Journal of Multimedia Information System VOL. 11, NO. 2, June 2024 (pp. 167-172): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2024.11.2.167

Demonstrating the Power of SHAP Values in AI-Driven Classification of Marvel Characters

Ho-Woong Choi[1*], Sardor Abdirayimov[2]

Abstract

The transparency and interpretability of machine learning models have become paramount in the era of Explainable AI (XAI). This study leverages SHAP (SHapley Additive exPlanations) values to elucidate the decision-making process of an XGBoost classifier trained to distinguish between 'good' and 'bad' Marvel characters based on their skill sets. This study highlights the nuanced interpretability SHAP values provide, bridging the gap between complex AI models and the subjective domains of storytelling and character development. It underscores the need for balanced datasets and careful model training to mitigate inherent biases. These findings contribute to the field of XAI, demonstrating the potential of SHAP values in complex classification scenarios and underscoring their role in advancing AI transparency and trustworthiness.

Key Words: XAI, SHAP, Model Interpretability, Character Classification.

I. INTRODUCTION

Marvel Studio is one of the dominant players of film industry. The evolution of Marvel Cinematic Universe (MCU) has arisen from superhero comics from 1960-1970s. The idea of bringing superheroes into single fictional universe was Marvel's empire-building model [1]. However, the release of "Guardians of the Galaxy (2014) indicated the slight change in the business model of Marvel studio [2]. Subsequently, studio has expanded its focus on creating new characters. Updated business model has brought the MCU to create phenomenal series of "Avengers". As a result, "Avengers: End Game (2019)" made a colossal impact on the global box office [3]. However, the primary plot of the MCU afterwards became more disorganized, incoherent, and complex. Haoran Qiu argues that The Phase Four of the Marvel Cinematic Universe features far from too much of same formulaic cinematography, which gradually reduce the interest of audience [4].

The concept of Shapley Additive exPlanations (SHAP) values, introduced by Scott Lundberg and Su-In Lee in 2017, epitomizes this pivot. SHAP values offers a coherent and individualized approach to elucidate the contribution of each feature in a models' prediction, akin to examining the motives and actions of each character in the MCU narrative. This study leverages the analogy of the MCU, utilizing SHAP values to classify Marvel characters into the archetypes of 'good' and 'evil' based on a range of attributes and skills. This unique application not only serves as an accessible introduction to the versatility of SHAP values but also allows for a deeper exploration of complex traits and moral alignments within a familiar context.

Our contribution through this novel intersection of AI interpretability and pop culture is twofold: 1) We provide a methodological bridge between complex AI systems and the general public, and 2) we demonstrate the utility of SHAP values in unpacking the layered elements of character-driven narratives. Furthermore, by analyzing the classification of Marvel characters with SHAP values, we offer readers a compelling glimpse into the power of explainable AI, reinforcing its significance in an increasingly automated world.

The paper progress from a review of the relevant background, through the methodology and analysis of SHAP values, to a discussion of the results, concluding with the study's limitation and potential direction for future research.

II. BACKGROUND AND RELATED WORK

SHAP values, inspired by Lloyd Shapley's cooperative

Manuscript received February 26, 2024; Revised April 02, 2024; Accepted April 27, 2024. (ID No. JMIS-24M-02-003)
Corresponding Author (*): Ho-Woong Choi, +82-10-8972-0517, techimpress@naver.com
[1]Department of Media Software, Sungkyul University, Anyang, Korea, techimpress@naver.com
[2]Department of AI and Big Data, Woosong University, Daejeon, Korea, 202112112@live.wsu.ac.kr

game theory, have transformed the way we interpret machine learning models [5]. KernelSHAP and TreeSHAP stand out in the SHAP framework for their efficiency in computing SHAP values but are part of a larger array of strategies that apply game theory principles to machine learning [6]. These techniques are particularly notable for their computational efficiency, enabling the practical application of SHAP values in various machine learning models, from simpler linear regressions to complex ensemble methods.

SHAP values have been used extensively in fields where critical decisions are made. In medical field, SHAP values help elucidate why AI models identify certain conditions, like cancer, by pinpointing influential features in predictions [7]. While SHAP values have been primarily utilized in critical areas, their application in character classification, particularly in a narrative or fictional context like the Marvel Universe, remains less explored. Current work is conducted to fill the gap by analyzing Marvel characters' classifications based on their skill sets, utilizing SHAP values to interpret these classifications.

III. METHODOLOGY

To facilitate an inclusive understanding of our experimental process, we have delineated the methodology through a structured flow chart illustrated in Fig. 1. As for data collection, we opted for a publicly available dataset, which includes Marvel characters and their skills, along with their alignment as 'good' or 'bad' in the movies [8]. Boolean features were converted to binary (0 and 1) for analysis. Addressing missing values was a significant step in data preprocessing. We postulated that the absence of information likely indicates a lack of skill, hence missing values were replaced with 0.

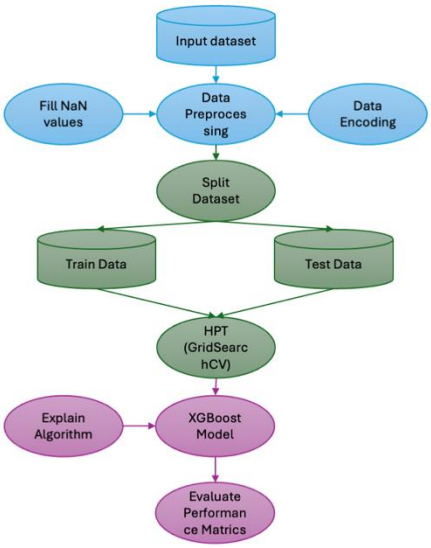

Fig. 1. Flowchart of our experiment.

The final composition of the dataset revealed a predominance of ‘good’ characters, constituting 68% of the data, with ‘bad’ characters making up the remaining 32%. We have employed SMOTE technique to overcome the unbalancing in the dataset [9].

For the classification task, the XGBoost Classifier was our algorithm of choice, elected for its well-documented robustness, computational efficiency, and the convenient tuning of its hyperparameters. XGBoost has demonstrated exemplary performance across a spectrum of classification challenges, adeptly managing various data types and intrinsic handling of missing values [10]. Hyperparameter optimization was conducted via GridSearchCV, which systematically worked through a range of combinations to pinpoint the optimal parameters, enhancing the model’s ability to generalize. We decided to allocate 80% of dataset for training purposes and reserving 20% to evaluate the model’s performance, respectively. The GridSearchCV process concluded with the identification of the most effective hyperparameters for our XGBoost model: a learning rate of 0.1, a max_depth of 5, n_estimators set to 300, and a colsample_bytree of 0.3. There parameters yielded the highest test score of 0.78.

The model effectively predicted true positives and true negatives, but it struggled with misclassification of false negatives. This outcome highlights the need to understand model limitations when interpreting SHAP values. The accuracy and reliability of SHAP values are contingent upon the overall performance and tuning of the underlying model.

IV. SHAP VALUES ANALYSIS

The computation of SHAP values was conducted using the open-source Python library SHAP, which is designed for explainability in machine learning [11]. The Tree Explainer module, a component of the SHAP library specifically optimized for tree-based models like XGBoost, was utilized to calculate the SHAP values for each prediction.

In Fig. 2, the SHAP summary plot for a true positive prediction reveals a contrasting scenario. The feature ‘Jump’, when present, is the most influential in swaying the model towards a true positive prediction, indicative of a ‘heroic’ alignment. ‘Invulnerability’ and ‘Accelerated Healing’ also contribute positively to the model’s confidence in making a true positive classification. Interestingly, the absence of features like ‘Super Strength’ negatively influences this outcome.

Fig. 3 offers a SHAP value summary plot detailing the model’s rationale when predicting a true negative outcome, which, in the context of the study, may refer to the classifi-

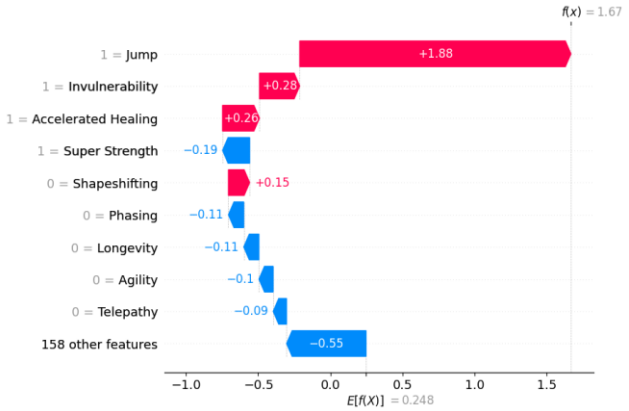

Fig. 2. SHAP summary: impact of features on a true positive classification with jump as the main positive driver.

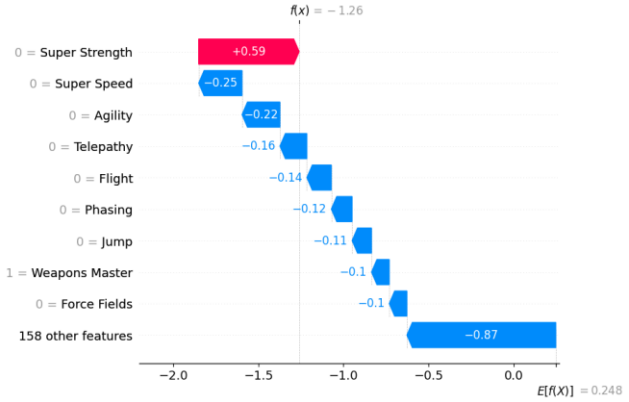

Fig. 3. SHAP value plot for true negative prediction highlighting the predominant influence of lacking super strength.

cation of a character as a ‘villain’ or possessing a negative trait. The plot explains that the absence of ‘Super Strength’ is the most significant positive driver for such a prediction, as evidenced by its high positive SHAP value. The features ‘Super Speed’, ‘Agility’, ‘Telepathy’, among others, when absent, also tend to influence the model toward a true negative classification, albeit to a lesser extent. Conversely, the presence of ‘Weapons Master’ has a slight negative impact on the prediction. Collectively, these feature effects illuminate the intricate interplay between various superhero attributes in shaping the classification decision.

These examples underscore the value of SHAP in model interpretation, revealing not just the features that influence predictions, but also highlighting the need for careful consideration of the model's baseline tendency. SHAP's detailed, instance-level explanations provide insights that are critical for understanding, validating, and improving the model's decision-making process.

V. RESULTS AND DISCUSSION

The application of the XGBoost classifier, supplemented by SHAP value analysis, yielded notable insights into the classification of Marvel characters as 'good' or 'bad'. The model achieved an accuracy of 74%, with precision, recall, and F1 scores reflecting a reasonable predictive performance given the complexity of the task.

The true positive predictions highlighted an interesting trend: characters with abilities like ‘jump’ and ‘invulnerability’ were most classified as 'good'. These findings align with common superhero tropes where such abilities are emblematic of heroism. Conversely, the analysis of false negatives revealed that the model occasionally misclassified 'bad' characters as 'good'. Notably, skills typically associated with villainy, such as 'Super Strength' and 'Energy Blasts', surprisingly contributed negatively to these misclassifications, suggesting that the model might have learned an unintended bias from the training data.

Based on the confusion matrix presented in Fig. 4, true negative predictions were less frequent, indicating a skew towards 'good' character classifications. This skewness in the model's predictions underscores the need for a more balanced dataset that encompasses a wider array of 'bad' character traits. The false positives were minimal, indicating that the model was generally successful at identifying 'bad' characters when the data presented a clear set of villainous traits.

The SHAP value analysis provided a granular view of the model's decision-making process, revealing that certain abilities have a stronger influence on character alignment predictions than others. This level of interpretability is crucial for understanding how the model processes information and which features it deems most important. The insights from SHAP analysis extend beyond model performance, suggesting a deeper narrative structure within the data. The analysis suggests that certain character abilities are cultur-

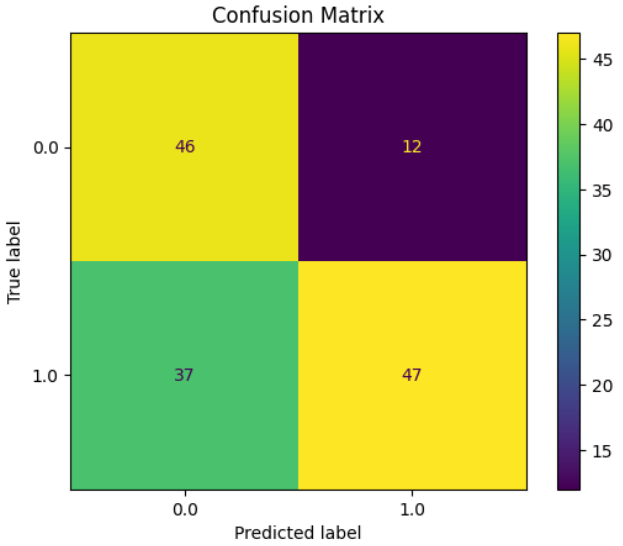

Fig. 4. Confusion matrix with 47 true positives and 41 true negatives.

ally and narratively associated with moral alignments in the Marvel Universe. These insights could be highly influential for the advancement of AI in narrative analysis and entertainment, guiding content creators in understanding audience perceptions of character traits.

We have selected Python 3.10 as the development language for machine learning. Anaconda environment was chosen for its flexible package administration [12]. The experiment was conducted on a 14-inch MacBook Pro with Apple M1 Pro chip and 16 GB RAM memory. The operating system is running on macOS. To facilitate programming, Visual Studio Code is used as the development environment. The code supporting the findings of this study is available upon request.

VI. LIMITATION AND FURTHER RESEARCH

The study's limitations include the limit size of dataset and the absence of new Marvel characters like Ms. Marvel. Another limitation is the potential overrepresentation of 'good' character traits in the dataset. Additionally, potential biases in the AI model's decision-making process could be assessed and mitigated in subsequent research. Exploring the effects of algorithmic biases on character classification could ensure a fairer representation of diverse character traits. While we employed SHAP tool, further research could use other interpretability tools, such as Local Interpretable Model-Agnostic Explanations (LIMA) and Shapley Interaction Quantification (SHAP-IQ) [13-14].

VII. CONCLUSION

To sum up, this research addresses the classification of marvel cinematic universe characters and interpretability of such classification. We employed XGBoost decision tree model on the preprocessed dataset of Marvel characters. This research contributes to the growing body of work in explainable AI, demonstrating the utility of SHAP values in interpreting complex classification models. The findings provide a foundation for further exploration into the character development, potentially guiding the creation of unique and resonant superhero personas. This work not only highlights the relevance of AI in media and the entertainment but also its role in enhancing our comprehension of complex storytelling elements.

REFERENCES

[1] S. Graves, "The marvel studios phenomenon: Inside a transmedia universe, Eds. Martin Flanagan et al. Bloomsbury, 2016. 268 pp. $120.00 cloth," *Popular Culture*, vol. 51, no. 3, pp. 812-814, Jun. 2018.

[2] Wikipedia, Guardians of the Galaxy, https://en.wikipedia.org/wiki/Guardians_of_the_Galaxy_(film), 2014.

[3] S. Mittermeier, "Avengers: Endgame," *Science Fiction Film and Television,* vol. 14, no. 3, pp. 423-429, 2021.

[4] H. Qiu, "Using content analysis to analyze issues in the development of the Marvel Cinematic Universe and their impacts," in *SHS Web of Conferences*, EDP Sciences, 2024, vol. 181, p. 04003.

[5] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 30, 2017.

[6] C. Molnar, Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/, 2020.

[7] J. Vrdoljak, Z. Boban, D. Barić, D. Šegvić, M. Kumrić, and M. Avirović, et al., "Applying explainable machine learning models for detection of breast cancer lymph node metastasis in patients eligible for neoadjuvant treatment," *Cancers*, vol. 15, no. 3, p. 634, 2023.

[8] R. Danniel, "Marvel Superheroes Stats and Info," https://www.kaggle.com/datasets/dannielr/marvel-superheroes?select=superheroes_power_matrix.csv, 2018.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[11] SHAP, SHapley Additive exPlanations, https://github.com/slundberg/shap.

[12] Anaconda, Distribution, https://www.anaconda.com/distribution/, 2024.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, " Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and data Mining*, 2016, pp. 1135-1144.

[14] F. Fumagalli, M. Muschalik, P. Kolpaczki, E. Hüllermeier, and B. Hammer, "Shap-iq: Unified approximation of any-order shapley interactions," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

Journal of Multimedia Information System VOL. 11, NO. 2, June 2024 (pp. 167-172): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2024.11.2.167

AUTHORS

Ho-Woong Choi received his B.S., M.S., and Ph.D. degrees from Seoul National University, Korea, in 1993, 1995, and 2003, respectively. From 2002 to 2016, he worked at Hyundai Heavy Industries Research Institute in Korea, where he was a principal researcher. At AIBB lab, he conducted research on NLP related to digital marketing and supply chain. Since September 2020, he has been working as an Industry-Academia Collaboration professor in the Department of Media Software at Sungkyul University. For over three decades, he has been a heavy user of Wolfram Mathematica and has participated in Wolfram Conferences as an invited speaker seven times. He has published three e-books on ChatGPT and a book on Wolfram Language sponsored by Wolfram Research.

Sardor Abdirayimov is an undergraduate student at Woosong University, where he is pursuing a B.S degree of AI and Big Data with a keen focus on the fields of computer vision and large language models. His research interests include computer vision, natural language processing (NLP) and MLOps. Sardor is passioned about studying the overlap of these domains to enhance the capabilities of artificial intelligence.