FINAL PROJECT REPORT

Project Title: Goodreads Book Rating Analytics & Prediction System

1. Introduction

In today's digital publishing ecosystem, online platforms such as Goodreads play a crucial role in shaping reader preferences and influencing book sales. Readers rely heavily on ratings and reviews to decide which books to read, while publishers and authors use these signals to evaluate market performance. However, understanding why certain books receive higher ratings and whether ratings can be predicted using measurable attributes remains a challenge.

This project presents an end-to-end data science solution that transforms unstructured web data into meaningful insights and predictive intelligence. By combining web scraping, data preprocessing, exploratory data analysis, machine learning, and interactive dashboards, the system enables stakeholders to understand rating behaviour and make informed decisions.

---

2. Problem Statement

The detailed business problem, objectives, and phase-wise problem statements for this project are defined in the approved document:

"Goodreads Book Rating Analytics & Prediction System – Problem Statement"
This report focuses on the implementation, analysis, results, and insights derived from that problem statement, without repeating it.

---

3. Project Objectives

The objectives of this project are aligned with the approved problem statement and are summarised as follows:

- To collect real-world book metadata from Goodreads using web scraping
- To clean and preprocess raw data into an analysis-ready format
- To identify patterns and drivers influencing book ratings
- To build and evaluate machine learning models for rating prediction
- To select the most reliable and interpretable model
- To communicate insights using an interactive Power BI dashboard
- To enhance explainability using AI-driven visuals

---

4. Data Description

1. Data Source
Platform: Goodreads
Category: Best Books Ever
Collection Method: Web scraping using Python

Video Explaination

2 Dataset Size
Approximately 1000 books

3 Raw Attributes Collected
Book name
Author
Genres list
Page format
First published year
Rating
Ratings count
Votes
Score
Book URL
The raw dataset contained unstructured text, missing values, and mixed data formats, requiring extensive preprocessing.

---

5. Methodology

The project followed a structured end-to-end data science pipeline, divided into clearly defined phases.
Phase 1: Data Collection (Web Scraping)
Since Goodreads does not provide a free public API for book metadata, data was collected using web scraping techniques.
Tools Used

- Python
- Requests
- BeautifulSoup
- Pandas

Process Explanation

- Multiple list pages were scraped to collect book-level information
- Each book's inner page was accessed to extract detailed attributes
- Data was stored in tabular format for further processing

This phase ensured access to real-world, large-scale data, forming the foundation of the project.

Phase 2: Data Cleaning & Preprocessing
The raw scraped data were unsuitable for direct analysis due to inconsistencies and noise.
1. Cleaning Steps Performed

- Removed text artefacts from numeric fields (ratings, votes, counts)
- Handled missing values using appropriate imputation
- Standardised genre information
- Separated the number of pages and book type
- Extracted numeric publication year
- Converted all columns to correct data types

Video Explaination

2. Feature Engineering

To improve analysis and model performance, new features were created:

- Book Age:

Current Year - First Published Year (More meaningful than raw year)

- Pages Category:

Short, Medium, Long, Very Long (Improves interpretability)

- Log Transformation:

Applied to ratings count and votes to reduce skewness

This phase produced a **clean, structured, ML-ready dataset**.

Phase 3: Exploratory Data Analysis (EDA)

EDA was conducted to understand patterns, trends, and relationships in the data.

Key Insights from EDA

- Certain genres (Fantasy, Comics, Religion) consistently receive higher ratings
- Hardcover and Kindle formats outperform other formats
- Reader engagement (votes and ratings count) positively correlates with ratings
- Medium to very long books show better rating stability
- Ratings have stabilised significantly after the mid-20th century

Correlation analysis confirmed that book ratings are influenced by identifiable factors, validating the feasibility of predictive modelling.

Phase 4: Model Building & Evaluation

1. Feature Exclusion

The **score** column was removed due to data leakage, as it is derived from votes and ratings and would artificially inflate model performance.

And also remove some irrelevant columns, such as book_name, ratings_count, and votes.

2. Target Variable

- eating

3. Models Implemented

★ Linear Regression

★ Random Forest Regressor

4. Evaluation Metrics

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R2 Score

Video Explaination

5. Model Performance Comparison

Model	MAE	RMSE	R^2 Score	Result
Linear Regression	**0.062**	0.097	0.839	Best Model
Random Forest	0.067	0.104	0.818	Good, but overfitted

6. Model Selection Justification

Linear Regression was selected as the final model because:

- It achieved the highest R^2 score
- It showed the lowest prediction error
- It generalised better than Random Forest
- It is simpler, interpretable, and easier to deploy

This indicates that the relationship between book attributes and ratings is largely linear.

Phase 5: Interactive Dashboard (Power BI)

An interactive Power BI dashboard was developed to translate analytical and predictive results into business insights.

Dashboard Components

- KPI cards (Average Rating, Total Books, Engagement)
- Genre-wise rating comparison
- Book format performance analysis
- Rating trends over publication years
- Engagement vs rating scatter plot
- Interactive slicers for dynamic filtering
- Detailed book-level table

The dashboard enables non-technical users to explore data and insights intuitively.

To enhance transparency and trust, an AI-focused dashboard page was added using Power BI AI visuals.

AI Visuals Used

- Key Influencers
- Decomposition Tree

AI Insights

- Very long books have the strongest positive influence on ratings
- Fantasy and Comics genres consistently drive higher ratings
- Optimal combinations of genre, format, and page length lead to peak ratings

These insights align closely with EDA and ML results, thereby reinforcing the model's reliability.

6. Results

- Book ratings are predictable using metadata alone
- Content depth and genre are the strongest drivers
- Engagement metrics support but do not directly cause ratings
- Linear Regression provides the best balance of accuracy and interpretability
- AI visuals successfully explain model behaviour

Video Explaination

7. Conclusion

This project successfully demonstrates an end-to-end data science workflow for analysing and predicting Goodreads book ratings. By integrating real-world data acquisition, machine learning, and interactive visualisation, the system provides actionable insights for publishers and authors. The final Linear Regression model achieved strong predictive performance and was supported by AI-driven explainability, making the solution both effective and transparent.

8. Recommendations

1. Focus on High-Performing Genres
Genres such as Fantasy and Comics consistently achieved higher average ratings across exploratory and AI analyses.
Recommendation: Publishers should prioritise content acquisition, marketing, and promotion strategies for these genres to maximise reader engagement and overall ratings.

2. Invest in Content Depth
Books categorised as Very Long demonstrated higher and more stable ratings, indicating that readers value detailed storytelling and in-depth narratives.
Recommendation: Authors and publishers should consider investing in richer content development, particularly for genres where long-form storytelling is well-received.

3. Prioritize Premium Book Formats
Hardcover and digital formats, such as Kindle, showed better rating performance compared to other formats.
Recommendation: High-potential titles should be released in premium and digital-friendly formats to improve reader perception and satisfaction.

4. Leverage Predictive Analytics Before Launch
The Linear Regression model achieved an R^2 score of 0.83, proving that book ratings can be reliably predicted using metadata alone.
Recommendation: Publishers can use this predictive model to estimate a book's expected rating before launch and optimise marketing, pricing, and distribution decisions accordingly.

5. Integrate Dashboard Insights into Business Workflow
The Power BI dashboard enables interactive exploration of ratings, genres, and trends.
Recommendation: Publishing teams should incorporate dashboard insights into regular performance reviews and planning meetings for continuous improvement.

Video Explaination