FRAUDULENT CLAIM DETECTION

A MACHINE LEARNING APPROACH TO DETECT INSURANCE FRAUD

PROBLEM STATEMENT

- Insurance fraud results in massive financial losses each year.
- The objective of this project is to build a model that can detect fraudulent insurance claims based on various customer and claim attributes.

DATASET OVERVIEW

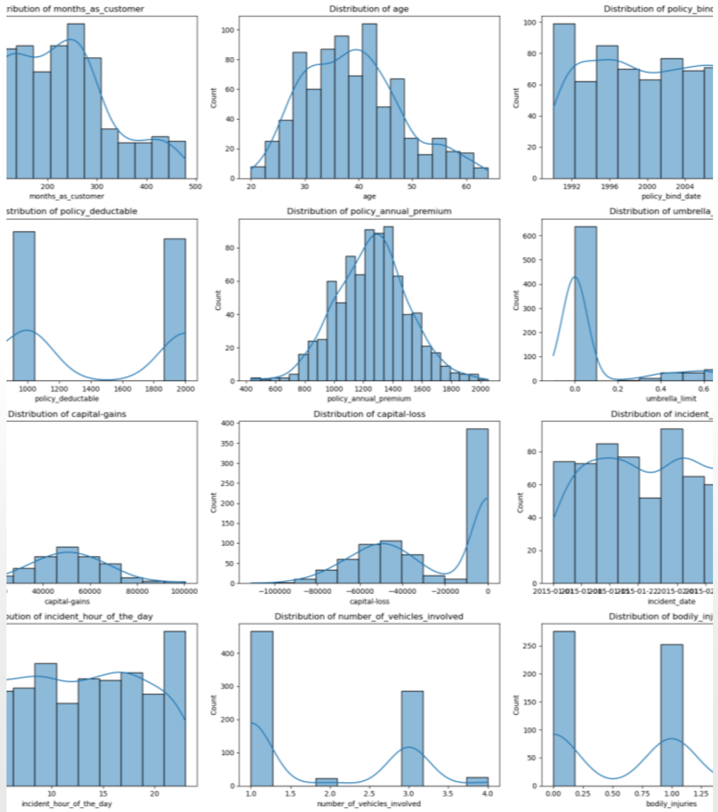
- - ~1000 insurance claim records
- - 40 columns
- - Target Variable: 'fraud_reported'
- - Mix of categorical, numerical, and date fields

DATA PREPROCESSING

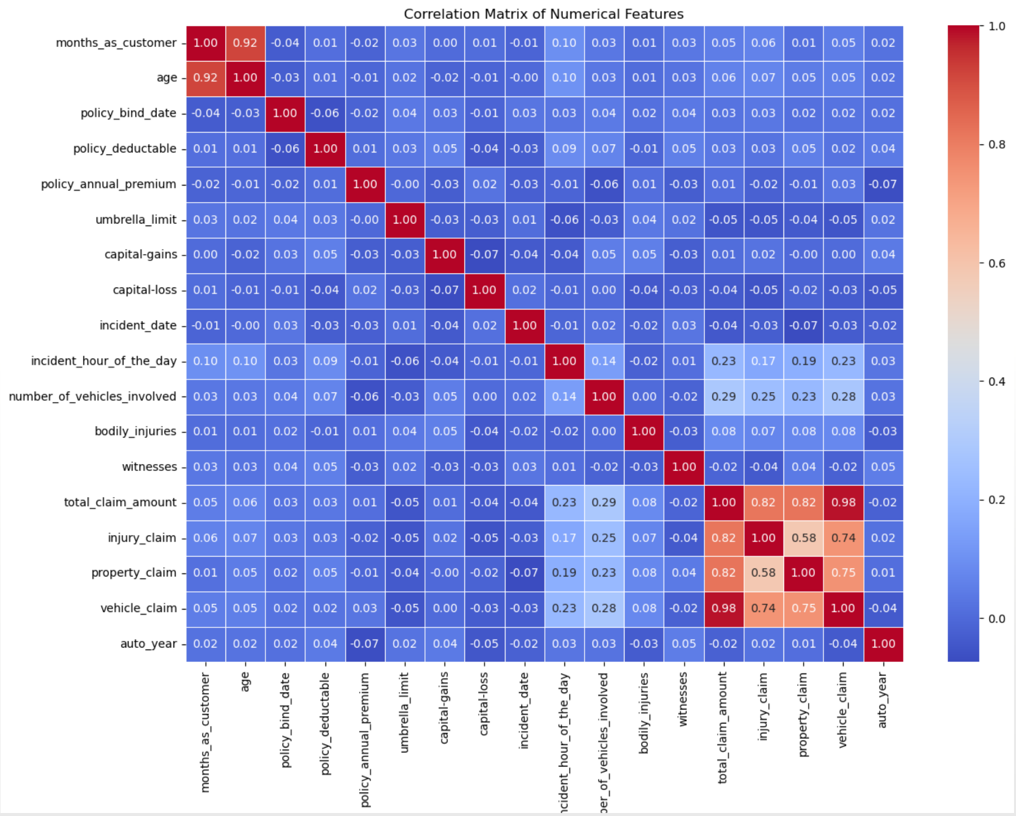
- - Removed redundant columns (_c39)
- - Converted date fields
- - Encoded categorical variables
- - Scaled numerical features
- - Handled class imbalance using resampling

TRAIN-VALIDATION SPLIT

- Training Set: 800 samples
- Validation Set: 200 samples
- Class Balance
  - Training: 75.25% not fraud, 24.75% fraud
  - Validation: 75.5% not fraud, 24.5% fraud
- The class distribution is preserved using stratified splitting.

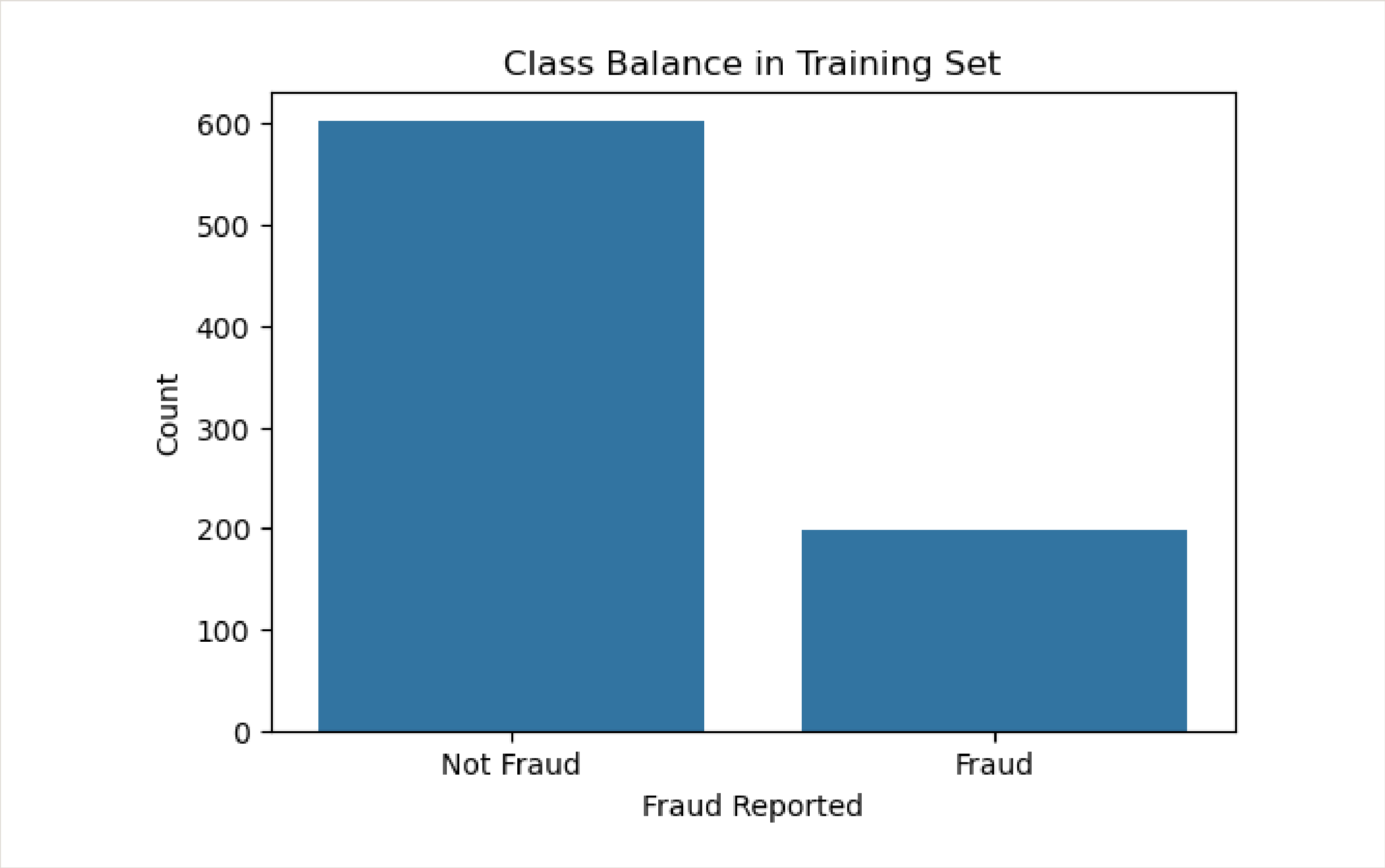


EXPLORATORY DATA ANALYSIS (EDA) ON TRAINING DATA

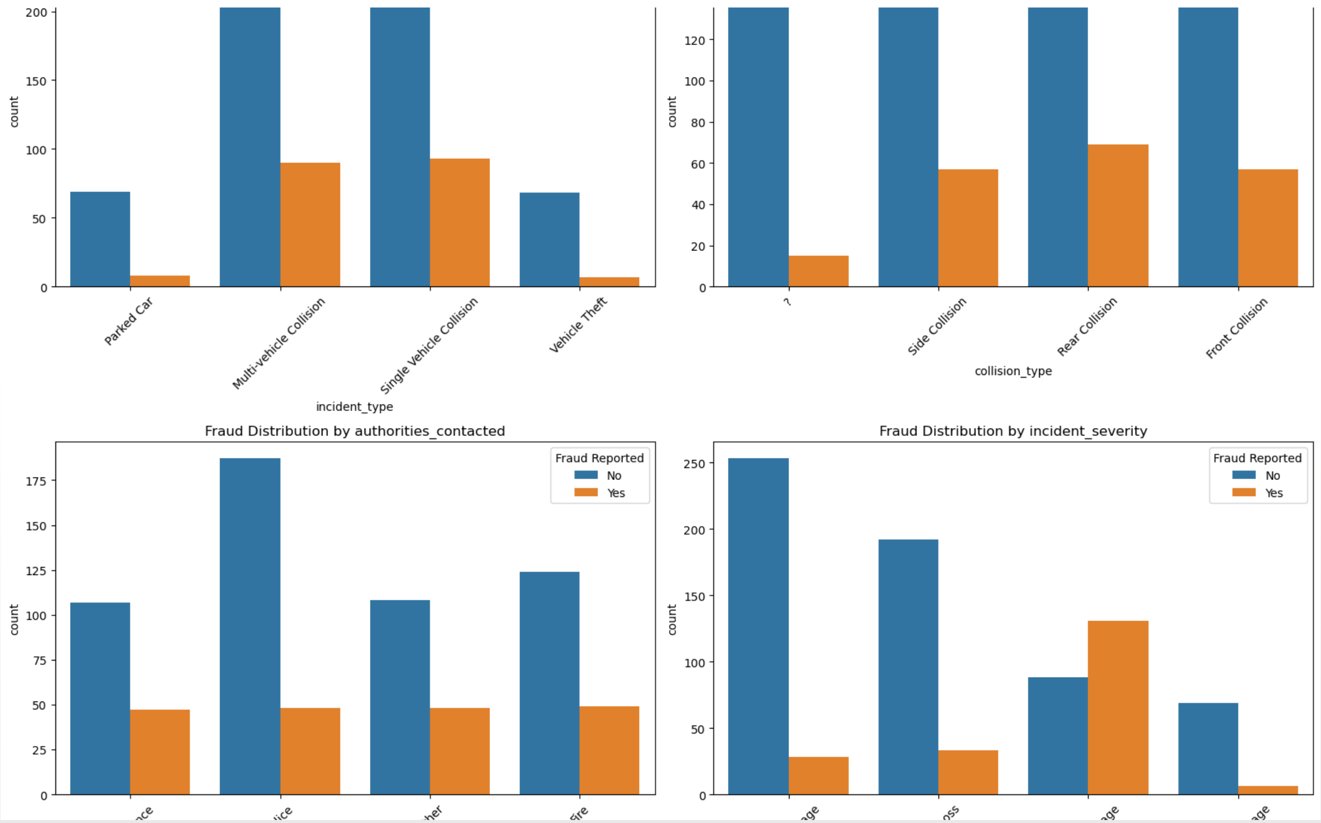


CORRELATION ANALYSIS

CLASS BALANCE CHECK

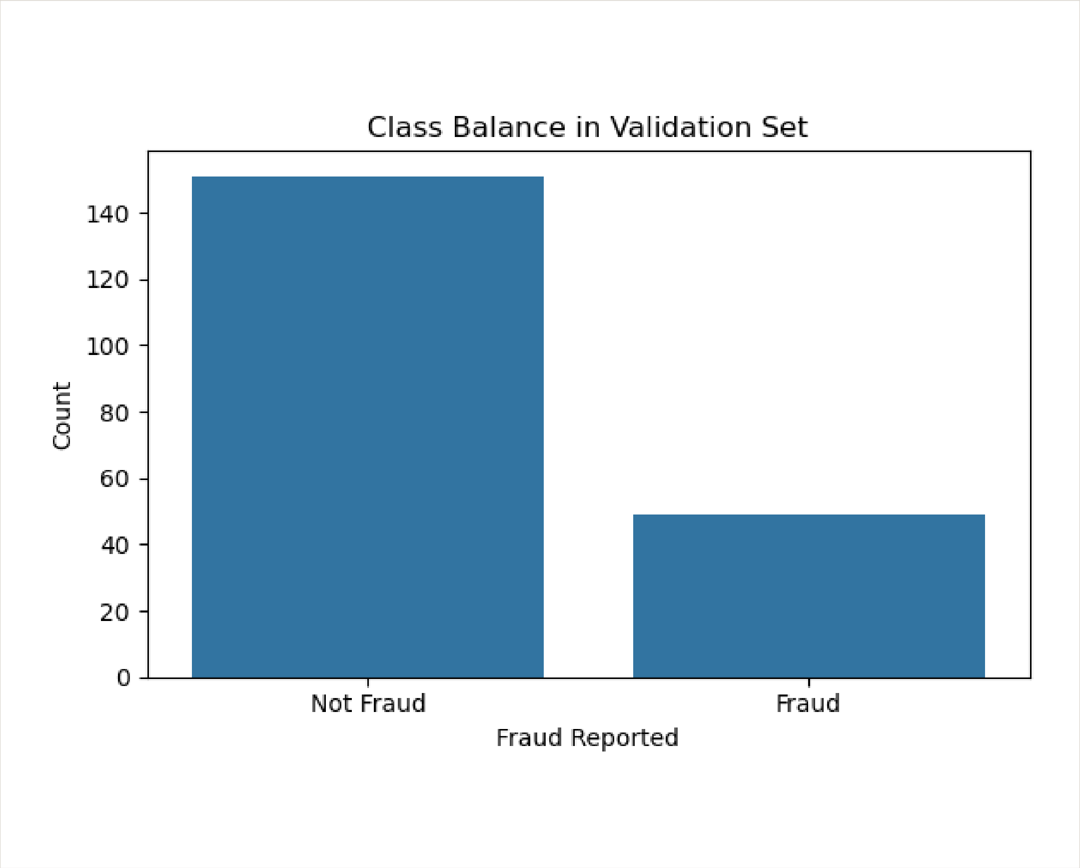


BIVARIATE ANALYSIS



BIVARIATE ANALYSIS INSIGHTS

- **Incident Type**: “Single Vehicle Collision” has a visibly higher fraud rate than other types.
- **Collision Type**: Some missing categories, but “Rear Collision” seems more associated with non-fraud.
- **Authorities Contacted**: Fraud is more prevalent when **no authorities** were contacted.
- **Incident Severity**: “Total Loss” has a disproportionately higher fraud rate.



EDA ON VALIDATION DATA (OPTIONAL)

- **Class balance** in the validation set mirrors the training set:
  - ~75% Not Fraud
  - ~25% Fraud

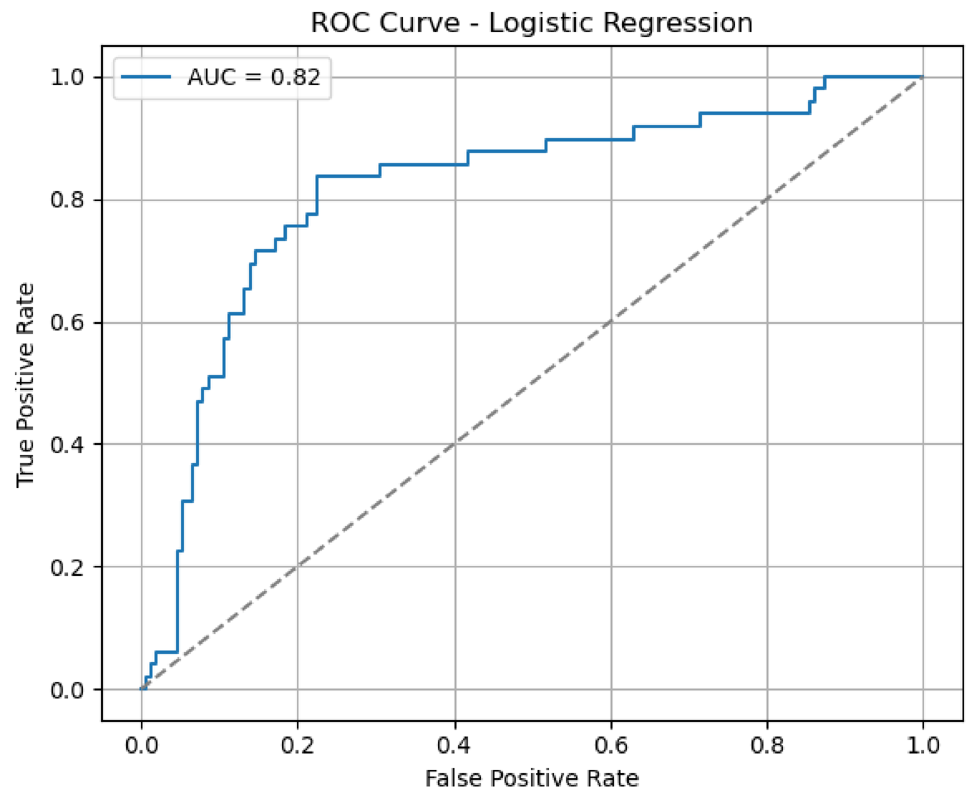
- This confirms that our **stratified split** preserved the class distribution.

FEATURE ENGINEERING

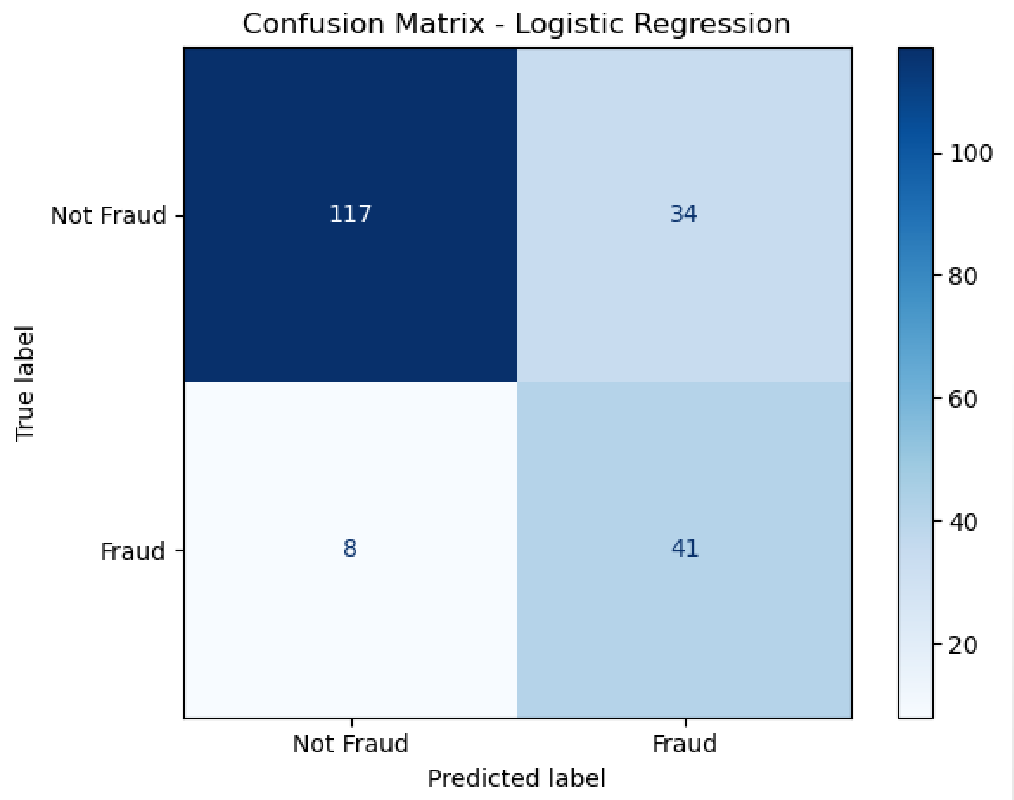
Key Steps Performed:

- Class imbalance fixed using upsampling (now 50/50 fraud vs not fraud).
- Date features were broken into year and month.
- Categorical features were one-hot encoded.
- Numerical features were standardized using StandardScaler.

Final feature space includes 161 columns after encoding and transformations.

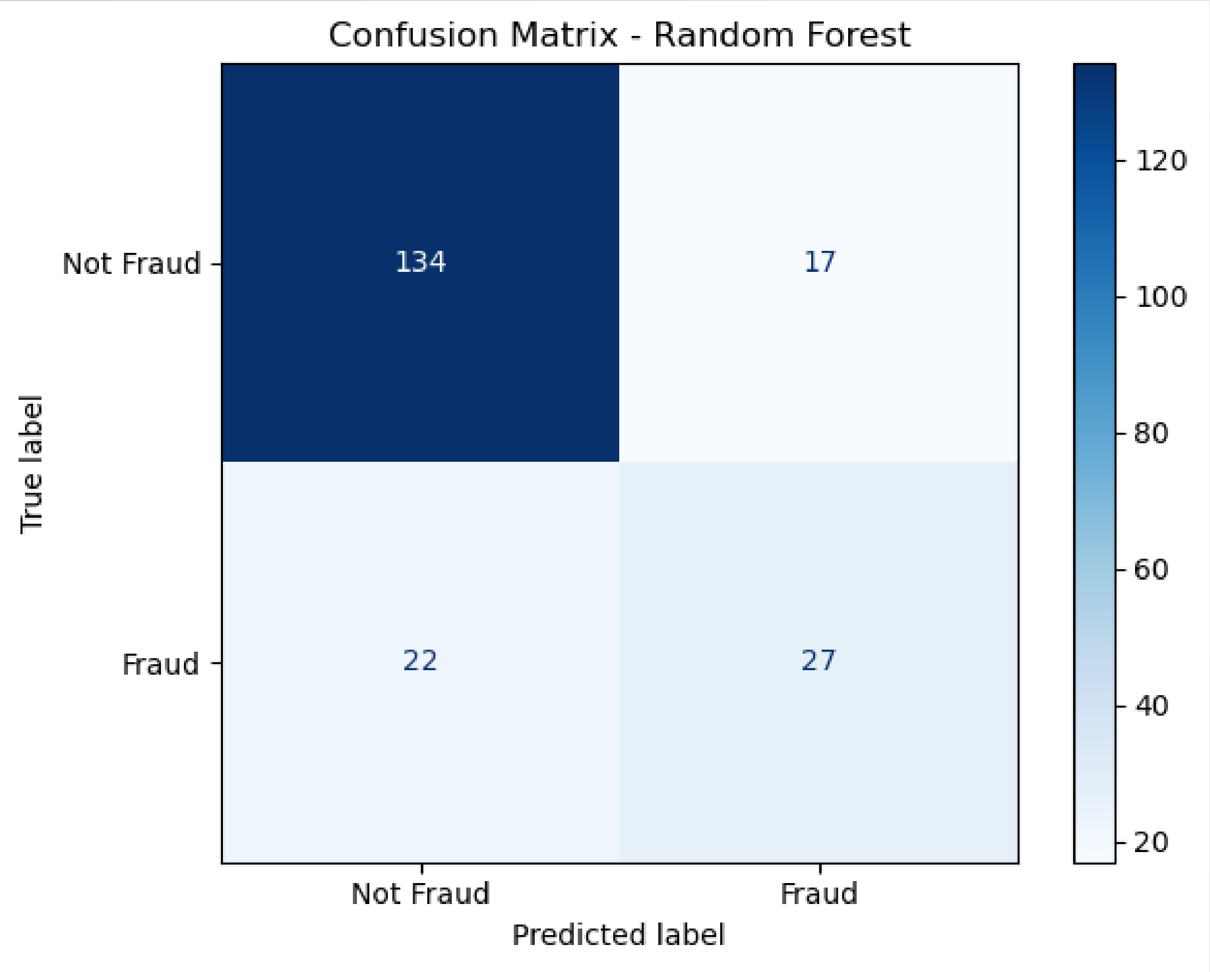


MODEL BUILDING



LOGISTIC REGRESSION MODEL

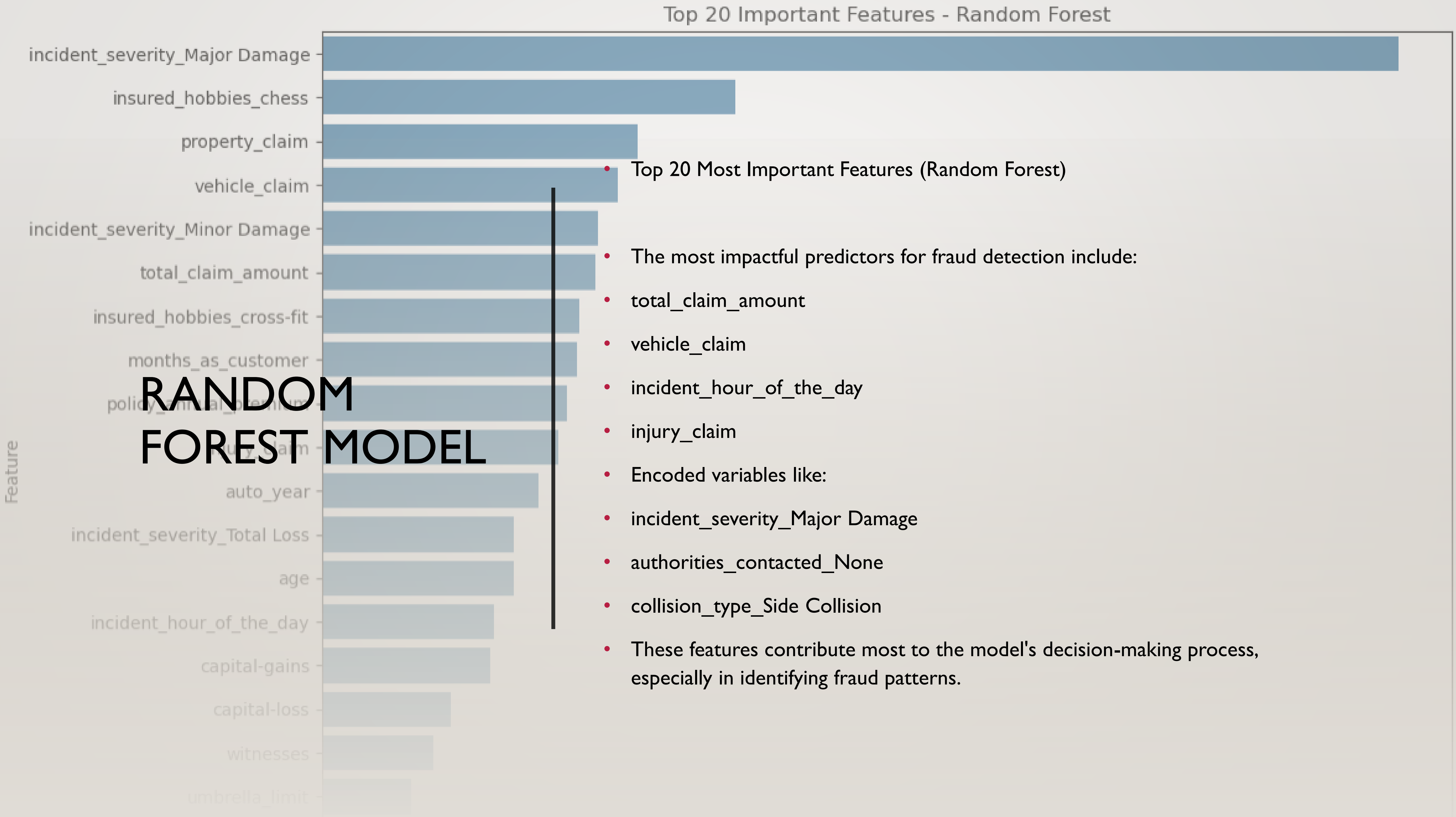
- Logistic Regression Results (at threshold = 0.25)
- Metric Not Fraud (0) Fraud (1)
- Precision 0.94 0.55
- Recall 0.85 0.84
- F1-score 0.85 0.66
- Overall Accuracy: 79%
- The model does well in detecting fraud (Recall = 0.84), though precision for fraud is moderate.



TRAIN A STANDARD RANDOM FOREST

- Random Forest Model – Evaluation Summary ROC AUC Score:
- AUC = 0.84, which indicates strong overall model performance.
- Metric Not Fraud (0) Fraud (1)
- Precision 0.86 0.61
- Recall 0.89 0.55
- F1-score 0.87 0.58
- Accuracy: 80.5%
- Key Observations:
- Excellent recall for Not Fraud
- Moderate fraud detection (precision = 0.62, recall = 0.55)

RANDOM FOREST MODEL



- Top 20 Most Important Features (Random Forest)
- The most impactful predictors for fraud detection include:
- total_claim_amount
- vehicle_claim
- incident_hour_of_the_day
- injury_claim
- Encoded variables like:
- incident_severity_Major Damage
- authorities_contacted_None
- collision_type_Side Collision
- These features contribute most to the model's decision-making process, especially in identifying fraud patterns.

MODEL EVALUATION

- Logistic Regression (Threshold = 0.25)
- Metric Not Fraud (0) Fraud (1)
- Precision 0.94 0.55
- Recall 0.85 0.84
- F1-Score 0.85 0.66
- Overall Accuracy: 79%
- Insight: Logistic Regression performs well in identifying frauds (high recall = 0.84), although the precision for fraud detection is moderate.

Random Forest Model

- ROC AUC Score: 0.84

Metric	Not Fraud (0)	Fraud (1)
Precision	0.86	0.61
Recall	0.89	0.55
F1-Score	0.87	0.58

- Overall Accuracy: 80.5%
- Insight:
- Strong generalization ability as seen from a high AUC score.
- Performs very well on non-fraud cases.
- Moderate performance on fraud cases, but better balanced than logistic regression in terms of precision.

Conclusion

- The models built (Logistic Regression and Random Forest) are effective at identifying fraudulent claims, with the **Random Forest** showing better overall performance and slightly higher accuracy.
- However, both models show a **trade-off between precision and recall** for fraud detection:
  - High recall ensures most frauds are caught.
  - Moderate precision indicates some false positives.
- **Random Forest** is the recommended model for deployment due to its robustness and superior evaluation metrics.
- Further improvements can be made using:
  - Advanced ensemble methods (e.g., XGBoost, LightGBM).
  - Feature selection and dimensionality reduction.
  - Cost-sensitive learning to penalize misclassification of fraud cases.