A/B Testing with Fairness Analysis

Simulated e-commerce experiment — Portfolio project

[Author]- Abhishek R

Synthetic dataset used to demonstrate experimentation + subgroup fairness

Abstract

This report evaluates a simulated A/B test comparing an existing recommendation system (Control) with a new personalization algorithm (Treatment) on a synthetic e-commerce dataset (N = 50,000). Overall, the Treatment increased conversions from **5.10%** (Control) to **7.20%** (Treatment), an absolute uplift of **+2.10 percentage points** (two-proportion z-test, $p < 0.001$). However, subgroup analysis and logistic regression with interaction terms reveal heterogeneous effects: older and high-income users receive the largest benefits, while young females (18–24) and low-income users see the smallest gains. Because the uplift is not uniform across demographics, I recommend delaying full deployment and addressing subgroup disparities through targeted retraining or fairness-aware model adjustments before rollout.

Introduction

This project simulates an A/B experiment to evaluate whether a new personalization algorithm improves purchase conversion on an e-commerce platform and whether any gains are shared equitably across users. The primary metric is **conversion** (binary: purchase or not) observed after users receive recommendations; the experiment compares the existing system (Control) to a new model (Treatment). Beyond the usual question of effectiveness, I place emphasis on fairness: a model that increases overall conversions but disproportionately benefits certain demographic groups can introduce ethical and business risks. The report proceeds as follows: experimental design and data, overall A/B results, subgroup fairness analysis, regression-based tests for heterogeneous effects, and recommendations.

Methods

Primary metric & hypotheses

Primary metric: conversion (binary: 1 = user purchased after seeing recommendation, 0 = no purchase).

Primary hypothesis (effectiveness):

H_0: No difference in conversion rate between Treatment and Control.

H_1: Conversion rate differs between Treatment and Control.

Fairness hypothesis:

H_0_fair: The treatment effect is equal across demographic subgroups.

H_1_fair: At least one demographic subgroup shows a different treatment effect.

Sample size & power

Assumptions used for sample-size planning: baseline (control) conversion = **5.0%**, target detectable absolute uplift = **+2.0 percentage points** (→ treatment = 7.0%), significance level **α = 0.05**, desired power **= 0.8**.

Using a two-sample proportions power calculation (Cohen's *h* via stats models / Normal and Power), the theoretical minimum required per group was **≈ 2,199 users**. To ensure adequate power for subgroup/fairness analysis, I simulated **25,000 users per arm** (total **N = 50,000**).

Randomization & assignment

Users were assigned to Control or Treatment via simple randomization (50/50) to mimic a randomized A/B experiment. Balance checks were performed in the analysis to confirm covariate balance.

Data generation (synthetic dataset)

Rationale: real A/B logs are sensitive and seldom public. I designed a realistic synthetic dataset to demonstrate both classical inference and fairness-aware evaluation. The simulation mirrors typical e-commerce behaviour and deliberately includes heterogeneous treatment effects so subgroup differences can be demonstrated and studied.

Demographics and distributions used:

age_bin: {18–24 (20%), 25–34 (35%), 35–44 (25%), 45+ (20%)}

gender: {Male 55%, Female 42%, Other 3%}

income_bin: {low 30%, mid 50%, high 20%}

Baseline conversion adjustments (relative to 5% baseline): small age/gender/income offsets were added so baseline rates vary realistically (e.g., younger users slightly higher baseline, older slightly lower).

Heterogeneous treatment effects (intentional):

Default uplift for Treatment: +0.02 (absolute).

Young females (age 18–24 & gender = Female): negative effect (–0.01) to simulate an unfair drop.

High-income users: additional positive effect (+0.01) (so they gain more).

Older users (35–44, 45+): small additional uplift (+0.01).

Outcomes were simulated as Bernoulli draws using the per-user conversion probability (baseline + treatment effect when assigned to Treatment).

Reproducibility & code

The simulation used a fixed random seed (seed = 42) for reproducibility. Full code and analyses are available in notebooks/AB_Fairness_Analysis.ipynb and the data file is data/ab_test_fairness_sim.csv.

Ethical note & limitations (brief)

Because the data are synthetic, results illustrate method and interpretation rather than real business outcomes. Synthetic data allowed safe demonstration of fairness checks that are often impossible with public real-world A/B logs. Limitations include simplified demographics and deliberately chosen effect sizes.

Dataset Summary : (N = 50000)

Data	Value
Total users	50,000
Control (count)	24,983
Treatment (count)	25,017
Overall conversion (Control)	1,275 / 24,983 = 5.1035%
Overall conversion (Treatment)	1,796 / 25,017 = 7.1791%
Absolute uplift (Treatment − Control)	+2.0756 percentage points
Age distribution	18–24: 10,000 (20.0%) • 25–34: 17,571 (35.1%) • 35–44: 12,500 (25.0%) • 45+: 9,929 (19.9%)
Gender distribution	Male: 27,453 (54.9%) • Female: 21,120 (42.2%) • Other: 1,427 (2.9%)
Income distribution	low: 14,834 (29.7%) • mid: 25,007 (50.0%) • high: 10,159 (20.3%)
Notes	Data simulated with fixed seed; heterogeneous treatment effects intentionally included.

Table 1 : Overview of the simulated experiment dataset used in this analysis.

Exploratory analysis & overall A/B test

I first verified randomization and measured the population-level effect of the new personalization algorithm. Conversion (binary: purchase or not) was compared between the Control and Treatment arms using descriptive rates, a two-proportion z-test, and 95% confidence intervals.

Control: 1,275 conversions / 24,983 users = 5.1035%

Treatment: 1,796 conversions / 25,017 users = 7.1791%

Absolute uplift (Treatment − Control): +2.0756 percentage points

Relative uplift: ≈ +40.6%

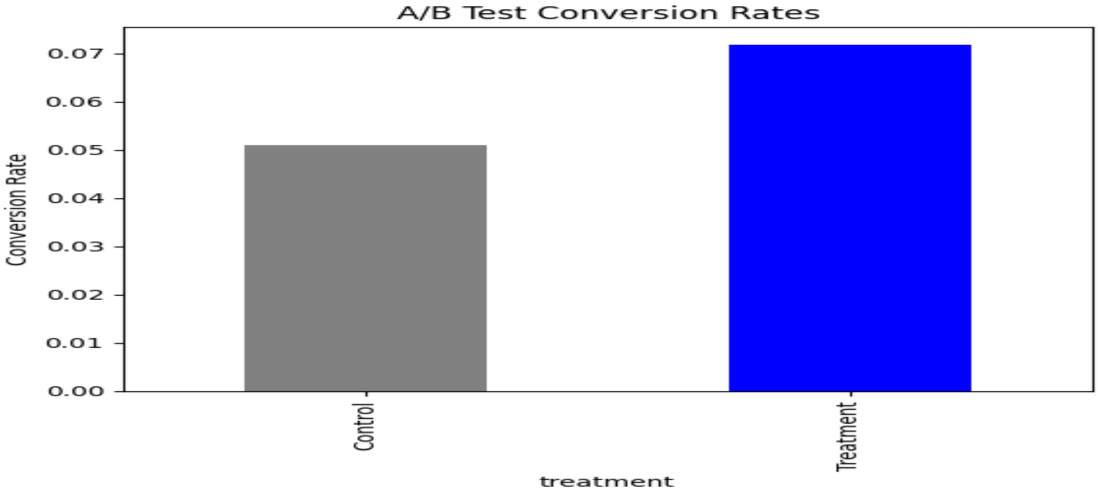

Figure 1 — Overall conversion rate by a/b (Control vs Treatment)

Statistical test & confidence intervals

A two-proportion z-test comparing conversion rates yields Z = −9.665 with p = 4.23×10^{-22} ($p < 0.001$), indicating the uplift is highly unlikely to be due to chance. Approximate 95% (Wilson) confidence intervals are: Control: [4.84%, 5.38%], Treatment: [6.87%, 7.51%]. The intervals do not overlap, supporting the statistical significance of the uplift.

Conclusion:

The Treatment arm produces a statistically significant increase in conversions (7.18% vs 5.10%, absolute +2.08 pp). This confirms the personalization algorithm improves conversions at the population level; next, I examine whether those gains are distributed equitably across demographic subgroups.

Subgroup fairness analysis

While the new personalization algorithm improves overall conversion, an important question is whether all demographic groups benefit equally. To investigate this, I compared conversion rates by age group, gender, and income level. For each subgroup, I computed Control and Treatment conversion rates, absolute lifts, and statistical significance using two-proportion z-tests. The results reveal heterogeneous treatment effects, with some groups gaining more than others.

By Age group

18–24: Control = 6.58%, Treatment = 7.12% (lift +0.54 pp, p = 0.282 — not significant)

25–34: Control = 5.19%, Treatment = 7.03% (lift +1.84 pp, p < 0.001)

35–44: Control = 4.79%, Treatment = 7.40% (lift +2.61 pp, p < 0.001)

45+: Control = 3.88%, Treatment = 7.24% (lift +3.36 pp, p < 0.001)

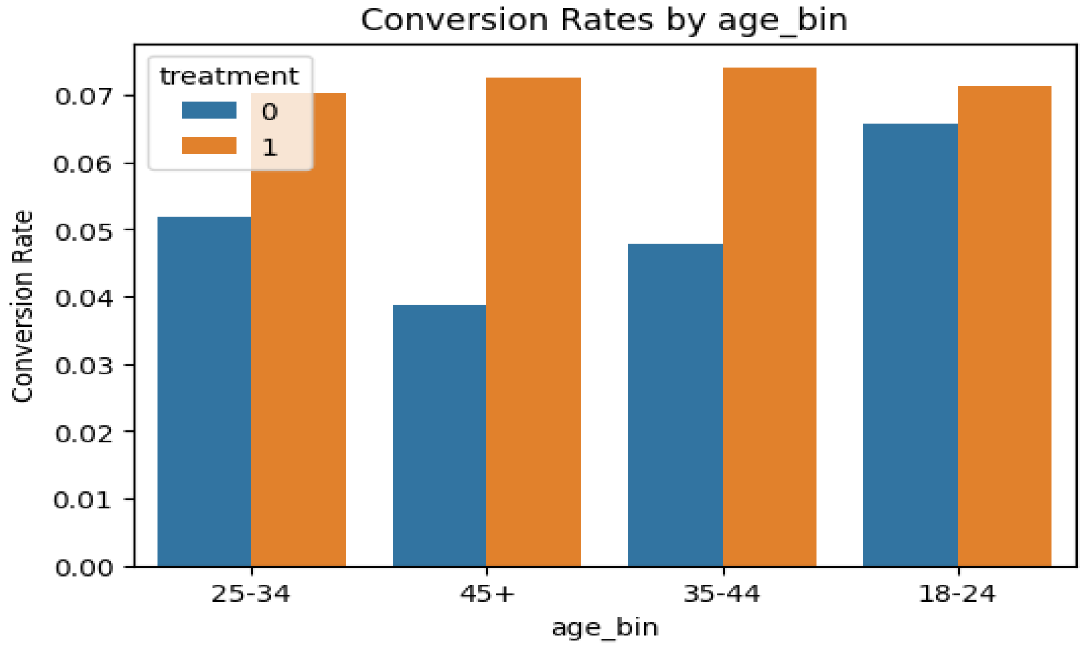

Figure 2 — Conversion rate by age group and arm. Older users benefit more strongly from the Treatment, with the 45+ group showing the largest lift.

By Gender

Female: Control = 4.77%, Treatment = 6.22% (lift +1.46 pp, p < 0.001)

Male: Control = 5.37%, Treatment = 7.89% (lift +2.52 pp, $p < 0.001$)

Other: Control = 5.01%, Treatment = 7.56% (lift +2.55 pp, $p \approx 0.048$)

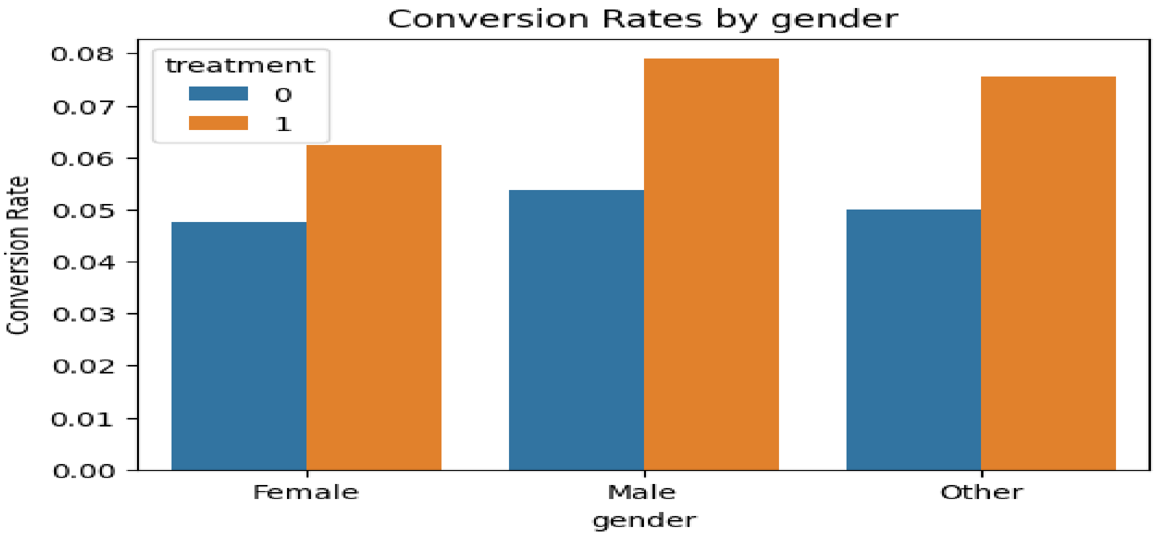

Figure 3 — Conversion rate by gender and arm. Male and "Other" users benefit more than Female users from the Treatment.

By Income level

High income: Control = 6.13%, Treatment = 9.45% (lift +3.32 pp, $p < 0.001$)

Mid income: Control = 4.99%, Treatment = 7.11% (lift +2.13 pp, $p < 0.001$)

Low income: Control = 4.59%, Treatment = 5.75% (lift +1.16 pp, $p = 0.0014$)

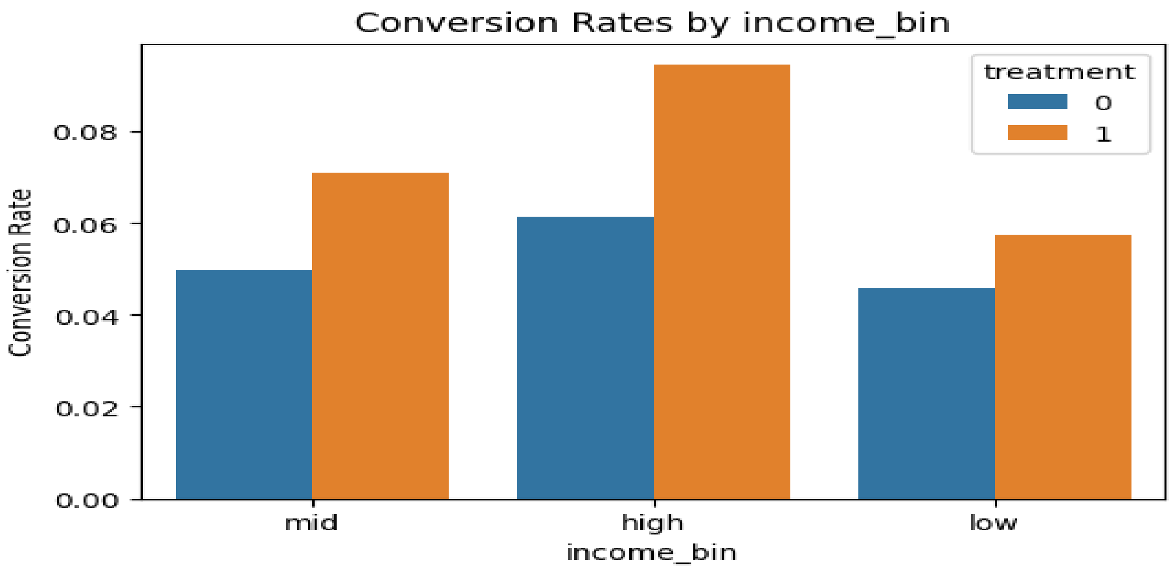

Figure 4 — Conversion rate by income group and arm. High-income users benefit most, while low-income users see much smaller gains.

Conclusion:

Subgroup analysis reveals that the Treatment effect is not uniform. Older users and high-income groups experience the strongest improvements, whereas young users especially 18–24year olds and low-income users benefit less or not significantly at all. Gender differences also emerge, with Female users showing smaller gains than Male users. These patterns highlight a fairness concern: while the algorithm increases overall conversions, it disproportionately advantages certain demographic segments.

Regression analysis

To formally test whether treatment effects vary across subgroups, I ran logistic regression models with conversion (converted) as the binary outcome. Logistic regression estimates the log-odds of conversion and allows me to include both main effects (e.g., treatment, demographics) and interaction terms (e.g., treatment × age group). This method provides a statistical test for heterogeneity while controlling for other variables.

Baseline model (treatment only)

Treatment coefficient = 0.363 ($p < 0.001$)

Odds ratio (OR) = 1.44

Interpretation: Users in the Treatment group were about 44% more likely to convert than those in Control, ignoring demographics.

Model with demographics (no interactions yet)

Demographic variables (age, gender, income) were added alongside treatment.

Treatment remained significant (coefficient ≈ 0.365, $p < 0.001$).

Some demographic differences emerged:

Male users had higher odds of conversion (OR ≈ 1.22 vs Female).

Low-income users had lower odds of conversion (OR ≈ 0.64 vs high income).

Interpretation: Even after adjusting for demographics, the Treatment still significantly improved conversions.

Full interaction model (treatment × demographics)

Adding interaction terms revealed heterogeneous effects:

Treatment × Age 25–34: OR ≈ 1.28 ($p = 0.016$)

Treatment × Age 35–44: OR ≈ 1.46 ($p = 0.001$)

Treatment × Age 45+: OR ≈ 1.78 ($p < 0.001$)

Treatment × Income low: OR ≈ 0.80 ($p = 0.032$)

The main Treatment effect (without interactions) became non-significant, meaning the benefit depends strongly on subgroup membership.

Interpretation: The Treatment provides greater benefit to older age groups, but smaller benefit to low-income users. Gender interactions were weaker and not consistently significant.

Predictor	Odds Ratio	p-value	Interpretation
Treatment (main)	1.11	0.34	Not significant after interactions
Treatment × Age 25–34	1.28	0.016	Higher uplift for 25–34 users
Treatment × Age 35–44	1.46	0.001	Higher uplift for 35–44 users
Treatment × Age 45+	1.78	<0.001	Largest uplift for 45+ users
Treatment × Income low	0.80	0.032	Reduced benefit for low-income users

Conclusion:

Regression analysis confirms the patterns seen in descriptive subgroup comparisons. The personalization algorithm clearly increases conversions, but not uniformly: older users benefit substantially more, while low-income users gain less. This suggests the algorithm may unintentionally widen disparities, raising a fairness concern despite its overall effectiveness.

Discussion

This experiment demonstrates that the new personalization algorithm delivers a clear overall benefit, raising conversion rates from 5.10% to 7.18% (absolute uplift of about +2 percentage points). For a business operating at scale, even a small increase of this magnitude can translate into substantial revenue gains. From a purely effectiveness standpoint, the Treatment is successful.

However, subgroup and regression analyses highlight that the uplift is not evenly distributed. Older users and high-income groups enjoy the largest gains, while 18–24-year-olds and low-income users benefit far less. Female users also experience a smaller uplift compared with males. Logistic regression with interaction terms confirms these patterns: the Treatment effect is strongest for older age groups (odds ratios up to 1.78) but weaker or absent for younger and low-income users (odds ratio ≈ 0.80). These results raise fairness concerns, as improvements skew toward advantaged groups.

From an ethical and practical perspective, deploying the algorithm in its current form could create or reinforce inequities. A system that systematically favors higher-income or older users may undermine trust and limit long-term business value, even if short-term metrics look positive. Businesses increasingly face scrutiny around fairness in AI systems; failing to address disparities risks reputational harm and regulatory challenges.

Overall, the findings suggest that while the personalization algorithm is effective at boosting conversions, it requires further refinement to ensure equitable benefits across demographic groups. Addressing subgroup disparities should be a priority before full deployment.

Recommendations

Based on the results of this simulated experiment, I recommend the following actions before deploying the new personalization algorithm at scale:

Delay full rollout until subgroup disparities are addressed. While overall conversion improves, the uneven benefit poses fairness and reputational risks.

Retrain or adjust the model using fairness-aware methods to reduce the gap between demographic groups, especially for younger and low-income users.

Collect richer feature data to better capture the needs of underperforming groups, enabling more inclusive personalization.

Run targeted follow-up experiments (e.g., A/B tests within low-income or young user segments) to validate adjustments and ensure equity.

Monitor fairness metrics continuously post-deployment, with dashboards and alerts that track conversion by subgroup, not just in aggregate.

Limitations

While this project provides valuable insights, several limitations should be noted:

Synthetic data :The dataset was simulated for learning purposes. Although realistic patterns were introduced, real-world noise, user behaviour, and confounders may differ.

Simplified demographics : Only age, gender, and income level were included. In practice, many other attributes (e.g., geography, device type, browsing history) influence conversions and fairness.

Designed treatment effects :Subgroup differences were intentionally built into the simulation to illustrate fairness issues. The exact magnitudes do not reflect real-world systems.

No business constraints modelled : Costs, revenues, and operational trade-offs were not incorporated; the focus was purely statistical and fairness driven.

Single experiment view : Results come from one simulated A/B test; multiple replications or longitudinal studies would be needed to confirm findings in practice.

Conclusion

This project set out to evaluate the impact of a new personalization algorithm using a simulated A/B test while also examining whether benefits were distributed fairly across user groups. The results show that the algorithm delivers a significant overall uplift in conversions (+2.08 percentage points), confirming its effectiveness at the population level.

However, deeper analysis reveals that the gains are not uniform: older and high-income users see the largest improvements, while younger and low-income users benefit less. Logistic regression with interaction terms confirmed these disparities, raising fairness concerns.

Taken together, the findings highlight an important lesson for real-world experimentation: measuring only overall effectiveness can obscure subgroup inequalities. A model that boosts total conversions may still disadvantage certain users. For businesses, this underlines the need to pair classical A/B testing with fairness analysis before rolling out new systems. Doing so not only protects trust and equity but also strengthens long-term business value.