[ReelGood]

[G4]

Data Science Capstone Project
Launch Report

Date:

[01/20/2025]

Team Members:

Name: Jaz Zhou

Name: Precious Orekha

Name: Alireza Hatami

Name: Caitlin Dunne

[The purpose of this report is initiating a new project. It provides an overview description of the project. It includes three major sections: The System/Product, The Team, and The Project Plan.]

The System /Product

System/Product Name: ReelGood

Introduction:

"You could download an entire movie in 3 seconds on 5G. That's gonna be fast. I mean, we'll still spend 45 minutes trying to decide which movie to watch."

-- Trevor Noah

Modern consumers are inundated with a huge selection of options, choice fatigue is real. From a business perspective, personalized recommendations play a critical role in enhancing user satisfaction and loyalty. For example, Amazon once reported that over 20% of its sales were driven by recommendations, while Netflix famously offered a substantial prize to improve its recommendation system by a mere 10%.

This highlights the importance of recommendation systems (RS), which have become a major area of research. Broadly speaking, there are two strategies to building such systems:

1. **Content-Based Filtering:** Creates a profile for each user or product to characterize its nature, then associates users with matching products. This method is straightforward and interpretable but requires domain knowledge and often yields lower accuracy.

2. **Collaborative Filtering:** Utilizes explicit feedback (e.g., ratings, reviews) and/or implicit feedback (e.g., clicks, purchase history, mouse movements) to generate recommendations. This approach is generally more accurate but is less explainable and suffers from what is called the cold start problem.

This project aims to explore collaborative filtering algorithms to develop an effective recommendation system. By focusing specifically on movies and leveraging the "Movies Dataset", the project seeks to create tailored recommendations. Inspired by the academic boost resulting from the Netflix Prize competition, using a similar data topic may provide valuable guidance for navigating the project.

Dataset

The dataset used in this project was downloaded from Kaggle. It includes metadata for all 45,000 movies listed in the Full MovieLens Dataset** and files with user ratings for these movies. The movies included were released on or before July 2017. Data points in the dataset encompass various details such as cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, etc.

Algorithms selected

For a collaborative filtering recommendation system, we have selected kNN and SVD to use with our model. Based on the results of these algorithms we've also indicated an interest in exploring MLP if we

find the results of these algorithms unsatisfactory. We also are exploring a novel algorithm we found in an academic paper from 2023 they call UISVD++. A link to this paper is below.

Link:

* https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/discussion?sort=undefined

** https://grouplens.org/datasets/movielens/latest/

***https://arxiv.org/pdf/2206.05654

Highlighted Features:

1. **Top-k Recommendations:** Provide users with personalized recommendations for the top-k movies.
2. **Filtering Recommendations:** Allow users to refine their recommendations by applying filters, such as selecting a specific genre, to get tailored top-k suggestions.
3. **Potential User Interface (UI):** Develop a user-friendly interface to make accessing and interacting with recommendations simple and engaging.

Sponsor or Proxy User:

None

Issues:

1. **Lack of Experience:** Limited expertise in handling large-scale recommendation systems necessitates further research and practice.
2. **Outdated Data:** The dataset may not reflect current movie trends or user preferences, necessitating updates or enrichment.
3. **Large Data Size:** The dataset's size could pose computational challenges, requiring optimization strategies for efficient processing.
4. **Inconsistent Metadata:** Missing or inconsistent metadata could impact accuracy, highlighting the need for thorough cleaning and sourcing of relevant information.

The Team

Team Name:

G4-ReelGood

Team Members and their specialties:

Individual experience:

- **Precious Orekha:** I am a data science master's candidate, in my second to last quarter. I have 3 years professional experience as a data scientist
- **Jaz Zhou:** I am a data science master's student in my second-to-last quarter. I also have experience as a graduate student in a research-focused program in another field, which gave me a strong background in research..
- **Alireza Hatami:** I am a second-year data science student. I previously studied civil engineering and earned my master's degree in business administration. I have more than 10 years experience in business management, market research, construction management
- **Caitlin Dunne:** I am a data science master's student, with three quarters until graduation. I have a BS in Electrical Engineering and have seven years experience working for a semiconductor manufacturer doing R&D test engineering.

Role of each member*:

Person	Responsibilities
Alireza Hatami, Precious Orekha	- **Data Acquisition** - **Data Preprocessing** - Handle missing values, remove duplicates, format columns, and normalize features.
Jaz Zhou, Alireza Hatami	- **EDA** - Analyze distributions, trends, and sparsity, correlation - Visualize patterns to uncover actionable insights for feature engineering
Precious Orekha, Caitlin Dunne	- **Advanced Feature Engineering** - Create complex features such as embeddings and interaction terms - **Integration and Data Splitting** - Combine preprocessed data and features into final datasets - Implement splitting strategies(e.g., stratified or chronological splits)

*Note: Tasks are interconnected, and it is challenging to completely separate responsibilities. Team members will collaborate closely and adjust work distribution as needed to support one another.

Team Communication:

- The team meets every Sunday from 3:00 to 3:30 PM on Microsoft Teams, which is also used for messaging.
- Git and Google Colab are used for sharing and collaborating on code.

Team Issues:

The team lacks experience with recommender systems. To address this, the project will start with initial weeks of learning key techniques to prepare before data processing. Using a movie dataset also makes it easier to follow the vast amount of Netflix Prize research papers and understand recommendation strategies.

Weekly Plan

Week	Focus	Tasks	Output/Deliverables
3	**Data Acquisition and Initial Preprocessing**	- Acquire and consolidate raw data from multiple sources. - Clean the dataset: remove duplicates, handle missing values, format columns.	cleaned_data.csv
4	**Data Preprocessing and Basic Feature Engineering**	- Normalize or scale numerical features. - Compute aggregates (e.g., user interaction counts, item popularity).	preprocessed_data_with_basic_features.csv
5-6	**EDA (Exploratory Data Analysis)**	- Analyze distributions, sparsity, trends, and correlations. - Visualize patterns (e.g., histograms, heatmaps).	eda_report.pdf
7-8	**Advanced Feature Engineering**	- Design and implement advanced features (e.g., embeddings, interaction terms). - Collaborate on refining feature ideas using EDA insights.	advanced_features.csv
9	**Collaboration and Final Adjustments Data Integration and Split**	- Address any remaining inconsistencies or missing tasks. - Integrate data and ensure proper train/test split.	train.csv, test.csv
10	**Presentation**	- Presentation - Compile and submit all deliverables.	presentation.ppt G4_ReelGood.zip

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	Writing	Editing
1	**Project**	Jaz Zhou	Caitlin Dunne
2	**Team**	Precious Orekha	Jaz Zhou
3	**Plan**	Alireza Hatami	Caitlin Dunne

Grading

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.