Predicting Yearly Science Fiction and Fantasy Awards

Zach Raines and Rohan Nair

Github

Objectives

Each year, science fiction and fantasy novels are eligible to win several major book awards, including the Hugo, Nebula, and Locus. Given a list of nominees each year, our goal was to see if we could construct a model (or models) that would predict which nominees would actually go on to win awards.

Stakeholders

- **Book publishers** looking for up-and-coming authors.
- **Media outlets** trying to decide which works to acquire rights to and where to allocate resources.

KPIs

- **Classifier models**: F_1 score
- **Regression models**: R^2 and root mean squared error

Both were compared to a naive baseline which assigned win probability by the proportion of nominations a book received for a year.

Data Collection and Processing

Data was collected from free APIs and data dumps. A list of all nominees and winners of eight (8) sci-fi and fantasy awards 1959-2024, as well as author bibliography information, was obtained from the WikiData SPARQL API. This was supplemented with book bibliographic data and descriptions from OpenLibrary, the Google Books API, and Wikipedia, and bestseller rankings for the year of publication from the New York Times.

As a proxy for world events, we also downloaded all front page headlines from the New York times for the same date range, using the NYT API.

The book data was then preprocessed, cleaned, and combined using duckdb.

Modeling

Two types of modeling were performed:

- **Regression modeling** to predict the number of awards for each nominee
- **Binary classifier models** to predict whether or not a nominee won an award

Results

Both model classes outperformed the baseline by a significant margin

- **Regression:** Linear, ridge, and lasso RMSE < .4 vs. Naive RMSE > .5
- **Classifier:** XGBoost F_1 of 0.5 vs 0.15 baseline

Future Directions

- **More sales data and book information.** Much of this data is collected by book publishers and not freely available. Access to this information would help us improve the predictive power of our models.
- **More robust topicality score.** The “topicality score” feature was computed only using New York Times headlines, but we could potentially get better results using more information.
- **Studying nomination-worthiness**. This would allow us to train and test our models on a significantly larger data set, since every published science fiction and fantasy book is, in theory, eligible for nomination in the years of its publication.