Analyzing Amazon Books Reviews

University of Pavia
Data Science and Big data Analytics course

Davide Ligari
Andrea Alberti
Cristian Andreoli

Dataset

- Kaggle dataset
- Two tables: Books data and Ratings
- Size: 3.86 GB
- Around 3 millions of reviews
- Ethical considerations

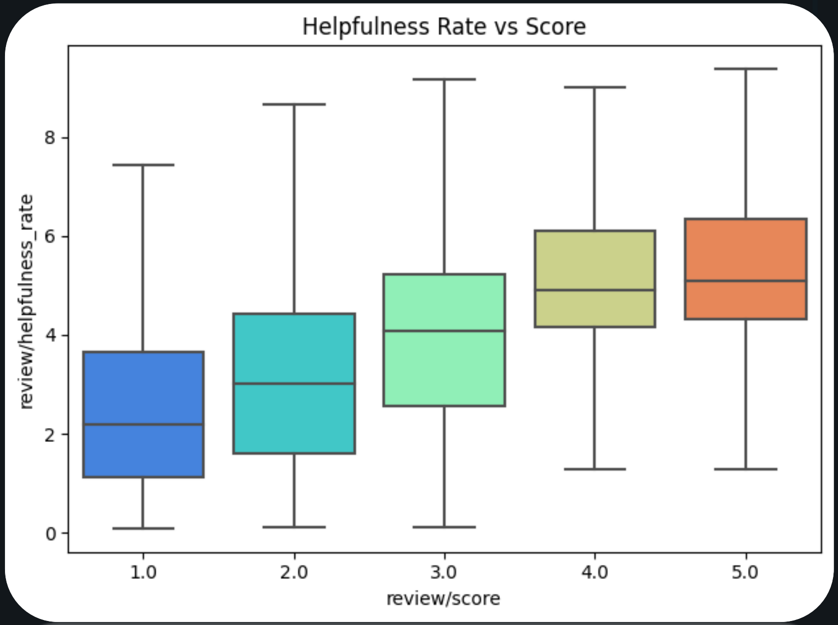
Data Table Schema:
Title
Description
Authors
Image
previewLink
Publisher
publishedDate
infoLink
categories
ratingsCount

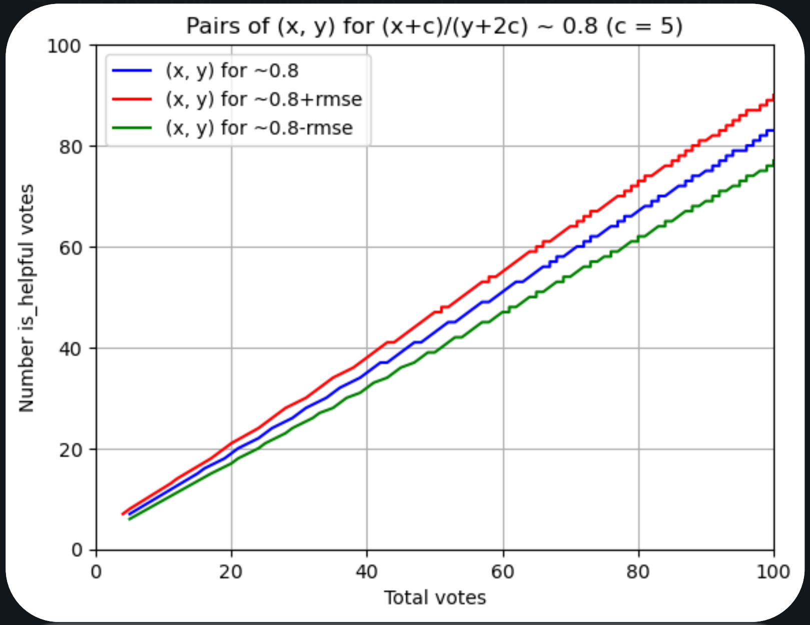
Ratings Table Schema:
Id
Title
Price
User_id
profileName
review/helpfulness
review/score
review/time
review/summary
review/text

Framing: Our Objectives

- Providing a scalable solution to the dataset exploration and analysis

- Developing a machine learning model able to predict the helpfulness of a review looking at its content

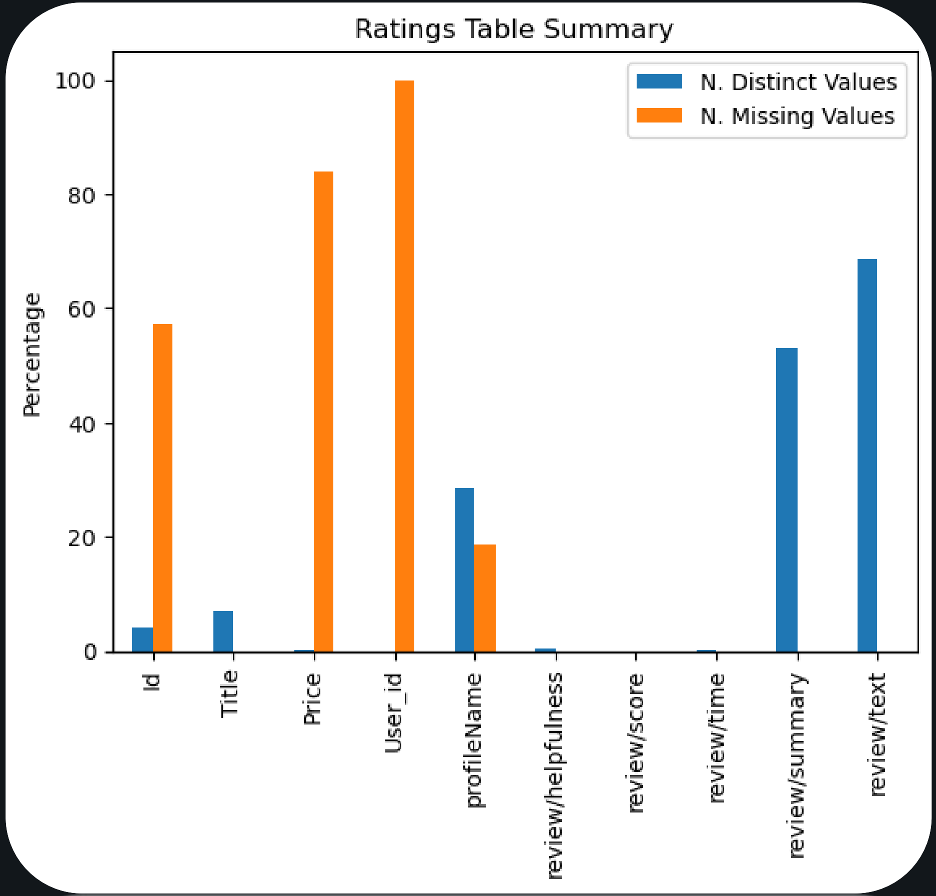


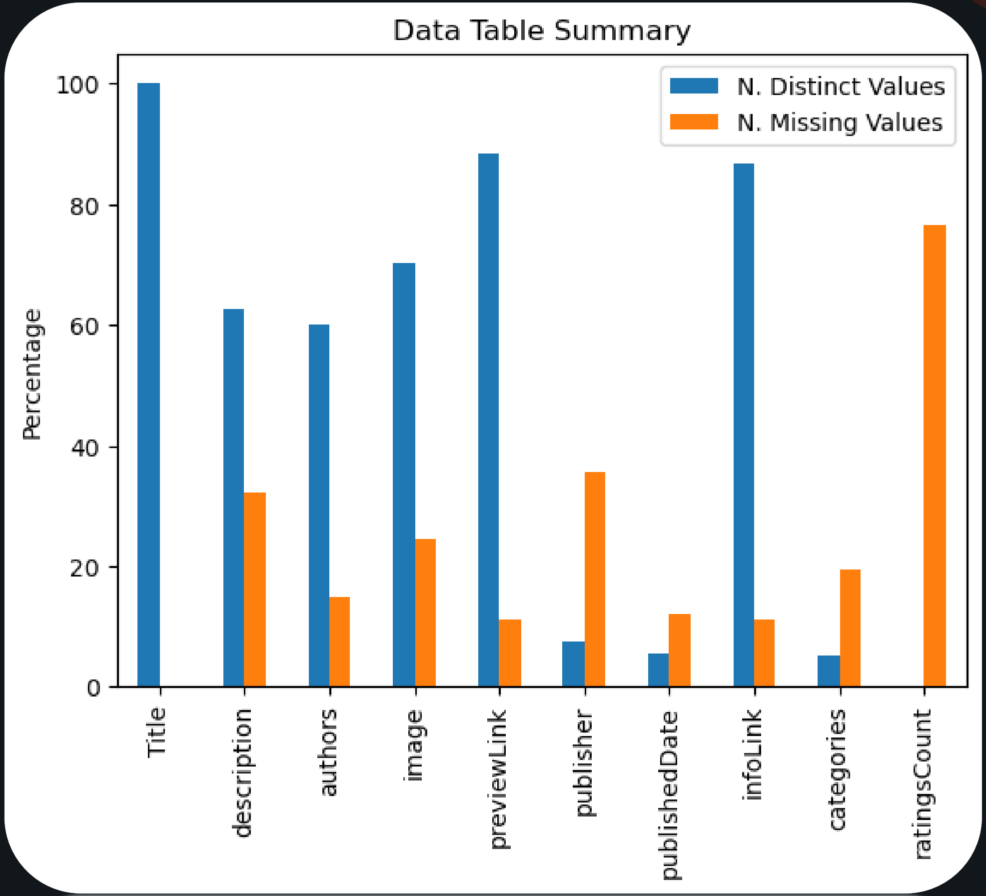


Workflow

- Load to HDFS
- Prior Analysis
- Data Cleaning
- MapReduce
- SandBox Creation

⬡ Prior Analysis ⬡





Data Cleaning

Methodology

- Duplicates deletion
- Unuseful columns deletion (those containing links)
- ‘Dangerous’ symbols deletion
- ‘Helpfulness’ columns splitting

review/helpfulness
7/7
10/10
10/11
7/7
3/3

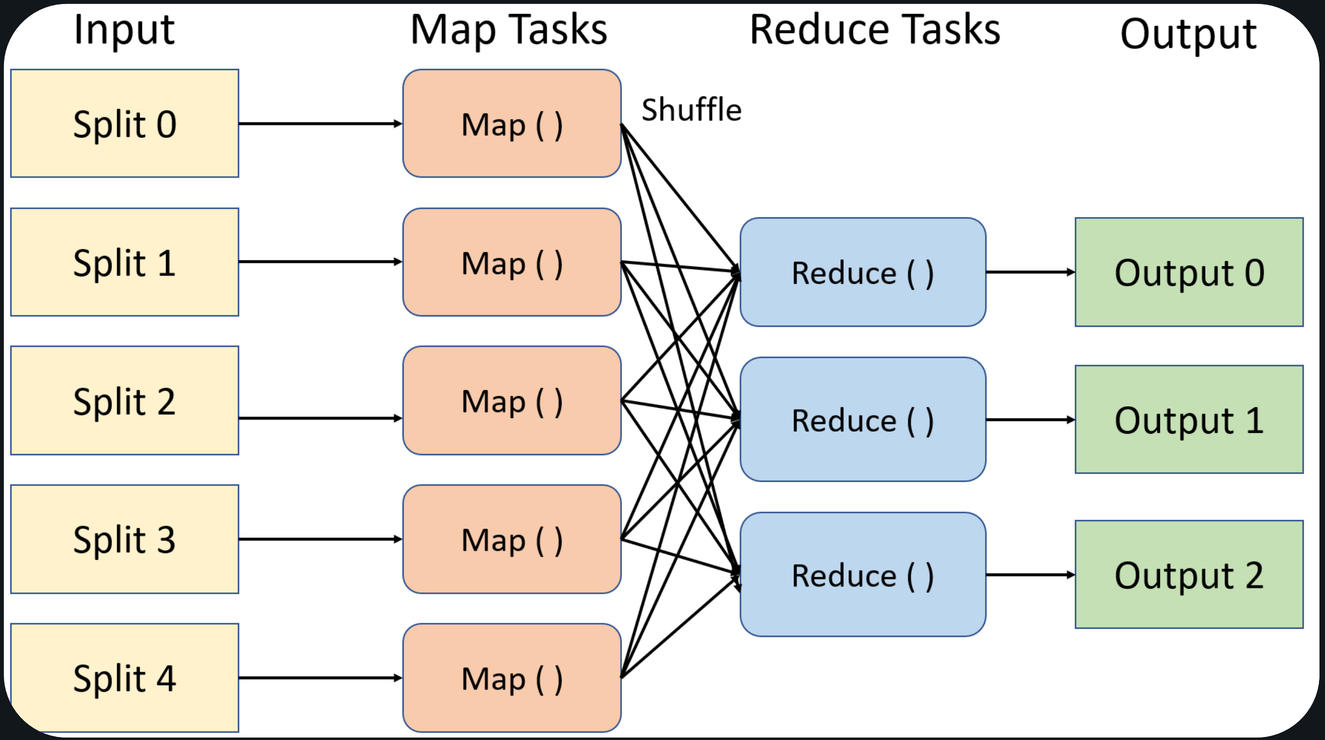
→

N_helpful	Tot_votes
7	7
10	10
10	11
7	7
3	3

MapReduce Job

Join the two tables

- Mapper creates a key-value structure
- Double key sorting (Title, second field)
- Reducer performs the join
- The table is stored in Hadoop



SandBox Creation

Methodology

- **Sandbox** on **MongoDB**
- Random sample
- **Rappresentative** subset

```
# Connect to MongoDB
import pymongo

client = pymongo.MongoClient('mongodb://localhost:27017/')
database = client['spark_db']
books = database['books_joined']
reviews = database['book_reviews']

# Load the data
df_joined = spark.read.csv("hdfs://localhost:9900/user/book_reviews/joined_tables",
header=True, schema=joined_schema, sep='\t')


# Select a random subset of the big data to import
N_to_sample = 300000
df_sample = df_joined.sample(withReplacement = False, fraction =
N_to_sample/df_joined.count(), seed = 42)

# Convert to a dictionary
df_sample_dict = df_sample.toPandas().to_dict(orient='records')

# Insert into MongoDB
books.insert_many(df_sample_dict)
```

Hypothesis Testing

Methodology

- **MongoDB** query to get data ready for analysis
- **SciPy** to compute metrics
- **Pandas** data manipulation
- **Seaborn** and **Matplotlib** for graphs

```
# Remove the samples which have no score or helpfulness data
pipeline_remove =
        {'$match':{
                'review/score':{'$exists':True},
                'N_helpful' :{'$exists':True, '$ne':0},
                'Tot_votes' :{'$exists':True, '$ne':0}
                }
        }

# Retain only the required fields
pipeline_project =
        {'$project':{
                'review/score':1,
                'review/helpfulness_rate':{
                        '$multiply':[
        {'$divide':['$N_helpful','$Tot_votes']},
                                {'$sqrt':'$Tot_votes'}
                        ]
                },
                '_id':0,
                'Tot_votes':1,
                'N_helpful':1
                }
        }

books_data = books.aggregate([pipeline_remove,pipeline_project])
```

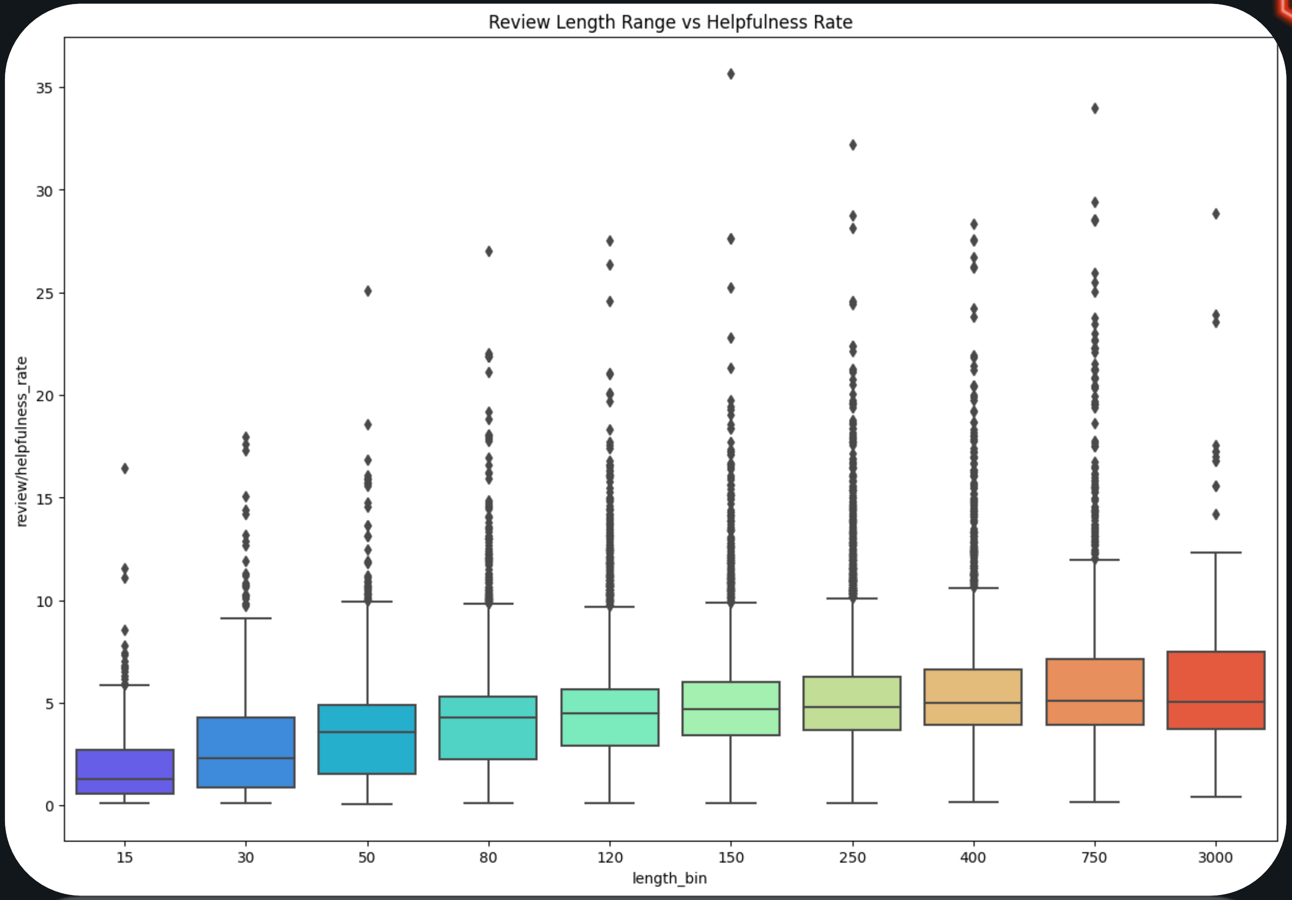
Hypothesis 1

Is the helpfulness correlated to the length of the review?

$$helpfulness\ score = \frac{x}{y}\sqrt{y}$$

- Spearman's correlation value: 0.331

- P-value < 0.05

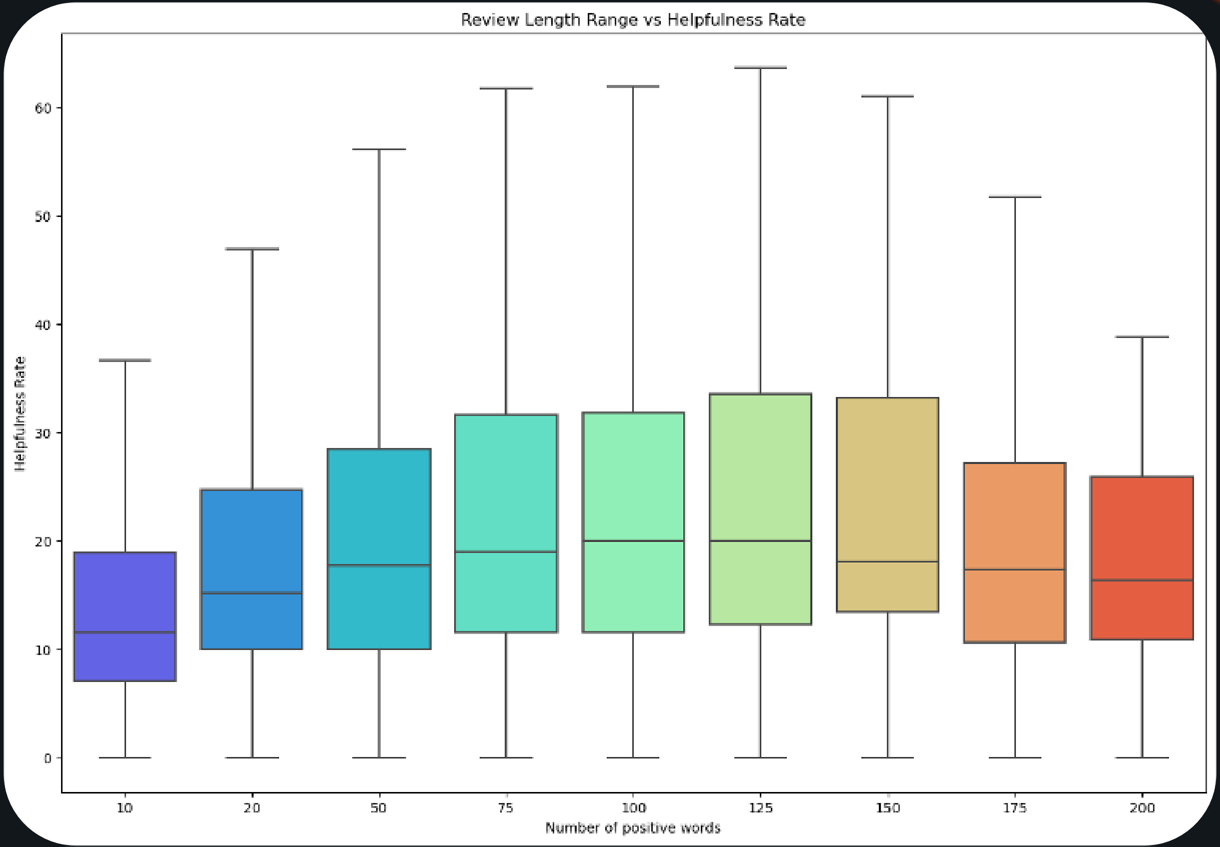


Hypothesis 2

Is the number of positive words correlated to helpfulness?

Multinomial NBC :
→ ***top 800 positive words***

- Spearman's correlation value: 0.318
- P-value < 0.05

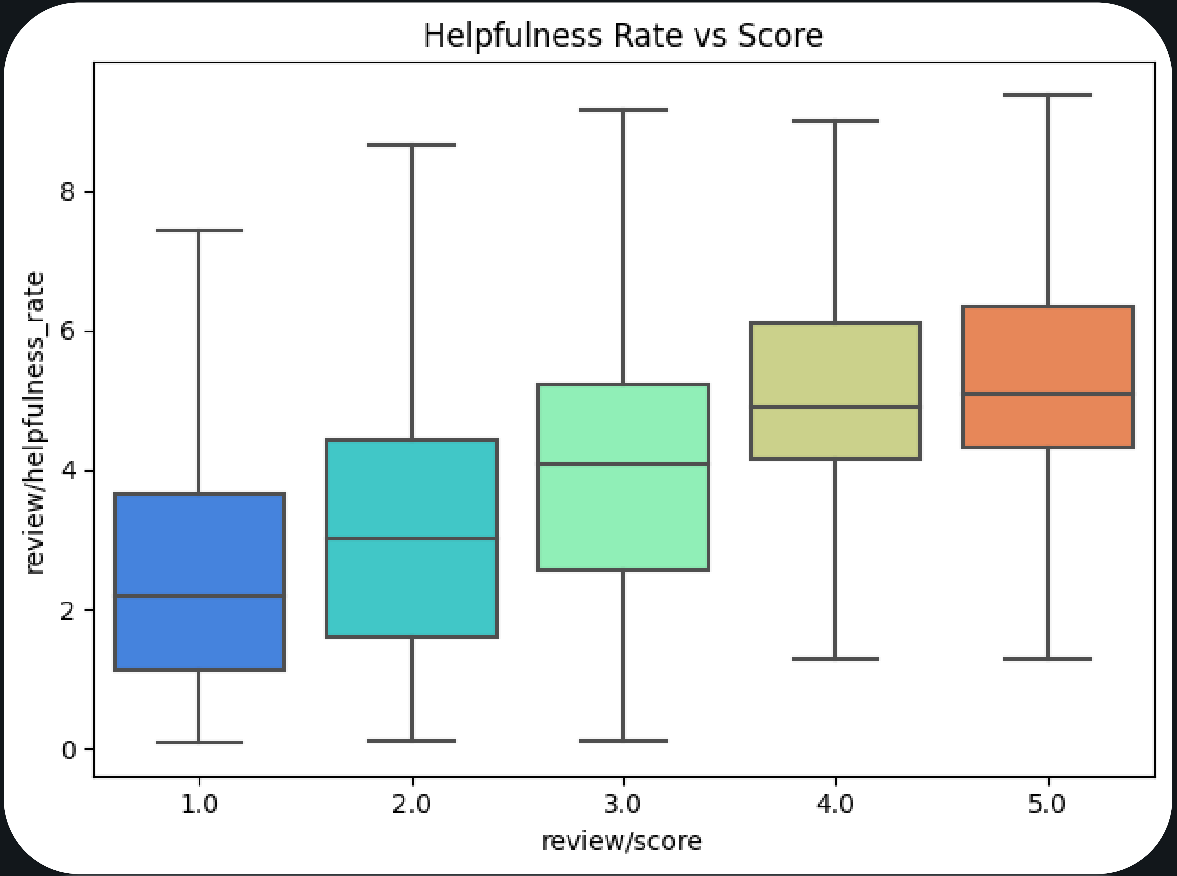


Hypothesis 3

Is there correlation between rating score and helpfulness?

$$Tot\ votes < 20$$
$$\rightarrow leads\ to\ bias$$

- Spearman's correlation value: 0.525
- P-value < 0.05



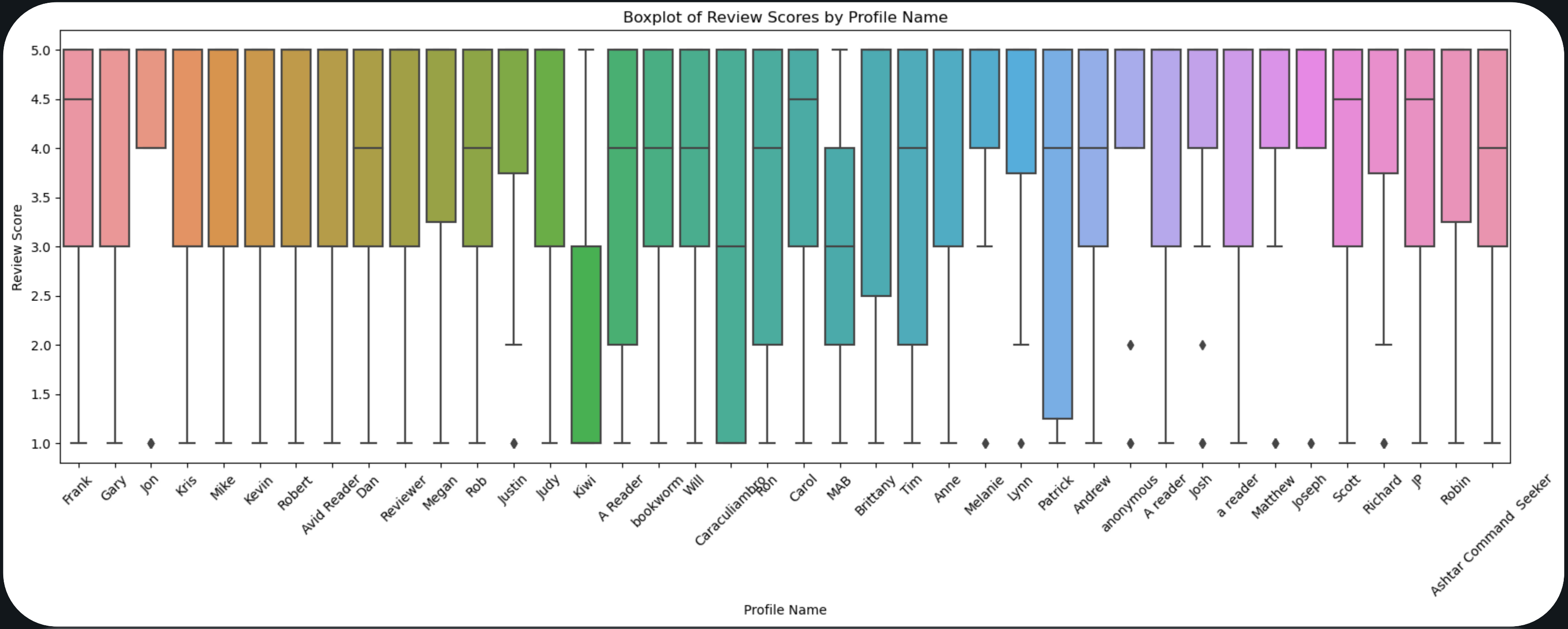
Hypothesis 4

Is the rating score influenced by the user?

$N.reviews < 20$
$\rightarrow$ *leads to bias*

ANOVA test

- F-statistic: 1.537
- P-value: 0.067



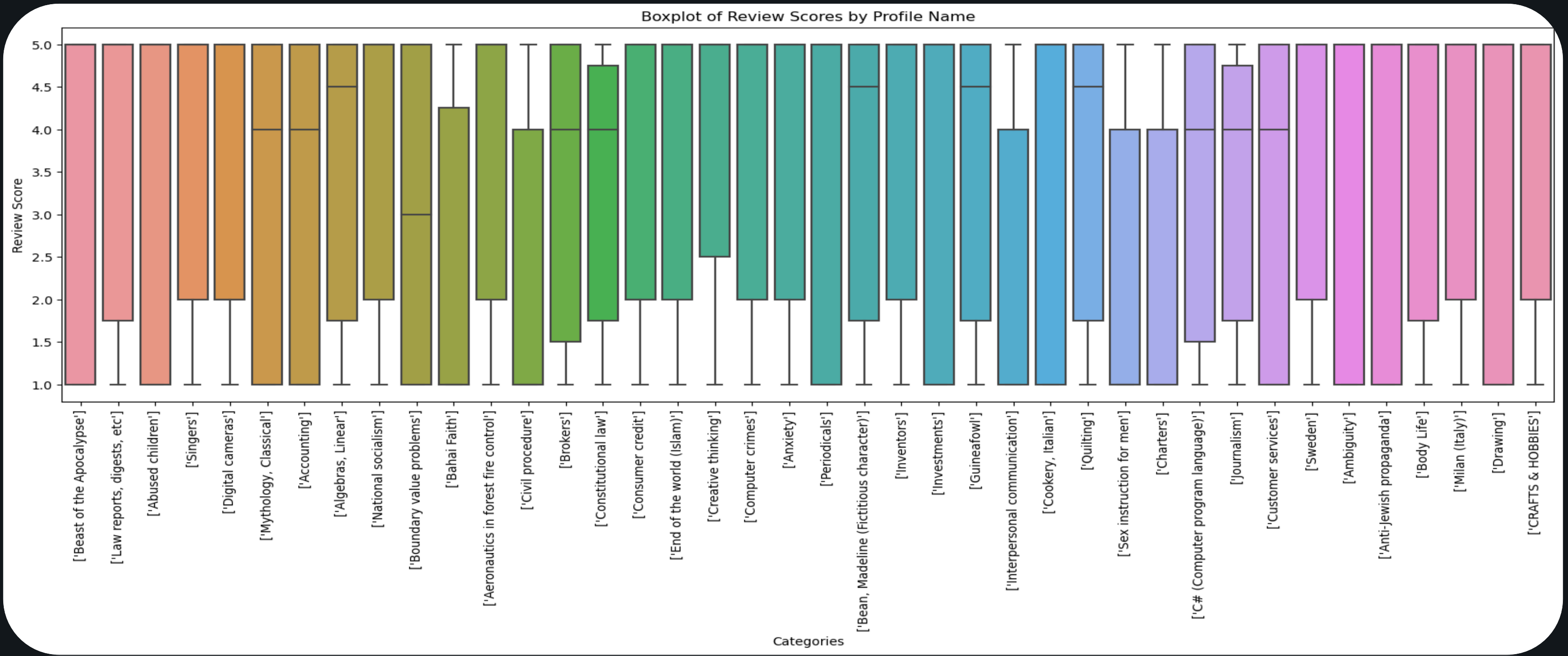
Hypothesis 5

Is the rating score influenced by the category of a book?

$N.reviews < 20$
$\rightarrow$ ***leads to bias***

ANOVA test

- F-statistic: 0.177
- P-value: 0.999

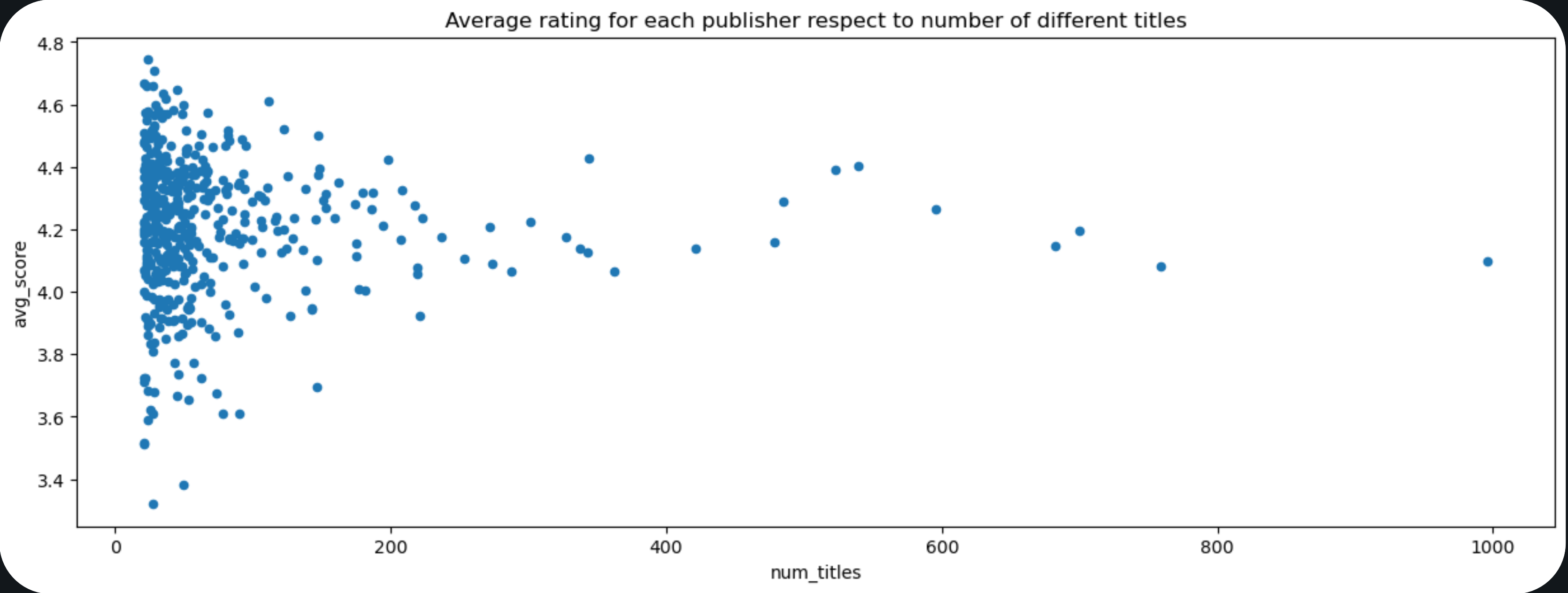


Hypothesis 6

Is there correlation between the number of books of a publisher and the review score?

$N.\,books < 20$
$\rightarrow$ ***leads to bias***

- Spearman's: -0.067
- P-value: 0.151

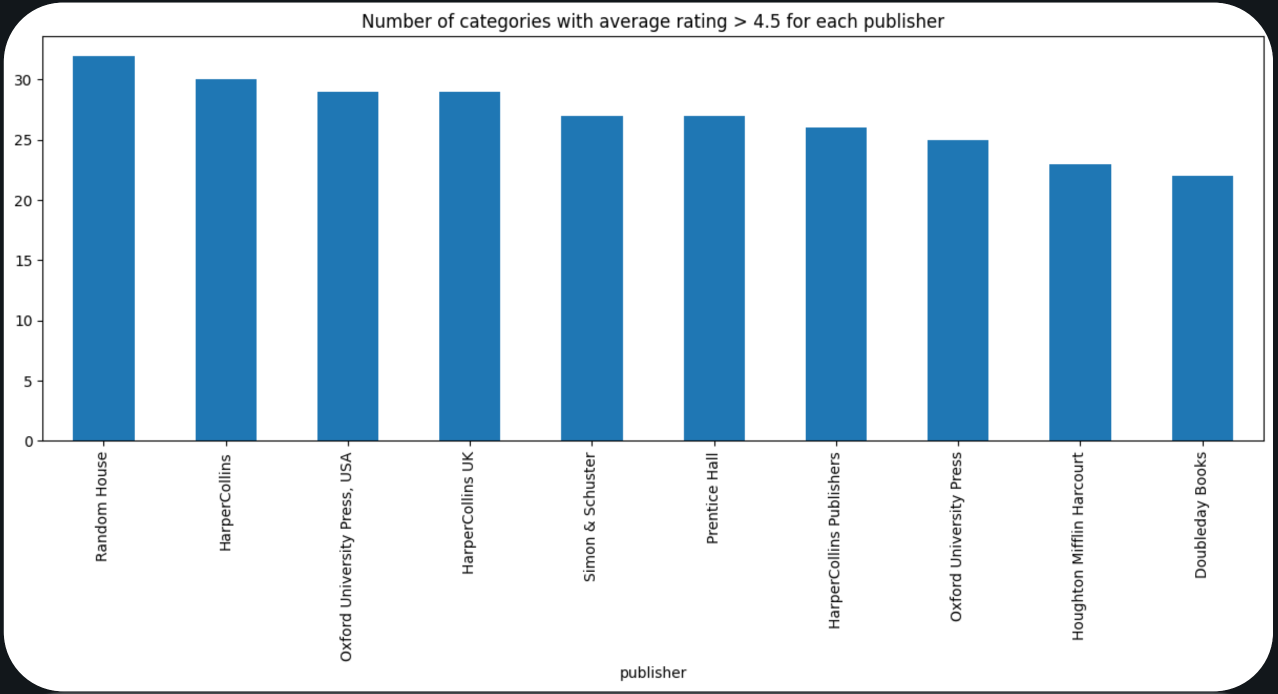


Curiosity

Which are the best publishers?

In which category are the best publishers focused?

Complex MongoDB query



Real Scenario

Goals

- Provide **scalable** solution
- Prove results **consistency**

Tools

- Spark DataFrame
- Pyspark.ml

Hypothesis 1

	Spearman Coeff
Hadoop	0.361
Sandbox	0.331

Hypothesis 2

	Spearman Coeff
Hadoop	0.318
Sandbox	0.318

Hypothesis 3

	Spearman Coeff
Hadoop	0.527
Sandbox	0.525

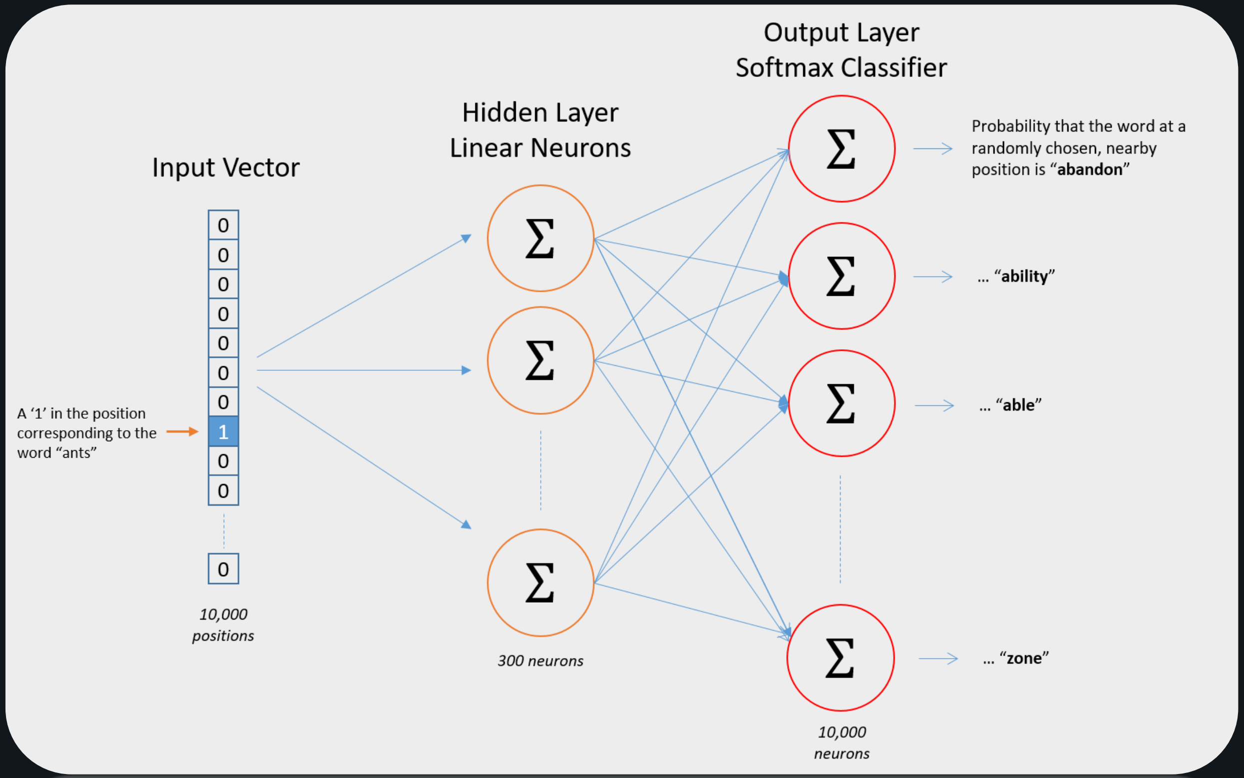
Helpfulness Prediction

- Features Extraction
- Model Selection
- Best Model Analysis

Features Extraction

Creation steps

- Word2Vec from Gensim
- Size = 30, Window = 5, Min count = 2
- Size = 150, Window = 5, Min count = 2
- Review = average of contained words



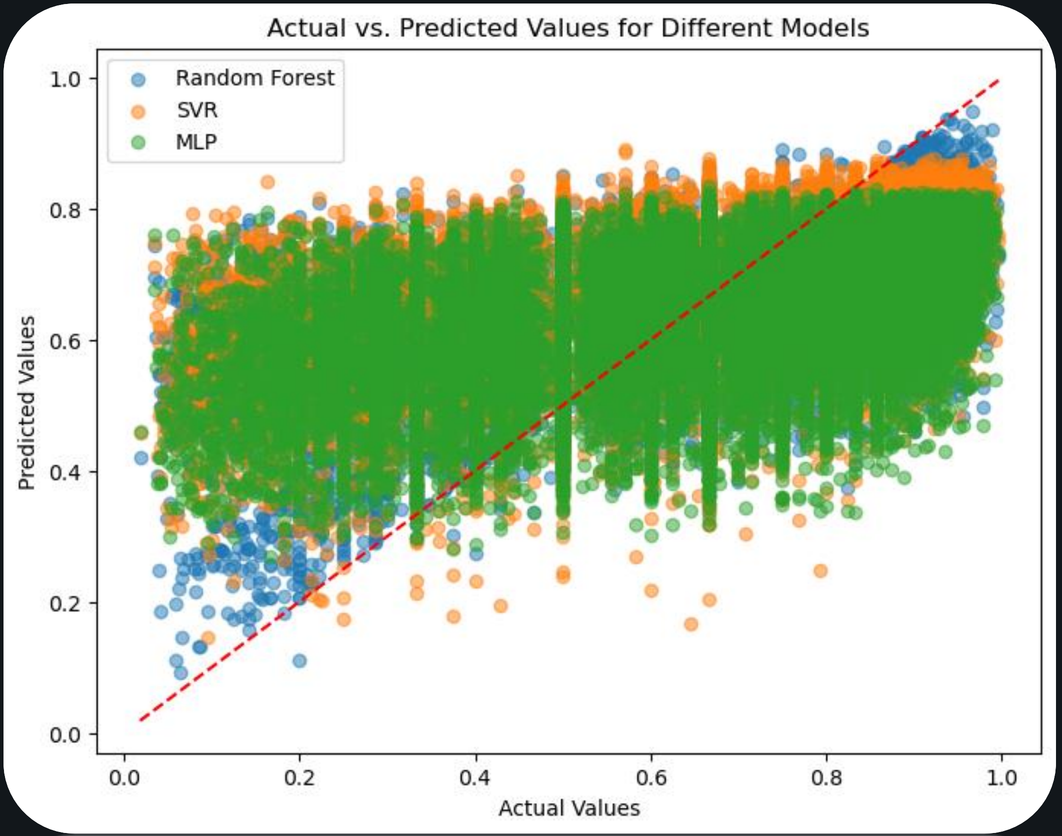
Model Selection

Trained Models

- Random Forest Regressor
- Support Vector Regressor
- MLP Neural Network

GridSearchCV → Hyperparameters selection

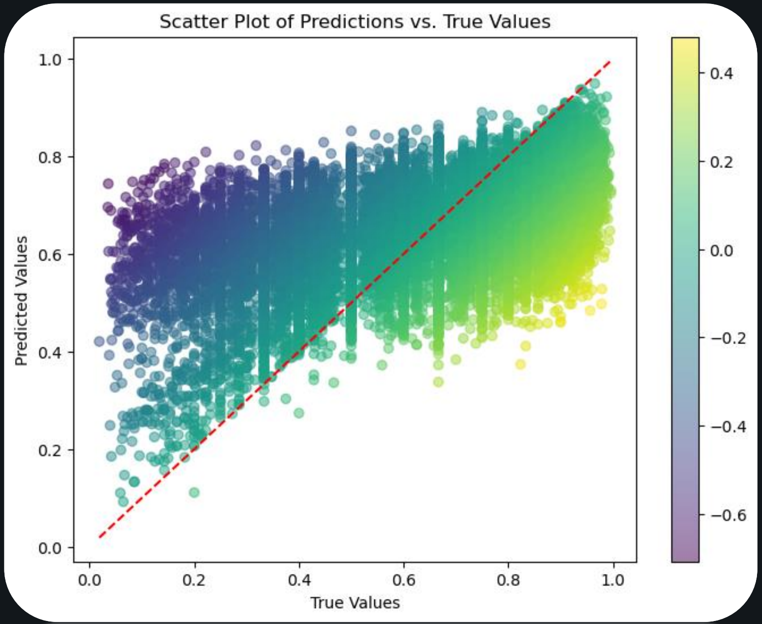
Model	MSE	RMSE	R^2
RF	0.0259	0.1609	0.2532
SVR	0.0279	0.1670	0.1955
MLP	0.0282	0.1680	0.1858

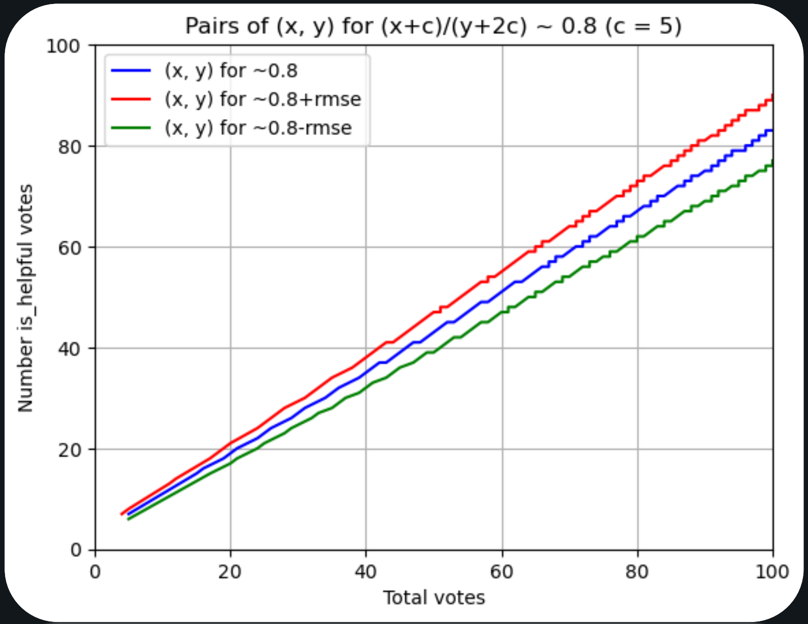


Best Model: Random Forest

- Size = 150 small improvement
- *Underestimate* when low and *Overestimate* when high

- Impact of RMSE on helpfulness votes
- 100 Total votes → ± 13 helpful votes





Conclusions

- **Importance of Scalable Systems:** Emphasizes the significance of scalable systems in data analysis.
- **Review Length and Sentiment:** Longer reviews, especially the ones with positive words, tend to be more useful, but excessively long reviews can be tedious.
- **User Preference for Positive Reviews**: Users find positive reviews more helpful.
- **Objective User Ratings:** User ratings appear to be unbiased and reflect objective evaluations of books.
- **Experience vs. Appreciation:** The experience of publishers does not necessarily correlate with higher appreciation from users.
- **Future Work:** Indicates a focus on feature engineering to enhance the model's performance.

THANK YOU!