ANALYSIS OF INFORMATION REVELATION IN BOOK POPULARITY

The analysis of Kullback-Liebler divergence (KLD) measures and their relationship with the popularity of English-language fiction books revealed several important insights. By examining a variety of KLD metrics—such as peak KLD, Author Frequency, Standard deviation, and range—we developed a detailed set of features that describe each book's narrative structure. Our initial analysis showed that specific KLD measures, like high KLD events and normalized KLD growth, though not overwhelmingly significant, still play a role in influencing book downloads. Other factors like author frequency, and several control features like the author's birth and death years, along with sentiment scores and genre categories, also significantly impact a book's popularity, indicating that both the story's structure and its metadata are crucial.

To refine our model and improve its accuracy, we tackled multicollinearity by using the variance inflation factor (VIF) method to identify and remove highly collinear features. Further refining through LASSO regression allowed us to clearly identify the most influential predictors for book downloads. This method highlighted that readers have a distinct preference for books with consistent narrative styles and positive emotional content. For example, features such as the range of KLD and the average sentiment showed that books with consistent and engaging narratives coupled with stable emotional undertones are more favored. Additionally, genres like action, science fiction, and horror emerged as particularly popular. Conversely, genres like adventure and biography are less influential. Additionally, the popularity of books from recently deceased authors (as indicated by author year of death variable) and those with frequent publications (as indicated by author frequency variable) highlight the importance of an author's legacy and activity in reader preferences.

Based on our analysis , in order to boost book downloads , readers can focus on enhancing narrative consistency and emotional engagement, and strategically market books in popular genres like action, fantasy, science fiction, and horror to capitalize on strong reader demand.

VARIABLE DESCRIPTION :

Variable	Description
peak_kld	The maximum Kullback-Liebler divergence (KLD) score within each book's narrative. This measure captures the highest point of information revelation in the narrative.
std_dev_kld	The standard deviation of KLD scores. This measures the variability /inconsistency in information revelation throughout the book.
range_kld	The range between the maximum and minimum KLD scores. It indicates the overall span of information variability within a book.
iqr_kld	The interquartile range of KLD scores, calculated as the difference between the 75th percentile and the 25th percentile. This metric helps understand the spread of the middle 50% of the data, reducing the impact of outliers.
high_kld_events	The number of KLD scores exceeding the 75th percentile threshold. It quantifies the frequency of high information revelation events throughout the narrative.
median_kld	The median of the KLD scores, tells the central tendency of narrative complexity or information revelation.
kld_peaks_count	The count of KLD scores that exceed a specific high threshold (0.3), indicating frequent peaks in narrative complexity.
kld_troughs_count	The count of KLD scores below a specific low threshold (0.1), indicating frequent troughs or lower levels of narrative complexity.
normalized_kld_growth	The total growth in KLD scores across the narrative, normalized by the length of the narrative. This measures the overall increase in information revelation from the beginning to the end of the book.

author_freq	Represents the frequency of books published by each author within the dataset. This variable reflects how actively an author produces content or how well-known they are.

REGRESSION TABLE

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.2129	0.019	169.735	0.000	3.176	3.250
peak_kld	0.2243	1.711	0.131	0.896	−3.129	3.578
std_dev_kld	1.1505	2.660	0.432	0.665	−4.064	6.365
range_kld	−1.0536	1.489	−0.707	0.479	−3.973	1.866
iqr_kld	2.0771	1.671	1.243	0.214	−1.197	5.352
high_kld_events	−0.0462	0.024	−1.965	0.049	−0.092	−0.000
median_kld	1.0770	1.610	0.669	0.504	−2.079	4.233
kld_peaks_count	−0.0303	0.025	−1.221	0.222	−0.079	0.018
kld_troughs_count	−0.3106	0.460	−0.675	0.500	−1.213	0.592
author_freq	0.1052	0.009	11.839	0.000	0.088	0.123
subj2_war	−0.0433	0.015	−2.886	0.004	−0.073	−0.014
authoryearofbirth	0.2641	0.114	2.313	0.021	0.040	0.488
authoryearofdeath	−0.4370	0.114	−3.828	0.000	−0.661	−0.213
subj2_adventure	−0.0692	0.015	−4.490	0.000	−0.099	−0.039
subj2_biography	−0.0117	0.013	−0.918	0.359	−0.037	0.013
subj2_romance	0.0321	0.016	2.065	0.039	0.002	0.063
subj2_drama	0.0031	0.012	0.245	0.806	−0.021	0.028
subj2_fantasy	0.1150	0.013	8.553	0.000	0.089	0.141
subj2_family	0.0053	0.014	0.382	0.702	−0.022	0.032
subj2_sciencefiction	0.1392	0.015	9.542	0.000	0.111	0.168
subj2_action	−0.0096	0.012	−0.775	0.439	−0.034	0.015
subj2_western	−0.0049	0.015	−0.328	0.743	−0.034	0.024
subj2_horror	0.1168	0.013	9.113	0.000	0.092	0.142
subj2_mystery	0.0355	0.016	2.198	0.028	0.004	0.067
subj2_crime	0.0087	0.013	0.683	0.495	−0.016	0.034
subj2_history	−0.0333	0.020	−1.660	0.097	−0.073	0.006
subj2_periodicals	0.0332	0.012	2.678	0.007	0.009	0.058
subj2_others	−0.1599	0.029	−5.454	0.000	−0.217	−0.102
speed	−0.0612	0.017	−3.607	0.000	−0.095	−0.028
sentiment_avg	−0.1638	0.014	−11.352	0.000	−0.192	−0.136
sentiment_vol	0.1602	0.020	7.822	0.000	0.120	0.200
wordcount	0.0472	0.019	2.495	0.013	0.010	0.084

IMPORTANT FEATURES SELECTED BY LASSO

```
Important features selected by LASSO:
 range_kld               -0.551514
authoryearofdeath        -0.393237
sentiment_avg            -0.160940
subj2_others             -0.139679
subj2_adventure          -0.076381
speed                    -0.057484
high_kld_events          -0.042202
subj2_war                -0.031254
subj2_biography          -0.026680
kld_peaks_count          -0.026448
subj2_history            -0.018765
subj2_western             0.005949
subj2_drama               0.008081
subj2_crime               0.011177
subj2_family              0.014193
subj2_periodicals         0.033127
subj2_mystery             0.035773
subj2_romance             0.041426
wordcount                 0.055602
subj2_fantasy             0.110040
subj2_horror              0.127143
subj2_sciencefiction      0.140087
author_freq               0.144049
sentiment_vol             0.152186
authoryearofbirth         0.234500
median_kld                1.321775
subj2_action             19.667207
dtype: float64
```