What Makes a Superhero Marketable?

Kevin Angotti

DSC 550, Data Mining

7 August 220

A Superhero Assessment for Future Adaptation

**Introduction**

When looking at the current outlook of the superhero landscape on the big screen, the Marvel Cinematic Universe has a stronghold on the competition, and one company alone controls the scene. Disney, a large production company, has capitalized on its success in adapting comic book characters to live-action movies. In an attempt to provide competition to Disney, Fox has taken their approach with their Detective Comics (DC) brand for live-action movies, though they have not capitalized on the success that Disney has had. Then there is the television and animated series that DC has had success in where Disney has struggled. Keeping that success going is a challenge, and companies like Netflix need to keep adapting to changes in the comic book scene to stay o top and keep their subscribers.

Disney has made advances in its streaming platform to allow for them to bridge the gap in TV and animated series. So, how can Netflix stay on top? By keeping the long run of popular TV series going. With the show Arrow ending, they need to find the next big hero to keep their content enjoyable and subscriber accounts going. I have proposed a way to dig through a list of superheroes that can be adapted to a television series based on keywords found in their backstories (history) and powers. This allows a company to make character decisions based on what makes them appealing to not only comic book fans, but also the average Netflix subscriber.

As Disney keeps producing Marvel movies for everyone's enjoyment, DC has had their hands full, trying to get it right. Their only success has been on the TV series side of things. However, Netflix has a few of their comic book TV series either coming to an end, or just recently ended, and a need to find the next show should be on the forefront. Having a tool that a company can use to filter out potential superheroes that an audience will follow and give a company like Netflix a step up on the competition and at the same time provide a potential show that will result in fewer failure rates.

The project asks will ask the question, can a tool such as a decision tree be created to choose a character from an extensive list of superheroes that can, in turn, be adapted into a television show or movie? If I can answer this question, any potential company to the likes of Netflix could use this tool for all their future projects pertaining to this type of content.

**Data**

The data for this project can be found on the Kaggle website. This data file I chose consists of over 1400 different types of superheroes. Most comprise of Marvel and DC superheroes; however, there are also ample amounts of from other bases or companies. The data in itself did not need much altering or cleaning outside of the dropping of NA or NaN values.

I chose this data set because it contained two columns of information that related to both a hero's history or backstory and their powers. Both of these fields will allow me to look at a hero differently than most data on a superhero does. I want to see if the history and powers of a hero can help determine if that hero will be popular enough to be marketable.

There were a few columns that had been dropped based on values that did not pertain to the information needed. For instance, a column which captured a link for a picture of the hero, or a column that identified skin color was not needed for analysis. If a row did not contain information on a hero's powers and or history, that row was also dropped. The resulting dataframe still contained over 900 superheroes for analysis. The author of this data set highlights a group of variables that relates to a hero(s), the "power sets" are what are considered the main attributes of a hero. This data set does provide an overall score, though; I will not be using this in my analysis outside of locating, which attributes a hero has or is considered the most important. The reason for this is that the overall score is derived from fan voting on the Superhero Database website. This allows fans to rank heroes based on their liking of said hero, thus not an accurate overall score. I created a variable called attribute score that sums up the attributes ad divides them by six or the number of attributes; this provides an average attribute score for better representation. The data set also contains Boolean data on the particular information that could be helpful if I can determine which traits are most important.

**Methodology**

I first took the overall score and plugged it into a linear regression model to determine which of the power set attributes had the most significance to that score. The resulting table, *Table 1*, shows that intelligence, speed, and strength are the most desirable traits a hero should possess. The power score attribute also had significance, but the t-value was not as high as the other three. In terms of significance, the further away from zero, these three attributes are with their t-values, the more they are considered desirable traits. With a big F statistic score, the linear model proves to be an excellent model to measure the overall score.

Table 1 Linear Regression Model for Overall Score

|  |
| --- |
| lm(formula = overall\_score ~ intelligence\_score + speed\_score + |
| strength\_score + combat\_score + durability\_score + power\_score, |
| data = Super\_Hero\_Dataset) |
|  |
| Residuals: |
| Min 1Q Median 3Q Max |
| -22.405 -8.357 -1.925 3.327 197.679 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) -35.61899 4.09468 -8.699 < 2e-16 \*\*\* |
| intelligence\_score 0.37722 0.04640 8.129 1.01e-15 \*\*\* |
| speed\_score 0.17526 0.02795 6.270 4.93e-10 \*\*\* |
| strength\_score 0.14575 0.02490 5.852 6.14e-09 \*\*\* |
| combat\_score -0.06186 0.02484 -2.490 0.012882 \* |
| durability\_score 0.02339 0.03065 0.763 0.445552 |
| power\_score 0.08965 0.02325 3.856 0.000121 \*\*\* |
| --- |
| Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 |
|  |
| Residual standard error: 18.9 on 1289 degrees of freedom |
| (66 observations deleted due to missingness) |
| Multiple R-squared: 0.2688, Adjusted R-squared: 0.2654 |
| F-statistic: 78.97 on 6 and 1289 DF, p-value: < 2.2e-16 |

A close up of a logo

Description automatically generated

Figure 1 Speed Rankings of Superheroes

You can see in figure one that speed is only ranked high for around 200 of the 944 superheroes in the data set. That means that most of these heroes do not have speed as one of their main power attributes. We know from the linear regression model that speed is a desired trait of a popular superhero.

A screenshot of a cell phone

Description automatically generated

Figure 2 Strength Rankings of Superheroes

Figure two, much more precise then figure 1, shows that superhero strength is only ranked high, just over 100 of the 944 superheroes. This makes the attribute, in a way, a highly coveted attribute in terms of who is genuinely a strong superhero.

A screenshot of a cell phone

Description automatically generated

Figure 3 Intelligence Rankings of Superheroes

Figure 3 represents the distribution of intelligence for the data set. From this figure, you can see that most of the superheroes retain the right level of intelligence. Though not as scares of an attribute, it is still considered in the linear model as a necessary trait. All three of these figures help show us that even if a hero is considered super, they might not have the attributes that are considered popular to an audience.

A screenshot of a cell phone

Description automatically generated

Figure 4 Attribute Score Rankings of Superheroes

In figure four, you can see that the distribution almost follows the normal distribution of a bell curve, providing a suitable data representation for analysis. Most of the superheroes attribute scores fall in the 300 to 400 range, making all possible filtered heroes relatively powerful.

The next step I took after identifying the key attributes that a hero needs to have; I looked to use those values and compare them to the history and power text variables from the data set. To increase keyword identification, I also looked at the superhero and villains website that displays vital terms that a hero or villain has, in terms of their common traits. Besides intelligence, speed, and strength, heroes also possess traits like skilled, endurance, fighting/martial arts, expertise, stealth, and superhuman strength.

From there, I began setting up the Natural Language Process, or NLP to locate some or all of these keywords from the history and powers textual information. I started by importing the nltk library in Python to create a dictionary from the two variables, history text, and powers text. I removed all punctuation from the data as well as stop words lie to, the, is, at, etc. this leaves just the text that mattered for analysis. I then used the frequency counter to count the number of times a particular word was returned in the results. This provided me a list to compare the keywords, too, and allowed me to choose which of the boolean data the data set also contains to target for the decision tree model. The TF IDF vectorizer function in Python allowed me to convert this textual data into a numbered vector to calculate the sparsity score of the dataframe. The results ended up being mostly sparse, with 99% and 100% in the powers and history text variables. That calculation tells me that the libraries for each are quite extensive. Why might that be important, a vast library means there is much text to sift through looking for those key attributes. This can either give us several choices or very few. The goal here is to have returned values that can be implemented into a decision tree to filter down the extensive list of heroes to one that provides a company with options.

**Results**

A decision tree will be the main factor in determining which hero(s) to target for potential marketing. The decision tree will be set up from the results of the NLP library. In conducting the NLP methodology, I was able to determine that key terms such as superhuman strength, energy absorption, stealth, durability, and intelligence were prevalent. To create the model, I used the attribute score, the six power attributes, and has energy absorption. The has energy absorption was chosen as the target variable; the sklearn package for train test split was used for the model, and a max depth of 4 was chosen.

The resulting tree produced an accuracy score of 85%, which you can see from *Table 2* below.

Table 2 Decision Tree Confusion Matrix

|  |
| --- |
| [[812 39] |
| [113 36]] |
| precision recall f1-score support |
|  |
| 0 0.88 0.95 0.91 851 |
| 1 0.48 0.24 0.32 149 |
|  |
| accuracy 0.85 1000 |
| macro avg 0.68 0.60 0.62 1000 |
| weighted avg 0.82 0.85 0.83 1000 |

The resulting tree *figure 5* will show the outcome and how a company can use this to filter down a large list of superheroes to one that is more concise to their needs. The model can also be altered, and keywords swapped for other desirable traits like has super strength or has stealth, etc., note this could alter results. The model I configured allows for the best outcome when concerning a hero's marketability or how well received they maybe when portrayed on the big screen. As you can see in *figure 5* below from the train test, set the attribute score has been identified as the root node. The root node is the base of the tree that each of the branches stems off. To move down the tree toward what is known as the leaf node or end of the tree. The answer to a yes, no question must be answered at each node or new branch.

A picture containing screen, computer, large, computer

Description automatically generated

Figure 5 Decision Tree

Looking at the attribute score a litter closer in *Figure 6* shown below, you can see that the second level focuses on combat score to the right and speed score to the left. We know that speed score is a sought-after attribute, so I would answer this question attribute score less than or equal to 477.5 as yes. This brings you to the speed score node in *Figure 7*, where the question, speed score less than or equal to 37.5?

A picture containing food

Description automatically generated

Figure 6 Decision Tree Levels 1 & 2

Looking at the third level with combat score and power score, I would answer this as no for two reasons. One, we want a speed score higher than 37.5 as it is one of the three desirable traits, and the gini score for the third level is higher on the right side.

A picture containing computer

Description automatically generated

Figure 7 Decision Tree Level 3

This brings us to *Figure 8,* where the question is power score less than or equal to 92.5? Both of the nodes in level 4 are desired traits, so we can then look at the gini scores intelligence has a much higher score ad is also had the more significant t value from the linear model so the answer will be no. Thus, moving us to the question is intelligence score less than or equal to 77.5? The answer here should also be no for two reasons intelligence is the highest rated trait, and even though the gini is lower for the right leaf, the sample sizes on that node have 25 of the 29 samples with the target trait of has energy absorption.

A close up of a logo

Description automatically generated

Figure 8 Decision Tree Levels 3, 4, & 5

**Conclusion**

when looking out the way, the decision tree provided the right amount of choices for a company such as Netflix to use when determining their next show based on a superhero character. I believe that with the use of this model, they will have a higher chance of that show succeeding then they would if they just chose a character at random. This model allows for a list of character suggestions based on desired traits from multiple sources the Superhero Database, the superheroes and villains character identifying traits, and the compiled superhero data set. Between all three of these sources, the chances of locating a character's views find appealing is increased substantially.

Loading my linear model into an Anova hypothesis test tells me that I should reject the null hypothesis that this model could not predict a superhero. It is thus allowing me to accept my hypothesis that a tool like a decision tree can be created to choose a character from an extensive list of superheroes. I feel that this model can help a company choose its next adaptation for television or movies.

**References**

1. Besomi, J. (2020). Superheroes NLP Dataset. Kaggle. Retrieved from <https://www.kaggle.com/jonathanbesomi/superheroes-nlp-dataset>

2. Superheroes and Villains Wiki. Fandom. Retrieved from <https://superheroesandvillains.fandom.com/wiki/Common_Traits(Hero)>

3. Superhero. Wikipedia. Retrieved from <https://en.wikipedia.org/wiki/Superhero>