Community-Clusterin §ports Betting

Eddie Kiernan

Introduction

- Sports betting networks provide a rich landscape for studying social interactions and behavioral patterns
- Goal of this project is to identify sports bettors who share similar betting behaviors using:
  - K-Means Clustering
    - Assign betters to distinct gambling archetypes based on summary-level user metrics
  - BigCLAM Community Detection
    - Assign community membership to betters based on contest/league-type participation details
- Real-world
  - Sportsbook- targeted promotions increase user engagement and maximize profits
  - Regulators - identifying fraud

Real-world

Sportsbook

- Companies can leverage community and cluster information derived from user behavior to recommend personalized promotions
- Targeted promotions are designed to increase user engagement and ultimately maximize profits
- As online sportsbook have become normalized/legalized, user behavior data has become a key driver of the sports betting industry's explosive growth over the past decade

1

2

Regulator

- Growing concern over the rise of fraud within the sportsbooks and broader sports betting ecosystem
- Identifying fraud is essential to protect the financial interests of platforms, but also to uphold the integrity of the industry and safeguard bettors.
- For example, federal investigation of a gambling ring tied to two NBA betting cases in 2023

Data

- Data retrieved from the Transparency Project, a research initiative under the Division on Addiction at the Cambridge Health Alliance, a Harvard Medical School teaching hospital
- Sourced from leading sportsbook DraftKings, from 2014-2015
- Two main datasets, both with 10,385 user enteris
  - TacklingData5All.csv - metrics across all DFS contests each player participated in, including NFL and non-NFL events
  - TacklingData6Play.csv - types of contests each player entered, particularly focusing on sport and contest type

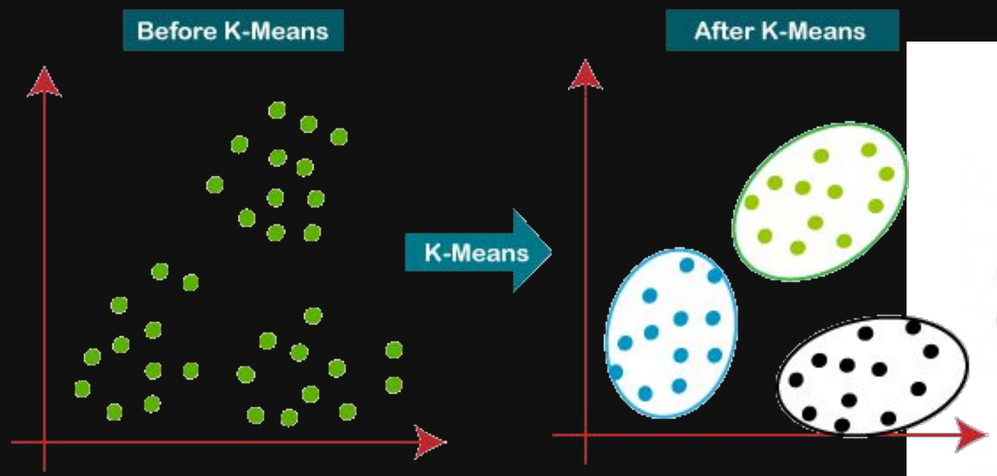
	UserID	DidNFL	DidNBA	DidOth	Cnt1	Cnt2	Cnt3	Cnt4	Cnt5	Cnt6
0	1	True	True	True	0	0	0	0	1	2
1	2	True	False	True	2	0	0	0	0	33
2	3	True	True	True	27	4	0	1	1	6
3	4	True	False	True	137	92	16	19	31	47
4	5	True	False	False	9	0	12	0	0	1
...	...	...	...	...	...	...	...	...	...	...
10380	12036	True	True	True	62	1	0	9	0	15
10381	12038	True	True	True	22	26	8	4	30	103
10382	12039	True	False	False	0	0	0	0	0	1
10383	12040	True	False	True	1	0	0	2	4	176
10384	12041	True	False	True	0	0	1	0	11	37

	Date1st	DateLst	nDays	nCont	nEntries	nLineups	TotFees	AvgBuyIn	TotWinnings	nUserUp	UserID
0	2014-08-28	2015-01-25	3	3	3	3	31.00	10.333333	0.00	0	1
1	2014-09-07	2015-01-25	22	35	35	35	23.50	0.671429	19.25	9	2
2	2014-09-21	2015-01-25	23	39	52	52	1479.00	37.051282	1278.50	15	3
3	2014-09-07	2015-01-25	24	342	355	348	1468.75	4.277778	1399.89	117	4
4	2014-09-28	2015-01-25	12	22	22	22	85.00	3.863636	76.80	10	5
...	...	...	...	...	...	...	...	...	...	...	...
10380	2014-10-02	2015-01-25	26	87	146	87	144.25	0.997126	109.44	39	12036
10381	2014-10-02	2015-01-25	46	193	261	259	6736.84	25.698653	4867.40	46	12038
10382	2014-10-05	2014-10-05	1	1	1	1	27.00	27.000000	0.00	0	12039
10383	2014-10-05	2015-01-25	31	183	317	316	3885.92	13.261858	2357.84	49	12040
10384	2014-10-04	2014-11-09	7	49	54	54	452.25	7.566327	254.50	11	12041

Methods

K-Means

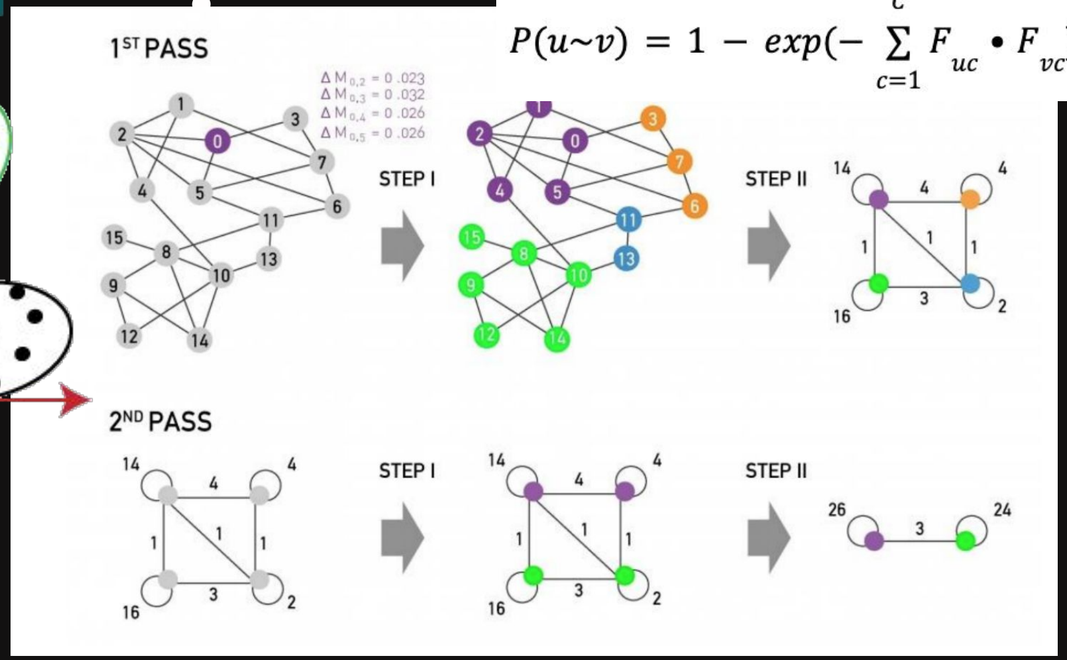
- Used for clustering betters into archetypes
- Uses TacklingData5All.csv
- Elbow method to determine the optimal number of clusters
- PCA used to reduce dimensionality for visualization



BigCLAM

- Used for community detection of betters based on play style
- Uses TacklingData6Play.csv
- Probabilistic model:
  - Poisson edge generation model - assumes edges between nodes arise independently with probabilities determined by the community affiliation strengths

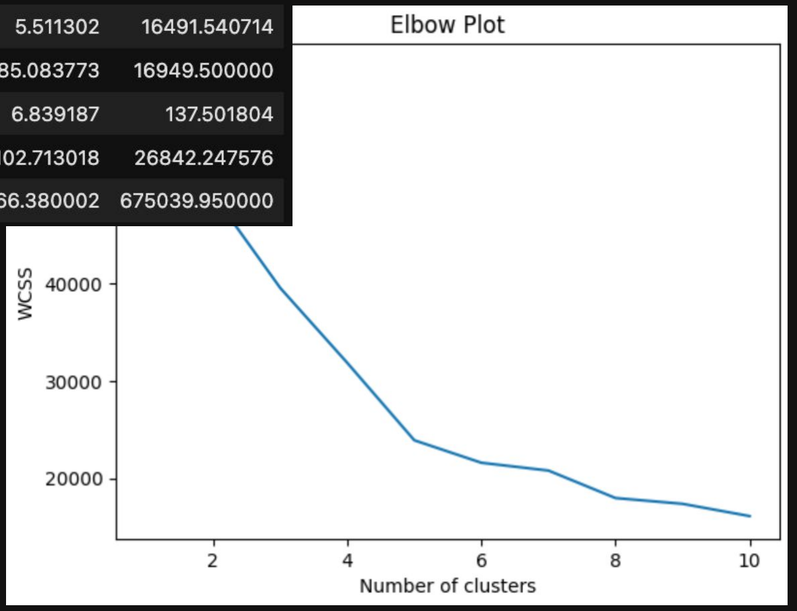
$$P(u \sim v) = 1 - exp(-\sum_{c=1}^{C} F_{uc} \bullet F_{vc})$$



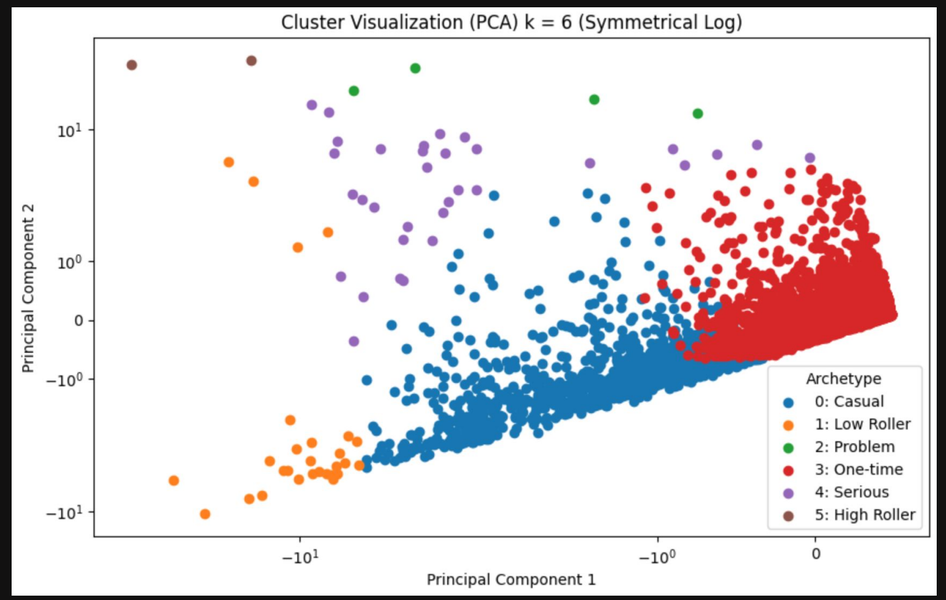
Results

K-Means

Cluster	nDays	nCont	nEntries	TotFees	AvgBuyIn	TotWinnings
0	64.130108	225.201001	327.057548	1444.286756	5.218119	1165.650842
1	93.464286	1945.178571	3375.392857	17122.866071	5.511302	16491.540714
2	12.000000	37.250000	40.000000	23912.247500	785.083773	16949.500000
3	11.553679	27.277991	32.750082	205.849739	6.839187	137.501804
4	73.666667	327.242424	591.181818	26362.224242	102.713018	26842.247576
5	40.500000	2073.000000	2182.500000	173130.875000	66.380002	675039.950000

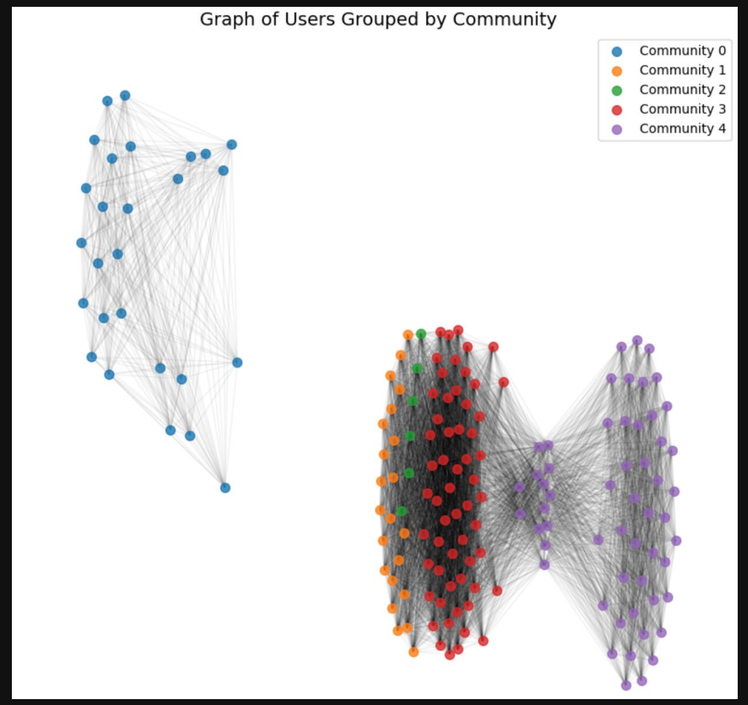


- Casual Bettors (0) - Users participate moderately, placing low-stakes bets. Likely bettors during the regular season or popular events, they exhibit mild engagement without high risk. More successful than one-time betters but not highly profitable.
- Low Rollers (1) - Highly active users, betting small amounts with moderate total spend. Despite low buy-ins, their total winnings are nearly breakeven, suggesting a high skill-to-stakes ratio. These are likely analytical, experienced grinders maximizing ROI from small stakes.
- Problematic Bettors (2) - Bet very infrequently but with extremely high stakes. Despite low participation, they win large amounts ($16,949), implying high volatility. This pattern aligns with sporadic, high-risk behavior, possibly by wealthy individuals seeking thrill bets on major events, or by problem gamblers during relapse phases. Risk-reward profile is extreme.
- One-Time Bettors (3) - Event-driven participants, very low spending Their winnings are modest , suggesting casual, one-off participation likely triggered by promotional offers or major televised sports events like the Super Bowl. Represent the least engaged and least successful segment.
- Serious Bettors (4) - Quite active, with substantial total spend and a moderate average buy-in. Their total winnings slightly exceed fees, suggesting competent and committed players. Representing season-long players or semi-professional users.
- High Rollers (5) - Elite cluster defined by extremely high stakes and volume. These users likely operate at near-professional or syndicate levels, placing many lineups and winning heavily. Represent the most successful, rarest, highest-value users.



Results

BigCLAM



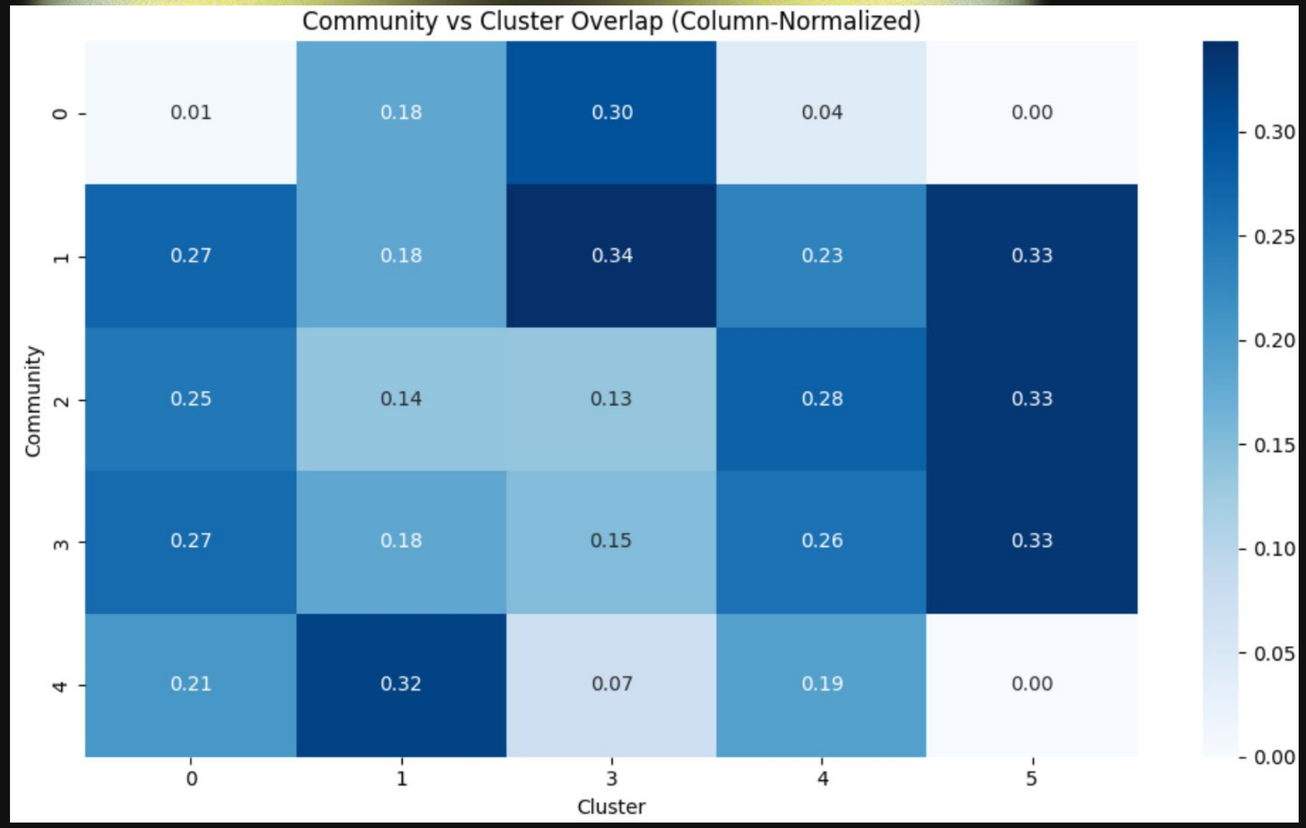
- Community 0 - Users who almost exclusively play in contests. About 39% of users participated in NBA league bets; minimal participation in other sports league bets. No overlap between this community and others.
- Community 1 - NBA participation ~90%, Conservative participation in contests with > 90% in Cnt6. Substantial overlap between this community and 2 and 3.
- Community 2 - All league participation. NBA participation ~100%, Conservative participation in contests with > 90% in Cnt6. Substantial overlap between this community and 1 and 3.
- Community 3 - All league participation. Moderate contributions across Cnt1–Cnt6, especially Cluster 6 (~80%). Semi-diverse contest participation. Substantial overlap between this community and 1, 2, and 4.
- Community 4 - All league participation. Most diverse contest participation. Heavy participation in Cnt1 (50/50 contests) at ~23%, much higher than other communities. And only ~50 participated in Cnt6 (other contests). Moderate overlap between this community and community 3.

	UserID	DidNBA	DidOth	Cnt1_Norm	Cnt2_Norm	Cnt3_Norm	Cnt4_Norm	Cnt5_Norm	Cnt6_Norm	Community
0	8239.0	0.0	0.0	0.074074	0.000000	0.000000	0.000000	0.000000	0.925926	0.0
1	1393.0	1.0	1.0	0.015915	0.007958	0.010610	0.031830	0.031830	0.901857	1.0
2	9444.0	1.0	1.0	0.032258	0.000000	0.019355	0.032258	0.064516	0.851613	2.0
3	9444.0	1.0	1.0	0.032258	0.000000	0.019355	0.032258	0.064516	0.851613	3.0
4	3692.0	1.0	1.0	0.224138	0.048851	0.048851	0.123563	0.057471	0.497126	4.0

Results

Cross-Analysis

- **Casual gamblers almost always bet on the NBA and other leagues and almost never are contest exclusive betters.**
- **Low rollers have diverse contest participation, more often playing more in 50/50 contests.**
- Problem gamblers have little to no discernible gambling pattern, likely due to their impulsive gambling nature.
- **One-time gamblers are all over the map. Split into distinct groups either not betting on leagues at all or other times betting on leagues frequently while consistently playing Cnt6 (other contests). This is likely due to one-time gamblers being targeted strongly by promotions and not betting long enough to exhibit patterns or congregate to a social community.**



- **Serious gamblers almost always bet on the NBA and other leagues and almost never are contest exclusive betters. Very similar to casual gamblers.**
- High rollers are difficult to derive insights from due to this archetype being incredibly rare to begin with. In general, it can be said high rollers often bet on leagues and on the NBA and other leagues and almost never are contest exclusive betters. Similar casual and serious gamblers.

Conclusion

Potential Improvements

Two main sources of improvement

- Better data
  - Data is from a decade ago and consists at times of limited data
  - Growth of DFS and legalization of sports betting in the last decade would no doubt produce better/more relevant results
- Utilization of machine learning techniques.
  - Modern DFS companies leverage machine learning techniques such as neural networks
  - Due to neural networks powerful computation capabilities and modern infrastructure, would likely also result in better results.

Key Takeaways

Gamblers exhibiting similar transaction and betting behaviors formed distinct clusters which were used to classify gambling into archetypes and assign community membership with community detection

Relations between these archetypes and communities lead to insights about betting patterns, and extrapolated with enough user data could be beneficial to to sportsbooks and regulators for identifying users and communities.

Thank You.