Character Name-Matching and Thematic Analysis

Our initial approach attempted to connect movies to comic themes by matching character names between the Marvel_characters database and comic descriptions. The hypothesis was straightforward: themes present in comics featuring specific characters would influence the success of movies featuring those same characters. We implemented logistic regression for this binary classification task, as it provides interpretable coefficients while handling the dichotomous outcome of high versus low success.

The model achieved a Kappa score of -0.14, indicating performance worse than random chance. This failure revealed several critical limitations in our approach. First, character name variations, aliases, and inconsistent descriptions created substantial matching errors. Second, the indirect pathway from character appearances to comic themes to movie success introduced multiple points of potential failure. Most significantly, this method assumed that any comic featuring a character would influence all movies featuring that character, disregarding the selective nature of comic adaptation where specific storylines, not entire character histories, inspire films. The complete absence of cultural themes in our results (0 movies) further suggested our matching methodology was fundamentally flawed.

Direct Comic-to-Movie Connections

Learning from the indirect nature of our first approach, we pivoted to utilizing the comic_to_movie table, which explicitly mapped source comics to their film adaptations. This direct connection should have eliminated the ambiguity inherent in character-based matching. However, exploration of this table revealed a significant constraint: only 13 of the 48 MCU properties had populated comic connections, severely limiting our analytical power. This limitation stems from the complex nature of comic adaptation. MCU films often loosely draw from dozens or even hundreds of different comic issues, making definitive source attribution challenging. Marvel Studios does not publicly disclose its source material, and while fan blogs attempt to document these connections, none provide comprehensive or verifiable data.

Given the reduced sample size, we implemented the theme-only approach by using Leave-One-Out Cross-Validation to maximize our training data for each fold. The initial logistic regression model using only comic themes (mentor, comedy, romance, origin story) achieved a Kappa of 0.03, marginally better than random classification. Hypothesizing that decision trees might better capture non-linear interactions between themes, we implemented CART models, which showed no improvement (Kappa = 0). Among individual themes, only romance showed any promise when tested in isolation.

We then pivoted to the Enhanced Feature Set. Despite the small sample size, we expanded our feature set to include character composition and team dynamics. We engineered features that capture power diversity (unique power types per movie), team affiliations (particularly the presence of the Avengers), and total character count. This enhanced model achieved a Kappa

of 0.28, representing fair agreement and our best performance across all movie success prediction attempts.

The dramatic improvement from the first model to the second revealed an important insight: structural elements of ensemble composition appeared more predictive than thematic content. Exploratory analysis confirmed this pattern; successful films averaged 27.9 characters compared to 10.3 for less successful films. This finding suggested that MCU success might be driven more by scale and interconnectedness than by fidelity to specific comic themes. However, with only 13 observations, we remained cautious about overfitting, acknowledging that our model's improved performance might not generalize to the broader MCU catalog.