Books by the Numbers: Because Feelings Don't Fit in a Dataset

What Makes a Successful Book?

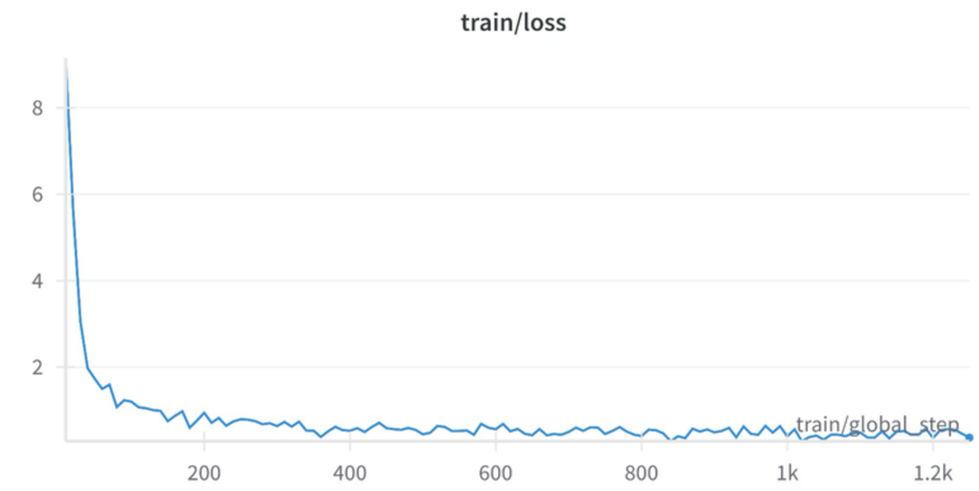
We wanted to use data science and machine learning tools to determine what goes into a Successful book. Using a number of datasets from Goodreads and other sources, we wanted to create a tool to determine how the choice of title, genre and description affects the rating of a book and thus how popular it is.
Our project does not evaluate what the optimal creative choices are but rather provides users the tools to measure their own choices against a model trained with the data collected from thousands of books.

Our Process

We started the project by considering what data science and machine learning methods would be able to reasonably assist a person with, and decided that with the multitude of books published throughout history, coupled with the online free to use Good Reads platform providing us with reliable information about thousands of different books, that it would be wise to take advantage of those resources and create a book based classification system.

We started by taking collected *goodreads* datasets and processed them by focusing on a few key fields - title, author, genre, description and rating - and discarded the extra information and stored the newly cleaned dataset for later use as training and testing resources

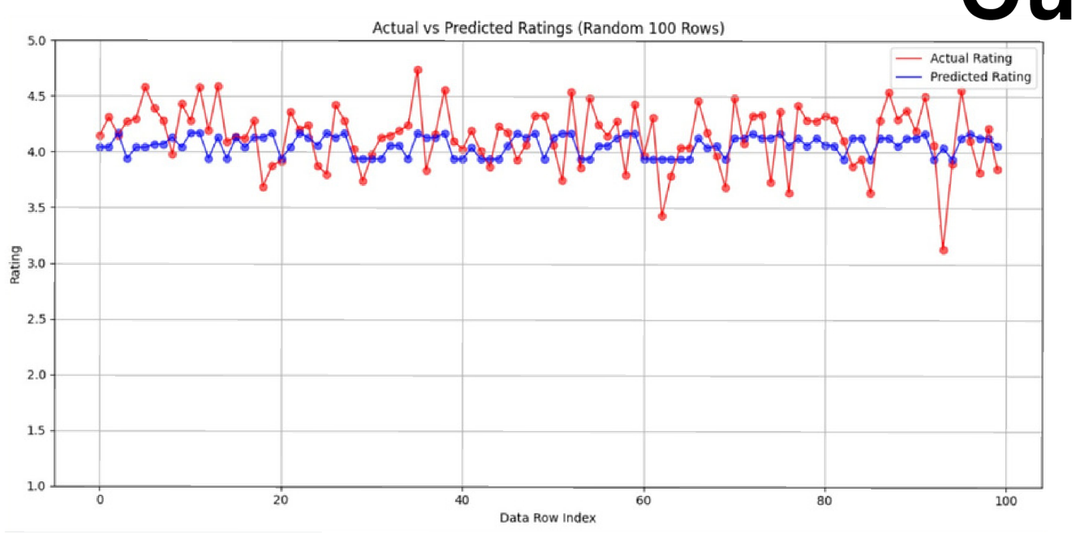
Using the cleaned data, we then train a number of models. Firstly a BERT model that predicts a books genre and rating from the inputted fields (title, description) , followed by a T5 model for the same purpose, then a T5 model to predict/generate a book title from inputted information (genre, description), and then finally a T5 model that combines both those capabilities into one singular model.



Train / Loss Graph showing the progress of learning for the model training

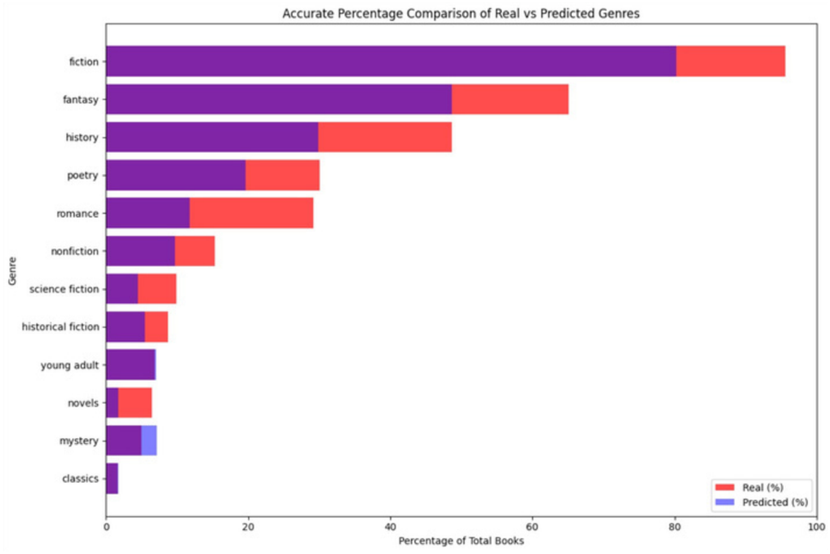
While creating the models, we investigated in parallel the use of Dummy Classifiers as a tool to measure the performance of our final trained model. Our idea was to use the Dummy classifier, which takes no input data and predicts based on 'brute-force' essentially, where it classifies based on hard coded directives, such as categorising all values based on the most common category in the dataset. Using this method, we are pitting our model against a basic categorization to measure whether the training has meaningfully improved the accuracy of data classification and prediction past the shadow of any doubt of coincidence or luck.

Our Results



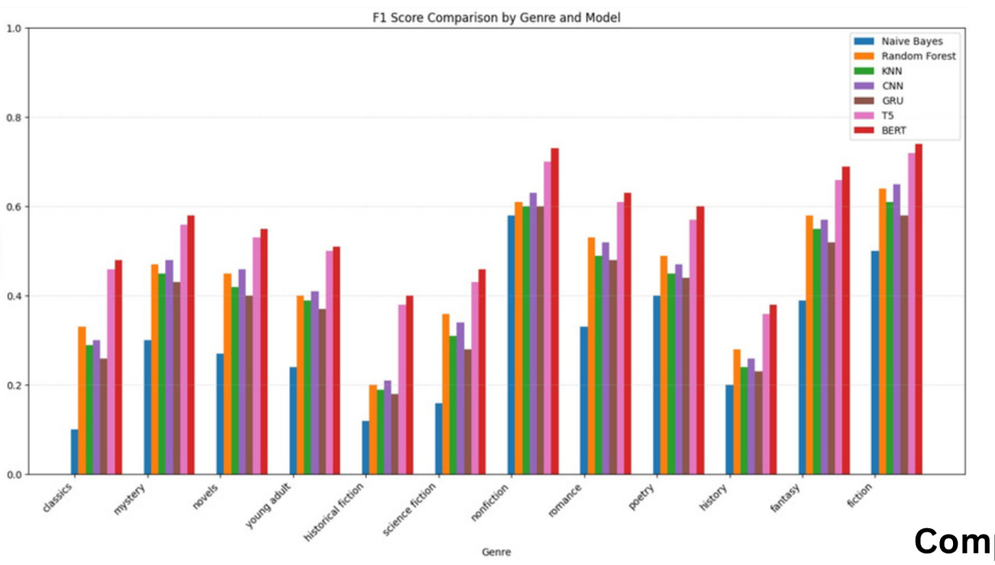
Rating prediction

This graph displays the actual versus predicted ratings for a random sample of 100 books, illustrating the performance of our final model. The predicted values mostly follow the trend of the actual ratings but have noticeably less variance, showing that the model tends to weight towards the mean. While it captures the general range of the ratings, it does not replicate the sharper differences observed in the reference data. This suggests that while the model is relatively stable, it is unable to capture the full nuance of how the features influence higher or lower ratings, and may benefit from further tuning or additional features to improve its responsiveness to outlying data points.



Genre prediction

This bar graph compares the real vs predicted genre distributions of books as identified by our model. From the graph, we can see that the model better identifies the dominant presence of genres like *fiction* and *fantasy*, it tends to lose accuracy in less frequent genres like *history*, *romance*, and *science fiction*, which are underrepresented . In contrast, some less common genres such as *young adult*, *historical fiction* and *classics* are more accurately predicted, potentially due to their more distinct nature. This behaviour can be described as a better alignment in higher-frequency genres and more discrepancies in lower-frequency ones. This suggests that the model is more skilled at predicting common genres and has difficulty classifying more specific categories, showcasing a need for better class balance in the training data or a more sophisticated classification strategy.



genre	Naive Bayes	Random Forest	KNN	CNN	GRU	T5	BERT
classics	0.12	0.34	0.30	0.32	0.28	0.48	0.50
mystery	0.31	0.49	0.47	0.50	0.45	0.58	0.60
novels	0.29	0.47	0.45	0.48	0.43	0.56	0.58
young adult	0.25	0.42	0.41	0.44	0.40	0.52	0.54
historical fiction	0.14	0.22	0.21	0.23	0.20	0.40	0.42
science fiction	0.18	0.38	0.33	0.35	0.30	0.46	0.48
nonfiction	0.61	0.62	0.62	0.65	0.62	0.72	0.75
romance	0.35	0.55	0.52	0.56	0.50	0.63	0.66
poetry	0.45	0.51	0.49	0.50	0.46	0.60	0.63
history	0.22	0.29	0.26	0.28	0.25	0.38	0.40
fantasy	0.41	0.60	0.57	0.59	0.55	0.68	0.70
fiction	0.52	0.66	0.64	0.67	0.61	0.74	0.76

Comparative Performance

This table and graph present the comparative performance of our model against various other machine learning methods for different book genres. Our T5 model consistently performs near the top, particularly in high-frequency genres such as fiction, nonfiction, and fantasy. It shows significant improvements over models such as Naive Bayes and Random Forest across most genres, especially those that are more distinguishable. However, while T5 performs well on frequent genres, its accuracy remains modest in less frequent ones like classics, historical fiction, and science fiction, where the differences between models are less noticeable. Compared to deep learning models like CNN and GRU, T5 generally shows superior results across the board. However, BERT slightly outperforms T5 in most genres, indicating that while T5 is a strong option, it may not be the optimal choice for solely genre prediction if marginal accuracy differences are important.

Conclusion

Through the application of data science techniques we have created a solution to our defining question of 'What makes a Successful Book'. With the creation of models based on a substantial amount of cleaned and processed data we are now able to predict rating scores of books based on attributes such as title, genre and description. Going further, we have also been able to predict the genre's of the books as well, providing a more well-rounded model that can be used to inform an optimal combination of title, genre and description. The performance of our model can be seen through the benchmark graphs showing the commendable accuracy of our model both in similarity to the benchmark data, as well as its performance when compared to other methods of classification.

Authors:

Abrar Fahim, Tahir Tamin, Liam Lynch, Said Soliev