The S14 Version Update and Its Effects on User Engagement, Monetization, and Retention: An Empirical Analysis

Group Members:

- Jian Lin (xxlin@ucdavis.edu)
- Zinnia Zeng (jyazeng@ucdavis.edu)
- Feini Pek(fpek@ucdavis.edu)

Abstract

In the rapidly evolving mobile application ecosystem, continuous product iteration is essential for sustaining user interest and maximizing monetization. Our project investigates the impact of the "S14" version update (v1.2.9 to v1.3.3) on a global weather forecasting application. Using a comprehensive dataset of daily user-level activities across four key markets (United States, India, Germany, and Iraq), we employ a multi-stage statistical framework combining bootstrap resampling, principal components analysis of behavioral metrics, regularized regression (OLS/Ridge/Lasso), classification trees, gradient boosting machines, and logistic regression to study both monetization and short-term user retention.

Our analysis reveals a significant phenomenon of market divergence. While aggregate global metrics suggest a successful increase in Average Revenue Per User driven by higher ad frequency, this mask a critical failure in the US market. The S14 update, characterized by a short-duration, high-frequency usage pattern, aligns well with the "snacking" behavior of Indian users and boosts revenue there. In contrast, in the US the update disrupts the "Home-Page-dependent" browsing flow, significantly reducing session depth. Lasso regression identifies a negative intrinsic coefficient for the version variable in the US, and decision-tree and GBM models highlight a "dead zone" in which users who fail to reach sufficient page depth generate near-zero revenue. Retention models show that 7-day retention is driven mainly by engagement level, timing (month), and market; after controlling for these factors, S14's aggressive monetization does not produce a large, systematic drop in short-term retention, though it increases risk in depth-oriented markets. We conclude that a one-size-fits-all update strategy is detrimental and recommend a bifurcated product roadmap: retaining S14 features for high-frequency markets while rolling back or redesigning the experience for engagement-focused markets such as the US.

1. Background

1.1. Background

The monetization of free-to-use mobile utility applications typically relies on a delicate balance between user experience (UX) and ad inventory pressure. Product managers often face a trade-off: increasing ad density may boost short-term revenue, but risks harming long-term retention and engagement depth. This report analyzes a specific major update, internally

codenamed "S14" (Version 1.3.3). Compared to the previous stable version (v1.2.9), S14 introduced significant changes to the user interface (UI), specifically adding a toolbar on the radar page and altering the logic for rewarded video ads and in-app purchase prompts.

1.2. Problem Definition

The primary objective of this study is to evaluate the net impact of the S14 update. While the topline revenue numbers appeared positive in initial A/B testing, deeper granularity is required to understand the sustainability of this growth. We structured our analysis to answer three specific questions:

- **Q1: Comparison of Versions (Behavioral Impact):** Did the update fundamentally alter how users interact with the app?
- **Q2: Revenue Drivers (Attribution):** Which specific user behaviors (e.g., visiting the Home Page vs. Radar Page) are the true drivers of revenue? Did the update align with or disrupt these drivers?
- **Q3: Market Heterogeneity:** Why did update perform differently across developed markets (e.g., US) and emerging markets (e.g., India)?
- **Q4: Long-term Sustainability (User Retention):** Does the S14 update's aggressive monetization strategy (e.g., higher ad density) negatively impact user retention?

1.3. Data Description

Data used in this project comes from Falcon Media, a technology company that develops utility-based mobile applications such as weather forecast tools and voice changer apps, serving users across multiple countries on Google Play. The company's business model relies on in-app advertising (IAA) and user engagement optimization. The internal dataset includes user activity logs, ad impression data, and retention metrics collected through a data analytics platform called DataTower. These data provide an opportunity to analyze user engagement patterns, advertising performance, and factors influencing user retention.

- User-Level Data
  - App usage metrics: Number of app opens per day, Total session duration, Total number of pages visited, Number of newly accessed pages.
  - Interaction metrics: Total number of clicks, Number of unlock-related clicks.
  - Advertising metrics: Ad impressions, Ad clicks, Advertising revenue.
  - Monetization metrics: Purchase events, Purchase amount.
- Derived Features
  - Average duration per session (avg_duration_per_session): Measures how long users stay in the app during a single session.
  - Average pages per session (avg_page_per_session): Captures browsing depth within each app visit.

- Ad click-through rate (CTR) (ad_click_through_rate):
  Defined as ad clicks divided by ad impressions, reflecting users' willingness to interact with ads.
- Ad density (ad_density): Defined as ad impressions per unit of usage time, measuring the intensity of ad exposure.

- Summary Statistics
  - This dataset contains 6782 observations and 39 variables, where each row is one day observation of a user.
  - This includes4 countries (Germany, India, Iraq, USA), and covers 2 different update periods (version 1.2.9 and 1.3.3).
  - Timeline covered: The data covers 06/28/2025 to 02/15/2025.

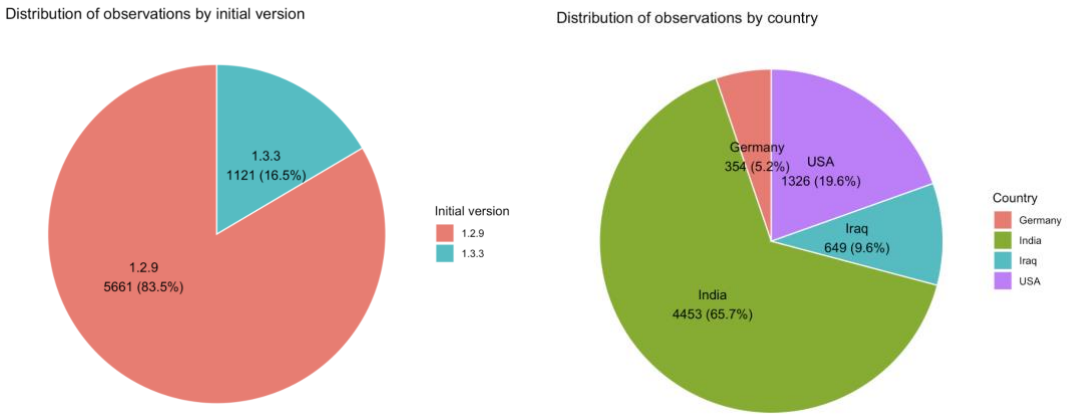

Figure 1: Distribution of observations

Based on the pie chart above, we observe that most users started on version 1.2.9 (5,661 observations, 83.5%), with a smaller share on the new version 1.3.3 (1,121, 16.5%). By country, the data are heavily concentrated in India (4,453 observations, 65.7%), followed by the USA (1,326, 19.6%), Iraq (649, 9.6%), and Germany (354, 5.2%)

2. Proposed Method

Because we have several related questions, we organize the methods section by question: for each question we describe the data preparation, models used, and key assumptions.

2.1 Q1: Comparison of Versions (Behavioral Impact)

2.1.1. Linear Regression

We first examined whether users on version 1.3.3 behaved differently from users on 1.2.9, after accounting for country and calendar effects. We focused on the following behavior metrics: number of sessions, average session duration, average pages per session, number of ad impressions, total clicks, ad density, and rewarded-ad ratio.

For each behavior metric, we fit a separate linear regression model with the metric as the outcome and version, country, and weekday as predictors. The coefficient on the 1.3.3 indicator

was interpreted as the adjusted difference in mean behavior between 1.3.3 and 1.2.9 users, holding country and calendar effects fixed.

For a subset of aggregate outcomes (e.g., total pages, total clicks, ad revenue), we also used a non-parametric bootstrap to do the A/B test.

2.1.2. Bootstrap

Given the non-normal, long-tail distribution typical of user session duration and ad revenue, standard parametric tests (like T-tests) may be unreliable. We utilize the Bootstrap method for robust inference.

- Algorithm: We generate B=3000 resampled datasets by sampling with replacement from the original user pool. For each iteration, we calculate the difference in means between the control group (Version 1.2.9) and the test group (Version 1.3.3).
- Objective: To construct 95% Confidence Intervals (CI) for key metrics. If the CI for the difference excludes zero, we reject the null hypothesis and confirm the update's impact is statistically significant.

2.1.3. Principal Component Analysis on Behavior

To summarize the multi-dimensional behavior patterns, we applied principal component analysis (PCA) to the standardized behavior variables. PCA produces orthogonal linear combinations (principal components) that capture most of the variability in behavior. The first component (PC1) loaded strongly on sessions, pages, impressions, clicks, and ad density and was interpreted as an overall “engagement and ad exposure” score. We attached the first two components (PC1 and PC2) to the user-level data and compared the distribution and mean of PC1 between versions using Welch two-sample t-tests and boxplots.

Main assumptions in this part were that users are independent, that the conditional mean of each behavior metric is approximately linear in version, country, and calendar variables, and that PCA is a reasonable way to summarize the covariance structure of the behavior variables. The bootstrap additionally assumes that resampling users approximates the sampling distribution of mean differences.

2.2. Q2: Revenue Drivers

2.2.1. Global Analysis

We first address the distributional challenges of ad revenue data. Since revenue is zero-inflated and highly skewed, we define a transformed outcome Y = log (1 + ad_revenue) for regression tasks. Additionally, we perform Principal Component Analysis (PCA) on standardized behavioral metrics to extract two latent features, PC1_engagement and PC2_pattern, which serve as condensed predictors to mitigate multicollinearity.

- Regularized Regression (Ridge/Lasso): We model the continuous log-revenue using glmnet. Predictors include initial version, country, weekday indicator, month, and PCA

components. We use cross-validation to select the penalty parameter and compare performance (R^2) against an unpenalized OLS baseline.

- High-Value Classification (GBM/Trees): We reframe monetization as a classification problem, defining "high-value" users as the top 25% by revenue. Using a 70/30 train-test split, we fit:
  - Classification Trees to obtain an interpretable segmentation.
  - Gradient Boosting Machines (GBM): Using Bernoulli loss and 5-fold CV to optimize tree count, evaluated via AUC and Relative Influence Scores.

  For the classification models, we assume that users are independent and that the relationship between predictors and high-value status is stable between the training and test sets, without imposing a specific parametric form.

2.2.2. Market-Specific Attribution & Threshold Optimization

While global analysis predicts overall revenue, it may mask heterogeneous drivers across markets. To address this, we conduct a deep-dive analysis:

- Comparative Attribution (Lasso Regression): We fit separate Lasso models for distinct geographic segments (specifically *US vs. India*). By isolating these subgroups, we identify market-specific revenue drivers (e.g., page_home vs. session_open) and quantify the specific impact of the S14 update (initial_version_1.3.3) within each context.
- Non-Linear Threshold Detection (Decision Tree Regressor)**:** To translate attribution insights into operational KPIs, we employ a shallow Decision Tree (max_depth=3) specifically for the US market. This model identifies critical non-linear tipping points (e.g., specific thresholds for page_home views) that trigger significant jumps in ARPU.

2.3. Q4: User Retention

For retention, we defined a binary outcome indicating whether a user was still active at day 7 (retained_D7 = 1 if retention_day ≥ 7, 0 otherwise). We used the same set of predictors as in the monetization models: version, country, month, weekday, PC, and PC2. Given this, we will perform the following methods:

- Logistic Regression

  This model will relate the log-odds of D7 retention to a linear combination of predictors. We estimated odds ratios and 95% confidence intervals for key predictors and evaluated predictive performance using test-set accuracy and AUC. This model assumes that, conditional on the predictors, users are independent and that the log-odds of retention are approximately linear in version, engagement, and contextual variables.
- High-Value Classification (GBM/Trees)

  To allow for non-linear effects and interactions, we also fit a classification tree and a Bernoulli GBM for retained_D7, again using a 70/30 train-test split. The tree provides a simple segmentation of retention risk (e.g., rules involving month, version, and

engagement), while the GBM provides a more flexible prediction rule and variable-importance ranking. Assumptions for these models are similar to those in the monetization classification: i.i.d. observations and a stable mapping from predictors to retention between the training and test samples.

3. Data analysis

3.1. Comparison of Versions (Behavioral Impact)

We began by evaluating the global impact of the S14 update using the Bootstrap method. The results depict a fundamental shift in the app's usage paradigm. We focused on four representative countries:United States and Germany, representing mature Western markets, alongside India and Iraq, representing emerging markets. Prior to the main analysis, the data underwent rigorous cleaning.

3.1.1. Bootstrap

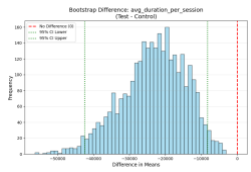

Figure 2 - US

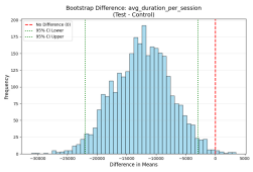

Figure 2 - German

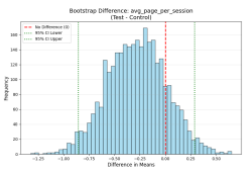

Figure 2 - Iraq

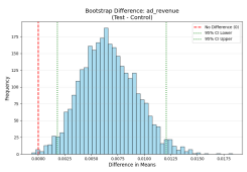

Figure 2 - India

The stratified Bootstrap analysis, conducted with 3,000 iterations per market, revealed a sharp divergence in performance, effectively bifurcating the markets into "Engagement-Sensitive" and "Frequency-Driven" clusters.

In the mature markets of the United States and Germany, the update proved detrimental to user experience without delivering the intended revenue lift. Specifically, the US market (Engagement-Sensitive) suffered a "lose-lose" outcome. The update significantly degraded the user experience: session duration plummeted by -23,310 ms and page depth declined by 0.78 pages, while this reduction in inventory time did not translate into increased monetization efficiency. Germany mirrored this "lose-lose" pattern, showing significant declines in both duration and page views with no corresponding gain in revenue. These results suggest that the S14 update is likely leading to reduced engagement that offset any potential gains from increased ad impressions.

In stark contrast, India emerged as the primary success story, validating a "Short & Frequent" usage model. It was the only market to demonstrate a statistically significant increase in ad revenue, with a positive average lift of + $0.0066. This revenue growth was driven by a surge in usage frequency; the session open rate increased by 0.47 times. While this confirms that Indian users are more tolerant of the new ad-heavy design, it is worth noting a trade-off: the ratio of rewarded ads declined significantly, suggesting that the

increased density of forced ads may be cannibalizing voluntary ad interactions. Also, **Iraq** presented a unique case of "Ad Tolerance". However, Iraq's tolerance did not translate into profit, with ad revenue showing no statistically significant change.

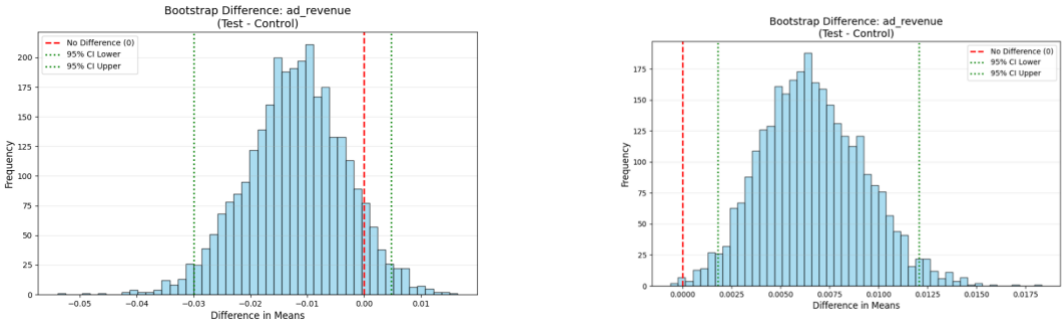

Figure 2: US (left) vs India on Revenue metrics

In summary, the empirical results confirm that version 1.3.3 is not a universally effective update. It succeeded in the frequency-driven Indian market but failed in the engagement-centric US and German markets, necessitating a deeper investigation (will be shown in 3.3) into the specific behavioral drivers of revenue in the subsequent attribution analysis

3.1.2. Regression and PCA

To quantify how behavior changed with the new version on a global level, we first fit separate linear regressions of each behavioral metric on the initial version indicator, adjusting for country and calendar effects (month / weekday). The coefficient for *initial_version1.3.3* in these models represents the mean difference between 1.3.3 and 1.2.9 users, holding country and time fixed; full estimates are reported in **Table A1**. Relative to 1.2.9, users on 1.3.3 opened about 0.46 more sessions on average ($p < 0.001$) and saw roughly 2.3 additional ad impressions per user-day ($p < 0.001$). At the same time, their sessions were shorter and lighter: average pages per session were lower by about 0.6 pages ($p < 0.001$), and average session duration was also reduced ($p \approx 0.01$). Total click volume and the rewarded-vs-non-rewarded mix did not differ significantly across versions. Overall, after accounting for country and calendar effects, version 1.3.3 encouraged more, but briefer, visits with higher ad exposure per user.

To summarize the joint behavioral pattern, we then performed a PCA on standardized behavior metrics. The first two components captured most of the variability in usage: PC1 explained about 40% of the variance and PC2 about 23% (cumulative 63%; **Table A2.1**). Loadings in **Table A2.2** show that PC1 places moderately large, negative weights on session_open, avg_page_per_session, ad_impression, total_click, and ad_density, so more sessions, more pages, and more ad activity all correspond to more negative PC1 scores. We therefore interpret PC1 as an overall *engagement & ad exposure* component (with the sign flipped: smaller values = more engaged). PC2 contrasts "many short, light sessions" (high session_open, low pages and duration) with "fewer, longer, page-heavy sessions with denser

ads", and is used mainly for descriptive interpretation.

Comparing PC1 across versions, 1.3.3 users had substantially lower (more engaged) scores than 1.2.9 users (mean PC1 ≈ −0.23 vs 0.05), and this difference was statistically significant ($t \approx 4.6$, $p < 0.001$). Together with the univariate regressions, these results suggest that the 1.3.3 update shifted behavior toward a mild higher-engagement, ad-heavier regime: users come back more often, interact with slightly fewer pages per visit, but are exposed to more ads overall. This summary of engagement patterns sets up our second question, where we investigate how these behaviors, and the new version itself, translate into monetization outcomes.

3.2. Revenue Drivers

3.2.1. Linear Models

To understand the revenue drivers on global level, we used OLS, ridge, and lasso regression to predict log-transformed ad revenue from version, country, month, weekday, and the two engagement components (PC1 and PC2). All three models achieved very similar performance (**Table B3**). Training RMSE was about 0.054 and test RMSE about 0.056 for every model, corresponding to roughly a 5–6% multiplicative error on the log scale. The models explained about 24–25% of the variance in log ad revenue on both the training and test sets (R^2: 0.24–0.25), and there was no evidence of strong overfitting: train and test metrics were almost identical.

The ridge model (**Table B1**) kept all predictors but shrank them strongly toward zero. The first engagement component (PC1) remained a positive predictor of revenue (roughly a 1% increase in revenue for a one-standard-deviation increase in engagement), while PC2 had a small negative effect. Country and month showed clear differences (e.g., higher revenue in June and in the US), capturing geographic and seasonal effects. Importantly, the coefficient for version 1.3.3 was essentially zero in the ridge model, and the lasso model (**Table B2)** shrank it exactly to zero, indicating that once engagement, country, and seasonality are accounted for, the version indicator itself does not add predictive power for ad revenue.

Although the models are stable and interpretable, their predictive power is modest: about three-quarters of the variation in user-level revenue remains unexplained. This is not surprising for monetization data. First, ad revenue is inherently noisy—auction dynamics, which specific ad is served; random user behavior, and other unobserved factors can cause large fluctuations that cannot be predicted from simple engagement summaries. Second, our feature set is limited: we do not include device type, time of day, ad placement, prior history, or user demographics, all of which could explain additional variance. Third, for interpretability we compressed seven behavior variables into only two principal components and used mostly linear effects. This helps with collinearity but inevitably discards some fine-grained patterns and non-linear relationships.

Given these limitations, it is reasonable that linear ridge and lasso regression provide only moderate improvements over OLS and similar overall performance. To better understand

monetization patterns, rather than chasing small gains in R^2 on continuous revenue, we next shift focus to a more business-oriented question: which users are "high value"? In the following section, we define a high-value segment based on the top quartile of ad revenue and apply classification trees and gradient boosting. These models allow non-linear effects and interactions and provide variable-importance summaries that are easier to translate into product and monetization strategies.

3.2.2. Tree Methods

A. Classification Tree

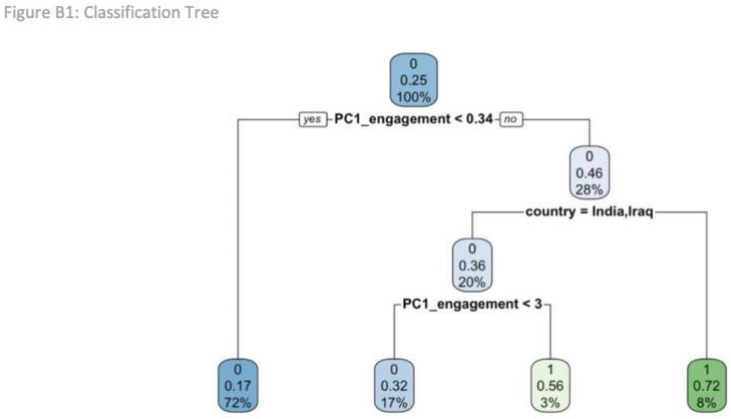

Figure B1: Classification Tree

To complement the linear models, we fit a classification tree to predict whether a user belongs to the top 25% of the ad-revenue distribution ("high-value"), using version, country, calendar variables, and the two PCA engagement components as predictors (**Figure B1**). The fitted tree is very shallow and easy to interpret. The root split is on the overall engagement score PC1_engagement: users with low–moderate engagement (PC1_engagement < 0.34) is almost always classified as low-value. Among the more engaged users (PC1_engagement ≥ 0.34), the tree next splits on country. Users from India or Iraq with high engagement form the main high-value leaf: in this subgroup, about 72% of users are high-value, compared to 17–32% in the other leaves. Notably, initial_version never appears as a split variable, suggesting that, once we account for engagement level and country, the version label itself does not provide additional discriminatory power for high vs. low revenue.

On the held-out test set, the classification tree achieves an overall accuracy of about 78%. It is quite conservative: specificity is high (≈95% of truly low-value users are correctly classified as low-value), but sensitivity is low (it only identifies about 27% of truly high-value users). When the model does predict "high-value," the prediction is reasonably reliable—roughly 65% of those predicted high-value are actually in the top-revenue group, while users predicted "low-value" are indeed low-value about 80% of the time. Overall, the tree highlights a clear pattern: high engagement combined with specific countries (India, Iraq) is the primary

route into the high-value segment, whereas the version update itself plays little direct role in this classification.

B. Gradient Boosting Machine (GBM)

To allow for nonlinear effects and interactions among predictors, we fit a gradient boosting machine (GBM) with Bernoulli loss to classify high-value users (top 25% of revenue). Using 5-fold cross-validation, we tuned the number of trees and selected a model with about 1,170 trees, where the cross-validated deviance was minimized (**Figure B2**). On the held-out test set, the GBM achieved an accuracy of about 0.79 and an AUC of 0.79, only slightly higher than the single classification tree but clearly better than chance. The relative influence summary (**Table B4)** indicates that PC1_engagement is by far the most important predictor (~45% of total influence), followed by PC2_pattern (~27%) and country (~22%), with month, version, and weekday contributing relatively little (<5% each). This confirms that overall engagement intensity and behavioral patterns are the main drivers of whether a user is high-value, while the version indicator itself has only a modest additional effect once behavior and country are considered.

3.3. Market Heterogeneity

To understand the mechanism behind the revenue changes, we moved from global aggregates to market-specific drivers.

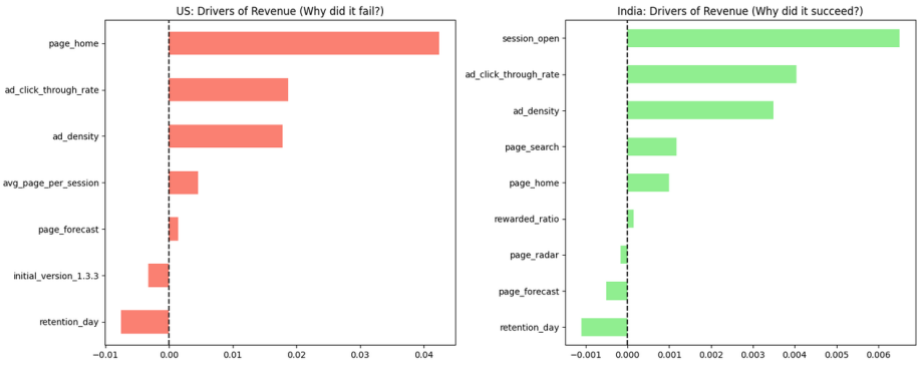

Figure 3: Lasso Coefficients Plot - US vs. India

3.3.1. The comparative Lasso analysis

The comparative Lasso analysis provides a definitive explanation for the market bifurcation, revealing two completely different "monetization metabolisms."

In the Indian market, the model identifies a clear "Frequency Driven" success story. As Figure shown above, the single strongest driver of revenue is session_open. This confirms that Indian users operate as "High-Frequency Flash Users"—their value is generated not by staying in the app for long periods, but by opening it repeatedly throughout the day.

Conversely, the US market operates on a "Home Page Dependent" logic. The Lasso results

show that US revenue is heavily reliant on home_page, with session_open fail to appear as a significant driver. This suggests that US users are purpose-driven: they open the app to check the Home Page summary and then leave. If this specific page experience is disrupted, the monetization loop breaks.

The most critical finding, however, is that the variable initial_version_1.3.3 exhibits a significant negative coefficient in the US. This result mathematically proves that the S14 update itself carries an intrinsic "revenue penalty" for US users. This implies that the specific UI/UX changes introduced in version 1.3.3 actively hinder the specific "Home Page flow" that US users value.

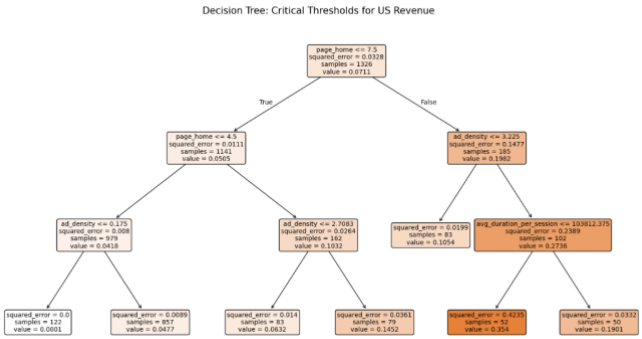

Figure 4: Decision Tree Diagram for US Market

3.3.2. The Decision Tree Analysis

To pinpoint the specific cause of revenue loss in the US, we utilized a Decision Tree which identified a critical profitability threshold at **7.5 home page views**. Users falling below this level generate near-zero revenue. The S14 update, by increasing navigation friction, prevented the majority of users from reaching this depth. Instead of browsing deeply as they did before, users were trapped in a "Dead Zone" below the 7.5-page cutoff, resulting in a large segment of low-engagement, non-monetizable sessions.

3.4. Long-term Sustainability (User Retention)

3.4.1. Logistic Regression

We modeled the probability that a user is retained at D7 using a logistic regression with retained_D7 (0/1) as the outcome and initial version, country, month of install, weekday indicator, and the two engagement components (PC1, PC2) as predictors. Odds-ratio estimates and confidence intervals are reported in **Table C1**.

After adjustment for engagement, country, and calendar effects, we find no clear evidence that version 1.3.3 changes D7 retention relative to 1.2.9. The coefficient for initial_version1.3.3 in the logistic regression is very imprecise (large standard error, $p \approx 0.92$; Table C1), and the corresponding odds ratio in Table C2 is numerically unstable, indicating that the model cannot reliably estimate a version effect. Therefore, there is no statistically reliable difference in

retention between the two versions once other factors are controlled for

Country and calendar effects are more pronounced. Relative to the reference country, users from Iraq have substantially lower retention, with an odds ratio of about 0.54 (≈46% lower odds of being retained at D7) (**Table C2**). India shows a trend toward higher retention (OR ≈ 1.4), but this effect is not statistically significant, and users from the USA are essentially similar to the reference country. For install month, July stands out as a strong negative predictor: users who installed in July have an odds ratio of roughly 0.28, corresponding to about 72% lower odds of D7 retention relative to the reference month. Estimates for rare install months (e.g., January, February, December, June) are unstable with extremely wide confidence intervals and are not interpreted substantively.

The overall engagement component PC1_engagement is the dominant predictor of retention. Each one-unit increase in PC1_engagement (corresponding to lower engagement, since higher engagement maps to more negative PC1 scores) is associated with an odds ratio of about 0.67. Equivalently, users who are one unit more engaged on this component have roughly 1/0.67 ≈ 1.5 times higher odds of being retained at D7. The second component PC2_pattern does not show a meaningful association with retention (odds ratio ≈ 1, p ≈ 0.84).

On the held-out test set, the logistic model achieves an accuracy of about 76% and an AUC of 0.85, indicating good discrimination between retained and non-retained users. Overall, these results suggest that engagement behavior (captured by PC1) and certain country/calendar effects matter much more for D7 retention than the version update itself.

3.4.2. Tree Methods

A. Classification Tree

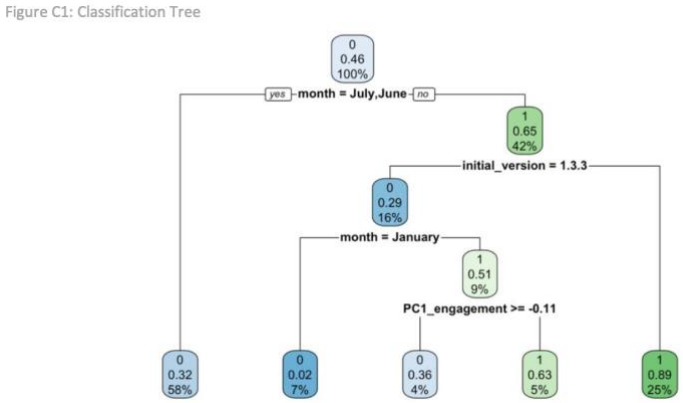

Figure C1: Classification Tree

To complement the logistic regression, we fit a classification tree to predict D7 retention from version, calendar month, weekday, and engagement components. The resulting tree (**Figure C1**) shows that calendar month is the primary driver: users who installed in June or July are sent to a leaf with an estimated D7–retention rate of about 32%, so these months are

associated with systematically lower retention. For installs outside June/July, version 1.3.3 becomes the key split: users on 1.3.3 have very high predicted retention (leaf probability ≈ 0.89), whereas users on 1.2.9 are further split by month and engagement. Among 1.2.9 users, January installs have extremely low retention (≈ 2%), while other months show moderate retention that is refined by PC1_engagement: low engagement (PC1 < −0.11) corresponds to ≈ 36% retention, and higher engagement to ≈ 63%. On the held-out test set, the tree achieves an accuracy of about 75% (913 non-retained and 449 retained users correctly classified), trading off some misclassification of retained users for a simple, interpretable set of retention rules.

B. Gradient Boosting Machine (GBM)

We next fit a boosted tree model for D7 retention using the same predictors as above (version, country, month, weekday, PC1_engagement, PC2_pattern), with Bernoulli loss and 5-fold cross-validation to choose the number of trees. Cross-validation selected about 1,400 trees (**Figure C2**). On the test set the GBM achieved an accuracy of 0.77 and an AUC of 0.86, slightly improving on both the logistic regression (AUC ≈ 0.85) and the single classification tree (accuracy ≈ 0.75).

The relative influence plot (**Table C2)** shows that most of the predictive signal comes from month (≈32%), initial version (≈30%), and overall engagement PC1 (≈26%), with smaller contributions from PC2_pattern (≈9%), and very little from country and weekday. In other words, the GBM confirms that when a user installs the app, which version they install, and how engaged they are during D0 are the key drivers of D7 retention, while country and day-of-week play only a minor role. Together with the logistic regression and tree results, this suggests that the update effect on retention is mediated largely through changes in engagement patterns and interacts with calendar timing, rather than having a simple uniform main effect.

4. Conclusion and Discussion

4.1. Summary of Findings

This study utilized a multi-stage statistical framework to evaluate the impact of the S14 version update. Our analysis leads to four primary conclusions regarding user behavior, revenue attribution, market heterogeneity, and retention dynamics:

- **Behavioral Transformation**
  The empirical results confirm that Version 1.3.3 is not a universally effective update; rather, it exposed a critical bifurcation in global user behavior. While the "Short & Frequent" model succeeded in India, driving revenue through increased open rates, it triggered an "Engagement Collapse" in mature markets like the US.
- **Mechanisms of Monetization**
  Monetization in the U.S. market remains structurally driven by page depth. By increasing UI complexity, S14 weakened this key revenue lever. Empirical evidence—including negative LASSO coefficients and decision-tree thresholds below 7.5 pages per session—supports the conclusion that the new design is misaligned with the preferences

of depth-oriented users, resulting in impaired revenue performance.

Complementing these results, our classification tree and GBM models highlight engagement intensity (PC1_engagement) as the single strongest predictor of high-value users, followed by country and usage frequency. The model achieves 78% accuracy, with high specificity (~95%) but moderate sensitivity (~27%), indicating that high-value users form a small, distinct segment that is correctly identified when predicted. Importantly, the version variable itself does not directly drive high-value classification once engagement and geography are accounted for, implying that monetization gains from S14 stem indirectly from higher user activity rather than a version-level effect.

- **Structural Market Divergence**

  Principal component loadings reveal a structural split across markets: developed markets load on engagement depth, while emerging markets load on engagement frequency. S14 optimized for frequency-driven behavior, succeeding in markets such as India and Iraq but underperforming in depth-driven markets like the U.S. and Germany. This asymmetry reinforces the need for market-tiered strategy alignment rather than one-size-fits-all design.

- **Retention Dynamics and Monetization Risk**

  Our retention models indicate that calendar effects and engagement dominate short-term user survival, while the S14 update itself does not significantly predict lower retention once these contextual factors are controlled. The logistic regression and tree-based models consistently show that users with moderate-to-high engagement outside of high-churn months (June–July) exhibit the highest retention probabilities. The GBM model (AUC ≈ 0.86) identifies month, version, and engagement as top predictors, but the direction of the version effect is small and context-dependent. Overall, there is no strong evidence that S14's higher ad density causes broad user churn—instead, retention risk is asymmetric: depth-oriented markets (e.g., U.S.) face elevated churn risk, while frequency-oriented markets (e.g., India) remain stable or improved.

4.2. Strategic Recommendations

Based on the evidence of market divergence and the identification of the "Dead Zone" in the US, we propose the following Bifurcated Product Strategy:

- **Tier-1 Markets (US/Germany):** Rollback and Redesign. Future monetization efforts should shift away from Ad Density towards Native Integration, which monetizes depth without interrupting the user experience.
- **Tier-2/3 Markets (India/Iraq):** Full Rollout and Optimization. Focus technical resources on optimizing APP Launch Speed. Since revenue in these markets is driven by session_open, reducing the latency of opening the app is the most direct lever to futher increase frequency and revenue.
- **Metrix Standardization:** Discontinue the use of "Global Averages" for decision-making.

All future A/B tests must be stratified by market tier to prevent high-volume, low-ARPU regions from masking performance degradation in high-value markets.

5. Code and Data Availability

https://github.com/Fei-p/Version-Update-Analysis

Appendix

Table A1: Version effect on behavior with adjusted country and time

term	estimate	std.error	statistic	p.value	metric
initial_version1.3.3	4.638441e-01	0.0734595	6.3142857	0.0000000	session_open
initial_version1.3.3	-1.147365e+04	4414.7215932	-2.5989518	0.0093712	avg_duration_per_session
initial_version1.3.3	-6.126746e-01	0.1333178	-4.5955960	0.0000044	avg_page_per_session
initial_version1.3.3	2.318048e+00	0.2349291	9.8670119	0.0000000	ad_impression
initial_version1.3.3	-1.369922e-01	0.1025397	-1.3359921	0.1815967	total_click
initial_version1.3.3	6.242515e-01	0.0897954	6.9519318	0.0000000	ad_density
initial_version1.3.3	-1.805000e-04	0.0005197	-0.3473152	0.7283666	rewarded_ratio

Table A2.1: Principal Components Variance Explained

PC	Std. Dev.	Prop. Var	Cum. Prop Var
PC1	1.677	0.402	0.402
PC2	1.273	0.232	0.634
PC3	0.995	0.141	0.775
PC4	0.929	0.123	0.898
PC5	0.627	0.056	0.954
PC6	0.498	0.035	0.990
PC7	0.265	0.010	1.000

Table A2.2: PCA Loadings for Behavioral Metrics

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
session_open	-0.2827947	-0.6290149	-0.0281890	0.1438806	0.4369379	0.3171935	-0.4597224
avg_duration_per_session	-0.1916157	0.2934988	-0.0397618	0.9352356	0.0162902	-0.0222903	0.0115361
avg_page_per_session	-0.4426036	0.4094497	0.0021150	-0.2044490	0.0381022	0.7443792	0.1977296
ad_impression	-0.5241120	-0.2948524	-0.0572626	-0.0378290	0.2163958	-0.3515465	0.6806222
total_click	-0.4884094	-0.2013688	-0.0239649	-0.0226084	-0.8068878	-0.0779820	-0.2503445
ad_density	-0.4091740	0.4722942	-0.0620392	-0.2456062	0.3298016	-0.4616869	-0.4726728
rewarded_ratio	-0.0821714	0.0006671	0.9949461	0.0238502	0.0265332	-0.0443856	-0.0093152

Table B1: Ridge regression coefficients for log-revenue model

Term	Coefficient
monthJune	0.2804
countryUSA	0.0324
(Intercept)	0.0322
countryIndia	-0.0235
countryIraq	-0.0151
PC1_engagement	0.0095
monthFebruary	-0.0075
monthJuly	0.0057
monthDecember	-0.0028
monthJanuary	0.0026
PC2_pattern	-0.0026
is_weekday1	0.0020
initial_version1.3.3	-0.0015

Table B.2: Lasso regression coefficients for log-revenue model

Term	Coefficient
monthJune	0.2775
countryUSA	0.0349
(Intercept)	0.0315
countryIndia	-0.0199
countryIraq	-0.0120
PC1_engagement	0.0096
monthFebruary	-0.0090
monthJuly	0.0034
PC2_pattern	-0.0024
monthDecember	-0.0023
is_weekday1	0.0012
initial_version1.3.3	0.0000
monthJanuary	0.0000

Table B.3: Model Performance Comparison (train/test)

model	RMSE_train	RMSE_test	R2_train	R2_test
OLS	0.0539597	0.0563183	0.2446324	0.2533381
Ridge	0.0539685	0.0563837	0.2443867	0.2516027
Lasso	0.0539849	0.0563981	0.2439284	0.2512197

Figure B2: Training vs CV Bernoulli Deviance

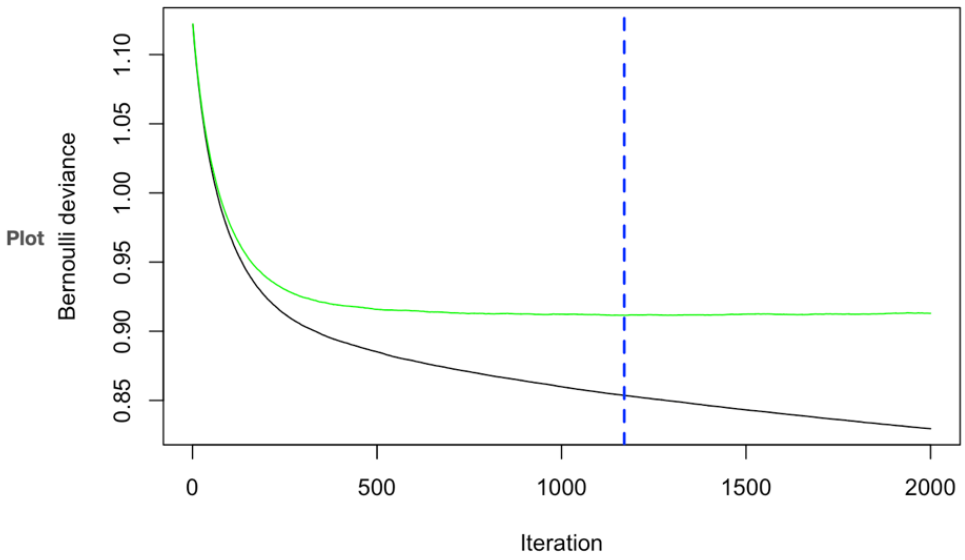

Table B4: GBM relative influence (variable importance)

Variable	Relative influence (%)
PC1_engagement	45.42
PC2_pattern	26.55
country	22.36
month	4.50
initial_version	0.84
is_weekday	0.33

Table C1: Logistic regression summary (logit link)

Term	Estimate	Std. Error	z value	P-value
(Intercept)	0.3110	0.2248	1.3832	0.1666
initial_version1.3.3	-21.9310	227.3934	-0.0964	0.9232
countryIndia	0.3210	0.2182	1.4711	0.1413
countryIraq	-0.6157	0.2321	-2.6522	0.0080
countryUSA	0.0405	0.2047	0.1979	0.8431
monthDecember	18.0183	604.1573	0.0298	0.9762
monthFebruary	21.9350	227.3934	0.0965	0.9232
monthJanuary	17.6541	227.3930	0.0776	0.9381
monthJuly	-1.2676	0.1763	-7.1916	0.0000
monthJune	-18.7238	2546.5245	-0.0074	0.9941
is_weekday1	-0.1289	0.0862	-1.4955	0.1348
PC1_engagement	-0.3989	0.0337	-11.8434	0.0000
PC2_pattern	0.0082	0.0400	0.2043	0.8381

Table C2: Odd-ratio

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.364770e+00	0.2248331	1.3831847	0.1666083	8.785221e-01	2.123207e+00
initial_version1.3.3	0.000000e+00	227.3933521	-0.0964451	0.9231671	0.000000e+00	0.000000e+00
countryIndia	1.378502e+00	0.2181972	1.4711336	0.1412550	9.013273e-01	2.121983e+00
countryIraq	5.402641e-01	0.2321451	-2.6522088	0.0079967	3.422293e-01	8.509922e-01
countryUSA	1.041348e+00	0.2047158	0.1979128	0.8431133	6.972917e-01	1.557249e+00
monthDecember	6.687015e+07	604.1572994	0.0298238	0.9762076	1.580225e+13	3.302324e+111
monthFebruary	3.359388e+09	227.3934222	0.0964629	0.9231530	1.084382e+98	2.717898e+115
monthJanuary	4.645829e+07	227.3929516	0.0776368	0.9381170	9.153817e+85	1.284886e+96
monthJuly	2.815187e-01	0.1762543	-7.1916347	0.0000000	1.987454e-01	3.967665e-01
monthJune	0.000000e+00	2546.5244962	-0.0073527	0.9941335	NA	7.636836e+97
is_weekday1	8.790961e-01	0.0861639	-1.4955331	0.1347754	7.427101e-01	1.041218e+00
PC1_engagement	6.710771e-01	0.0336788	-11.8434066	0.0000000	6.273895e-01	7.159376e-01
PC2_pattern	1.008199e+00	0.0399630	0.2043236	0.8381006	9.327012e-01	1.090943e+00

Figure C2: Training vs CV Bernoulli Deviance

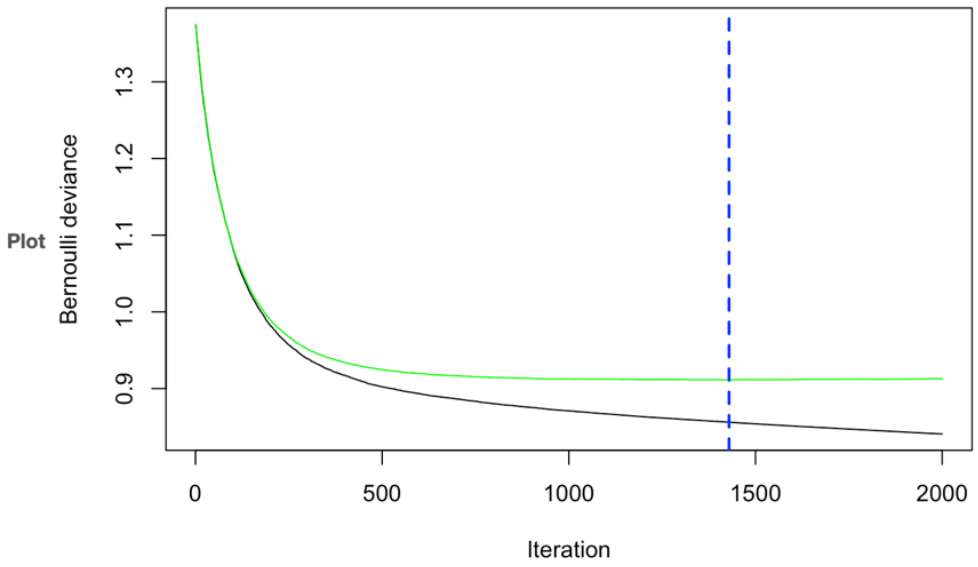

Table C2: GBM relative influence (variable importance)

Variable	Relative influence (%)
month	31.84
initial_version	30.16
PC1_engagement	26.21
PC2_pattern	8.64
country	2.71
is_weekday	0.44