Team GTA VI

Cluster Segmentation and Machine Learning: User profiles and percentile behavior

Rockstar Games Datathon, Spring 2022, NYU

Our team

Jaden Stryker, CS-DS, EGD 2023
Rahul Sawhney, CS-DS, EGD 2023
Mateus Aragao, CS-DS, EGD 2023
Carlos Figueroa, ECON-DS, EGD 2023

Table of Contents

Overall experience

- Started exploring the data by finding the unique account_id values in each dataset. We notice that gen_player_stats contained the most accurate amount of individual accounts tracked during that time span of 3 months.

- Both item_spend and activity_played are missing unique account_id entries, because someone user neither bought stuff in the game or played online mode during that period.

- We found some ambiguous columns among the datasets, such as time_spent and daily_playtime, which we thought it would add up to be equal, but turned out they referred to different things (since you can do multiple activities at the same time, and that sums into time_spent by double counting)

- While going through basic data description, we found a lot of noise between variables, which was expected for a time series data analysis

- By using SQL and python modeling packages, we started our data cleaning process early on, and built up our main goals after finishing that process

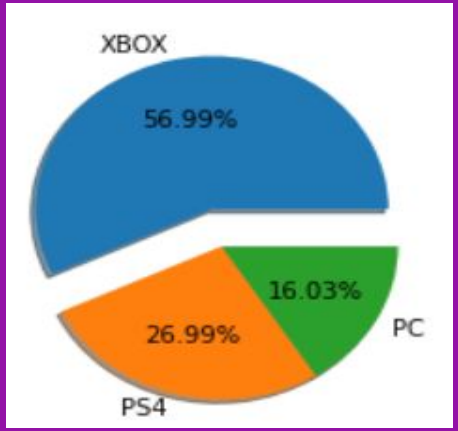
Player Persona

Total amount of players
9527

Players engaging in activities
9288

Players engaging in item purchasing
8990

Consoles they are using

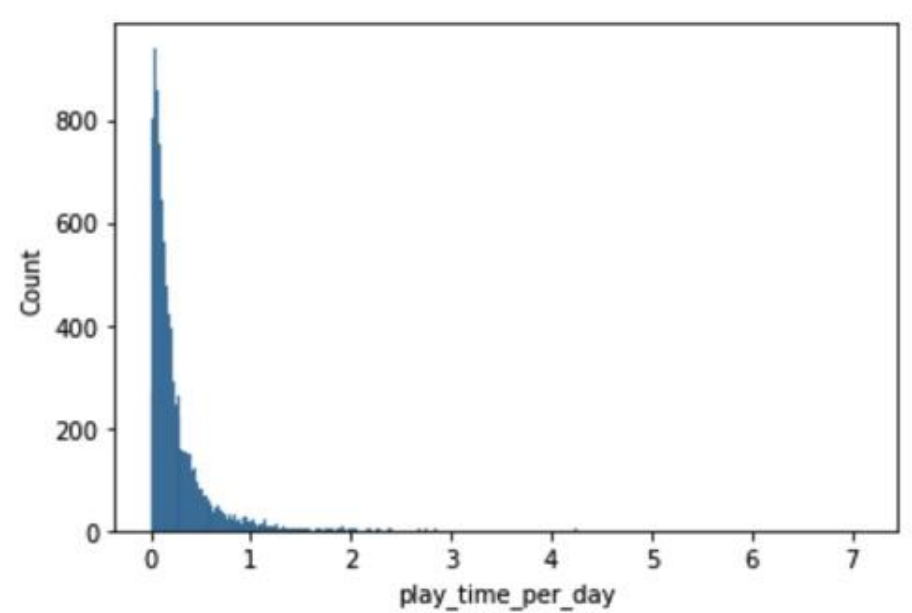


Amount of real money invested by console type

	platform_id	amount_of_users	total_amount_spent
0	PS4	66922	42367.721584
1	XBOX	141314	35761.831574
2	PC	39748	9291.335308

Player Persona (Time)

Distribution of the number of players against hours in a day



Weekly details of time played

count	9527.000000
mean	1.817319
std	2.482100
min	0.001155
25%	0.521480
50%	1.040716
75%	2.123878
max	49.664263
Name:	play_time_per_week,

Details on the overall time frame

count	9527.000000
mean	17.821177
std	24.560485
min	0.014523
25%	5.589537
50%	10.542793
75%	20.591039
max	638.540530
Name:	total_time_playing

Player Persona (Time)

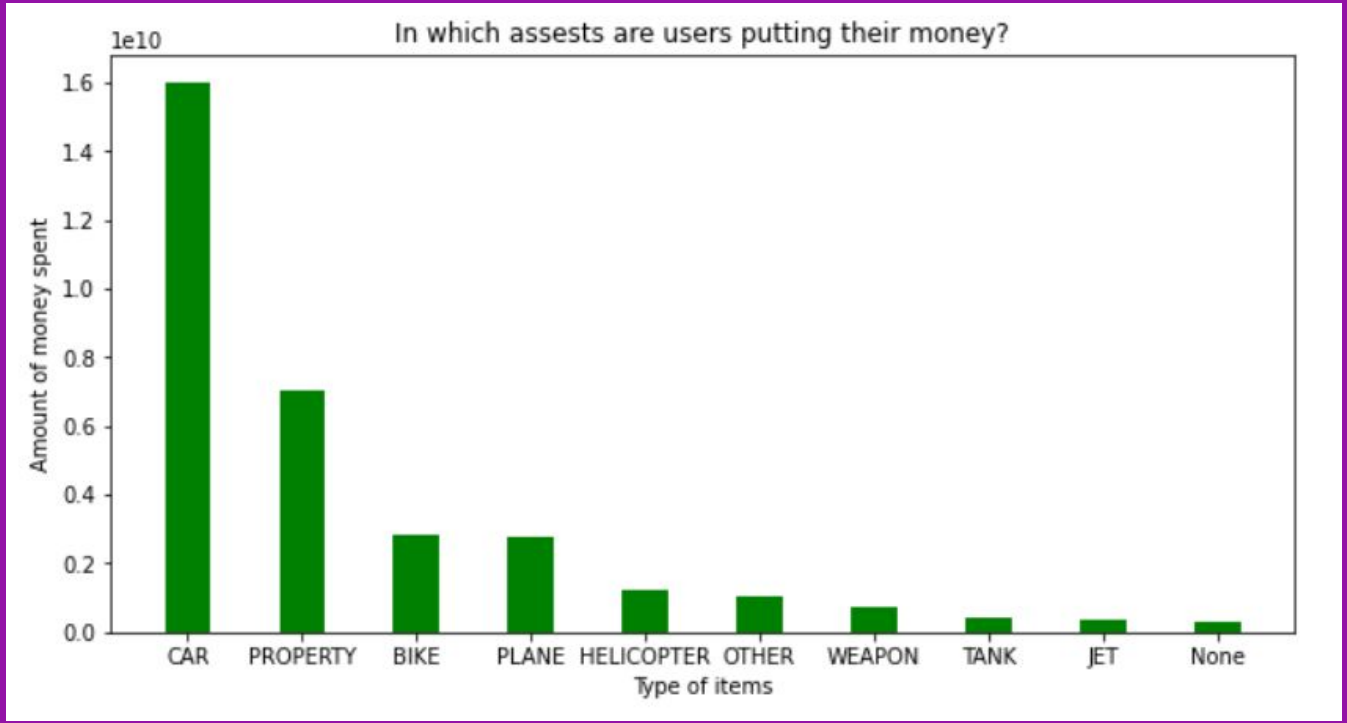
Activities which users spent most of their time (hours)

	activity_type	time_spend_overall
0	Heist	948907.075460
1	Biker	262039.277039
2	Executive	257703.264687
3	Freeroam - Business Battles	142145.552604
4	Race	124376.986262
5	Gunrunner	116972.507441
6	Casino	30574.836939
7	Nightclub Owner	22249.734214

Time spent growing in time (weekly)

```
occur_date
2020-09-06    11820.965338
2020-09-13    13714.129314
2020-09-20    12696.244820
2020-09-27    12059.030282
2020-10-04    11476.344573
2020-10-11    11154.415193
2020-10-18    11295.758984
2020-10-25    11776.030412
2020-11-01    12615.702073
2020-11-08    13033.201169
2020-11-15    14287.873922
2020-11-22    15026.673507
2020-11-29    16843.601905
2020-12-06     1982.383889
Freq: W-SUN, Name: daily_playtime,
```

Player Persona (spending)



Player Persona (spending)

Ten most purchased items

	item_type	specific_item	Amount_spent
0	BIKE	Oppressor Mk II	1.549018e+09
1	PROPERTY	None	8.930873e+08
2	CAR	Deluxo	7.372080e+08
3	PROPERTY	The Aquarius	6.855424e+08
4	PLANE	P-996 LAZER	5.056593e+08
5	CAR	Declasse Scramjet	5.024494e+08
6	CAR	Vigilante	4.951871e+08
7	PLANE	Buckingham Luxor Deluxe	4.172936e+08
8	TANK	TM-02 Khanjali	4.063395e+08
9	JET	Mammoth Hydra	3.640508e+08

Spending through time

occur_date	
2020-09-06	2.082469e+09
2020-09-13	2.404106e+09
2020-09-20	2.292415e+09
2020-09-27	2.125884e+09
2020-10-04	2.218644e+09
2020-10-11	2.570580e+09
2020-10-18	2.445121e+09
2020-10-25	2.179266e+09
2020-11-01	2.343618e+09
2020-11-08	2.791468e+09
2020-11-15	2.612376e+09
2020-11-22	2.877836e+09
2020-11-29	3.444001e+09
2020-12-06	3.183812e+08

Freq: W-SUN, Name: money_spent,

1

Goals & Strategy

Focused on the Cluster Segmentation and Machine Learning challenge.

Our main goals

1

Who enjoys the game?

Can we classify who enjoys the game with the data we have?

2

Engineer insightful features

By creating different variables that emcompasses the main characteristics of the original dataset

3

Segment customers through ML

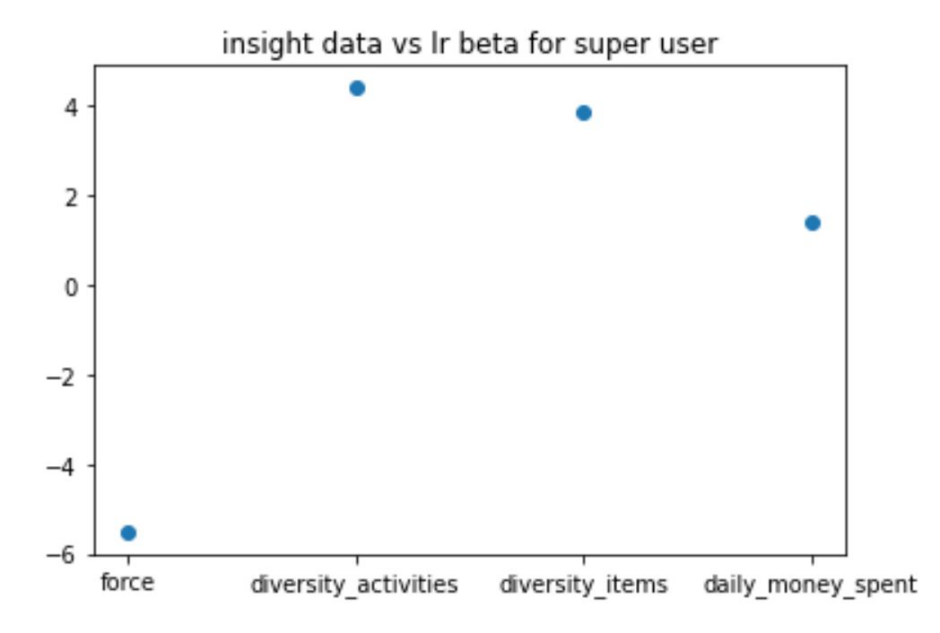
Use these features to generate more accommodate categories of user profile

Feature Engineering

- Play_time_per_day: Measures how long players spend in-game per day
- Diversity_activity: Represents how diverse the events choices of each player are
- Diversity_items: Representing how diverse the purchase choice of each player are
- Modal_activity: Each players most engaged in activity
- Modal_item: Each players most acquired item
- Force: Measuring how often a player is taking a break from the game
- Daily_money_spent: Measures how much in game money players are spending per day
- Deposit: Represents whether the player has spent real money on the game

Super Users and Force

- Goal: Ability to predict whether a user is a super user.
- Super User 90% percentile of time spent per day on average
- Super-users 35 minutes per day
- Average users: 15 minutes as day
- Correlation force-time greater for average user



Descriptive statistics on each cluster, in terms of EVC balance

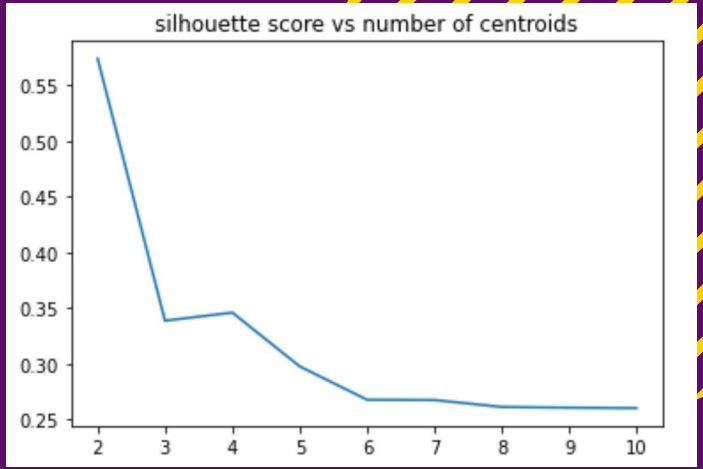
	play_time_per_day	force	diversity_activities	diversity_items	daily_money_spent	deposit	cluster
count	8184.000000	8184.000000	8184.000000	8184.000000	8.184000e+03	8184.0	8184.0
mean	0.245827	3.221918	0.529264	0.246925	5.204020e+04	0.0	0.0
std	0.347261	2.388838	0.238706	0.155533	1.834281e+05	0.0	0.0
min	0.000165	0.000000	0.000000	0.000000	0.000000e+00	0.0	0.0
25%	0.070412	1.584963	0.375000	0.166667	4.909796e+03	0.0	0.0
50%	0.139141	3.000000	0.500000	0.250000	1.470003e+04	0.0	0.0
75%	0.283212	4.954196	0.750000	0.333333	3.413623e+04	0.0	0.0
max	7.094895	9.998590	1.000000	1.000000	6.272233e+06	0.0	0.0

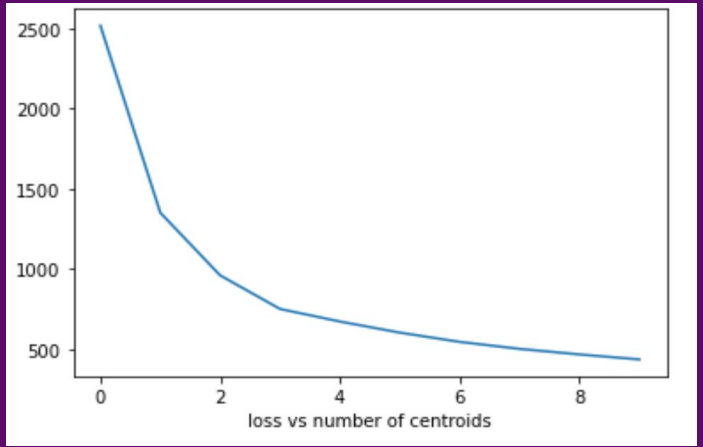
```
c1.describe()
```

	play_time_per_day	force	diversity_activities	diversity_items	daily_money_spent	deposit	cluster
count	1343.000000	1343.000000	1343.000000	1343.000000	1.343000e+03	1343.0	1343.0
mean	0.343650	3.141457	0.586188	0.333457	5.570615e+04	1.0	1.0
std	0.385935	2.588082	0.228987	0.145267	1.555633e+05	0.0	0.0
min	0.002289	0.000000	0.000000	0.000000	0.000000e+00	1.0	1.0
25%	0.107774	1.000000	0.375000	0.250000	1.068986e+04	1.0	1.0
50%	0.218332	2.807355	0.625000	0.333333	2.448588e+04	1.0	1.0
75%	0.422063	5.000000	0.750000	0.416667	5.205404e+04	1.0	1.0
max	3.398076	8.997179	1.000000	0.916667	3.135173e+06	1.0	1.0

Hyperparameter tuning on clusters and logistic regression

- Min and max scales to normalized distribution
- Elbow distribution and Silhouette score
- Class Imbalance
- Weight class for logistic regression
  AUC: .55 -> .77





Recipe for a Super User

Goal

How to create a "super users" based on what we noticed from the dataset

- **Diversity and engagement**
  - Advertise and incentivise exploring the world
  - Reward systems for diversity
- **Repetition in playtime**
  - Create weekly challenges for average user
  - Not imposing penalties on super users

Extra insights and future questions to dig into after our findings

- Which activity does a user first do after not playing for a week or more(coming back to the game)?
- We could identify the behavior of binge players that have high amounts of hours spend but at the same time, weeks of separation between each time they play, and see how they engage with the game.
- Is there a threshold of engagement until the user invests real money in the game?

Thanks!

Do you have any questions?

cdf5579@nyu.edu
+1 9174779237