Capstone Project Report - Metacritic Game Review Sentiment Analysis

Problem

What makes a good RPG game appealing to a gamer? What did gamers enjoy and dislike the most in other game genres on the Xbox One, Playstation 4 and Nintendo Switch consoles? With the next generation version of these consoles on the rise, in this project I aimed to conduct sentiment analysis on user reviews of games and examine what people said about certain titles. With the knowledge uncovered I hoped to find weaknesses and strengths presented in certain game genres and classify whether a user enjoyed a game based on their review.

Context

With the next generation gaming consoles of Microsoft's Xbox Series X and Sony's Playstation 5 recently being released, I reflected on some of the games I enjoyed the most on their predecessors and the Nintendo Switch. Some of my favorite titles included The Witcher 3: Wild Hunt, Borderlands, Kingdom Hearts series and the Dishonored series. If you are not familiar with game genres all of my favorite titles are considered role playing games (RPG) and action adventure games.

Personally, the aspects I enjoyed about these games was the versatility a player is given to customize and equip their character with different skills and armor, the open world environment and the mystical and fantasy-like characteristics of these games. I wanted to see what other gamers thought about these games and games in other genres. I decided I could perform sentiment analysis on game reviews submitted by the average gamer and not critics to get a sense of what the public actually felt.

Clientele

The model developed may be of use to different game developers. These companies may use the model to assess the sentiment of consumers and digest their feedback shortly after the release of DLC. If gamers hated the DLC they may use this knowledge to implement a patch fixing a technical deficiency or avoid using certain features in upcoming DLC. Additionally, if gamers enjoyed certain aspects of the game such as the emotes developers can focus on creating more, as in the case of Fortnite where their emotes (dancing celebrations that are used as taunts or boastful celebrations) have created a million dollar monthly revenue for Epic.

The Data

The data was scraped off Metacritic, a website where people review movies, music, video games and other types of media. For this project 15 reviews were scraped for all game titles across the Xbox One, PS4 and Switch for games with at least 15 reviews. The features scraped included:

- **title:** Title of the game
- **platform:** Console reviewer played the game on
- **metascore:** Average score given to the game by various game critics
- **metasentiment:** Overall critic sentiment classification
- **average_userscore:** Average score given to the game by users
- **average_usersentiment:** Overall user sentiment classification
- **developer:** Developer of game
- **genre:** Genre of game
- **number_of_players:** Number of players that can play the game
- **esrb_rating:** Entertainment Software Rating Board (ESRB) rating
- **release_date:** Release date of game
- **username:** Metacritic username of the game reviewer
- **userscore:** Individual user rating
- **usersentiment:** Individual user sentiment classification
- **review:** Text review left by user
- **review_date:** Date review was left by user

There were various rows with missing values for various features but as these rows all contained the null values and only accounted for 2% percent of the data, they were dropped. The only feature with a significant amount of missing values was the ‘number_of_players’ columns. About 16% of the data did not have a value assigned to this feature.

To address the null values in the ‘number_of_players’ column, the titles for the games having a null value were explored. There were a mixture of singleplayer and multiplayer games missing values for the number of players. Each title was googled to see if the game was a single or multiplayer game. Then, to reduce the amount of different, unnecessary amounts of values for the ‘number_of_players’ columns of the dataframe, the game titles for ‘No Online Multiplayer’ games were also explored. Mostly all, if not all were single player games. For simplicity, the ‘number_of_players’ column was then converted to a binary column where a game was either considered a single player game (values of ‘No Online Multiplayer’ and ‘1 Player’) or multiplayer game (all other values).

Using the lang_detect library each review was tagged with an abbreviation of the language it was written in. Different review density features were created for EDA and to potentially be used for modeling such as part of speech counts, word counts, sentence counts, punctuation counts, etc. Finally, the reviews were cleaned in the following order:

1. Transformed into lower case
2. Stripped of digits
3. Expanded contractions
4. Emojis transformed into words
5. Stripped of punctuation
6. Stripped of white space
7. Filtered from stop words
8. Lemmatized

Exploratory Data Analysis

In the exploratory data analysis portion of the project the features of the data were visually analyzed to see if any trends or correlations could be seen before conducting feature selection. A few questions about the data were explored and delved into further if the initial visual plot revealed any outstanding observations. The following diagrams show the most prominent findings.

How are the user review scores distributed in the data?

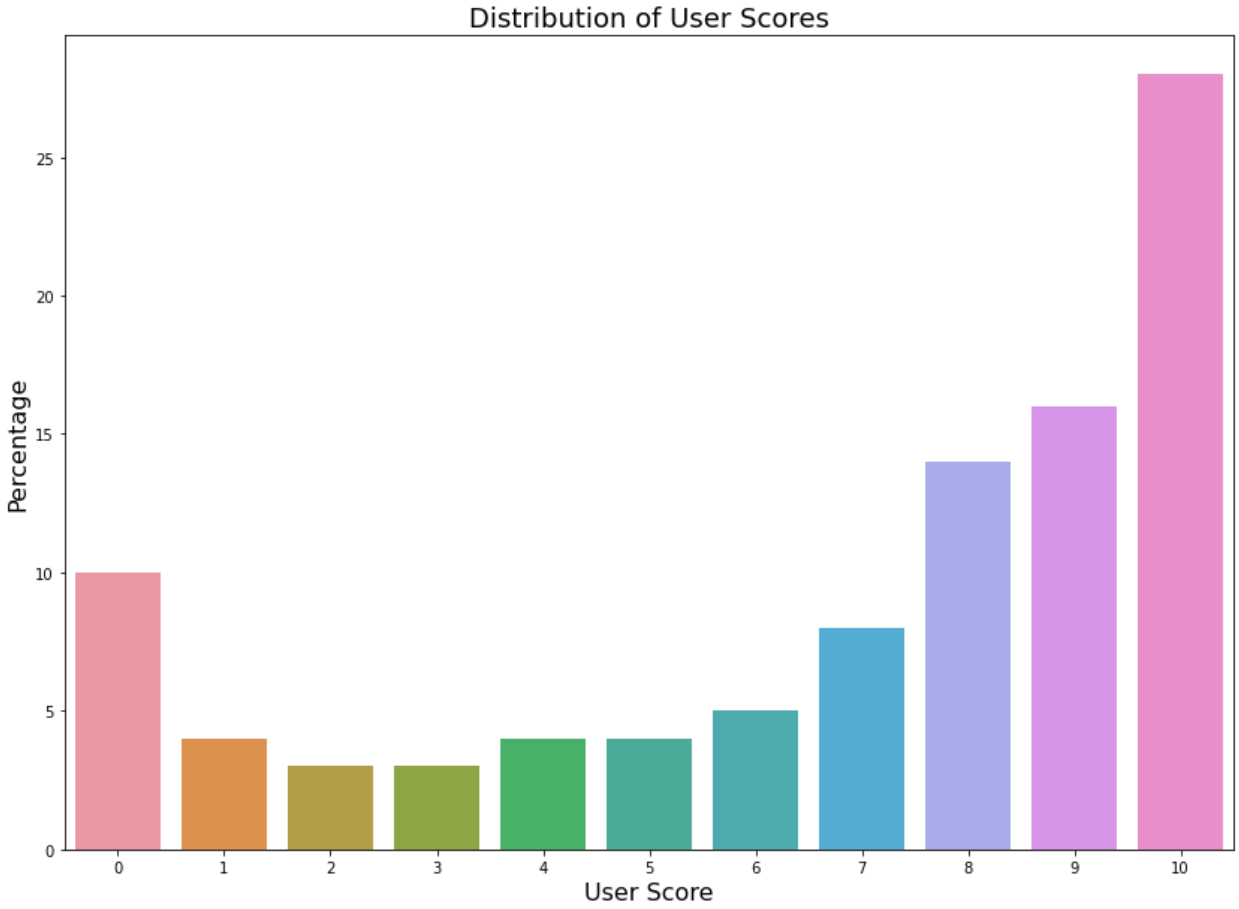

Figure 1: Distribution of user scores

Metacritic classifies a review as positive, mixed and negative by the user's score. User scores range from 0 - 10. The following ranges distinguish the user's sentiment:

- Positive: 7.5 -10
- Mixed: 5 - 7.49
- Negative: 0 - 4.99

Most of the reviews in the dataset were positive, accounting for about 57% of the total reviews shown in Fig. 1. A whopping 10% of all reviews received a 0. It seems like people tend

to enjoy video games released on console or they tend to be easily pleased with the attributes game developers add to their games.

Are positive user reviews longer than negative ones?

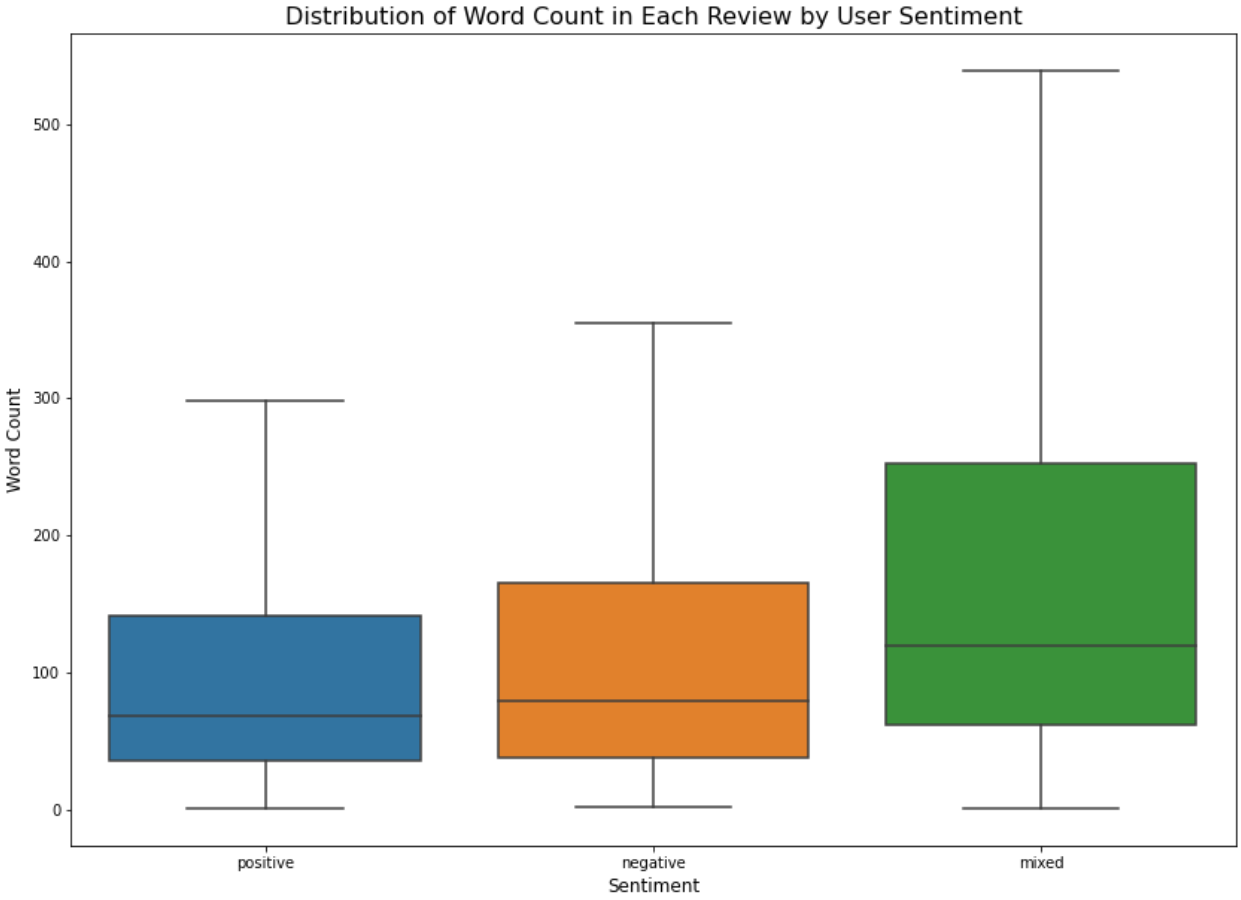

Figure 2: Variability of word counts in reviews by user sentiment

On average reviews where the user had mixed feelings about a game, encompass the most words among all reviews (Fig. 2). This could be due to the gamer expressing what he enjoyed in the game and what he disliked. With more words a reviewer with mixed feelings toward a game can vividly express why they experienced conflicting emotions.

There was a statistically significant difference among positive and negative reviews (p-value < 0.001). Positive reviews were generally the shortest type of reviews with the average positive review having somewhere between 70 and 90 words. The negative reviews may have been written in similar fashion. Perhaps accounting for the larger number of word usage from positive reviews, with additional excerpts explaining how the game could be improved.

---------- **Positive Example** ----------

"A definitive release was not necessarily needed but extremely welcomed. The added content made an already flawless game better. Playing through the game a second time was somehow more fun. I will always highly

recommend this game. The best side scrolling game Xbox One currently has to offer. Beautiful graphics, fluid controls and solid game play that almost forces you to finish in one sitting it's so riveting."
**Median word count = 69*

---------- **Negative Example** ----------

"It's not a bad game but the gameplay is an outdated one. It by no means a 10 out of 10: - "Realism" is annoying. I hate watching loot, skin and cook animations 100 times. - Parking a horse or picking up something is frustrating as you have to stand in the exact right position to do so. - 75% of the game time it takes riding to the destination to play the game a bit. - The cover system, menu system and controls are retrograde"
**Median word count = 79*

Which is the most common part of speech among positive and negative reviews? Do positive reviews have more adjectives in them than negative ones?

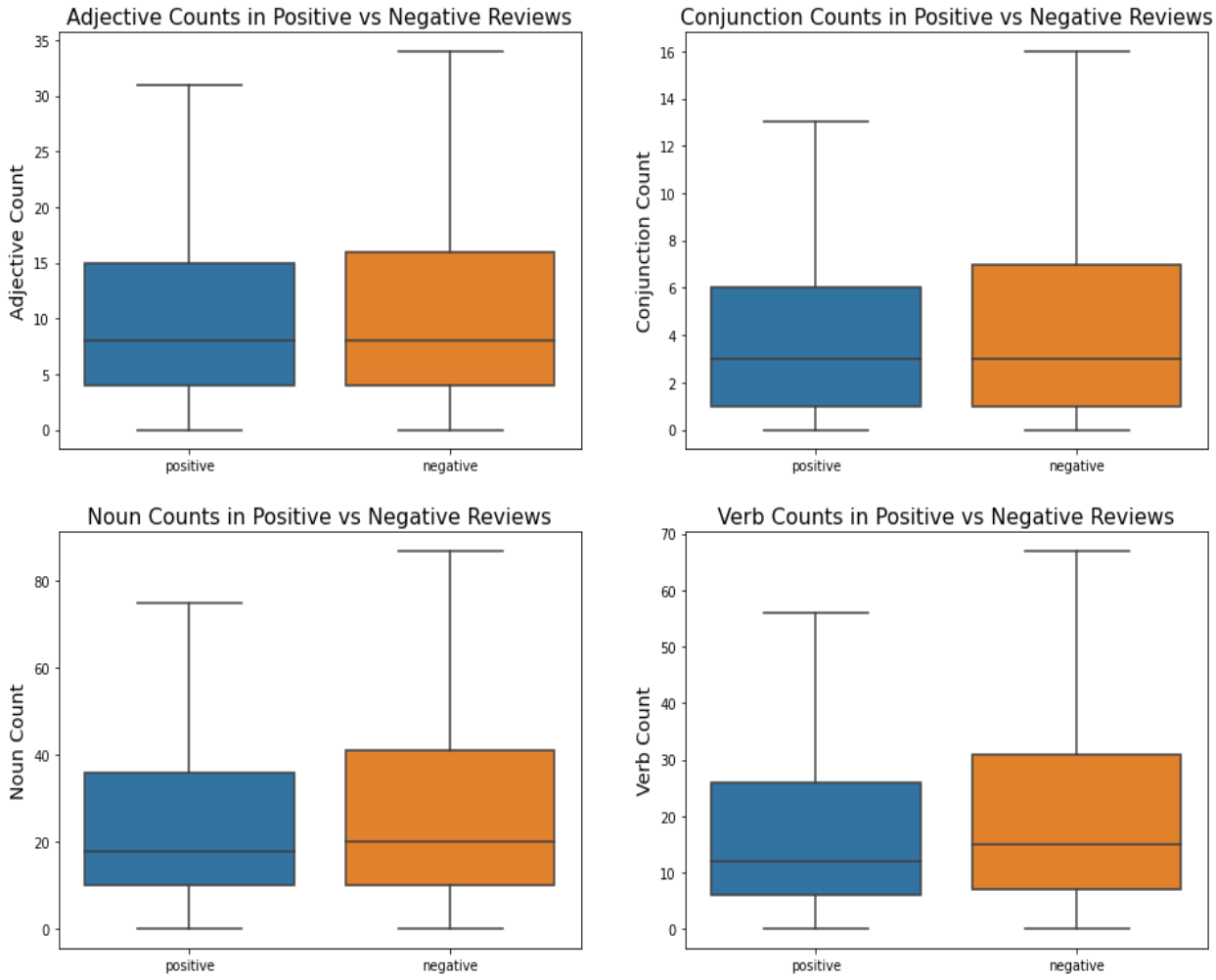

Figure 3: Variability of different parts of speech counts among positive and negative reviews

According to the Mann-Whitney U test, there is evidence to suggest that there is a statistical difference in the noun and verb counts between positive and negative reviews (p value < 0.001). However, no statistically significant differences were found among adjective and

conjunction counts in these reviews. From Figure 3, it is evident among the different parts of speech examined conjunctions were the least represented of the group, which made sense as they are not used too often in sentences.

---------- **Positive Example (Nouns)** ----------

"This game would have got more praise from the public if it was a little cheaper but quite rightly a lot of people are complaining that it's effectively a full price game without the full package, i.e a campaign. However it is an absolutely epic shooter, it allows people to jump straight in and have fun without having to unlock this weapon/peak etc. It can be as simplistic or competitive as the community makes it, my only fear is the console generation will become bored of it very quickly, I suspect the PC community will truly embrace it for what it is. Overall it runs fantastically well, looks good and most of all is fun!"

**Median noun count = 18*

---------- **Negative Example (Nouns)** ----------

"Boring and overrated. I have no idea what people see in this mediocre game. It might be fun the first match or two, but beyond that it falls flat. It's way too easy to score from center, teammates are more interested in hitting you than the other team, and you get flung so far that you never really have a chance to get the ball once you have been hit. Matchmaking is a mixed bag. I was on a team of all rookies playing against all semi-pros and masters. Avoid this game."

**Median noun count = 20*

---------- **Positive Example (Verbs)** ----------

"Outstanding. Outstanding. Outstanding. There's not much else left to say. This game offers everything its predecessor did and more. Fluid gameplay, brilliant music. And now online daily and weekly challenges, as well as adorable collectables, loads of Origins inspired unlockable levels, AND probably best of all - Rhythm based levels focused around pieces of music that show how truly seamless the level design is. A must have for your Xbox One!"

**Median word count = 12*

---------- **Negative Example (Verbs)** ----------

"Game is about you looking for your plastic.... I mean baby with whom you spend a total of 5 minutes with, Baby looks like plastic and gum. Game world big but graphics are from the 360. Not much innovation and they actually made the conversation system obsolete by not having any variations in most conversations no matter which option you choose. Very lazy. Wish I didnt get it online as I would have returned the disk. If youre a fallout fan you'll have fun. If this is your first attempt at exploring the world.. Try Fallout one first."

**Median word count = 15*

Do user scores differ among game genres?

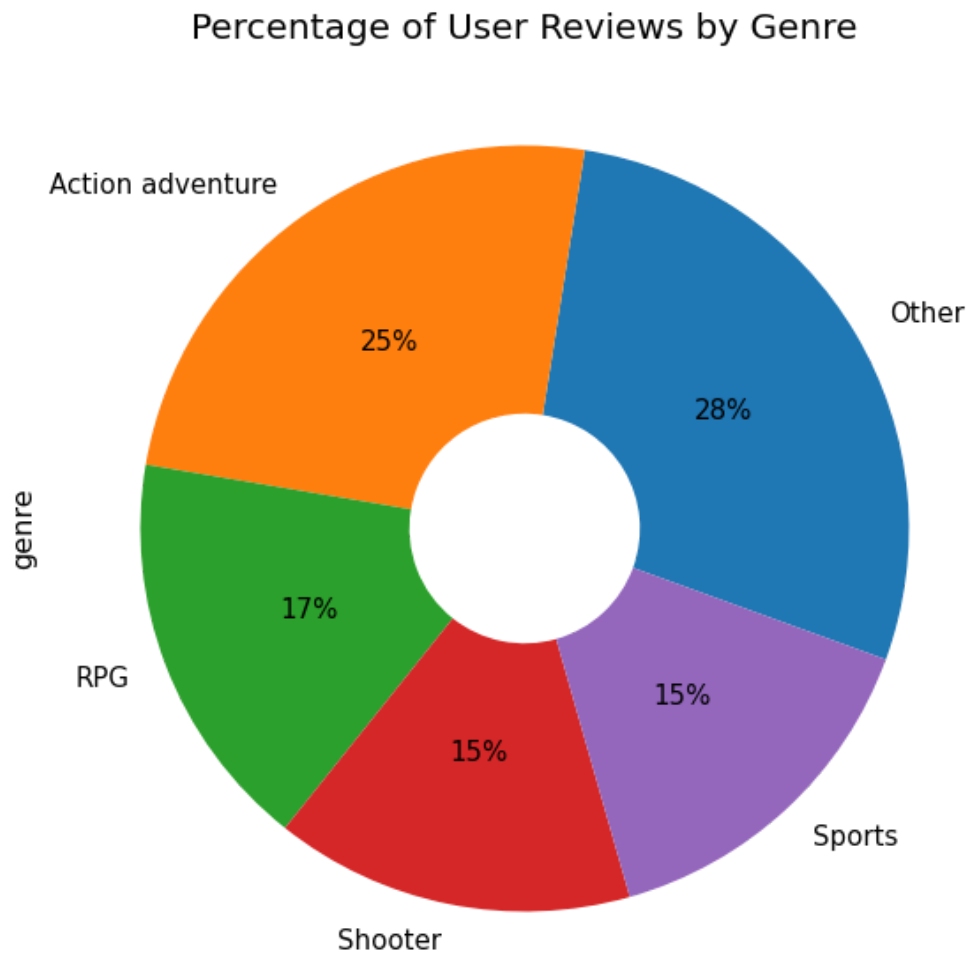

Figure 4: Percentage of user reviews by genre in the sample data

Figure 4, shows that games in other genres were the most prominent in the sample data with 28% representation. Reviews for action adventure games followed accounting for 25% of the population. RPG, shooter and sport game reviews all had about equal representation each ranging from 15% to 17%.

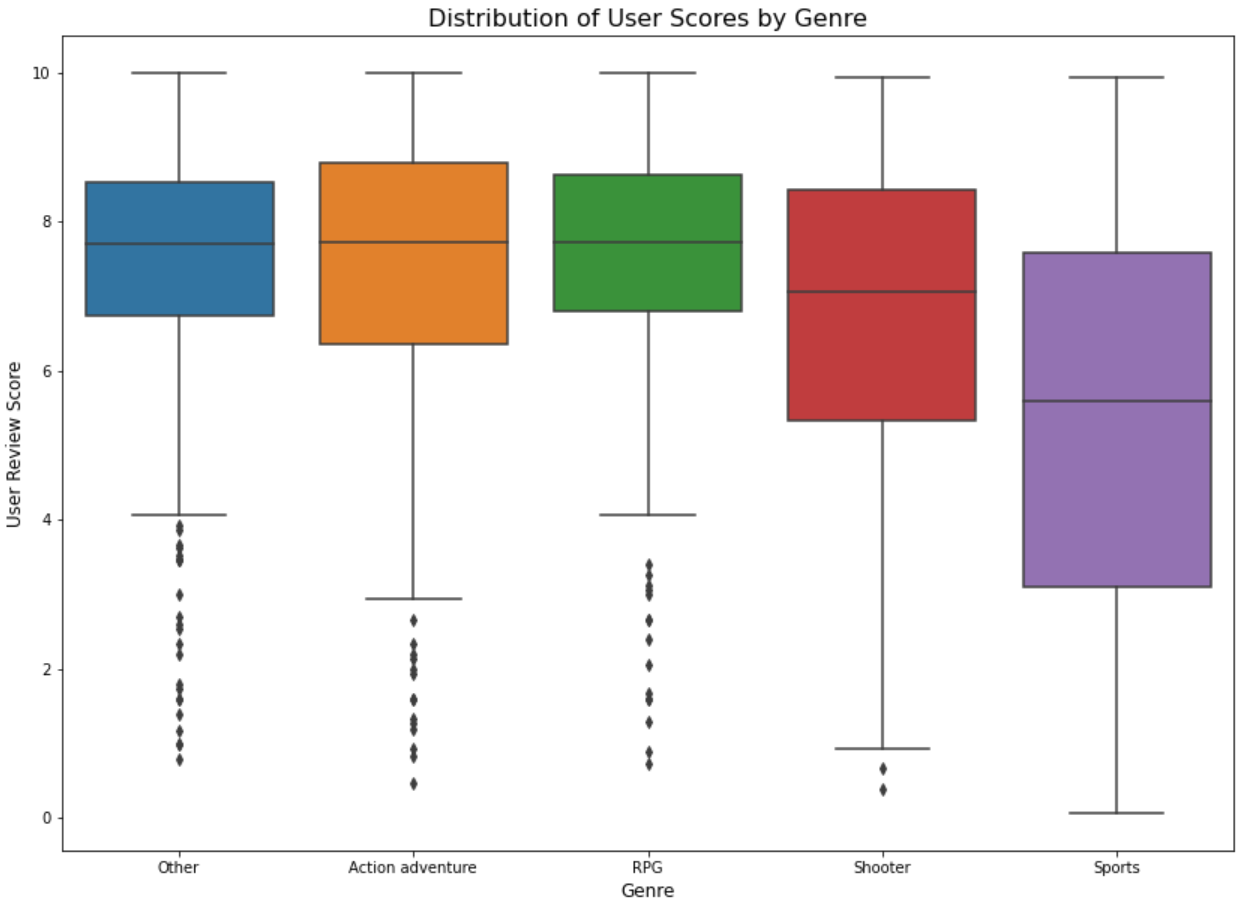

Figure 5: Variability of user scores by game genre

According to the Kruskal-Wallis H test, there is evidence of a game's genre affecting the score a user gave a game (p-value < 0.001). Figure 5, shows that games in the sports genre generally tend to receive much lower scores than any other genre. The majority of which lie in the mixed and negative review areas. The median of RPG, action adventure and games in other genres were around 7.8, corresponding to a positive review.

What are the most predictive words in game reviews by genre and sentiment?

Method:

In order to get the most predictive words of positive and negative reviews by genre, a term-document matrix was generated from the pre-processed text of a certain genre. Using the term-document matrix and the target sentiment feature, a predictive model was trained. An identity matrix the length of the vocabulary, essentially a list of documents each being one word long for each word in the vocabulary, was created to be predicted on. The trained model was used to predict on the identity matrix, generating a list of probabilities for each word. High probabilities corresponded to a positive review and vice versa.

Results:

Figure 6: Most predictive words in positive (left) and negative (right) reviews in the action adventure genre

Positive action adventure games were strongly represented by words that seemed to equate a great action adventure game to a cinematic-like experience (Fig. 6). On the contrary, negative reviews complained about features that were either broken, boring or flat out awful.

Figure 7: Most predictive words in positive (left) and negative (right) reviews in the RPG genre

Figure 7 depicts positive RPGs as characterized by the typical features RPG games contain such as quests, combat mechanics and battles. Reviewers also seemed to enjoy the voice and acting RPGs have. Along with quests, negative reviews of RPGs were plagued by

different features of RPGs such as skills, attacks and effects all telling of what characters in these games are able to do or not do. Load and save were words also common in negative RPG reviews, which is perhaps due to the fact that these games require larger amounts of memory to track a player's progress.

Figure 8: Most predictive words in positive (left) and negative (right) reviews in the shooter genre

Shooter games shared many of the same words for both positive and negative reviews. Yet, in negative shooter game reviews the word campaign was very prominent. Players were either dissatisfied with what shooter campaigns offered or if they even offered a campaign (FIg. 8). Money and content were also mentioned, gamers may not have liked the quantity they spent for the content they received.

Figure 9: Most predictive words in positive (left) and negative (right) reviews in the sports genre

Positive sport games seemed to be received by racing games. I assumed that could have been the case as car, track, course were largely represented in the word cloud. In negative sport games EA pops out immediately. Gamers may not be too fond of sports games developed by EA, shown in Figure 9, along with the virtual balls, shots and AI.

Figure 10: Most predictive words in positive (left) and negative (right) reviews in the other genres

Figure 10 reveals that games in other genres that received positive reviews were described by words that could be telling of some of those genres. For instance, puzzle and fighting are considered to be genres on their own, albeit less popular genres. Not surprisingly, Nintendo is telling of positive reviews in other genres. Nintendo develops very fun games for audiences of all ages. On the other hand, negative reviews simply consisted of negative adjectives. Concepts of money is the only negative theme found among negative reviews in other genres.

Modeling

The aim of this project was to predict the sentiment of a gamer towards a game based on their review. The target variable, user sentiment, contained three different categories - positive, mixed and negative. For this project, the only categories of interest were positive and negative as such the mixed reviews were considered to be negative. The following are the six different models that were considered:

- K-nearest Neighbors (KNN)
- Logistic Regression
- Linear Support Vector Machines (SVM)
- Multinomial Naive Bayes
- Random Forest
- Gradient Boosting

Table 1: ROC-AUC scores of the six different models

Model	ROC-AUC
KNN	0.521
Logistic Regression	0.923
Linear SVM	0.919
Multinomial Naive Bayes	0.900
Random Forest	0.899
Gradient Boosting	0.866

Word features by themselves yielded the best ‘ROC-AUC’ scores than when using density count features, dummy variables of categorical features, or a combination of these features. The top two performing models were Logistic Regression and linear SVM both using word features transformed by a count vectorizer, TF-IDF pipeline with a minimum document frequency of 2. The best parameter inverse of regularization strength value for the Logistic Regression model was 1 and 0.01 for the linear SVM model.

Model Evaluation

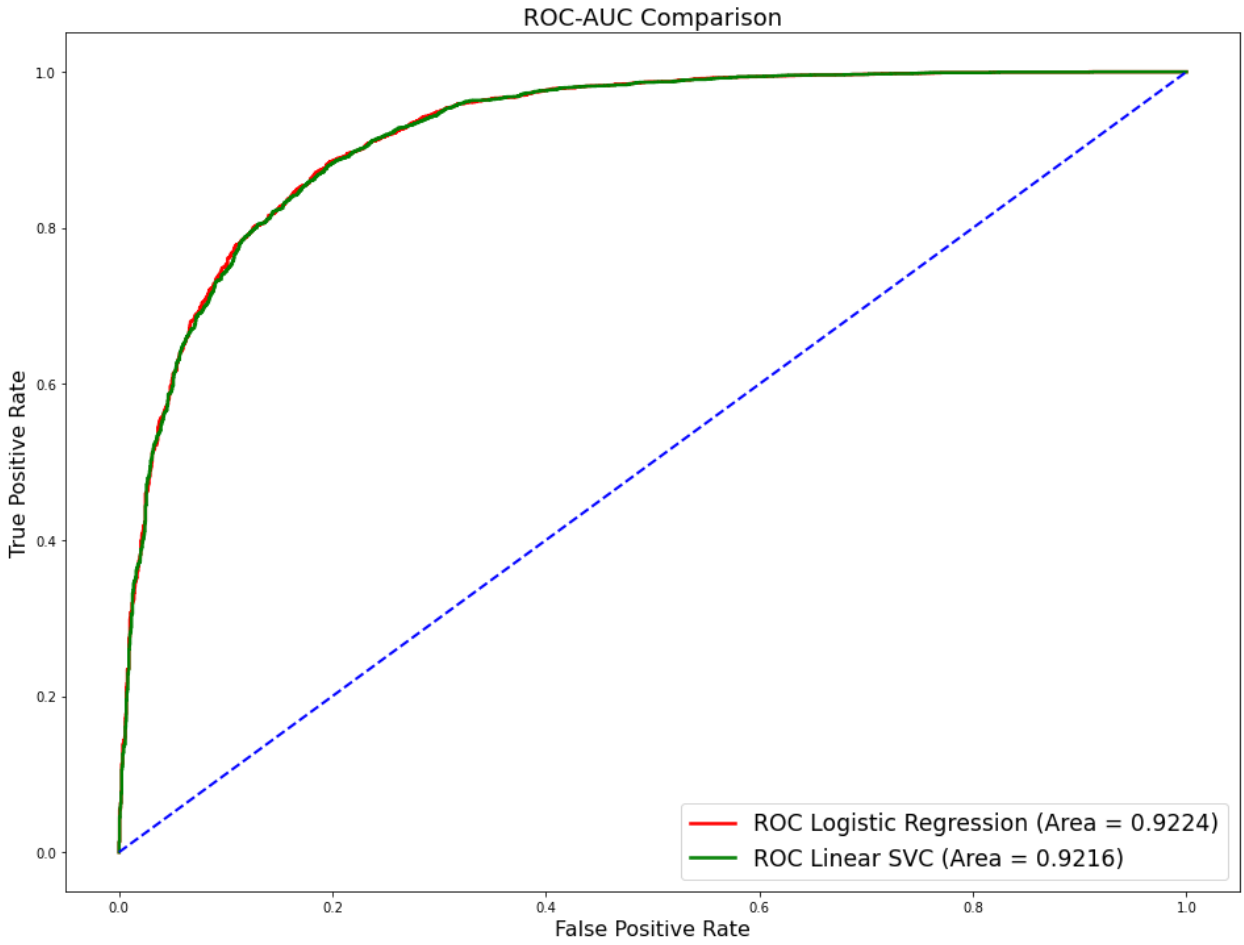

Figure 11: ROC-AUC comparisons between best performing models

From the ROC-AUC curves (Fig. 11) it is difficult to assess which model stands out from the rest. The decision was made to move forward with the Logistic Regression model as it slightly outperformed the Linear SVC.

Choosing A Metric

Business Case 1

One potential business case for usage of this model would be that a company wanted to assess a general response to a release by customers on social media. In order to accurately depict this aggregate score, it is important that positive and negative reviews are equally weighted. To accomplish this and accommodate for the unbalanced dataset, the threshold for the highest balanced accuracy was determined. This way both sensitivity and specificity would be weighted equally.

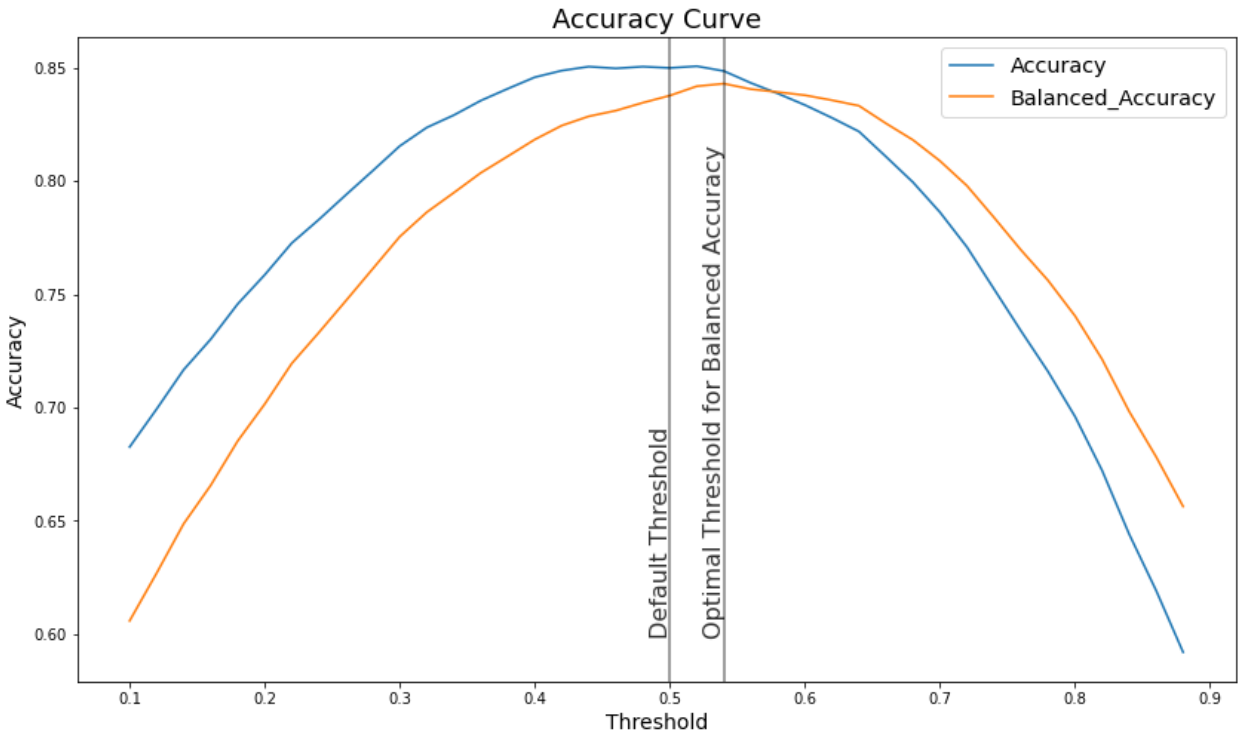

Figure 12: Accuracy curves at different thresholds

At a probability threshold value of 0.54, the model's balanced accuracy is optimized (Fig. 12). At this threshold the misclassified positive and negative reviews should be relatively the same in proportion to their sample size. Equal value is kept in having an accurate prediction of both positive and negative sentiments. With this balance, a company may assess how their game fares against other games within their genres accurately. Letting them know what about their game excites the average gamer or what features of their games can be improved.

Although there are better thresholds for overall accuracy of the model, you risk misclassifying positive/negative reviews at a greater rate than the other. A final confusion matrix was developed (Fig. 13) to see the classifications of the model with this new threshold.

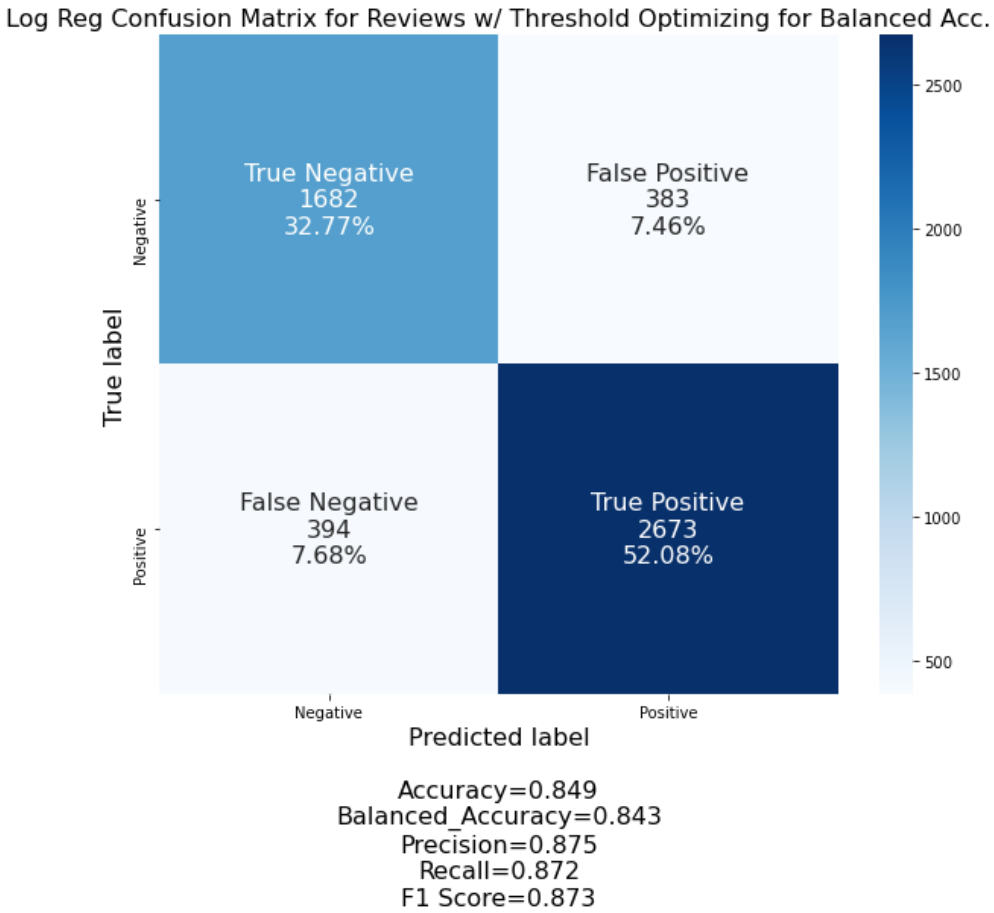

Figure 13: Confusion matrix for logistic regression model optimized for balanced accuracy

Business Case 2

If a company were interested in finding potential influencers on social media that genuinely enjoyed a game and were willing to promote or advertise the product, the model optimized for precision would be able to find those influencers. A high precision allows the model to minimize the number of false positives. Ensuring the positive predictions made by the model were truly all positive reviews, at the cost of having some positive reviews predicted as negative reviews. With a beta = 0.25, the F score, recall and precision values for the model were plotted at different threshold values.

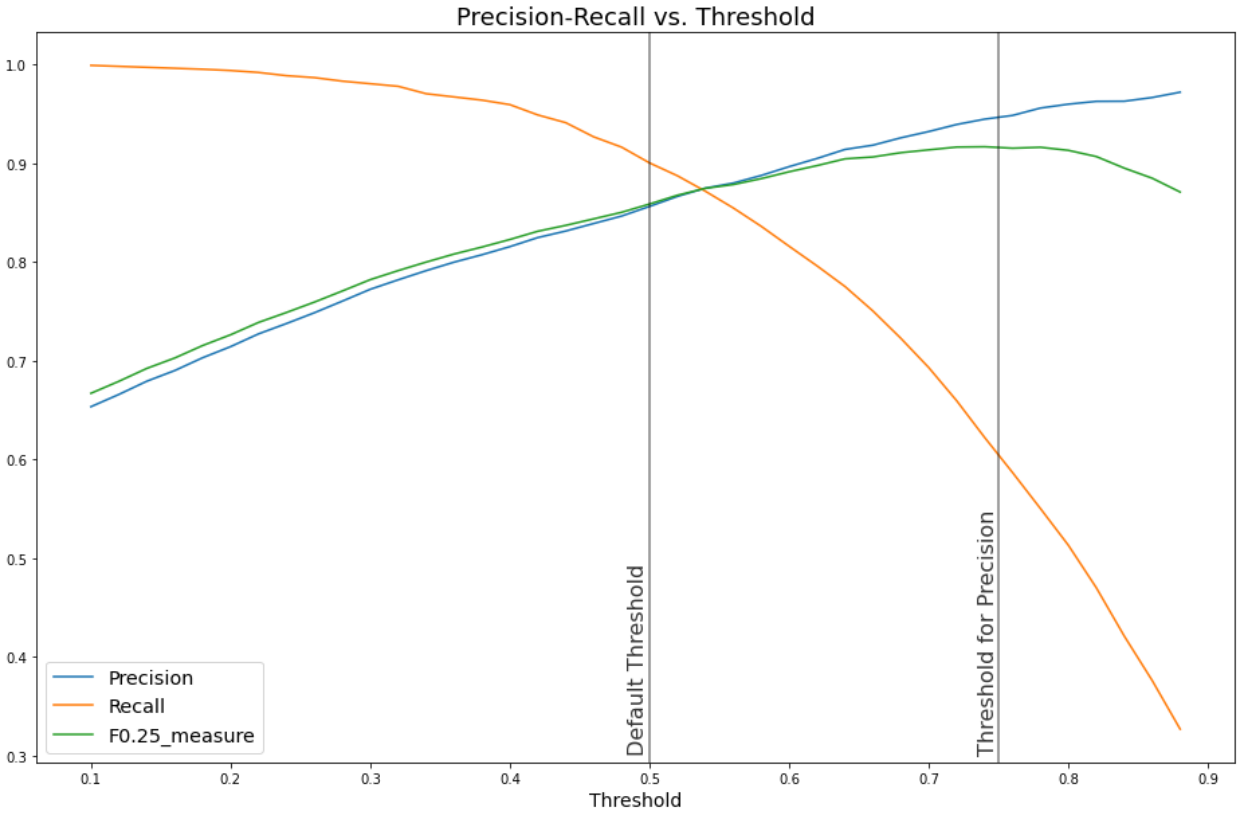

Figure 14: F-beta score, precision and recall of logistic regression model at different thresholds

At a probability threshold value of 0.75, the model's precision is optimized (Fig. 14). At this threshold the false positives are minimized, allowing the model to yield positive predictions in which all are really positive reviews. Figure 15 shows the classification report of the model optimized for precision.

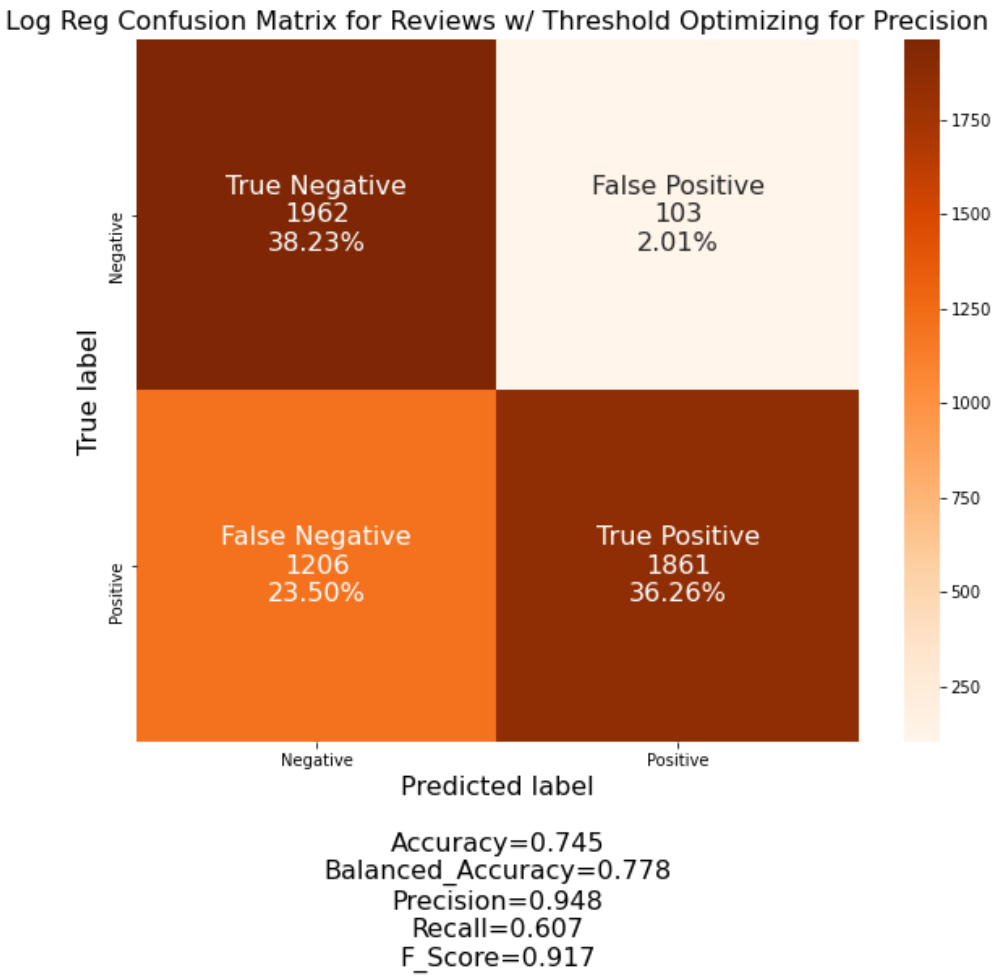

Figure 15: Confusion matrix for logistic regression model optimized for precision

Conclusion

Generating word clouds of the most predictive words in positive reviews by genre was very telling about characteristics most important to gamers. I learned that gamers that play RPG games enjoy the combat style and different quests offered in the games, but detest the screen loading and saving time. I also learned players are tired or very unsatisfied with sport games developed by EA. Perhaps, due to yearly releases of sport titles with little change in game fluidity and game graphics.

The best performing model was a Logistic Regression model, achieving an ROC-AUC score of around 92%. The model was optimized for balanced accuracy in one business case. With a threshold of 0.54 the highest balance accuracy of 84.3% was attained. Allowing game developers to accurately assess their games versus their competitors or to evaluate how consumers receive their game. With reliable feedback gained from the model, these companies can strategize and roll out plans to improve deficiencies and maximize profit.

In the second business case the F score was adjusted to favor precision with a beta = 0.25. With the adjusted F score and a threshold of 0.75 the model achieved a precision of about 95%, making the model an efficient predictor of positive reviews greatly reducing the number of false positives. A high precision model would allow companies to find potential influencers on social media who love their game and would be willing to promote or advertise their product.

In conclusion, a Logistic Regression model was used to correctly predict as many positive and negative reviews through the use of methods like text pre-processing, cross-validation grid searching, thresholding and word vectorizers. However, there is still a lot of room for improvement and some next steps to consider are listed below:

- collecting more reviews, only about +20,000 reviews were used in this analysis all retrieved from the same site
- performing topic modeling of reviews
- incorporating more in depth text processing techniques

.

References

Movie Reviews, TV Reviews, Game Reviews, and Music Reviews. (n.d.). Retrieved November 21, 2020, from https://www.metacritic.com/

Sarkar, D. (2018, December 04). A Practitioner's Guide to Natural Language Processing (Part I) - Processing & Understanding Text. Retrieved November 21, 2020, from https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72

Says:, S., Says:, V., Says:, S., Says:, M., Says:, D., Says:, B., . . . Says:, U. (2017, September 13). Basic evaluation measures from the confusion matrix. Retrieved November 21, 2020, from https://classeval.wordpress.com/introduction/basic-evaluation-measures/