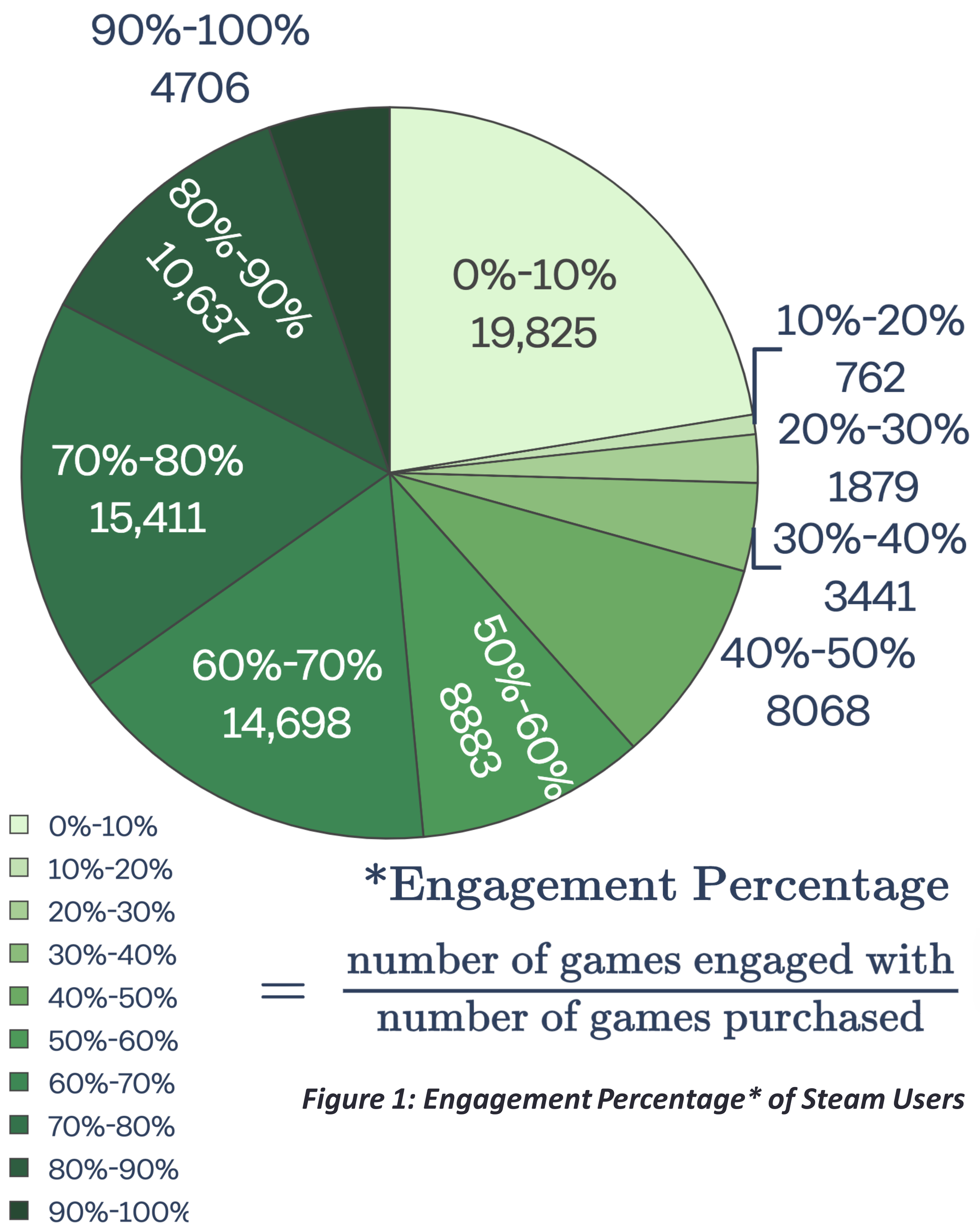
Predicting Video Game Playtime Before Purchase

Chengzhu Duan, Ethan Lan, Ganesh Valliappan, Fengbo Xia, Professor Julian McAuley

UCSD CSE Computer Science and Engineering

BACKGROUND/MOTIVATION

- E-commerce companies recommend items that users are more likely to **buy**.
- These companies are motivated by profits; they may prioritize **selling products** over ensuring **positive user engagement** with products.



$$\text{*Engagement Percentage} = \frac{\text{number of games engaged with}}{\text{number of games purchased}}$$

Figure 1: Engagement Percentage of Steam Users*

OBJECTIVE

Our goal is to explore **what factors determine user engagement** and why users might purchase products they do not intend to use. To accomplish this, we aim to build a reasonably accurate predictor that will **predict if a user will play a game before purchasing**.

Companies: $P(Buy \mid Interests)$

vs.

Our Goal: $P(Play \mid Purchased)$

If they buy it, what would be the probability that they **engage with the game**?

SOLUTION

Our approach is to build a **binary classifier** to predict whether the user will engage with a game the user has not purchased yet. For our project, we used two machine learning models.

Logistic Regression Model

Feature Selection of User-Game Interactions

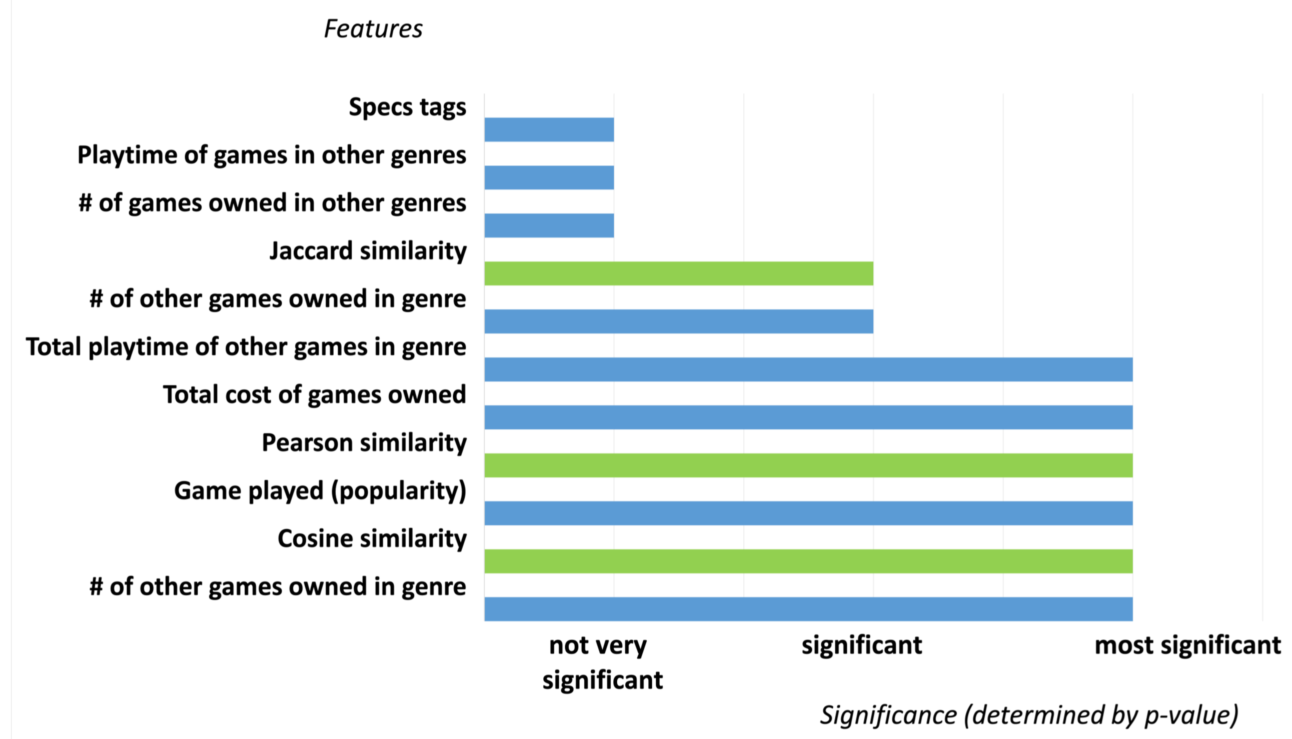


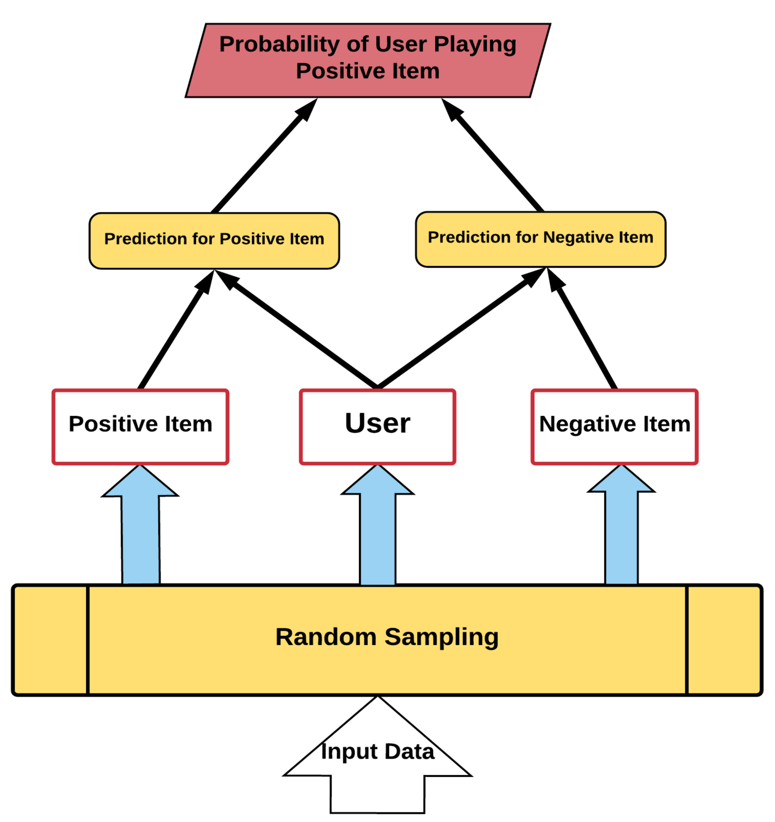
Figure 2: Significance of Features – Significance (α) ≤ 0.00001 for most significant features; α ≤ 0.001 for significant features; α ≤ 0.05 for not very significant features

Feature Engineering of Similarity Measures

Features	Accuracy
Jaccard Similarity	87.1%
Cosine Similarity	87.9%
Pearson Correlation	88.1%

- The similarity measures are calculated between the game of interest and the user's most played game in two weeks and most played of all time.
- Many of the significant features fit our expectations – users tend to play games in genres they like and games that are popular.
- It is surprising that total cost of games is more significant than the total number of games owned.

Latent Factor Model



- **Bayesian Personalized Ranking** with a sigmoid activation function
- **Maximize the difference** between the prediction value of a user's positive item and that of a user's negative item

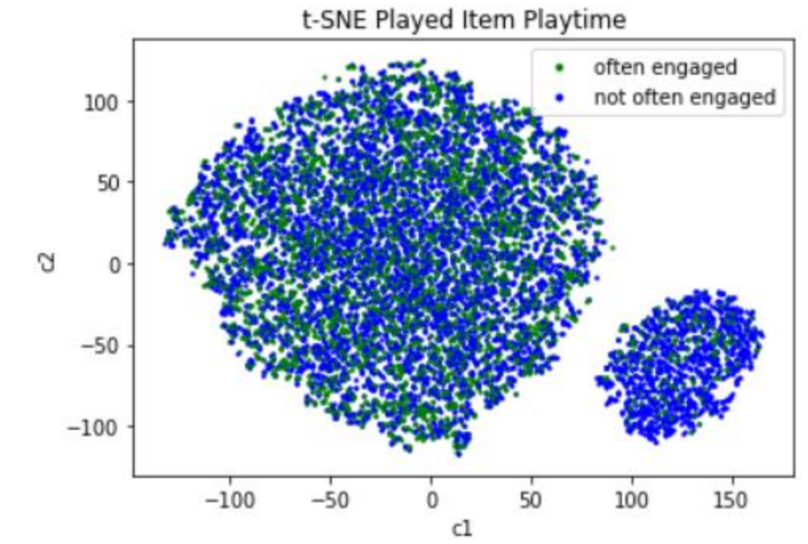


Figure 3: t-SNE for played items engagement Engaged if Mean Playtime < Median Playtime

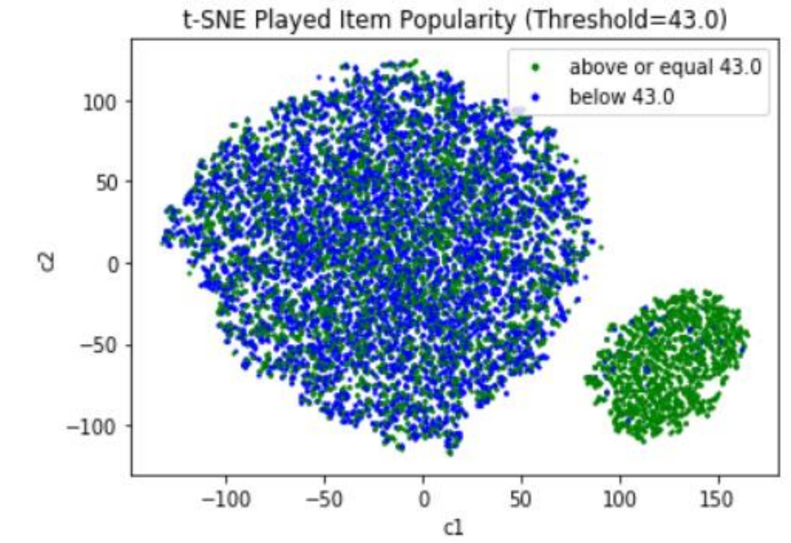


Figure 4: t-SNE for played items popularity Popularity Threshold = Median Popularity (43)

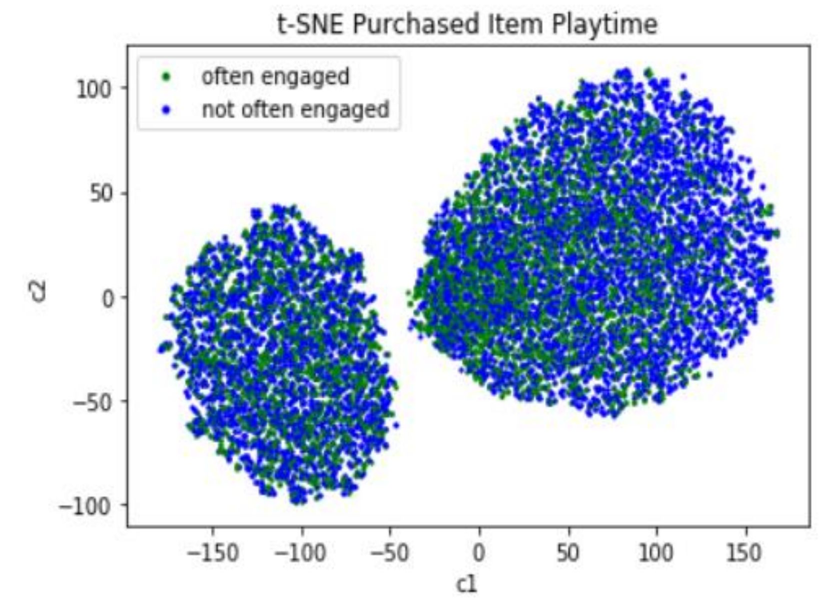


Figure 5: t-SNE for purchased items engagement Engaged if Mean Playtime < Median Playtime

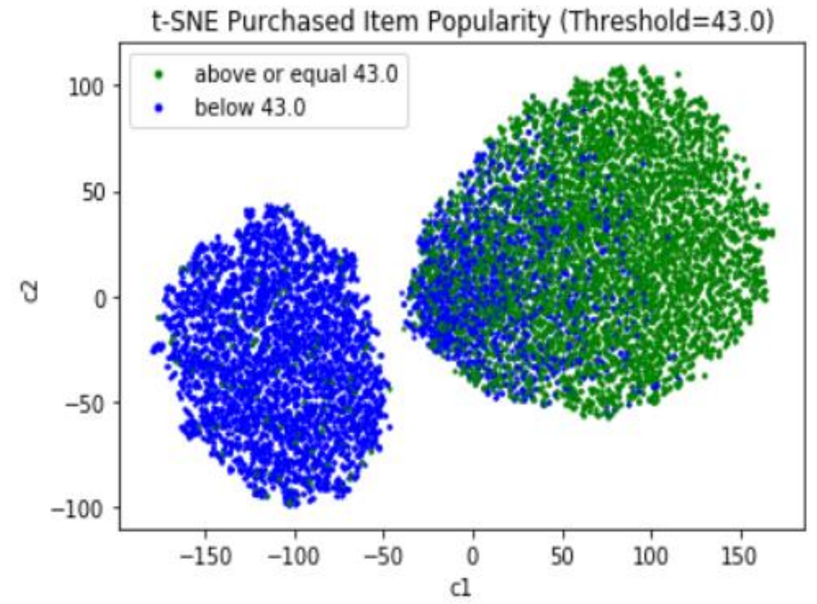


Figure 6: t-SNE for purchased items popularity Popularity Threshold = Median Popularity (43)

RESULTS

Model	Objective	Accuracy
Logistic Regression	Predict Playtime with Implicit Feedback	91.1%
BPR	Determine Whether a User will Purchase a Game	93.5%
BPR	Determine Whether a User will Play a Game	71.6%

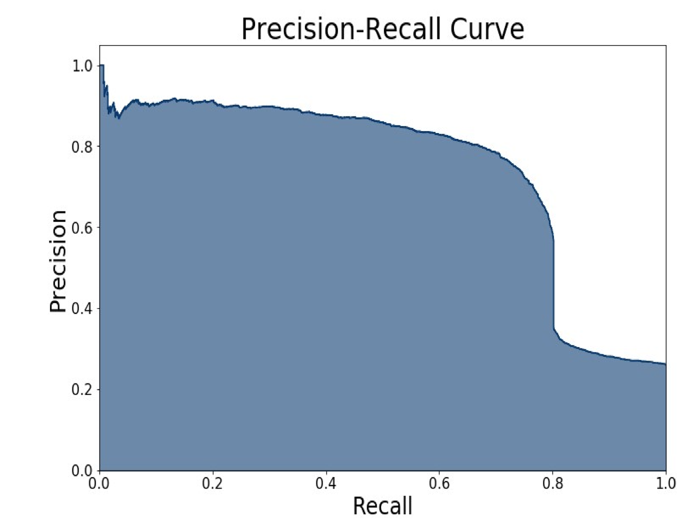


Figure 7: Precision-Recall curve for logistic regression model with all features.

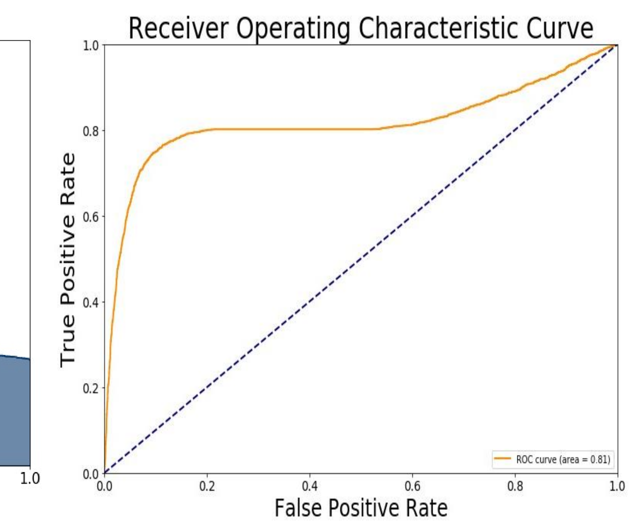


Figure 8: ROC curve for logistic regression model with all features.

The curves show that the logistic regression model is successful in predicting imbalanced classes (25% of games that are engaged with over all games owned).

DISCUSSION

- The logistic regression model in our approach shows that game genres have a large impact on predicting the user's engagement.
- The BPR performs **better** on predicting **user interest** (influence by *game popularity*) than predicting **user engagement** (influenced by *game playtime*).
- Our future work will focus on constructing a recommender system that combines both **BPR** and **feature engineering** and recommends users a list of games on the market which they are most likely to engage with in the future.

ACKNOWLEGEMENTS

We want to thank Professor Julian McAuley for his advising and this research opportunity. We also want to thank Professor Christine Alvarado, Vignesh Gokul, and Mengting Wang for their guidance.