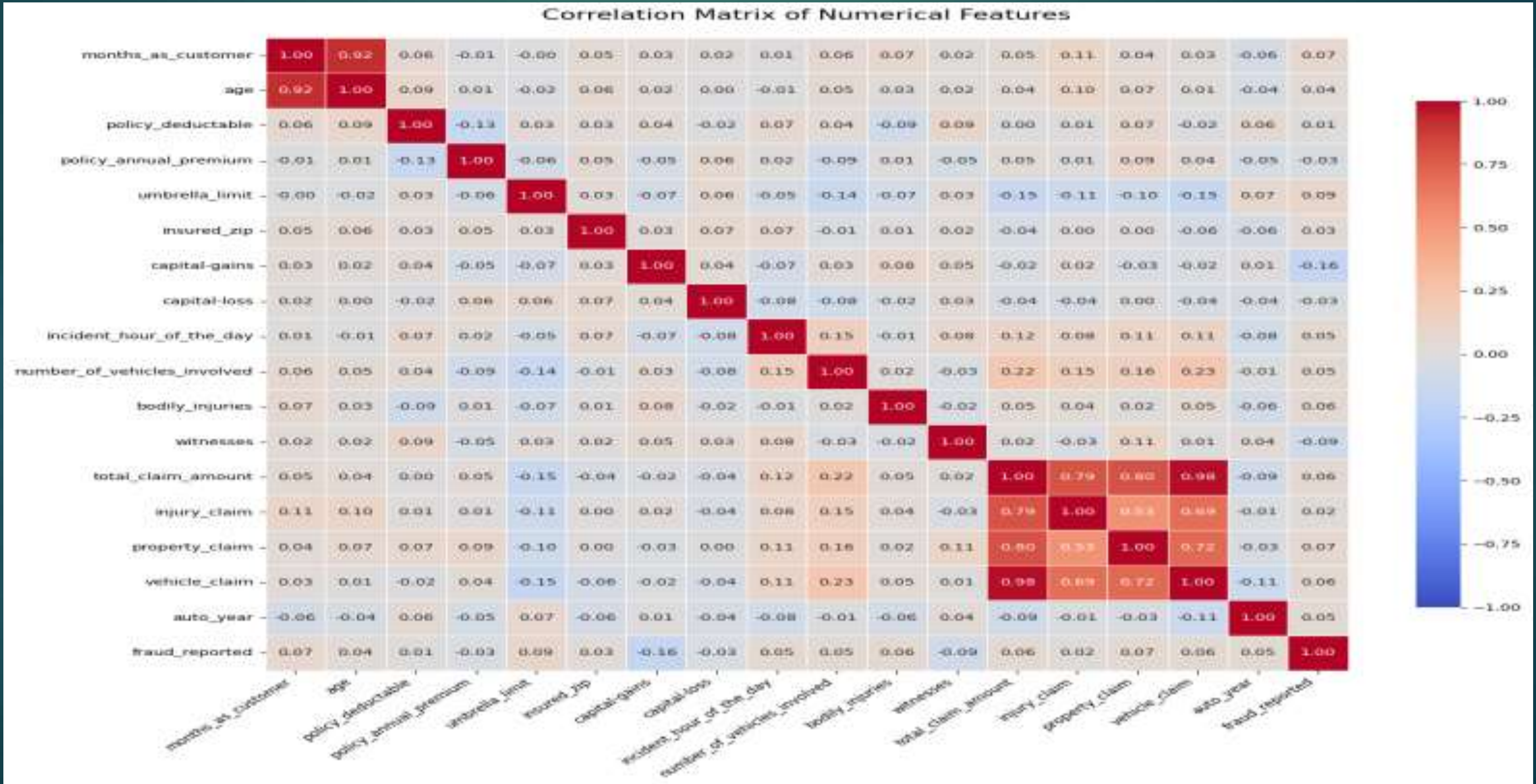
Exploratory Data Analysis

Numerical values

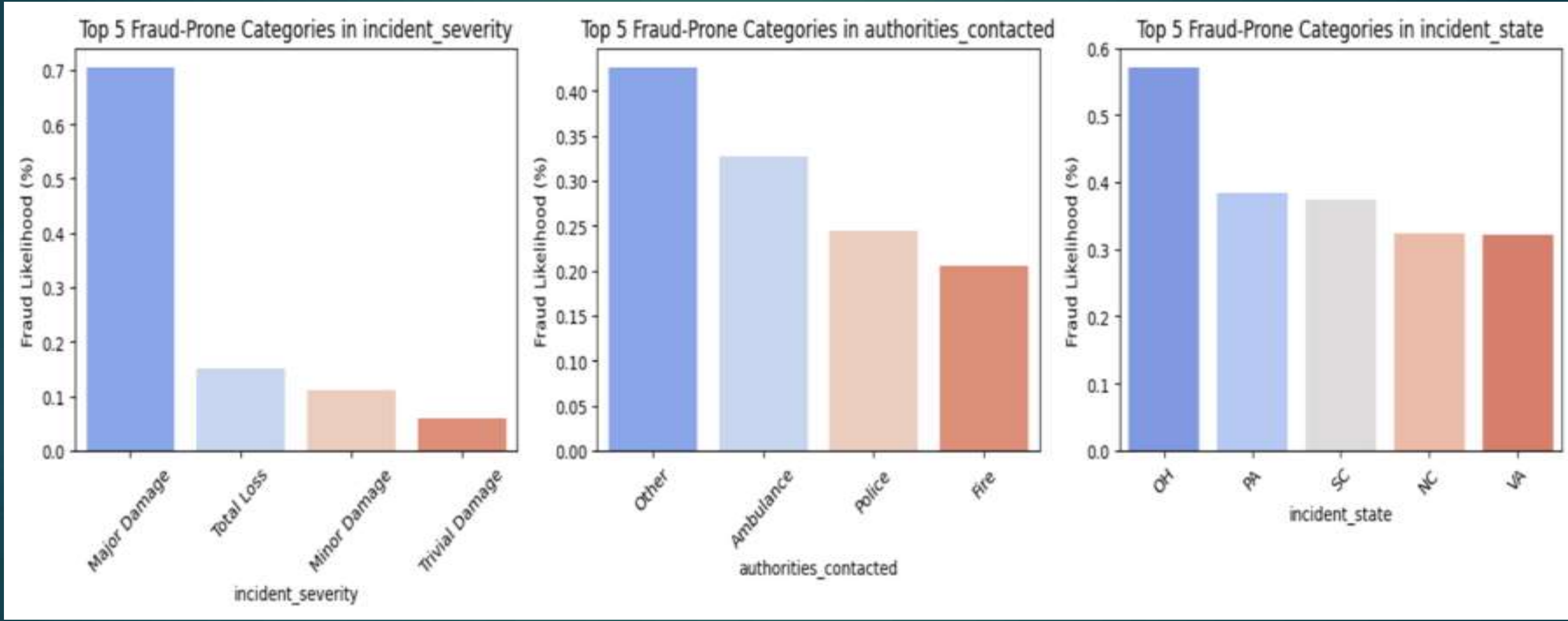




Correlation



Categorical values



The overall fraud reported is 25.51% with a moderate imbalance between fraudulent claims and genuine claims

Total claim amount is a useful predictor for fraud. However, it is highly correlated with property claim, injury claim and vehicle claim indicating that ratios might be more useful than the sub claim components of total claim amount.

Months as a customer and age are interchangeable due to high correlation

Incident hour, insured zip are low impact features

Incident severity, incident type, education level, occupation, auto make are the most predictive categorical features

Model Development

Chosen Model - Logistic Regression

Both Random Forest and Logistic Regression performanced comparably well.

Selected Model: Logistic Regression

82.3% recall which indicates that it is good at catching fraud cases

89.2% specificity which indicates that it is good at rejecting non-fraud cases

78.8% f1-score indicating an overall balanced performance

Metric	Logistic Regression	Random Forest
Sensitivity (Recall)	0.823	0.835
Specificity	0.892	0.876
Precision	0.756	0.733
F1 Score	0.788	0.781

Since both models deliver comparable performance, Logistic Regression is the preferred choice due to its greater interpretability and faster training time

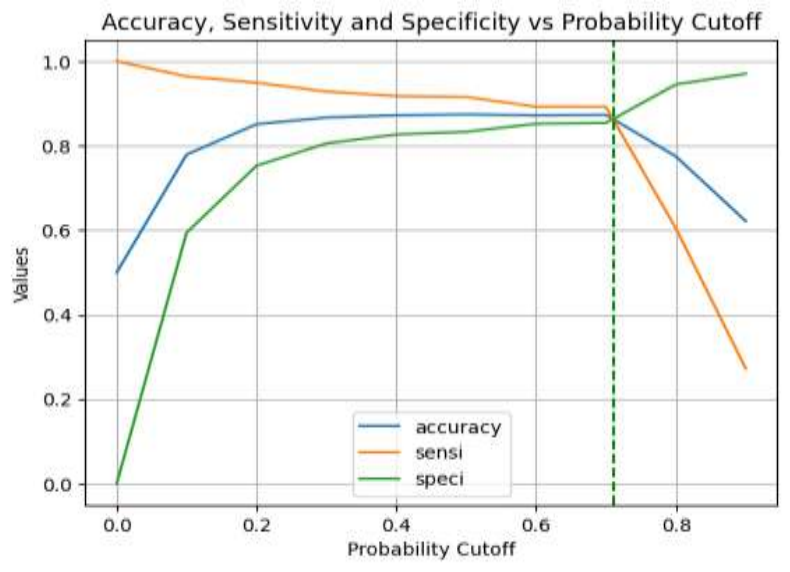
Optimization

Threshold cutoff 0.7

At a threshold of 0.7, the model strikes a balance between recall and specificity, whilst maintaining a high accuracy

A higher threshold ensures that only high-confidence fraud cases are classified as fraud

This reduces false positives



Model Performance

Tests with 273 claims

65 Fraud claims correctly flagged

14 Fraud cases missed

173 legitimate claims correctly flagged

21 legitimate claims incorrectly flagged

These results indicate the model is effective at detecting fraud while maintaining a low positive rate, balancing precision and recall well

Conclusion

Recommendations

Deploy model for use with real claim data

Continuously monitor model performance and results

Next Steps

Continuously improve model through updating it periodically

Explore more advanced models for improved accuracy and insights

Questions?

How can we analyse historical claim data to detect patterns of fraud?

Historical claim data was subjected to exploratory data analysis to uncover underlying patterns and anomalies. Steps taken included univariate analysis, bivariate analysis, target likelihood analysis and correlation analysis.

These steps combined with data cleaning and feature engineering enabled us to detect patterns that distinguish fraudulent claims from legitimate claims

Features that are the most predictive of fraud?

- incident_occurred_in_state
- insured_occupation_handlers-cleaners
- insured_occupation_priv-house-serv
- insured_hobbies_board-games
- insured_hobbies_camping
- insured_hobbies_chess
- insured_hobbies_cross-fit
- insured_hobbies_polo
- insured_hobbies_reading
- insured_relationship_other-relative
- incident_severity_Minor Damage
- incident_severity_Total Loss
- incident_severity_Trivial Damage
- incident_state_PA
- age_category_young

These features were retained based on their strong relationship with the fraud outcome observed during model training. Some reflect behavioral patterns while others emerged as key predictors contributing to the models performance

Can we predict likelihood of fraud of incoming claim?

The model confirms that historical data can be used to predict the likelihood of fraud in incoming claims. It estimates the probability of a claim being fraudulent, allowing for informed decision-making. Through threshold tuning, the model balances false positives and false negatives ensuring the predictions align with the organisation’s risk tolerance.

Insights that can be drawn from the model that can help improve the fraud detection process

The analysis of historical claim data combined with feature engineering and model-based selection not only provides a robust mechanism to predict fraud likelihood for incoming claims but also yields valuable insights. These insights guide improvements in both the operational workflow and strategic decision making helping reduce financial losses from fraudulent claims

Thank you

By,
Pradeep Singh Jethodi
Lebohang Danster
Kankanala Saisudheer