Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 156-166. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2848

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

EXPERIMENTAL METHODS AND A/B TESTING IN FINTECH: APPLIED GUIDE IN PYTHON

Alexander Haro-Sarango*[1], Julian Coronel-Reyes[2], Johnny Triviño-Sanchez[3], Ibeth Molina-Arcos[4], Gerardo Villarreal-Terán[5] and Gladys Proaño-Altamirano[6]

[1, 4, 5, 6]Instituto Superior Tecnológico España, Ambato, Ecuador.
[2, 3]Universidad Agraria del Ecuador, Guayaquil, Ecuador.

[1]https://orcid.org/0000-0001-7398-2760, [2]https://orcid.org/0000-0002-7883-5388, [3]https://orcid.org/0009-0008-2151-4867
[4]https://orcid.org/0000-0001-9650-1317, [5]https://orcid.org/0009-0003-1384-5184, [6]https://orcid.org/0000-0001-6809-7687

Email: *alexander.haro@iste.edu.ec

ARTICLE INFO

Article History
Received: October 23, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Experimental methods,
A/B testing,
FinTech,
Python,
user conversion.

ABSTRACT

This paper provides an applied guide to the use of experimental methods, particularly A/B testing, in the FinTech sector, leveraging Python-based tools. It examines how these techniques validate business hypotheses, optimize interfaces, personalize financial services, and enhance critical metrics such as conversion and retention. Using a rigorous experimental design, the study evaluates treatment effects on conversion rates, deposits, and user activity through t tests, z tests, bootstrap resampling, and regression models. Results confirm a significant uplift in conversion rates but show no substantial effects on deposits or post-conversion engagement, thereby delimiting the scope of impact. Heterogeneous effects across age segments are identified, with strategic implications for targeted personalization and risk management. The article highlights the importance of incorporating ethical safeguards, transparency, and regulatory compliance when experimenting with sensitive financial data, and proposes a replicable, scalable methodological framework for researchers and practitioners.

I. INTRODUCTION

Over the past decade, the FinTech sector has undergone a significant transformation driven by advances in artificial intelligence, data analytics, and experimental methodologies. The convergence of digital financial services and information technologies has given rise to a new generation of financial products that are more accessible, personalized, and efficient [1]. In this context, the capacity of FinTech firms to make decisions grounded in empirical evidence has become a key source of competitive advantage [2]. Experimental methods and A/B testing in particular have emerged as fundamental tools to validate business hypotheses, optimize user interfaces, personalize financial services, and improve key metrics such as customer retention and customer lifetime value (CLV). A/B testing enables the comparison of two or more variants of a digital product such as a web interface, a personalized interest rate, or a marketing campaign with the aim of determining which one delivers superior performance with respect to a predefined set of metrics[3], [4].

Originating in the domain of classical experimental design, this methodology has been adopted and adapted by high-growth technology companies such as Amazon, Netflix, and Google, and it now constitutes standard practice in the FinTech environment [5]. Nevertheless, the rigorous implementation of A/B testing in financial settings presents substantial challenges: selection bias, interference between users, issues of external validity, and ethical dilemmas associated with the manipulation of financial products [6], [7]. The situation becomes even more complex when experiments are conducted in real time with highly sensitive data, where errors can entail economic and regulatory consequences [8]. In addition, the explosion of data and the adoption of microservice architectures demand that these experiments be automated, reproducible, and scalable a requirement that has been facilitated by modern Python tools such as statsmodels and SciPy, as well as specialized libraries like Evidently, Optimizely, and PyAB. The integration of these technologies has enabled both startups and established banks to implement continuous experimentation systems [9], [10].

This article provides an applied guide to implementing experimental methods and A/B testing in the FinTech context using Python, with the objective of bridging the gap between statistical theory and its practical execution in production systems. The guide is grounded

in recent studies, industry best practices, and real-world use cases in the digital financial industry. It also discusses extensions such as multivariate testing, Bayesian testing, and adaptive experimentation, offering a comprehensive and up-to-date perspective for researchers, developers, and data scientists in FinTech.

II. LITERATURE REVIEW

A/B testing, also known as a randomized controlled experiment, is a widely used technique in applied research to establish causal inferences between a manipulated variable and an observed outcome. Its rise in digital environments is attributable to platforms' ability to randomly assign treatments to large user populations and to record resulting behavior with precision [3]. By directly comparing two variants (A and B), the approach assesses which one yields superior performance on predefined metrics such as conversion rate, user retention, or profitability. On platforms such as Amazon, Google, or Netflix, this technique has been fundamental to the continuous improvement of interfaces and personalized services. [3] and [8] describe how controlled experimentation becomes an operating system for innovation when integrated with automated tooling and rigorous statistical principles. These advances have been transferred to the FinTech domain, where precise measurement of the impact of algorithmic changes, pricing strategies, or design interventions is essential.

The FinTech sector presents particularities that complicate the application of A/B testing: regulatory compliance, the sensitivity of financial data, high interdependence among users, and the need for explainability in automated decisions [9]. and [10].notes that A/B tests can be useful for evaluating alternative credit-scoring tools, financial inclusion strategies, or digital identity validation mechanisms. Nevertheless, such experiments must be conducted under ethical oversight, especially when treatments affect sensitive decisions such as credit approval, assignment of credit limits, or insurance personalization. [11-13] caution that, although randomized experiments are methodologically robust, they can generate unintended consequences if interference effects among users or historical algorithmic biases are not addressed. In this regard, [14], [15] propose the design of experiments on social networks using cluster randomization techniques that control spillovers among connected nodes a particularly useful practice in FinTech applications based on reputation systems or peer-to-peer lending.

Beyond the classical hypothesis-testing approach based on p-values, authors such as [16] and [17] have proposed Bayesian frameworks for experimentation that allow for a more intuitive interpretation of results and the dynamic updating of evidence as data are collected. This is especially relevant in FinTech, where contexts can change rapidly, and decisions must adapt to emerging user behaviors. Bayesian methods yield posterior distributions for the parameters of interest, thereby facilitating probabilistic decision-making. [17] introduces the multi-armed bandit approach, which dynamically adapts treatment allocation to maximize expected value, thus reducing exposure to ineffective alternatives. [18] also highlights the role of Bayesian A/B testing as an accelerator in product-validation processes. These approaches are ideally suited to financial contexts in which risk is inherent to decision-making, and it is necessary not only to identify the superior alternative but also to quantify uncertainty and minimize the associated losses.

From a computational standpoint, Python has become the dominant language for experimentation in data science due to its flexibility, active community, and rich library ecosystem. Tools such as statsmodels, SciPy, PyAB, Evidently, and Bayesian PyAB enable the implementation of A/B tests, analysis of key metrics, automated visualizations, and real-time monitoring of experiment validity [19]. Additionally, commercial platforms like Optimizely and Google Optimize offer APIs that integrate with Python workflows, facilitating the execution of experiments within production data pipelines. [20], [21] propose microservice-based continuous experimentation architectures, which are especially relevant for FinTechs seeking to scale products across multiple market segments. Tools such as MLflow, Dagster, and Airflow make it possible to combine experimental design with DevOps and MLOps practices, thereby facilitating reproducibility, auditing, and traceability crucial aspects for compliance with financial regulations such as GDPR, PSD2, and banking-supervision requirements.

The use of A/B testing in FinTech raises important ethical challenges. Experimental decisions can affect a user's eligibility for credit, insurance, or investment services; consequently, transparency and fairness become essential pillars [6], [22]. Recent literature emphasizes the need to audit algorithms and to ensure that tests do not reinforce existing biases or generate indirect discrimination. [22] introduce the concept of "algorithmic audits" to evaluate the social impact of automated systems, while [23] analyze mechanisms for explainability and accountability in financial algorithmic contexts. These perspectives are fundamental when implementing experimental methods that affect sensitive variables such as interest rates, credit limits, or approval rates. Therefore, the specialized literature broadly supports the use of A/B testing in FinTech, provided that robust methodological safeguards, appropriate technological tools, and ethical principles aligned with fairness and user protection are adopted.

III. MATERIALS AND METHODS

From an anonymized dataset of 10,000 customers of a Segment-1 cooperative savings and credit institution in Ecuador, the study adopts a classical A/B randomized experiment to estimate the causal effect of a digital Fintech intervention (e.g., a redesigned onboarding flow, a promotional banner, or a new deposit screen). Each observation corresponds to a unique user exposed once during the study window. Assignment to control A or treatment B was performed via simple randomization with probability $P(B) = 0.5$, without replacement and at the user level. Under protocol compliance, the treatment indicator $D_i = \mathbf{1}\{i \in B\}$ is independent of the potential outcomes $(Y_i(0), Y_i(1))$. This independence guarantees identifiability of the Average Treatment Effect (ATE) as

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]. \quad (1)$$

The primary outcome is the binary conversion $Y_i^{\text{conv}} \in \{0,1\}$, which flags whether the user completes the target action (e.g., onboarding or first deposit). Secondary outcomes capture usage intensity: deposited amount among converters $Y_i^{\text{dep}} \geq 0$, number of logins login_count_i, and active days days_active_i.

Pre-treatment covariates used for balance checks and precision gains include age_i, self-reported $income_i$, and historical activity metrics prior to the intervention. Duplicates were removed, missingness was documented, and monetary variables were winsorized at the

top 1% to mitigate outliers while preserving mean interpretability. Before effect estimation, balance between groups is validated through mean-comparison tests on pre-treatment covariates. For each continuous variable $X \in \{age, income,$ login_count, days_active$\}$ we apply a two-sided Welch t-test:

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}, \tag{2}$$

And report the associated p-value. Systematic non-significance (e.g., $p > 0.05$) supports proper randomization and thus the experiment's internal validity. Overlaid histograms by group are inspected to rule out substantial imbalances not captured by the mean. The main hypothesis tests whether conversion rates differ between A and B. Let $p_A = \mathbb{E}[Y^{\text{conv}} \mid D = 0]$ and $p_B = \mathbb{E}[Y^{\text{conv}} \mid D = 1]$. The two-proportion z-statistic is

$$\hat{z} = \frac{p_B - p_A}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}, \qquad \hat{p} = \frac{n_A p_A + n_B p_B}{n_A + n_B} \quad , \tag{3}$$

Under $H_0: p_B = p_A$. We report point estimates, a 95% confidence interval (CI), and a two-sided p-value. Business magnitude is also communicated as the relative lift,

$$\text{lift} = \frac{p_B - p_A}{p_A}. \tag{4}$$

For the deposited amount among converters, we estimate a conditional mean effect by comparing Y^{dep} in the sub-sample $\{i: Y_i^{\text{conv}} = 1\}$. Given its continuous and skewed nature, we contrast mean differences via Welch's t-test and accompany the analysis with boxplots to visualize median and dispersion; when appropriate, robustness is checked with a 10% trimmed mean or a bootstrap difference in medians. To characterize treatment heterogeneity, the sample is stratified ex-post by age bands $S \in \{< 30,\ 30\text{–}50,\ > 50\}$ using binning on $[18{,}30{,}50{,}70)$. For each segment we estimate

$$\hat{\tau}_S = \mathbb{E}[Y^{\text{conv}} \mid D = 1, S] - \mathbb{E}[Y^{\text{conv}} \mid D = 0, S \tag{5}$$

These analyses are descriptive and interpreted cautiously to avoid multiple-comparison pitfalls; when a segment shows signal, we complement with bootstrap CIs and recommend future ex-ante stratification in the sampling design. Uncertainty around the ATE in conversion is further quantified via nonparametric bootstrap with independent resampling within A and B. For iterations $b = 1, \ldots, B$ (with $B = 5000$), we draw with replacement $A^{(b)}$ and $B^{(b)}$ matching original sizes and compute $\hat{\tau}^{(b)} = \bar{Y}_{B^{(b)}} - \bar{Y}_{A^{(b)}}$. The 95% CI is given by $[q_{2.5},\ q_{97.5}]$ of the empirical distribution $\{\hat{\tau}^{(b)}\}$, providing a normal-assumption-free interval particularly useful with low rates or unequal sizes. To gain precision and control residual noise, we estimate a linear probability model of conversion on treatment and pre-treatment covariates:

$$Y_i^{\text{conv}} = \beta_0 + \beta_1 D_i + \gamma_1\, age_i + \gamma_2\, income_i + \gamma_3\, \text{login\_count}_i + \gamma_4\, \text{days\_active}_i + \varepsilon_i. \tag{6}$$

Under randomization and protocol compliance, the causal interpretation of β_1 coincides with the ATE, and we report OLS estimates with heteroskedasticity-robust (HC3) standard errors. As a sensitivity check, we also fit a reduced model

$$Y_i^{\text{conv}} = \beta_0 + \beta_1 D_i + \eta_i, \tag{7}$$

To verify that adding controls improves efficiency without altering the sign or order of magnitude. For continuous outcomes (Y^{dep}, login_count, days_active) we apply analogous OLS specifications; if heavy tails persist, we test log-transformations on $(Y^{\text{dep}} + c)$. Placebo and bias validations include tests on variables that should not respond to treatment, such as pre-intervention income. A Welch t-test on $income$ between A and B is expected to be non-significant; otherwise, we diagnose potential randomization issues or attrition. We additionally inspect missingness rates by column and the distribution of sample sizes by group to rule out differential measurement bias. Inference adopts $\alpha = 0.05$ with two-sided tests, and Holm–Bonferroni adjustments are noted when evaluating multiple secondary outcomes. For ex-ante statistical power, we use the standard approximation for a difference in proportions with 1:1 allocation. Given significance α and power $1 - \beta$, the Minimum Detectable Effect (MDE) around a baseline rate p satisfies

$$\text{MDE} = \left(z_{1-\alpha/2} + z_{1-\beta}\right)\sqrt{\frac{4\,p(1-p)}{n}}, \tag{8}$$

Where n is the total sample size. This quantifies whether $n = 10{,}000$ is adequate for the business-relevant lift. Although the analysis is ex-post, reporting the MDE contextualizes the sensitivity of findings. The workflow is implemented reproducibly in Python: pandas for data handling, scipy.stats for t- and z-tests, statsmodels for OLS with robust errors, and plotly for interactive figures. The pipeline proceeds through reading and cleaning, balance verification, the primary conversion test, analysis of financial and activity metrics, age-based heterogeneity, bootstrap uncertainty quantification, and robustness regressions, followed by placebo and experiment-quality checks. Code is organized in self-contained cells to facilitate auditability and replication. Regarding validity, randomization and balance checks support internal validity; external validity is restricted to the cooperative's clientele and time period analyzed. The intervention is purely digital and does not alter contractual conditions or fees, so participant risk is minimal.

Privacy and data-security principles are observed through irreversible anonymization, attribute minimization, and exclusive use for product-evaluation purposes, consistent with the local personal-data protection framework. Finally, interpretation is aligned with

Fintech decision-making: if the p-value is below 0.05 and the ATE CI excludes zero with business-relevant magnitude, the *B* variant is recommended; if effects are null or uncertain, we favor the lower-cost alternative or iterate the design incorporating segment learning with ex-ante stratification. This protocol offers an applied, mathematically grounded, and operational guide for product teams in cooperative financial institutions, scalable to continuous experimentation cycles with robust causal rigor.

IV. RESULTS AND DISCUSSIONS

It is the logical presentation of the results that demonstrate the true contribution of the research and also justify the conclusion. The parts indicated above, Introduction, Material and Methods serve to explain how the results are obtained. Use tables and figures for your explanation. The results must be written in the past, in a brief and simple way. Avoid redundancy. Analyzing and interpreting the results is part of the discussions, their importance, achievements and limitations, highlighting the innovative aspects of the practical applications of the study and the conclusions derived from them, delimiting unresolved issues. If necessary, recommendations can be proposed. In this first stage of the experimental analysis, the aim is to verify that the random assignment of users to groups A (control) and B (treatment) was effective that is, that both groups are comparable prior to any causal analysis.

To this end, several observable variables are examined: age, income, number of logins (login_count), and days active on the platform (days_active). Validation is performed by comparing the means of these variables across groups using an independent-samples t test. This procedure assesses whether there are statistically significant differences in the means of the variables considered. The p-value obtained in each t test indicates the probability of observing a difference as large as the one found if, in fact, no difference exists in the population. In this case, the p-values were 0.2378 for age, 0.9431 for income, 0.4110 for login count, and 0.8552 for days active. All of these values are well above the conventional 0.05 threshold, so we conclude that there are no significant differences between groups. The analysis is complemented by a descriptive summary of the means for each variable in both groups.

For example, the average age in group A was 34.94 years, whereas in group B it was 34.72 years. Similarly, average income was virtually identical: 30,052 for A and 30,066 for B. The average number of logins was 5.01 in A versus 4.98 in B, and days active were 14.98 and 14.95, respectively. These figures indicate that both groups exhibit nearly identical characteristics, reinforcing the validity of the randomization. To visualize these results, histograms generated with the Plotly library were used. These plots display how each variable is distributed within each group and allow any notable overlaps or differences to be observed. The distributions for age, income, logins, and days active show a high degree of visual similarity between groups A and B, confirming the statistical findings (see Table 1).

Table 1: Step 1 – Validation of balance between groups.

Variable	p-value	Mean Group A	Mean Group B
age	0.2378	34.94	34.72
income	0.9431	30,051.88	30,065.77
login_count	0.4110	5.01	4.98
days_active	0.8552	14.98	14.95

Source: Authors, (2026).

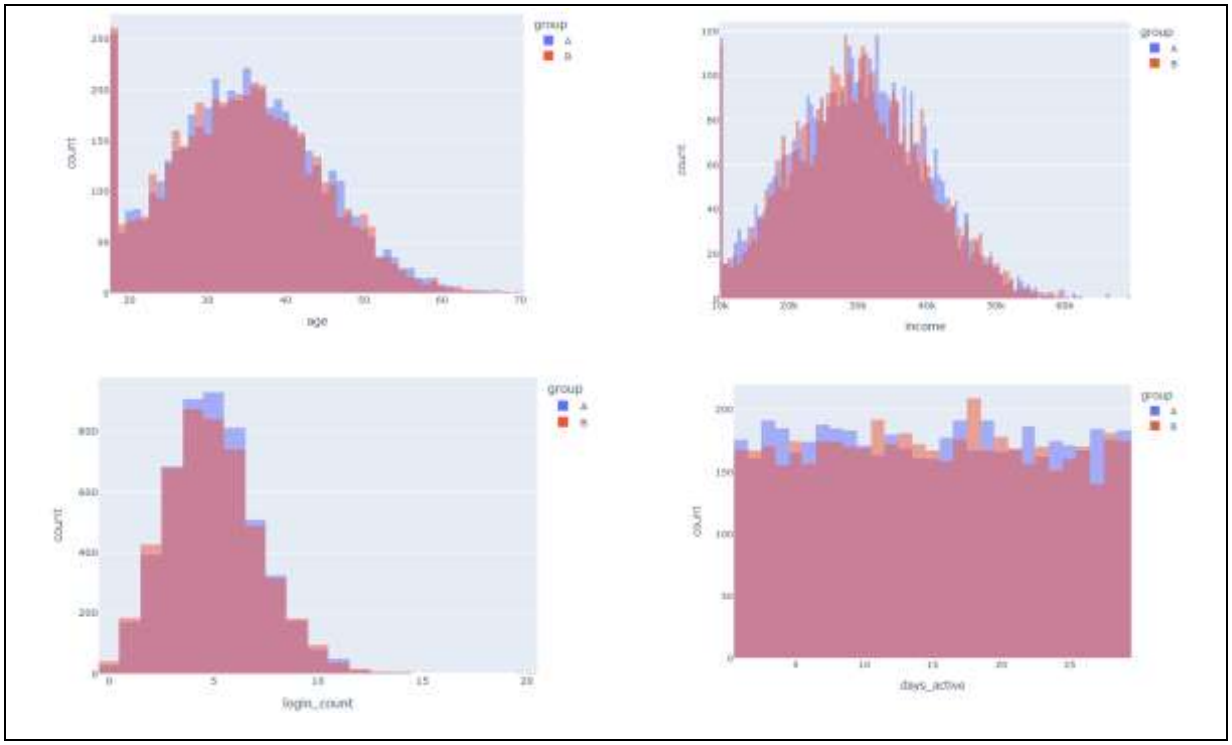

Figure 1: Step 1 – Distribution among groups.
Source: Authors, (2026).

In the second step of the analysis, we evaluate whether the treatment applied to group B had a significant effect on conversion; that is, whether the proportion of users who performed the desired action (e.g., opening an account or completing a registration) is higher relative to group A. To this end, we compute the conversion rate in each group and then apply a formal statistical test. In this case, a z test for the difference in proportions specifically designed to compare the success rates of two independent groups was used. The results show that the conversion rate in group A was 0.0985, whereas in group B it was 0.1509. This difference represents a relative increase of more than 50%. To determine whether this difference is statistically significant, we calculated the z statistic, which was −7.9389. The associated p-value was less than 0.0001, indicating that the probability of observing so large a difference by chance alone, if no true effect existed, is virtually zero. Consequently, we reject the null hypothesis of equal proportions and conclude that the treatment had a significant effect on conversion. The observed difference is also depicted with a bar chart, which clearly shows that the height of the bar corresponding to group B is substantially greater than that of group A. This provides a clear and intuitive visualization of the treatment's positive effect. In sum, this analysis offers strong statistical evidence that the intervention applied to group B significantly increased the user conversion rate (see Table 2).

Table 2: Step 2: Conversion analysis and hypothesis testing.

Metric	Group A	Group B	Test Statistic	p-value
Conversion Rate	0.0985	0.1509	z = -7.9389	0.0000

Source: Authors, (2026).

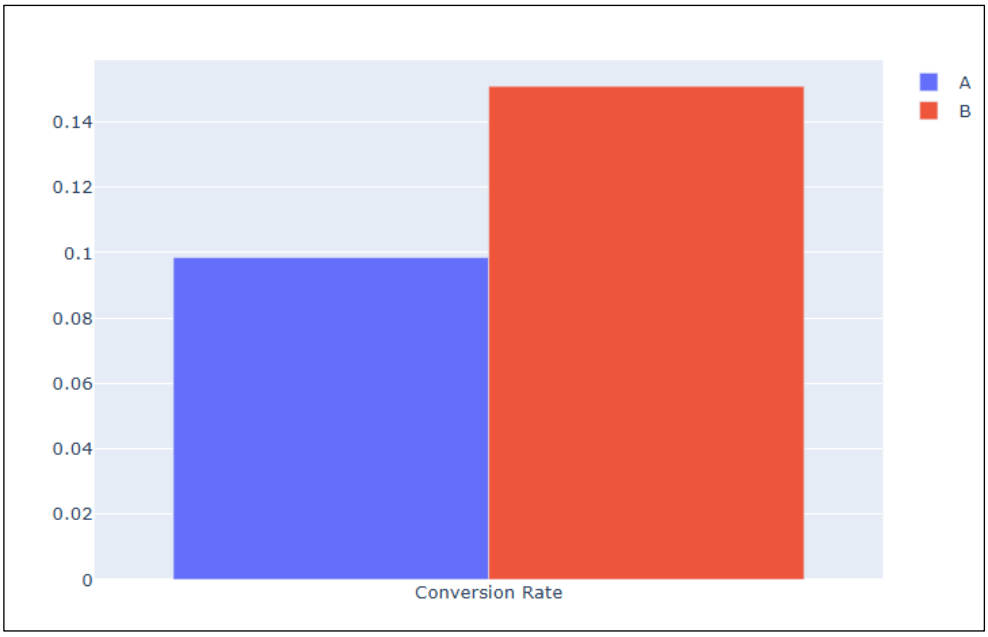

Figure 2: Step 2: Conversion rate by group.
Source: Authors, (2026).

In this third step, we analyze whether the treatment affected not only conversion but also users' subsequent financial and operational behavior specifically, the amount deposited, the number of times they logged into the platform, and the days they remained active. These indicators are crucial for assessing conversion quality: it is not sufficient to convert more users if they subsequently do not interact or generate revenue. To begin, an independent-samples t test was applied to compare the deposited amount between groups A and B, considering only users who actually converted. The result was a t value of −0.4683 and a p value of 0.6397, indicating no statistically significant difference between the two groups in terms of deposited amount. Although the group B distribution appears visually to have a slightly higher median, variability in the data and overlap between distributions render this difference inconclusive from a statistical standpoint.

Boxplots make this pattern clear. In the deposited-amount plot, both the median and the dispersion are similar across groups. A handful of outliers appear on both sides as is to be expected with financial data but there is no evidence of a systematic treatment effect. With respect to the number of logins (login_count) and days active (days_active), the boxplots likewise show very similar distributions for groups A and B, both in terms of median and interquartile range. This suggests that post-conversion behavior, in terms of platform use, was not altered by the treatment. Although group B achieved a significantly higher conversion rate, the users who converted in that group did not deposit more money nor were they more active than those in group A. This is an important finding because it helps distinguish between superficial conversion and conversion that yields real business value (see Table 3).

Table 3: Step 3: Analysis of Financial Metrics (Amount and Activity).

Metric	Test Statistic	p-value
Deposited Amount	t = -0.4683	0.6397

Source: Authors, (2026).

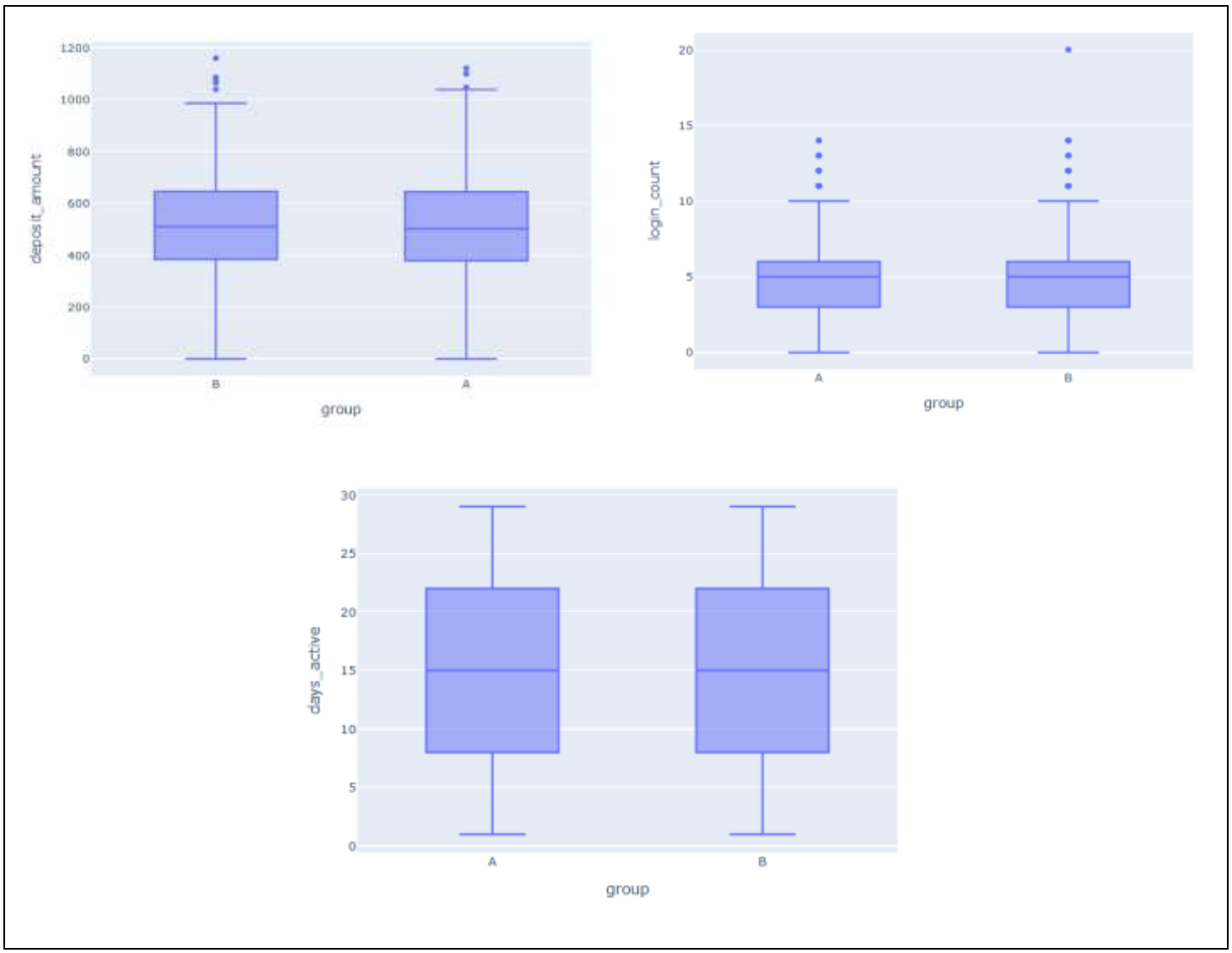

Figure 3: Step 3: Conversion rate by group.
Source: Authors, (2026).

At this stage (see Table 4), we investigate whether the treatment's impact varies by users' age segment, thereby identifying potential heterogeneous effects. To this end, the age variable was partitioned into three brackets: under 30, 30–50, and over 50. We then computed the conversion rate for each age group, disaggregated by experimental arm (A or B). This segmentation makes it possible to detect whether the treatment was more effective for certain demographic profiles. The results show that group B consistently outperformed group A across all three age segments. Among users under 30, the conversion rate was 9.4% in group A and 16.1% in group B, an absolute difference of 6.8 percentage points. In the 30–50 segment, the rates were 10.0% for A and 14.5% for B, a 4.5-point difference. Among users over 50, group B reached a conversion rate of 16.2%, compared with 8.9% for group A the largest observed difference (7.3 percentage points).

A grouped bar chart reinforces this reading. In every segment, the bars corresponding to group B are notably higher than those for group A, indicating that the treatment's positive effect holds across age ranges. However, the impact is slightly larger at the extremes (<30 and >50), suggesting that these groups may be more sensitive to the change introduced by the experiment. This segmented analysis is valuable because it reveals that, although the treatment is beneficial overall, its magnitude varies across user subgroups. Such information can be used to design more personalized campaigns or to allocate the treatment more efficiently in future deployments. It also rules out the possibility that the observed effect on overall conversion is driven by a single segment, which adds robustness to the conclusions.

Table 4: Step 4: Treatment heterogeneity analysis (segmented).

Segment	Conversion A	Conversion B	Difference
< 30	0.094	0.161	0.068
30–50	0.100	0.145	0.045
> 50	0.089	0.162	0.073

Source: Authors, (2026).

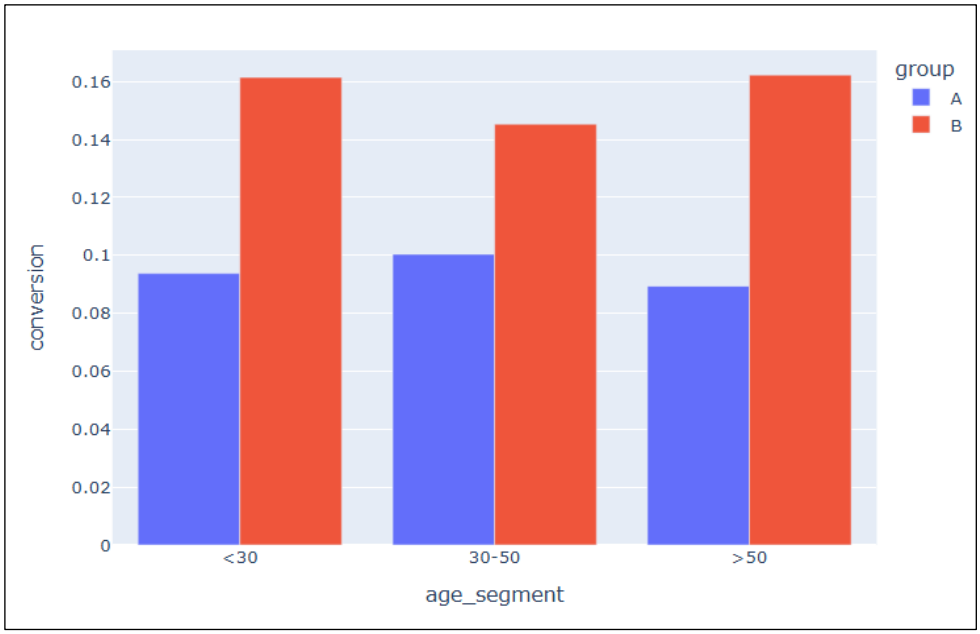

Figure 4: Step 4: Distribution in bar chart Treatment heterogeneity analysis (segmented).
Source: Authors, (2026).

In this fifth step, we apply a resampling technique bootstrap to robustly estimate the average treatment effect (ATE) on conversion together with a confidence interval. Unlike traditional parametric tests, bootstrap does not assume normality or homoscedasticity; instead, it draws many random samples with replacement from the original data and computes the statistic of interest at each iteration. Here, 5,000 iterations were performed, computing the difference in conversion rates between group B and group A for each resampled dataset. This procedure yields an empirical distribution of the ATE, from which a 95% confidence interval can be obtained by extracting the 2.5th and 97.5th percentiles. The resulting mean ATE was 0.0524 that is, an absolute difference of 5.24 percentage points in favor of group B. The confidence interval ranged from 0.0391 to 0.0652, indicating that, with 95% confidence, the true treatment effect lies within this interval. The accompanying plot depicts the distribution of conversion-rate differences across all simulations. The curve is approximately normal, centered around the mean value (0.0524), with red dashed lines marking the confidence-interval bounds. This visualization shows that virtually all simulations yielded a positive effect, reinforcing confidence in the robustness of the result (see Table 5).

Table 5: Step 5: Bootstrap-based confidence intervals for the treatment effect.

Metric	**Estimate**	**95% CI (Lower – Upper)**
ATE (Bootstrap)	0.0524	0.0391 – 0.0652

Source: Authors, (2026).

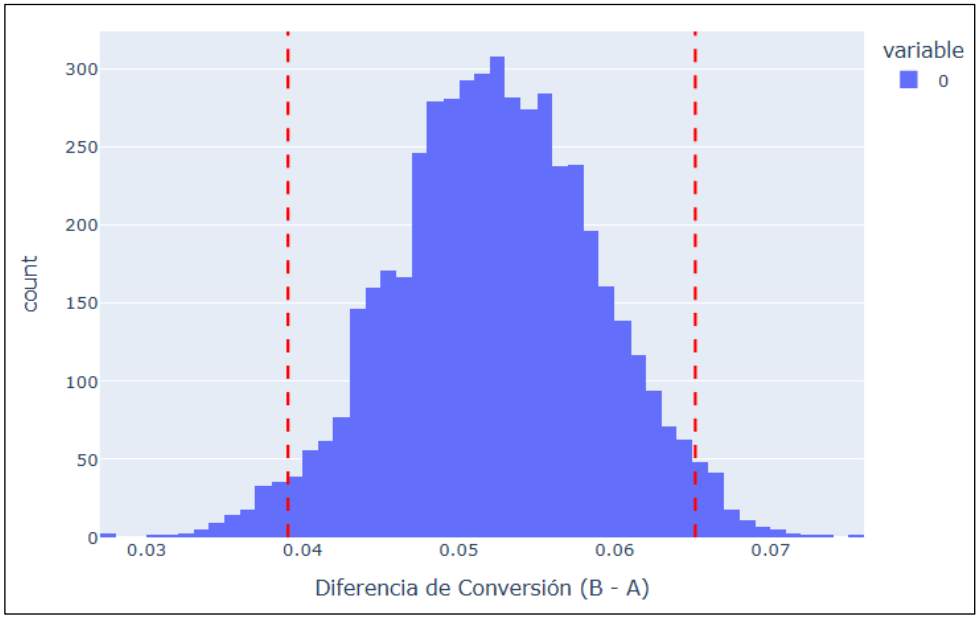

Figure 5: Step 5: Histogram of Bootstrap-based confidence intervals for the treatment effect.
Source: Authors, (2026).

In this step (see Table 6), an ordinary least squares (OLS) regression model is used to estimate the causal effect of the treatment on conversion while controlling for additional variables. Although conversion is a binary variable (0 or 1), OLS can be employed as an initial approximation to the average treatment effect, and it also allows inspection of the influence of other explanatory variables. In this model, the dependent variable is conversion, and the predictors are the assigned group (group), age, income, number of logins, and days active. The most relevant coefficient is that associated with group B (C(group)[T.B]), which has an estimated value of 0.0523 with a standard error of 0.0066. This implies that, holding the other variables constant, belonging to group B increases the probability of conversion by approximately 5.23 percentage points—a result that is highly consistent with the findings in the previous steps.

In addition, the p-value associated with this coefficient is less than 0.0001, confirming that the effect is statistically significant. This p-value measures the probability that so large a difference would occur by chance if group B truly had no effect. As for the remaining variables, none is statistically significant: age has a negative coefficient but a p-value of 0.207; income has a p-value of 0.337; the number of logins is irrelevant (p = 0.383); and days active likewise shows no effect (p = 0.278). This indicates that, within this model, the only factor clearly associated with conversion is assignment to the treatment group. The model's R-squared is very low (0.007), which is common in models with binary outcomes and reflects that most individual variability is not explained by the included covariates. However, the objective here is not to predict conversion but to estimate the causal effect of the treatment, which is clearly obtained.

Table 6: Step 6: Estimating the causal effect with regression (OLS).

Field			**Value**			
Dep. Variable			conversion			
Model			OLS			
Method			Least Squares			
Date			Sat, 19 Jul 2025			
Time			00:35:36			
No. Observations			10000			
Df Residuals			9994			
Df Model			5			
Covariance Type			nonrobust			
R-squared			0.007			
Adj. R-squared			0.006			
F-statistic			13.57			
Prob (F-statistic)			3.21e-13			
Log-Likelihood			-3066.6			
AIC			6145.0			
BIC			6188.0			
Variable	**coef**	**std err**	**t**	**P>|t|**	**0.025**	**0.975**
Intercept	0.1040	0.019	5.541	0.000	0.067	0.141
C(group)[T.B]	0.0523	0.007	7.942	0.000	0.039	0.065
age	-0.0004	0.000	-1.262	0.207	-0.001	0.000
income	3.247e-07	3.38e-07	0.959	0.337	-3.39e-07	9.88e-07
login_count	-0.0013	0.001	-0.873	0.383	-0.004	0.002
days_active	0.0004	0.000	1.086	0.278	-0.000	0.001
Statistic			**Value**			
Omnibus			4113.569			
Prob(Omnibus)			0.000			
Skew			2.254			
Kurtosis			6.143			
Durbin–Watson			1.992			
Jarque–Bera (JB)			12586.243			
Prob(JB)			0.000			
Cond. No.			1.81e+05			

Source: Authors, (2026).

In this seventh step (see Table 7), the goal is to validate the robustness of the experiment, ensuring that the results obtained are not artifacts of methodological errors, hidden biases, or spurious effects. To this end, three complementary strategies are applied: a placebo test, a review of experimental integrity, and a sensitivity analysis with simplified models. The placebo test consists of applying the same type of statistical comparison used to measure the treatment effect, but to a variable that should not be influenced by the treatment. In this case, income was chosen as the placebo variable, since there is no reason to expect that belonging to group A or B would alter users' income. A t test was conducted to compare income across groups, yielding a t value of −0.0714 and a p value of 0.9431. Such a high p-value indicates that there is no significant difference between groups, which validates that the experimental design is not generating systematic false positives. Next, a quality review of the data and assignment was carried out. No missing values were found in any key variable of the experiment, indicating that there was no dropout or data loss in the process.

The only exception was the age_segment column used solely for the segmented analysis which contains 517 null values due to users whose ages fell outside the defined range. Furthermore, the distribution of users across groups was very close to 50%, with 50.76% in group A and 49.24% in group B, confirming that random assignment was executed correctly.

Table 7: Step 7: Robustness checks (placebo, integrity review, and sensitivity analysis).

Variable	**Test Statistic**	**p-value**
Income	t = -0.0714	0.9431
Column	**Missing Values**	
user_id	0	
group	0	
age	0	
income	0	
conversion	0	
deposit_amount	0	
login_count	0	
days_active	0	
age_segment	517	
Group	**Proportion**	
A	0.5076	
B	0.4924	

Source: Authors, (2026).

Subsequently, a sensitivity verification (see Table 8) was conducted using a reduced regression model that included only two variables: group and age. The objective is to assess whether the coefficient for group B remains stable when other covariates are removed. The results show that the coefficient for group B remains exactly 0.0523, with a p-value less than 0.0001, indicating that the treatment effect does not depend on the inclusion of additional variables in the model. This stability reinforces confidence that the observed effect is robust and not conditioned by model specification.

Table 8: Step 7 (part 2) – Sensitivity check with partial regression.

Field	**Value**
Dep. Variable	conversion
Model	OLS
Method	Least Squares
Date	Sat, 19 Jul 2025
Time	00:39:49
No. Observations	10000
Df Residuals	9997
Df Model	2
Covariance Type	nonrobust
R-squared	0.006
Adj. R-squared	0.006
F-statistic	32.49
Prob (F-statistic)	8.64e-15
Log-Likelihood	-3068.0
AIC	6142.0
BIC	6164.0

Variable	**coef**	**std err**	**t**	**P>|t|**	**0.025**	**0.975**
Intercept	0.1135	0.013	8.806	0.000	0.088	0.139
C(group)[T.B]	0.0523	0.007	7.948	0.000	0.039	0.065
age	-0.0004	0.000	-1.248	0.212	-0.001	0.000

Statistic	**Value**
Omnibus	4115.683
Prob(Omnibus)	0.000
Skew	2.255
Kurtosis	6.145
Durbin–Watson	1.992
Jarque–Bera (JB)	12599.341
Prob(JB)	0.000
Cond. No.	143.0

Source: Authors, (2026).

V. CONCLUSIONS

The experimental program provides convergent evidence that the treatment materially improves user conversion while preserving internal validity. Random assignment yielded well-balanced groups across observable covariates (age, income, login_count, days_active), and classical tests failed to detect any pre-treatment imbalances. In the primary outcome, the treatment produced a large and statistically precise lift in conversion rising from 9.85% in the control arm to 15.09% in the treated arm corroborated by a z-test with $p<0.0001$. Bootstrap estimation further quantified this effect with a mean average treatment effect (ATE) of 5.24 percentage points and a narrow 95% confidence interval [3.91%, 6.52%], indicating that the uplift is not an artifact of parametric assumptions and is robust to resampling uncertainty. Causal estimation with a linear probability model (OLS) that controlled for user demographics and baseline activity yielded a treatment coefficient of 0.0523 (SE=0.0066; $p<0.0001$), essentially identical to the nonparametric bootstrap estimate. Taken together, the randomized comparison, the resampling analysis, and the regression adjustment converge on the same quantitative conclusion: the intervention consistently increases the probability of conversion by about five percentage points.

Robustness checks placebo testing on income, integrity audits of missingness and assignment ratios, and a reduced-form sensitivity model found no evidence of hidden bias, data leakage, or specification dependence, reinforcing the credibility of the causal claim. At the same time, secondary outcomes delineate the boundary of the treatment's impact. Among converters, the intervention did not increase deposited amounts nor post-conversion engagement (logins and active days); distributions and inferential tests show no statistically meaningful differences between arms on these metrics. This pattern implies that the treatment primarily improves top-of-funnel behavior (initial conversion) without yet translating into deeper behavioral or financial value. In managerial terms, the intervention appears to drive "shallow" conversions rather than high-value activations, a distinction that is crucial for revenue forecasting, lifetime-value modeling, and budgeting of acquisition spend. Heterogeneity analyses by age segment show that the uplift is broad-based but not uniform: effects are directionally larger in users under 30 and over 50, with mid-age users showing a smaller yet still positive gain.

This profile suggests immediately actionable targeting policies. In the short run, firms could prioritize deployment in the more responsive segments to maximize near-term ROI while designing complementary nudges aimed at upgrading mid-segment users from initial conversion to meaningful activation. In the medium term, these findings motivate segment-specific experimentation (e.g., multivariate tests tuned to frictions and value propositions salient for each cohort). These results carry two strategic implications. First, acquisition and activation should be treated as distinct optimization problems: the tested treatment is a strong lever for conversion, but it should be paired with post-onboarding interventions (e.g., personalized prompts, financial education micro-journeys, product bundling, or rate-structure trials) to unlock deposit growth and durable engagement. Second, evaluation should migrate from single-metric hypothesis tests to multi-objective experimental design that jointly tracks conversion, early-value proxies (first-week deposits, feature adoption), and forward-looking business KPIs (predicted CLV, retention hazard).

Doing so will align experimental success criteria with financial materiality. Methodologically, the study's triangulation across classical tests, bootstrap inference, and covariate-adjusted models is a strength, as is the suite of robustness diagnostics. Future extensions could corroborate the estimates with generalized linear models (logit/probit), apply Bayesian A/B frameworks for continuous monitoring and decision thresholds, and incorporate adaptive allocation (multi-armed bandits) to improve sample efficiency while respecting guardrails for fairness and regulatory compliance. Because the treatment does not yet shift value-creation metrics, uplift modeling at the user level (treatment effect heterogeneity on deposits/engagement) and experiment-within-experiment designs (sequential interventions after conversion) are natural next steps. Finally, given the financial context, governance remains central. Any scale-up should include pre-registered analysis plans, segment-level fairness audits, and continuous monitoring for interference and spillovers (especially where social or referral effects may arise). By coupling disciplined experimentation with ethical safeguards and value-focused KPIs, organizations can translate the demonstrated conversion uplift into sustained, equitable, and auditable business impact.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.
Methodology: Alexander Haro-Sarango, Julian Coronel-Reyes.
Investigation: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez.
Discussion of results: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.
Writing – Original Draft: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.
Writing – Review and Editing: Alexander Haro-Sarango, Julian Coronel-Reyes.
Resources: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos.
Supervision: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.
Approval of the final text: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.

VII. REFERENCES

[1] Gai K, Qiu M, Sun X. A survey on FinTech. J Netw Comput Appl. 2018;103: 262–273. doi:10.1016/j.jnca.2017.10.011

[2] Gomber P, Kauffman RJ, Parker C, Weber BW. On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. J Manag Inf Syst. 2018;35: 220–265. doi:10.1080/07421222.2018.1440766

[3] Kohavi R, Tang D, Xu Y. Trustworthy online controlled experiments: A practical guide to a/b testing. 2020. Available: https://books.google.com.ec/books?hl=es&lr=&id=Gu-

CEAAAQBAJ&oi=fnd&pg=PP1&dq=Trustworthy+Online+Controlled+Experiments:+A+Practical+Guide+to+A/B+Testing+%5BInternet%5D.+Cambridge&ots=w57WGTgP5l&sig=OsWBsb-RluhO6B8SMgfL1cj3nsk.

[4] Zou T, Xiong F, Li S, Zhang W. Understanding the determinants of firms' usage of A/B testing: A technology–organization– environment framework. IEEE Trans Eng Manage. 2025;72: 378–400. doi:10.1109/tem.2024.3415500

[5] Malovaná S, Janků J, Hodula M, Green G, Joy M, Von Rüden L, et al. Macroprudential policy and income inequality: The trade-off between crisis prevention and credit redistribution. ijcb.org; 2025 [cited 22 Oct 2025]. Available: https://www.ijcb.org/journal/ijcb25q4a5.pdf

[6] Pearl J. Lord's paradox revisited – (oh Lord! Kumbaya!). J Causal Inference. 2016;4. doi:10.1515/jci-2016-0021

[7] Årnes A, Fjeld MK, Stigum H, Nielsen C, Stubhaug A, Johansen A, et al. Does pain tolerance mediate the effect of physical activity on chronic pain in the general population? The Tromsø Study. Pain. 2024. doi:10.1097/j.pain.0000000000003209

[8] Nudurupati SS, Tebboune S, Garengo P, Daley R, Hardman J. Performance measurement in data intensive organisations: resources and capabilities for decision-making process. Prod Plan Control. 2024;35: 373–393. doi:10.1080/09537287.2022.2084468

[9] Kuai L, Wei H. The role of serendipity in narratives: How serendipitous story promotes product interest. Psychol Mark. 2025;42: 1346–1360. doi:10.1002/mar.22181

[10] Chen X, Lin A, Webber S. "We do not always enjoy surprises": investigating artificial serendipity in an online marketplace context. J Doc. 2025;81: 403–422. doi:10.1108/jd-01-2024-0011

[11] Hashi I. Machine Learning–Driven Fintech Solutions for Credit Scoring and Financial Inclusion in the Gig Economy. International Journal of Innovative Science and Research Technology. 2025;10: 1805–1820. Available: https://www.academia.edu/download/124446866/IJISRT25AUG1023.pdf

[12] Óskarsdóttir M, Bravo C, Sarraute C, Vanthienen J, Baesens B. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Appl Soft Comput. 2019;74: 26–39. doi:10.1016/j.asoc.2018.10.004

[13] Alamsyah A, Hafidh AA, Mulya AD. Innovative credit risk assessment: Leveraging social media data for inclusive credit scoring in Indonesia's fintech sector. J Risk Fin Manag. 2025;18: 74. doi:10.3390/jrfm18020074

[14] Ugander J, Karrer B, Backstrom L, Kleinberg J. Graph cluster randomization: network exposure to multiple universes. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM; 2013. doi:10.1145/2487575.2487695

[15] Holtz D, Lobel F, Lobel R, Liskovich I, Aral S. Reducing interference bias in online marketplace experiments using cluster randomization: Evidence from a pricing meta-experiment on Airbnb. Manage Sci. 2024. doi:10.1287/mnsc.2020.01157

[16] Camlin CS, Sheira LA, Kwena ZA, Charlebois ED, Agot K, Moody J, et al. The effect of a social network-based intervention to promote HIV testing and linkage to HIV services among fishermen in Kenya: a cluster-randomised trial. Lancet Glob Health. 2025;13: e669–e678. doi:10.1016/S2214-109X(24)00539-4

[17] Phénix T, Ginestet É, Valdois S, Diard J. Visual attention matters during word recognition: A Bayesian modeling approach. Psychon Bull Rev. 2025;32: 1165–1203. doi:10.3758/s13423-024-02591-4

[18] Taddy M. Business data science. 2019. Available: http://222.254.35.8/handle/TLU/8894

[19] Lin C, Ouyang Z, Wang X, Li H, Huang Z. Preserve integrity in realtime event summarization. ACM Trans Knowl Discov Data. 2021;15: 1–29. doi:10.1145/3442344

[20] Niu W, Huang L, Chen M. Spanning from diagnosticity to serendipity: An empirical investigation of consumer responses to product presentation. Int J Inf Manage. 2021;60: 102362. doi:10.1016/j.ijinfomgt.2021.102362

[21] Bao Z, Zhu Y. Understanding online reviews adoption in social network communities: an extension of the information adoption model. Inf Technol People. 2025;38: 48–69. doi:10.1108/itp-03-2022-0158

[22] Bahiru TK, Sinshaw NT, Moges TH, Singh DK. Auditing and mitigating bias in gender classification algorithms: A data-centric approach. arXiv [cs.CV]. 2025. doi:10.48550/arXiv.2510.17873

[23] Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N. "it"s reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2018. doi:10.1145/3173574.3173951