Using unsupervised learning to create a book recommendation system

BookBots
UoB data analytics bootcamp

Introduction

Books help us to explore the world in many different ways.

Challenge of finding next book.

Power of new innovative system, data analysis and artificial intelligence.

Introduction to personalized book recommendation system

Data Preparation

- Data cleaning: removing any incomplete and duplicated records
- Separated multiple authors into individual entries
- Calculation of total no of authors and categories
- Identifying most frequent authors and categories

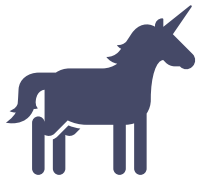
Enhancing Book Discoverability through Clustering

- Challenges in book discovery.
  - Information Overload
  - Diverse Preference
  - Limited Visibility

- Importance of Personalized Recommendation in Improving User Experience

- **Our Objective**: To recommend books based on user preferences and book characteristics.

Why clustering?



It uncovers hidden patterns in book features, aiding in discovering similar titles.

Enables dynamic recommendation without needing explicit user input.

Adaptable and scalable approach for a growing dataset.

Methodology Overview

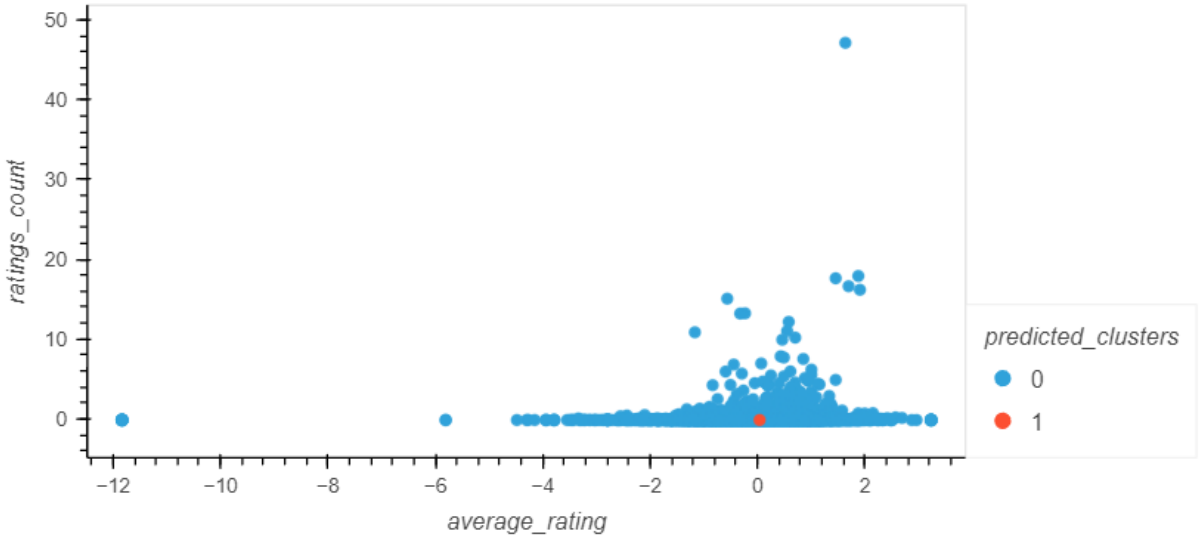
- **Data Preparation:** Dropping unnecessary columns, one-hot encoding, scaling.
- **Finding the Best K:** use of the Elbow and Silhouette methods for optimal cluster number.
- **Implementation and Evaluation:** Running K-means model, optimising with PCA, and assessing the model using The Calinski-Harabsz Index.

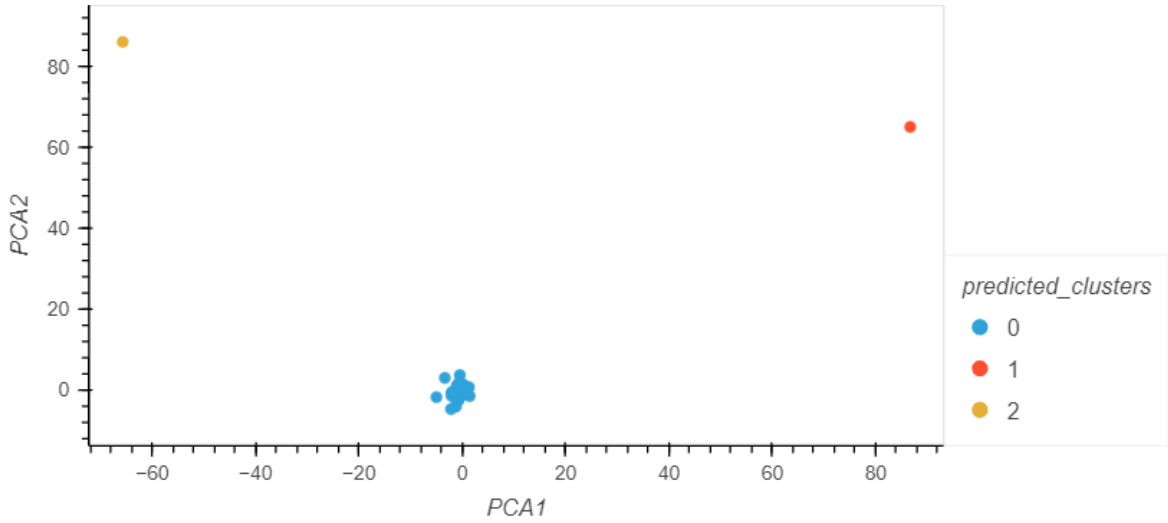
Model optimisation

Methods Used to Optimise the K-Means Algorithm Model:

- Dropped the “authors” column
- Dropped the “categories” column
- Filtered the “published year”
- Binned the “published year”
- Eliminated outliers

Model optimisation: dropping authors column





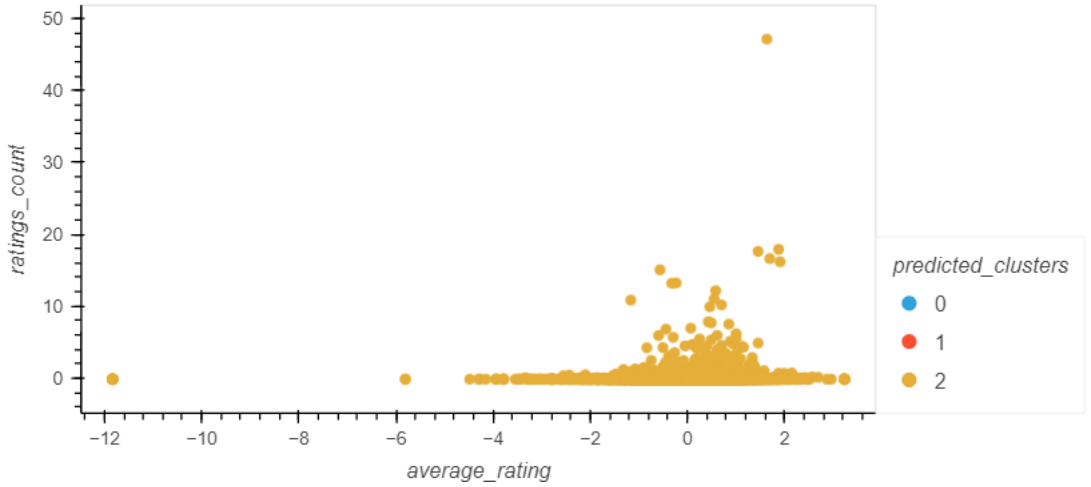
Elbow/Silhouette Method:

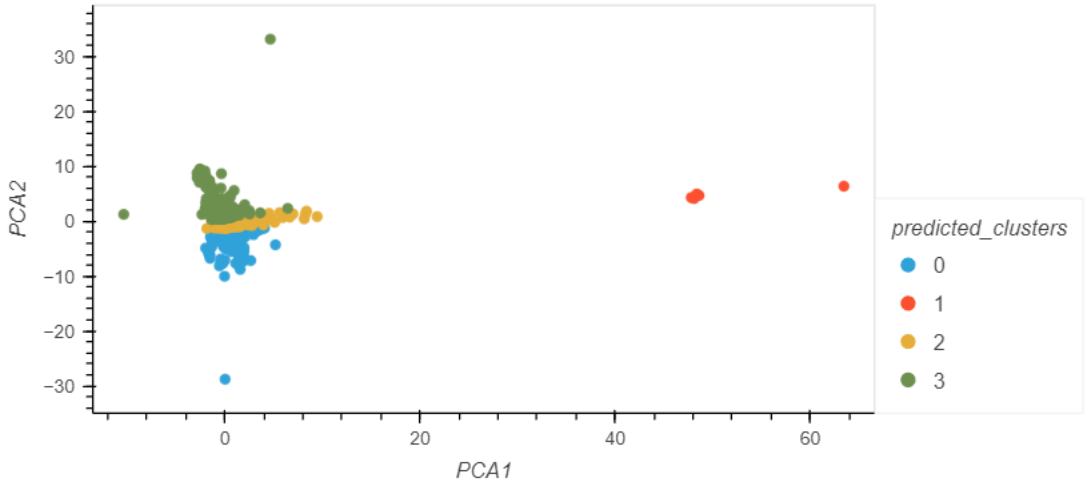
- The silhouette method suggests the optimum k-value is 2
- Running a K-means model with 2 clusters results in a Calinski-Harabasz Index score of 20.89.

PCA:

- The optimum k-value is 3.
- Running the K-means algorithm increases the Calinski-Harabasz Index to 177156.05
- Result seems highly determined by 2 outlier books.

Model optimisation: dropping categories column





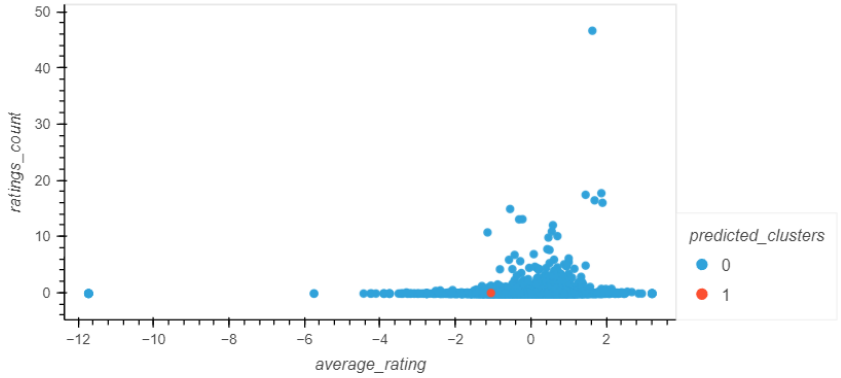
Elbow/Silhouette Method:

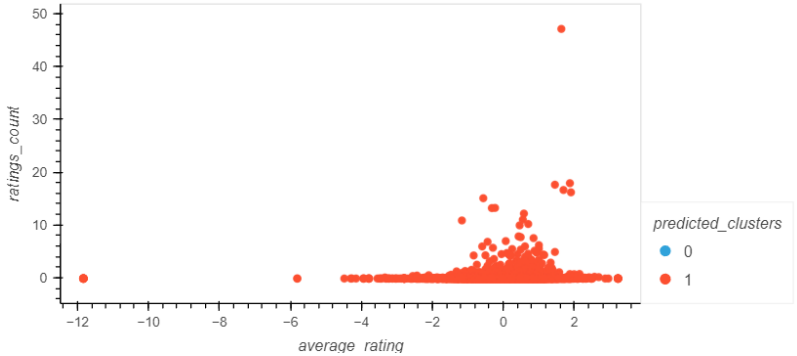
- The silhouette method suggests 3 as the best value for k.
- Running a K-means model with 3 clusters resuts in a Calinski-Harabasz Index score of 1.95.

PCA:

- The optimum value for k is 4
- Running the K-means algorithm increased the Calinski-Harabasz Index to 5600.92.

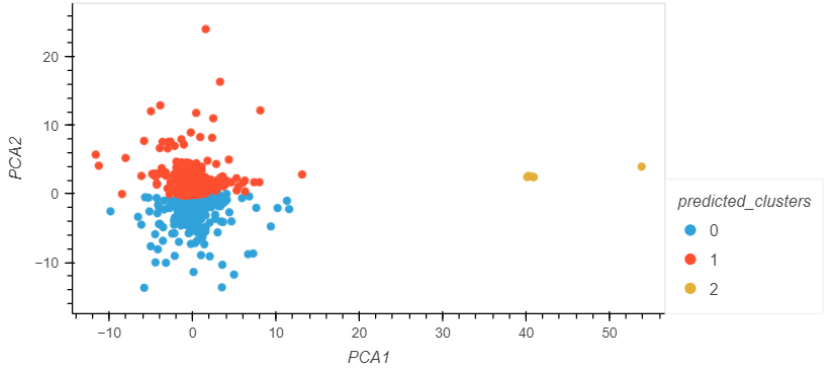
Model optimisation: filtering and binning published year

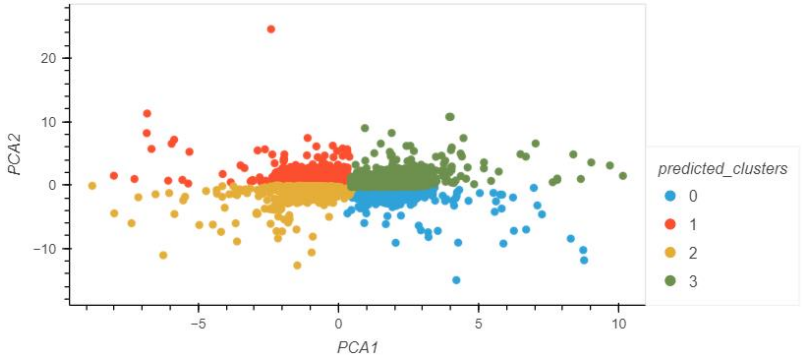




Elbow/Silhouette Method:

- Running a K-means model resulted in a Calinski-Harabasz Index score of 2.95 and 2.35 for the filtered and binned models respectively.

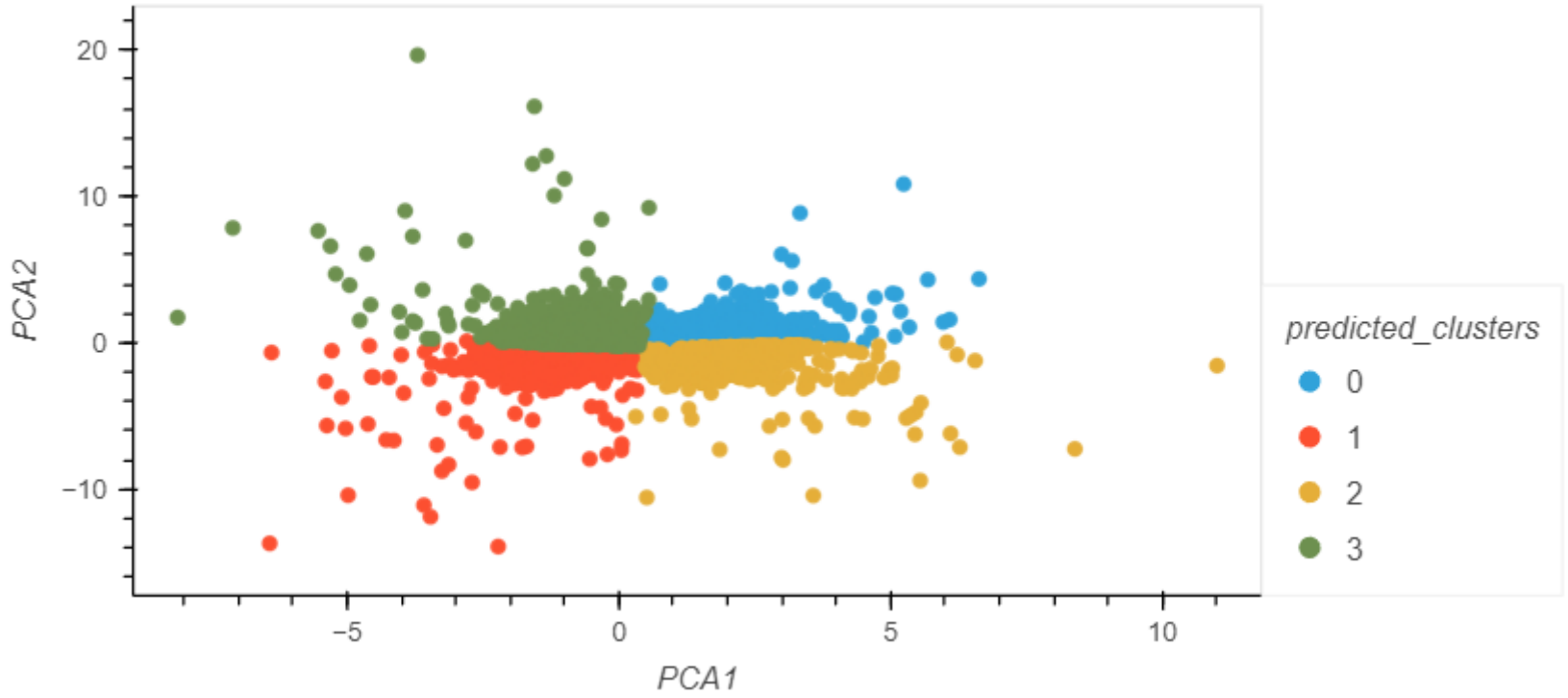




PCA:

- Running the K-means algorithm on filtered and binned data resulted in a Calinski-Harabasz Index score of 4010.15 (k=3), and the binned data in a score of 4031.18 (k=4).

Model optimisation: eliminating outliers



- Running the k means algorithm on the pca model gives a Calinski-Harabasz Index of 6247.65
- Results not biased by outlier books and clusters are more accurate

Book recommendation website

Software and libraries used:

- Created database using SQLite to serve an API
- Set up API route to serve the data using Flask
- Created book recommendation website using a combination of HTML, CSS and JavaScript

How it works:

- User selects a book
- JavaScript calculates the 5 closest books (Euclidean distance) in the same cluster, and displays recommendation

Conclusions

Limitations:

- Complex dataset (huge number of features after one-hot-encoding)
- No data on user interactions and preferences
    - Does not allow for collaborative filters

BUT

We have built a successful skeleton for a personalized book recommendation system: with just a few tweaks (appropriate data), we believe this could be a very accurate and successful system.

The BookBots team

Ammarah Ahmad

Catarina Ferreira

Kashfi Khalid

Sadek Ahmed

THANK YOU!