Avengers Death Predictions

By: Kevin Kimmel

1. Introduction

The first issue of the Avengers comic book was released in September of 1963. Since then, there have been nearly 200 members to join the team. Although being on the team has its perks, such as saving the world multiple times, it can get very dangerous as well. This isn't super surprising considering the level of threats the Avengers deal with on a daily basis. But just how dangerous is it? And do certain types of heroes die more or less often than others? These are the kinds of questions that this report will aim to answer. The methods of doing so includes Linear Regression and Decision Trees.

2. Related Work

There has already been some analysis on these topics in the past. FiveThirtyEight has an article titled "Joining the Avengers is as Deadly as Jumping off a Four-Story Building" where the author analyzes how many people have died as an Avenger and compares it to the likelihood of dying when jumping off a building. The comic book world is a little strange with how people die, as many people come back from the dead either because who we thought had died was actually an imposter, the person traveled through time to stay alive, or countless other reasons. Because of this, FiveThirtyEight had to do some carful classifications when it came to what counted as a death and what didn't. In it's final results, they found that 69 people had died at least once out of the 173 members that they analyzed. This is a mortality rate of about 40%. To put this in perspective, they compared this data to results from a case report by Christopher Kepler on how deadly jumping off a building was. Kepler

found that jumping off a five story building had a death rate of about 50%. So joining the Avengers team is comparable to jumping off a five story building. Now it isn't as clear cut as that however, because since these Avengers are dying in a comic book world, they can also come back to life. Out of the 69 Avengers who died at least once, 46 of them came back to life in one way or another. That means if you were to die as an Avenger, you have two-thirds of a chance to be resurrected. The analysis doesn't stop there though however, as some characters died *again* after coming back to life the first time. In fact, FiveThirtyEight counts up to four possible resurrections in their dataset. If we combine all of the deaths into one, we get 89 total deaths. Comparing that to the number of total resurrections, which is 57, then you would have a 64% chance to come back to life after dying as an Avenger. Those are some pretty good odds!

3. Dataset

The main data set for this report is the same one that was used in the FiveThirtyEight report. There are 173 members that were analyzed. Each one has the URL to their Marvel Wiki page, their real name/alias, how many times they have appeared in the comics, whether or not they are a current member of the Avengers, their gender, the date they were introduced, the amount of years since joining the team, whether or not they are an honorary member or a full time member, and then finally the amount of deaths and resurrections that they had.

3.1 Feature Selection

There were not many features to choose from that were actually somewhat informative to do this analysis, so the ones that ended up being chosen were the number of appearances from the characters and the amount of years it has been since they joined the team. Intuitively, it seems that

the more times a character appears, the more likely they are to die, so it will be interesting if the numbers correlate with that intuitiveness. For the Decision Tree, all of the features were used.

4 Methods

There were many data science methods that were considered for this analysis, however after much consideration, Linear Regression and Decision Trees were chosen to be the main methods of analysis for this data.

4.1 Splitting Train and Test sets

The first thing that needed to be done was split the dataset into a training and testing set so that an error rate could be calculated. As there were 173 rows in the dataset, it was decided that the training set would consist of 100 random selections. The training set was then plotted on a graph, color coded depending on whether that character has died or not.

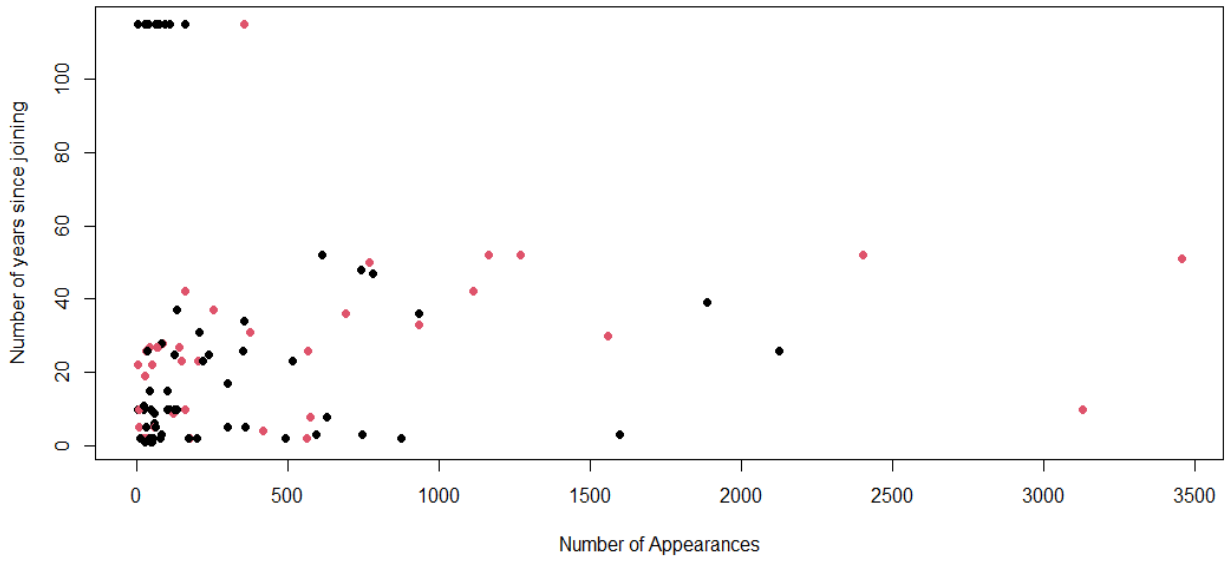

Figure 1: Plot of training set

4.2 Linear Regression

Using linear regression as a tool for analysis, we are able to find if there is any correlation between the chosen features and the classification.

4.3 Decision Tree

A decision tree can be used to determine which features are most important in determining a classification. The features that are closer to the root of the tree are more important while the features that are closer to the leaves are less so.

5 Results

5.1 Linear Regression Results

In order to determine how effective this model was in predicting whether or not a character died, Mean Square Error (MSE) and R^2 were both used. The MSE for this dataset was approximately 0.36 or 36% meaning that the model was incorrect 36% of the time. This is not that great of an error rate, and the R^2 value is also not great coming in at 0.022. This makes sense when looking at the scatter plot of the training set as there wasn't really a visual correlation there and the colors seemed to be all mixed together. This seems to mean that it doesn't really matter how long or short you've been in the Marvel Universe or on the Avengers team, that doesn't necessarily mean you are more or less likely to die.

5.2 Decision Tree Results

Due to the lack of good results from the linear regression model, a decision tree model was used to find what features were most important.

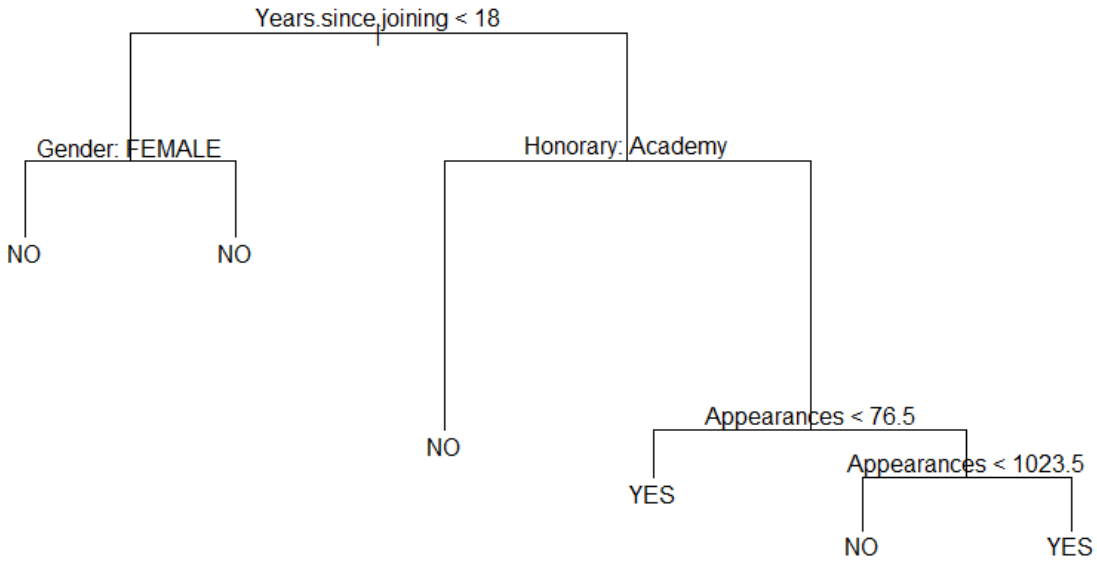

Figure 2: Classification tree

Looking at this decision tree, it appears that the amount of years since joining the team is the most important factor as it is in the root node. If the amount of years is more than 18 then it classifies the character has having not died before. If the amount of years since joining is less than 18 then it looks at the status of the character. In this dataset, a character could have three different statuses, a full member of the team, an honorary member of the team, or a member of the academy. According to this model, if the character was not in the academy it classified it as not having died, and if they were then it looked at the number of appearances to make its decision. After running the testing dataset through this model, the error came out to be about 33%, which is pretty similar to the Linear Regression error.

6. Conclusion

The FiveThirtyEight article brought up some interesting points on how many Avengers have died since the formation of the group. And while it would've been interesting to see if there were any trends in those deaths and to see if they could've been predicted, using the two methods that were used in this report it seems that isn't the case. Both the Linear Regression and the Decision Tree had errors of about 33% which, while it isn't awful, it could definitely be better. The Decision Tree is still probably the better method to use as the R^2 value from the Linear Regression was very low. Future work may include trying other types of methods, using different features, and potentially even updating the dataset to be a bit more descriptive.

References

https://fivethirtyeight.com/features/avengers-death-comics-age-of-ultron/