CIS545 Project Proposal

Meet Gandhi, Shikha Reddy Mothe, Aditya Jayant Ganapathiraju

March 21, 2023

1 Introduction

Steam is a popular online PC game store. There are thousands of games sold on steam that engage thousands of users. Some games instantly reach a large community of players, while others remain unpopular. It is difficult to fully understand the dynamics of a game's popularity after its release and how players would react to it, which substantially impacts the net profit and popularity of the platform. Such information is affected by many factors and usually needs some predictive modeling.

2 Data Source

We are using data collected from the Steam gaming platform by Brigham Young University for the 2016 ACM Internet Measurement Conference. It is a massive data set containing both user and game-level information. More details on the available tables and their schema are available here.

3 Project Plan

In this project, we aim to answer the following questions, which could help newly launched games:

1. Are there any user and game segments? User segmentation refers to dividing users into groups based on similar characteristics or behavior, and game segmentation involves grouping games based on their attributes or popularity. Understanding these segments can help develop marketing strategies, competitive analysis, etc. While we do not plan to do any marketing or competitive analysis, we want to understand the segments existing in the data. Further, these segments might help the machine learning models we develop next. For these segments, we are considering one of the clustering algorithms like K-Means.

2. How likely is a player to buy a game? Whenever a developer releases a new game on steam, we know its attributes, such as genre, developer, publisher,

price, etc. We plan to combine these with user-level attributes like their demographics, which group/groups they are a part of, their friends on the platform, etc., to predict their likelihood of buying the game. For this, we plan to use a supervised classification model like Logistic Regression, Decision Trees, etc.

3. What will be the game's Metacritic rating? Given game-level attributes like genre, publisher, developer, price, type, etc., we want to see if we can effectively predict its Metacritic rating. Here, we will be using a regression model, starting with Linear Regression and moving to other complex models as time permits.

4 Value Proposition

Having answers to the questions defined in the project plan can help game developers and publishers better understand user behavior and preferences. Knowing the likelihood of a player buying a game and its potential Metacritic rating can help them tailor their games and marketing strategies for their user base. Further, it would also allow them to optimize game development efforts to create more successful games with improved user engagement and satisfaction and avoid losses due to the game becoming inert on the platform.

5 Anticipated Obstacles & Challenges

A few potential challenges could be:

- The steam dataset is massive, with 11 tables comprising varied information about users and games on the platform. Combining these tables to infer logical relationships could be a bit tricky.
- Identifying potential predictors for the models described above and engineering them would require us to establish and validate various hypotheses.

6 Group Members & Duties

- **Aditya:** Exploratory data analysis of the dataset to understand the relationships between different schemas and deriving useful information.
- **Meet:** User segmentation, the model for predicting the likelihood of a user buying a new game.
- **Shikha:** Game level segmentation, the Metacritic model.

7 Requested TA: Vatsal Jain