ML-Driven Fraud Detection for Staged Auto Accidents in Standard Auto Coverage

Project Description

Detect and reduce fraudulent accident claims using advanced ML algorithms.

Business Problem

Problem Description

The business problem centers on the inability to accurately distinguish between **genuine accidents and staged events** designed to trigger fraudulent claims.

- Insurers struggle with claims that, despite having similar accident characteristics, contain **subtle fraud patterns**.
- These anomalies include:
    - **Inconsistent damage reports**
    - **Atypical repair cost trajectories**
    - **Deviations from historical claim behavior**
- Example: A claim categorized under **Collision & Comprehensive coverage** may initially appear legitimate but could exhibit **data inconsistencies** typical of **staged accidents** upon deeper inspection.
- This problem is **compounded** by the **vast and diverse data** collected during the claims process—such as:
    - **Vehicle repair estimates**
    - **Accident timestamps**
    - **Historical claim patterns**
    - **Vendor repair histories**
- Given the complexity, a **machine learning (ML)-powered approach** is needed to **generate risk scores** and quickly **identify fraudulent claims**.

OKR (Objective & Key Result)

Objective

Optimize fraud detection accuracy to substantially **reduce financial losses** from **staged accident claims** within one year.

Key Result

- **Increase the Accuracy of Fraud Detection by 25%** using **enhanced ML algorithms** by **Q4**.

Segments

Segment Name	Description	Purpose	Volume
Common Routine Claims	Standard claims that follow common, expected patterns with average submission delays and claim amounts. Rules: FNOL Submission Delay between 24–72 hr; Claim Amount < $10,000; Historical Claim Frequency normal; Policy Age any; Denied Claims History = 0.; fraud flag = false	Confirm that claims with normative behavior do not require additional scrutiny.	20
Consistent Older Vehicle Claims	Claims from older vehicles that exhibit damage patterns consistent with wear and tear. Rules: Vehicle Age > 10 yrs; Policy Age > 3 yrs; Claim Amount within normal range for older vehicles; Historical Claim Frequency normal; Denied Claims History = 0., fraud flag: false	Validate claims against expected depreciation and repair trends of older vehicles.	8
New Policy High Risk	New policy claims with moderate amounts and some risk indicators, not as extreme as urgent alerts. Rules: FNOL Submission Delay < 2 hr; Claim Amount between $15,000–$25,000; Policy Age < 1 yr; Accident Timestamp any; Vehicle Age < 5 yr; Policyholder age >20 and <40: mixed profiles; Denied Claims History = 0; fraud flag: true	Identify potentially risky new policy claims that require further review.	3
Long-Term Loyal Customers	Claims from customers with long-standing policies and consistent, low-frequency claims. Rules: Policy Age > 5 yrs; Historical Claim Frequency < 1/year; FNOL Submission Delay > 12 hr; Claim Amount ≤ $15,000; Denied Claims History = 0; fraud flag: false	Reward stability and reduce unnecessary investigations for loyal policyholders.	10
Early Morning New Policy Alerts	New policy claims occurring in the early morning hours with high amounts and short delays, indicating potential staging. Rules: FNOL Submission Delay < 12 hr; Claim Amount > $22,000; Accident Time of the day = 05:00 - 08:00; Policy Age < 1 yr; Vehicle Age < 3 yr; Policyholder Demographics: Age < 30; Denied Claims History = 0.; fraud flag: true	Detect high-risk new policy claims filed in the early hours.	1.2

Suspiciously Fast Filers	Claims filed extremely quickly after the incident, raising suspicions of pre-planned staging. Rules: FNOL Submission Delay < 1 hr; Claim Amount > $20,000; Policy Age < 1 yr; Denied Claims History >= 1.fraud flag: true	Detect claims with abnormally fast FNOL submissions that may be artificially expedited.	2.5
Frequent Low-Amount Filers	Policyholders filing claims frequently with low claim amounts, possibly indicating systematic abuse. Rules: Historical Claim Frequency ≥ 4/year; Claim Amount < $5,000; FNOL Submission Delay < 2 hr; Denied Claims History ≥ 1.fraud flag:true	Detect potential abuse from high-frequency, low-dollar claims.	5
Daytime High-Value Alerts	High-value claims occurring during daytime hours with amounts well above average. Rules: FNOL Submission Delay < 24 hr; Claim Amount > $35,000; Accident Time = 12:00 - 18:00; Policy Age < 2 yrs; Vehicle Age < 4 yr; Denied Claims History = 0.; fraud flag = true	Flag claims that, despite typical filing hours, exhibit unusually high financial figures.	2
Mixed Demographic Mid-Risk	Claims with a mix of demographic factors where risk is moderate due to varying socio-economic profiles. Rules: Policyholder Demographics: Age between 30–50; Claim Amount between $12,000–$22,000; Historical Claim Frequency moderate (1–2/year); Denied Claims History = 0., fraud flag: true	Highlight claims with ambiguous demographic profiles that require a closer look.	6
Urban Moderate-Risk Claims	Claims from high-density urban areas with moderate risk profiles and claim amounts. Rules: Accident Location within urban core; Claim Amount between $10,000–$20,000; FNOL Submission Delay < 24 hr; Policy Age < 3 yrs; Historical Claim Frequency ~1–2/year; Denied Claims History = 0., fraud flag: false	Identify urban claims that deviate moderately from typical patterns due to local risk factors.	7
Urgent New Policy Alerts	New policy claims filed almost immediately with very high amounts and high risk, often occurring late at night. Rules: FNOL Submission Delay < 6 hr; Claim Amount > $25,000; Historical Claim Frequency ≥ 4/year; Accident Timestamp 22:00 - 06:00; Accident Location >15 mi from cluster; Policy Age < 1 yr; Vehicle Age < 3 yr; Vehicle Make/Model = 'HighRisk'; Policyholder Demographics: Age < 30; Denied Claims History = 0; fraud flag: true	Flag rare, extremely high-risk claims for immediate investigation.	1

Stable Long-Term Claims	Claims from long-term policyholders with stable histories and proven low-risk behavior over time. Rules: Policy Age > 5 yrs; Historical Claim Frequency < 1/year; Accident Timestamp = 10:00-18:00; Claim Amount within expected range; Denied Claims History = 0.; fraud flag: false	Validate low-risk claims from established customers with a strong track record.	10
Suburban Standard Claims	Claims from suburban areas that follow routine patterns with standard claim amounts and submission behaviors. Rules: Accident Location in suburban area; Claim Amount between $8,000–$15,000; FNOL Submission Delay between 24–72 hr; Policy Age ≥ 2 yrs; Historical Claim Frequency < 1/year; Denied Claims History = 0., fraud flag = false	Recognize normal claims in suburban settings that typically do not warrant further investigation.	12
Routine Low-Risk Claims	Claims meeting expected norms with longer submission delays, lower amounts, and consistent historical behavior. Rules: FNOL Submission Delay ≥ 24 hr and <= 72hr; Claim Amount < $10,000; Policy Age ≥ 2 yrs; Historical Claim Frequency < 1/year; Denied Claims History = 0.; fraud flag = false	Recognize routine claims that are unlikely to be fraudulent.	20
Young Frequent Claimers	Claims from young policyholders who file frequently and quickly, with moderately high amounts. Rules: FNOL Submission Delay < 12 hr; Claim Amount > $20,000; Historical Claim Frequency ≥ 3/year; Accident Timestamp any; Accident Location within 10 mi; Policy Age < 1 yr; Vehicle Age < 5 yr; Vehicle Make/Model = 'HighRisk'; Policyholder Demographics: Age 18–30; Denied Claims History = 1; fraud flag: true	Identify potential fraud among high-frequency, younger filers.	1.5

Schema

Data Element Name	Data Type	Description	Constraints	Is Target	Format	Unit
Claim_ID	string	Unique identifier for each claim.	Not Null, Unique, Required	No		
Policyholder_ID	string	Unique identifier for each policyholder.	Not Null, Unique, Required	No		
Accident_Timestamp	datetime	Date and time when the accident occurred.	Not Null, Required	No	YYYY-MM-DD HH:MM:SS	
Accident_Latitude	float	Latitude coordinate of the accident location.	Not Null, Required	No	Decimal degrees	degrees
Accident_Longitude	float	Longitude coordinate of the accident location.	Not Null, Required	No	Decimal degrees	degrees
Accident_Altitude	float	Altitude of the accident location, if available.	Not Null	No	Decimal meters	meters
Collision_Angle	float	Angle at which the collision occurred.	Not Null, Required	No	Degrees	degrees
Vehicle_Speed	float	Speed of the vehicle at the time of the accident.	Not Null, Required	No		mph
Repair_Cost_Estimate	float	Estimated repair cost provided in the claim.	Not Null, Required	No		USD
Labor_Cost_Estimate	float	Estimated labor cost for repairs.	Not Null, Required	No		USD
Parts_Cost_Estimate	float	Estimated cost for parts required for the repair.	Not Null, Required	No		USD

Reported_Damage_Severity	string	Severity of the damage as reported in the claim.	Not Null, Required	No		
Vendor_ID	string	Unique identifier for the repair vendor handling the claim.	Not Null, Required	No		
Vendor_Avg_Repair_Cost	float	Average repair cost typically charged by the vendor.	Not Null, Required	No		USD
Policy_Coverage_Type	string	Type of policy coverage.	Not Null, Required	No		
Claim_Submission_Timestamp	datetime	Date and time when the claim was submitted.	Not Null, Required	No	YYYY-MM-DD HH:MM:SS	
Submission_Delay	float	Time delay between accident occurrence and claim submission in hours.	Not Null, Required	No		hours
Historical_Claim_Frequency	integer	Number of claims filed by the policyholder per year.	Not Null, Required	No		claims/year
Historical_Avg_Claim_Cost	float	Average claim cost historically incurred by the policyholder.	Not Null, Required	No		USD
Duplicate_Claim_Flag	boolean	Indicator whether the claim is a duplicate of a previously filed claim.	Not Null, Required	No		
External_Benchmark_Repair_Cost	float	External benchmark cost for repairs in similar accident scenarios.	Not Null, Required	No		USD

Cost_Ratio	float	Ratio of the claimed repair cost to the external benchmark repair cost.	Not Null, Required	No		
Adjusted_Damage_Severity	string	Severity of damage after adjustments based on repair estimates.	Not Null, Required	No		
Similar_Claims_Count	integer	Count of similar claims identified based on matching criteria.	Not Null, Required	No		claims
Fraudulent_Claim_Flag	boolean	Indicator of whether the claim is confirmed or suspected to be fraudulent.	Not Null, Required	Yes		
Denied_Claims_History	integer	Count of previously denied claims for the policyholder.	Not Null, Required	No		claims
Policyholder_Age	integer	Age of the policyholder.	Not Null, Required	No		years
Policyholder_Gender	string	Gender of the policyholder.	Not Null, Required	No		
Policyholder_Income	float	Annual income of the policyholder.	Not Null	No		USD
Policyholder_Employment_Status	string	Employment status of the policyholder.	Not Null	No		
Policyholder_License_Age	float	Number of years the policyholder has held a driving license.	Not Null, Required	No		years

Number of records:

Number of records to generate: 100000

Column Config

Data Element Name	Values	Distribution
Claim_ID	N/A	N/A
Policyholder_ID	N/A	N/A
Accident_Timestamp	2023-01-01T00:00:00 to 2023-12-31T23:59:59	Uniform
Accident_Latitude	-90 to 90	Uniform
Accident_Longitude	-180 to 180	Uniform
Accident_Altitude	-430 to 8848	Uniform
Collision_Angle	0 to 360	Uniform
Vehicle_Speed	0 to 200	Uniform
Repair_Cost_Estimate	0 to 50000	Gamma
Labor_Cost_Estimate	0 to 10000	Gamma
Parts_Cost_Estimate	0 to 40000	Gamma
Reported_Damage_Severity	Low, Medium, High	N/A
Vendor_ID	N/A	N/A
Vendor_Avg_Repair_Cost	0 to 30000	Gamma
Policy_Coverage_Type	Comprehensive, Collision, Liability	N/A
Claim_Submission_Timestamp	2023-01-01T00:00:00 to 2023-12-31T23:59:59	Uniform
Submission_Delay	0 to 720	Gamma
Historical_Claim_Frequency	0 to 10	Poisson
Historical_Avg_Claim_Cost	100 to 50000	Gamma
Duplicate_Claim_Flag	true (2%), false (98%)	Binomial
External_Benchmark_Repair_Cost	100 to 50000	Gamma
Cost_Ratio	0 to 10	Uniform
Adjusted_Damage_Severity	Low, Medium, High	N/A
Similar_Claims_Count	0 to 50	Poisson

Fraudulent_Claim_Flag	true (12.8%), false (87.2%)	Binomial
Denied_Claims_History	0 to 5	Poisson
Policyholder_Age	16 to 100	Uniform
Policyholder_Gender	Male, Female, Other	N/A
Policyholder_Income	10000 to 1000000	Gamma
Policyholder_Employment_Status	Employed, Unemployed, Retired	N/A
Policyholder_License_Age	0 to 70	Uniform