Team 27

Books recommendation system

Team members:

- Arsenii Pavlov
- Ivan Beltsov
- Andrei Pavlov
- Timofei Ivlenkov
- Bulat Latypov

Introduction

In the digital age of book publishing, understanding reader preferences through ratings is crucial for platforms and publishers. This project leverages machine learning to predict book ratings (1–5 stars) using a rich dataset from Amazon containing 3 million user reviews across 212,404 unique books (1996–2014).

Business Objectives

We aim to enhance user experience on discovering new books by providing personalized recommendations by predicting his review of the book. Also we want to increase platform engagement and drive higher user retention and interaction by improving the relevance of suggestions. Translate accurate recommendations into increased book purchases or downloads directly benefiting publishers and online retailers.

Data Description

The dataset employed in this project consists of two primary files—Books_rating.csv and books_data.csv—which together form a rich source of information for building a robust and personalized book recommendation system. These datasets are complementary: one captures user-generated feedback and behavior, while the other provides detailed metadata about the books themselves.

1. **Books_rating.csv**

This file contains over 3 million user reviews for books, representing a diverse and extensive collection of reader feedback accumulated over nearly two decades (1996–2014). Each row in the dataset represents a single review submitted by a user for a particular book. The structure of the data includes:

Id: A unique identifier for the book being reviewed.
Title: The name of the book.
Price: The listed price of the book, when available.
User_id: A unique identifier for the user who submitted the review.
profileName: The name (or pseudonym) of the user.
review/helpfulness: A metric indicating how helpful other users found the review, expressed as a ratio
review/score: The numeric rating given to the book, ranging from 0 to 5.
review/time: A Unix timestamp indicating when the review was posted.
review/summary: A brief summary or title of the review.
review/text: The full text of the review, which often provides valuable insights into the reader's opinions and preferences.

This dataset not only provides a basis for collaborative filtering but also allows for sentiment analysis, trend detection over time, and the identification of user profiles based on review behavior.

2. **books_data.csv**

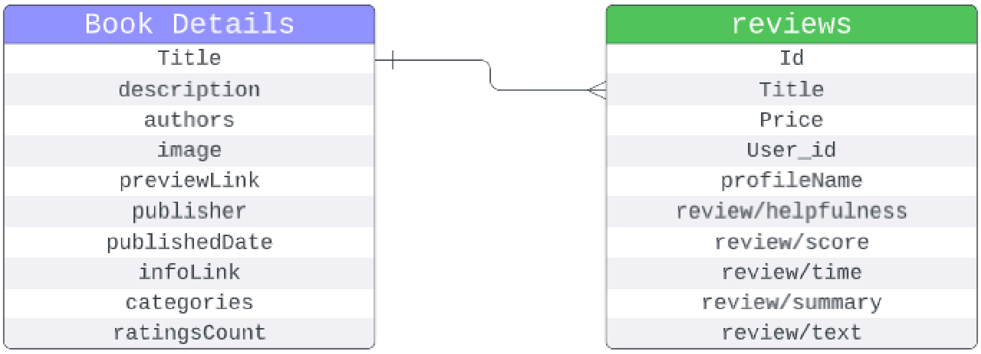
This file offers detailed metadata for 212,404 unique books, most of which are referenced in the reviews dataset. The metadata includes bibliographic and descriptive fields that are essential for content-based recommendation methods. Key attributes include:

Title: The book's title, which serves as a key for merging with the reviews.
description: A summary or synopsis of the book's content, useful for NLP-based content analysis.
authors: The author(s) of the book.
image: A URL to the book's cover image.

previewLink: A link to a Google Books preview of the book, when available.
publisher: The publishing company.
publishedDate: The original publication date.
infoLink: A Google Books link providing more detailed information about the book.
categories: The genre or subject classification of the book (e.g., fiction, history, science).
ratingsCount: The total number of ratings the book has received, which can be used as a proxy for popularity.

The inclusion of rich descriptive fields enables deeper user-book profiling and supports hybrid recommendation approaches by combining collaborative and content-based methods.

The data can be described as diagram:

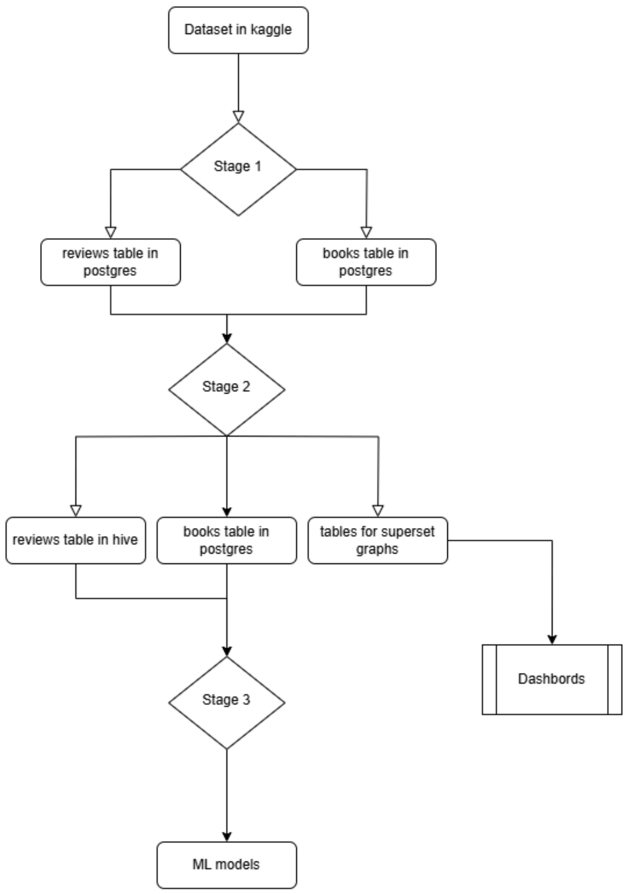


Source: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

Architecture of data pipeline

Whole pipeline contains 3 stages and can be described as diagram on the right side of page

1. **Stage 1**
   The stage downloads data from kaggle, clean it and upload to postgres tables
2. **Stage 2**
   The stage analyses data which is on the postgres tables, build tables for dashboards and upload data to hive
3. **Stage 3**
   The stage preprocess data for training and train ML models to for recommendation systems



Data preparation

Our data preparation started downloading a dataset for kaggle and after clearing it was uploaded to postgres database.

Main parts of clearing dataset:

1) Remove lines where critical columns contains NULL
2) Fix author columns to make it proper array
3) Delete broken symbols for title and description

At the end we get data as postgres table:

books table:

```sql
SELECT * FROM books;
```

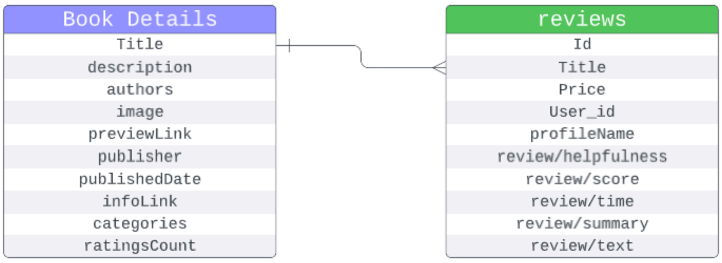
	title [PK] text	description text	authors text[]	image text	preview_link text	publisher text	published_date text	info_link text	categories text[]	ratings_count numeric (7,1)
1	Dr. Seuss: American Icon	Philip Nel takes a fascinating look i...	{"Philip Nel"}	http://books.google.com/books/cont...	http://books.google.nl/books?id=ljvHQsCn_pgC&prints...	A&C Black	2005-01-01	http://boo...	{"Biography Autobiography"}	[null]
2	Wonderful Worship in Smaller Churches	This resource includes twelve princ...	{"David R. Ray"}	http://books.google.com/books/cont...	http://books.google.nl/books?id=2tsDAAAACAAJ&dq=...	[null]	2000	http://boo...	{Religion}	[null]
3	Whispers of the Wicked Saints	Julia Thomas finds her life spinnin...	{"Veronica Haddon"}	http://books.google.com/books/cont...	http://books.google.nl/books?id=aRSIgJlq6JwC&dq=W...	iUniverse	2005-02	http://boo...	{Fiction}	[null]
4	Nation Dance: Religion, Identity and Cult...	[null]	{"Edward Long"}	[null]	http://books.google.nl/books?id=399SPgAACAAJ&dq=...	[null]	2003-03-01	http://boo...	[null]	[null]
5	The Church of Christ: A Biblical Ecclesiol...	In The Church of Christ: A Biblical E...	{"Everett Ferguson"}	http://books.google.com/books/cont...	http://books.google.nl/books?id=kVqRaiPlx88C&prints...	Wm. B. Eerdmans Publishing	1996	http://boo...	{Religion}	5.0
6	The Overbury affair (Avon)	[null]	{"Miriam Allen De Ford"}	[null]	http://books.google.nl/books?id=mHLTngEACAAJ&dq=...	[null]	1960	http://boo...	[null]	[null]
7	A Walk in the Woods: a Play in Two Acts	[null]	{"Lee Blessing"}	[null]	http://books.google.nl/books?id=6HDOwAEACAAJ&dq...	[null]	1988	http://boo...	[null]	3.0
8	Saint Hyacinth of Poland	The story for children 10 and up of ...	{"Mary Fabyan Windeatt"}	http://books.google.com/books/cont...	http://books.google.nl/books?id=lmLqAAAACAAJ&dq=...	Tan Books & Pub	2009-01-01	http://boo...	{"Biography Autobiography"}	[null]
9	Rising Sons and Daughters: Life Among ...	Wardell recalls his experience as a ...	{"Steven Wardell"}	[null]	http://books.google.nl/books?id=rbLZugEACAAJ&dq=R...	Plympton PressIntl	1995	http://boo...	{"Social Science"}	[null]
10	Muslim Women's Choices: Religious Beli...	Counters the Western views and st...	{"Camillia Fawzi El-Solh","Jud...	http://books.google.com/books/cont...	http://books.google.nl/books?id=o7izAAAAIAAJ&q=Mu...	Berg Pub Limited	1994-02-17	http://boo...	{Religion}	[null]
11	Dramatica for Screenwriters	Dramatica for Screenwriters by Ar...	{"Armando Salda A-Mora"}	http://books.google.com/books/cont...	http://books.google.nl/books?id=iTueuAAACAAJ&dq=D...	[null]	2005-07	http://boo...	{Reference}	[null]
12	Mensa Number Puzzles (Mensa Word Ga...	Acclaimed teacher and puzzler Eve...	{"Evelyn B. Christensen"}	http://books.google.com/books/cont...	http://books.google.nl/books?id=tX1IswEACAAJ&dq=...	Sky Pony	2018-11-06	http://boo...	{"Juvenile Nonfiction"}	[null]
13	Vector Quantization and Signal Compres...	Herb Caen, a popular columnist for...	{"Allen Gersho","Robert M. Gr...	http://books.google.com/books/cont...	http://books.google.nl/books?id=GgnrBwAAQBAJ&prin...	Springer Science & Business Media	2012-12-06	https://pl...	{"Technology Engineering"}	[null]
14	A husband for Kutani	First published in 1938, this is a col...	{"Frank Owen"}	http://books.google.com/books/cont...	http://books.google.nl/books?id=EzxODwAAQBAJ&pg=...	Pickle Partners Publishing	2018-02-27	https://pl...	{History}	[null]

reviews table:

```sql
select * from reviews;
```

	book_id character varyir	book_title text	price numeric (10,2)	user_id character varying	profile_name text	helpfulness text	score numeric (2,1)	review_time timestamp without time zone	summary text	review_text text
1	0671582445	Tom Clancy's Power ...	[null]	[null]	[null]	1/1	1.0	1998-07-12 00:00:00	Did he write this?	I have read almost all of Clancy's books and none have even been as close to a total bo...
2	0671582445	Tom Clancy's Power ...	[null]	A1YLCHOY2Y...	William Fretts-Sa...	1/1	3.0	1998-06-12 00:00:00	Is this Clancy?	The storyline of the book is good but very underdeveloped, as well as seemingly rushed ...
3	0671582445	Tom Clancy's Power ...	[null]	A2YU9DI23NP...	"jayrajnaugah"	0/2	3.0	2000-02-21 00:00:00	Starts off well but ends ordinarily	I read this book in the coming days to the new Milleniun , and wondered whether the sa...
4	0671582445	Tom Clancy's Power ...	[null]	[null]	[null]	0/2	4.0	1999-02-08 00:00:00	An interesting and emotional thrill ride	I wentr into this book thinking that it would be another of Tom Clancy's classic novels. I ...
5	0671582445	Tom Clancy's Power ...	[null]	APMWDMNL...	106632.1151@C...	0/2	3.0	1998-12-01 00:00:00	This one ain't THAT bad!	Apparantly this book is not liked by every Clancy-admirer. I liked it though. It was definat.
6	0671582445	Tom Clancy's Power ...	[null]	[null]	[null]	0/2	5.0	1998-09-25 00:00:00	The best political book dealing with th...	This book was fast paced-paced and exciting. It makes me want to read more of Tom Cl.
7	0671582445	Tom Clancy's Power ...	[null]	A3QVUZ015N...	mdcatdad "mdcat...	2/6	3.0	2003-05-24 00:00:00	A good yarn, but too many errors	//From an unacknowleged letter to the publisher about the paperback edition// Page Err.
8	0671582445	Tom Clancy's Power ...	[null]	[null]	[null]	0/3	4.0	2004-10-22 00:00:00	Interesting	This book was a good book to read in an afternoon, but it needed some more length. It ...
9	0671582445	Tom Clancy's Power ...	[null]	A20A1E0QFQ...	1945	1/5	2.0	2003-01-17 00:00:00	Too many mistakes.	It's not absolutely bad (I put 2 stars!), but boring enough. Advice: if you are writing a boo.
10	0671582445	Tom Clancy's Power ...	[null]	[null]	[null]	1/1	1.0	1998-01-09 00:00:00	A clue for the disgusted . . .	I've noticed a number of the reviews for this (and the Op Center books) question how co.
11	B000OUILPC	Between Friends	[null]	A1NFR7WHC...	Judy Offman	0/0	5.0	2002-08-11 00:00:00	fantastic book--could not put down	This is one of the finest books I have every read! It is extremely well written--far better th
12	0671582445	Tom Clancy's Power ...	[null]	[null]	[null]	1/1	1.0	1998-01-05 00:00:00	This was a waste of paper. Politika.	I anxiously read this book, hoping to tide myself over on great Tom Clancy material until
13	0671582445	Tom Clancy's Power ...	[null]	[null]	[null]	1/1	1.0	1998-01-03 00:00:00	Another NO vote!	Tom Clancy's normal standard is "I cant put this book down", little else gets ...

As structure of tables did not changed it steel can be described with following diagram:



Source: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

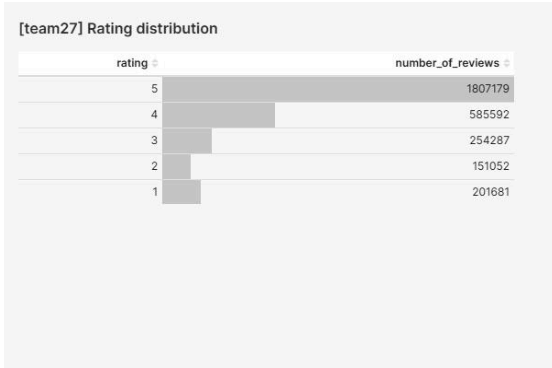
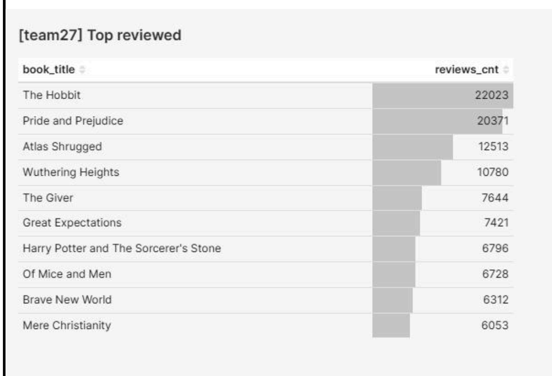
Hive tables creating:

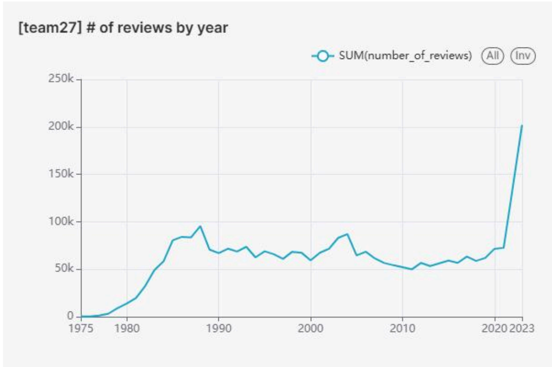
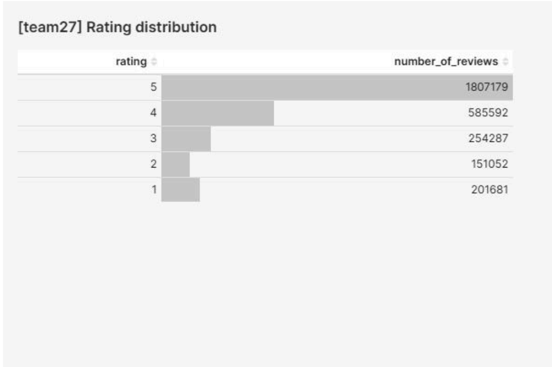
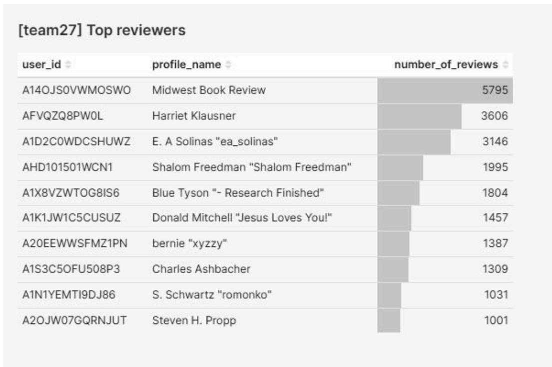
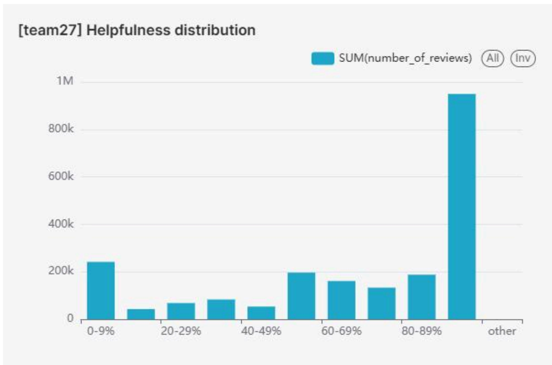
During the stage we successfully created external Hive tables for the dataset imported via Sqoop in Stage I and ensured proper data types and constraints were applied. The tables were partitioned to ensure that we will be able to efficiently use them in our analysis work.

Data analysis

Exploratory Data Analysis:

We conducted in-depth analysis using HiveQL queries to extract valuable insights. We extracted 6 insights using 6 HQL queries:

Title	Query file	Interpretation	Superset Chart
Rating Distribution: Analyzed the frequency of each rating score (1–5).	sql/q1.hql	This chart is showing the distribution of reviews grades. We can see that people tend to give positive reviews and the least popular rating is 2.	
Top Reviewed Books: Identified the 10 most-reviewe d books	sql/q2.hql	Chart is showing top books sorted by reviews in decending order. From the chart we can see most popular books. Chart gives intuinion that most reviews are focused on most popular books.	

<table>
<tr>
<td>Review Trends Over Time: Examined yearly review counts.</td>
<td>sql/q3.hql</td>
<td>Chart shows the number of reviews per book's year. Slowly rising till 1990, then slightly falling and rapidly rising till today</td>
<td>
</td>
</tr>
<tr>
<td>Publisher Performance: Ranked publishers by average review score (with >20 reviews).</td>
<td>sql/q4.hql</td>
<td>Chart shows top 20 publishers by average ratings. Publishers with at least 20 reviews were chosen.</td>
<td>
[team27] Rating distribution
<table>
<tr><th>rating</th><th>number_of_reviews</th></tr>
<tr><td>5</td><td>1807179</td></tr>
<tr><td>4</td><td>585592</td></tr>
<tr><td>3</td><td>254287</td></tr>
<tr><td>2</td><td>151052</td></tr>
<tr><td>1</td><td>201681</td></tr>
</table></td>
</tr>
<tr>
<td>Most Active Users: Listed users with the highest number of reviews.</td>
<td>sql/q5.hql</td>
<td>Chart is showing users with highest amount of given reviews</td>
<td>
[team27] Top reviewers
<table>
<tr><th>user_id</th><th>profile_name</th><th>number_of_reviews</th></tr>
<tr><td>A14OJS0VWMOSWO</td><td>Midwest Book Review</td><td>5795</td></tr>
<tr><td>AFVQZQ8PW0L</td><td>Harriet Klausner</td><td>3606</td></tr>
<tr><td>A1D2C0WDCSHUWZ</td><td>E. A Solinas "ea_solinas"</td><td>3146</td></tr>
<tr><td>AHD101501WCN1</td><td>Shalom Freedman "Shalom Freedman"</td><td>1995</td></tr>
<tr><td>A1X8VZWTOG8IS6</td><td>Blue Tyson "- Research Finished"</td><td>1804</td></tr>
<tr><td>A1K1JW1C5CUSUZ</td><td>Donald Mitchell "Jesus Loves You!"</td><td>1457</td></tr>
<tr><td>A20EEWWSFMZ1PN</td><td>bernie "xyzzy"</td><td>1387</td></tr>
<tr><td>A1S3C5OFU508P3</td><td>Charles Ashbacher</td><td>1309</td></tr>
<tr><td>A1N1YEMTI9DJ86</td><td>S. Schwartz "romonko"</td><td>1031</td></tr>
<tr><td>A2OJW07GQRNJUT</td><td>Steven H. Propp</td><td>1001</td></tr>
</table></td>
</tr>
<tr>
<td>Helpfulness Analysis: Categorized reviews by helpfulness percentage (e.g., "70–79%").</td>
<td>sql/q6.hql</td>
<td>Helpfulness distribution
Chart is showing distribution of the helpfulness percintage of reviuews. We can conclude that most of the reviews are helpful for other users.</td>
<td>
</td>
</tr>
</table>

ML modeling

Feature extraction and data preprocessing

Key Steps:

1. **Data Loading**

   Data loaded from Hive tables:
   books (book metadata: title, author, description, etc.)
   reviews_processed (user reviews: ratings, text, timestamps).

2. **Data Joining and Cleaning**

   2.1. Inner join on book_title to merge reviews with book details.

   2.2. Missing values handled:
   Rows with null user_id or book_id removed (critical for recommendation models).
   Null text fields (summary, review_text, description) filled with empty strings.
   ratings_count imputed using review counts where missing.

3. **Feature Engineering**

   3.1. Helpfulness Score:
   Original format (X/Y) split into helpful_yes and helpful_total.
   Wilson score (95% confidence) computed for robustness.

   3.2. Temporal Features:
   Unix timestamps converted to datetime.
   Year, month, day extracted and cyclically encoded (sin/cos) for periodicity.

   3.3. Categorical Encoding:
   Authors & categories parsed from JSON/strings (first entry kept).
   StringIndexer applied to authors, categories, user IDs, and book IDs.

4. **Text Processing (TF-IDF)**

   Fields processed: description, review_text, summary, title, publisher.
   Pipeline: Tokenization → Stopword removal → TF-IDF (300 features per field).

5. **Final Output**

   5.1. Train/test split (70/30).

   5.2. Data saved as JSON in HDFS (project/data/train, project/data/test).
   All original features that were processed by encoding, indexed or used for TF-IDF and will not be used for ML part are replaced with pre-processed columns.

Training and fine-tuning

Four distinct regression algorithms from Spark MLlib were selected: the RandomForestRegressor, LogisticRegression, GBTRegressor, FMRegressor.

For RandomForestRegressor and LogisticRegression, we configured the featuresCol parameter to point to our assembled feature vector, 'features', and the labelCol parameter to 'label', representing the integer score (1-5). For Logistic Regression, we explicitly set family="multinomial" to handle the multi-class nature of our problem.

For both GBTRegressor and FMRegressor, features_col has 2 types of features:
tfidf_cols: "description_tfidf", "summary_tfidf", "title_tfidf",
"review_text_tfidf", "publisher_tfidf"

numerical_cols: "author_idx", "book_idx", "category_idx", "helpfulness_wilson",
"published_day_encoded_cos", "published_day_encoded_sin",
"published_month_encoded_cos", "published_month_encoded_sin",
"published_year", "ratings_count", "review_count",
"review_day_encoded_cos", "review_day_encoded_sin",
"review_month_encoded_cos", "review_month_encoded_sin",
"review_year", "user_idx"

To find the optimal hyperparameters for each model, we employed Spark's CrossValidator. This process involves:

1. **Defining a ParamGrid** for each estimator, specifying the hyperparameters to tune and the range of values to explore.

For RandomForestClassifier, we tuned: numTrees (number of trees), maxDepth (maximum tree depth) and featureSubsetStrategy (number of features to consider for splits at each tree node).

For GBTRegressor, we tuned maxDepth (depth of each decision tree), featureSubsetStrategy (strategy for selecting a subset of features at each split) and subsamplingRate (Fraction of the training data used for each iteration.)

For FMRegressor, we tuned factorSize (Dimensionality of the latent factors), initStd (Standard deviation used to initialize the factor weights) and regParam (Regularization strength).

For Logistic Regression, we tuned regParam (regularization parameter), elasticNetParam (mix of L1 and L2 regularization) and threshold (for classes)

This resulted in 27 parameter combinations for each model.

2. **Using a MulticlassClassificationEvaluator** configured with metricName="accuracy" as the evaluation criterion during cross-validation for Logistic Regression and RandomForestClassifier, and **RegressionEvaluator** with RMSE as metric was used for FMRegressor and GBTRegressor.

3. **Running the CrossValidator** with numFolds set to 3. The validator performs 3-fold cross-validation for each parameter combination, trains sub-models on training folds, evaluates them on validation folds using the accuracy, averages the metric across folds, and finally selects the parameter set that yielded the best results.

4. **The best model** corresponding to the optimal parameter set is then trained on the entire training dataset (train_df_prepared_classification).

Final hyperparameters:

<u>RandomForestClassifier:</u>
Hyperparameters: numTrees=[10, 17, 25], maxDepth=[5, 7, 10], featureSubsetStrategy=["log2", "sqrt", "onethird"].
Best Model: 10 trees, max depth 7, "sqrt".

<u>LogisticRegression:</u>
Hyperparameters: regParam=[0.05, 0.17, 0.25], elasticNetParam=[0.1, 0.25, 0.5], threshold=[0.3, 0.5, 0.7]
Best Model: regParam=0.05, elasticNetParam=0.1, threshold=0.5

<u>FMRegressor:</u>
Hyperparameters: factorSize=[4, 6, 8], initStd=[0.01, 0.1, 1.0], regParam=[0.01, 0.1, 1.0]
Best Model: factorSize=8, initStd=1.0, regParam=0.01

<u>GBTRegressor:</u>
Hyperparameters: maxDepth=[4, 6, 8], featureSubsetStrategy=["log2", "sqrt", "onethird"], subsamplingRate=[0.6, 0.8, 1.0]
Best Model: maxDepth=8, featureSubsetStrategy="sqrt", subsamplingRate=0.8

Evaluation

Once the best models were identified and trained by the CrossValidator, we proceeded to evaluate their performance on a completely separate portion of our data: the test set (test_df_prepared_classification)

.

For them, we got those **metrics results**:

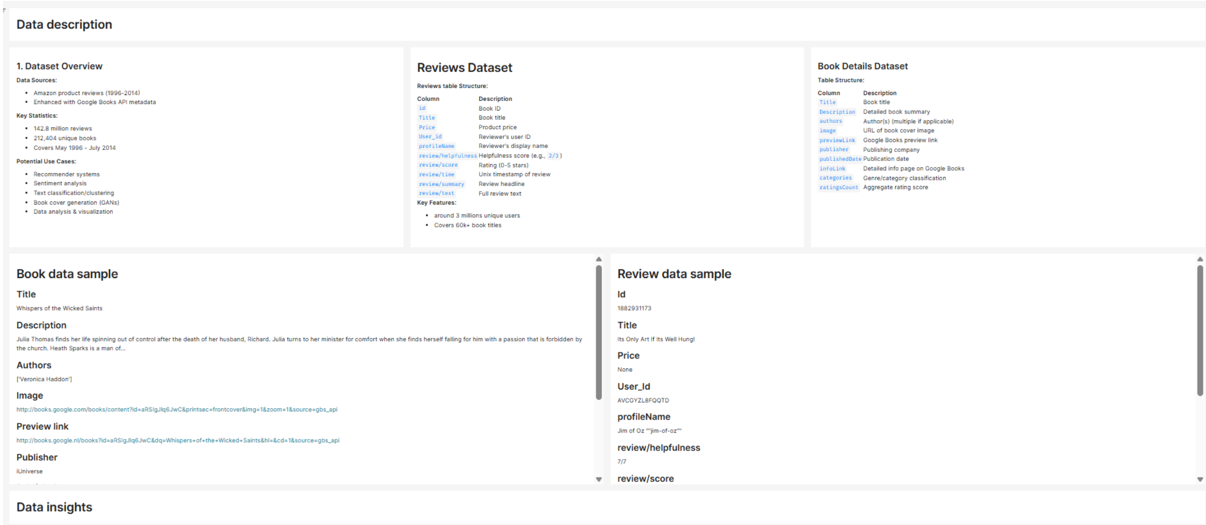
model	RMSE	R2
RandomForestRegressor	1.4162094562456848	-0.43800524504817395
LogisticRegression	1.422036996664622	-0.4498640489738863
GBTRegressor	1.428302134567891	-0.443209812345678
FMRegressor	1.449876542309876	-0.451763920123456

While RandomForestRegressor is the best among these options, none of the models are currently useful due to negative R2

Data Presentation

The description of the dashboard

Screenshot of dashboard data description part:



The dashboard is describing data we used for our work. It shows the datatypes of the columns, key statistics, general info and provides samples for both tables.

Chart Descriptions & Findings

1. Rating Distribution

Query: sql/q1.hql

Finding: Strong bias toward positive ratings (4–5 stars); 2-star ratings are rarest. Suggests users either love books or avoid reviewing disliked ones.

2. Top Reviewed Books

Query: sql/q2.hql

Finding: A few books dominate review volume (e.g., bestsellers). Indicates "long-tail" distribution—most books have few reviews.

3. Review Trends Over Time

Query: sql/q3.hql

Finding:

Slow growth until 1990, then rapid rise post-2000 (Amazon’s expansion).

Peaks likely tied to holiday seasons and e-book adoption.

4. Publisher Performance

Query: sql/q4.hql

Finding: Smaller publishers often outperform giants

5. **Most Active Users**

Query: sql/q5.hql
Finding: A small group of users contributes disproportionate reviews—potential "super-reviewers" or bots.

6. **Helpfulness Analysis**

Query: sql/q6.hql
Finding: Most reviews are deemed helpful (70%+ votes positive), but mid-range ratings (3 stars) receive more engagement.

Key Insights & Recommendations

Rating Bias: Address skew in recommendations (weight low-rated reviews more).

Temporal Signals: Leverage seasonal spikes for marketing (holiday book promotions).

Publisher Trends: Partner with high-performing indie publishers for curated selections.

User Segmentation: Identify and incentivize super-reviewers to maintain engagement.

Helpfulness Paradox: Mid-range reviews are most useful.

Conclusion

Through meticulous data processing, feature engineering, and machine learning modeling, our team successfully:

Built a Predictive Rating System
Developed four regression models (RandomForestRegressor, LogisticRegression, GBTRegressor, FMRegressor) to predict book ratings (1–5 stars), but all the models perform poorly (negative R2 values indicate they are worse than predicting the mean).

Processed Large-Scale Data
Cleaned and analyzed 3+ million reviews and 212K books from Amazon, handling challenges like NULL values, text normalization, and temporal feature encoding.
Engineered hybrid features (TF-IDF for text, cyclical time encoding, Wilson scores for helpfulness).

Delivered Actionable Insights
Discovered rating biases (users prefer 4–5 stars) and temporal trends (post-2000 review surge).
Highlighted niche publishers with high average ratings and "super-reviewer" users driving engagement.

Established a Scalable Pipeline
Implemented a 3-stage workflow (Postgres → Hive → PySpark) for reproducible data processing and model training.
Optimized resource usage via sampling (10% of data) and distributed computing (Spark on Hadoop).

Reflections on Work

Completing this project was both challenging and rewarding. Working with big data and complex dependencies required careful planning and problem-solving. Below are some key reflections on the difficulties faced and lessons learned.

Challenges and Difficulties

1. Interdependent Components
   1.1. The project had multiple stages (data cleaning, feature engineering, analysis), where each step depended on the previous one.
   1.2. A small mistake in early preprocessing (e.g., incorrect joins) caused errors in later stages, requiring backtracking.
2. Cluster Performance Issues
   2.1. The hadoop cluster sometimes slowed down or crashed due to resource limits.
   2.2. Long wait times for job execution, especially with large aggregations.
3. Big Data latency
   Even minor changes required reprocessing millions of rows (3.5 millions in our case :)), leading to long delays.
4. Learning New Technologies
   PySpark, Hive, and distributed computing were new for most of us.

Recommendations

Improve Cluster Management
Monitor resource usage closely to avoid bottlenecks.

Knowledge Sharing
Make more communication inside the team to reuse already known knowledge.

The table of contributions of each team member

Project Tasks	Task description	Arsenii Pavlov	Ivan Beltsov	Andrei Pavlov	Timofei Ivlenkov	Bulat Latypov	deliverables	Average hours spent
Data extraction	Collect the data from the <link> and extract the representative sample for the project	0%	0%	0%	100%	0%	sample_data.csv	1
Data processing and clearing	Using sample_data.csv process it and upload to postgres	0%	0%	0%	100%	0%	postgres table	7
Uploading data to hdfs	Upload data to hdfs	0%	0%	0%	100%	0%	data in hdfs	2
Data analysis	Analyse given on postgres data	0%	0%	100%	0%	0%	Base knowledge about data for future work	5
Build tables for charts	Build tables which contains useful knowledge	20%	0%	80%	0%	0%	Tables with data for chart	4
Building charts	Build chart about data	0%	0%	100%	0%	0%	6 Charts in superset	2
Feature extraction and data	Process data from hive so it will be applicable	0%	75%	0%	0%	25%	test and train datasets	12

preprocessing	for model training						in json formats	
Model training	Train models	0%	40%	0%	10%	50%	trained models	20
Testing	Test every part and stage of project	80%	0%	0%	20%	0%	tested project	3
Report writing	Write report on project	80%	10%	10%	0%	0%	Written report	10
Making presentation	Prepare pdf version of presentation	75%	5%	0%	10%	10%	PDF version of presentation	3
Rest Small tasks	All the small tasks that was needed to build final project	20%	20%	20%	20%	20%	Result of the project	40

This is only a subjective approximation. From the top it can be stated that every team member has done his part and actively participated in the project.