Learning to Spot the Lie: A High-Recall Random Forest Approach to Vehicle Insurance Fraud Detection

BY Rusi Rothschild

Overview

The Problem and Its Impact

Project Goals and Objectives

Dataset and Key Features

EDA and Preprocessing

Modeling Strategy and Workflow

Model Performance Comparison

Final Model and Feature Importance

Business Use and Value

Next Steps

Appendix

Understanding the Problem and its Impact

Vehicle insurance fraud is a massive, growing problem that costs the U.S. industry over **$40 billion annually**, raising honest customers' premiums by **$400–$700 per household**. This creates **financial strain** and **erodes trust** in insurers.

Traditional methods are slow, inconsistent, and struggle to detect **increasingly more complex fraud schemes**.

A smarter, **data-driven** approach is needed — one that **flags suspicious claims early**, detects hidden patterns, and supports investigators in **reducing fraud-related costs and risks**.

Project Goals and Objectives

- Build a machine learning model to detect fraudulent claims
- Improve accuracy compared to traditional methods
- Prioritize recall to minimize undetected fraud
- Identify key features most predictive of fraud

Dataset and Key Features

This dataset includes 15,420 auto insurance claims with detailed policy, driver, and accident info — used to build a fraud detection model.

Total Claims: 15,420

- About 6% of claims are labeled as fraud

Total Features: 33

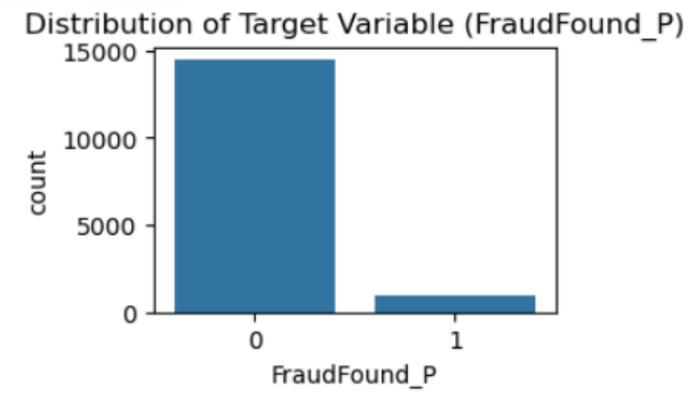
- Mix of numerical and categorical variables
- Features include vehicle, driver, policy, and accident details

Target Variable: FraudFound

- Binary Label: 1 = fraud, 0 = legitimate
- Class imbalance requires special modeling strategies

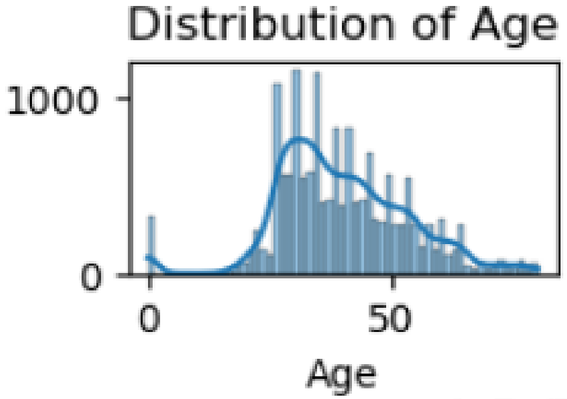
Exploratory Data Analysis - Class Balance and Distributions

Class Imbalance Bar Chart



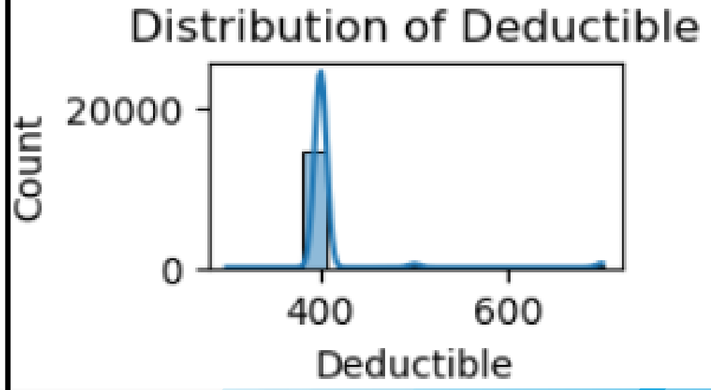
The dataset is highly imbalanced, with only ~6% of claims labeled as fraud. This required special handling during model training

Numerical Variable Distribution: Agee



Claimants are mostly between 30–50 years old. Fraud appears slightly more common among younger policyholders.

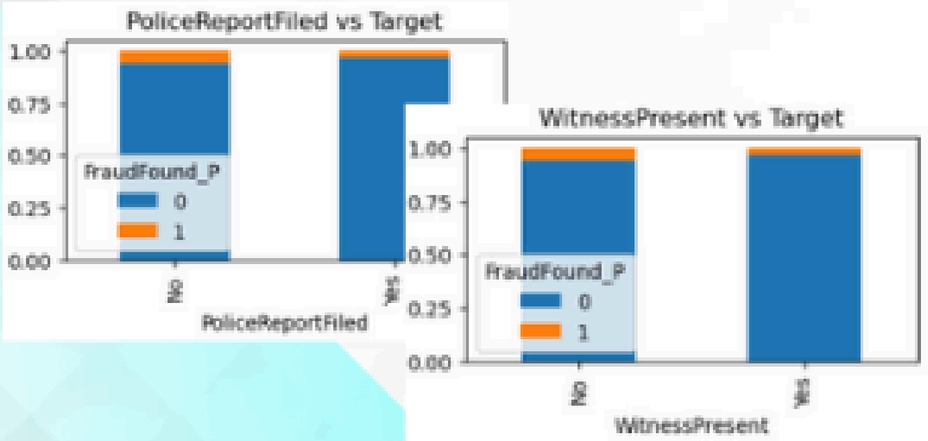
Constant Feature: Deductible



Deductible shows no variation (mostly 400) — likely not useful for modeling.

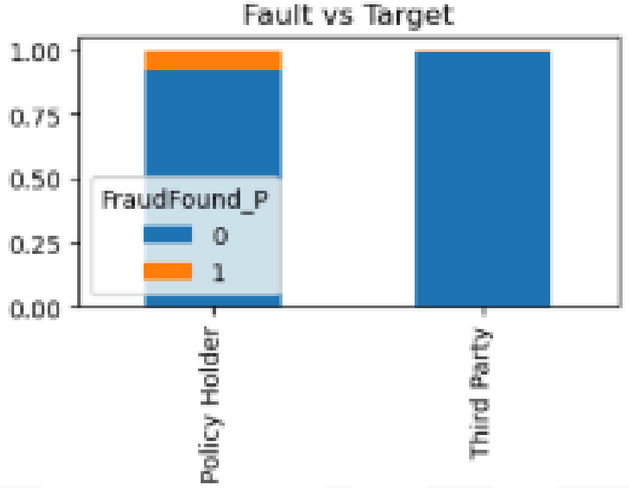
Exploratory Data Analysis - Feature Patterns and Key indicators

Suspicious Claims Often Lack Documentation



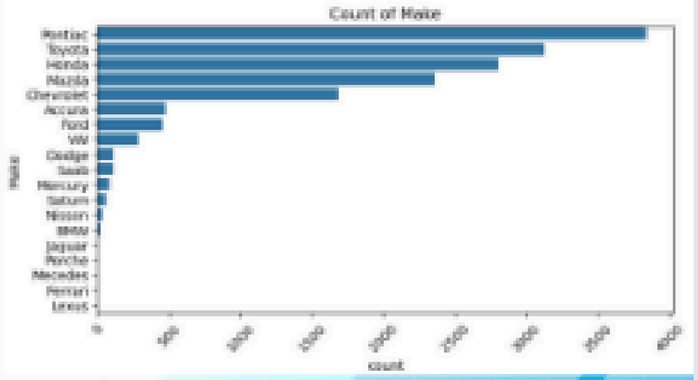
Fraud was more common when no witness was present and no police report was filed — these may be weak spots in claim verification.

Fraud More Likely When Policyholder Is at Fault



Claims where the policyholder was listed as at fault had a noticeably higher fraud rate — suggesting fault assignment may be a useful early fraud signal.

High-Cardinality Variables Pose Modeling Challenges



Some features (like Make) had many unique values, requiring label encoding during preprocessing.

Feature Engineering and Preprocessing

Feature Encoding

- Ordinal encoding for ordered variables (ex: Month, DayOfWeek, VehiclePrice)
- Binary encoding for yes/no features (PoliceReportFiled, WitnessPresent, etc.)
- Nominal features were encoded two ways depending on the model type:
  - Label encoding for tree-based models (ex: Random Forest, XGBoost)
  - One-hot encoding for logistic regression (to avoid implying order)

Preprocessing Steps

- Removed ID fields (PolicyNumber, RepNumber)
- Flagged irregular values (ex: 0 in MonthClaimed)
- Managed high-cardinality features (ex: Make)

Modeling Strategy and Workflow

Modeling Approach:

- Tested multiple supervised ML models
- Performed feature selection using RFE
- Used SMOTE and class_weight='balanced' to handle class imbalance
- Tuned hyperparameters using grid/random search
- Prioritized recall to reduce missed fraud cases
- Evaluated using precision, recall, FI-score, and AUC

Models Tested

XGBoost

Decision Tree

Random Forest

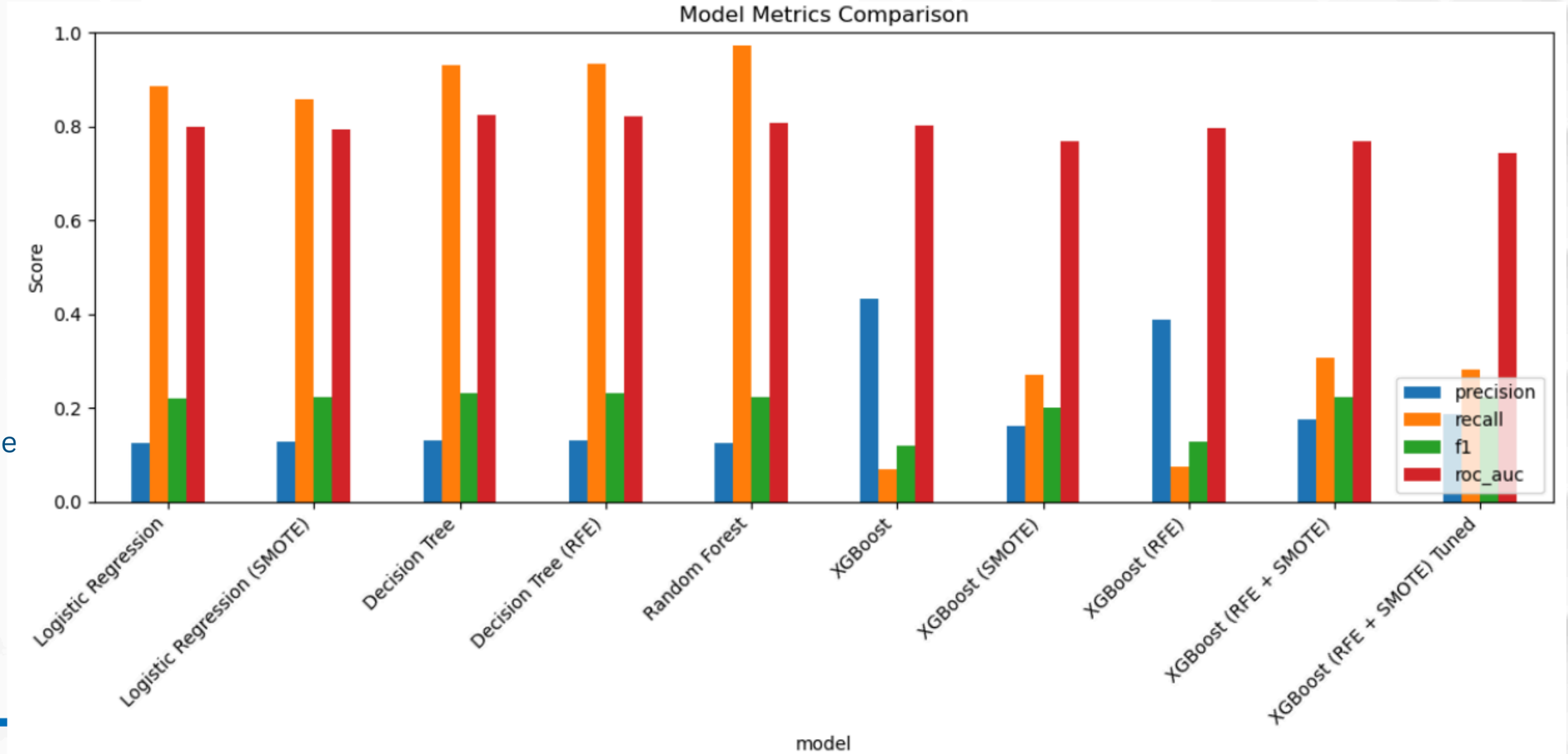
Logistic Regression

Lazy Predict

Model Performance and Comparison

Model Insights:

- **SMOTE** did not improve performance for **Logistic Regression**
- **XGBoost** had the **highest precision**, but recall was low— and recall was the priority
- **RFE + SMOTE** improved XGBoost slightly, but **hyperparameter tuning** did not boost recall
- **Random Forest** achieved the **highest recall** and strong overall performance

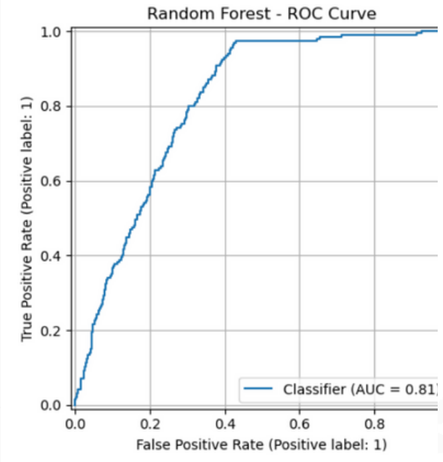


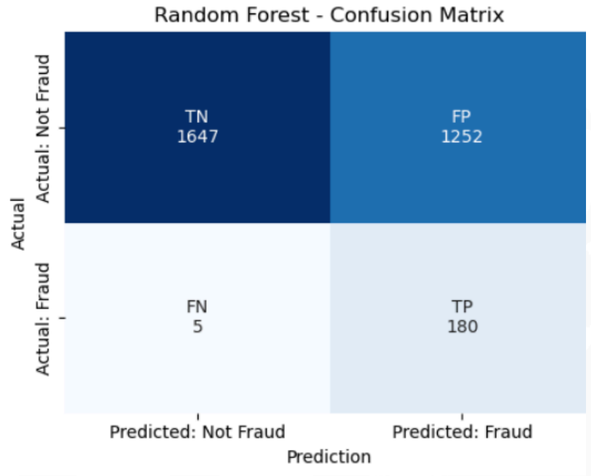
Final Model: Random Forest

The Random Forest model successfully identified nearly all fraud cases, making it the best fit for a high-recall fraud detection system.

Final Model Evaluation:

- Precision: 0.13
- Recall: 0.97
- F1-score: 0.22
- AUC: 0.81



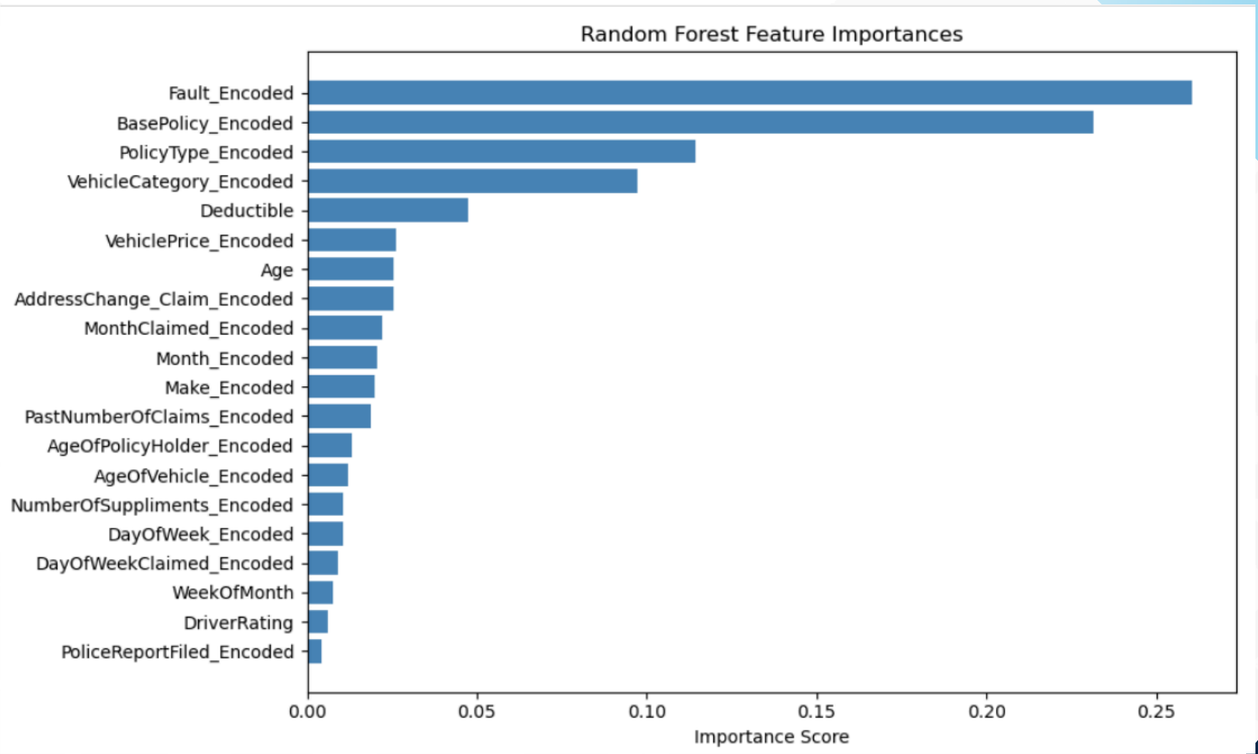


- Threshold tuned using precision-recall curve to maximize fraud detection
- Small drop in precision was acceptable given project goals

Final Model: Feature Importance

Top Feature Insights:

- **Fault** was the strongest predictor — fraud more common when policyholder was at fault
- **Base Policy** and **Vehicle Category** also showed strong influence
- Several categorical features contributed to the prediction, including **Make** and **Address Change**
- **SHAP** analysis from XGBoost confirmed similar feature rankings



Business Use and Value

1

Reduce Financial Losses
By flagging high-risk claims early, insurers can allocate investigative resources more efficiently and prevent costly payouts.

2

Protect Honest Policyholders
Reducing fraud helps stabilize or lower premiums for non-fraudulent customers—preserving fairness and customer trust.

3

Support Smarter Claim Prioritization
The model acts as a pre-screening tool, sending high-risk claims for review without replacing human oversight.

4

Improve Investigative Efficiency
Key fraud signals (ex: fault, policy type, vehicle value) can guide targeted reviews—saving time and improving hit rates.

5

Enable Transparent, Explainable AI
The model prioritizes interpretability—supporting compliance, stakeholder trust, and ethical deployment.

Next Steps

Expand the Dataset
Include more recent and diverse claims to improve generalizability.

Add Richer Inputs
Incorporate unstructured data (ex: text, images) for deeper fraud signals.

Validate on External Data
Test the model on other insurers or time periods to assess real-world robustness.

Enable Collaborative Detection
Explore privacy-safe data sharing to strengthen fraud detection across companies.