Steam Game Recommender System

By: Blake Hernandez

Problem Statement

- There is fierce competition in the world of digital entertainment nowadays, with so many easily accessible sources, it is important to keep users engaged with your platform lest they look for entertainment elsewhere.

- For a long time Steam was the only large digital game store, but recently there has been competition popping up. The Epic Games store has been making headway in the digital sales market, and new game subscription services like Xbox Game Pass have also been on the rise.

Data

1.) Kaggle:
   - a.) A dataset contained user playtime data in minutes for different games in their steam libraries.
   - b.) A separate dataset with feature data for a number of games in Steam, including game genres, tags, developer, and price among other things.

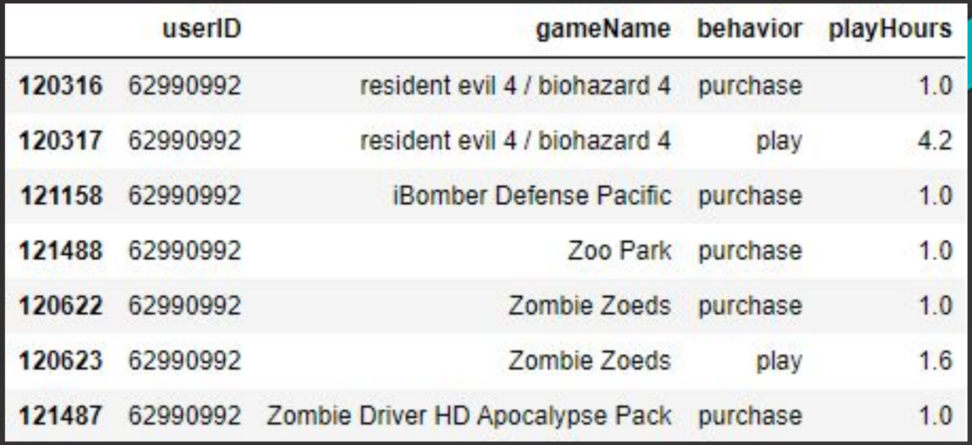
2.) SteamWorks API:
   - a.) Was used to get unique appids for steam games, and pull additional playtime data.

3.) SteamSpy API:
   - a.) Provided additional feature data for games.

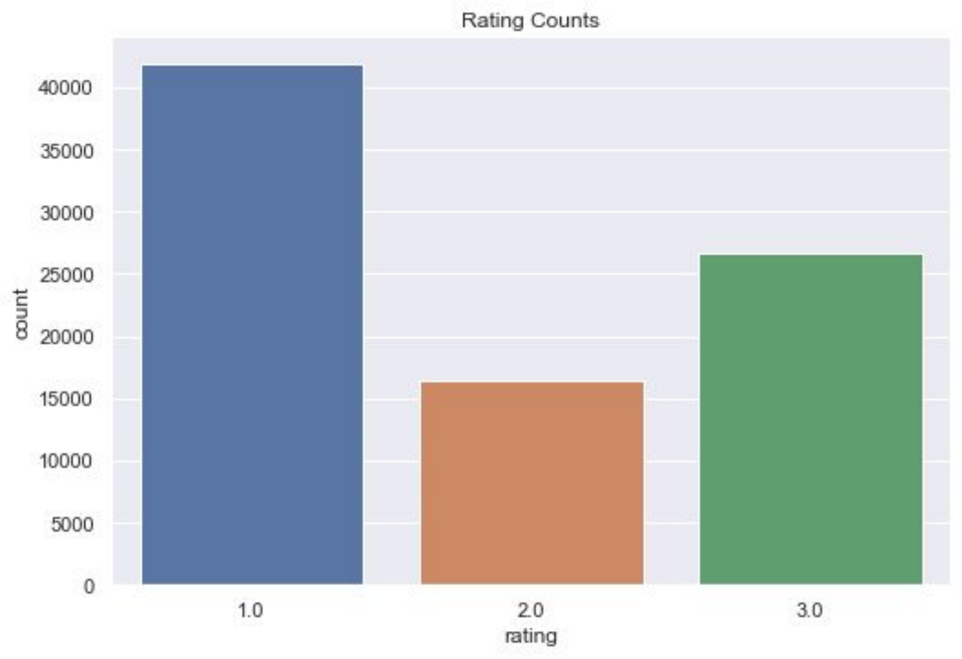
Data Wrangling

- Raw data came with multiple actions. The purchase and play actions had to be separated and re-consolidated into a single row for each user/item interaction.

- The play hour data is a form of implicit rating. For some of the models experimented with, explicit ratings were needed. After play hours were scaled on a game by game basis, explicit ratings were assigned to each user/game interaction based on the games scaled hours relative to the user's average play hours across all games.



	userID	gameName	behavior	playHours
120316	62990992	resident evil 4 / biohazard 4	purchase	1.0
120317	62990992	resident evil 4 / biohazard 4	play	4.2
121158	62990992	iBomber Defense Pacific	purchase	1.0
121488	62990992	Zoo Park	purchase	1.0
120622	62990992	Zombie Zoeds	purchase	1.0
120623	62990992	Zombie Zoeds	play	1.6
121487	62990992	Zombie Driver HD Apocalypse Pack	purchase	1.0

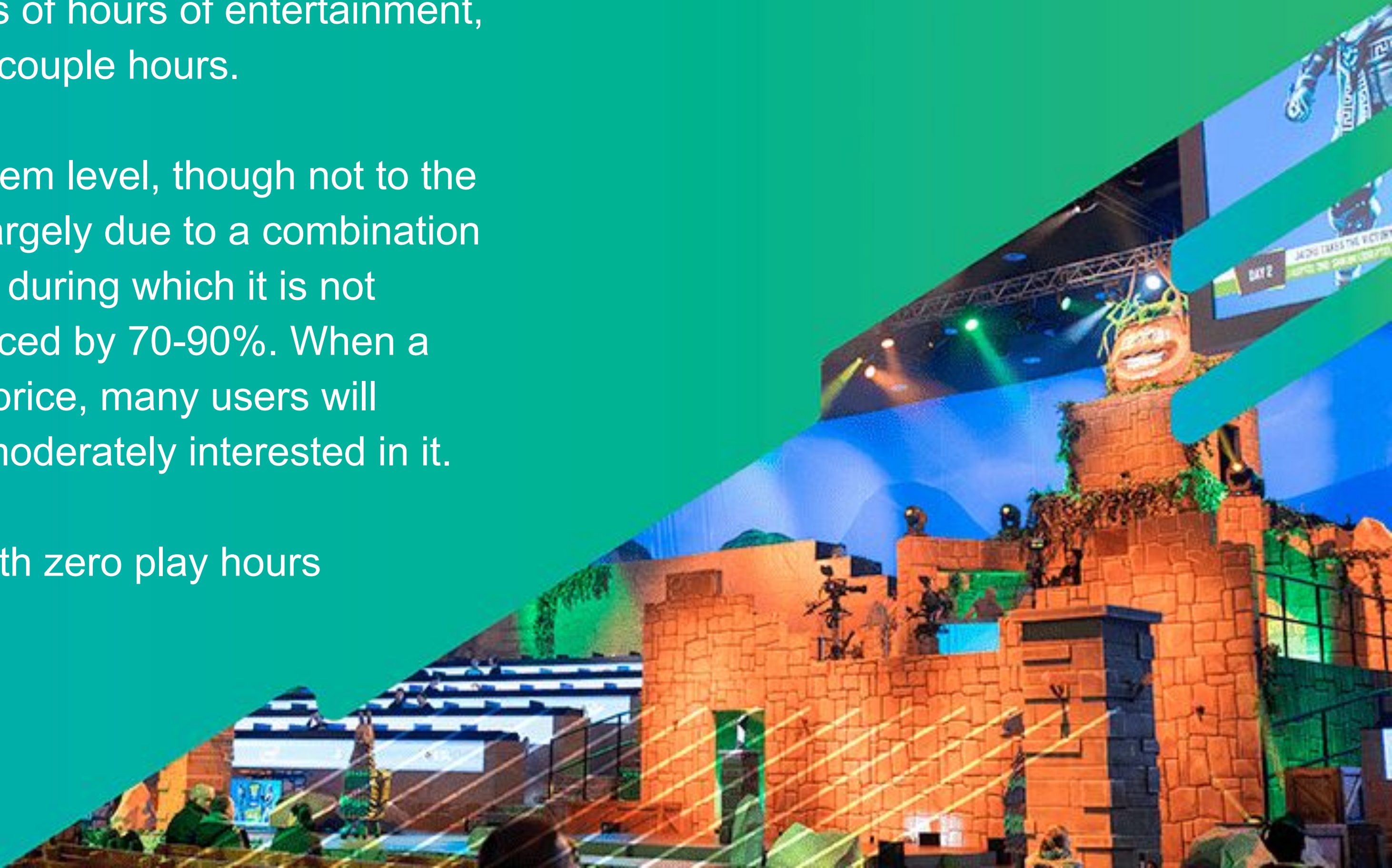


EDA

- EDA revealed that the play hour data was heavily skewed to the left on a population level. The primary reason for this is that not all games are meant to be played for the same amount of time. Some online games are designed to provide hundreds if not thousands of hours of entertainment, while others are meant to only deliver a short couple hours.

- This skew was also present on an individual item level, though not to the same extent as the population. This skew is largely due to a combination of free games and the notorious Steam sales, during which it is not uncommon for games to have their price reduced by 70-90%. When a game is free or costs only a fraction of its full price, many users will download it on a whim, even if they are only moderately interested in it.

- A significant portion of the data was games with zero play hours



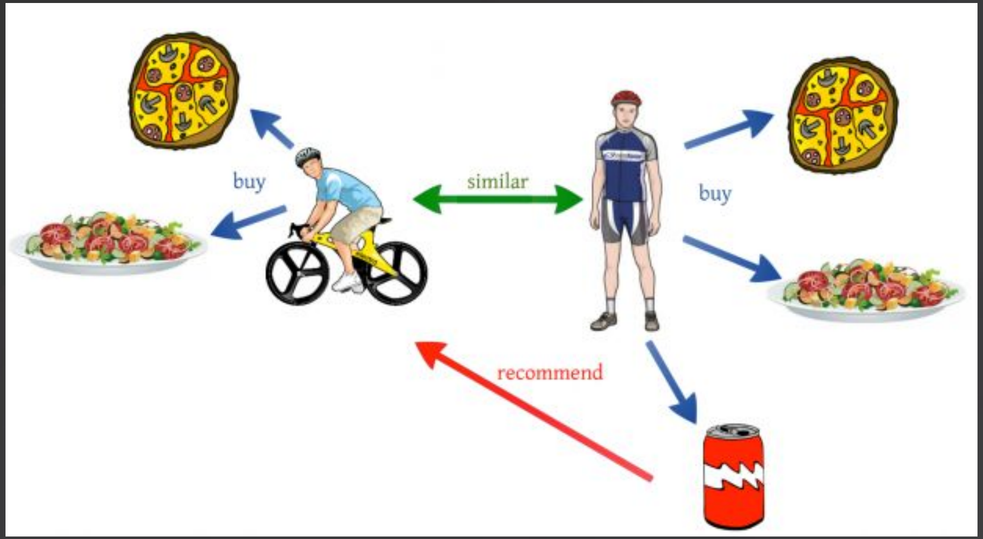
EDA Continued

EDA Decisions

- Capped outliers based on 1.5 IQR
- Considered but ultimately decided not to drop zero play hour games
- Reduced the data size, both to speed up and increase matrix density for training.

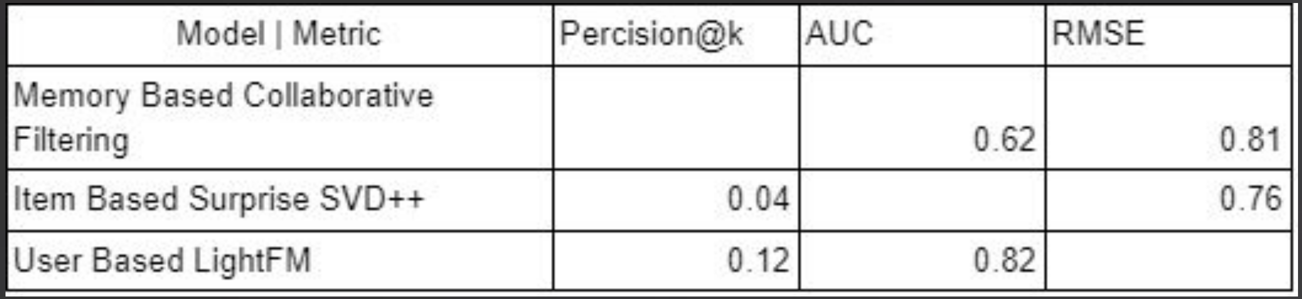
Modeling Methods

- Explored both user and item based collaborative filtering models, and a hybrid model.
- User based collaborative filtering ultimately yielded the best results.
- Three primary models were used:
  - Memory Based collaborative filtering using similarity matrix
  - SVD++ Matrix Factorization using the Surprise library
  - LightFM model, both basic and hybrid



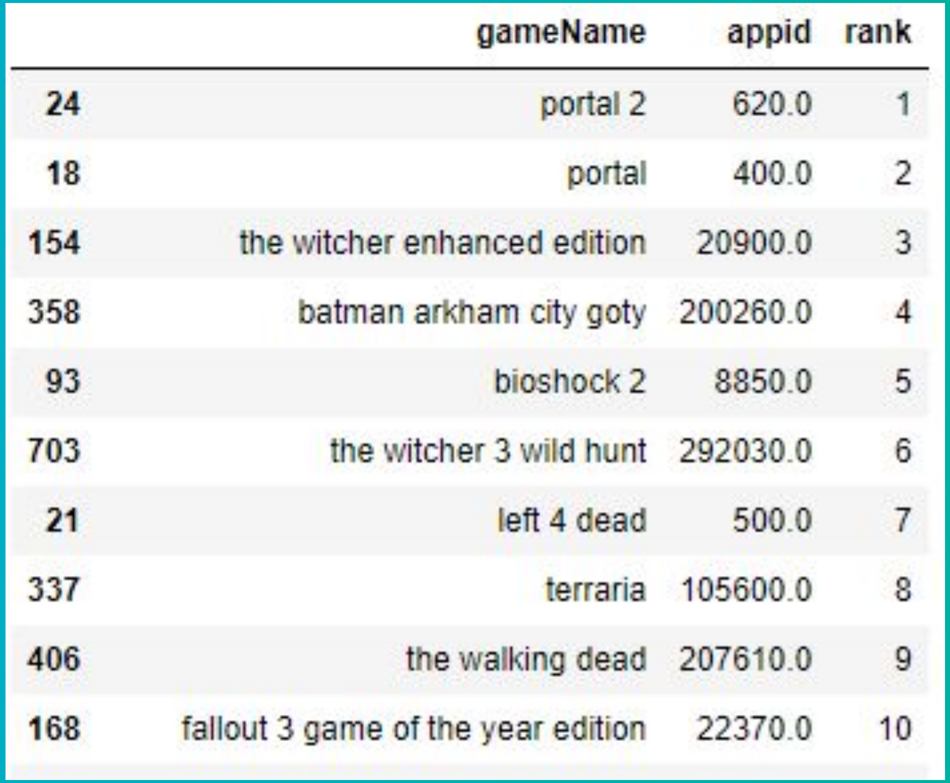
Final Model

- The best performing model was the basic collaborative filtering LightFM model. It outperformed the memory based and SVD++ models by a wide margin.
- Surprisingly the basic model performed better than the hybrid.
- Loss function used was WARP and learning schedule was Adagrad.



Model | Metric	Percision@k	AUC	RMSE
Memory Based Collaborative Filtering		0.62	0.81
Item Based Surprise SVD++	0.04		0.76
User Based LightFM	0.12	0.82	

Model Recommendations



	gameName	appid	rank
24	portal 2	620.0	1
18	portal	400.0	2
154	the witcher enhanced edition	20900.0	3
358	batman arkham city goty	200260.0	4
93	bioshock 2	8850.0	5
703	the witcher 3 wild hunt	292030.0	6
21	left 4 dead	500.0	7
337	terraria	105600.0	8
406	the walking dead	207610.0	9
168	fallout 3 game of the year edition	22370.0	10

Future Improvements

- Develop an algorithm to determine best tradeoff between matrix density and data loss.

- Property address the cold start problem. Either develop a model that can effectively deal with it or create two separate models, one for new users and one for users who have passed a certain threshold of interactions.

- Seek out user feedback for model recommendations

Data Sources

Kaggle Steam interaction data:
https://www.kaggle.com/datasets/tamber/steam-video-games

Kaggle Steam game feature data:
https://www.kaggle.com/datasets/nikdavis/steam-store-games

SteamSpy API
https://steamspy.com/api.php

SteamWorks API:
https://partner.steamgames.com/doc/webapi_overview

Questions?