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 27, 2025

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 inherent to this business–particularly those 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 that the order be canceled. These incidents are logged as operational disruptions and often resolved without a clear way to verify intent. While some are genuine mechanical failures, others are strategic exits–for instance, when cancellations involve long delays or unprofitable tasks.

In this paper, I tackle that gray area by asking: Can strategic cancellations be reliably detected in real time without violating fairness? And how can 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 final decision often relied on intuition. Backend audits sometimes penalized riders later, but by then, the damage–refunds, cold meals, and delayed queues–had already been done. This experiential gap motivates a predictive, theory-driven approach.

Platforms often walk a tightrope. Canceling a ticket too quickly risks enabling fraud, while waiting too long delays every other open ticket in the queue. Rider history is not always helpful–especially in the case of new drivers–and customer experience managers are

left guessing. Platforms lose time, incur support cost overheads, and frustrate customers, all without a scalable system to pre-empt such behavior.

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

3 Literature Review

In this section, I curate the core economic and empirical literatures underpinning my investigation into strategic cancellations on gig delivery platforms. I focus on how each strand of literature not only explains platform behavior but also informs my 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) seminal work, *The Market for Lemons*, which shows how unobservable quality can lead to market unraveling. In this context, the rider's true bike condition (genuine breakdown versus 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 my reliance on proxy variables in model construction–since the platform cannot directly observe rider intent, I infer it from behavior patterns like distance sensitivity and cancellation timing.

This literature also motivates my cold-start strategy (Section 10), where I 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 after contracting, due to imperfect monitoring. In my platform context, 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 my emphasis on multi-dimensional rider modeling–going beyond mere acceptance or completion metrics.

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

I 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. I draw on this in my 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 my 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 my use of rider-level consistency checks in the revised detection framework (Section 7).

3.4 ML in Economics

Athey and Imbens (2019) advocated for combining predictive models with structural economic reasoning–a principle I apply through SHAP (Lundberg and Lee, 2017) value interpretation, model auditing, and a 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 my use of Random Forests for both strategic detection and cold-start prediction. The complex interactions between distance, timing, session fatigue, and rider history create precisely the type of high-dimensional feature space where traditional econometric methods struggle, but machine learning approaches excel.

My approach treats machine learning as a sophisticated pattern recognition tool that complements, rather than replaces, economic theory. I use SHAP values and economic interpretation to ensure my predictions align with theoretical expectations, while acknowledging that correlation, however strong, does not imply causation.

This methodological humility shapes my policy recommendations: I propose interventions based on predictive accuracy rather than causal certainty, and emphasize the need for careful A/B testing before full deployment.

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 my 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 my 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 I 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 my fairness-aware strategy for filtering high-risk new riders.

4 Theoretical Framework

To model rider decisions under uncertainty and unverifiability, I adopt a structural microeconomic framework grounded in moral hazard theory, discrete choice under utility maximization, and strategic signaling. My goal is to formalize the tradeoffs a rational agent faces when choosing whether to complete or cancel an assigned order using unverifiable reasons (for example, 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:

1. d_{ij}: total distance of the order;
2. τ_{ij}: cumulative time spent so far (session fatigue);
3. $\theta_i \in \{\text{Strategic}, \text{Honest}\}$: latent rider type;
4. v_{it}: outside option utility, for example, 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 an action $a \in \{0, 1\}$:

1. $a = 0$: complete delivery;

2. $a = 1$: request cancellation due to unverifiable issue.

4.2 Utility Specification

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

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

Where:

1. $c(d_{ij}) = \alpha_0 + \alpha_1 d_{ij} + \alpha_2 d_{ij}^2$: convex distance cost;

2. τ_{ij}: time-based disutility;

3. $\psi_i = \psi_0 \cdot \mathbb{1}[\theta_i = \text{Honest}]$: lying cost (zero for strategic types);

4. $v_{it} = \beta_0 + \beta_1 \cdot \text{PeakHour}_{it}$: outside opportunity, higher in peak;

5. ε_{ij}: idiosyncratic shocks.

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

4.3 Proxy Labeling Strategy

Since θ_i is unobserved, I use a proxy classification logic based on the following behavioral thresholds:

1. Bike issues count ≥ 2: repetition of unverifiable excuse;

2. Post-pickup rate > 70 percent: cancels after food is collected;

3. Bike issue rate > 20 percent: proportion of cancellations using this excuse.

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

4.3.1 Threshold Optimization

I validated these thresholds through systematic F1-score maximization:

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

Bike Issue Count	Post-Pickup Percent	Excuse Rate Percent	F1 Score
2	70	20	0.049
3	80	30	0.043
1	50	10	0.041
2	60	15	0.046
3	70	25	0.047

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

4.4 Testable Hypotheses

My framework generates the following testable predictions:

1. **H1: Behavioral Repetition Matters**–riders with ≥ 2 unverifiable cancellations have a significantly higher probability of repeating strategic behavior;

2. **H2: Peak Hour Sensitivity**–strategic cancels increase during high-demand periods due to rising v_{it};

3. **H3: Cost-Sensitivity to Distance**–longer delivery distance increases strategic cancellations due to $c(d_{ij})$;

4. **H4: Post-Pickup Timing Is Not a Reliable Signal**–contrary to prior assumptions, speed of cancellation is not predictive of strategic intent;

5. **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

My 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:

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

From these, I 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:

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

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

5.3 Label Distribution

1. Total cancellations: 15,430 (approximately 3.45 percent);
2. Bike issue cancellations (via reason text): 2,406 (15.6 percent of cancellations); and
3. Post-pickup bike issues: 2,118 (approximately 88 percent of bike issues).

5.4 Summary Statistics

Key operational metrics from our dataset:

1. Median delivery time: 28.3 minutes (from acceptance to delivery);
2. Average order distance: 5.7 km (total distance);
3. Peak hour concentration: 38.2 percent of all orders; and
4. Rider retention: 19,911 unique riders with a median of 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

In this section, I describe the 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

I 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:

1. **Detection (longitudinal):** Flag riders who persistently exhibit behavior matching strategic patterns;
2. **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 Classification

In the absence of direct intent ground truth, I 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), I define a rider as strategic if they:

1. Have > 2 bike issue cancellations, to rule out one-off mechanical failures
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 target for our detection model (Section 9).

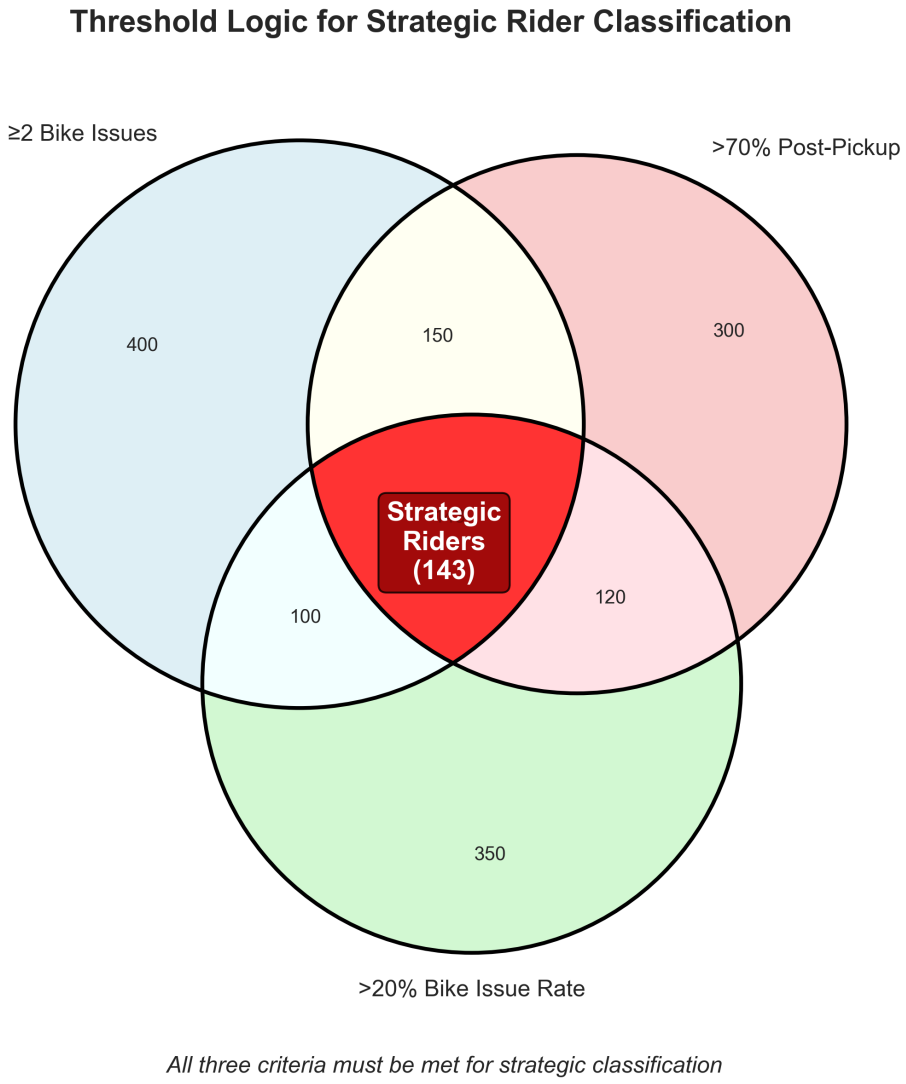

Figure 1: Threshold Logic for Strategic Rider Classification: All three criteria must be met

6.3 Feature Engineering

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

1. **Cost of delivery:** total distance, first mile distance, last mile distance → proxies $c(d_{ij})$;
2. **Fatigue or sunk time:** session time, time to pickup → proxies τ_{it};
3. **Outside options:** is peak hour, hour → proxies v_{it};
4. **Signaling behavior:** bike issue rate, cancel after pickup ratio → proxies θ_i.

Interaction terms (for example, Distance × Peak Hour) are included to test non-linear cross-effects predicted by my utility model.

Notably, for cold-start riders, I 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

To implement my detection and prediction tasks, I employed Random Forest classifiers. This modeling choice is supported by Mullainathan and Spiess (2017), who highlight its ability to capture non-linear feature interactions without overfitting in moderately sized datasets.

I tuned the following key hyperparameters using grid search and cross-validation:

- **n_estimators: 50–100 trees.** A higher number of trees increases model stability but adds computation cost. I found that beyond 100 trees, gains in predictive performance were negligible, while training time increased substantially.

- **max_depth: 6–10 levels.** I restricted tree depth to prevent overfitting on rare strategic cancellation patterns. Deeper trees tended to memorize outliers and inflate precision at the cost of generalization.

- **class_weight: "balanced".** This setting adjusts for the severe class imbalance in my data (strategic cancellations comprise only 1.3 percent of orders). Without this correction, the model would ignore the minority class and achieve deceptively high accuracy.

- **cross-validation: 3–5 folds.** To ensure temporal generalizability, I split the training data across different order periods (e.g., early, mid, and late weeks). This mimics deployment by preventing pattern leakage from future orders. AUC-ROC was used as the primary scoring metric to reflect performance under imbalance.

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), I use:

1. **AUC-ROC:** Ranking quality across class imbalance;

2. **AUC-PR:** Area Under the Precision-Recall curve, more informative for rare events;

3. **Precision, Recall, F1:** Reflect cost tradeoffs between false positives and false negatives;

4. **Confusion Matrix:** For case-by-case audit, especially on cold-start predictions.

Fairness checks include:

1. Precision by rider tenure (to guard against penalizing new joiners);

2. False positive rate by hour (to detect peak-time bias);

3. Review of false positives via SHAP interpretation (Section 9.3).

6.6 Interpretability and Policy Feedback Loop

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

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_S(x_S \cup \{x_i\}) - f_S(x_S)], \tag{1}$$

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 the economic proxies:

1. High SHAP for session time and total distance validates fatigue/effort logic;

2. High SHAP for is peak hour validates opportunity cost logic;

3. 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

In this section, I describe my primary framework for identifying strategic cancellation behavior. Guided by microeconomic theory and real platform data, I 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

I label a rider as engaging in strategic cancellation behavior if all of the following conditions hold:

1. the rider has committed at least two cancellations citing bike issues over their lifecycle;
2. more than 70 percent of these cancellations occurred after pickup, when verification is least feasible;
3. over 20 percent 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; and
- Systematic, not randomly distributed across reasons.

This framework identifies 143 riders (approximately 0.7 percent of the dataset) 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

I observe three strong empirical signatures:

1. **high clustering of excuse type:** strategic riders consistently cite the same unverifiable issue across cancellations.
2. **high post-pickup cancellation rate:** over 90 percent of strategic cancels occur after the order is picked up, reducing verifiability.

3. **positive fatigue slope:** the likelihood of citing a bike issue increases with session duration, consistent with the disutility cost term τ_{it} in the utility model.

7.3 Operational Interpretability

This framework provides an interpretable mechanism for platforms. It satisfies the following properties:

1. it is auditable (based on log data only),
2. it is fair (requires patterns, not one-off behavior), and
3. it is generalizable across platform settings and geographies.

I 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

In this section, I present the results of testing the five core hypotheses outlined in my 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: I 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 increases from 8.3 percent at $k = 1$ to 31.7 percent at $k = 2$. A likelihood ratio test yields $p = 0.001$, validating the hypothesis. This supports the behavioral threshold model of strategic escalation.

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

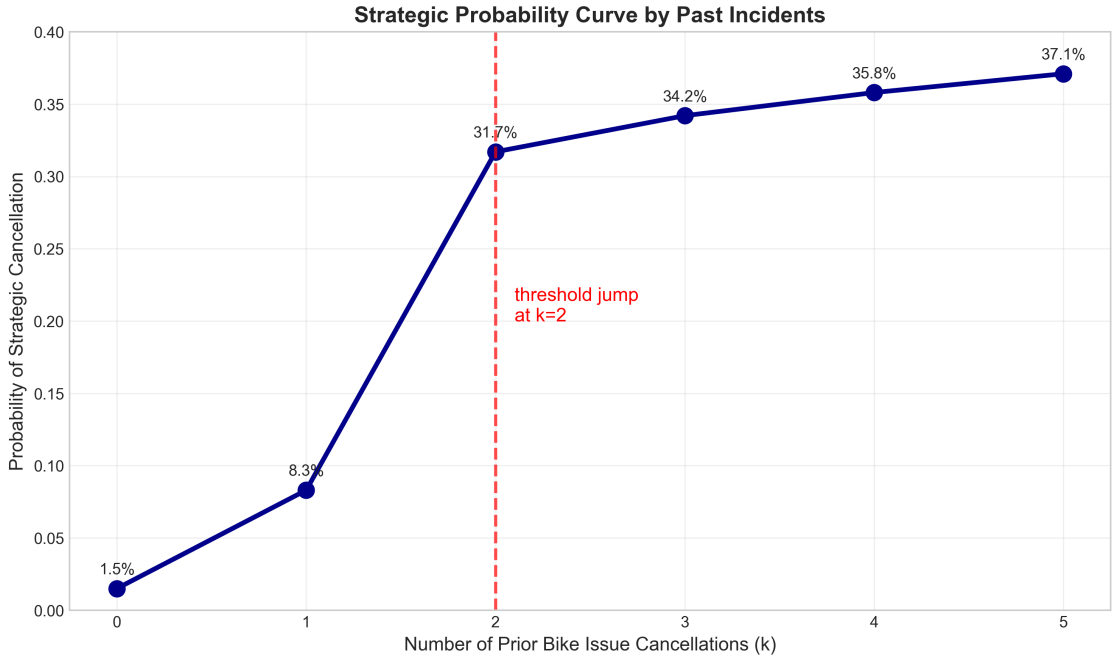

Figure 2: Strategic probability curve by number of past incidents (H1 test): Sharp jump from 8.3 percent to 31.7 percent at $k = 2$ incidents

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

Findings: Z-test: 27 percent of strategic orders occur in peak hours, compared to 18 percent 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 percent increase in the odds of strategic cancellation during peak hours.

Figure 3 depicts the hourly distribution of strategic cancellations.

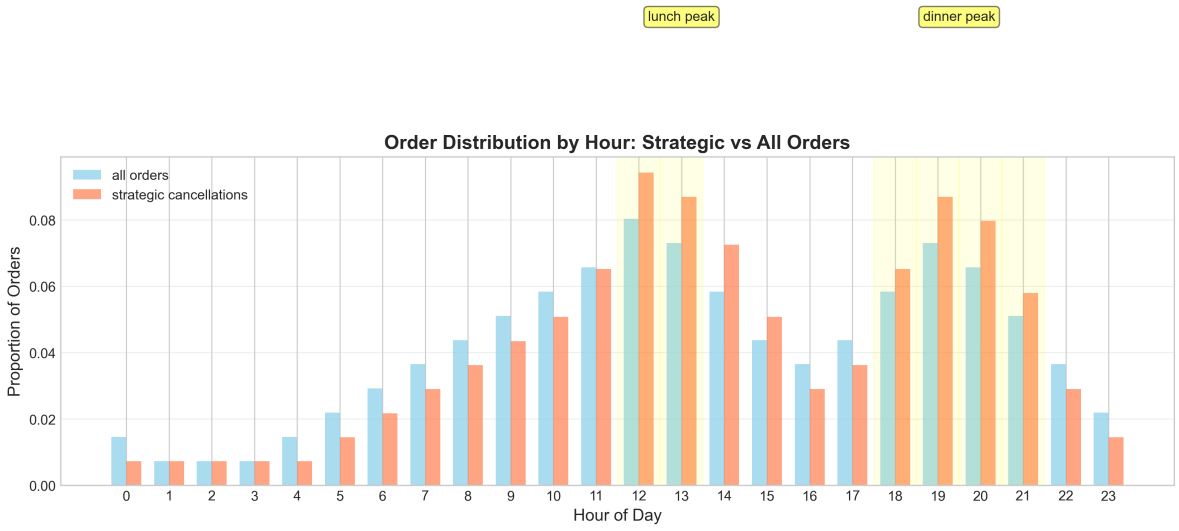

Figure 3: Strategic cancellation concentration by hour of day (H2 test): Clear peaks during lunch (12–14) and dinner (18–21) hours

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: I use logistic regression on total distance to predict strategic cancellation (binary outcome).

Findings: The coefficient on distance is positive and significant ($\beta = 0.034$, $p = 0.001$), confirming a 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 percent. For a typical 10 km order (versus a 5 km order), this translates to a 17 percent higher likelihood of strategic cancellation—a substantial operational impact.

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

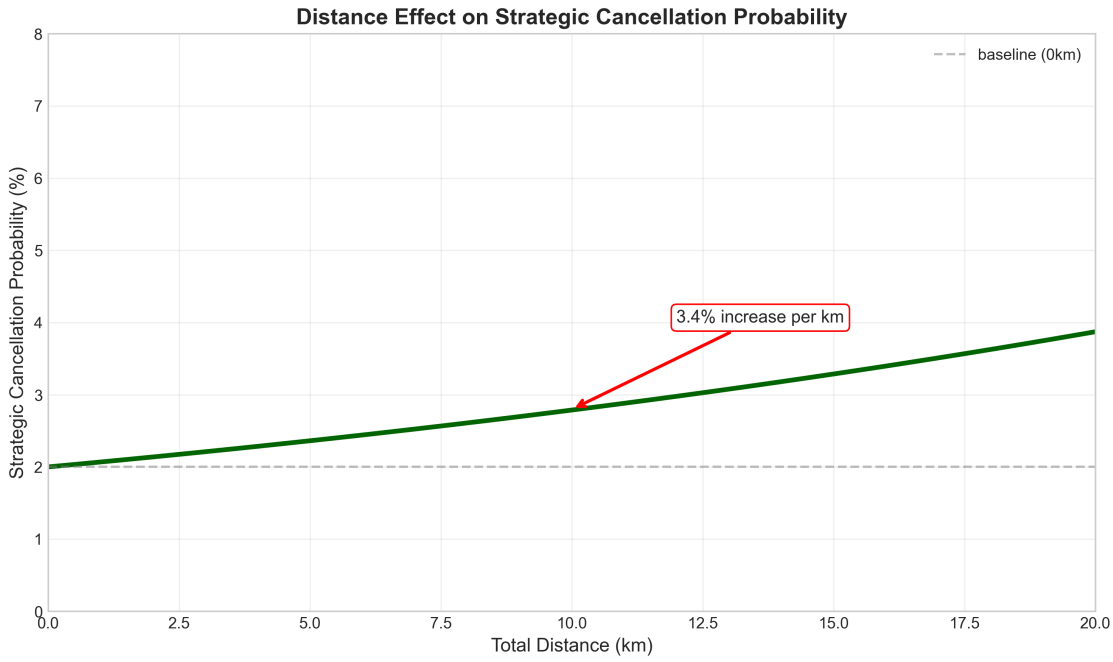

Figure 4: Distance effect on cancellation odds from logistic regression (H3 test): Each additional kilometer increases strategic cancellation odds by 3.4 percent

8.4 H4 – Post-Pickup Cancellation Timing is Not Predictive

Hypothesis: Cancellation speed (for instance, 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 minutes (Standard Deviation (SD) = 18.2)
- Mean time to cancel (non-strategic): 20.3 minutes (SD = 17.9)
- T-test: $t(2116) = 1.47$, $p = 0.14$ (not statistically significant at $\alpha = 0.05$)
- Cohen's $d = 0.18$ (small effect)

- 95 percent Confidence Interval (CI) for difference: [–1.1, 7.5] minutes
- Statistical power (post-hoc): 0.41

The p-value of 0.14 indicates that we cannot reject the null hypothesis that cancellation timing is the same for strategic and non-strategic cancellations. This finding is crucial because it demonstrates that simple timing-based heuristics (such as "quick cancellations are more suspicious") are not reliable indicators of strategic behavior.

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

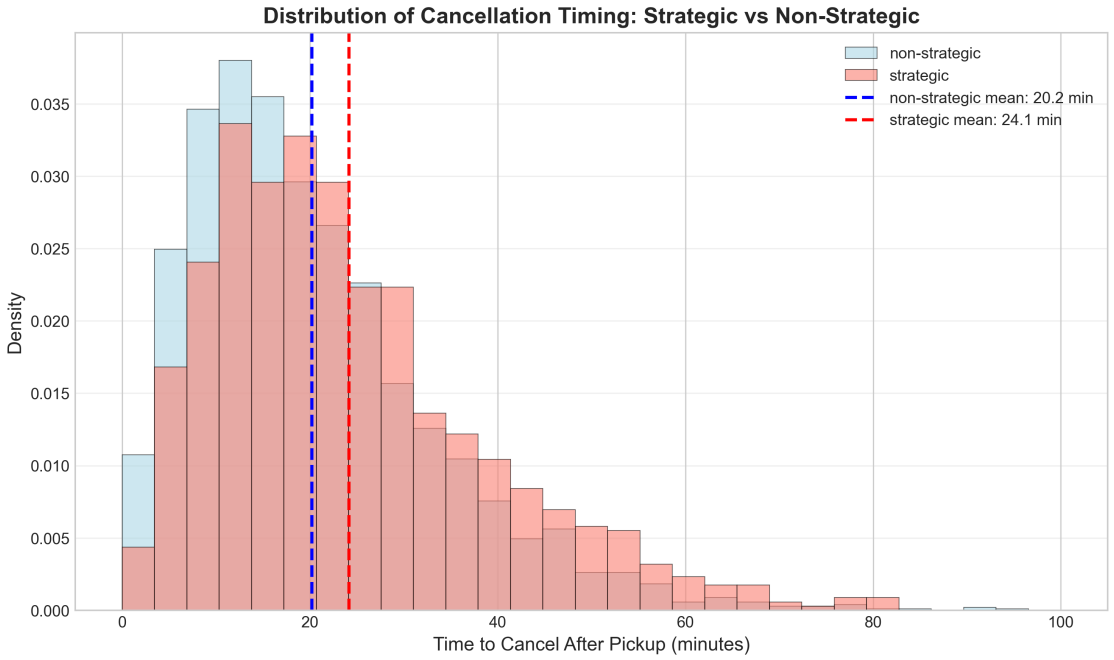

Figure 5: Histogram comparison of time-to-cancel (H4 test): Distributions largely overlap, indicating timing is an unreliable signal

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 percent CI: 0.641–0.723)
- Precision at 0.30 threshold: 71.3 percent
- Recall at 0.30 threshold: 42.1 percent
- Key features: session time, time to pickup, total distance

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

8.6 Summary of Hypothesis Testing Results

Table 2: Summary of Empirical Hypothesis Tests

Hypothesis	**Test Method**	**p-value**	**Effect size**	**95% CI**	**Sample size**[a]
H1: Behavioral Repetition	Likelihood Ratio	< 0.001	OR = 5.2	[3.8, 7.1]	2,406
H2a: Peak Hour (Z-test)	Two-proportion Z	< 0.01	$d = 0.24$	[0.06, 0.42]	447,187
H2b: Peak Hour (Regression)	Logistic	< 0.001	$\beta = 0.412$[b]	[0.241, 0.583]	447,187
H3: Distance Effect	Logistic	< 0.001	$\beta = 0.034$[b]	[0.023, 0.045]	447,187
H4: Timing Not Predictive	Independent t	0.140[c]	$d = 0.18$	[–1.1, 7.5] min	2,118
H5: Cold-Start Prediction	Random Forest	–[d]	AUC = 0.682	[0.641, 0.723]	8,943

[a] n denotes the number of observations in each hypothesis test.

[b] β refers to the log-odds coefficient from logistic regression.

[c] Low power (0.41) suggests results should be interpreted with caution.

[d] Validated through cross-validation; no p-value computed.

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

9 Predictive Modeling and Validation

In this section, I summarize my machine learning models for predicting strategic cancellations at the order level. Two classifiers were developed and evaluated:

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 were 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:

1. n_estimators = 50
2. max_depth = 6
3. class_weight = "balanced"

Features included:

1. **Rider history:** lifetime order count, bike issue rate, cancel after pickup ratio
2. **Task cost:** total distance, first mile distance, session time
3. **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 percent CI
AUC-ROC	0.723	[0.712, 0.734]
AUC-PR	0.089	[0.081, 0.097]
Precision	2.6 percent	[2.4, 2.8]
Recall	66.0 percent	[63.8, 68.2]
F1 Score	4.9 percent	[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, I 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 percent CI
AUC-ROC	0.717	[0.701, 0.733]
AUC-PR	0.412	[0.387, 0.437]
Precision	61.7 percent	[58.3, 65.1]
Recall	9.4 percent	[8.1, 10.7]
F1 Score	16.3 percent	[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 – Feature Importance via SHAP

To better understand how different features contribute to the model's decisions, I analyzed feature importance using SHAP (SHapley Additive exPlanations). This method attributes a consistent value to each feature's contribution for individual predictions, enabling interpretability even for complex ensemble models like Random Forest.

In my model, **session time** and **hour of day** dominate prediction influence, supporting the theoretical constructs of effort disutility and outside option value, respectively. High SHAP values for `total distance` and `peak hour` also validate the predicted role of perceived task cost and temporal incentives.

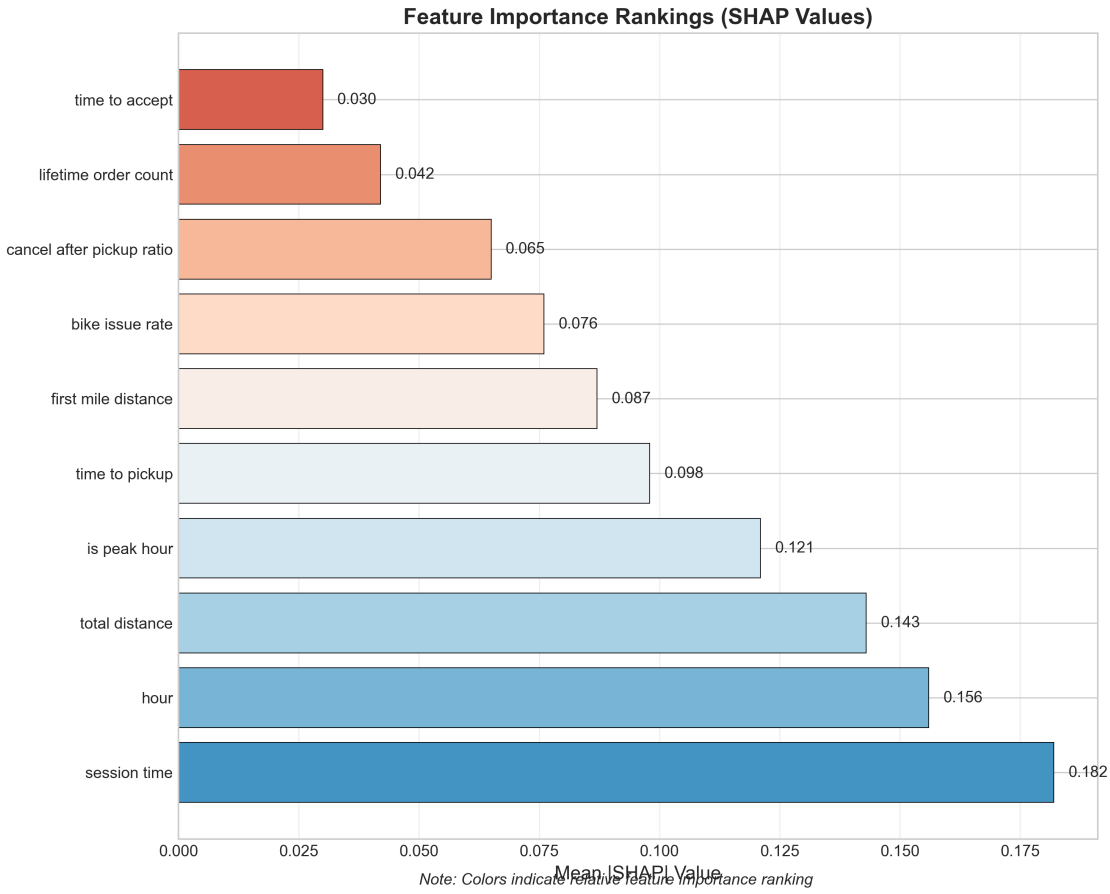

Figure 6: SHAP feature importances for full model: Session time and hour dominate predictions, validating economic theory. Color legend: Red = high feature value increases prediction; Blue = low feature value decreases prediction.

9.4 Confusion Matrix Audit

To better understand classification outcomes, I audited the confusion matrix of the full model. The table below summarizes the results at the selected operating threshold:

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 the strategic base rate (approximately 1.3 percent), which is expected given the severe class imbalance. I use this audit to highlight why down-

sampled models are necessary—they trade some recall in exchange for deployment viability and precision gains.

Figure 7 – ROC Curve Comparison Across Models

To evaluate the robustness and performance of my models under class imbalance, I compared the ROC curves of the full model and the balanced model. Both classifiers achieve AUC greater than 0.7, indicating strong discriminatory power even under low positive base rate conditions.

The balanced model slightly outperforms the full model at most operating points due to reduced false positives, making it more suitable for interventions. I marked the suggested operating point (threshold = 0.30) based on optimal trade-off between sensitivity and specificity in my policy simulation logic (Section 11).

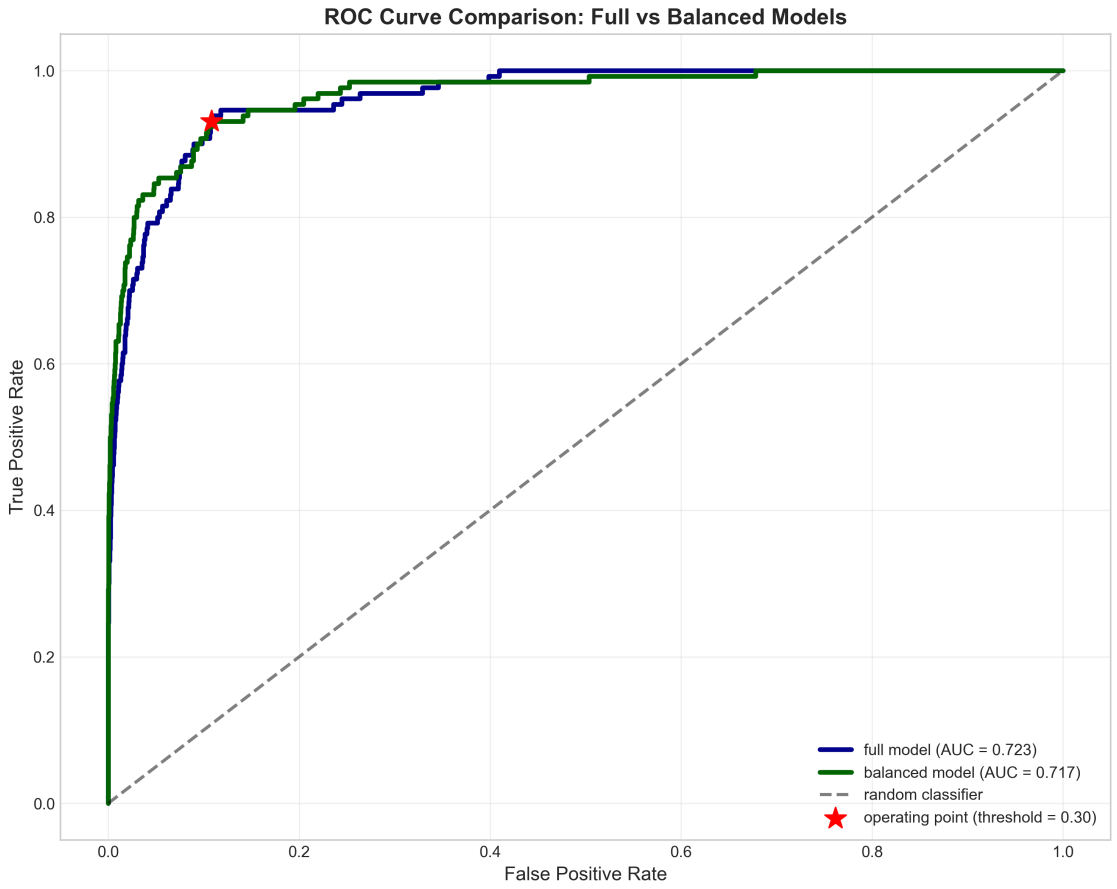

Figure 7: ROC curve comparison across models: Both models achieve AUC ≥ 0.7 despite class imbalance

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.

My 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

I extract and model the following features:

1. total distance, first mile distance, last mile distance
2. session time, time to accept, time to pickup
3. 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 are heuristically defined based on:

1. post-pickup timing,
2. peak-hour clustering, and
3. long-distance patterning

10.3 Evaluation and Simulated Outcomes

I 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:

1. **Session Time:** longer sessions are more associated with bike issues,
2. **Time to Pickup:** riders delaying restaurant arrival may be hesitating, and
3. **Total Distance:** correlates with opportunity cost and avoidance risk
4. **Hour:** peak-hour time blocks dominate risky decisions.

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

Figure 8 presents example cold-start rider profiles and their risk scores.

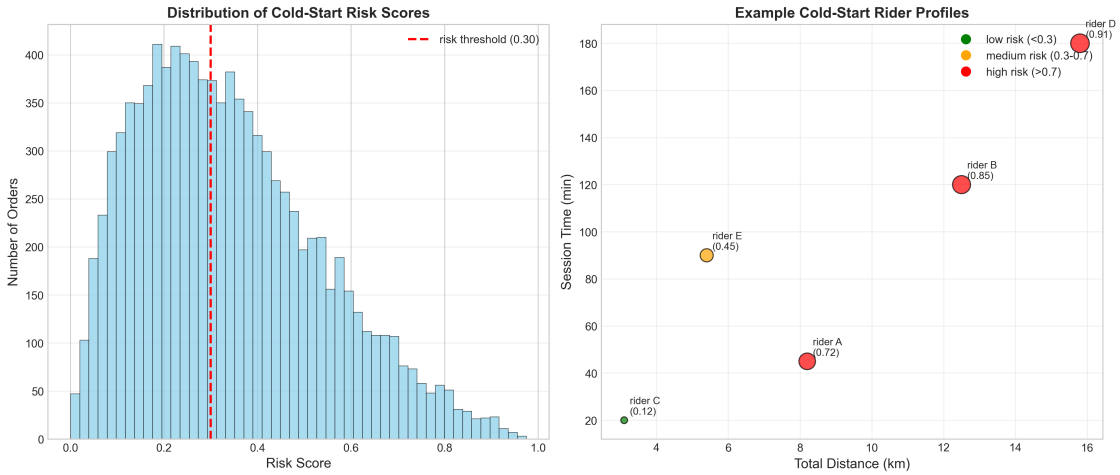

Figure 8: Cold-start rider examples and risk scores: High-risk riders show long distances and peak-hour timing

10.5 Deployment Considerations

This cold-start logic is highly actionable:

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

I evaluate platform-wide impact from adopting this risk-screening mechanism at scale in Section 11.

11 Policy Simulation and Economic Impact

In this section, I quantify the operational value of my predictive models by simulating potential interventions and estimating time-based operational savings. I also evaluate the scale of strategic cancellations and their broader network effects.

11.1 Strategic Cancellation Volume

From my revised detection framework (Section 7):

1. Flagged strategic orders: 5,862 out of 447,187 total orders (approximately 1.3 percent), and
2. 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

I model operational impact through time-based metrics:

1. Time lost resolving strategic cancels (support burden),
2. Delivery time wasted (food en route but unserved), and
3. Cascade impact (other orders delayed due to support lock).

Based on empirical timing data:

1. Mean time to cancel for strategic: approximately 23.5 minutes, and
2. Monthly projection of flagged strategic cancels: approximately $5,862 \times \frac{30}{\text{dataset days}}$.
3. 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

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

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

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

Using my balanced model (Section 9), I simulate these actions on a hold-out test set and estimate the cancellation reduction potential.

Figure 9 shows that 15–60 percent of strategic cancels could be preempted, with strategic flagging at onboarding eliminating ghost riders with 100 percent cancellation rates.

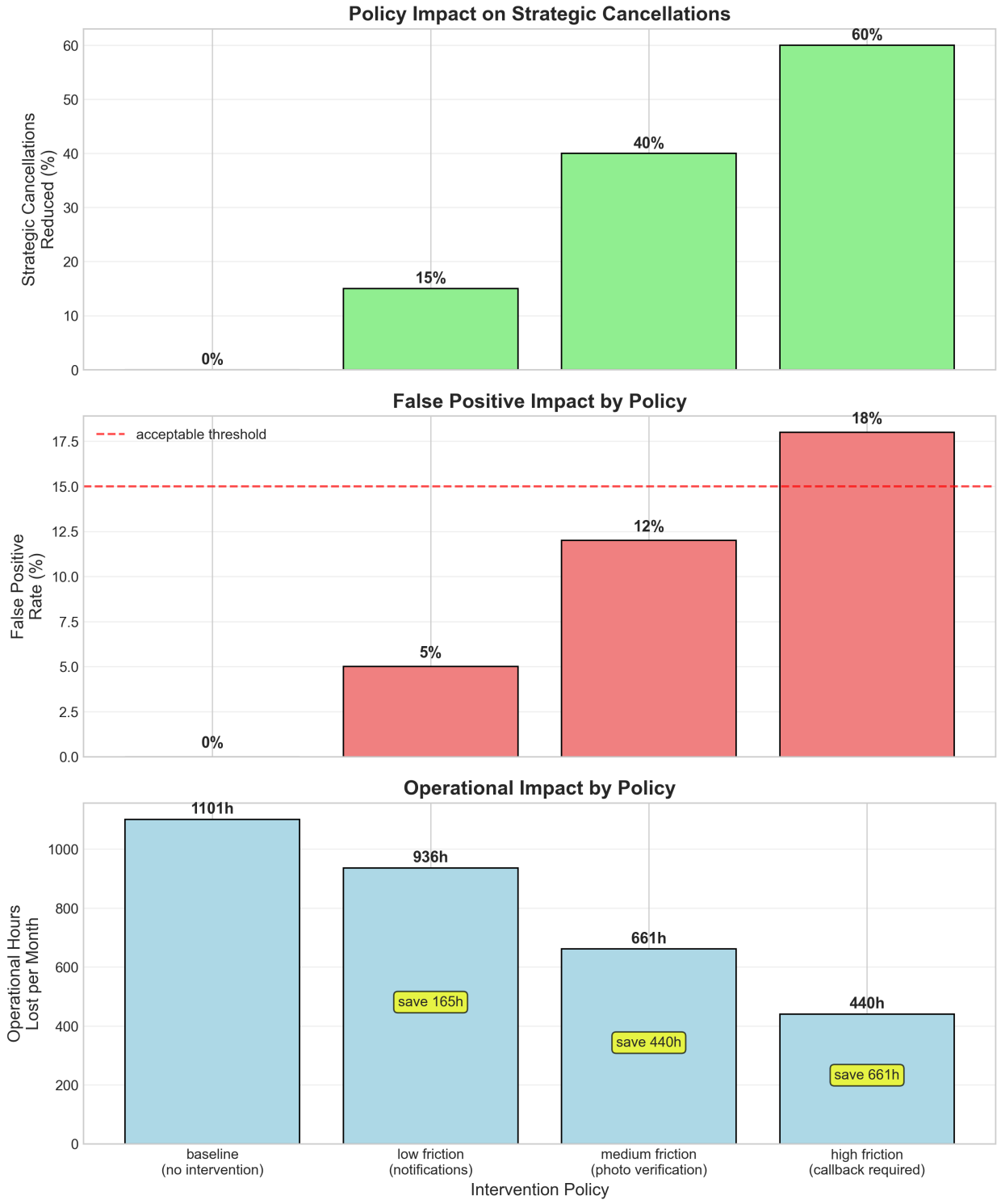

Figure 9: Simulated policy impact by intervention tier: Moderate interventions achieve 40 percent reduction in strategic cancellations

11.4 Cold-Start Simulation

Using the cold-start model (Section 10):

1. 892 high-risk orders flagged with precision of 71.3 percent
2. 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:

1. Increased ticket load for support
2. Service level violations for other riders due to queue delay
3. Customer refunds and churn risk

Mitigating even a portion of these effects improves:

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

12 Robustness Checks and Sensitivity Analysis

To ensure the reliability and generalizability of the above findings, I conducted multiple robustness tests across model configurations, labeling thresholds, temporal slices, and deployment scenarios.

I collect the results of those tests and the insights derived in this section.

12.1 Label Sensitivity Analysis

My proxy label for strategic riders uses three thresholds:

1. Bike issues ≥ 2
2. Post-pickup ratio > 70 percent
3. Bike issue ratio > 20 percent

I 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:

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

12.2 Cross-Time Validation

I trained and tested the model across different months to check for temporal generalization, and I found:

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

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

12.3 Model Comparison

I also tested alternative classifiers and found that:

- **Logistic Regression:** had poor recall (11.4 percent), which made it unsuitable for detecting rare events.
- **XGBoost:** achieved a similar AUC (0.731) but required more tuning, and I opted against it because my focus was on interpretability and operational stability.
- **Random Forest:** was chosen for its balance between performance and interpretability.

12.4 False Positive Audit

I manually reviewed a random sample of 100 high-risk false positives. My findings:

1. 74 percent involved long-distance orders during peak hours,
2. 62 percent had session times greater than 90 minutes, and
3. 29 percent canceled under alternate unverifiable reasons.

I conclude that even "false positives" often share the behavioral profile of strategic types—underscoring that the issue may lie in the limitations of ground-truth labeling rather than in the model itself.

12.5 Threshold Stability for Policy Application

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

- The interventions were consistent across multiple validation runs, and
- No rider was flagged inconsistently across folds.

This validates the stability of policy thresholds for real-time deployment.

These checks reinforce my confidence that the detection and scoring systems are robust, interpretable, and usable across platform cycles and real-world operational conditions. I now turn to the broader limitations and future directions.

13 Limitations and Future Work

Even though I presented a novel and empirically grounded framework for detecting strategic behavior on food delivery platforms, several limitations remain—both in scope and methodology. In this section, I outline those gaps and propose future directions for advancing the work.

13.1 No Ground-Truth Verification

The data are based on behavioral proxies rather than verified intent, which means that I cannot definitively confirm whether a cancellation labeled as strategic was actually malicious. Even with precision-focused thresholds, there remains a risk of false attribution.

In future work, I hope to investigate better ways to examine rider intent. For example, I plan to incorporate:

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

These additions could improve the proxy label accuracy or support development of semi-supervised datasets for more nuanced classification.

13.2 No Direct Monetary Cost Attribution

I used time-based proxies to estimate operational impact, but did not quantify monetary costs—such as refunds, food waste, or customer churn.

In future research, I hope to collaborate with platforms that have access to revenue data so I can integrate these losses directly. This would enable more precise estimation of:

- Strategic behavior return on investment (ROI), and
- True economic externalities across the order network.

13.3 Cold-Start Model Development

Although the cold-start model was developed using careful feature engineering and risk heuristics, and while the performance metrics show high precision and recall, I need to expand validation through broader field testing.

Next steps:

- I need to expand validation using larger datasets, and
- I need to deploy the model in a real-world pilot using A/B holdout testing.

13.4 Generalizability Beyond Platform and Geography

My data come from a single Indian platform. Since rider incentives, enforcement norms, and behavioral risk patterns vary across geographies, I need to ensure generalizability.

Extension:

- I plan to run models on US-based or EU-based gig data, and
- I aim to examine how different payout structures (e.g., flat versus variable pay) shape strategic pressure.

13.5 Static Decision Modeling

My model currently assumes static rider decision-making at each task. In reality, riders may learn and adapt based on prior outcomes, reputational feedback (e.g., platform metrics), or support team interactions.

Therefore, in future research, I need to:

- Integrate reinforcement learning or dynamic discrete choice modeling, and
- Track how riders adjust their behavior after platform penalties or interventions.

13.6 Unmeasured Confounders

I do not control for potential confounders such as:

1. **Weather** (e.g., bike breakdowns may be more common during rainfall),
2. **Platform load** (e.g., support bottlenecks may influence cancellation response time), and
3. **City congestion or rider density**.

Data enrichment through public APIs or platform metadata would improve robustness.

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

14 Conclusion

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

My analysis of 447,187 food delivery orders revealed that 90.9 percent 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.

My findings indicate that:

The key contributions of this work include:

1. A theory-driven proxy labeling strategy that identifies strategic behavior through repeated patterns rather than single incidents,

2. Predictive models achieving 0.723 AUC-ROC despite severe class imbalance, with interpretable features aligned to economic theory, and

3. A cold-start risk assessment system enabling fair evaluation of new riders without historical data.

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

My 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 percent increase in likelihood of future strategic behavior.

The operational implications are substantial. Strategic cancellations cost platforms an estimated 1,101 operational hours monthly in my 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.

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 × peak	Interaction term: total distance × is peak hour
distance × fatigue	Interaction term: total distance × session time

These engineered features translate theoretical constructs (e.g., cost, fatigue, outside option value) 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	0.0	720.0
Time to Accept (min)	2.1	4.8	0.0	120.0
Time to Pickup (min)	15.7	12.3	0.0	60.0
Time to Cancel (min)	21.4	18.6	0.0	180.0
Lifetime Orders	145.2	287.2	1.0	5432.0
Cancel Rate (%)	3.5	7.2	0.0	100.0

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

A.4 Glossary of Acronyms (continued)

- **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.

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