GAME SUCCESS ANALYTICS

Predicting Game Success with ML Models

A Data-Driven Journey through Steam Metrics

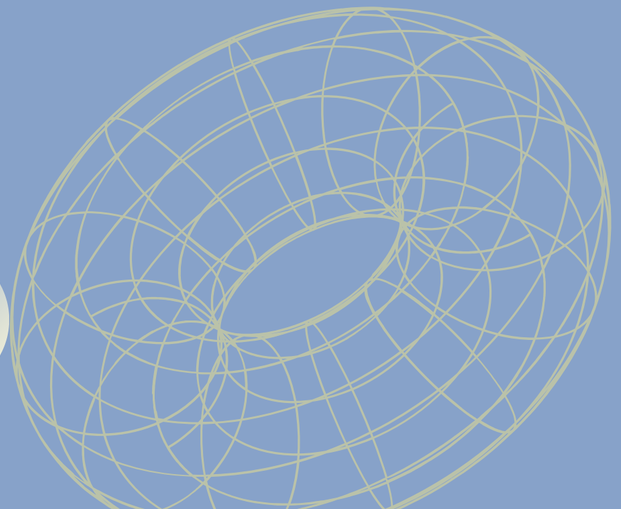
Project Overview

Project Goals: "Develop predictive models for key game performance metrics: Copies Sold, Wishlists, and Bayesian Score"

Dataset Overview: "Initial dataset of 98,350 Steam games, filtered to 7,029 games with complete metrics"

Business Value: "Enable developers and publishers to forecast game performance based on game characteristics"

Approach Summary: "Regression modeling with gradient boosting (CatBoost) for three target variables"

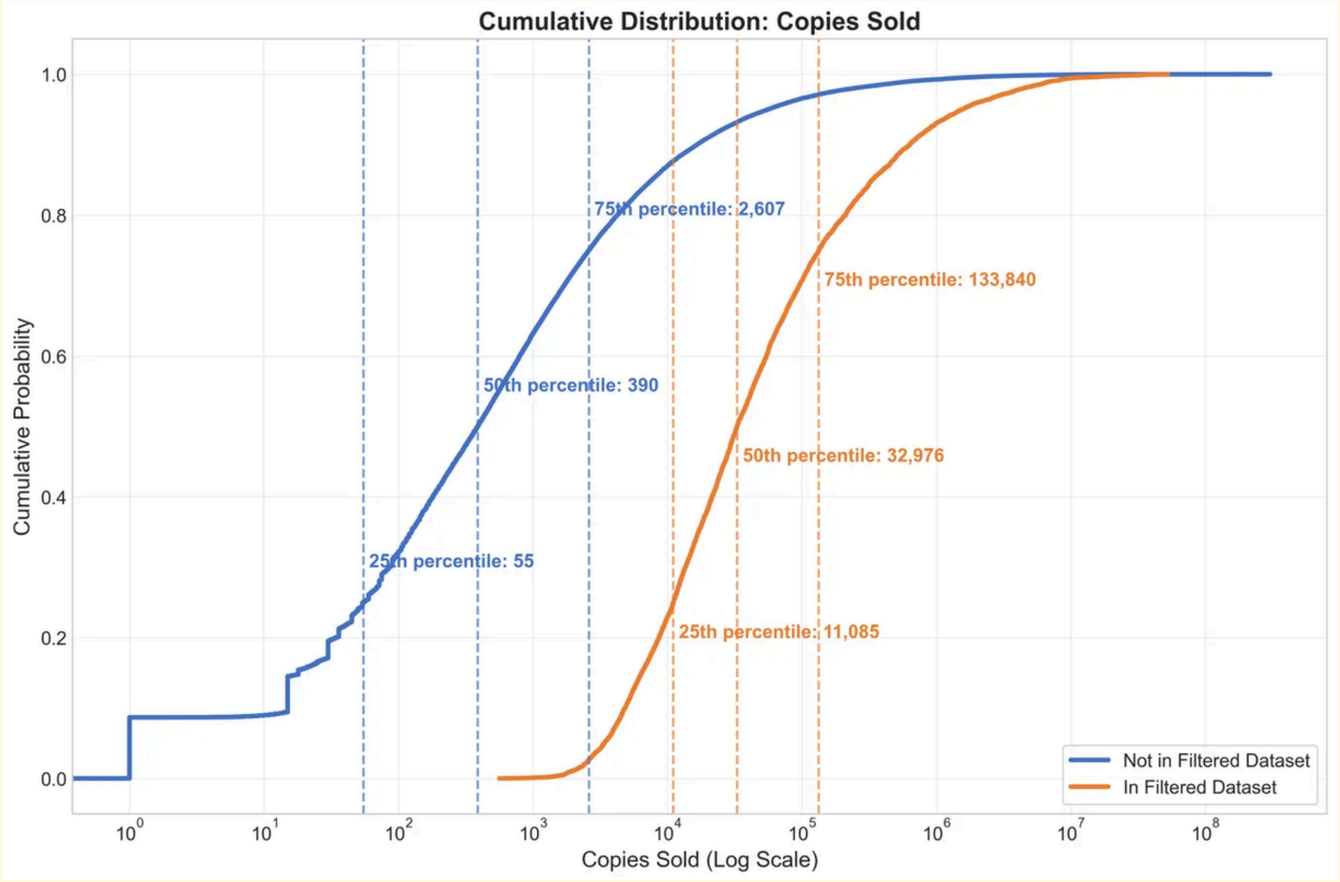


Data Analysis & Selection Bias Discovery

Discovery narrative: Initial data exploration revealed significant selection bias

Key statistics: Filtered dataset represents only 7.15% of the total Steam market

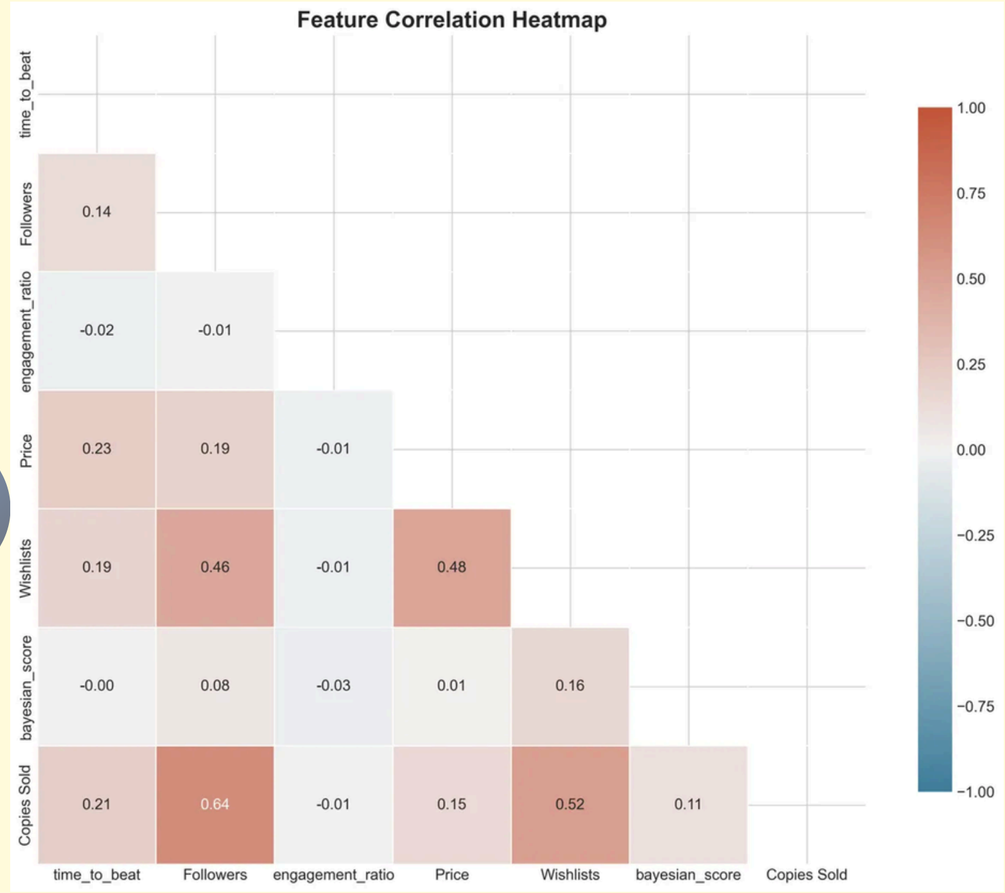
Implications: Games in filtered dataset dramatically outperform market average



Understanding Selection Bias in Datasets

How bias can skew analytical outcomes

Selection bias occurs when certain data points are **overrepresented** or **underrepresented** in a dataset. This can lead to misleading conclusions, making it crucial to identify and address bias before conducting further analysis. Proper filtering and visualization techniques can help uncover these discrepancies.



Data Preprocessing: Techniques and Strategies

Data Cleaning Process: Filtered for games with complete time_to_beat metrics

- Handled missing values in *engagement* metrics

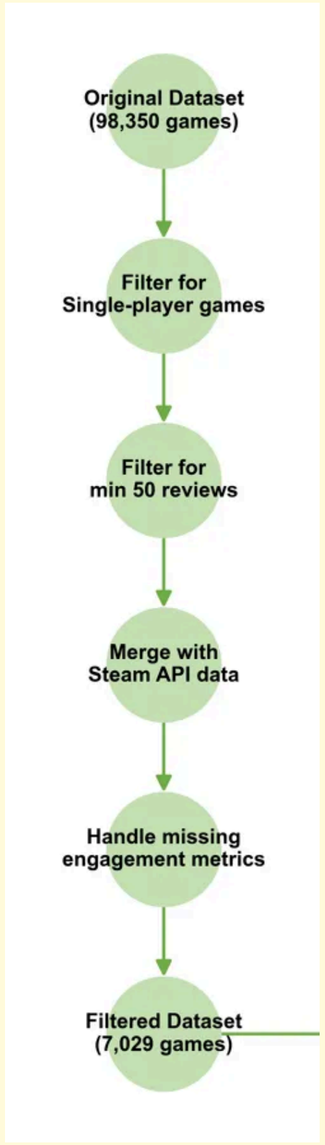
Created engineered features:

- completion_rate_avg = avg_playtime / time_to_beat
- engagement_ratio (uncapped version of completion rate)
- playtime_deviation (difference between avg and median playtime

Feature Engineering: One-hot encoded categorical features:

- 400+ game tags (e.g., "Open World", "RPG", "Multiplayer")
- 17 genres
- 40+ categories
- Publisher class (AAA, AA, Indie, Hobbyist)

Normalization & Transformation: Standard approach towards handling numerical features



Data Preprocessing: Techniques and Strategies

Data Cleaning Process: Filtered for games with complete time_to_beat metrics

- Handled missing values in *engagement* metrics

Created engineered features:

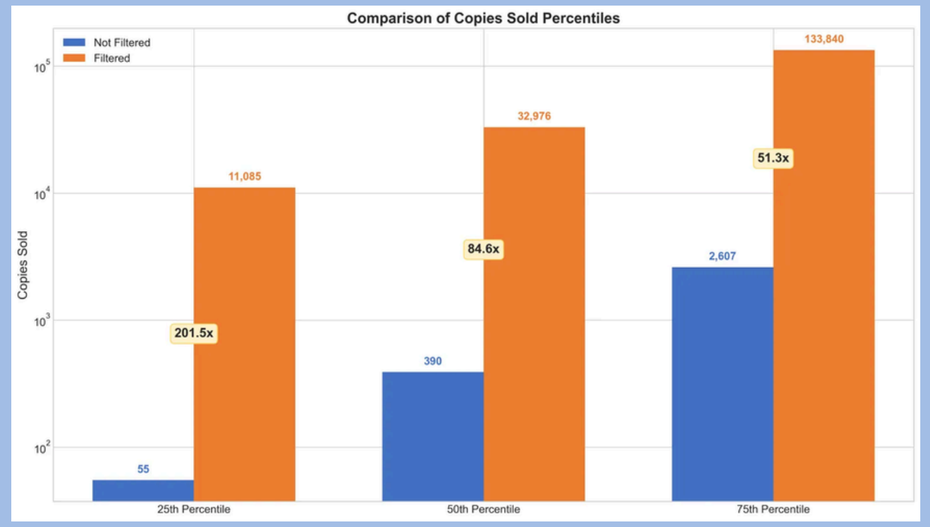
- completion_rate_avg = avg_playtime / time_to_beat
- engagement_ratio (uncapped version of completion rate)
- playtime_deviation (difference between avg and median playtime

Feature Engineering: One-hot encoded categorical features:

- 400+ game tags (e.g., "Open World", "RPG", "Multiplayer")
- 17 genres
- 40+ categories
- Publisher class (AAA, AA, Indie, Hobbyist)

Normalization & Transformation: Standard approach towards handling numerical features

Modeling Strategy: CatBoost



Tested Models:

- Linear Regression (baseline)
- Random Forest
- XGBoost
- CatBoost ← Final choice

Training Strategy

- 80/20 train-test split
- Cross-validation with 5 folds
- Hyperparameter tuning with grid search
- Feature importance analysis
- Multiple target variable modeling (separate models)

Model Performance Metrics Overview

Evaluating the accuracy of our predictive models

Target Variable	R^2 Score	MAE
Copies Sold	0.83	171,311
Wishlists	0.75	34,115
Bayesian Score	0.39	6.59

Most important features:

- time_to_beat
- Followers
- engagement_ratio
- Steam Workshop support
- Publisher Class_AAA
- Tags_Open World

Target VariableR^2 ScoreMAEMSECopies Sold0.83171,3116.06e+11Wishlists0.7534,1151.64e+10Bayesian Score0.396.5978.69

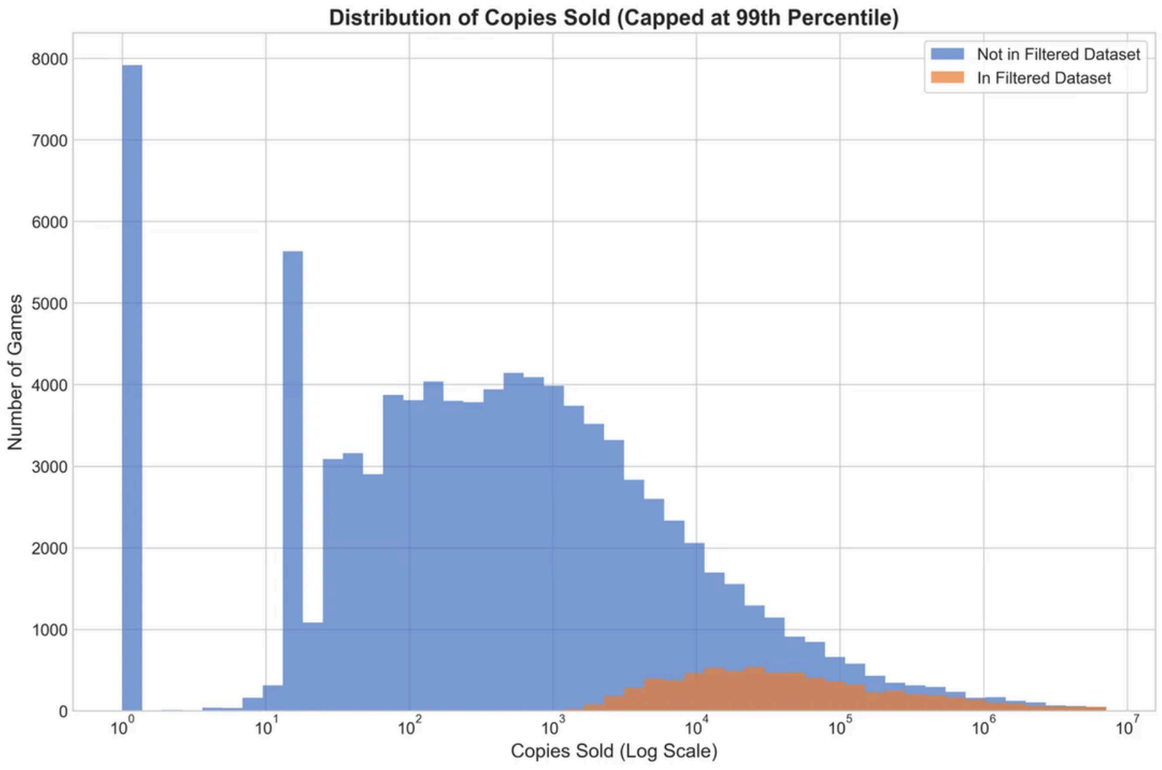
Model Limitations: Understanding Bias Impact

Dataset Bias Implications

- Models trained on highly successful games (top 7% of market)
- Even the 25th percentile in our dataset represents the ~90th percentile in the market
- Limited applicability to newer or less established games

Technical Challenges

- High feature dimensionality (500+ features after encoding)
- Heavy reliance on categorical features
- Limited data for some publisher classes
- Engagement metrics only available for successful games

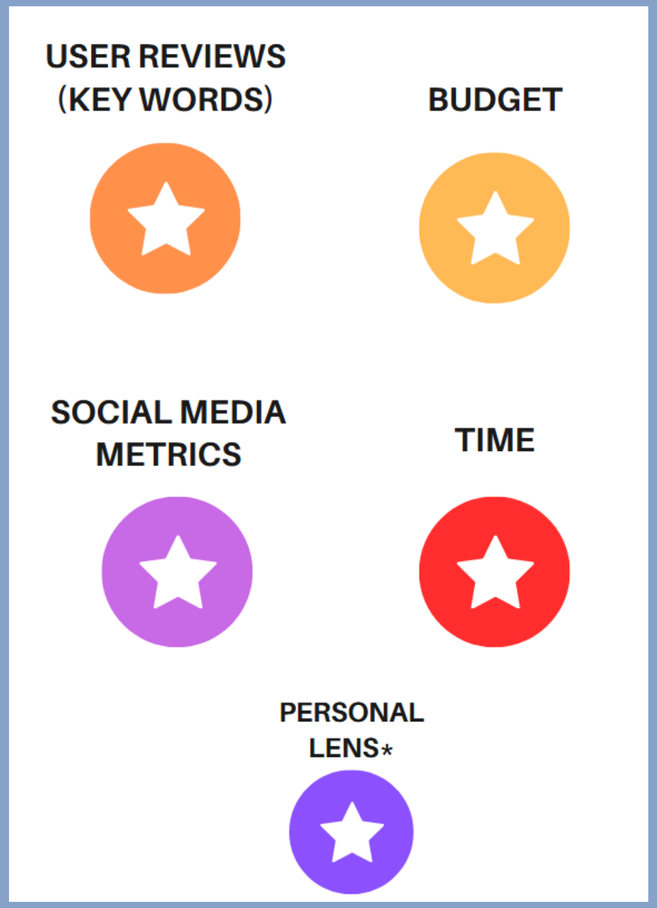


Proposed Future Improvements: Two-Stage Model

Combining classification and regression for better predictions

First Stage: Predict if a game will be in top ~7% of market
Second Stage: For "successful" games, predict precise metrics

Future steps:



We already had key features.

But now with NLP, Reviews can be measured for keywords. Time can be used to track growth of a game over prolonged periods and spot patterns in trends or fads. These can then be cross referenced with Twitch and Youtube APIs and their tags.

This project will develop into using RNNs in the coming weeks.

Thank you and Questions!

Niche is Nice!