Fraud Detection ML Model Documentation

Real-time ensemble system for transaction fraud prevention

Version: 1.0

Author: AI/ML Engineering Team

Category: Machine Learning | Fraud Detection

Last Updated: December 12, 2025

This document provides comprehensive technical documentation for Monzo's real-time fraud detection machine learning model. This system processes all transactions through ensemble ML models to identify fraudulent activity patterns, achieving 99.2% legitimate transaction approval rate while detecting 87% of fraud attempts within milliseconds.

Model Overview

Model Name: Monzo Fraud Detection System v4.8

Model Type: Ensemble (Random Forest + Gradient Boosting + Neural Network)

Purpose: Real-time fraud detection and transaction risk scoring

Deployment Date: November 2024

Latency Requirement: <50ms per transaction (P95)

Owner: Financial Crime & ML Engineering Team

Regulatory Classification: High-risk AI system under FCA requirements

System Architecture

The fraud detection system employs a three-stage ensemble architecture combining multiple model types for optimal accuracy and speed:

Stage 1: Fast Rules Engine (5ms)

Deterministic rules catch obvious fraud patterns instantly:

• Card not activated but transaction attempted

• Geographic impossibility (two transactions in distant locations within minutes)

• Blocked merchant category codes

• Spending limits exceeded

• Known compromised merchant detection

Action: Immediate block for ~8% of transactions, remaining 92% proceed to ML models

Stage 2: Random Forest Classifier (15ms)

Parameter	Value	Rationale
Algorithm	Random Forest	Fast inference, handles non-linear patterns
Number of Trees	300	Balanced accuracy vs. latency
Max Depth	12	Captures complex fraud patterns
Features Used	187	Transaction + behavioral + network features
Min Samples Split	50	Prevents overfitting on rare patterns
Feature Sampling	0.7	Randomization for robustness

Stage 3: Gradient Boosting + Neural Network Ensemble (30ms)

For transactions not blocked by Random Forest, two additional models provide final risk scoring:

XGBoost Model: 500 trees optimized for sequential pattern detection (account takeover, testing small amounts before large fraud)

Neural Network: 5-layer feedforward network (256→128→64→32→1 neurons) specializing in anomaly detection using embeddings for merchant/location

Ensemble Logic: Weighted voting (RF: 40%, XGBoost: 35%, NN: 25%) produces final fraud probability score 0-1

Training Data Specifications

Dataset Composition

Training Period: 12 months rolling window (December 2023 - November 2024)

Total Transactions: 4.2 billion transactions

Fraud Cases: 1,847,392 confirmed fraud transactions (0.044% base rate)

Legitimate Transactions: 4,198,152,608 (99.956%)

Class Imbalance Handling: Strategic undersampling of legitimate transactions + SMOTE for fraud minority class + class weights (1:2000 ratio)

Data Labeling Methodology

Confirmed Fraud: Transactions confirmed fraudulent through customer disputes, chargebacks, or detected by Operations team

Labeling Lag: 90-day window to allow chargeback cycles to complete (affects training freshness)

False Positive Feedback Loop: Legitimate transactions blocked by model reviewed by humans; confirmed legitimate cases added to training data with high weight

Quality Control: Random 5% sample of confirmed fraud reviewed by senior analysts quarterly

> ■ ***NOTE:*** *The 90-day labeling lag means the model trains on fraud patterns 3+ months old. To compensate, we retrain weekly and heavily weight recent fraud patterns.*

Feature Engineering

The model uses 187 engineered features across five categories:

Transaction Features (48 features)

• Amount (absolute, log-transformed, bucketed)

• Currency and FX rate (if international)

• Merchant category code (MCC) + high-risk MCC flags

• Online vs. in-person transaction

• Chip + PIN vs. contactless vs. magstripe vs. CNP

• Time of day (bucketed: night 12am-6am, morning 6am-12pm, etc.)

• Day of week

Velocity Features (52 features)

• Transaction count in last 1h, 6h, 24h, 7d, 30d

• Total amount spent in last 1h, 6h, 24h, 7d, 30d

• Number of unique merchants in last 24h, 7d

• Number of declined transactions in last 24h

• Sudden spending spike vs. 30-day average

• Time since last transaction (seconds)

Geographic Features (31 features)

• Country of transaction

• High-risk country flag (based on fraud rates)

• Distance from home location (km)

• Distance from last transaction location

• Travel velocity (km/hour between transactions - detects geographic impossibility)

• First time in this country indicator

• Number of countries visited in last 7d

Behavioral Features (38 features)

• Merchant familiarity (previously used, frequency)

• MCC familiarity (spending history in this category)

• Amount deviation from typical spend at this merchant

• Card usage patterns (preferred card for online vs. in-store)

• ATM withdrawal patterns

• App session activity (active session within 1 hour indicator)

Network Features (18 features)

• Device fingerprint match with historical devices

• IP address risk score (VPN, proxy, Tor detection)

• IP geolocation mismatch with transaction country

• Device type (iOS, Android, Web)

• Merchant fraud rate (% of transactions at this merchant flagged as fraud historically)

■■ WARNING: Network features rely on device fingerprinting and IP analysis which can change legitimately (VPN usage, new device, travel). Model trained to accept these changes when other behavioral signals are normal.