Game of the Year (GOTY) Candidate Prediction Model

Analytics & Discovery Informatics

Project Proposal & Results – Fall 2024 – Syed Hashim Raza

Introduction

This project aims to predict Game of the Year (GOTY) nominations for video games specifically on the Steam platform using machine learning techniques. In an era where the gaming industry stands as a cultural powerhouse, predicting GOTY nominees has become a compelling pursuit for both developers and fans alike. Steam is a leading digital distribution platform, developed by Valve Corporation. Steam serves as a vast marketplace where millions of gamers worldwide access, purchase, and play games across a variety of genres. Beyond just hosting games, Steam provides a community-driven environment, featuring reviews, ratings, achievements, and in-depth player statistics, making it an invaluable resource for gathering data on gaming trends and audience preferences.

As an avid gamer and someone who closely follows the industry's evolution, I am captivated by the dynamics that shape a game's journey to critical acclaim. What makes a game resonate so deeply with an audience? How do elements like design, storytelling, and gameplay converge to a title of GOTY status? This project seeks to answer these questions by building a supervised learning model that leverages features extracted from Steam's dataset to classify games as potential GOTY nominees.

By diving into this project, I not only hope to enrich my understanding of the factors that drive a game's success but also harness machine learning as a tool to reveal patterns and insights within the vast and ever-changing gaming landscape.

The Approach

Problem Framing

The central problem this project addresses is the prediction of GOTY nominations for video games. This relates to a generating process where various game characteristics and performance metrics contribute to a game's likelihood of being nominated. The goal is to construct a predictive model that uses features extracted such as genre, gameplay mode, player engagement statistics, and user-generated reviews and ratings, while the objective function is to maximize the accuracy of predictions. The underlying assumption is that these variables collectively represent aspects of a game's appeal and critical reception, which are fundamental in determining its GOTY nomination potential.

Data Framing

The dataset includes features capturing each game's characteristics, player engagement, and reception, specifically:

- Game Characteristics (genre, multiplayer options, release dates)

- Player Statistics
- Review Scores
- Release Information (year, season)
- Price and Discount data

The target variable is the binary "*GOTY_Nomination*" label. To optimize for GOTY prediction, player statistics may be transformed to reflect engagement intensity, while text features like genre will be encoded. The data will be structured into a feature matrix X and a target vector y, then split into training and testing sets. This setup allows for effective model training, parameter tuning, and performance assessment, enabling the project to identify the most predictive GOTY nomination factors.

Objective Framing

The goal of this prediction model is to classify games as potential GOTY nominees accurately. To evaluate model performance, we will use the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), as our primary metric. AUC-ROC is ideal for assessing the model 's accuracy in distinguishing between GOTY nominees and non-nominees, even if the dataset is imbalanced. Additionally, precision and recall metrics will help monitor the model's capability in identifying true nominees. The combination of metrics will guide the training process, helping prediction accuracy and generalizability.

Data collection and processing

The data collected for this project is via two different sources, SteamAPI and SteamSpy. SteamSpy is a third-party analytics tool that aggregates public data about games on Steam, providing valuable statistics on player engagement and ownership. It offers high-level aggregated metrics like estimated player count, playtime distributions, and genre popularity. These metrics are particularly useful for identifying overall game popularity and engagement trends. However, since SteamSpy data is based on estimation and aggregation, it may lack granularity in real-time player statistics and specific gameplay details.

Steam API is the official API, provided by Valve, enables more granular data collection about individual games on Steam, such as real-time player count, detailed user reviews, and game-specific metadata like pricing, release date etc. It requires an API Key, which was easy to obtain, as only a steam account is required. However, it does not provide aggregate metrics on ownership or player demographics like SteamSpy, as it focuses more on individual game characteristics. Another challenge via the Steam API is the request limitations due to which extracting data becomes a very time-consuming process.

Using both sources enriches the dataset by combining SteamSpy's broader audience insights with detailed, real-time game attributes from the Steam API. Together these sources provided a balanced dataset for training an ML model that leverages both game-specific details and broader engagement trends.

There are 101,035 games listed on Steam as of June 2024. Using all the game data listed on Steam would result in an unimaginably large dataset with many datapoints potentially irrelevant for predicting GOTY nominations. By narrowing down the dataset to the 5,000 games with high player counts, the dataset focuses on popular games that are likely to be candidates for critical acclaim.

This approach balances dataset size and relevance and aligns it with the ML application's objective of predicting high-impact nominations. For our purposes, the dataset was narrowed down to 5,000 games with the highest player count. That task proved to be impossible via the Steam API since it does not offer a built-in feature for ranking or filtering games by player count. Steam API data is accessible on a per-game basis, meaning you need to know the specific game IDs beforehand to make requests. SteamSpy was utilized in this regard since they offer aggregates and estimates broader player metrics across all games, allowing it to rank by popularity. This allows requests to be made using an efficient filtering method. Using those filter, 5,000 games and their data was extracted via SteamSpy.

An advantage of extracting data via API calls is that we can request for specific information that we require and ignore the data that we do not require for the purposes of this project. By reading through the official API documentation, keywords were used in the requests and the data was output in a JSON file, which was transformed into a csv file format for further processing.

From the data collected, there were a total of 22 feature columns which includes a mix of categorical, numerical and date data types associated with a single game, and a total of 5,000 observations (games). A few feature columns were completely null, so those were removed. Null values were checked in other feature columns as well. Different decision-making processes were made with each feature with missing values.

Genre features had 1049 null values out of 5,000. Those null values were imputed with 'Unknown'. Furthermore, the genre column has multi tags for each game out of a total of 23 total genre categories. Since this is useful data for our model, the genre column was broken down into individual tags, and one-hot encoding was done to create 23 separate feature columns for each game in the list. Similarly, the Release Date feature was also broken down using the months to create 4 feature columns for the seasons, also one-hot encoded for modeling purposes. The release year was also extracted to a separate feature column.

The owners of a game feature were extracted in a range form, since SteamSpy gives an estimated number as an output. That feature was converted to the average number for simplicity purposes. Multiplayer feature of a game had 4 separate feature columns, which broke down the different variations of multiplayer options, but essentially gave the same form of information, so those 4 feature columns were converted to a single feature column for simplicity purposes.

Some genre feature observations were in non-English symbols, and some tags were in a different language. That was also processed to remove the symbols, and translate the different language tags into English, keeping the integrity of the datapoints.

Another important step taken was updating the release date column via the Steam API. The code to request the information and update the column was ran again since SteamSpy did not have that data. After updating the dataset via Steam API data retrieval, some games still did not have release data information, either because of them being very old, or unpopular that it was never updated in the Steam database. This step was critical for temporal integrity for the model's accuracy.

Since the top 5,000 games were pulled via SteamSpy due to reasons shared earlier, there were too many game duplicates in the dataset. All the duplicates were removed, leaving us with a total dataset of 933 games.

Finally, our dataset did not have a target label column. That was an issue for prediction modeling accuracy. To engineer that target label, 'The Game Awards' online source was used to manually find out the past GOTY candidates from 2011 till 2024. Using the corresponding unique AppID of those past candidates, the target label column was created in a binary form with 1 as candidate game and 0 for non-candidate game.

For a detailed process of the data cleaning and target label creation process, reference the hex project proposal Dataset Summary where all the code and steps are shared with detailed comments to explain each step of the process to recreate raw dataset and process it to the final dataset that will be used in the project.

Literature Review

The application of machine learning in predicting video game success, particularly in forecasting Game of the Year (GOTY) nominations, is a developing area of research. This project builds upon various predictive methodologies from recent studies, particularly those that explore player behavior, game features, and temporal trends.

Akhmedov and Phan (2021) explored predictive modeling for Dota 2 match outcomes, applying Linear Regression, Neural Networks, and Long Short-Term Memory (LSTM) models. By using real-time data collected via Game State Integration, they achieved up to 93% accuracy. This study is closely related to the GOTY prediction model as it demonstrates how combining complex machine learning models with highly granular data can improve outcome predictions in games with dynamic player interactions (Akhmedov & Phan, 2021). The use of LSTM offers valuable insights into how sequential data might improve GOTY predictions by considering player engagement patterns and game performance metrics over time.

Similarly, Owen and Baker (2018) introduced a feature engineering framework, IDEFA, aimed at optimizing predictive models for serious games by identifying and extracting key player behavior attributes. Their work is crucial for this project's focus on feature engineering, as they found that well-engineered features significantly enhance model interpretability and accuracy. For GOTY prediction, this suggests that curated features such as genre, release date, and user ratings can provide a strong foundation for classification models (Owen & Baker, 2018). This framework's systematic approach to feature selection aligns well with the project's data processing needs.

In addition, Chen et al. (2017) focused on player skill decomposition in multiplayer online battle arenas (MOBAs). They examined player engagement and skill indicators to predict game outcomes, emphasizing the impact of player-centric metrics. This study's approach to analyzing player performance data is particularly relevant to GOTY prediction as it highlights how game-specific features like playtime and multiplayer options contribute to success (Chen et al., 2017). By incorporating player engagement statistics, GOTY predictions can better reflect user satisfaction and critical acclaim, two key aspects in award recognition.

Other studies provide further context. A project on CodeSpeedy used XGBoost to predict video game sales, emphasizing model tuning for higher accuracy (CodeSpeedy, n.d.). Another IEEE Xplore study applied hybrid feature selection methods for video game sales forecasting, underscoring the importance of tailored feature engineering (IEEE Xplore, n.d.). Additionally, a GitHub project used decision trees, SVM, and KNN for video game genre prediction, demonstrating model adaptability for categorical data and supporting the GOTY model's objective to differentiate success across genres (GitHub, n.d.).

These studies collectively underscore the importance of feature engineering, model complexity, and the role of qualitative success factors in predicting GOTY nominations. Insights from Akhmedov and Phan (2021), Owen and Baker (2018), and Pfau et al. (2022) directly inform the structure of this project by guiding the selection of features and advanced models suitable for identifying GOTY candidates.

Methodology

This GOTY prediction model employs both baseline and complex models to classify games as potential nominees. The modeling process includes structured feature engineering, player engagement metrics, and robust evaluation metrics to ensure model accuracy and interpretability.

1. **Feature Engineering**
   Based on insights from Owen and Baker (2018), this project emphasizes feature engineering to capture critical aspects of a game's success. Key features include genre, multiplayer options, release date, average playtime, and concurrent users. For genre, one-hot encoding is applied to create a distinct feature for each category, allowing the model to recognize genre-based patterns. The release date feature is transformed to capture seasonal effects, encoded by quarter, and separated by year to capture potential GOTY timing advantages. Engagement metrics, such as average playtime and user reviews, are calculated to reflect sustained interest and satisfaction, crucial indicators in award candidacy.

2. **Model Selection**
   Three models were chosen to build this predictive framework, incorporating both baseline and complex techniques:

   - **Baseline Models**: Logistic regression and decision trees serve as baseline models. Logistic regression is selected for its straightforward interpretability in binary classification tasks, allowing clear insights into feature importance, which is beneficial for understanding key GOTY indicators. Decision trees are included to capture non-linear relationships, particularly interactions between player engagement and genre, as noted in Chen et al. (2017).

   - **Advanced Model – Gradient Boosting Machine (GBM)**: Building on the literature's insights into complex model application, GBM is employed for its superior accuracy and effectiveness with imbalanced datasets. GBM captures intricate patterns, particularly those driven by player engagement metrics, as seen in Akhmedov and Phan's (2021) model for real-time prediction. This approach supports our objective to balance model precision with depth in capturing GOTY-related characteristics.

3. **Sequential Data Consideration**
   Inspired by Akhmedov and Phan's (2021) LSTM approach, we considered sequential models for analyzing temporal dynamics in player engagement. Although not implemented in the current model, future iterations may incorporate LSTM to enhance model performance based on evolving player metrics.

4. **Evaluation Metrics**
   To thoroughly assess model effectiveness, we use AUC-ROC as the primary metric for distinguishing between GOTY nominees and non-nominees, with precision, recall, and F1

score as supplementary metrics. These evaluation measures, supported in the literature for balancing interpretability and accuracy, enable a comprehensive understanding of the model's predictive capabilities and ensure robustness against imbalanced class distributions.

This methodology aligns with the literature's emphasis on feature engineering, player engagement analysis, and model evaluation, making it well-suited for the complex task of predicting GOTY nominations.

Practical Application of the Base Model

The logistic regression model can be used to support decision-making in game development and marketing by identifying features most likely to influence a game's GOTY nomination potential. This insight can guide developers in prioritizing resources and design elements that align with award-winning attributes.

For example, if the model shows that high user ratings and longer average playtime have a strong positive association with GOTY nominations, developers might focus on enhancing gameplay quality and engagement. Additionally, if multiplayer options or specific genres are correlated with increased nomination likelihood, marketing teams can tailor campaigns to highlight these elements.

Feature Explanation and Association

Based on initial assumptions, several features are expected to impact the model's predictions:

1. **Review Scores** (positive association): Higher positive reviews likely indicate quality and enjoyment, factors associated with GOTY nominations.
2. **Player Engagement (**positive association): Metrics like average playtime and concurrent players often reflect sustained interest, suggesting high replay value and engagement, which aligns with GOTY standards.
3. **Genre** (variable association): Certain genres might correlate strongly with GOTY nominations based on recent trends (eg., RPGs or action-adventure games).
4. **Release Timing** (positive association): Games released in certain seasons or close to the award cycle may gain more attention, enhancing nomination chances.

This model's insights can help developers and marketers align product features and promotion strategies with factors that increase a game's likelihood of critical recognition.

Model Fitting and Results

Initial results and data modification requirement

Initially, I expected ensemble methods such as Random Forest and Gradient Boosting to perform well despite the dataset's imbalance. However, after fitting the models on the unfiltered dataset, the results were suboptimal and displayed severe problems due to the imbalance of the target label.

These weak performances prompted a critical reassessment of the dataset. I concluded that predicting GOTY (Game of the Year) nominations requires a focus on high-engagement games, as GOTY contenders are typically critically acclaimed, highly popular among players and reviewers, and represent a very competitive subset of games. Each year, only 4-5 games receive GOTY nominations, underscoring the need to filter out low-engagement games to improve the dataset's relevance and balance.

To address this, I implemented a data filtration process aimed at retaining high-engagement games while ensuring all GOTY nominations remained in the dataset. The filtering criteria included:

1. **Owners**: Games with at least 3,500,000 owners.
2. **Median Playtime**: Games with a median playtime of at least 769 minutes.
3. **Reviews**: Games with at least 75,952 positive reviews.
4. **Average Playtime**: Games with an average overall playtime of at least 2275 minutes.

After applying these filters, the dataset quality improved significantly for the prediction model. Removing low-engagement games reduced noise and ensured the model focused on games that had made a substantial market impact. This process not only improved data quality but also helped balance the target label, a critical step for building a predictive model for GOTY nominations.

By focusing on high-quality, relevant data points, the refined dataset provided a more representative sample of successful games. This allowed the model to better identify the characteristics and engagement metrics associated with GOTY nominations, increasing its likelihood of delivering meaningful and accurate predictions.

Model Fitting After Results

After filtering the dataset to include only high-engagement games while retaining all GOTY nominations, predictive models were fitted using logistic regression, decision trees, random forest, and gradient boosting. Filtering improved the dataset by reducing noise and focusing on meaningful game characteristics associated with GOTY nominations.

Methodology

1. **GridSearchCV with Cross-Validation:**
   - All models underwent hyperparameter tuning using GridSearchCV with 5-fold cross-validation. This ensured robust evaluation of model performance on different training and validation splits.
2. **Threshold Adjustment:**

- After fitting, the default classification threshold (0.5) was adjusted for each model to maximize the F1-Score, using precision-recall curves. This adjustment addressed the moderate imbalance in the target variable and ensured a better balance between precision and recall.

3. **Evaluation Metrics:**
   - Models were evaluated based on accuracy, recall, precision, F1 Score, and AUC-ROC.

Updated Results

Model	Threshold	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.344	0.833	0.667	0.800	0.727
Decision Tree	1.000	0.833	0.667	0.800	0.727
Random Forest	0.265	0.833	1.000	0.667	0.800
Gradient Boosting	0.344	0.833	0.667	0.800	0.727

Discussion of Results

1. **Logistic Regression:**
   - Logistic Regression achieved balanced performance after threshold adjustment, with an F1-Score of 0.727 and an accuracy of 0.833. The adjusted threshold improved recall to 0.667 while maintaining strong precision (0.800). Despite its linear nature, the model provided a reliable baseline.
2. **Decision Tree:**
   - Decision Tree delivered consistent performance, achieving an F1-Score of 0.727 with recall and precision both at 0.667 and 0.800, respectively. Its hyperparameter tuning (e.g., max_depth and min_samples_split) and ability to capture non-linear relationships made it well-suited for the task.
3. **Random Forest:**
   - Random Forest performed well after threshold adjustment, achieving the highest recall (1.000) and an F1-Score of 0.800. However, its slightly lower precision (0.667) highlights the trade-off between minimizing false positives and capturing all GOTY nominations.
4. **Gradient Boosting:**
   - Gradient Boosting matched the Decision Tree in F1-Score (0.727) and precision (0.800) while balancing recall (0.667). Its sequential learning and ability to handle

imbalanced datasets proved effective, making it a strong contender with potential for further optimization.

Key Observations

- **Threshold Adjustment Impact:**
    - Adjusting thresholds improved recall across all models, addressing the issue of missed GOTY nominations in earlier results.
- **Random Forest Strengths:**
    - Random Forest excelled at recall, ensuring all GOTY nominations were identified, though this came at the cost of precision.
- **Gradient Boosting Potential:**
    - Gradient Boosting demonstrated versatility and promise for further fine-tuning, balancing precision and recall effectively.
- **Logistic Regression Baseline:**
    - While not optimized for the dataset's non-linear structure, Logistic Regression provided valuable insights as a baseline model.

Conclusions

The refined methodology and threshold adjustments highlighted the strengths of tree-based models (Decision Tree and Gradient Boosting) for predicting GOTY nominations. These models effectively handled non-linear relationships and feature interactions, making them better suited for the task. Random Forest demonstrated exceptional recall, while Gradient Boosting achieved balanced performance across metrics.

Future improvements could include ensemble approaches that combine these models' strengths or further optimization of hyperparameters for Gradient Boosting. Aligning model selection with dataset complexity remains critical for enhancing predictive accuracy.

References

Akhmedov, K., & Phan, A. H. (2021). Machine learning models for DOTA 2 outcomes prediction. arXiv preprint arXiv:2106.01782. Retrieved from arxiv.org

Owen, V. E., & Baker, R. S. (2018). Fueling Prediction of Player Decisions: Foundations of Feature Engineering for Optimized Behavior Modeling in Serious Games. Technology, Knowledge and Learning, 25, 225–250. Retrieved from link.springer.com

Chen, Z., Sun, Y., El-Nasr, M. S., & Nguyen, T. H. D. (2017). Player skill decomposition in multiplayer online battle arenas. arXiv preprint arXiv:1702.06253. Retrieved from arxiv.org

Predicting video game sales using Machine Learning in Python. (n.d.). CodeSpeedy. Retrieved from codespeedy.com

Predicting Video Game Sales Based on Machine Learning and Hybrid Feature Selection Method. (n.d.). IEEE Xplore. Retrieved from ieeexplore.ieee.org

Pfau, J., Debus, M., Juul, J., Hammar, E. L., Canossa, A., & Seif El-Nasr, M. (2022). *Predicting Success Factors of Video Game Titles and Companies*. In B. Göbl, *ICEC 2022, LNCS 13477*, pp. 269–282. Springer. https://doi.org/10.1007/978-3-031-20212-4_22

Data-Mining-for-Video-Game-Sales-Prediction-Using-Decision-Tree-SVM-and-KNN-Algorithms. (n.d.). GitHub. Retrieved from github.com