ISyE 6740 - Spring 2023
Final Report

Team Member Names: Almat Yeraly

Project Title: Analysis of the Relationship Between Steam Indie Games' Descriptions and Their Success

1 Problem Statement

Independent (indie) video games are video games that are typically developed by a small team of developers. There are a lot of factors that influence the success of indie games, mainly innovative gameplay, a good story, and a unique gaming experience. Additionally, indie games are often published digitally and marketed through word of mouth, as small teams do not have the finances for physical publishing and/or large marketing campaigns.

Steam is one of the most popular digital stores where one can find indie games. When shopping for an indie game on Steam, customers see a brief trailer, screenshots, and a short text description of the game (see Figure 1 below). As with any new product, a text description is needed to lure new customers and influence their decision to buy the product. Additionally, text descriptions are the perfect way for developers and publishers to describe the gaming experience that potential buyers can expect from the game. For example, in the Figure 1, the text descriptions mentions making a choice between being a good or bad detective.

The focus of this project is to analyze the relationship between text descriptions of indie games and their success on Steam by clustering and exploring whether it is possible to distinguish successful from non-successful games based on text descriptions alone. Additionally, the relationship has been further analyzed using variable importance in several classification models, mainly tree based ensemble classifiers.

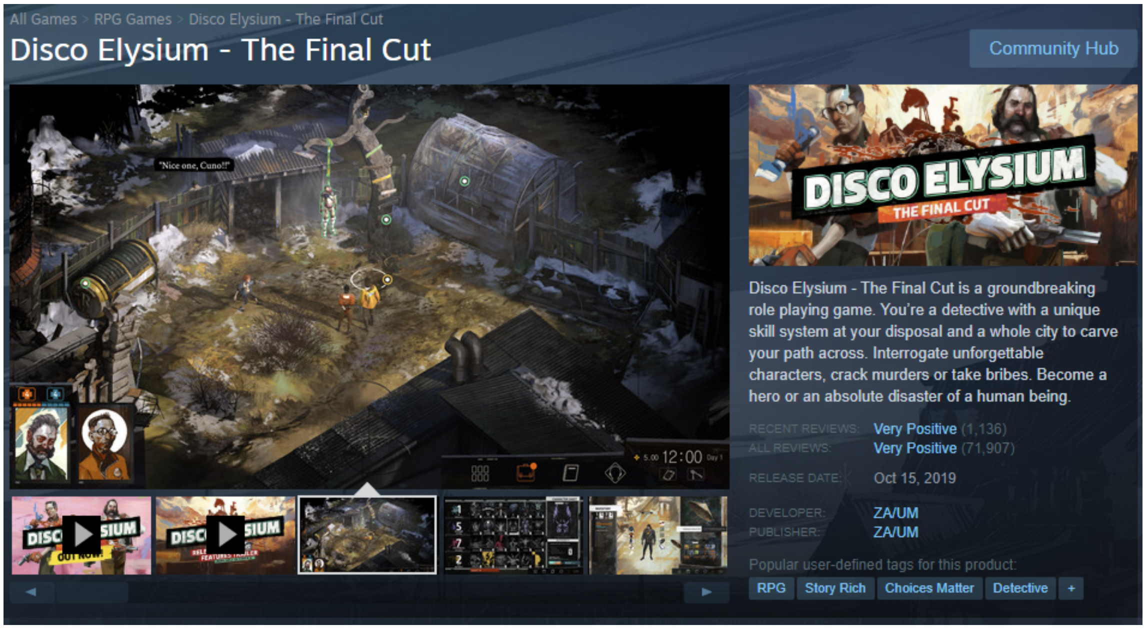
I hypothesized that there is a clear distinction between positively reviewed and non-positively reviewed games based on text descriptions, which can aid in quantifying the gaming experience. However, the findings suggest otherwise.

2 Dataset

For this project, I used the dataset called SteamGames (71k games) found on kaggle.com. The original dataset contains over 71,000 records with 39 columns, but I reduced the size reduced to 27,249 records and 20 columns due to filtering out non-indie and non-English games. After the cleaning, there are 18,354 (69.4%) games that are classified as positively reviewed. A game is classified as positive if it has at least 10 reviews of which 70% are positive. The column names and their descriptions can be seen in Table 1 below. Columns *Num Reviews, Positive Ratio, Recommendations Ratio, and Positive Class* were derived from the original dataset using columns *Positive, Negative, and Recommendations.*

Having all games with at least 10 reviews adds a lot of noise to the analysis, thus I split split the reduced dataset into three datasets based on the number of reviews: the first dataset had 10-100 reviews (n = 16,937 with 64.7% positive class), second 101-1000 (n = 7489 with 72.7% positive class), third 1001 and above (n = 2823 with 89.8% positive class). The positive class ratios differed in the given ranges, which is another reason for splitting the reduced dataset.

Figure 1: Example of an indie game on Steam


All Games > RPG Games > Disco Elysium - The Final Cut
Community Hub
"Nice one, Cuno!"
12:00
Popular user-defined tags for this product:
RPG
Story Rich
Choices Matter
Detective
+

Disco Elysium - The Final Cut

Disco Elysium - The Final Cut is a groundbreaking role playing game. You're a detective with a unique skill system at your disposal and a whole city to carve your path across. Interrogate unforgettable characters, crack murders or take bribes. Become a hero or an absolute disaster of a human being.

RECENT REVIEWS: Very Positive (1,136)
ALL REVIEWS: Very Positive (71,907)
RELEASE DATE: Oct 15, 2019
DEVELOPER: ZA/UM
PUBLISHER: ZA/UM

Table 1: Explanation of the dataset columns

Column	Explanation
AppID	Unique ID
Name	Name of the video game
Release date	Date the game was released
Price	Price of the game
DLC count	Number of additional contents
About the game	Text description of the game
Metacritic score	Score given to the game by Metacritic, 0 if missing
Positive	Number of positive reviews
Negative	Number of negative reviews
Achievements	Number of achievements in the game
Recommendations	Number of times game was recommended in the review
Average playtime forever	Average hours the game was played since release
Median playtime forever	Median hours the game was played since release
Categories	Categories provided by Steam
Tags	Tags provided by Steam
Num Reviews	Number of reviews, derived from Positive and Negative
Positive Ratio	Ratio of Positive reviews to Num Reviews
Recommendations Ratio	Ratio of Recommendations to Num Reviews
Positive Class	Binary class if game is positively reviewed, 1 is if positive

3 Methodology

3.1 Data Pre-Proccesing

For both classification and clustering models, the Term Frequency - Inverse Document Frequence (TF-IDF) of each text description and the genres as dummy columns were used as the predictor variables. TF-IDF is a numerical measure of how important a word is in a collection of texts and is one of the most popular techniques in information retrieval.

Before calculating the TF-IDF for each data point, the text descriptions were pre-processed. The pre-processing steps were:

1. Transform every letter to lowercase letters.
2. Remove punctuation and stop words such as ”is, as, was, of,” etc.
3. Tokenize each word in text.
4. Lemmatize each token.
5. Calculate TF-IDF for each data point.

In the calculation step, both 2-grams and 3-grams were explored. However, with over 27,000 data points overall, terms that had a document frequency lower than 0.1% were ignored in order to reduce computational complexity

3.2 Modeling

I analyzed each dataset using different modeling techniques for different purposes.

1. With TF-IDF, words with the highest mean TF-IDF values were extracted to see if there are any significant differences between positive and negative review classes.
2. Each dataset was reduced to three principal components to plot 2D plots and their classes to evaluate the distinction between review classes.
3. Spectral clustering, K-Means, XGBoost, and random forest and their accuracy rates were explored to evaluate whether text descriptions can capture the success of games. Since the response variable is a binary variable, one would correctly classify games randomly without any prior knowledge of them 50% of the time. Thus, the accuracy rates of the models will be compared to the baseline of 50%.
4. Additionally, to further understand text descriptions, I analyzed the XGBoost’s and random forest’s variable importance. XGBoost’s variable importance is measured by how many times a variable was used in a tree split. Random forests’s variable importance is measured by mean decrease in Gini impurity.

4 Results and Evaluation

This section will focus on each item from the modeling list above. Additionally, during modeling, there was not a significant difference in the results between 2-grams and 3-grams, thus the presented results will be of 3-grams.

4.1 TF-IDF

Table 2 below shows words with the highest mean TF-IDF values grouped by their review class and dataset. As mentioned above, this analysis was done with the purpose of evaluating whether there is a clear distinction in the text descriptions between classes. The main takeaway from this analysis is that there is a clear distinction in games with fewer reviews; however, the distinction fades as the number of reviews goes up. Additionally, the fact that the number of positively reviewed games goes up with the number of reviews is another indicator that the distinction between games fades, as games that become mainstream tend to share a lot of similarities.

Another takeaway from this analysis is that, people generally prefer simple indie games, which can be seen from two 3-grams that are shared between all datasets in the positive column: turn based combat and point click adventure. Also, it seems that games that are negative with fewer reviews, aim higher in the complexity of their games which can be seen by first person shooter and action adventure game in the negative column.

Table 2: TF-IDF words with the highest mean grouped by review class and dataset

Dataset	**Negative**	**Positive**
10-100	first person shooter day night cycle single player mode first person horror action adventure game	turn based combat full controller support fast paced action point click adventure single player campaign
101-1000	first person shooter day night cycle real time strategy turn based combat single player campaign	point click adventure turn based combat full controller support fast paced action first person shooter
1001 and above	single player mode first person shooter new explore new open world sandbox fast paced action	first person shooter point click adventure turn based combat real time strategy full controller support

4.2 PCA

The initial hypothesis was that there would be some visible distinction between the review classes in 2D plots of three principal components; however, dimensionality reduction did not produce any meaningful results as illustrated by Figure 2. The explained variances were 41.8%, 77.4%, and 70% for the 10-100, 101-1000, and 1001 above datasets, respectively.

4.3 Clustering and Classification

The results from both clustering and classification models were meaningless, as shown in Table 3. Overall, the clustering models could not exceed the 50% baseline and the classification models did so only slightly.

XGBoost and random forests used the default threshold of 0.5 for classifying data points; however, changing the threshold also did not have a significant effect on the overall accuracy rates. It is important to note that while the accuracy rates for classification models are higher for games with more reviews, it is also because the datasets were more imbalanced. Thus, the classification models defaulted to predicting as positive majority of the time, which in turn increased the accuracy rates.

Confusion matrices for each model were omitted in this section since the results were not anything of note; nevertheless, they are included in the appendix at the end of the report.

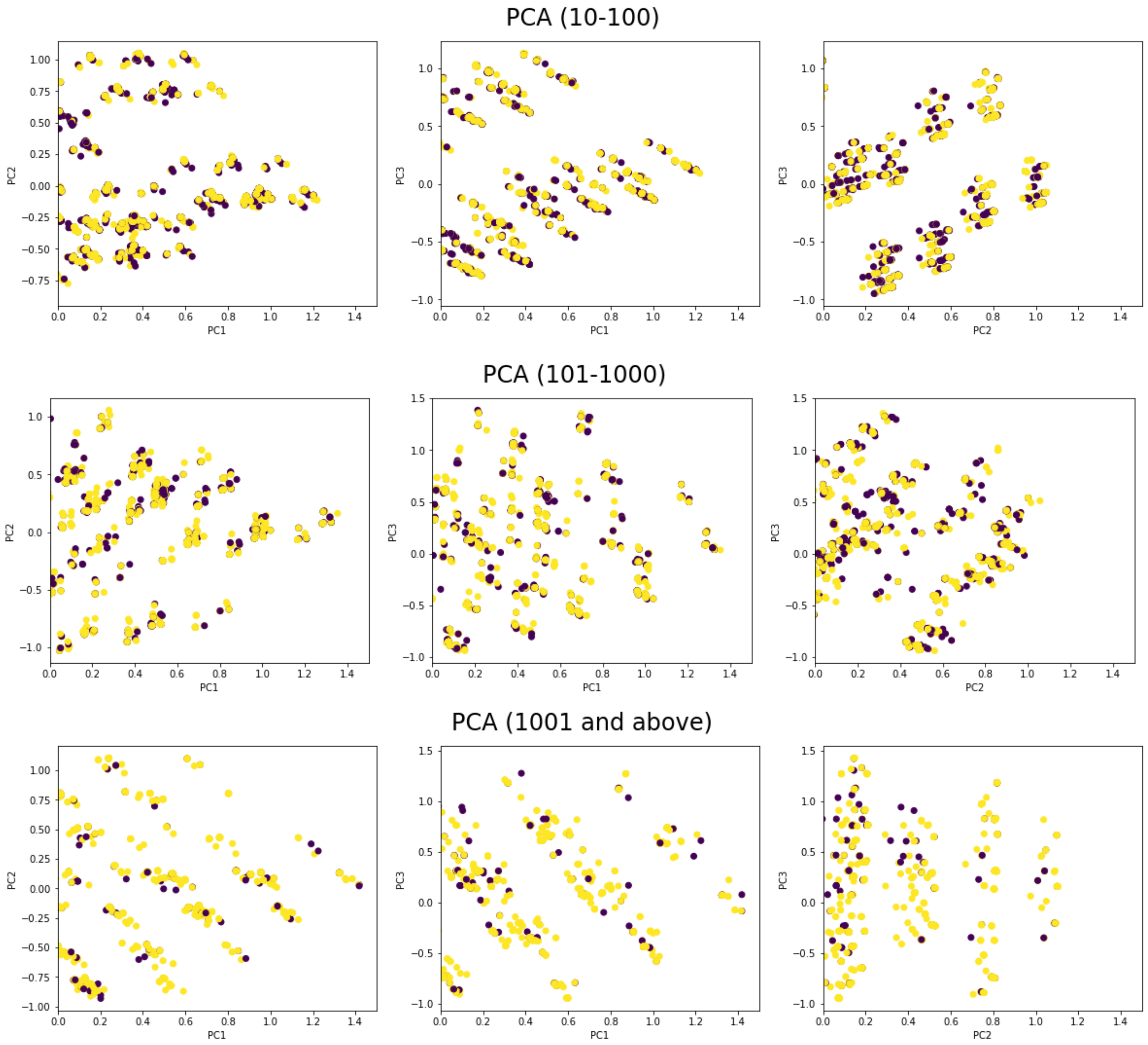
Table 3: Overall Accuracy Rates of Clustering and Classification Models

	Spectral Clustering	K-Means	XGBoost	Random Forest
10-100	48.9%	52.3%	63.1%	63.7%
101-1000	45.3%	44.2%	64.1%	68.9%
1001 and above	66.1%	46.7%	78.1%	87.6%

4.4 Classification Variable Importance

The bar plots below show the top-15 most important variables used in XGBoost and random forests modeled for each dataset. The variables that are dark red are the computed 3-grams and the green are the genre dummy variables. The key takeaway from these plots is that the text descriptions do not capture the success of games well. In all of the plots, the genres are far more important than the 3-grams.

Figure 2: 2D Plots of 3 Principal Components Grouped by Review Class



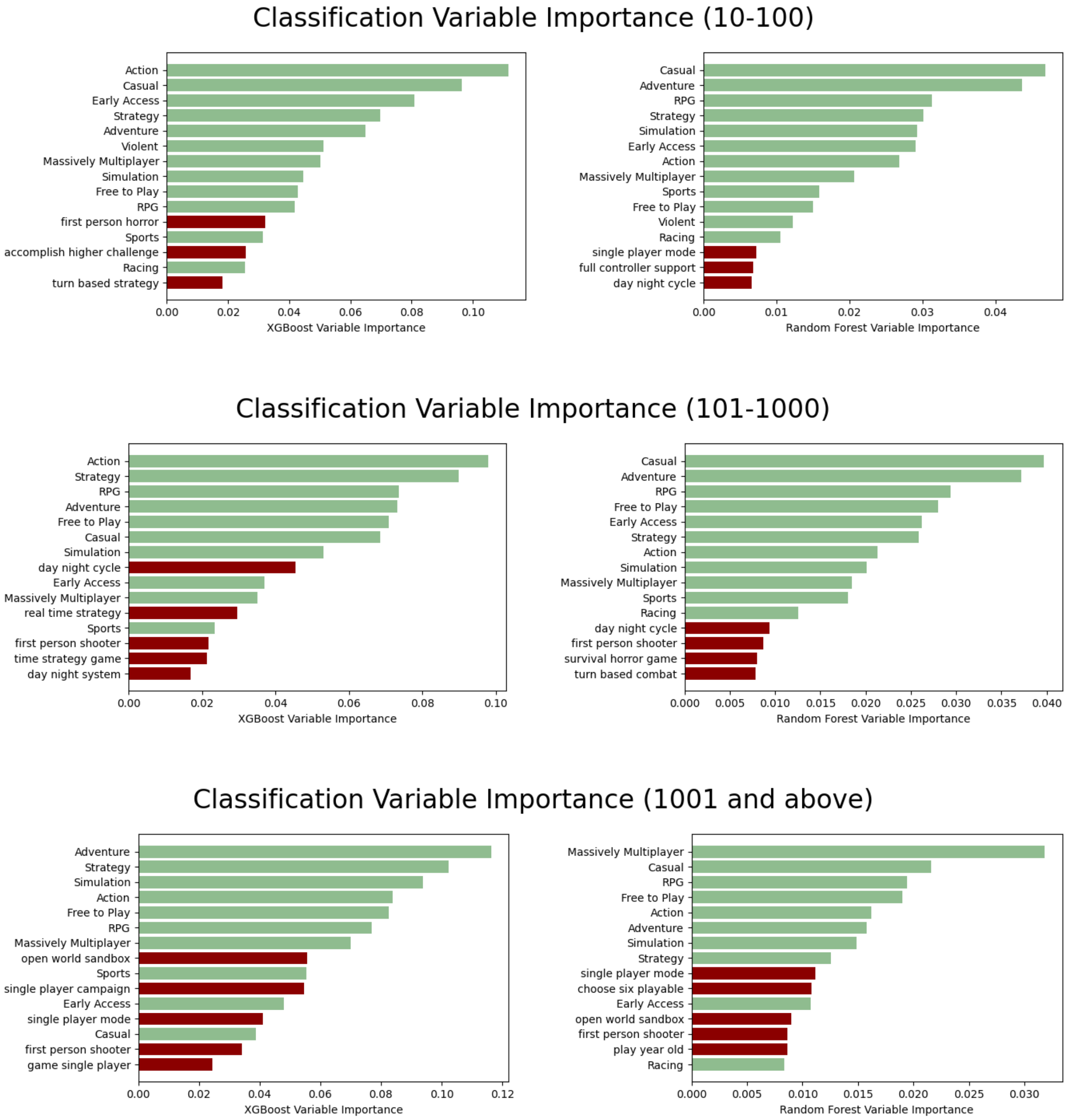
PCA (10-100)

PCA (101-1000)

PCA (1001 and above)

An interesting result that is worth mentioning is that the most important 3-grams match well with the highest mean TF-IDF values from above. Additionally, another argument could be made in favor of the finding that simpler games are preferred by looking at the most important genres: Casual and/or Adventure are consistently in the top-5 most important variables.

Figure 3: XGBoost and Random Forest Variable Importance Plots



5 Conclusion

The initial hypothesis of this project was that text descriptions are good predictors to identify successful games. This hypothesis, however, was disproven. The principal component analysis showed that there is no clear visual distinction between negatively and positively reviewed classes. Clustering models, generally, were not able to exceed the baseline, and the classification models exceeded it only slightly. Nevertheless, even though the outcome of the project was not expected, the highest mean TF-IDF values and the most important variables in the classification models still produced some interesting insights into which games are preferred in the community.

To expand this study, more variables could be explored, such as the average playtime, price, or DLC counts. Additionally, other variables that capture user experience could provide more insight, for example, using reviews written by gamers. These variables could add additional explanatory power and help determine the success of an indie game.

6 Appendix

Table 4: Confusion Matrices for 10-100 Dataset

Spectral clustering	Predicted negative	Predicted positive
Actual negative	3013	2974
Actual positive	5676	5274

K-Means	Predicted negative	Predicted positive
Actual negative	2908	3079
Actual positive	4999	5951

XGBoost	Predicted negative	Predicted positive
Actual negative	608	1198
Actual positive	677	2599

Random forest	Predicted negative	Predicted positive
Actual negative	417	1389
Actual positive	435	2841

Table 5: Confusion Matrices for 101-1000 Dataset

Spectral clustering	Predicted negative	Predicted positive
Actual negative	1277	764
Actual positive	3330	2118

K-Means	Predicted negative	Predicted positive
Actual negative	1022	1019
Actual positive	3161	2287

XGBoost	Predicted negative	Predicted positive
Actual negative	261	365
Actual positive	441	1180

Random forest	Predicted negative	Predicted positive
Actual negative	101	525
Actual positive	173	1448

Table 6: Confusion Matrices for 1001 and above Dataset

Spectral clustering	Predicted negative	Predicted positive
Actual negative	88	199
Actual positive	758	1778

K-Means	Predicted negative	Predicted positive
Actual negative	147	140
Actual positive	1364	1172

XGBoost	Predicted negative	Predicted positive
Actual negative	4	93
Actual positive	9	741

Random forest	Predicted negative	Predicted positive
Actual negative	37	60
Actual positive	134	616