|  |
| --- |
| Term Project Overview |
| DSC630-T301 Predictive Analytics |

|  |
| --- |
| By: Gabriel Avinaz, Joshua Greenert, and Mithil Patel  3-2-2023 |

**Introduction**

To enhance users' shopping experience and help them make correct purchases, Steam Technologies LLC wanted to develop a gaming recommendation system. Not only will it increase sales, but it would assist in developing the company's credibility for consumers' gaming needs. The need only becomes even more relevant after seeing such a market share held by other gaming platforms, and how often users are playing the games that they offer through their respective services. Therefore, the competitor within the gaming industry that builds a model that can predict what games users would want to play — after users play other games — would lead by the example that other companies would follow.

All our data used throughout this project were obtained through Kaggle.com from users uploading data collected through Steam Technologies LLC. Kaggle is a brilliant source for datasets, especially for budding data scientists. The data, specifically obtained from Kaggle, contains Steam user reviews where users have reviewed games they’ve played or purchased to encourage — or discourage — others from purchasing the game, and Steam games on the platform with short descriptions about the games themselves. Both datasets were selected because of their massive size and relevance to the upcoming project. They are useful in solving the problem because the sheer size of the data allows us to feed machine learning models easily to build predictions based on what all is encompassed.

**Methods/Results**

When we first started this project, we explored the data by graphing what we could see. We wanted to grasp a better understanding of what data we had, and how we could use this data to tell a story. From the graphs we created, we could quickly see that some games had excessively more reviews than others. Was this because of the playability? Was this a result of fandom on another level where users only rated games based on ones that others were rating? Being gamers ourselves, it was easy to identify why the sample we grabbed showed a massive amount of support since the games listed with the most reviews were typically providing a lower skill gap. However, that wouldn’t be as transparent to the layman.

We created bar charts to drill down into the data so we could see how many users left reviews, how fruitful were the reviews that were left, and other various scatterplots to determine the number of positive/negative reviews. We saw quickly that this was going to be a larger challenge than we had initially thought. Several users left comments that didn’t help at all like “dis gud'' or “this is a game”. When we noticed there were a greater number of positive reviews than negative ones, we became concerned that users may have been paid, or encouraged, to write them; we cover this train of thought more in our ethical considerations.

Nevertheless, we began preparing the data by removing all other languages besides English. We started to analyze the columns that we found useful, and removed the ones that we decided couldn’t provide us further insight into the data. For the categorical variables that remained besides the user reviews, we made dummy variables to simplify the process for our models and to give us more options in terms of models we could select at a later stage. For the user reviews, we used stop-words and lemmatization to remove or change any words that wouldn’t be significant in assisting the process. The process to fully correct all user reviews took about 6 hours in total! After coming to this point, we decided that it was best to set up a checkpoint where the data would be exported so that we could start the next steps without performing all the steps discussed previously.

After considerable discussion in our meeting, we decided to create two models, and then compare the results of each to determine which was the better one to finalize. We chose a collaborative model using Nearest-Neighbors (NN) — an unsupervised version of a machine learning model used to calculate the closest items to a target — and a content-based model using the Term Frequency-Inverse Document Frequency (TF-IDF) — an unsupervised machine learning model that utilizes the importance of words by calculating their frequency (Logunova, 2020; Ramadhan, 2021). Based on its familiarity and efficiency, we decided to use cosine similarity for both models. With both models using a metric that provides accuracy data, we could then compare the results afterwards to see how similar/dissimilar they were.

**Lessons Learned**

After completing all the work for both models, we felt there were several important lessons along the way. First, massive amounts of data present some significant processing challenges. Because of these processing challenges, we found that alternative techniques had to be used to enhance performance and reduce break points to not melt our CPUs in the process. This also further stressed the importance of using a version control system (VCS) like GitHub to store our data and ensure its safety/accuracy. Next, we found that the results of our models were different in the way they were required to be viewed; for instance, a lower score on KNN is better for accuracy while a higher score on TF-IDF is better for accuracy. Moreover, our collaborative model found similarities in users, and the games they played, while the content-based model found similarities in the games themselves. Finally, and most importantly, we learned that Subject Matter Experts (SMEs) are crucial to the process of model development. Domain knowledge was essential for our analysis so we could determine whether the results we received back from each model were good or bad; without this knowledge, we wouldn’t have understood if our model was making similar game suggestions or if the model was simply choosing at random.

**Recommendations**

Based on our analysis, the unsupervised Nearest-Neighbors (NN) model was the most optimal and is ready for deployment. However, there is some possible work that can be done to improve the model. For instance, the current model produces a generated list of game suggestions that doesn’t consider the age of the user, or the rating that the user should be set to. A filter could be established that recommends games based on the user’s personal information and settings to avoid potential conflict from consumers. Additionally, a web interface will need to be constructed and connected with Steam Technologies LLC to display the recommendations in a user-friendly view while matching the theme/style of the company for brand recognition.

**Ethical Implications**

Due to the presence of popular games in our dataset, there could potentially be a question surrounding biases since popular games are likely to have more reviews which increases the likelihood of getting recommended by the recommendation system. Another ethical concern involves the authenticity of the review dataset, as review datasets can easily be manipulated or faked by gaming companies to promote their product which can also lead to biased recommendations. However, the ethical concerns were mitigated by using a large dataset to account for biases to create an unbiased and fair recommendation system during the model development phase. An ethical implication to consider if the model was live in production would be the social effect the recommendation would have on users. Since the recommendations are generated based on similarity, users may not be exposed to different game genres. For example, a user searching for violent games will only be exposed to violent games, which can result in a multitude of other problems. The recommendation system can generate a completely random game along with the top similar results to ensure users are being exposed to different genres. Furthermore, age restriction filtering can be implemented to ensure younger gamers get age-appropriate recommendations.

**References**

Logunova, I. (2022, September 20). KNN Algorithm in Machine Learning. Serokell Blog. <https://serokell.io/blog/knn-algorithm-in-ml>

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)

Milano, Silvia, et al. “Recommender Systems and Their Ethical Challenges - Ai & Society.” SpringerLink, Springer London, 27 Feb. 2020, <https://link.springer.com/article/10.1007/s00146-020-00950-y>

Ramadhan, L. (2021, January 20). TF-IDF Simplified. Towards Data Science. <https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530>

Statista. (2022, February 9). Video games - worldwide. Retrieved February 14, 2023, from <https://www.statista.com/outlook/dmo/digital-media/video-games/worldwide#:~:text=The%20largest%20segment%20is%20Mobile,to%20US%24137.80%20in%202023>

Tristan. “All 55,000 Games on Steam (November 2022).” Www.kaggle.com, [www.kaggle.com/datasets/tristan581/all-55000-games-on-steam-november-2022](http://www.kaggle.com/datasets/tristan581/all-55000-games-on-steam-november-2022)