Dynamic Abuse Prevention System for AV Ride-Hailing: Integrating ML-Powered Risk Scoring with an Operational Response Framework

BUSINESS PROBLEM

In traditional ride-hailing, the human driver serves as both a custodian of the vehicle and a natural deterrent to misconduct. Autonomous ride-hailing removes this oversight entirely, leading to abuse vectors including payment fraud, vandalism, denial-of-service behaviors, and unauthorized access.

Because abuse is typically identified only during post-trip or depot-level reviews, a latency gap makes attribution to a specific rider inherently difficult. Left unaddressed, these behaviors reduce vehicle availability, degrade service quality, and increase costs that compound as the fleet scales.

DATA SOURCES

Operational event logs - millions of records and thousands of tracked variables capturing the full ride lifecycle from app interaction through trip completion to post-trip activity ~120 behavioral signals selected across four abuse domains: operational availability, financial integrity, physical integrity, and trip/account patterns.

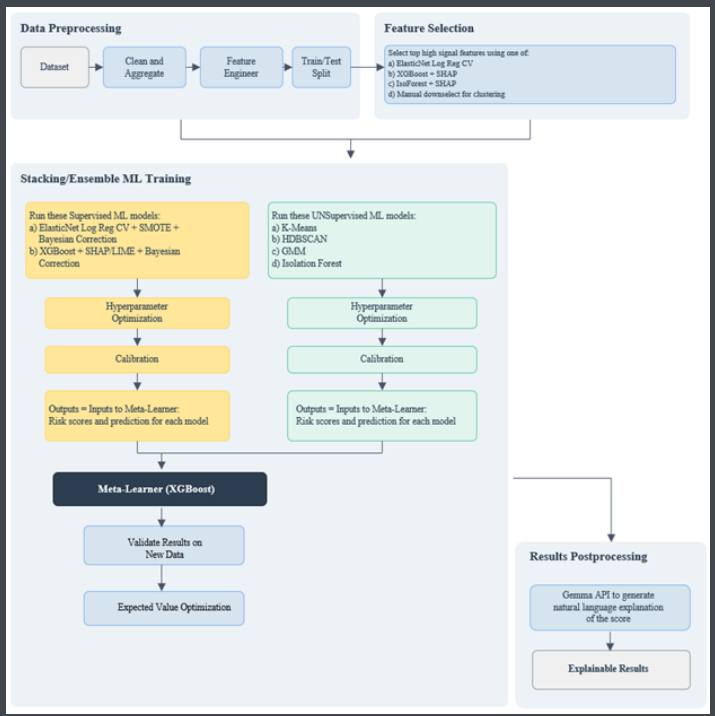
All personally identifiable information excluded.

Data Types and Format

Boolean, Categorical (nominal), Count (integer), Numeric, Relational Data, Structured, Text

APPROACH

To shift from reactive investigation to proactive prevention, we build a machine learning system that scores every rider's abuse likelihood based on behavioral history. The system combines methods that learn from confirmed abuse cases with others that detect unusual behavior deviating from normal patterns, enabling detection of both known and novel abuse types. The model trains exclusively on non-personally identifiable data from operational ride logs. An explainability module translates each risk score into a plain-language summary. Resulting risk scores feed into a tiered response framework.



Anastasiia Holubova
MIT Leaders for Global Operations, Class of 2026

Leaders for Global Operations

IMPACT

For the first time, operations and security teams gain a unified, daily view of rider risk, replacing fragmented, reactive investigation with a system that surfaces potential abuse before losses occur.

The model achieves a PR-AUC of 0.96 and ROC-AUC of 0.99, which means it reliably distinguishes abusive from legitimate riders: in practice, it correctly flags roughly 9 out of every 10 abusive riders while misidentifying fewer than 1 in 200 legitimate users. Analysis of predictions confirms that irregular, spiky patterns in how riders use the platform are the strongest predictors of abuse, more so than how frequently they ride or how they pay. Medium-risk riders receive targeted verification steps such as confirming their payment method or completing an identity check, while high-risk riders are surfaced to investigators alongside a plain-language summary explaining the specific behaviors that triggered the flag. The system is a proof-of-concept validated on historical data and has not yet been tested in a live environment. The system investment economics is very strong and justifiable with returns expected in the range of 100-200x and an almost immediate break-even period within the first month of operation. At higher actioning rates, returns grow proportionally as more abuse is prevented before it occurs. Remaining classification errors are concentrated among new accounts with limited ride history, which future iterations can address by incorporating signals specific to early account behavior. Beyond abuse prevention, risk scores can also inform product and marketing decisions such as selecting reliable user groups for new feature rollouts.

DRIVERS

With the rapid growth of autonomous ride-hailing operations teams face increasing pressure to maintain service quality and rider accountability across expanding fleets. Recent advances in ensemble machine learning, explainable AI, and real-time fraud detection frameworks now make it possible to predict and prevent platform abuse at scale.

BARRIERS

There are two main challenges. First, verified abuse constitutes less than 1% of all riders, meaning very few confirmed examples to learn from. Second, the system must rely exclusively on behavioral data for fairness, produce explanations investigators can act on, scale to millions of riders, and avoid misidentifying reliable riders.

ENABLERS

Despite the sensitivity of the problem, the company entrusted this work to the thesis project and provided support to move quickly. When data access restrictions created roadblocks, leadership resolved them promptly. Operations, engineering, and product teams collaborated to ensure the project continued beyond thesis.

ACTIONS

We first conducted deep-dives into operational data with subject matter experts to identify ~120 behavioral signals indicative of abuse. We then iteratively tested multiple detection models, combining them into a system producing a risk score and plain-language explanation per rider. Once validated, we designed the tiered response framework with operations, legal, and security teams, then produced a roadmap covering live testing and experiments.

INNOVATION

While the ML methods used are well-established, this research makes three contributions. First, it applies stacked ensemble architectures to AV ride-hailing abuse prevention, where no targeted academic research exists. Second, it achieves high accuracy using exclusively non-PII data on highly imbalanced datasets. Third, it integrates ML predictions directly into human-in-the-loop operational frameworks for large-scale commercial deployment.

IMPROVEMENT

The model catches ~88% of abusive riders, but actual prevention depends on how effectively teams act on each flag. At moderate capacity, the system prevents over half of total losses, delivering a projected 164x return in its first year with break-even in under one week, with costs driven by investigator labor not infrastructure. ~10% of abuse remains undetected due to novel patterns among new accounts, being a clear target for future iteration.

BEST PRACTICES

Invest significant time in data understanding and feature engineering before building models. With rare positive cases, addressing imbalance at every pipeline stage matters more than any single modeling choice. Each model should earn its place through measurable contribution. Engage stakeholder early, otherwise it won't survive the proof of concept

OTHER APPLICATIONS

The core methodology of combining different models into a stacked ensemble with an integrated response framework can be applied broadly to abuse detection and fraud prevention in other industries. Within the company, risk scores can also support adjacent functions such as informing feature rollout decisions and enriching customer health metrics.