WWE Superstar Popularity Tier Prediction using Multi-class Classification

Jay Arre Talosig
College of Computing and Information Technologies
National University - Philippines
Manila, Philippines
talosigjp@students.national-u.edu.ph

Moira Mercado
College of Computing and Information Technologies
National University - Philippines
Manila, Philippines
mercadomn@students.national-u.edu.ph

Queen Maegan Pedido
College of Computing and Information Technologies
National University - Philippines
Manila, Philippines
pedidoqr@students.national-u.edu.ph

James Adrian Castro
College of Computing and Information Technologies
National University - Philippines
Manila, Philippines
castrojb@students.national-u.edu.ph

***Abstract*—The world of World Wrestling Entertainment (WWE) features a diverse roster of superstars, categorized by their prominence and fan engagement into tiers such as Main Eventer, Midcard, and Enhancement Talent. Predicting these tiers is crucial for talent management, storyline development, and marketing strategies. This research presents a comprehensive machine learning system to classify WWE superstars into popularity tiers based on performance metrics and career statistics. We implemented and evaluated several models, including Random Forest, Support Vector Machine (SVM), and Gradient Boosting. The tuned Random Forest classifier demonstrated superior performance, achieving a test accuracy of 70.27%, outperforming other models. Key predictive features identified were the number of championship reigns and matches. This study establishes a data-driven framework for understanding and predicting superstar positioning within the WWE ecosystem, while highlighting the challenges of working with limited and imbalanced sports entertainment data.**

***Index Terms*—WWE, popularity prediction, multi-class classification, machine learning, Random Forest, sports analytics**

I. INTRODUCTION

World Wrestling Entertainment (WWE) operates not just as a sports organization but as a global entertainment powerhouse. Its roster of superstars is strategically positioned in different tiers of popularity, which directly influences their screen time, storyline significance, and commercial value. These tiers are informally classified by fans and analysts as **Main Eventer** (top-tier stars who headline major events), **Midcard** (reliable performers who are consistently featured but not in the main event), and **Enhancement Talent** (superstars used primarily to make others look strong).

Currently, the classification of a superstar into one of these tiers is subjective, based on fan perception and internal company booking decisions. A data-driven approach to predict these tiers can provide valuable insights for various stakeholders: WWE can optimize talent investment, media analysts can quantify a star's draw, and fans can gain a deeper understanding of the business.

This study aims to automate this classification process by leveraging machine learning. The primary objectives are:

- To collect and preprocess a dataset of WWE superstars with relevant features such as match counts, championship reigns, and win-loss records.
- To train and evaluate multiple machine learning models for the multi-class classification task of predicting popularity tiers.
- To identify the most significant features that contribute to a superstar's tier classification.

II. REVIEW OF RELATED WORKS

The application of data analytics and machine learning in sports, often referred to as "sabermetrics" in baseball or more broadly as sports analytics, has seen widespread adoption for performance prediction and player valuation [1]. However, its application in sports entertainment, particularly professional wrestling, is a nascent field.

In traditional sports, studies have successfully used player statistics to predict outcomes like match results [2] and player performance tiers. Machine learning models such as Random Forests and SVMs are frequently employed for these classification tasks due to their ability to handle non-linear relationships [3].

Within the context of WWE, public data analysis is often conducted by fan communities and journalists, but formal academic research is limited. Studies that exist often focus on the socio-cultural aspects of wrestling fandom rather than quantitative performance analysis. This research gap presents an opportunity to apply established sports analytics methodologies to the unique, storyline-driven world of WWE. By treating superstar metrics as features in a classification model, this work bridges the gap between traditional sports analytics and entertainment analytics.

III. METHODOLOGY

The implementation of this project followed a standard machine learning pipeline, from data acquisition to model deployment.

A. Data Collection

The dataset was manually curated from the online wrestling database, **Cagematch.net**, a comprehensive source for professional wrestling match results and performer statistics. The initial dataset contained information for 185 WWE superstars after preprocessing. The features collected for each superstar included:

- **Matches:** Total number of matches in the superstar's career.
- **Wins / Losses / Draws:** Count of match outcomes.
- **Championships:** Total number of championship reigns.
- **Years Active:** The number of years the superstar has been performing.

The target variable, the **Popularity Tier**, was labeled based on a combination of fan consensus and observed booking patterns into three classes: **Enhancement Talent (0)**, **Midcard (1)**, and **Main Eventer (2)**.

B. Data Preprocessing

To ensure the quality and suitability of the data for model training, a series of preprocessing steps were undertaken.

Initial Sanity Check and Cleaning The dataset was inspected for null and duplicated values. Any null or duplicated entries were identified and subsequently removed to ensure data integrity.

Feature Engineering New features were created to enhance the model's predictive power:

- **Win Percentage:** Calculated as Wins / Matches.
- **Loss Percentage:** Calculated as Losses / Matches.
- **Draw Percentage:** Calculated as Draws / Matches.

Data Splitting and Feature Scaling The cleaned dataset was split into training and testing sets using an 80-20 ratio, ensuring stratified sampling. The features were standardized using `StandardScaler` for models sensitive to feature scales.

C. Experimental Setup

The entire project was developed in a Jupyter Notebook environment using Python. Key libraries included `pandas` for data manipulation, `scikit-learn` for machine learning models and preprocessing, and `matplotlib` and `seaborn` for visualization.

D. Training Procedure

The baseline models were first trained and evaluated using default parameters. Subsequently, hyperparameter tuning was performed for the most promising models using **GridSearchCV** with 5-fold cross-validation to find optimal hyperparameters and ensure model generalizability.

E. Evaluation Metrics

The models were evaluated using comprehensive metrics:

- **Accuracy:** The proportion of total correct predictions.
- **Precision:** The ability to avoid false positives.
- **Recall:** The ability to find all positive samples.
- **F1-Score:** The harmonic mean of precision and recall.

IV. ALGORITHM

Six classification algorithms were selected for this study:

A. Random Forest

An ensemble learning method that constructs multiple decision trees and outputs the mode of their classes. It is robust against overfitting and effective for high-dimensional data.

B. Support Vector Machine (SVM)

A discriminative classifier defined by a separating hyperplane. The kernel trick was used with Radial Basis Function (RBF) kernel to handle non-linear decision boundaries.

C. Gradient Boosting

An ensemble technique that builds trees sequentially, with each tree correcting the errors of the previous one.

D. K-Nearest Neighbors (KNN)

An instance-based learning algorithm that classifies data points based on neighbor classifications.

E. Logistic Regression

A linear model for classification that uses a logistic function to model class probabilities.

F. Decision Tree

A non-parametric supervised learning method that learns simple decision rules from data features.

V. RESULTS AND DISCUSSION

A. Exploratory Data Analysis

Initial EDA revealed class imbalance with Midcard tier being most populous (103 superstars), followed by Main Eventer (43) and Enhancement Talent (39). This imbalance presented classification challenges.

B. Baseline Model Performance

As shown in Table I, Random Forest and K-Nearest Neighbors achieved the highest baseline accuracy of 67.57%. Logistic Regression and Decision Tree performed poorest.

C. Hyperparameter Tuning Results

Hyperparameter tuning was conducted for top-performing models. The tuned Random Forest model (RF_Tuned) achieved the highest performance with 70.27% accuracy, representing a 2.7 percentage point improvement over baseline.

TABLE I
PERFORMANCE OF BASELINE MODELS ON TEST SET

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.6757	0.6543	0.6757	0.6535
K-Nearest Neighbors	0.6757	0.6847	0.6757	0.6319
Gradient Boosting	0.6486	0.6327	0.6486	0.6290
Support Vector Machine	0.6486	0.5277	0.6486	0.5614
Logistic Regression	0.6216	0.6181	0.6216	0.5661
Decision Tree	0.5405	0.5718	0.5405	0.5477

TABLE II
PERFORMANCE OF TUNED MODELS ON TEST SET

Model	Accuracy	Precision	Recall	F1-Score
RF_Tuned	**0.7027**	**0.6980**	**0.7027**	**0.6721**
SVM_Tuned	0.6757	0.7554	0.6757	0.6177
GB_Tuned	0.6216	0.6032	0.6216	0.6067

D. Feature Importance Analysis

Analysis revealed **Championships** as the most significant predictor, followed by **Matches** and **Win Percentage**, aligning with wrestling logic where top-tier stars receive more opportunities.

E. Discussion of Results

The results demonstrate feasibility while highlighting challenges. The 70.27% accuracy indicates meaningful patterns were captured, though significant predictive challenges remain due to:

- **Class Imbalance:** Midcarders outnumbering other tiers affected minority class identification
- **Intangible Factors:** Missing qualitative elements like charisma, mic skills, and social media presence
- **Subjectivity:** Creative decisions in wrestling that defy pure statistics

The moderate performance range (54-70%) reflects the unique challenges of sports entertainment compared to traditional sports analytics. The dominance of championship and match counts confirms their importance, but also suggests the need for more sophisticated features in future work.

VI. CONCLUSION

This study successfully developed a machine learning system for classifying WWE superstars into popularity tiers. The tuned Random Forest model achieved the best performance with 70.27% accuracy, demonstrating that quantitative career statistics provide meaningful indicators of superstar positioning.

Limitations include the small dataset size (185 superstars), class imbalance, and absence of important qualitative factors. The manual labeling process also introduced some subjectivity.

Future work will focus on:

- Expanding the dataset with more wrestlers from different eras
- Incorporating NLP techniques for social media sentiment analysis
- Collecting additional features like pay-per-view main events and merchandise sales
- Applying advanced techniques for handling class imbalance
- Exploring deep learning models with more extensive data

This work establishes a foundation for data science in professional wrestling analytics and demonstrates both the potential and limitations of quantitative approaches in sports entertainment.

REFERENCES

[1] M. E. Glickman and H. S. Stern, "Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers," Columbia University Press, 2014.

[2] T. Ueda and A. Murai, "A machine learning approach for predicting the outcome of English Premier League matches," PLOS ONE, vol. 15, no. 1, 2020.

[3] C. Sheehan and Y. Xu, "Predicting player performance in professional Dota 2," in Proceedings of the MIT Sloan Sports Analytics Conference, 2020.