Milestone 2: Data Selection and Proposal

Gabriel Avinaz, Joshua Greenert, Mithil Patel

Bellevue University

DSC630-T301 Predictive Analytics (2233-1)

Prof. Andrew Hua

December 11, 2022

Introduction

Since antiquity, games have been an integral aspect of human society – especially for cultural development and social interaction. With the advancement of human civilization, the way games are being played has drastically changed over the eons as modern games are primarily played electronically (video games). With over 3 billion players worldwide and approximately 200 billion dollars in revenue, the video gaming industry is constantly looking to attract new customers and boost playtime. As a result, we will implement a collaborative recommendation system to assist users in finding similar games that may interest them. By utilizing a collaborative filtering system, we can use historical review information to determine whether a product might interest one of our customers. We will experiment with two approaches to determine which application would be better for an e-commerce implementation.

What types of model(s) do you plan to use and why, and how do you intend to evaluate your results?

Our first approach will implement memory-based collaborative filtering, which utilizes the collective review information of other users to find similar games. We'll be implementing both item-item and user-item methods to determine which provides a better result for our users. These two methods offer predictions based on what similar users like and what users who like a particular game are likely to enjoy. We'll use several distance measurement algorithms to determine which provides a more accurate recommendation such as: cosine similarity, Pearson Correlation, and K-nearest neighbors. The models created using these methods should perform fastest while providing acceptable accuracy. For our model-based collaborative filtering, we will train a few machine learning (ML) algorithms to make recommendations and find similar games. Our project will focus on utilizing matrix factorization algorithms to make this determination but will be looking to implement a deep learning approach given the time. The process of applying different ML algorithms is fairly simple, so we'll be looking to implement the most we can: SVD, PMF, and NMF, then compare our results with the memory-based results.

Our second approach to creating a recommendation system for our games library is by implementing a content-based filtering system utilizing natural language processing and available user reviews. This process will use a vectorized matrix of user reviews to find games that have been reviewed in a similar way. We can implement many of the techniques from our memory-based collaborative filtering process into this one and measure performance across each of our distance-measuring algorithms. We can utilize a standard train-test methodology from many of the of the models we will be implementing while cross-referencing our suggested games with user-rated games. We also have user information in our data set as to whether they would make a recommendation for the game they are reviewing. We can utilize this data in testing our results accuracy.

What do you hope to learn?

We hope to learn consumers' interests and buying patterns which can help our company optimize revenue by focusing resources on developing games customers will more likely enjoy. Additionally, we intend to learn appropriate methods to develop recommendation systems for alternative enterprises and future outsourcing. Finally, while going through this process, we expect to learn data from our models that can be used for other points of interest for the company and our personal career development.

Are there any risks or ethical implications with your proposal?

Fortunately, all the personal information of each individual review has been stripped from the data provided from the Steam API. Instead of the actual user's profile information, a review_id value is used instead. However, the comments may contain personal or sensitive information which will be appropriately handled or removed to ensure the safety and security of all users. Besides potential user-provided information, there are no additional risks or ethical concerns within this project.

What is your contingency plan?

In the event that our findings aren't conclusive, our dataset proves too massive to work with, or our data doesn't express a proper recommendation system as intended, we have collected another dataset to use instead. This database revolves around used cars posted to Craigslist and would have a similar strategy to our current proposal. We would use the dataset to review models for recommendations for users based upon their interests and preferences. From that information, we would predict cars that users would be interested in while also being local to their respective area based on the information from the dataset. All parties have agreed that if our project doesn't appear feasible by the end of week 3, we will be shifting gears to this alternative to salvage the time we have remaining.

Additional important information

As the project progresses, we plan to implement a deep learning algorithm if we are able to add it into the scope; at the moment, we anticipate our limited time will be a factor that prevents us from doing so. With a deep learning algorithm, we would be able to enhance our accuracy and predict better recommendations for our users.

References

M., M. (2021) Steam reviews dataset 2021, Kaggle. Available at: https://www.kaggle.com/datasets/najzeko/steam-reviews-2021?select=steam_reviews.csv (Accessed: December 10, 2022).