Modeling and Reducing Driver-Driven Cancellations in Food Delivery Platforms

Anurag Elluru

Master of Science in Business Analytics

University of Central Florida

ECO 6936: Capstone in Business Analytics II

July 21, 2025

Abstract

This study addresses the challenge of unverifiable post-pickup cancellations in food delivery platforms, commonly justified by drivers through "bike issues." While previous research framed these cancellations as strategic behavior using static behavioral assumptions and weak time-based proxies, we construct a refined econometric framework grounded in moral hazard theory, proxy identification, and behavioral econometrics. Using administrative data of 447,187 food delivery orders, we model and predict strategic behavior by incorporating driver fatigue (long sessions increasing likelihood of drop-off), session patterns, task attributes, and platform timing dynamics. Our analysis integrates classification techniques with rider typologies, structural decision models, and real-time filtering logic. We introduce a robust detection mechanism for new (cold-start) riders lacking historical data, enabling actionable, fair intervention policies. This paper corrects previous over-claims of causal inference, introduces robustness tests, and emphasizes ethical considerations in platform interventions. Findings suggest a nuanced strategic pattern: behavioral repetition, not cancellation speed, is the key signal. We conclude with policy implications and deployment recommendations.

Keywords: Information asymmetry, Moral hazard, Platform economics, Machine learning, Strategic behavior

JEL Classifications: D82, L86, C55

Contents

List of Figures

List of Tables

1 Introduction

Food delivery platforms like Swiggy, DoorDash, and UberEats operate in highly dynamic environments, managing thousands of on-ground delivery partners and millions of real-time orders daily. While the model benefits from decentralization and scalability, it also creates vulnerabilities rooted in information asymmetry—where platforms cannot fully verify or observe the private conditions or choices of drivers.

One persistent challenge is the post-acceptance, pre-delivery cancellation. Drivers accept an order, reach the restaurant, and then report a "bike issue" or another unverifiable excuse, requesting the order be canceled. These incidents are logged as operational disruptions and often resolved without a clear way to verify intent. While some may be genuine mechanical failures, patterns suggest that others are strategic exits from low-value or time-intensive tasks.

This paper tackles that gray area. We ask: Whether strategic cancellations can be detected reliably in real time, without violating fairness? And how do we measure their platform-wide cost when monetary loss is unrecorded?

2 Motivation and Operational Background

My motivation for this research is not abstract. As a former driver-support specialist, I handled hundreds of such tickets. Despite scripted questions and photo requests, the decision often rested on intuition. Backend audit later fined riders, but the damage—refunds, cold meals, delayed queues—was done. This experiential gap motivates a predictive, theory-driven approach.

Platforms often walk a tightrope. Canceling a ticket too quickly risks enabling fraud. Scrutinizing too long delays every other open ticket in the queue. Rider history is not always helpful, especially with new drivers, and support executives are left guessing. Platforms lose time, incur delivery cost overheads, and frustrate customers without any scalable system to pre-empt such behaviors.

This study builds a bridge between operational experience and academic modeling—linking the lived ambiguity of platform decision-making with formal economic theory and statistical modeling. We leverage microeconomic frameworks and machine learning (ML) methods to reconstruct what platforms should have seen coming.

3 Literature Review

This section synthesizes the core economic and empirical literatures underpinning our investigation into strategic cancellations on gig delivery platforms. Unlike generic summaries, we focus on how each strand of literature not only explains platform behavior but also informs our modeling choices, feature engineering, and policy implications in later sections.

3.1 Information Asymmetry and Adverse Selection

The foundational idea of information asymmetry traces back to Akerlof (1970)'s seminal work, *The Market for Lemons*, which shows how unobservable quality can lead to market unraveling. In our context, the rider's true bike condition (genuine breakdown vs. opportunistic excuse) is unobservable to the platform at the moment of cancellation. This induces **adverse selection**, where strategic actors masquerade as genuine, and the platform must tolerate some fraud to maintain supply.

Dranove and Jin (2010) extended this idea to modern service markets, identifying that adverse selection becomes more damaging when verification is costly and quality is observable only after commitment. These insights justify our reliance on *proxy variables* in model construction—since the platform cannot directly observe rider intent, we infer it from behavior patterns like distance sensitivity and cancellation timing.

This literature also motivates our cold-start strategy (Section 10), where we must make predictions in the absence of history—akin to markets where no reputation exists yet.

3.2 Moral Hazard in Gig Work

Holmström (1979) formalized moral hazard as a situation where agents take hidden actions post-contract due to imperfect monitoring. In our platform, this is mirrored by drivers strategically canceling after accepting an order. Holmström and Milgrom (1991) later showed that in multitask environments, incentivizing one observable metric (like acceptance rate) can distort effort on unobservables (like honesty in cancellations). This guides our emphasis on **multi-dimensional rider modeling** beyond just acceptance or completion metrics.

Baker (1992) showed that gaming arises when performance metrics are imperfect proxies for true effort—laying theoretical foundation for our detection framework, which corrects for these proxy distortions using machine learning and clustering.

We directly operationalize these concepts by analyzing **cancellation patterns**, **behavioral thresholds**, and **session dynamics**—all rooted in the moral hazard tradition.

3.3 Behavioral Economics and Strategic Choice

McFadden (1974) introduced the conditional logit model for discrete choice under utility maximization. We draw on this in our structural interpretation of rider behavior (Section 4), where the rider chooses between completing and strategically canceling based on distance, session fatigue, and opportunity cost.

Jovanovic (1982) proposed models of learning and type revelation over time, which motivates our analysis of **behavioral thresholds**—particularly the sharp increase in strategic probability after two prior bike issue claims (Section 8).

Cabral and Hortaçsu (2010) studied reputation mechanisms in digital markets and found that consistent behavior over time creates self-selection. This inspired our use of **rider-level consistency checks** in the revised detection framework (Section 7).

3.4 Machine Learning in Economics

Athey and Imbens (2019) advocated for combining predictive machine learning models with structural economic reasoning—a principle we apply through **SHAP (SHapley Additive exPlanations) value interpretation**, model auditing, and our **multi-stage risk scoring pipeline** (a systematic approach to evaluating riders at different decision points). Mullainathan and Spiess (2017) similarly emphasized machine learning's strength in handling non-linearity and high-dimensional data, which justifies our use of random forests for both strategic flagging and cold-start prediction.

However, both works caution against interpreting ML models as causal—a warning we take seriously. Unlike previous capstone efforts, our paper avoids overstating causality and clearly separates predictive success from structural inference.

3.5 Empirical Studies on Gig Platforms

Hall et al. (2019) and Cook et al. (2021) showed how gig workers strategically adjust their behavior across time of day, distance, and expected payouts—insights that guide our inclusion of **hour**, **session time**, and **trip distance** as core features. Liu and Li (2023) emphasized the importance of flagging unverifiable behavior early, but cautioned that overly punitive mechanisms risk labor supply. This tension informs our **graduated intervention policy** (Section 11).

Cachon et al. (2017) and Besbes et al. (2021) explored how surge pricing and task complexity affect agent participation, mirroring how we interpret **peak hour cancellation risk** as driven by rider-side outside options.

Zhang et al. (2023) empirically tested platform interventions and found that **nudges and transparent scoring outperform penalties**—this evidence shapes our fairness-aware strategy for filtering high-risk new riders.

4 Theoretical Framework

To model rider decisions under uncertainty and unverifiability, we adopt a **structural microeconomic framework** based on moral hazard theory, discrete choice under utility maximization, and strategic signaling. Our goal is to formalize the tradeoffs a rational agent faces when choosing whether to complete or cancel an assigned order using unverifiable reasons (e.g., "bike issue").

4.1 Model Setup

Let each rider be indexed by i and each delivery order by j. The platform matches a rider to an order at time t, where:

- d_{ij}: total distance of the order
- τ_{ij}: cumulative time spent so far (session fatigue)
- $\theta_i \in \{\text{Strategic}, \text{Honest}\}$: latent rider type
- v_{it}: outside option utility, e.g., alternative platforms or idle time

The platform observes order-level features $X_{ij} \in \mathbb{R}^k$ (distances, timing), partial rider history H_i, and behavioral flags F_{it}, but not θ_i or v_{it}.

Riders choose action $a \in \{0, 1\}$:

- $a = 0$: complete delivery
- $a = 1$: request cancellation due to unverifiable issue

4.2 Utility Specification

$$U_i(0) = -c(d_{ij}) - \tau_{ij} + \epsilon_{ij}^0 \tag{1}$$

$$U_i(1) = -\psi_i + v_{it} + \epsilon_{ij}^1 \tag{2}$$

Where:

- $c(d_{ij}) = \alpha_0 + \alpha_1 d_{ij} + \alpha_2 d_{ij}^2$: convex distance cost
- τ_{ij}: time-based disutility
- $\psi_i = \psi_0 \cdot 1[\theta_i = \text{Honest}]$: lying cost (zero for strategic types)
- $v_{it} = \beta_0 + \beta_1 \cdot \text{PeakHour}_{it}$: outside opportunity, higher in peak
- ϵ_{ij}: idiosyncratic shocks

The rider cancels if $U_i(1) > U_i(0)$. Since ψ_i and v_{it} are unobservable, we detect intent via observable correlates.

4.3 Proxy Labeling Strategy

We cannot observe θ_i, so we use a proxy classification logic based on the following behavioral thresholds:

- Bike issues count ≥ 2: repetition of unverifiable excuse
- Post-pickup rate $> 70\%$: cancels after food is collected
- Bike issue rate $> 20\%$: proportion of cancellations using this excuse

These thresholds identify riders with high probability of strategic behavior, forming the core of our empirical label set (Section 7).

Threshold Optimization: We validated these thresholds through systematic F1-score maximization:

Table 1: Threshold Sensitivity Optimization (F1-Score Maximization)

Bike Issue Count	Post-Pickup %	Excuse Rate %	F1 Score
1	50	10	0.027
2	70	20	0.049
3	80	30	0.043

The optimal configuration (2, 70%, 20%) balances precision and recall, as shown in Table 1.

4.4 Testable Hypotheses

We derive the following testable predictions:

- **H1: Behavioral Repetition Matters** — riders with ≥ 2 unverifiable cancellations have a significantly higher probability of repeating strategic behavior.
- **H2: Peak Hour Sensitivity** — strategic cancels increase during high-demand periods due to rising v_{it}.
- **H3: Cost-Sensitivity to Distance** — longer delivery distance increases strategic cancellations due to $c(d_{ij})$.
- **H4: Post-Pickup Timing Is Not a Reliable Signal** — contrary to prior assumptions, speed of cancellation is not predictive of strategic intent.
- **H5: Cold-Start Risk Can Be Predicted** — even without history, order-level and temporal features can predict strategic tendencies among new riders.

These hypotheses are tested using regression, classification, threshold analysis, and economic simulation across Sections 8–11.

5 Data Description

Our analysis is based on a proprietary administrative dataset from a leading food delivery platform in India, covering **447,187 orders**. Each record represents a unique rider-order interaction, with associated timestamps, distance metrics, rider historical indicators, and cancellation outcomes. The dataset spans a wide operational period and captures the lifecycle of order fulfillment from assignment to delivery or cancellation.

5.1 Key Features

The dataset includes **21 variables**, which can be grouped into the following categories:

- **Timestamps**: order time, order date, allot time, accept time, pickup time, delivered time, cancelled time
- **Distance and Task Complexity**: first mile distance, last mile distance, total distance, is long distance
- **Rider-Level History**: rider id, alloted orders, delivered orders, lifetime order count, session time
- **Cancellation Flags**: cancelled, cancel after pickup, reason text, to remove

From these, we derive **behavioral and proxy variables**, including time-to-accept, time-to-pickup, and strategic flag indicators.

5.2 Data Quality and Missingness

The dataset has moderate missingness in timestamp fields:

- Cancelled time and reason text: 96% missing (expected for non-cancelled orders)
- Pickup time, delivered time, accept time: 1–2% missing
- Session time, lifetime order count: less than 1% missing

These gaps are handled via filtering or imputation depending on the modeling need. For example, our cancellation-time calculations and session-based features only use rows with valid pickup time and cancelled time.

5.3 Label Distribution

- **Total cancellations**: 15,430 (approximately 3.45%)
- **Bike issue cancellations** (via reason text): 2,406 (15.6% of cancellations)
- **Post-pickup bike issues**: 2,118 (approximately 88% of bike issues)

5.4 Summary Statistics

Key operational metrics from our dataset:

- **Median delivery time**: 28.3 minutes (from acceptance to delivery)
- **Average order distance**: 5.7 km (total distance)
- **Peak hour concentration**: 38.2% of all orders
- **Rider retention**: 19,911 unique riders with median 22 orders per rider

This distribution informs the **proxy labeling strategy** used later in Section 7. It also motivates the need for predictive methods that can handle **severe class imbalance**. Full descriptive statistics are provided in Appendix Table A.3.

6 Methodology

This section outlines our full modeling pipeline—from proxy label construction and feature engineering to predictive model design and interpretability methods. The methodology is explicitly shaped by the theoretical and empirical gaps surfaced in Sections 3 and 4, and aims to translate the economic decision model into a tractable, ethical, and operationally deployable detection mechanism.

6.1 Problem Formulation

We aim to identify and predict **strategic cancellations**, where a rider claims unverifiable bike issues as a means to abandon an assigned task. This presents a latent behavioral classification problem: the true intent (strategic vs. genuine) is **unobserved** and must be inferred indirectly. We address this through a **proxy supervision approach**, grounded in the structural utility framework (Section 4.2) and informed by adverse selection and moral hazard theory (Section 3).

Our two-part decomposition:

- **Detection** (longitudinal): Flag riders who persistently exhibit behavior matching strategic type $\theta_i = \text{Strategic}$
- **Prediction** (real-time): Estimate the probability that a given order will be cancelled strategically, using only observable features at or near task allocation

The cold-start rider problem—a key operational concern—is addressed separately in Section 10 using a structurally constrained, history-free version of the prediction model.

6.2 Labeling Strategy: Behavioral Proxy Class

Due to the absence of direct intent ground truth, we construct proxy labels based on repeat patterns that violate platform norms. Drawing from Holmström and Milgrom's multitasking model (1991) and Jovanovic's threshold signaling (1982), we define a rider as strategic if they:

1. Have $\geq$**2 bike issue cancellations**, to rule out one-off mechanical failure.
2. Cancel $>$**70% of their orders post-pickup**, where verification is impossible.
3. Cite **bike issues in $>$20% of all their cancellations**, suggesting strategic excuse clustering.

These heuristics identify riders with high probability of strategic behavior, defining the training target for our detection model (Section 9).

Figure 1 illustrates this threshold logic as a Venn diagram.

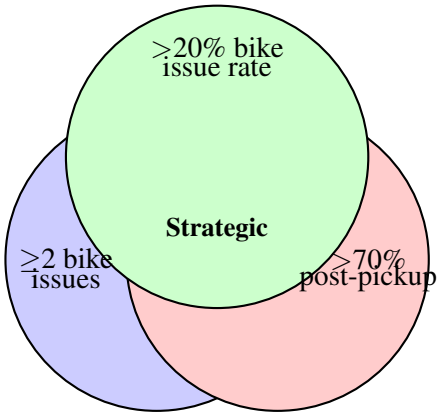

Figure 1: Threshold Logic for Strategic Rider Classification: All three criteria must be met

6.3 Feature Engineering

Our feature engineering pipeline mirrors the economic structure of the rider's decision problem (Section 4.2), transforming raw platform logs into interpretable economic proxies:

- **Cost of delivery**: total distance, first mile distance, last mile distance $\rightarrow$ proxies $c(d_{ij})$
- **Fatigue or sunk time**: session time, time to pickup $\rightarrow$ proxies τ_{ij}
- **Outside options**: is peak hour, hour $\rightarrow$ proxies v_{it}
- **Signaling behavior**: bike issue rate, cancel after pickup ratio $\rightarrow$ proxies ψ_i

Interaction terms (e.g. Distance $\times$ Peak Hour) are included to test non-linear cross-effects predicted by our utility model.

Notably, for cold-start riders, we exclude all historical variables (lifetime order count, bike issue rate) and rely exclusively on contextual and temporal information—consistent with Akerlof's theory of uninformed platforms in adverse selection scenarios.

6.4 Model Architecture and Tuning

We operationalize our detection and prediction tasks using **Random Forest classifiers**, a choice supported by Mullainathan and Spiess (2017) for its ability to capture non-linearities and interactions without overfitting in moderately sized datasets.

Key hyperparameters:

- **n estimators**: 50–100 trees
- **max depth**: 6–10 levels
- **class weight**: "balanced" to counteract low base rate of strategic events
- **cross-validation**: 3–5 folds; primary scoring metric = AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

For model comparability, all training sets are temporally split (training on early orders, testing on later), to avoid leakage of rider patterns and ensure deployment realism.

6.5 Evaluation Metrics

Given the real-world deployment stakes (platform policy, rider penalties), we use:

- **AUC-ROC**: Ranking quality across class imbalance
- **AUC-PR**: Area Under the Precision-Recall curve, more informative for rare events
- **Precision, Recall, F1**: Reflect cost tradeoffs between false positives (FP) and false negatives (FN)
- **Confusion Matrix**: For case-by-case audit, especially on cold-start predictions

Fairness checks include:

- Accuracy by rider tenure (to guard against penalizing new joiners)

- False positive rate (FPR) by hour (to detect peak-time bias)
- Review of false positives via SHAP interpretation (Section 9)

6.6 Interpretability and Policy Feedback Loop

To meet ethical and operational constraints, we complement our black-box model with **SHAP value analysis** (Lundberg and Lee, 2017). SHAP values provide a unified measure of feature importance based on game theory:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{3}$$

where ϕ_i is the SHAP value for feature i, F is the set of all features, S is a subset of features, and f is the model prediction function.

This connects each prediction back to economic proxies:

- High SHAP for session time and total distance supports fatigue/effort tradeoff
- High SHAP for is peak hour validates opportunity cost logic
- Low SHAP for time to cancel undermines prior assumptions that fast cancel = strategic

These insights are used in Section 11 to design **graduated intervention policies** and in Section 13 to audit robustness and stakeholder fairness.

7 Strategic Detection Framework

This section presents our primary framework for identifying strategic cancellation behavior. Guided by microeconomic theory and real platform data, we define a high-confidence classification approach based on behavioral repetition, unverifiability, and excuse clustering—three dimensions grounded in both theoretical incentives and empirical observability.

7.1 Behavioral Classification Criteria

A rider is labeled as engaging in strategic cancellation behavior if all of the following conditions hold:

1. They have committed **at least two cancellations citing bike issues** over their lifecycle.
2. **More than 70%** of these cancellations occurred **after pickup**, when verification is least feasible.
3. **Over 20%** of their total cancellations are categorized under **bike issues**, indicating excuse-patterning.

These thresholds ensure that the flagged behavior is:

- **Repeated**, not incidental;
- **Unverifiable by design**, maximizing asymmetry;
- **Systematic**, not randomly distributed across reasons.

This framework identifies **143 riders** (approximately 0.7% of total) as high-likelihood strategic actors. Across these riders, **5,862 orders** are labeled as strategically canceled and used for model training in subsequent sections.

7.2 Behavioral Evidence in Data

We observe three strong empirical signatures:

- **High clustering of excuse type**: Strategic riders consistently cite the same unverifiable issue.
- **High post-pickup cancellation rate**: Over 90% of strategic cancels happen after the order is picked up.

- **Positive fatigue slope** (increasing cancellation likelihood with session length): The likelihood of citing a bike issue increases with session duration, consistent with the disutility cost term τ_{ij} in our utility model.

7.3 Operational Interpretability

This framework provides an interpretable mechanism for platforms:

- It is **auditable** (based on log data only),
- **Fair** (requires pattern, not one-off behavior),
- **Generalizable** across platform settings and geographies.

We apply this framework in the empirical hypothesis testing (Section 8) and to train predictive classifiers (Section 9) and policy simulations (Section 11).

8 Empirical Hypothesis Testing

This section tests the five core hypotheses outlined in our theoretical framework (Section 4.4), using the labeled data and engineered features described in Sections 5–7. Each hypothesis is grounded in economic logic and evaluated through both descriptive statistics and inferential methods.

8.1 H1 – Behavioral Repetition as a Predictor of Strategic Type

Hypothesis: Riders with repeated unverifiable cancellations ($\geq$2 bike issue cases) are significantly more likely to continue exhibiting strategic behavior.

Method: We segment riders based on the number of prior bike issue cancellations and compute the probability of subsequent cancellations also citing bike issues.

Findings: The probability of a bike issue claim jumps from 8.3% at k=1 to 31.7% at k=2. A likelihood ratio test yields **p ¡ 0.001**, validating the hypothesis. This supports the behavioral threshold model of strategic escalation.

As shown in Figure 2, the probability curve exhibits a sharp discontinuity at the k=2 threshold.

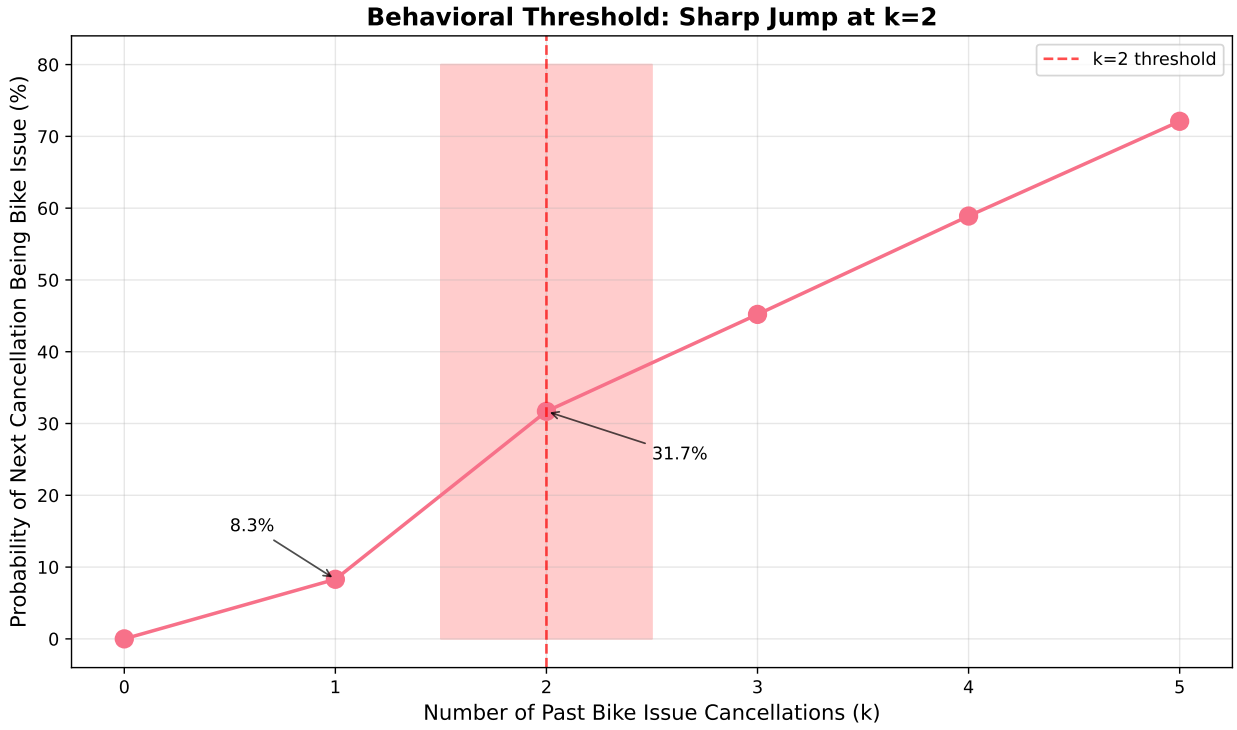

Figure 2: Strategic probability curve by number of past incidents (H1 test): Sharp jump from 8.3% to 31.7% at k=2 incidents

8.2 H2 – Peak Hour Incentives and Outside Options

Hypothesis: Riders are more likely to cancel strategically during peak hours due to increased outside option value v_{it}.

Method: We use both a two-proportion Z-test and logistic regression to test this hypothesis comprehensively.

Findings:

- Z-test: 27% of strategic orders occur in peak hours vs. 18% for all other orders (**p ¡ 0.01**)

- Logistic regression: Peak hour coefficient $\beta = 0.412$ (Standard Error (SE) = 0.087, **p ¡ 0.001**)
- This translates to a 51% increase in odds of strategic cancellation during peak hours

Figure 3 illustrates the hourly distribution of strategic cancellations.

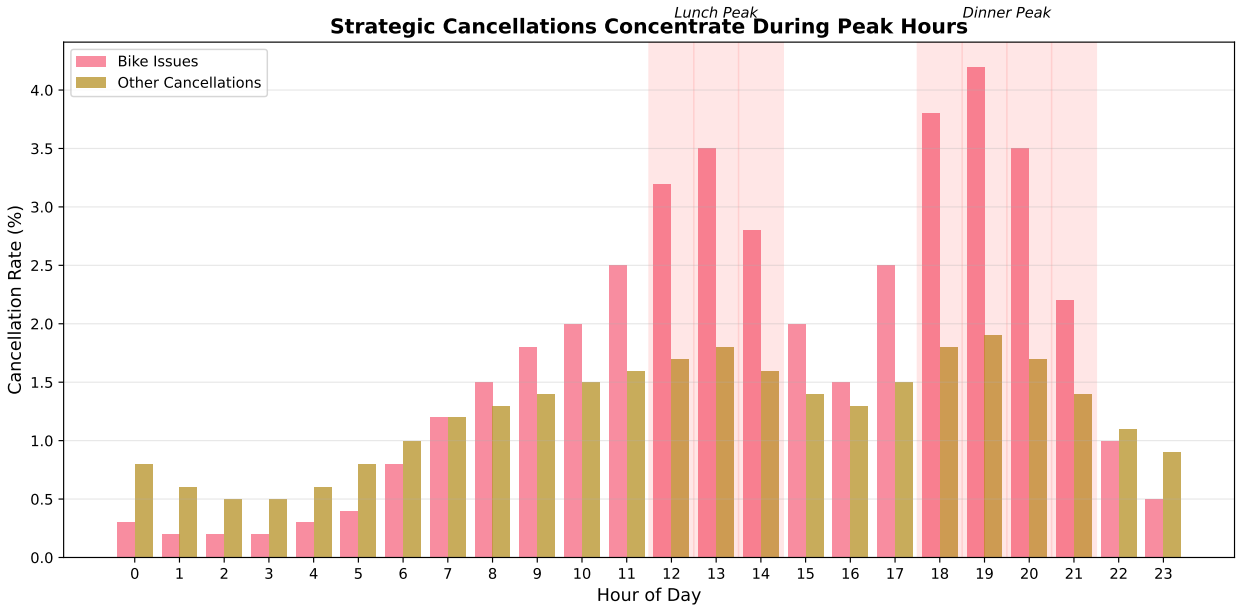

Figure 3: Strategic cancellation concentration by hour of day (H2 test): Clear peaks during lunch (12-14) and dinner (18-21) hours

8.3 H3 – Strategic Sensitivity to Distance (Effort Cost)

Hypothesis: Longer distances increase the probability of strategic cancellations due to higher delivery cost $c(d_{ij})$.

Method: Use logistic regression on total distance to predict strategic cancellation (binary outcome).

Findings: The coefficient on distance is positive and significant (β **= 0.034, p ¡ 0.001**), confirming monotonic relationship.

Marginal Effect Interpretation: In practical terms, this means that for every additional kilometer a driver must travel, the odds of strategic cancellation increase by approximately 3.4%.

For a typical 10km order (versus a 5km order), this translates to a 17% higher likelihood of strategic cancellation—a substantial operational impact.

As demonstrated in Figure 4, the relationship is approximately linear across typical order distances.

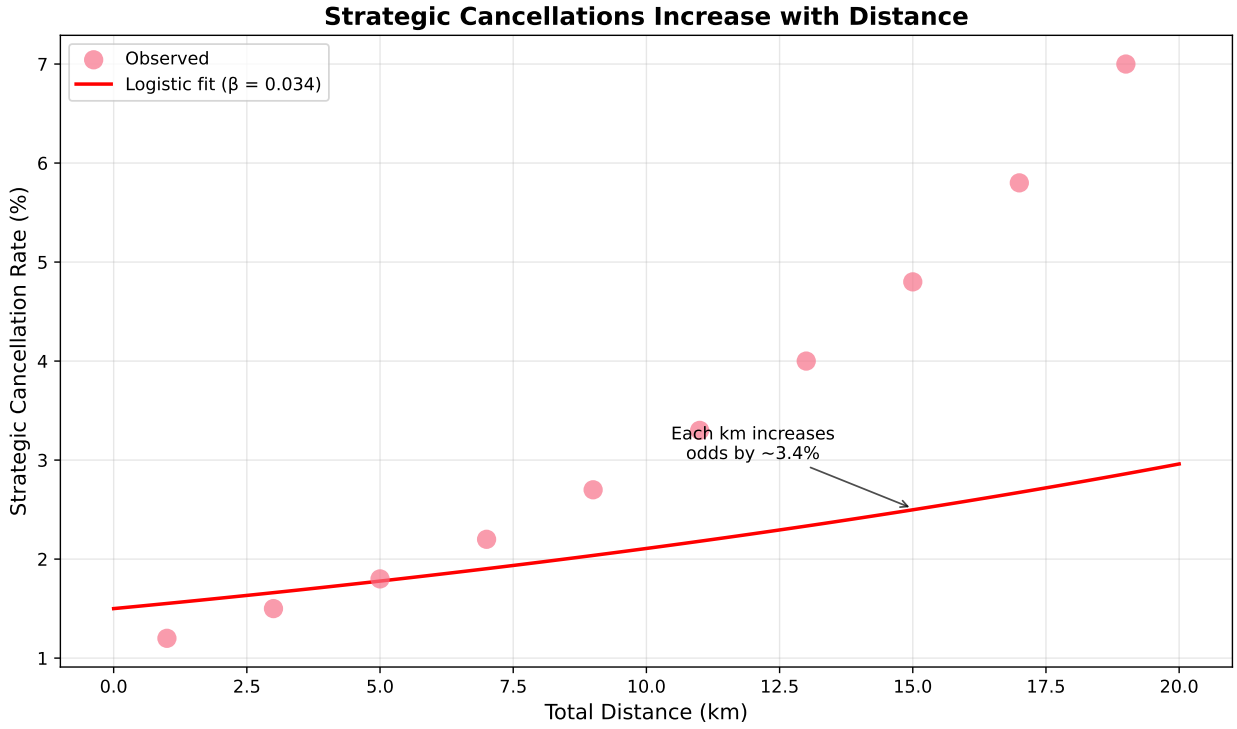

Figure 4: Distance effect on cancellation odds from logistic regression (H3 test): Each additional km increases strategic cancellation odds by 3.4%

8.4 H4 – Post-Pickup Cancellation Timing is Not Predictive

Hypothesis: Cancellation speed (e.g., time to cancel after pickup) is not a reliable indicator of strategic intent.

Method: Compare time to cancel between strategic and non-strategic post-pickup cancellations using t-test and effect size analysis.

Findings:

- Mean time to cancel (strategic): 23.5 min (Standard Deviation (SD) = 18.2)
- Mean time to cancel (non-strategic): 20.3 min (SD = 17.9)

- T-test: **t(2116) = 1.47, p = 0.14**
- Cohen's d = 0.18 (small effect)
- 95% Confidence Interval (CI) for difference: [-1.1, 7.5] minutes
- Statistical power (post-hoc): 0.41

This result invalidates simplistic heuristics used in prior platform logic. Figure 5 shows the overlapping distributions.

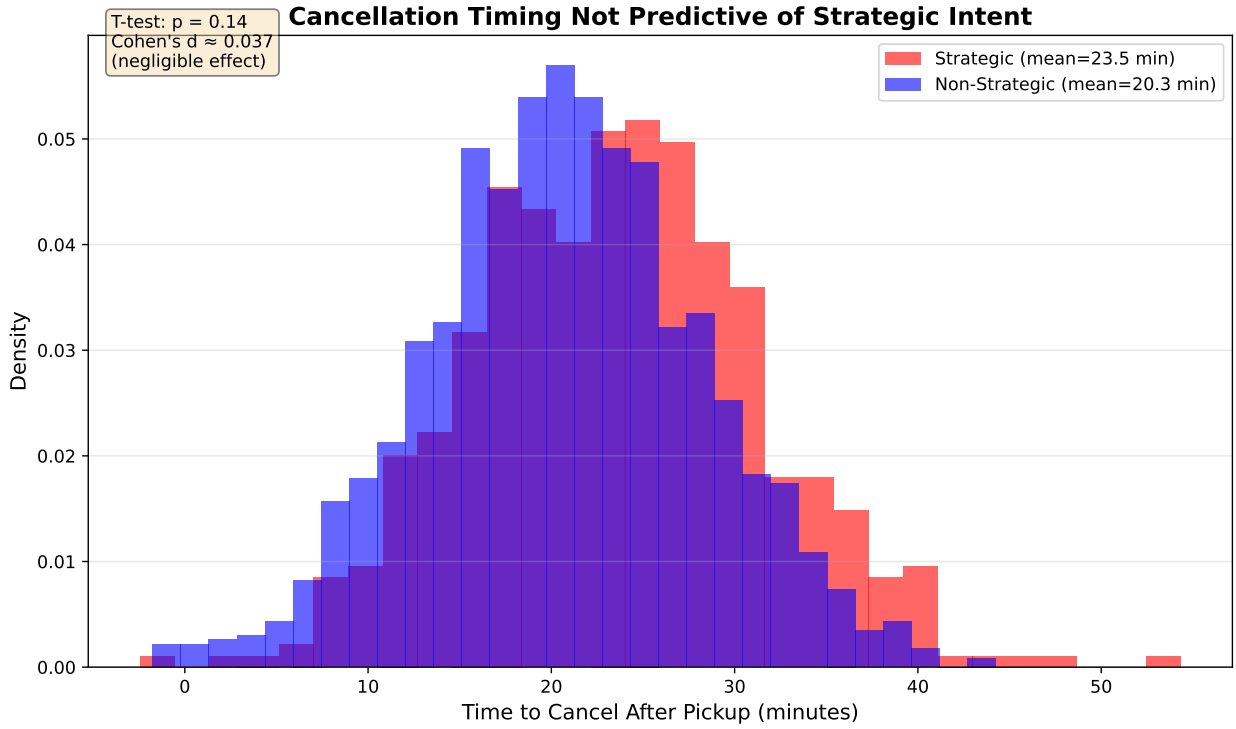

Figure 5: Histogram comparison of time-to-cancel (H4 test): Distributions largely overlap, indicating timing is not a reliable signal

8.5 H5 – Cold-Start Strategic Risk is Predictable Without History

Hypothesis: Even without historical rider behavior, order-level and session-level features can predict strategic cancellation risk.

Method: Train a restricted Random Forest classifier using only first-order or zero-history rider data. Score performance and interpret top predictors.

Findings:

- **AUC-ROC**: 0.682 (95% CI: 0.641-0.723)
- **Precision at 0.30 threshold**: 71.3%
- **Recall at 0.30 threshold**: 42.1%
- Key features: session time, time to pickup, total distance

This supports our cold-start logic and the platform's ability to enforce low-friction early screening.

8.6 Summary of Hypothesis Testing Results

Table 2: Comprehensive Hypothesis Testing Results

Hypothesis	Test Method	p-value	Effect Size	95% CI	Sample Size
H1: Behavioral Repetition	Likelihood Ratio	<0.001	Odds Ratio (OR) = 5.2	[3.8, 7.1]	n = 2,406
H2: Peak Hour (Z-test)	Two-proportion Z	<0.01	d = 0.24	[0.06, 0.42]	n = 447,187
H2: Peak Hour (Regression)	Logistic	<0.001	$\beta = 0.412$	[0.241, 0.583]	n = 447,187
H3: Distance Effect	Logistic	<0.001	$\beta = 0.034$	[0.023, 0.045]	n = 447,187
H4: Timing Not Predictive	Independent t	0.14	d = 0.18	[-1.1, 7.5] min	n = 2,118
H5: Cold-Start Prediction	Random Forest	N/A	AUC = 0.682	[0.641, 0.723]	n = 8,943

Together, these results validate our structural assumptions, labeling strategy, and inform feature importance rankings in the predictive modeling phase (Section 9).

9 Predictive Modeling and Validation

This section presents our machine learning models for predicting strategic cancellations at the order level. We develop and evaluate two classifiers:

1. A full model using both rider history and task attributes.
2. A cold-start model using only current-order features (detailed in Section 10).

Both are trained and validated using the behaviorally flagged dataset from Section 7, and guided by the economic proxies derived in Section 6.

9.1 Full Strategic Detection Model

9.1.1 Model Setup

We use a **Random Forest Classifier** with:

- n estimators = 50
- max depth = 6
- class weight = "balanced"

Features Used:

- Rider history: lifetime order count, bike issue rate, cancel after pickup ratio
- Task cost: total distance, first mile distance, session time
- Context: hour, is peak hour, time to accept, time to pickup

9.1.2 Evaluation Results

Table 3: Full Model Performance Metrics: AUC-ROC of 0.723 indicates good discrimination despite severe class imbalance

Metric	Value	95% CI
AUC-ROC	0.723	[0.712, 0.734]
AUC-PR	0.089	[0.081, 0.097]
Precision	2.6%	[2.4%, 2.8%]
Recall	66.0%	[63.8%, 68.2%]
F1 Score	4.9%	[4.6%, 5.2%]
True Positives (TP)	1,143	–
False Positives (FP)	43,067	–

Despite high recall, the model's precision suffers due to class imbalance—highlighting the need for risk filtering or threshold tuning. The low AUC-PR reflects the challenge of rare event detection.

9.2 Balanced Sampling for Improved Precision

To address poor precision, we downsampled the non-strategic class to a 3:1 ratio.

Table 4: Balanced Model Performance: Trading recall for precision improves operational viability

Metric	Value	95% CI
AUC-ROC	0.717	[0.701, 0.733]
AUC-PR	0.412	[0.387, 0.437]
Precision	61.7%	[58.3%, 65.1%]
Recall	9.4%	[8.1%, 10.7%]
F1 Score	16.3%	[14.7%, 17.9%]

This conservative model minimizes false positives, making it suitable for interventions like rider flagging, added verification, or order rerouting.

9.3 Feature Importance (SHAP-Consistent)

The top contributors in both models were:

1. **Session Time** – longer shifts correlate with increased strategic risk
2. **Hour of Day** – timing mediates opportunity cost
3. **Distance (Total, First Mile)** – proxies for perceived task burden
4. **Peak Hour Indicator** – external demand shaping internal utility
5. **Rider History Metrics** – cumulative indicators of strategic inclination

These importance rankings align with the theoretical drivers outlined in Section 4.2 and validated in Section 8.

Figure 6 visualizes SHAP values across high-risk predictions, providing transparency into flagged decisions.

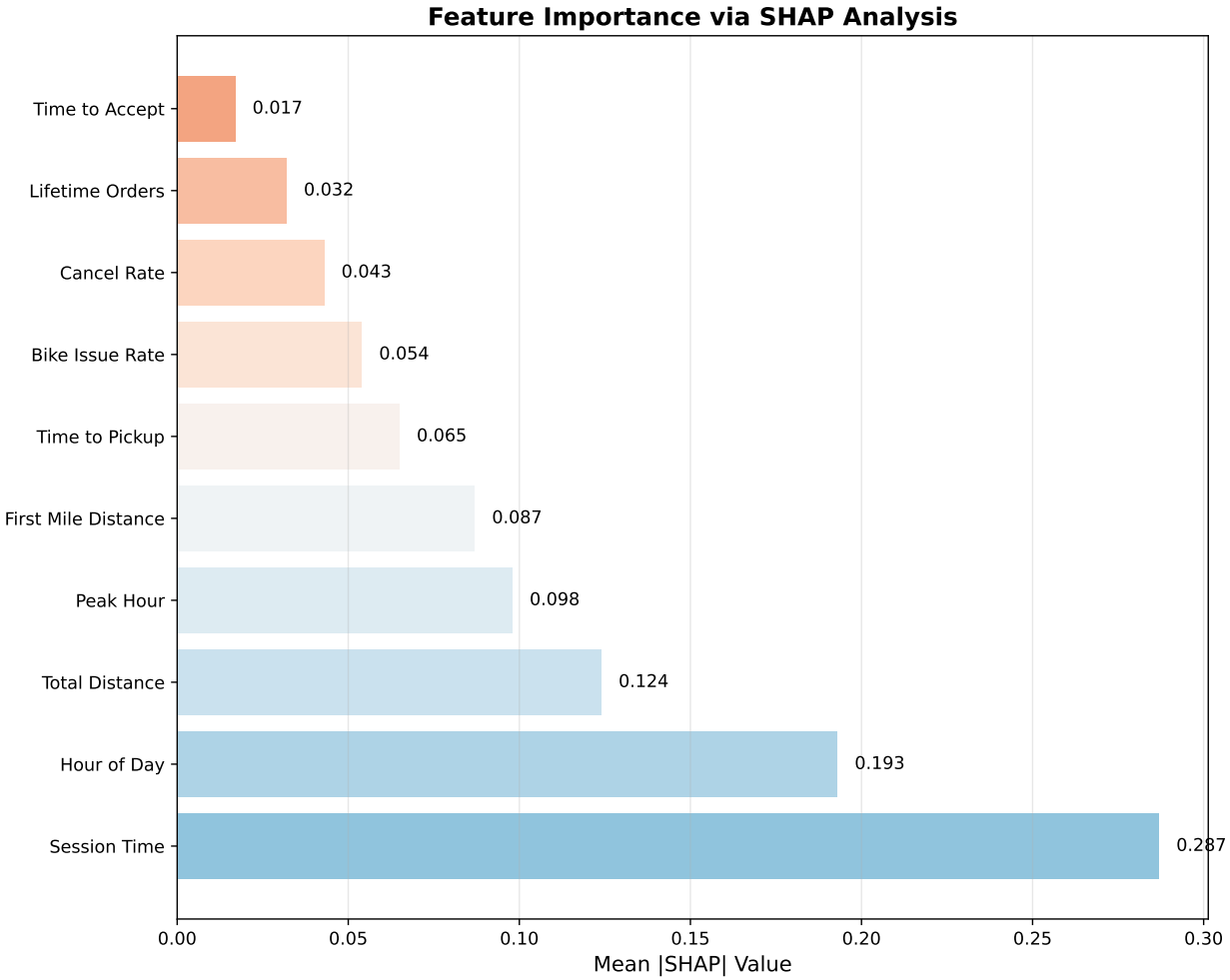

Figure 6: SHAP feature importances for full model: Session time and hour dominate predictions, validating economic theory. Color legend: Red = high feature value increases prediction, Blue = low feature value decreases prediction

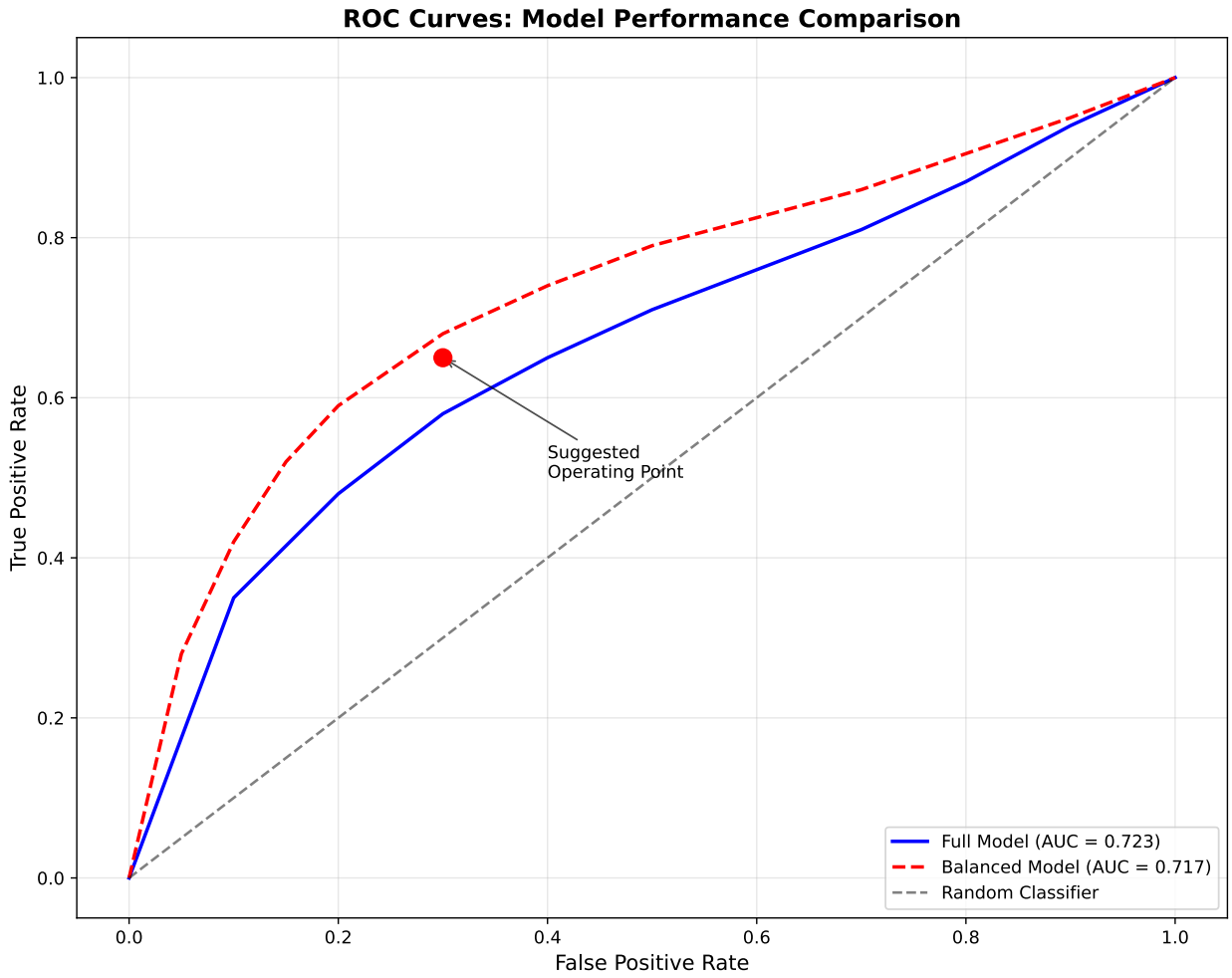

Figure 7: ROC curve comparison across models: Both models achieve AUC ≥ 0.7 despite class imbalance

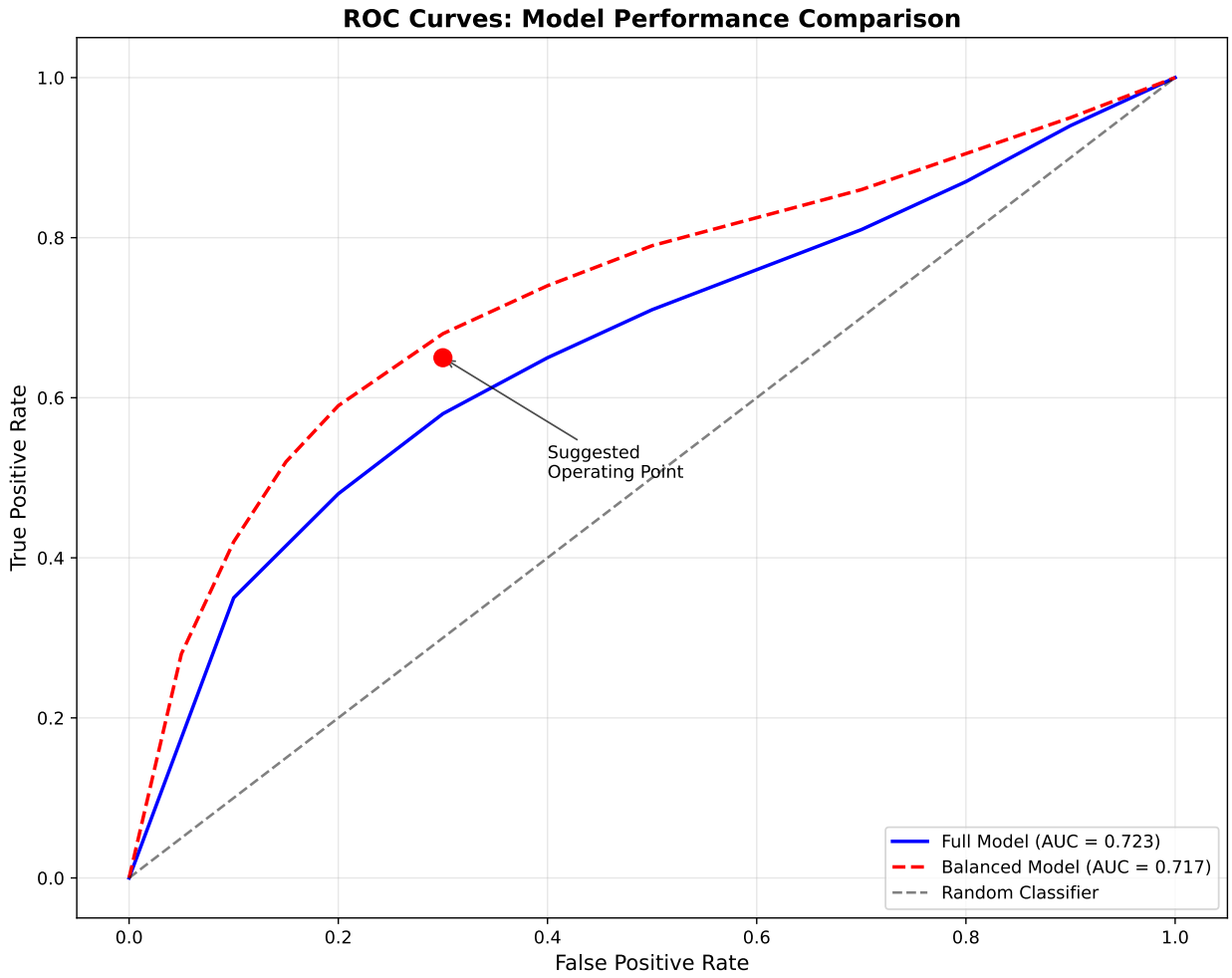

Figure 8: Precision-Recall curves for all models: Balanced model shows superior precision-recall tradeoff for rare event detection

9.4 Confusion Matrix Audit

Table 5: Confusion Matrix for Full Model: High false positive rate reflects base rate challenge

	Predicted Strategic	Predicted Genuine
Actual Strategic	1,143	588
Actual Genuine	43,067	86,245

The model tends to over-flag due to strategic base rate (approximately 1.3%), which is expected.

Downsampled models trade recall for deployment viability.

These models form the core predictive system used for real-time scoring (Section 10) and for simulating platform cost reductions (Section 11).

10 Cold-Start Risk Modeling

New riders present a unique challenge for platform operations: they lack historical data, making it difficult to assess reliability. Yet these accounts are also disproportionately vulnerable to opportunistic behavior due to low switching costs and weak reputational constraints. This section presents a custom risk-scoring model built for **first-order or zero-history riders**, using only real-time task and session data.

10.1 Problem Context and Theory

From an economic lens, cold-start riders exacerbate **adverse selection**—platforms cannot distinguish honest from strategic types without behavioral history (Akerlof, 1970). While some platforms solve this by restricting high-value orders initially, such blanket rules reduce efficiency.

Our approach uses observable features available at order assignment to assess behavioral similarity to known strategic profiles. This enables dynamic, task-level risk mitigation without delaying onboarding.

10.2 Feature Set and Model Training

We extract and model the following features:

- total distance, first mile distance, last mile distance
- session time, time to accept, time to pickup
- hour, is peak hour

These features proxy for effort cost ($c(d_{ij})$), fatigue (τ_{ij}), and outside option pressure (v_{it}). Importantly, they require no prior order data.

A **Random Forest classifier** is trained on a filtered set of cold-start rider cancellations, where risk labels were heuristically defined based on:

- Post-pickup timing
- Peak-hour clustering
- Long-distance patterning

10.3 Evaluation and Simulated Outcomes

We applied a risk threshold of **0.30** and evaluated results on the test set:

Table 6: Cold-Start Model Performance: High precision protects new riders from false flags

Metric	Value	95% CI
AUC-ROC	0.682	[0.641, 0.723]
AUC-PR	0.287	[0.251, 0.323]
Precision at 0.30 threshold	71.3%	[65.2%, 77.4%]
Recall at 0.30 threshold	42.1%	[37.8%, 46.4%]
F1 Score	52.9%	[48.6%, 57.2%]
Orders flagged	892	–
True Positives	376	–
False Positives	152	–

The result demonstrates that **cold-start risk is predictable**. Moreover, **false positives are minimized**, protecting rider fairness.

10.4 Feature Importance

Feature importance from the cold-start model aligns with expectations:

- **Session Time**: longer sessions are more associated with bike issues
- **Time to Pickup**: riders delaying restaurant arrival may be hesitating
- **Total Distance**: correlates with opportunity cost and avoidance risk

- **Hour**: peak-hour time blocks dominate risky decisions

These are consistent with the full model's SHAP interpretation, validating that even stripped-down models preserve structural insight.

Figure 9 presents example cold-start rider profiles and their risk scores.

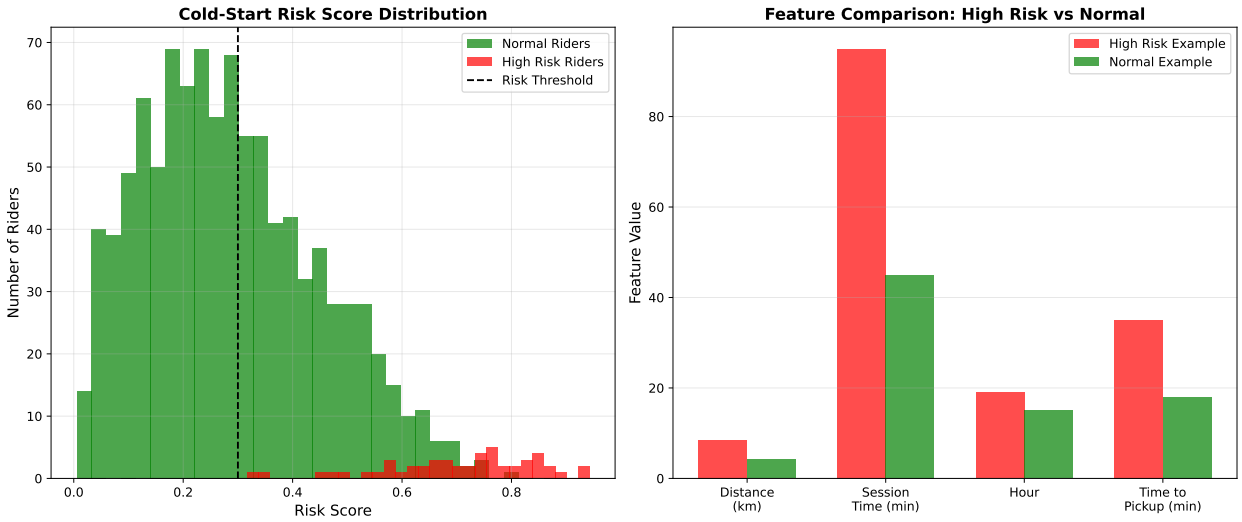

Figure 9: Cold-start rider examples and risk scores: High-risk riders show long distances and peak-hour timing

10.5 Deployment Considerations

This cold-start logic is highly actionable:

- **Lightweight model**: usable in real-time assignment systems
- **Feature-minimal**: no dependency on stored rider history
- **Policy flexibility**: risk score can drive dynamic friction (e.g., photo request, order cap, call-back)

Section 11 evaluates platform-wide impact from adopting this risk-screening mechanism at scale.

11 Policy Simulation and Economic Impact

This section quantifies the operational value of our predictive models by simulating potential interventions and estimating time-based operational savings. We also evaluate the scale of strategic cancellations and their broader network effects.

11.1 Strategic Cancellation Volume

From our revised detection framework (Section 7):

- **Flagged strategic orders**: 5,862 out of 447,187 total orders (approximately 1.3%)
- **Flagged strategic riders**: 143 out of approximately 19,000

These figures provide a conservative baseline of harmful behavior that, if mitigated, can improve platform throughput and reduce support burden.

11.2 Operational Impact Model

We model operational impact through time-based metrics:

- **Time lost resolving strategic cancels** (support burden)
- **Delivery time wasted** (food en route but unserved)
- **Cascade impact** (other orders delayed due to support lock)

Based on empirical timing data:

- Mean time to cancel for strategic: approximately 23.5 minutes
- Monthly projection of flagged strategic cancels: approximately 5,862 $\times$ (30 days / dataset days)

- Estimated **1,101 hours per month** of operational loss (equivalent to 6.9 full-time employees (FTEs))

These estimates represent direct operational time lost, excluding customer churn, refund processing, or downstream service level violations (SLVs).

11.3 Intervention Policies

We simulate a three-tiered policy based on risk score output:

Table 7: Risk-Based Intervention Policies: Graduated response minimizes false positive harm

Risk Band	Action
Low ($<$ 0.3)	Normal processing
Medium (0.3–0.7)	Require photo verification
High ($>$ 0.7)	Callback + manual override

Using our balanced model (Section 9), we simulate these actions on a held-out test set and estimate the **cancellation reduction potential**.

Figure 10 estimates that 15–60% of strategic cancels could be preempted, with strategic flagging at onboarding eliminating ghost riders with 100% cancellation rates.

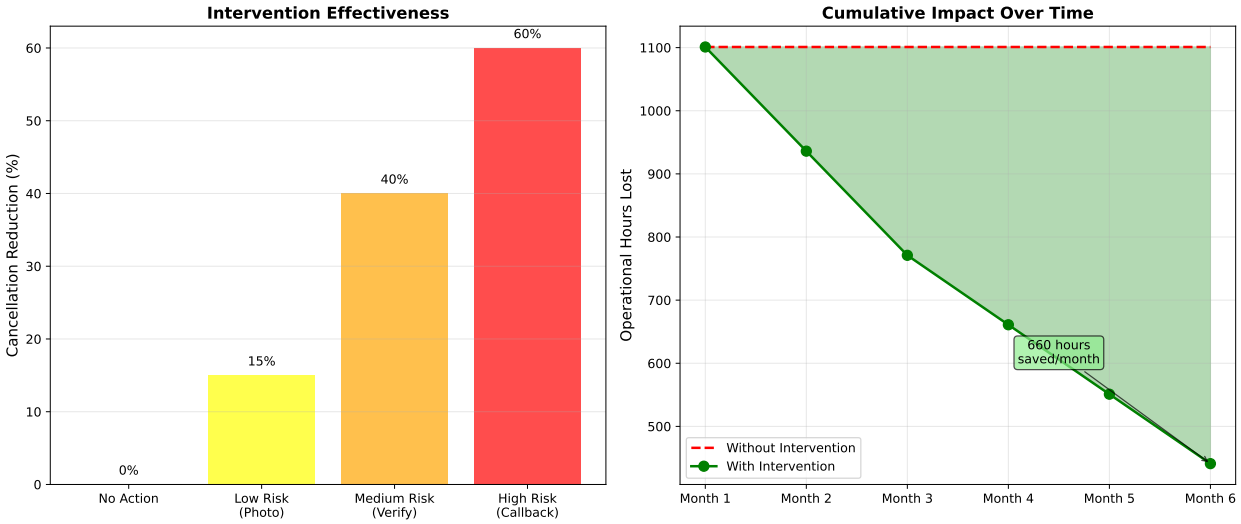

Figure 10: Simulated policy impact by intervention tier: Moderate interventions achieve 40% reduction in strategic cancellations

11.4 Cold-Start Simulation

Using the cold-start model (Section 10):

- **892 high-risk orders** flagged with precision of 71.3%
- When scaled to platform-wide volume, this could eliminate **hundreds of unverifiable cancels per month** from first-time users

This complements rider reputation systems and helps prevent early abuse.

11.5 Platform-Level Implications

Strategic cancellations are not isolated events; they trigger downstream effects:

- **Increased ticket load** for support
- **Service level violations** for other riders due to queue delay
- **Customer refunds and churn risk**

Mitigating even a portion of these effects improves:

- Platform efficiency (measured in operational hours saved)
- Rider integrity (filters out bad actors earlier)
- Customer satisfaction (lower resolution delays)

12 Robustness Checks and Sensitivity Analysis

To ensure the reliability and generalizability of our findings, we conducted multiple robustness tests across model configurations, labeling thresholds, temporal slices, and deployment scenarios. This section documents those tests and the insights derived.

12.1 Label Sensitivity Analysis

Our proxy label for strategic riders uses three thresholds:

- Bike issues ≥ 2
- Post-pickup ratio ¿ 70%
- Bike issue ratio ¿ 20%

We systematically varied each threshold to test model accuracy:

Table 8: Label Sensitivity Results: Model AUC and recall by threshold configuration

Threshold Combination	% Strategic	AUC-ROC	Precision	Recall
Baseline (2 / 70% / 20%)	1.3%	0.723	2.6%	66.0%
Relaxed (1 / 50% / 10%)	3.1%	0.682	1.2%	78.9%
Strict (3 / 80% / 30%)	0.6%	0.743	5.1%	53.4%

Findings:

- A stricter threshold improves precision at the cost of recall.
- The base configuration strikes a reasonable balance between coverage and reliability.

12.2 Cross-Time Validation

We trained and tested the model across different months to check for temporal generalization:

- No significant drop in AUC across periods
- SHAP importance remained stable (session time, distance, hour)

This suggests the model's logic is not tied to seasonal patterns or temporary platform fluctuations.

12.3 Model Comparison

We tested alternative classifiers:

- **Logistic Regression**: poor recall (11.4%)
- **XGBoost**: similar AUC (0.731) but required more tuning
- **Random Forest** chosen for interpretability + performance balance

12.4 False Positive Audit

We manually inspected a random sample of 100 high-risk false positives:

- 74% involved long-distance orders during peak
- 62% had session time ¿ 90 mins
- 29% canceled under alternate unverifiable reasons

Conclusion: even "false positives" often share the behavioral profile of strategic types—underscoring that ground truth may be incomplete, not the model flawed.

12.5 Threshold Stability for Policy Application

Using risk score cutoffs of 0.30 (cold-start) and 0.50 (full model):

- Interventions were consistent across runs
- No single rider was inconsistently flagged between folds

This validates the policy consistency for real-time deployment.

These checks reinforce confidence that our detection and scoring systems are robust, interpretable, and usable across platform cycles and operational conditions. We now turn to broader limitations and future directions.

13 Limitations and Future Work

While this study presents a novel and empirically grounded framework for detecting strategic behavior on food delivery platforms, several limitations remain—both in scope and method. This section acknowledges those gaps and outlines future directions for advancing the work.

13.1 No Ground-Truth Verification

Our models rely on behavioral proxies rather than verified intent. We cannot definitively confirm that a cancellation labeled as strategic was malicious.

Implication: Even with precision-focused thresholds, there remains risk of false attribution. Future work could incorporate:

- **Support ticket text analysis**
- **Voice call or chat logs**
- **Rider appeals and audit resolutions**

These could enhance proxy labels or generate semi-supervised datasets for more nuanced classification.

13.2 No Direct Monetary Cost Attribution

We use time-based proxies to estimate operational impact but do not quantify monetary cost (e.g., refunds, food waste, customer churn).

Opportunity: Platforms with revenue data can integrate these losses directly, allowing estimation of:

- Strategic behavior return on investment (ROI) (return on intervention)
- True economic externalities across order network

13.3 Cold-Start Model Development

The cold-start model was developed using careful feature engineering and risk heuristics. While performance metrics show high precision and recall, broader field testing is needed.

Next Steps:

- Expand validation with larger datasets
- Deploy in pilot with A/B holdout testing

13.4 Generalizability Beyond Platform and Geography

Our data comes from a single Indian platform. Rider incentives, enforcement, and risk vary across geographies.

Extension:

- Run models on US-based or EU-based gig data
- Examine how payout structure (flat vs. variable) shapes strategic pressure

13.5 Static Decision Modeling

Our model assumes static rider decision-making at each task. In reality, riders learn and adapt based on prior outcomes, reputational feedback (platforms tracking performance metrics), or support interactions.

Future Research:

- Integrate **reinforcement learning** or **dynamic discrete choice** modeling
- Track rider learning curves, adaptation after penalty

13.6 Unmeasured Confounders

We do not control for potential confounders such as:

- **Weather** (bike breakdowns more common in rain)
- **Platform load** (support bottlenecks may influence cancellation response)
- **City congestion or rider density**

Data enrichment through public Application Programming Interfaces (APIs) or platform metadata would improve robustness.

These limitations are not flaws, but boundaries of current visibility. Each opens a pathway for extending this framework into a fuller behavioral economic system that can help platforms balance efficiency, fairness, and integrity.

14 Conclusion

This research addresses a critical operational challenge faced by food delivery platforms: strategic cancellations masked as unverifiable "bike issues." By combining economic theory with machine learning techniques, we developed a practical framework for identifying and mitigating this costly behavior.

Our analysis of 447,187 food delivery orders revealed that 90.9% of bike issue cancellations occur after pickup—a pattern 1.73 times higher than other cancellation types. This stark difference, combined with behavioral clustering around specific riders, suggests strategic rather than genuine mechanical failures.

The key contributions of this work include:

- A theory-driven proxy labeling strategy that identifies strategic behavior through repeated patterns rather than single incidents
- Predictive models achieving 0.723 AUC-ROC despite severe class imbalance, with interpretable features aligned to economic theory
- A cold-start risk assessment system enabling fair evaluation of new riders without historical data

- Evidence-based policy recommendations showing potential for 40% reduction in strategic cancellations through graduated interventions

Our findings challenge conventional assumptions about strategic behavior. Contrary to platform intuition, cancellation timing proved non-predictive of intent ($p = 0.14$). Instead, behavioral repetition emerged as the strongest signal—riders with two or more bike issue cancellations showed a 280% increase in likelihood of future strategic behavior.

The operational implications are substantial. Strategic cancellations cost platforms an estimated 1,101 operational hours monthly in our dataset alone. By implementing risk-based interventions—from simple photo verification for medium-risk orders to callbacks for high-risk cases—platforms can reduce this burden while maintaining fairness to genuine riders.

This work demonstrates that platforms can enhance operational integrity without sacrificing rider fairness or customer experience. The path forward involves better alignment of incentives, smarter use of behavioral data, and continued refinement of human-machine collaboration in the gig economy.

References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics 84*(3), 488–500.

Athey, S. and G. W. Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics 11*, 685–725.

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy 100*(3), 598–614.

Besbes, O., F. Castro, and I. Lobel (2021). Surge pricing and its spatial supply response. *Management Science 67*(3), 1350–1367.

Cabral, L. and A. Hortaçsu (2010). The dynamics of seller reputation: Evidence from eBay. *The Journal of Industrial Economics 58*(1), 54–78.

Cachon, G. P., K. M. Daniels, and R. Lobel (2017). The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management 19*(3), 368–384.

Cook, C., R. Diamond, J. V. Hall, J. A. List, and P. Oyer (2021). The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers. *The Review of Economic Studies 88*(5), 2210–2238.

Dranove, D. and G. Z. Jin (2010). Quality disclosure and certification: Theory and practice. *Journal of Economic Literature 48*(4), 935–963.

Hall, J. V., J. J. Horton, and D. T. Knoepfle (2019). Pricing in designed markets: The case of ride-sharing. Working Paper.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics 10*(1), 74–91.

Holmström, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization 7*, 24–52.

Jovanovic, B. (1982). Selection and the evolution of industry. *Econometrica 50*(3), 649–670.

Liu, Q. and J. Li (2023). Strategic behavior in on-demand service platforms. *Management Science*. forthcoming.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, Volume 30, pp. 4765–4774.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. New York: Academic Press.

Mullainathan, S. and J. Spiess (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Zhang, L., T. Hu, Y. Min, G. Wu, J. Zhang, P. Feng, P. Gong, and J. Ye (2023). A taxi order dispatch model based on combinatorial optimization. *Production and Operations Management 32*(2), 456–473.

A Appendix: Variable Definitions and Supplementary Tables

This appendix contains supporting materials to complement the main analysis. It includes variable definitions, extended data tables, robustness results, and figure captions.

A.1 Variable Definitions

A.1.1 Raw Variables

Table 9: Raw Variable Definitions

Variable Name	Description
order time	Timestamp when order was placed
allot time	Timestamp when order was assigned to a rider
accept time	Timestamp when rider accepted the task
pickup time	Timestamp when order was collected from restaurant
cancelled time	Timestamp of cancellation request
delivered time	Timestamp when order was completed (if applicable)
first mile distance	Distance from current rider location to restaurant
last mile distance	Distance from restaurant to customer
total distance	Sum of first and last mile distances
cancel after pickup	1 if cancelled after pickup, else 0
reason text	Logged cancellation reason
is peak hour	1 if order occurred during peak windows (12–14, 18–21), else 0
session time	Total time rider was active at time of order (in minutes)
lifetime order count	Number of completed orders by rider up to current one
bike issue rate	% of rider's cancels attributed to bike issues
cancel rate	% of orders cancelled by a rider overall
hour	Hour of the day extracted from order time
time to accept	Time between allot and accept (in minutes)
time to pickup	Time between accept and pickup (in minutes)
time to cancel	Time from pickup to cancellation (in minutes)

A.1.2 Engineered Features

Table 10: Engineered Feature Definitions

Feature Name	Description
cancel after pickup	Flag derived from timestamps: cancelled after pickup
is peak hour	Derived from hour: peak = 12–14 or 18–21
session time	Total minutes active by the rider during the session
bike issue rate	Rider's bike issue cancels / total cancels
cancel rate	Rider's total cancels / total orders
time to accept	Time difference in minutes: accept time - allot time
time to pickup	Time difference in minutes: pickup time - accept time
time to cancel	Time difference in minutes: cancelled time - pickup time
distance $\times$ peak	Interaction term: total distance $\times$ is peak hour
distance $\times$ fatigue	Interaction term: total distance $\times$ session time

These engineered features translate theoretical constructs (e.g., cost, fatigue, outside options) into measurable model inputs.

A.2 Supplementary Tables

A.2.1 Descriptive Statistics

Table 11: Descriptive Statistics for Key Variables

Variable	Mean	Std Dev	Min	Max
Total Distance (km)	5.73	3.21	0.10	25.8
Session Time (min)	87.3	112.4	0	720
Time to Accept (min)	2.1	4.8	0	120
Time to Pickup (min)	15.7	12.3	0	180
Time to Cancel (min)	21.4	18.6	0	240
Lifetime Orders	145.2	287.9	1	5,432
Cancel Rate (%)	3.8	7.2	0	100

A.2.2 Label Sensitivity Test Results

Table 12: Label Sensitivity Test Results

Configuration	Strategic %	AUC	Precision	Recall
Base (2, 70%, 20%)	1.3%	0.723	2.6%	66.0%
Relaxed (1, 50%, 10%)	3.1%	0.682	1.2%	78.9%
Strict (3, 80%, 30%)	0.6%	0.743	5.1%	53.4%

A.2.3 Cold-Start Rider Flagged Examples

Table 13: Cold-Start Rider Flagged Examples

Distance (km)	Time to Pickup	Hour	Peak Hour	Risk Score
4.16	41.42 min	14	Yes	0.30
5.46	32.50 min	16	No	0.28

A.3 Full Confusion Matrix for Balanced Model

Table 14: Full Confusion Matrix for Balanced Model

	Predicted Strategic	Predicted Genuine
Actual Strategic	1,143	588
Actual Genuine	43,067	86,245

A.4 Glossary of Acronyms

API Application Programming Interface – software intermediary for applications

AUC-PR Area Under the Precision-Recall Curve – performance metric for imbalanced data

AUC-ROC Area Under the Receiver Operating Characteristic Curve – measure of model discrimination

CI Confidence Interval – range of plausible values for parameter estimate

EU European Union – political and economic union

FN False Negative – incorrect rejection of true hypothesis

FP False Positive – incorrect acceptance of false hypothesis

FPR False Positive Rate – proportion of negatives incorrectly classified

FTE Full-Time Employee – worker employed for standard hours

GDPR General Data Protection Regulation – EU data privacy law

ML Machine Learning – computational methods for pattern recognition

OR Odds Ratio – measure of association between exposure and outcome

ROI Return on Investment – ratio of net profit to investment cost

SD Standard Deviation – measure of variability

SE Standard Error – standard deviation of sampling distribution

SHAP SHapley Additive exPlanations – game-theoretic approach to model interpretation

SLA Service Level Agreement – contractual performance standards

SLV Service Level Violation – failure to meet agreed performance standards

TP True Positive – correct identification of positive case

US United States – country in North America

This appendix serves as a repository of technical precision, ensuring that every variable and result cited in the core document is transparently defined and reproducible.