

Fraudulent claim detection case study


Submitted by:

1. Pavithra Sri S

Date: 13.04.2025



Contents

1. Problem statement and Objective
 2. Solution approach
 3. Dataset overview and Manipulations
 4. Exploratory data analysis
 5. Model Building
 6. ROC curve and optimal cutoff.
 7. Evaluation metrics and observation
 8. Business recommendations and conclusion
- 



Problem statement

Global Insure faces significant financial losses due to fraudulent insurance claims, which are currently identified through manual, time-consuming inspections—often after payouts. To improve efficiency and reduce losses, the company aims to implement a data-driven model that can classify claims as fraudulent or legitimate early in the approval process.

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.

Solution approach



Data Preparation:

- Import Libraries.
- Load the data.

Data cleaning:

- Handle null values.
- Identify and handle redundant values and columns.
- Fix Data Types.

Train-Validation Split:

- Define feature and target variables
- Split the data

EDA on training data:

- Perform univariate and bivariate analysis.
- Perform univariate analysis.

Feature Engineering:

- Perform resampling and feature Creation
- Combine values in Categorical Columns
- Dummy variable creation and Feature scaling.

Model Building:

- Build Logistic Regression Model and find the optimal cutoff.
- Build Random Forest Model and hyper parameter tuning.

Prediction and Model Evaluation

- Make predictions over validation data using logistic regression model and random forest model.
- Comparison of model.

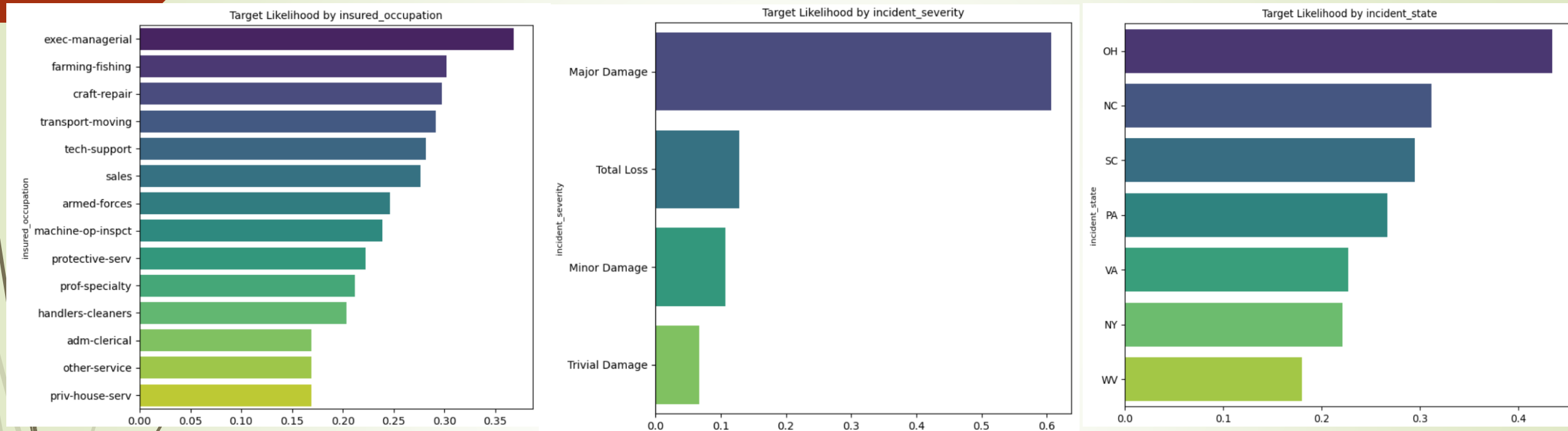
Dataset overview and Manipulation

- Data set has 40 columns and 1000 rows initially with categorical and numeric data.
- Target Variable: Fraudulent vs. Legitimate claim.
- Identify and drop any columns that are completely empty and have illogical or invalid values like '_c39'.
- Removed columns where a large proportion of the values are unique or near-unique like 'policy_number', 'insured_zip', 'incident_location'.
- Columns 'policy_bind_date' and 'incident_date' is convert to date time data type from Object.
- Columns that exhibit high correlation is removed like age and months_as_customer.
- Performed Resampling technique to tackle calss imbalance - [('N', 526), ('Y', 526)]
- Year and month extraction from column 'policy_bind_date' and binnig of incident_hour_of_the_day.
- Grouped low-frequency categories under Single category for the columns like 'auto_model' etc. Numerical columns are standardized using Standard scaler().
- Log transformation on column 'Umbrella limit'.
- Dummy variables are created for easy interpretation and analysis.

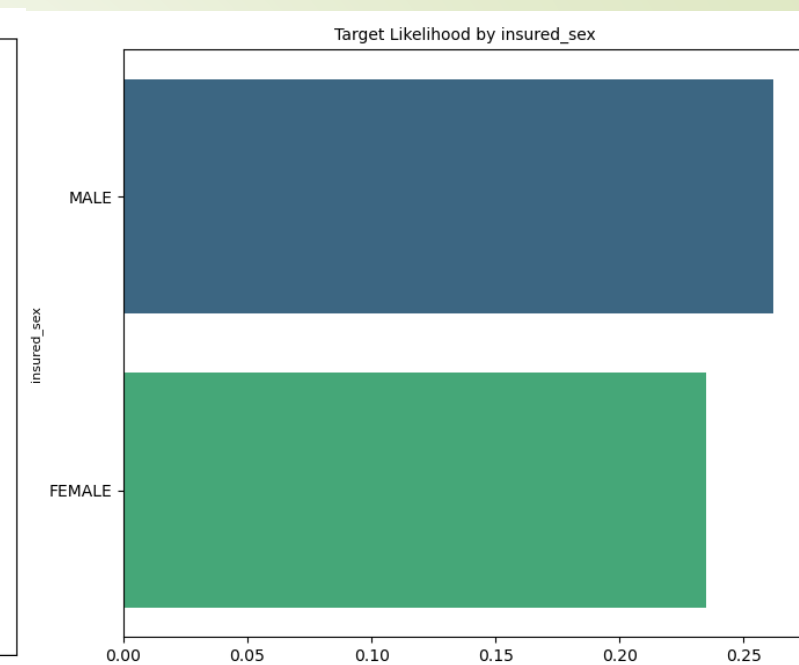
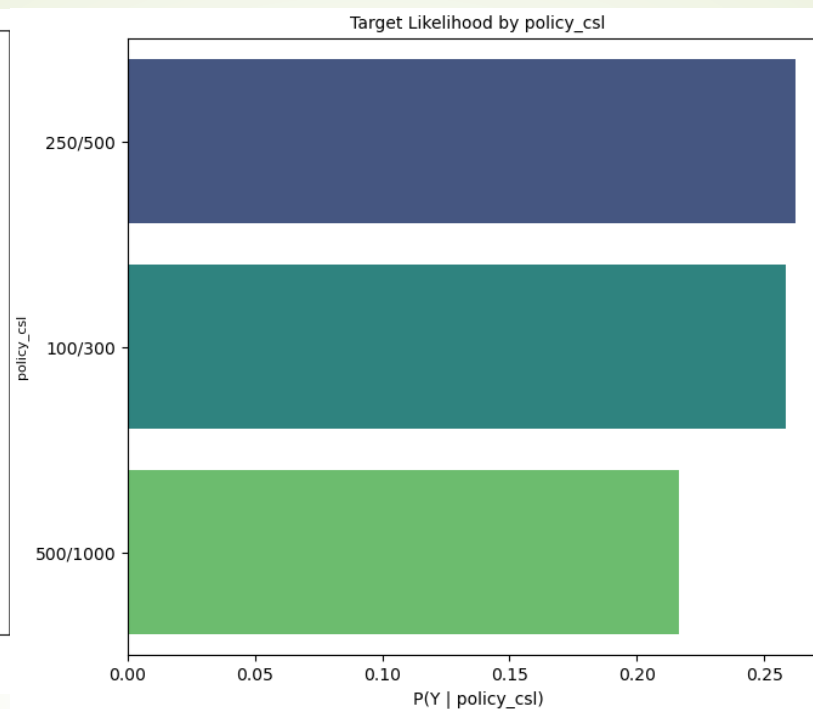
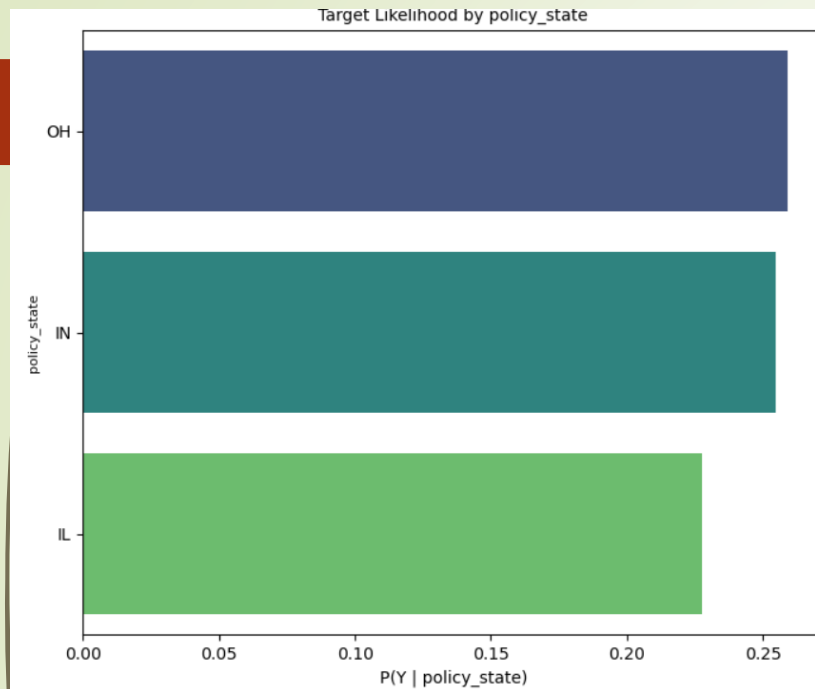
Outcome:

After all the manipulation being done, we retained with 114 columns and 1052 rows for the model building after resampling.

Exploratory Data analysis – Likelihoods analysis



- Incident severity with Major Damage has the highest likelihood of being associated with fraud, followed by Total Loss.
- Certain occupations like executive-managerial and farming-fishing show a stronger correlation with fraudulent claims.
- State-wise, claims from Ohio (OH) have the highest fraud likelihood, with North Carolina (NC) and South Carolina (SC) also showing elevated risks.



Features with less importance:

- ***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%.
These may not significantly contribute to the model's performance and can be considered for removal.



- **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.

Model Building

- Split the dataset into training (70%) and validation (30%) sets for model evaluation.
- Scaled features using standard Scaler to normalize the values and ensure consistent scaling across variables.
- Applied Recursive Feature Elimination (RFE) to select the top 15 most relevant features, reducing overfitting.
- Built the logistic regression model by removing features with a p-value > 0.05 and VIF > 5 to ensure stability and avoid multicollinearity.
- Made predictions on both the training and testing datasets to evaluate model performance.
- Used the ROC curve to determine the optimal cutoff probability and adjusted the threshold for better conversion predictions.
- Evaluated the model using accuracy, sensitivity, specificity, and other relevant metrics to assess its effectiveness in predicting lead conversions.

Model Building

- Split the dataset into training (70%) and validation (30%) sets for model evaluation.
- Scaled features using standard Scaler to normalize the values and ensure consistent scaling across variables.
- Made predictions on both the training and testing datasets to evaluate model performance.

Logistic regression	Random forest
<ul style="list-style-type: none">• Trained on the training dataset using default hyperparameters.• Performance optimized by adjusting probability cutoff values (0.5 and 0.44).• Calculated Accuracy, Precision, Recall, Specificity, and F1 Score on both training and validation sets.• Shows signs of overfitting: high training performance but drop in validation metrics.	<ul style="list-style-type: none">• Built initial Random Forest model without hyperparameter tuning.• Applied GridSearchCV to tune parameters (e.g., n_estimators, max_depth, etc.).• Initial model (untuned) showed perfect performance on training data — clear overfitting.• After hyperparameter tuning, performance improved significantly on the validation set, balancing generalization and accuracy.

Evaluation metrics and observation

Metric	Logistic Regression	Random Forest (Tuned)
Accuracy	72%	80%
Recall (Sensitivity)	0.59	0.65
Precision	0.45	0.59
Specificity	0.76	0.85
F1 Score	0.51	0.62

1. Random Forest (tuned) clearly outperforms logistic regression across all metrics.
2. Recall and F1 Score are especially critical for fraud detection.
3. Recall: Ensures more fraudulent claims are caught early.
4. F1 Score: Balances the trade-off between catching fraud and minimizing false positives.

Training vs Validation Performance:

1. Logistic Regression shows overfitting, with a strong drop in validation performance.
2. Random Forest (without tuning) performs perfectly on training data - a sign of severe overfitting.
3. Tuned Random Forest generalizes well to unseen data, making it the most reliable model.

Business recommendations and conclusion

Recommended Model : Random Forest with Hyperparameter Tuning

- 1.High generalization to unseen claims
- 2.Strong fraud detection capability (recall = 0.65)
- 3.Good balance of precision and recall (F1 = 0.62)


Business Impact:

- 1.Increases early detection of fraudulent claims
- 2.Reduces financial losses
- 3.Supports efficient and data-driven claim triaging
- 4.Focus on the main features suggested to reach the objective.

Top predictive features suggested to achieve this objective

Important features to be considered

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

- 
- The Random Forest classifier ranks feature importance based on their contribution to accurate fraud prediction.
 - Top predictors include:
 1. **incident_severity_Minor Damage and incident_severity_Total Loss** – severity levels of the incident play a crucial role.
 2. **months_as_customer** – longer customer relationships may correlate with reduced fraud risk.
 3. **Types of claims such as injury_claim, vehicle_claim, and property_claim** also show strong influence, suggesting fraud may be tied to specific claim patterns.
 - Other notable features include policy_annual_premium, collision_type_UNKNOWN, and location-based features like incident_state_WV and incident_city_Springfield.
 - These insights help focus attention on high-impact variables, improving model interpretability and guiding further fraud investigation strategies.



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

- Data preprocessing, EDA, and likelihood analysis were used to uncover trends.
- Fraudulent patterns were detected using features like **occupation, incident severity, and state**.
- Machine learning models (Logistic Regression & Random Forest) were applied to classify fraud.

2. Which features are most predictive of fraudulent behaviour?

- The Random Forest model identified key predictors such as incident_severity_Minor Damage, months_as_customer, injury_claim, policy_annual_premium.
- These features suggest fraud is more likely in minor damages and specific claim types.

3. Can we predict the likelihood of fraud for an incoming claim, based on past data?

- Yes.
- The tuned Random Forest model achieved **80% accuracy** with a **recall of 0.65**, enabling early and effective fraud flagging.

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

- Focus on key indicators like claim severity and customer tenure.
- Automate early detection using the trained model for higher efficiency.
- Prioritize manual investigation for high-risk profiles based on model output.



Thank you !!!