Predicting the Success of Indie Games on Steam Using Metadata and Machine Learning Models

Derek W. Graves

Northwest Missouri State University, Maryville MO 64468, USA
S573443@nwmissouri.edu and derek.graves4@outlook.com

Abstract. This study aims to predict the commercial success of indie games on the Steam platform by analyzing game metadata. Utilizing data collected from the Steam API, machine learning techniques, including Random Forest and Logistic Regression, were implemented to identify significant attributes contributing to game popularity. By exploring factors such as gameplay features, pricing strategies, and developer information, this research aims to offer practical insights for indie developers looking to enhance player engagement and maximize their games' market success.

Keywords: machine learning · data analytics · Steam · indie games

1 Introduction

The purpose of this research is to analyze and predict the success potential of indie games on Steam based on available metadata. By utilizing the Steam Web API to gather data, machine learning models—such as Random Forest and Logistic Regression—were applied to explore the attributes that contribute most significantly to a game's popularity. The ultimate aim is to provide developers with actionable insights to optimize their games for improved audience engagement.

1.1 Research Goals

The primary goals of this study include:

- Identifying which game metadata features are correlated with the success of an indie game.
- Applying predictive machine learning techniques to forecast game success based on these features.
- Providing meaningful insights that indie game developers can use to enhance the reception of their games.

2 Data Collection

The dataset for this study was collected using the Steam Web API, which offers comprehensive metadata for all games on the platform, including gameplay features, pricing, developer details, and user engagement metrics. Documentation from the Steam API Documentation was referenced to understand the API's various parameters and response formats.

2.1 Source of Data

The dataset consists of data from indie games released between 2010 and 2024. Data was gathered from various regions, including North America, Europe, and Asia, capturing a diverse set of game genres such as action, role-playing, puzzle, and adventure. This broad spectrum helps ensure that the analysis captures different aspects of indie game success.

2.2 Data Extraction Procedure

The dataset was extracted using Python, leveraging the Requests library to interact with the Steam API. The specific procedure involved:

- **API Integration:** Metadata was retrieved by calling the Steam Web API through the `Requests` library in Python, which provided access to detailed information about each game.
- **Rate Limiting and Retry Logic:** To avoid exceeding API limits, pauses were added after every 10 requests. A retry mechanism was implemented using the `Retry` feature from the `urllib3` package, which helped mitigate server timeouts and manage transient errors effectively.
- **Data Filtering:** The dataset was filtered to focus only on completed indie games, explicitly excluding adult content, demos, downloadable content (DLC), and games still in early access.
- **Data Storage:** The collected data was saved in CSV format, containing both the full dataset and a balanced subset that focused on games with varying levels of popularity for further analysis.

2.3 Data Format and Volume

The dataset comprises 300 game records, each containing seven attributes: `AppID`, `Game Name`, `Release Date`, `Developer`, `Genres`, `Price ($)`, and `Recommendations`. Games were classified into three popularity tiers—low ($\leq$50 recommendations), moderate (50–500 recommendations), and high (more than 500 recommendations). Any entries with incomplete metadata were excluded to maintain data quality.

2.4 Dataset Attributes

Table 1 provides an overview of the dataset attributes:

Column Name	Description	Data Type	Example Value
AppID	Unique identifier for each game	Integer	440
Game Name	Title of the game	String	Team Fortress 2
Release Date	Date when the game was released	String	October 10, 2007
Developer	Developer(s) of the game	String	Valve
Genres	Genres associated with the game	String	Action, Free-to-Play
Price ($)	Price of the game in USD	Float	19.99
Recommendations	Number of recommendations received	Integer	50000

Table 1. Attributes of the Indie Games Dataset

2.5 Other Considerations

Challenges during data collection included handling incomplete metadata for certain games and managing the rate limits imposed by the Steam API. Future work could consider integrating data from other platforms, such as the Epic Games Store, to broaden the analysis and improve model performance.

3 Data Pre-processing and Cleaning

The data pre-processing phase included encoding categorical variables (such as genres) and normalizing numerical attributes (e.g., price and recommendations). Specifically:

- **Handling Missing Data:** Any game entries lacking essential metadata were removed.
- **Categorical Encoding:** Techniques like one-hot encoding were used to transform genre information into a format suitable for machine learning models, as outlined by Bellavista et al. [1].
- **Normalization:** Numerical features, such as price and recommendation count, were normalized to improve model accuracy.

4 Model Development and Performance Assessment

The dataset was split into training and testing subsets in an 80:20 ratio. Random Forest and Logistic Regression models were then used for predictive analysis:

- **Random Forest:** Selected for its robustness and ability to manage datasets with high variance. It was also suitable for capturing non-linear relationships in the data.

- **Logistic Regression:** Chosen for its interpretability, especially useful for understanding the relationship between features and game success. This model worked well for lower-dimensional datasets, as discussed by Kirasich et al. [2] and Lounela [3].

4.1 Training Details

The training data was used to fit both models. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. Hyperparameters for both models were optimized using grid search, which helped fine-tune the models and enhance prediction accuracy.

5 Results and Discussion

Upon training the models, insights into feature importance were gained. For instance, the Random Forest model showed that the number of recommendations and the price were critical in predicting game success. Visual representations, such as feature importance plots, will be incorporated to further illustrate model performance and findings.

6 Limitations and Future Work

The study faced limitations, including potential biases in the metadata due to the exclusion of certain game types (e.g., adult content and early access). Furthermore, constraints on niche game availability may limit the generalizability of the findings. Future research could involve incorporating data from additional platforms (e.g., Epic Games Store) and utilizing deep learning models to improve predictive accuracy.

7 Conclusion

This study explored the use of machine learning in predicting the success of indie games on Steam. By identifying which features most significantly influence game popularity, actionable insights are provided to indie developers to help them optimize player engagement and commercial outcomes.

Additional Resources

For more details, please refer to the project resources below:

- Overleaf Report
- GitHub Repository
- GitHub Data Directory
- Steam API Documentation

References

1. Bellavista, P., Corradi, A., Stefanelli, C.: Mobile agent middleware for mobile computing. Computer **34**(3), 73–81 (2001). https://doi.org/10.1109/2.910896
2. Kirasich, K., Smith, T., Sadler, B.: Random forest vs logistic regression: Binary classification for heterogeneous datasets. SMU Data Science Review **1**(3), Article 9 (2018), bluehttps://scholar.smu.edu/datasciencereview/vol1/iss3/9, creative Commons License
3. Lounela, K.: On identifying relevant features for a successful indie video game release on steam. Master's Programme in Department of Information and Service Management (2024), bluehttps://aaltodoc.aalto.fi/items/d578980e-71fa-4618-b500-dff30bbac490