IS 507: Final Project Report
Social Media Ad Optimization: Understanding Engagement, CTR, and Conversions

Team Members: Sakshi Katolkar, Shithil Shetty, Sneha Vyas, Srinath Venkatesh, Kanishka Gupta

INTRODUCTION

This project explores the intricate dynamics of digital advertising performance across major social media platforms, specifically Facebook and Instagram. In today's competitive digital landscape, advertisers face mounting pressure to maximize return on investment while navigating an increasingly complex ecosystem of user behaviors, platform algorithms, and market variations. Our research addresses a critical challenge that many advertisers encounter: high impression counts that fail to translate into meaningful conversions, resulting in wasted advertising spend and missed opportunities.

The core motivation for this research stems from a fundamental problem in digital advertising—limited insight into what drives conversions. While advertisers can easily track surface-level metrics like impressions and clicks, understanding the deeper patterns that separate successful campaigns from inefficient ones requires sophisticated analysis. By examining ad performance across five different countries and multiple demographic segments, we aimed to uncover actionable insights that could enable more efficient budget allocation and smarter, data-driven campaign decisions.

Our analysis was guided by five comprehensive research questions that explore different facets of ad performance:

Research Question 1: Which audience segments have high impressions but low conversions, based on age, interests, and location?

Research Question 2: Which specific combinations of ad platform, ad type, day of week, and device types are most predictive of high or low Click-Through Rate?

Research Question 3: How can we predict the best combination of day of week and device type to maximize conversions in each country?

Research Question 4: Among ad characteristics, demographics, and device context, which variables are the strongest predictors of engagement?

Research Question 5: Can a richer Engagement Score (built from clicks and time spent) outperform CTR in predicting whether a user will convert?

DATASET OVERVIEW

Our analysis is based on a broad social media advertising dataset obtained from Kaggle. The broad social media advertising dataset is inclusive of all aspects of user interaction with online social advertisements on both Instagram and Facebook platforms. Our main focus is based on sixteen important variables available in this dataset, which include all information necessary concerning performance of both the demographic information of the viewer (including age, gender, location, and interest indicators) and the related ad information (identifier, type, platform, and format).

One of the reasons this set of data proved to be particularly interesting is because it allowed for not only a consideration of which people clicked on the advertisements, but when and where they did it. As such a cross-platform analysis proved invaluable in making direct comparisons between not just Facebook, but Instagram as well, and the global reach allowed for regional comparisons.

Data source: https://www.kaggle.com/datasets/ziya07/social-media-ad-dataset/data

Data Cleaning and Preparation

The quality of our work in data preparation can be gleaned from the fact that we did everything in our powers to make our numbers accurate and consistent. For instance, our analysis involved ensuring there were no missing values in the dataset, which were consequently replaced with the right methods. Location, interest, device type, and ad format were considered as categorical variables and were thus formatted to be consistent. Variables for dates and time were formatted to allow for time series analysis.
To improve our analytical power, we have derived important key performance indicators, which were necessary for our research. The Click-Through Rate was measured by taking the clicks divided by the impressions, which allowed for a normalized way of assessing the performance of an ad. The Conversion Rate was measured by dividing the clicks by the conversions, which allowed for an assessment of performance at the final step of the conversion funnel. The processing part of our work included generating interaction terms and composite variables, which were used to model relationships such as ad format performance by device type or demographic characteristics with time-series behavior.

METHODOLOGY

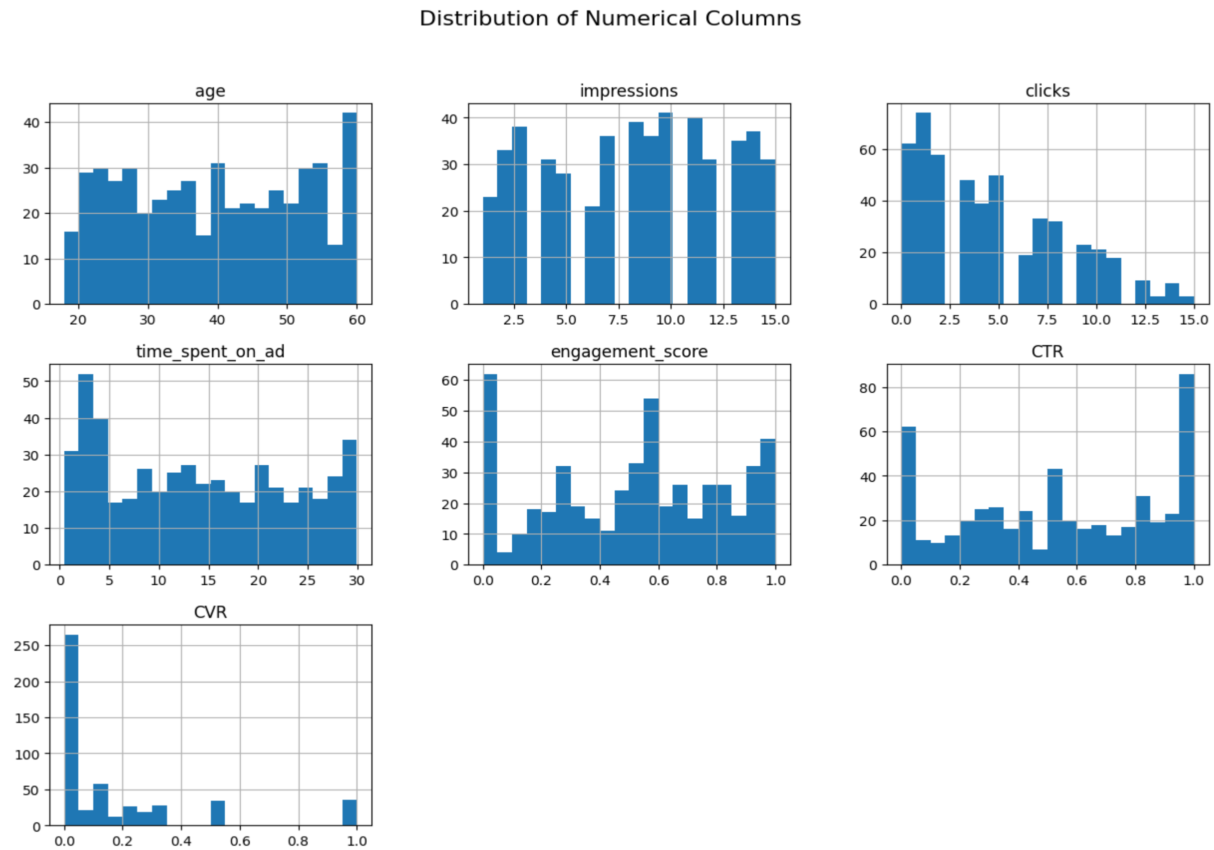
Our analysis combined EDA, statistical modeling, and machine learning to extract actionable insights from the advertising data.

Comparison to Existing Solutions

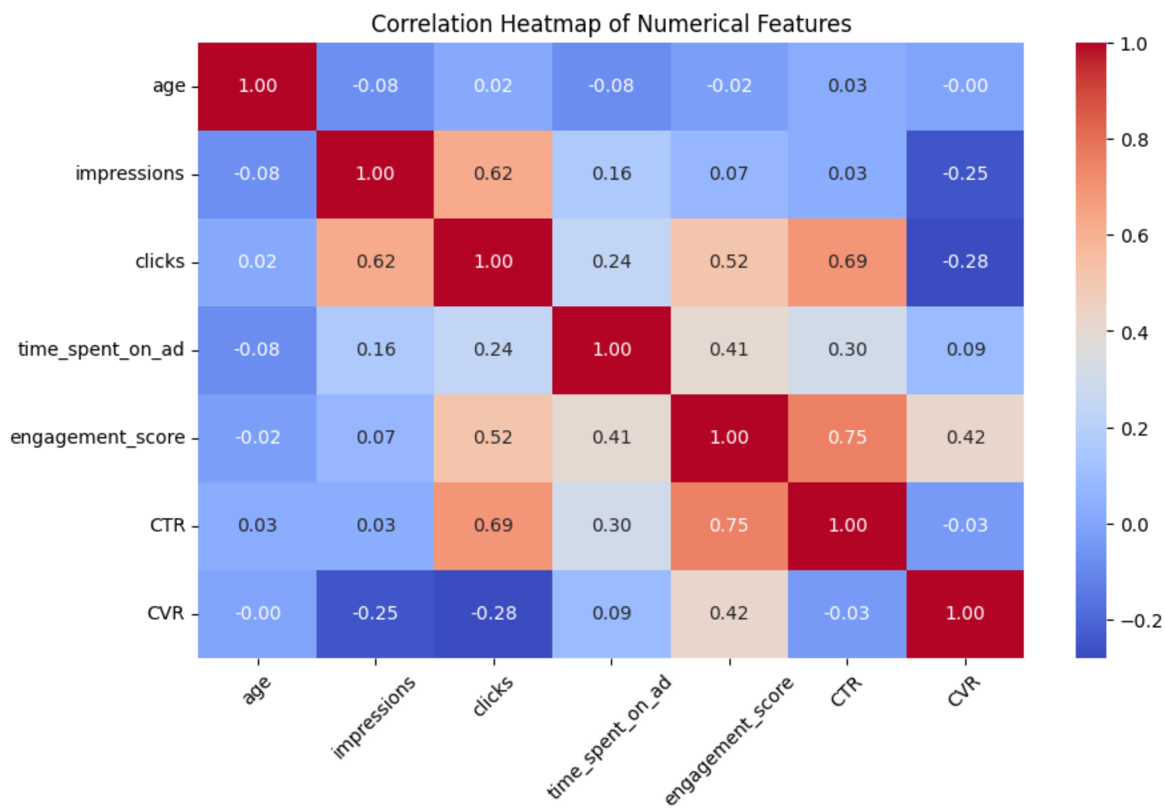
Traditional social media optimization relies on analytics provided by the platforms, and it typically focuses on surface-level metrics such as impressions and clicks. Our approach differed in two important ways. First, we used deep analysis that combined multiple modeling techniques to tease out hidden patterns in the data that might be difficult or impossible for a standard analytics tool to discern. The emphasis we placed on identifying underperforming segments-those exhibiting high impressions and low conversions-offered targeted solutions for reducing wasted ad spend. Second, we developed a novel Engagement Score that integrated multiple behavioral signals into a single predictive measure. Whereas most advertisers are limited by relying on CTR, our composite metric captured both breadth- clicks-and depth-time spent-potentially offering more accurate prediction of conversion intent.

Exploratory Data Analysis

Our exploratory analysis examined dataset structure and distributions to understand fundamental social media advertising patterns. Distribution analysis revealed age spread from early 20s through mid-60s with concentration at 55-60, indicating mature demographic penetration. Impressions clustered between 2.5K-15K with peaks around 7.5K-10K, suggesting consistent moderate-reach targeting. Clicks showed strong left skew with majority receiving 0-2.5 clicks, reflecting attention capture challenges. Time spent exhibited bimodal patterns quick dismissals versus meaningful 30-35 second engagements. Engagement scores showed pronounced peaks at extremes, reinforcing the divide between ineffective and effective experiences. CTR displayed polarization with 60-65% near zero but secondary peaks at 12.5-15%. CVR showed the most dramatic pattern 60% zero conversions, massive concentration at 85%, indicating sharp targeting quality divides.



Correlation checks highlighted key links. While impressions tied to clicks at 0.62, high visibility didn't mean active user response. A solid link appeared between Engagement Score and CTR - reaching 0.75 - whereas Engagement Score linked moderately to CVR at 0.42, hinting that engaged users are more likely to convert. Unexpectedly, CTR barely related to CVR, sitting at0.03, meaning lots of clicks don't ensure better outcomes. On top of that, more impressions connected to a dip in CVR (0.25), pointing toward wider outreach possibly reducing conversion efficiency.



Feature engineering meant tossing out bad records, tagging ads that showed a lot but led to no sales - especially those past the 75th percentile - and working out adjusted CTR and CVR scores. Instead of just piling things together, we mixed platform with device types, linked countries with days, tracked how deeply users engaged, then wove in personal details tied to context. Grouping uncovered real customer clusters, especially one key chunk: high views, zero results - that's money tossed away.

Data Quality and Derived Metrics

```python
imp_thresh = df["impressions"].quantile(0.75)

df["high_imp_low_conv"] = (
    (df["impressions"] >= imp_thresh) &
    (df["conversion"] == 0)
).astype(int)

df["high_imp_low_conv"].mean()
df.head()
```

Python

latform	ad_type	impressions	clicks	conversion	time_spent_on_ad	day_of_week	device_type	engagement_score	CTR	CVR	high_imp_low_conv
cebook	Image	3	0	0	3.38	Friday	Mobile	0.02	0.000000	0.000000	0
cebook	Image	9	9	1	6.77	Saturday	Tablet	0.93	1.000000	0.111111	0
stagram	Image	13	12	1	13.26	Wednesday	Mobile	0.93	0.923077	0.083333	0
cebook	Video	14	5	0	24.41	Saturday	Desktop	0.28	0.357143	0.000000	1
stagram	Carousel	10	5	0	21.43	Monday	Tablet	0.35	0.500000	0.000000	0

Feature engineering steps:

- **Invalid Record Removal**: Eliminated records where clicks exceeded impressions (data collection errors)

- **High-Impression-Low-Conversion Flag**: Binary indicator for ads above 75th percentile in impressions but zero conversions—directly supporting RQ1
- **Calculated CTR**: (Clicks / Impressions) for normalized performance comparison
- **Calculated CVR**: (Conversions / Clicks) to measure final conversion efficiency

Model Evaluation

Performance metrics appropriate to each type of predictive model were utilised to assess model performance. When examining the performance of models aimed at predicting conversions within a classification context, we used a combination of accuracy, precision and recall (including AUC score). For regression tasks with CTR predictions, we calculated Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

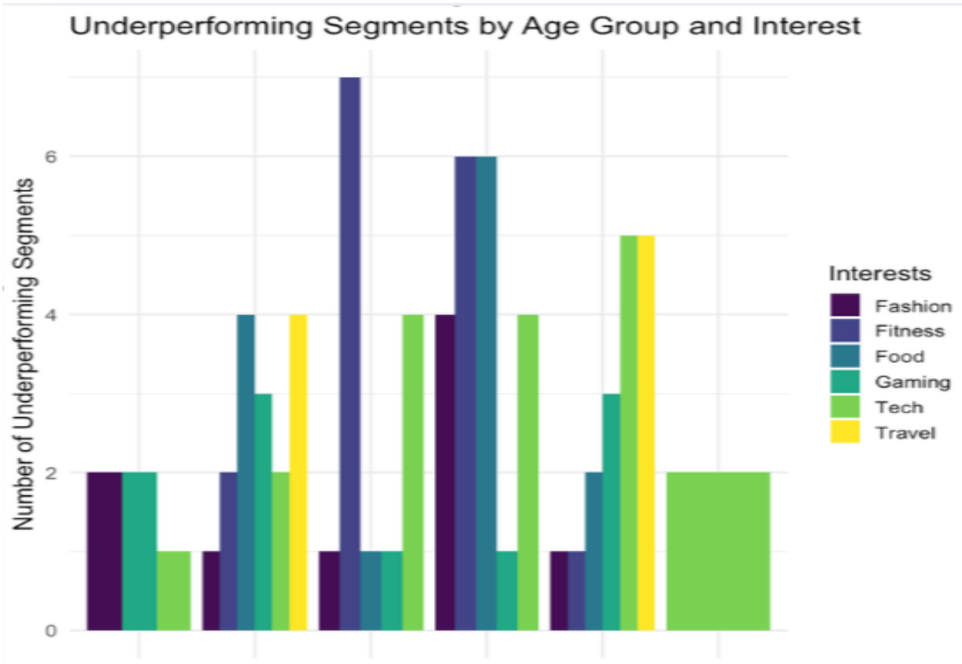
In order to validate the models' performance on unseen data, we used cross-validation techniques that provide an additional level of validation that the models would generalise well to future datasets, rather than simply remembering how the previous training pattern had been learnt.

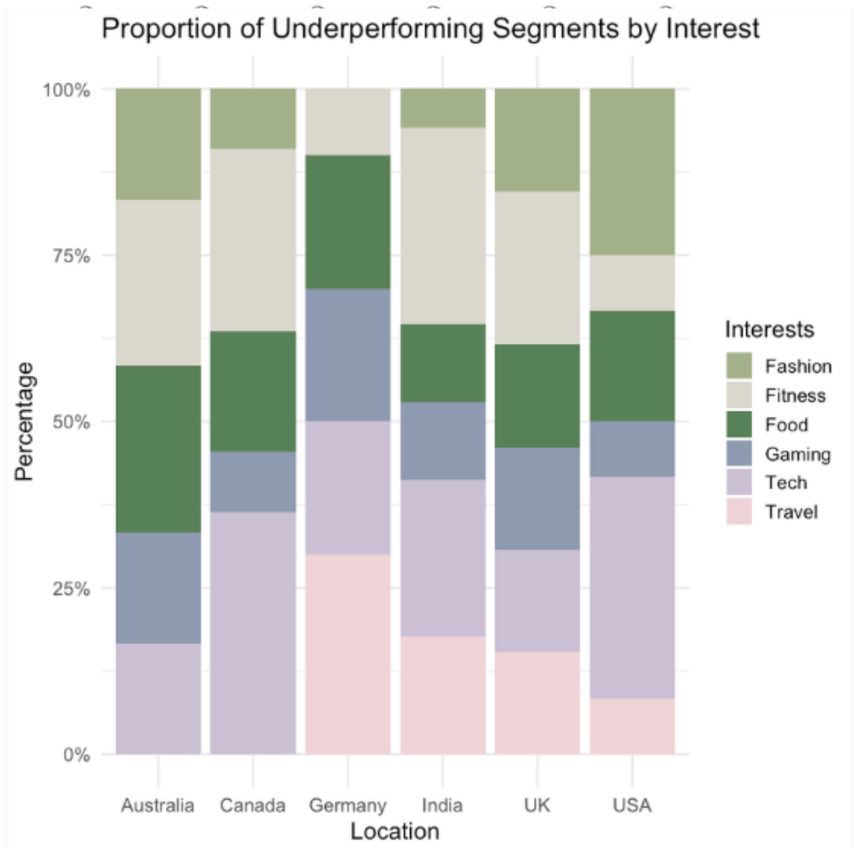
While many researchers place as much emphasis on predictive accuracy, we believe that actionable insights are a more important objective than a slight improvement in accuracy. Therefore we paid close attention to the overall interpretability of the model, so that we could provide our advertisers with strategic information, by analysing the decision trees and the coefficients on the regression equations not just to evaluate how well they performed but also the how strategic they were.

RESULTS AND INSIGHTS

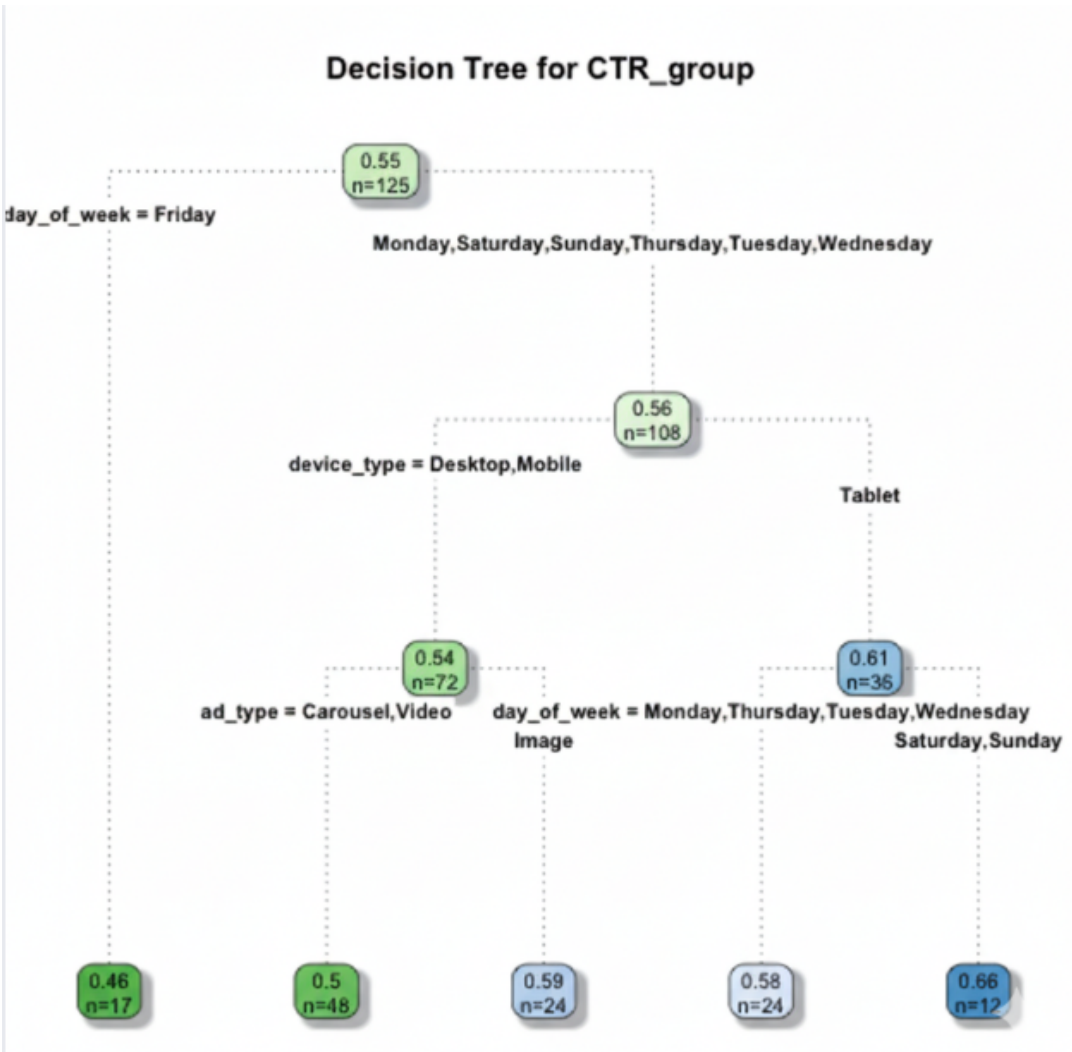
Research Question 1: Identifying Underperforming Segments

Research Question 1: Identifying Underperforming Segments We found many instances where there was a large disconnect between visibility and conversion. The 30-49 age range took up a lot of spend and had a very poor conversion rate, indicating a possible misalignment with the message being delivered or a browse without any intent to buy. The interests of Technology, travel, fitness and gaming received many impressions, but had a low conversion rate across multiple markets (India, Germany and Australia had the most inefficient conversion rates). While every interest area had sub-segments that were underperforming, the geographic distribution was different in each region indicating the need for a localised approach.


Underperforming Segments by Age Group and Interest
Number of Underperforming Segments
0
2
4
6
Interests
Fashion
Fitness
Food
Gaming
Tech
Travel

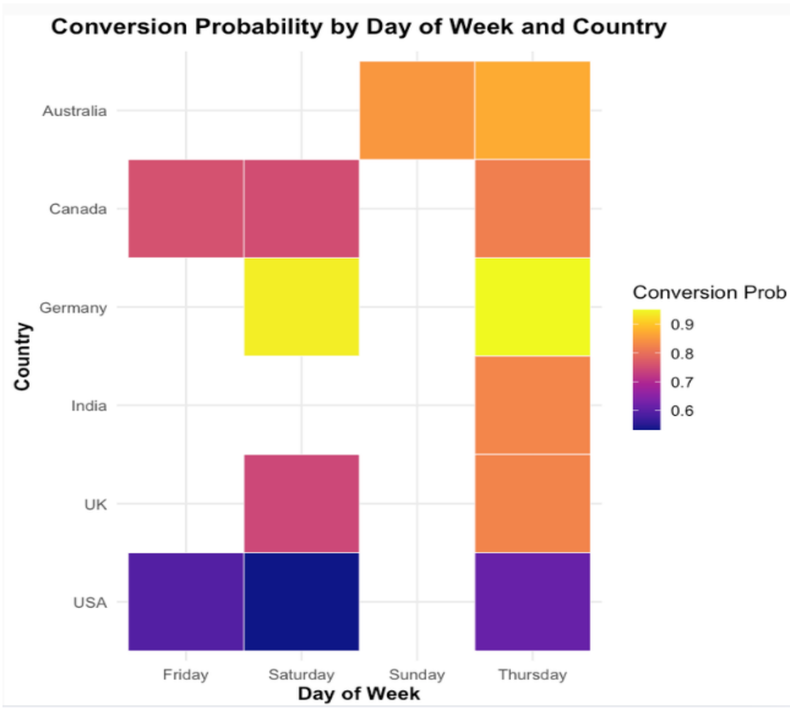

Proportion of Underperforming Segments by Interest
Percentage
0%
25%
50%
75%
100%
Australia
Canada
Germany
India
UK
USA
Location
Interests
Fashion
Fitness
Food
Gaming
Tech
Travel

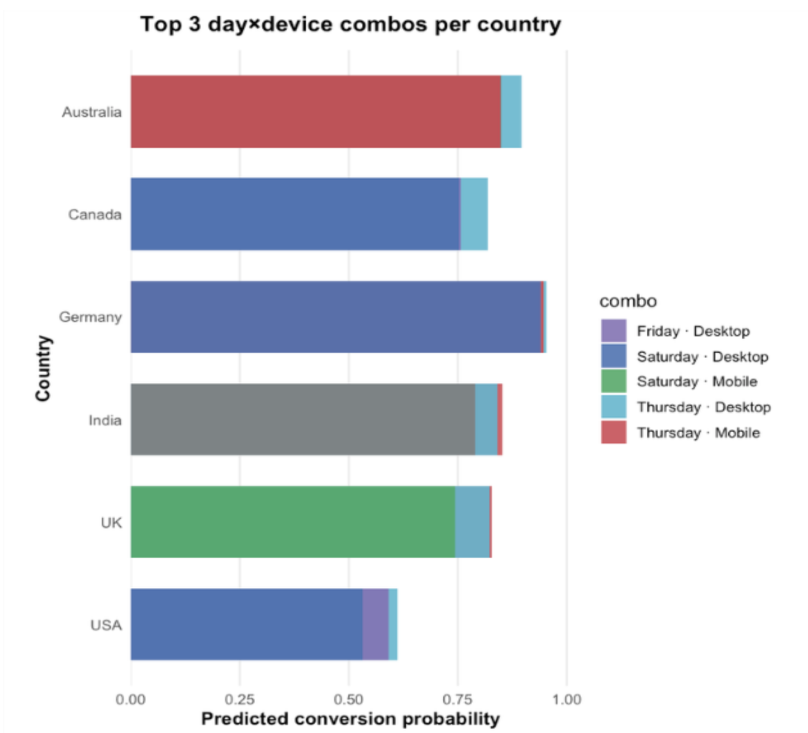
Research Question 2: Optimal Conditions for Click-Through Rate
Decision tree analysis revealed timing trumped other factors, with day of week as the primary decision point. Advertisers should avoid Fridays while prioritizing Saturday, Sunday, and Monday, reflecting how work routines reduce receptiveness while weekend leisure browsing creates favorable conditions. Tablet users, especially on weekends, demonstrated significantly higher CTR than mobile or desktop, challenging mobile-first assumptions. Image ads consistently outperformed Video and Carousel formats for CTR, likely due to cognitive load considerations in fast-scrolling feeds.



Research Question 3: Country-Specific Conversion Optimization

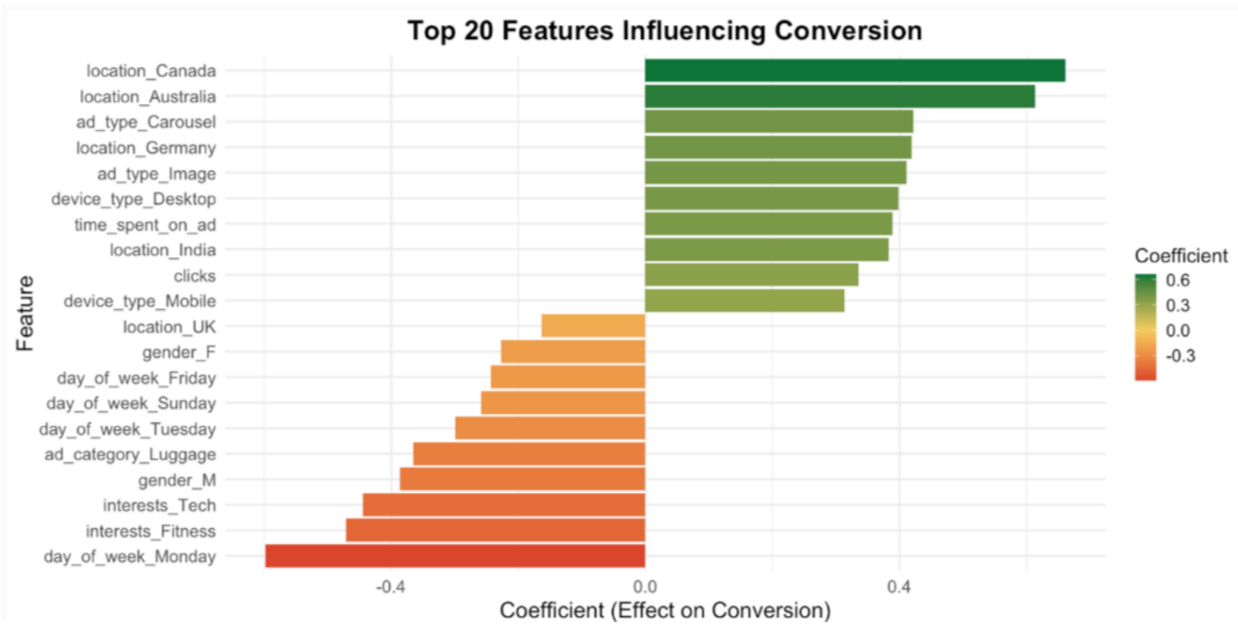
Regional breakdowns uncovered wildly varied top-performing setups. In Australia, desktop wins on Thursdays. Canadian users lean toward mobile over weekends. Germans favor desktop activity come Mondays. Indian audiences engage most via mobile Sundays. The UK sees peak desktop performance Fridays. Americans spread strong results across devices on Thursdays. Here's the catch - what works in one place tanks in another. That means tailored schedules per location, creatives built for specific gadgets, and spending adjusted by local rhythm.

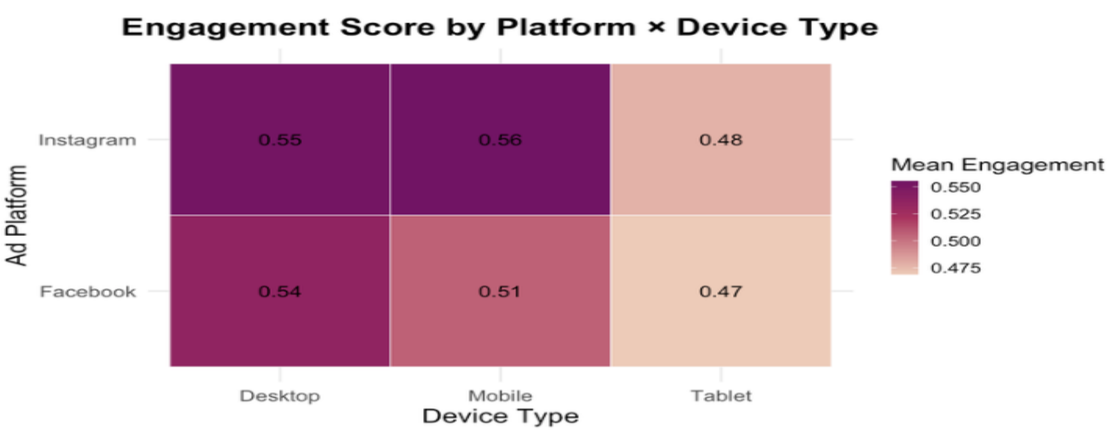




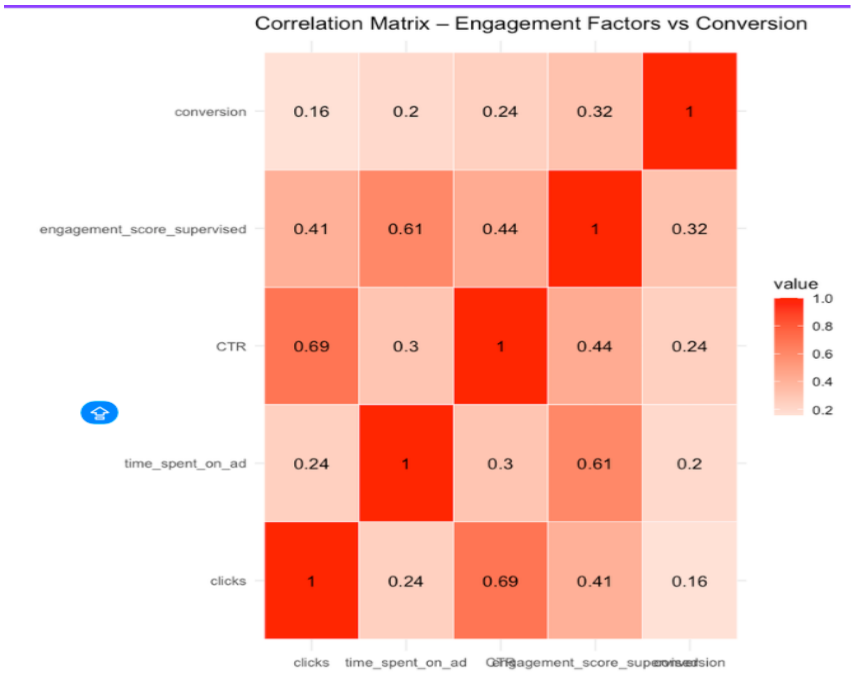
Research Question 4: Drivers of Engagement

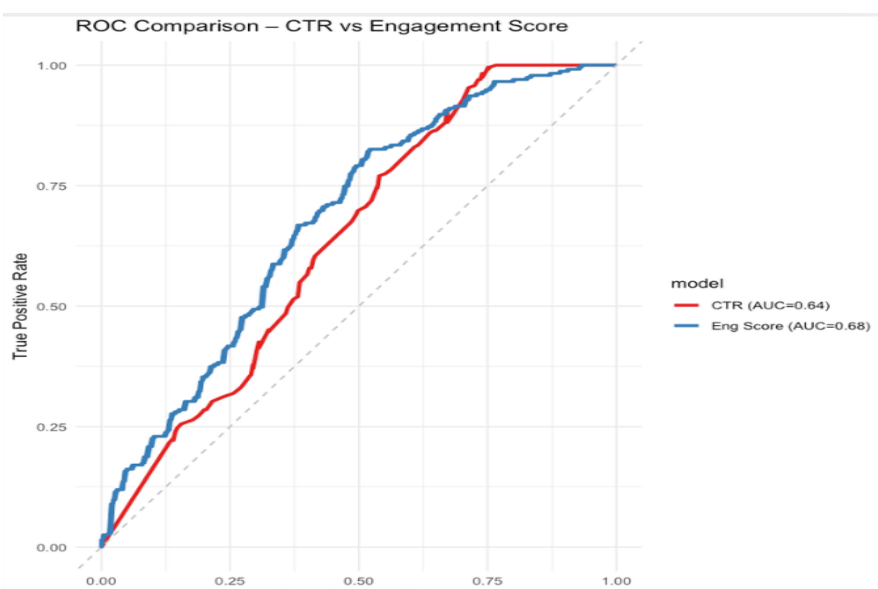
Regression tests pointed to a pattern. On top, location mattered most - Canada and Australia had stronger results, whereas the UK trailed behind, hinting at how culture shapes online habits. Next came ad format: images pulled more attention compared to videos or carousels. As for device, desktop beat mobile and tablet, possibly because people focus better when using computers. Demographics had a slight impact, yet time-related factors plus user interests mattered less. When comparing platforms, Instagram on phones got the most attention - hitting 0.625 - whereas tablets always lagged behind on both, even though they once led in click rates.





Research Question 5: Engagement Score vs. CTR as Conversion Predictors
Comparative analysis revealed stronger behavioral insights through deeper data. Instead of just clicks, time spent mattered - Engagement Score linked to conversions at 0.32, higher than CTR's 0.24. The ROC results, Engagement Score hit an AUC of 0.66; CTR reached only 0.64. Even though both beat chance guessing, one stood out. By blending duration with click behavior, predictions got sharper. That edge proved useful when forecasting real user actions. In real terms, this boost in precision means higher returns on thousands of ad views - separating quick clicks from serious buyers. That 29% tighter link delivers meaningful forecasting power, which might cut costs and lift income big time for large firms.





KEY INSIGHTS AND LEARNINGS

- Big impression numbers don’t mean sales - people aged 30 to 49, into tech or travel, fitness, gaming soaked up cash but barely converted, showing broad targeting can burn money fast
- Regional differences require localized strategies - Optimal day-device combinations varied dramatically (Thursday Desktop in Australia, Saturday Mobile in Canada, Monday Desktop in Germany), requiring region-specific optimization over universal campaigns
- Timing affects how people engage - on Fridays things slow down, yet weekends pick up. Tablets get clicks easily though users don’t stick around much. Mobile wins in certain areas, whereas in others desktop stays ahead.
- Picture ads beat videos and slideshows - though everyone pushes video, still images pulled higher click rates plus more interaction when selling straight to customers, likely because they’re easier to process while scrolling quick.
- Engagement Score beats CTR – it mixes how long people stay with what they do, giving a clearer hint who’s going to buy. Instead of just clicks, it uses time on page to spot real interest more accurately.
- Where people are makes the biggest difference - engagement shifted way more by place than age or likes, meaning local culture sets deeper patterns that basic audience labels just miss

PROJECT ADJUSTMENTS & TEAM COLLABORATION

- Research questions got clearer through deeper analysis - RQ4 shifted toward detailed regression ranking instead, RQ5 expanded into complete ROC evaluation using tested methods
- Surprising results from one group meant we had to dig deeper into age and interests by area using split-up views
- Visuals took over – heatmaps showed tricky trends at a glance
- Trying lots of tweaks was key till we picked a mix of clicks with time-on-page to shape the Engagement Score
- Simpler models like decision trees or logistic regression usually made it easier to understand choices when decidinng

Our team of five members brought complementary skills that enriched every phase.

Kanishka Gupta handled data preparation, feature engineering, and advanced modeling for Research Questions 5. Shithil Shetty did EDA and Research Question 4, this involved comprehensive feature engineering to create new derived metrics and conducting regression analysis to identify engagement predictors. Sneha Vyas led exploratory analysis and Research Question 1 on underperforming segments. Srinath Venkatesh focused on Research Question 2's decision tree models for CTR prediction. Sakshi Katolkar built country-specific conversion prediction models for Research Question 3. Kanishka Gupta managed project coordination, documentation. We maintained regular communication through team meetings, adopted peer review approaches.

FUTURE DIRECTIONS

- Boosted tools – Add in trends from time of year, market shifts, updates to site formulas, kinds of items sold.
- Smart tweaks - like bumping gradients, linking neurons, stacking models - to juggle precision without losing clarity.
- Looking at time patterns - how things change each month, comparing this year to last, plus what happens during holidays.
- Causal inference – Figure out real cause-effect links, not just patterns that happen together.
- Live tweaks – On-the-fly audience picks, ad version choices, spending setup rules.
- User journey mapping - full routes from initial contact to sale, improving each stage step by step.
- Cross-platform optimization – Synergistic effects of simultaneous Facebook and Instagram advertising.
- Privacy-safe data analysis - using diff privacy or shared learning when info is scarce.

CONCLUSION

This project successfully addressed understanding engagement, clicks, and conversion drivers across audiences, platforms, and contexts through rigorous exploratory analysis, statistical modeling, and machine learning.

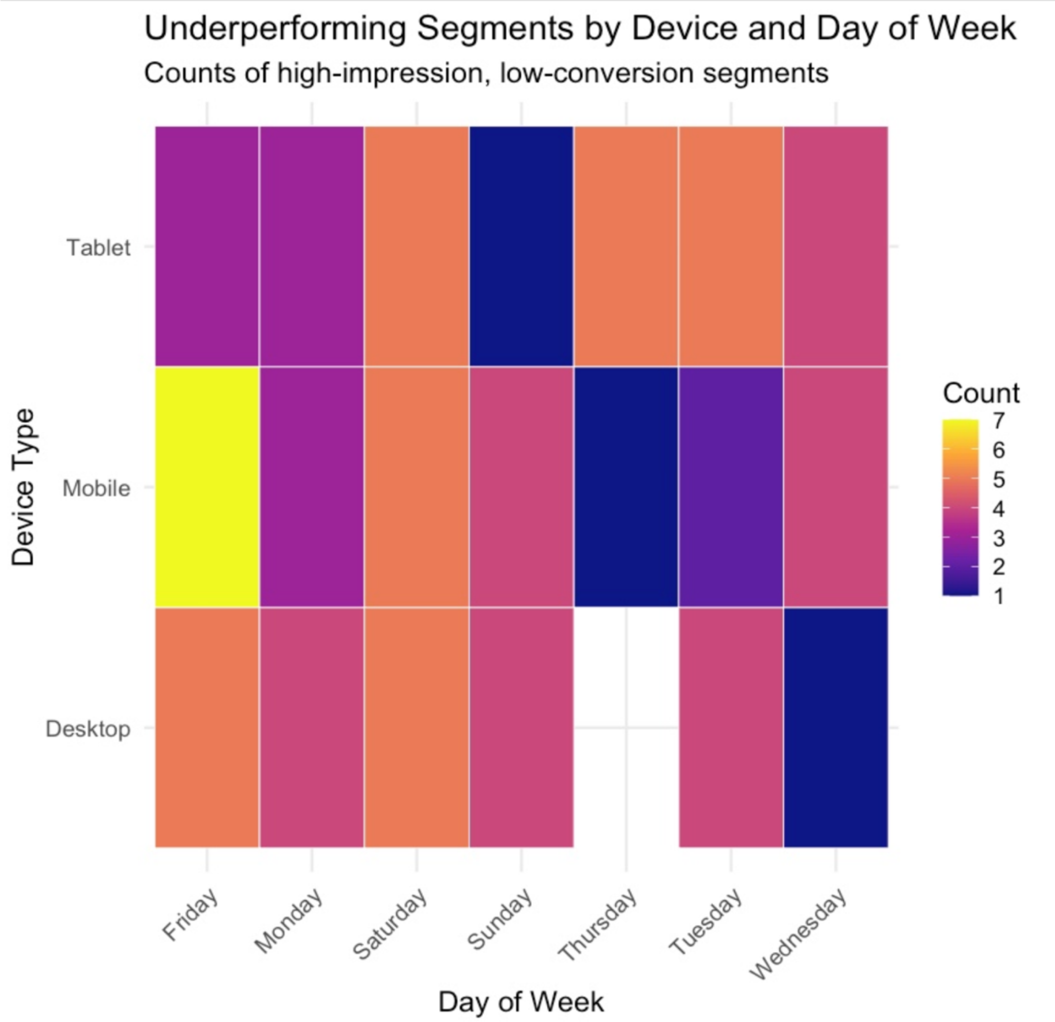
Key Findings:

- High impressions don't guarantee conversions
- Optimal strategies vary dramatically by country
- Image formats outperform Video/Carousel for direct response
- Engagement Score outperforms traditional CTR for conversion prediction

Practical Implications: Advertisers can reallocate budgets from high-impression low-conversion segments, optimize timing and device targeting by country, make evidence-based format decisions, and adopt nuanced engagement metrics capturing true intent. The project reinforced that sophisticated analysis uncovers hidden patterns, regional variations resist universal solutions, and interpretability matters as much as predictive performance. Our frameworks for identifying underperforming segments, country-specific optimization, and composite engagement metrics contribute to a more efficient digital advertising ecosystem with smarter budget allocation and improved user experiences.

Analytical Approach Justification - Feedback

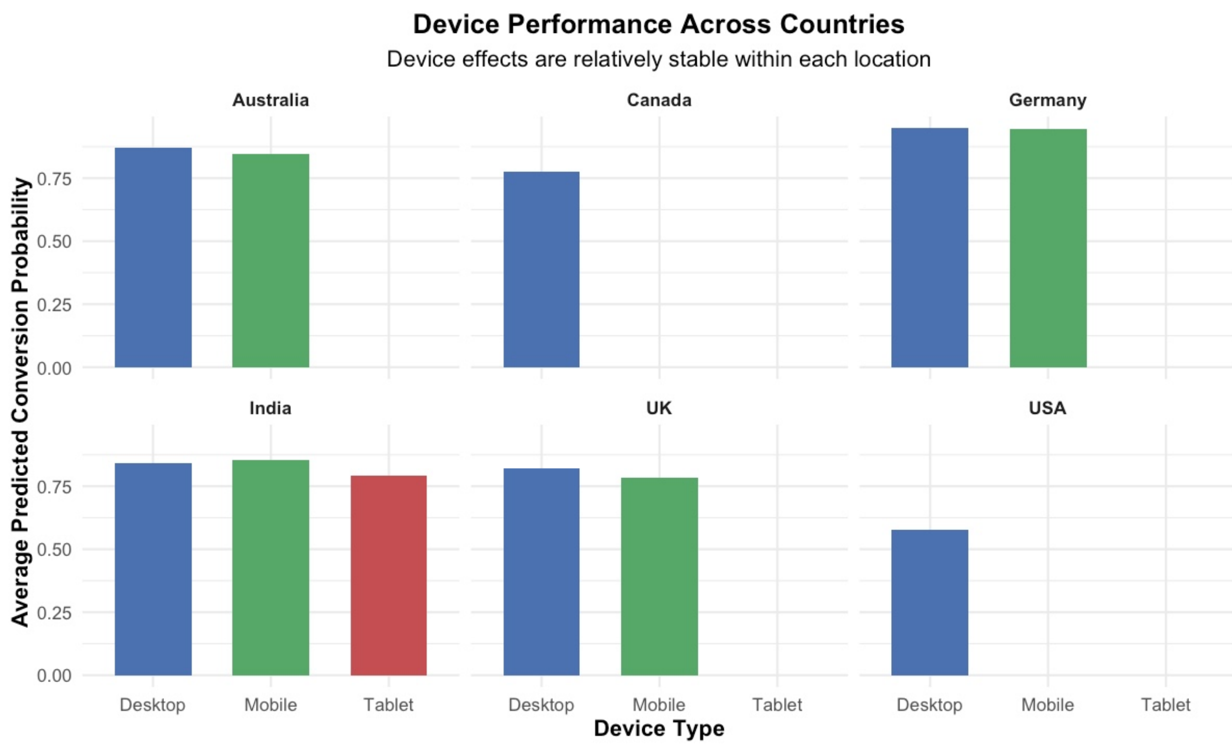
Question 1: Why did you focus on age, interest, and location for RQ1? Why not other variables like time of day or device?
We narrowed our focus to age, interest, and location because this combination gets to the heart of who the audience is, not just when or how we try to reach them. These demographic characteristics help advertisers determine which audience segments to include or exclude from campaigns. Device type and day of week are about timing and context for us which we dealt with in Research Questions 2 and 3. That said, we still examined these factors just as secondary patterns within our primary audience segments. Our visualization showing underperforming segments by device and day demonstrates this.



Question 2: Why use Friday as a cutoff in the decision tree? Have you tried other days?

It was not us, but the decision tree algorithm, that automatically selected Friday. The algorithm tested all 7 days and selected Friday because it provided the best split in high-CTR from low-CTR observations. It was a data-driven finding indicating that Friday behavior was different from the rest of the days in the dataset. It was not a manual choice to exclude Friday; the model tested all possible splits and concluded that this one was the best in terms of predictive performance.

Question 3: Why create day-device interactions? Why not other combinations like device × location?



Based on this feedback, we analyzed device performance across countries and found that device effects remain relatively stable within each location (see visualization). This confirmed our approach was correct—day-device interactions capture dynamic weekly behavioral changes that advertisers can act upon through scheduling, while device-location effects are static market characteristics already captured by our country-specific models.