
Project Overview

You are working as a Data Science Consultant for a digital publishing analytics firm aiming to support publishers, bookstores, and online reading platforms in making data-driven decisions. With millions of books on platforms like Goodreads, it is difficult to predict how well a book will perform, identify what drives high ratings, and understand patterns across genres, authors, and formats.

The leadership wants to know:

- What factors influence book ratings?
- Can we predict a book's rating before release?
- Which genres or authors consistently perform well?
- How do pages, book format & reader engagement impact ratings?

This project builds an end-to-end analytics and prediction system -from web scraping to dashboard & machine learning -to generate actionable business insights.

BUSINESS PROBLEM

Publishing houses and e-commerce bookstores often struggle with estimating a book's reception before launch, leading to uncertain marketing investments, suboptimal inventory planning, and missed revenue opportunities.

- Core Business Problem
  Stakeholders lack a predictive system to estimate a book's expected rating and identify key performance drivers using publicly available metadata.

Solving this problem helps in:

- Forecasting reader response before a book launch
- Better marketing & promotional planning
- Stock and distribution decisions for bookstores
- Improving recommendation systems for reading platforms

PROJECT OBJECTIVE

To design and develop an end-to-end data science solution using Goodreads metadata to explore trends, understand rating influencers, and build a predictive model that forecasts a book's rating using engineered features from scraped data.

PROBLEM STATEMENTS BY PHASE

- Phase 1 - Data Acquisition

Goodreads does not provide direct structured access to book metadata. To obtain analysis-ready data, a custom scraping workflow is required to collect book information from multiple pages.

- Phase 2 - Data Cleaning & Preprocessing

The scraped data contains messy formats (commas, text strings, symbols, multi-value genres). This prevents direct analysis and machine learning. Cleaning and converting data into numeric & categorical formats is required.

- Phase 3 - Exploratory Data Analysis

Raw values cannot reveal what drives book ratings. Visual exploration is needed to identify which factors have the greatest influence on success.

- Phase 4 - Machine Learning

There is no automated mechanism to estimate the expected book rating. A supervised regression model must be built to forecast rating and compare performance across algorithms.

- Phase 5 - Dashboard & Reporting

Stakeholders cannot interpret CSV/ML output. A visual decision system (Power BI) is required to summarise KPIs, model outputs, and performance drivers.

MACHINE LEARNING PLAN

We will evaluate two models:

Model	Reason to Include	Expected Outcome
Linear Regression	Baseline benchmark model	Understand linear relationships & baseline performance
Random Forest Regressor	Handles nonlinearity + categorical data, real-world applicability	Higher accuracy & business-ready deployment candidate

Model Comparison Metrics

- MAE
- RMSE
- R^2 Score

Output Needed in Report

- Which model performs better
- Why does it perform better
- Practical use case for business