

"NICHE IS NICE! "

Making a successful indie game.



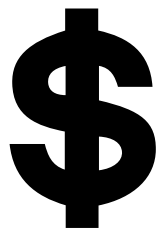
WHY PREDICT GAME PERFORMANCE?

Videogame Sales between 1985 to 2020*

YEAR

GENRE

GLOBAL SALES



01



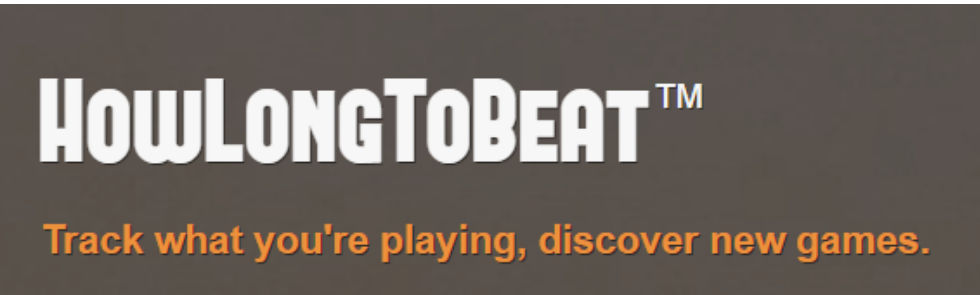
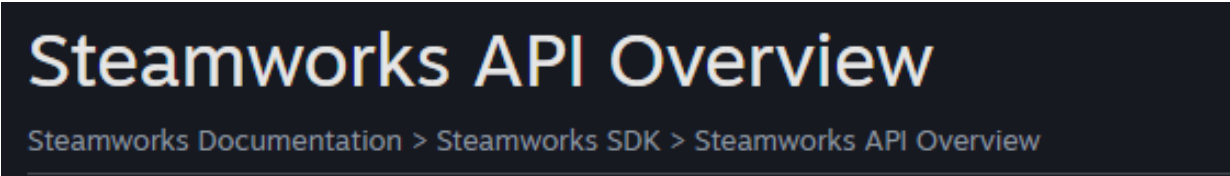
Thousands of games launch yearly—most go unnoticed.

Developers need insight into what drives visibility and sales.

Can we predict key outcomes like wishlists, scores, and sales before launch?

HYPOTHESIS

There are some KPIs that our studio should measure in playtests off the get go.



HowLongToBeat Games Scraper (2/16/2025)

Dataset of game playtime estimates scraped from HowLongToBeat.

Steam Games Dataset 2025

The largest public dataset of 90 000+ games published on Steam.

games_march2025_cleaned.csv (468.64 MB)

Single-Player	Polled	Average	Median	Rushed	Leisure
Main Story	493	73h 34m	69h 44m	47h 49m	123h 19m
Main + Extras	2.5K	118h 1m	110h	70h 7m	294h 9m
Completionist	486	186h 43m	160h	116h 10m	536h 40m
All PlayStyles	3.5K	121h 18m	110h	69h 49m	507h 49m

genres,id,name,platforms,summary,storyline,rating,main,extra,completionist,review_score,review_count,people_polled

[5],274203,Short 'n Quick,[6],this is a techstyled map taking place in a warehouse,Missing,Missing,nan,nan,nan,nan,nan,nan

[5],256819,Caverns of Darkness,[6],The final rift was closed and the Hell War was over at last The citizens of the Earth c

[26, 31]",232860,"Nu, pogodi! Vypusk 1: Pogonya",[6],The wolf decides to take revenge on the Hare for all past grievances

[12, 15, 16]",228979,Power Dolls 5,[6],The fifth entry in Kogado Studios allfemale mech tactical RPG series Power Dolls,M

[2],225648,Kapsyljakt med Anki & Pytte,[6],A pointandclick game based on the Swedish childrens program Anki Pytte As the

CONCLUSION+ NEXT STEPS

What are our new hypotheses? What data do we need?

USER REVIEWS



BUDGET



PRICING



MONTHLY SALES



MULTI-GENRE



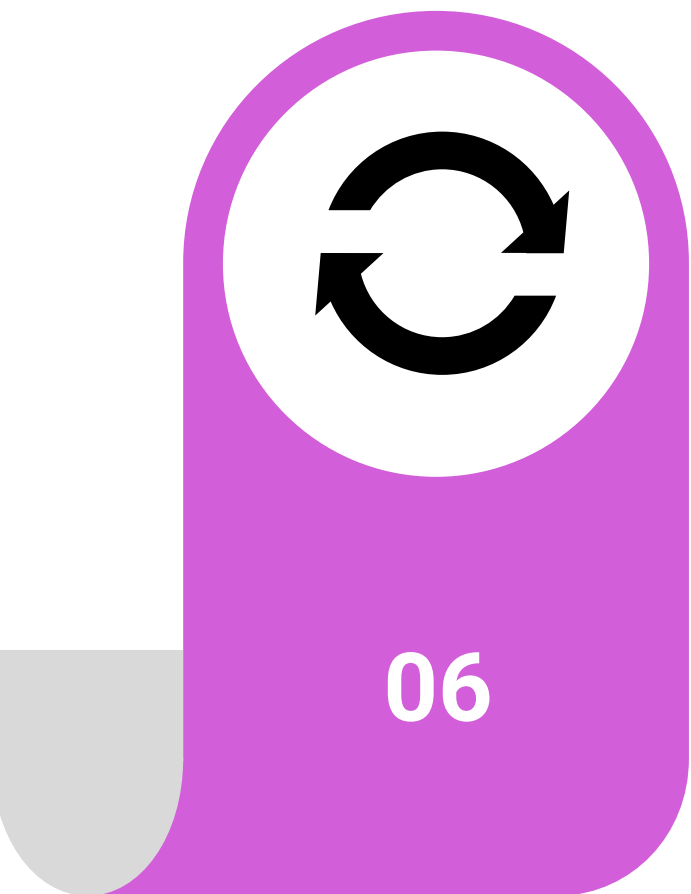
LARGER
DATABASE



06

CONCLUSION+ NEXT STEPS

What are our new hypotheses? What data do we need?



USER REVIEWS



BUDGET



PRICING



YEARLY* SALES



MULTI-GENRE



PLAYTIMES

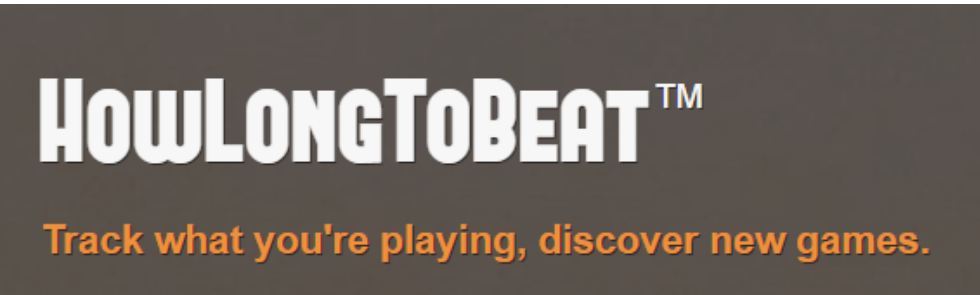
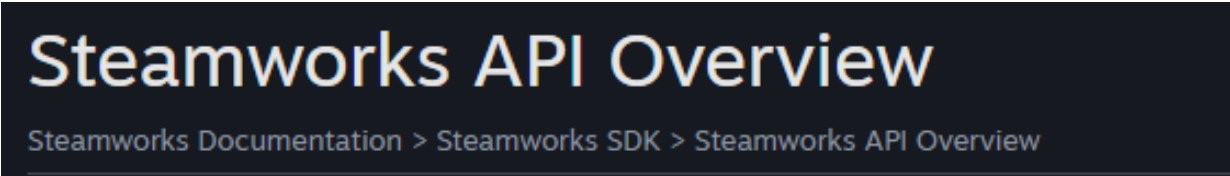


LARGER
DATABASE



HYPOTHESIS

There are some KPIs that our studio should measure in playtests off the get go.



HowLongToBeat Games Scraper (2/16/2025)

Dataset of game playtime estimates scraped from HowLongToBeat.

Steam Games Dataset 2025

The largest public dataset of 90 000+ games published on Steam.

games_march2025_cleaned.csv (468.64 MB)

Single-Player	Polled	Average	Median	Rushed	Leisure
Main Story	493	73h 34m	69h 44m	47h 49m	123h 19m
Main + Extras	2.5K	118h 1m	110h	70h 7m	294h 9m
Completionist	486	186h 43m	160h	116h 10m	536h 40m
All PlayStyles	3.5K	121h 18m	110h	69h 49m	507h 49m

genres,id,name,platforms,summary,storyline,rating,main,extra,completionist,review_score,review_count,people_polled

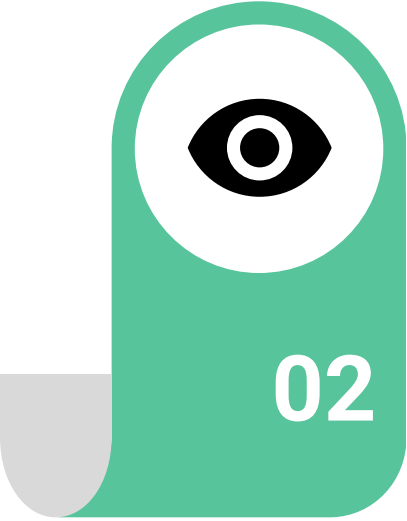
[5],274203,Short 'n Quick,[6],this is a techstyled map taking place in a warehouse,Missing,Missing,nan,nan,nan,nan,nan,nan

[5],256819,Caverns of Darkness,[6],The final rift was closed and the Hell War was over at last The citizens of the Earth c

[26, 31]",232860,"Nu, pogodi! Vypusk 1: Pogonya",[6],The wolf decides to take revenge on the Hare for all past grievances

[12, 15, 16]",228979,Power Dolls 5,[6],The fifth entry in Kogado Studios allfemale mech tactical RPG series Power Dolls,M

[2],225648,Kapsyljakt med Anki & Pytte,[6],A pointandclick game based on the Swedish childrens program Anki Pytte As the



THE DATA BEHIND IT ALL

Fusion of seperate large databases

~500 one-hot encoded features

Structured data only — no text/NLP

Key features used:

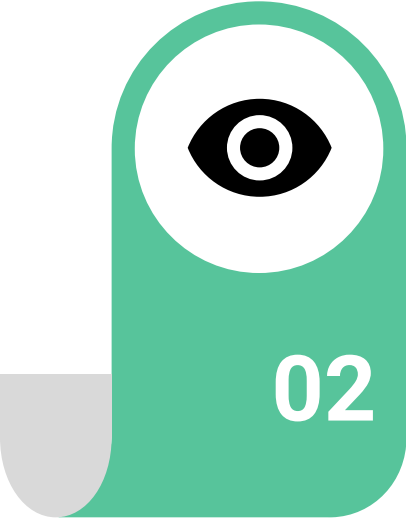
Time to beat

Price

Engagement ratio

Publisher class (AA/AAA/Indie)

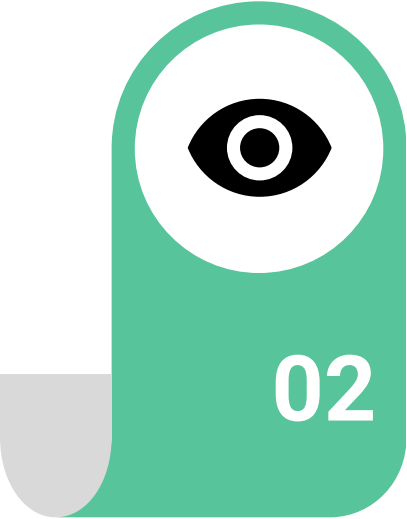
Tags, genres, categories



MACHINE LEARNING PIPELINE

Training the Models

1. **Drop target columns (Wishlists, Bayesian Score, Copies Sold)**
2. **Filter numeric features**
3. **Train individual CatBoost models for each target**
4. **Save models using joblib**
5. **Test and evaluate with R2, MAE, and MSE metrics**

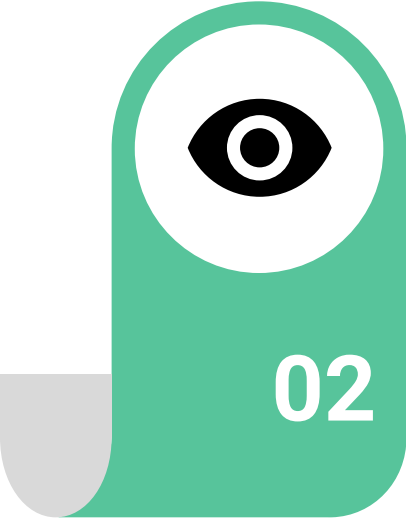


MACHINE LEARNING PIPELINE

Training the Models

1. Drop target columns (Wishlists, Bayesian Score, Copies Sold)
2. Filter numeric features
3. Train individual CatBoost models for each target
4. Save models using joblib
5. Test and evaluate with R2, MAE, and MSE metrics

Target	R2 Score	MAE	MSE
Wishlists	0.752458	34115.164881	1.635226e+10
bayesian_score	0.386534	6.591071	7.868539e+01
Copies Sold	0.834776	171311.519471	6.055316e+11



STREAMLIT APP PREVIEW

Interactive Streamlit App

1. Enter game details via sliders and dropdowns
2. Get instant predictions for:
3. Wishlists
4. Score
5. Copies Sold
6. Built with streamlit, deployed on Streamlit Cloud

Input Game Features

Time to Beat (minutes)

60.00 - +

Price (\$)

19.99 - +

Followers

50 - +

Engagement Ratio

0.90 - +

Select Tags

2D x 4X x

Action RTS x

Select Genres

Adventure x Indie x

RPG x

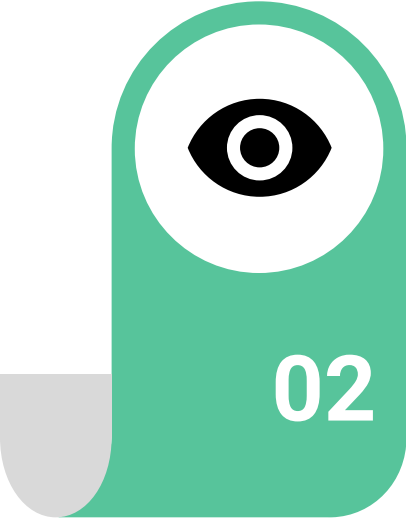
Select Categories

Single-player x Co-op x

In-App Purchases x

Select Publisher Class






Indie

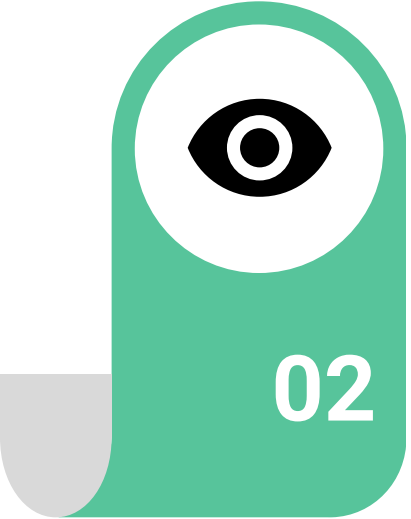


- Full preprocessing + modeling notebook
- Trained .pkl CatBoost models
- Cleaned datasets
- Streamlit app (app.py)
- **more steps to come...**

GITHUB

W.I.P







 Cleancsv	TOO MUCH STUFF	5 hours ago
 catboost_info	bugs bug bugs	1 hour ago
 .gitattributes	Extra bits	last week
 Clean_Database.ipynb	Fixing more shit!!	1 hour ago
 DataBase.ipynb	TOO MUCH STUFF	5 hours ago
 GraphsCoreCurves.ipynb	TOO MUCH STUFF	5 hours ago
 ML_Friday.csv	TOO MUCH STUFF	5 hours ago
 ML_final.csv	TOO MUCH STUFF	5 hours ago
 ML_ready.csv	TOO MUCH STUFF	5 hours ago



- Full preprocessing + modeling notebook
- Trained .pkl CatBoost models
- Cleaned datasets
- Streamlit app (app.py)
- **more steps to come...**

GITHUB

W.I.P

 Cleancsv	TOO MUCH STUFF	5 hours ago
 catboost_info	bugs bug bugs	1 hour ago
 .gitattributes	Extra bits	last week
 Clean_Database.ipynb	Fixing more shit!!	1 hour ago
 DataBase.ipynb	TOO MUCH STUFF	5 hours ago
 GraphsCoreCurves.ipynb	TOO MUCH STUFF	5 hours ago
 ML_Friday.csv	TOO MUCH STUFF	5 hours ago
 ML_final.csv	TOO MUCH STUFF	5 hours ago
 ML_ready.csv	TOO MUCH STUFF	5 hours ago