Beyond Lifetime Value: A Customer Journey Analysis to Fan Engagement and Spending in Professional Sports

by

Christina Elizabeth Antonakakis

B.S. Business Analytics and Computer Science, Economics, and Data Science, Massachusetts Institute of Technology (2022)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE, ECONOMICS, AND DATA SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

© 2025 Christina Elizabeth Antonakakis. All rights reserved.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by:	Christina Elizabeth Antonakakis Department of Electrical Engineering and Computer Science January 17, 2025
Certified by:	Christina Chase Lecturer, MIT EECS & MechE, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Beyond Lifetime Value: A Customer Journey Analysis to Fan Engagement and Spending in Professional Sports

by

Christina Elizabeth Antonakakis

Submitted to the Department of Electrical Engineering and Computer Science
on January 17, 2025 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE, ECONOMICS, AND DATA SCIENCE

ABSTRACT

This research examines engagement and spending behaviors in a professional sports ecosystem, introducing Customer Journey Analysis (CJA) as a dynamic alternative to traditional customer lifetime value (LTV) models. Through an analysis of over 930,000 net new fans acquired from July 1, 2021, to June 30, 2024, this study identifies critical patterns in acquisition channels, spending behaviors, and engagement metrics over multiple seasons. Notably, the findings highlight the significant influence of early touchpoints, such as ticket purchases and email interactions, on fan progression. Metrics like email open rates and multi-channel engagement emerge as strong predictors of future spending, revealing nuanced insights into fan behavior.

This research emphasizes the importance of integrating behavioral and financial metrics to sustain fan involvement. By transitioning from static LTV models to a multi-dimensional CJA framework, actionable strategies are proposed for optimizing engagement channels, improving retention, and driving long-term revenue growth. Key findings reveal that predictive modeling and customer segmentation analysis are instrumental in identifying high-potential fans and distinct audience profiles. Tailored retention strategies, including personalized follow-ups and exclusive engagement incentives, address churn risks while fostering ramp-up and loyalty across diverse fan groups. Future work should explore tenured fan behaviors and incorporate diverse data sources, such as in-venue spending and team performance metrics, to deepen understanding of fan evolution across different lifecycle stages.

Thesis supervisor: Christina Chase
Title: Lecturer, MIT EECS & MechE

Acknowledgments

First and foremost, I would like to thank God for providing me with the strength, wisdom, and perseverance to complete this journey. Without His guidance and grace, none of this would have been possible. I am deeply grateful to my parents, Joe and Linda, my brother Joseph, and my sister Kathleen for their unwavering encouragement and love throughout this journey. Your belief in me has been a constant source of motivation. To my boyfriend, Christopher, thank you for your patience, understanding, and unwavering presence during the most challenging moments of this process. What a journey these past 5.5 years at MIT have been, plus a detour to Dallas along the way. I could not have done this without you by my side.

I would also like to express my deepest gratitude to Professor Christina Chase, whose expertise, guidance, and thoughtful mentorship have been instrumental in shaping this thesis. Working with you has been an absolute pleasure, and your support and encouragement have left a lasting impression on me. To the dedicated MIT Sports Lab, thank you for fostering such a dynamic and inspiring environment where I was surrounded by so many brilliant individuals. Being part of this community was an honor, and I am grateful for the opportunity to learn and grow alongside everyone. I am incredibly thankful for the financial backing I received through my role as a Teaching Assistant for the 6.100 class over the course of three semesters. Thank you to Professors Ana Bell, Andrew Wang, John Guttag, and Stefanie Mueller for providing me with the opportunity to give back to the class that encouraged me to become passionate about computer science and data science.

A special thank you to Coach Sonia Raman for coming into my life at basketball camp during the summer of 2014. You not only recognized my ability to succeed as a student at MIT, but also to make a meaningful impact as an athlete on your women's basketball team. Your compassion and guidance continue to inspire me, and I am immensely grateful for the friendship we still share today. I would also like to thank Coach Lucia Robinson-Griggs, with whom I had the honor of concluding my basketball career—your leadership and passion have left an enduring mark on my journey at MIT. To Scott Alessandro, my advisor at Sloan, your insights and guidance have been invaluable over the last several years, far beyond what you may even realize today.

Finally, to all the friends and teammates I met along the way, thank you for being such an integral part of this adventure and for the incredible memories we have created along the way. Attending my dream school has been a blessing, but it is the remarkable people who have truly made MIT so special. Your camaraderie, laughter, and lasting friendship have made my time here unforgettable and I am very grateful to have shared this experience with you.

Contents

Chapter 1

Introduction

1.1 Motivation

Fan engagement and loyalty are essential drivers of sustained revenue and brand valuation in the sports industry. However, traditional metrics like Customer Lifetime Value (CLV) often reduce the complexity of the fan journey to short-term financial transactions while overlooking the broader and more nuanced ways fans interact with teams over time. While CLV models remain a useful tool for quantifying monetary contributions, they fail to capture the holistic nature of fan relationships, which are shaped by a combination of behavioral, emotional, and financial factors.

This thesis focuses specifically on net new fans entering the team's ecosystem—those engaging for the first time through a variety of acquisition channels. These include game attendance, ticket purchases, online merchandise platforms, in-venue Wi-Fi registration, mobile app usage, youth sport's camps, and other sources. Understanding how these fans progress within the ecosystem is critical for developing strategies that not only foster retention but also encourage growth in engagement and spending habits. For net new fans, their initial touchpoints with the team are particularly influential in shaping their future interactions. These early engagements can spark deeper involvement and lead to a progression through various stages of fandom, such as reviewing and participating in email content, purchasing merchandise, attending a few games games, or even becoming a season ticket member.

This research introduces Customer Journey Analysis (CJA), a multi-season framework designed to track and understand the evolution of these new fans. By broadening the scope of analysis beyond immediate financial returns, this approach identifies strategies that support fan progression and deepen their relationship with the team. The focus is on fostering a lasting connection that integrates both behavioral and financial dimensions, moving beyond retention to unlock opportunities for ramping up participation. By focusing on net new fans, this analysis offers actionable insights for fostering long-term loyalty and sustaining engagement over time.

1.2 The Role of Customer Journey in the Sports Industry

The sports industry presents unique challenges in understanding and managing fan engagement due to its inherently cyclical nature. Unlike other industries, where customer allegiance tends to remain stable, fan behavior often fluctuates due to seasonality, team performance, or external factors, such as venue changes or impactful global events. These dynamics make understanding fan engagement an evolving process, requiring an intertemporal approach that extends beyond the scope of traditional CLV models.

The CJA framework captures a wide range of fan interactions and behaviors over multiple seasons. Unlike approaches that narrowly emphasize monetary contributions, this model seeks to understand the progression of fan behaviors—from attending games physically to

engaging digitally—highlighting how non-transactional touchpoints can also drive brand loyalty and future revenue. For example, fans who engage through mobile apps or respond to email campaigns may significantly impact the team's ecosystem, even without existing financial contributions. By incorporating these varied touchpoints, the analysis provides a more comprehensive view of fan engagement.

This framework's adaptability is especially important as fan preferences evolve over time. Whether influenced by shifts in behavior or external factors, tracking the progression of net new fans provides an opportunity to refine engagement strategies proactively. By emphasizing the importance of sustained relationships and analyzing fan progression, this thesis aims to empower teams with insights to foster deeper connections and build long-term loyalty across an ever-changing landscape.

1.3 Related Literature

The concept of CLV has long been a cornerstone for understanding fan contributions in professional sports, traditionally measured through ticket sales and renewals. Early research highlighted the influence of service quality on customer retention and spending, focusing narrowly on direct financial metrics (e.g., season ticket renewals) while excluding broader elements of fan behavior such as digital engagement or secondary market participation [1]. Conducted before the rise of social media and the complexities of the COVID-19 pandemic, this foundational work reflects an earlier stage in the evolution of CLV models.

Recent advancements in fan analysis enhance traditional CLV frameworks to account for long-term, multi-dimensional engagement. These models differ by incorporating both non-financial behaviors (e.g., emotional investment, participation in digital channels, virtual events) and financial contributions, resonating with the objectives of this thesis. Through the use of this dual-metric model, these newer approaches emphasize fan segmentation as a critical tool for targeting marketing strategies and personalizing experiences. For instance, high-engagement fans may be ideal candidates for upsell opportunities, while lower-engagement fans could benefit from tailored interventions to foster participation [2].

Another critical theme in contemporary analysis is the link between retention and revenue growth. It highlights the challenges teams face when analyzing fan behaviors beyond renewals, particularly for franchises lacking a long-standing and loyal fanbase [2]. For example, player dynamics, such as high-profile trades or retirements, can significantly impact retention and cancellations. This thesis parallels these approaches by going beyond transactional data to analyze fan progression and sustained interaction over time.

Research in the application of CLV in professional sports often centers on retention strategies for season ticket holders but remains largely tethered to financial metrics, overlooking digital engagement or secondary market behaviors [3]. Similarly, analyses in recent years have emphasized the need to prioritize fan engagement as a reliable growth driver, particularly in the wake of the COVID-19 pandemic [4]. The shift to digital interactions during live event disruptions demonstrated the importance of sustained fan relationships and long-term loyalty. This aligns closely with the objectives of this thesis in advocating for predictive engagement models that assess fan value over time.

Machine learning approaches, such as segmentation using Recency, Frequency, and Monetary (RFM) models, further contribute to this evolving landscape by identifying high-value

fan segments [5]. While effective for immediate returns, these methods often overlook fan progression and non-financial engagement, reinforcing the need for a more holistic approach.

Collectively, these works illustrate the shift from traditional, transaction-focused CLV models toward engagement-driven strategies that emphasize long-term fan value. The framework developed in this thesis builds on this progression by integrating both sets of metrics into a comprehensive CJA. This approach not only identifies opportunities for re-engagement and upsell strategies but also offers a multi-season perspective on fan loyalty, bridging gaps in existing literature and advancing the understanding of how fans interact with professional sports organizations.

1.4 Intentions and Research Questions

This research aims to uncover the factors driving fan engagement in professional sports and explore how predictive modeling can enhance loyalty and revenue growth. This approach examines both overarching trends across the fanbase and specific behaviors at the individual level. To better understand fan behavior, this research aims to address the following questions:

- Who are the fans? What are their demographic characteristics (e.g., age, gender, location)?
- How are fans acquired? What are the primary acquisition channels, and how does engagement vary across these channels?
- What spending patterns exist? How do ticket and merchandise purchases influence long-term loyalty and engagement?
- What is the nature of email campaign engagement? How do metrics like open rates correlate with broader outcomes like overall spend?

To identify future opportunities for fan engagement, predictive modeling was applied to analyze individual behaviors and engagement patterns. This modeling focuses on the following questions:

- Which fans are prime candidates for ramp-up? How can past financial and non-monetary behaviors signal opportunities for increased engagement?
- How can targeted strategies prevent churn and re-engage fans? What interventions can drive participation and sustain loyalty?

By addressing these questions, this thesis bridges historical and predictive insights to uncover broader trends in fan behavior while identifying opportunities for targeted interventions. This approach not only sustains fan loyalty but also highlights actionable strategies to ramp up engagement and drive long-term value among key segments of the fanbase.

1.5 Partnership and Anonymization

This research was developed in partnership with a professional sports organization, which provided access to proprietary datasets and unique insights into fan engagement within a

real-world context. This collaboration enabled the creation of methodologies tailored to the sports industry, offering practical applications for understanding fan behavior and enhancing engagement strategies.

To ensure confidentiality and adhere to legal obligations, the analysis, graphs, and attributes have been fully anonymized to protect the privacy of the organization and its fanbase. Despite this fact, the methodologies and findings remain relevant and applicable to other sports organizations and customer engagement contexts. By focusing on scalable frameworks and transferable concepts, this thesis delivers actionable strategies for strengthening fan relationships across diverse ecosystems.

1.6 Research Scope and Key Considerations

This research was developed with several key considerations and constraints, each of which shaped the scope and focus of the analysis to address the team's context and fanbase dynamics.

One important consideration is the team's recent venue change, which represents a pivotal shift in fan engagement patterns. This relocation likely influenced attendance and ticket purchasing behaviors, creating a natural starting point for this analysis. Additionally, the selected time period of interest—July 1, 2021, to June 30, 2024—captures a critical post-COVID era during which the sports industry was working to rebuild attendance and re-engage fans following the pandemic's disruptions. By focusing on this timeframe, the research aims to reflect how net new fans entered the team's ecosystem in a period of transition and recovery.

This thesis also places a deliberate emphasis on fan ramp-up, given the absence of robust legacy data for long-term fans. Instead of attempting to analyze decades-long behavior, the analysis centers on new fans who entered the sports ecosystem during the specified timeframe. This focus highlights opportunities to deepen engagement and foster loyalty among fans as they progress through various stages of the customer journey.

Unlike a traditional CLV approach, this analysis does not incorporate customer acquisition cost (CAC). When consulting business stakeholders, it was determined that CAC was not a key priority for this research. Instead, the focus remains on understanding fan engagement, spending patterns, and opportunities to intervene without factoring in the financial cost of acquisition.

At the same time, the analysis is shaped by certain data limitations. Demographic data, for example, is limited in scope and sourced primarily from external APIs or voluntary submissions. While it provides some context for understanding fan behaviors, it is excluded from predictive modeling due to concerns about data quality and coverage. Similarly, merchandise data reflects only online purchases, excluding in-arena and physical retail sales. Food and beverage spending is also incomplete, capturing only transactions linked to fan accounts, further limiting the analysis of in-venue spending patterns.

Challenges also arise with certain acquisition channels. For instance, SMS campaigns began in April 2023, resulting in a small sample size and limiting their representation in the analysis. Outdoor plaza along with food and beverage activations require authentication, which may understate their role as acquisition sources due to incomplete data capture.

These considerations are essential for contextualizing the findings presented in this thesis. While the analysis provides meaningful insights into fan behavior and engagement, these

constraints demonstrate the importance of refining and expanding the dataset to unlock additional opportunities for analysis. These limitations and potential future enhancements are revisited in Chapter 9, where recommendations for iterative improvements are discussed in greater detail.

Chapter 2

Data Overview

2.1 Data Sources and Structure

This analysis draws on nine datasets to comprehensively capture fan behavior across ticketing, attendance, digital channels, and online merchandise. Each dataset contributes unique insights into the customer journey, emphasizing both transactional and non-transactional interactions. This section provides an overview of the data sources, their origins, and the scope of their contributions.

2.1.1 Customer Data

The customer dataset offers a foundational view of fan behavior across multiple touchpoints. It links unique emails to engagement activities, including ticketing, merchandise purchases, email campaigns, Wi-Fi sign-ins, mobile app downloads, and more. This dataset is constructed by aggregating data from various source tables through a union operation, consolidating interactions across channels into a single view.

Each fan's initial entry into the team's ecosystem was determined by taking the earliest recorded engagement date across all touchpoints. Subsequent interactions are also recorded, enabling the construction of a dynamic timeline of customer engagement. This dataset focuses on net new customers—approximately 930,000 fans—who entered the ecosystem during the analysis period of July 1, 2021, to June 30, 2024. It captures both initial acquisition channels and progression into additional engagement touchpoints.

Demographic data complements the customer table by adding attributes such as age, gender, marital status, and address. However, these attributes rely on voluntary submissions and external API feeds, resulting in incomplete data. Therefore, online merchandise data was used to enrich locational fields, including zip code, city, and state, to mitigate this gap. Despite these efforts, demographic attributes were only included in uncovering directional overarching trends, but excluded from the models due to their sparsity and bias.

2.1.2 Ticketing Data

Ticketing data provides a comprehensive view of fan transactions across both primary and secondary markets. It captures ticket types (e.g., individual, mini-plans, and season tickets), seat locations, resale activity, and forwarded tickets.

- **Primary Market:** The primary market refers to ticket sales made directly by the organization through official sales channels, such as the team's website, box office, or its authorized ticketing partner.
- **Secondary Market Resale:** The secondary market resale involves tickets resold by the original purchaser to another party. While this can occur on various platforms, this analysis specifically focuses on resale activity facilitated by the team's ticketing partner.

- **Secondary Market Forwarded Tickets:** Forwarded tickets refer to transfers of purchased tickets to another individual without a resale transaction. These transfers allow fans to share tickets without an associated monetary value.

Both primary market data and secondary market activity are sourced from Ticketmaster. Forwarded tickets presented a unique challenge, as their monetary value was not explicitly recorded. Section 2.2 discusses the imputation technique employed to address this limitation.

2.1.3 Attendance Data

Attendance data bridges the gap between ticket purchases and in-venue participation, verifying whether tickets were scanned. This dataset distinguishes between customers who resell tickets as intermediaries and those who purchase with the intent to attend. Such differentiation, especially within the secondary market, provides valuable insights into varying levels of fan engagement. Attendance records, sourced from internal venue management systems, complement ticketing data by linking transactions to actual event participation.

2.1.4 Event Data

The event dataset includes all events held at the team's home venue, detailing event type, date, and time. This dataset was used to identify and filter all sports games that occurred within the time period of interest, excluding non-sporting events. This ensured the analysis focused on sports-related attendance and ticketing, aligning with the research objectives.

2.1.5 Online Merchandise Data

Online merchandise data captures digital purchasing behaviors, segmented by product categories such as apparel and accessories. Beyond spending patterns, as mentioned above, this data enriches locational attributes, filling demographic gaps in the customer table. Despite its limited scope to online sales, this dataset offers valuable insights into fan behaviors outside of ticketing activities, complementing the broader analysis.

2.1.6 Email Campaign Data

The email campaign datasets provide detailed metrics on customer interactions with email communications, including counts of delivered emails, unique opens, and unique clicks. The raw dataset, detailing individual campaigns for each customer, was used to create aggregated interaction metrics for each customer and season. Delivered, unique opens, and unique clicks were grouped into categories such as newsletters, merchandise promotions, post-purchase confirmations, event-related information, and other.

The raw engagement dataset was further processed to capture customers' entry points into the email ecosystem, integrating their first email engagement date with newsletters and/or merchandise as well as their ticketing and merchandise behavior. Customers were auto-opted into merchandise promotions if they purchased online merchandise prior to voluntarily opting in. Similarly, customers were auto-opted into newsletters if they purchased tickets or merchandise before voluntarily opting in. This processing identified customer intent, distinguishing between active engagement (voluntary opt-ins) and passive engagement (auto-opt-ins).

Subscription statuses and engagement data, including unique opens and unique clicks, were processed into a row-by-row customer view by season. This structured dataset allowed for dynamic, longitudinal analyses of customer participation and engagement, revealing how email interactions evolved as customers progressed through the ecosystem.

2.2 Caveats and Data Fixes

2.2.1 Lack of Representativeness in Demographics

One of the primary challenges in this analysis is the sparsity of voluntary demographic attributes, such as age, gender, marital status, and location information. These attributes rely on self-reported data or APIs that collect information on an opt-in basis, leading to incomplete coverage.

Improved Coverage

To address the locational gaps, billing information from customers who purchased online merchandise was incorporated. This enrichment significantly increased coverage for locational attributes such as country, state, city, and zip code, as shown in Table 2.1. For instance, zip code coverage rose to approximately 341,000 rows, covering approximately 36.7% of the dataset, while other locational attributes saw similar improvements. These enhancements provide a stronger foundation for analyzing location-based behaviors.

Attribute	Before Enrichment	After Enrichment	Improvement
Country	43k (4.6%)	336k (36.2%)	+31.6%
State	39k (4.2%)	306k (33.0%)	+28.8%
City	49k (5.3%)	305k (33.0%)	+27.7%
Zip Code	93k (10.0%)	341k (36.7%)	+26.7%

Table 2.1: Improvement in Locational Attributes Coverage Before and After Enrichment

While these improvements addressed gaps in locational attributes, demographic fields such as age and gender remain sparse, as they are not included in the online merchandise data. Additionally, the enrichment process introduced potential bias by over-representing customers who engage in online merchandise purchases, which limits the generalizability of findings.

Sample Size Estimation

To evaluate the dataset's adequacy for analyzing locational and demographic attributes, sample size estimation was performed. This ensures that the dataset contains a sufficient number of observations to accurately estimate population proportions within a specified confidence level and margin of error. The formula assumes an infinitely large population,

which is a reasonable approximation in this context. The required sample size is calculated using Cochran's formula for proportions:

$$n = \frac{z^2 \hat{p}(1-\hat{p})}{e^2}$$

where:

- n is the required sample size
- z is the Z-score corresponding to the desired confidence level ($z = 1.96$ for a 95% confidence level)
- $\hat{p}$ is the estimated proportion of the population (assuming maximum variability when $\hat{p} = 0.5$)
- e is the desired margin of error ($e = 0.05$ for a 5% margin)

Although the dataset contains 930,000 fans, the population size has little impact on the required sample size due to the stabilizing properties of large populations. For such datasets, the sample size remains consistent, and any adjustments from the finite population correction are negligible. Ultimately, this calculation yields a required sample size of 384 rows, which is easily exceeded by the dataset's locational and demographic attributes. However, the reliance on voluntary submissions for demographic data and enrichment through online merchandise records introduces biases, limiting the representativeness of the sample. Nevertheless, the dataset provides sufficient data for exploratory analyses despite these limitations.

The enrichment of locational attributes significantly improves data coverage, offering a more robust foundation for analysis. However, the sparsity of demographic attributes requires cautious interpretation of findings, particularly in generalizing results to the broader population.

2.2.2 Email Campaign Data Inconsistencies

Inconsistencies in email engagement metrics emerged as a notable challenge during the analysis. Manual examination revealed instances where the recorded number of email opens exceeded the number of emails delivered for certain customers. Similarly, there were cases where clicks outnumbered opens. These anomalies resulted in engagement rates exceeding 100%, although they were limited to a negligible subset of the dataset.

To address these issues, engagement rates for affected customers were capped at 100% to ensure all metrics remained within logical boundaries. By truncating these values, all customers remained included in the analysis, preserving the reliability of the dataset and preventing disproportionate influence from anomalous data points on overall engagement statistics.

The root cause of these discrepancies are not fully understood but are likely attributable to factors such as duplicate records, incorrect timestamps, or errors in system logging. Further investigation is required to identify and resolve the underlying issues.

2.2.3 Quantifying the Value of Forwarded Ticket

As mentioned in Section 2.1, forwarded tickets' monetary value was not explicitly recorded in the data. Table 2.2 summarizes the per capita food, beverage, and merchandise values attributed to forwarded tickets for each season. The sum of these values was used in the imputation technique to quantify the value of a forwarded ticket.

Season Year	**Food and Beverage ($)**	**Merchandise ($)**
2021-2022	23.24	11.74
2022-2023	26.35	13.46
2023-2024	29.73	12.39

Table 2.2: Per Capita Food, Beverage, and Merchandise Values by Season

This approach was important because the team did not want to make assumptions about how customers obtained their tickets—whether they paid their friend offline, received the ticket as a gift, or through other means—but recognized that attending the game still held value. Assigning a value of $0 would have underestimated the significance of their presence at the arena.

Chapter 3

Understanding the Beginning of the Customer Journey

3.1 Demographic and Locational Information

Understanding the demographic and geographic makeup of the customer base establishes a crucial foundation for examining engagement and spending behaviors. The following exploration highlights trends across age, gender, marital status, and geographic distributions.

The age distribution, depicted in Figure 3.1, demonstrates a broad appeal across demographics, with the largest cohort (35–44 years) representing 26.3% of the customer base. This is followed by 25–34 years (23.7%) and 45–54 years (19.6%). Notably, younger fans (less than 24 years old) account for 9.8%, while older age groups (55 and older) comprise 20.6%, respectively. These figures align with patterns typical in professional sports, where middle-aged individuals, who are often at peak earning potential and social aptitude, dominate both engagement and spending.

To address gender distribution, Figure 3.2 shows a nearly even split, with men representing 51% of the customer base and women 48.5%, while a small fraction (<1%) identify as non-binary or prefer not to disclose their gender. Marital status, illustrated in Figure 3.3, reveals a similarly balanced distribution: single customers make up 54% of the population, while married customers account for 46%. These proportions demonstrate the team's ability to connect with a diverse audience, underscoring the inclusivity of its fan engagement strategy.

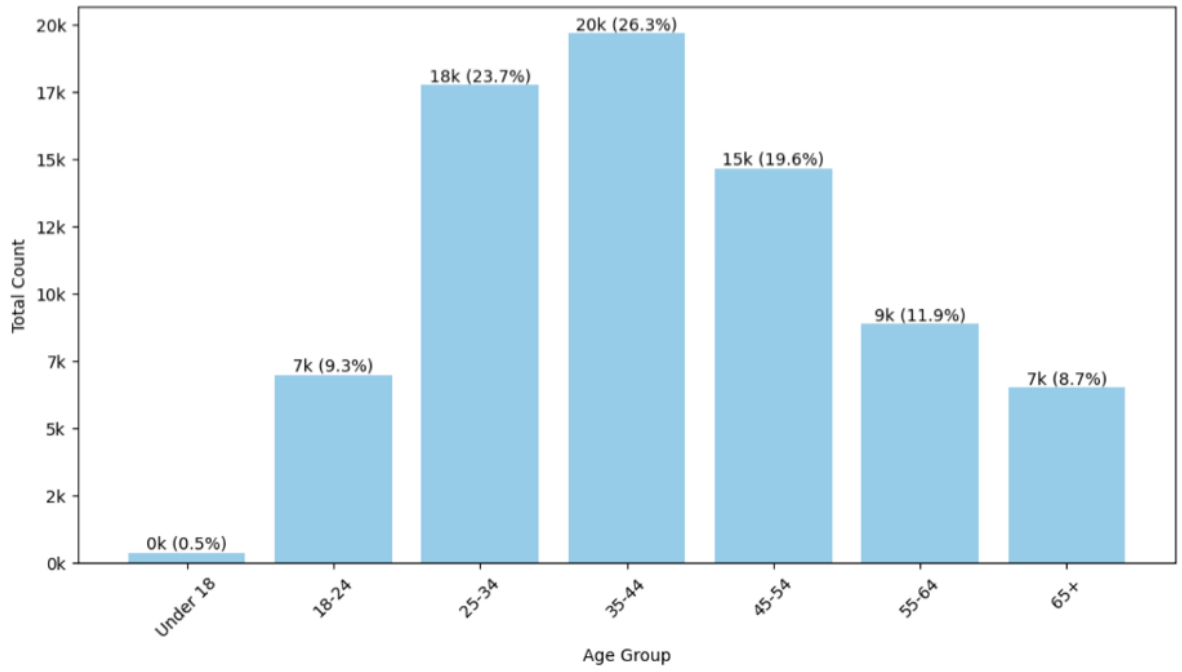

Figure 3.1: Age Distribution

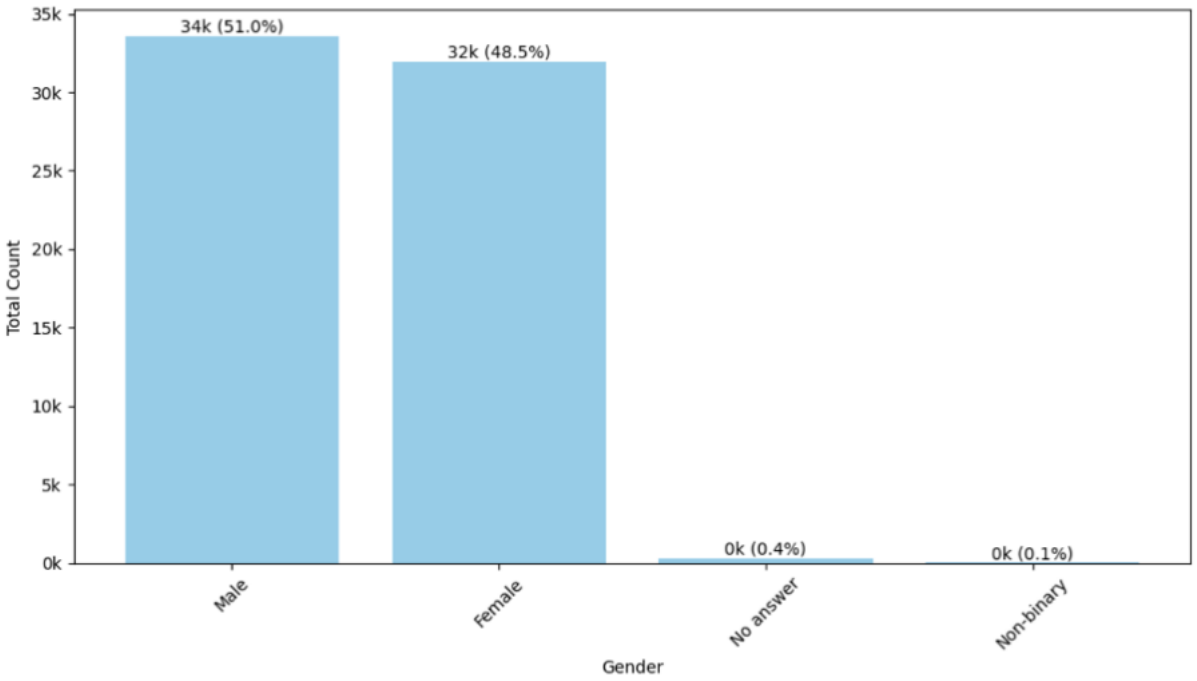

Figure 3.2: Gender Distribution

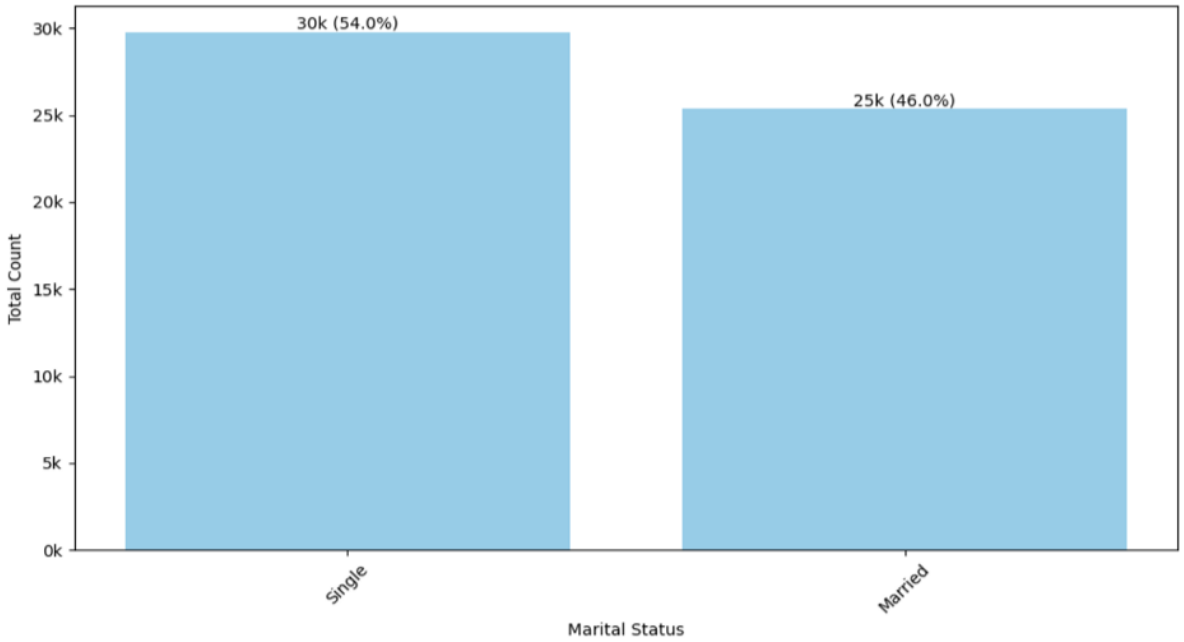

Figure 3.3: Marital Status Distribution

Geographic analysis, as shown in Figures 3.4 through 3.6, reveals a predominantly domestic fanbase, with 96.1% residing in the United States. International customers, though a smaller segment (3.9%), are primarily from Canada (1.3%) as well as countries such as Australia, Germany, and the UK. This small but notable international representation reflects its reach beyond domestic boundaries.

Domestically, while regional loyalty is evident, with 38.1% of customers residing in the team's home state (State A), there is significant representation from neighboring states and major cities across the nation (Figure 3-5). At the county level, acquisition is highest near the team's home area (Counties A through D), reflecting the localized nature of fan loyalty (Figure 3-6). While proprietary restrictions limit more specific analysis, these patterns underscore the team's strong local connections while hinting at potential for strategic outreach into broader markets.

These demographic and geographic insights collectively lay the groundwork for understanding how different customer segments engage with and contribute to the team's ecosystem.

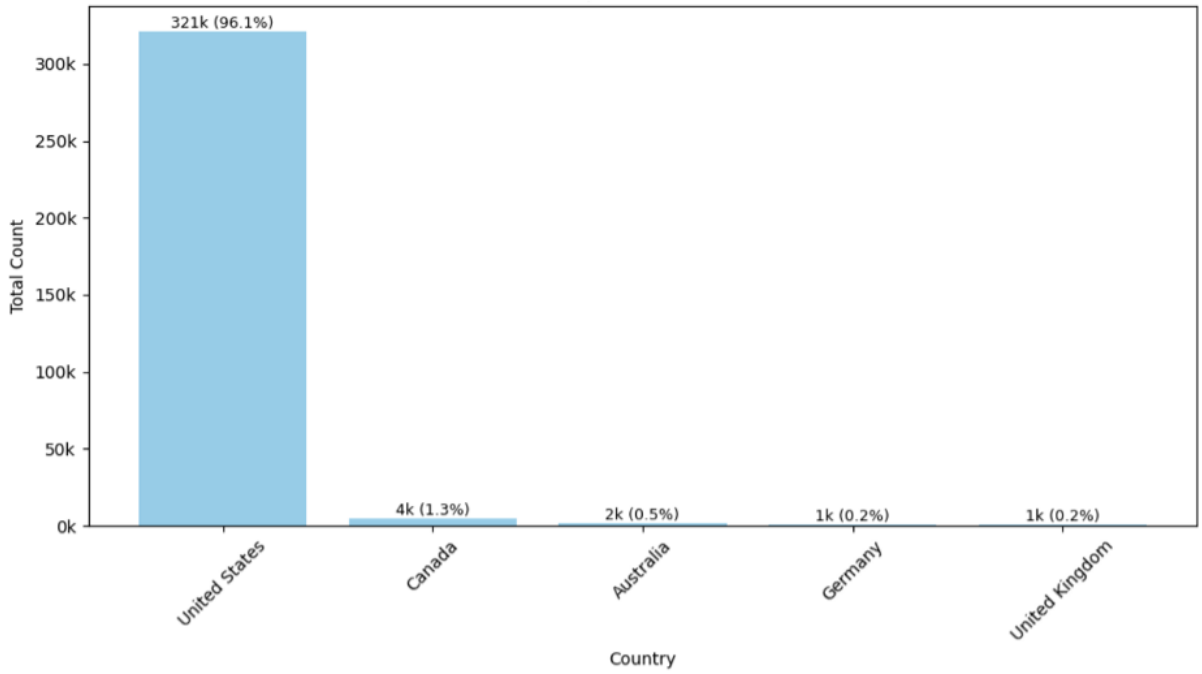

Figure 3.4: Top 5 Countries

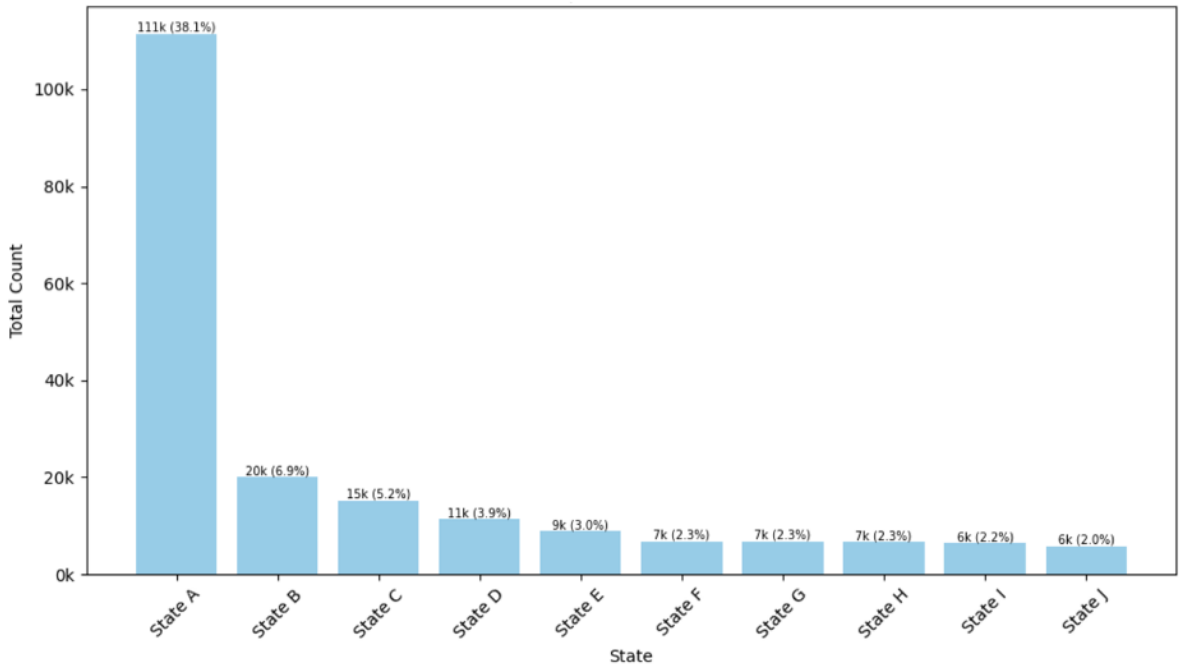

Figure 3.5: Top 10 US States

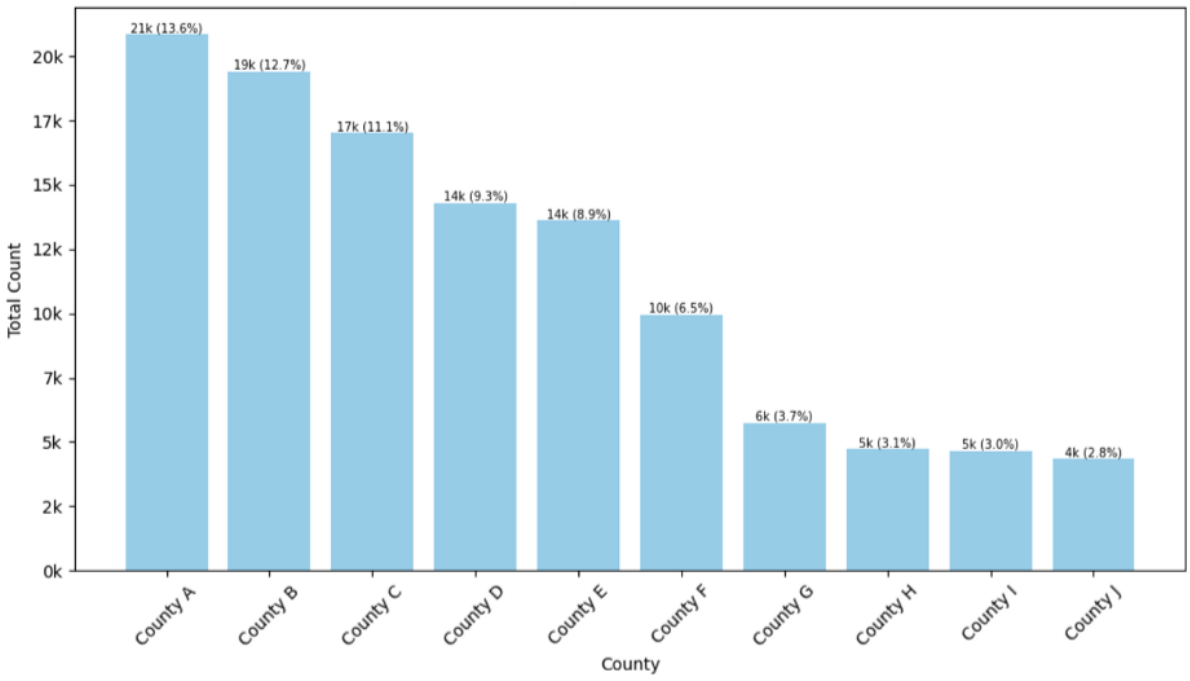

Figure 3.6: Top 10 Home-State Counties

3.2 Attended Events Beyond Sports

For this analysis, the focus is on net new customers who entered the sports segment of the database within the past three years. While the primary scope of the study is on customers' engagement with sports-related activities, examining whether their first recorded interaction with the venue was through a non-sporting event provides valuable context for understanding alternative entry points into the ecosystem. By doing so, this analysis sheds light on how the venue's broader entertainment offerings contribute to customer acquisition for the sports side of the business.

The data revealed that 6.2% of these net new customers engaged with the venue through a non-sporting event at some point. Of this subset, 54.5% attended a non-sporting event prior to their first sports-related activity. This highlights the role of non-sporting events as a potentially meaningful entry point for customers who might not otherwise have engaged with the sports ecosystem directly. These findings suggest that the venue's appeal as a broader entertainment destination helps to expand the pool of potential sports customers.

To better understand these dynamics, Figure 3.7 illustrates the distribution of non-sporting event attendees based on their first recorded interaction with the sports side of the business. While distinct groups exist within the venue customer base, there is some overlap in the proportion of individuals who engage with both sports and non-sports offerings. This reinforces the idea that non-sporting events do not just operate independently but also serve as complementary pathways for introducing new audiences to sports-related activities.

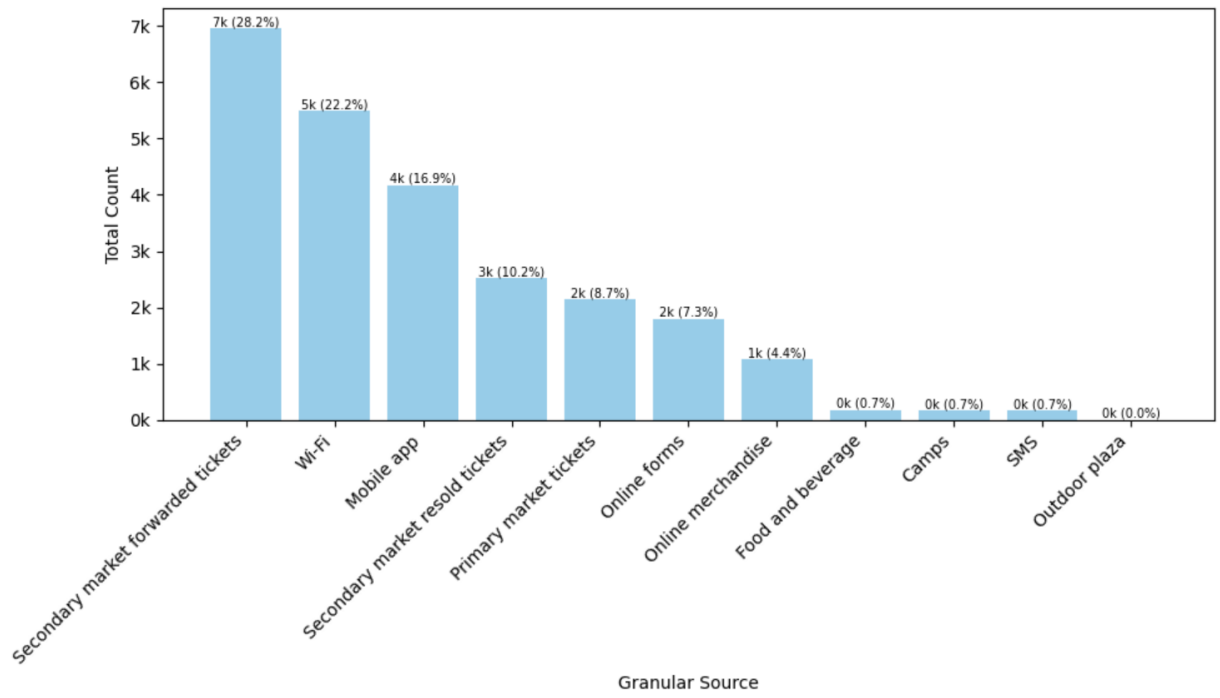

Figure 3.7: First Touch-Point for Fans with Prior Non-Sporting Event Attendance

3.3 Engagement Channel Information

3.3.1 Acquisition Channel Overview

This distribution underscores the diversity of entry points into the customer ecosystem, each playing a pivotal role in fostering engagement. As illustrated in Figure 3.8, ticket-related channels dominate, constituting nearly 40% of total acquisitions, and online merchandise sales contribute to roughly 30% of acquisitions.

Although smaller channels such as Wi-Fi, camps, and SMS contribute a less significant share to the total, their value lies in capturing niche segments that might otherwise remain outside the ecosystem. These channels serve as touchpoints for acquiring a diverse range of customers and reaching groups that are typically harder to engage.

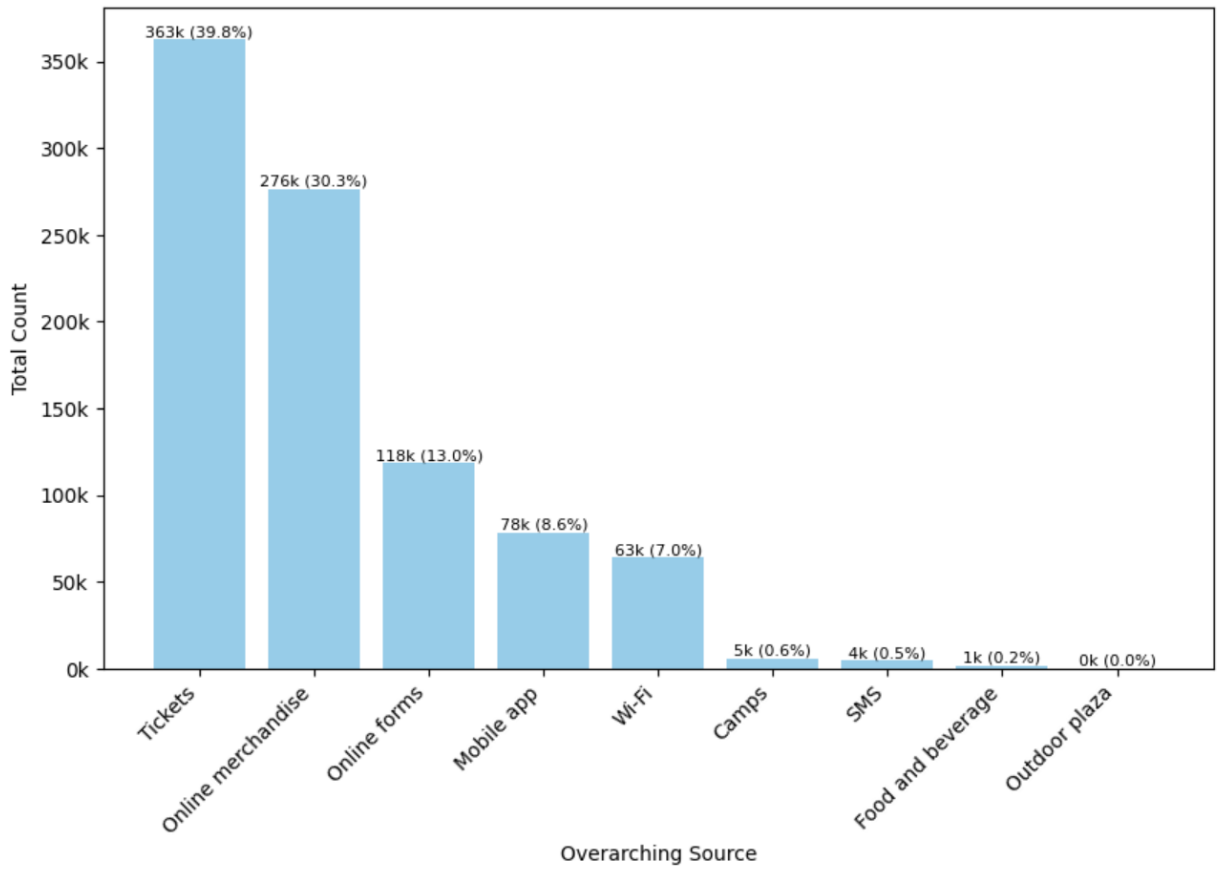

Figure 3.8: Overarching Acquisition Channel Distribution

3.3.2 Customer Acquisition Trends and Channel Dynamics

The temporal dynamics of customer acquisition exhibit pronounced seasonality, as reflected in the trends shown in Figure 3.9. Peaks in net new customers are particularly evident in Q2 of most fiscal years, coinciding with heightened fan enthusiasm driven by the team's competitive performance. During periods of success, such as playoff runs, the organization observes a tangible uptick in acquisitions, likely due to increased engagement through ticket purchases, merchandise sales, and broader fan involvement. Off-seasons or periods reflect noticeable declines, speaking to the inherent sensitivity of customer acquisition to performance and seasonal factors.

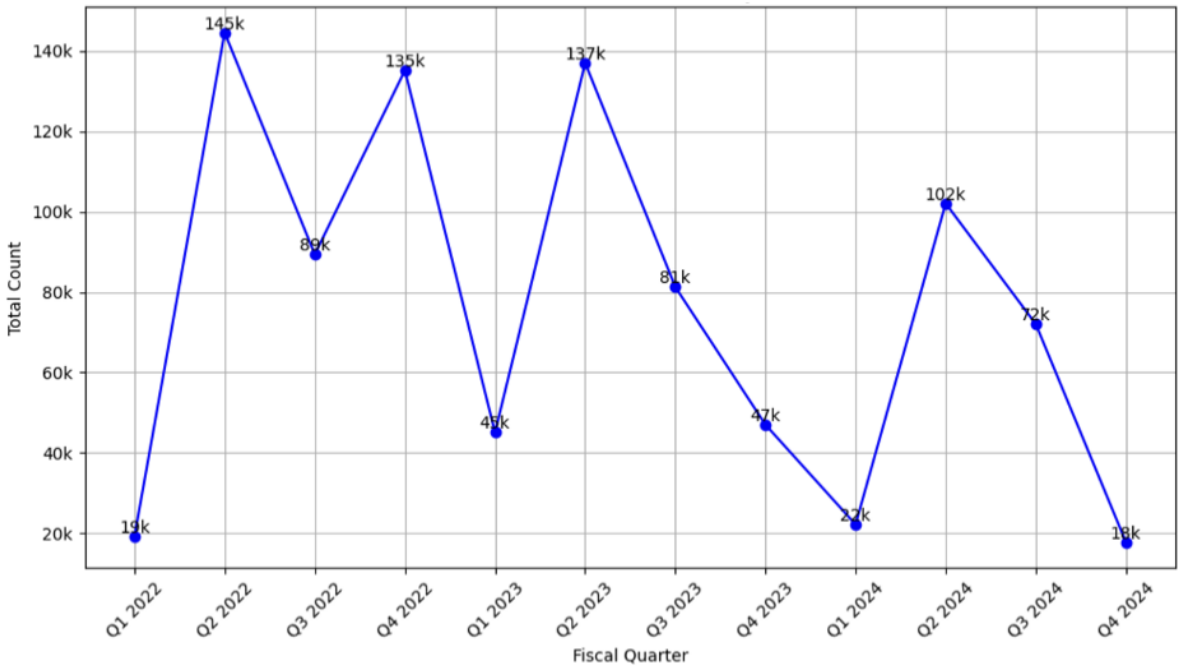

Figure 3.9: Volume of Fan Acquisition by Quarter

Figure 3.10 delves deeper into acquisition channels, illustrating their relative contributions during these periods of flux. Ticket sales consistently emerge as the predominant acquisition pathway, particularly during high-engagement periods. This pattern shows the central role of live game attendance in driving initial customer interactions. Meanwhile, online merchandise sales show a marked increase during specific quarters, likely driven by events such as holiday shopping seasons or championship successes that galvanize fan excitement.

Conversely, digital channels such as online forms and mobile apps demonstrate contributions throughout the entirety of the fiscal year, positioning themselves as complementary acquisition pathways. These channels enable ongoing engagement, especially during the off-season when direct interactions like ticket sales decline.

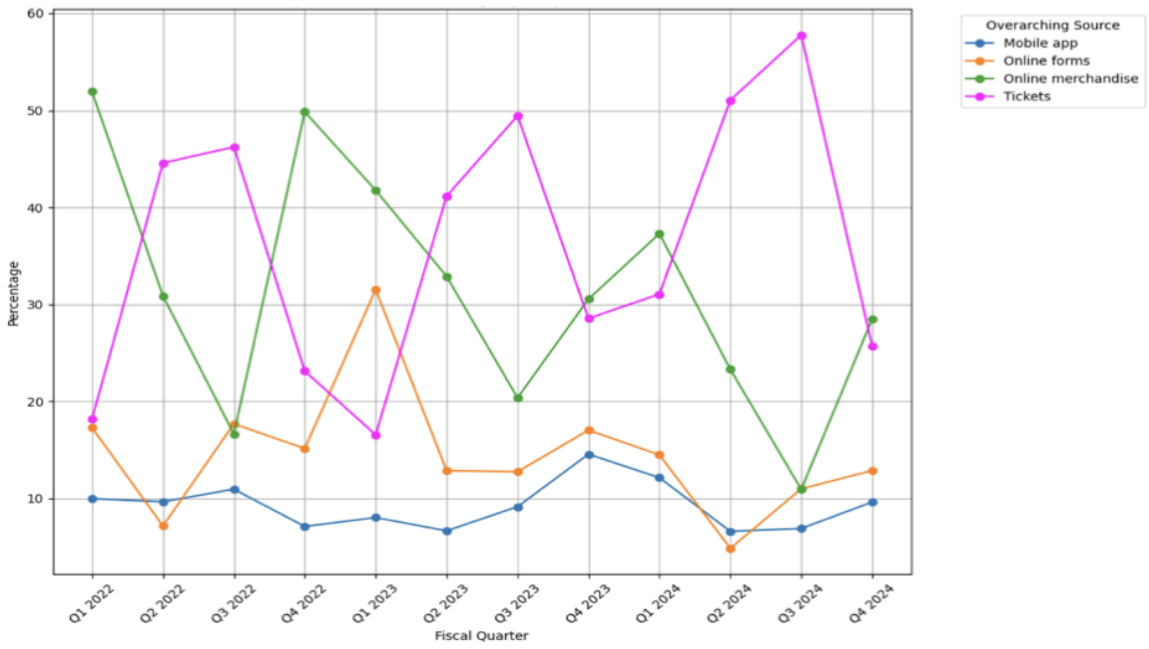

Figure 3.10: Percentage of Fans by Top Acquisition Channels

These findings highlight the importance of tailoring acquisition strategies to align with seasonal dynamics and channel-specific strengths. By leveraging the unique capabilities of each channel and aligning them with the broader fan journey, the organization can optimize both peak and off-peak acquisition efforts, enhancing overall engagement and retention.

3.3.3 Engagement Channels per Customer

As shown in Figure 3.11, most customers engage with only one or two channels, with 80.1% interacting through a single channel and another 14% through two. Beyond this, engagement falls off sharply—only 2.9% of customers use three channels, and engagement across four or more channels is extremely rare. This pattern suggests that while most customers engage at a basic level, multi-channel participation is relatively uncommon.

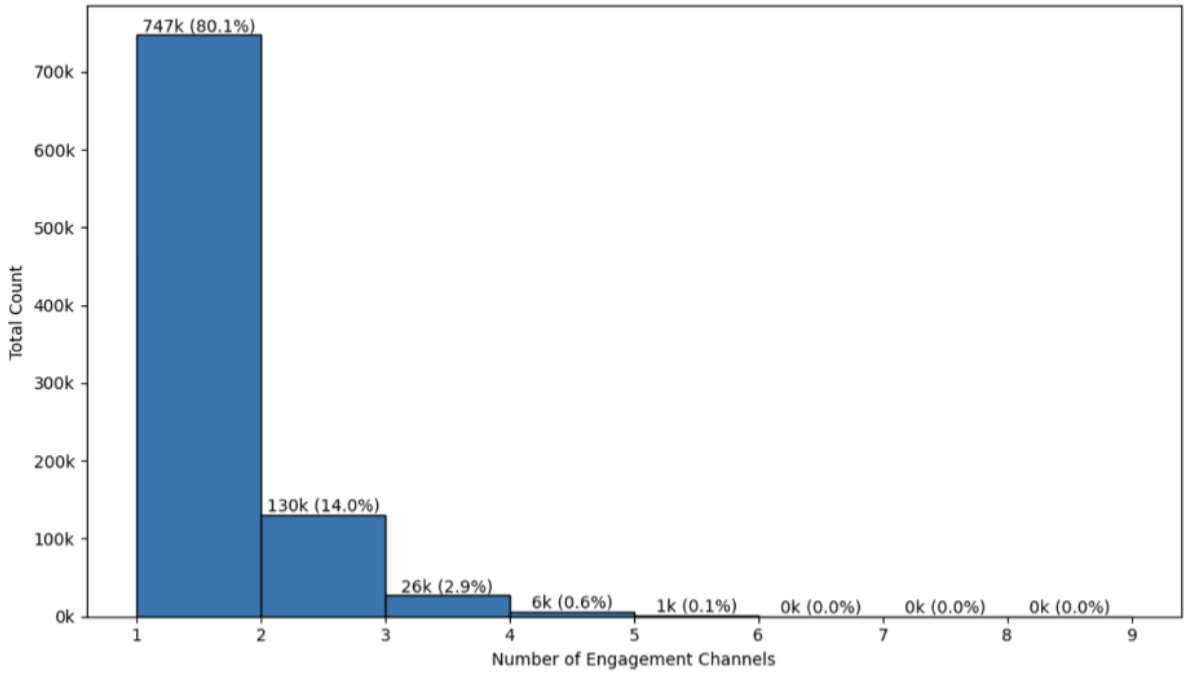

Figure 3.11: Distribution of Customers by Number of Engagement Channels

3.3.4 Sequence of Engagements

The sequence of engagements refers to the chronological order in which customers interact with various channels, such as purchasing tickets, accessing Wi-Fi, or using a mobile app. Customers using two to three channels exhibit distinct and consistent engagement patterns over the three-year period. Figure 3.12 shows that "Tickets, Wi-Fi" and “Tickets, Mobile App” account for 43.2% of the observed interactions. This may reflect a typical game-day behavior, such as a customer looking to access digital tickets or to improve their in-stadium experience.

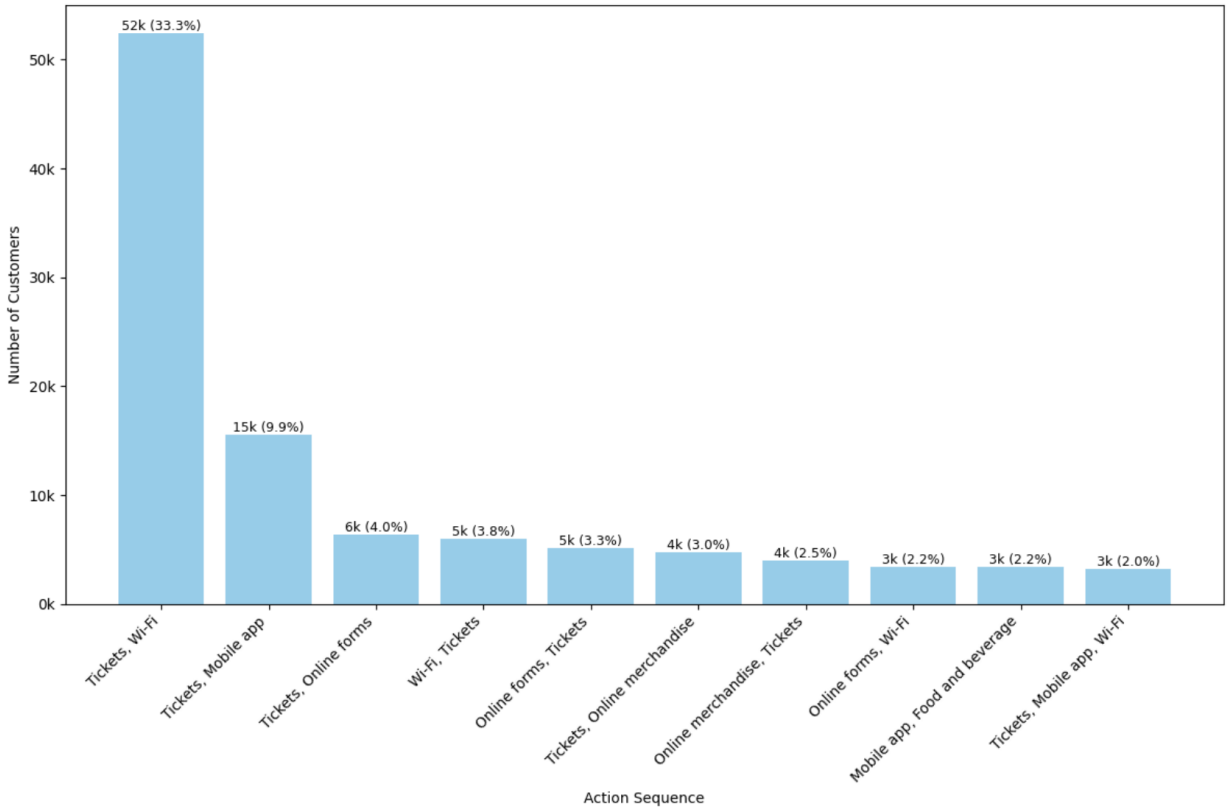

Figure 3.12: Top 10 Engagement Sequences for Customers Using 2–3 Channels

Shorter gaps—such as those seen in "Mobile App, Food and Beverage" or "Tickets, Mobile App" (0 days or 1 day in between, respectively)—reflect immediate, game-day activities like concession ordering or ticket use. These timing gaps were calculated by determining the median number of days between the first engagement and the subsequent engagement. In contrast, sequences like "Tickets, Online Forms" exhibit a noticeably longer median delay of 75 days, indicating engagement types that are less tied to game-day activities and more likely related to follow-up actions or longer-term behaviors.

The heat map in Figure 3.13 highlights how the popularity ranking of these sequences changes across fiscal quarters, with "1" being the most popular and "5" being the least popular. Sequences like "Tickets, Wi-Fi" consistently rank as the most popular, maintaining a "1" ranking across all quarters. "Tickets, Mobile App" also ranks highly, frequently appearing in the top three. In contrast, sequences such as "Tickets, Online Forms" and "Online Forms, Tickets" show greater variation in their rankings, ranging from "2" to "5" depending on the quarter, reflecting less consistency compared to the more stable patterns seen in "Tickets, Wi-Fi."

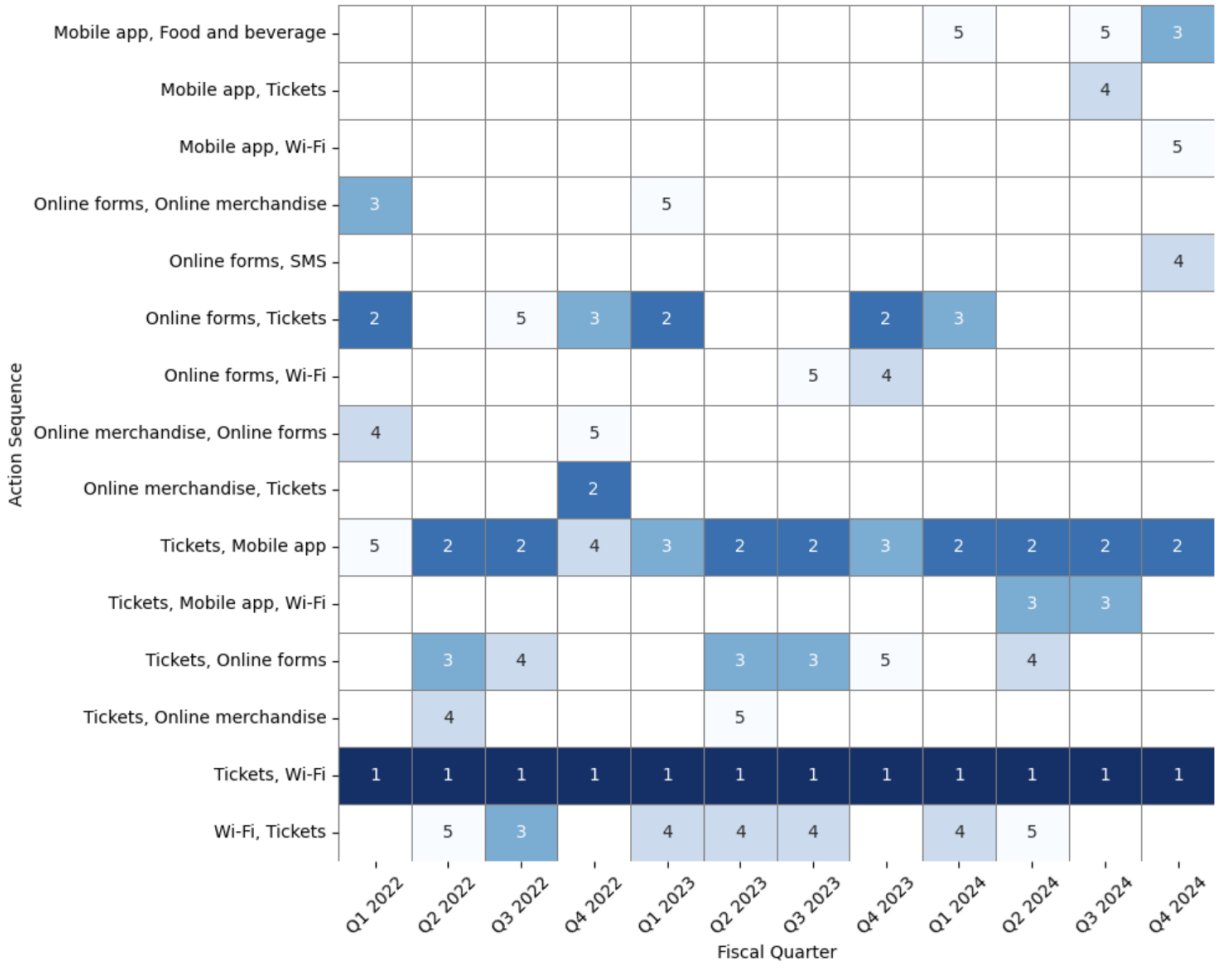

Figure 3.13: Ranking of Top Engagement Sequences by Fiscal Quarter

Chapter 4

Intent Metrics

Aggregate spending data was analyzed to uncover key insights into customer behavior and engagement patterns. This section begins by examining online merchandise spending trends, with a focus on understanding "intent metrics." Intent metrics, defined as spending-related indicators such as ticket purchases and online merchandise sales, provide a direct measure of customer engagement and financial commitment. Next, ticketing behavior is explored, including acquisition channels, first-year activity, and multi-year lifecycle patterns, offering a comprehensive view of customer interactions over time.

4.1 Merchandise Spend

Analyzing merchandise spending behavior provides insights into how customers engage with the ecosystem beyond ticket purchases. The analysis explored two dimensions of spending patterns: revenue by acquisition channel and category-level preferences, using winsorized averages to minimize the impact of outliers.

Figure 4.1 highlights the median merchandise revenue by source. The outdoor plaza has the highest median revenue ($155), followed by online forms ($108), camps ($105), food and beverage ($104), and SMS ($104). Ticketing channels and Wi-Fi generated median revenues of $98 and $89, respectively, while the mobile app and online merchandise have the lowest medians, at $85 and $81, respectively.

However, channels such as outdoor plaza, camps, and food and beverage represent smaller sample sizes, which warrants a cautious interpretation of these findings. Similarly, ticketing channels require additional consideration: merchandise purchases captured in this dataset reflect online activity only, potentially underrepresenting the full spending behavior of ticketing customers who may prefer to shop in-person at events or team stores. As noted in Chapter 1, further exploration of offline spending is necessary to provide a comprehensive understanding of this group.

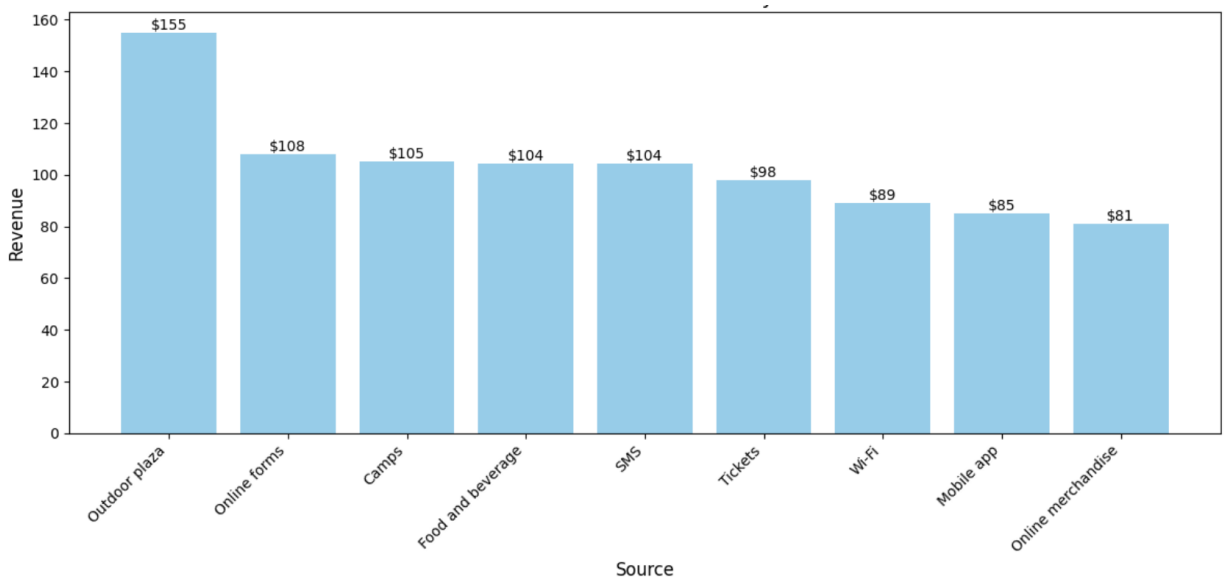

Figure 4.1: Median Online Revenue by Overarching Source

To further explore merchandise spending, Figure 4.2 examined the average number of items purchased in each product category, using winsorized data to account for outliers. This adjustment was necessary to manage extreme cases, such as individual customers purchasing an unusually high number of jerseys, which could otherwise skew the results. The heat map shows that apparel is the most frequently purchased category across all acquisition sources, with customers buying a range of 1 to 2.4 items. Jerseys and accessories follow as secondary categories, showing steady averages across acquisition channels.

Customers who enter through online merchandise directly tend to purchase fewer items per transaction. This trend may reflect the nature of online shopping, where transactions often focus on specific items rather than larger bundles.

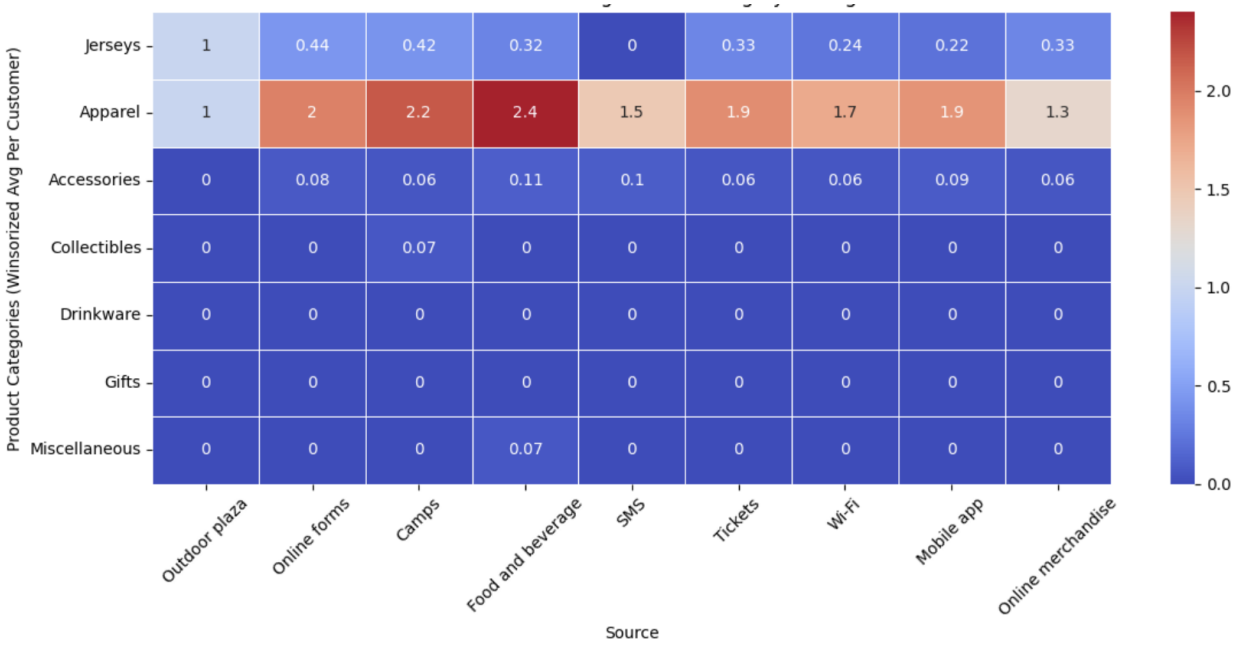

Figure 4.2: Winsorized Overarching Product Category Average Items Bought

4.2 Ticket Behavior and Lifecycle Analysis

4.2.1 Acquisition Channel Distributions

The data offers insight into customer entry points into the ticketing ecosystem, with Figure 4.3 highlighting the distribution of acquisition channels. The top three sources—secondary resale or forward activities, primary purchases, and forwarded tickets—dominate. Following these, online forms appear as the fourth most significant channel, which includes activities such as presale ticket promotions through newsletters and digital campaigns.

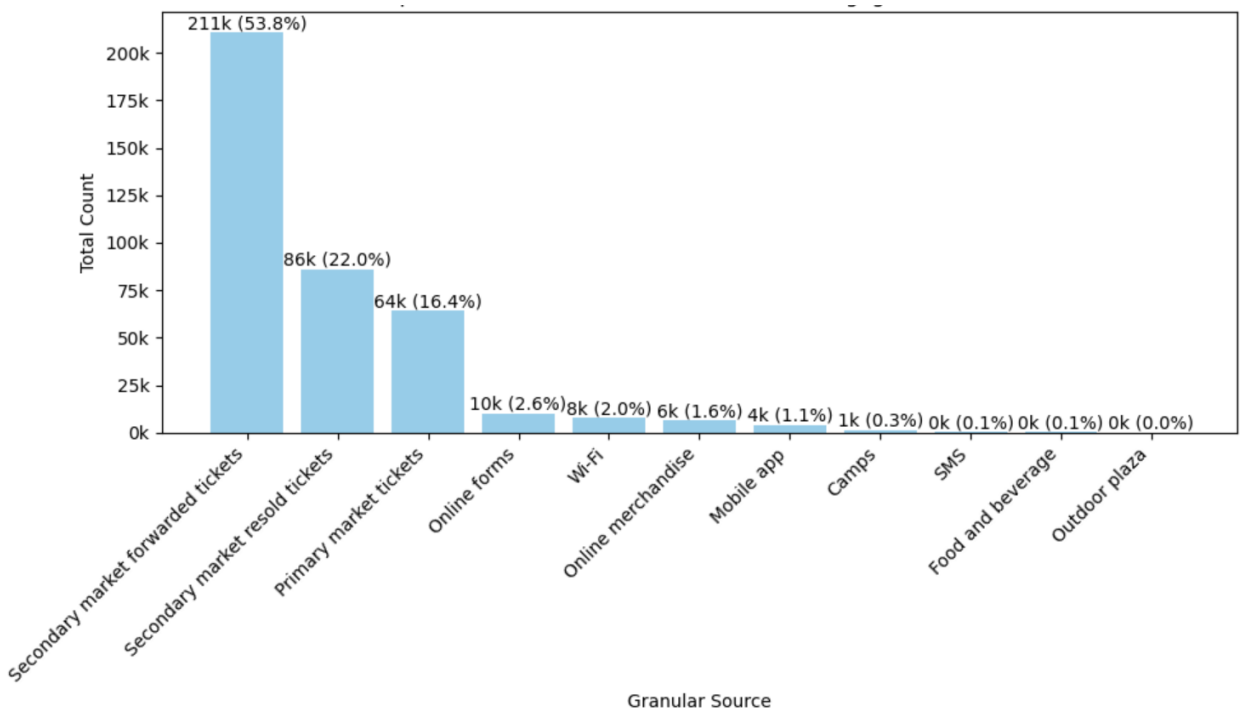

Figure 4.3: Acquisition Channel Distribution for Ticket-Engaged Fans

Revenue patterns by acquisition channel, shown in Figure 4.4, provide insights into ticketing behavior based on median values, highlighting the varied ways customers engage with the ticketing ecosystem. Secondary resale channels generate the highest median ticket revenue at $760, reflecting the higher prices typical of the secondary market, where supply and demand drive costs above the more controlled pricing of the primary market.

Primary market channels have a median ticket revenue of $375, reflecting the organization's structured pricing strategy. These controlled prices aim to maintain accessibility for fans, particularly during regular-season games to balance affordability with revenue generation.

Forwarded tickets, with a median revenue of $175, capture the indirect contributions of attendees who receive tickets through redistribution. Smaller channels, such as mobile apps and SMS, show consistent median ticket revenues of $134–$140, contributing less to overall revenue due to lower transaction volumes.

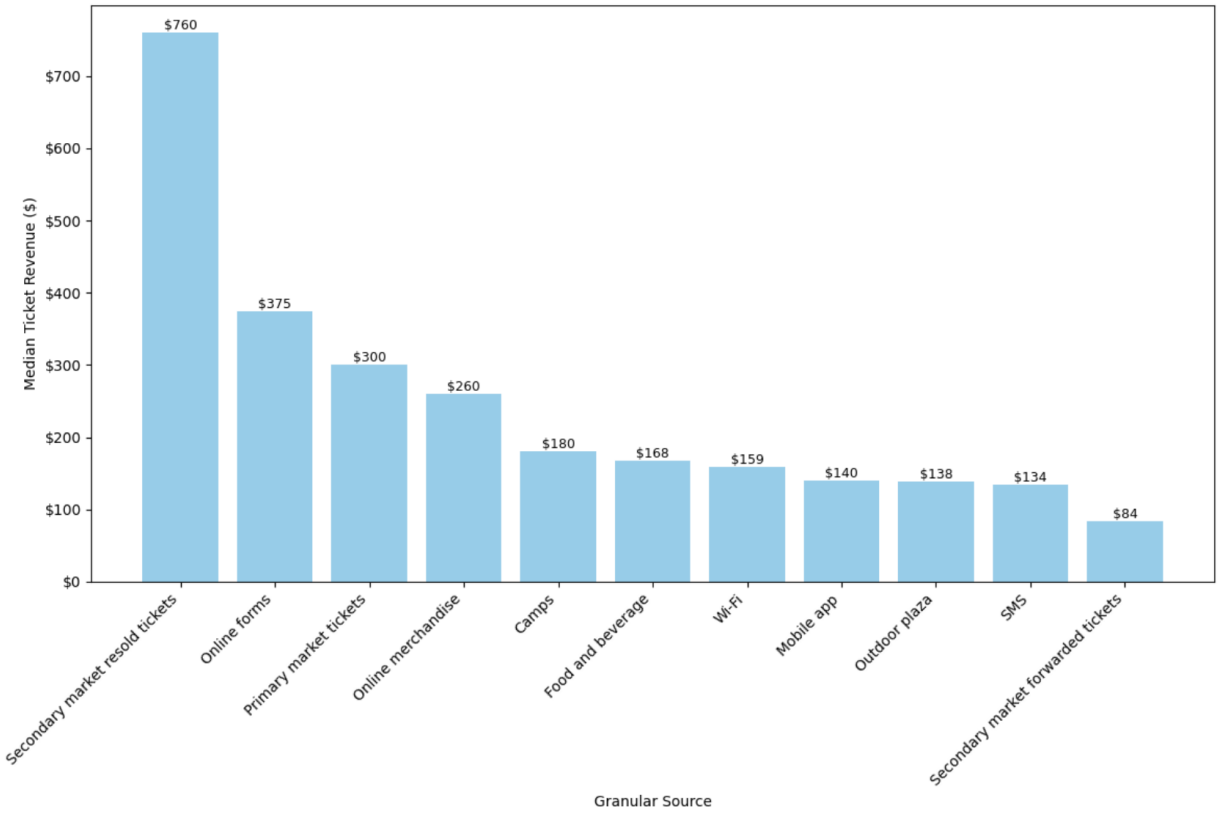

Figure 4.4: Median Ticket Revenue by Granular Source

Mean attendance percentages across acquisition channels, as shown in Figure 4.5, demonstrate consistently high levels of event participation, highlighting the effectiveness of ticketing efforts in driving attendance. To account for the distribution of the data, where a long tail of lower attendance values skews the averages downward, the decision to focus on mean attendance percentages rather than medians was made. The mean provides a more detailed comparison across channels, and it reveals subtle variations in attendance that median values might overlook.

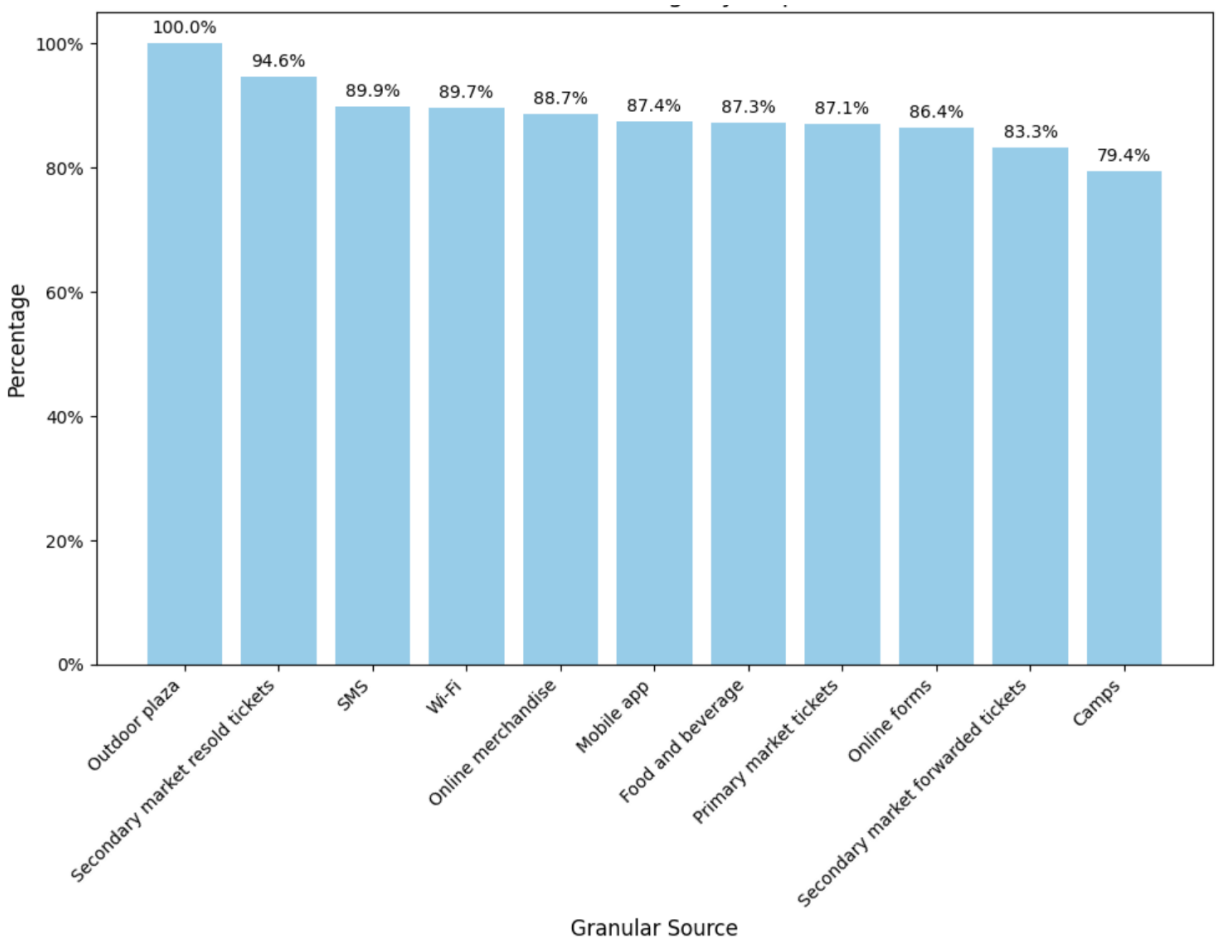

Figure 4.5: Mean Attendance Percentage by Granular Source

4.2.2 First-Year Ticket Activity

In the past three years, 96.1% of customers in our population who engaged with tickets did so within their first year of joining. This highlights the first year as a pivotal window for introducing customers to ticketing and establishing a lasting connection with the brand.

Figure 4.6 breaks down the types of ticket market activities customers participated in during their first year. The data shows that most customers engaged with just one type of activity—either forwarded, resold, or primary tickets—indicating that customer behavior in the ticket market is predominantly single-channel, with little overlap between these categories.

Forwarded ticket activity was the most common, with 51.5% of customers falling into this group. While these customers may not have purchased their tickets directly, their attendance at events creates opportunities to further integrate them into the brand ecosystem. This group represents a chance to convert attendees into paying customers by leveraging their exposure to the arena experience.

Following this, 25.1% of customers engaged exclusively in secondary resale activity, suggesting a meaningful reliance on resale platforms for ticket access. Another 13.1% made primary ticket purchases, representing a smaller but crucial cohort of customers engaging

directly through official sales channels.

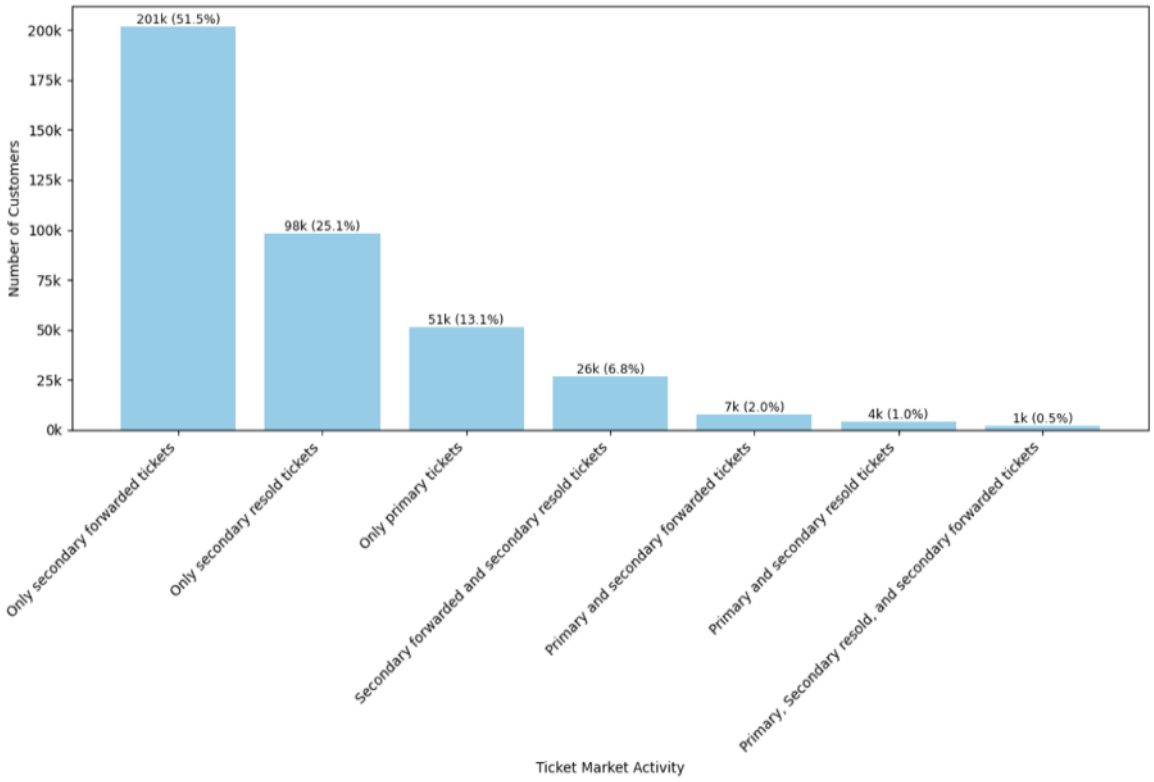

Figure 4.6: Distribution of Ticket Market Activity in First Season

4.2.3 Customer Lifecycle and Revenue Transitions

The analysis of customer ticket transitions between seasons provides critical insights into patterns of churn, inactivity, upgrades, downgrades, and retention. By leveraging both transition matrices and bar graphs, this section aims to classify and quantify the nature of these movements, providing a comprehensive view of customer lifecycle stages and revenue dynamics.

This analysis uses five terms to define fan behavior: churn, upgrades, downgrades, inactivity, and no change. These classifications are derived from logical comparisons of revenue across seasons, allowing for nuanced insights into fan behavior and engagement over time:

- *Churn* is defined as fans who did not generate revenue in the prior season and continue to generate no revenue in the current season.
- *Upgrades* refer to fans who increase their spending compared to the prior season or transition from no revenue to generating revenue in the current season.
- *Downgrades* describe fans whose revenue has declined compared to the previous season but who remain active by maintaining some level of spending.

- *Inactivity* captures fans who previously generated revenue but have ceased all revenue contributions during the current season (i.e., prior season's revenue is positive and current season's revenue is equal 0).

- *No change* refers to fans who sustain consistent revenue levels across consecutive seasons.

Transition matrices were used to illustrate multi-season fan engagement dynamics for those acquired in 2021–2022 season, capturing transitions from 2021–2022 to 2022–2023, and from 2022–2023 to 2023–2024. These diagrams highlight pathways such as churned customers re-entering the ecosystem or consistently upgrading customers sustaining their trajectory over multiple seasons.

The transition matrix in Table 4.1, based on a population of approximately 160,000 fans, provides critical insights into fan behavior across two consecutive periods for primary ticket and secondary resale data. Churn-churn and inactive-churn dominate the transitions, with 95.4% of fans who churned between 2021–2022 and 2022–2023 remaining churned in the subsequent period; only 4.6% of this group upgraded by generating revenue in 2023–2024. Similarly, among fans classified as inactive in the first period, 95.8% then churned, with just 4.2% upgrading. Fans who downgraded showed slightly more variation, as 12.8% churned, 70.8% remained inactive, and 16.1% upgraded their revenue levels in the following season. For fans that had no change from 2021-2022 to 2022-2023, 88.9% became inactive, 6.7% continued to be no change, and 2.2% upgraded in the next transition. Fans who upgraded displayed poor retention potential, with 78.4% becoming inactive in the second transition period.

	2022–2023 to 2023–2024				
2021–2022 to 2022–2023	**Churn**	**Downgrade**	**Inactive**	**No change**	**Upgrade**
Churn	95.4	0.0	0.0	0.0	4.6
Downgrade	0.0	12.8	70.8	0.3	16.1
Inactive	95.8	0.0	0.0	0.0	4.2
No change	0.0	2.2	88.9	6.7	2.2
Upgrade	0.0	13.2	78.4	0.1	8.3

Table 4.1: Transitions of Primary and Secondary Ticket Resale Revenue for Fans Beginning in 2021–2022

Table 4.2 provides an overview of transitions within the secondary forwarded tickets group, comprising approximately 125,000 fans. Churn remains the most prominent transition, with 92.6% of fans who churned in the first period continuing to churn in the second period. Similarly, the inactive group demonstrates a high retention of inactivity, with 94.4% of these fans starting as inactive then churning. 26.3% of fans who downgraded in the first period upgraded their spending in the second period. Additionally, 16.9% of fans who upgraded in the first period continued to upgrade into the second period.

	2022–2023 to 2023–2024			
2021–2022 to 2022–2023	**Churn**	**Downgrade**	**Inactive**	**Upgrade**
Churn	92.6	0.0	0.0	7.4
Downgrade	0.0	11.5	62.2	26.3
Inactive	94.4	0.0	0.0	5.6
Upgrade	0.0	13.2	69.9	16.9

Table 4.2: Transitions of Secondary Forwarded Tickets for Fans Beginning in 2021–2022

For customers who joined in 2022–2023, their transition from their first season to the next reinforced the observations from the transition matrices above. Figures 4.7 and 4.8 emphasize churn and inactivity for both primary and secondary ticket resale revenue as well as secondary forwarded ticket revenue.

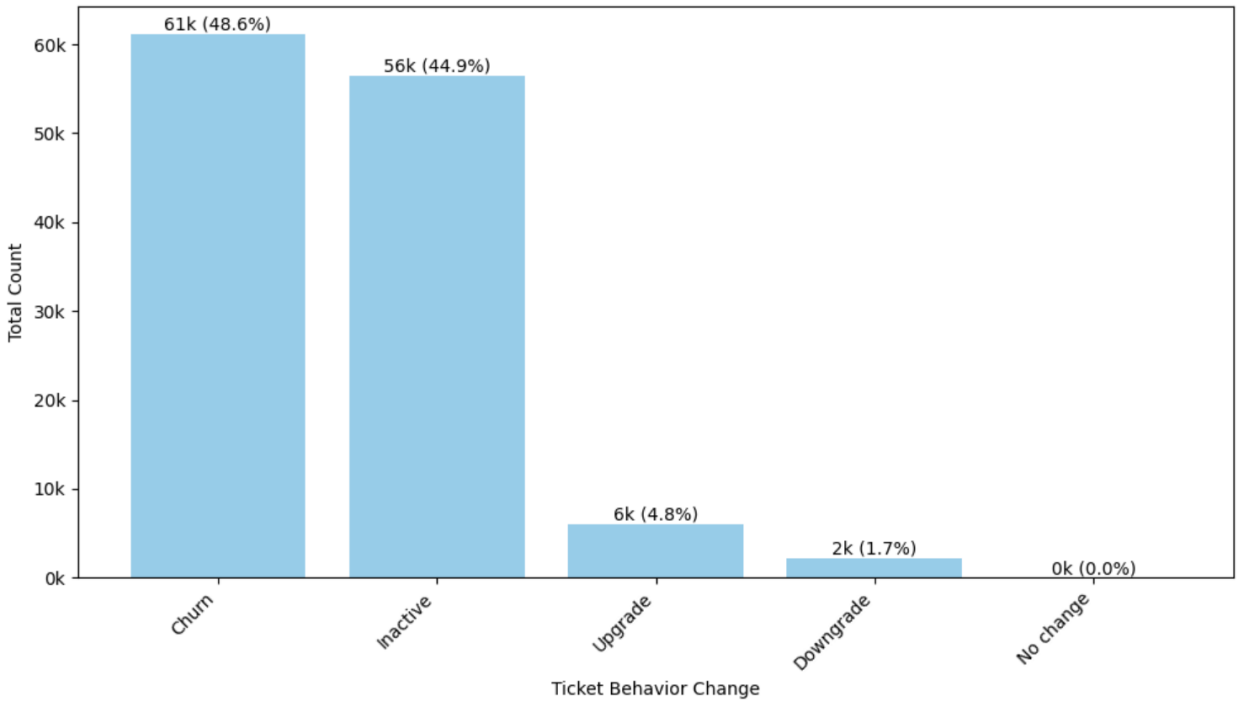

Figure 4.7: Transition of Primary and Secondary Ticket Resale Revenue for Fans Beginning in 2022-2023

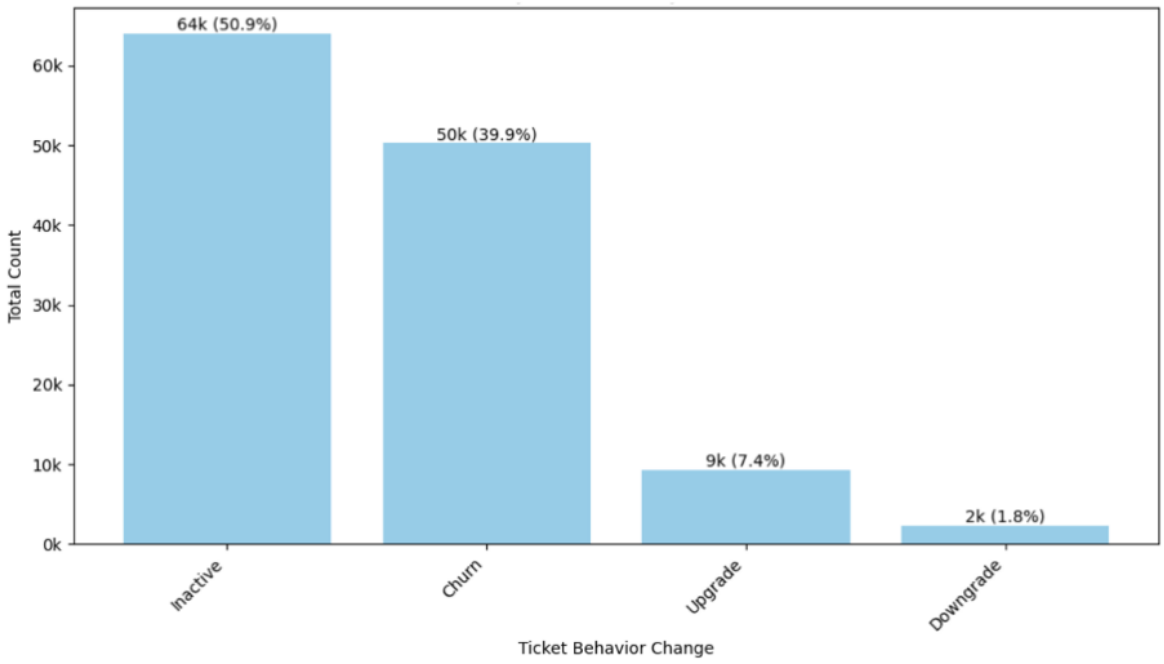

Figure 4.8: Transition of Secondary Forwarded Ticket Revenue for Fans Beginning in 2022-2023

Chapter 5

Email Campaign Metrics

5.1 Distribution of Email Engagement Metrics

To explore patterns in email engagement, key metrics such as deliveries, unique opens, and unique clicks were analyzed at both the individual and aggregate levels. The data were grouped by season and displayed using boxplots to highlight distributional characteristics. Because raw counts tended to exhibit skewness and limited variance, a logarithmic transformation was applied to compress extreme values while preserving meaningful variation.

Figure 5.1 illustrates the distribution of email deliveries by campaign type. Newsletters emerged as the main communication method, demonstrating the highest median delivery count and the smallest interquartile range. This reflects their role in disseminating regular updates while also serving transactional purposes, such as promoting pre-sale and playoff opportunities. Merchandise and event-related emails, on the other hand, displayed greater variability in delivery volume, likely tied to their more targeted and behavior-specific content.

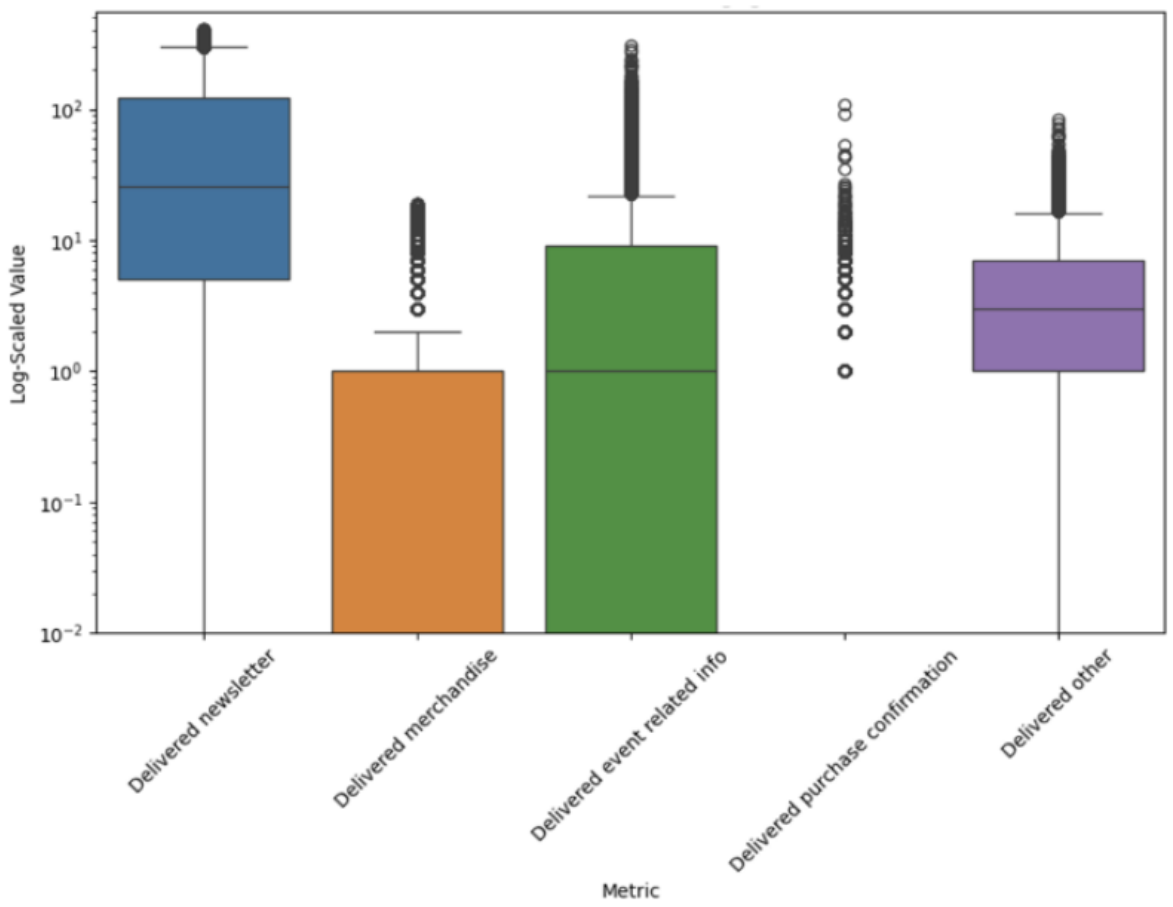

Figure 5.1: Distribution of Delivered Email Engagement Metrics

The distribution of unique opened emails, shown in Figure 5.2, further emphasizes the importance of newsletters. These emails consistently engaged a broad audience, evidenced by their high median and relatively narrow interquartile range. Conversely, the "other" category showed broader variability, a reflection of its role in encompassing diverse and occasionally niche content. Merchandise and event-related emails also exhibited higher variability in opens, highlighting their selective appeal.

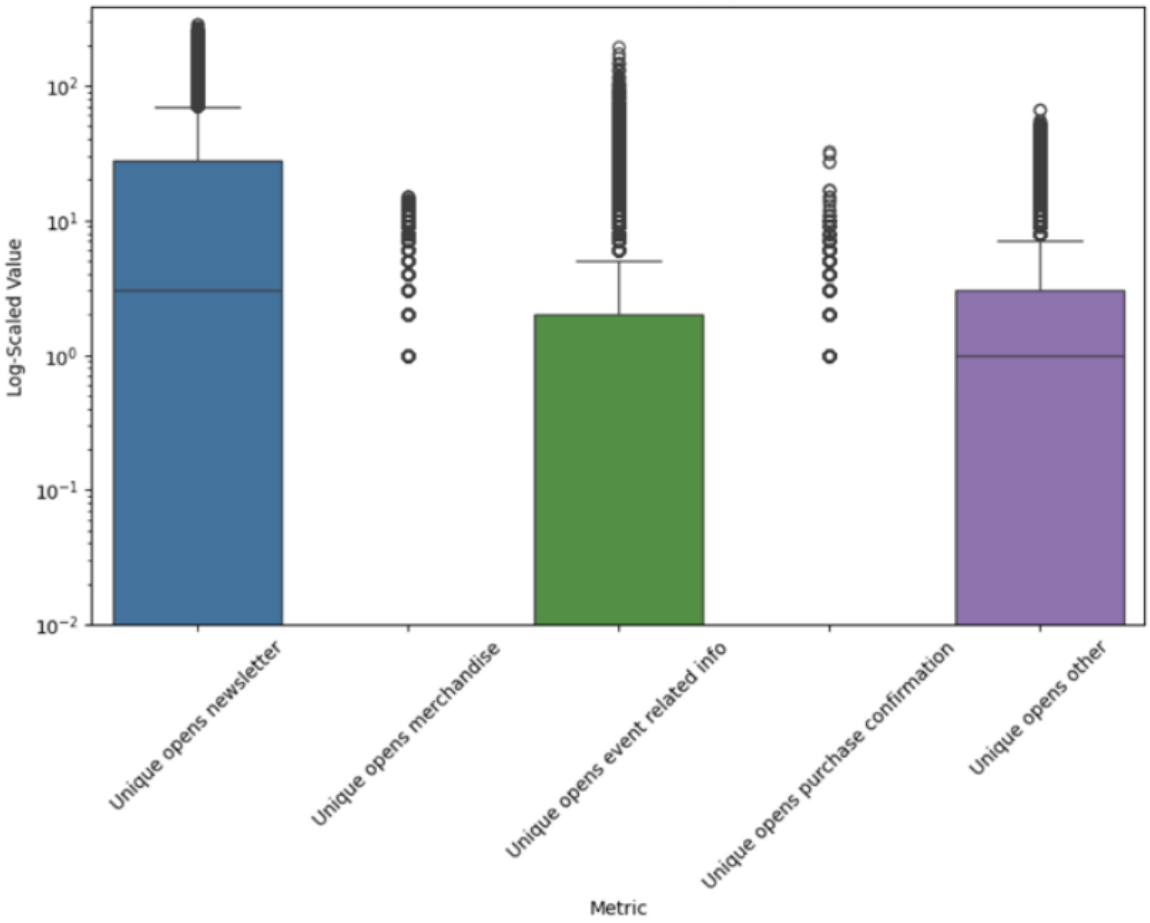

Figure 5.2: Distribution of Opened Email Engagement Metrics

Figure 5.3 presents the distribution of unique clicks, revealing a noticeable decline in engagement as customers move from opening emails to interacting with their content. Despite their wide reach, newsletters experienced sparse clicks, as did merchandise and event-related emails.

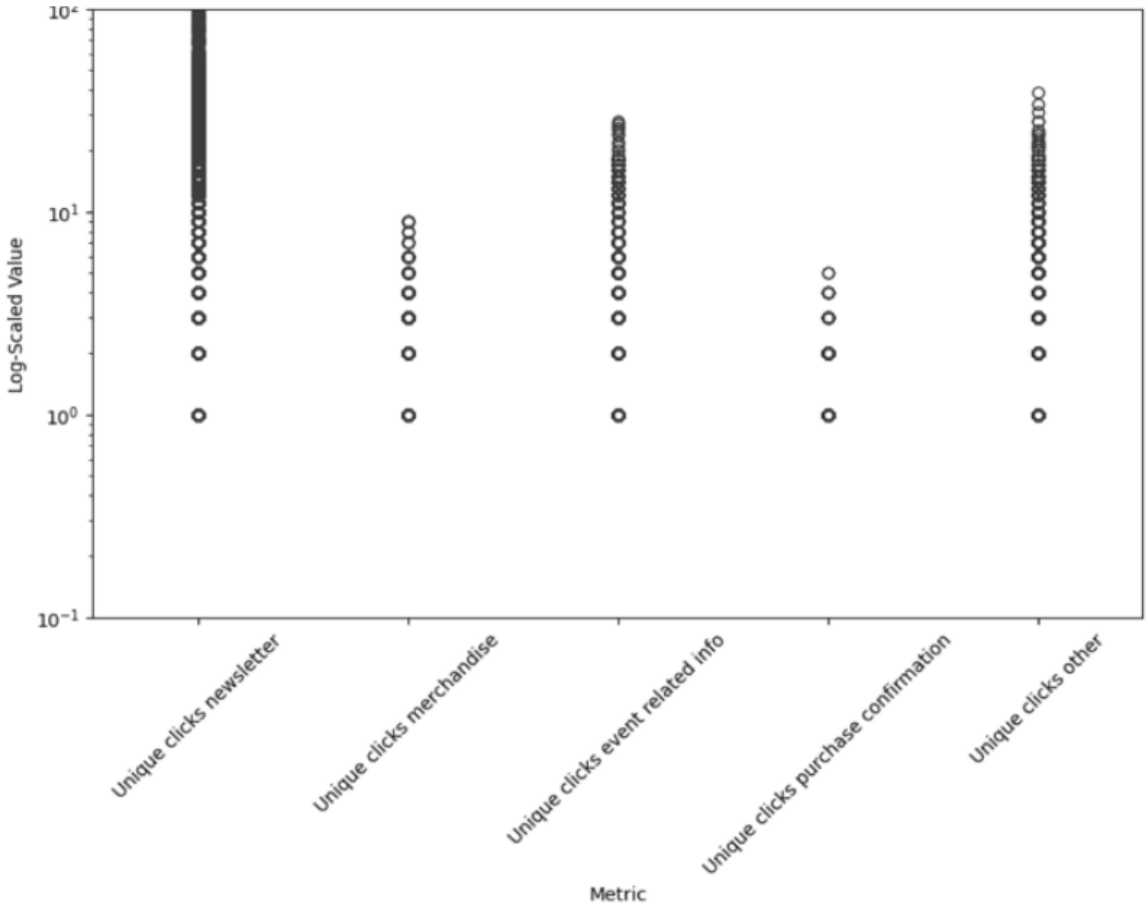

Figure 5.3: Distribution of Clicked Email Engagement Metrics

5.2 Email Engagement Rates by Granular Source

Figure 5.4 displays a heatmap summarizing median engagement rates (open rates, click-through rates, and open-to-click rates) across acquisition sources. Smaller sources, such as camps and outdoor plaza events, exhibited the highest open rates (0.403 and 0.424, respectively). However, these findings should be interpreted cautiously given the limited sample sizes. Broader channels, such as ticket purchases and online merchandise, also reported strong open rates (0.386 and 0.250), highlighting their relevance as key acquisition points. More passive channels, including mobile app sign-ins and Wi-Fi registrations, showed lower open rates (0.139 and 0.197). These results suggest that passive acquisition sources may lack the immediacy or relevance of more interactive touchpoints. Across all sources, click-through rates and open-to-click rates were low, indicating a gap in converting email opens into more meaningful engagement.

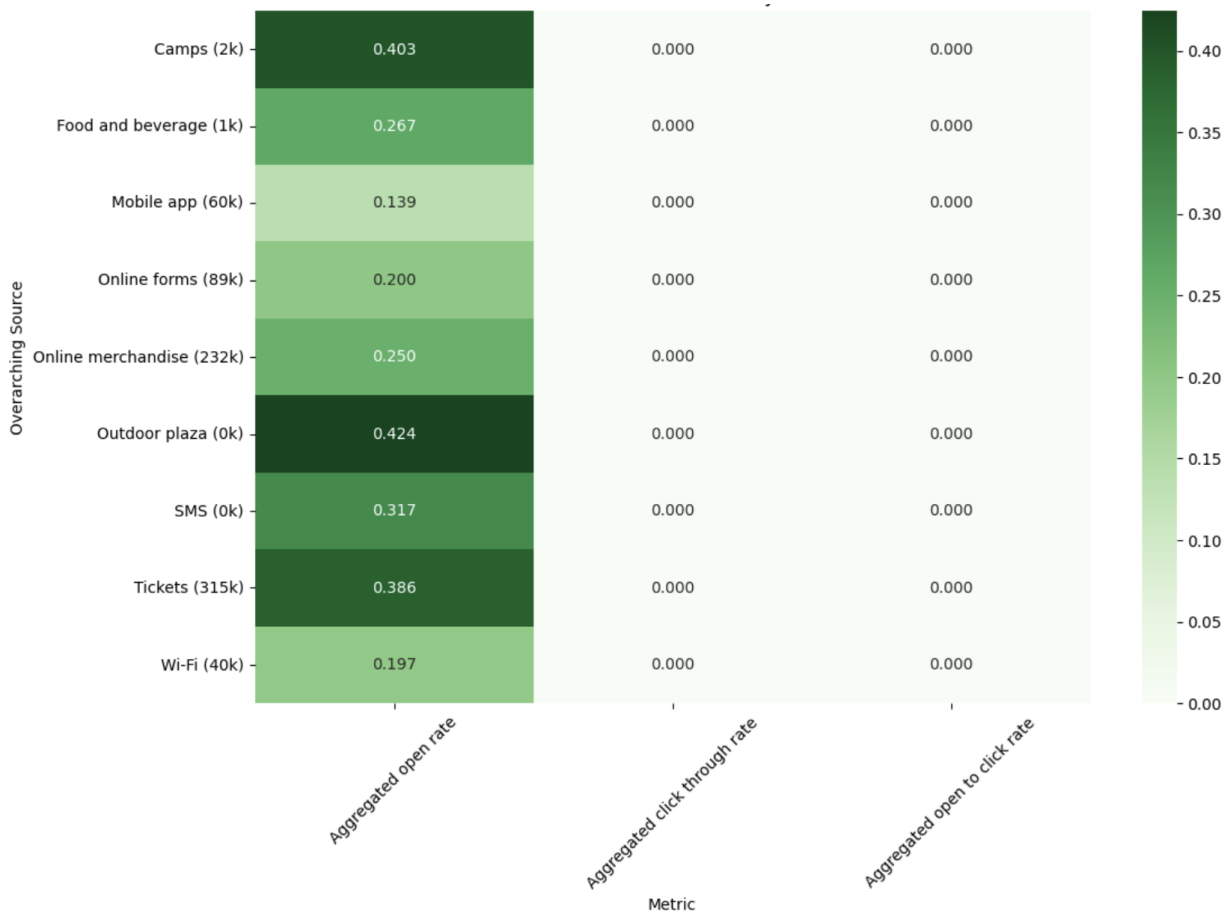

Figure 5.4: Median Email Rate Metrics by Overarching Source

5.3 Voluntary Opt-In Metrics by Source

Newsletter and merchandise opt-ins were also analyzed to understand customer subscription behaviors across acquisition sources. The breakdown focuses only on customers who opted in, regardless of whether they were automatically or voluntarily subscribed. The team's internal data also illustrates a strong connection between fan enthusiasm and opt-in rates, with voluntary subscriptions positively correlating with periods of greater team success.

Figure 5.5 illustrates the percentage of customers voluntarily opting into newsletters across various acquisition sources. The highest opt-in percentages are seen in secondary market forwarded tickets, mobile app, and Wi-Fi, with over 90% of customers in these categories voluntarily subscribing. However, while these categories have high opt-in rates, the total customer counts are relatively small, whereas online merchandise and primary market tickets represent larger customer populations but have significantly lower voluntarily opt-in rates.

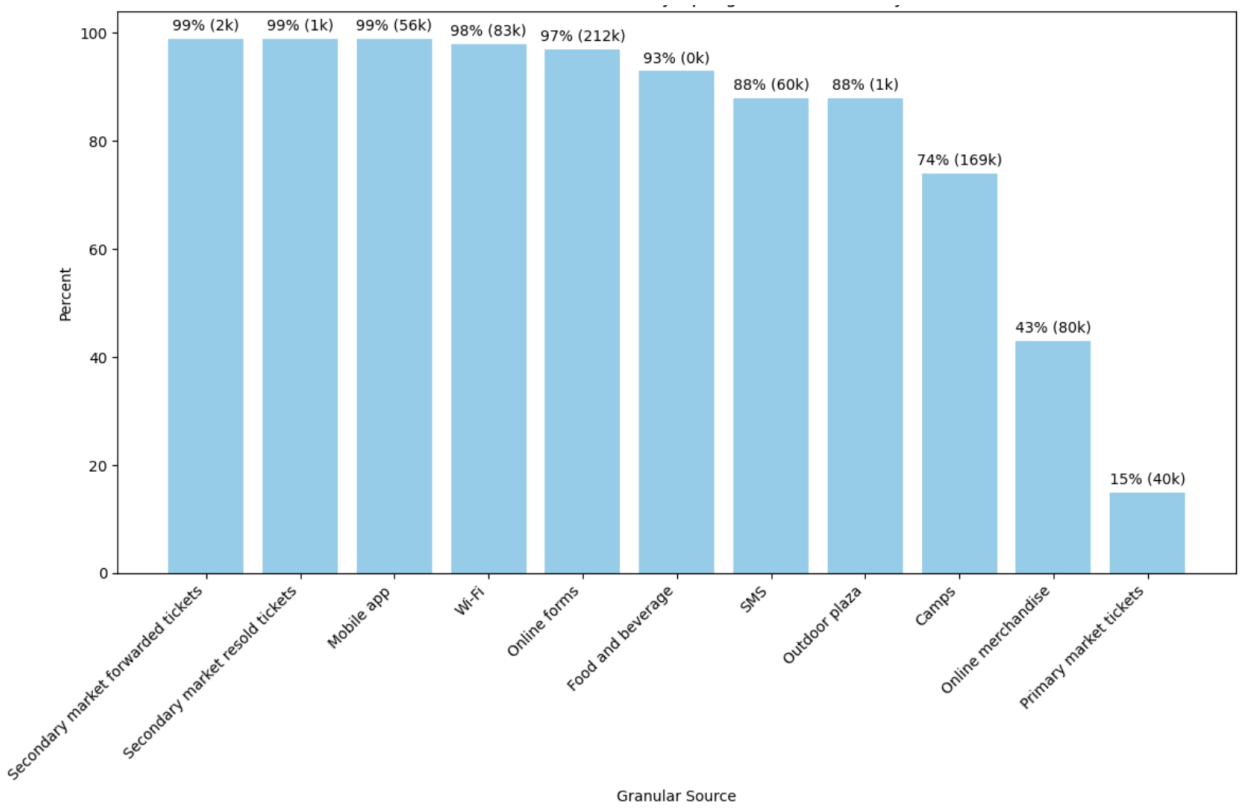

Figure 5.5: Percentage of Fans Voluntarily Opting into Email Newsletters by Granular Source

Figure 5.6 depicts the percentage of fans voluntarily opting into merchandise emails across different acquisition sources. It shows a more balanced distribution between voluntary and auto opt-ins, with the exception of online forms and SMS, which have notably higher voluntary opt-ins, and online merchandise, which shows a stronger reliance on auto opt-ins.

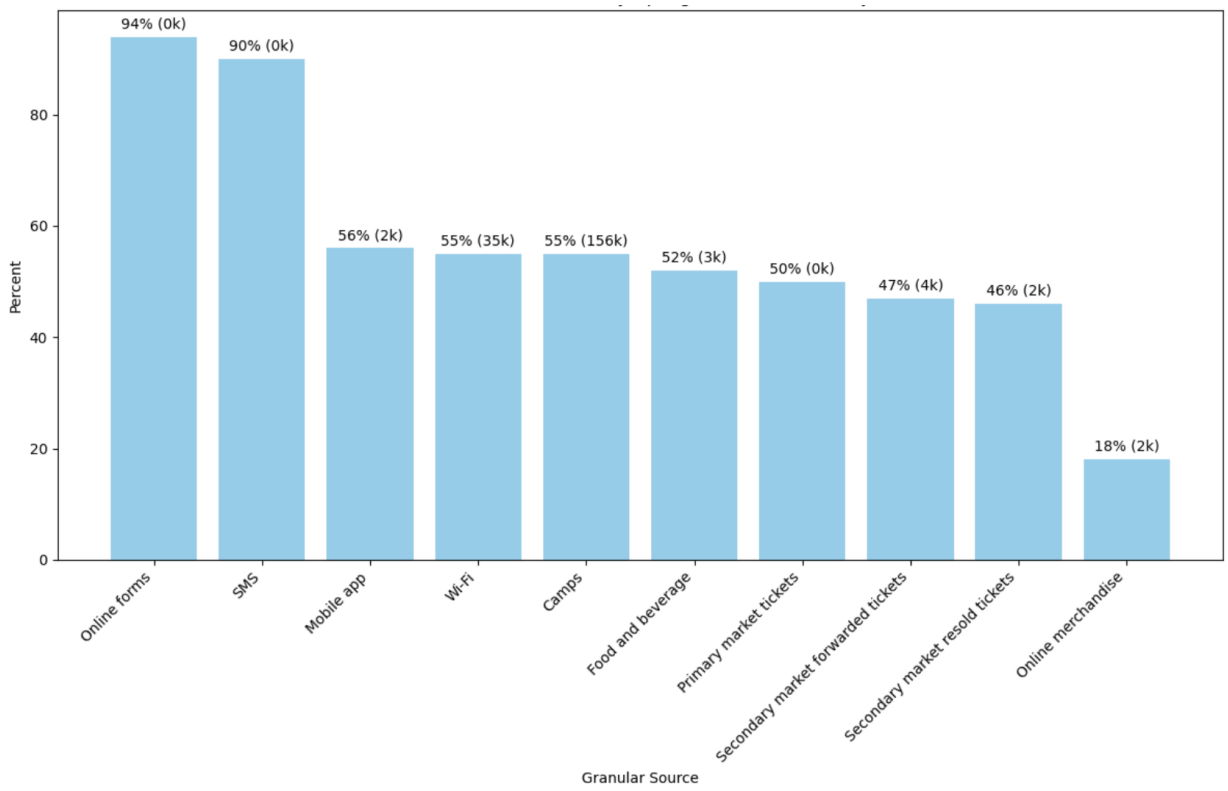

Figure 5.6: Percentage of Fans Voluntarily Opting into Merchandise Emails by Granular Source

Chapter 6

Machine Learning Methodology

Chapter 6 details the analytical approaches applied to understand fan behavior and predict spending patterns, emphasizing their role in optimizing the customer journey. Clustering was employed to identify distinct customer segments, while classification focused on forecasting year-over-year spend levels. Together, these methods address the immediate objectives of this thesis and establish a scalable framework for analyzing customer evolution within the team's ecosystem. This chapter also covers key preprocessing steps, highlights model selections, and reflects on the limitations and challenges encountered, providing a road map for future data-driven initiatives.

6.1 Model Formulation

6.1.1 Clustering

The clustering approach was chosen to identify distinct fan segments and behavior patterns by analyzing a range of features related to email engagement and historical spending behaviors. There were approximately 750,000 fans included in his model, selected based on prior email engagement. Key metrics considered included email delivery, unique open rates, and unique click volumes across various categories such as newsletters, merchandise promotions, event-related information, post-purchase confirmations, and other communications. Additionally, binary flags indicated whether a fan was auto-subscribed or voluntarily opted into the team's newsletter and merchandise promotions. To further enrich the dataset, additional features were incorporated, including the number of years of email activity and spending metrics such as primary ticket purchases, secondary resale activity, forwarded tickets, and online merchandise transactions. These features were chosen for their relevance in understanding customer behavior and their potential to inform actionable marketing strategies. By combining engagement patterns that reflect fan activity and spending interest that signal financial commitment, the clustering approach offers a holistic perspective on customer interactions with the team.

To simplify the dataset while retaining key patterns in customer behaviors, Principal Component Analysis (PCA) was applied as a preprocessing step. By reducing the number of features, PCA addressed challenges such as noisy data and correlations between variables. The resulting components highlighted the most important patterns, ensuring that the clustering process prioritized meaningful insights. This approach improved both the efficiency of the clustering algorithm and the interpretability of the final segments, making the results more actionable for business strategies [6].

Following PCA, the clustering analysis utilized the reduced dimensions and cluster summaries to interpret and classify customer behaviors. The PCA loadings highlighted which original features contributed most strongly to each principal component, while the cluster summaries provided detailed insights into the average characteristics of each cluster

in terms of the original features. These methods ensured that the identified clusters were both statistically robust and meaningful for business applications.

The optimal number of clusters was determined using the elbow method. This method identifies the point at which the rate of decrease in within-cluster inertia—an indicator of how tightly grouped the points are within each cluster—begins to level off. Smaller inertia values indicate tighter, more cohesive clusters, which are generally desirable. However, adding more clusters beyond the elbow point provides diminishing returns as the improvement in inertia becomes negligible. Balancing interpretability and cluster compactness, the analysis segmented customers into three distinct groups. The specific characteristics of each cluster, including their behavioral and spending patterns, are explored in detail in Section 7.1.3.

6.1.2 Classification

The classification model was developed to predict customers' spending buckets for the following year by leveraging their prior season's spending behaviors, email engagement patterns, and other key features [7]. This approach relied on the assumption that spending patterns from one season could reliably predict spending behavior in the next. The primary goal of the model was to provide insights into expected spending levels, enabling targeted marketing strategies, optimized revenue opportunities, and improved customer retention.

To enhance interpretability and minimize noise, the model focuses on year-over-year spend patterns rather than aggregating all historical behaviors. Given the variability in spending patterns across the fanbase, the most recent season's behavior serves as the most relevant and reliable indicator of future purchasing intent. As more data becomes available in future seasons, the model can be enhanced to capture broader behavioral patterns, improving its robustness while remaining accessible to stakeholders.

The decision to approach this problem as a classification task, rather than a regression problem, considered both practical and conceptual points. Spend buckets, which group customers into ranges of total spending, align more directly with the real-world needs of the business, where actionable insights often stem from understanding behavioral segments rather than predicting precise dollar amounts. For example, distinguishing between a $100 customer and a $10,000 customer provides far greater strategic value than focusing on marginal differences, such as $110 versus $120. Furthermore, classification mitigates the influence of outliers and noise in spending data by grouping extreme values into predefined categories, reducing their impact on model predictions. This approach not only aligns closely with stakeholder priorities but also ensures the resulting insights are actionable and tailored to the objectives of the analysis.

There were about 550,000 fans included in this model and they were selected based on prior financial engagement, demonstrating spending in at least one category such as primary ticket purchases, secondary market resales, forwards, or online merchandise purchases. These spending behaviors formed the foundation of the model, providing insights into customer commitment and loyalty. Among the spending-related metrics, the model considered detailed ticketing activity, such as the number of individual tickets purchased, the number of mini plans bought, and the total number of games and seats reserved. Additionally, the online merchandise data captured both the total spend and the types of merchandise purchased, further enriching the understanding of customer purchasing behaviors.

On top of these spending metrics, the model incorporated a range of non-spending features to provide a more comprehensive view of customer behavior. These included many of the same email engagement metrics used in the clustering model, such as the volume of emails delivered, opened, and clicked, and their respective rates, as well as whether or not they were subscribed to receiving the team's newsletter and merchandise promotions that year. These features offered valuable insights into customer interaction with team communications across various channels, complementing the spending data.

Other relevant predictor variables included things like attendance percentage, which measured the proportion of games attended relative to those purchased, reflecting customer engagement and participation. These features collectively provided a holistic view of customer behavior, blending spending and non-spending elements to inform predictions.

The target variable for the classification model was spend buckets, which represent ranges of total customer spending in the following season. To address the skewness in spending data, a log transformation was applied to normalize the distribution, as shown in Figure 6.1. This transformation highlighted the concentration of customers in lower spend ranges while still preserving the relative differences across higher spend buckets.

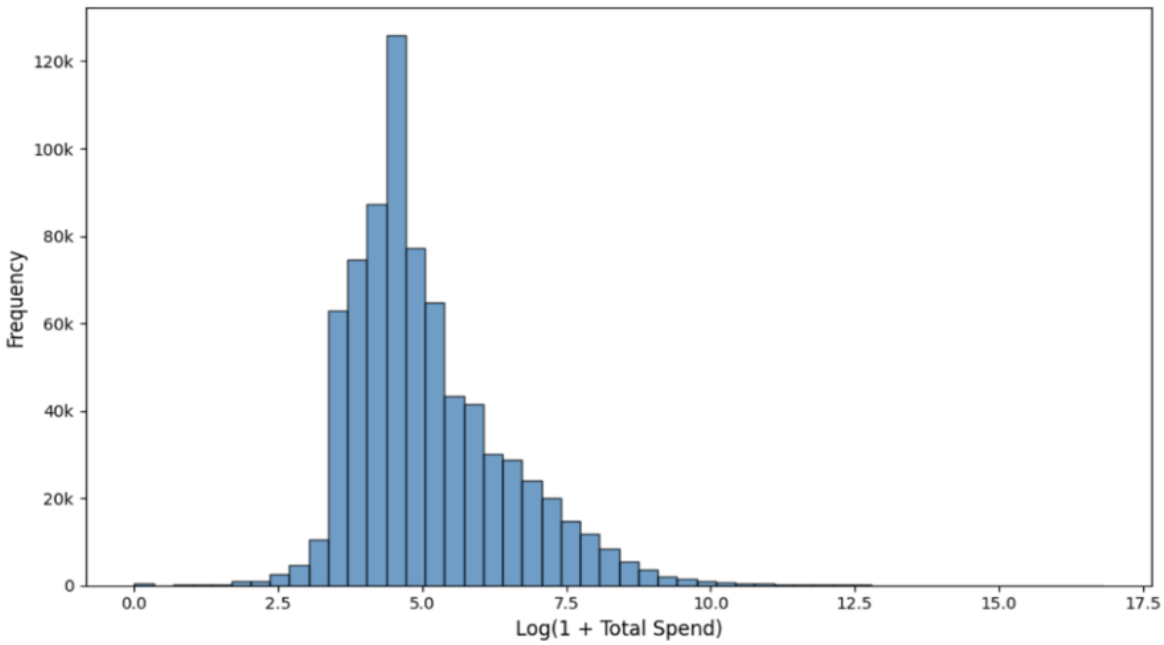

Figure 6.1: Log-Transformed Total Spend Bucket Histogram

The spend buckets were then defined as follows: \$0, \$1–\$100, \$101–\$250, \$251–\$1,000, \$1,001–\$2,500, \$2,501–\$5,000, and \$5,001+. The classification task involved assigning each customer to one of these spend buckets, enabling the model to focus on predicting future spending levels based on the prior season's behavior. By normalizing the skewed spending data, the log transformation facilitated more balanced model training and improved the interpretability of predictions.

6.2 Model Selection

6.2.1 Clustering

The unsupervised learning technique employed was MiniBatchKMeans, a scalable variant of the traditional KMeans algorithm. This method was selected due to its computational efficiency and ability to handle large datasets while maintaining interpretable and actionable cluster outputs. Traditional KMeans recalculates distances and reassigns clusters for every data point in each iteration, which can be computationally intensive and memory-consuming for large datasets. MiniBatchKMeans addresses this challenge by updating clusters using a randomly selected subset of the data at each iteration, significantly improving efficiency. This approach builds on the principles of the k-means++ initialization method, which optimizes cluster center selection to enhance convergence and clustering performance [8].

While MiniBatchKMeans may not always find the exact same cluster centers as traditional KMeans, its computational advantages make it well-suited for this use case. The primary aim of clustering in this analysis was to reveal customer heterogeneity in engagement and spending patterns, where the approximations provided by MiniBatchKMeans were sufficient for deriving actionable insights.

6.2.2 Classification

To evaluate customer behavior and predict key outcomes, multiple machine learning algorithms were tested to ensure a robust analysis that balances predictive accuracy, computational efficiency, and interpretability. These models were selected based on their ability to handle a variety of dataset characteristics, including high-dimensional features, class imbalances, and potential non-linear relationships.

The classification models evaluated included CART (Classification and Regression Trees), Random Forest, gradient-boosting algorithms like LightGBM and XGBoost, Neural Networks, and AutoML tools such as AutoGluon. This diverse selection allowed for benchmarking performance across a range of methodologies:

- *CART* was chosen for its simplicity and interpretability, serving as a baseline to assess the performance of more complex models.
- *Random Forest*, as an ensemble method, was included for its ability to reduce overfitting while offering insights through feature importance scores.
- *Gradient-boosting algorithms* like LightGBM and XGBoost were evaluated for their state-of-the-art performance, scalability, and ability to model complex interactions in the data.
- *Neural Networks* were included to explore their capability to capture highly non-linear patterns, particularly in datasets with rich, complex features.
- *AutoML tools*, such as AutoGluon, were tested to provide a comparison against manually tuned models, leveraging automated processes to identify potential solutions efficiently.

Each model was evaluated on key metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to ensure comprehensive benchmarking. The rationale for this approach was to identify the model that best aligns with the goals of this analysis: strong predictive performance, practical interpretability, and scalability for future applications. This evaluation provided insights into the relative strengths of different algorithms, ultimately guiding the selection of the most appropriate model for further refinement and analysis, as detailed in Chapter 7.

6.3 Feature Engineering

6.3.1 Clustering

The feature engineering process for the clustering model was designed to capture critical patterns of email engagement, spending, and subscription behaviors, with a focus on reflecting the temporal and dynamic nature of fan interactions [9].

A weighted aggregation framework was implemented to account for customer behaviors over the three seasons analyzed (2021–2022, 2022–2023, and 2023–2024) and resulted in one row per customer. Initial weights of 0.2, 0.3, and 0.5 were applied to these seasons, prioritizing recent behaviors based on the hypothesis that they are the most predictive of future engagement. For customers who entered the ecosystem after the first season, the weighting coefficients were dynamically adjusted and re-allocated evenly. For example, if a customer joined in 2022–2023, the weight assigned to the first season was redistributed across the remaining seasons, resulting in weights of 0.4 for 2022–2023 and 0.6 for 2023–2024. This methodology ensured that customer behaviors were appropriately contextualized while maintaining the temporal relevance of the analysis.

Given the diverse scales of the features, ranging from email counts and rates to ticket spending and flags for subscription origin, a standard scaler was applied to normalize the data. This step ensured that no single variable disproportionately influenced the clustering outcomes due to differences in scale, allowing the model to treat all features equitably.

By consolidating the data into one row per customer and integrating both transactional and behavioral insights, the feature engineering process created a robust foundation for clustering. The resulting dataset effectively balanced granularity and temporal relevance, enabling meaningful segmentation that aligned with the research's broader goals of understanding and fostering fan engagement.

6.3.2 Classification

Feature engineering was a pivotal step in preparing the dataset for classifying year-over-year spend buckets. The process involved addressing multi-collinearity, refining key features, and incorporating transactional, temporal, and behavioral insights to ensure the data aligned with the analysis objectives.

To ensure the model's robustness and interpretability, the first step was to mitigate multi-collinearity. Highly correlated variables can distort model performance and inflate the significance of certain predictors. As seen in Figure 6.2, a correlation matrix was used to visualize relationships between features and flag those with a Pearson correlation coefficient above 0.9 for further examination. Many of these flagged columns were linear combinations

of other variables, such as total primary individual ticketed games, which was the individual ticketed games and other ticketed events.

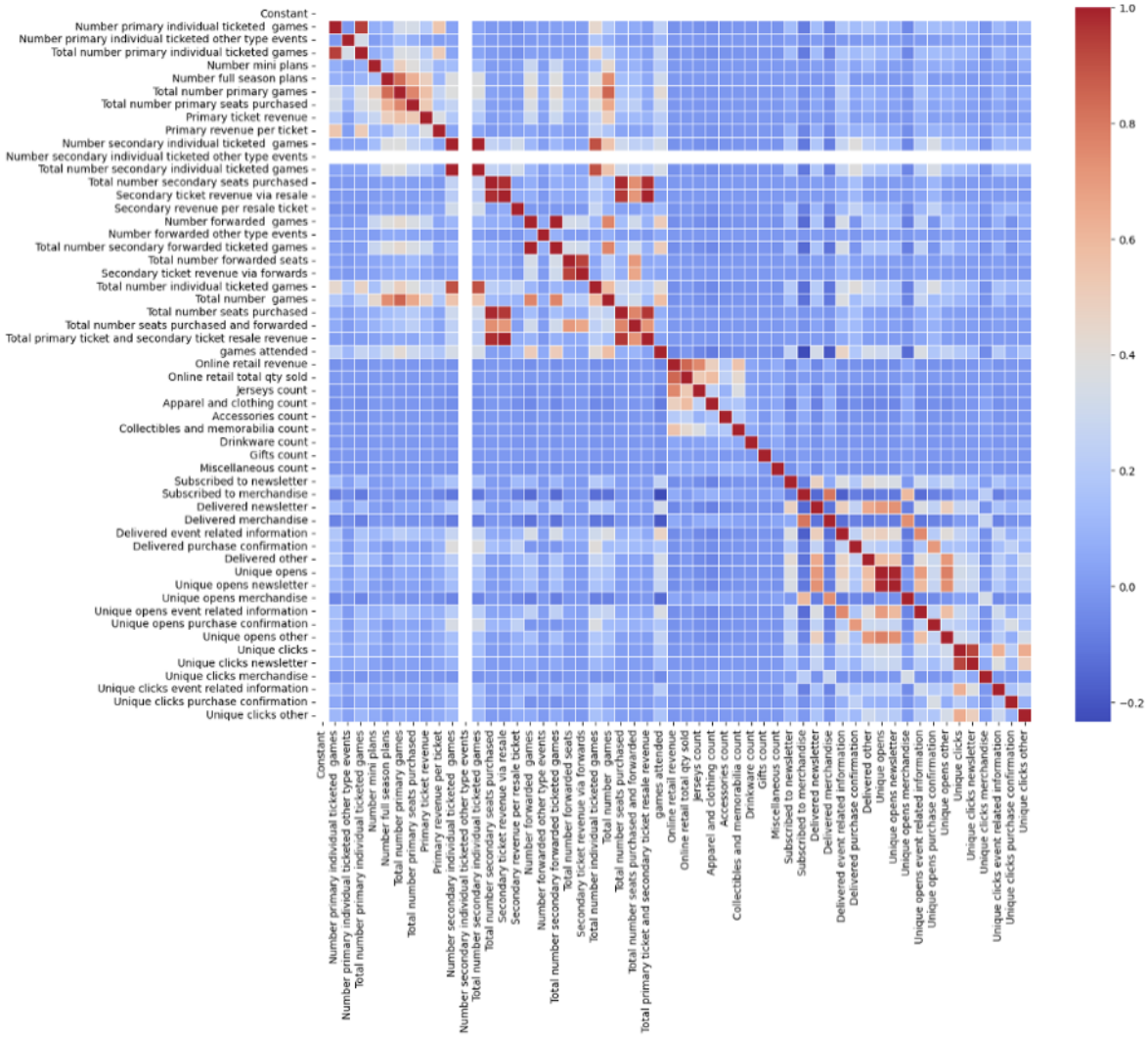

Figure 6.2: Correlation Matrix of Independent Variables for Clustering Model

In addition to the correlation matrix, Variance Inflation Factor (VIF) values were calculated to quantify multicollinearity among predictors. There were 14 features that exceeded the common threshold of 10 and were therefore removed. Examples of excluded variables include "Total number games", "Total number primary individual ticketed games", and so on. This two-step process ensured the final feature set struck a balance between predictive performance and interpretability, retaining only those variables that provided distinct and meaningful contributions to the analysis.

Categorical features, such as acquisition sources, were one-hot encoded to provide insights into different fan entry points. Lag features, including components of the prior season's spend bucket (e.g., primary market ticket spend, resale ticket spend, forwarded tickets, and online merchandise spend), added historical context by linking the past season's behaviors to future

spending patterns. Furthermore, email engagement metrics, such as CTR and delivery volume, contributed a behavioral dimension, illustrating how digital email interactions influenced loyalty and spending.

To support model training and validation, the dataset was divided into training, validation, and test sets using stratified sampling to ensure balanced representation across spend buckets. The training set, comprising 60% of the data, was used for model fitting and feature refinement, whereas the validation set (30%) served as a benchmark for interim evaluations during hyperparameter tuning and model selection. The test set (10%) was held out for final validation, ensuring unbiased assessments of the selected model's generalizability.

The feature engineering process resulted in a dataset that integrated transactional, temporal, and behavioral features to create a comprehensive view of fan progression. This refined dataset aligned with the overarching objectives of the research: identifying opportunities to ramp up fan engagement, fostering deeper connections, and driving sustained loyalty over multiple seasons.

6.3.3 Handling Missing Values

Handling missing values was an essential step in preparing the dataset, particularly for features tied to ticket purchases, email engagement, and behavioral patterns. Missingness in the data primarily stemmed from customers with sporadic interactions or from features that were not universally applicable.

For monetary spend components, missing values were imputed as zeros to reflect the absence of spending. This ensured that customers who made no purchases during the period were correctly represented in the dataset, preserving the distinction between true inactivity and missing data.

Derived metrics, including email open rates, click-through rates, open-to-click rates, and attendance percentages, presented a challenge. Missing values in these features often arose from division by zero, such as when no emails were delivered, or no tickets were purchased. To address this, these missing values were replaced with zeros to ensure compatibility with modeling requirements, while separate missingness indicator variables were introduced for each of these features. These indicators allowed the model to distinguish between true zero behavior and cases where the original values were undefined.

The imputation approach balanced practicality with the need to retain interpretability and accuracy. This ensured that the model could incorporate nuanced patterns of customer behavior without introducing bias or distorting the results.

6.4 Classification Model Parameters

The classification model aimed to predict customers' spending buckets for the following season by evaluating a variety of machine learning algorithms. This section focuses on the process of assessing model performance through default and optimized parameter settings, emphasizing the importance of hyperparameter tuning in uncovering the full potential of candidate algorithms.

6.4.1 Default Parameters

To establish a baseline, all models were first evaluated using their default parameter settings. This initial step provided a fair comparison of out-of-the-box capabilities and highlighted strengths and weaknesses across algorithms. The key insights from this phase included:

- *Gradient-boosting models*, such as LightGBM, XGBoost, and CatBoost, demonstrated strong predictive performance, even with default configurations. These models effectively captured the complexity of customer behavior and spending patterns.
- *Ensemble methods*, such as Random Forest, showed reasonable accuracy and recall but did not match the overall effectiveness of gradient-boosting algorithms.
- *Simpler models*, such as CART and Naive Bayes, struggled with the complexity of the dataset and performed poorly in both accuracy and AUC-ROC.
- *Neural networks* required significant hyperparameter tuning to become competitive, with default configurations yielding suboptimal results.

A detailed comparison of default parameter results is presented in Section 7.2.1.

6.4.2 Hyperparameter Tuning

To maximize model performance, hyperparameter tuning was conducted for each algorithm. This process focused on optimizing critical parameters such as learning rate, tree depth, and the number of estimators to achieve a balance between model complexity and generalization, while minimizing the risk of overfitting. Grid search and cross-validation served as the primary strategies for systematically identifying the best combinations of parameters.

The tuning process demonstrated improvements in key metrics, with notable gains in AUC-ROC, precision, and recall observed for several algorithms. The specific parameter settings and performance metrics for each tuned algorithm are detailed in Section 7.2.1, where they are summarized in Table 7.3. These results underscore the importance of hyperparameter tuning in maximizing model performance and ensuring robustness.

Chapter 7

Analysis of Results and Findings

7.1 Clustering Results

7.1.1 Principal Component Analysis

PCA was applied to reduce the dataset's high dimensionality, retaining meaningful patterns while enhancing computational efficiency. By transforming correlated features into orthogonal components, it captured the most significant axes of variation, simplifying the interpretation of clustering results.

Variance explained is a key concept: components with higher variance contribute more significantly to capturing and understanding the underlying structure of the data. As shown in Table 7.1, the first four principal components collectively account for roughly 50% of the total variance, with diminishing returns beyond this point. Retaining these four components ensures that the analysis captures a substantial portion of the dataset's variability while avoiding unnecessary complication.

Principal Component	Explained Variance Ratio	Cumulative Variance Ratio
PC1	0.197999	0.197999
PC2	0.153768	0.351767
PC3	0.077720	0.429487
PC4	0.065537	0.495024

Table 7.1: Explained Variance Ratios of Principal Components for Clustering Model

The two-dimensional scatter plot (Figure 7.1) visualizes the clustering results just using the first two principal components. While these components explain approximately 35% of the variance, they provide a clear and interpretable representation of the clustering structure. Distinct groupings emerge, suggesting meaningful differences among fan segments based on the features captured by these components.

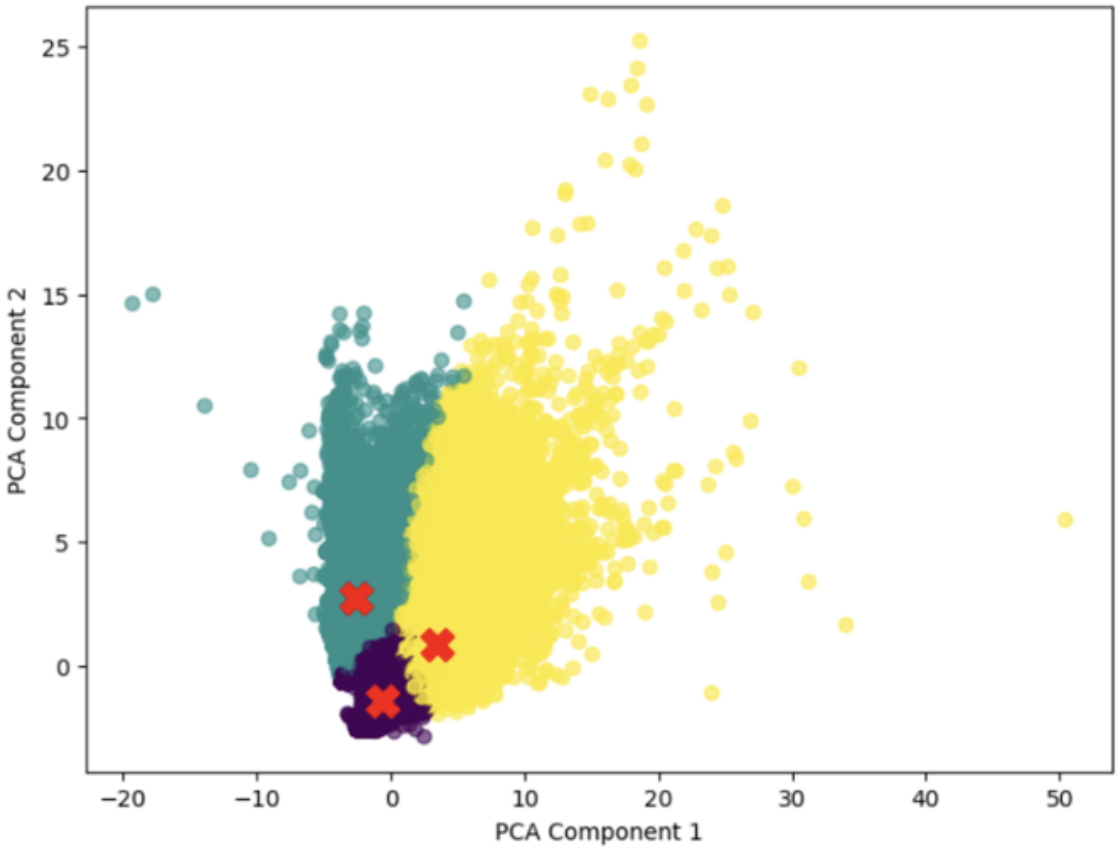

Figure 7.1: 2-D MiniBatch K-Means Clusters with Centroids

To understand the composition of each principal component, the heat map in Figure 7.2 highlights the contributions of individual features, showing how they load onto each component and shaping the dimensions driving the data. The first principal component is heavily influenced by weighted email delivery and engagement metrics, underscoring the importance of digital activity in differentiating fan behaviors. In contrast, the second principal component is dominated by merchandise opt-in behaviors, suggesting a distinct axis tied to online retail engagement. The heatmap also shows that certain features, like newsletter opt-ins, contribute meaningfully across multiple components, signaling their broader importance in defining engagement patterns. The heatmap translates reduced dimensionality into clear insights, preserving the dataset's nuance while highlighting the key factors driving segmentation.

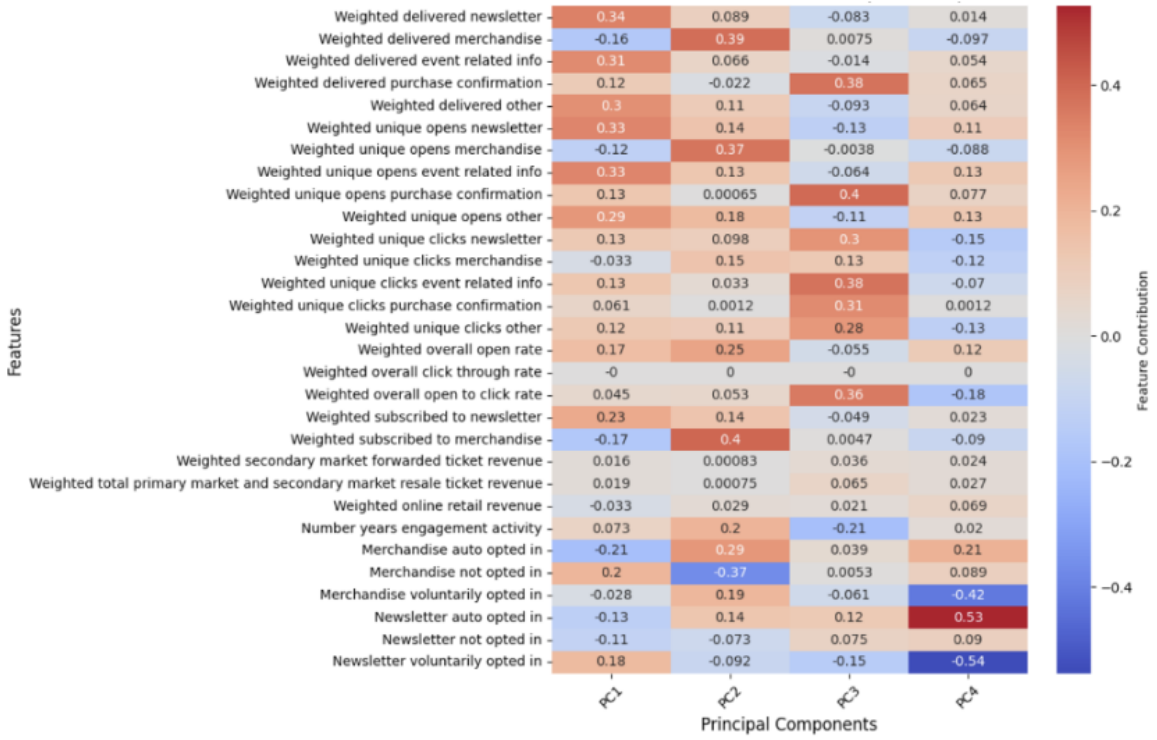

Figure 7.2: Feature Contributions to Principal Components for Clustering Model

While PCA reduced the dimensionality, it is important to note that not all variance in the data is captured. However, the feature loadings provide a valuable lens for interpretation. Identifying the key variables associated with each principal component allows us to infer the characteristics defining clusters, as discussed in Section 7.1.3. This approach maintains the dataset's nuance while providing a meaningful framework for understanding segmentation at a broader level.

7.1.2 Optimal Number of Clusters

To determine the most meaningful number of clusters, the elbow method was employed (Figure 7.3). This technique evaluates the reduction in 'inertia'—a measure of how tightly the data points are grouped within each cluster. By identifying the point where this reduction starts to level off, we were able to choose the number of clusters that maximized interpretability and coherence while avoiding overfitting.

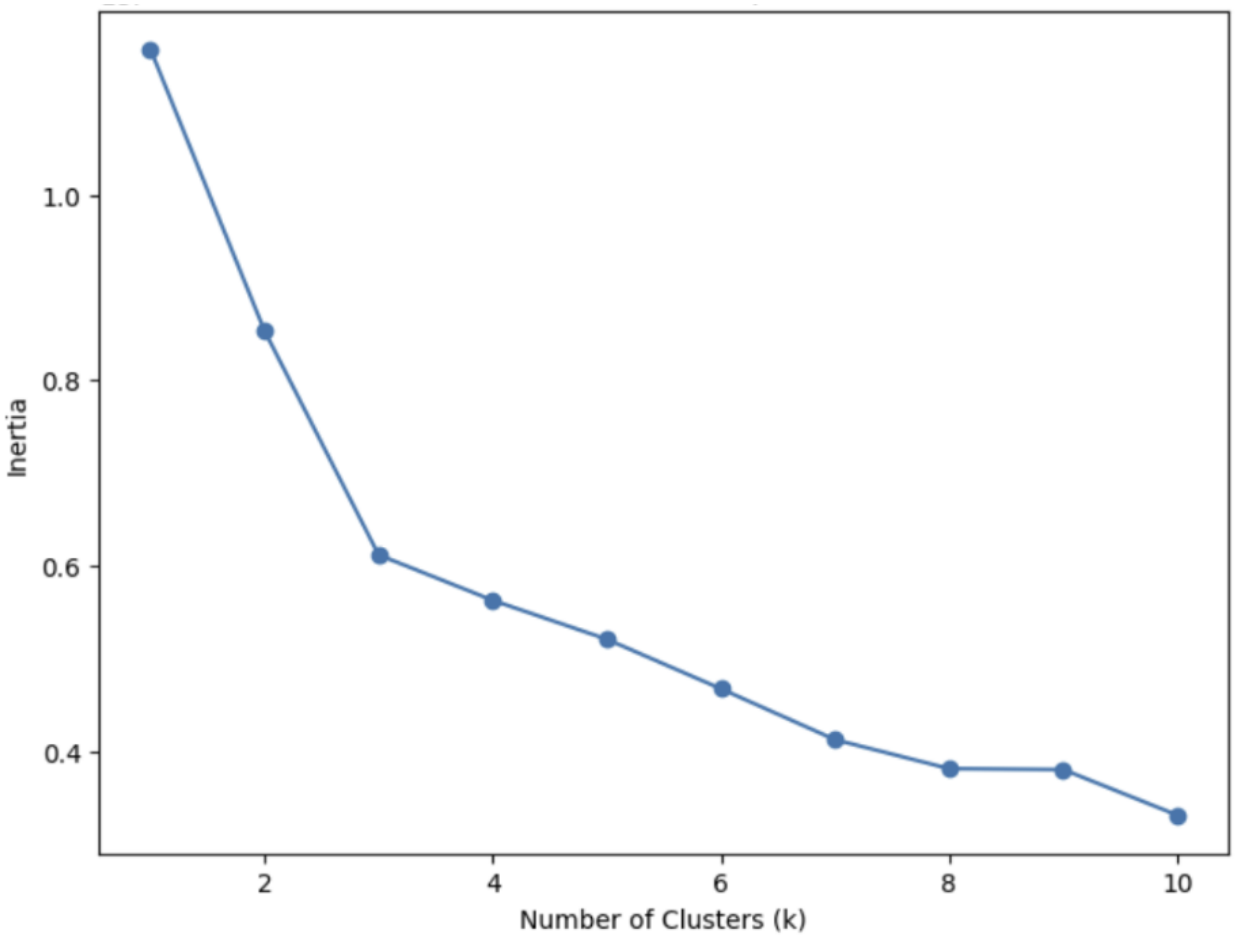

Figure 7.3: Elbow Plot to Determine the Number of Clusters, k

To validate this choice, the clustering algorithm was run with four clusters too. However, this did not provide additional interpretive or practical benefits over the three-cluster solution. With three segmentations, distinct profiles are effectively captured while maintaining simplicity and interpretability, making it the best choice for the analysis.

7.1.3 Cluster Profiles

This approach breaks down the fanbase into distinct segments, each with unique engagement and spending behaviors. As shown in the first bar plot (Figure 7.4), Cluster 1 represents 55% of the population, making it the largest group, while Clusters 2 and 3 account for 22% and 23%, respectively. This distribution highlights key differences within the fanbase and sets the stage for a deeper exploration of each cluster's characteristics.

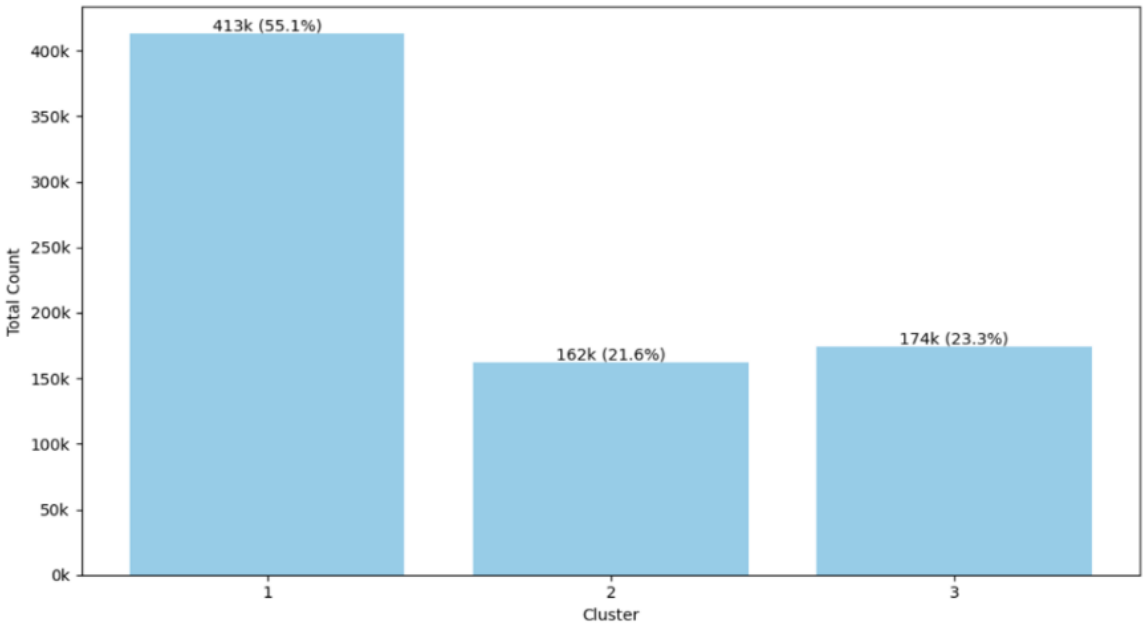

Figure 7.4: Number of Fans Assigned to Each Cluster

It is important to interpret the cluster metrics in context. Metrics such as ticket revenue, email engagement rates, and behavioral indicators are presented as weighted averages to account for differences in how long customers have engaged with email campaigns. These figures are not meant to reflect absolute values on a per-customer basis but instead provide a relative measure of trends within and between clusters. By incorporating these averages alongside engagement duration, the analysis allows for a more balanced and meaningful comparison of behaviors across the fanbase.

The following cluster profiles delve into the unique traits of each group, including engagement patterns, spending behaviors, and opportunities for targeted strategies:

Cluster 1: Limited Engagement, Moderate Ticket Spend

Cluster 1 shows low engagement across most channels. Email interactions are minimal, with overall open rates below 15%. However, this group demonstrates significant interest in newsletter communications, with 75% voluntarily opting in, suggesting a preference for team-specific updates. Nonetheless, merchandise campaigns perform poorly, with 95% of customers not being opted in.

In terms of spending, average ticket revenue for this group is $166, spread across both primary and secondary resale markets, while merchandise spending remains low at $11. Forwarded ticket activity, averaging $62, indicates some exposure to the game-day experience, likely through gifting or shared attendance. On average, these customers have been active in email campaigns for 1.67 years.

Retention strategies that align with this group's interests could help boost engagement. For instance, messaging focused on newsletter content, coupled with low-commitment offers like resale incentives or exclusive promotions, may encourage deeper involvement and longer-term

loyalty.

Cluster 2: Limited Engagement, High Merchandise Spend

Cluster 2 is characterized by its strong focus on online merchandise. This is reflected in their high opt-in rates for merchandise communications, with 72% auto-subscribed and 28% voluntarily subscribed. General email campaign engagement, however, remains low, highlighting a specific preference for retail-related content.

Merchandise revenue for this group averages $47, significantly outpacing ticket revenue at $18. Minimal forwarded ticket activity aligns with their lower engagement in game attendance. With an average of 2.24 years of email campaign engagement, this group's steady interest in merchandise provides a clear opportunity for targeted promotions. Combining exclusive product launches with insider content could increase engagement and deepen customer loyalty.

Cluster 3: High Engagement, High Ticket Spend, Minimal Merchandise Focus

Cluster 3 is the most engaged and highest-spending segment. Average ticket revenue on the primary and secondary resale markets is $551. Similarly, the average for secondary ticket revenue via forwards is also the highest by far between clusters at $496. Merchandise spending, however, is negligible, averaging just $3.

This group demonstrates strong engagement with email campaigns, with an overall open rate of 57% and 88% voluntarily subscribing to newsletter communications. On average, these customers have been active in email campaigns for 2.34 years, with 87% maintaining engagement for at least two seasons. Their established presence within the ecosystem positions them as key contributors to the fanbase.

For this cluster, leveraging their high-ticket engagement through premium campaigns or advanced secondary markets offers may drive additional value. Additionally, connecting merchandise incentives to attendance milestones could address the gap in online retail engagement for this otherwise highly active group.

Lastly, the side by side bar chart (Figure 7.5) examines the geographic distribution of each cluster. As noted in Section 2, incomplete zip code data limits full representativeness, but it does reveal trends: Clusters 1 and 2 have a higher proportion of non-local customers, while Cluster 3 skews more local. These geographic differences may influence engagement patterns and spending behaviors.

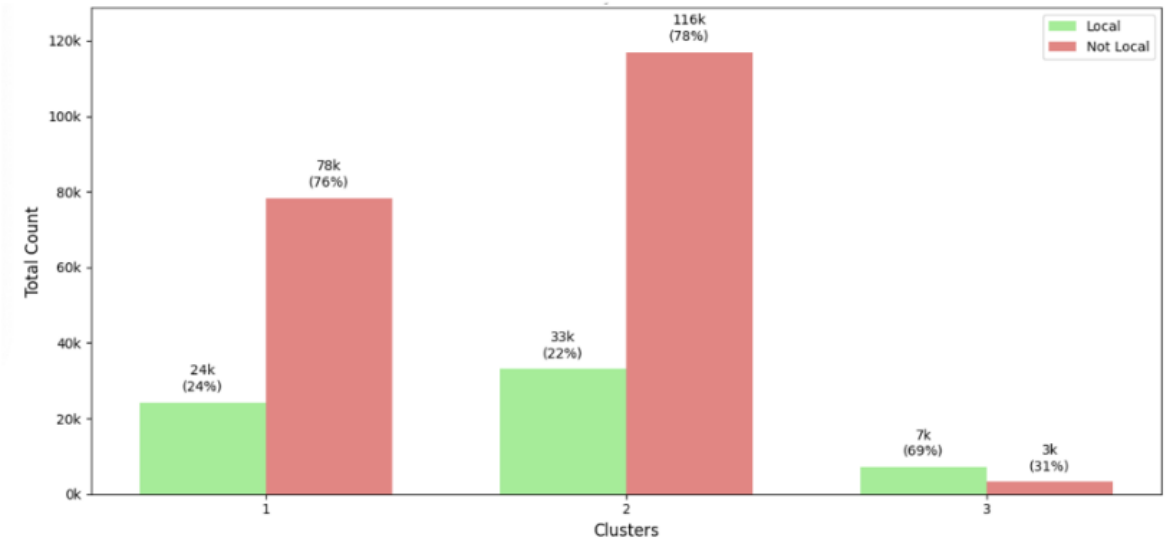

Figure 7.5: Fan Distribution by Cluster and Local Presence

7.2 Classification

7.2.1 Metrics and Model Comparison

The evaluation of classification algorithms began with default parameter benchmarking to establish a baseline. Among the models tested, LightGBM emerged as the strongest candidate, achieving an AUC-ROC of 0.955 (Table 7.2). Its out-of-the-box performance consistently outperformed competitors such as XGBoost and CatBoost, which also demonstrated robust capabilities. Neural networks and Random Forest models delivered solid predictive performance but fell short of the precision and recall achieved by gradient-boosting models. Simpler algorithms, such as CART and Naive Bayes, struggled to capture the complexity of the data, with Naive Bayes performing the worst across all metrics.

LightGBM's superior performance can be attributed to its computational efficiency, driven by features such as histogram-based growth and leaf-wise optimization [10]. These advantages made it particularly well-suited for this analysis, especially considering the scalability required for future dataset expansion. Additionally, its ability to provide feature importance scores enhanced interpretability, offering actionable insights into customer behavior.

Although XGBoost performed comparably and remains a reliable alternative with widespread adoption, LightGBM's combination of speed, scalability, and transparency ultimately positioned it as the preferred model. This selection aligned well with the project's emphasis on practical, stakeholder-oriented solutions.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
LightGBM	0.868	0.877	0.868	0.868	0.955
XGBoost	0.866	0.876	0.866	0.866	0.955
CatBoost	0.866	0.876	0.866	0.866	0.954
Random Forest	0.864	0.870	0.864	0.864	0.942
Neural Network	0.845	0.854	0.845	0.845	0.949
CART (Decision Tree)	0.829	0.830	0.829	0.829	0.877
K-Nearest Neighbors	0.818	0.821	0.818	0.818	0.915
Naive Bayes	0.320	0.402	0.320	0.225	0.692
TensorFlow Neural Network	0.797	0.816	0.797	0.787	0.932

Table 7.2: Performance Metrics of Classification Models with Default Parameters

Hyperparameter tuning was conducted to enhance the model's robustness, with a particular focus on improving recall and F1-score. As shown in Table 7.3, the optimal configuration for LightGBM—learning rate of 0.1, maximum depth of 6, and 200 estimators—yielded meaningful gains across key metrics.

Model	Best Parameters
LightGBM	{'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 200}
Random Forest	{'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200, 'oob_score': True}
XGBoost	{'learning_rate': 0.2, 'max_depth': 6, 'n_estimators': 200}
CatBoost	{'depth': 10, 'iterations': 200, 'learning_rate': 0.2}
Decision Tree	{'criterion': 'gini', 'max_depth': 15, 'min_samples_leaf': 5, 'min_samples_split': 2}
Neural Network	{'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': (64, 32), 'learning_rate_init': 0.001}
K-Nearest Neighbors	{'n_neighbors': 10, 'weights': 'distance'}
Naive Bayes	{'var_smoothing': 1e-07}

Table 7.3: Best Parameters for Classification Models.

The tuned model achieved an AUC-ROC of 0.9562 on the validation dataset, with notable improvements in both recall and precision (Table 7.4). These results, validated through cross-validation, underscore the model's robustness and reliability.

While Random Forest showed noticeable improvements, achieving an AUC-ROC of 0.951, it fell short in precision and recall. Similarly, XGBoost and CatBoost delivered incremental gains but did not surpass LightGBM's performance, reinforcing its position as the optimal choice for this analysis.

Model	Cross-Validated Accuracy	Test Accuracy	Precision	Recall	F1-Score	AUC-ROC
LightGBM	0.868	0.870	0.879	0.870	0.870	0.955
Random Forest	0.868	0.870	0.881	0.870	0.870	0.951
XGBoost	0.867	0.868	0.878	0.868	0.867	0.955
CatBoost	0.863	0.866	0.876	0.866	0.865	0.954
Decision Tree	0.851	0.854	0.861	0.854	0.854	0.941
Neural Network	0.837	0.842	0.856	0.842	0.841	0.948
K-Nearest Neighbors	0.816	0.823	0.827	0.823	0.823	0.921
Naive Bayes	0.333	0.341	0.391	0.341	0.239	0.631

Table 7.4: Performance Metrics of Classification Models

7.2.2 Feature Importance

The factors influencing the model's predictions were analyzed using LightGBM's feature importance metrics. Figure 7.6 highlights the top ten normalized feature importance scores, grouping predictors into non-spend-related (light blue) and spend-related (dark blue) categories; non-spend features capture engagement behaviors, while spend-related features reflect financial interactions. The term "current" refers to behavior from the prior season, serving as a predictor for spending in the subsequent season. The normalized feature importance scores provide a clear ranking of predictors, with values scaled to sum to one for easy comparison. While raw importance scores could offer complementary insights, the normalized values effectively highlight the key drivers, offering a concise view of the variables shaping the model's predictions.

"Delivered newsletter current" emerged as the most influential feature, suggesting that customer interactions, even without recent spending, are critical for differentiation. Spend-related features, including "Online retail revenue current" and "Secondary ticket revenue via forwards current," highlight the importance of recent purchasing activity in shaping customer behaviors.

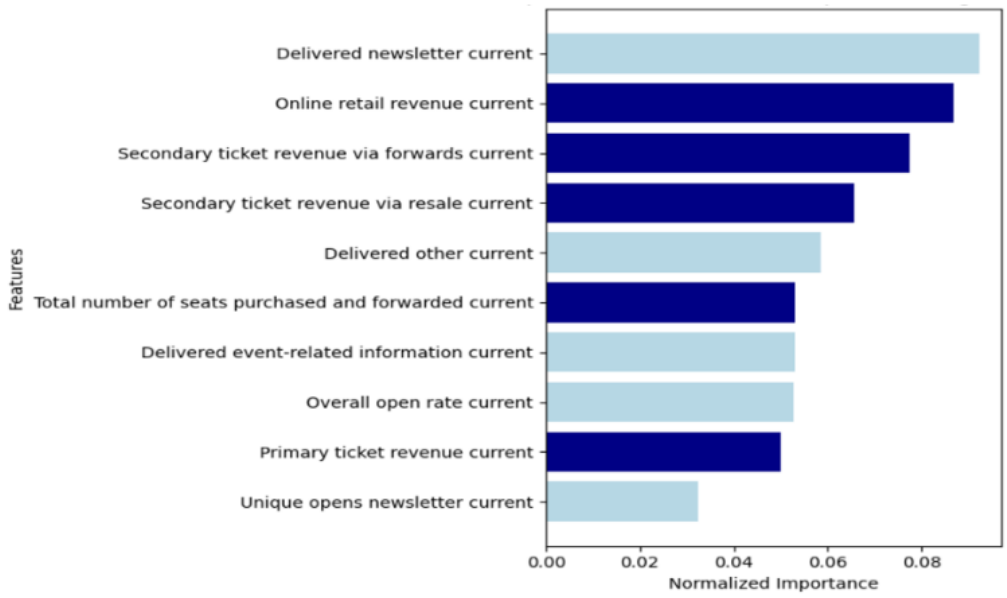

Figure 7.6: Top 10 Normalized Feature Importance for LightGBM Model

These findings emphasize the value of integrating engagement and financial data to analyze customer behavior. This balanced approach can inform targeted strategies, enhances model performance, and align with a broader customer journey analysis by capturing how touchpoints and behaviors shape progression through the fan lifecycle.

7.2.3 Model Performance Across Spend Buckets

Figure 7.7 displays a confusion matrix comparing predicted and actual spend categories. The model excelled in mid- and high-spend tiers ($2,501–$5,000 and $5,001+), with predictions

closely aligning with true labels. For the highest spend tier, it achieved high precision, correctly identifying over twenty-three thousand customers. Predictions in mid-spend brackets ($251–$1,000 and $1,001–$2,500) also showed strong consistency.

However, challenges arose in lower spend categories ($0 and $1–$100), where predictions occasionally fell into adjacent tiers. This may reflect the subtler patterns in lower spend behaviors, which are harder to distinguish. Future iterations could address this by incorporating more nuanced behavioral features or refining segmentation strategies.

Overall, the model's strong performance in higher spend categories highlights its value for CJA, supporting strategies to ramp up these segments. Insights from errors in lower spend tiers offer opportunities for improvement, enabling better identification of emerging spending behaviors.

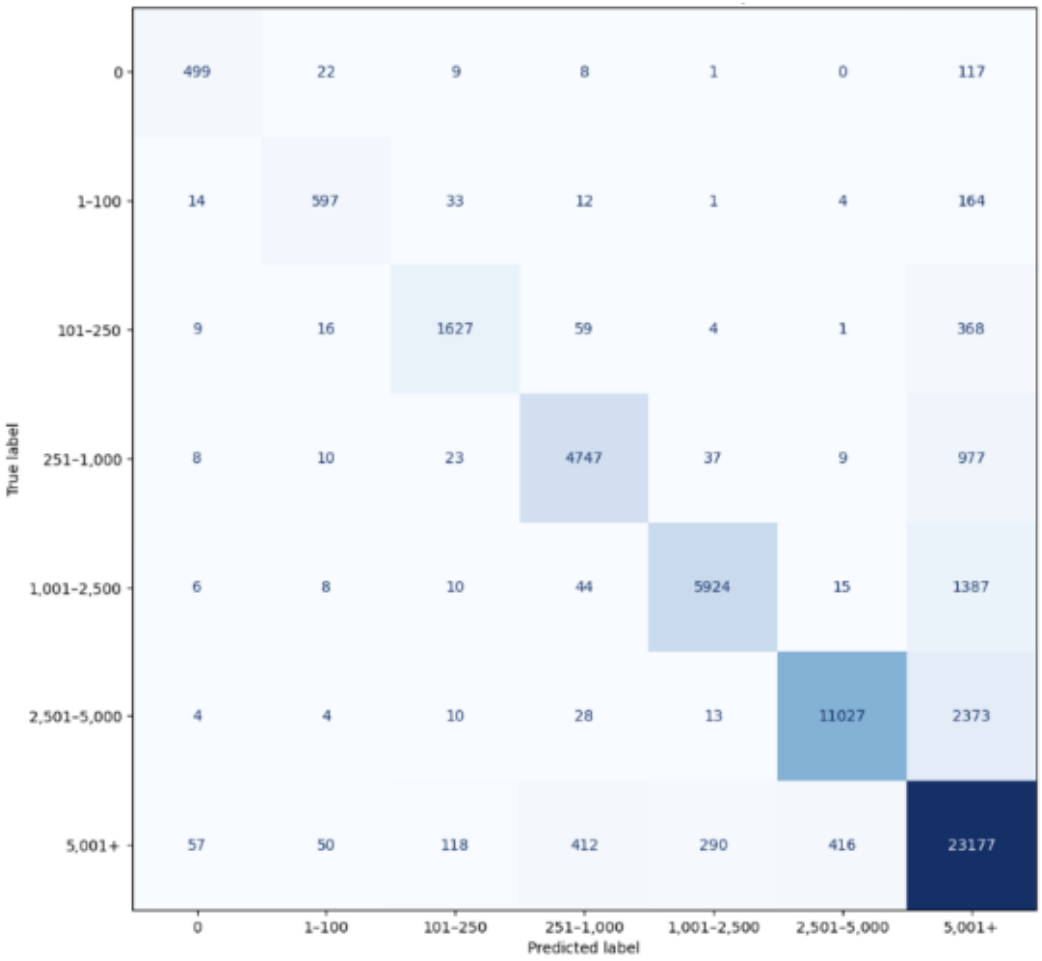

Figure 7.7: Confusion Matrix for Predicted vs. Actual Spend Buckets in Classification Model

7.3 Limitations of this Analysis and Challenges

While the clustering and classification methodologies demonstrated strong performance and provided actionable insights, several limitations should be acknowledged to contextualize results and guide future improvements.

7.3.1 Clustering Challenges

While clustering proved effective in segmenting the fanbase, the lack of predefined labels (as it is an unsupervised approach) introduced challenges in validating the cluster assignments. Without ground truth data, cluster quality was assessed using interpretive metrics like inertia and expert domain knowledge. This limitation means that while the segments are meaningful, their accuracy could benefit from future validation with external data or through supervised methods.

The clustering process was also sensitive to feature scaling and centroid initialization, though PCA mitigated some issues by reducing dimensionality and emphasizing key patterns. Nonetheless, the potential for instability highlights the need for careful preprocessing and iterative refinement in future analyses.

7.3.2 Classification Challenges

The classification model faced challenges due to the imbalanced distribution of spend buckets. The concentration of customers in lower spend categories made it harder for the model to distinguish behaviors, introducing potential biases in underrepresented tiers despite strong overall metrics like AUC-ROC.

Additionally, relying on prior season data effectively captured recency effects but may have limited generalizability to atypical or emerging behaviors. Future iterations could address this by incorporating regularization techniques or expanding training data diversity as more data becomes available.

7.4 Additional Consideration: AutoML for Classification

AutoML frameworks like AutoGluon were evaluated as alternatives to manually-tuned models for classification tasks. AutoGluon's Weighted Ensemble L2 model achieved a marginally higher test accuracy (87.94%) than LightGBM (87.71%), but the 0.23% improvement came with significant trade-offs in computational demands and interpretability.

The ensemble combined base models, including LightGBM, XGBoost, and CatBoost, optimized through AutoGluon's automated tuning. Despite its strong predictive performance, the reliance on complex ensemble techniques limited transparency into feature contributions and model behavior, making it less suitable for providing actionable insights. Additionally, the high computational cost of training and predictions further reduced AutoGluon's practicality for deployment. Given the minimal performance gains and the emphasis on delivering interpretable, efficient, and stakeholder-aligned models, the manually-tuned LightGBM was ultimately selected for its balanced approach to these priorities.

While AutoGluon remains a promising tool for automation and benchmarking, its utility

in this context was constrained by the specific needs of the analysis. Future use cases may benefit from AutoML's capabilities in scenarios where interpretability and computational efficiency are less critical concerns.

Chapter 8

Strategic Insights and Business Implications

This chapter serves as a discussion section, synthesizing the findings from preceding analyses to provide actionable insights and strategic recommendations. Building on the CJA introduced in Chapter 1 and leveraging results from Chapters 2 through 7, these insights provide a targeted roadmap for enhancing fan engagement and loyalty within the team's ecosystem. The proposed strategies focus on translating data-driven findings into practical applications that address both behavioral and financial dimensions of the fan journey.

8.1 Retention and Churn Prevention Strategies

Lifecycle analysis revealed a significant churn risk among fans after their first season, with patterns varying slightly across cohorts and ticket engagement types. For fans who joined in 2021-2022, transition matrices reveal that churn—defined as maintaining no activity after an initial lack of engagement—dominated movements across both primary and secondary ticket resale markets. In the forwarded ticket segment, churn was also the most common outcome, though inactivity—where fans initially generated revenue but did not re-engage—also represented a significant proportion of transitions. Few fans upgraded their spending in subsequent seasons, although some demonstrated potential for future engagement. These trends highlight the need for targeted retention strategies to mitigate churn, address inactivity, and convert high-potential fans into loyal participants.

Fans who joined in the 2022-2023 season exhibit ticket activity patterns during their transition from the first to second season that align with the trends observed in the 2021-2022 cohort. Across this cohort of ticket-engaged fans, churn and inactivity remained the most prominent outcomes, though the patterns varied by initial engagement type. In the primary and secondary resale markets, churn dominated, while in the forwarded ticket segment, inactivity was more common. Although some fans upgraded their spending in subsequent seasons, they represented only a small portion of the overall population.

In addition, as highlighted in Chapter 4, first-season ticket activity patterns further underscore the importance of these strategies. The majority of fans engaged primarily through secondary channels, such as forwarded or resold tickets, with fewer relying solely on primary ticket purchases. This reliance on secondary markets during initial interactions points to a critical need for entry-level engagement strategies. Personalized follow-ups, exclusive offers, or incentives designed to transition secondary-market users into direct or multi-channel participation could play a key role in fostering stronger connections with the team.

These insights, drawn from lifecycle transitions and first-season activity patterns, emphasize the importance of a data-driven approach to retention. By addressing churn risks, re-engaging inactive fans, and nurturing upgrade opportunities, the team can create a robust strategy for fostering loyalty and engagement across its ecosystem.

8.2 Channel Optimization and Cross-Selling

Findings from Chapters 4 and 5 revealed significant differences in channel performance. Email engagement emerged as a strong predictor of ticket purchases, with click-through rates correlating directly with transactional activity. To leverage high-performing channels, embedding merchandise promotions within targeted email campaigns can drive cross-selling opportunities. Conversely, underperforming channels, such as forwarded tickets, require focused strategies. For instance, encouraging voluntary newsletter opt-ins for forwarded ticket users can integrate them into broader engagement funnels and foster deeper participation.

Outdoor plaza events and youth sports camps demonstrated high email open rates, despite representing relatively small populations within the fanbase. Investing in these channels through additional data capture tools, such as QR codes or on-site surveys, is valuable for identifying and engaging new segments. By enhancing the granularity of data collected at these touchpoints, teams can develop more tailored engagement strategies aligned with observed fan behaviors, expanding the potential reach and impact of these initiatives.

As discussed in Chapter 3, the median time between specific engagement sequences, such as "Tickets, Online Forms," reveals opportunities to streamline interactions. For instance, the time lag between purchasing tickets and completing online forms suggests a potential gap in engagement. To address this, the team could improve post-game surveys by prompting fans to sign up for online forms. These surveys can simultaneously gather valuable feedback while encouraging fans to explore additional engagement opportunities, such as subscribing to newsletters or learning about upcoming events.

8.3 Customer Segmentation Insights

Chapter 7 identified three distinct fan segments: occasional fans, merchandise-focused fans, and core fans. These clusters, defined by email engagement and spending patterns, offer a foundation for customizing outreach initiatives. Occasional attendees, who are infrequent in both financial and non-financial involvement, could benefit from personalized promotions like discounted tickets or online merchandise offers. Merchandise-focused fans, who display consistent engagement through online retail purchases, might be incentivized through loyalty programs or bundled merchandise-and-ticket experiences. Core fans, representing the most engaged audience, could be targeted with exclusive perks, such as VIP events or priority access to games, to deepen their loyalty. While this group is the smallest segment, they are the most local. This provides an opportunity to foster community connections through hyper-localized campaigns, such as exclusive neighborhood events or tailored outreach that reinforces their strong ties to the team.

8.4 Year-Over-Year Classification Model for Fan Behavior

The predictive model introduced in Chapters 6 and 7 provides valuable tools for strategy development by analyzing predicted spend bucket classifications among fans. The year-over-year classification model identified key drivers influencing whether fans are likely to increase, maintain, or decrease their spending. This model enables teams to allocate resources toward

high-priority areas, such as targeting fans in lower spend buckets with tailored promotions or renewal discounts. Additionally, if the model predicts a fan to fall into a higher spend bucket but they have not made purchases a few months into the season, targeted interventions could include follow-up calls from sales representatives. These personalized interactions could help re-engage fans, highlight ticketing or promotional opportunities, and align their behavior with the predicted spend potential.

To further enhance our strategic approach, it is essential to understand the underlying factors that drive these spending behaviors. The analysis of feature importance provides critical insights into which variables most significantly influence fan spending classifications. By identifying and prioritizing these key drivers, we can refine our targeting strategies and optimize resource allocation for maximum impact.

8.5 Community Engagement Initiatives

Chapter 3 highlighted the role of localized events in engaging diverse demographics. Initiatives such as family-friendly gatherings or meet-ups near universities can serve as entry points for new fans while strengthening community ties. Digital outreach campaigns tailored to these local events further enhance their impact.

The analysis also underscored the importance of seamless first interactions, such as forwarded tickets or app usage. Optimizing these touchpoints ensures initial experiences lay a strong foundation for continued engagement.

These strategies—rooted in lifecycle transitions, engagement sequences, and customer segmentation—provide a comprehensive framework for enhancing fan engagement and loyalty. By addressing immediate opportunities and preparing for long-term fan dynamics, the team can foster a stronger, more connected ecosystem.

Chapter 9

Conclusion and Future Work

This thesis explores the progression of new fans within a professional sports ecosystem. While there are inherent limitations, the findings provide a solid foundation for understanding fan engagement and spending patterns. To build on this analysis, the following recommendations are proposed to capture additional dimensions of the fan journey. These suggestions aim to deepen insights into engagement without broadening the scope of this research.

- **Social Media Engagement Analysis:** Social media plays a pivotal role in fan engagement due to its accessibility, scale, and versatility. As a non-transactional platform, it provides a low barrier to entry, allowing fans to interact without financial commitment through the options of liking, sharing, or commenting. Platforms such as TikTok, Instagram, X (formerly Twitter), and YouTube offer the opportunity to reach millions of users, including casual fans and potential new followers. Social media also facilitates content distribution, enabling teams to share highlights, behind-the-scenes footage, and live updates to sustain fan interest beyond game days. Additionally, it provides targeted advertising opportunities to promote tickets, merchandise, or events while gathering valuable behavioral data. By fostering direct interaction through polls, contests, and real-time responses, social media strengthens connections between fans and the team by reinforcing loyalty and community.

- **Post-Event Surveys:** Post-event feedback surveys provide a valuable tool for assessing fan satisfaction and uncovering actionable insights. These surveys capture both quantitative ratings, such as Net Promoter Scores (NPS), and qualitative feedback through open-ended responses. A high NPS from a casual fan, for instance, presents an opportunity to upsell them on future events or premium experiences. Conversely, negative feedback allows for targeted interventions to address issues and prevent churn. By combining structured ratings with detailed fan suggestions, post-event surveys offer a comprehensive view of the fan journey and guide enhancements to the overall experience.

- **Broader Non-Transactional Data Usage:** Beyond social media—other forms of non-transactional data, such as app usage metrics or content consumption—offer a comprehensive view of fan activity. These insights can inform indirect monetization strategies, including affiliate marketing and co-branded campaigns.

- **Seasonality and Performance Metrics:** Examining the effects of seasonality and team performance provides deeper insights into engagement dynamics. Understanding how playoff runs, major organizational changes (e.g., venue relocation), or off-seasons influence fan behavior can inform targeted strategies to optimize engagement during different times of the year. For example, analyzing changes in fan activity during significant wins or losses may offer actionable insights for re-engagement campaigns.

Additionally, integrating performance metrics with seasonality data can highlight how specific milestones, including playoff appearances or team successes, influence fan behavior and engagement trends over time.

Future research incorporating lifecycle patterns across decades and insights from tenured fans could reveal strategies for sustaining long-term engagement. These patterns could be correlated with external factors such as team success, marketing efforts, and economic conditions. While this study focuses on newer participants to explore ramp-up and progression, incorporating insights from long-tenured fans would offer a complementary perspective and deepen the understanding of long-term engagement dynamics.

In conclusion, this thesis highlights the dynamic nature of fan engagement and its essential role in sustaining loyalty and long-term success within the sports industry. By introducing the CJA framework, this research goes beyond traditional metrics, namely CLV, to capture the broader and more nuanced ways fans interact with teams. This approach emphasizes the importance of understanding both non-financial and financial dimensions of fan relationships, particularly for net new fans whose initial touchpoints critically shape their progression within the team's ecosystem.

The findings provide a clear roadmap for fostering deeper connections with fans, ramping up their participation and sustaining enduring loyalty between them and the team. By analyzing fan interactions across multiple seasons and acquisition channels, this research demonstrates the value of a holistic approach to engagement—one that integrates non-transactional touchpoints and their impact on long-term brand valuation. These insights offer teams practical tools to create meaningful connections that extend beyond individual seasons and support scalable strategies for professional sports organizations.

References

[1] *Service Quality and Customer Lifetime Value in Professional Sport Franchises*. ProQuest Dissertations & Theses. Accessed 7 Dec. 2024.

[2] K. Strang, ed. *Fandom Analytics*. Accessed 7 Dec. 2024. Springer, 2023.

[3] J. Jensen. "Applying Customer Lifetime Value to Major League Baseball Season Tickets". In: *Journal of Applied Sport Management*. Accessed 7 Dec. 2024.

[4] *Why Sports Companies Should Focus on Engagement & Lifetime Value over Short-Term Revenue*. FT Strategies. Accessed 7 Dec. 2024.

[5] *Unlocking High-Value Football Fans: Unsupervised Machine Learning for Customer Segmentation and Lifetime Value*. ResearchGate. Accessed 7 Dec. 2024.

[6] I. T. Jolliffe. *Principal Component Analysis*. 2nd. Springer, 2002.

[7] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[8] D. Arthur and S. Vassilvitskii. "k-means++: The Advantages of Careful Seeding". In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (2007), pp. 1027–1035.

[9] M. Kuhn and K. Johnson. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2019.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[11] OpenAI. *ChatGPT*. Accessed 7 Dec. 2024. 2024.