An Analysis on Biases in Comic Characters

Morgan McCarty

15 August 2023

1. What question is being answered?

   - The question here is what biases exist in comic characters and how can we determine them?
   - This especially focuses on those biases which are not clear and plain to see.
   - While it might be obvious that male characters vastly outnumber female characters in the comic world, the exact consequences and the extent of this are relatively hard to determine.
   - For example: in comics certain colors of hair are some of the leading indicators as to what gender a character is (see charts).

   (a) What is the target - categorical / continuous / group-id (for unsupervised datasets)

      - The target of the models created here are all of the categorical fields in the dataset. Each field was treated as a separate target for a unique Random Forest model.
      - For reference these fields are: "ALIGN" (the ideological alignment of good, evil, and neutral), "SEX" (male, female, etc.), "EYE" (color), "HAIR" (color), "GSM" (gender and sexuality minorities), "ALIVE" (living status), and "ID" (hidden or known identity).
      - The dataset is available at this link.

2. What techniques are being used for modeling?

   - The primary technique used is Random Forests. Random Forests were chosen as they are interpretable which is very important when dealing with data which could have strong societal implications.

   (a) Is there a progression from high bias to low bias models?

      - The inital models were a straight implementation of an Random Forest without any hyper parameter tuning. This was done to get a solid baseline of what could be expected from the data.
        e.g. the model for hair and eye color performed terribly as there are so many categories for each, the data is sparse, and there is no clear correlation between the color of a character's hair (or eyes) and anything else (or so it seemed...).
      - After the initial models were created grid-search with Stratified K-Fold cross validation was used to tune the hyper parameters of the Random Forests. This reduced the bias of the models and increased their performance.

   (b) Are justifications provided for using specific models?

      - As stated before Random Forests were chosen as they are interpretable and can be used to determine feature importance.

3. Complexity of the dataset:

- The dataset itself is not terribly complex, however many of its features have so many categories that the least frequent categories have only a few examples (see the beginning of the notebook pdf where the data was counted).

(a) Are raw features used or is feature engineering applied?

- Raw features were not used as it was necessary to convert the categorical features into numerical features. The target feature was converted into a numerical feature using a LabelEncoder and the other features were converted into numerical features using a OneHotEncoder.
- Additionally some of the fields were dropped as they did not seem to have relevence to the task at hand.
- Finally, the "FIRST APPEARANCE" field was converted into a numerical feature through the use of a custom function.

(b) How is the dimensionality of the dataset handled?

- The dimensionality of the dataset was handled through the use of a OneHotEncoder. This was done as the categorical features did not have any inherent order to them and using a LabelEncoder would have introduced a false sense of order to the data.
- A LabelEncoder was necessary for the target features however as otherwise it would have been impossible to train the model.

4. End-to-end implementation of the prediction pipeline:

- The prediction pipeline is as follows:
  (a) Load the data
  (b) Apply the preprocessing steps to create three separate datasets: Marvel, DC, and Combined
  (c) Create a Random Forest model for each of the target features
  (d) Apply GridSearchCV to each of the models to tune the hyperparameters
  (e) Evaluate the models
  (f) Create SHAP plots for each of the models
  (g) Create probability plots for each of the models

(a) Implementation done completely with pre-processing

- The preprocessing steps are all done in the notebook.

(b) No leakage between training / test sets

- As everything is done with GridSearchCV there is no leakage between the training and test sets. Prior to that separation a separate test set was taken for validation as well.

5. Evaluation strategies:

(a) Correct evaluation methodology used for evaluation that reflects dataset nuances

- All nuances in the data were looked at to the best extent to improve the quality of the models.

6. What metrics are used for tuning models?

- The metrics used for tuning were GridSearchCV and Stratified K-Fold cross validation.

(a) Correct metric selection for hyperparameter tuning

- As the GridSearchCV was used to tune the hyperparameters the metric used was the accuracy score, other scores were also analyzed.

7. Visualization of results:

   - Two main chart types were created, SHAP plots and probability plots.

   (a) Charts reflecting model performance

      - The probability plots reflect the performance of the models. Additionally the SHAP plots help explain why the models performed the way they did.

   (b) All relevant metrics visualized during training and testing

      - The accuracy score (as well as other major scores) were visualized during training and testing.