M244 Final Project

What Games Attract You?

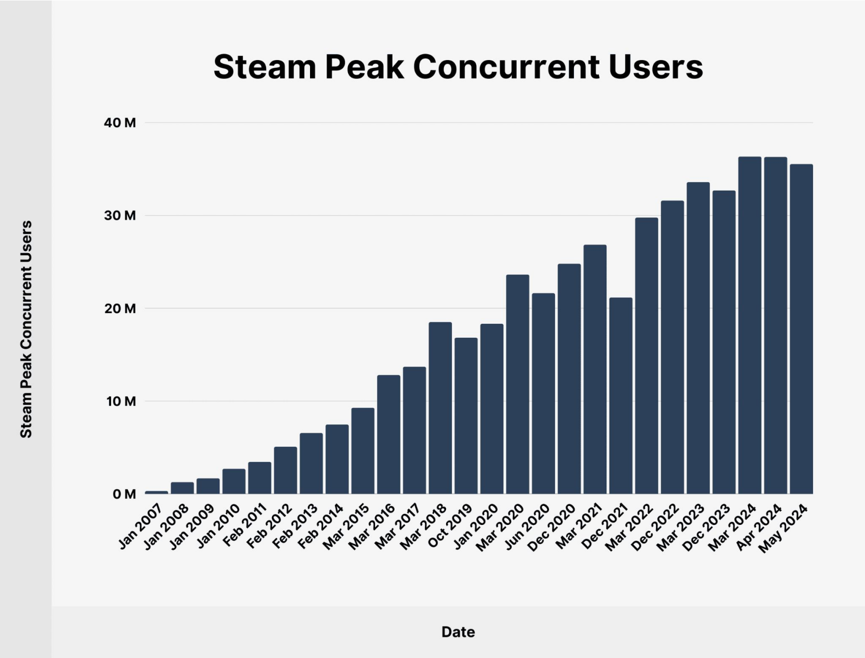
Boxiong & Yifei

Introduction

Video games have emerged as one of the most dynamic and influential forms of entertainment, reshaping how people engage with media and spend their leisure time (Polcyn, 2018).

As the world's leading PC gaming platform, Steam has revolutionized game distribution and player interaction and boasts a large and active player base.

By analyzing the steam dataset, we hope to provide data-driven insights that can help developers create more compelling gaming experiences while enabling players to make better-informed choices.



Research Question

How is **median gameplay time** of a game affected by other variables?

We expect:

- Higher-priced games will correlate with longer playtime, reflecting deeper content or premium quality
- Games with higher positive review rates will sustain longer engagement, as player satisfaction likely enhances retention
- Moderate popularity of games measured by the peak concurrent users maximizes playtime

Data

Data

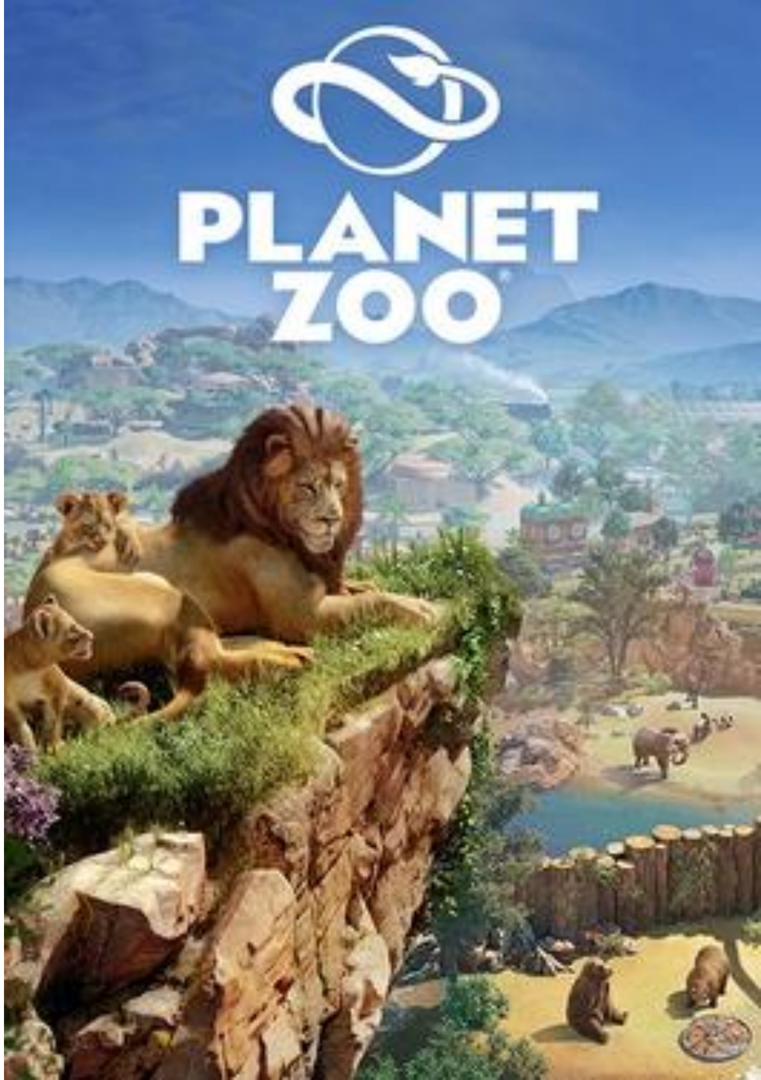
- Downloaded from Kaggle
- **110,000+** PC game titles on Steam
- JSON format, keyed by Steam AppID

Core Fields:

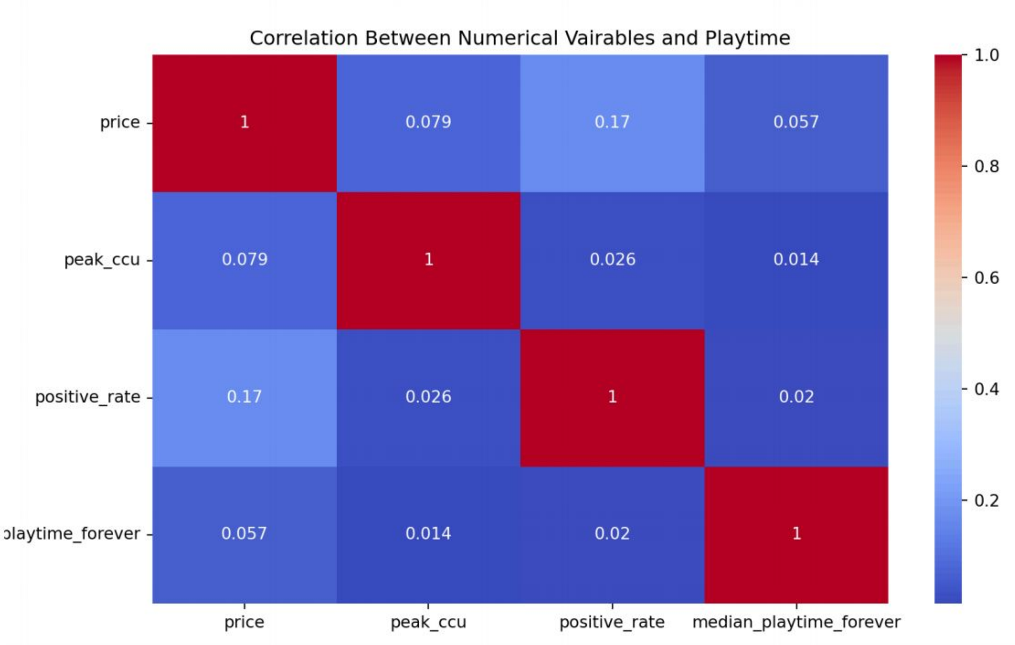
- **Identifiers & Metadata**: `appID`, `name`, `release_date`, `price`
- **Ownership & Playtime**: `estimated_owners`, `average_playtime_forever`, `average_playtime_2weeks`
- **User Feedback**: `positive`, `negative`, `user_score`, `metacritic_score`
- **Platform Support**: `windows`, `mac`, `linux`
- **Additional**: `dlc_count`, `achievements`, `recommendations`, `supported_languages`

Data Wrangling

- Positive Review Rate: Positive / (Positive + Negative)
- Compatible Systems: Sum of system dummies (1, 2, 3)
- Publishers: Kept only the ten most frequent values
- Genres: Kept only the 'main' genre for each game, and took only the ten most frequent values
- Release Year: Year component of Release Date

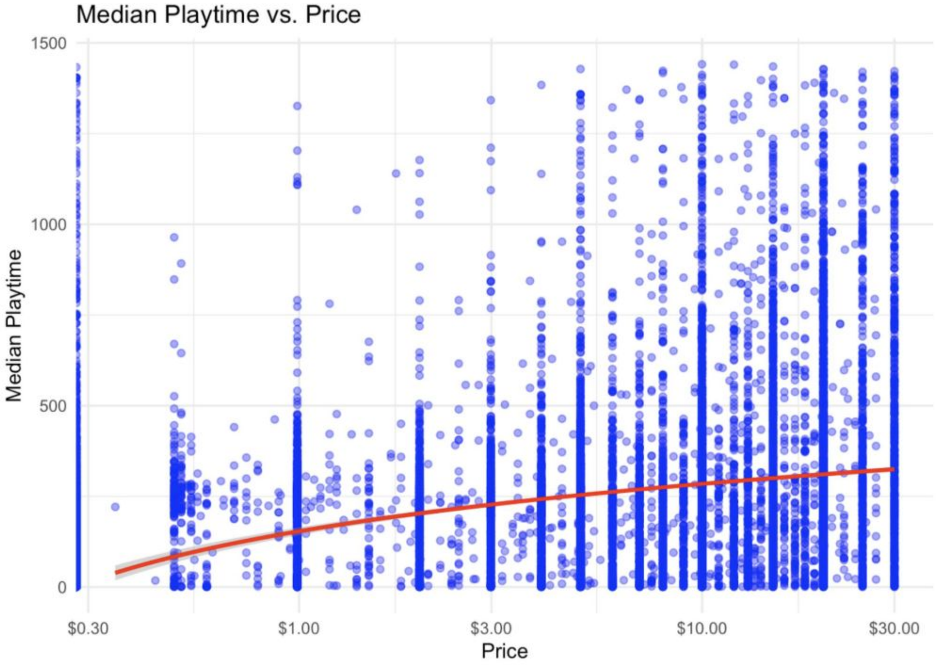


Highlights from EDA



Price, peak concurrent users, and positive review rates all have a positive correlation with median playtime of games, but very weak.

Highlights from EDA



- Linear relationship
- Cluster of data points representing free-to-play games concentrated on the left
- Red regression line would become steeper if remove those free games

Methodology

All models shared the same preprocessing pipeline:

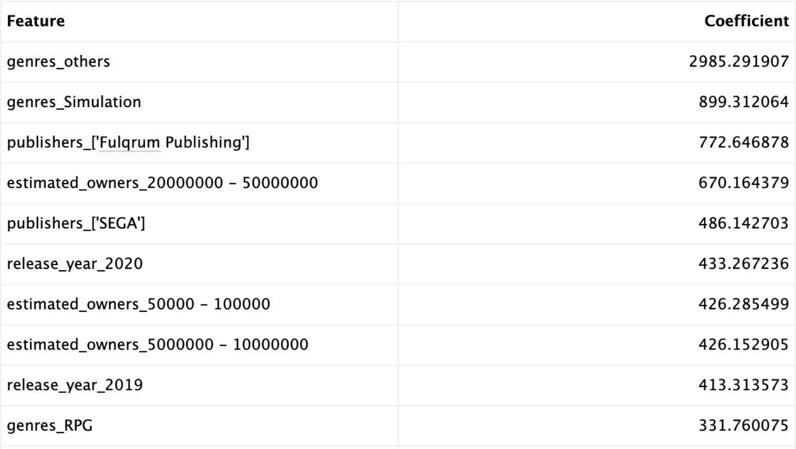
- Categorical variables were one-hot encoded.
- Numerical features were scaled.

Model 1: **Linear Regression**

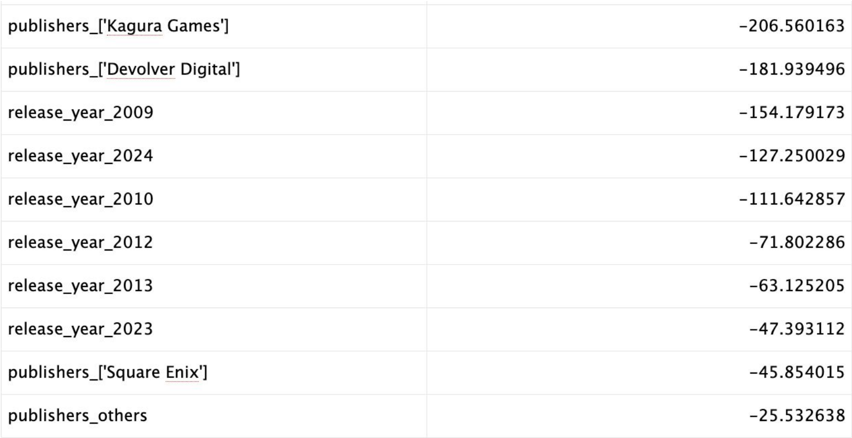
Model 2: **Lasso**

Model 3: **Elastic Net Cross Validation**

Results - Regression



Feature	Coefficient
genres_others	2985.291907
genres_Simulation	899.312064
publishers_['Fulqrum Publishing']	772.646878
estimated_owners_20000000 - 50000000	670.164379
publishers_['SEGA']	486.142703
release_year_2020	433.267236
estimated_owners_50000 - 100000	426.285499
estimated_owners_5000000 - 10000000	426.152905
release_year_2019	413.313573
genres_RPG	331.760075



publishers_['Kagura Games']	-206.560163
publishers_['Devolver Digital']	-181.939496
release_year_2009	-154.179173
release_year_2024	-127.250029
release_year_2010	-111.642857
release_year_2012	-71.802286
release_year_2013	-63.125205
release_year_2023	-47.393112
publishers_['Square Enix']	-45.854015
publishers_others	-25.532638

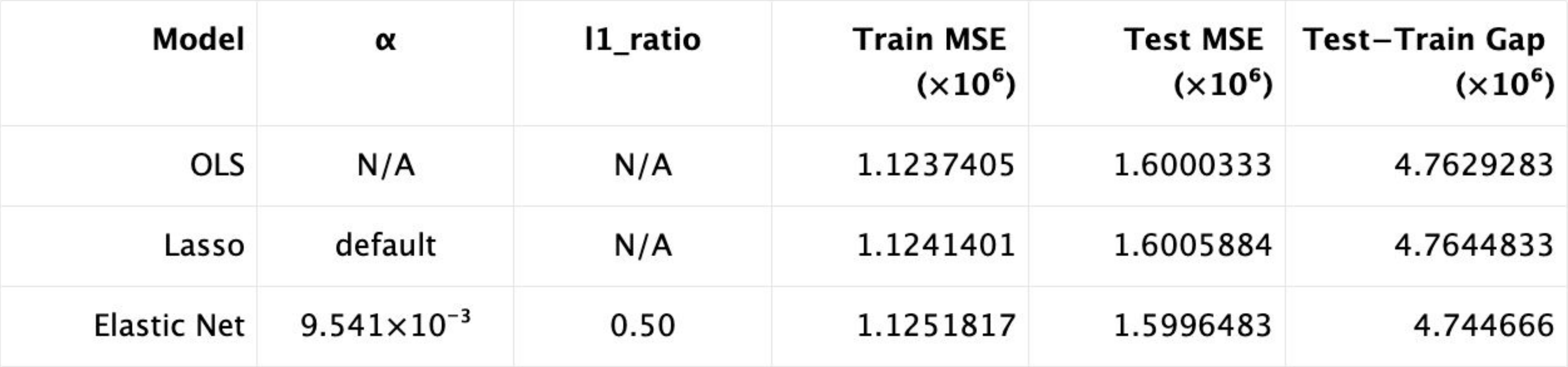
Top positive drivers :

- **genres_others**: +2,985
- **genres_Simulation**: +899
- **publisher = Fulqrum Publishing**: +773

Top negative drivers:

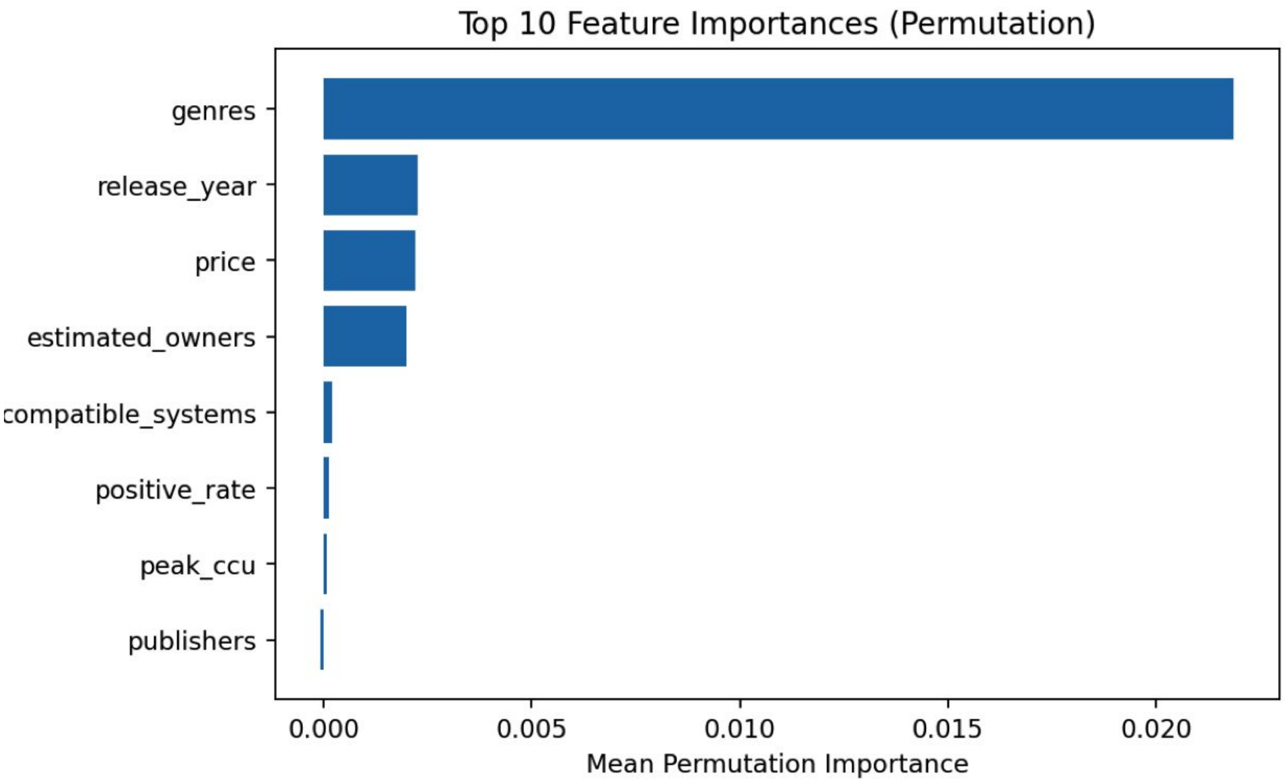
- **publisher = Kagura Games**: −207
- **publisher = Devolver Digital**: −182
- **older release years (2009, 2010, 2012, 2013)** each −60 to −155

Results - Regression



Model	α	l1_ratio	Train MSE $(\times 10^6)$	Test MSE $(\times 10^6)$	Test−Train Gap $(\times 10^6)$
OLS	N/A	N/A	1.1237405	1.6000333	4.7629283
Lasso	default	N/A	1.1241401	1.6005884	4.7644833
Elastic Net	9.541×10^{-3}	0.50	1.1251817	1.5996483	4.744666

Results-Regression

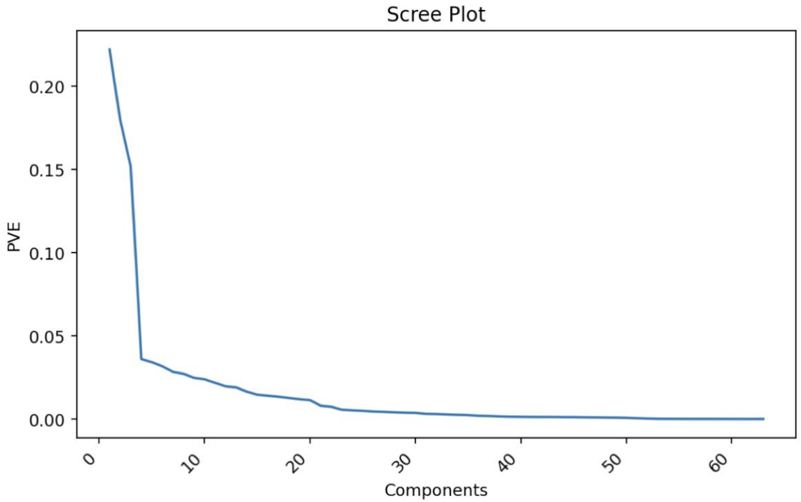


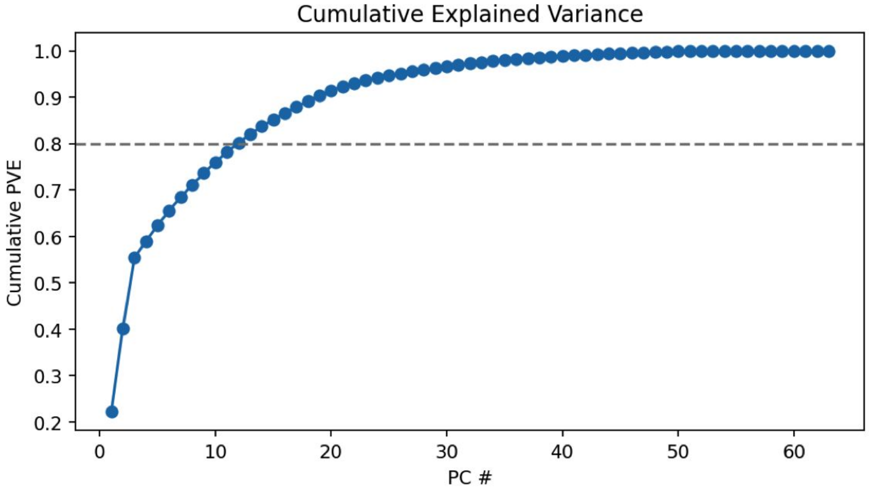
Model relies heavily on genre information to predict playtime.

Release year and price signals are secondary drivers.

Peak CCU and positive rate contribute least once genre and price are known.

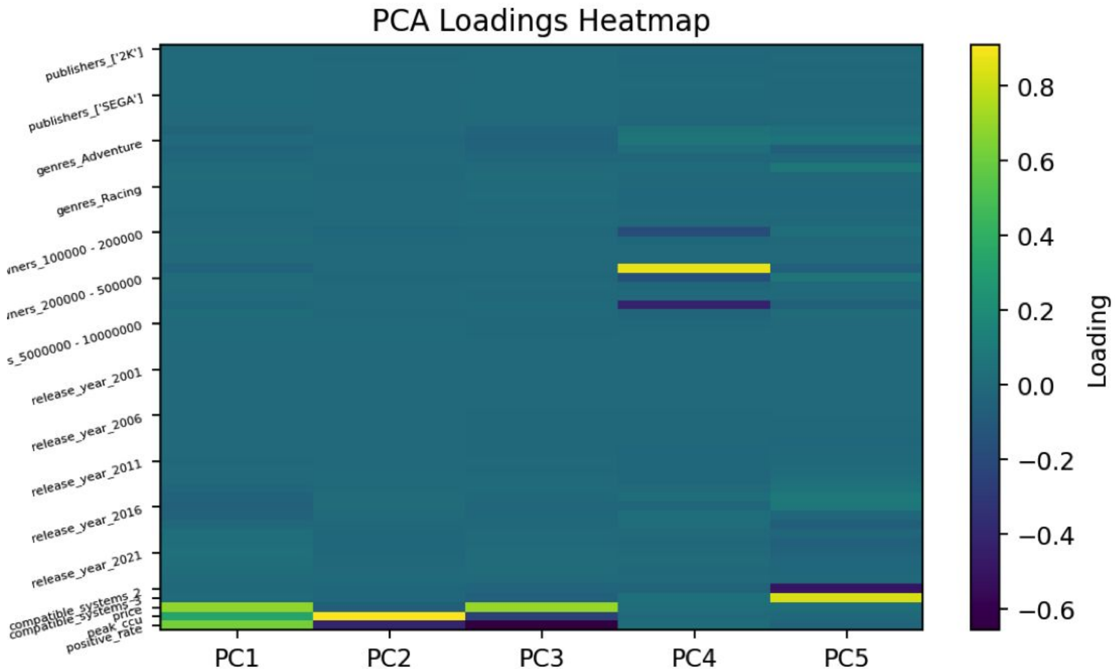
Results - PCA



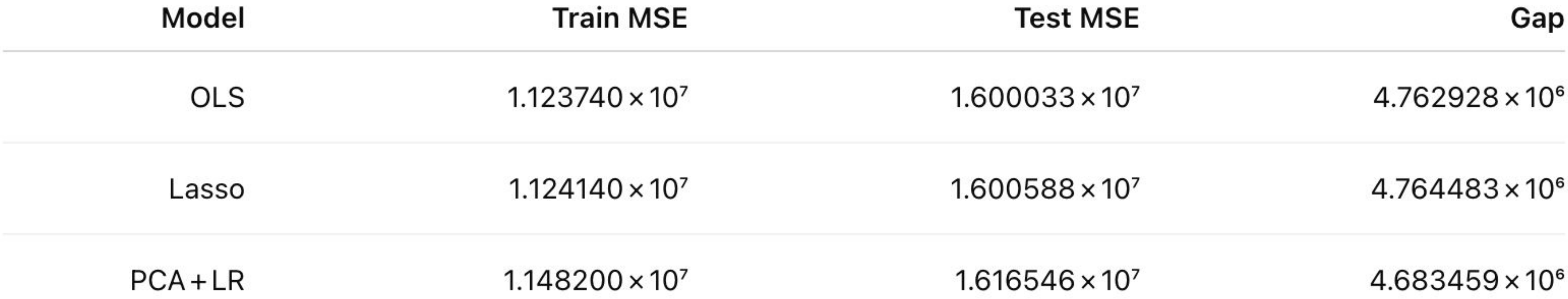


Chose the number of components = 10

Results-PCA


PCA Loadings Heatmap
publishers_['2K']
publishers_['SEGA']
genres_Adventure
genres_Racing
wners_100000 - 200000
wners_200000 - 500000
s_5000000 - 10000000
release_year_2001
release_year_2006
release_year_2011
release_year_2016
release_year_2021
compatible_systems_2
compatible_systems_3
price
peak_ccu
positive_rate
PC1
PC2
PC3
PC4
PC5
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Loading

Results-PCA



Model	Train MSE	Test MSE	Gap
OLS	1.123740×10^7	1.600033×10^7	4.762928×10^6
Lasso	1.124140×10^7	1.600588×10^7	4.764483×10^6
PCA + LR	1.148200×10^7	1.616546×10^7	4.683459×10^6

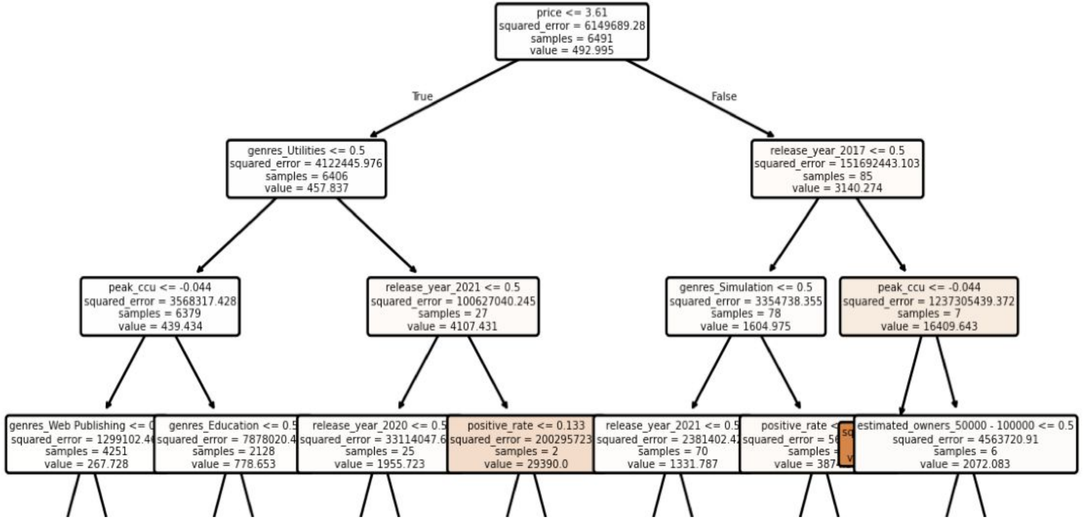
Methodology - Nonlinear

Model 4: **Random Forest**

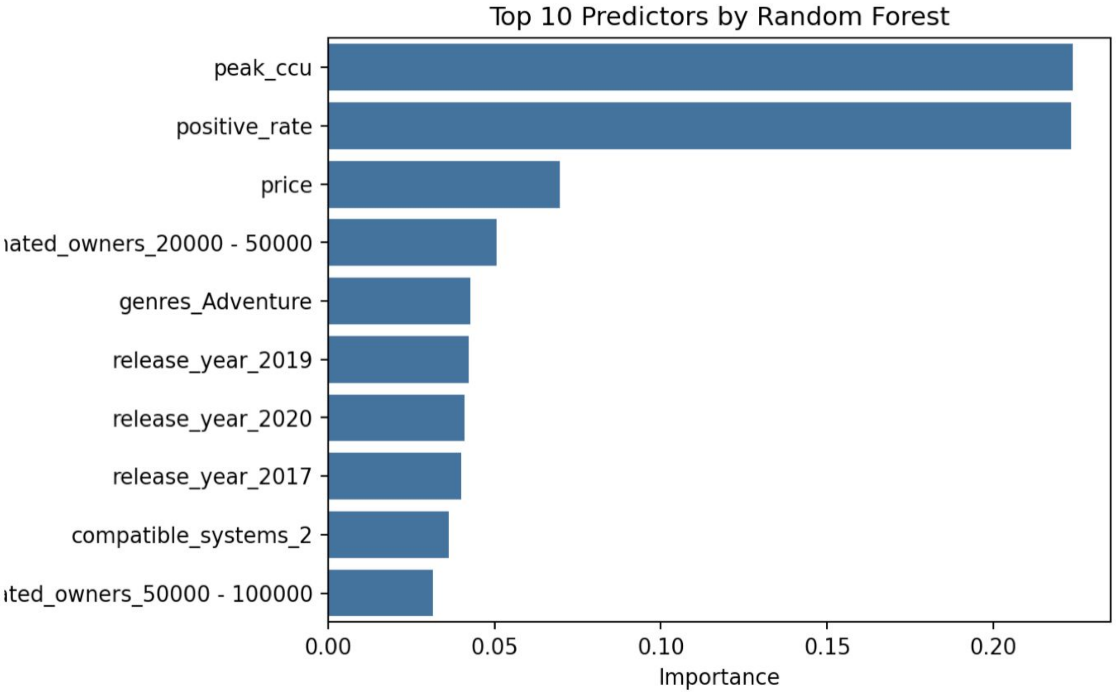
Use features like price, release_year, genres, and positive_rate to split the data

Starting with price <= 3.61

Further moving down to genres, release year, and other key predictors



Results-Random Forest

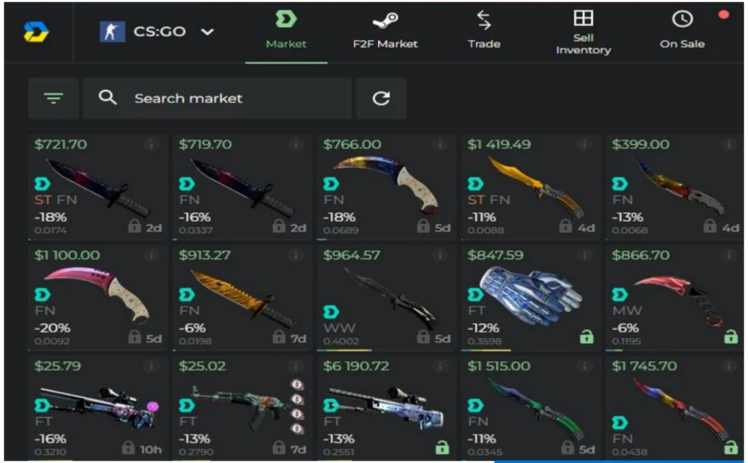


- **Peak concurrent users(peak_ccu)** and **positive review rates(positive_rate)** are the most influential predictors of median playtime

- Price, release year, genres, compatible systems, and estimated owners have less influence on median playtime.

Limitations

- Many free-to-play games rely heavily on in-game purchases
    - Counter-Strike: Global Offensive (CSGO)
        - most widely played FPS games on Steam
        - free to download
        - players spend money on cosmetic weapon skins
        - online markets for players to trade skins



$721.70 ST FN -18% 0.0174 2d	$719.70 FN -16% 0.0337 2d	$766.00 FN -18% 0.0689 5d	$1 419.49 ST FN -11% 0.0088 4d	$399.00 FN -13% 0.0068 4d
$1 100.00 FN -20% 0.0092 5d	$913.27 FN -6% 0.0198 7d	$964.57 WW 0.4002 5d	$847.59 FT -12% 0.3598	$866.70 MW -6% 0.1195
$25.79 FT -16% 0.3210 10h	$25.02 FT -13% 0.2790 7d	$6 190.72 FT -13% 0.2551	$1 515.00 FN -11% 0.0345 5d	$1 745.70 FN 0.0438

- Steam is not the only platform where games were sold or played
    - other platforms like the Nintendo Switch, Xbox, PlayStation
    - or other digital stores like the Epic Games Store

- **Positive review rates** by dividing the number of positive reviews by the total number of reviews
    - misleading for games with very few reviews, sometimes only one or two
    - a single review can skew the ratio to 100% or 0%

Future Work

Cross-Validations to choose more appropriate alpha

More careful grouping and engineering of variables

Incorporating the total number of reviews as a separate feature

Filter out all games that have in-game purchases

More non-linear models

Ethical Considerations

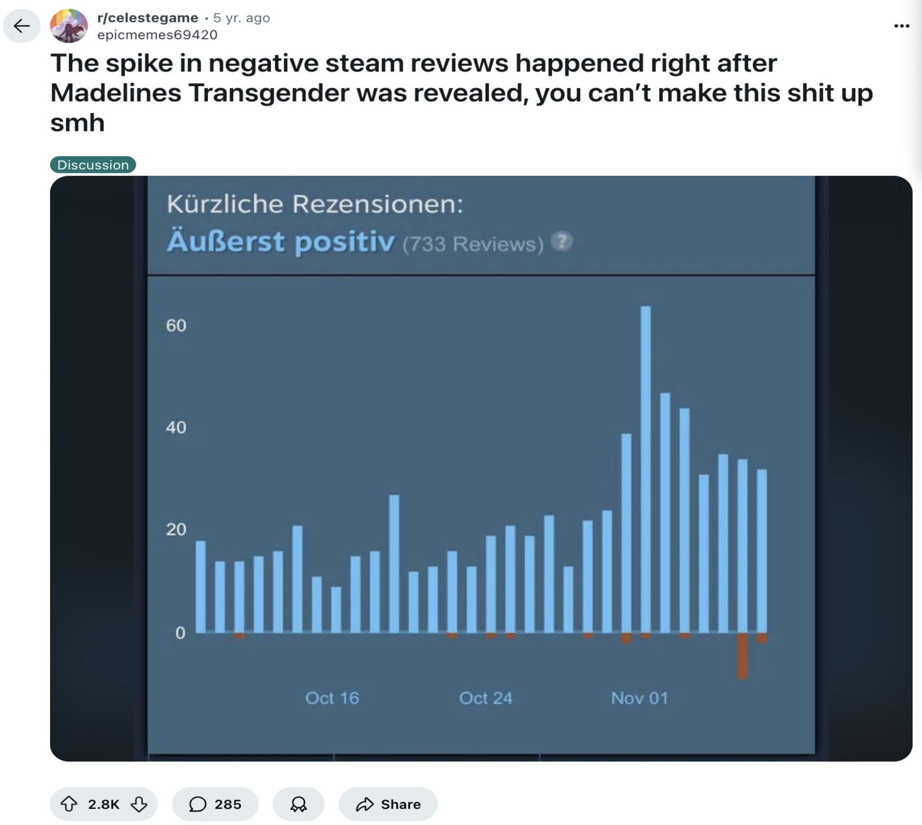
This Steam dataset is sourced from Steam's official aggregated data.

Players' personal information is not collected, so there are **no confidentiality issues** related to individual privacy exposure.

Our decision to treat **positive reviews** as a key variable could be problematic

- Games with LGBTQ+ themes, female protagonists, or minority developers are often subject to review bombing by toxic communities
- Example: **Celeste**!
- If we interpret low review scores as an indicator of poor game quality without context, we risk reproducing and legitimizing cultural bias through data analysis

Review Bombing: Celeste



r/celestegame • 5 yr. ago
epicmemes69420

The spike in negative steam reviews happened right after Madelines Transgender was revealed, you can’t make this shit up smh

Discussion

- Developed and published by the indie studio *Maddy Makes Games*.

- Despite receiving critical acclaim and winning awards for gameplay and storytelling, the game was targeted with waves of negative reviews due to the developer’s transgender identity

Thank You

Questions?