



CAN WE PREDICT THE NEXT GAMING BLOCKBUSTER?

SHIJIN, ESRA, FELIX

START >

INDEX



1. PROJECT TITLE
2. PROJECT OVERVIEW
3. DATA OVERVIEW
4. WHY CLASSIFICATIONS?
5. FEATURE ENGINEERING
6. MODELING APPROACH
7. CONFUSION MATRIX
8. KEY INSIGHTS
9. TRENDS BY YEAR
10. REAL-WORLD APPLICATIONS
11. CHALLENGES AND LESSONS
12. CONCLUSIONS
13. FUTURE WORK
14. THANK YOU & Q&A

PROJECT OVERVIEW

Challenge: Can we predict whether a video game will be a commercial hit using only pre-release data?

Top Franchise. Big Budget. Hype Overload.
Guaranteed hit

Risky Bet. Good Genre, Unknown Publisher.
Might surprise us

Low Budget. Niche Genre. Weak Platform.
No one would buy it

Goal: Predict if a game will sell more than 1 million copies based on genre, platform, release year, and publisher.

DATA OVERVIEW

Combines historical sales data with game attributes
to reveal what makes a title a hit —
before it even launches.

Source

Two Kaggle Video Games datasets merged

Size

59,000+ records

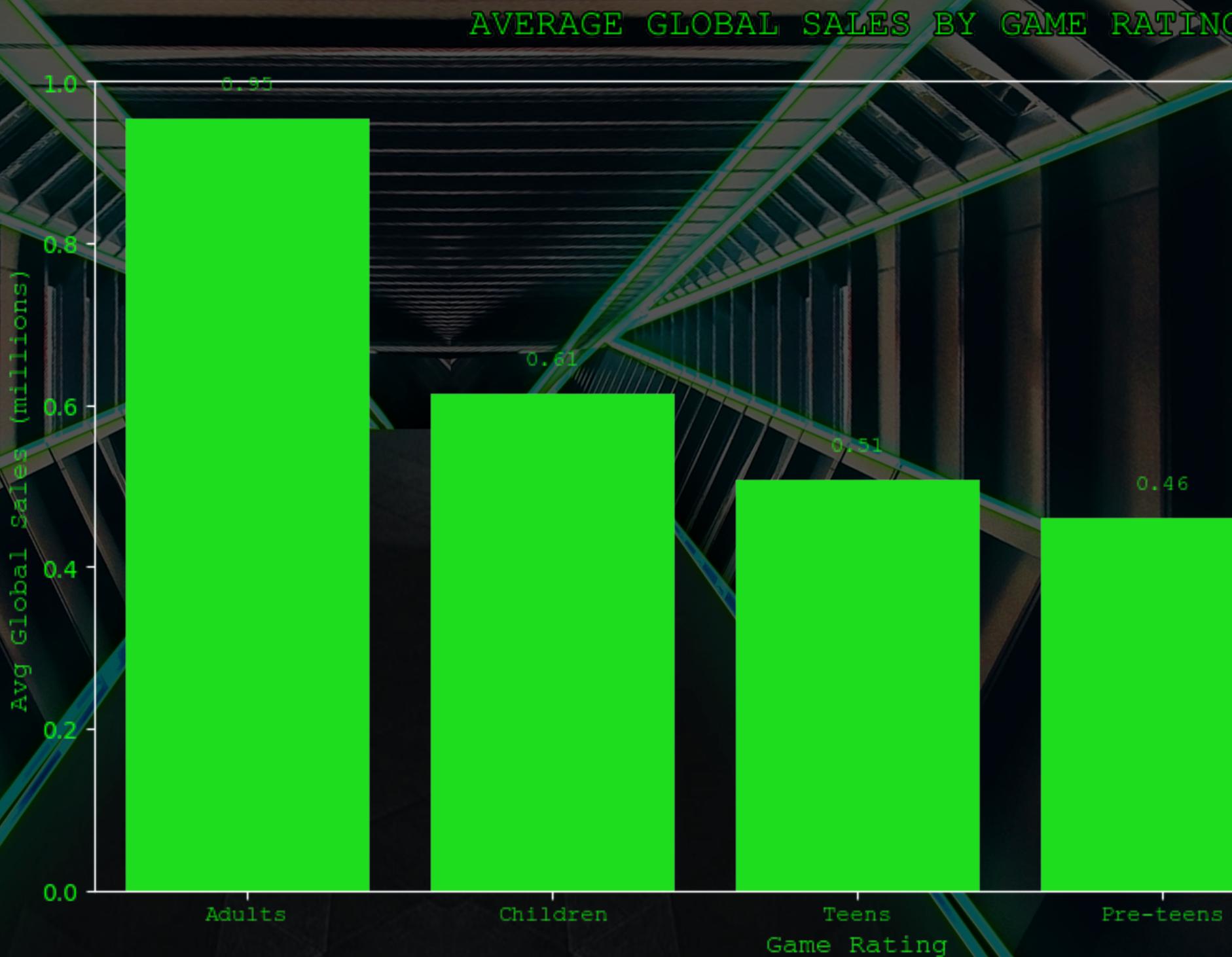
Key Features

Genre, Platform, Publisher,
Rating, Sales data

Cleaning Steps

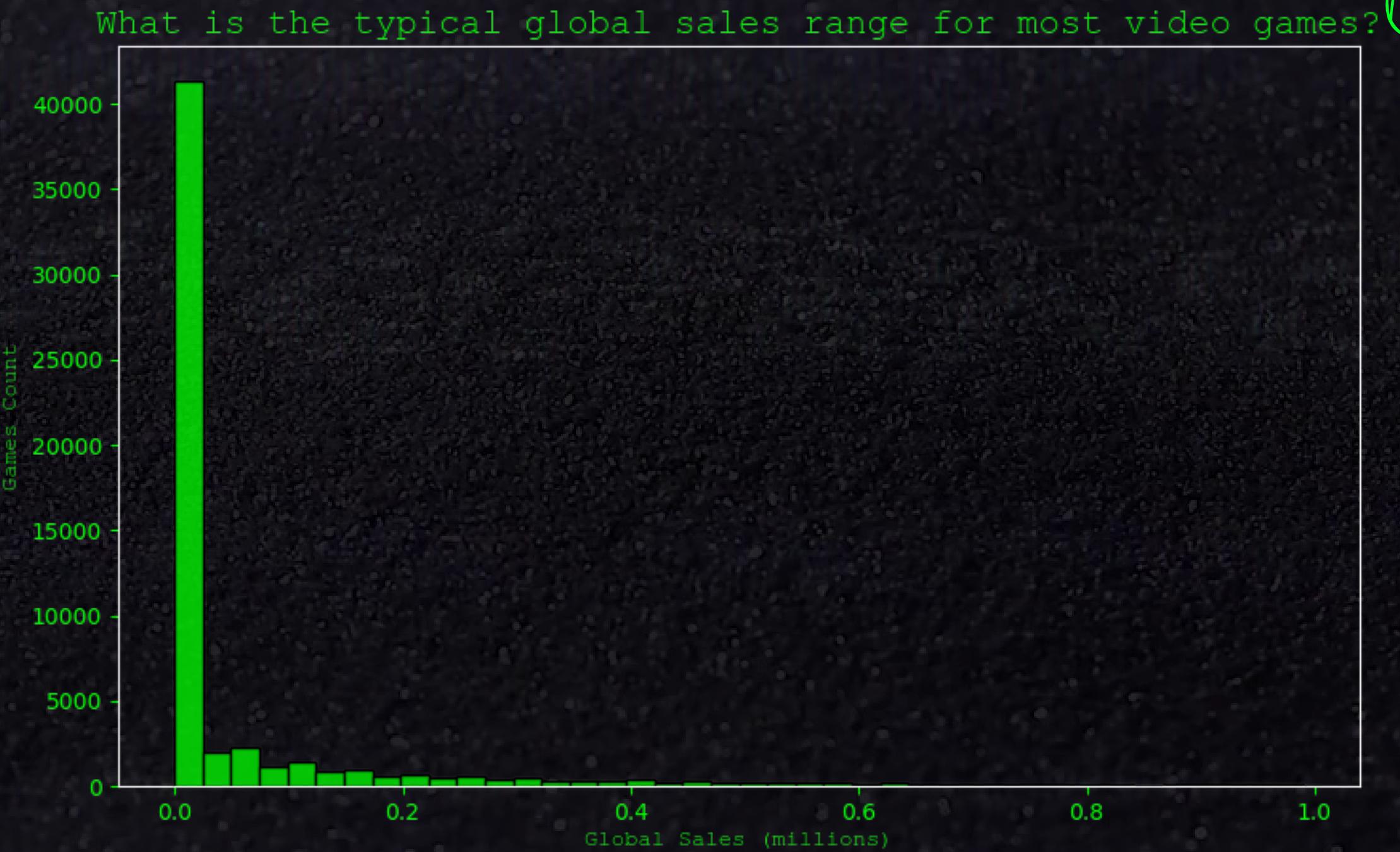
Dropped duplicates, standardized
genres, removed
missing/inconsistent data.

Contrary to common belief, mature-rated games outsell child-friendly titles on average.



WHY CLASSIFICATION?

- Most games sell under 1M copies
- Sales vary too much to predict accurately
- Focus: identify potential hits
- "Hit" = more than 1M units sold



FEATURE ENGINEERING

What we did to make the data usable and predictive:

1. Created binary target variable `is_hit`
2. Applied one-hot encoding to categorical variables
3. Converted year to integer, grouped rare genres into "Other"
4. Dropped unneeded columns like game names and developer

MODELING APPROACH

MODELING LAYER

EZ Mode (A)

SAKS LANG (B)

Ha? Hatdog (C)

Split Combo (D)

Final Score (E)

DESCRIPTION

Logistic Regression as baseline

Random Forest Classifier

Gradient Boosting (XGBoost)

80% Training / 20% Test

Final Score (E)

CONFUSION MATRIX

Confusion Matrix

True Label \ Predicted Label	Low Sales	High Sales
Low Sales -	7682	3
High Sales -	16	4175

KEY INSIGHTS

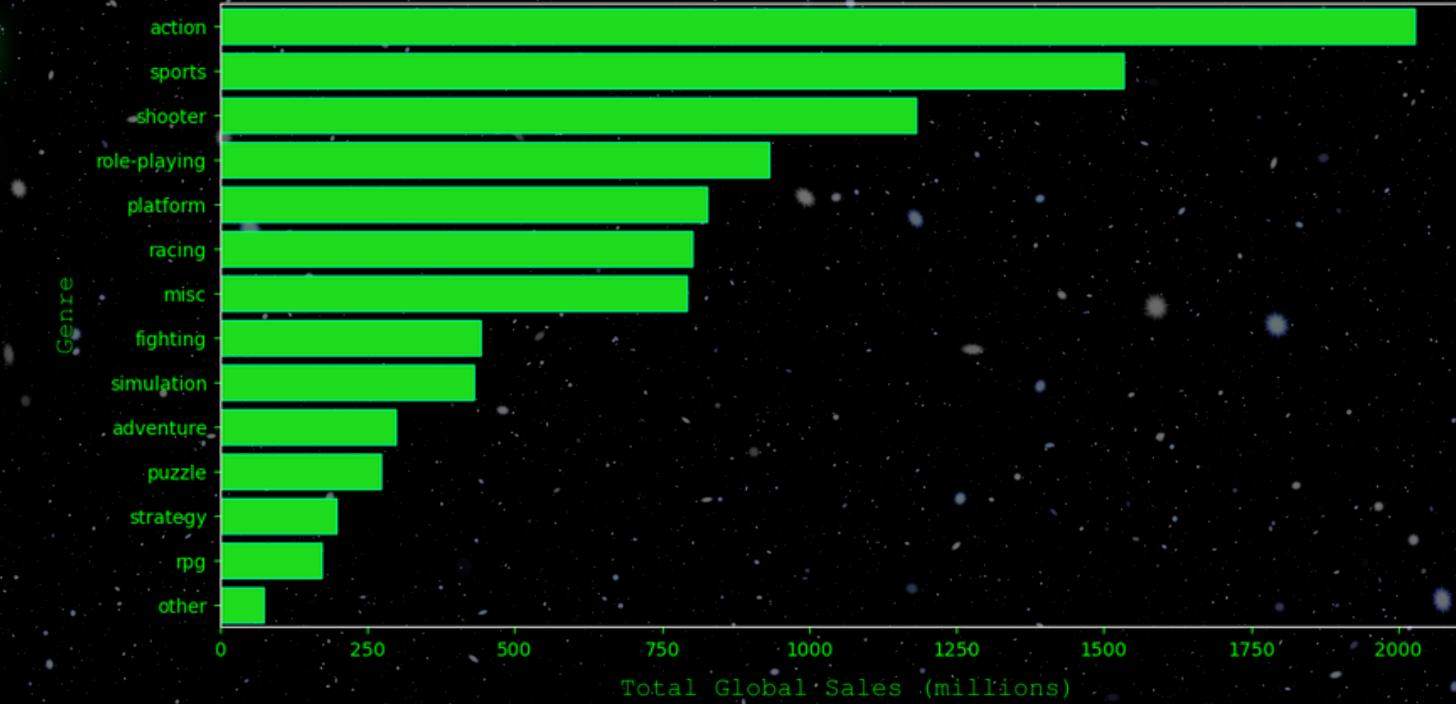
Top-selling games are driven by early design choices, not luck.

Top Predictors: Genre, Platform, Publisher

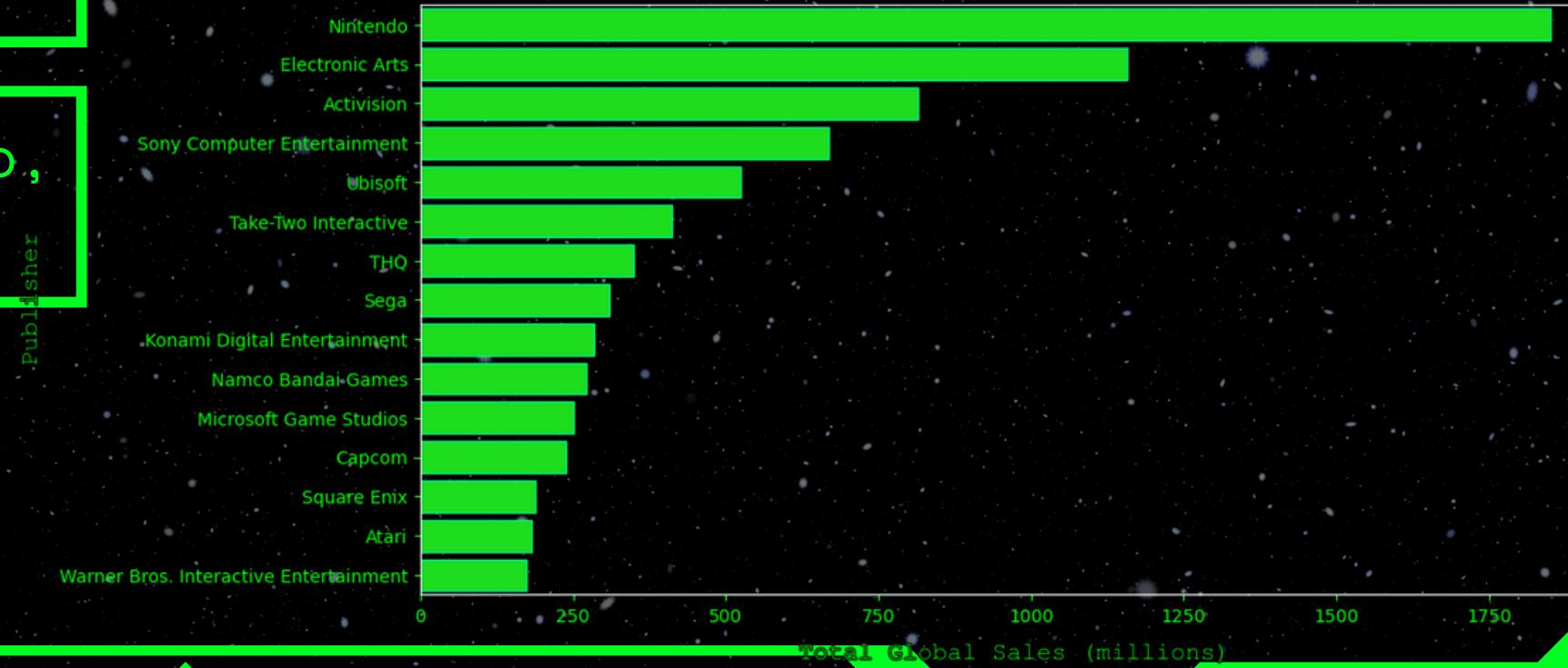
Winning Genres: Action, Sports, Shooter dominate in hit games

Big Publishers Win: EA, Nintendo, Ubisoft push most blockbusters

TOTAL GLOBAL SALES BY GENRE



TOP 15 PUBLISHERS BY GLOBAL SALES



TRENDS BY YEAR



Peak period:

2005 TO
2015



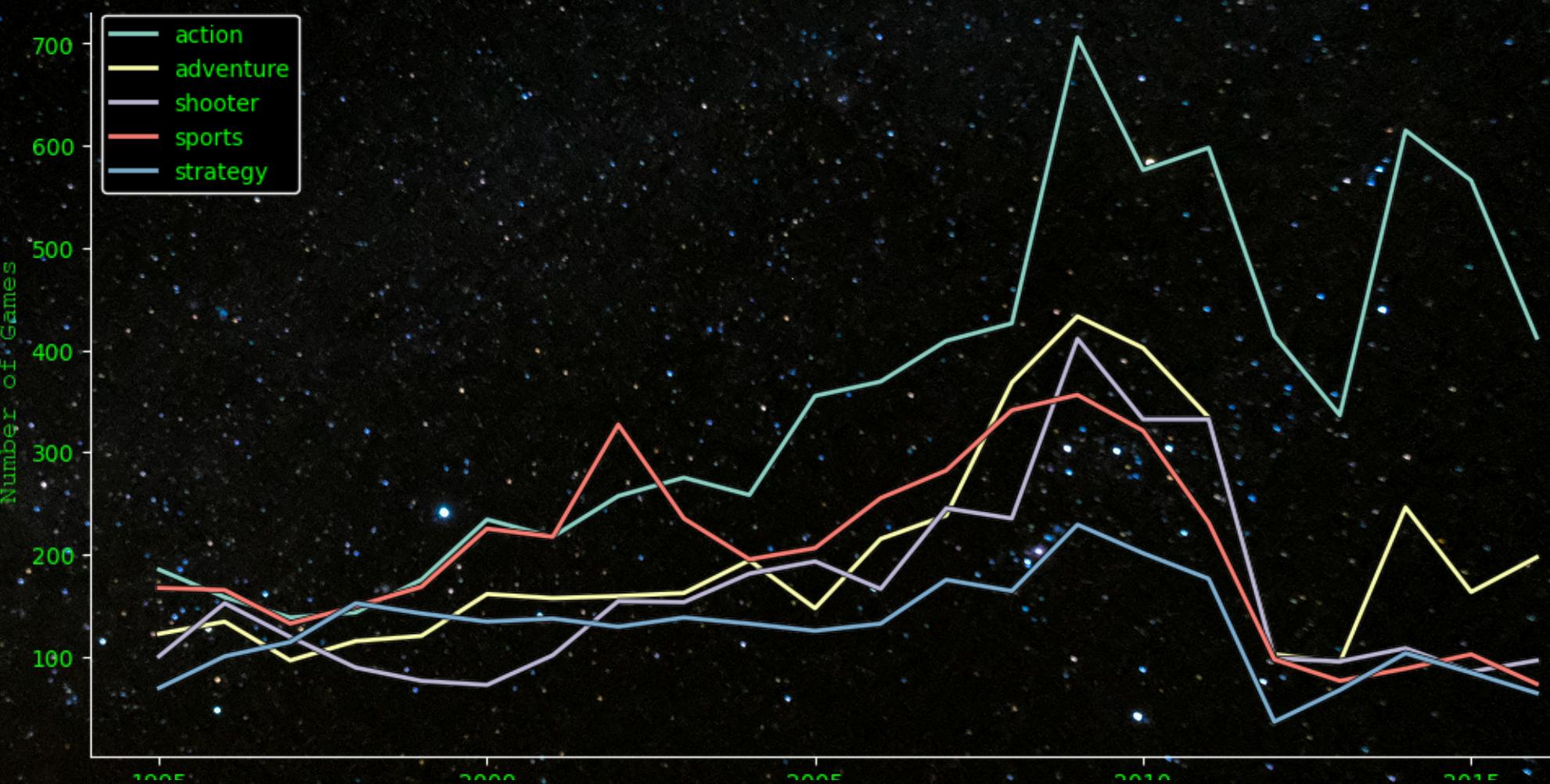
Sales declining
after 2020

COVID
EFFECT



Fewer launches but

HIGHER
PRODUCTION
VALUE



REAL-WORLD APPLICATIONS

- Predict sales potential before launch
- Optimize marketing strategy
- Improve budget allocation and platform selection

PRESS START

CHALLENGES AND LESSONS

Our modeling journey had 3 major bosses to beat...

- 1 Cleaning genres and platforms was tedious
- 2 Regression model underperformed due to skewed data
- 3 Learned importance of feature engineering and class balancing

CONCLUSION

Key Findings:

- Most games sell under 1 million copies
- Platform and genre strongly influence success
- Mature-rated games (M) show higher average sales
- Each top platform has its own best-selling game

ML Model:

- Random Forest classifier achieved 99% accuracy
- This classification report shows perfect precision and recall for both classes.

Business Insight:

- Early game features can guide smart decisions on marketing, budget, and platform focus, even without review scores.

FUTURE WORK

What improvements could enhance future sales predictions for video games?

A: Add user reviews and search trends

C: Apply deep learning/NLP on game descriptions

B: Train regression to predict sales volume

D: All of the above



THANK YOU

Any questions?