Investing in the Future

Early Access Games

Problem Statement:

The team of Data scientist has been task to help game publishers make better choices when
selecting games to invest in based on the game’s description.

Outline

Introduction to Investing in Games

Collecting Data from Steam

EDA

Model Deployment

Conclusion

Investing in Games

Benefits of Investing in Early Access Games

Early access games offer companies the opportunity to invest in a game before it is officially released. This allows companies to get in on the ground floor and reap the rewards of a successful game before anyone else.

- Monetary (ROI)
- Build relationships with developers
- Build portfolio and visibility in the industry
- Gain insights into gaming market
- Being first to have access to new and innovative games & technologies

The Risk/Costs of Investing

Resources are finite

No guarantee of success

Time consuming process

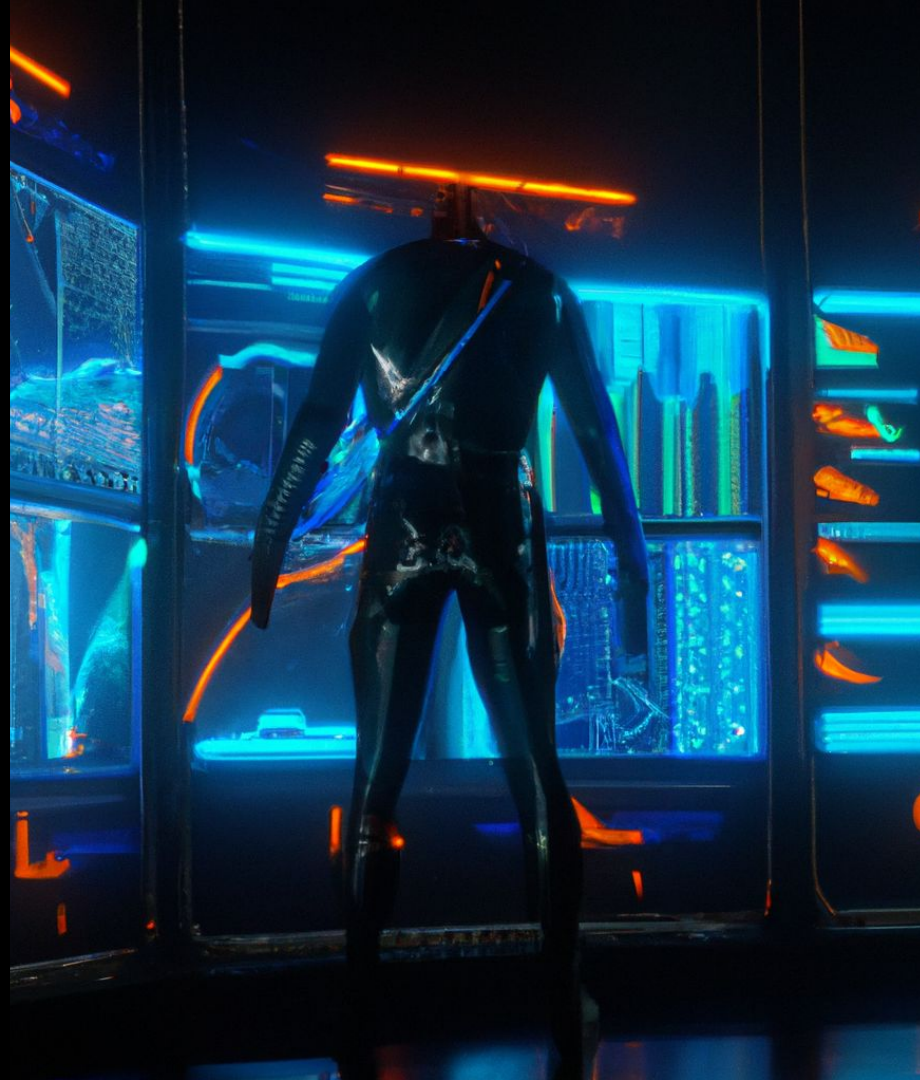
- 2 yrs of refinement before release

Data Collection, Preparation, & Analysis

Data Collection

- Steam Valve
  - Scraped using Selenium and BeautifulSoup
- List of 50k games
  - Game Title
  - Urls
  - Prices
  - Released Date
  - Descriptions
  - Reviews
  - User-defined Tags

STEAM®



Data Wrangling

Original Dataset: 50K Row, 7 Col

- Number of duplicate: 0
- Majority data types: string
- All columns have missing data (<0.01%)

Removed Missing date 49.1K Row, 7 Col

Drop all rows that contain any Na

Clean Datatype 49.1K Row, 7 Col

Convert respective columns data to:

- Datetime
- Price in float type
- Clean description text
- Convert all string data to lowercase

Creating new features: 49.1K Row, 10 Col

Creating columns that Review type, Percentage of positive reviews, Total Number of reviews

Filtering desired data: 26K Row, 10Col

Example:

About 40% data which did not have desired target variable are dropped (games with no reviews or insufficient number of reviews to be given a ‘Review type’ label)

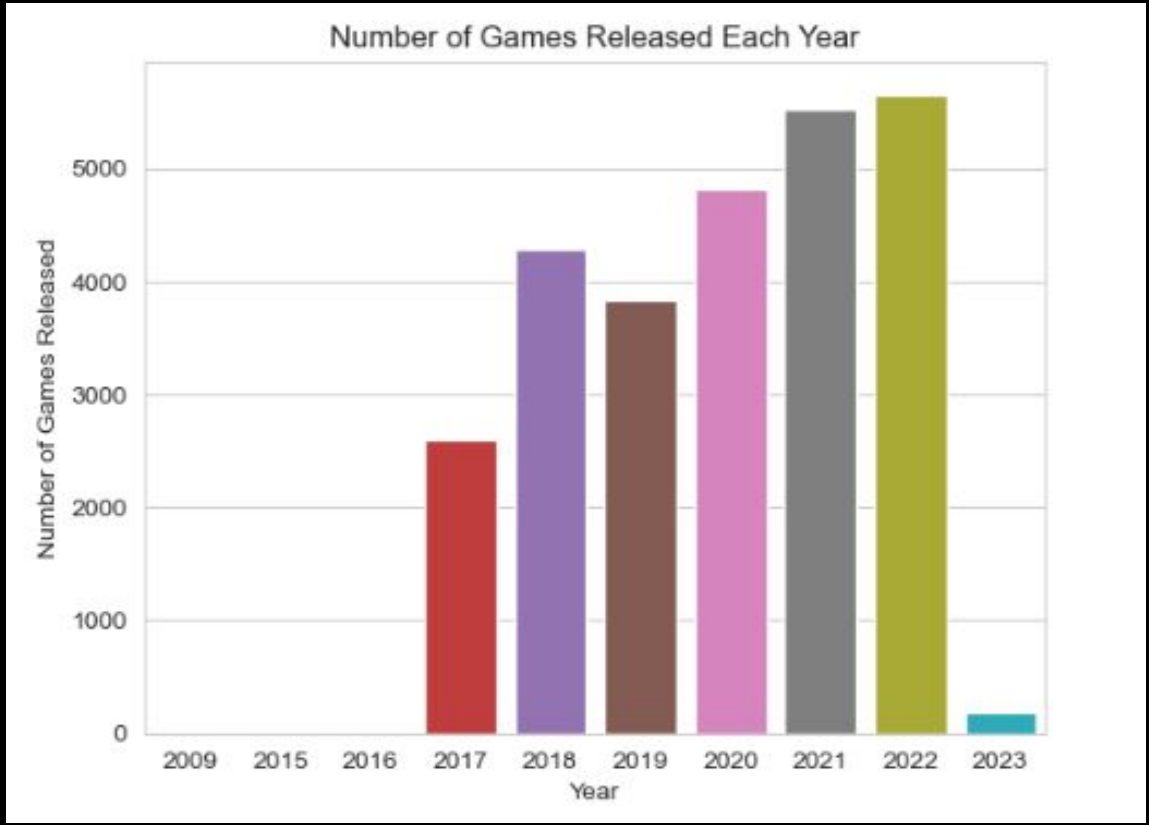
EDA

Identifying patterns and trends

Exploring our data

Over the years, there has been an increase in game development.

Steam has proactively started to record data since 2017.

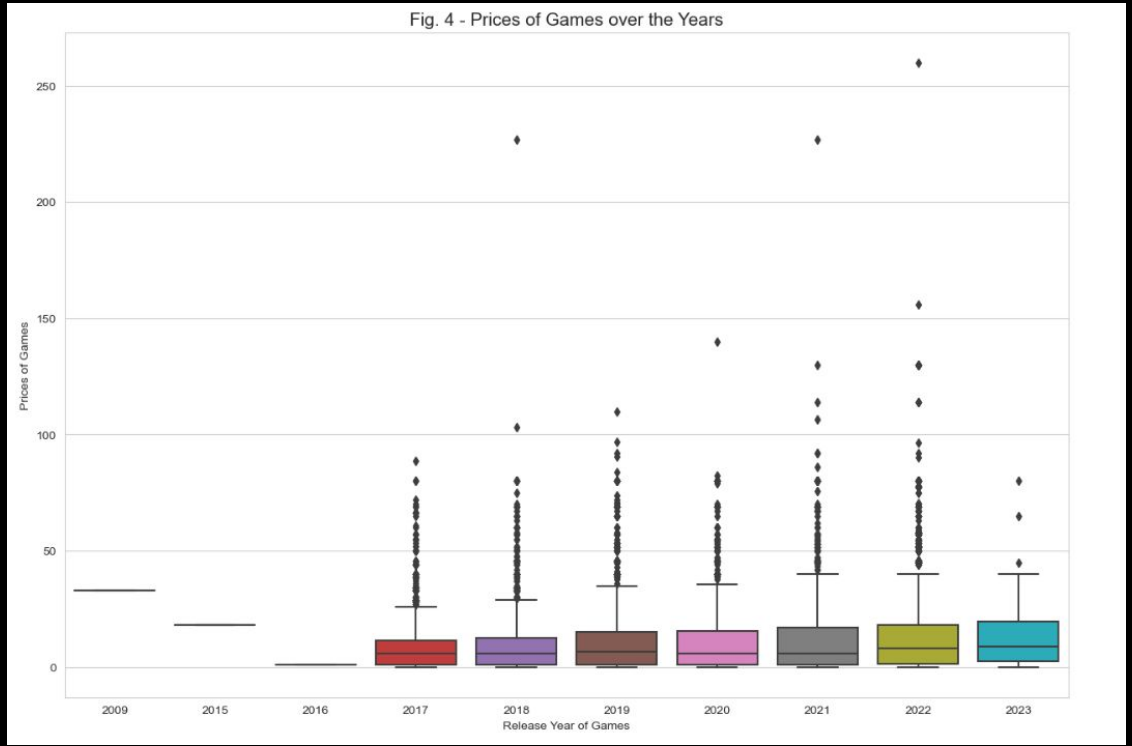


WashingtonPost article 2020- https://www.washingtonpost.com/video-games/2020/05/12/video-game-industry-coronavirus/
Forbes article 2021 on how the gaming industry has leveled up during the pandemic - ttps://www.forbes.com/sites/forbestechcouncil/2021/06/17/how-the-gaming-industry-has-leveled-up-during-the-pandemic/?sh=30c984e7297c

Price of Games

Discounting the years 2016 and before as well as 2023 (since we're only at the 1st month), the ceiling prices of games seems to have increase year on year, with more games crossing the $100 mark.

Games are getting more expensive!!!

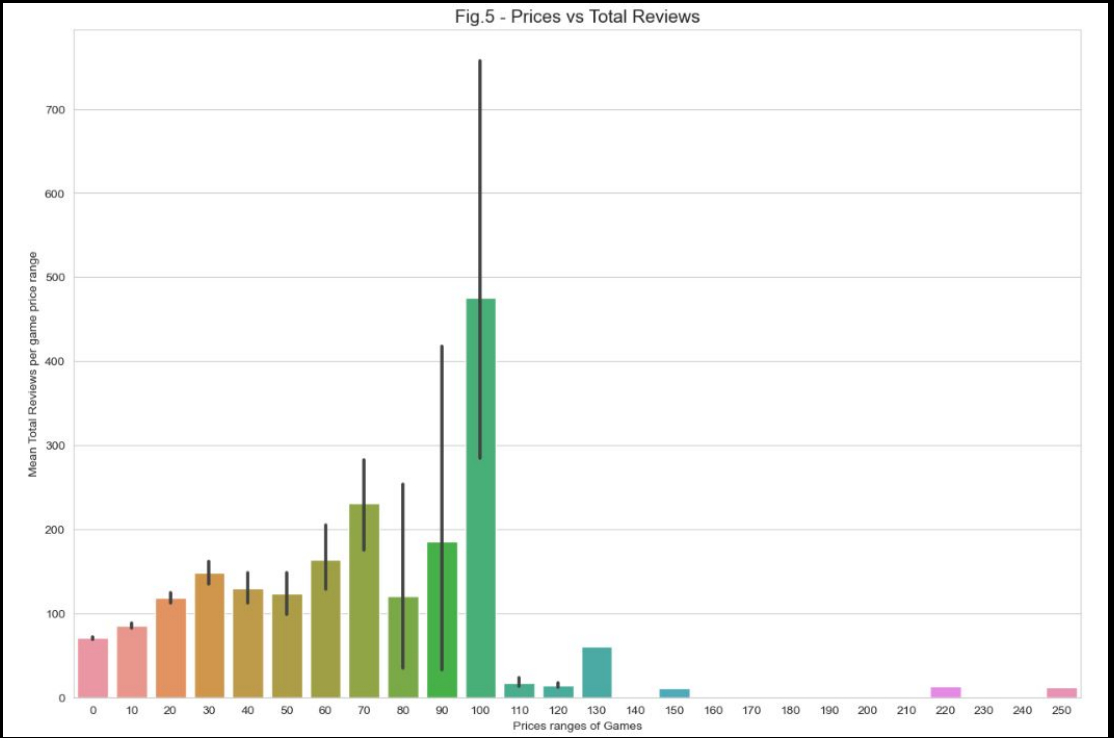


Price & Total Reviews

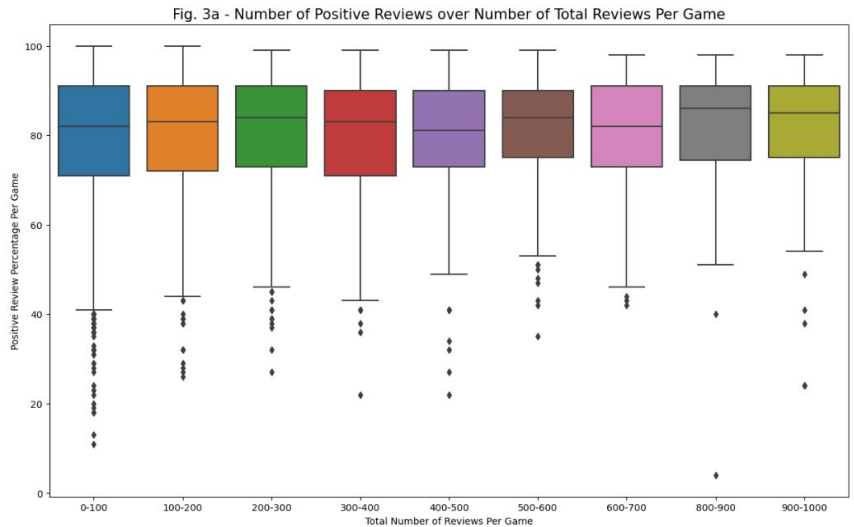
$20 - $100 Games
- > 100 reviews

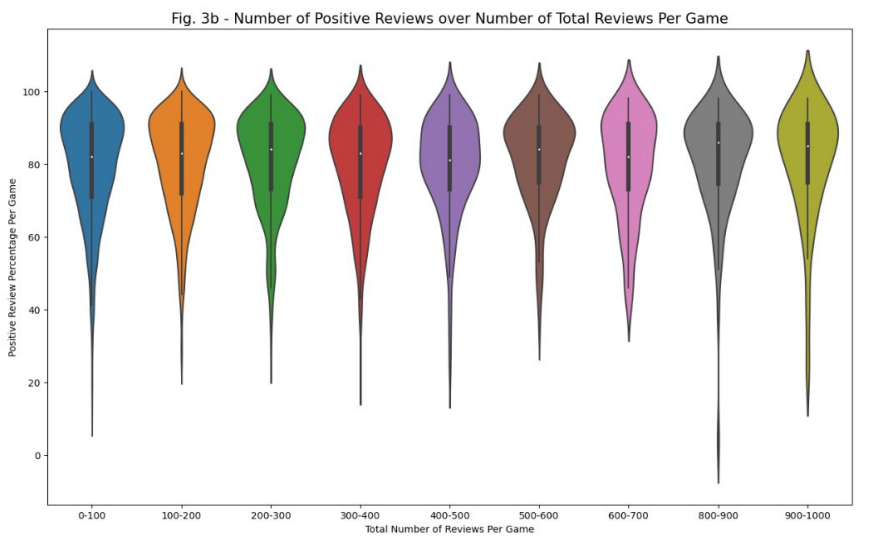
The number of reviews falls drastically when a game costs more than $100. It could be due to the high price that leads to less people buying the games, hence resulting in a minimal count of reviews.

There are most number of reviews in games that costs in the $70 range and $100 range(the most number of reviews).



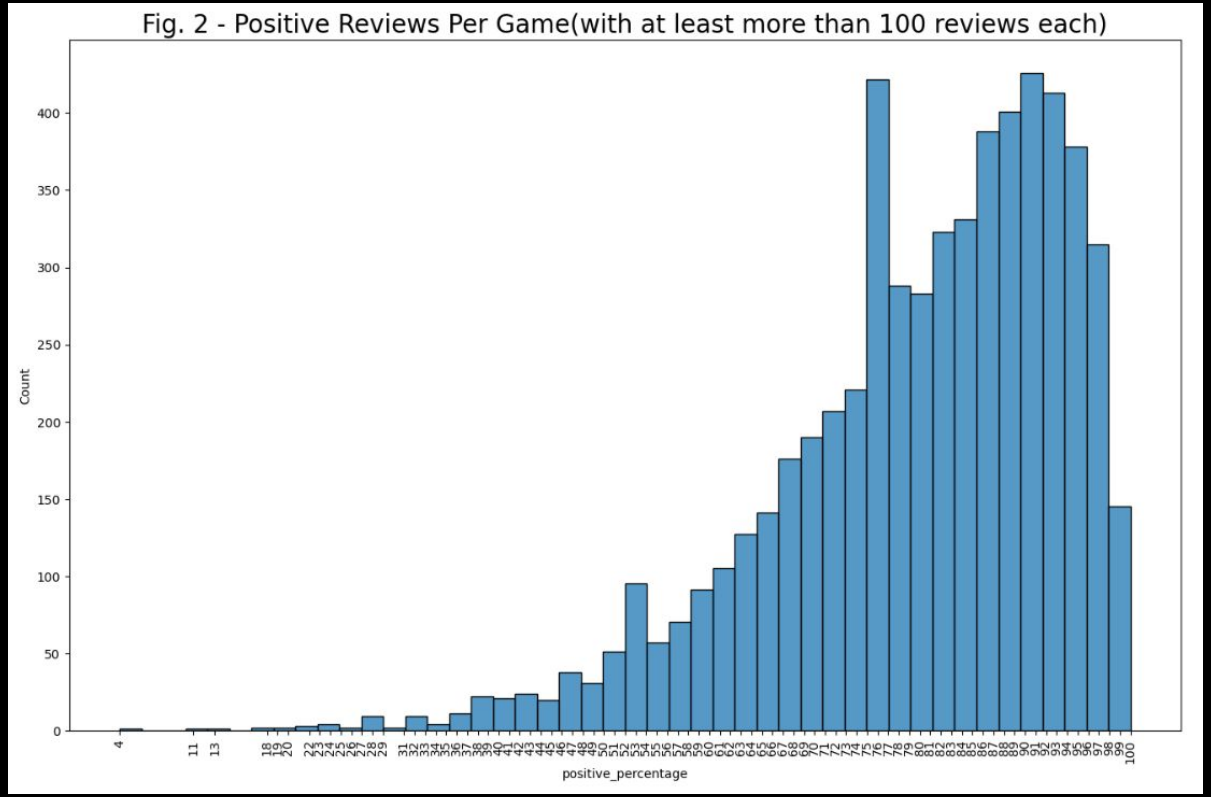
Number of Positive Reviews vs Total Reviews Per Game





Positive Reviews Per Game

In Fig.2, there is a sharp increase in the counts of 75% positive reviews.



OLS Summary

It is found that the p-value of the game price is >0.05.

As such, we infer that the price has no significant effect on the percentage of positive reviews a game receives..

OLS Regression Results

Dep. Variable:	positive_percentage	**R-squared:**	0.035
Model:	OLS	**Adj. R-squared:**	0.035
Method:	Least Squares	**F-statistic:**	326.3
Date:	Tue, 31 Jan 2023	**Prob (F-statistic):**	4.08e-208
Time:	00:12:10	**Log-Likelihood:**	-1.1545e+05
No. Observations:	26910	**AIC:**	2.309e+05
Df Residuals:	26906	**BIC:**	2.309e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3767.3405	132.539	-28.424	0.000	-4027.125	-3507.556
price	0.0075	0.009	0.874	0.382	-0.009	0.024
r_year	1.9029	0.066	28.999	0.000	1.774	2.031
total_reviews	0.0113	0.001	13.744	0.000	0.010	0.013

Omnibus:	3677.448	**Durbin-Watson:**	1.956
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5489.971
Skew:	-1.006	**Prob(JB):**	0.00
Kurtosis:	3.922	**Cond. No.**	2.49e+06

Defining our target variable:

70%

Steam labelled the game's review type 'Positive' when there is at least 70% of the reviews being positive ones.

10 reviews

Steam give games a 'Negative', 'Mixed' or 'Positive' review type only if the game has minimum 10 reviews'

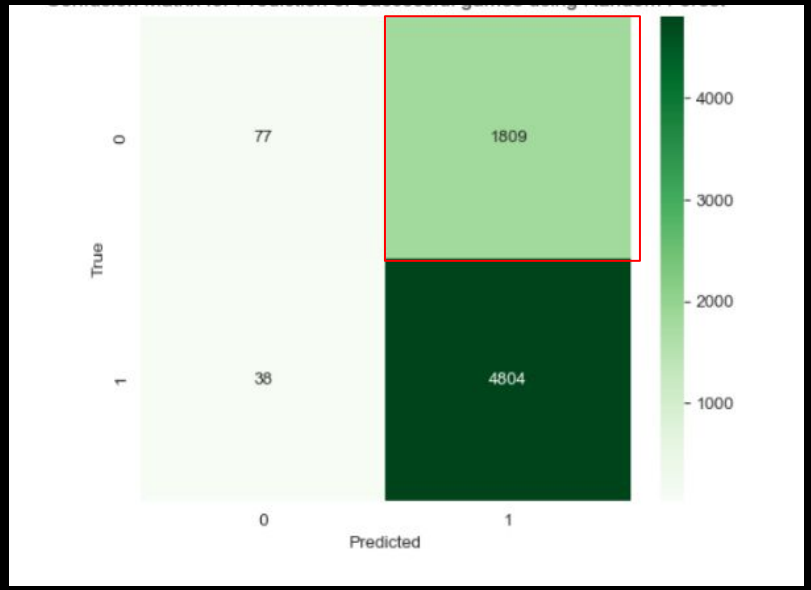
'Successful' Game

We defined games with at least 70% positive reviews (a.k.a labeled 'Positive' for its review type) to be successful game

Modelling

Baseline Model: Random Forest

As compared to our proportion of “success” to “unsuccessful” games, roughly 72% : 28%, our baseline random forest model without stop word removal had a precision score of 73%.



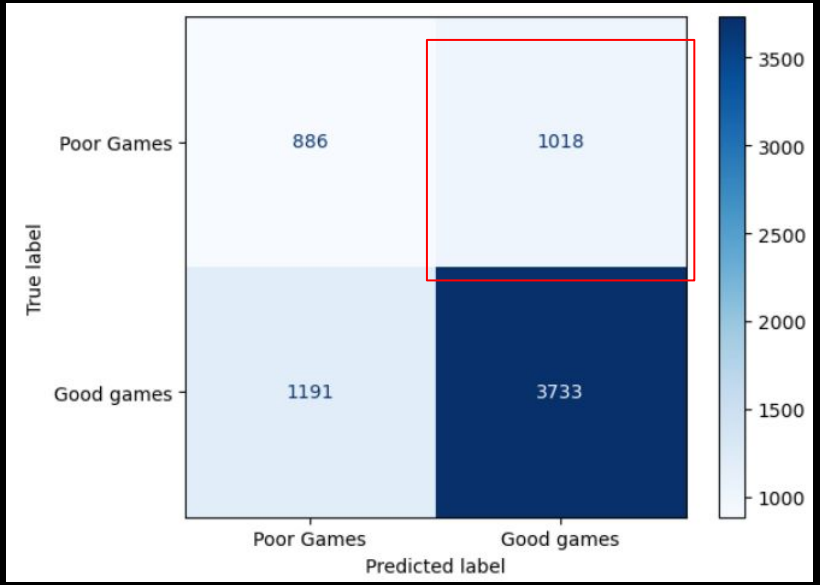
	precision	recall	f1-score	support
0	0.67	0.04	0.08	1886
1	0.73	0.99	0.84	4842
accuracy			0.73	6728
macro avg	0.70	0.52	0.46	6728
weighted avg	0.71	0.73	0.63	6728

Processing involved:

- CountVectorizer (unigram)
- No Stemming
- ~40K features

Model A:
Naive Bayes

Significantly more better performance than Baseline model. However, the number of False Positive can be improved.



	precision	recall	f1-score	support
0	0.43	0.47	0.45	1904
1	0.79	0.76	0.77	4924
accuracy			0.68	6828
macro avg	0.61	0.61	0.61	6828
weighted avg	0.69	0.68	0.68	6828

Processing involved:

- CountVectorizer (unigram)
- No Stemming
- ~40K features

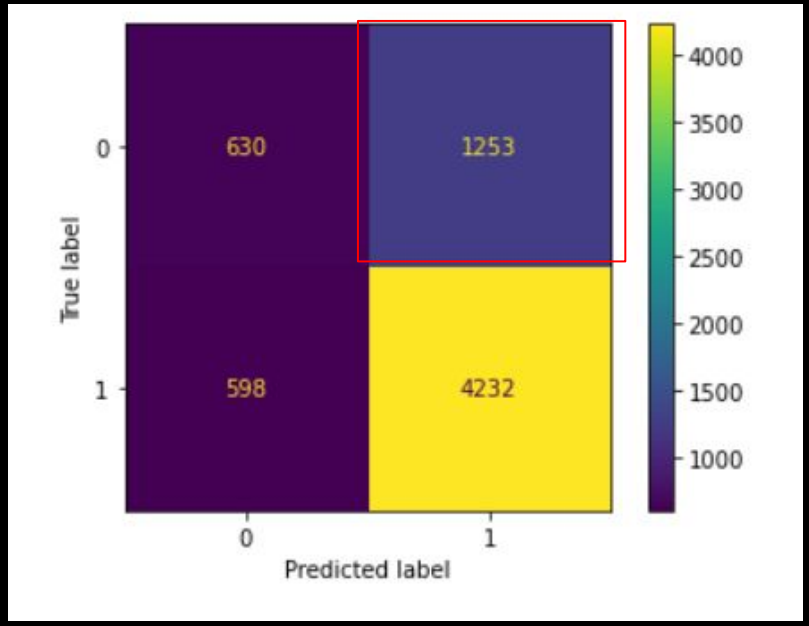
Model B:
Logistic regression with game-tags

Significantly more False negatives as compare to our other models but a strong recall score for predicting class 1 (‘successful’).

Incorporated the user tags as a feature for this model.

<u>Processing involved:</u>

- CountVectorizer (unigram,bigram,trigram)
- Stemming (nltk)
- Stopwords
- ~20K features



	precision	recall	f1-score	support
0	0.51	0.33	0.41	1883
1	0.77	0.88	0.82	4830
accuracy			0.72	6713
macro avg	0.64	0.61	0.61	6713
weighted avg	0.70	0.72	0.70	6713

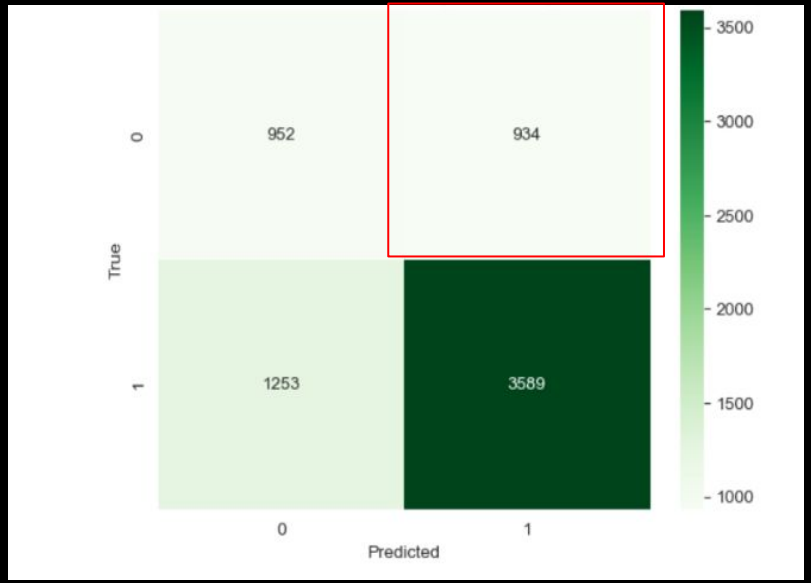
Model C:
SpaCy Lemmatization with Complement Naive Bayes

Our focus in on the precision score of 1 which for this model was 0.79, one of our stronger precision scores.

We used Spacy to lemmatize the dataset.

<u>Processing involved:</u>

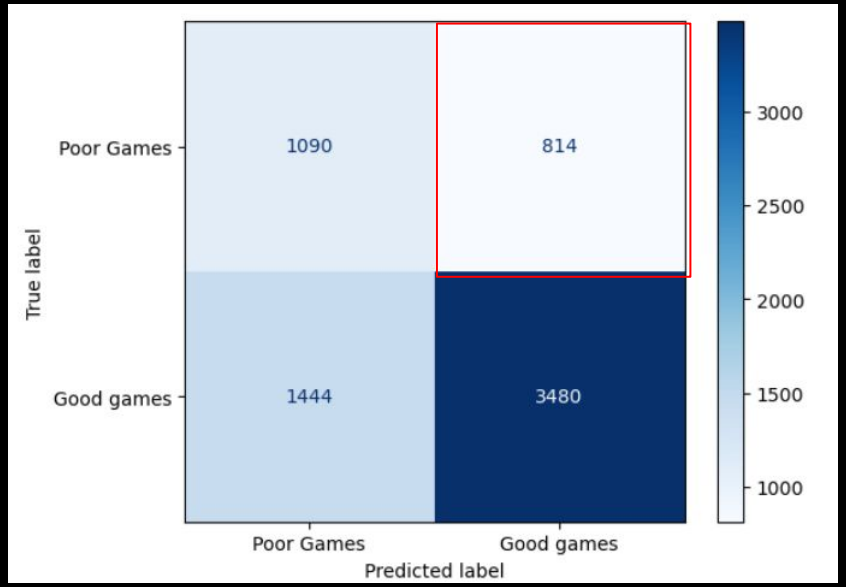
- CountVectorizer (unigram,bigram,trigram)
- Stemming (Spacy)
- Stopwords
- 40K features



	precision	recall	f1-score	support
0	0.43	0.50	0.47	1886
1	0.79	0.74	0.77	4842
accuracy			0.67	6728
macro avg	0.61	0.62	0.62	6728
weighted avg	0.69	0.67	0.68	6728

Model D:
Complement Naive Bayes remove of common terms

Model D approach is based on the suspicion that top common N-Grams in the two classes of games is impacting our model performance adversely.



	precision	recall	f1-score	support
0	0.43	0.57	0.49	1904
1	0.81	0.71	0.76	4924
accuracy			0.67	6828
macro avg	0.62	0.64	0.62	6828
weighted avg	0.70	0.67	0.68	6828

Processing involved:

- CountVectorizer (unigram)
- Stemming (nltk)
- Removed top common Unigrams, Bigrams, and Trigrams found in both class of games
- 9K features

Summary of our model findings:

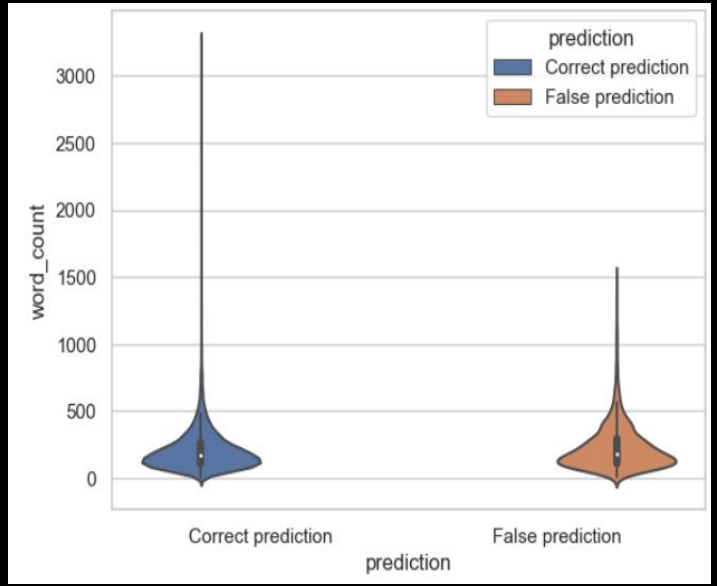
1. The model has achieved the goal to improve investors chance of investing a ‘Successful Game’ based on the game’s description.
2. Based on the dataset used and the model performance, that chance increased from
   ~72% to 81%!

Limitation:

1. It was suspected that wrong predictions were a result of game’s description with low word count.
   - Post modelling analysis showed that it is not necessarily true.
2. There is an emerging trend that games description are utilising gif and images more instead of text.
3. The gaming industry is ever-changing and our model needs to adapt to new trends

Comparing description word count between correctly and wrongly predicted games

prediction	count	mean	std	min	25%	50%	75%	max
Correct prediction	4571.0	205.62	156.96	2.0	109.0	170.0	256.0	3266.0
False prediction	2257.0	229.39	175.81	7.0	112.0	179.0	292.0	1496.0



Future Exploration:

1. There are other game features that can influence a game's success such as publishers, developers, system requirements etc that typically not captured in the game's description.
    - These features can be further explored to incorporate into the model to boost its performance in selecting potential successful games
2. Alternatively, more analysis can be done to identify features associated with poorer performing games, to help investors identify risky games to avoid (*Risk Management*).
    - More granular analysis (focus on negative reviews on specific games).