FRAUDULENT CLAIM DETECTION

Abhishek Ujwal and Gopal Ijagude and

Fraudulent Claim Detection

21-Oct-2025

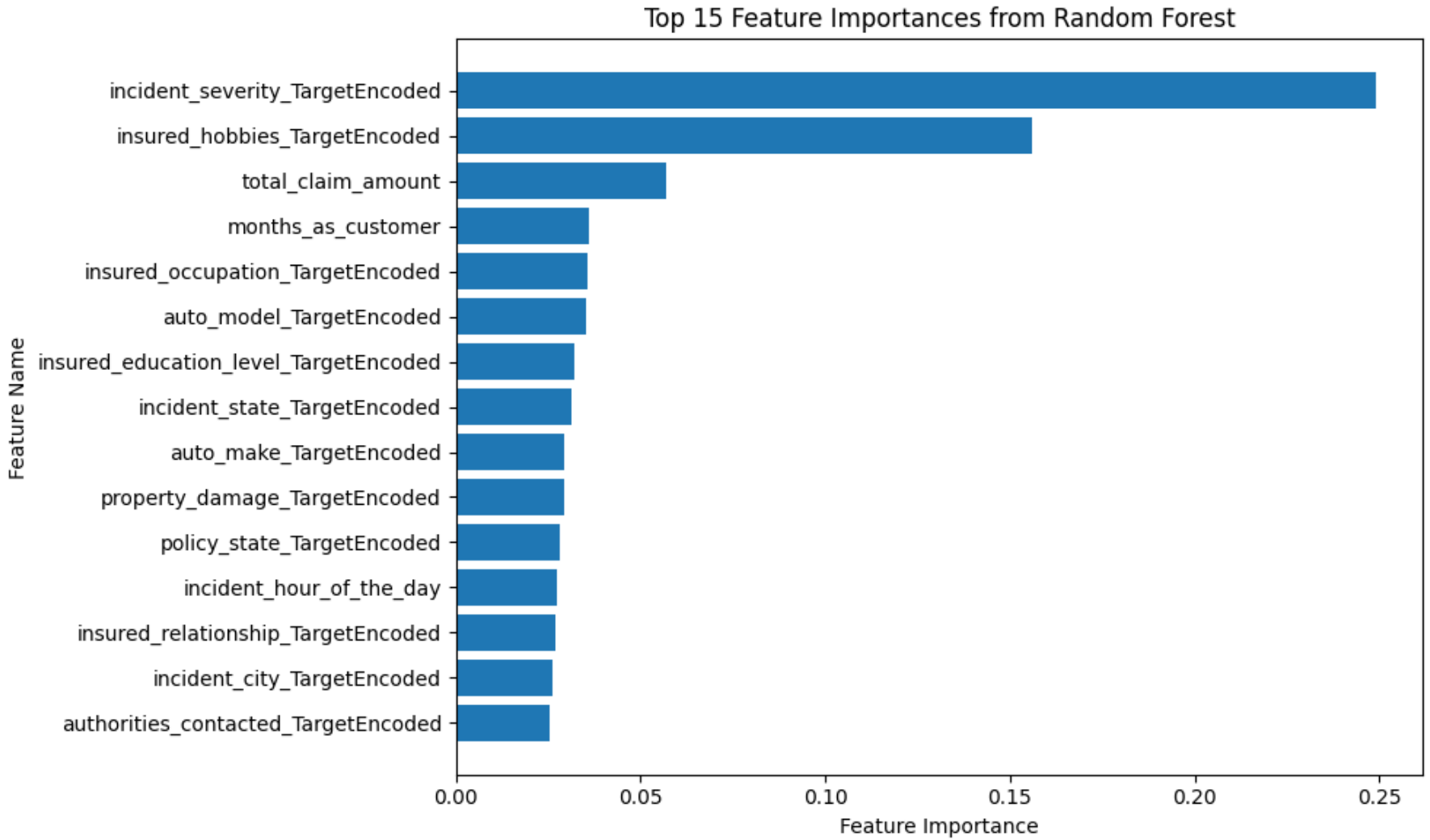
1) HOW CAN WE ANALYSE HISTORICAL CLAIM DATA TO DETECT PATTERNS THAT INDICATE FRAUDULENT CLAIMS?

- **Approach:**
- **Data Cleaning & Preparation:** Remove redundant, identifier, and illogical columns/values; handle missing data; convert data types appropriately.
- **Exploratory Data Analysis (EDA):** Use univariate and bivariate analysis, correlation matrices, and class balance checks to understand feature distributions and relationships with the target (fraud_reported).
- **Feature Engineering:** Apply target encoding for categorical variables, create dummy variables, and scale numerical features.
- **Model Building:** Train machine learning models (Logistic Regression, Random Forest) to learn patterns that distinguish fraudulent from legitimate claims.
- **Pattern Detection:** Feature importance and model coefficients highlight which variables are most associated with fraud.

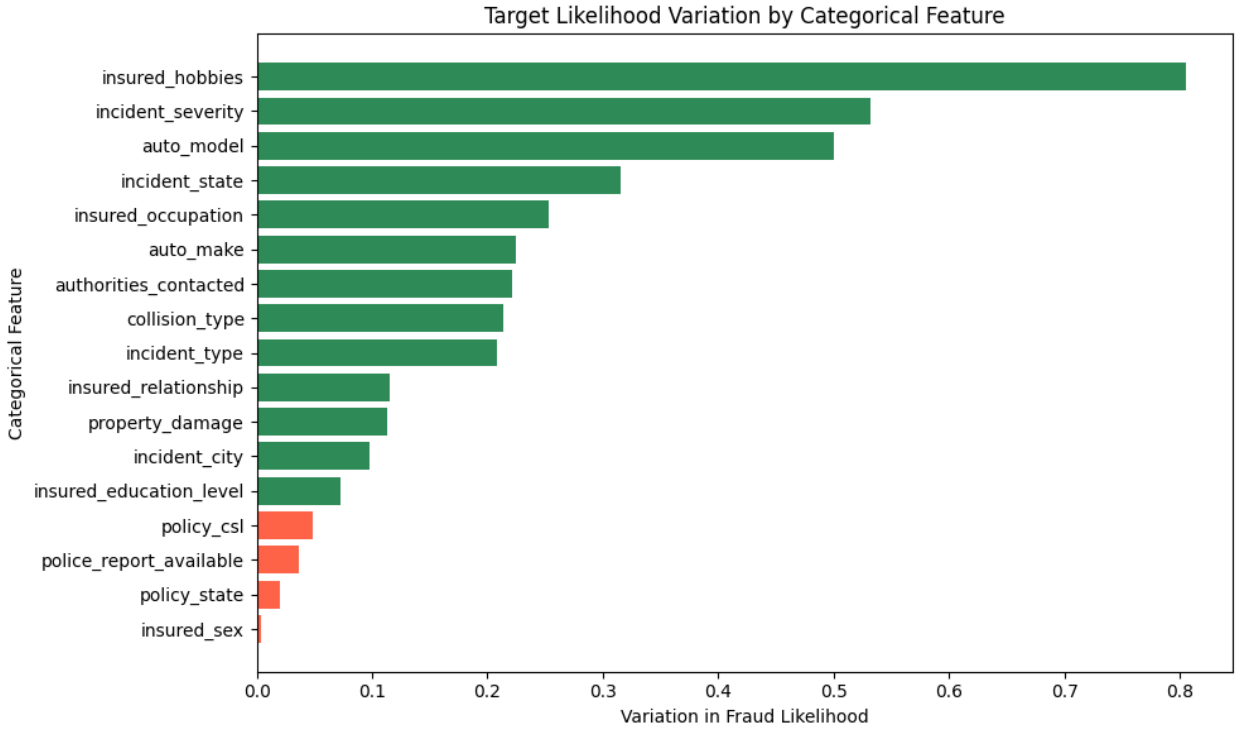
2) WHICH FEATURES ARE MOST PREDICTIVE OF FRAUDULENT BEHAVIOUR?

- **Findings from Feature Importance (Random Forest):**
- **Top predictive features** (as per the feature importance plot and DataFrame) typically include:
  - incident_severity
  - Insured_hobbies
  - total_claim_amount
  - These features have the highest importance scores and thus contribute most to the model's ability to detect fraud.

IMPORTANT FEATURES



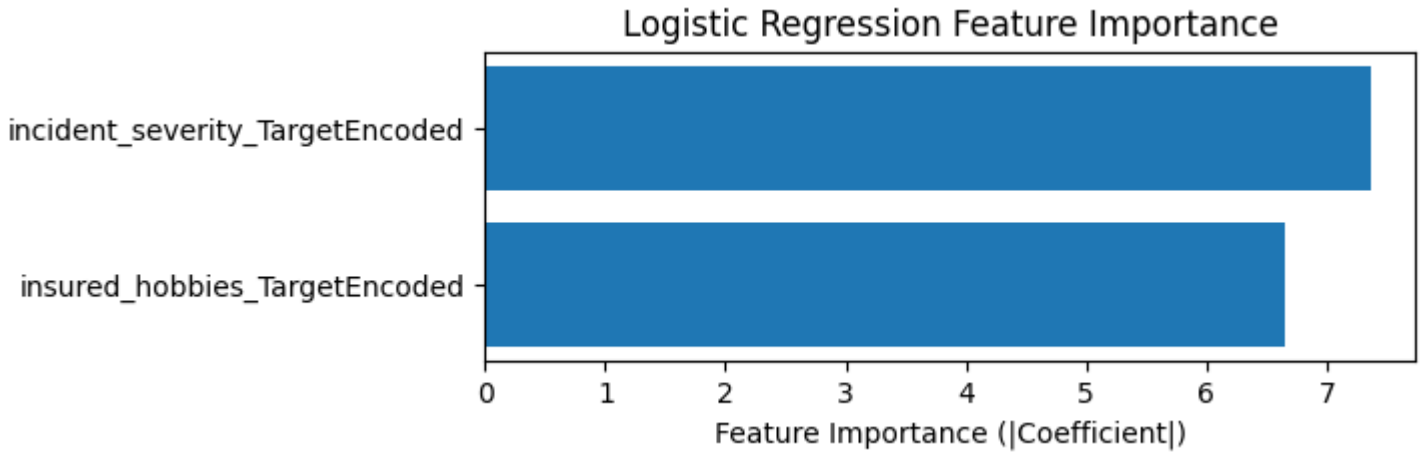
IMPORTANT CATEGORICAL FEATURES BASED ON DUMMY VARIABLES WITH RANDOM FOREST



IMPORTANT FEATURES BASED ON LOGISTIC REGRESSION

- **Findings from Feature Importance (Logistic Regression):**
- **Top predictive features** (as per the feature importance plot and DataFrame) typically include:
  - incident_severity
  - Insured_hobbies
  - These features have the highest importance scores and thus contribute most to the model's ability to detect fraud.

IMPORTANT FEATURES BASED ON TARGET ENCODING AND LIKELIHOOD ANALYSIS FOR LOGISTIC REGRESSION



3) CAN WE PREDICT THE LIKELIHOOD OF FRAUD FOR AN INCOMING CLAIM, BASED ON PAST DATA?

- **Yes.**
- The notebook demonstrates that both Logistic Regression and Random Forest models can be trained on historical data to predict the probability of fraud for new claims.
- The models output a probability score for each claim, which can be thresholded (using an optimal cutoff) to classify claims as fraudulent or not.
- Model evaluation on validation data shows good accuracy, sensitivity, and specificity, confirming the approach is effective.

4) WHAT INSIGHTS CAN BE DRAWN FROM THE MODEL THAT CAN HELP IN IMPROVING THE FRAUD DETECTION PROCESS?

- **Key Insights:**
- **Feature Importance:** The most influential features (claim amounts, incident details, and certain categorical encodings) should be prioritized in manual reviews and automated checks.
- **Process Optimization:** Early identification of high-risk claims allows for targeted investigation, reducing manual workload and financial losses.
- **Data Quality:** Removing redundant and non-informative features improves model performance and interpretability.
- **Balanced Data:** Handling class imbalance (e.g., with RandomOverSampler) ensures the model is not biased toward the majority class, improving fraud detection rates.
- **Continuous Improvement:** Regularly updating the model with new data and re-evaluating feature importances can adapt the system to evolving fraud patterns.

- **Summary:**
  By systematically cleaning, analyzing, and modeling historical claim data, the notebook provides a robust framework for detecting fraudulent claims, identifying key predictive features, and supporting actionable improvements in the fraud detection process.