Analyzing Determinants of Video Game Success: A Data Mining Approach to User Preferences, Publisher Impact, and Performance Metrics

Flora Chen
The Department of Computer Science
Emory University
Atlanta, GA
flora.chen@emory.edu (Student ID: 2432195)

Jingxuan Zhang
The Department of Computer Science
Emory University
Atlanta, GA
jzh883@emory.edu (Student ID: 2539111)

Jihan Lee
The Department of Comptuer Science
Emory University
Atlanta, GA
jihan.lee@emory.edu (Student ID: 2453072)

***Abstract*—This research employs data mining to analyze video game attributes and player engagement metrics, aiding strategic decision-making in game development, user experience design, and marketing. Through comprehensive data collection and preprocessing of video game attributes, we apply association mining, clustering, classification, regression, and anomaly detection techniques. Rigorous evaluation and validation of models ensure their reliability, while analysis and visualization of results empower stakeholders in the dynamic video game market.**

***Index Terms*—Data Mining, Video Games, User Experience Design, Marketing Strategies, Association Mining, Clustering, Regression, Anomaly Detection, Evaluation, Visualization.**

I. Introduction

A. Motivation

In the fast-evolving video game industry, understanding consumer behavior and preferences is crucial. With a diverse range of genres, platforms, and gaming experiences available, developers and publishers constantly seek insights into gamer engagement. Our project aims to harness gaming data to inform strategic decision-making in game development, user experience design, and marketing efforts. By analyzing consumption patterns, including genres, player counts, and publishers, alongside metrics like playtime and review scores, we aim to provide actionable insights for future gaming projects.

B. Problem Statement

This project aims to understand gaming consumption and performance metrics to identify factors contributing to a game's popularity and financial success. It focuses on user preferences for game genres, the impact of game publishers on player engagement, and how these factors correlate with performance indicators like review scores and sales figures.

C. Significance of the Problem

The significance of the problem lies in its potential to transform game development, marketing, and optimization strategies. Insights gained can enable developers and publishers to better align projects with audience preferences, optimize marketing efforts, and allocate resources effectively. Understanding industry trends aids in staying competitive, ultimately enhancing the gaming experience for users and benefiting the gaming industry as a whole.

II. Related Work

Data mining has become a pivotal tool across various industries, including the video game sector, to glean insights from vast datasets to inform decision-making processes. Research in this area explores multiple dimensions, from consumer behavior analysis and strategy optimization to ethical considerations and privacy concerns. For instance, studies such as those by Coll (2013) [1] and Danna & Gandy (2002) [2] delve into how retail companies leverage data mining through loyalty programs to understand customer preferences and the social implications of consumer profiling. These methodologies are readily applicable to the video game industry, where understanding player behavior and preferences is crucial for game development and marketing strategies.

In the realm of gaming, specific studies like those by Bosc et al. (2014) [3] and Suh & Alhaery (2015) [4] showcase the application of data mining to uncover winning strategies in real-time strategy games and predict customer behavior, respectively. These insights are instrumental for game developers and marketers in crafting games that engage and retain players. Furthermore, the work by Aziz et al. (2018) [5] emphasizes the role of data mining in analyzing factors leading to video game sales success, highlighting the importance of leveraging user-generated content and gameplay data to drive business analytics and enhance the gaming experience. Collectively, these studies underscore the multifaceted applications of data mining in not only enhancing player engagement and satisfaction but also in navigating the complex ethical landscape of data use within the video game industry.

III. Intended Proposed Approaches

- Data Collection and Preprocessing:
  Gather comprehensive data on video game attributes and engagement metrics from reputable gaming databases and platforms. Preprocess the data to ensure the quality and reliability of the dataset for analysis in the context of the video game industry.

- Dimensionality Reduction and Feature Selection:
  Use feature importance ranking to select relevant attributes indicative of player behavior, game performance, and market trends. Employ dimensionality reduction methods like feature engineering and aggregation to reduce dataset complexity.
- Association Mining:
  Apply association mining techniques like Frequent Itemset Mining and the Apriori algorithm to identify patterns and associations between video game attributes and player behaviors. Discover frequent itemsets representing common combinations of game genres, platforms, and publishers, unveiling insights into player preferences, cross-platform availability, and licensing trends in the video game market.
- Clustering Methods:
  Use K-means and hierarchical clustering to segment video games based on attributes and player engagement metrics like platform, region, and genre. Group games with similar characteristics to enable targeted analysis and personalized recommendations for game design, marketing strategies, and platform optimization.
- Classification and Regression Methods:
  Develop classification and regression models to predict video game performance indicators, using input variables like game genres, platforms, and publishers. Train models to categorize games based on player preferences and engagement metrics, revealing factors influencing a game's success in the competitive market.
- Anomaly Detection Methods
- Evaluation and Validation:
  Evaluate the performance of the proposed approach using appropriate metrics, such as accuracy, precision, etc., to assess the predictive power and reliability of the models. Validate the findings to ensure the robustness and generalizability.
- Analysis and Visualization of Results

IV. Intended System Design

A. System Architecture

The system architecture is designed to be modular and scalable, facilitating easy integration of new data sources and analytical modules as needed. It consists of four main components:

- Data Ingestion Module: Automates the collection of data from various sources, including public APIs, databases, and manually curated datasets.
- Data Processing and Storage Module: Cleans, transforms, and stores the ingested data in a structured format suitable for analysis. This module will use a combination of relational and NoSQL databases to accommodate the variety of data types and structures encountered.
- Analysis and Machine Learning Module: Performs statistical analysis and machine learning on processed data to uncover patterns, correlations, and trends. It will also include models to predict future trends based on historical data.
- Visualization and Reporting Module: Generates interactive dashboards and reports to present the findings in an accessible manner to stakeholders.

B. Programming Languages

- Python: The primary language due to its extensive support for data analysis and machine learning through libraries like Pandas, NumPy, scikit-learn, TensorFlow, and Matplotlib. Python's readability and broad community support make it ideal for this project.
- SQL: For database queries and management, SQL will be used to interact with relational databases where structured data is stored.

C. Software Tools

- Jupyter Notebook: For interactive development and testing of data analysis and machine learning models.
- GitHub: For version control and code management, ensuring that all project contributors can collaborate effectively.
- MySQL: For storing structured data such as user profiles, game metadata, and performance metrics.

D. Data Sets

The project will leverage several datasets, including:

- Video Games CSV File
- Video Game Sales

References

[1] Coll, S. (2013). Consumption as biopower: Governing bodies with loyalty cards. Journal of Consumer Culture, 13, 201 - 220. https://doi.org/10.1177/1469540513480159.

[2] Danna, A., & Gandy, O. (2002). All That Glitters is Not Gold: Digging Beneath the Surface of Data Mining. Journal of Business Ethics, 40, 373-386. https://doi.org/10.1023/A:1020845814009.

[3] Bosc, G., Kaytoue-Uberall, M., Raïssi, C., Boulicaut, J., & Tan, P. (2014). Mining Balanced Sequential Patterns in RTS Games. , 975-976. https://doi.org/10.3233/978-1-61499-419-0-975.

[4] Suh, E., & Alhaery, M. (2015). PREDICTING CROSS-GAMING PROPENSITY USING E-CHAID ANALYSIS. UNLV gaming research & review journal, 19, 3.

[5] Aziz, A., Ismail, S., Othman, M., & Mustapha, A. (2018). Empirical Analysis on Sales of Video Games: A Data Mining Approach. Journal of Physics: Conference Series, 1049. https://doi.org/10.1088/1742-6596/1049/1/012086.

Appendix

1) Task Assignment
   - Flora: Writing introduction, abstract, analysis, writing report
   - Jingxuan: System design, model development
   - Jihan: Related literature review, preprocessing, visualization, writing report
2) Time Schedule
   - Week 1: Project Planning and Data Collection
   - Week 2-3: Data Preprocessing and Exploration
   - Week 4-6: Model Development and Evaluation
   - Week 7-8: Analysis and Visualization
   - Week 9: Final Report and Presentation