An Analysis on Social Biases in Comic Characters

Morgan McCarty

Background on the Dataset

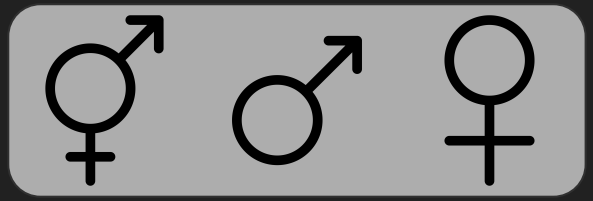
- The data set has two portions
  - Marvel (from the Marvel Fandom)
  - DC (from the DC Fandom)
- There are 13 data fields of which 8 are fully categorical (technically 9 including the name), 2 are partly (years), and 3 are metadata (including the name)
- The DC and Marvel datasets are exactly the same minus two tiny differences in the name of one column and the formatting of dates
- It was created by FiveThirtyEight for an analysis on women in comics in **2014**

https://github.com/fivethirtyeight/data/tree/master/comic-characters

Why Should we Care About Biases in Comics?

- The Marvel Cinematic Universe has grossed over 30 billion USD
- Batman has grossed over 29 billion USD
- Spider-man has grossed over 25 billion USD
- What was once a relatively niche topic is now a mainstream facet of life
- Children grow up idolizing superheroes

Step 1 - Biases Before Learning

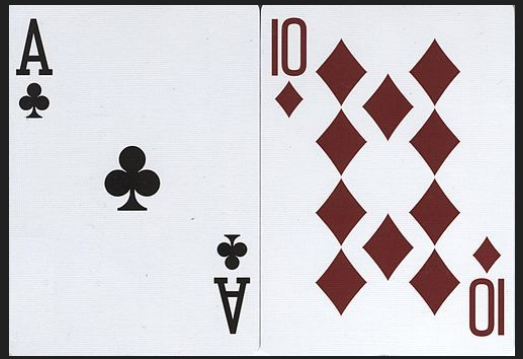


- Before we even discuss the models and their architecture as well as what they can show we need to look at some data
- Combined between both Marvel and DC there are:
  - 16421 Male Characters
  - 5804 Female Characters
  - 68 Characters who fall under a different sex/gender identity (as well as 979 NaN characters)
- Immediately it's clear that there is a massive gender gap within comic book characters (2.829x as many men as women, and for characters who are neither Male nor Female the gap is even larger)
- As of 2014 there were only **2** transgender characters in Marvel or DC

Step 2 - The Model (High Level)

- How can we design a model which can look at various social biases without introducing biases of its own?
- Random Forest:
  - Easily interpretable and easily visualizable
  - Can build a lot of these to see what causes different results

Step 3 - Data Preprocessing



- Unfortunately the data is not perfect as such there are three main steps
  a. Convert the dates to a standardized number between datasets
  b. One-hot encode any categorical feature that is not the target feature and label encode the target vector
  c. Replace NaNs with the mean value for the class
- With seven categorical features, and three datasets, 21 Random Forests were made
- As such a standardized process is essential
- If there was a NaN target feature, its row was removed for that model

Step 4 - Training

- As with any set of models being built disjointly, but at the same time, it was very important to make sure the data was properly separated
- Separated data two ways
  - 20% of the data was immediately separated away to make a validation set
  - On the remaining 80%: GridSearchCV with StratifiedKFold (the number of folds depended on the feature)
- Data is doubly protected from leaking between the train and test

Step 4 Continued - Grid Search

- Random Forests have several hyperparameters
- Five key hyperparameters were chosen to improve the performance of the model
  - “n_estimators”, “min_samples_split”, “min_samples_leaf”, “criterion”, “max_depth”
  - Key points: “criterion” is the evaluation function for splits, “n_estimators” is the number of Trees in the forest
- More hyperparameters or a larger grid range of these could have been considered, but the time to evaluate the best parameters rose very quickly

Step 4 Continued - StratifiedKFold

- StratifiedKFold was applied to reduce the bias and variance within the models
- The number of folds was set to `min(n, 5)` where “n” is the number of class examples of the smallest class
  - If “n” were less than 2 (i.e. 1) then the StratifiedKFold and GridSearch had to be skipped (this occurred on several of the categories)

Step 5 - Results (Accuracy)

- While more metrics were looked at within the code accuracy is the easiest to view across all 21 Random Forests
- Marvel: ALIGN - 0.6030, SEX - 0.7314, EYE - 0.4735, HAIR - 0.3033, GSM - 0.8333, ALIVE - 0.7679, ID - 0.5781
- DC: ALIGN - 0.6243, SEX - 0.6937, EYE - 0.4602, HAIR - 0.2972, GSM - 0.8461, ALIVE - 0.7519, ID - 0.6530
- Combined: ALIGN - 0.6062, SEX - 0.7481, EYE - 0.4038, HAIR - 0.2954, GSM - 0.7741, ALIVE - 0.7707, ID - 0.5714

accuracies above .7 are highlighted

Step 5 Continued - Interpreting the Results

- The numbers don’t show much without some context
- ALIVE is the only binary classification topic (i.e. Living Characters or Deceased Characters) [though as discussed earlier SEX is divided fairly binarily]
- EYE and HAIR both have the most classes (26 and 28 respectively)

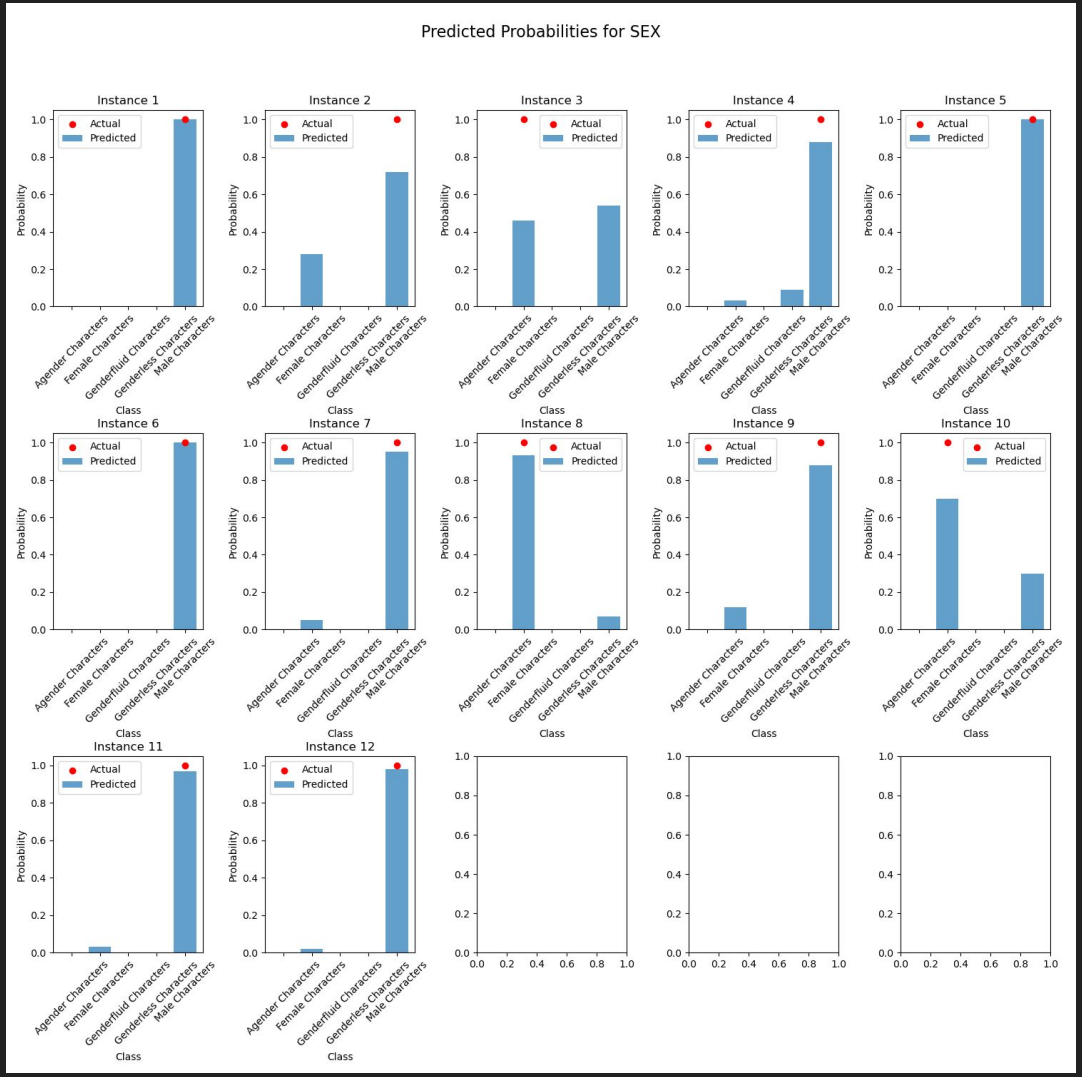
Step 5 Continued - Interpreting the Results (SEX and GSM)

- For the purpose of analyzing biases, SEX and GSM are the two most important features
- Let’s look at some charts!

Step 5 Continued - Interpreting the Results (SEX and GSM)

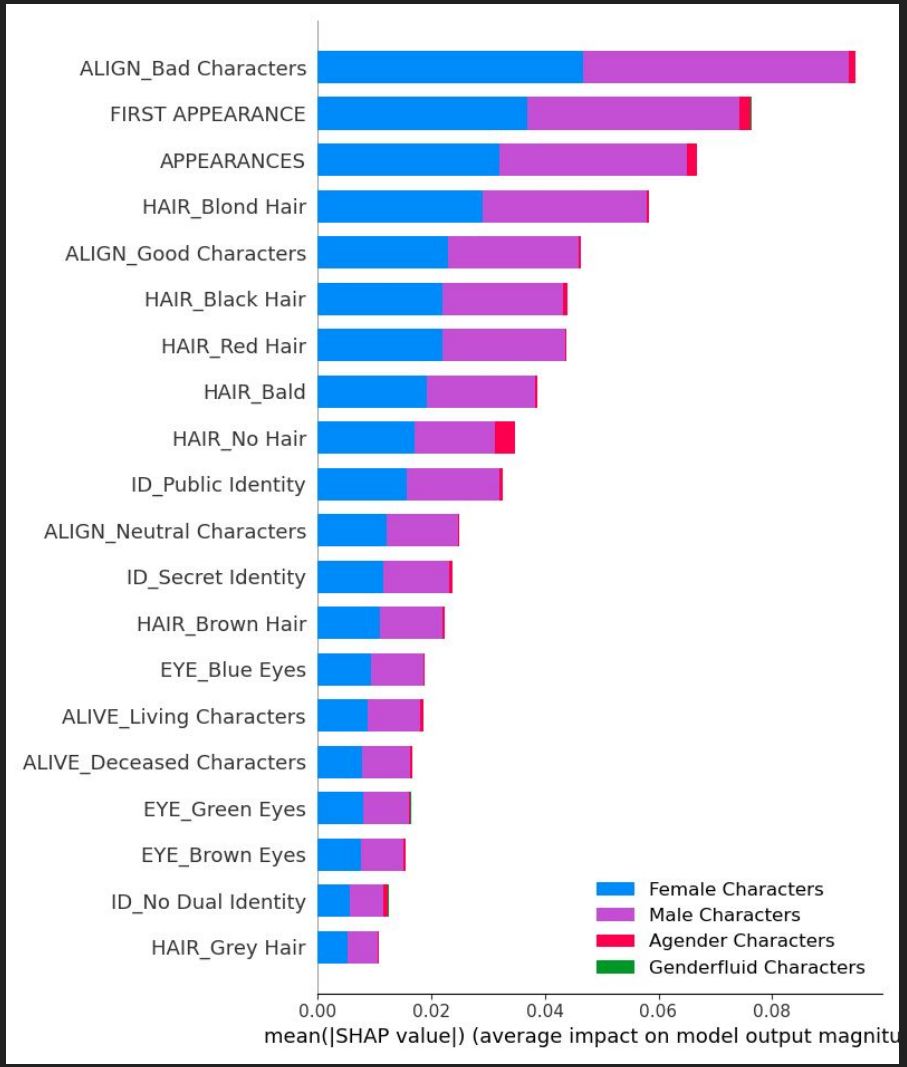
- These are 12 random instances in which you can see the probability distributions created for each character
- Male is dominant throughout as expected

n.b. there was an unfortunate error in the name transcription for Marvel and DC which caused the name to be pretty much random - combined was just left as “instance”



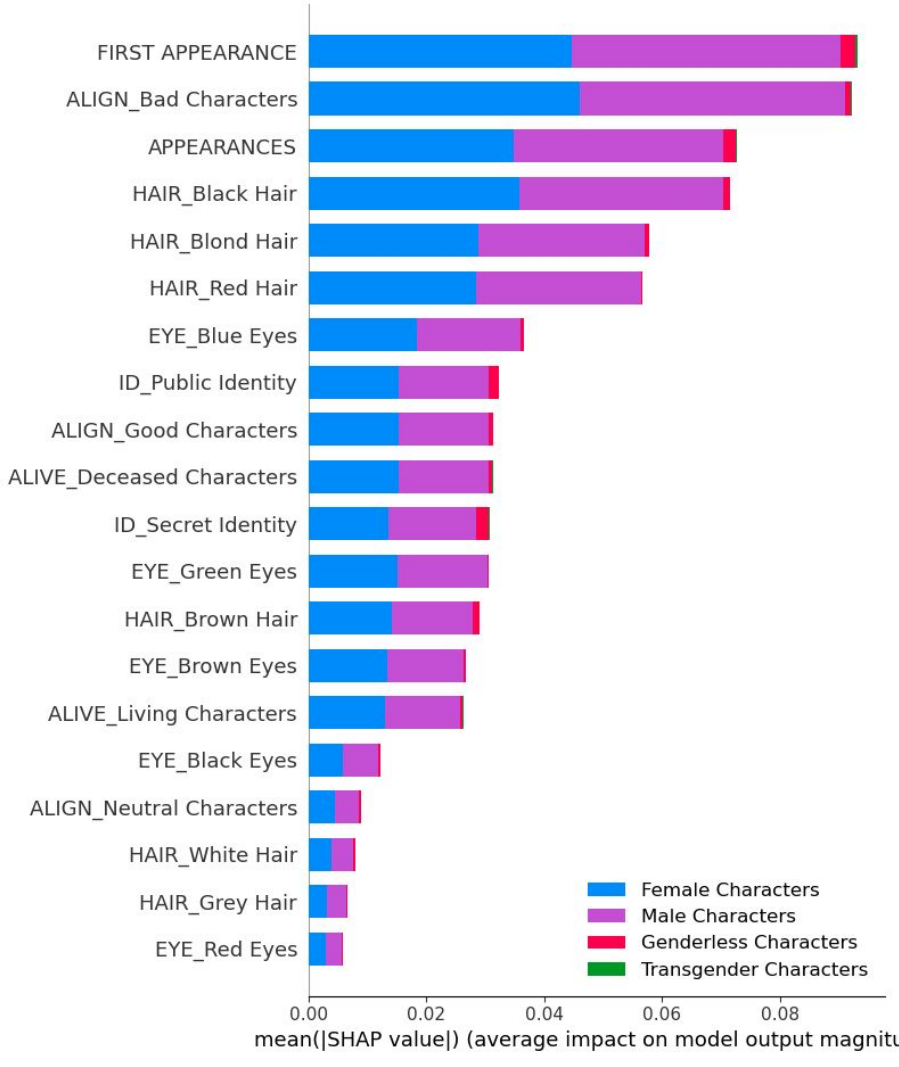
Step 5 Continued - Interpreting the Results (SEX and GSM): Marvel

- Almost immediately something about this chart jumps out
- The biggest contributor to Agender Characters is having no hair
- Looking at the data this can be partially explained by the presence of the symbiotes (aliens) [which also correlates to the villain presence of agender]
- However this reduces non-binary representation within Marvel to hairless individuals



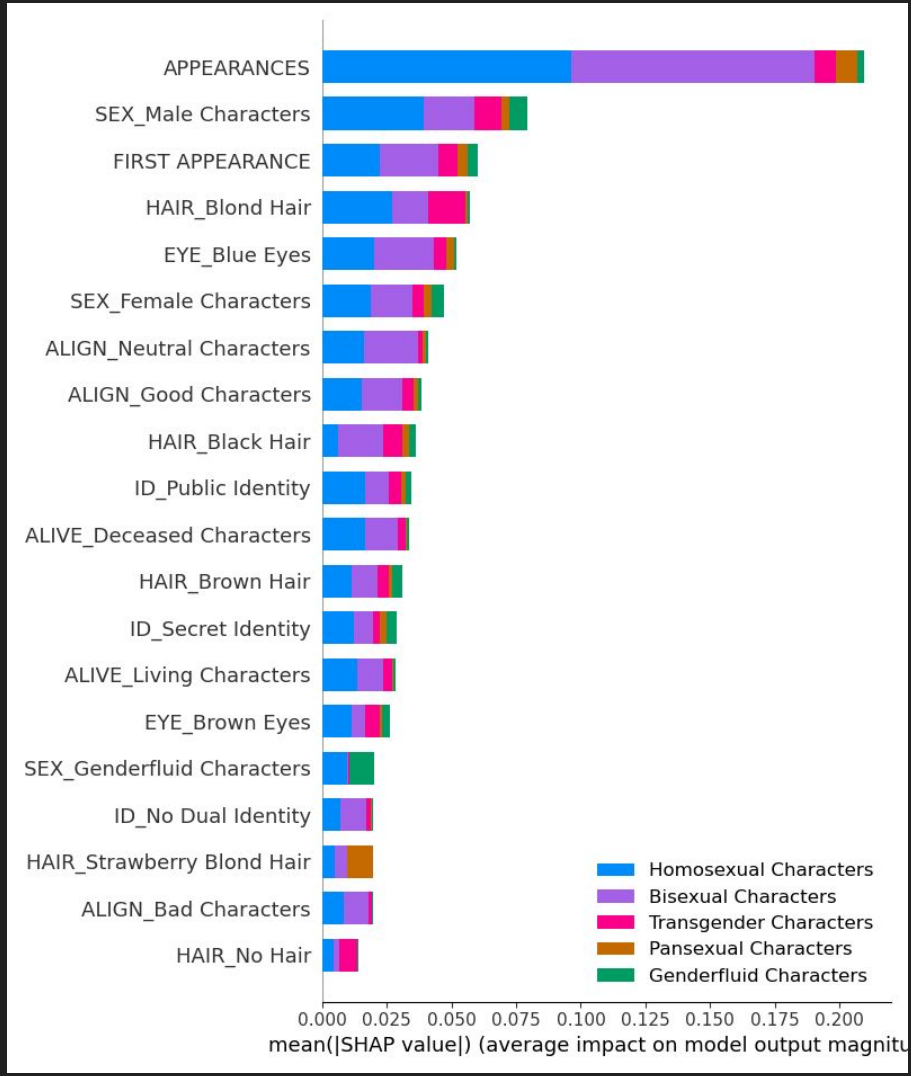
Step 5 Continued - Interpreting the Results (SEX and GSM): DC

- DC graph gives us another chance to consider similar things to the Marvel graph
- Value shown on graphs is mean absolute value of the importance of a feature for its class
- Hair color is a strong indicator for sex in DC
- Features which should be strongly discorrelated to sex seem very important here
- Some features do make sense, however, like the first appearance having a strong role due to the increased presence of female characters as time has passed



Final Notes on Results (GSMs in Marvel)

- Gender and Sexual Minorities are another important consideration
- Very few within the dataset
- Two of the highest contributing factor to GSMs: number of appearances, first appearance
- GSMs a recent change for Marvel
- High mean absolute value of appearances shows large room for improvement



Ending Notes

- The data and predictions from the data show a few key points
  - Marvel and DC have been improving over time with regards to inclusivity, but both have a **long** way to go until they are properly inclusive of women and GSMs
  - SEX and GSM as categories both are too heavily influenced by non-gendered physical characteristics (eye color, hair color)
- As such it holds that comics are very gender and inclusivity biased