ASPIRE: Air Shipping Recommendation for E-commerce Products via Causal Inference Framework

Abhirup Mondal
mabhirup@amazon.com
Amazon
Bengaluru, India

Anirban Majumder
majumda@amazon.com
Amazon
Bengaluru, India

Vineet Chaoji
vchaoji@amazon.com
Amazon
Bengaluru, India

ABSTRACT

Speed of delivery is critical for the success of e-commerce platforms. Faster delivery promise to the customer results in increased conversion and revenue. There are typically two mechanisms to control the delivery speed - a) replication of products across warehouses, and b) air-shipping the product. In this paper, we present a machine learning based framework to recommend air-shipping eligibility for products. Specifically, we develop a causal inference framework (referred to as Air Shipping Recommendation or ASPIRE) that balances the trade-off between revenue or conversion and delivery cost to decide whether a product should be shipped via air. We propose a doubly-robust estimation technique followed by an optimization algorithm to determine air eligibility of products and calculate the uplift in revenue and shipping cost.

We ran extensive experiments (both offline and online) to demonstrate the superiority of our technique as compared to the incumbent policies and baseline approaches. ASPIRE resulted in a lift of +79 bps of revenue as measured through an A/B experiment in an emerging marketplace on Amazon.

CCS CONCEPTS

• **Computing methodologies** → **Causal reasoning and diagnostics**; • **Information systems** → Recommender systems.

KEYWORDS

Causal Inference, Individual Treatment Effect, Average Treatment Effect, Doubly-Robust Estimation, Recommendation Systems, E-Commerce

ACM Reference Format:
Abhirup Mondal, Anirban Majumder, and Vineet Chaoji. 2022. ASPIRE: Air Shipping Recommendation for E-commerce Products via Causal Inference Framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539197

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539197

1 INTRODUCTION

In the absence of sufficient data, business decisions are often based on intuitive, broad-brush "back of the envelop" rules, without detailed evaluation of the impact of the decisions. While intuition is important for strategic decision making, following scenarios necessitate data-driven decisions: 1) a manufacturer providing *all* accessories at a 30% discount, to increase the sales of a car, without measuring the impact of the discount, 2) an e-commerce company selling *all* low-cost products only in quantities greater than 3 to minimize cost of shipping, without gauging the impact of this constraint, and 3) a struggling airline broadly offering a $50 discount on *all* segments to entice customers.

At times, organizations conduct small scale pilots to measure the impact of an intervention (e.g., discount, etc.), as the difference in the primary business metric, after and before the pilot period. Even in such situations, the decisions based on a limited pilot are generalized much beyond the scope of the pilot. For instance, results from a pilot on one category of products are generalized across all categories. Simple approaches such as these suffer from inaccuracies due to selection bias. Essentially, for the above decisions, a prudent approach would be to address the *what if* question to arrive at the appropriate decision. For instance, in the first example above, the manufacturer should answer the question - *what* is the impact on revenue *if* they provided a discount of 30% on accessories? A principled data driven response would help determine the appropriate quantum of discount to maximize the revenue per unit discount.

We can generalize the above scenarios within a decision framework consisting of an optimization problem wherein the decision maker aims to maximize (or minimize) a function F (e.g., revenue, profit) under a constraint C (e.g., cost, capacity). The function F depends on quantities Q (e.g., conversion, click) that need to be estimated. Since the quantities Q cannot be estimated directly, due to the bias in limited data collected through small scale pilots, Q have to be estimated within a counterfactual setting. In this paper, we present a decision framework that applies counterfactual reasoning to determine whether a particular product should be eligible for delivery via air-transport (referred to as air-shipping). In this problem setup, the optimization (F) maximizes revenue given the air-capacity constraint (C), by estimating the customer's likelihood of purchasing the product (Q), when the air-shipping option is promised.

Speed of delivery is crucial to the success of e-commerce platforms. Econometric studies have shown that faster delivery promise results in 20% increase in units sold [14], leading to an increase in revenue. Specifically, at the time a customer visits the product page, based on her address, the fastest delivery option (or promise) is shown (Figure 1). Faster delivery also reduces the chances of

Figure 1: Delivery promise as shown on the product search page on Amazon. The promise is calculated by back-end system at the time of customer visit and depends on various factors: customer's location, air eligibility of the product, location of the warehouse etc.

customer pain points such as product damages, pilferage, etc. However, customer's reaction to delivery speed often depends on the product at hand. Certain products such as smartphones, gift cards are more *speed sensitive* than others, i.e., showing faster promise on these products results in greater uplift in conversion. As a result, it is judicious to provide faster delivery promise to speed sensitive products.

From a supply chain and logistics perspective, there are primarily two mechanisms to control speed of delivery. On one hand, we can ensure the products are in-stock (or replicated) across multiple warehouses, such that a warehouse close to the customer can promise 1-day or 2-day (popularly called *fast-track*) (FT) delivery. In the absence of fast-track promise, the *standard* promise defaults to 3-5 days. However, replication is an expensive operation due to the periodic re-stocking costs across multiple warehouses. Furthermore, in emerging marketplaces, where third-party sellers are more common, replication is not a preferred option since sellers may have to bear the shipping costs. On the other hand, air shipping is the preferred choice to meet 1-day or 2-day (fast-track) promises, especially when the product is at a warehouse far from the customer's location.

Since replication involves setting up warehouses, for emerging marketplaces, a significant fraction of shipments are air-shipped. The existing policy allowed air-shipping products that were above a certain price point. As a result, the costs fluctuated significantly as the range of products changed, specifically during holiday seasons. In turn, the required air capacity saw a high variance, resulting in either over or under-utilization of the available air capacity.

We present ***ASPIRE*** (Air ShipPIng REcommendation), a causal modeling framework to determine offline (not real-time), the air eligibility of products based on their historical performance. *Given air capacity (in terms of tonnage), the model estimates conversion uplifts of products and determines which products to fly subject to replication and shipping cost constraints. ASPIRE* has the following advantages.

(1) **Efficient allocation of air capacity:** The existing strategy allowed fast promise for products above a particular price. With *ASPIRE*, we will be able to divert the air-capacity to products where it matters the most.
(2) **Better utilization of air capacity during high-traffic sales or peak periods:** With *ASPIRE*, we will be able to dynamically refresh air-eligibility of products based on the available capacity. This eliminates ad-hoc adjustment to the price threshold during peak periods.
(3) **Allows multiple distribution channels:** Different distribution channels cater to products of varying price ranges. A single price threshold biases air eligibility to a few channels. With *ASPIRE*, we can seamlessly unify all channels within a single framework.

We make the following contributions

(1) We consider the problem of air capacity allocation across a large number of products. Through algorithmic selection of air-eligibility, our goal is to efficiently manage air capacity and drive revenue improvements. To the best of our knowledge this is the first paper that addresses this problem systematically.
(2) We develop a causal modeling framework to estimate the impact of air-eligibility on conversions. The proposed approach addresses the impact of confounding factors such as replication, available air capacity and demand forecasts.
(3) We conduct extensive offline and online (A/B test) experiments to measure the performance of *ASPIRE*. A/B test indicates that the ML based policy results in a 79 basis points[1] improvement in revenue as compared to the incumbent rule based policy.

2 RELATED WORK

Counterfactual reasoning has been applied to a wide range of problems within medicine [11], engineering and e-commerce [1, 12]. We apply counterfactual reasoning to a novel problem of identifying products that should be eligible for air-shipping.

Within counterfactual estimation, multiple techniques have been proposed in order to deal with bias in data. Adjustment by subclassification for observational studies goes back to the early work of Cochran [3]. Subclassification divides the dataset into strata such that the population within each stratum looks homogeneous, conditioned on the confounders. However, this approach was limited to a single confounding variable and could not be extended to multi-variate case.

Conditioning on multiple confounders were considered in the seminal work of Rubin et al [15, 16]. They showed that if the treatment assignment is *strongly ignorable*, i.e., all confounders are observed and if the propensity score is correctly specified then treatment assignment and the vector of covariates are independent, conditioned on the propensity score. It implies that if we bin the propensity score, within each bin, the treatment and control samples look homogeneous. This approach was further extended to estimating causal effects using subclassification [17].

The propensity score, in general, needs to be estimated from data and one can not expect to produce completely homogeneous subclasses. Thus model based covariate adjustment along with subclassification on the propensity score is often used in conjunction [4, 18]. Since then propensity score matching techniques have been extended to the case of continuous treatments [7, 9], multiple treatments [5] and various other scenarios.

[1] 1 basis point or 1 bps = 0.01%.

Doubly Robust Regression [6, 8] techniques combine direct regression and propensity score estimation with the guarantee that the estimator is unbiased if at least either of the regression model or the propensity model is correctly specified. In this work, we apply the doubly robust estimator. Recent works have focused on leveraging state-of-the-art predictive ML models for causal estimation - e.g. neural networks in TARNet [19], arbitrary ML models in DoubleML [2].

3 ASPIRE CAUSAL INFERENCE ALGORITHM

ASPIRE balances the trade-off between conversion/revenue and the cost of air-shipping to decide whether a product should be shipped via air. Although revenue is our target metric, the algorithm can be easily extended to incorporate other objectives in the decision making process e.g. reduction in customer contacts following a purchase or maximizing downstream revenue.

3.1 Notation and Problem Setup

We denote the set of potential treatments or interventions by $\mathcal{T}$. Let $\mathcal{X}$ be the set of context vectors and $\mathcal{Y}$ be the set of potential outcomes. In our setting, $\mathcal{T}$ is binary, indicating standard or fast promise, $\mathcal{Y}$ is binary as well to indicate the event of conversion. Table 1 summarizes our notations.

Table 1: Table of Notations.

m	number of products
n	dataset size, i.e., number of product page views
$a_{1\cdots m}$	products under consideration
$x_{1\cdots n}$	context feature vectors (confounders) corresponding to product page views
$y_{1\cdots n}$	conversion outcome (binary)
$t_{1\cdots n}$	treatment applied (standard/fast promise) on each product page view
$r_{1\cdots m}$	air recommendation for each product
$w_{1\cdots m}$	product weights
$p_{1\cdots m}$	product prices
$d_{1\cdots m}$	demand forecast of products
B	available air-capacity (tonnage)
c_a^t	conversion rate of product a under treatment t
Δ_a	average treatment effect (ATE) for product a

For a context x (represented by features corresponding to the product, customer, merchant, etc.) and each possible treatment $t \in \mathcal{T}$, let $y_t(x) \in \mathcal{Y}$ be the potential outcome for the context. The fundamental problem of causal inference is that only one of the potential outcomes is *factual*, i.e., observed. In our problem setting, this corresponds to the promise that was shown to the customer at the time of visit to the product page and the observed conversion outcome. Even if we show the customer a different promise at a later point of time, the system is not in the same state or context. Therefore, the *counterfactual* outcomes are generally unknown.

For the binary treatment set, the quantity $y_1(x) - y_0(x)$ is often of primary interest. This is known as Individual Treatment Effect or ITE for a given context x. However, as we have access to only the factual outcome, ITE is generally unknown. Another quantity that is of interest is the Average Treatment Effect or ATE which is defined as $\mathbb{E}_{x\sim p(x)}[y_1(x) - y_0(x)]$.

A naive approach for estimating ITE is by *direct regression* where we train a model $f : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{Y}$ on the factual observations $(x_i, t_i, y_i)_{i=1}^n$. The model can be used to estimate the ITE as

$$ITE(x_i) = \begin{cases} y_i - f(x_i, 1 - t_i), & t_i = 1 \\ f(x_i, 1 - t_i) - y_i, & t_i = 0 \end{cases} \tag{1}$$

It is important to note how the task of causal inference is different from standard supervised learning task. Let's assume that factual data follows distribution $\mathcal{P}^F$ and counterfactual data is sampled from $\mathcal{P}^{CF}$. For a supervised learning task, both the training and test data points are sampled from $\mathcal{P}^F$. On the other hand, for causal inference tasks, the training or observation data is sampled from $\mathcal{P}^F$ whereas, the test data points are from $\mathcal{P}^{CF}$. In general, the distributions $\mathcal{P}^F$ and $\mathcal{P}^{CF}$ can be quite different and therefore supervised learning algorithms (including direct regression approach) can't be used for causal impact estimation.

In our first attempt, we trained a LightGBM [10] based classifier to predict conversion. However, it was observed that the treatment variable was not present in the list of top-95% significant features which led to gross underestimation of ITE and ATE estimates. This behavior can be attributed to the presence of *confounder variables* in the feature set. These are the features (e.g., product price, postal code, merchant rating, customer Prime membership status, etc.) that control both the treatment assignment and the outcome. To estimate ITE or ATE, we need to condition on the known confounders, which we achieve through propensity score matching.

3.2 Propensity Score Matching

Propensity score is the probability of treatment assignment conditioned on observed confounders i.e. $p(t \mid x)$. Rubin et al. [15, 16] defined treatment assignment to be *strongly ignorable* if the following conditions hold,

$$(y_0, y_1) \perp\!\!\!\perp t \mid x$$
$$0 < p(t \mid x) < 1$$

The first condition states that treatment assignment is independent of the potential outcomes given the set of covariates (and no unobserved confounders). The second condition states that every sample has a non-zero probability to be assigned the treatment. If treatment assignment is strongly ignorable then conditioning on propensity score allows one to obtain unbiased estimates of ITE and ATE.

There are various propensity score matching techniques to remove the effect of confounding. For our work, we use the subclassification approach which involves stratifying samples into mutually exclusive subsets based on their estimated propensity scores. Quantiles of the propensity score are used to define bin-boundaries for subclassification. If the propensity score is correctly specified, the distribution of covariates for the treatment groups will be very similar within each stratum.

Let $\bar{y}_{0,k}$ and $\bar{y}_{1,k}$ be the sample average of the control and treatment outcomes for stratum k. Then the ATE estimated over K bins

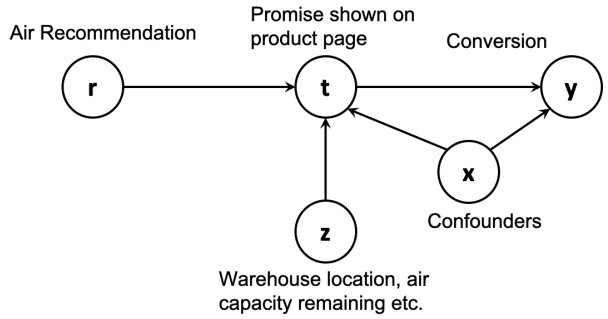


Figure 2: Causal relationship involving air recommendation (r), promise shown on product page (t) and conversion (y). Apart from air recommendation, the promise depends on various latent factors (z), e.g., air capacity remaining at the time of customer visit to the product page, location of nearest warehouse etc.

can be written as,

$$\frac{1}{K}\sum_{k=1}^{K}\left(\bar{y}_{1,k}-\bar{y}_{0,k}\right) \tag{2}$$

3.3 Doubly Robust Estimation

Doubly robust estimator [6, 8] combines direct regression and propensity score matching techniques to create an unbiased estimator for ITE and ATE. Let $f_0, f_1 : \mathcal{X} \rightarrow \mathcal{Y}$ be outcome prediction models for the treatment groups $t = 0$ and $t = 1$, respectively. The doubly robust estimator of ITE given context x is defined as

$$\begin{aligned}\mathcal{DR}(x,y,t) = &\left(\frac{t\cdot y}{p(t\mid x)} - \left(\frac{t-p(t\mid x)}{p(t\mid x)}\right) f_1(y\mid x)\right)\\ &-\left(\frac{(1-t)y}{1-p(t\mid x)} + \left(\frac{t-p(t\mid x)}{1-p(t\mid x)}\right)\cdot f_0(y\mid x)\right)\end{aligned} \tag{3}$$

In our work, we train the classification model for each bin defined on the propensity score, e.g., if there K bins, we train a classifier for each bin and treatment (i.e., all-together there are $2K$ models).

3.4 Estimating Causal Impact of Air Eligibility on Conversion

Through Equation 3, we are able to estimate the causal impact of fast promise on conversion. However, our objective is to estimate the impact on conversion conditioned on air recommendation i.e. to determine $\Pr(y = 1 \mid r = 1, x) - \Pr(y = 1 \mid r = 0, x)$ (refer to Figure 2). Note that these two estimands are different due to presence of latent factors e.g. location of nearest warehouse (factor z in Figure 2).

We make the assumption that $p(t = 1 \mid r = 1) = 1$ since air recommendation allows us to show fast track promises for all the customer visits irrespective of the delivery zip code. Assuming that the joint distribution of the variables y, r and t follows the graphical model in Figure 2, we derive an expression for the lift in conversion due to air eligibility in terms of conversion lift due to fast promise:

$$\begin{aligned}&p(y=1\mid r=1,x) - p(y=1\mid r=0,x)\\ &= p(y=1\mid t=1,x)\cdot p(t=1\mid r=1)\\ &- p(y=1\mid t=1,x)\cdot p(t=1\mid r=0)\\ &- p(y=1\mid t=0,x)\cdot p(t=0\mid r=0)\\ &= p(y=1\mid t=1,x)\cdot (p(t=1\mid r=1)\\ &- p(t=1\mid r=0))\\ &- p(y=1\mid t=0,x)\cdot (1-p(t=1\mid r=0))\\ &= p(y=1\mid t=1,x)\cdot (p(t=1\mid r=1) - p(t=1\mid r=0))\\ &- p(y=1\mid t=0,x)\cdot (p(t=1\mid r=1) - p(t=1\mid r=0))\\ &= (p(y=1\mid t=1,x) - p(y=1\mid t=0,x))\\ &\cdot (p\,(t=1\mid r=1) - p\,(t=1\mid r=0))\\ &= \mathcal{DR}(x,y,t)\cdot (p\,(t=1\mid r=1) - p\,(t=1\mid r=0))\end{aligned} \tag{4}$$

We observe that the lift in conversion due to air recommendation has an intuitively pleasing factorization into two components: $\mathcal{DR}(x, y, t)$ which represents the lift in conversion due to fast promise and $p(t = 1 \mid r = 1) - p(t = 1 \mid r = 0)$ which represents the lift in fast promises shown due to air recommendation. This also leads to the understanding that products which are well replicated will not see a huge change in conversion due to air recommendation as they would have minimum impact on the uplift of fast promises.

The factor $p(y = 1 \mid t = 1, x) - p(y = 1 \mid t = 0, x)$ is estimated via Equation 3. The only unknown quantity is $p(t \mid r)$. There are two counterfactual cases to consider: 1) moving a new product into the air recommended list, 2) purging an existing air eligible product.

Algorithm 1: Algorithm for Air Shipping Recommendation of Products

Input: Context features $x_{1\cdots n} \in \mathcal{R}^k$; treatment $t_{1\cdots n} \in \{\text{fly, no-fly}\}$; outcome $y_{1\cdots n} \in \{\text{conversion, no-conversion}\}$; air-capacity B; product weight $w_{1\cdots m}$; price $p_{1\cdots m}$; demand forecasts $d_{1\cdots m}$.

Output: Air shipping recommendation of products.

Train propensity model $g(x) = p(t \mid x)$ using data-set $\{(x_i, t_i)_{i=1}^n\}$

Perform quantile binning of propensity score into K bins.

Train conversion model $f_b(x, t) = p(y \mid x, t, b)$ for each bin b=1, $2 \cdots K$ using data-set $\{(x_i, t_i, y_i) \mid g(x_i) \in \text{Bin b}\}$

Calculate Doubly-robust estimate of conversion ATE $\Delta_{1\cdots m}$ and conversion rates $c^t_{1\cdots m}$ using Equation 6

Reverse sort products based on benefit scores $\left(\frac{p_i\cdot\Delta_i}{w_i\cdot c_i^{\text{fly}}}\right)_{1\cdots m}$

FLY $:= \emptyset$, $k := 1$, $B^{res} := B$.

while $\left(B^{res} - d_k \cdot w_k \cdot c_k^{fly}\right) \geq 0$ **do**

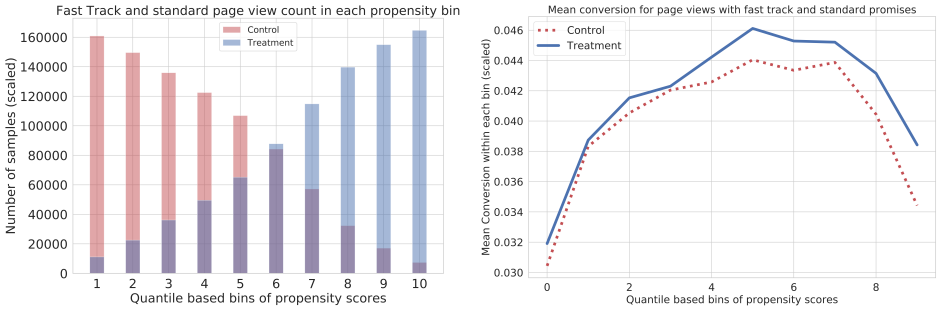
  FLY $:=$ FLY $\cup \{\text{product}_k\}$

  $B^{res} := B^{res} - d_k \cdot w_k \cdot c_k^{\text{fly}}$

  $k := k + 1$

end

Return FLY as air shipping recommendations.



(a) Counts of treatment and control page views across propensity bins. **(b) Naive estimates of conversion rate within each bin.**

Figure 3: Propensity model is used to estimate the likelihood of offering a fast promise at the time of product page view. Plots show behavior of standard (control) and fast promise (treatment) page views across propensity bins. The y-axis is scaled for confidentiality reasons.

For case 2, we can estimate the fraction of demand that offered fast promise through ground transport. This is the demand that would remain unaffected by the change. For case 1, assuming the product demand remains unchanged due to air recommendation, we approximate $p(t \mid r, \text{product} = a)$ as,

$$\sum_k p(t \mid pincode = k) \cdot p(pincode = k \mid r, \text{product} = a) \quad (5)$$

To measure the impact of air eligibility at product level, we group together page views across bins and calculate the average impact. Specifically, if G_a denote the set of views for product a, then the *speed sensitivity* is defined as the uplift in conversion due to air recommendation and is calculated as

$$\Delta_a = \frac{1}{|G_a|} \sum_{(x,y) \in G_a} (p(y = 1 \mid r = 1, x) - p(y = 1 \mid r = 0, x)) \quad (6)$$

3.5 Air Capacity Allocation

Air recommendation of products allows us to show faster promise to customers and results in increased revenue. However, it also increases the delivery cost. Let p_a be the price of the product and w_a be its weight and c_{air} is the cost of air shipping per unit weight. Given fixed air capacity (in terms of tonnage), we are interested in maximizing the total revenue earned through air recommended products.

Let $\lambda_a \in \{0, 1\}$ denotes air recommendation for product a. The expected revenue (for product a) can be written as

$$\begin{aligned} \mathbb{E}(y \mid a) = \lambda_a \cdot p_a \cdot p(y = 1 \mid r = 1, x) \\ + (1 - \lambda_a) \cdot p_a \cdot p(y = 1 \mid r = 0, x) \end{aligned} \quad (7)$$

where the expectation is taken over all customer visits on the product page. Note that λ_a is the only unknown and other quantities are either known or estimated from data. Using Equation 6, the expression can be rewritten as $\mathbb{E}(y \mid a) = \lambda_a \cdot p_a \cdot \Delta_a + p_a \cdot p(y = 1 \mid r = 0, x)$ where the second term doesn't depend on λ_a and can be discarded from the optimization problem we describe next.

The expected air shipping weight for product a can be written as $\mathbb{E}(C \mid a) = \lambda_a \cdot w_a \cdot p(y = 1 \mid r = 1, x)$. Given total air capacity B, air recommendation can be defined in terms of the following optimization problem

$$\begin{aligned} &\max_{\lambda_{1 \cdots m}} \sum_a \mathbb{E}(y \mid a) \\ &\text{s.t.} \sum_a \mathbb{E}(C \mid a) \leq B \end{aligned} \quad (8)$$

This is an instance of the well known 0-1 Knapsack [2] problem. We use a simple greedy algorithm that allows us to scale to millions of products. Algorithm 1 outlines the high level steps of our recommendation algorithm.

4 EXPERIMENTS

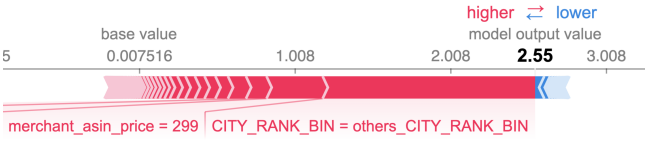
We conduct multiple offline and online experiments to evaluate the performance of *ASPIRE*. For offline evaluation, we study performance of the propensity and conversion models in predicting factual and counterfactual outcomes and obtain offline estimates of expected impact in comparison to baseline policies.

In the online setup, we run an A/B test in an emerging marketplace on Amazon and compare *ASPIRE* with the incumbent policy for air shipping products. A/B test indicates that *ASPIRE* results in +79 basis points improvement in revenue. We also present how our policy changes the delivery promises shown across product page views, thereby positively impacting conversion rates, purchase patterns and other business metrics.

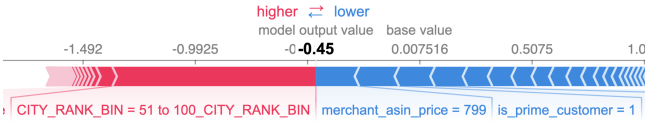
4.1 Propensity and Conversion Models

To estimate the conversion uplift due to fast track promises (refer to Equation 6), we build a propensity model for product page views. The model estimates the likelihood of observing a fast track promise conditioned on confounder variables. We use more than 40MM page views from an emerging marketplace over a period of three months in 2019 and train a LightGBM [10] model. Quantile binning is used on the propensity scores (10 bins are used) and for each bin we train a LightGBM model to estimate conversion. Figure 3a shows the quantile based binning on propensity scores. We observe that,

[2] https://en.wikipedia.org/wiki/Knapsack_problem



(a) Low price and city being low tier leads to lower prediction for a fast track promise ($Pr(T = 1) = 0.13$)



(b) High price and Prime customer status increases the chance of fast track promise, while a low city tier decreases the chance of fast track promise leading to mid range probability of fast track promise ($Pr(T = 1) = 0.47$)

(c) High price and a city being a top tier one increases the chance of fast track promise ($Pr(T = 1) = 0.91$)

Figure 4: Local explanations using SHAP values for samples from different propensity bins: (a) Low propensity bin, (b) Mid propensity bin, (c) High propensity bin.

within each bin there is decent overlap between populations of page views with fast-track promise and with standard promise which results in the conversion models being more robust towards selection bias.

The propensity model achieves an ROC-AUC of 0.86 on out-of-time test set and is observed to be stable in terms of performance across different time periods. Some important features for propensity modeling are *city tier*, *Prime membership status of customer*, *price and category of the product*, etc. Other than treatment allocation, these features also control the event of conversion; as a result, they form the confounding variable set as shown in Figure 2.

Aside from observing feature importances through importance gain, which provides us with a global interpretation of the propensity model, we also run analysis on individual predictions using the SHAP [13] values. SHAP allows us to visualize feature attributions as "forces", where each feature value is a force that either tries to increase or decrease the prediction and these forces balance each other out to arrive at the actual prediction for the instance. In Figure 4, the higher score for the model output corresponds to a lower probability of getting fast track promise. We see that in Figure 4b, the *product price* of 799 and *Prime customer* status increases the chance of a fast track promise whereas the *city* not being a top tier city decreases the chance of getting a fast track promise.

The within-bin conversion models achieve ROC-AUC of 0.79 on an average as measured on out-of-time test sets. Some important features for conversion modeling are *Prime status of customer*, *product price and category*, etc. All models (including the propensity

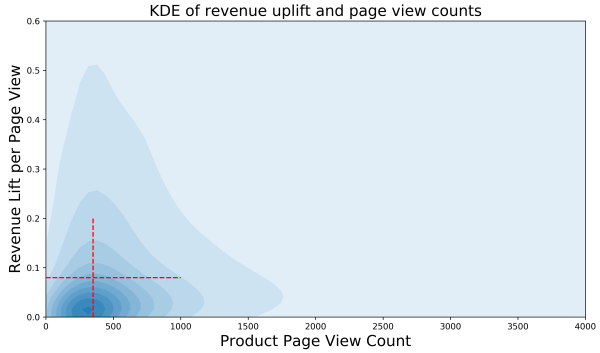


Figure 5: Kernel Density Estimate (KDE) of the joint distribution of page view counts and uplift in revenue.

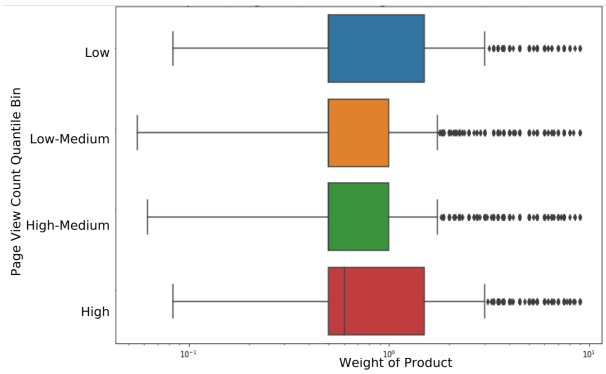


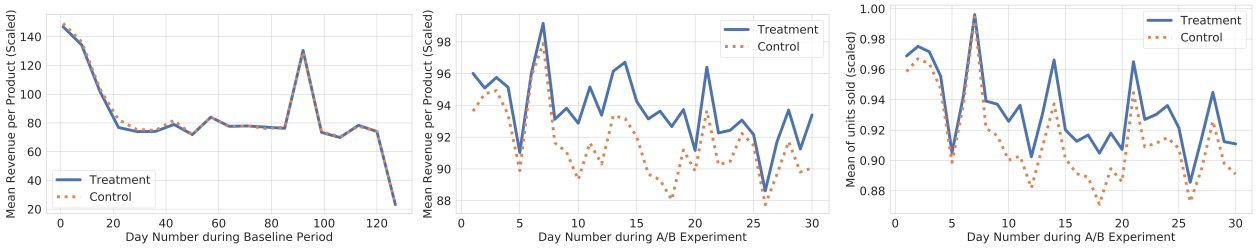
Figure 6: Distribution of product weights across view counts.

model) are calibrated on an out-of-sample validation dataset. The propensity and conversion models are used to calculate benefit scores and rank products as outlined in Algorithm 1 (step 5). It is observed that products such as physical gift cards and travel cards consistently have the highest benefit scores. This is intuitively appealing as these products have high price, high sensitivity to delivery speed and they are of extremely low weight.

We also obtained the ITE estimates using recent techniques such as DoubleML [2] and TARNET [19]. We did not observe significant improvement as compared to the above approach.

4.2 Offline Comparison with Baseline Policies

In this section, we compare *ASPIRE* with other baseline policies for air shipping recommendation. The policies we considered rank products based on their historical a) page view count (proxy for popularity), b) units sold (proxy for high demand), and c) revenue. These baseline policies are simple, intuitive and widely used in e-commerce decision making. Our methodology is the following: we rank products as per the given policy and greedily allocate air capacity to each product till the overall capacity is exhausted. Since incremental revenue and weight are the deciding factors in recommendation, we present detailed offline analysis on how the page view count based policy performs on both these aspects. The



(a) Revenue prior to the experiment period. (b) Revenue during the experiment period. (c) Units sold during the experiment period.

Figure 7: Metrics on the control and treatment cohorts during pre-experiment and experiment period. The cohorts are balanced on price, view count, units and revenue prior to the A/B test (only the revenue stats are shown).

analysis for the other policies are similar and not repeated here. For

Table 2: Offline revenue lift estimates for *ASPIRE* over baseline policies.

Baseline policy	Offline lift estimated for *ASPIRE*
Historical page view count	+14 bps
Historical units sold	+22 bps
Historical revenue	+18 bps

the view count policy, we rank products based on their historical view count over a month. The expected air capacity is estimated using the counterfactual conversion models. Note that the factual and counterfactual conversion estimates are independent of choice of the policy and are fixed throughout our evaluation. Table 2 summarizes the estimated revenue impact of different policies. To analyze the performance of the baseline policy, Figure 5 plots the Kernel Density Estimate (KDE) of joint distribution of revenue uplift and page view count. Note that the set of products which lie in the top left quadrant are better candidates for air shipping than the products lying in the bottom right quadrant. However, the page view count based baseline would select the later ones resulting in sub-optimal recommendations. Figure 6 shows the distribution of weights across bins of product view counts (4 quantile bins are used). The baseline policy selects all products from the highest bin and doesn't even consider the potentially much lower weight products present in the other bins, thereby choosing a sub-optimal set of product in terms of weight.

4.3 Online Experiment

We design an A/B test using live traffic to compare the performance of *ASPIRE* with a baseline rule based policy. The baseline policy allows a product to fly if its price exceeds certain threshold. For this experiment, the products are divided into two homogeneous cohorts (control and treatment) such that their distribution of price, page-view count, revenue and units sold on a weekly basis are matched over a historical period of few months.

On the treatment cohort of products, we use *ASPIRE* for air recommendation whereas the baseline policy is applied on the control cohort. To ensure the treatment and control cohorts are indeed *balanced* during the pre-experiment period, we performed a

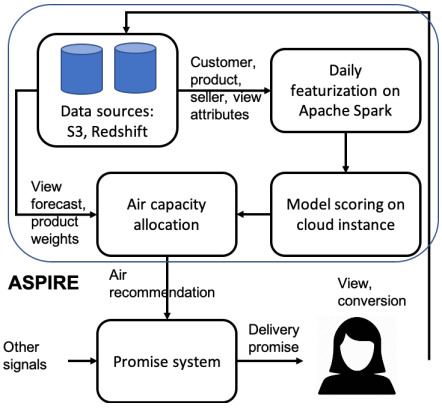


Figure 8: High-level system architecture of *ASPIRE*.

statistical test for the equality of mean revenue generated by the two populations. Stated more formally, let the (unknown) mean revenue of the control and treatment populations are represented by μ_C and μ_T respectively. We test for,

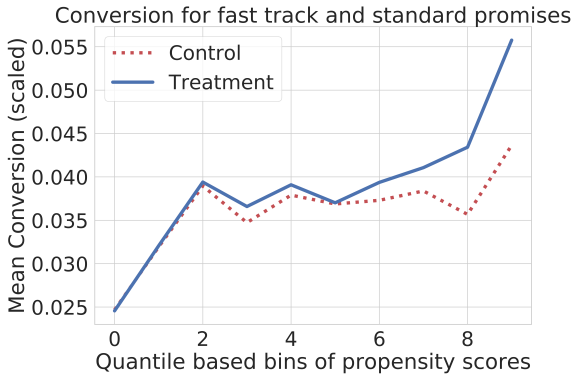
$$H_0 : \mu_C = \mu_T$$
$$H_1 : \mu_C \neq \mu_T \tag{9}$$

We run Wald's test [3] to check equality of means. The null hypothesis is accepted with a p-value of 0.88 which implies the population means are equal during the pre-experiment period. Figure 7a shows the weekly trend of revenue for control and treatment products during the pre-experiment period. The plots for other metrics are similar and not repeated here.

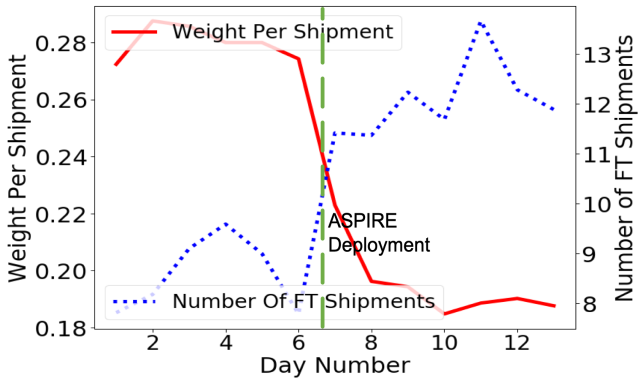
The A/B test was run for an entire month in 2019. Figure 7b and Figure 7c show the weekly trends on revenue and units observed during the experiment period. Experimental evidence suggests that *ASPIRE* achieved statistically significant ($p < 0.03$) +79 bps improvement in revenue. Table 3 highlights the lift observed on several key business metrics.

Interestingly, we observe a lift in page view count by +242 bps. This was counter-intuitive as faster delivery promise was expected to drive more conversion and not have any direct impact on page views. However, this happened because product search prioritizes search results that can be delivered faster to the customer along with

[3] https://en.wikipedia.org/wiki/Wald_test



(a) Conversion rates for fast-track and standard promise post *ASPIRE*.



(b) Weight per shipment and number of fast-track shipments.

Figure 9: Effect of *ASPIRE* on Treatment set during A/B experiment on (a) subclassification plot and (b) weight per shipment and number of fast-track shipments

.

Table 3: Impact of *ASPIRE* on key business metrics.

Metric	Lift
Revenue	**+79 bps**
Conversion Rate	**+182 bps**
Page View Count	**+242 bps**
1-day Tier1 city fast track promise	**+82 bps**
2-day Tier1 city fast track promise	**+264 bps**
Weight per shipment	**-297 bps**

other relevance signals. Therefore, products with better promise are discovered more easily by customers leading to significant increase in page views.

4.4 Implementation of A/B Experiment

Figure 8 shows the high-level architecture diagram of the system used to implement *ASPIRE* for the A/B experiment. The customer, seller, product and view level attributes are pulled from backend data sources (s3/Redshift cluster) to generate features using Apache Spark [20] and scored through the ML models on a cloud instance. The air-recommendation from *ASPIRE* is consumed by the Promise system to surface delivery promises to customers. The promise information and conversion data is fed back to the database for feature generation and model retraining.

In Figure 9, we highlight the impact on the Treatment set due to *ASPIRE* during the A/B experiment. For example, in Figure 9a shows the observed change in the subclassification plot as compared to pattern observed in prior time period (Figure 3b). It shows that we are able to provide fast track promise for products which have higher lift in conversion due to fast track promise. Figure 9b shows the sharp decline (increase) in weight per shipment (fast track promises resp.) due to *ASPIRE* on the Treatment set.

5 CONCLUSION AND FUTURE WORK

In this work, we presented a causal estimation framework (referred to as *ASPIRE*) to give air shipping recommendation to products. We presented a comprehensive set of offline and online (A/B test) experiments to measure the performance of *ASPIRE*. The A/B test results indicate that the ML based policy leads to significant incremental revenue as compared to the incumbent rule based policy.

There are several ways we can extend the functionality of *ASPIRE*. First of all, we can optimize product selection based on downstream effects, e.g., future potential revenue, profits, etc. Since the capacity constraint is global, it is often the case that some air lane capacities are over or under-utilized. This can be avoided by suitably adjusting the optimization criteria in Section 3.5. Finally, we can improve *ASPIRE*'s performance by extending it to an online decision making policy to selectively upgrade the shipping speed at a product page view level. We consider these extensions as future work.

REFERENCES

[1] Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, and others. 2013. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.* 14 (2013), 3207–3260.
[2] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, and others. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (2018), C1–C68.
[3] W. G. Cochran. 1968. The Effectiveness of Adjustment By Subclassification in Removing Bias in Observational Studies. *Biometrics* (1968).
[4] R. B. D'Agostino. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* (1998).
[5] Rhian M. Daniel, Bianca L. De Stavola, and Simon N. Cousens. 2011. Gformula: Estimating Causal Effects in the Presence of Time-Varying Confounding or Mediation using the G-Computation Formula. *The Stata Journal* 11, 4 (2011), 479–517.
[6] Marie Davidian. 2005. Double Robustness in Estimation of Causal Treatment Effects. *https://www4.stat.ncsu.edu/ davidian/double.pdf* (2005).
[7] Christian Fong, Chad Hazlett, and Kosuke Imai. 2018. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics* 12 (2018), 156–177.
[8] Michele Jonsson Funk, Daniel Westreich, and et. al. 2011. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology* (2011).
[9] Keisuke Hirano and Guido W. Imbens. 2004. The Propensity Score with Continuous Treatments. In *Applied Bayesian Modeling and Causal Inference from*

Incomplete Data Perspectives.

[10] Guolin Ke, Qi Meng, Thomas Finley, and et. al. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st NIPS*. 3149–3157.

[11] Samantha Kleinberg and George Hripcsak. 2011. A review of causal inference for biomedical informatics. *Journal of biomedical informatics* 44 (07 2011), 1102–12. https://doi.org/10.1016/j.jbi.2011.07.001

[12] Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. 2015. Counterfactual Estimation and Optimization of Click Metrics in Search Engines: A Case Study. In *WWW '15 Companion*.

[13] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*. 4765–4774.

[14] Xiaosong Peng and Guanyi Lu. 2017. Exploring the Impact of Delivery Performance on Customer Transaction Volume and Unit Price: Evidence from an Assembly Manufacturing Supply Chain. *Production and Operations Management* (2017).

[15] Paul R. Rosenbaum and Donald B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* (1983).

[16] Paul R. Rosenbaum and Donald B. Rubin. 1984. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *J. Amer. Statist. Assoc.* (1984).

[17] Donald B Rubin. 1997. Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine* (1997).

[18] Donald B. Rubin and Neal Thomas. 2000. Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *J. Amer. Statist. Assoc.* (2000).

[19] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. In *Proceedings of the 34th ICML* (Sydney, NSW, Australia) *(ICML'17)*. 3076–3085.

[20] Matei Zaharia, Reynold S. Xin, Patrick Wendell, and others. 2016. Apache Spark: A Unified Engine for Big Data Processing. *Commun. ACM* 59, 11 (Oct. 2016), 56–65. https://doi.org/10.1145/2934664