PREDICTING GAMEPLAY CHOICES ON STEAM

Aimee Wong
University of California, San Diego
aiwong@ucsd.edu

Christina Mak
University of California, San Diego
cymak@ucsd.edu

ABSTRACT

Oftentimes consumers purchase items, but do not use them. In regards to video games and *Steam*, user often purchase many games during sales and in packages but often do not play all games immediately or ever. This may be due to lack of time, impulse purchasing, package deal, or other factors. In our work, we are looking to determine if a user would actually play a given game beyond purchasing said game. Our goal is to learn what prompts users to play particular games and why users may purchase games but not end up playing them.

1 DATASET AND EXPLORATORY ANALYSIS

Our data is from *Steam*, the digital distribution platform for games, video streaming, and multimedia. We used a dataset collected by Apurva Pathak, Kshitiz Gupta, and Julian McAuley. This dataset has a focus on the Australian Steam community "GameAus" containing 88,310 unique users and 10,978 unique games. The dataset specifies users and the games said user has purchased. The dataset also specifies games and their features, such as genre and price.

1.1 Statistics and Properties

Looking through the data, we found basic statistical data regarding users and their purchases. This information can be seen in the table below.

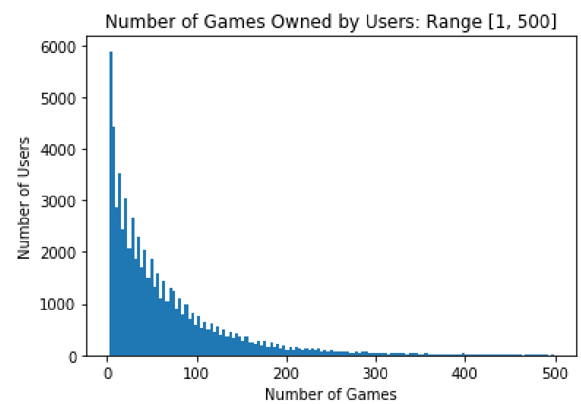
During the exploratory analysis of the dataset, there were several interesting trends and findings. We found that the mean number of games a user owned was 58.35 and the median number owned was 36. The number of games owned was skewed to the right with some users having a very large number of games, the max being 7,762. However, we also found that there were 16, 806 users with no games on the account.

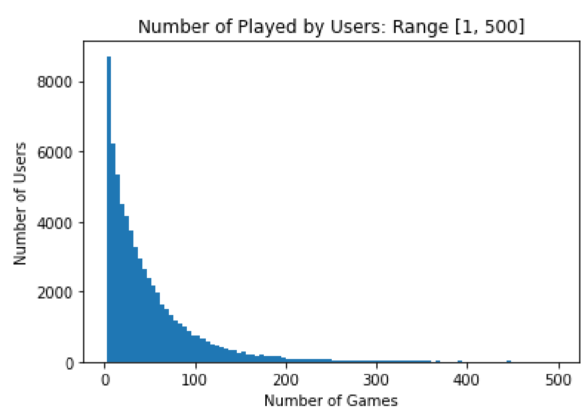
Table 1: *User and Item Statistics*

Feature	Count
Users	88,310
Total Games	10,258
Total Free Games	227
Median # of Games Owned per User	26
Mean # of Games Owned per User	58.35
Average Ratio of Games Played to Games Owned	0.547

The trend observed in the number of games owned and number of games by users were similar with the number of played games being predictably lower.

Figure 1 and 2: *# of Games Owned & Played by Users*





1.2 Notable Trends and Aspects

After loading the dataset, we looked into the trends of the user's purchased and played games. We found that the average ratio of games played to games owned was 0.55. In the figure below, we charted ratio of games played to games owned by users, excluding users that owned zero games. From the chart, it can be seen that the majority of users that have games, play at least half of their owned games. Out of the 88,310 users in the dataset, 404 users played 100% of their games, given that they owned at least 1 game. On average, these users owned 4.40 games with 57 being the highest number of games one of those users owned. Since the set of users who played all of their games makes up only 0.46% of all users, this means that most users own games that they have not played.

Figure 3: *Ratio of Games Play to Owned Histogram*

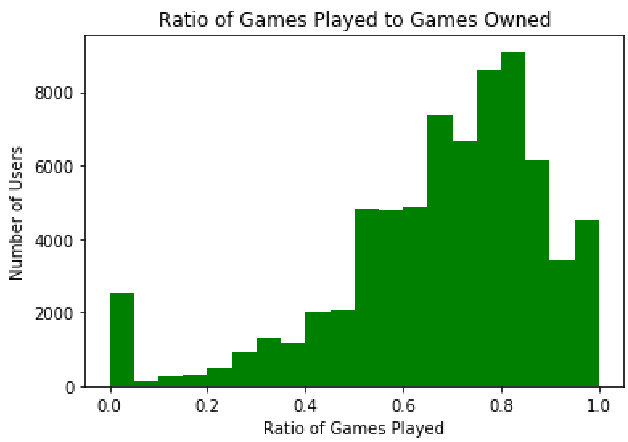
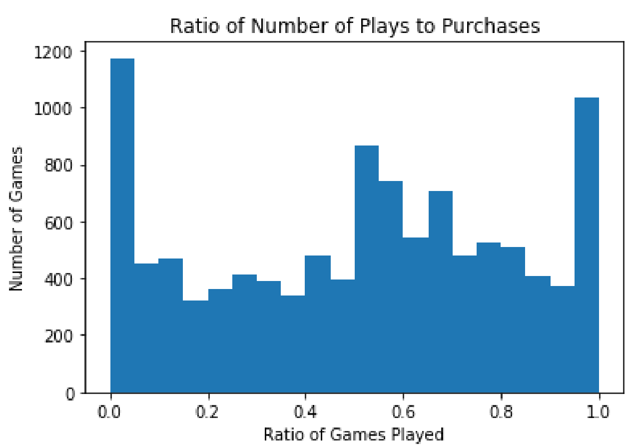
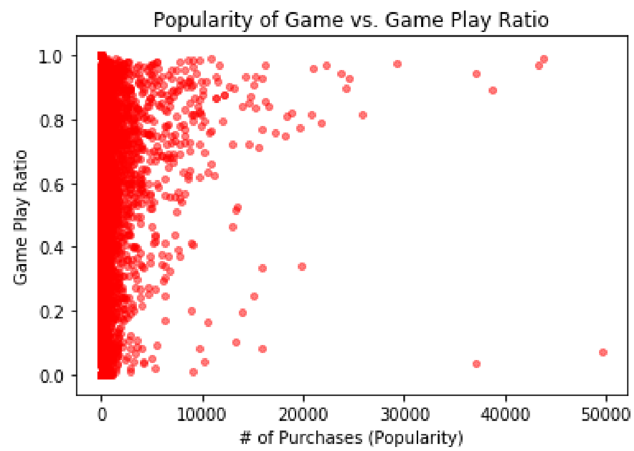


Figure 4: *Ratio of Game Plays to Purchases Histogram*



We then looked at the ratio of number of times a game is played versus the number of time the game is purchased. It is interesting to see that there are many games with close to no plays and games that are always played when purchased. Otherwise, play frequency of games seem to vary greatly.

Figure 5: *Popularity of a Game vs. Game Play*



Finally, we mapped the relationship between the popularity of a game--represented by the number of purchases—and the play rate of said purchases—the number of plays divided by the number of purchases. There is no trend in games with less than 10,000 purchases. Beyond 10,000 purchases, an increase in number of purchases seems to be positively correlated with a greater play rate. However, there are several outliers as well.

1.3 Dataset Discrepancies

For the dataset, there are several assumptions that we made due to some data discrepancies and aspects. In our dataset of games, we have 10, 258 games; however, amongst users, there are 10, 978 unique games. This means that we have 720 games without features. These games will not be excluded when the data is used. Additionally, the most common number of games owned by users was zero, with 16,806 users. It is possible that these accounts are private or dummy/spam accounts, so we would like to note so. However, these accounts will also be included in the data used as well.

2 PREDICTIVE TASK AND MODEL EVALUATION

2.1 Task: Gameplay Prediction

Our task is to predict if a given user will play a given game. While analysing the data, we noticed that most users have games that they have purchased, but have no gameplay hours. Of all the users in the dataset, only 404 users have played all of the sames that they own, assuming that the user owns at least 1 game. Therefore, we have decided to create a predictor to determine whether a user would play a game, having at least 1 hour of gameplay in a given game.

2.2 Features

There are several features in the dataset that we considered using for the predictive task. These were chosen based off our exploratory analysis. We decided to focus on game genres, games reviews, and if a given game is played often when owned.

The first feature we are using is genre. Each game can have one or more genre tags. Examples of genre tags include: Action, Mystery Strategy, Indie, and Casual. We chose to use genre as a feature as a user will likely play a game with the same or some of the same genre tags, the intuition being that users will want to play games similar to games that they have already played.

The second feature being used is the review rating. Each game has reviews as well. Review tags include: Very Positive, Mostly Positive, Positive, Mixed, Negative, Mostly Negative, and Very Negative. The review of a game can be used as a feature a games with more positive reviews will more likely be played. The opposite as well, games with more negative reviews are less likely to be played.

The third feature being used is the play rate of a game. Even if a game has a lot of sales, a game might not be played very often for any given reason. The play rate is a good indicator to determine how likely the average user would play a game, as compared to just purchasing said game.

We decided to not include all of these features when designing our model as we did not think that they were as significant.

2.3 Data Processing

In terms of data processing, there were a few things that needed to be done to generate our features. While processing the data initially, we ran into an issue with the JSON file containing the items/games. 720 items were missing from the file. We parsed the item data and the user data so that each item and each user was represented as a dictionary of its features.

We used 50,000 pairs of users and played games for the training set, 10,000 pairs for the validation set, and 10,000 pairs for the test set.

2.4 Prediction Significance and Model Evaluation

2.4.1 Baselines

Some simple baselines that can be used in compare our predictor model to:

1. If a given game has negative or mixed reviews, the user will not play the game.
2. Of the people who have purchased the game, if there are more people who have played the game than not, then the user will play the game. Otherwise, the user will not play the game.
3. If a user has played a game of the same genre, then the user will play the game.

2.4.2 Testing and Validation

In order to test and validate our model, we have developed a validation and test set of user and game pairs. Our goal is to test our model on pairs in which the user has purchased and played the game and pairs in which the user has not purchased or played the game. We will use the validation set to tune our model. We are looking at overall accuracy to determine the success of our model, the number of correct predictions over the number of predictions made.

3 MODEL PROPOSITION

3.1 Models Used

We chose to use an SVM with the following features: whether the user has played a game of that genre before, the max Jaccard similarity between the the game and the other games the user has played, and the average Jaccard similarity between the game and the other games the user has played. We chose this model because support vector machines are effective for classification tasks and it would be a good place to start our model.

To optimize the hyperparameter, we ran our model on the validation set with different hyperparameters to choose the one that performed the best on the validation set.

We did not run issues due to scalability. There were only minor differences in the accuracy as we increased the size of the data we used. Since the accuracy of the predictor on the training set was 0.9193 and the accuracy on the test set was 0.9102, overfitting was not an issue.

3.2 Baselines and Other Models

We started with a simple predictor that only looked at the review information. If the game is known to not have any negative reviews, then it predicted that the user will play it.

We also tried a model that predicted that a user would play a game if the user has played a game that had a genre in common with the game in question.

Another baseline model predicted that the user would purchase the game if the game was more often played than not, i.e. its play/purchase ratio is greater than 0.5.

We attempted to use different features for our SVM model. We tried adding a feature for whether the reviews contained negative tags. This feature was unsuccessful in increasing the accuracy while the genre features did well.

The advantages of our first three models were that they were easy to train and the second and third ones performed fairly well. The disadvantages of these models is that they were not as accurate as the model we chose. The weakness of the model we chose is that it is more expensive to train. Adding more features and increasing the size of the data set made the model take longer to train.

4 RELEVANT LITERATURE

4.1 Data Set Characteristics

Our dataset comes from a paper published by Apurva Pathak, Kshitiz Gupta, and Julian McAuley on *Steam* and Bundle Recommendations. Their paper was centered around generating personalised bundles of games for users and how "good" a given bundle is based of game similarity, bundle size, and other features. The dataset does focus on on the Australian Steam community "GameAus", so the dataset is a subset of the greater *Steam* community. It is possible that the dataset involves users that are more involved in gaming as they are a part of a particular *Steam* community. Their particular behaviour could be slightly different than that of a larger, more diverse dataset derived from international *Steam* accounts. For our purposes, we have omitted the use of the review text and bundle data as we decided to not use them as features. We decided to do so as our focus in on the individual game. While the bundle data could be a useful feature to characterise the similarity between games, we decided to focus on genre similarity features.

4.2 Similar Datasets, Works, and Conclusions

There are several papers that perform analysis on *Steam* and gamer behaviour. One paper entitled "Condensing Steam: Distilling the Diversity of Gamer Behaviour" published by Mark O'Neill, Elham Vaziripour, Justin Wu, and Daniel Zappala of Brigham Young University focuses on analysing varying gamer behaviour and the social connections that form around them. Their dataset was much more comprehensive with over 100 million user accounts and 380 million games. In their work, they noticed that most users exhibit "mild" behavior while there are few, but "very extreme" outliers. This is found across multiple traits: gameplay time, game collection size, game achievements, and more. Additionally, they determined that users with similar behaviour tend to befriend and interact with one another more.

Another paper we looked into is entitled "Methods for classifying usability qualities and problems for action games from user reviews using text mining", published to IEEE by Artinat Wattanaburanon and Nakorthip Propoon. The focus of the paper determining whether a review is positive or negative and extracting relevant information that can be used to improve the game. We looked into this paper because of its use of the *Steam* dataset and game-specific features and characteristics. The focus of the paper is on sentiment classification and determining if portions of a review are regarding the game usability qualities or game problems that could be addressed.

The final paper we looked in is entitled "Machine Learning for Predicting Success of Video Games". This paper is a Master's Thesis by Michal Trneny of Masaryk University. The dataset set was generated from *Steam* and gathered information regarding games, prices, and popularity charts. The goal of the paper was to study the factors that affected a video game's success and estimate how successful a game would be based on its features. They concluded that there is a high correlation between core features known prior to release and the average number of players within the first two months of release. Many models and methods were used to predict the average number of concurrent players, using both regression and classification. Random Forest and SVM yielded the best results for this task. Gaussian Process and a linear model performed slightly worse and RPART scored significantly worse.

4.3 Methods and Model Choice

In Pathak, Gupta, and McAuley's paper on bundling, the problem was formulated and used Bayesian Personalized Ranking (BPR) to rank bundles. The greedy algorithm and BPR was used to generate bundles based off item features and bundle size. The use of bundle size and item features affected the outcome of the bundles. With the use of the same dataset, we were able to the item features and genres to determine item similarity.

In Trneny's paper on predicting video game success, he used and compared a number of methods. He used several different regression methods such as Random Forest, Gaussian Process, and SVM. In the particular paper, SVM was able to reach 0.82 correlation in predicting the average number of players for a given game. There was a focus on the genre of the games as well a game core features. In our work, we also focus on genre and used review data to determine if a particular game is popular.

Since most of the papers regarding Steam datasets were not looking at similar predictive tasks, we also looked at papers solving similar tasks. One paper entitled "A case study in a recommender system based on purchase data", published in 2011, describes several implementations of collaborative filtering using a user's purchase history. There was a focus on domain knowledge in the paper, emphasising the use of purchase history rather than ratings. This is related to our task as we looked into user behaviour when playing games. One point in which our task differs is trying to differentiate between whether a user will purchase a game or purchase and play a game.

In terms of methods, our model choice was influenced by work that we did in class. We used the methods from the visit prediction task. We looked into using collaborative filtering for our task as well, but decided to go with a simpler model initially.

4.4 Conclusion Analysis

The conclusion that we drew from our analysis of the *Steam* dataset is in line with the conclusions of the other papers. However, since each of the papers focused on different aspects of *Steam*-- user behaviour, reviews, and game success respectively-- the conclusions drawn were on different matters. We did come to the conclusion that the majority of users have "mild" behaviour with several extreme outliers. The outliers have extreme behaviour and deviate greatly from the norm.

5 RESULTS

5.1 Methods and Model Evaluation

Table 2: *Model Results*

Data set	Accuracy
Training	0.9193
Validation	0.9172
Testing	0.9102

Our model performed fairly accurately compared to the other models we attempted to use. The strongest features used similarity in genres, indicating that the games users play tend to be similar in genre.

Our simple reviews model, which looked at whether the item had a negative review tag, had an accuracy of 0.5695.

Our model using the play/purchase ratio performed with an accuracy of 0.7177.

Our initial SVM model had only one feature that looked at whether the user has played a game with a genre in common with the given item. Its accuracy was 0.8554. Since adding Jaccard similarity features increased the accuracy, it means that not only do users tend to play genres they've already played, but also play games with similar sets of genres.

5.2 Model Parameters

We used the validation set to find the optimal value of the hyperparameter. We tested the hyperparameters of 0.01, 0.1, 1.0, 100.0, and 1000.0. For all values of the hyperparameter, the accuracy was the same.

5.3 Significant Features

The features that worked well were whether the user has played a game with a genre in common with the given item, the max Jaccard similarity between the item and the user's played items, and the average Jaccard similarity between the item and the user's played items.

When choosing features for our SVM, the genre features did well but the reviews feature made no difference in the accuracy of the predictor.

5.4 Task Conclusion

One thing we set out to do was determining whether or not a user would play a game.

Based off our predictors and models we came to several conclusions. Users tend to play games with genres they already played. Since games can have many genres, users also tend to choose games that have sets of genres similar to the sets of genres of games they already played.

6 CONCLUSION AND FUTURE WORKS

After working with the *Steam* dataset to determine glay play for users, we have learned a number of things regarding user behaviour and trends. The biggest determinant of game play was the type of games that the user played, therefore genre was a major feature. The play rate of games was an indicator of determining if a game would be played but had a lesser role. We also learned that play rate varies greatly among games, but a higher popularity tends to mean a higher play rate. Overall, an individual user's likelihood of playing a game can mostly be determined by user's gaming history.

Further steps would be in creating a classifier to differentiate between unplayed and played purchases. In our work, we essentially predicted the likelihood of a user playing a game based off the subset of played games information. So, we focused solely looking at the played games. There would be additional value add in a model that is able to determine if a user will purchase a game and not play it as that information would be valuable to the game creators and *Steam*.

7 REFERENCES

[1] A. Pathak, K. Gupta, J. McAuley, *Generating and personalizing bundle recommendations on Steam.* SIGIR. 2017

[2] M. O'Neil, E. Vazirirpour, J. Wu, D. Zappala. *Condensing Steam: Distilling the Diversity of Gamer Behavior*. 2016

[3] A. Wattanaburanon, N. Propoon. *Method for classifying usability qualities and problems for action games from user reviews using text mining.* IEEE. 2016.

[4] M. Trneny. *Machine Learning for Predicting Success of Video Games.* Masaryk University.2017

[5] B. Pradel, S. Sean, J. Delporte, S. Guerif, C. Rouveirol, N. Usunier, F. Fogelman-Soulie, F.

Dufau-Joel. *A case study in a recommender system based on purchase data.* KDD'11. 2011