PREDICTING WELL-RATED BOOKS

BY: STEPHANIE CHO

DECEMBER 2020 | SPRINGBOARD

THE PROBLEM

- Fact: Reading is good
- 27% of US adults didn't read a book in 2018 (Pew Research Center)
- 43 million of US adults possess low literacy skills (National Center for Education Statistics)

- How can we encourage more reading?

TO WHOM DOES IT MATTER?

- School education systems
- Parents
- Adults
- Publishing companies
- Authors

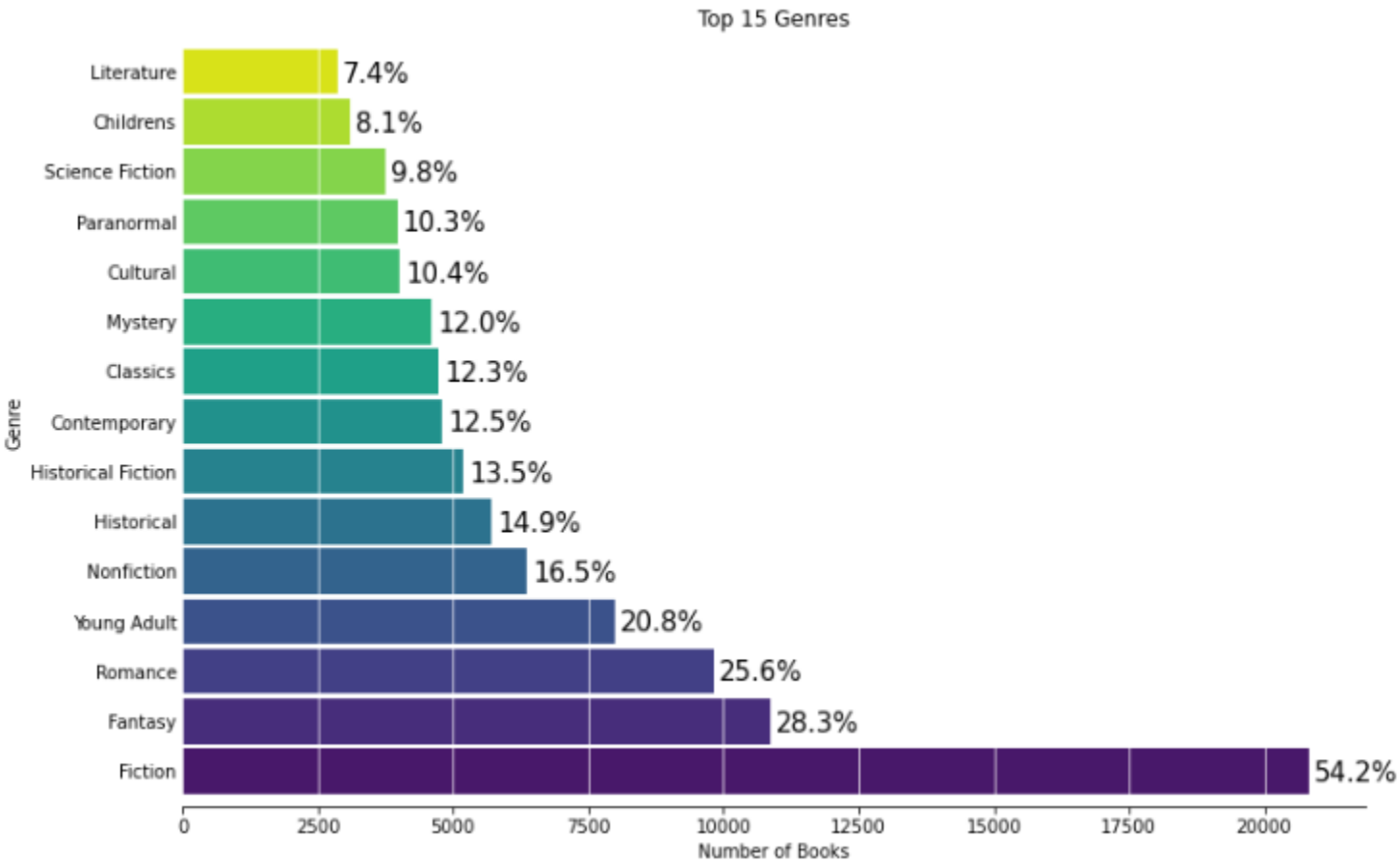
HOW TO ENCOURAGE MORE READING?

MACHINE LEARNING

METHODS

- Data: Goodreads Best Books List from Kaggle
- Factors: Genres, Book pages, Formats, Editions, Authors, Review ratings...
- Type: Supervised learning
- Problem: Binary classification (0 for 'bad' book and 1 for 'good' book)
  - Balanced data

DATA EXPLORATION



ALGORITHMS

- Baseline dummy classifier
- Logistic Regression
- Random Forest
- Ada Boost
- KNN

MODEL COMPARISONS

MODEL	F-1 SCORE [0, 1]	PREDICTION TIME (s)	FIT TIME (s)
Baseline	0.51, 0.50	0.001995	0.000996
Logistic Regression	0.00, 0.67	0.333113	0.002986
Random Forest	0.66, 0.64	56.12945	3.572085
Ada Boost	0.64, 0.62	8.274668	0.652254
KNN	0.60, 0.57	1.787197	23.685205

ASSUMPTIONS, LIMITATIONS, DISCLAIMERS

- Used only this Goodreads dataset to draw conclusions from
- Only physical books were analyzed (no audiobooks)
- The dataset contains mostly at least average books (lowest rating was a 2.09 with a mean of 3.99)
- Underfitting model

FUTURE IMPROVEMENTS

- NLP applied to books' descriptions feature of the original dataset or add features regarding languag-es
- ML Vision to analyze book covers
- Add more genres or deal with missing features dif-ferently
- Complement with other book datasets
- Expand into a book recommendation system

CONCLUSIONS

- Out of 4 supervised classification models, the Random Forest model provided the best results
- AUC for ‘good’ books = 0.708
- Underfitting
- But with more ideas, the model can be improved in the future

THANK YOU

Questions?