13TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 1 - 2, 2019 | BOSTON, MA
presented by ESPN

The Horizon of Sports is Digital: Using Fantasy Sports, e-Sports and Electronic Gambling to Find Next Generation of Ticket Buyers

Business of Sports
Paper ID - 13378
Authors: Peter Lenz, Matthew Sabban and Peter Ibarra

1. Introduction

Sports teams and leagues are constantly looking for their next-generation of ticket buyers. As we progress further into the digital age, the existing methods of converting fringe followers into ticket purchasers are losing their effectiveness. Over the last couple decades, large TV contracts and increased availability for media consumption have made up for this potential loss in in-stadium revenue. [10] However, as linear TV is increasingly threatened by streaming and over-the-top (OTT) service providers [13], sports properties have begun investing large amount of resources into finding new audiences based in the digital environment; namely Fantasy Sports, e-Sports, and e-Gambling. [7][8] This research analyzes the effectiveness of this outreach and tests the hypothesis that these three emerging fields are good investments for sports teams seeking to expand their current pool of in-stadium ticket buyers. We test this by overlapping samples of digital behaviors between NFL and MLB attendees with behavioral models of e-Sports enthusiasts, e-gamblers and fantasy sports players to find which are most likely to convert into purchasers of in-stadium experiences.

2. Methodology

As part of it's day-to-day business, Dstillery maintains a catalog of hundreds of behavioral audience segments [3]. Since sport fandom is a detectable digital signal [5] we selected a set of audiences from Dstillery's catalogue comprising of three behavioral segments (Fantasy Sports, e-Sports, and e-Gambling) and two location-based segments (MLB & NFL Attendees) to represent sports fans and the behaviors we were interested in testing for physical conversion at stadiums. We utilized a random sampling process to down sample our 350 million-device universe into a test audience. We then performed an agglomerative clustering to identify mutually exclusive subpopulations based on similar behaviors. We then calculated behavioral index profiles for each subpopulation against national and seed population baselines. These profiles served as the basis for our analysis and conclusions. Each step of the methodology is described in detail below.

In order to promote reproducibility we selected our analysis methodology prior to collecting any data.

2.1. Data Sources

Data for our experiment originated from three primary sources: real-time bid requests from ad-monetized sites, non-monetized web traffic from third party data providers and app usage data from software development kit (SDK) integrations. This unique combination of data is crucial to our being able to accurately sample online and mobile location behaviors [2]. Real time bid requests (BRQs) occur when a device appears on an ad-monetized website or an app. When that event occurs, a call is sent from the site/app publisher as an opportunity to fulfill an advertisement slot. The call contains information such as an advertising device identifier (cookie, IDFA, AAID), a timestamp, an IP address, the publisher's name, the ad category and location data. Not all fields are available in all instances of a BRQ. Third party desktop and mobile app data streams are acquired via licensing agreements from applications where users have opted-in to provide visitation data in return for the functionality of the application. These data sets allow us to have a broader understanding of online behavior beyond ad-monetized sites and apps. For convenience, we use the term "BRQ" to refer to records collected both through the RTB bidstream and through SDK integrations, as the type of information collected in each is similar.

In order to maintain a coherent view of a user across multiple screens, we maintain a probabilistic network of connections between digitally connected devices, also known as a device graph.[2] With this graph, we can determine which mobile devices are connected to which desktop devices. This provides a more robust view of a user's online behavior as they switch devices and locations throughout the day.

Dstillery applies a suite of data quality [4] and anti-fraud [6] filtering during the data collection process. These have been described in a set of publications [1][2][15] and U.S. Patents.

2.2. Behavioral Model Based Device Selection

To generate behavioral models, as used in our seed population, we select a set of websites (the 'seed set') that are highly indicative of a distinct behavior. For example, *www.mlb.com*, *www.baseball-reference.com*, and *www.fangraphs.com* are URLs whose content focuses on the collection and sharing of player statistics for Major League Baseball. In building a behavioral model on this hypothetical "Sabermetricians" seed set, our system calculates predictive indices for each URL observed in the profiles of devices that visit the seed set websites. The top half-million of these predictive indices forms the features of our behavioral model. [1]

For this experiment, we selected 10000 devices from each of the Fantasy Sports, e-Sports, and e-Gambling behavioral models.

2.3. Location Based Device Selection

Location, shared as a latitude/longitude, is one of the key signals collected from mobile apps through BRQs or our third party licensing agreements. For each stadium in this study, we identified the center of the stadium to represent the centroid and created a geofence around it; translating a single point into a specifically crafted polygon [15]. The geofence is purposefully designed to contain the entirety of the stadium grounds.

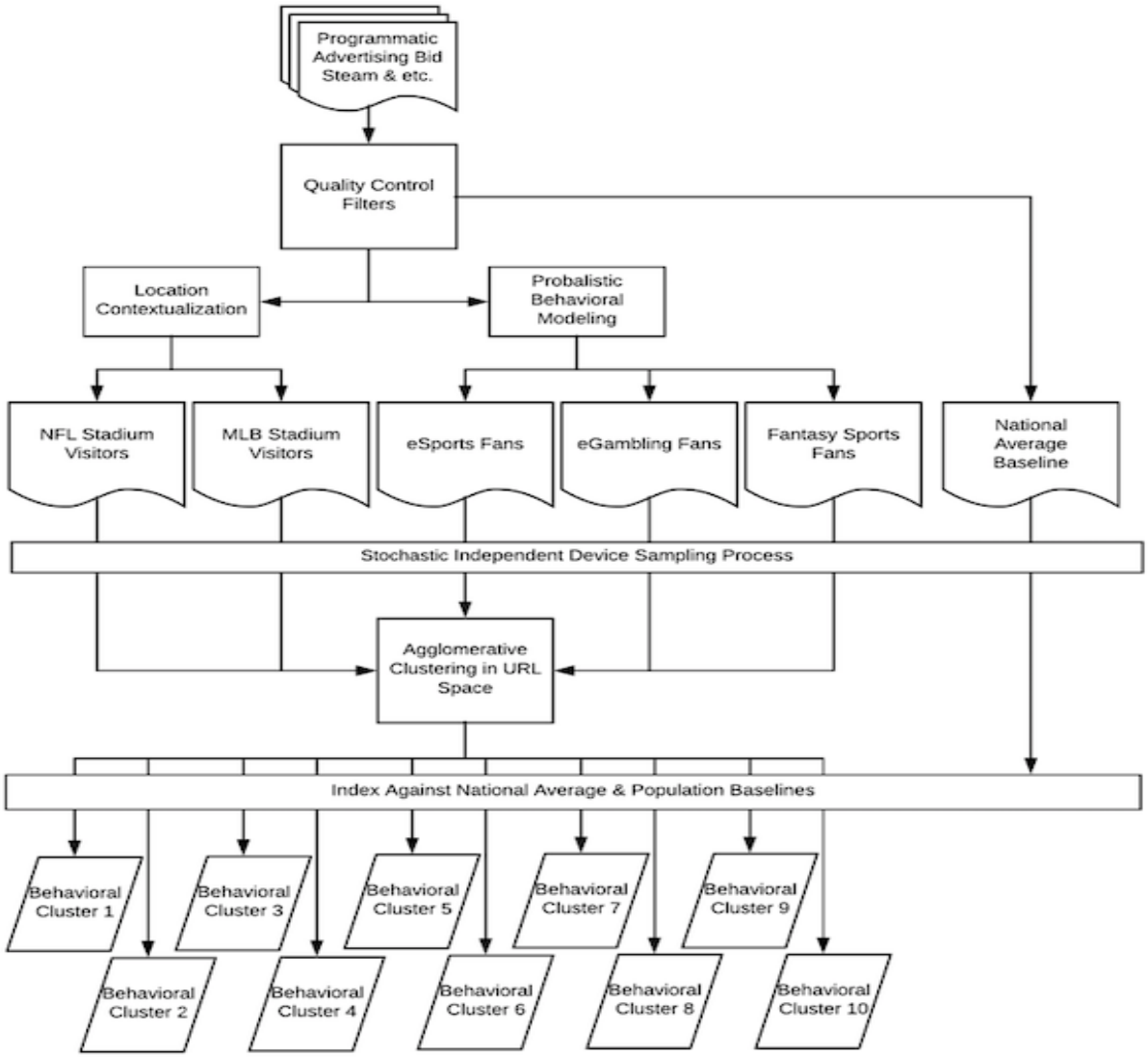

Figure 2.1

For every mobile BRQ and third party data point containing geospatial data, we matched the shared location data to our stadium geofence. If the location data fell into the polygon dimensions, we qualified that data for inclusion in our experiment. In the interest of privacy and following the automated location quality and contextualization step, the raw location data was deleted, leaving only the point of interest data, in this case "MLB Stadium" or "NFL Stadium". [4]

We randomly selected 5000 devices observed at NFL and MLB stadiums, during game times, for a combined 10000 device group of game attendees. Selecting from multiple sports was important to prevent the behaviors related from one sport from dominating the results.

2.4. Stochastic Independent Device Sampling Process

Dstillery's system is programmed to evaluate which devices can be used in modeling. Our observational period ran from August 30, 2018 to September 15, 2018 that resulted in 40,070,299 unique device IDs with qualifying characteristics. Due to hardware constraints, the clustering technique outlined in section 2.5, can only consider 40000 devices and requires a down sample of our total population. The number of devices we selected per audience is outlined in figure 2.2. We then applied a pseudo-random number generator to each device to generate an ordered list of devices against the pseudo-random number and selected the first *N* devices for each characteristic.

We used this same technique to randomly select an additional 40000 devices observed during the observation period to serve as a baseline for comparison. We refer to this sample as the national baseline.

Audience	Type	Device Count	Comment
e-Sports Fans	Probabilistic Behavioral	10000	
e-Gambling Fans	Probabilistic Behavioral	10000	
Fantasy Sports	Probabilistic Behavioral	10000	
MLB Stadiums	Location-based	5000	Only collected during game time
NFL Stadiums	Location-based	5000	Only collected during game time

Figure 2.2

2.5. Agglomerative Clustering in URL Space

Prior to our experiment we created a 128-dimensional URL space, structured such that distance within this space was representative of how related two URLs are in terms of sequence of device visitation. The closer URLS are in the embedding space, the more similar the URL's are in their content. Conversely, the farther apart two URLs are, the less related the topics. [14] To create this, we took the time stamped, ordered histories of 430648822 devices and ran them through a Convoluted Neural Network using the Skipgram training algorithm. This resulted in a data set where each input URL has a 128-dimension vector that represents that URLs location within the embedded space.

To reallocate our seed population into subpopulation groups, we used each device's full visitation history to calculate the device's location in the embedded URL space described above. We did this by averaging the vectors of the visited URLs within the 128-dimensional space to generalize the content the device is interested in. Once placed into the same embedded URL space, we can use the same distance properties to understand the similarity of device's behaviors. [9]

From that, we calculated a distance matrix of each device to every other device in the URL space using cosine similarity. We then performed an agglomerative clustering on the distance matrix using Ward's method with ten randomly selected devices as cluster seeds. This resulted in ten mutually exclusive clusters of similarly behaving devices.

The methodology of this step is patent pending and a separate manuscript exploring this technique in fuller depth is being prepared for publication.

2.6. Behavioral index profiles

We use the behavioral audiences described in section 2.2 as a set of descriptors in order to understand and describe each subpopulation. For a given subpopulation, we create a behavioral index profile by calculating the index of that subpopulation against a set of behavioral audience segments, relative to a baseline population. We build these profiles using two different baseline populations: 1) the seed population itself and 2) the national baseline described in section 2.4. In English, the behavioral index of a given subpopulation for a behavioral audience fills in the blank in this sentence: "A device in this subpopulation is X times more likely to be a member of this behavioral audience, compared to a baseline population." The calculation of the index is described below.

We begin by assigning the behavioral audience memberships of each device through the modeling process described in section 2.2. The empirical probability of behavioral audience membership is determined by examining segment memberships for each unique user seen within our seed population. [2] We then calculate the percentage of devices from a subpopulation that are members of each behavioral segment. That is, the probability P of inclusion in a behavioral segment j for a subpopulation i is given by:

$$P(j \mid i) = \frac{Z_{j,i}}{N_i}$$

where $Z_{j,i}$ is the number of devices found in both behavioral segment j and subpopulation i, and N_i is the total number of users in subpopulation i. This probability can be interpreted as the propensity of behavioral affinity j for users in subpopulation i.

Similarly, we calculate $P(j)$, the baseline probability of inclusion in behavioral segment j, for the relevant baseline. We can then calculate an behavioral index for this behavior and subpopulation relative to the baseline:

$$Ij,i = P(j \mid i) / P(j)$$

This process is done twice for each subpopulation and behavior, once for each baseline (national and seed population). In this way, we create two observation profiles, referred to as National and Population observations, respectively. By indexing against the independent, 40000 device random sample described in section 2.4, we discover differences between our subpopulation and the national audience. By indexing subpopulations against the seed population, we are able to discover nuances between the subpopulations.

3. Results

The agglomerative clustering analysis on the forty thousand devices produced ten mutually exclusive subpopulations. The percent breakout for each subpopulation is displayed in figure 3.1. The 'label' column indicates the subpopulation number. The '# of Devices' column displays the sum of devices within the subpopulation. The ' Percent Size' column calculates the percent of devices for the subpopulation

Subpopulation Label	# of Devices	Percent Size
1	132	0.33%
2	8044	20.1%
3	524	1.3%
4	631	1.6%
5	8406	21.0%
6	4288	10.7%
7	1222	3.1%
8	1359	3.4%
9	8549	21.4%
10	6842	17.1%

Figure 3.1

Visualizing the clustering between the subpopulations is shown in figure 3.2. The x-axis represents the devices relationship to one another as determined by the measured distance along the y-axis as described in section 2.5. At 0.0, each device resides in its own space. The agglomerative clustering algorithm calculates the spatial relationship between the devices/subpopulations and groups similar sets together.

In the first part of our analysis, we calculate the size, of each subpopulation to ensure minimum thresholds are met to provide statistically significant results in subsequent analysis. Subpopulations 1, 3 and 4 do not meet our 1000 device threshold required for producing significant results and are disregarded for the purposes of this study.

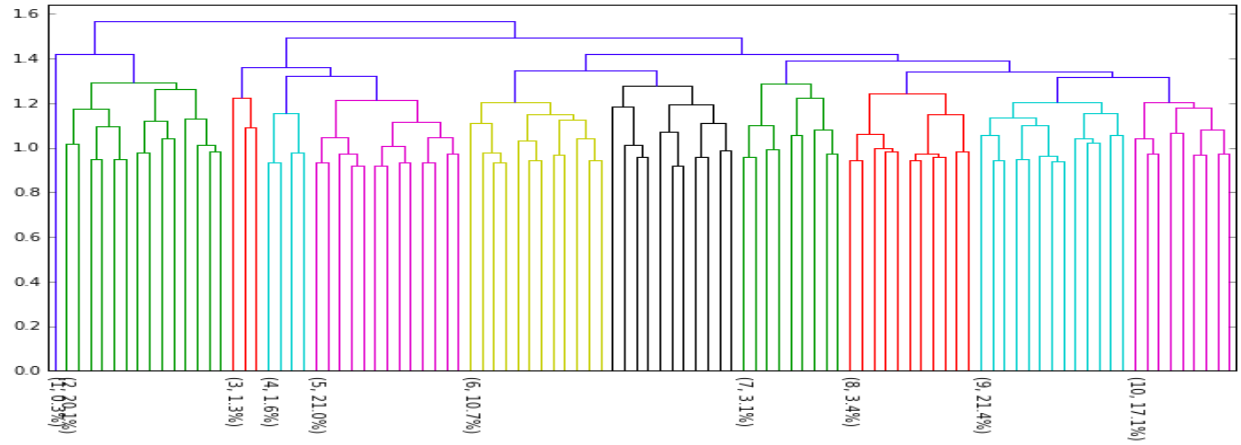

Figure 3.2

Subpopulations were further broken down to understand the percent of devices that derived from each of the five seed-set audiences. As figure 3.3 shows, the percent by audience breakout varied significantly across the subpopulations. Broadly, subpopulation 2 had the most distinct behavioral profile, as it is the subpopulation with the highest percent of devices from a single seed seet audience (e-Sports). This suggests the behavioral composition of the e-Sports seed-set is distinct from the other four seed-set audiences (Sports Gambling, Daily Fantasy Sports, MLB & NFL Attendees). The rest of the subpopulations showed a more diverse audience composition.

Subpop #	NFL Attendees	MLB Attendees	Daily Fantasy Sports	Sports Gambling	e-Sports
2	1.4%	1.7%	3.6%	13.5%	79.8%
5	16.3%	16.1%	20.4%	42.6%	4.6%
6	5.2%	16.8%	33.2%	37.5%	7.3%
7	11.0%	8.3%	30.0%	46.9%	3.8%
8	33.9%	7.6%	14.4%	27.2%	16.9%
9	16.7%	17.6%	40.5%	24.1%	1.1%
10	15.2%	14.8%	24.3%	14.7%	31.0%

Figure 3.3

Our last goal is to measure behavioral segment participation rates for each of the subpopulations to measure the potential effectiveness of driving new ticket purchasers.

4. Discussion

In each of the below figures, the national observations were derived from indices relative to the national average. The indices against the seed-set population are shown as Population Observations. Both sets of indices are needed to provide insight on the propensities of a subpopulation and to understand the niche differences between the subpopulations. The behavior labels within the observations are derived from Dstillery's list of publicly available audiences. [3]

Subpopulation #	**National Observation**	**Population Observation**
Subpopulation 2	• Video Games (9.91x), Role Playing Games (9.58x) • College Sports (1.01x) • Fantasy Sports (1.0x) • Professional Sports (0.64x)	• Japanese Anime and Manga (4.65x) • College Sports (0.85x) • Professional Sports (0.48x)

Figure 4.1

Subpopulation 2 consists of e-Sports enthusiasts, gamers, and followers of "Geek Culture". They actively follow and play competitive video games from major publishers (AAA titles), tabletop card games, and Massive Multiplayer Online games (MMOs). This population is diverse in their video gaming interests, not skewing towards any particular genre. In addition, Subpopulation 2 is an avid enthusiast of Science Fiction/Fantasy books, movies, and comics. Lastly, Subpopulation 2 is interested in "*Otaku*" content, being fans of Japanese Comics and Animation (Manga/Anime). [11]

Subpopulation 2 Conclusion

As Subpopulation 2 under-indexes across multiple sports properties, it is not a fit for driving new ticket purchasers.

Subpopulation #	**National Observation**	**Population Observation**
Subpopulation 5	• Horse racing (5.23x) • Fantasy Sports (2.45x) • Collegiate Sports (3.33x) • Golf (3.2x), Bowling (2.14x) • Finance (1.5x), Luxury Travel (1.2x), Cigar Aficionado (1.94x) • Video Games (0.80x)	• Conservative Politics (1.49x) • Collegiate Sports (1.56x) • Video games (0.38x)

Figure 4.2

Subpopulation 5 consists of hardcore gamblers whose highest indexing activity is Horse Racing and Horse Betting. Subpopulation 5 engages with collegiate and professional sports content as a likely means to research their gambling interests. They are interested in easy going recreational activities such as golf and bowling, implying an older demographic. Other interests of Subpopulation 5 include investment and financial news, luxury travel, and cigar aficionado.

Subpopulation 5 also indexes highly against Fantasy Sports, possibly as another means to gamble. Video games and general e-Sports culture severely under-indexes against this group.

Subpopulation 5 Conclusion

Subpopulation 5 exhibits traits that make them good candidates for future ticket purchasers.

Subpopulation #	National Observation	Population Observation
Subpopulation 6	• Horse racing (4.86x), Gambling (1.5x) • BBQ (2.88x), Country Music (2.1x) • Sports (2.44x), Sports Apparel (2.13x) • Fantasy Sports (1.75x) • Vacation and Travel (2.31x) • Video Games (0.5x)	• Video Games (0.40x) • e-Sports (0.32x)

Figure 4.3

Subpopulation 6 consists of gambling enthusiasts that demonstrate high affinity towards activities such as Horse Racing, Horse Betting and Gambling. Subpopulation 6 follow sports at both the collegiate and professional level and are Fantasy Sports enthusiasts, perhaps as means to further their gambling interests. Overall, subpopulation 6 exhibits very similar traits to subpopulation 5 however; there is lower engagement with investment and finance activities. Subpopulation 6 maintains a high affinity towards Vacation and Travel but under-index for video games and e-Sports.

Subpopulation 6 Conclusion

Subpopulation 6 is a good candidate for in-stadium ticket purchasers.

Subpopulation #	National Observation	Population Observation
Subpopulation 7	• Hunting/Trapping (10.0x) • Trucking (9.26x) • Guns (9.2x) • Fishing (8.67x) • Country Life (8.33x) • Sports (2.33x) • Fantasy Sports (0.74x) • Video Games (0.63x)	• Trucking (7.39x) • Hunting/Trapping (7.0x) • Power Tools (6.45x)

Figure 4.4

Subpopulation 7 is interested in Sports and outdoors, over-indexing on content related to hunting, off-road trucking and fishing. They are fans of various sports properties at both the collegiate and professional level but under-index for Fantasy Sports, video games and e-Sports. Similar to subpopulations 5 and 6, gambling indexes highly for this audience.

Subpopulation 7 Conclusion

As shown in figure 3.3, subpopulation 7 does not consist of many MLB/NFL attendees despite the interest and fandom of sport entities. Due to the subpopulation's small size, it is inconclusive whether they are prime candidates for ticket buying.

Subpopulation #	National Observation	Population Observation
Subpopulation 8	• Sneakers (8.98x), Celebrity News (6.12x), Boxing (5.78x) • Sports (5.53x) • Fantasy Sports (3.34x) • Video Games (1.25x) • E-Sports (0.99x)	• Sports (7.95x) • Sneakers (6.85x) • Celebrity News (6.3x)

Figure 4.5

Subpopulation 8 shows interest in Sports and Fantasy Sports more than the average consumer. They are interested in Video Games but not the types of games that are prevalent among e-Sports, as seen in subpopulation 2. The distinguishing feature of this group is the over-indexing for sneaker and celebrity news, suggesting more mainstream consumer tendencies.

Subpopulation 8 Conclusion

While subpopulation 8 consists of attendees to NFL events and is fans of local sports team, it is inconclusive on whether or not this population is ideal for becoming ticket purchasers due to its size.

Subpopulation #	National Observation	Population Observation
Subpopulation 9	• Fantasy Sports (5.96x) • Professional Sports (3.71) • Collegiate Sports (4.33x) • Finance (2.53x), Investment (2.32x) • Luxury Travel (1.43x) • Video Games (1.0x)	• Finance (3.11x) • Sports (2.32x)

Figure 4.6

Subpopulation 9 consists of diehard Fantasy Sports fans indexing at the highest among all the subpopulations. This group follows sports, from the local to national level, at the highest propensity. Similar to subpopulation 5, this group shows a high affinity towards finance and investment, indicating disposable income for various activities. Unique to this subpopulation, there is overlapping propensity for both sports and casual video game behaviors.

Subpopulation 9 Conclusion

The high-indexing interest in sports properties and fantasy sports make this subpopulation a prime candidate for investment.

Subpopulation #	National Observation	Population Observation
Subpopulation 10	• Video Games (2.24x) • Libertarian (3.22x), Conservative (2.99x) • Fantasy Sports (2.18x) • Sports (2.08x) • MMOs (1.79x)	• Computers, DIY/Computer Parts (1.5x) • Sports (1.08x), Fantasy sports (0.85x)

Figure 4.7

Subpopulation 10 plays video games more than the average user. They stick to Massively Multiplayer Online games (MMOs) or other big title video games from major publishers but are not necessarily the hardcore competitive gamers observed in subpopulation 2. This audience is full of enthusiasts in the Information Technology sector, researching the latest hardware for their gaming systems. Affinities indicate that subpopulation 10's prefers PC gaming but is also consumers of home video game consoles. Compared to the national average, subpopulation 10 is technically savvy and avid followers of politics.

Subpopulation 10 Conclusion

In relation to sports properties, these users have interest in Fantasy Sports and general Sports news, at 1.5x the national index. Results show subpopulation 10 indexes highly for their local sports teams and is an ideal audience for becoming ticket purchasers.

5. Conclusion

In summary, the e-Sports centric subpopulation 2 showed the lowest behavioral overlap with in-stadium attendees. Based on this research, we do not believe e-Sports is an ideal channel for finding future ticket purchasers. In contrast, subpopulations 5-10 resulted in a diverse behavioral composition of fantasy sports, e-gamblers, and in-stadium attendees. We conclude that e-gambling, as demonstrated in subpopulations 5 and 6, is the best audience for investment in finding future in-stadium ticket purchasers with Fantasy Sports being a viable alternative.

We present a forward-looking methodology into how teams should spend their investment resources and expand the understanding of how a new market can be targeted with precise and effective messaging. An in-depth understanding of a potential acquisition audience, before the allocation of money and resources, will only increase the likelihood of success. We believe our experiment is a first step in shedding light onto an area of investment that is difficult to quantify. Finally, we envision applications beyond the study of audiences. We've demonstrated the model's ability to measure the interests of a subpopulation without the need for first party data sources. Adding additional data sources, such as purchase data, would only increase the effectiveness of this methodology and allow us to provide increased specificity in the analysis of results.

6. Acknowledgement

We would like to thank Dstillery and, most especially, the entire Dstillery Data Science team, including our Chief Data Scientist, Dr. Melinda Han Williams, for their invaluable feedback and insight. Their willingness to assist and answer any questions we had in this project ensured our results were of the highest quality. We also thank Abby Beltrani for her insight into the business of sports marketing industry. Lastly, we thank our families for their constant and unwavering support.

References

1. Wojan, Timothy, Crown, Daniel, Slaper, Timothy, Lenz, Peter, Bianco, Alyssa. "Are the Problem Spaces of Economic Actors Increasingly Virtual? What Geo-located Web Activity Might Tell Us about Economic Dynamism". *65th Annual Meetings of the North American Regional Science Council*. NARSC, 2018.
2. Slaper, Timothy, Alyssa Bianco, and Peter Lenz. "Digital Vapor Trails: Using Website Behavior to Nowcast Entrepreneurial Activity." *2nd International Conference on Advanced Research Methods and Analytics (CARMA 2018)*. Editorial Universitat Politècnica de València, 2018.
3. "FULL AUDIENCE TAXONOMY" *dstillery.com* Dstillery, 2018. https://audiences.dstillery.com/taxonomy
4. Williams, Melinda Han, Ori M. Stitelman, and Rodney Alan Hook. "Evaluating authenticity of geographic data associated with media requests." U.S. Patent No. 9,948,733. 17 Apr. 2018.
5. Kim, Joon K., and Kevin Hull. "How fans are engaging with baseball teams demonstrating multiple objectives on Instagram." *Sport, Business and Management: An International Journal* 7.2 2017: 216-232.
6. Stitelman, Ori M., et al. "Methods, systems and media for detecting non-intended traffic using co-visitation information." U.S. Patent No. 8,719,934. 6 May 2014.
7. Fischer, Eric "Millennials Put Ticket Strategies To Test" *Sports Business Daily*, 2015. https://www.sportsbusinessdaily.com/Journal/Issues/2015/06/08/In-Depth/Ticketing-main.aspx
8. Shaikin, Bill, Mitchell, Houston. "MLB Becomes Third Major Sports League To Form Partnership with MGM" *Los Angeles Times*, 2018. https://www.latimes.com/sports/mlb/la-sp-mlb-mgm-partnership-20181127-story.html
9. Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. "A simple but tough-to-beat baseline for sentence embeddings." *International Conference on Learning Representations*, ICLR, 2017.
10. Leitch, Will. "Nobody's Going To Sports in Person Anymore. And No One Seems To Care." *New Yorker Magazine*, 2018. https://nymag.com/intelligencer/2018/07/nobodys-going-to-sports-in-person-and-no-one-seems-to-care.html
11. Hiroki, Azuma, Jonathan E. Abel, and Shion Kono. "Otaku: Japan's database animals." *Trans. Jonathan E. Abel and Shion Kono. Minneapolis: University of Minnesota Press*, 2009.
12. Dalessandro, Brian, et al. "Scalable hands-free transfer learning for online advertising." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.
13. Keating, Cal. "Over the Top or over the Heads of Sports Broadcasting: Sports and Entertainment Content Licensing and Distribution in a New Era." *Sports Law. J.* 25, 2018: 177.

14. Abadi, Martín, et al. "Tensorflow: a system for large-scale machine learning." *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. Vol. 16. 2016.
15. Lenz, Peter, Ibarra, Peter. "Using Digital Signals To Measure Audience Brand Engagement At Major Sports Events: The 2015 MLB Season." *10th MIT-Sloan Sports Analytics Conference.* MIT, 2016.