Using Machine Learning to Learn the Law of Alive Characters in Marvel Comics

DV2578-Machine Learning Project Report

Rong Peng
Blekinge Institute of Technology

I. Introduction

From 1947 Martin Goodman founded Marvel Comics, more and more fans are crazy about this famous brand name because of its well known characters such as Spider-Man, Iron Man, Captain American. In 2018, the popular film *Avengers: Infinity War* based on the Marvel Comics released on theaters around the world. The die-hard audience was shocked by the bonkers ending, some audience even cried during watching the film, due to Thanos wins the war and many superheroes like Loki, Spider-Man, Black Panther have been killed. Heroism has died more than half, dark forces is powerful.

People can't help but ask what are the Marvel team tricks? What kind of comic characters could live to the end? What are their features?

Sadly, the father of Marvel, Stan Lee, died in November 2018. Besides mourning and memory, how will the Marvel Comic team develop the storyline could continue attracting the die-hard fans and keep the box office miracle?

This paper tries to figure out answers of the above questions, and proposes models to find the law of Marvel Comic characters alive features through analyzing the provided data coming from Marvel Wiki[1] which scrapped by FiveThirtyEight[2].

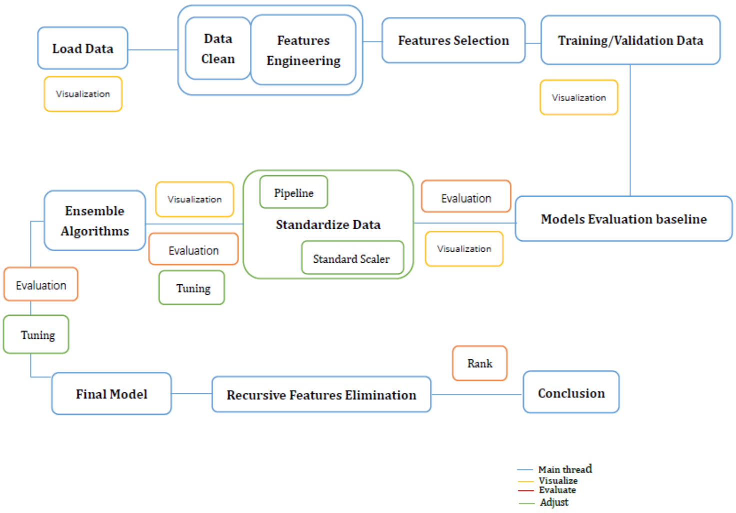
And the paper is organized as follows: Section II gives main procedure and related work, Section III & IV presents the experiment, involving data collecting and pre-processing, features selection, models evaluation, tuning, evaluation .etc. Further work and limitations are discussed in Section V. Finally Conclude in Section VI.

II. Related Work

A. Main Ideas

Given the problem what kind of characters living to the end, features engineering is very important. The main work can be separated into three tasks. Firstly, Pre-process the original data and do the features engineering. Secondly, carry out some models as baseline and train. The binary classification problem that whether the character is alive or not, can use models like the LogisticRegresssion, the LinearDiscriminantAnalysis, KNN, SVM to evaluate and compare. During the process, I will also normalize the data, try to make the data fit the Gaussian distribution[3], and use Pipeline to sequentially apply a list of transforms and a final estimator. At the same time, using some ensembles algorithms such as the AdaBoostClassifier, the ExteaTreesClassifier to train and determine whether they can improve accuracy. Thirdly, using the Recursive Features Elimination (RFE) and combining with the final selected models to obtain the features ranking, print out the relationship then we can obtain predictive conclusion. Fig1 illustrates the main steps of my work.

Fig1. Main procedure



B. Measure

From Flach.P [4]we can know the simplest case we have only two classes which are usually referred to as positive and negative, which is often called binary classification. He also mentioned[4] assessing classification performance can use contingency table or confusion matrix, as Fig2.

Fig2.Confusion matrix

Real Situation	Predictive Result	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The real situation and the predictive result can be divided into true positive, false positive, true negative, false negative this four types.

The following formulas measure based on the confusion matrix.

$$Recall = \frac{TP}{TP+FN} \quad Precision = \frac{TP}{TP+FP}$$

$$F1 = \frac{2TP}{2TP+FP+FN} \quad Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Generally, Precision and Recall are for a certain class. Precision indicates how many predictions are correct in the sample you are predicting positive. The Recall of the positive sample indicates how many positive samples are positively predicted by you. In general, F1-score is used to synthesize precision and recall as an evaluation indicator.

Accuracy indicates how many proportions of your sample predictions are correct.

Peter Flach[4]has summarized different quantities and evaluation measures for classification. In this paper measures models by comparing the Recall, Precision,F1-Score,Accuracy results.

C. *Recursive Features Elimination (RFE)*

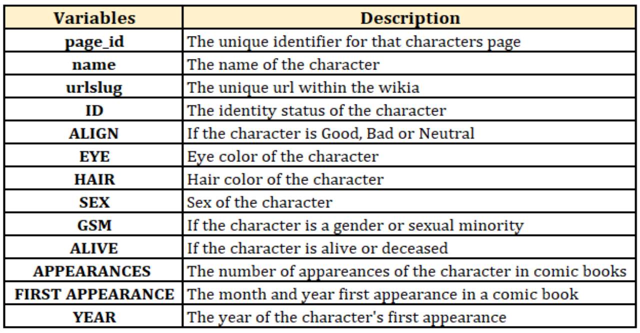
RFE algorithm is used for feature selection and belongs to one of the packaging method feature selection algorithms. The recursive elimination feature method uses a machine learning model to perform multiple rounds of training. After each round of training, the features corresponding to several weight coefficients are eliminated, and the next round of training is performed based on the new feature set. RFE with cross-validation, which can automatically adjusts the number of features selected by cross-validation.[5]

The last step in the procedure, selected the final model as the base model, and through RFE with cross-validation selecting the ranking top three features that have the greatest impacts on the prediction results, thus analyze association.

III. METHOD & EXPERIMENT

The full data was provided by FiveThirtyEight[2], they scrapped from Marvel Wiki[1]. The data contains the following variables as Fig3:

Fig3.Full data columns name[2]



Variables	Description
page_id	The unique identifier for that characters page
name	The name of the character
urlslug	The unique url within the wikia
ID	The identity status of the character
ALIGN	If the character is Good, Bad or Neutral
EYE	Eye color of the character
HAIR	Hair color of the character
SEX	Sex of the character
GSM	If the character is a gender or sexual minority
ALIVE	If the character is alive or deceased
APPEARANCES	The number of appareances of the character in comic books
FIRST APPEARANCE	The month and year first appearance in a comic book
YEAR	The year of the character's first appearance

The experiment takes the features engineering. Some features needs to dropout. And the experiment processing steps can be list as following:

STEP 1: Pre-processing and Visualization
STEP 2: Baseline Models Training
STEP 3: Standard data and Parameter Tuning
STEP 4: Ensembles Models Training and Tuning
STEP 5: Evaluation and Comparison
STEP 6: Recursive Features Elimination

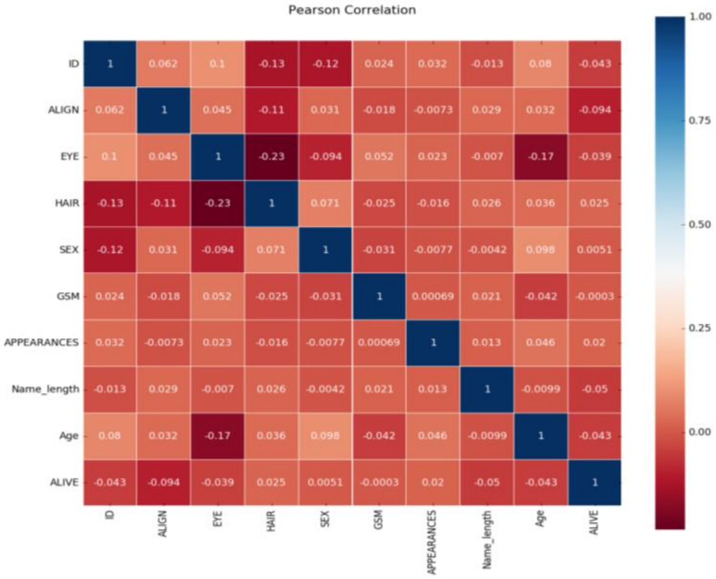
The first five steps aims at extracting the models, the last step aims at finding the most relevant attributes.

3.1 Pre-processing and Visualization

Loading the full data, and remove the blank record in “ALIVE”. Meanwhile generate the new features “Name_length” and “Age”, to record the characters length of their names, and record from the first appearance year to 2019, how old they are.

During the feature engineering, I separate data into several intervals, as Fig4 presents. Then in the features selection, I dropout some unnecessary attributes but remain 'ID', 'ALIGN', 'EYE', 'HAIR', 'SEX', 'GSM', 'APPEARANCES', 'Name_length', 'Age', 'ALIVE' features, using Pearson Correlation to visualize, as Fig5 presents. Lastly, using 10-fold cross-validation to divide randomly, gaining the training and validation dataset.

Fig5.Pearson Correlation

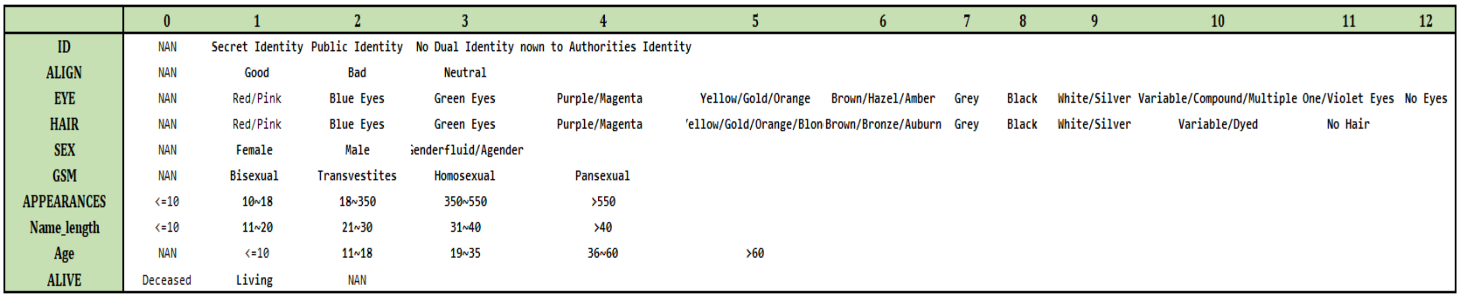


	ID	ALIGN	EYE	HAIR	SEX	GSM	APPEARANCES	Name_length	Age	ALIVE
ID	1	0.062	0.1	-0.13	-0.12	0.024	0.032	-0.013	0.08	-0.043
ALIGN	0.062	1	0.045	-0.11	0.031	-0.018	-0.0073	0.029	0.032	-0.094
EYE	0.1	0.045	1	-0.23	-0.094	0.052	0.023	-0.007	-0.17	-0.039
HAIR	-0.13	-0.11	-0.23	1	0.071	-0.025	-0.016	0.026	0.036	0.025
SEX	-0.12	0.031	-0.094	0.071	1	-0.031	-0.0077	-0.0042	0.098	0.0051
GSM	0.024	-0.018	0.052	-0.025	-0.031	1	0.00069	0.021	-0.042	-0.0003
APPEARANCES	0.032	-0.0073	0.023	-0.016	-0.0077	0.00069	1	0.013	0.046	0.02
Name_length	-0.013	0.029	-0.007	0.026	-0.0042	0.021	0.013	1	-0.0099	-0.05
Age	0.08	0.032	-0.17	0.036	0.098	-0.042	0.046	-0.0099	1	-0.043
ALIVE	-0.043	-0.094	-0.039	0.025	0.0051	-0.0003	0.02	-0.05	-0.043	1

3.2 Baseline Models Training

To form the first impression on which models will fit the dataset, using some baseline models to train at first. Then I used some models like LogisticRegression, LinearDiscriminantAnalysis, KNeighborsClassifier, GuassianNB, SVC to train and evaluate their accuracy. Result finding shows the LogisticRegression and LinearDiscriminantAnalysis are the best performance.

Fig4. Features Engineering

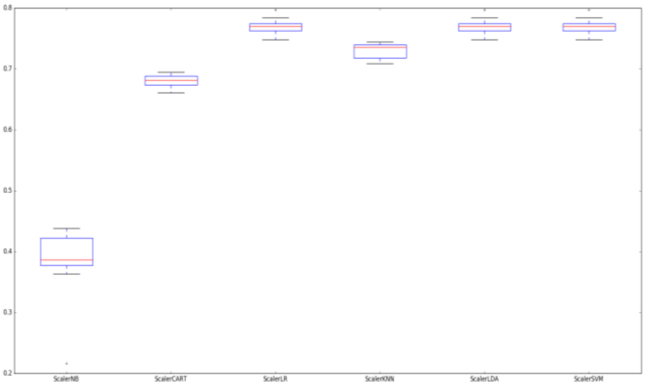


	0	1	2	3	4	5	6	7	8	9	10	11	12
ID	NAN	Secret Identity	Public Identity	No Dual Identity	nown to Authorities Identity								
ALIGN	NAN	Good	Bad	Neutral									
EYE	NAN	Red/Pink	Blue Eyes	Green Eyes	Purple/Magenta	Yellow/Gold/Orange	Brown/Hazel/Amber	Grey	Black	White/Silver	Variable/Compound/Multiple	One/Violet Eyes	No Eyes
HAIR	NAN	Red/Pink	Blue Eyes	Green Eyes	Purple/Magenta	'ellow/Gold/Orange/Blon	Brown/Bronze/Auburn	Grey	Black	White/Silver	Variable/Dyed	No Hair	
SEX	NAN	Female	Male	;enderfluid/Agender									
GSM	NAN	Bisexual	Transvestites	Homosexual	Pansexual								
APPEARANCES	<=10	10~18	18~350	350~550	>550								
Name_length	<=10	11~20	21~30	31~40	>40								
Age	NAN	<=10	11~18	19~35	36~60	>60							
ALIVE	Deceased	Living	NAN										

3.3 Standardize data and Parameter Tuning

Standardized data is a means of efficiently processing Gaussian-distributed data. The output is 0 as the median and the variance is 1, and is used as the input to the algorithm that assumes that the data fits the Gaussian distribution. I used StandardScaler and transform methods to realize. After that, visualize their condition in box plot.

Fig6. Models Box Plot



Tuning for the best performance model, I chose to use GrideSearchCV, to find the optimal estimator. GridSearchCV performs a Cartesian set combination based on the hyperparameter range listed by the user, using each set of hyperparameter training models, and picking the hyperparameter combination with the smallest validation set error. In this experiment I found when C=1, LogisticRegression accuracy has improved by tuning.

3.4 Ensembles Models Training and Tuning

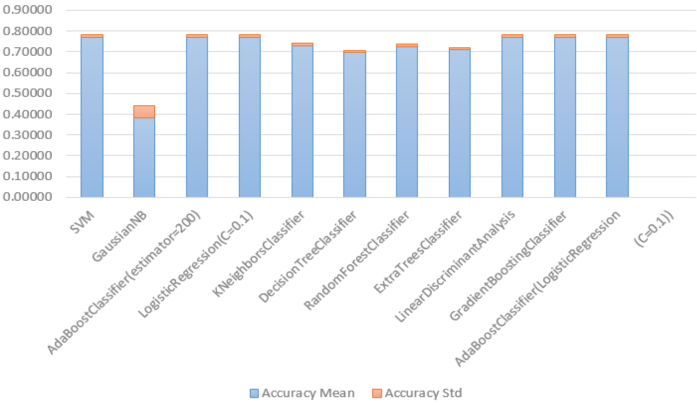
Some ensembles algorithms tend to show better results. I Used some models like AdaBoostClassifier, RandomForestClassifier,ExtraTreesClassifier, GradientBoostingClassifier,AdaBoostClassifier(LogisticRegression(C=0.1)) to train and also tune the optimal by GrideSearchCV and compare the accuracy.

3.5 Evaluation and Comparison

Algorithms use accuracy as their scoring and compare. Their accuracy mean and accuracy standard deviation indicated the performance, as the Fig7, if the model perform well, the accuracy mean should be high but the standard deviation should be low, which indicates the degree of dispersion of a data set. It is obvious LR do the best.

Fig7. Accuracy Comparison

Models	Accuracy Mean	Accuracy Std
SVM	0.76879	0.01324
GaussianNB	0.38136	0.06005
AdaBoostClassifier(estimator=200)	0.77035	0.01325
LogisticRegression(C=0.1)	**0.77038**	0.01324
KNeighborsClassifier	0.72950	0.01484
DecisionTreeClassifier	0.69835	0.00947
RandomForestClassifier	0.72568	0.01149
ExtraTreesClassifier	0.71133	0.01008
LinearDiscriminantAnalysis	0.76996	0.01324
GradientBoostingClassifier	0.77019	0.01358
AdaBoostClassifier(LogisticRegression (C=0.1))	0.76996	0.01324



Consider the LogisticRegression performs the best, and the second one is AdaBoostClassifier. Keep comparing their Precision, Recall and F1-score, as Fig8 ,in general to see, AdaBoostClassifier is slightly inferior.

Fig8. Final models Comparison

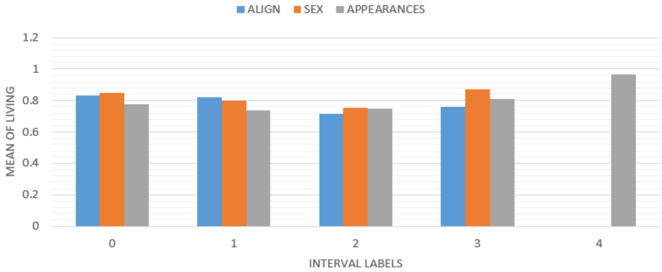
Models	Accuracy	Precision	Recall	F1-score
AdaBoostClassifier	0.77008	0.77031	0.99960	0.87011
LogisticRegression	0.77038	0.77038	1.00000	0.87030

3.6 Recursive Features Elimination

When the final model LogisticRegression is determined, using RFE method to do the features selection and get the ranking and support. The top three features are ALIGN, SEX, and APPEARANCES. Ranking the second is Name_length. It is mentioned that all the three features had already divided into intervals, and labels as one to three, APPEARANCE has fourth label.

Through calculating the average of the living characters in different intervals of the three features to visualize their association, results grouped by three features, as Fig9 shows.

Fig9. LR-RFE Selected Features



For ALIGN and SEX, 0 labels for the blank value, for APPEARANCE, 0 labels for the appearance number less than 10 times. The blank record accounts a part, because in some plots, small roles were not mentioned about whether it is a human or an animal or other kind of aliens, and that's hard to define its role nature and gender. However, we cannot ignore these hard-to-define roles, sometimes they can live longer than the major superheroes like Spider-Man.

For ALIGN, 1 means Good characters, 2 means Bad characters, 3 means Neutral role. It can been see that besides the roles of unknown nature, the Good characters could live longer.

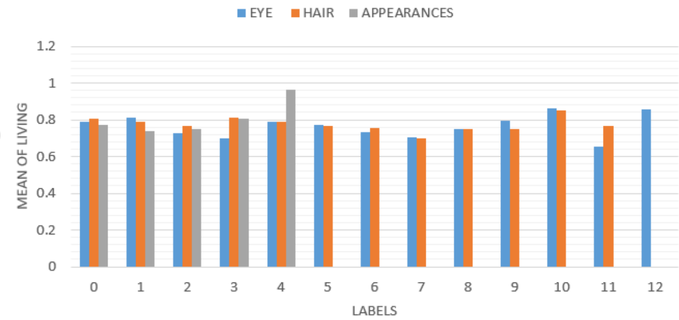
For SEX, 1 means female, 2 means male, 3 means other genders(Agender etc.). Although it seems minority sex characters can live longer, they are still minority groups and sometimes for adding rich real-life elements to comics. Therefore, besides considering the minority sex, should also consider the second ranking gender—female.

For APPEARANCE, 1 labels for appearance times between 10 to 18, 2 labels for 18 to 350 times, 3 labels for 350 to 550 times, 4 labels for more than 550 times. When appearance times achieves more than 550 then the characters seems living longer.

Then can have a temporary conclusion that the character meet up the conjunction : ALIGN(Good)∩SEX(Minority|female)∩APPEARANCE(>550 times) can be most likely a living role.

AdaBoostClassifier performs also well, then we can use the same way to see what are the top three attributes impacting AdaBoosting most. After running the RFE method, the top three features are EYE, HAIR, APPEARANCES. The relation presents as Fig10:

Fig10. AB-RFE Features Selection



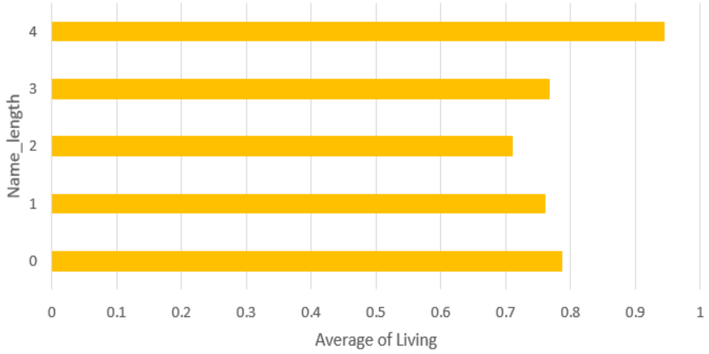
From the Fig10 we can see, the type 10 of HAIR and EYE would live longer, type 10 is multiple color hair/eyes. The APPEARANCE times should also more than 550.But in reality, with colorful eyes and hairs characters didn't have chances to show up more than 150 times, then limits the appearance times interval between 18 to 350. Then the second temporary conclusion could be like: EYE(Variable/Multiple/Compound) ∩ HAIR (Variable/Dyed) ∩ APPEARANCE(18~350 times), which means large probability the character alive

IV. Results & Analysis

From the experiment, it seems that we can get the conclusion we want. But from a realistic perspective and from audience perspective, the reference value of APPEARANCE times is not very significant, because the more a character can show up the more likely the role will not deceased in recent. Then what others can be refered?

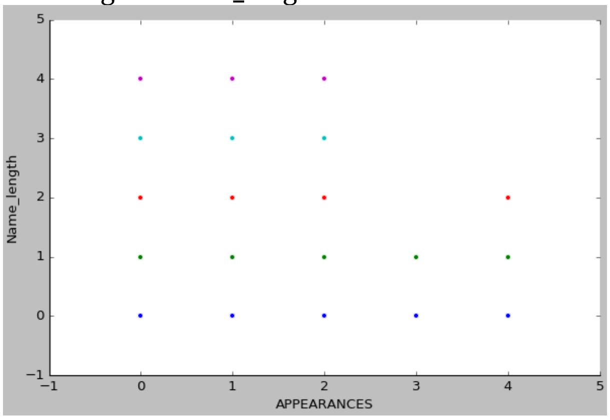
Settle for second best, both the LR and AB models ranks Name_length as the second impacted feature. What surprised me is that if the length of name is longer, maybe more than 40 characters, then the role is possible living, as Fig11 to see.

Fig11. Name_length & ALIVE



I supposed that is it that the name length is too long for the authors to remember they have created such roles, then such roles have less chance to appear in the comic plots, they can save life. However, the length of their names doesn't affect their appearance quite a lot, as Fig12 shows:

Fig12. Name_length & APPEARANCE



Therefore, my conclusion of the most likely living characters should meet one of the following conditions:

1. Character should be a Good role
   Character sex should be minority or female

2. Character EYE should be multiple/variable /compound color eyes
   Character HAIR should be variable/dyed color hair

If one of the above conditions are met, then the third features you can choose APPEARANCE or Name_length. If you choose APPEARANCE times, select characters with more showing times. If you choose Name_length, select with a long name. Then this time, you can have a idol who is most likely living to the end.

V. Discussion

In the future work, I will use some *Hot deck Imputation* or *Combinatorial Completer* methods to fill the blank values. This time I kept them as Unknown and labeled as 0 (most of them).

Also I hope deepen the research on some new features like Skills, Clothing Color, Education etc., to record roles' special stunt, dressing style and learning ability. I can also try to learn the relation between characters first appeared Season with the ALIVE result in the future work.

VI. Conclusions

This paper, I present an experiment to learn the law of alive characters in Marvel Comic. The two models LogisticRegression and AdaBoostClassifier perform well in this binary classification problem, both of them can achieve 77% accuracy. Features like good identity, minority gender or female gender, colorful hair, colorful eyes, appearance times or even name length could influence a role's lifespan.

From the paper, this time our crazy Marvel Comic fans can re-select their idols wisely. If select wisely, you will not cry in the cinema but laugh to the end. After seeing through Marvel's tricks, Marvel Comic team could change with new ideas and do the opposite, for example saving some bad characters life who with short name, blue eyes, black hair, the out of unexpectedness design will attract more and more fans.

Reference

[1] FANDOM Comic Community,Marvel Database[DB/OL], March 2005, [Jan14th,2019], http://marvel.wikia.com/wiki/Marvel_Database

[2] FiveThirtyEight, FiveThirtyEight Comic Characters Dataset[DB/OL], August 24,2018,[Jan14th,2019], https://www.kaggle.com/fivethirtyeight/fivethirtyeight-comic-characters-dataset/home

[3] Rasmussen, C. E. (2004). Gaussian processes in machine learning. In *Advanced lectures on machine learning* (pp. 63-71). Springer, Berlin, Heidelberg.

[4] Flach, Peter. (2012). The Art and Science of Algorithms that Make Sense of Data,pp.49-57,pp344-357

[5] scikit-learn dvelopers(BSD License),[P/OL], 2017-2018,[Jan14th,2019], https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_with_cross_validation.html#sphx-glr-auto-examples-feature-selection-plot-rfe-with-cross-validation-py