Game Recommendation system

" The Gradients"

Team Members:

Abjasree S, Poojan Smart, Niranjan Solanki, Vaishnav Panuganti

01 Problem Statement

02 Dataset and EDA

03 Evaluation metrics and Models

04 Results and Future Scope

Problem Statement

- Scrape the data necessary for the model and clean it
- Build recommendation system for recommending games
- Visualizing whether there are any clusters so that we can exploit that property too in building the recommendation system

Dataset

- Data: https://store.steampowered.com/
- User Data: https://steamcommunity.com/games/steam/members
- We have also obtained the user data from the reviews given for each game

Dataset

Game Dataset

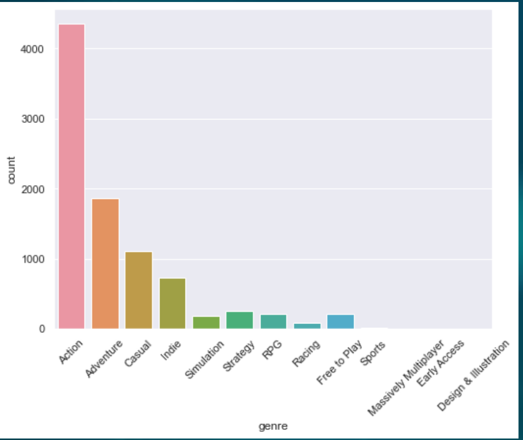
	game_link	price	title	tags	genere
0	https://store.steampowered.com/app/292030/The_...	₹ 800	The Witcher® 3: Wild Hunt	Open World, RPG, Story Rich, Atmospheric	rpg
1	https://store.steampowered.com/app/435150/Divi...	₹ 989	Divinity: Original Sin 2 - Definitive Edition	Tactical RPG, Turn-Based Strategy, RPG, Explo...	rpg
2	https://store.steampowered.com/app/413150/Star...	₹ 479	Stardew Valley	Farming Sim, Life Sim, Pixel Graphics, RPG	rpg
3	https://store.steampowered.com/app/230410/Warf...	Free to Play	Warframe	Free to Play, Action RPG, RPG, Action	rpg
4	https://store.steampowered.com/app/1085660/Des...	Free To Play	Destiny 2	Free to Play, Open World, Looter Shooter, FPS	rpg

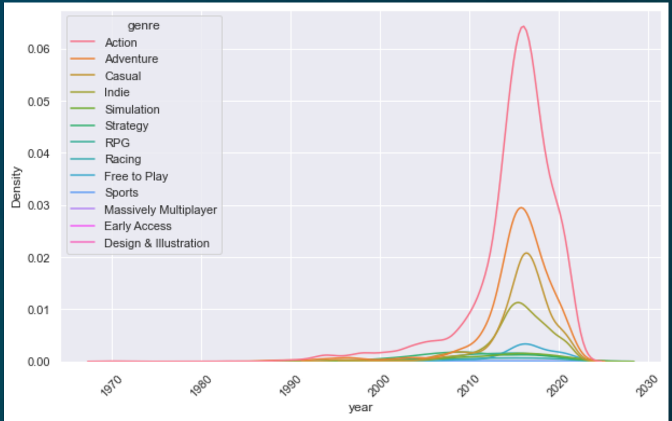
User Dataset

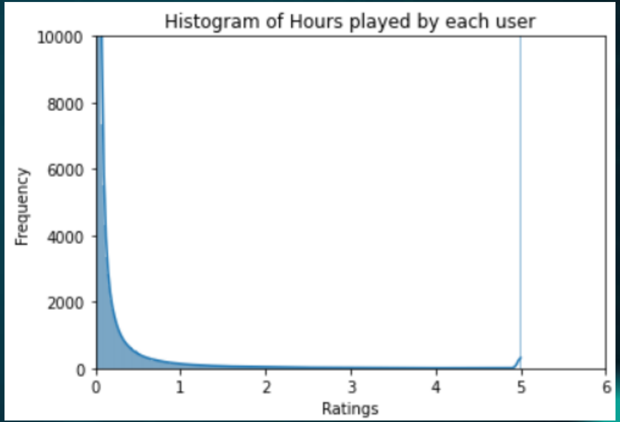
	user	appid	hours_forever	name
0	id/flutterdim	268500	1,466	XCOM 2
1	id/flutterdim	582010	1,434	Monster Hunter: World
2	id/flutterdim	200510	674	XCOM: Enemy Unknown
3	id/flutterdim	262060	578	Darkest Dungeon®
4	id/flutterdim	544750	516	SOULCALIBUR VI

- The appid is the common feature here and we obtain it by cleaning the url in the game dataset
- Hours_forever is used as the implicit feedback and it has either been normalized or rescaled to 0-5

EDA







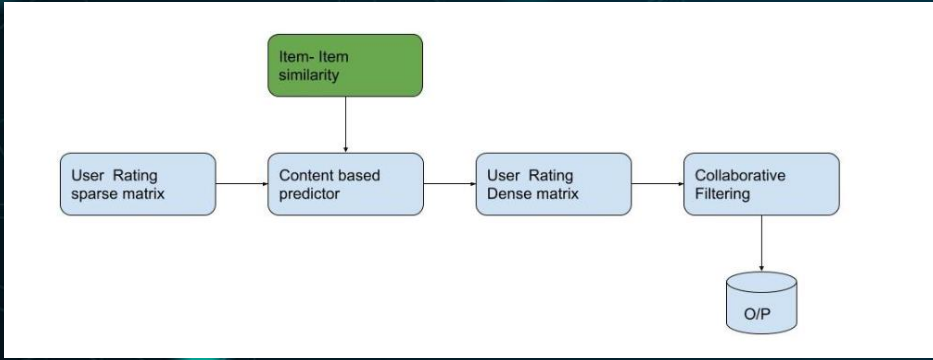
- Action most popular genre
- No. of games in each genre increased over the years
- Most of the values in hours_forever is towards 0

Evaluation metric and Models

Hit-rate: Measure of intersection b/w recommended games and played games

(Ranges from 0 to 1) 0->Bad 1->Good

Content Boosted Collaborative Filtering



Sparse Matrix

40	50	60	70	130	220	240	300	...	1877960	1888430
NaN	NaN	NaN	NaN	NaN	0.00509	0.00376	0.00664	...	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	0.11521	1.03687	...	NaN	NaN
NaN	NaN	NaN	NaN	NaN	0.01876	0.27893	0.02233	...	NaN	NaN
NaN	NaN	NaN	NaN	NaN	0.02503	NaN	NaN	...	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
NaN	NaN	NaN	NaN	NaN	0.04371	NaN	NaN	...	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN

Dense Matrix

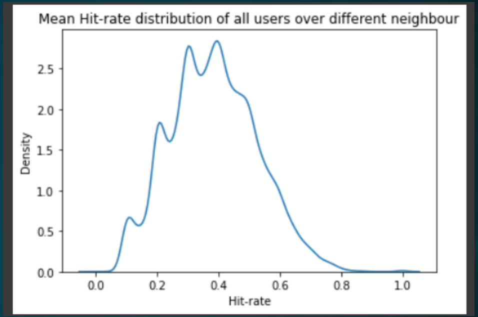
40	50	60	70	130	220	240	300	...	1877960	1888430
0.015592	0.007023	0.007023	0.007023	0.015750	0.005090	0.003760	0.006640	...	0.007023	0.007023
0.020619	0.010285	0.010207	0.010207	0.014869	0.010920	0.010207	0.010271	...	0.010207	0.010207
0.022061	0.020194	0.020194	0.021070	0.024925	0.021586	0.115210	1.036870	...	0.027962	0.020194
0.021581	0.013417	0.009501	0.009683	0.012839	0.018760	0.278930	0.022330	...	0.009501	0.009147
0.029087	0.026935	0.027649	0.029325	0.028971	0.025030	0.026935	0.034133	...	0.027649	0.026935
...	...	...	...	...	...	...	...	...	...	...
0.030537	0.059863	0.005725	0.010695	0.008874	0.043710	0.005529	0.005832	...	0.005529	0.005529
0.011925	0.006746	0.006746	0.006746	0.013121	0.006746	0.006746	0.007319	...	0.006935	0.006746
0.018108	0.025765	0.022527	0.021485	0.015446	0.016321	0.014668	0.014732	...	0.017381	0.014847

Models and Results

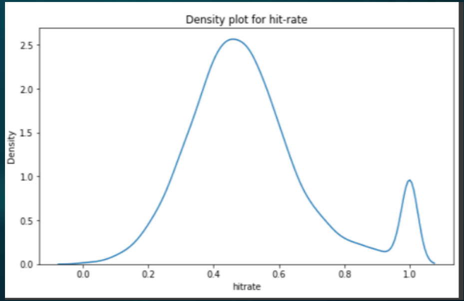
Base Models	MSE
Baseline Popular Model	0.086
CF Model	0.096
Simple CB Model	0.269
Hybrid Model (mean(CF, CB))	0.13
User-user model (cosine sim)	0.000047

Models	MSE
CF Model (cosine sim weighted avg)	0.005
User centric Content based (with hyper parmeter tuning)	0.12
CF-Matrix factorization (Embedding)	0.02
Content Boosted Collaborative Filtering	0.03

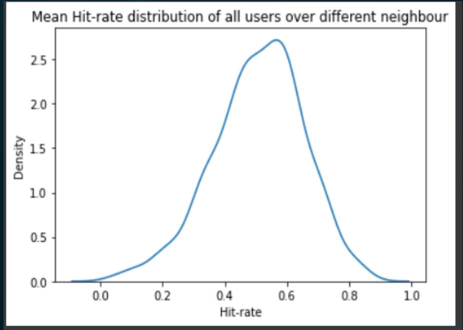
User-user based CF
(Hit-rate distribution)



Matrix factorization
(Hit-rate distribution)



Content-boosted
Filtering
(Hit-rate distribution)



Conclusion and Inference

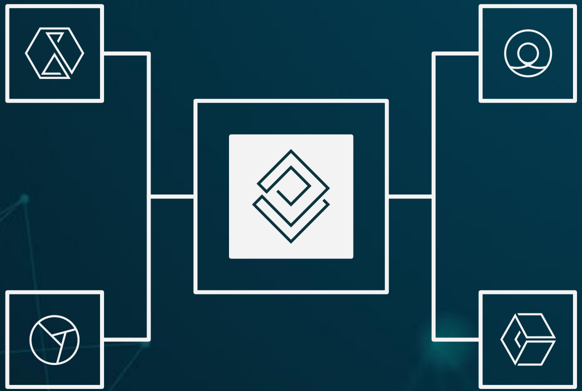
Performance

0.51 hit-rate is decent progress but it tells us that we are not able to make models fully personalizable.

Best Model

base model : user-user based filtering (0.41 hit-rate)

complex model: Matrix factorization and content-boosted filtering (0.51 hit-rate)



Eval. Metric

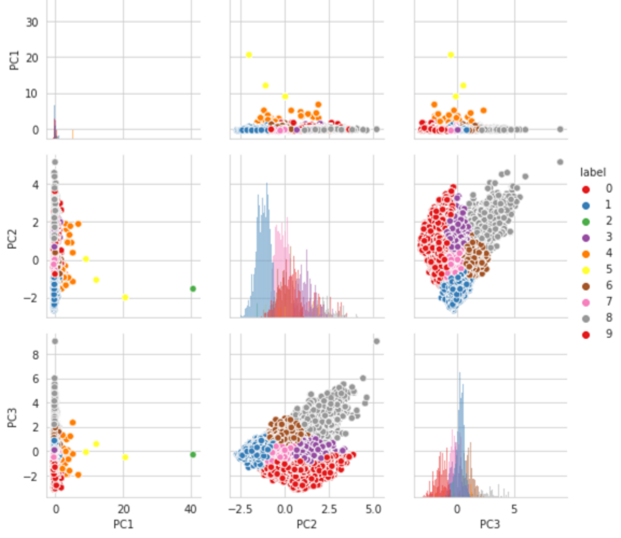
MSE can be used as loss function but to interpret results, we can not rely on mse. Hit-rate alone can't explain the model.

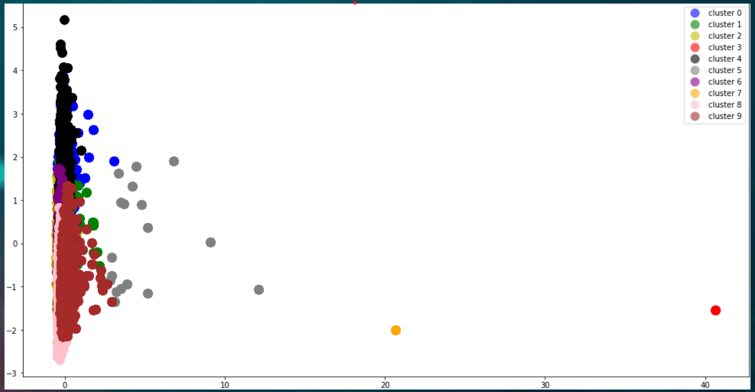
Online Evaluation

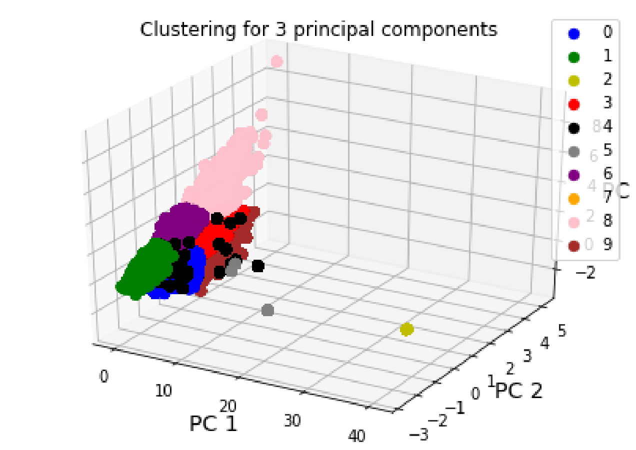
Our system currently only uses past user's data, so we haven't incorporated things like new games released, trends in community etc.

Clustering on Game's Features using PCA

- PCA n_components = 130 (explains > 90 % variance)
- KMeans k = 10 - From edge plot
- Can clearly Observe clusters which can be used for further modelling







IMPROVEMENTS

Variety of user's data

We can add variety of users profiles to the dataset. Use steam API to get user's profile information. Also use other models in content boosted models.

Use other eval. Metrics

Currently we only used hit-rate to evaluate model performance, we can user ranked hit-rate, novelty etc. metrics.

Game related features

It will be good if we get more game related data like length of the game or story line of the game, since the content boosted filtering when used alone wasn't giving good results

Cluster Based Models

We can find the content based prediction from transformed data exploiting the cluster information and also use other cluster based models

THANKS!

Team Members

Abjasree S
Poojan Smart
Niranjan Solanki
Vaishnav Panuganti