

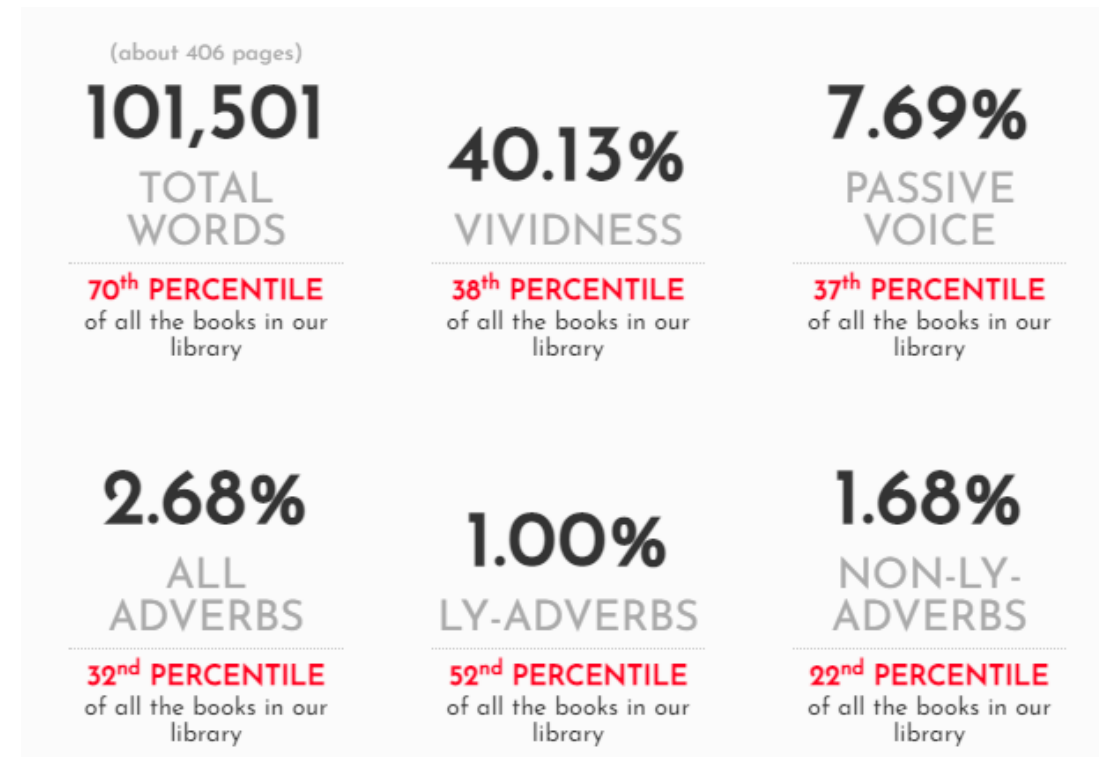


Genre prediction

A multilabel
classification model

Problem statement

- A now-offline website called Prosecraft.io contained analysis of tens of thousands of books.
- Is there a relationship between these analyses and the genres of the books?
- Could it be made into a predictive model?



Data Wrangling

A vertical line is positioned to the right of the text. In the bottom right corner, there is a yellow triangle pointing towards the center, partially overlapping a light gray border.

The list of books

- As of March 3, when the data was collected, Prosecraft contained 25,090 books
- A list of them could be found on the page for word-count percentiles

```
URL = "http://prosecraft.io/analysis/word-count/percentile/"
page = requests.get(URL)
```

```
page.text[:1000]
```

```
'<html>\n<head>\n  <meta charset="UTF-8">\n  <title>Prosecraft: Percentiles of Word Count</title>\n  <script>\n    var directory = [{"a": "Simon Jimenez", "t": "The Vanished Birds", "w": 197, "e": "jpg", "v": 124205.0}, {"a": "Jonathan P. Brazee", "t": "The Price of Honor", "w": 200, "e": "jpg", "v": 77253.0}, {"a": "Michael Carter", "t": "The Mathematical Murder of Innocence", "w": 190, "e": "jpg", "v": 37688.0}, {"a": "Anthony Boucher", "t": "The Case of the Baker Street Irregulars", "w": 197, "e": "jpg", "v": 80557.0}, {"a": "Charlie Dalton", "t": "Zombie Nation", "w": 188, "e": "jpg", "v": 64396.0}, {"a": "C. C. Harrington", "t": "Wildoak", "w": 197, "e": "jpg", "v": 55602.0}, {"a": "T. M. Logan", "t": "The Holiday", "w": 195, "e": "jpg", "v": 101767.0}, {"a": "Howard Sounes", "t": "Fab: An Intimate Life of Paul McCartney", "w": 225, "e": "jpg", "v": 225785.0}, {"a": "Nora Roberts", "t": "Honest illusions", "w": 197, "e": "jpg", "v": 163279.0}, {"a": "P. D. Cacek", "t": "Second Chances", "w": 194, "e": "jpg", "v": 91571.0}, {"a": "Robertson Davies", "t": "Fifth Business", "w": 196, "e": "jpg"}]
```

```
[{'a': 'Simon Jimenez',  
  't': 'The Vanished Birds',  
  'w': 197,  
  'e': 'jpg',  
  'v': 124205.0},
```

- After cleaning up the HTML with BeautifulSoup, the list was combined into a JSON.

The JSON was cleaned and turned into URLs, and the website was scraped for linguistic data.

Getting the Prosecraft Data

```
def get_info(title, author):  
    '''Given a title and author of a book in the list,  
    returns a dictionary of prosecraft's analysis about the book.'''  
  
    #Get rid of special characters in URL  
    chars_to_remove = [':', "'", '.', ',', '"', ' ', '"]  
    info = {'title': title, 'author': author}  
    URL = f"{author}/{title}/"  
    for char in chars_to_remove:  
        URL = URL.replace(char, '')  
    URL = URL.replace('&', 'and').replace(' ', '-').lower()  
    URL = "http://prosecraft.io/library/" + URL  
  
    #Get data from Prosecraft and turn it into a dict  
    page = requests.get(URL)  
    soup = BeautifulSoup(page.content, "html.parser")  
    headings = soup.find_all("div", {"class": "book-info-metric-heading"})  
    values = soup.find_all("div", {"class": "book-info-metric-value"})  
    for heading, value in zip(headings, values):  
        info[heading.text] = float(value.text.strip('%').replace(',',''))  
    return info
```

Getting the genres

Next, it was time to get the genre lists from Goodreads. This was more of a challenge .

```
def get_genres(title, author):
    browser.get('http://www.goodreads.com/search?q=&qid=')
    search_book = browser.find_element(By.ID, value='search_query_main')
    search_book.send_keys(title)
    search_book.submit()
    sleep(5)
    itemqueue = browser.find_elements(By.XPATH, value="//table/tbody/tr[contains(@itemtype, 'http://schema.org/Book')]")
    img = browser.find_elements(By.CLASS_NAME, value="bookCover")
    book_list = list()
    for i in range(len(itemqueue)):
        book_list.append(itemqueue[i].text.split('\n'))
        book_list_ap = list()
    for i in range(0, len(book_list)):
        book_list_ap.append((book_list[i][0],book_list[i][1],img[i].get_property("src")))
    for book in book_list_ap:
        if f'by {author}' in book[1]:
            book_id = book[2].split('/')[0].split('.')[0]
            break
    book_url = f'https://www.goodreads.com/book/show/{book_id}'
    browser.get(book_url)
    genres = browser.find_elements(By.XPATH, value="//span[contains(@class, 'BookPageMetadataSection__genreButton')]")
    return [genre.text for genre in genres]
```

	author	author_clean
176	Alva No'	Alva No'
233	César Aira	Cesar Aira
329	Maj Sjöwall & Per Wahlöö	Maj Sjowall & Per Wahloo
388	María José Ferrada	Maria Jose Ferrada
588	Eva García Sáenz	Eva Garcia Saenz

A few missing titles

It turns out that 410 books (1.6%) were missing data from Prosecraft due to special characters. A new function fixed that.

Missing genres

- Books not found would have to be searched for
- Books with no genres would have to be dropped

Books found from Prosecraft	24,857 (99%)
Prosecraft books not found (yet) on Goodreads	2,362 (9.5%)
Books found on Goodreads with no genres listed	2,157 (8.7%)



Goodreads second pass

- A longer, more involved function was created (searched for title and author, excluded middle initials, etc)
- With this, 40% of the missing books were able to be fixed!
- Of the books found on the second pass 17.0% had no genre data on Goodreads

A breakthrough!

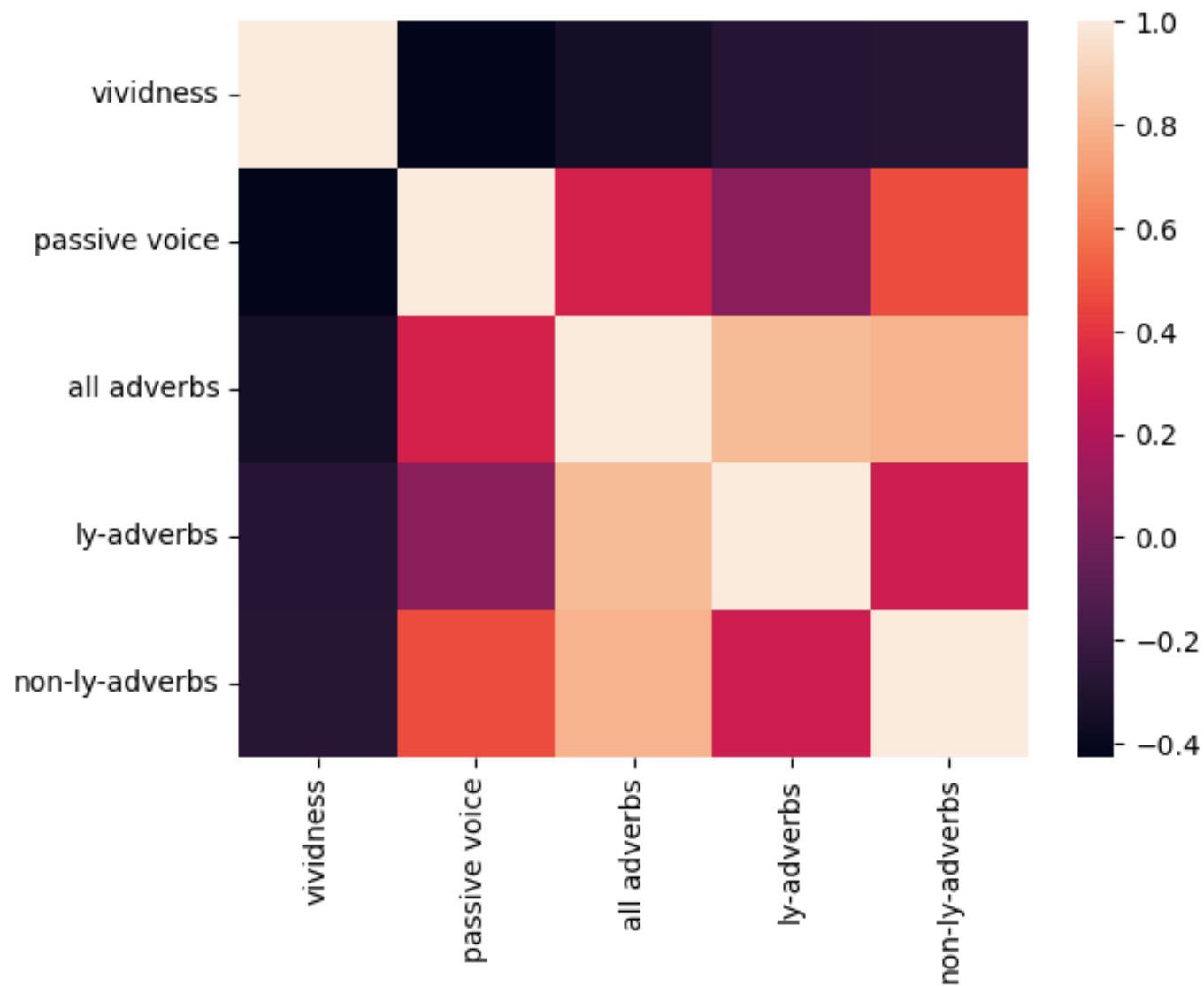
- From watching the program run, it became apparent that some of the missing values were simple typos
- Using the Levenshtein library, I was able to find small errors, like Michael Ben-Naftali to Michal Ben-Naftali.
- This enabled me to find another 40% of the missing data points.

An offer to send a list of typos to Benji Smith, the Prosecraft creator, changed the course of the project.

He was so enthusiastic about the project that he sent some extra data: Wordcounts for each unique word in each book.



Exploratory
Data
Analysis

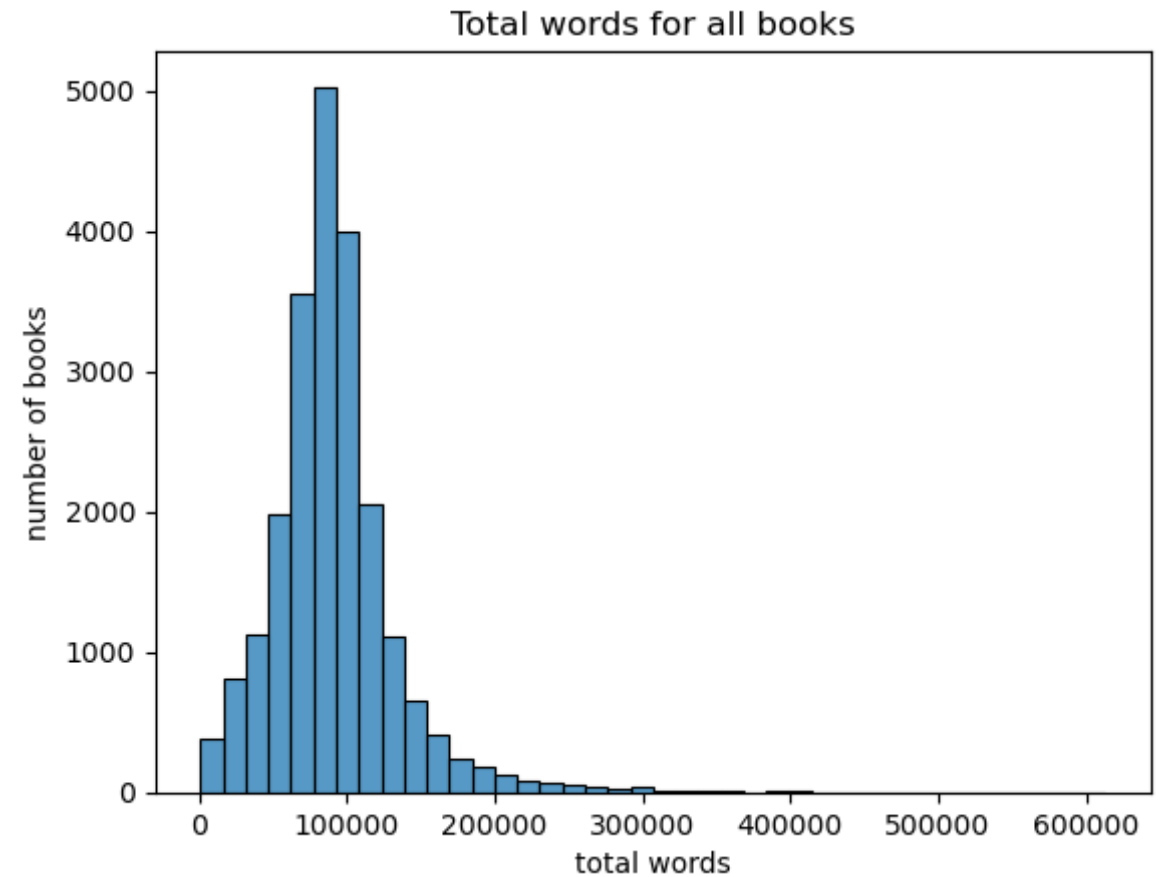


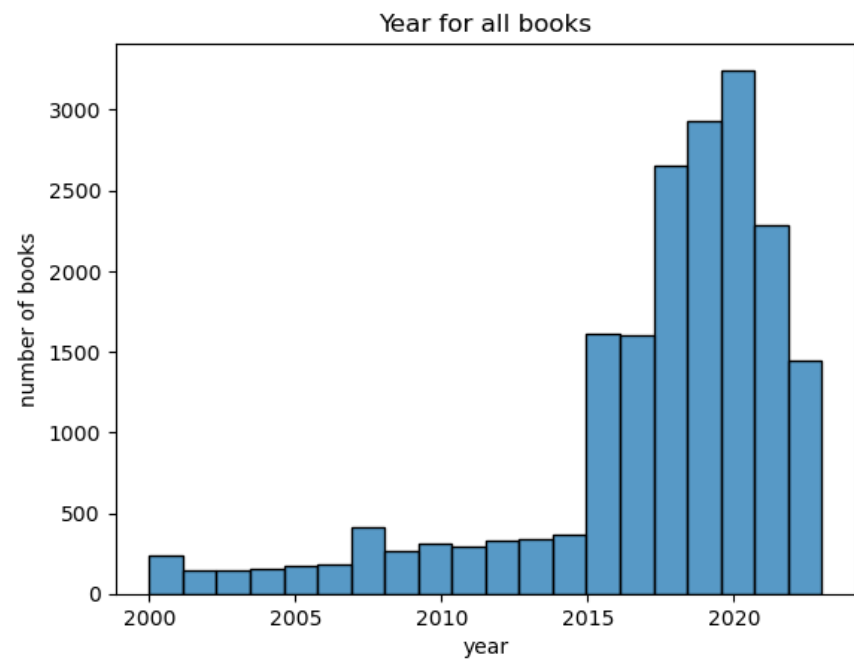
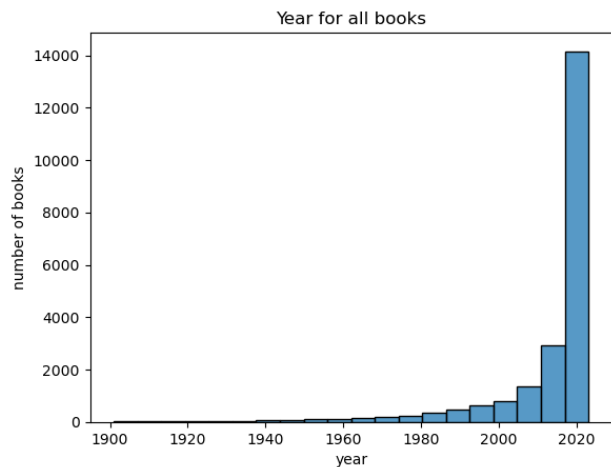
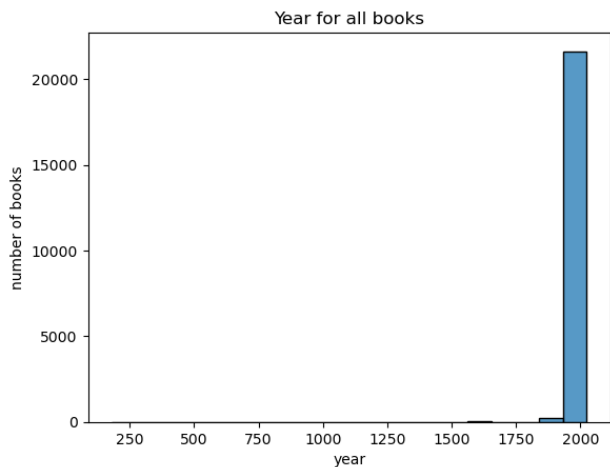
Correlations
in linguistic
data

Correlations in all quantitative data


	total words	vividness	passive voice	all adverbs	ly-adverbs	non-ly-adverbs	year
total words	1.000000	-0.001394	-0.018643	0.004145	0.023499	-0.017973	0.011426
vividness	-0.001394	1.000000	-0.425165	-0.348603	-0.282485	-0.279284	0.058717
passive voice	-0.018643	-0.425165	1.000000	0.323722	0.064872	0.468020	0.056973
all adverbs	0.004145	-0.348603	0.323722	1.000000	0.817468	0.793055	-0.099414
ly-adverbs	0.023499	-0.282485	0.064872	0.817468	1.000000	0.297553	-0.036549
non-ly-adverbs	-0.017973	-0.279284	0.468020	0.793055	0.297553	1.000000	-0.126100
year	0.011426	0.058717	0.056973	-0.099414	-0.036549	-0.126100	1.000000

Histogram of Wordcount

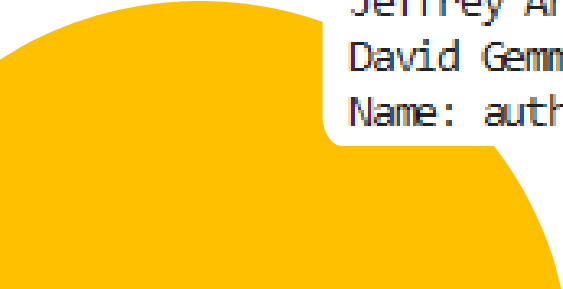




Publication year



Franklin W. Dixon	146
Carolyn Keene	83
Debbie Macomber	80
Stephen King	71
Danielle Steel	68
Dean Koontz	63
Nora Roberts	58
Piers Anthony	46
John Sandford	39
John Grisham	39
Lorelei James	38
R. L. Stine	38
Sandra Brown	38
Sherrilyn Kenyon	37
William Shakespeare	36
Tom Clancy	34
Ruth Rendell	34
Agatha Christie	33
Jeffrey Archer	32
David Gemmell	32
Name: author, dtype: int64	



Author counts

- The average author in the dataset had 1.79 books
- 20 authors had more than 30 books
- The top two authors, Dixon (pseudonym used for The Hardy Boys) and Keen (pseudonym for Nancy Drew), made up 1% of the data with their children's mystery fiction series.

Analyzing the target feature



Total books with genre data	22022
Average genres per book	6.0
Total number of distinct genres	688
Number of genres with only one associated book	117
Number of genres with at most ten associated books	333
Books lost after all but top genres are eliminated	211

Sample of
“genres”
with just
one book

	count
M M Sports Romance	1
Mysticism	1
Green	1
M M Mystery	1
American Fiction	1
Banks	1
Cthulhu Mythos	1
Prayer	1
Bigfoot	1
Romania	1

	count
Mauritius	1
Paranormal Urban Fantasy	1
Victorian Romance	1
Transport	1
Wonder Woman	1
Romanian Literature	1
Golden Age Mystery	1
Romanticism	1
Spider Man	1
Georgian	1

Genres with more than 1000 books

Fiction	15167
Mystery	7840
Thriller	6067
Fantasy	5670
Audiobook	5477
Mystery Thriller	3934
Crime	3765
Science Fiction	3483
Contemporary	3328
Romance	3234
Nonfiction	3106
Suspense	2937
Historical Fiction	2917

Adult	2679
Young Adult	2543
Historical	1997
Horror	1595
Adventure	1343
Paranormal	1301
History	1212
Literary Fiction	1200
Science Fiction Fantasy	1188
Magic	1131
Novels	1104
Biography	1077
Classics	1058
LGBT	1009

Genre correlation

	index	Fiction	Mystery	Thriller	Fantasy	Audiobook	Mystery Thriller	Crime	Science Fiction	Contemporary	Romance	Nonfiction	Suspense	Historical Fiction	Adult	Young Adult	Historical	Horror	Adventure	Paranormal	History	Literary Fiction	Science Fiction Fantasy	Magic	Novels	Biography	Classics	LGBT
	index																											
	Fiction	1.000000	0.451045	0.358278	0.280345	0.286807	0.245797	0.217182	0.186919	0.197732	0.150722	0.000000	0.180919	0.174458	0.156458	0.137208	0.112547	0.089273	0.075427	0.042065	0.000000	0.078658	0.073976	0.049581	0.072130	0.000000	0.061779	0.048858
	Mystery	0.872577	1.000000	0.828316	0.097577	0.290689	0.481378	0.433036	0.055867	0.139796	0.082908	0.004082	0.344770	0.135969	0.102041	0.081122	0.091709	0.080990	0.046939	0.048214	0.002551	0.020536	0.010332	0.014413	0.023087	0.001148	0.036480	0.019005
	Thriller	0.895065	0.811933	1.000000	0.061645	0.320092	0.550849	0.453766	0.080930	0.128400	0.040547	0.002143	0.434811	0.064282	0.107796	0.047140	0.029998	0.116038	0.054557	0.027031	0.001648	0.013186	0.009890	0.001319	0.024065	0.000659	0.008241	0.017307
	Fantasy	0.749912	0.134921	0.065961	1.000000	0.171781	0.012169	0.011111	0.337586	0.047443	0.231217	0.000000	0.010229	0.117989	0.149030	0.289947	0.072840	0.133510	0.118871	0.189065	0.000000	0.018695	0.156614	0.195414	0.027337	0.000000	0.047090	0.085714
	Audiobook	0.794230	0.416104	0.354574	0.177835	1.000000	0.269673	0.219281	0.137667	0.204674	0.130729	0.171444	0.187877	0.114661	0.188424	0.067555	0.084170	0.051488	0.046376	0.024283	0.054044	0.072668	0.055505	0.035969	0.029578	0.080701	0.015337	0.025561
	Mystery Thriller	0.947636	0.959329	0.849517	0.017539	0.375445	1.000000	0.524911	0.027453	0.136756	0.038638	0.000254	0.510014	0.063803	0.121505	0.051601	0.033045	0.064311	0.019319	0.014489	0.000254	0.012201	0.001017	0.000508	0.014743	0.000000	0.023386	0.014743
	Crime	0.874900	0.901726	0.731208	0.016733	0.318991	0.548473	1.000000	0.016733	0.074635	0.028685	0.026029	0.367596	0.087649	0.052058	0.018592	0.060823	0.040372	0.020983	0.006906	0.015936	0.008234	0.003187	0.002125	0.020186	0.011155	0.030013	0.012483
	Science Fiction	0.813953	0.125754	0.140970	0.549526	0.216480	0.031008	0.018088	1.000000	0.036750	0.078668	0.000000	0.032156	0.048808	0.111111	0.167959	0.018375	0.130060	0.109101	0.031582	0.000000	0.016078	0.241459	0.034166	0.058857	0.000000	0.059432	0.068045
	Contemporary	0.901142	0.329327	0.234075	0.080829	0.336839	0.161659	0.084435	0.038462	1.000000	0.389724	0.009014	0.127704	0.093450	0.279748	0.134916	0.032452	0.029447	0.007512	0.029447	0.000601	0.245793	0.001803	0.005108	0.145433	0.003906	0.015024	0.075120
	Romance	0.706865	0.200989	0.076067	0.405380	0.221398	0.047001	0.033395	0.084725	0.401051	1.000000	0.007112	0.094820	0.176252	0.220470	0.270563	0.142239	0.010513	0.025356	0.180581	0.000309	0.027211	0.011132	0.128942	0.022573	0.004020	0.018553	0.080396
	Nonfiction	0.000000	0.010303	0.004185	0.000000	0.302318	0.000322	0.031552	0.000000	0.009659	0.007405	1.000000	0.000644	0.000000	0.027688	0.000000	0.036703	0.004829	0.024791	0.003220	0.383773	0.001932	0.000000	0.003542	0.000000	0.344173	0.025113	0.027688
	Suspense	0.934287	0.920327	0.898195	0.019748	0.350358	0.691181	0.471229	0.038134	0.144705	0.104188	0.000681	1.000000	0.032346	0.110317	0.015322	0.010895	0.073885	0.040858	0.021110	0.000681	0.005107	0.004086	0.001382	0.012598	0.000340	0.002724	0.004767
	Historical Fiction	0.907096	0.365444	0.133699	0.229345	0.215290	0.086047	0.113130	0.058279	0.106616	0.195406	0.000000	0.032566	1.000000	0.141584	0.104902	0.622215	0.046623	0.053137	0.030854	0.000000	0.137127	0.009256	0.019198	0.082962	0.000000	0.068564	0.047309
	Adult	0.885778	0.298619	0.244121	0.315416	0.385218	0.178425	0.073162	0.144457	0.347518	0.266144	0.032102	0.120941	0.154162	1.000000	0.036954	0.119448	0.085106	0.028369	0.046286	0.008959	0.113475	0.063457	0.059724	0.035834	0.013438	0.004106	0.092572
	Young Adult	0.818325	0.250098	0.112466	0.646481	0.145497	0.079827	0.027527	0.230043	0.176563	0.344082	0.000000	0.017696	0.120330	0.038930	1.000000	0.074322	0.084153	0.177350	0.104601	0.000000	0.009438	0.053480	0.169878	0.009044	0.000000	0.099882	0.116398
	Historical	0.854782	0.360040	0.091137	0.208810	0.230846	0.065098	0.114672	0.032048	0.054081	0.230346	0.057086	0.018024	0.908863	0.160240	0.094642	1.000000	0.032549	0.048072	0.029044	0.056585	0.105158	0.005508	0.018027	0.043565	0.027541	0.038057	0.046570
	Horror	0.848903	0.427586	0.441379	0.474608	0.176803	0.158621	0.095298	0.284013	0.061442	0.021317	0.009404	0.136050	0.085266	0.142947	0.134169	0.040752	1.000000	0.034483	0.203135	0.006270	0.027586	0.033229	0.015047	0.029467	0.003135	0.048903	0.048276
	Adventure	0.851824	0.274013	0.246463	0.501862	0.189129	0.056590	0.058824	0.282949	0.018615	0.061057	0.057334	0.089352	0.115413	0.056590	0.335815	0.071482	0.040953	1.000000	0.014147	0.034996	0.004468	0.111690	0.098287	0.030529	0.029784	0.162323	0.009680
	Paranormal	0.490392	0.280546	0.128057	0.823982	0.102229	0.043812	0.019965	0.084550	0.075327	0.448885	0.007686	0.047656	0.069178	0.095311	0.204458	0.044581	0.249039	0.014604	1.000000	0.003075	0.002306	0.014604	0.289792	0.003075	0.000769	0.004612	0.030746
	History	0.000000	0.016502	0.008251	0.000000	0.244224	0.000825	0.049505	0.000000	0.001650	0.000825	0.983498	0.001650	0.000000	0.019802	0.000000	0.093234	0.008251	0.038779	0.003300	1.000000	0.001650	0.000000	0.002475	0.000000	0.339934	0.039604	0.010726
	Literary Fiction	0.994167	0.134167	0.066667	0.088333	0.331667	0.040000	0.025833	0.046667	0.681667	0.073333	0.005000	0.012500	0.333333	0.253333	0.020000	0.175000	0.036667	0.005000	0.002500	0.001667	1.000000	0.003333	0.000000	0.349167	0.000833	0.072500	0.076667
	Science Fiction Fantasy	0.944444	0.068182	0.050505	0.747475	0.255892	0.003367	0.010101	0.707912	0.005051	0.030303	0.000000	0.010101	0.022727	0.143098	0.114478	0.009259	0.044613	0.126263	0.015993	0.000000	0.003367	1.000000	0.147306	0.067340	0.000000	0.079125	0.039562
	Magic	0.064898	0.099912	0.007073	0.979664	0.174182	0.001768	0.007073	0.105217	0.015031	0.368700	0.009726	0.003537	0.049514	0.141468	0.381963	0.031830	0.021220	0.116711	0.310345	0.002653	0.000000	0.154730	1.000000	0.003537	0.000000	0.003537	0.041556
	Novels	0.990942	0.163949	0.132246	0.140399	0.146739	0.052536	0.068841	0.185688	0.438408	0.066123	0.000000	0.033514	0.219203	0.086957	0.020833	0.078804	0.042572	0.037138	0.003623	0.000000	0.379529	0.072464	0.003623	1.000000	0.000000	0.216488	0.022645
	Biography	0.000000	0.008357	0.003714	0.000000	0.410399	0.000000	0.038997	0.000000	0.012071	0.012071	0.992572	0.000929	0.000000	0.033426	0.000000	0.051068	0.004643	0.037140	0.000929	0.382544	0.000929	0.000000	0.000000	0.000000	1.000000	0.028784	0.040854
	Classics	0.885633	0.270321	0.047259	0.252363	0.079395	0.086957	0.106805	0.195652	0.047259	0.056711	0.073724	0.007561	0.189036	0.010397	0.240076	0.071834	0.073724	0.206049	0.005671	0.045369	0.082231	0.088847	0.003781	0.225898	0.029301	1.000000	0.011342
	LGBT	0.734390	0.147671	0.104063	0.481065	0.138751	0.057483	0.046581	0.234886	0.247770	0.257661	0.085233	0.013875	0.136769	0.245788	0.293360	0.092170	0.076313	0.012884	0.039643	0.012884	0.091179	0.046581	0.046581	0.024777	0.043608	0.011893	1.000000

Cutting down genres

Formats (not genres)	Audiobook, Novels, Fiction, Nonfiction, Short stories
Topics (not genres)	Magic, LGBT
High correlation (split up among component parts)	Mystery Thriller, Science Fiction Fantasy
Books disqualified after these genres were removed	1,532
Books remaining in dataset	20,490

Breaking down books by genre

- After creating summary statistics for each genre, it could be seen that most of the statistics were separated from each other by well under one standard deviation.
- It was time to incorporate the wordcount data.



Preprocessing



(1000, 394352)
(1000, 389858)
(1000, 404815)
(1000, 363286)
(1000, 376105)
(1000, 372208)
(1000, 362448)
(1000, 365104)
(1000, 380801)
(1000, 377421)
(1000, 374421)
(1000, 379593)
(1000, 386220)
(1000, 361970)
(1000, 366176)
(1000, 380285)
(1000, 381545)
(1000, 383617)
(1000, 389048)
(1000, 322558)
(490, 214767)

Wordcount data by chunks

- The dataframe was broken into chunks of 1,000 books
- The wordcount data was added.
- The resulting dataframes were quite large
- Since most of the columns did not match up, combining them into one far larger dataframe crashed the program
- This was the time I switched to Cloud Computing

A sample chunk

- Using just one chunk of 1,000 books as an example, pruning words that appear only once in each book caused the number of columns to decrease from 394,352 to only 162,978
- This eliminates 60% of the columns

```
def get_word_counts_from_json(author, title):  
    file_path = f'../word-counts/{author}/{title}/word-counts.json'  
    word_counts = read_json(file_path)  
    # Remove words that only appear once  
    pruned_word_counts = {word: number for word, number in word_counts.items() if number > 1}  
    return pruned_word_counts
```

Eliminating words across books

What happens if words appearing in less than 5 books in a chunk of 1,000 are eliminated?

It turns out, the original features are reduced by a factor of 10.

A list of words that appear at least 1 times includes 162968 words.

The first 10 words are ['0', '0-10', '0-2-1-7', '0-3', '0-99', '00', '000', '000-Year', '000-a-year', '000-acre']

The last 10 words are ["ó'cuinn", 'óig', 'ôi', 'ösana', 'ù', 'ü', 'über', 'übersecret', 'Öfunar', 'Ötsuka']

A list of words that appear at least 2 times includes 76563 words.

The first 10 words are ['0', '00', '000', '000-acre', '000-foot', '000-mile', '000-pound', '000-ton', '002', '007']

The last 10 words are ['è', 'é', 'éclair', 'écus', 'émigré', 'émigrés', 'était', 'étoile', 'êtes', 'über']

A list of words that appear at least 3 times includes 59206 words.

The first 10 words are ['0', '00', '000', '01', '02', '03', '04', '0400', '045', '05']

The last 10 words are ['École', 'Émile', 'Île', 'à', 'ça', 'è', 'éclair', 'émigré', 'émigrés', 'êtes']

A list of words that appear at least 4 times includes 50531 words.

The first 10 words are ['0', '00', '000', '01', '02', '03', '04', '05', '06', '0600']

The last 10 words are ['zoos', 'zu', 'zucchini', 'zygote', 'École', 'Île', 'à', 'ça', 'éclair', 'émigré']

A list of words that appear at least 5 times includes 45013 words.

The first 10 words are ['0', '00', '000', '01', '02', '03', '04', '05', '06', '07']

The last 10 words are ['zooms', 'zoos', 'zucchini', 'zygote', 'École', 'Île', 'à', 'ça', 'éclair', 'émigré']

The target feature as a dataframe

	Mystery	Thriller	Fantasy	Science Fiction	Crime	Contemporary	Romance	Suspense	Young Adult	Historical	Horror	Adventure	Paranormal
0	0	0	1	1	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	0	0	0	0	0	0
2	1	0	0	0	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	1	1	0	0	0
4	1	1	0	0	1	0	0	1	0	0	0	0	0

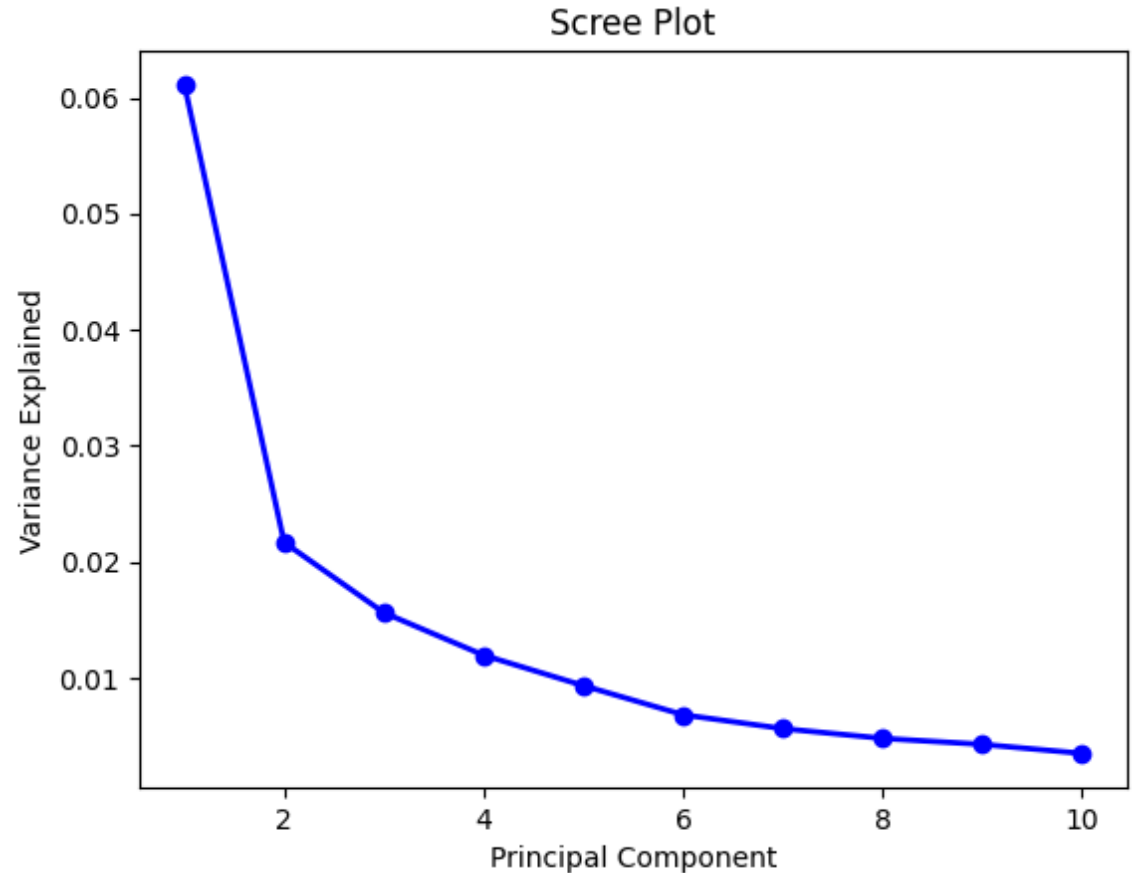


Scaling

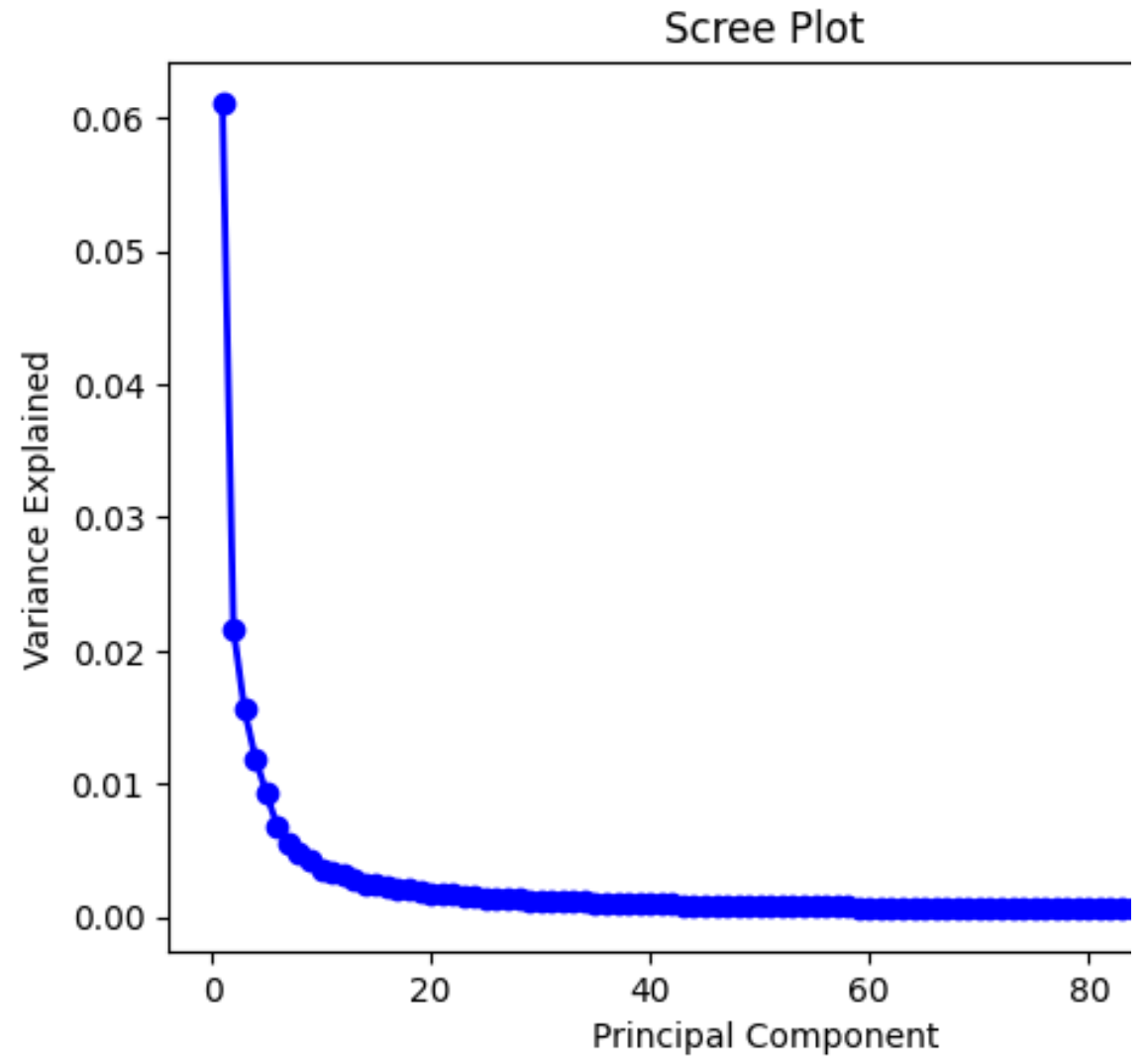
- The numerical data other than wordcounts were scaled with a StandardScaler
 - The wordcount data was scaled with a MinMax Scaler, as some words were far more common than others and did not necessarily merit an outsized effect on the model. This would also keep the 0's as 0.
-

Principal Component Analysis

- After Cross-validation was unable to fit a model, it was time for PCA
- An initial Scree plot (10 components) was less than encouraging
- No component captured more than 6% of the variation



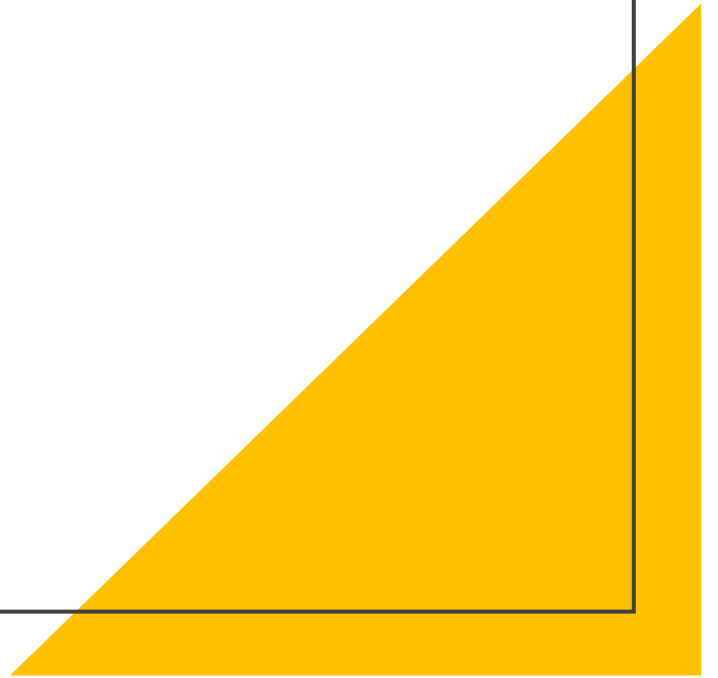
- Increasing the number of components to 100 did not help significantly, with a cumulative sum of explained variance of only 23%



Sparse PCA

After applying Sparse PCA, the model fit successfully.

Modeling



- As a basis for comparison, a Random Model was fit with default settings

Hamming Loss: 0.0953

F1 Score (Micro): 0.5814

F1 Score (Macro): 0.4745

Untuned Random Forest model

```
# Set up the sample space
```

```
param_dist = {  
    'n_estimators': [50, 100, 150, 200],  
    'max_depth': [None, 10, 20, 30, 40],  
    'min_samples_split': [2, 5, 10],  
    'min_samples_leaf': [1, 2, 4],  
    'bootstrap': [True, False]}
```

```
Hamming Loss: 0.0943
```

```
F1 Score (Micro): 0.5942
```

```
F1 Score (Macro): 0.4975
```

Tuned Random Forest model

A large yellow triangle is positioned in the bottom right corner of the slide, partially overlapping the white background and the black border of the main content area.

```
# Set up the sample space
xgb_param_dist = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 4, 5, 6],
    'min_child_weight': [1, 2, 3],
    'gamma': [0, 0.1, 0.2],
    'subsample': [0.6, 0.7, 0.8],
    'colsample_bytree': [0.6, 0.7, 0.8],
    'lambda': [0, 1, 2],
    'alpha': [0, 1, 2],
    'scale_pos_weight': [1, 2, 3],
    'eval_metric': ['logloss', 'auc'],
    'tree_method': ['auto', 'exact', 'approx', 'hist'],
    'grow_policy': ['depthwise', 'lossguide'],
}
```

Hamming Loss: 0.1064

F1 Score (Micro): 0.6523

F1 Score (Macro): 0.6047

XB Boost model

Full results of Random Forest

	precision	recall	f1-score	support
Mystery	0.77	0.74	0.76	1585
Thriller	0.74	0.70	0.72	1323
Fantasy	0.82	0.66	0.73	1181
Science Fiction	0.74	0.42	0.54	726
Crime	0.68	0.42	0.52	729
Contemporary	0.72	0.35	0.47	702
Romance	0.76	0.38	0.51	642
Suspense	0.77	0.23	0.35	610
Young Adult	0.80	0.30	0.43	502
Historical	0.79	0.38	0.51	599
Horror	0.84	0.07	0.13	298
Adventure	0.89	0.21	0.34	255
Paranormal	0.81	0.08	0.14	273
History	0.91	0.73	0.81	250
Literary Fiction	0.75	0.08	0.14	267
Biography	0.77	0.43	0.55	210
Classics	0.83	0.68	0.75	176
Memoir	0.76	0.43	0.55	192
micro avg	0.77	0.48	0.59	10520
macro avg	0.79	0.40	0.50	10520
weighted avg	0.77	0.48	0.56	10520
samples avg	0.65	0.50	0.54	10520

Full results of XGBoost

	precision	recall	f1-score	support
Mystery	0.70	0.88	0.78	1585
Thriller	0.66	0.85	0.74	1323
Fantasy	0.72	0.76	0.74	1181
Science Fiction	0.56	0.68	0.61	726
Crime	0.52	0.73	0.61	729
Contemporary	0.56	0.65	0.60	702
Romance	0.59	0.66	0.62	642
Suspense	0.48	0.66	0.56	610
Young Adult	0.61	0.55	0.58	502
Historical	0.56	0.61	0.58	599
Horror	0.36	0.32	0.34	298
Adventure	0.46	0.36	0.40	255
Paranormal	0.45	0.27	0.34	273
History	0.84	0.81	0.83	250
Literary Fiction	0.48	0.40	0.43	267
Biography	0.60	0.68	0.64	210
Classics	0.77	0.80	0.78	176
Memoir	0.71	0.68	0.70	192
micro avg	0.61	0.70	0.65	10520
macro avg	0.59	0.63	0.60	10520
weighted avg	0.61	0.70	0.65	10520
samples avg	0.62	0.71	0.63	10520

Significant features

Once the model had been created, it was possible to look at the features each model had deemed significant. A small sample of significant non-PCA features is shown below.

Mystery

Random forest significant features: ['passive voice', 'total words', 'vividness score']

XGB significant features: ['passive voice', 'total words', 'publication year']

Thriller

Random forest significant features: ['passive voice', 'total words', 'vividness score']

XGB significant features: ['total words', 'passive voice', 'publication year', 'vividness score']

Fantasy

Random forest significant features: ['vividness score', 'total words']

XGB significant features: ['total words', 'vividness score', 'publication year', 'passive voice']

Closer examination of significant features

- Next, a comparison was generated of the average by genre of each significant feature (as used in the XG Boost model) compared to the average from all books. Another sample is below.

Average passive voice for Mystery: 8.39 vs 8.03 for all books ($z = 0.29$)

Average total words for Mystery: 88237.07 vs 92713.29 for all books ($z = -0.1$)

Average publication year for Mystery: 2013.31 vs 2010.13 for all books ($z = 0.08$)

Average total words for Thriller: 91491.5 vs 92713.29 for all books ($z = -0.03$)

Average passive voice for Thriller: 8.37 vs 8.03 for all books ($z = 0.26$)

Average publication year for Thriller: 2014.46 vs 2010.13 for all books ($z = 0.11$)

Average vividness score for Thriller: 46.78 vs 47.7 for all books ($z = -0.08$)

Average total words for Fantasy: 101863.69 vs 92713.29 for all books ($z = 0.21$)

Average vividness score for Fantasy: 51.91 vs 47.7 for all books ($z = 0.35$)

Average publication year for Fantasy: 2010.94 vs 2010.13 for all books ($z = 0.02$)

Average passive voice for Fantasy: 7.89 vs 8.03 for all books ($z = -0.11$)



Observations

- Some features were found significant (feature importance > 0.04) despite having z-scores close to 0 ($|z| \leq 0.04$)
 - Total words for Thriller, Romance
 - Publication year for Fantasy, Science Fiction, Young Adult, Horror
 - -ly adverbs for Horror, non-ly-adverbs for Adventure, all adverbs for Paranormal
- Some features were found with $|z| \geq 0.5$
 - Vividness for Horror ($z=0.58$)
 - Passive voice for History ($z= -1.29$)
 - Publication year for classics ($z= -2.02$)


Opportunities for improvement

- Re-do beginning notebooks with current knowledge
- Drop “all adverbs” from predictive features, although multicollinearity should not affect predictions (Paul, 2006)
- Use a more balanced dataset (more older books, more genre variety, fewer Hardy Boys)
- Use publishers’ genre data rather than Goodreads shelves (increase validity of labels and decrease bias toward more popular books)





Note on Prosecraft.io

- Near the end of this project, the data source, Prosecraft.io, came under criticism for unethical data sourcing, and was shut down by the creator, Benji Smith
 - This project has used the data for educational purposes only, not to be published or profited from. I am committed to ethics in Data
- 

Note on Goodreads

- Goodreads genre data is based off of users' shelves, and the data has not been vetted by professionals
- As such, there are often errors. For instance, adult Fantasy by female authors tends to be misclassified as Young Adult (The Guardian, 2019)
- It is to be hoped that since all but the most frequent genres of relatively popular books were dropped, there will be enough data to average out most of these inaccuracies.

Citations

- Paul, R. K. (2006). Multicollinearity: Causes, Effects and Remedies *Indian Agricultural Statistics Research Institute*.
<http://apps.iasri.res.in/seminar/AS-299/ebooks/2005-2006/Msc/trim2/3.%20Multicollinearity-%20Causes,Effects%20and%20Remedies-Ranjit.pdf>
- *Women write fantasy for grown-ups, too*. (2019, January 30). The Guardian.
<https://www.theguardian.com/books/shortcuts/2019/jan/30/women-write-fantasy-for-grown-ups-too>

Acknowledgements

- I'd like to thank my mentor at Springboard, Chris Esposito, for his help with both the general ideas and specific implementation for this project.
- Thanks also to Benji Smith for the original dataset, his encouragement and enthusiasm, and the gift of the extra data which so improved my model.