# What makes a book a betseller?

by Kliton Andrea

Mentor
Matt Fornito

# Introduction

**GOAL:**

Predict the features that make a book successful

**Client:**

Publishing companies

**Data sources:**

- Open Library dump
- Goodreads

# Overview

1. Data source and data retrieval
2. Data cleaning and data wrangling
3. Exploratory Data Analysis
4. Hypothesis testing
5. Feature selection
6. Prediction models
7. Conclusion and recommendations

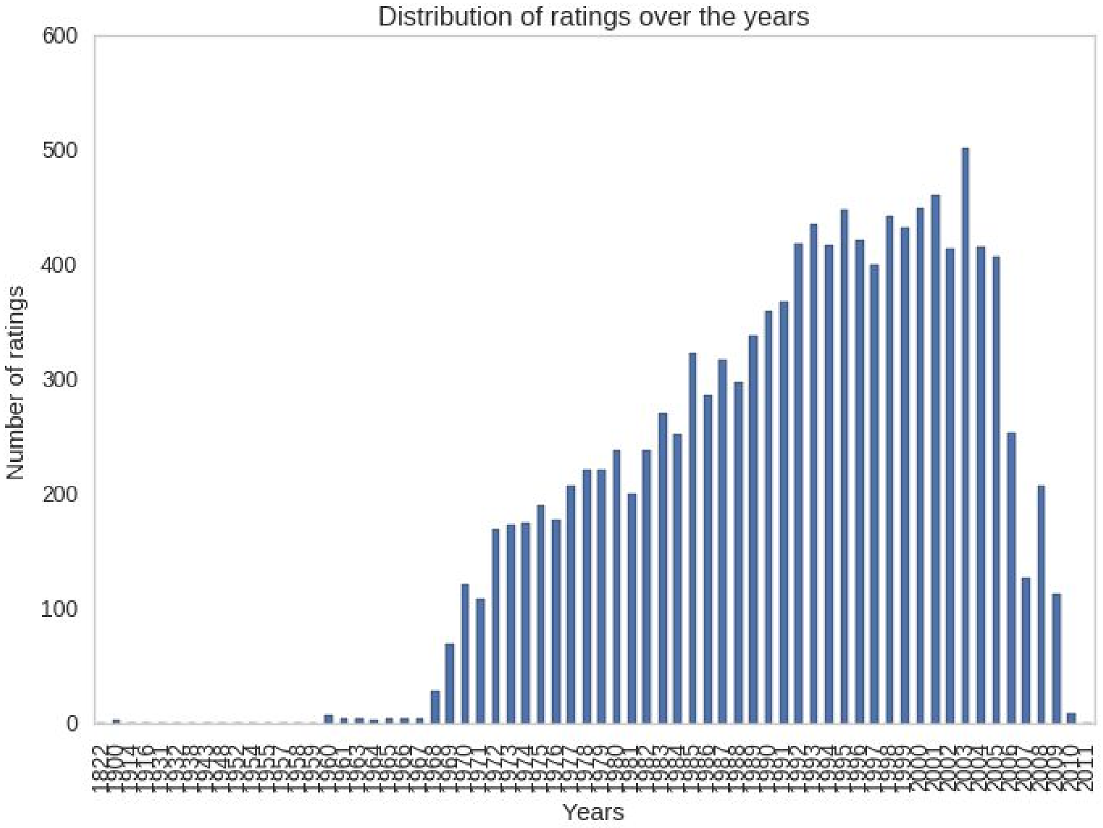
# Data Retrieval process

- Open Library dumps (https://openlibrary.org/developers/dumps)
  - Downloadable zip files (json format; 29602272 records * 5 files )
  - Contains Book Information
  - Contains Author Information
- Goodreads (https://www.goodreads.com)
  - Developer API (xml format)
  - Book information
  - Information about authors
- Amazon book details and reviews
  - Web scraping with Beautiful Soup based on the identifier from
  - Amazon policy limitation on web pages scraping does not allow to get all the information
  - Not usable

# Data Cleaning and Wrangling

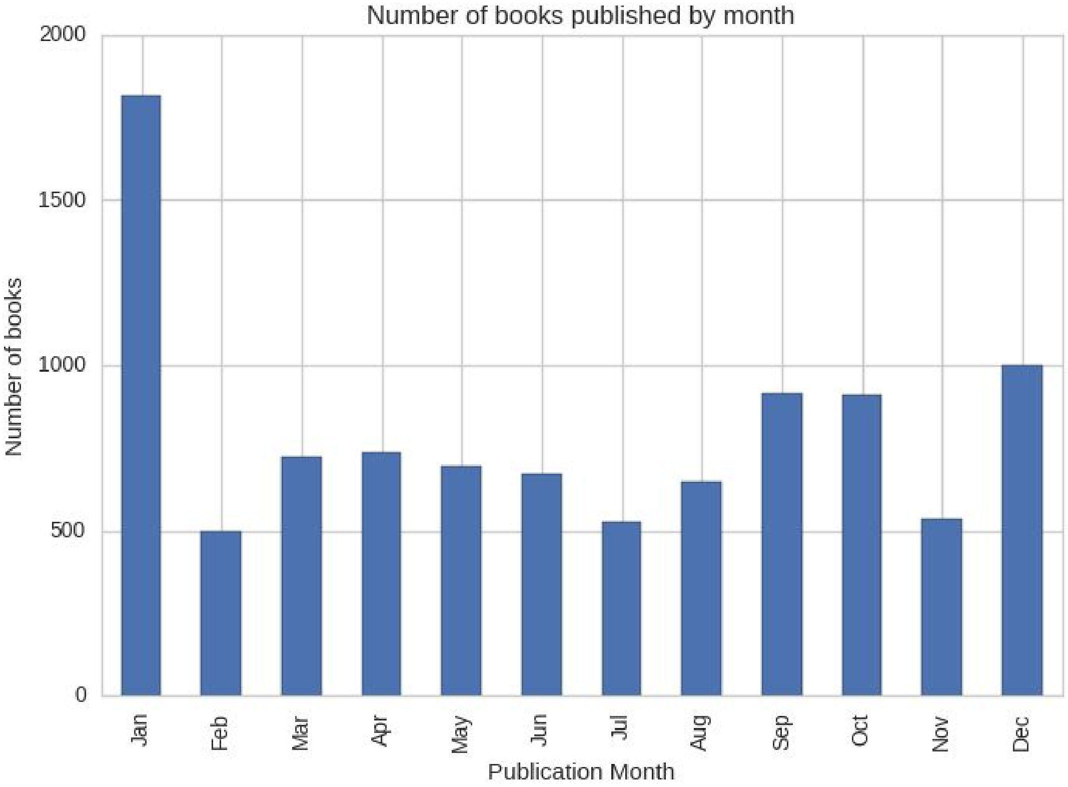
- Raw data converted to pandas DataFrame
- Removed most of the columns related to various book identifier, but ISBN
- Merged 4 DataFrames based in ISBN and authors
- Average ratings merged from Goodreads
- Records with no ratings removed
- Filtered out all books that are not related to art literature (Dewey code starting with ‘8’)
- Parsed and normalized date of birth from OL dumps
- Transformed Dewey classification codes to genres

# EDA: Book ratings over the years



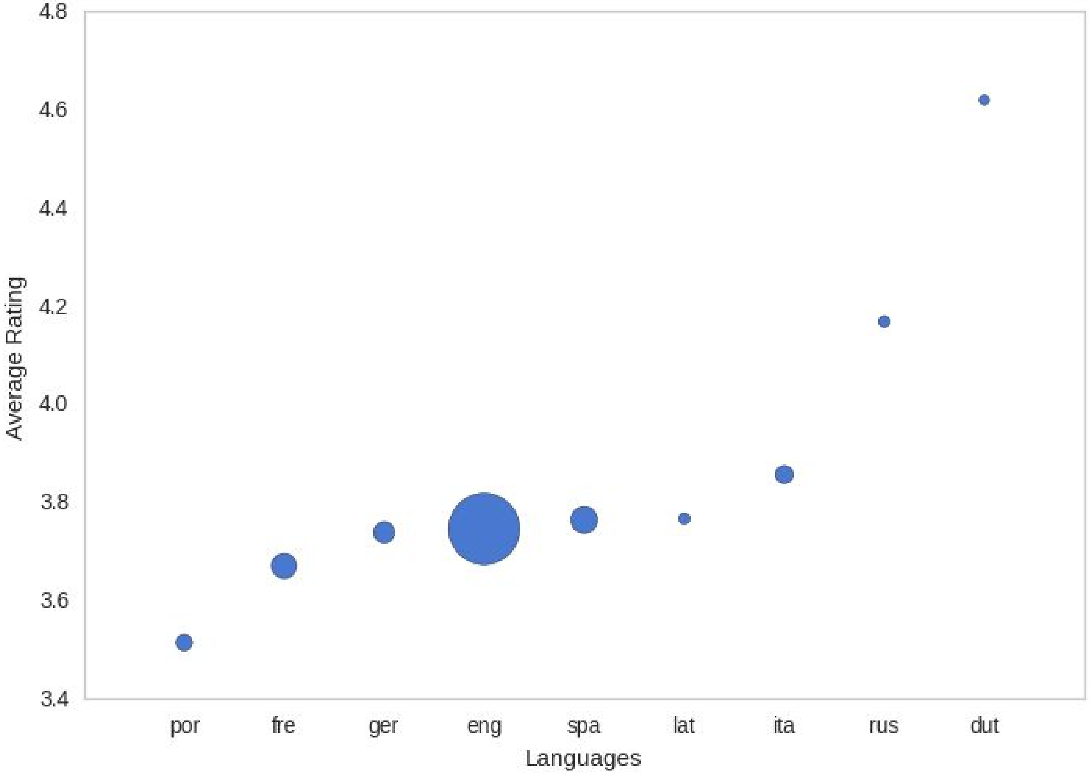
- This is an histogram of the number of books published every year and that have a rating.
- Peak year is 2003

# EDA: Book ratings per months



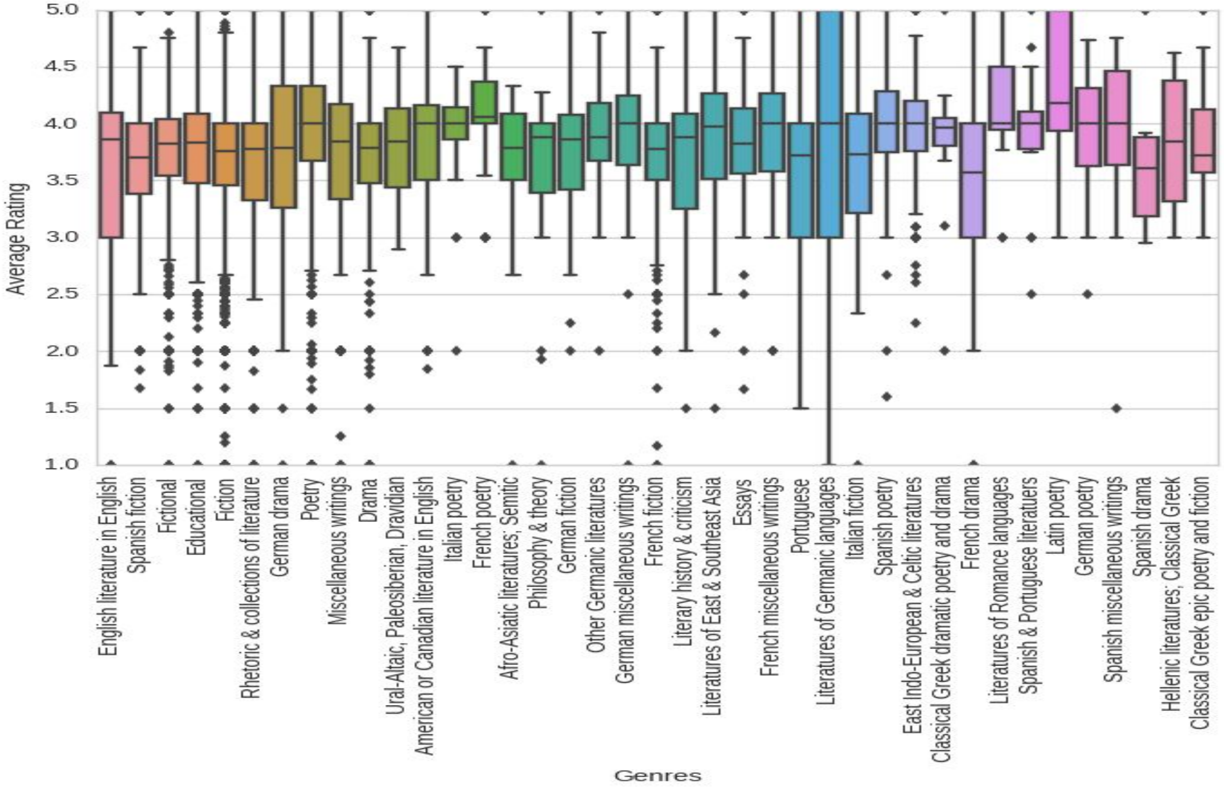
- There is a periodic sequence of book editions
- The high number of ratings on January actually is an outlier value. It could be a result of miscoded value.

# EDA: Average rating by language



- The diameter of a dot is weighted by the number of books in its respective group
- Dataset consist mostly of English books which is the ‘heaviest’ group in the plot

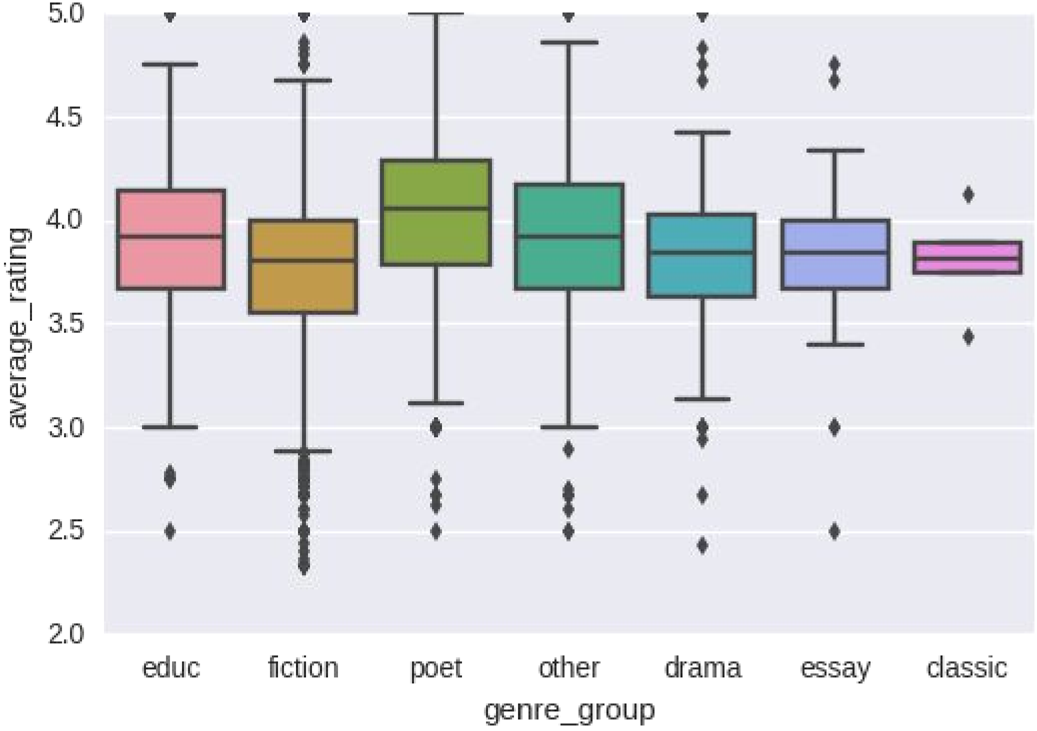
# EDA: Average rating by genres



Grouped the genres as per Dewey classification.

Boxplots show the average rating distribution for each genre book.

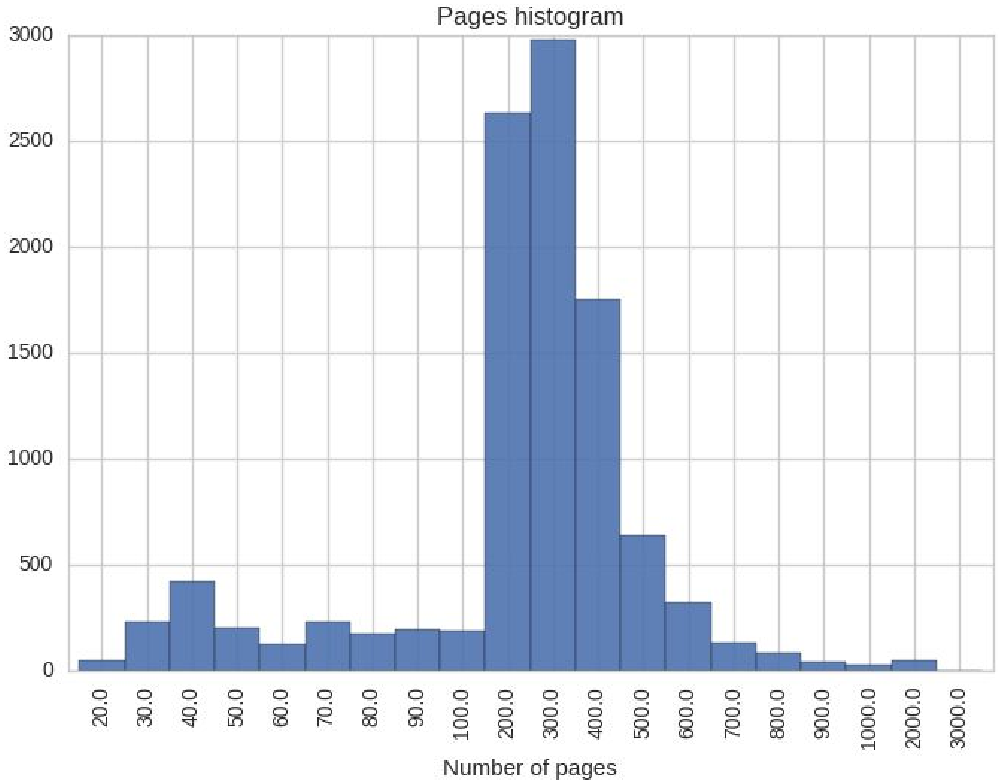
# EDA: Average rating for each genre group II



Reduced the number of genres to following:

1. Educational
2. Fiction
3. Poetry (best average rating)
4. Drama
5. Essay
6. Classic
7. Other

# EDA: Distribution of books by number of pages



200-500 pp. is the most popular range.

# Hypothesis testing

- There is no statistical significance in average ratings between genders in general.
- Average rating for Drama genre is higher for women, but the difference is not statistically significant.
- Women write more Educational books than men. However men get higher ratings and it does not seem to be a random effect.

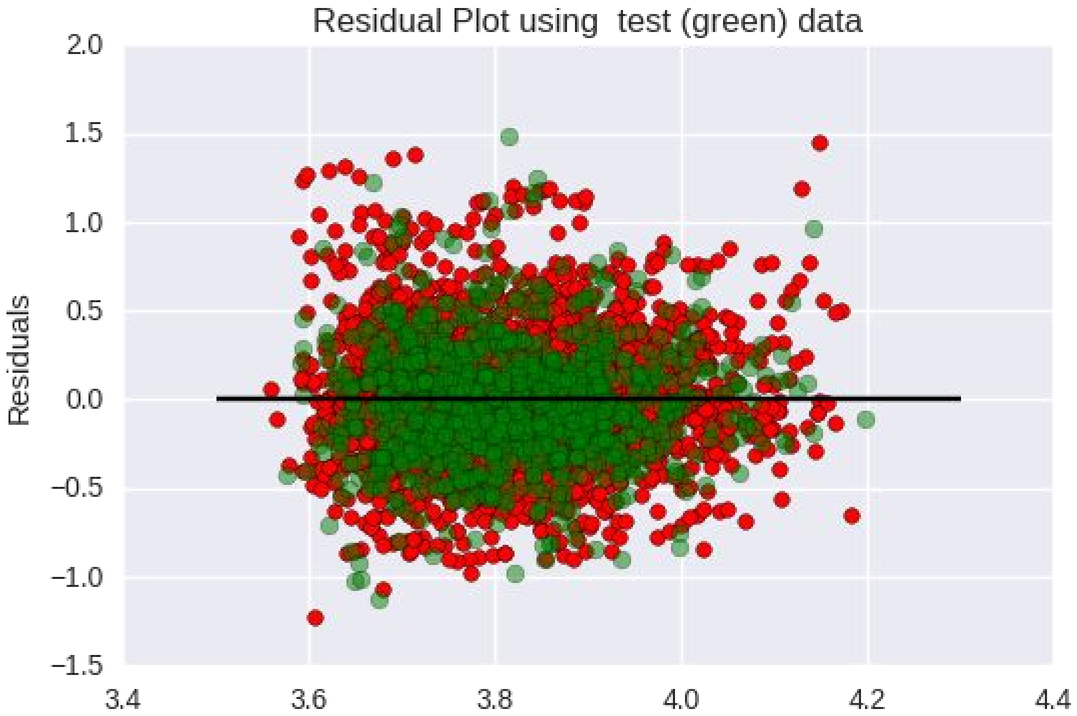
# Feature Selection

Selected features:

1. Fiction genre (categorical: [0, 1])
2. Number of pages (log transformed)
3. Number of words in the title (log transformed)
4. Number of contributors
5. Interval in years between the author date of birth and book publication date
6. Illustrated or not

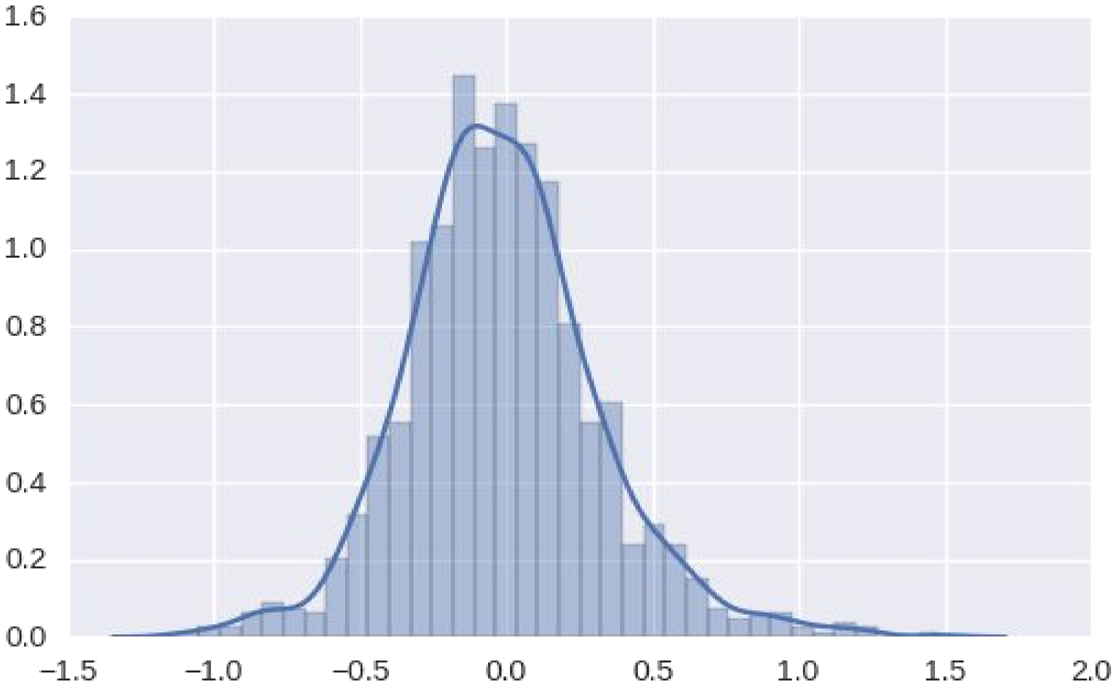
Dependent variable: average book rating

# Prediction Models: Ordinary linear regression



OLS MSE = 0.11

# Prediction Models: Ordinary linear regression

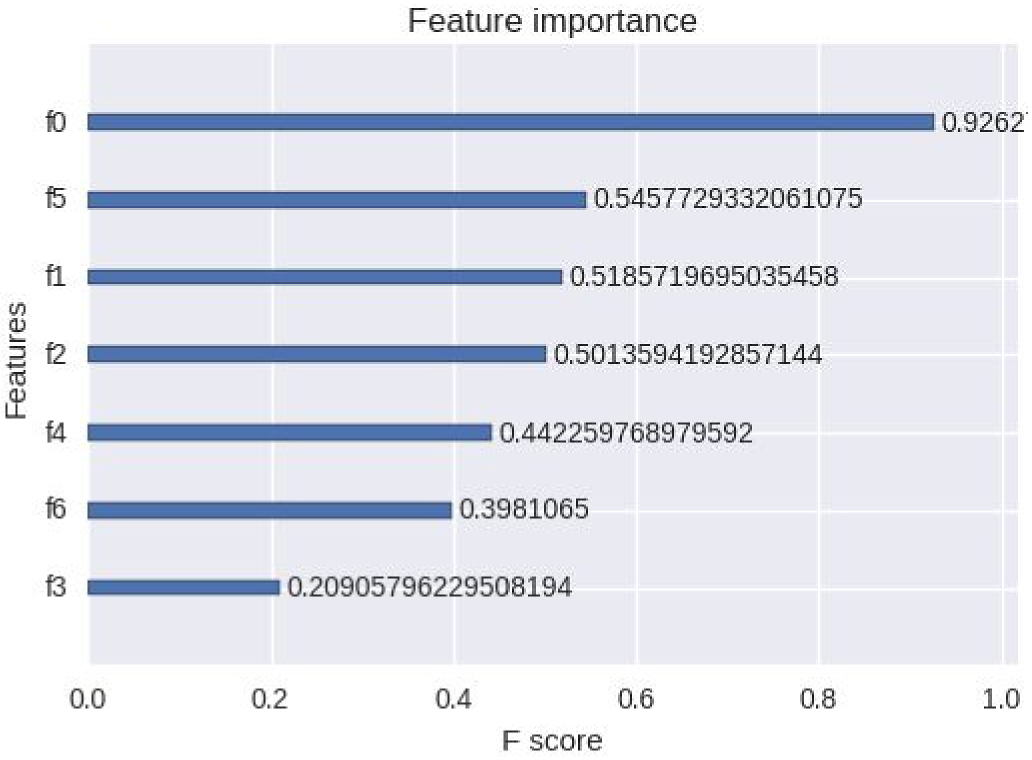


Residuals’ histogram

# Prediction Models: Various regressor

| **Model** | **Interc.** | **Fiction** | **Pages** | **Title** | **Contr.** | **Years** | **Fans** | **Ill.** | **MSE** |
|---|---|---|---|---|---|---|---|---|---|
| OLS | 3.47 | -0.16 | 0.01 | 0.14 | 0.03 | 0.11 | 0.03 | -0.03 | 0.11 |
| Ridge | 3.49 | -0.16 | 0.01 | 0.13 | 0.03 | 0.10 | 0.03 | -0.03 | 0.11 |
| BayesRidge | 3.5 | -0.16 | 0.01 | 0.13 | 0.03 | 0.10 | 0.03 | -0.03 | 0.11 |
| SVR | 3.42 | -0.17 | 0.02 | 0.15 | 0.08 | 0.14 | 0.03 | -0.02 | 0.11 |
| XGBoost | | | | | | | | | **0.1075** |

# XGBoost feature importance



f0 - is fiction

f1 - number of pages

f2 - no. words in title

f3 - no. of contributors

f4 - author maturity

f5 - no. of fans

f6 - is illustrated

# Recommendations and conclusions

1. Genre is an important feature to predict a success of a book. An author is likely to have a better rating if he chooses a genre different from fiction.
2. The popular authors are better positioned for success. Number of fans is the second most important feature according to XGBoost model.
3. A collection of literature works is most likely to get a good average rating. That is because a collection is compiled on successful past literature works.
4. Poetry is one of the highest rated genres in average.

# Next steps

1. The datasets lacks literature metrics, i.e. number of metaphors, idioms, euphemisms, etc.
2. Missing information on the structure of the story, number of persons in the narratives, number of topics in literature work.
3. It is possible to retrieve the information above using text analysis. On the other side it is time consuming and beyond the scope of this work.
4. More data about authors, their status, psychological metrics about their characters.

Thank you!