AVENGERS RISK AND ALIGNMENT ANALYSIS

FINAL REPORT

1 INTRODUCTION

Comic books have been a part of pop culture for a very long time. Many of them have a cult following and what could be a better example than Marvel. It can be argued that Avengers are the best-known characters from the Marvel universe. And with every issue the lore around these characters becomes more extensive. This project focuses on risks that come with joining the Avengers as stated by FiveThirtyEight [1]. The problems that I tackle on my own are- The analysis of the social network of the Marvel Universe [2] and prediction of the alignment of comic book characters based on their physical features and abilities using random forest classifier. All data analysis has been performed in python, using packages like pandas, NumPy, sklearn, matplotlib , networkx and seaborn.

2 RELATED WORK

2.1 Data

The data used in this part of the project was obtained from FiveThirtyEight's github [3]. The dataset has details about the deaths and resurrections of Marvel comic book characters between the time they joined the Avengers and April 30, 2015. This puts the number of characters at 173. The data is collected by Walt Hickey by crawling thorough marvel wiki pages.

2.2 Method

Existing literature [1] claims that being an Avenger is about as dangerous as jumping from 5 story building. Jumping from a five-story building puts the risk of death at about 50% [4]. To verify the results that are mentioned by FiveThirtyEight, risk analysis for each category was performed. In the dataset the Comic characters have been granted one of three designations as mentioned in the "Honorary" column of the dataset – honorary, full, probationary, academy. To perform the categorical analysis, extra columns that had the total deaths and resurrections experienced by each character (each row) were added to the original dataset, since the original dataset keeps track of each time a death/revival occurs but not the total. Then the data was grouped by categories and values that were greater that 1 were counted. For this section one row is one count, the magnitude of the total death total resurrection does not play a role in this categorical analysis. The result can be seen in Figures 1 & 2.

The magnitude of deaths is shown in Figure 3. This also brings out an interesting outlier, a character who has died 5 times.

STATUS	TOTAL	DEATH	REVIVAL
Full	138	61	42
Honorary	16	5	3
Probationary	2	1	1
Academy	17	2	0

Table 1: Death table

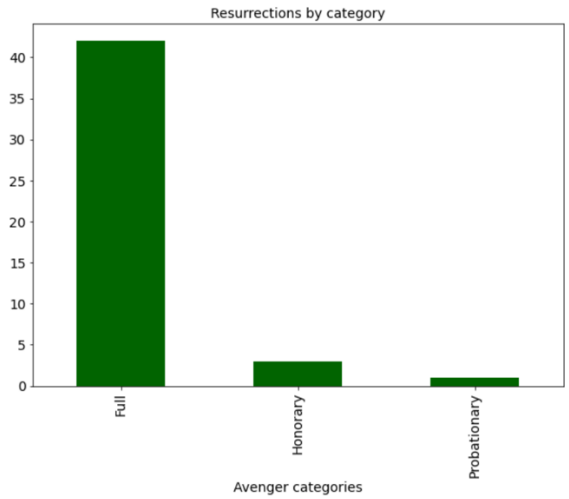

Figure 1: characters who were resurrected at least once by category

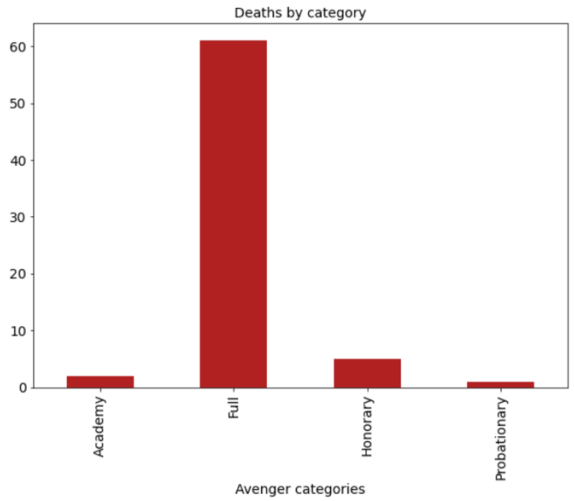

Figure 2: characters who were killed at least once by category

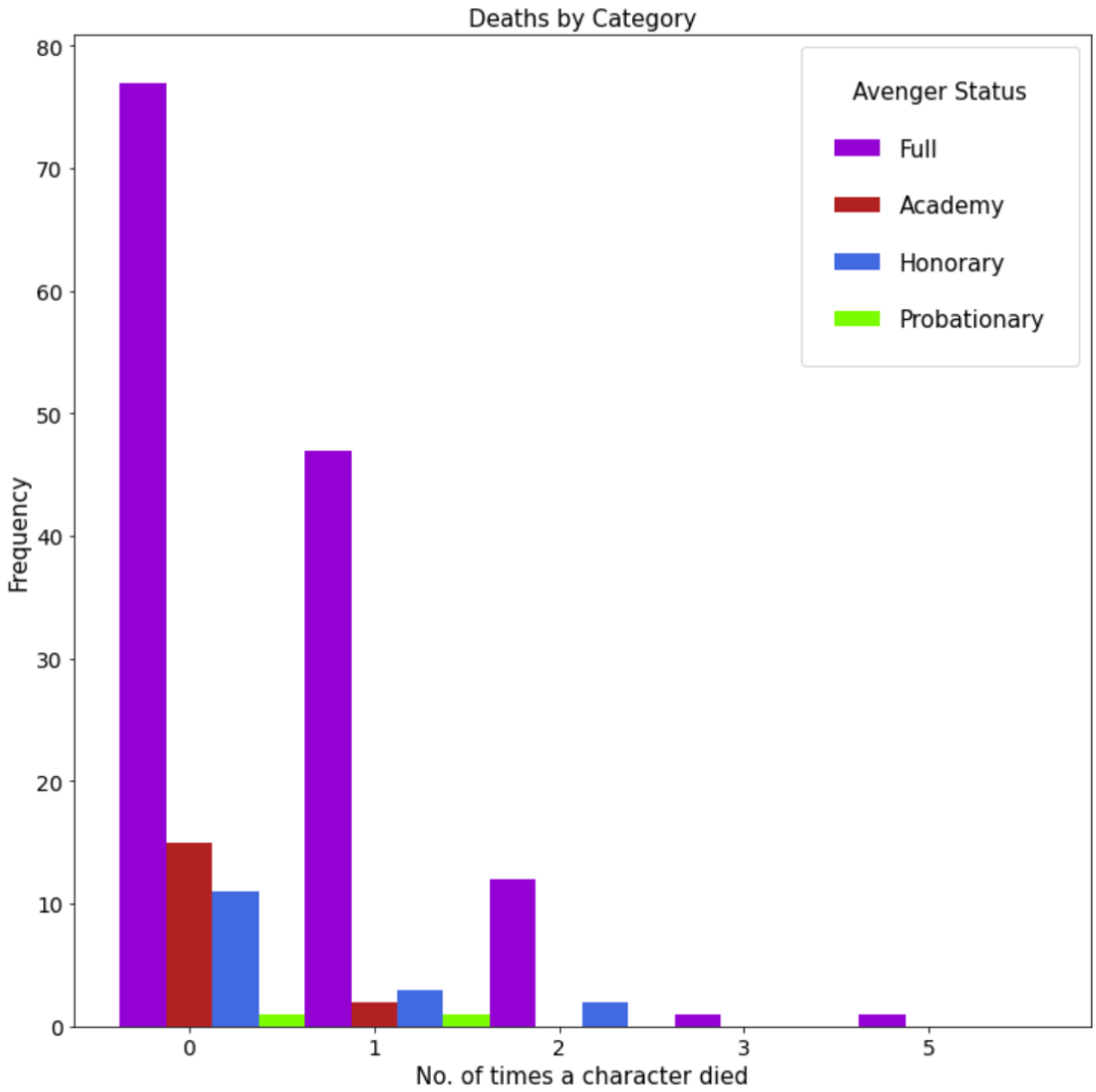

Figure 3: Death distribution by category and count

3 EXTRA MODELS

The data and the analysis in the FiveThirtyEight article does surface level statistical analysis, the most intensive part of the process that was followed by Walk Hickey is the data collection itself. The questions that this section focusses on are related to the social network of the Marvel Universe and the classification of characters as “good”, “bad”, “neutral”

3.1 Social Network analysis

This analysis is performed to find characters who have the highest degree centrality, closeness and betweenness.

3.1.1 Data

For social network analysis datasets corresponding to edges between characters, character nodes and character names was obtained from Kaggle [5].

3.1.2 Method

A data object that has connection information between the characters was formed using networkx. This object would be used in the calculation of graph properties.
First, to find out who the most well connected Marvel character is, the degree centrality was calculated. Degree centrality is the fraction of nodes that a specific node is connected to [6]. After a dictionary that had all 6426 nodes and their degrees was created, it was sorted and the top 10 values were displayed.
Second, to find out who the most important characters are the eigenvector centrality was calculated. A node is more important if it is connected to many important nodes [7]. In this case as well the top 10 characters were calculated.
Thirdly, to find which character to go to find the character we are looking for the quickest, the betweenness of the nodes was calculated. Betweenness is the percentage of shortest paths that a node is a part of [6]. Then the top 10 characters were extracted.

CHARACTER	DEGREES
CAPTAIN AMERICA	0.30
SPIDER-MAN	0.27
IRON MAN	0.24
THING	0.22
MR. FANTASTIC	0.21
WOLVERINE	0.21
HUMAN TORCH	0.21
SCARLET WITCH	0.21
THOR	0.20

Table 2: Top 10 degree centrality characters

CHARACTER	INFLUENCE
CAPTAIN AMERICA	0.1168
IRON MAN	0.1025
SCARLET WITCH	0.1008
THING	0.1008
SPIDER-MAN	0.1002
MR. FANTASTIC	0.0997
VISION	0.0985
HUMAN TORCH	0.0985
WOLVERINE	0.0984

Table 3: Top 10 influential characters

CHARACTER	BETWEENNESS
SPIDER-MAN	0.0735

CAPTAIN AMERICA	0.0570
IRON MAN	0.0372
WOLVERINE	0.0357
HAVOK	0.0357
DR. STRANGE	0.0292
THING	0.0254
HAWK	0.0248
HULK	0.0239

Table 4: Top 10 characters who know other characters

3.2 Alignment Classification

The question was if a character can be classified to the right alignment based on their abilitie s, physical attributes and gender. Because if that were the case then it could be argued that th e villain or the heroes have an advantage when they face each other.

3.2.1 Data

The datasets [8] that were used had information about the physical attributes, including gend er, height, weight and ratings that were fan given, this included intelligence, strength, speed, durability, power, combat. A merged dataset that had all these attributes was used in the class ification process. "ID", "SkinColor", "Race", "EyeColor", "HairColor" were dropped from th e merged dataset because they had a lot of missing information as well as more than 8 classes each. Multiple classes did not have many samples each, therefore they would made made the model biased.

3.2.2 Method

Since the scatter plots of the dataset didn’t show any clear decision boundaries or clusters or l inear relationships, a tree-based approach was chosen. The random forest classifier from skle arn was used. The testing size was 20% and the training size was 80%. The best n_estimator was determined to be 125 trees. An analysis on the most important features was also perform ed.

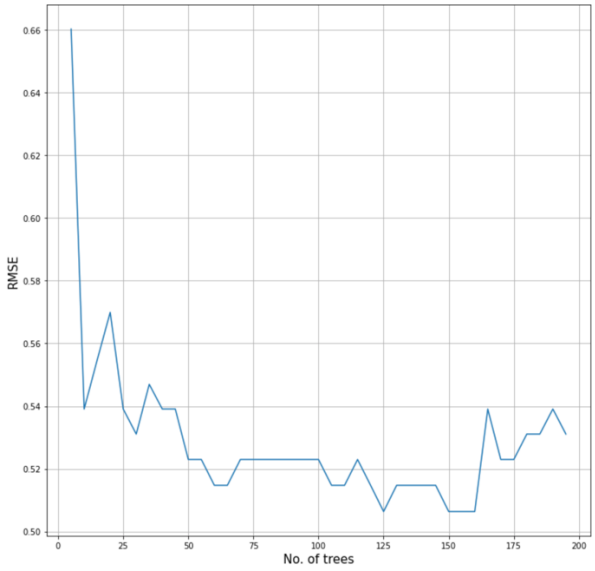

Figure 4: Best number of trees based on RMSE

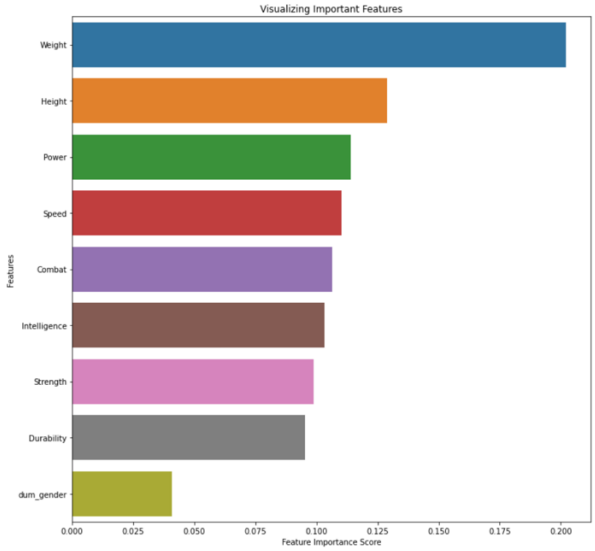

Figure 5: Important features

4 RESULTS

It is more likely for comic characters who do not have full avengers status to stay dead after they are killed. “Academy” status comes with 100% mortality rate, out of 2 academy status who died none were resurrected. Being a “Full” avenger has the most perks. Out of 61 “Full” avengers who died at least once, 42 were resurrected at least once, that is a mortality rate of 33% (Table 1). Figure 3 shows that as an avenger of any status, you have a 39.9% chance of getting killed at least once. This is close to the figure of *about 40%* that is mentioned in the FiveThirtyEight article. A character named Jocasta is an outlier in the dataset. She has been killed and resurrected 5 times. This has been possible because she is a robot made by Ultron and has been turned on/off multiple times over the course of the comics.

The top 10 characters from each category of the social network analysis have a lot of overlap. This makes sense because in a social setting the graph classification criterion will be highly correlated. Meaning the mst popular charcter will be the one with the most connections and they will also be the person who has the shortest connections to another person. Not surprisingly, Captain America, Iron Man and Spider Man are always in the top 5 (Table 2-4).

Random forest with 125 trees has the lowest RMSE of 0.555 and an accuracy score of 72%. Removing gender, which was the least important of all the features, did not change the accuracy score at all. Given the poor accuracy of the model it can be said that the percentage

of overpowered heroes and villain comic characters is balanced i.e the not all villains are weak and not all superheroes have unbitable powers.

5 REFERENCES

1- WaltHickey. "Joining The Avengers Is As Deadly As Jumping Off A Four-Story Building." *FiveThirtyEight*, FiveThirtyEight, 12 May 2015, fivethirtyeight.com/features/avengers-death-comics-age-of-ultron/.
2- Wang, Xingya. "Marvel Universe Visualization." *Information Visualization*, 10 Jan. 2019, studentwork.prattsi.org/infovis/projects/marvel-universe-visualization/.
3- Fivethirtyeight. "Fivethirtyeight/Data." *GitHub*, github.com/fivethirtyeight/data/tree/master/avengers.
4- G, Suresh. "Https://Medwinpublishers.com/NNOA/NNOA16000183.Pdf." *Nanomedicine & Nanotechnology Open Access*, vol. 5, no. 2, 2020, doi:10.23880/nnoa-16000183.
5- Sanhueza, Claudio. "The Marvel Universe Social Network." *Kaggle*, 28 Jan. 2017, www.kaggle.com/csanhueza/the-marvel-universe-social-network.
6- Telatnik, Mitchell. "How To Get Started with Social Network Analysis." *Medium*, Towards Data Science, 27 May 2020, towardsdatascience.com/how-to-get-started-with-social-network-analysis-6d527685d374.
7- *Eigenvector Centrality*, www.sci.unich.it/~francesc/teaching/network/eigenvector.html.
8- R, Danniel. "Marvel Superheroes." *Kaggle*, 29 July 2018, www.kaggle.com/dannielr/marvel-superheroes.