JOHNS HOPKINS UNIVERSITY

EN.625.742
Theory of Machine Learning

The MUIR ISLAND Team

Analysis of the X-men Clairmont Run Dataset

Kevin Bohlin
Shahvaiz Hanif
Nick Lines
Marius Tudor

The story so far…

1.

The Claremont Run Dataset

A quick review

[2]

The Claremont Run Data

- Chris Claremont saved the X-Men comic and over 16 years wrote about 8,360 pages of X-men stories.
- Claremont's X-Men introduced a diverse cast struggling through societally relevant moral issues.
- From these comics, the Claremont Run group extracted data about comic covers, collaborators, character features, locations, and so forth [1].

2.

Predicting comic collectibility via Regression

Can we guess how many certifications a Claremont issue will see?

Introducing Comic Book Grading Data

- Unfortunately the Claremont Run Dataset does not include any metric for popularity or success of an issue.
- We mined data from the Certified Guaranty Company (CGC)'s comic registry for Claremont X-men issues.
- This data shows counts of each grade certified for each issue.

GRADING SCALE

10.0 Gem Mint
9.9 Mint
9.8 Near Mint / Mint
9.6 Near Mint+
9.4 Near Mint
9.2 Near Mint-
9.0 Very Fine / Near Mint
8.5 Very Fine+
8.0 Very Fine
7.5 Very Fine-
7.0 Fine / Very Fine
6.5 Fine+
6.0 Fine
5.5 Fine-
5.0 Very Good / Fine
4.5 Very Good+
4.0 Very Good
3.5 Very Good-
3.0 Good / Very Good
2.5 Good+
2.0 Good
1.8 Good-
1.5 Fair / Good
1.0 Fair
0.5 Poor

Fusing the data

- **8** Data tables
- **229,010** Total cells
- Different issue coverage

- **1** Feature matrix
- **455** Possible features
- **199** Observations

- **80%** Training
- **20%** Testing

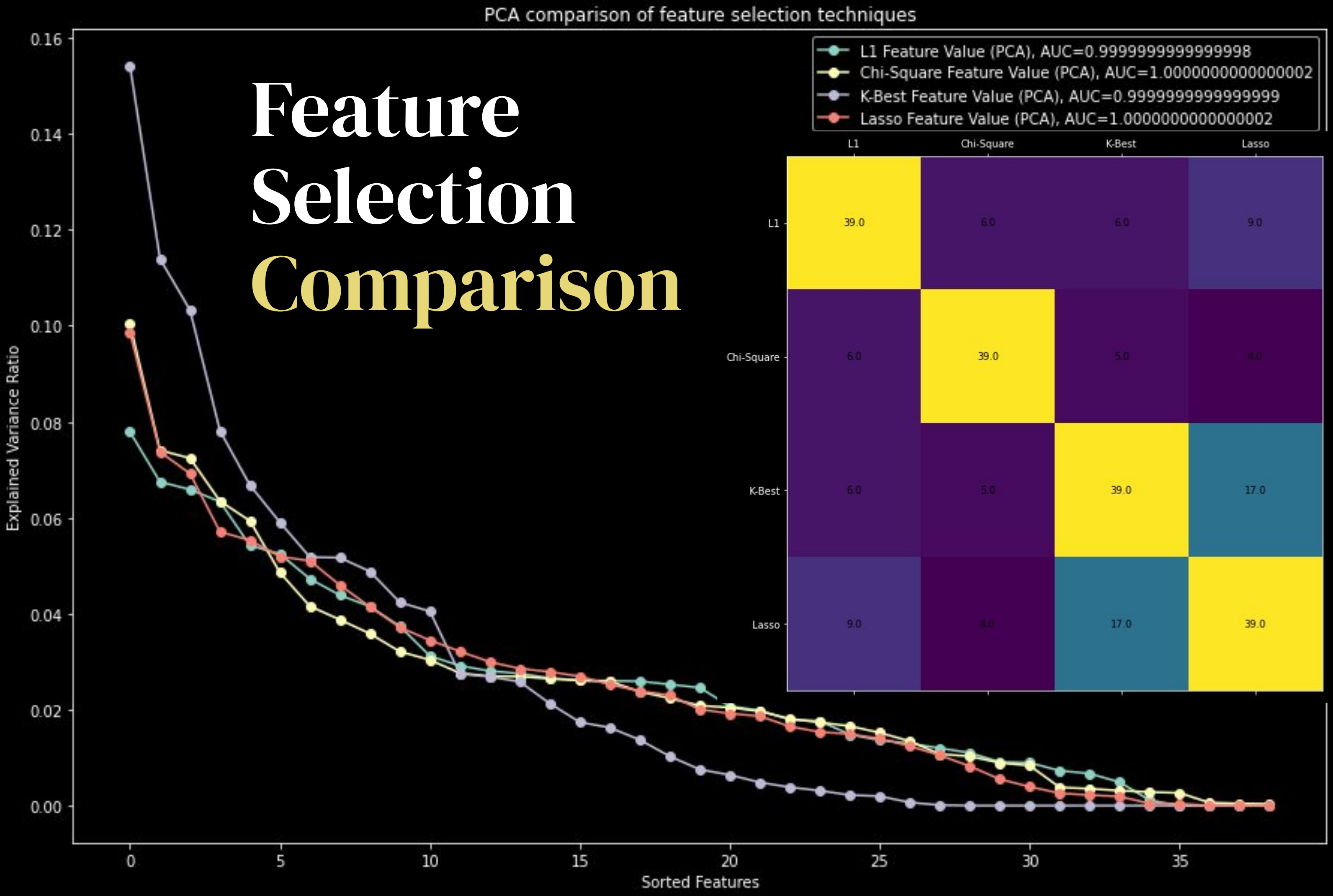
Feature Selection

Commonly used for Classification tasks

Commonly used for Regression tasks

L1-SVM Linear support vector machine penalized by L1 norm, similar to Lasso.	**Chi-Square** Identifies dependence between features. Features must be non-negative.
K-Best ANOVA f value Weeds out features with similar distributions (means).	**Lasso** Drops superfluous features coefficients to zero in linear regression model.

Feature Selection Comparison

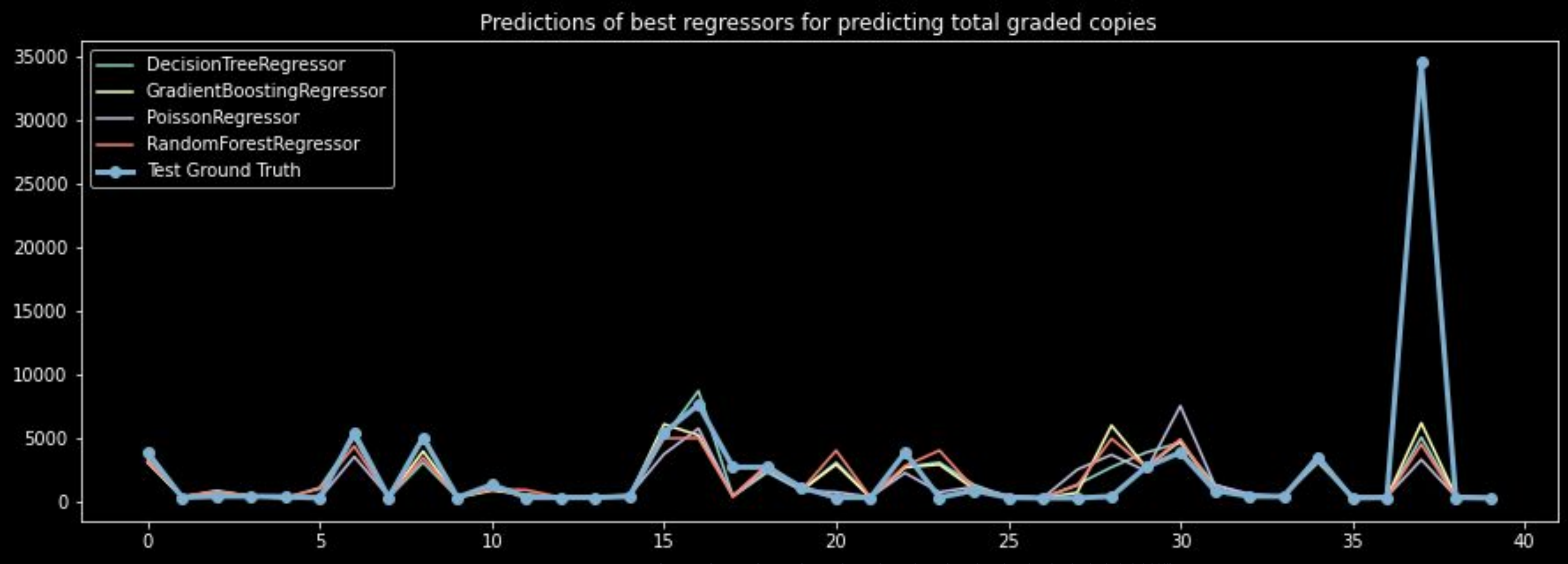


Lasso Features

'cover_artist_John Romita',
'cover_artist_Paul Smith',
'cover_artist_uncredited',
'cover_features_Banshee',
'cover_features_Captain America',
'cover_features_Professor Xavier',
'cover_features_Sabertooth',
'cover_features_Sebastain Shaw',
'depicted_Gambit = Name Unknown',
'editor_Marv Wolfman',
'editor_in_chief_Archie Goodwin',
'editor_in_chief_Marv Wolfman',
'issue',
'location_Baxter Building',
'location_Blackbird, towards X-Mansion',
'location_Boat, Drake Passage, South of Cape Horn',
'location_Cassidy Keep, Ireland',
'location_Circus Wagon in Orbit',
'location_Disco, Delano Street, Lower Manhattan',
'location_Downtown Calgary, Alberta',
'location_Downtown Calgary, Alberta, Canada',
'location_Hollywood Mall',
"location_Jean Grey's Apartment, Greenwich Village, NYC",
'location_Jinguchi Maru Ship, Drake Passage, South of Cape Horn',
"location_Magneto's Underground Base, Antarctica",
'location_Morlock residence under New York',
'location_Pryde Residence, Deerfield, Illinois',
'location_Reisz Residence, Cairo Illinois',
'location_Spaceship',
'location_Stampede Park, Calgary, Alberta*',
'narrative_Editor narration',
'penciller_John Byrne',
'penciller_John Romita Jr.',
'penciller_Paul Smith',
'speech_Angel = Warren Worthington',
'speech_Gambit = Name Unknown',
'speech_Rogue = Name Unknown',
'thought_Ariel/Sprite/Shadowcat = Kitty Pryde',
'thought_Marvel Girl/Phoenix = Jean Grey'

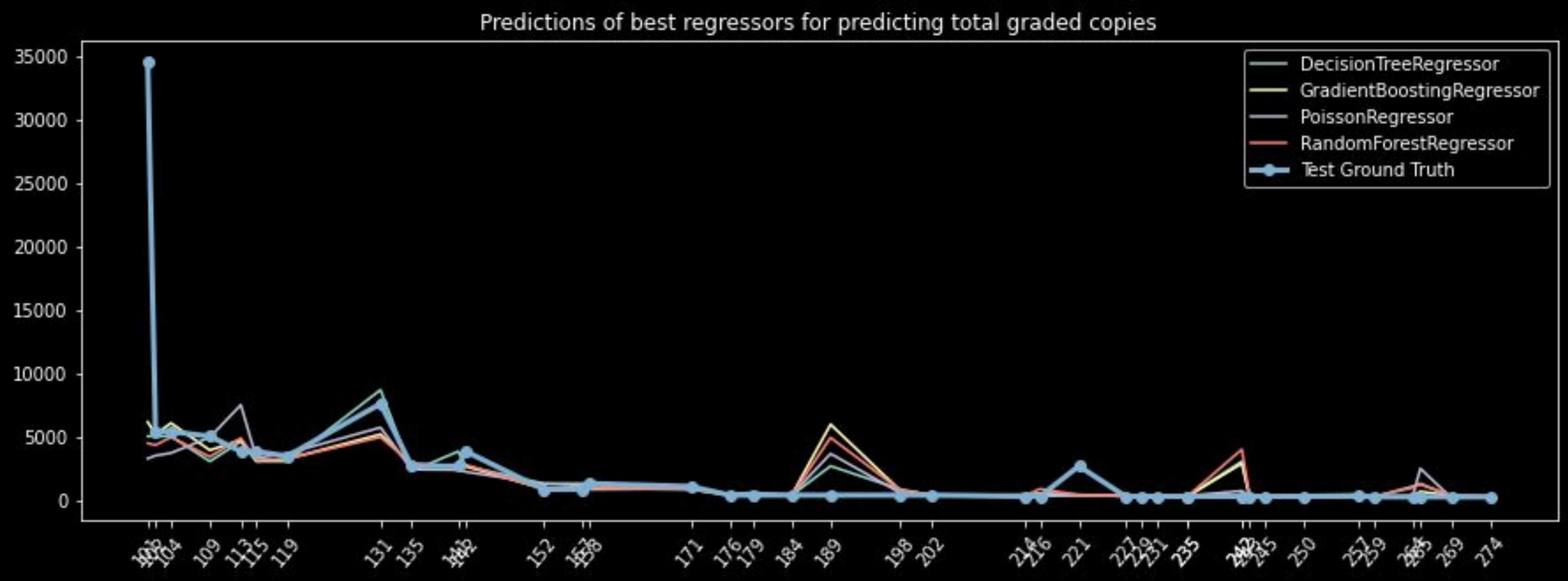
Applying lots of regressors!

Out of **46** regression techniques tried, the best performing in the RMS Error-sense were Gradient Boosting, Decision Tree, Random Forest, and Poisson Regression.



Why so bad?

The first Claremont issue is WAY more popular than any others, and it ended up in the test-set. This outlier really messes things up. But we correctly predict a rise in Issue 131.



Issue embedding for a simple recommender

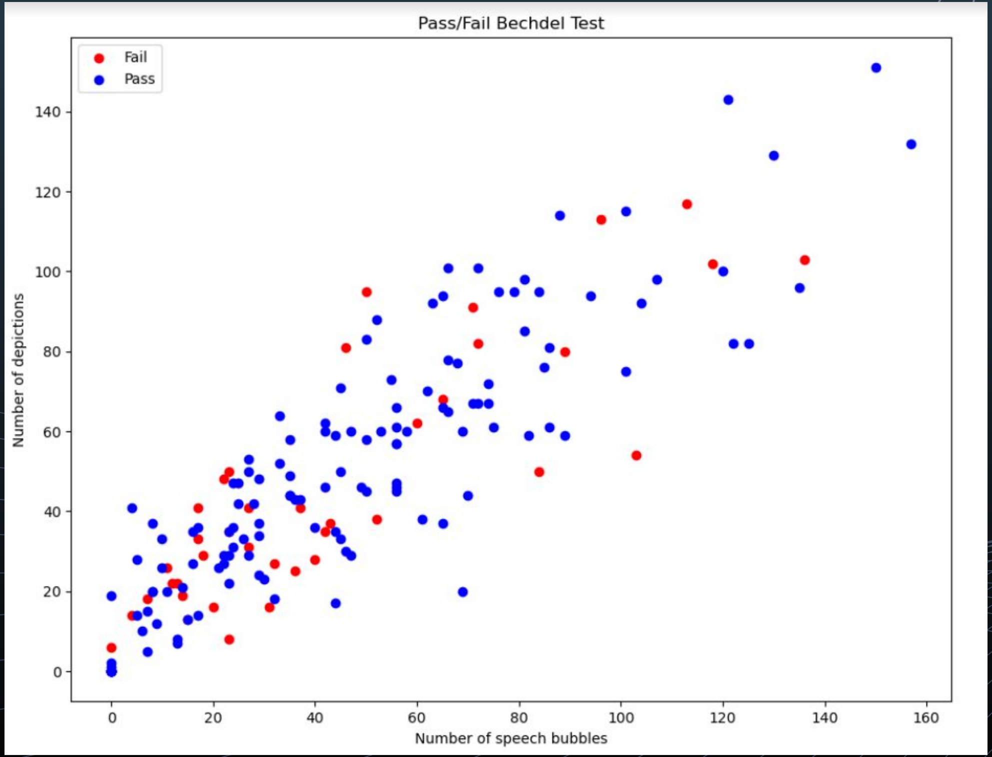
3.

Bechdel Test Predictions

Bechdel Classifier

(1) Can we predict whether an issue passes Bechdel test based on depictions of female characters?
   (a) Using female speech as predictors
   (b) Using both male and female speech as predictors
(2) Constructed several K-Nearest Neighbor models
   (a) N_neighbors = [3, 5, 7, 10] with 3-fold and 5-fold cross validation
(3) All models gave similar results
   (a) Average CV error: ~30%

Why so poor?



(1) Speech and depiction highly correlated
  (a) Makes sense
(2) No natural clustering of pass/fail with speech/depiction
  (a) When considering female only, male only, and both together
(3) Why?
  (a) Proliferation of female heroes

Point-Biserial Correlation

(1) Measures correlation between continuous predictor and binary response - equivalent to Pearson correlation
   (a) r = 0.04 (no correlation)

(2) Split predictor data (X) into two groups
   (a) Predictor data corresponding to "pass" (1) and "fail" (0)

Mean of "pass" data

Mean of "fail" data

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}}$$

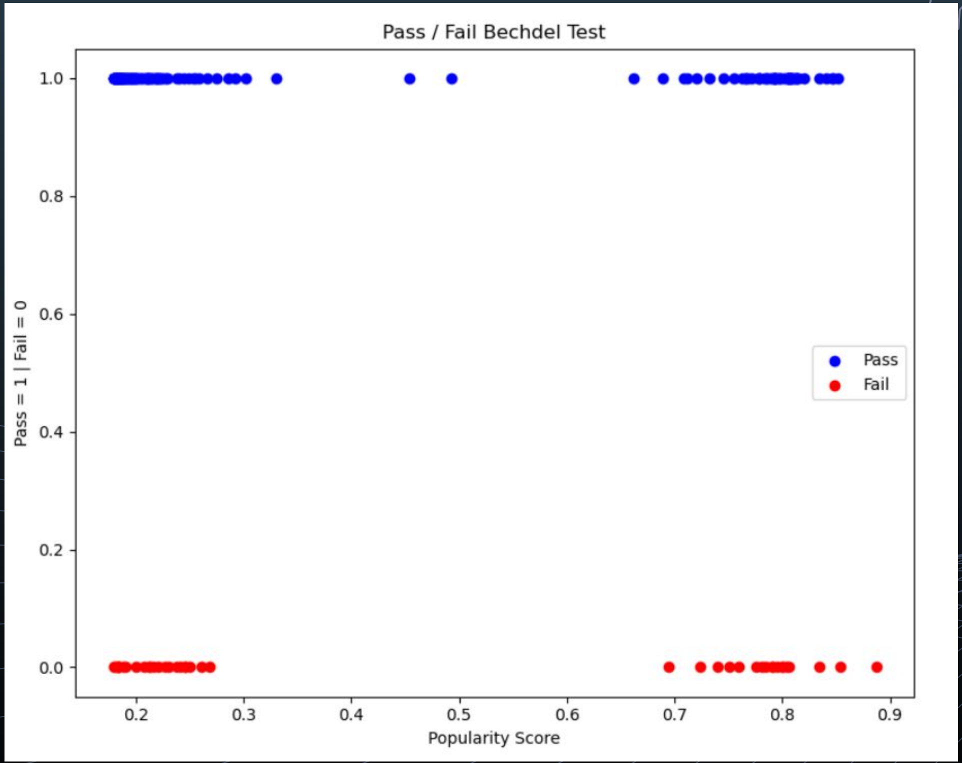
$$s_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Does Bechdel Inform Popularity/Collectability?

(1) Need a normalized “popularity” score based on collectability
   (a) Assumption: Linear relationship between chronology and number of gradings
   (b) Find OLS linear fit, compute residuals, normalize via scaled sigmoid function (for convenience)
      (i) Results in uniformly distributed collectability scores

(2) Can score predict Bechdel result?
   (a) Hypothesis: High Popularity/Collectability => Pass Bechdel Test

Logistic Regression Classification



(1) Trained/tested with 85/15 split
  (a) Test Error: 39%
  (b) Avg CV Error: 37%
(2) Only slightly better than random guessing
(3) Point-Biserial Correlation: r = -0.05
  (a) No relationship

4.
Character Classification

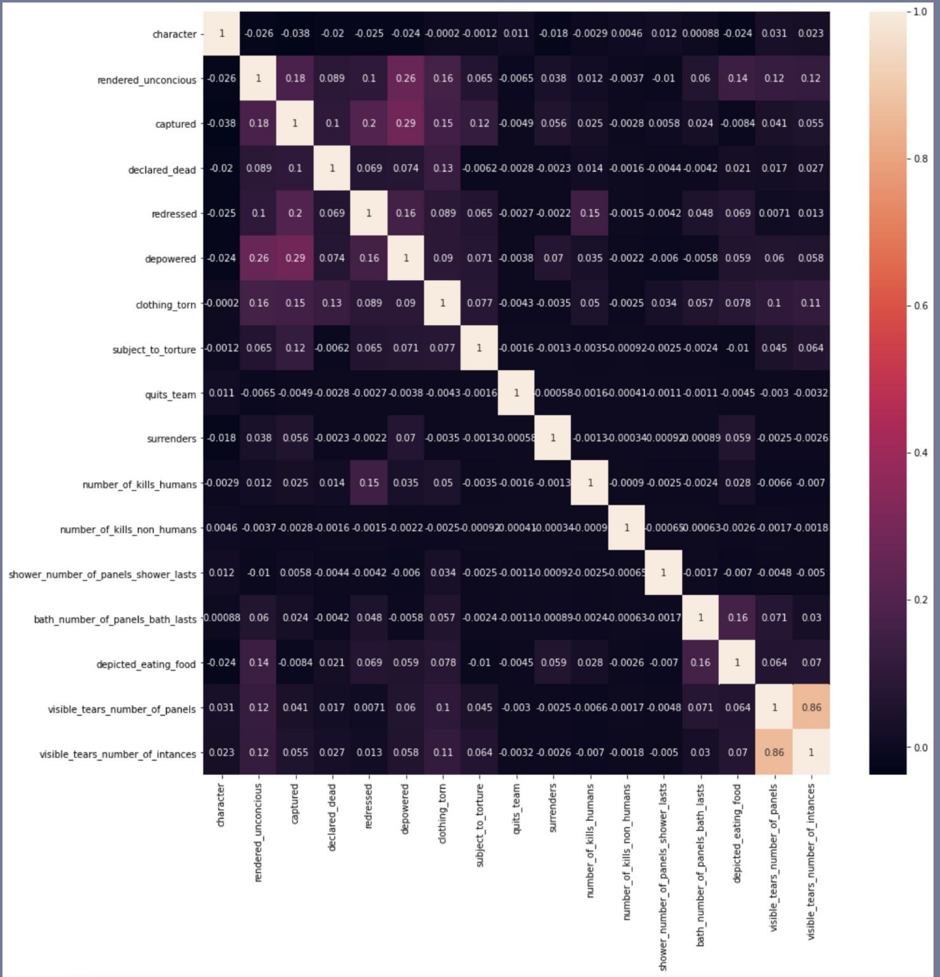
Classification of characters using Random Forest

Independent Variables

16 variables were used to predict the X-Man character that is depicted in the issue.

Correlation Matrix

The correlation matrix shows very low correlation values between the independent variables, indicating no issues with multicollinearity.



Classification of characters using Random Forest

75/25 Test Train Split

Using a 75/25 test-train split, we achieved an accuracy of 0.9373

Random Forest

Using a random forest classifier, we predicted the X-man character from 16 variables. The model was tested using 10 fold cross validation.

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Average
0.9026	0.9234	0.8717	0.8931	0.9477	0.8646	0.9002	0.9192	0.9240	0.8904	**0.9038**

Results

The model demonstrated strong prediction power. There is limited overfitting as indicated by a small spread between the .9373 accuracy and the 10-fold CV accuracy value of .9038. The model does not suffer from multicollinearity.

5.
Possible Future Work

Possible Future Work

POPULARITY REGRESSION

- Remove outlier comic issues, linearly adjust scores, and re-attempt regression prediction of issue popularity
- Restructure the popularity regression problem as a time series analysis problem, forecasting or backcasting popularity

BECHDEL TEST ANALYSIS

- Build more holistic models incorporating data that'd ostensibly be disregarded as "irrelevant" to Bechdel Test result
- Try to find and use more accurate data on comic issue popularity

Questions?

[4]

References

1. Data: https://www.kaggle.com/jessemostipak/uncanny-xmen, https://www.claremontrun.com/ ,
2. The Fall of the Mutants cover (p. 4) By Chris Claremont - Marvel Database Wikia, Fair use, https://en.wikipedia.org/w/index.php?curid=53741038
3. The Certified Guaranty Company (CGC), comic registry. https://comics.www.collectors-society.com/registry/comics/
4. https://knowyourmeme.com/photos/892897-x-men

All unreferenced figures were original work. Copies of these and our Python Jupyter Notebooks available upon request.

APPENDIX

Material discussing the original data and our previous work on it.

Show Me the Data!

- *character_visualization* - counts of character speech, thought, narrative, and visual depictions
- *characters* - descriptions of character actions (dies, captured, changes outfit, etc.)
- *xmen_bechdel* - whether or not an issue of Uncanny X-Men met the Bechdel test
- *comic_bechdel* - whether or not an issue of a non-Xmen comic met the Bechdel test
- *covers* - data visible on the comic's cover
- *issue_collaborators* - data about other collaborators involved in creating the issue
- *location* - locations that appear in each issue

Data Assumptions

(1) Assumptions/Biases in the data?
   (a) Ambiguity in measuring qualitative variables (i.e. what exactly constitutes a character being “redressed” or “rendered unconscious”)
   (b) Some columns included are entirely empty - maybe this isn’t all the data?

(2) What other data would be nice to have in this dataset?
   (a) A metric for which issues were most popular (could glean what features of Claremont’s comics made them so popular)

Exploratory data analysis

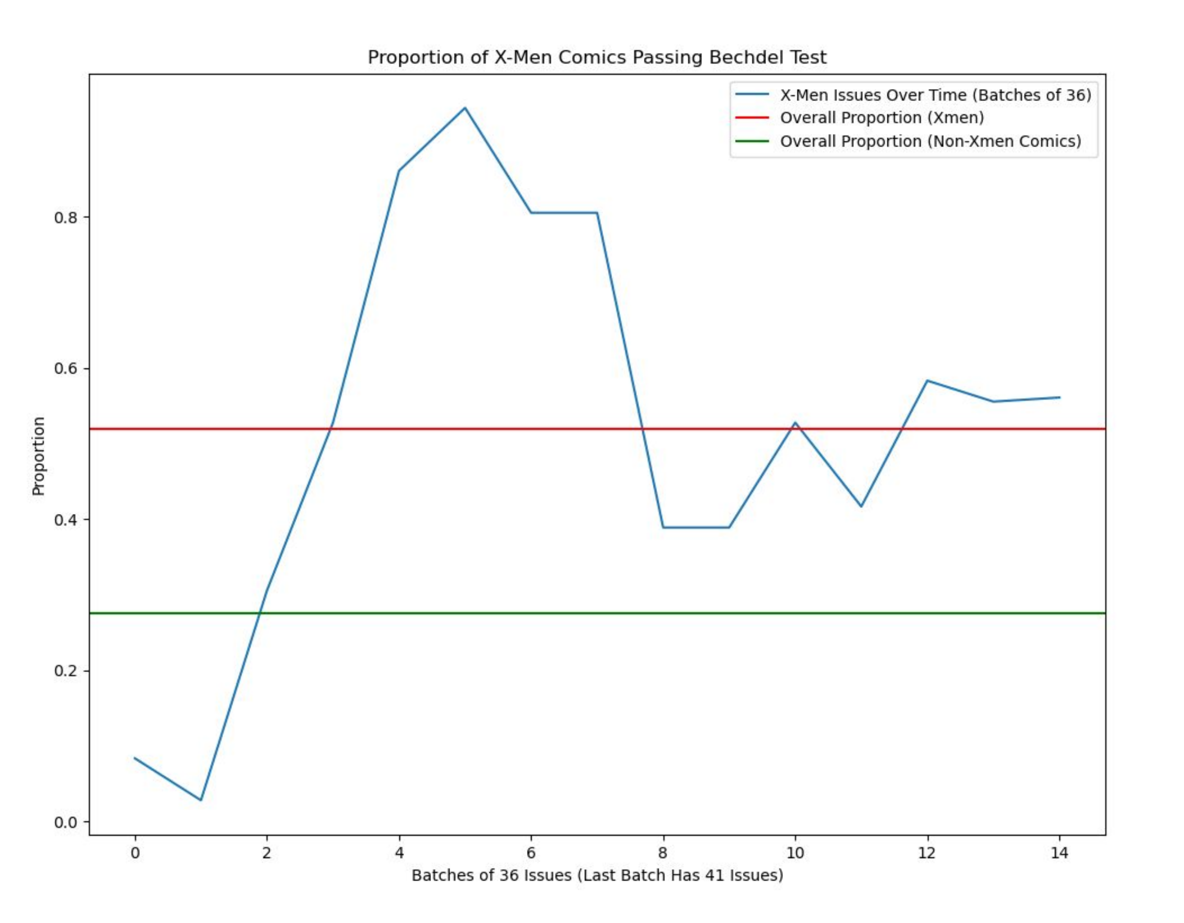
Some results and visualizations

Bechdel Test Visualization

A quick look at the Bechdel test over time

Bechdel comparison of X-Men with other comics

X-axis can be viewed as the number of years into the Claremont Run



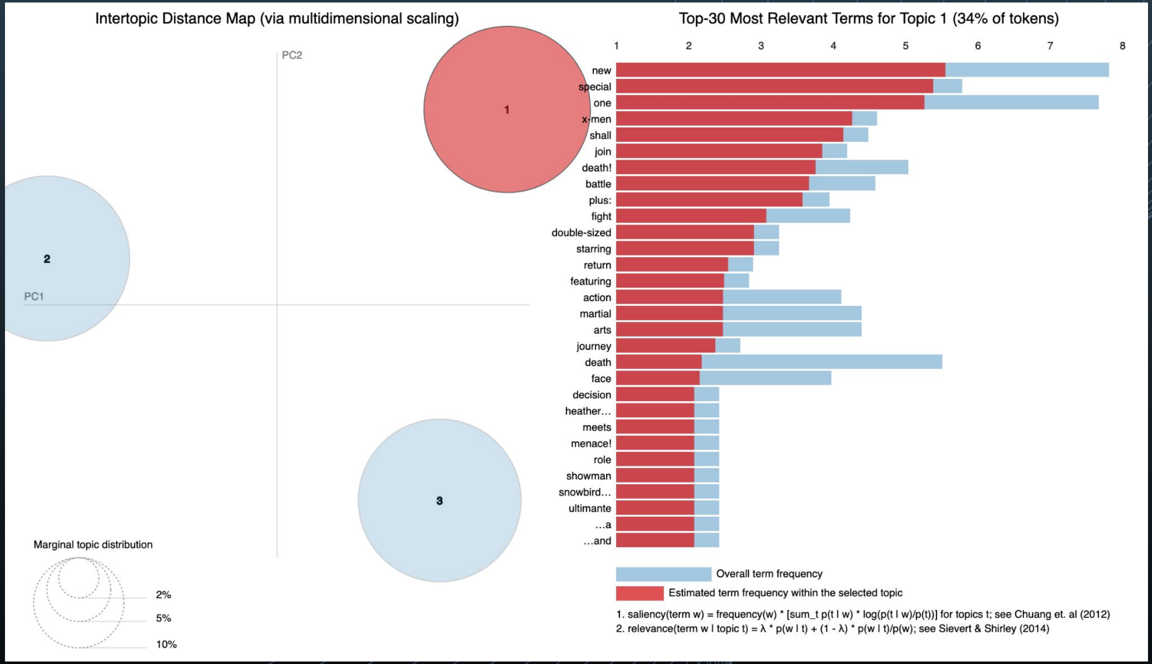
Topic Modeling to visualize main themes

Latent Dirichlet Allocation

- LDA finds hidden topics from within a set of documents
- A set of documents is a mixture of topics, and each topic is a mixture of words. Both of these mixtures follow the dirichlet distribution
- Thus, a topic is a probability distribution of a set of words

Main topics

There were 3 topics extracted from the comic titles. The primary topic contained words such as *death*, *fight*, *martial arts*, and *battle*

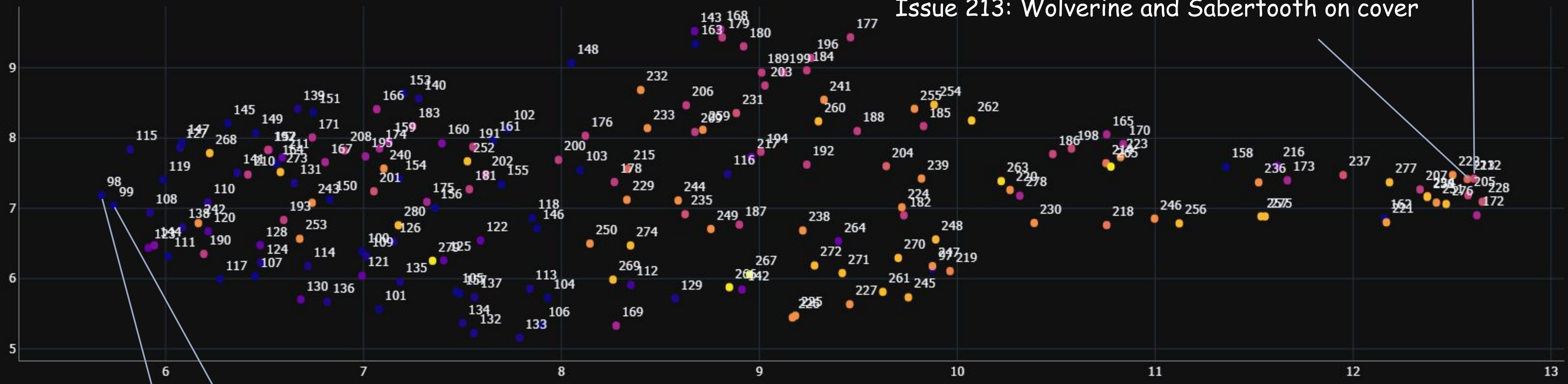


Alternative: Cover Sentence Embedding

Using the `all-mpnet-base-v2` sentence embedding model, we embed the covers as 768-D vectors. Similar issues end up close to each other.

Issue 212: “Wolverine Vs Sabertooth”

Issue 213: Wolverine and Sabertooth on cover



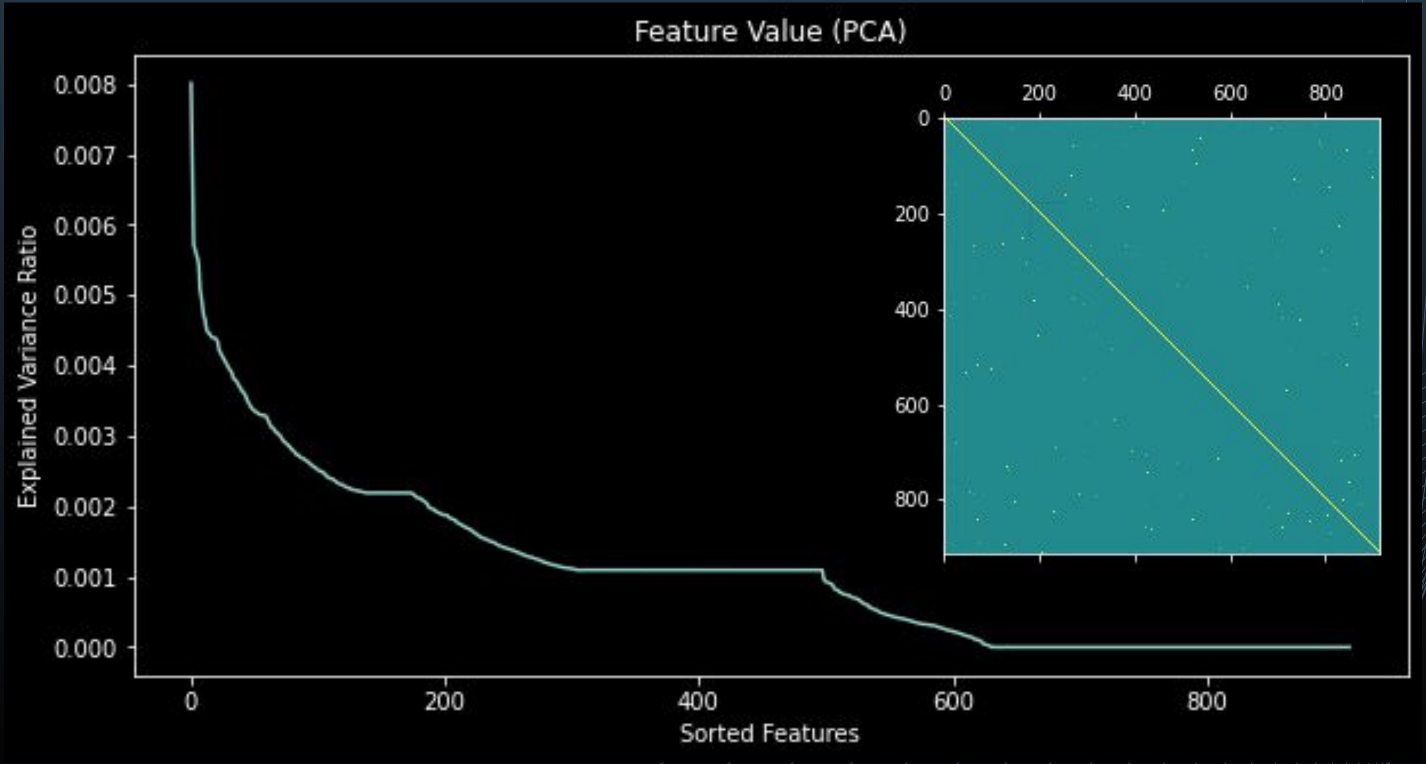
Issue 99: “Even if we escape the energy-spheres of the Sentinels -- how can we survive in the vacuum of space?”

Issue 98: “The Sentinels are Back! 'Nuff Said!”

Character Dimensionality reduction

Here we form character vectors by summing all event vectors for the same character, and project down to two dimensions using UMAP. Note that villains and a guest Avenger end up close to each other.

Character feature selection using correlation and PCA



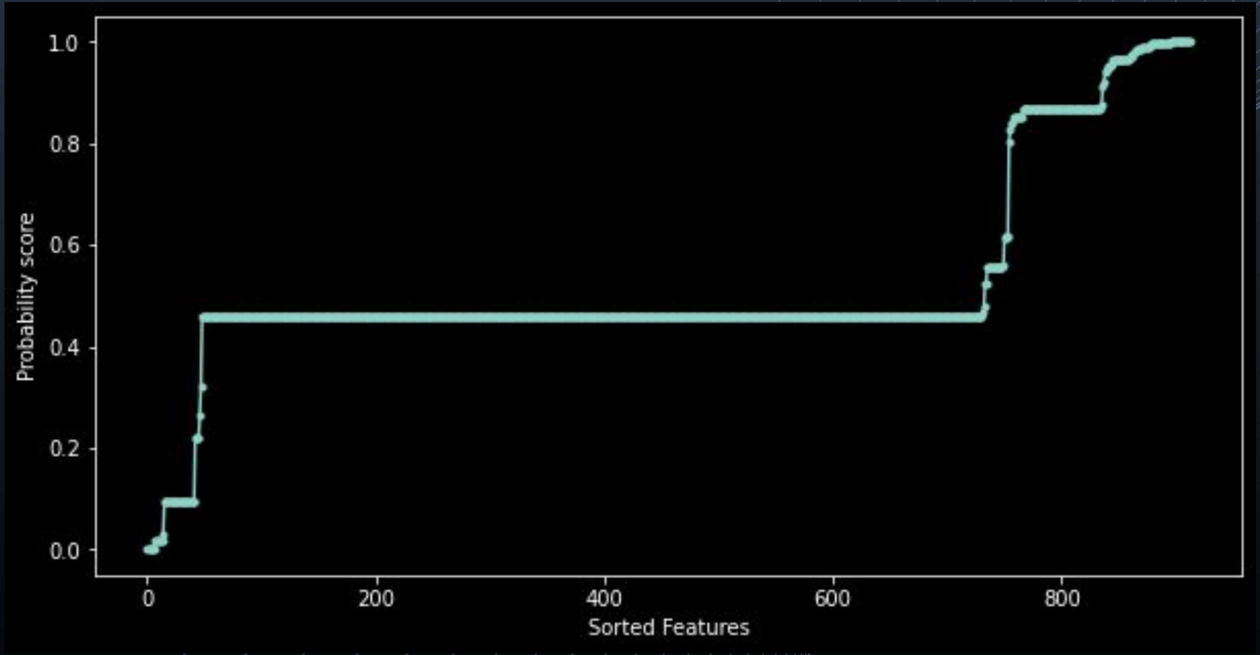
The standardized data shows that variance is spread widely over the features, and feature correlations are very limited. Many features will be needed to train a model to predict characters from event features.

Character feature selection K-best features (Chi-square scoring)

Lower values are better. This actually suggests that things plateau after the first ~50 features.

Best features include:

- Did the character kiss Lilandra
- Did they kiss Cyclops
- Did they walk arm in arm with Jean Grey
- Did they hug Madelyn Pryor



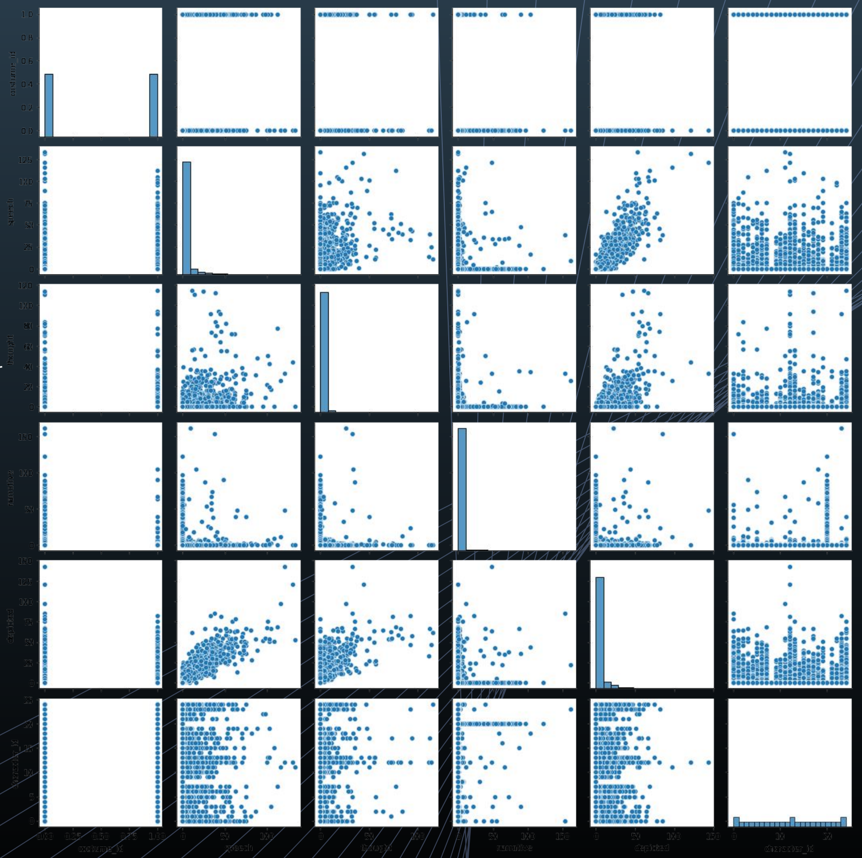
Classification of characters

Problem Statement

Can we classify/predict the character from their actions?

Data

- There are 5 predictors (number of speech bubbles, number of thought bubbles, narrative statements, number of depictions, and whether or not they appeared in costume)
- The target variable is the X-man character (there are 25 characters)



Variable correlation

Using pairwise plots, we inspected for any variables that should be dropped. All of the variables demonstrated some correlation with the target variable, so none of the variables were dropped.

Classification of characters

Logistic Regression

Using multinomial logistic regression, we predicted the X-man character from the 5 variables. The model was tested using 10 fold cross validation.

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Average
0.1061	0.1122	0.1193	0.1051	0.1010	0.0989	0.0928	0.0918	0.0836	0.0510	**0.0962**

Linear Discriminant Analysis

We also used LDA to predict the X-man character. The model also was tested using 10 fold cross validation

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Average
0.1102	0.1173	0.1163	0.1061	0.0969	0.0959	0.0867	0.0765	0.0775	0.0489	**0.0932**

Results

Both models demonstrated very poor prediction power. More analysis is needed to understand how to improve the model performance.

Upcoming work and conclusion

What comes next

Predictive Power:

Previews of coming Attractions

Hero/Villain Classifier

Based on certain representative features, can we identify the heroes and villains?

Dimensionality Reduction

Any predictor variables that are highly correlated, thus creating redundancy?

Other Models for Prediction

Predicting character actions/depictions from other actions/depictions. Predicting which comic issues got the best reception or were most popular.

Feature and graph engineering

- To classify the characters better, maybe we should translate relationship connections to counts (e.g. instead of "kissed character A and kissed character B" have a column for kisses with a value of 2.

- Much of the character feature data shows relationships to other characters (usually outside the standard list). Maybe we could make a social network graph to aid in predictions?