CMSE 381 Final Project

96

Charlie Hultquist

April 18, 2021

1 Introduction

In Five Thirty Eight's article, "Joining the Avengers Is As Deadly As Jumping Off A Four-Story Building," [1] Walt Hickey, the author, tells us about the dangers and rewards of being one of America's most well known super heroes. Compiling information from the Marvel Wikia, he reports the number of deaths, and the number of returns from death, for 173 characters in the series. The data compiled and the article are a great start, but there is much more to be analyzed here. This project aims to dive further into this dataset, and answer the question: **What character demographics serve as good predictors of how often an MCU character dies?**. I employ various range of methods – LASSO regression, logistic regression, random forests, and boosting trees – to answer this guiding question. These methods determine that number of character appearances, and the character's total years serve as good predictors of character death.

2 Related Works

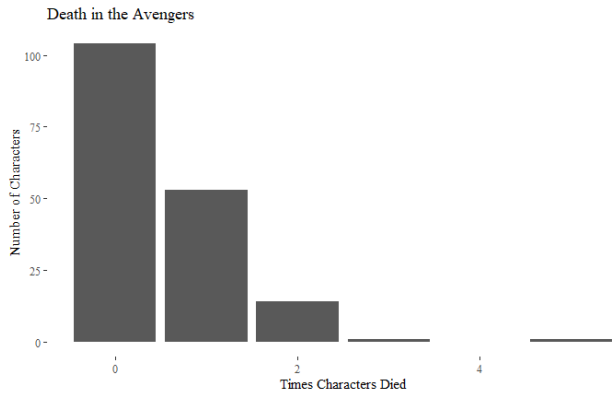
Hickey's analysis of the dataset lies mostly in producing summary statistics about the data. He correctly claims that 40% of avengers have died, about once every seven months. Despite the high death rate, there is a remarkable 66% chance of an avenger returning from their first death. This report's corresponding `R` notebook contains independent verification of these claims. Another analysis using this data was conducted by Joy Harjanto in 2019, titled "As An Avenger, Dying Is Part of the Job" [2]. Harjanto goes further than Hickey and produces nice visualizations of the dataset, which are recreated below. She also performs a t-test to conclude there is no significant difference in the number of appearances of avengers who have died or not. Seyda Zinnur Kalkan also performs an analysis of the dataset on Kaggle, titled, "Basic Data Science with Marvel Characters" [3]. In it, she find that appearances and year are weakly negatively correlated, and also produces summary plots of the data similar to Harjanto's.

The models produced by Hickey, Harjanto, and Kalkan share some nice statistics and visualizations, but don't analyze the data to it's full potential. For instance, there are a few variables that are in the data that go unanalayzed: gender, number of appearances, years since joining, and avengers status. We can try to use these variables, as well as total death and return count, as predictors and responses in various models. All three reports produced nice summaries, but stopped short of developing prediction models or interpreting such models.

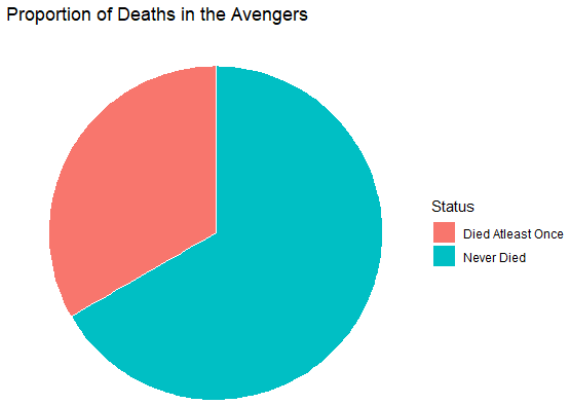
The other major fault I find in the analyses is that "total deaths" is a faulty metric, as it doesn't really capture if a character actually dies. Therefore, I introduce another variable that discusses if a character has died and not returned. This further allows me to approach character from both a classification (has the character fully died) or regression (how many times has the character "died") standpoint.

3 Dataset

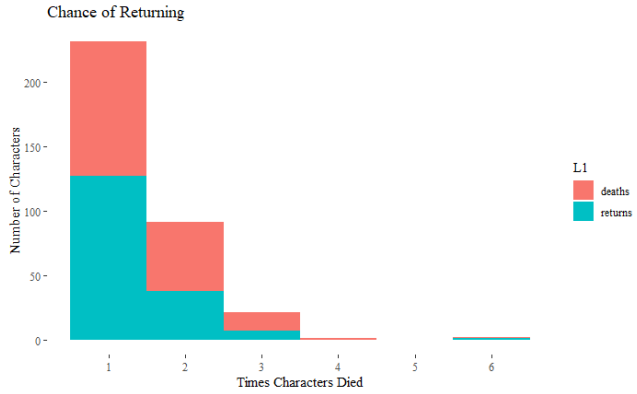
The original dataset from FiveThirtyEight was difficult to use, as it didn't total the deaths and recoveries from the characters; it used the last 12 columns as a running tally of the total deaths and recoveries. This was cleaned up and organized into three predictors: `deaths` (total number of deaths), `returns` (total number



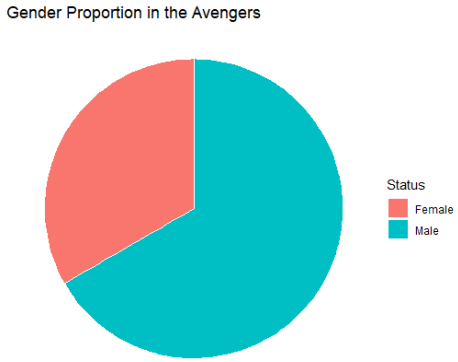
(a) Total Avengers Death Counts



(b) Most Avengers Haven't Died



(c) You have a better chance of recovery after death 1



(d) Most Avengers are Male

Figure 1: Summary Graphs of the Avengers Dataset

of returns from death), and `gone` (a boolean noting if the character *actually* died, i.e. more deaths than returns).

Together, these new variables, along with `Appearances` (total number of comic appearances), `Gender`, `Years.since.joining`, and `Status` (honorary or full member) are the variables that I will use as predictors and responses. The below graphs in Figure 1 summarize the data nicely. Some are recreations of Harjanto's work.

The other major modification made to the dataset was the removal of characters without accurate data on start date. A large number of Avengers are listed as joining in 1900, which I believe is a fill in for no joining date found (the Marvel Wikia lacks this data for these characters, and no date is given for their creation in the dataset's original Full/Reserve Column either). Because I am using this as a predictor, I remove these characters from the data - they are essentially null values, without being labelled as "N/A". All together, this removes 14 of the original 173 characters.

4 Methodology

4.1 What Does Death Mean?

Before attempting to develop models to predict the likelihood of death, I first address the issue of what it really means to "die". Given the fictional nature of the Avengers, multiple deaths can make sense to

discuss, as many characters do revive rom their deaths. Using a stricter, more realistic, definition of death, "gone and hasn't come back", it seems natural to death as a boolean value answering "has this character completely died?", instead of the numerical value answering "how many times has this character died?". Thus, in addition to predicting how many deaths a character has gone through, I also attempt to do perform some classification regression with the response variable, `gone`, discussed earlier. When in the classification setting, only one of predictors `deaths` and `returns` are used at a time, since together they were used to compute the `gone` variable. The methods chosen to study these questions were selected because of their interpretability in order to find *which* predictors were important in determining death.

4.2 Lasso Regression

A LASSO regression is performed for the regression setting, using the predictor variables `Gender`, `Status`, `Appearances`, and `Years.since.joining`. Because there are relatively few predictors, I also include fourth-degree polynomials of the variables `Years.since.joining` and `Appearances`. Lasso regression will regularize this fit and avoid over-fitting with large polynomials. Fourth-degree polynomials were selected, as larger polynomials tended to produce a LASSO model with incredible large λ's and predicted each character's number of deaths as simply the average number of deaths, $y_i = \overline{y}$. Lasso regression was specifically chosen over ridge regression because of it's ability to give sparse models [4]. The best λ for the model is selected by 10-fold cross-validation. The training and testing error and R^2 are reported for these methods, and we use the sparse model obtained by lasso to make conclusions about the importance of each predictor. The `R` package `glmnet` [5] is used to perform this regression.

4.3 Tree Ensembles

Both Random Forests and Gradient-Boosted ensembles are used as predictive models, given their high predictive power [4]. These tree ensembles are used both in the regression setting (predicting deaths per appearance), and classification setting (predicting if character has permanently died). In the classification setting, the Gini index is used to grow trees. The variables `Gender`, `Status`, `Appearances`, and `Years.since.joining` are used as predictors for these models. The `R` packages `randomForest` [6] and `gbm` [7] are used to perform this prediction.

The regression and classification random forests were done with 100 trees considering $\sqrt{p} = 2$ predictors at each split. The regression boosting is done with 100 trees and a shrinkage of $\lambda = 0.001$, while the regression classification is done with 25 trees and a shrinkage of $\lambda = 0.001$. These parameters were selected by trial and error to achieve lowest test MSE or misclassification rate.

The training and testing accuracy is reported for classification trees, and R^2 and mean square error are reported for regression trees. In the regression setting, I obtain importance of each predictor as total amount that the RSS (8.1) is decreased due to splits over the given predictor, averaged over all trees. In the context of classification, I used the total Gini index decrease instead of RSS.

4.4 Logistic Regression

A logistic regression is performed for the regression setting three, first using the variables `Gender`, `Status`, and `Years.since.joining`, and `deaths`, second using the same variables, except with `returns` in place of `deaths`, and lastly using neither `deaths` or `returns`. The reasoning for splitting up these variables is discussed above. The base `R` function `glm` is used to perform this regression. The training and testing accuracy are reported for these two fits, and I conclude about the importance of each predictor, β by its ability to reject the null hypothesis $H_0 : \beta = 0$.

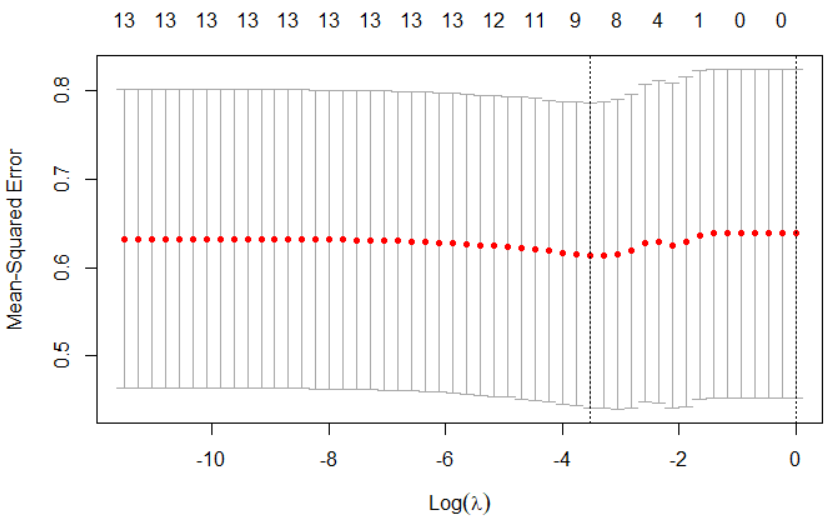


Figure 2: Various Cross-Validation MSE, using 10 folds. The optimal λ was selected to be 0.023

Predictor	Coefficient
Intercept	0.616
Gender=MALE	-0.200
Honorary=Full Avenger	0.149
Honorary=Honorary Avenger	-0.262
$(\text{Years.Since.Joining})^1$	2.234
$(\text{Years.Since.Joining})^2$	-0.379
$(\text{Years.Since.Joining})^3$	0.631
$(\text{Years.Since.Joining})^4$	0.134
$(\text{Appearances})^1$	**0.000**
$(\text{Appearances})^2$	1.012
$(\text{Appearances})^3$	-0.667
$(\text{Appearances})^4$	**0.000**

Table 1: Caption

5 Results and Discussion

5.1 Regression Results

Figure 2 shows the output of the cross-validation to choose the optimal λ, finding it to be $\lambda_{best} = 0.023$. Comparatively, most of the predictor coefficients were on the order of 0.1. Using this optimal parameter, a LASSO regression was fit to testing data. The LASSO prediction's coefficients are summarized in Table 1.

The LASSO prediction contains some interesting information about the relation of different parameters to the number of deaths. For example, all other variables held constant, male avengers have less deaths than female avengers, and generally Years since joining is positively related to number of deaths. The LASSO feature removal property was successful, removing the linear and the fourth-degree Appearances predictor. Overall the LASSO results can suggest `Years.Since.Joining` is a more important predictor than Appearances, and Gender and Honorary status can be important features. The overall results of the fit are reported in Table 2.

The variable importance from the regression fit using Random Forest is shown in Figure 3. It is clear that Years.since.joining is an important variable. The results also indicate that number of appearances and

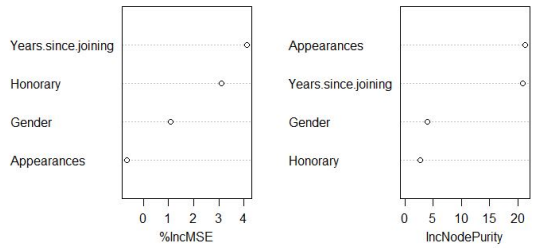


Figure 3: Variable importance measures from Random Forest regression fit. The left plot demonstrates the increase in error when a given variable is excluded from the model. The right plot demonstrates the mean decrease in RSS error that results from splits using a given variable. In both, larger values indicate a variable is more important

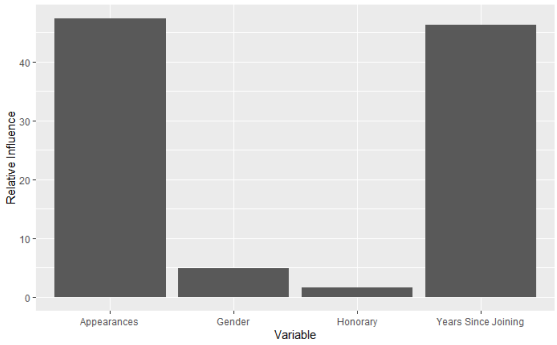


Figure 4: Variable importance measures from Boosting Tree regression fit. The plot demonstrates the mean decrease in RSS error that results from splits using a given variable. Larger values indicate a variable is more important

Method and Parameters	Test MSE	Test R^2	Train MSE	Train R^2
LASSO ($\lambda = 0.023$)	0.52	0.25	0.52	0.41
Random Forest ($m = \sqrt{p} = 2$)	0.51	0.27	0.22	0.87
Boosting Tree (Shrinkage=0.001)	0.54	0.22	0.34	0.73

Table 2: Testing and training MSEs and R^2 for the various regression methods

honorary are important variables by some metrics. The variable importance from the regression fit using Boosting is shown in Figure 4. This seems to agree with the second plot in Figure 3 – Appearances and Years since joining seem to be the most important variables in predicting number of deaths. The overall results of both tree ensembles are reported in Table 2.

Overall, it is difficult to say any method was really that great at predicting number of deaths. The testing MSE is about 0.5, and the R^2 value about 0.25 for each of the three results. Results were limited by the small number of predictors and there only being 5 values of "deaths," making the regression setting somewhat difficult. Despite, the mediocre predictive power, we found that across all models, `Years.since.joining` was an important predictor. Logically, this makes sense, as an character who has been in the avengers longer will have had more chances to die. However, when studying "death likelihood", we now move to the classification setting, where we predict if a character has died and not returned.

5.2 Classification Results

The variable importance from the regression fit using Random Forest is shown in Figure 5. It is clear that Appearances is the most important variable in predicting whether a character has actually died. The variable importance from the regression fit using Boosting is shown in Figure 6. This seems to agree with results from random forest: Appearances is the most important variables in predicting `gone`. In the boosting tree, years since joining is another equally important variable. The overall results of both tree ensembles are reported in Table 4.

Logistic regression results appear in Table 3. Variables with coefficients β for which the null hypothesis, $H_0 : \beta = 0$ could be rejected at a significance level of 0.05 are bolded in the tables. Only logistic regression when including death produces coefficients significantly different from 0. Both data with only `returns` and data with neither `returns` or `deaths` perform the same, suggesting the number of returns from death a character has had is unrelated to if the character has fully died. Across all fits, it seems that Years since joining is a good predictor of if a character has permanently died. Across the board, the longer a character has been in the comics, the less likely they are to have permanently died, perhaps the opposite of expectation.

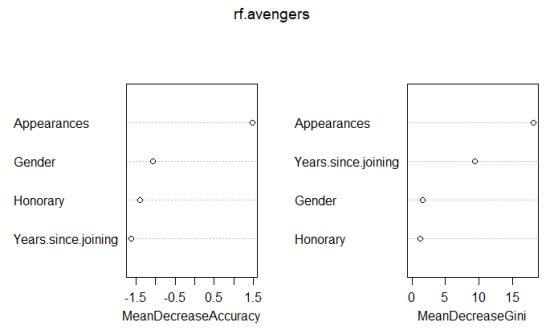


Figure 5: Variable importance measures from Random Forest classification fit. The left plot demonstrates the decrease in accuracy when a given variable is excluded from the model. The right plot demonstrates the mean decrease in total Gini index (measuring node impurity) decrease that results from splits using a given variable. In both, larger values indicate a variable is more important.

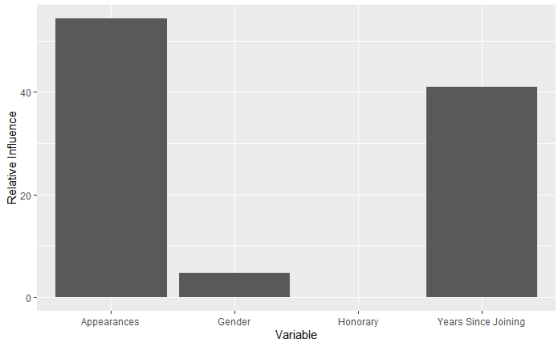


Figure 6: Variable importance measures from Boosting Tree classification fit. The plot demonstrates the mean decrease in total Gini index (measuring node impurity) decrease that results from splits using a given variable. Larger values indicate a variable is more important.

	With Deaths		With Returns		With Neither	
Predictor	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Intercept	-2.27	0.188	-0.835	0.516	-0.834	0.516
Honorary=Full Avenger	-0.822	0.610	-0.423	0.739	-0.422	0.739
Honorary=Honorary Avenger	0.474	0.803	-0.327	0.829	-0.328	0.828
Appearances	0.00031	0.527	0.00015	0.727	0.00015	0.727
Years.since.joining	-0.073	**0.005**	-0.02	0.296	-0.02	0.257
Gender=MALE	1.17	0.123	0.342	0.534	0.341	0.527
deaths	2.49	$\mathbf{3.28\times10^{-6}}$	-	-	-	-
Returns	-	-	0.0038	0.992	-	-

Table 3: Results for Logistic Regression with Different Variable Sets. Results with $p < 0.05$ are bolded.

The accuracy of each fit is reported in Table 4.

The results from the trees and the logistic regression are somewhat in contrast. Both boosting and random forest report that `Appearances` is the most important predictor, followed by `Years.since.joining`. However, the logistic regression result reports `Years.since.joining` is the most significant, and `Appearances` is unlikely to be significant. This is presumptively due to the linear nature of logistic regression - it produces a linear decision boundary, while the decision boundary for appearances can be a variety of shapes. If the `gone` response is very non-linear in the `Appearances` variable, logistic regression wouldn't find it to be significant. The trees are able to outperform the logistic regression because of this.

6 Conclusions and Discussion

Overall, the most important variables in predicting deaths are not personal descriptors of characters like gender or avengers status, but rather their history of existence in the comics. Together, total number of appearances, and years since conception of the character predict the total number of deaths (both fake and permanent) deaths a character has gone through. If a new character were introduced to the Marvel Comic Universe, it would be difficult to predict how likely the character is to die, as we simply wouldn't have good data on how often this character would appear. Number of deaths *should* be correlated to years in the MCU, as a character that has existed longer will have had time to die more often.

Method and Parameters	Test Accuracy	Train Accuracy
Logistic Regression w/ Deaths	0.84	0.90
Logistic Regression w/ Returns	0.78	0.84
Logistic Regression w/ Neither	0.78	0.84
Random Forest ($m = \sqrt{p} = 2$)	0.86	0.92
Boosting Tree (Shrinkage=0.001)	0.86	0.86

Table 4: Testing and training accuracy rates for the various classification methods

Overall, the analysis could be improved with more data on the characters. The Marvel Wikia has much more information on many of the characters: birthplace, species, superpowers, age, and more. Further data collection was outside the scope of this analysis, but if the data were compiled, it could provide more interesting results. Although the results that "death" is strongly related to "years existed" and "appearances" are unsurprising, they loosely confirm expectations that Marvel's writers don't write more deaths for characters of different status, gender, and don't "kill off" an over-proportional amount of recent characters.

References

[1] Hickey (2015). Joining The Avengers Is As Deadly As Jumping Off A Four-Story Building. *FiveThirtyEight*.

[2] Harjanto (2019). As An Avenger, Dying Is Part of the Job. *TowardsDataScience.com*.

[3] Kalkan (2019). Basic Data Science with Marvel Characters. *Kaggle.com*.

[4] James, Witten, Hastie, Tibshirani (2017). An Introduction to Statistical Learning. *Springer Texts*.

[5] Friedman J, Hastie T, Tibshirani R (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.

[6] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. *R News* 2(3), 18–22.

[7] Greenwell, B., Boehmke, B., Cunningham, J., Developers, G. (2019). gbm: Generalized boosted regression models. *R package version*, 2(5).