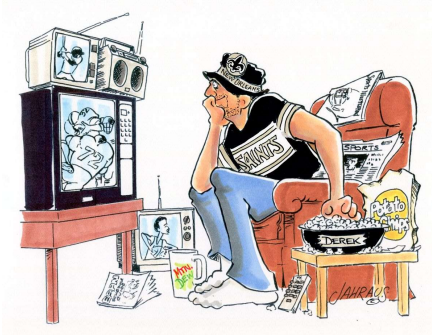


Predicting Sport Fan Behaviors

Erdos Institute Data Science Bootcamp, Summer 2024

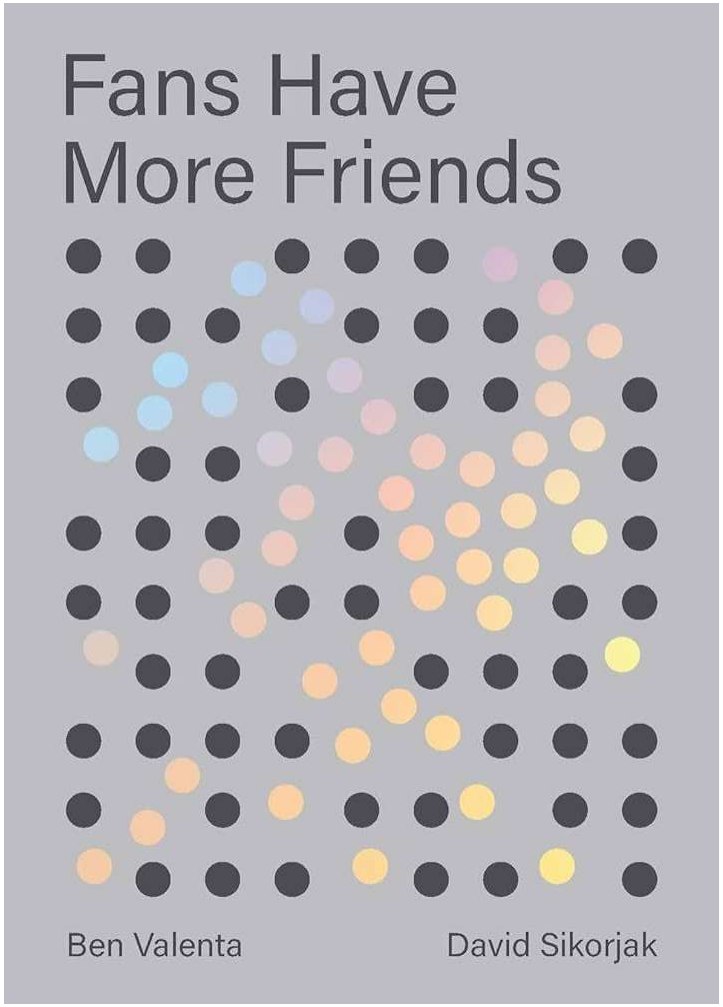
Jacob Kepes, Ryan Moruzzi Jr., Rudy Perkins,
Deepisha Solanki, Sujoy Upadhyay, & Mariah Warner

Research Question/Project Goal & Stakeholders

- **Which sports fans engage in which types of fan behaviors/activities/consumption?**
- Why does this matter?
  - Targeted Marketing Strategies
  - Increased/Better Fan Engagement
- Stakeholders
  - Sports Teams/Leagues
  - Media Companies/Streaming Services (FOX, ESPN, Amazon, NBC, etc.)
  - Sports Betting Companies (DraftKings, FanDuel, etc.)

Data

- Survey data collected in 2021-2022 by Fox Sports Analytics via Prodege
- Representative sample of U.S. Adults
- Following data cleaning, N=10,362
- Over 100 survey questions



Variables of Interest - Outcomes

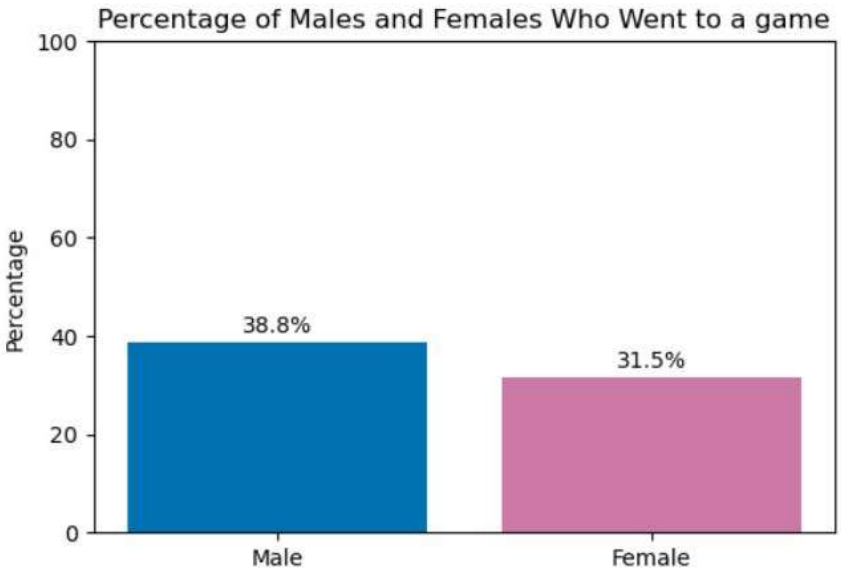
VL1: Thinking of the last year, which of the following activities have you done in conjunction with the sports you follow?		
Values: 0-1		
	0	Unchecked
	1	Checked
	[VL1r1]	Went to a game
	[VL1r2]	Watched a game at a sports bar
	[VL1r3]	Watched a game at a friend's home
	[VL1r4]	Placed a bet through a sportsbook/casino/bookie
	[VL1r5]	Listened to sports radio
	[VL1r6]	Called into a sports radio show
	[VL1r7]	Wore my team's jersey
	[VL1r8]	Watched games at home
	[VL1r9]	Talked about games in person/over the phone
	[VL1r10]	Talked about games online
	[VL1r11]	Played in a fantasy sports league
	[VL1r12]	Purchased a multi-game ticket package (e.g., season tickets, partial plan) for a professional sports team
	[VL1r13]	Bet in a group pool, e.g., football squares, survivor leagues
	[VL1r14]	Played daily fantasy

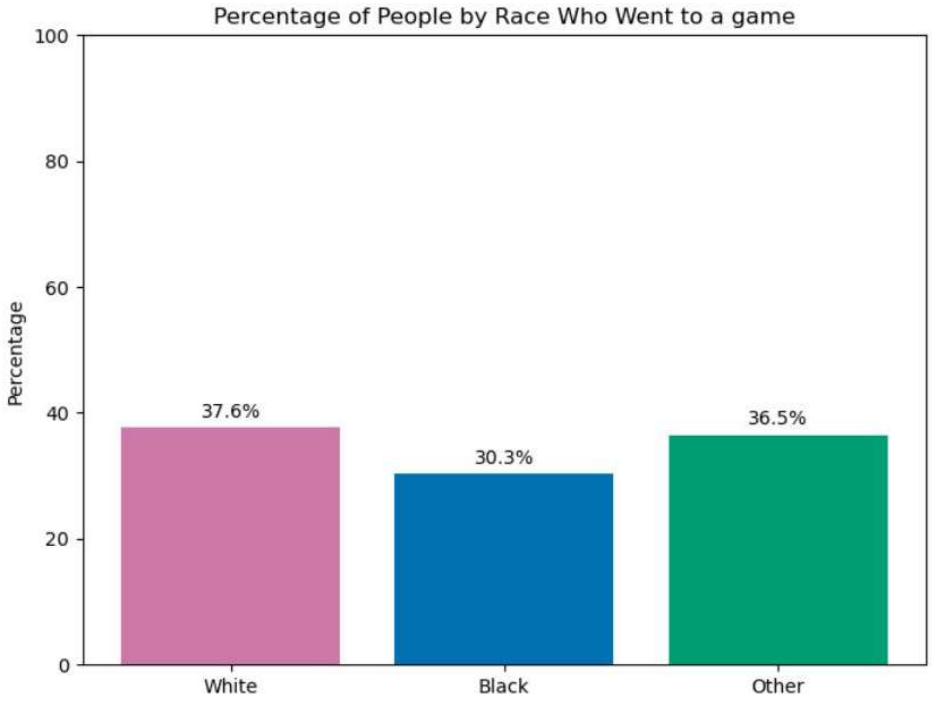
Variables of Interest - Features/Independent Variables

- Gender (male, female)
- Race (white, Black, other)
- **Age**
- **Income (6 level categorical variable)**
- **Overall fan magnitude**
- Hours spent watching sports per week
- NFL team interest

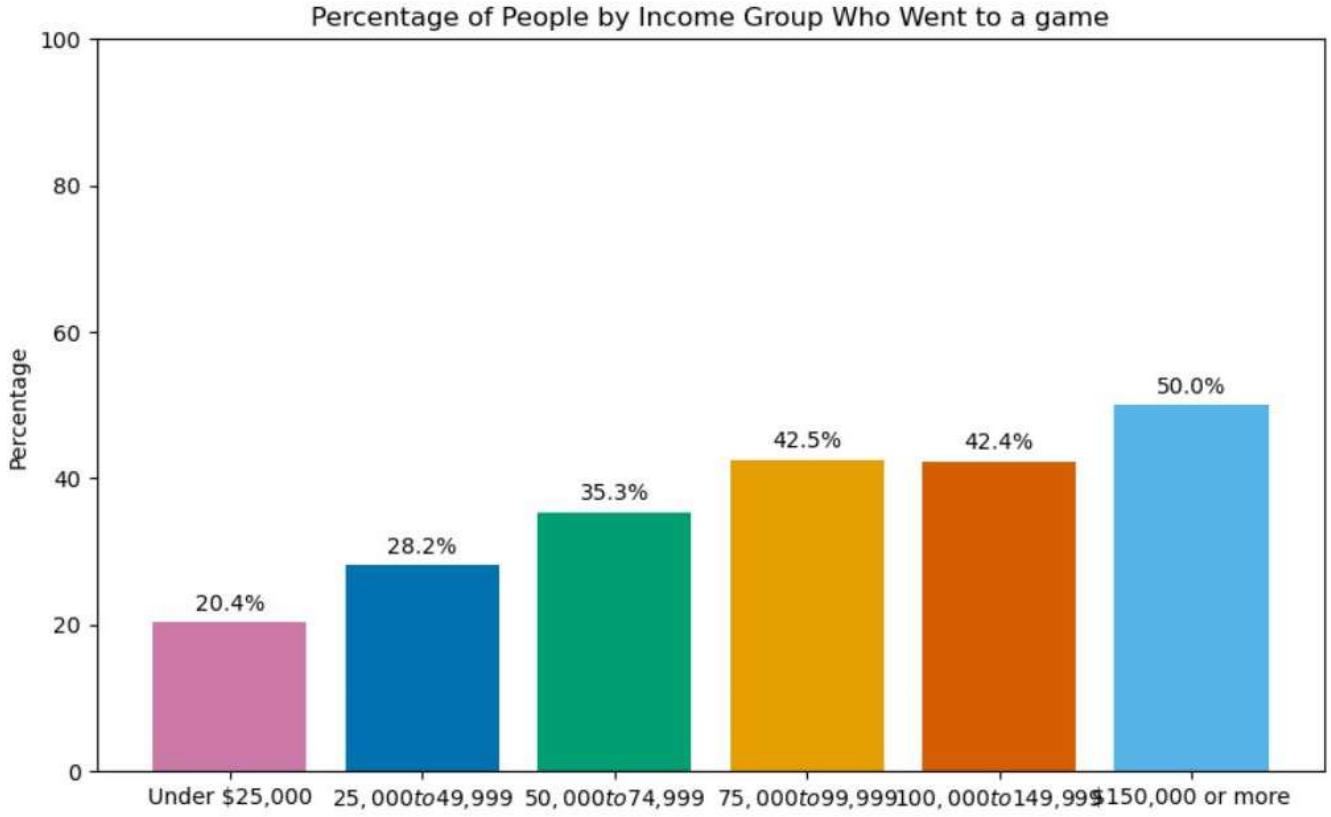
S15: How much of a fan are you of the following sports?		
Values: 1-6		
	1	1[pipe:prog_GCTP]Not a fan of this sport
	2	
	3	
	4	
	5	
	6	6[pipe:prog_GCTP]Obsessed fan of this sport
	[S15r1]	NFL
	[S15r2]	NBA
	[S15r3]	College Football
	[S15r4]	College Basketball
	[S15r5]	MLB
	[S15r6]	NHL
	[S15r7]	International Soccer
	[S15r8]	MLS
	[S15r9]	Combat Sports
	[S15r10]	NASCAR
	[S15r11]	Formula 1

Exploratory Data Analysis

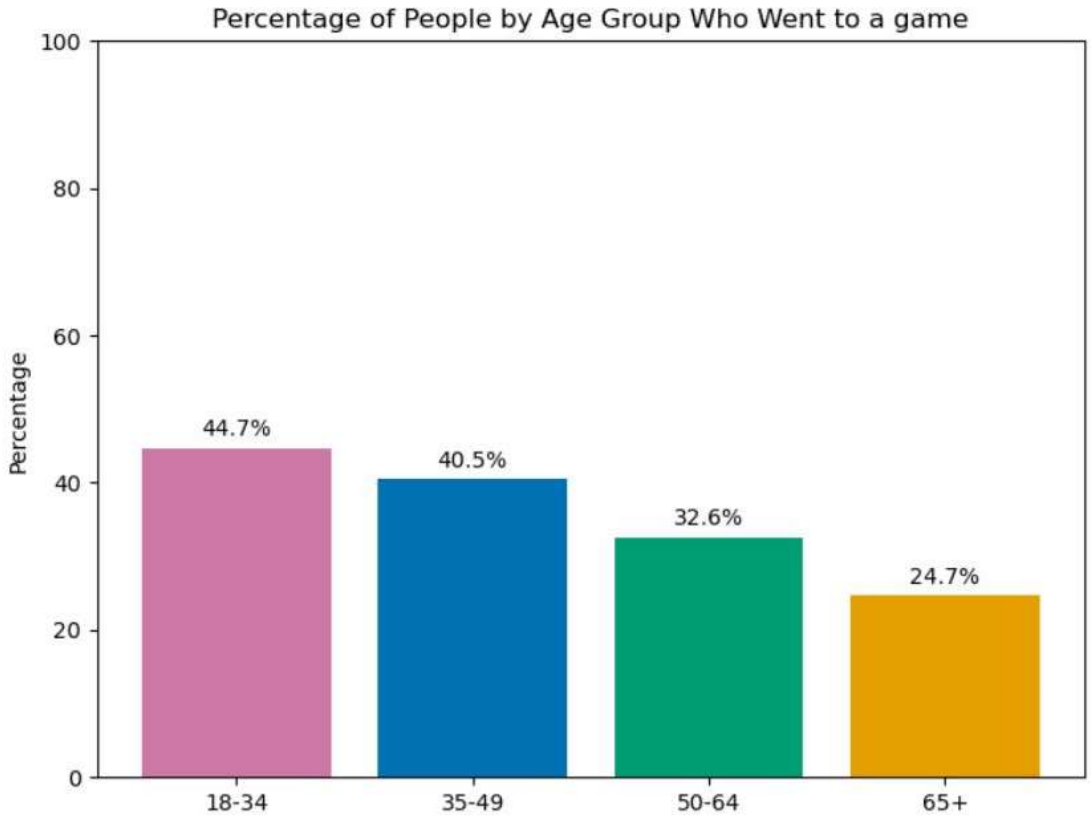




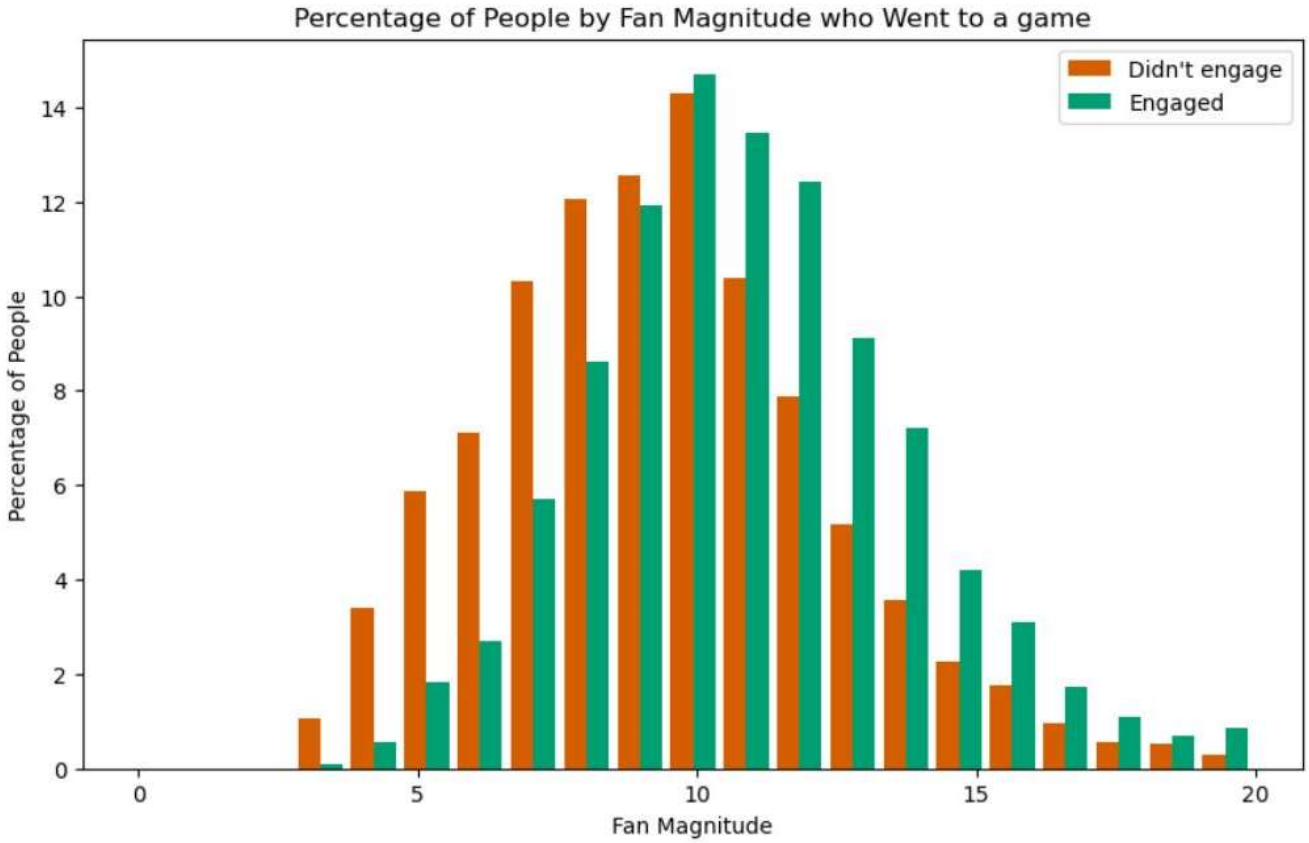
Exploratory Data Analysis cont.



Exploratory Data Analysis cont.



Exploratory Data Analysis cont.



Modeling Approach

- Took 80:20 train / test split (stratified over VL1s/Fan Behaviors)
- Restricted to most relevant continuous features suggested by EDA: S2 (Age), D4 (Income Bracket), Fan Magnitude, (self reported fan score), etc…
- Model output: For each of the fan behaviors selected, we trained a model to predict whether or not an input will engage in that fan behavior/VL1.
- Classification models we trained: Log. Reg., Knn, RandomForests, Adaboost, XGBoost, a few feed forward NNs.
- Compared models' accuracy scores to a random coin-toss baseline using stratified 5-fold cross validation.
- All performed similarly, with Knn being worst.
- Adaboost had the best accuracy scores by a very slim margin.

Final Model Performance

- Choosing AdaBoost as our model, we tuned the hyperparameters for each fan behavior [VL1] question.
- Our model predicted almost all fan behavior responses with 64% accuracy or greater.
- The model performed the best to the question:

[VL1r12]: Thinking of the last year, did you purchase multi-game tickets in conjunction with the sports you follow?

Which was predicted with 92.76% accuracy. What was more interesting was that each feature, Age, Income, and Fan magnitude, were of almost equal importance in the model for this question.

Final Model Performance for each VL1 question

VL1: Thinking of the last year, which of the following activities have you done in conjunction with the sports you follow?	Prediction accuracy	Feature importance: [Age, Income, Fan Magnitude]
VL1r1: went to a game	65.60%	[0.2229, 0.0457, 0.7314]
VL1r2: watched a game at a sports bar	64.06%	[0.3467, 0.1822, 0.4711]
VL1r4: placed a bet through a sportsbook/etc	78.39%	[0.3857, 0.1486, 0.4657]
VL1r5: listened to sports talk radio	63.48%	[0.2109, 0.0145, 0.7745]
VL1r7: wore my team's jersey	59.14%	[0.3707, 0.0185, 0.6108]
VL1r10: talked about games online	72.12%	[0.3333, 0.1867 0.48])
VL1r11: played in a fantasy sports league	76.51%	[0.30 , 0.32, 0.38]
VL1r12: Purchased multi-game tickets	92.76%	[0.332, 0.288, 0.38]
VL1r13: Bet in a group pool	79.31%	[0.2733, 0.06, 0.6667]
VL1r14: Played daily fantasy	83.36%	[0.416, 0.16 , 0.424]

Conclusions/Takeaways

Our predictive model, focusing on age, income and self-reported fandom (Fan_mag), significantly outperforms random chance in predicting fan behaviours. Importantly, Fan_mag emerged as the most crucial variable.

Key insights for marketing strategies:

- Inclusive Marketing
- Target High Fan_mag
- Age-Specific Campaigns
- Income-Based Promotions

Recommendations for advertisers:

Develop predictive models based on demographic information of fans to optimise ad targeting and resource allocation. By leveraging these insights, stakeholders can enhance fan engagement and achieve better marketing outcomes for sports organisations.

Next Steps/Future Directions

- Rather than yes or no predictions for each of the VL1s (fan behaviors), have our models predict probabilities.
- Given that our initial goal was to quantify fandom using various demographics, stratifying model(s) by gender, race, etc. to see how our results hold up along these lines may yield interesting results for markets that advertisers are not currently catering to.
- With the increased attention to sports betting and online behavior, more data could improve a model's prediction accuracy and may be a worthwhile feature to explore for advertisers.