CSC 373 Data Mining: Assignment 5

by Alejandro Gonzalez Rubio and J Pinheiro

Wednesday, March 19, 2025

1 Understanding

Before modeling, we conducted an initial analysis to understand the structure and challenges of the dataset.

Dataset Overview:

- 7.79 million total records.
- Most important feature 'hours' shows a large range from 0 to over 42,000.
- Review data includes both game specific entries and general account ones.

Data Quality Issues:

- 99% of entries had missing values in at least one field.
- Some reviews lacked required features such as `text`, `hours`, or `early_access`.

Challenges:

- Large dataset size (7.7 million records). Subsetting might be neccesary.
- Wide range of playtime values required transformation for better modeling.
- Missing values and extreme outliers needed to be handled to ensure model reliability.

2 Working with the Data

Initial dataset: 7.79 million records

Cleaning:

- Removed entries with missing values in required fields: `text`, `hours`, and `early_access`

Post-cleaning dataset: 7.74 million records

Next step:

- Applied an 80/20 split for training and development
- This split was used in all three tasks: Estimation, Classification, and Recommendation

3 Modeling

We are going to perform three data mining tasks:

3.1 Part 1: Estimation

Goal: To estimate the number of hours a user spent playing a game using regression based on features from review data.

Model Used: Linear Regression.

Approach: A classification model (Multinomial Naive Bayes) was first trained on discretized playtime labels to predict playtime bins from review text. These predictions were then used as one of the features in the regression model. We removed outliers above the 90th percentile and applied a log transformation to stabilize variance in the target variable.

Features Used: Multinomial Naive Bayes prediction on review text (discretized hours), Length of the review text, Early access.

3.2.1 Part 2: Regular Classification

Goal: To classify the playing hours using various classification models and identify the model with the highest accuracy.

Models Compared: Dummy Classifier, Multinomial Naive Bayes, Random Forest, Gradient Boosting.

Features Used: Review text (processed with TF-IDF), Length of the review text, Early access.

3.2.2 Part 2: Time-Based Classification

This approach explores the performance of Multinomial Naive Bayes when trained on data from one time period and tested on data from another.

Goal: To assess whether the model's classification performance changes when trained on reviews from earlier years and evaluated on reviews from later years, and vice versa.

Model Used: Multinomial Naive Bayes.

Features Used: Review text (processed with TF-IDF), Length of the review text, Early access indicator

Training and Testing Scenarios (Based on Year):

1. Train on data from before/during 2014, test on data from during/after 2015.
2. Train on data from during/after 2015, test on data from before/during 2014.

3.3 Part 3: Recommendation

Goal: To predict the number of hours a user is likely to spend playing a game using collaborative filtering.

Model Used: Alternating Least Squares (ALS) from PySpark.

Approach: Used a 5% sample of the data with usernames and product IDs indexed numerically. Applied log transformation and removed outliers above the 90th percentile.

Features Used: Username (indexed), Product ID (indexed).

Model Training: Performed cross-validation over different iteration values to select the best ALS model, which was then used to predict playing hours on the dev set.

4 Results

Part 1: Estimation (Linear Regression)

- MSE on dev set: 11.83
- Underprediction rate: 5.3%
- Overprediction rate: 94.7%

Part 2.1: Regular Classification

- Dummy Classifier Accuracy: 49.9%
- Multinomial Naive Bayes Accuracy: 58.8%
- Gradient Boosting Accuracy: 59.4%

Part 2.2: Time-Based Classification

- Train before/during 2014 → test after 2015: 64.4% accuracy
- Train after 2015 → test before/during 2014: 65.3% accuracy

Part 3: Recommendation (ALS)

- RMSE on dev set: 306.30
- Underprediction rate: 85.5%
- Overprediction rate: 14.5%