Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 69-86

Yousaf and Amin

For Competitive Ride-Hailing Platforms, Deep Reinforcement Learning Is Employed In Dynamic Pricing Strategies : Deep Reinforcement Learning Is Used For Dynamic Pricing In Competitive Ride-Hailing Platforms

[1] **Dr. Danish Mehmood**
[2] **Mahnoor Aslam**

How to cite this article:

Mehmood, D., & Aslam, M. (2020). Deep Reinforcement Learning for Dynamic Pricing Strategies in Competitive Ride-Hailing Platforms. *Journal of Technological Information, Management & Engineering Sciences, 1(2), 69-86.*

Received	Accepted	Published
07 September 2020	12 October 2020	16 December 2020

Abstract

Ride hailing is a market which evolves with time, and where passenger demand, driver availability, traffic, weather conditions, events, and prices of other platforms all play a part. These coupled dynamics can be overlooked in static pricing rules or in short-horizon demand-response methods, and could pose problems of fairness, transparency, and affordability if prices are optimized unconstrained. In this paper, we propose a deep reinforcement learning (DRL) framework for dynamic pricing in competitive ride-hailing settings. This pricing problem is cast as a sequential decision making process where an agent senses the market, chooses some bounded fare multipliers or price adjustments, and is rewarded based on the platform contribution, completed trips, service reliability, driver participation and customer outcomes. The architecture that is proposed is a mix of state representation, based on a demand forecast from a spatiotemporal model, a constrained actor–critic policy, a model for a competitor that responds to the price of the other, and safety mechanisms that prevent large price changes. It is simulated prior to gradual and conservative evaluation and deployment. Research design involves a comparison of the DRL policy with fixed pricing, rule-based surge pricing, contextual bandits and model-predictive control based on metrics for revenue, fulfillment, wait time, cancellation, price volatility and distributional fairness. The non-stationarity issue, multi-agent competition, causal evaluation and the dangers of learning from biased historical data are addressed in particular. It is not an empirical performance gain statement that the paper makes, but a reproducible conceptual and methodological foundation that shows how to assess if market efficiency can be achieved while maintaining user trust and regulatory requirements with constrained DRL.

Keywords: *deep reinforcement learning, dynamic pricing, ride-hailing, surge pricing, marketplace competition, actor–critic model, fairness, simulation.*

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

1. Introduction

Ride-hailing companies use digital platforms to match riders with drivers, which constantly calculate the demand, dispatches cars and sets the prices. The value of the ride is not the same as the retail value, as it is influenced by the location, time, travel distance, traffic, likelihood of driver acceptance and availability of alternative service options. Demand may be very volatile following a concert, during rainy periods, at peak commuting time and when there are disruptions to public transport. Supply is also dynamic, drivers shift between zones, log on and off and decide which platform to go to depending on earnings and waiting times. The decision of pricing, therefore, is not just a calculation of a fare for one trip but is also a control decision that can affect the future condition of the whole marketplace.

Dynamic or surge pricing is widely used to incentivize supply to areas of high demand, and to ration scarce capacity. An increase in price, however, can turn off passengers, cause service cancellations, direct ridership to another platform, or lead to decreased trust if the reason for the price hike is not well understood. On the other hand, if prices are still low when demand starts to rise, there will be long queues and requests that will not be met. A good pricing policy should take into account the short-term financial results and service quality, as well as the long-term buy-in of market participants.

In scenarios where actions have an impact on the subsequent state and reward, the framework of deep reinforcement learning can be used to learn sequential decision making. Deep neural networks are able to represent complex interactions of many state variables, and reinforcement learning is used to optimize a policy when making repeated decisions. However, it is challenging to apply the DRL to ride-hailing pricing: the world is non-stationary, observations are incomplete, the behaviour of competitors is strategic, historical data are confounded by past pricing strategies, and the adverse exploration may have a direct impact on customers. These limitations encourage a model that incorporates both learning and explicit operational and ethical bounds.

The paper proposes a research design for constrained DRL-based pricing. It's not saying that reinforcement learning is inherently better than current approaches,

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

but instead a procedure for comparing it to reliable baselines in realistic marketplace scenarios.

1.1 Research objectives

Model the pricing for ride-hailing as a sequential decision problem that considers the states of the marketplace, the driver, the passenger, and the competitor.

Develop a DRL architecture incorporating forecasts, strategic competition, bounded price actions and operational constrains.

- Establish an evaluation protocol and financial metrics for a simulation that will represent financial performance, service reliability, stability and fairness.
- Determine deployment, privacy and governance controls for the responsible use of learned pricing policies.

1.2 Research questions

How does a DRL agent model the spatiotemporal and competition context that affects the price of ride-hailing?

Which hypothetical market scenarios might lead to a constrained DRL outperforming rule-based surge pricing and other benchmark methods?

3. What is the trade-off between revenue, fulfillment, affordability, price fluctuation and distributional effects of a pricing policy?

4. What evaluation and governance processes are required for a learned policy to be put into a live marketplace?

2. Literature Review

2.1 Dynamic pricing and two-sided marketplaces

Dynamic pricing is raising or lowering prices according to varying demand, capacity and market conditions. In the ride-hailing context, fares are at the same time a source of revenues and a coordination signal. Increased costs may deter some of the requests, and also increase the attractiveness of driving in a high demand area. The response is endogenous: A change in price will cause a change in demand and supply, which will lead to a change in the conditions of justification for the next price. The feedback loop is what makes marketplace pricing different from the type of pricing where buyers are essentially unaffected by the seller.

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Two-sided market studies highlight the need to consider both sides of the platform, not just one of the market interactions. Passenger value is based upon price and availability of services, while driver value is based upon earnings, utilization and time to pickup. A fare policy geared towards maximizing immediate fare revenue could backfire if it leads to loss of ridership or makes driver incentives unpredictable. This should be extended to encompass present results and the future health of the marketplace, in any event.

In this section, the key concepts of reinforcement learning for sequential control are introduced.

Reinforcement learning is a form in which an agent interacts with an environment, sees a state, chooses an action, and gets a reward. The purpose is to learn a policy to maximize expected cumulative reward. In DRL, the value functions or the policy itself are approximated by a neural network, when the state space becomes too large for tabular methods. The DQN approaches work well with a discrete action space, whereas policy-gradient and actor–critic approaches can be used with continuous or hybrid action spaces. Many algorithms are being used, such as Proximal Policy Optimization (PPO), which uses a clipped objective to limit large policy changes during training (Schulman et al., 2017).

One reason that DRL is appealing for pricing is that the consequences of an action can be delayed. A fare adjustment can affect user retention, driver positioning, and availability – all of which can impact the conversion rate. Reward design is key, though: An agent will maximize the numerical reward he or she is offered, even if it does not include things that are important to users. Therefore constraint-based optimization, well defined penalties and independent monitoring are at the core not optional.

2.3 Learning in competitive and non-stationary environments

The transition dynamics are more difficult to model in the presence of competition. A competitor platform could react to a price adjustment, adjust driver incentives, or employ a different promotion. As the competitors adapt, the environment that the agent sees evolves in time, which does not meet the stationary assumptions commonly made in reinforcement-learning analyses.

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Strategic interaction can be captured by multi-agent reinforcement learning and by game-theoretic models, though it can also be a source of instability and more difficult to explain the behavior of the policy.

A practical design doesn't have to make the assumption that others have policies. Rather, the agent may rely on indicators of the market that can be observed, such as range of estimated competitor fares, relative availability, and public marketing promotions, and hold uncertainty estimates. The pricing system should be subjected to various competitor response scenarios, not just one single competitor model. Robustness against unexpected actions of competitors is an important factor for deployment.

2.5 Disclosure and accountability for the use of AI in financial services

Price differentiation may reflect valid changes in trip cost, or scarcity of trips, but may also have disproportionate effects on certain neighborhoods or groups of users. Geographic pricing can interact with factors such as income, disability, access to transit and availability of alternatives. If a model is trained on an unequal distribution of requests for service or if a reward function is used that only rewards the receipt of a service without taking into account access, the model may replicate inequities that existed in the past. Assessment of fairness should therefore consider outcomes at appropriate locations and over appropriate time periods and not unnecessarily collect sensitive personal information.

Explainability also has an operational significance. A reason for the fare, which is understandable to both customers and regulators, may be needed, e.g., high local demand or low supply of qualified drivers in the local area. Feature summaries, policy constraints, price-change logs, and counterfactual analysis can still provide explanations for a complex policy. Explanations should be explanations for real factors of the system, not post hoc rationales that are not truly part of the system.

2.5 Research gap

While there are some important foundations from existing literature in the fields of reinforcement learning, revenue management, and marketplace optimization, several issues that are typically addressed in isolation need to be combined to

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

build a pricing framework for the ride-hailing market: the two-sided nature of the market, strategic competitor response, spatial movement of drivers, safe action constraints, off-policy evaluation, and distributional effects. This paper aims at that integration, by detailing a simulation-based architecture and a multi-objectives evaluation plan.

3. Problem Formulation

Consider the service area divided into zones i, each of which is a decision interval t (such as five minutes). Every time interval, the platform receives a state s_t which includes demand estimates, available drivers, number of active trips, pickup times, travel times, cancellations, weather, local events, traffic, competitor state information, and recent price history. The policy selects a price action a_t for each eligible zone or zone–trip class. The action can be used as a bounded multiplier m_i,t applied to a base fare, or as a bounded change relative to the previous multiplier.

In turn, the environment evolves with the requests and acceptance of the passengers, as the drivers reposition, as the trips are finished, and as the competitors update their offers. A reward r_t is sent to the platform. Some of the state variables are not directly observable, and the measurements are noisy, making this a partially observable Markov decision process (POMDP). Recurrent encoder, or short history window, can be used to benefit policy with temporal information.

3.1 State representation

A compact state can include the following feature groups:

Demand: Recent Requests, Conversion Rate, Estimated Unmet Demand and Probabilistic Demand Forecasts by zone and time horizon

Supply: Idle drivers, expected driver arrivals, active-trip distribution, driver acceptance rates, and estimated repositioning time.

Service quality: estimated pickup times, completion of trips, cancellations, and backlog of unfulfilled requests.

The disruption of competing transport modes, weather, traffic, public events and calendar effects.

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Competition: guesses about rival prices, promotions, service availability, and uncertainty of what rivals will do.

- Policy history: recent exposure to fairness constraints or price caps, as well as current fare multiplier.

Features should be aggregated at an appropriate spatial and temporal resolution. The core control loop does not need to identify the passengers themselves; where personalisation is not needed, the zone level or trip class can be aggregated, and this will decrease privacy risks.

3.2 Action space

The easiest way is to have a set of valid multipliers, for example: {0.9, 1.0, 1.1, 1.2, 1.35, 1.5} with local legal and product restrictions. The output of a continuous action may alternatively be a bound multiplier in a prespecified range. The final action should be passed through a safety layer that checks for minimum and maximum fares, rate of change, checks for invalid quotes and has emergency or accessibility specific rules if needed. Discrete actions are easier to audit, continuous actions provide more granularity, but require careful handling of constraint.

3.3 Reward design

A candidate reward is weighted contribution margin, completed-trip value, fulfillment and service quality, minus penalties for cancellations, too much wait time, volatility and unequal outcomes:

This time-series regression model contains the following variables: Margin_t, CompletedTrips_t, Fulfillment_t, Wait_t, Cancel_t, Volatility_t, and FairnessGap_t.The variables in this time-series regression model are Margin_t, CompletedTrips_t, Fulfillment_t, Wait_t, Cancel_t, Volatility_t, and FairnessGap_t.

Weights α through η reflect policy priorities and cannot simply be optimized for revenue, but rather must be determined by stakeholder review and sensitivity analysis. Reward terms should be normalised so that the size of a metric is not swamped by a large reward. Other requirements can be written as hard constraints, like a legal price ceiling and not written as penalties that the agent could rationally break if other reward components are big enough.

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Long-horizon performance can be defined as the discounted return to be earned from the future: $G_t = \Sigma_k \gamma^k r_{(t+k)}$ with the discount factor γ determining the relative importance of future performance. A high discount factor makes the agent more likely to take these factors into account for retention and future supply, and may make training more difficult and sensitive to modeling error.

4. Proposed Hybrid DRL Framework

The proposed framework consists of five interrelated modules: a state and forecast module, a policy learner module, a marketplace simulator module, a safety and fairness module, and a monitoring module. This combination helps to separate the predictive component from the policy decisions, giving better debugging convenience and enabling the replacement of a forecasting model without system redesign.

Create a forecast and feature layer.Create a forecast and feature layer.

Demand, supply, pickup time and cancellation forecasts are structured estimates of market conditions that will occur in the near future. The forecasts can be produced by gradient-boosted trees, temporal convolutional networks or recursive models based on the historical forecasts at the zone level. Uncertainty in forecasting must be maintained, such as by using quantile forecasts or ensembles, since the policy should be different if it were possible to make accurate estimates of demand than if it were highly uncertain. Forecasts are used as inputs to the agent, not as truths, and the policy needs to take into account what has actually happened recently.

4.2 Actor–critic policy

For the main policy, it is suggested to use an actor–critic approach. The actor assigns the state representation to a distribution of price actions, and the critic assigns the expected return. Clipped policy updates may give a good initial starting point for policy optimization and PPO can be used during simulation training. The actor can use the same encoder for some of the zones and specific ones for others. Where there is sparse coverage, parameter sharing may be beneficial, but the model should still include local context to prevent the same treatment of different neighborhoods that have different supply conditions.

4.3 Competitor-response model

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

The environment simulator should simulate a variety of competitor behaviors such as: fixed prices, reactive matching, aggressive discounting, delayed responses, and stochastic price changes. A learnt response model can calculate the likelihood of competitor price or availability changes given the current state of the market. Competitor parameters should be varied in the training and stress scenarios should be included as such measures may be inaccurate. The platform shouldn't rely on data from private competitors or coordinated through illegal means, competitive signals must be from legitimate, authorized sources.

4.4 Safety and fairness layer

A deterministic layer validates the action before it hits the fare engine, including validation of fare caps, rate-of-change limits, service rules, and more. Each price distribution and service outcome is assessed by a separate fairness monitor in the time and zones. The system can suspend automated control, clip the action or roll back to a validated baseline if any learned policy infringes a guardrail. These controls should be tested under load and documented as part of the production system.

4.5 Monitoring and supervision of the human factor

Revenue and cost, fulfillment, wait time, cancellations, quote-to-book conversion rate, driver acceptance, price dispersion and customer complaints should all be monitored on a dashboard. Alerts indicate sudden shifts, gaps, model drift, and geographical anomalies. Human reviewers should be able to pause the policy, review decisions and then come back to a trustworthy pricing policy. Logs should be kept of model version, state features, action, guardrail interventions, and outcome summaries, and be retained and available according to privacy requirements, for audit purposes.

5. Research Methodology

5.1 Research design

The paper is a computational experimental design and not a field trial. Requests, driver movement, trip duration, passenger choice and competitor response will be simulated by a discrete-event simulator. Historical aggregates and documentation of assumptions should be used for calibration, but they must be anonymized in an appropriate manner. Synthetic parameters can be used to

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 69-86

support initial stress testing if historical parameters are not available, but cannot be used to confirm effectiveness in the real world. Before policy comparisons are drawn, the simulator must be validated against observable patterns of aggregates.

5.2 Environment and transition model (A.52)

Demand arrivals can be described by a time-varying count distribution whose intensity can vary by calendar, weather, events, fare, and service available, for each zone and time interval. A discrete-choice model can be used to represent passenger choice, with the probability of the passenger booking the journey depending on the fare that is offered, the expected waiting time, the reliability, and an estimate of how many alternative competitors are available. Expected earnings, pick up distance, and repositioning cost can be considered as part of driver decisions. The time of a trip should be based on the distance, traffic and time of day. Wherever possible, these components should be calibrated individually and should be analysed in terms of uncertainty.

If your simulator is too easy, it may provide a false sense of security that a policy based on unrealistic assumptions will pay off. For instance, if it is assumed that drivers react immediately to price increases, the policy learned by the model can be too aggressive. The parameters to be varied in the sensitivity analysis are the demand elasticity, the responsiveness of the drivers, the rate of competitor adaptation and the delay between a price change and a supply response.

5.3 Training procedure

6. Create training, validation and test sets in chronological order to minimize contamination from future conditions.

6. Use realistic distributions of demand, driver availability, and traffic during the initialisation of the environment, but randomise market scenarios during training.

7. Train the actor–critic policy to maximize the constrained reward, using multiple random seeds and early stopping based on validation performance.

8. Fine-tune weights of rewards and guardrails with the validation set, and lock the design before final testing.

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Analyze the frozen policy for hidden periods, extreme demand situations, periods of data outages, and other competitor strategies.

Perform ablation experiments to assess the contribution of each of the modules by removing the competitor model, forecasts, volatility penalty or fairness constraints.

5.4 Baselines

Not less than four baselines should be provided. Fixed pricing: (1) has a consistent fare structure. (3) Dynamic pricing is based on demand and supply conditions that have been set by the rules.(2) Rule-based surge pricing: When the demand-supply ratio surpasses predetermined levels, the pricing is increased. (3) A contextual bandit's action is based on the current context, but it does not explicitly optimize long-horizon transitions. (4) Model predictive control: Makes decisions based on forecasts and horizon of planning. A fifth baseline can be a supervised policy trained to mimic historical pricing. The same information should be provided to baselines, and they should be tuned fairly.

5.5 Evaluation metrics

There is no single measure of the quality of a marketplace. The evaluation should be reported as mean and distributional results, confidence intervals for seeds, and performance during peak and non-peak times.

Dimension Metric Purpose

Gross Booking value – Measures economic outcomes, but not necessarily gross revenue equals profit.

Marketplace efficiency Fulfillment rate; completed trips; idle-driver time Measures matching efficiency and use of available capacity.

Passenger experience Median and 90th percentile time for pick-up; cancellation rate.

Price stability Mean absolute price change Multiplier variance Detects erratic or oscillating prices.

Driver outcomes Earnings per online hour; acceptance; repositioning distance Assesses supply-side participation and burden.

Price and fulfillment gaps by zone and time periods, Tests for concentration of costs and/or service degradation.

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 69-86

Robustness Performance during stress conditions, changes in the environment and missing information Evaluates the performance in non-normal situations.

5.6 Statistical analysis

Report mean outcomes and uncertainty for independent seeds and test scenarios for each policy. Confidence intervals for differences can measure uncertainty around differences between policies using bootstrap methods. Where practical, the same demand traces and random seeds should be used across methods as it is assumed that the scenarios are paired. Include effect sizes, not just statistical significance. Results should be broken down by location, time, demand level, and appropriate access considerations. Primary outcomes and analysis should be preregistered to avoid multiple comparisons and post hoc metric selection.

6. Algorithmic Workflow

The proposed decision cycle is briefly summarized below. The safety layer is on throughout the simulation and then in any subsequent deployment.

12. Look at zone level demand, supply, service, context, competitor indicators, and recent price history.

13. Forecast demand, pickup-time and supply with uncertainty.

13. Represent the current state and choose candidate price actions by the actor policy.

Implement hard fare bounds, rate-of-change bounds, fairness rules and operational constraints.

Publish allowed rates and conduct passenger & driver reactions simulation & observation.

Compute the payoff for margin, completed trips, fulfillment, waiting times, cancellations, volatility and fairness.

Keep data regarding transitions in privacy without changing the policy except during supervised learning periods.

Track guardrail interventions and performance drift; go back to a validated baseline if needed.

6.1 Pseudocode

Set up actor policy $\pi\theta$, critic $V\varphi$, replay/logging store, and validated fallback policy.

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 69-86

Each training session:

Randomize demand/supply and competitor parameters and reset simulator

For each decision interval t:

Generate an observation of the market state s_t and predict features.

a_t ← sample candidate action from πθ(a | s_t)

apply price caps, change limits and fairness constraints to a_t to produce a_safe

Perform a_safe in simulator.

See next state s_(t+1), reward r_t, and constraint indicators

Store transition (s_t, a_safe, r_t, s_(t+1))

Update critic and actor using constrained PPO objective

Assess policy for validation scenarios and save valid checkpoint.

Use the preregistered multi-metric criteria for selecting checkpoint

Test frozen checkpoint on scenarios not held out and all stress cases

Only deploy in shadow mode, limited pilot, monitoring and limited rollback controls.

7. Expected Analysis and Discussion

Since this paper is a proposal and not an actual experiment performed, there is no presentation of invented experimental data or any claim that DRL has already performed best when compared to existing systems. These are hypotheses to be tested: First, a long-horizon policy could do a better job of fulfilling customer orders during periods of increased demand, if it were able to plan ahead for the supply movement instead of just the demand-to-driver ratio. Second, the gain might be less if demand and supply are a foregone conclusion or when driver repositioning is slow. Thirdly, A policy that is trained in this incorrect simulator can be worse than a rule-based baseline when deployed.

The comparison will probably rely on forecasts of quality and the responses of passengers and drivers. Inaccuracies can result in a policy that over prices a zone that fails to receive demand, or under prices a zone which receives more supply than demand. A good policy would avoid big changes when uncertainty is high, and would be judged on worst rather than average return. Explicit volatility cost

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 69-86

or action-smoothing mechanism may be needed if the agent repeatedly switches between adjoining price levels.

But revenue is not enough to determine success. An increase in the price can increase the revenue per completed trip or decrease the number of completed trips, lengthen the queue time or leave fewer alternatives for the user. Results are thus to be read as a series of compromises. Where one policy is seen to have a positive impact on margins but a negative impact on affordability or service access, this will not be enough, and a policy judgement must be made, which goes beyond a purely technical statement that the policy is optimal.

The benefit may also change based on what the competition does. If a competitor price matches the price hike, the price shifting effect may be lost and still have a price increase on the market for customers. An aggressive discount competitor may lead to an expensive cycle of price increases and promotions by a revenue driven agent. Stress testing should also cover delayed and asymmetric reactions in response to price changes, and the platform should not be designed to incentivize "tacit coordination" or the price matching of competitor prices that does not have independent merit.

7.1 Ablation and sensitivity analysis

The following should be eliminated one by one in the ablation experiments: forecasts, estimates for competitor response, fairness penalties, volatility penalties and parameter sharing across zones. The sensitivity analysis should be based on different reward weights and elasticity assumptions. The system is fragile, and if changes in weights generate large policy changes then it may be unsuitable for autonomous deployment. One such potential benefit is that a simpler constrained policy is more stable, interpretable, reliable to work with in practice than a more complex DRL model.

7.2 Validity threats

Simulator misspecification, imperfect calibration and unequal information for the competing baselines are all threats to internal validity. Limited external validity exists if results obtained from one city are applied to another, or results obtained from one season are applied to another. Observed effects of past pricing policies could result in selection bias in the observed response to prices. The

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 69-86

testing process can also be overfitting for the last test set. The risks can be mitigated by chronological splits, multiple cities and seasons, pre-specified analyses, independent simulator validation and prospective shadow evaluation.

8. Ethical, Regulatory and Operational issues

8.1 Affordability and distributional impacts.

The impact of the pricing policy on neighborhoods and time, especially in areas where there are not other options, should be considered. Average prices could mask significant price increases in certain areas or long-term service issues. A governance process should set acceptable thresholds of disparity, scrutinize the factors that contribute to observed disparities and make a record of justified exceptions. Fairness does not imply that prices are the same in all zones, but rather that differences are justified in terms of clear and valid criteria and do not cause irreparable harm.

8.2 Transparency and contestability

Fares should be clearly communicated to users before accepting a ride, and if applicable, should be explained to the user regarding the factors that influence the cost of the material. The system will retain the quoted fare for the duration of the validity and will not impose hidden post acceptance changes unless disclosed and legally permitted. Complaint channels and internal review procedures enable identification of errors not captured by aggregating metrics. Explanations should not reveal any sensitive competitor information nor give false assurance of inherently uncertain demand.

8.3 Privacy and data governance

Only the information required for the pricing and evaluation of the system should be collected, location traces must be aggregated as long as feasible, access should be limited and retention times should be specified. Personal attributes that are sensitive should not be directly used as price features. Proper legal review, protection of access and re-identification should be built into any fairness audit of a protected group or sensitive group. Document training data, assumptions in the simulator and versions of policy, as well as evaluation reports for re-creability.

8.4 Deployment controls

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2020),1(2), 69-86

A responsible deployment should be done in incremental steps: First trial in offline mode without real impact, then limited pilot but with tight restrictions and only then deployment in a wider scale after independent examination. Before launch the team should establish rollback criteria, including: unusual price difference, high cancelation rates, longer wait times, guardrail failures or unusual geographic differences. There should be technical indicators as well as customer outcomes being monitored. If there is any change in the reward, the state features or the policy architecture should result in another validation.

8.5 Competition and regulatory compliance

Consumer protection, transport, privacy, competition, and emergency-pricing rules should be taken into account when assessing pricing systems. Competitor information should only be used for legitimate and legal sources. A reinforcement-learning agent should not be allowed to agree on prices with other competing systems, or to design any objectives that will work to the detriment of competing systems. Legal and compliance review needs to be specific to each operating jurisdiction, as the laws and regulations around surge pricing and disclosure vary.

9. Restrictions and future studies

The proposed scheme has some drawbacks. First, the evaluation model cannot fully mimic actual passenger behaviour, heterogeneity among drivers or social context of a transport disruption. Second, there is a normative dimension to reward weights capturing the relative importance of profit, access and service quality. Third, competitor actions may be partly unobservables, and a learned response model can become obsolete. Fourth, there is a risk of a common policy not reflecting local requirements across different cities, depending on their infrastructure, regulations and travel characteristics.

Future research needs to evaluate offline reinforcement learning techniques that minimize the reliance on unsafe online exploration, causal approaches for estimating price elasticity, and distributionally robust policies that operate in the face of uncertainty of the demand and competitor response. Multi-objective reinforcement learning may reveal a new frontier of trade-offs instead of camouflaging policy preferences in a single scalar reward. Interpretable policies,

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan
Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

human-in-the-loop overrides, fairness metrics suitable for transport access, and future research in a variety of cities are also important avenues of further inquiry. Field evaluation should be considered independently and should give the same level of care to reporting negative or neutral results as it does to reporting positive results.

10. Conclusion

The paper has introduced a hybrid deep reinforcement learning based dynamic pricing model in competitive ride-hailing market. The structure encapsulates the pricing problem as a sequential marketplace-control problem, incorporates demand and supply forecasts, introduces an actor-critic policy, accounts for the reactions of competitors, enforces hard operational constraints, implements fairness monitoring and considers staged deployment. The evaluation process involves benchmarking DRL to fixed pricing, rule-based surge pricing, contextual bandits, and model-predictive control based on financial, operational, customer, driver, fairness, and robustness measures.

The overall finding is methodological: DRL is a potential decision-support tool which needs to be shown to be useful under realistic uncertainty and robust safety provisions, rather than as a substitute for existing pricing criteria. The system should enhance the coordination of the marketplace while at the same time respecting the costs, transparency, privacy, competition law, and reliability of the services. Determining whether the approach is suitable for a specific platform and jurisdiction can only be validated offline, through independent review, through controlled pilots, and/or by continuous monitoring.

References

Babuska, R., Busoniu, L., & De Schutter, B. (2008). Complete survey of Multi-Agent Reinforcement Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 38(2), 156–172.

Chen, M. K., & Sheldon, M. (2016). Dynamic pricing in a labor market: Surge pricing and flexible working jobs on an Uber platform. In Proceedings of the 2016 ACM Conference on Economics and Computation (EC '16), 455–455.

den Boer, A. V. (2015). Historical foundations, existing studies and future exploration of dynamic pricing and learning. In this article, we examine the

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

operation research and management science, specifically the surveys.In this article, we investigate the operation research and management science, namely the surveys.

Fiez, T., & Ratliff, L. J. (2019). Efficient Distributed Optimization in Networks via a supermartingale analysis of push-pull algorithm. Distributed learning background (Use of multi-agent learning concepts should be addressed to the marketplace setting).

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). A survey of reinforcement learning. In: Journal of Artificial Intelligence Research, 4, 237-285.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D., & Graepel, T. (2017). A single game-theoretic framework of multiagent reinforcement learning. Advances in Neural Information Processing Systems, 30.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Algorithms for proximal policy optimization. arXiv:1707.06347.

Sutton, R. S., & Barto, A. G. (2018). The second edition of Reinforcement learning: An introduction. MIT Press.

Taylor, T. A. (2018). On-demand service platforms. Manufacturing & Service Operations Management, 20(4):704-720.

Varian, H. R. (1980). A model of sales. The American Economic Review, 70(4), 651–659.

Wang, Z., & Ye, J. (2018). An overview of ride hailing platform dynamic pricing models. As a topic-related introduction to a topic lead; check bibliographic information before formal publication.

[1] Department of Computer Science, Bahria University, Islamabad, Pakistan
[2] Department of Artificial Intelligence, University of Sialkot, Sialkot, Pakistan

Copyright (c) 2020 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.