Predicting Yearly Science Fiction and Fantasy Awards

Zach Raines and Rohan Nair

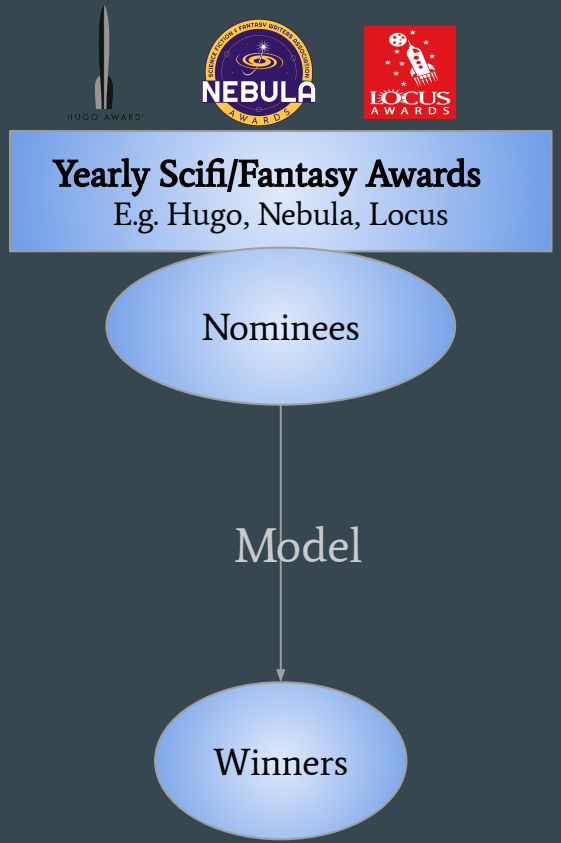
Erdős Institute Data Science Boot Camp

Problem Statement

Goal: Given a short list of nominees for yearly awards, predict which will be the winners.

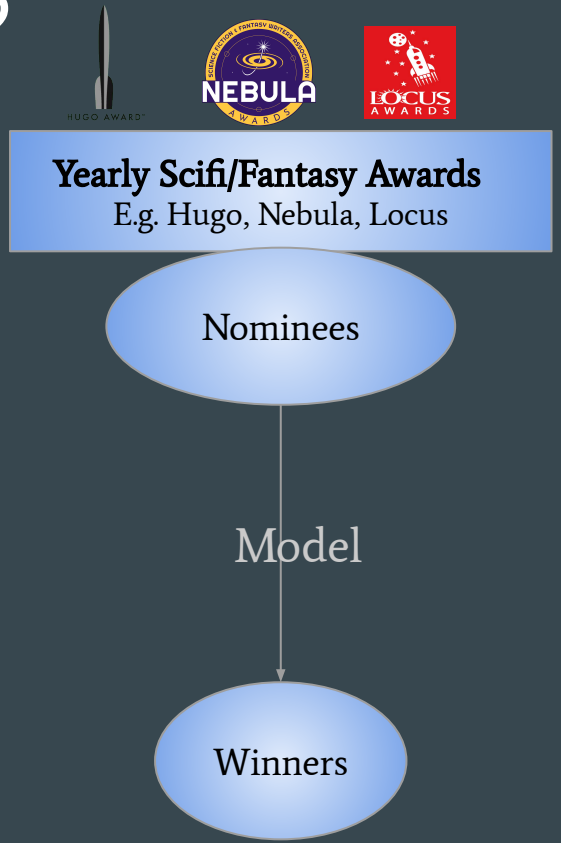
Nominees : a subset novels published within the preceding year

Winners : chosen from the nominees at a later date



Problem Statement: Why?

- Predicting winners allows publishers and other media producers to better position themselves
- An opportunity to support new talent
- A hard problem: little data is publicly available



Data

Book Data

- Book metadata from
  - WikiData
  - OpenLibrary
  - Google Books API
  - Wikipedia
  - Goodreads
- Author Metadata
  - Wikidata
- Bestseller status
  - New York Times

Automated data pipeline (Snakemake, duckdb, pandas, SentenceTransformers)

Tabular dataset of: All nominees from 1959-2024

Training 1959-2018

Test 2019-2024

World State

- Historical news headlines
  - New York Times

Features

Features

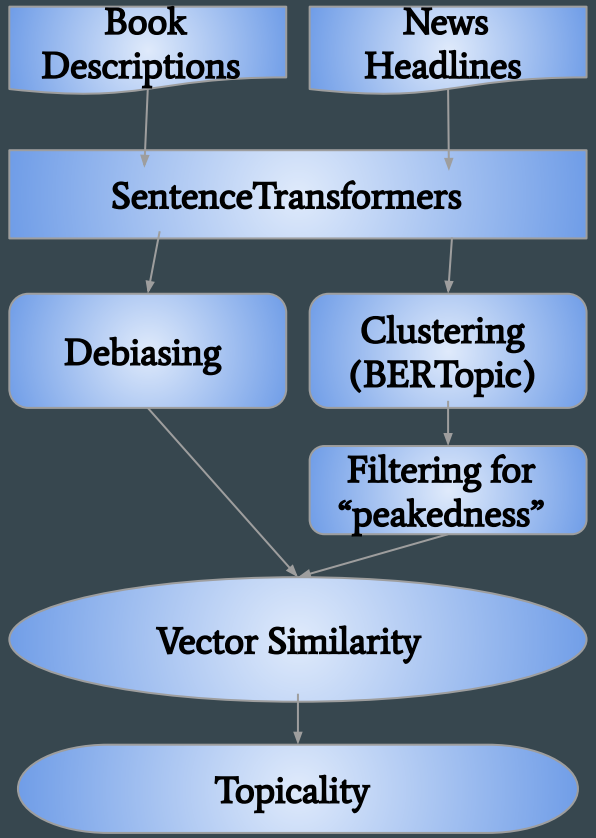
- Bibliographic Data
  - Title, Author, ...
- Biographical Data
  - Age, Gender, Country of origin, ...
  - Awards as of year
- Bestseller status
- Number of nominations
- **“Topicality Score”**

To handle the effect of competition between entries in the same year:

Simplest approach: scale features relative to cohort (nominees in same year)

Topicality score

- Generated news embeddings from the NYT front page
- BERTopic used for topic extraction and clustering
- Localized topics filtered by stationarity and kurtosis
- Book descriptions compared to yearly topics

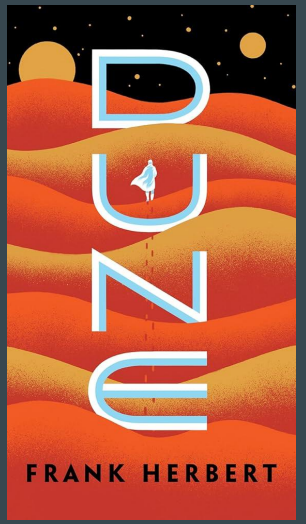


Modeling Approaches

Baseline model: "Naive" approach, using multinomial draw - predict winner based on how many nominations they have in a given year.

Can we do better than this?

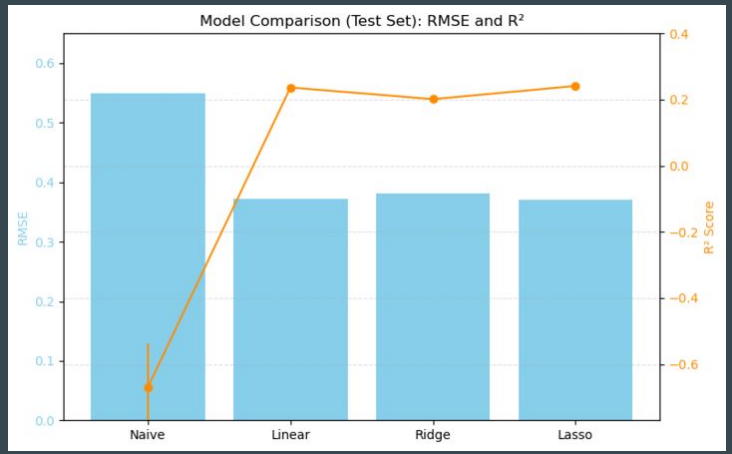
- Linear regression
- Ridge regression
- Lasso regression
- Logistic regression
- Decision trees
- Random forest
- XGBoost



Results

Can we do better than this? We can!

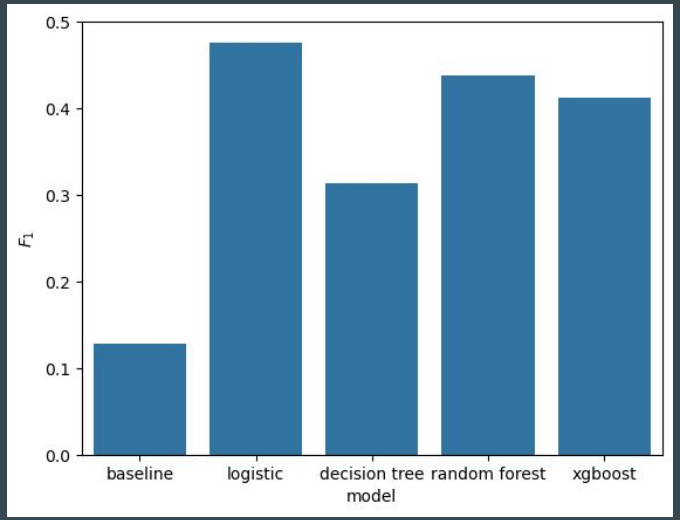
For linear, ridge, and lasso regressions, we use RMSE and R^2 scores to quantify predictive power over the naive baseline.



Results

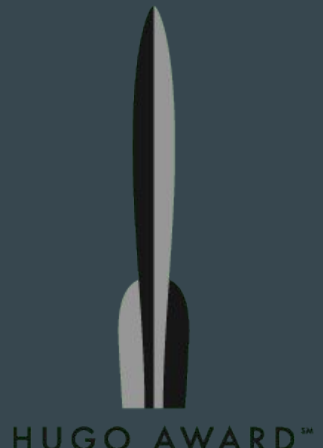
Can we do better than this? We can!

For logistic regression, decision trees, random forest, and XGBoost, we computed F1 scores over the naive baseline.



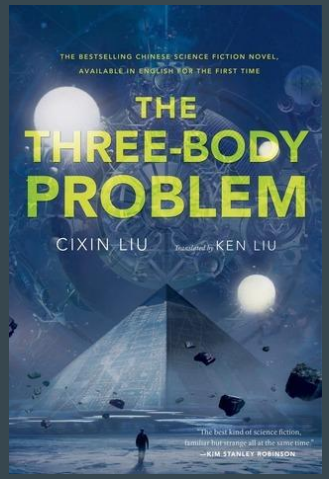
Conclusion

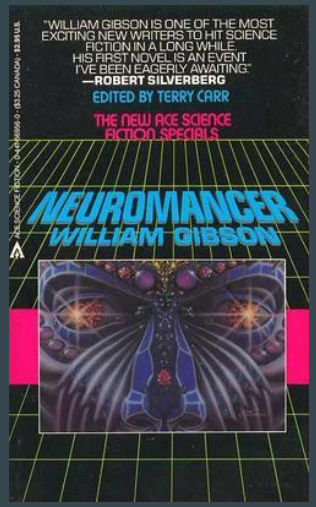
More to “winningness” than number of nominations - topicality, gender, age, and other features also play a role.



Challenges and Future Directions

- Data access: useful sales data is owned by publishers, not freely available.
- More robust topicality score.
- Extend models to determine what makes book “nomination”-worthy in the first place.





Acknowledgements

Thank you to our project mentor, Alec Traaseth, and to the whole Erdős Institute Data Science team for the opportunity!