Project Proposal

Dylan Tucker, dst833, 11235055

ABSTRACT

In this paper, I discuss a dataset of steam reviews and genre tags for video games. I will construct a model for determining the probability of a game being recommended by users based on its genre tags. The model will be devised using linear regression.

CCS Concepts

• Information • Data Analytics •Statistics • Linear Regression

1. INTRODUCTION

The dataset I will be using was sourced from the website Kaggle under a CCO: public domain license. This dataset was originally used to predict user recommendations based on user reviews. The dataset contains a list of reviews for video games in the format of title, year of release, user review and a user recommendation. A second dataset is also provided which contains an overview of each game featured in the first dataset. The second dataset is formatted so that each row contains a title, the name of the developer, the name of the publisher, genre tags, and an overview/description. A linear regression will be performed to predict a game's probability of being recommended by users based on its genre and content tags.

2. Motivation and challenges

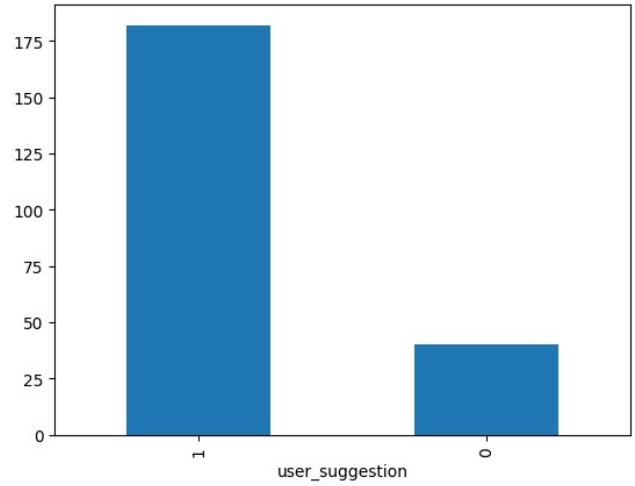

Figure 1. The total number of user suggestions in the dataset, where one is a thumbs up recommendation by a user and a zero is thumbs down.

Figure 1 shows the range of user suggestions for each game. The motivation for this analysis is to determine if there are user preferences for specific genres of video games. For example, do first-person shooter games review more favorably than horror games? Can a model be trained to predict how a game will be received based on its genre tags? These are some of the questions I hope to answer in this analysis.

The challenge will be to match each game with its relevant genre tags and user recommendations; both data points are separated into two datasets. The genre tags for each game are formatted as a string and will need to be spliced into an array before its data can be matched with each user review. After the game reviews are matched with their genre tags, I can split the dataset into training and test sub-datasets. The training data will be roughly 90% of the total dataset, while the testing data will be the remaining 10%. The model can then be built that utilizes genre and user recommendations to predict how favorably a game will be reviewed. This could help game developers predict which type of games are reviewed favorably by users. If a game is reviewed well it could translate into higher sales. Understanding how genre tags affect game reviews could allow developers to make informed decisions about players' interest in certain genres.

3. Input/Output and Linear Regression

The input for the model will be the training dataset, which will consist of the average of the user recommendations for a game and its genre/content tags. User recommendations are a binary value of one for recommend, or zero for do not recommend. The average of a user recommendation will be a value between zero and one, with a value closer to one being recommended by more users than a value closer to zero. The output will be a series of predictions using test data that will attempt to predict the average of user recommendations from the test data. The test data is a small portion taken from the overall dataset. This task will be performed using linear regression as the genre will be the only variable with impact on user suggestion that will be considered by the model. The model's prediction score will represent how accurate the model was at predicting the average user suggestion. An R2 score will also be provided to identify how well the regression line approximates the data. Various graphs and data visualizations will also be constructed to represent the model and its data.

4 Data Sample

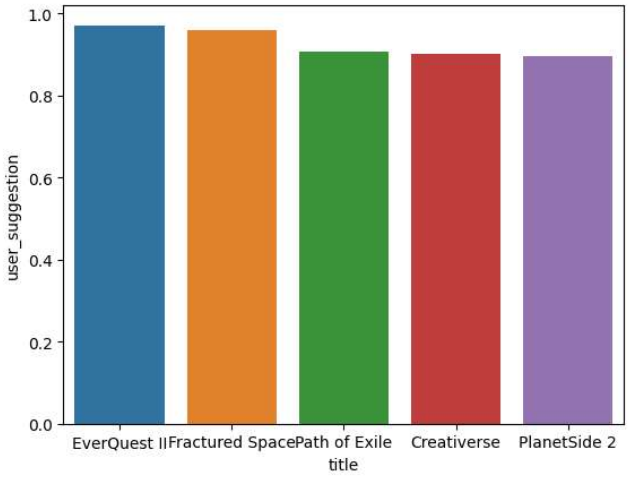

Figure 2. The top 5 highest recommended games in the dataset.

Figure 2 shows a range of different games paired against the average of their user suggestion. Planet side 2 is a massively multiplayer online (MMO) first-person shooter, while EverQuest II is an MMO role playing game (RPG). The aim for this analysis will be to determine if there is a trend amongst the user suggestions for first-person shooters or RPG's that can be used to predict the average user score for a game within the same genres.

5 Proposed Results

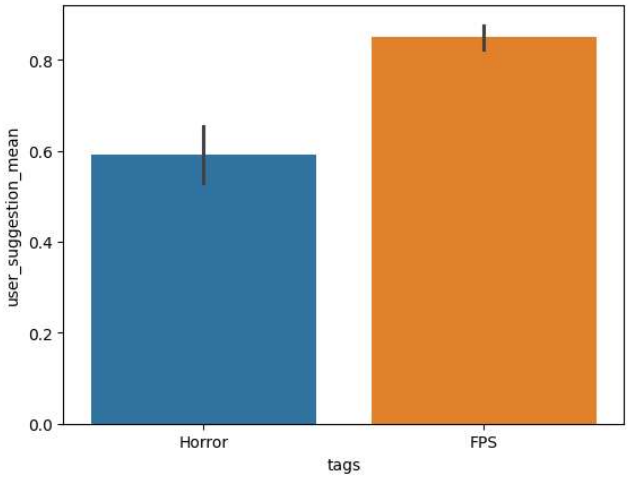

Figure 3 was made using fabricated or synthetic data that represents a sample of the proposed results for this analysis.

6. Conclusion

Understanding how genre can affect player reception of a video game could help developers decide which genre of games are likely to be reviewed more favorably. Building a model that could make accurate predictions on how it will be received by players could help achieve this goal. If the model is not accurate, it could show that genre tags have little or no affect on how a game is received by players.