Machine Learning in the Milieu

Predicting Genre from Text

Alexander White

Deloitte AI Academy Capstone

January, 2023

Using NLP to aid in Publishing

Natural Language Processing (NLP) is a branch of Machine Learning focused on deriving intelligent insights from computer reading of texts.

This study was an effort to develop a model which would correctly predict the most appropriate genre for a book from the contents of its summary.

Why use NLP to predict genre?

- Genre is one of the main ways people find new books.
- Books that are labeled with appropriate genres are likely to find more success with their target audience.
- While everyone "knows" what genre is, it can be difficult to define the genre of a text.

UNIT SALES OF PRINT BOOKS, 2020–2021
(in thousands)

	2020	2021	CHANGE
Total	757,939	825,745	8.9%
Category			
Adult Nonfiction	308,823	322,564	4.4%
Adult Fiction	138,840	174,190	25.5%
Juvenile Nonfiction	77,865	75,059	-6.2%
Juvenile Fiction	184,178	201,868	9.6%
Young Adult Fiction	23,691	30,974	30.7%
Young Adult Nonfiction	3985	4,316	8.3%
Format			
Hardcover	226,369	249,788	10.3%
Trade Paperback	418,260	457,218	9.3%
Mass Market Paperback	39,420	38,215	-3.1%
Board Books	43,996	49,820	13.2%

SOURCE: NPD BOOKSCAN

Table from Publisher's Weekly, "Print Books Had a Huge Sales Year in 2021" by Jim Milliot, 2022

Our goal is to help publishers and editors like **you** to quickly and efficiently process the many books you receive.

Developing the model

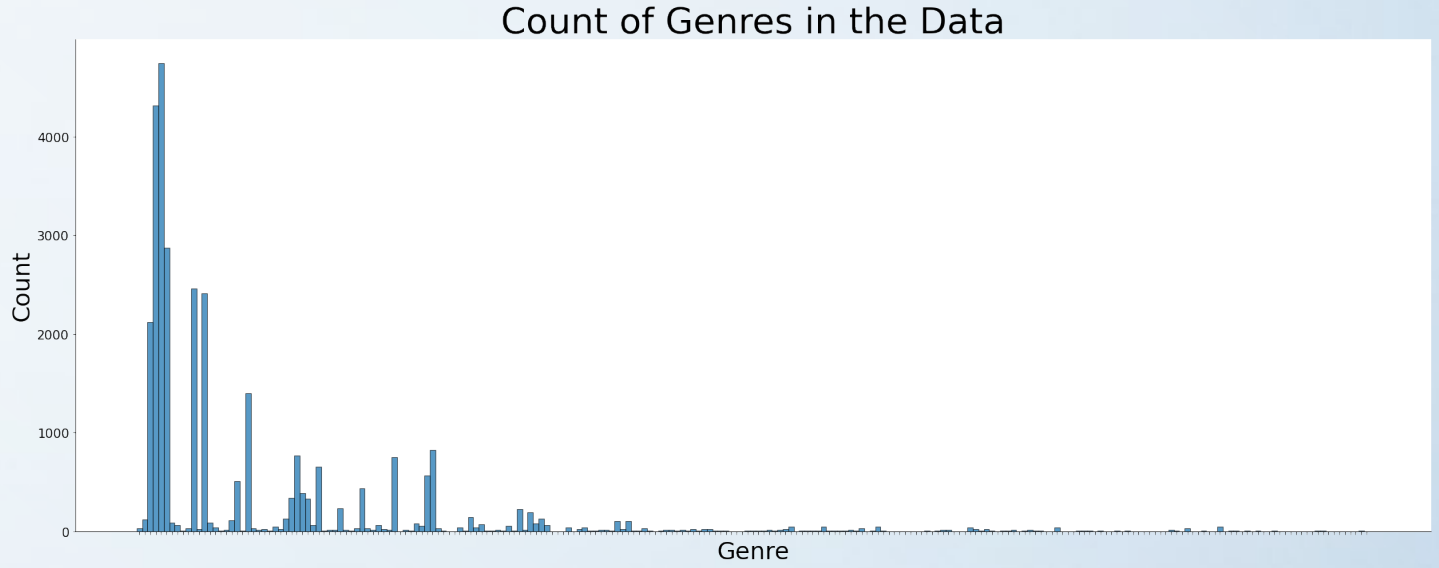
- Our initial hypothesis was that each genre would have its own subset of commonly used words.
- We ingested 12,000+ book summaries from a Kaggle dataset, originally sourced from Wikipedia and Freebase.
- The summaries were broken into standardized words (for example: am, are, were → is), which were counted and analyzed for each book.
- From these counts, we trained our model to predict the associated genre of a book.

Defining the Data Setting

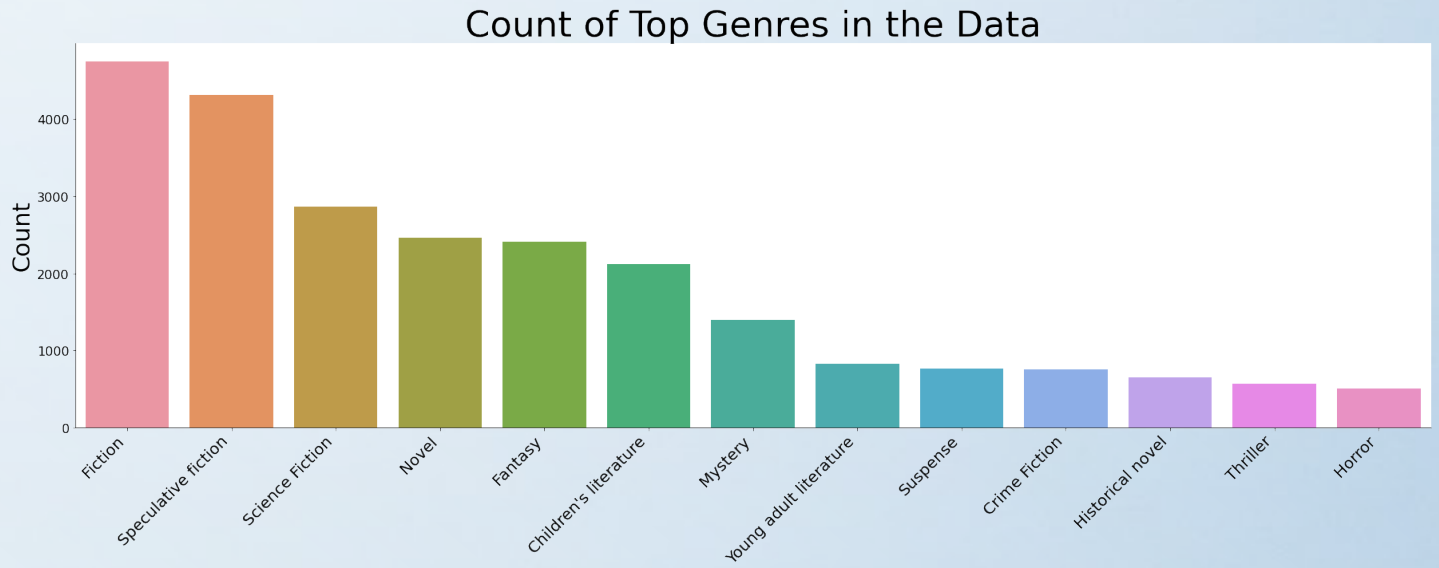
The initial data was severely unbalanced, so we focused the modeling on only those genres with more than 500 samples in our set.

Genre	Count
Fiction	4747
Speculative fiction	4314
Science Fiction	2870
Novel	2463
Fantasy	2413
…	
Pastiche	1
Marketing	1
Anti-nuclear	1
Fictional crossover	1
Collage	1

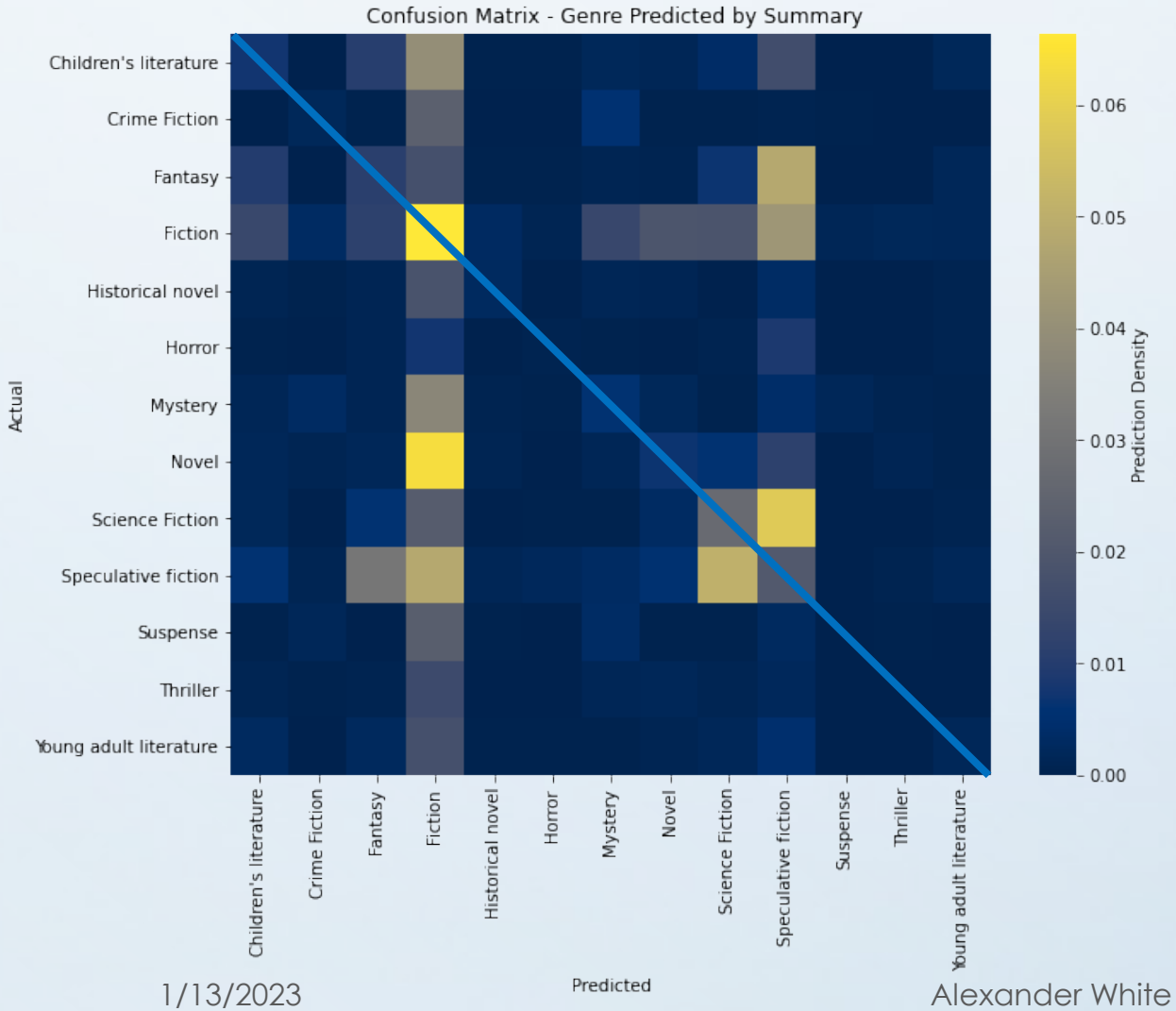
Initial



Selected

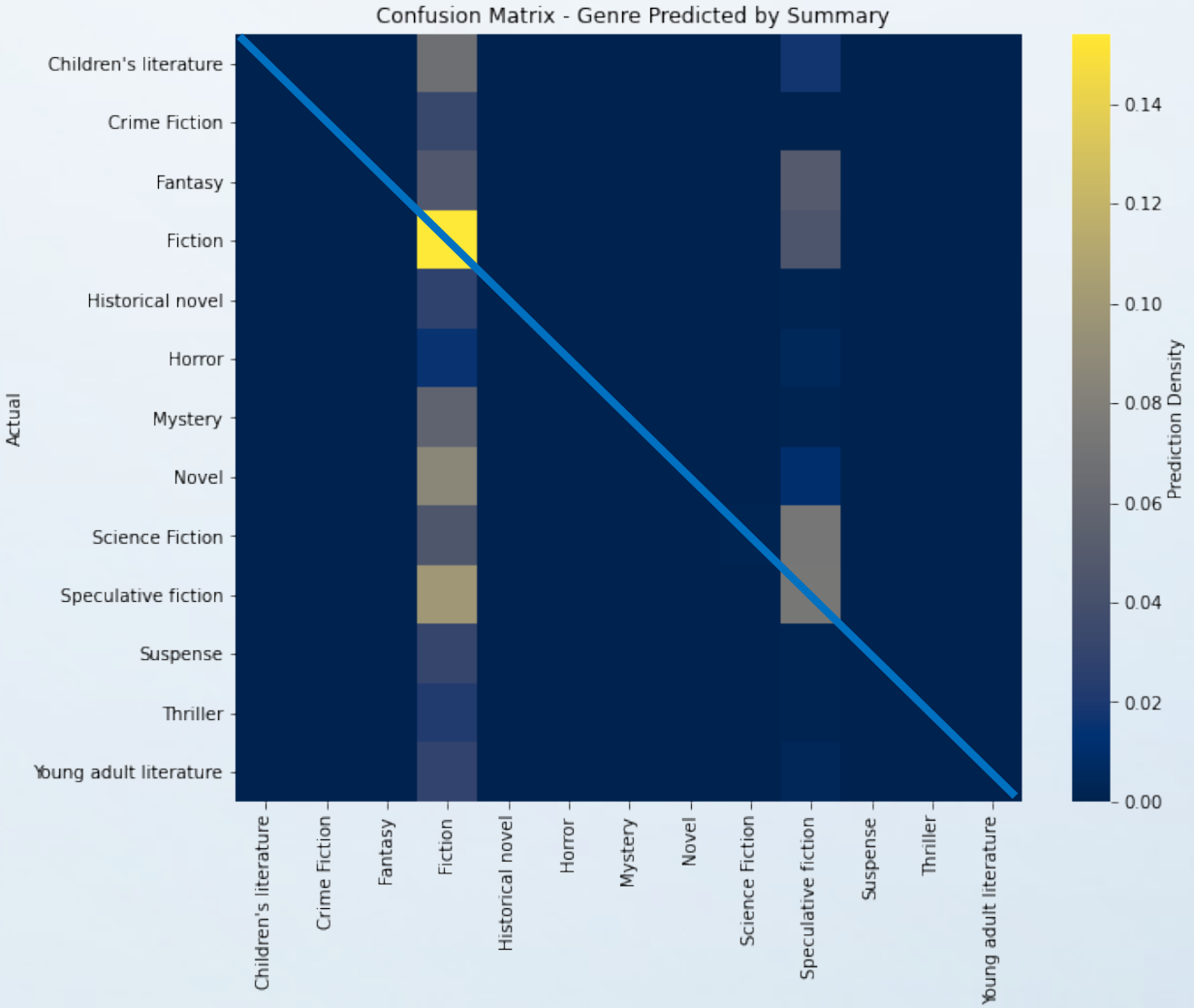


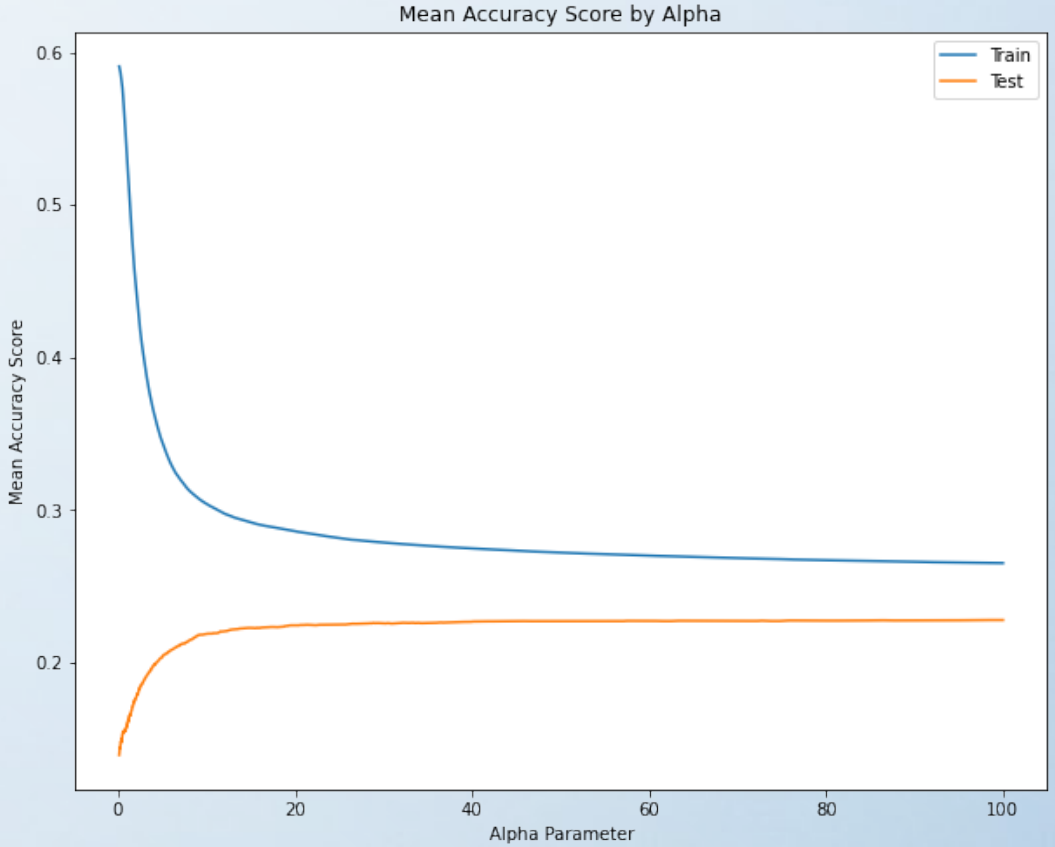
Initial modeling and predictions



genre	precision	recall	f1-score	support
Children's literature	0.15	0.09	0.11	619
Crime Fiction	0.15	0.06	0.08	241
Fantasy	0.14	0.11	0.13	713
Fiction	0.17	0.33	0.22	1454
Historical novel	0.26	0.09	0.13	219
Horror	0.1	0.03	0.04	147
Mystery	0.15	0.1	0.12	431
Novel	0.16	0.07	0.1	705
Science Fiction	0.23	0.23	0.23	881
Speculative fiction	0.09	0.12	0.1	1262
Suspense	0	0	0	235
Thriller	0.03	0.01	0.01	173
Young adult literature	0.16	0.05	0.08	241
accuracy			0.15	7321
macro avg	0.14	0.1	0.1	7321
weighted avg	0.15	0.15	0.14	7321

Further training caused loss of precision





Final Model Selection

- The initial model performed under our expectations, but did work.
- Further training the model increased the accuracy but reduced the utility.
- Using our complement naïve Bayes model with Term Frequency – Inverse Document Frequency data, we can predict a book's genre from its summary about 1 in 5 times.

The Significance: Time is Money

What can this do for you?

- Reduce the workload on editors
- Provide publishers with rapid suggestions for genre tagging of texts
- Help identify “tricky” books
- Grow with use

What limitations still exist?

- Only considers word use, not order or style
- Suggestions are not flawless and accuracy is currently very low
- Only works for English language texts
- Only suggests a single genre

Next Steps

Given more time, money, and data, we can significantly improve the accuracy and utility of our service.

- Examine alternative modeling techniques (such as categorial)
- Train with larger datasets
- Examine word order and context
- Include more genres
- Add other features such as sentiment analysis, plot summarization, automatic editorial suggestions, and many more

Conclusions

- Publishers and editors can make use of Natural Language Processing to analyze books, even for less concrete ideas such as genre.
- This can save man-hours and effort as well as assisting in classifying unusual texts.
- The field still has plenty of room for growth and could be extended to many other aspects of the editorial and publishing business.

"[Genre] is what tells you where to look in a bookstore…[it's] telling you which aisles to not bother going down. That's the simplicity of book shelving in bookstores. It tells you what not to read."

-Neil Gaiman, 2013

Thank you for your time.
Are there any questions?