Scientific Computing, Modeling & Simulation
Savitribai Phule Pune University

Master of Technology (M.Tech.)
Programme in Modeling and Simulation

Mini Project Report

Book Recommender System

Mayuri Adhav
MT2101

Academic Year 2022-23

Scientific Computing, Modeling & Simulation
Savitribai Phule Pune University

Certificate

This is certify that this report titled

Book Recommender System

authored by

Mayuri Adhav (MT2101)

describes the project work carried out by the author under our supervision during the period from July 2022 to Nov 2022. This work represents the project component of the Master of Technology (M.Tech.) Programme in Modeling and Simulation at the Department of Scientific Computing, Modeling & Simulation, Savitribai Phule Pune University.

Mihir Arjunwadkar, Head
SCMS-SPPU, Pune, India

Scientific Computing, Modeling & Simulation
Savitribai Phule Pune University

Author's Declaration

This document titled

Book Recommender System

authored by me is an authentic report of the project work carried out by me as part of the Master of Technology (M.Tech.) Programme in Modeling and Simulation at the Department of Scientific Computing, Modeling & Simulation, Savitribai Phule Pune University. In writing this report, I have taken reasonable and adequate care to ensure that material borrowed from sources such as books, research papers, internet, etc., is acknowledged as per accepted academic norms and practices in this regard. I have read and understood the University's policy on plagiarism (`http://unipune.ac.in/administration_files/pdf/Plagiarism_Policy_University_14-5-12.pdf`).

Mayuri Adhav
MT2101

Abstract

Now-a-days, everyone depending on reviews by others in many things such as selecting a movie to watch, buying products, reading a book. Recommender systems are used for that purpose only. A recommender system is a kind of filtering system that predicts a user's rating of an item. Recommender systems recommend items to users by filtering through a large database of information using a ranked list of predicted ratings of items. Online Book recommender system is a recommender system for ones who love books. When selecting a book to read, individuals read and rely on the book ratings and reviews that previous users have written. In this paper, Hybrid Recommender system is used in which Collaborative Filtering and Content-Based Filtering techniques are used. Collaborative techniques such as Clustering in which data-points are grouped into clusters. Algorithms such as K- means clustering and Gaussian mixture are used for clustering. The better algorithm was selected with the help of silhouette score and used for clustering. Matrix Factorization technique such as Truncated-SVD which takes sparse matrix as input is used for reducing the features of a dataset. Content Based Filtering System used TF- IDF vectorizer which took statements as input and return a matrix of vectors. RMSE (Root Mean Square Error) is used for finding the deviation of an absolute value from an obtained value and that value is used for finding the fundamental accuracy.

Keywords: Book Recommender System, Matrix Factorization, Clustering, K-Means, Gaussian Mixture, Root Mean Square Error.

.

Acknowledgements

I am indebted to **Prof. Mihir Arjunwadkar** for their guidance on a daily basis during this project. Prof. Mihir Arjunwadkar supported me a great deal on analysis and design of algorithms. I am also thankful to all of them for sharing their own wonderful reference articles with me and making me capable enough of carrying out such work independently in the future.

I would like to thank our parents, friends, and classmates for their encouragement throughout our project period. At last, but not the least, we thank everyone for supporting us directly or indirectly in completing this project successfully.

Contents

Chapter 1

Introduction

Now-a-days, online rating and reviews are playing an important role in books sales. Readers were buying books depend on the reviews and ratings by the others. Recommender system focuses on the reviews and ratings by the others and filters books. In this paper, Hybrid recommender system is used to boost our recommendations. The technique used by recommender systems is Collaborative filtering. This technique filters information by collecting data from other users. Collaborative filtering systems apply the similarity index-based technique. The ratings of those items by the users who have rated both items determine the similarity of the items. The similarity of users is determined by the similarity of the ratings given by the users to an item. Content-based filtering uses the description of the items and gives recommendations which are similar to the description of the items. With these two filtering systems, books are recommended not only based on the user's behaviour but also with the content of the books. So, our recommendation system recommends books to the new users also. In this recommender system, books are recommended based on collaborative filtering technique and similar books are shown using content based filtering. The required dataset for the training and testing of our model is downloaded from Good-Reads website. Matrix Factorization technique such as Truncated-SVD which takes sparse matrix of dataset is used for reduction of features. The reduced dataset is used for clustering to build a recommendation system. Clustering is a collaborative filtering technique that is used to build our recommendation system in which data points are grouped into clusters. . In this paper, we used two methods i.e., K-means and Gaussian mixture for clustering the users. The better model is selected based on the silhouette score and used for clustering. Silhouette score or silhouette coefficient is used to calculate how good the clustering is done. Negative value shows that clustering is imperfect whereas positive value shows that clustering was done perfectly. Difference between the mean rating before clustering and after clustering is calculated. Root Mean square Error is used to measure the error between the absolute values and obtained values. That RMSE value is used to find the fundamental accuracy.

1.0.1 Introduction to Machine learning

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. It gives the computer that makes it more similar to humans i.e. ability to learn. Machine learning is used in many streams than anyone would accept. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers

learn automatically without human intervention or assistance and adjust actions accordingly. Machine Learning is a sub-area of artificial intelligence, whereby the term refers to the ability of IT systems to independently find solutions to problems by recognizing patterns in databases. In other words: Machine Learning enables IT systems to recognize patterns on the basis of existing algorithms and data sets and to develop adequate solution concepts. Therefore, in Machine Learning, artificial knowledge is generated on the basis of experience.

In order to enable the software to independently generate solutions, the prior action of people is necessary. For example, the required algorithms and data must be fed into the systems in advance and the respective analysis rules for the recognition of patterns in the data stock must be defined. Once these two steps have been completed, the system can perform the following tasks by Machine Learning: • Finding, extracting and summarizing relevant data • Making predictions based on the analysis data • Calculating probabilities for specific results

Basically, algorithms play an important role in Machine Learning: On the one hand, they are responsible for recognizing patterns and on the other hand, they can generate solutions. Algorithms can be divided into different categories:

1.0.2 Supervised learning

In the course of monitored learning, example models are defined in advance. In order to ensure an adequate allocation of the information to the respective model groups of the algorithms, these then have to be specified. In other words, the system learns on the basis of given input and output pairs. In the course of monitored learning, a programmer, who acts as a kind of teacher, provides the appropriate values for a particular input. The aim is to train the system in the context of successive calculations with different inputs and outputs to establish connections. Supervised learning is where you have input variables (X) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output. Y = f(X) The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (Y) for that data. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance. Techniques of Supervised Machine Learning algorithms include linear and logistic regression, multi-class classification, Decision Tree and Support Vector Machine. Supervised Learning problems can be further grouped into Regression and Classification problems. The difference between these two is that the dependent attribute is numerical for regression and categorical for classification:

• Regression:

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

• Classification:

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories. In short classification either predicts categorical class labels or classification data based on the training set and the values(class labels) in classifying attributes and uses it in classifying new data. There are number of classification models. Classification models include Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Tree, One-vs.-One and Naïve Bayes.

1.0.3 Unupervised learning

In unsupervised learning, artificial intelligence learns without predefined target values and without rewards. It is mainly used for learning segmentation (clustering). The machine tries to structure and sort the data entered according to certain characteristics. For example, a machine could (very simply) learn that coins of different colors can be sorted according to the characteristic "color" in order to structure them. Unsupervised Machine Learning algorithms are used when the information used to train is neither classified nor labeled. The system does not figure out the right output but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data. Unsupervised Learning is the training of Machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Unsupervised Learning is classified into two categories of algorithms:

• Clustering: A clustering problem is where you want to discover the inherent grouping in the data such as grouping customers by purchasing behavior.

• Association: An Association rule learning problem is where you want to discover rules that describe large portions of your data such as people that buy X also tend to buy Y.

1.0.4 Application of Machine learning

1) Virtual Personal Assistant

Siri, Alexa, Google Now are some of the popular examples of virtual personal assistants. As the name suggests, they assist in finding information, when asked over voice. Machine learning is an important part of these personal assistants as they collect and refine the information on the basis of your previous involvement with them. Later, this set of data is utilized to render results that are tailored to your preferences.

Virtual Assistants are integrated to a variety of platforms. For example: • Smart Speakers: Amazon Echo and Google Home • Smartphone: Samsung Bixby on Samsung S8 • Mobile Apps: Google Allo

• Imagine a single person monitoring multiple video cameras! Certainly, a difficult job to do and boring as well. This is why the idea of training computers to do this job makes sense.

• The video surveillance system, nowadays are powered by AI that makes it possible to detect crime before they happen. They track unusual behaviour of people like standing motionless for a long time, stumbling, or napping on benches etc. The system can thus give an alert to human attendants, which can ultimately help to avoid mishaps. And when such activities are reported and counted to be true, they help to improve the surveillance services. This happens with machine learning doing its job at the backend. Social Media Services: From personalizing your news feed to better ads targeting, social media platforms are utilizing machine learning for their own and user benefits. • People You May Know

• Face Recognition

2)Search Engine Result Refining:

Google and other search engines use machine learning to improve the search results for you. Every time you execute a search, the algorithms at the backend keep a watch at how you respond to the results. If you open the top results and stay on the web page for long, the search engine assumes that the results it displayed were in accordance to the query. Similarly, if you reach the second or third page of the search results but do not open any of the results, the search engine estimates that the results served did not match requirement. This way, the algorithms working at the backend improve the search results.

3)Clustering

Clustering is an unsupervised learning method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Clustering is very important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. This algorithm must make some assumptions which constitute the similarity of points and each assumption make different and equally valid clusters.

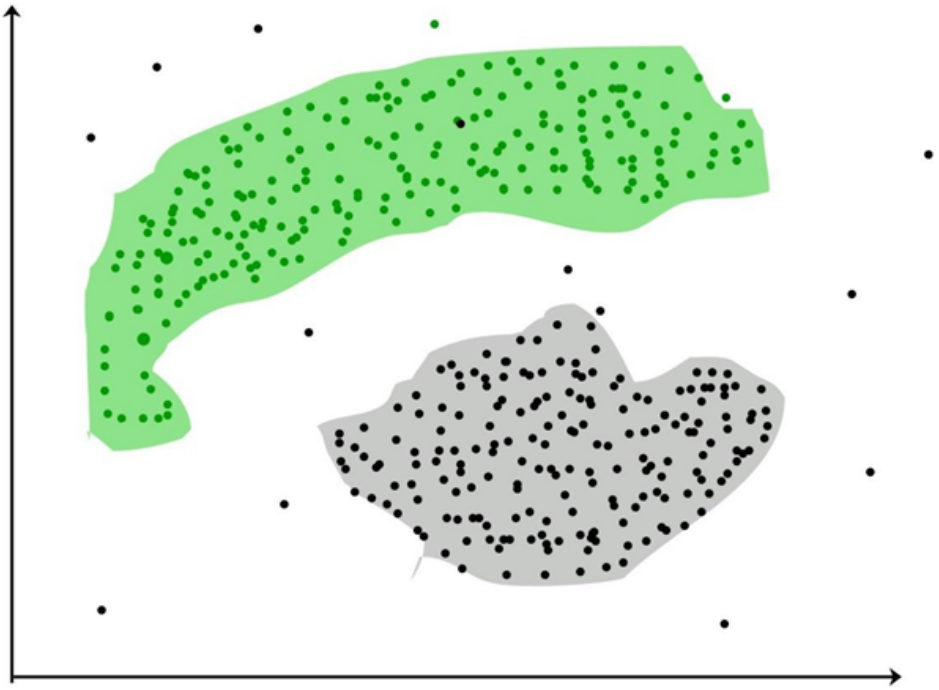

Figure 1.1: Clustering of data point

1.0.5 Clustering Method

1]Density-Based Methods:

These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and ability to merge two clusters.

Example: DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure) etc.

2]Hierarchical Based Methods:

The clusters formed in this method forms a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two categories: 1]Agglomerative (bottom up approach) 2]Divisive (top down approach)

Examples: CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies) etc.

3]Partitioning Methods:

These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter

example K-means, CLARANS (Clustering Large Applications based upon Randomized Search) etc.

4]Grid-based Methods:

In this method the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operation done on these grids are fast and independent of the number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (Clustering In Quest) etc.

1.0.6 Matrix Factorization

Matrix factorization is a way to generate latent features when multiplying two different kinds of entities. Collaborative filtering is the application of matrix factorization to identify the relationship between items and users entities. With the input of users' ratings on the shop items, we would like to predict how the users would rate the items so the users can get the recommendation based on the prediction. Matrix Factorization is a technique to discover the latent factors from the ratings matrix and to map the items and the users against those factors. Consider a ratings matrix R with ratings by n users for m items. The ratings matrix R will have n $\times$ m rows and columns.

Matrix Factorization is a significant approach in many applications. Curse of dimensionality is a phenomenon which occurs in high dimensional space that hardly occur in lower dimensional space. Due to higher number of dimension model gets sparse. Higher dimensional space causes problem in clustering (becomes very difficult to separate one cluster data from another), search space increases, complexity of model increases. We can reduce the dimension by following two ways: • Feature selection: Selecting important features which are relevant to model (it avoids the curse of dimensionality) • Feature extraction: Transformation of high dimensional space into lower dimensional space by using various methods such as PCA, TSVD, T- SNE etc.

In this paper, we used Feature extraction for reducing the features and used method i.e. Truncated SVD for dimensionality reduction.

1.0.7 Silhouette Score

Silhouette score or silhouette coefficient is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. It refers to a method of interpretation and validation of consistency within clusters of data. Silhouette Score is a metric used to calculate the goodness of a clustering technique. The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance. Where,

Silhouette Score = (b-a)/max (a, b)

a = average intra-cluster distance i.e. the average distance between each point within a cluster.

b = average inter-cluster distance i.e. the average distance between all clusters.

This value ranges from -1 to 1. Positive value indicates that mean clusters are well apart from each other and clearly distinguished so we require a ¡¡ b. 0 indicates that mean clusters are indifferent, or we can say that the distance between clusters is not significant. Negative value indicates that mean clusters are assigned in the wrong way. We can also increase the likelihood of the silhouette being maximized at the correct number of clusters by re-scaling the data using feature weights that are cluster specific.

1.0.8 Motivation of the work

In the present world, all products are buying based on the reviews and ratings by the others. There are so many products with high rating and reviews but we only put our interest in some products which we like. Recommender system works on this principle only i.e. it recommends products based on the interest of the users. Our idea is to create recommender system that recommends books based on the user's interest i.e. we recommends books which are similar to the books that user already liked. It can also recommend books which are liked by similar users. Similar users are those who liked the books which are liked by the current user.

We also add another feature i.e. we recommends books which are independent of the users interest. With this feature, we can recommend books to the new users also. Book recommendation sites that were available online now a days shows the books which are recommended by the system. Here, we are also recommending books based on the description of the book. We will get books which are similar to the book we selected in this system. For that purpose only, we built Hybrid recommender system. Hybrid recommender system is a combination of Collaborative Filtering system and Content Based Filtering system.

1.1 Types of Book Recommender System

A recommendation system is usually built using 3 techniques which are content-based filtering, collaborative filtering, and a combination of both.

1.1.1 Content-Based Filtering

The algorithm recommends a product that is similar to those which used as watched. In simple words, In this algorithm, we try to find finding item look alike. For example, a person likes to watch Sachin Tendulkar shots, so he may like watching Ricky Ponting shots too because the two videos have similar tags and similar categories.

Only it looks similar between the content and does not focus more on the person who is watching this. Only it recommends the product which has the highest score based on past preferences.

1.1.2 Collaborative-based Filtering

Collaborative based filtering recommender systems are based on past interactions of users and target items. In simple words here, we try to search for the look-alike customers and offer products based on what his or her lookalike has chosen. Let us understand with an example. X and Y are two similar users and X user has watched A, B, and C movie. And Y user has watched B, C, and D movie then we will recommend A movie to Y user and D movie to X user.

Youtube has shifted its recommendation system from content-based to Collaborative based filtering technique. If you have experienced sometimes there are also videos which not at all related to your history but then also it recommends it because the other person similar to you has watched it.

1.1.3 Hybrid Filtered Method

Hybrid Filtering Method It is basically a combination of both the above methods. It is a too complex model which recommends product based on your history as well based on similar users like you.

There are some organizations that use this method like Facebook which shows news which is important for you and for others also in your network and the same is used by Linkedin too.

1.2 Collecting Data

I have collect the data from Kaggle in which three datasets are present i.e. Books Dataset, Ratings Dataset, Users Dataset.

	ISBN	Book-Title	Author	Year	Publisher	Image-URL-S	Image-URL-M
0	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/0195153448.0...
1	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/0002005018.0...
2	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...
3	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux	http://images.amazon.com/images/P/0374157065.0...	http://images.amazon.com/images/P/0374157065.0...
4	0393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton & Company	http://images.amazon.com/images/P/0393045218.0...	http://images.amazon.com/images/P/0393045218.0...

Figure 1.2: dataset

Book dataset contains various feature ISBN,Book-Title,Book-Author Year-Of-Publication,Publisher,Ima URL-S some feature are not important so we remove that feature from our dataset.book dataset consist of (271360, 8) data.top 10 author are given below top 10 publisher are given below

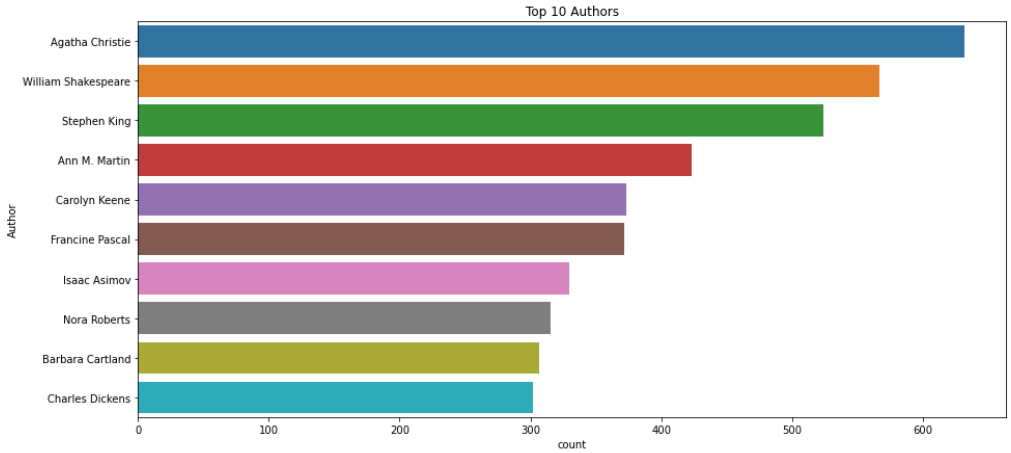

Figure 1.3: top 10 author

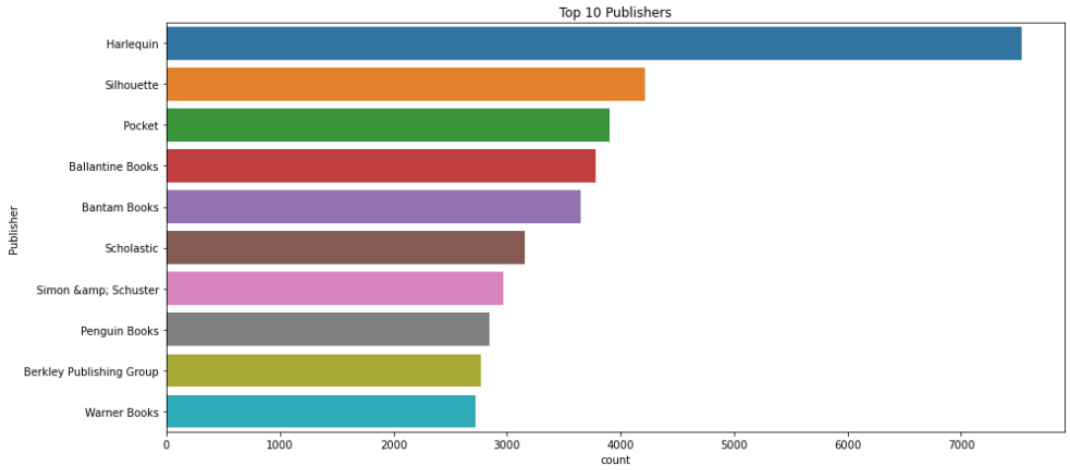

Figure 1.4: top 10 publisher

user data set consist of (278858, 3) data user data set have three different column User-ID,Location,Age.we have to consider age column because people below 7 age cant read book.so we have to consider the criteria that people above the age 20 who read the book.

	User-ID	Location	Age
0	1	nyc, new york, usa	NaN
1	2	stockton, california, usa	18.0
2	3	moscow, yukon territory, russia	NaN
3	4	porto, v.n.gaia, portugal	17.0
4	5	farnborough, hants, united kingdom	NaN

Figure 1.5: User

1.2.1 Finding Outliers

The Book-Crossing dataset comprises 3 files. Users : Contains the users. Note that user IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values. Books : Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title, Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services. Note that in the case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavors (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon website. Ratings : Contains the book rating information. Ratings

(Book-Rating) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

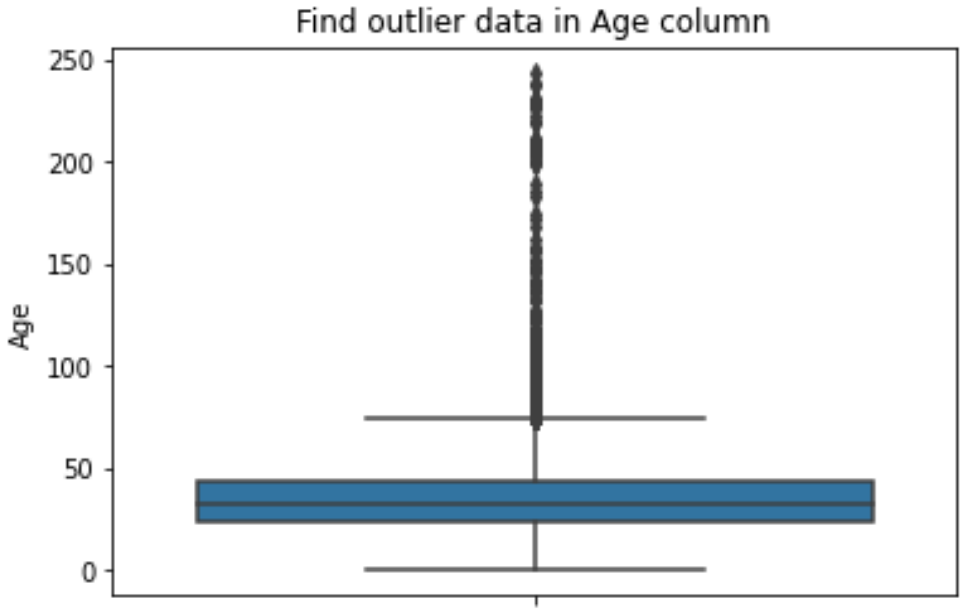

Figure 1.6: User

Chapter 2

Building Model

2.1 Building Model

"the overall structure of the system and the ways in which the structure provides conceptual integrity". The system architecture to build a recommendation system involves the following five major steps. Data Acquisition Data Pre-processing Feature Extraction Training Methods Testing Data

Dataset was collected from kaggle in which three datasets are present i.e. Books Dataset, Ratings Dataset, Users Dataset. In Step Datasets were pre-processed to make suitable for developing the Recommendation system. Feature extraction is performed in which Truncated-SVD is used to reduce the features of the dataset and Data splitting is done in which training dataset and testing dataset are divided into 80:20 ratio.Content Based Filtering System is developed in which book description is taken as an input and Collaborative Filtering System is developed by building a model using K-Means Algorithm over Gaussian Mixture after comparing with Silhouette scores.Testing of model with test data is performed. After acquiring the data our next step is to read the data from the csv file into python notebook. Python notebook is used in our project for data pre-processing, features selection and for model comparison. we have read data from csv file using the inbuilt python functions that are part of pandas library

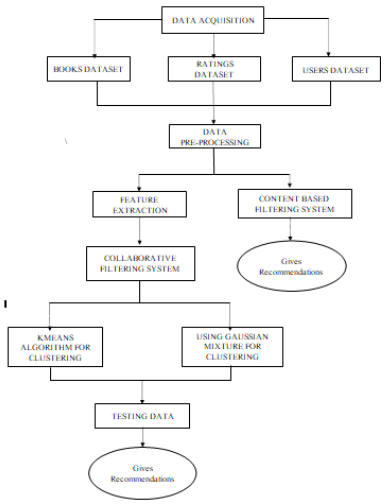

Figure 2.1: System Architecture

2.1.1 Data Acquisition:

The goal of this step is to find and acquire all the related datasets or data sources. In this step, the main aim is to identify various available data sources, as data are often collected from various online sources like databases and files. The size and the quality of the data in the collected dataset will determine the efficiency of the model. The Books dataset is collected from the kaggle website.After acquiring the data our next step is to read the data from the csv file into python notebook. Python notebook is used in our project for data pre-processing, features selection and for model comparison. I have read data from csv file using the inbuilt python functions that are part of pandas library.

2.1.2 Data Preprocessing:

I have to check null values and remove them as they may affect efficiency. Identifying duplicates in the dataset and removing them is also done in this step

User-ID Book-Title	254	2276	2766	2977	3363	3757	4017	4385	6242	6251	...	274004	274061	274301	274308	274808	275970	277427	277478	277639	278
1984	9.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	N
1st to Die: A Novel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
2nd Chance	NaN	10.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	0.0	N
4 Blondes	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
84 Charing Cross Road	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	10.0	NaN	NaN	NaN	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Year of Wonders	NaN	NaN	NaN	7.0	NaN	NaN	NaN	NaN	7.0	NaN	...	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	N
You Belong To Me	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
Zen and the Art of Motorcycle Maintenance: An Inquiry into Values	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	0.0	...	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	N
Zoya	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N

Figure 2.2: Finding Null value in Dataset

drop the null values from the dataset and then replaced with empty strings so that we can use it to fit for TF-IDF Vectorizer model

2.1.3 Feature Extraction:

After pre-processing the acquired data, the next step is to reduce the features i.e. Dimensionality reduction. The reduced features should be able to give high efficiency. We used Matrix Factorization technique such as Truncated SVD which takes sparse matrix as input for reduction of features.

2.1.4 Splitting dataset into train and test dataset:

divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose, if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.Usually, dataset will be split into train and test in the ratio of

8:2 i.e., 80 percent of data is used for training and 20 percent of data is used for testing the model fill missing value with 0

User-ID	254	2276	2766	2977	3363	3757	4017	4385	6242	6251	...	274004	274061	274301	274308	274808	275970	277427	277478	277639	2784
Book-Title																					
1984	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1st to Die: A Novel	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2nd Chance	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4 Blondes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
84 Charing Cross Road	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Year of Wonders	0.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	7.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
You Belong To Me	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Zen and the Art of Motorcycle Maintenance: An Inquiry into Values	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Zoya	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Figure 2.3: Fill Null value with 0

2.1.5 Training Methods:

I have our training and testing data. The next step is to identify the possible training methods and train our models. We have used two different clustering methods for training models. After that based on the silhouette score of each model, we would decide on which model to use finally.

2.1.6 K-Means Clustering Algorithm

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters in such a manner that each dataset belongs to only one group that has similar features. Here K defines the number of pre-defined clusters. We have to associate each cluster with a centroid in this algorithm. The sum of distances between the data point and their corresponding clusters should be minimized. The unlabeled dataset is taken as an input and the dataset into k-number of clusters is divided, and the process is repeated until it does not find the best clusters. We have to predetermine the k value in this algorithm. Elbow method is used to find the value of k which decides the number of clusters. This method uses the Within Cluster Sum of Squares (WCSS) value that defines the total variations within a cluster. The Formula for calculating the value of WCSS for n clusters is as follows:

WCSS = Pi in Cluster1distance (Pi C1)2 +Pi in Cluster2distance (Pi C2)2+.....+Pi in Cluster n distance(Pi Cn)2 The basic steps involved in K-Means Clustering algorithm is as follows: Step-1: Select the number K which gives the number of clusters. Step-2: Select random K number of points or centroids.

Step-3: Each data point to their nearest centroid should be assigned, which forms the predefined K clusters. Step-4: Calculate the variance and place a new centroid for each cluster. Step-5: We have to repeat the step-3, each data-point to the new closest centroid of each cluster should be reassigned. Step-6: If reassignment happens, then go to step-4 or else go to step-7. Step-7: Stop. In scikit-learn python library, sklearn.cluster.KMeans module is used for carrying out K-Means Clustering. I have to specify the number of clusters a parameter for this function.

2.1.7 Guassian-Mixture

Gaussian Mixture models are powerful clustering algorithms. It assumes that there are a certain number of Gaussian distributions where each distribution represents a cluster. This model groups the data points together into a single distribution. These models used the soft clustering technique for assigning data points to Gaussian distributions. In a one dimensional space, the probability density function of a Gaussian distribution (univariate) is as follows: P(x — µ, 2) = N (x; µ, 2) = 1/ Z [exp((x µ) 2/ 22)] Where, Z is the normalization constant i.e., Z = 22, µ is the mean i.e., µ = E[x], and 2 is the variance of the distribution i.e., 2 = var[x]. In a multidimensional space, the probability density function of a Gaussian distribution (multivariate) is as follows: P(x — µ,) = N (x; µ,) = 1/Z [exp(1/2 (xµ)T1(xµ))] Where, X is the input vector, is the 2D mean vector, and is the 2×2 covariance matrix Thus, we would have K (number of clusters) Gaussian distributions.

2.2 Implementation of single valu decomposition and negative matrix factorization

2.2.1 Implementation

The goal of the recommender system is to predict user preference for a set of items based on the past experience. Two the most popular approaches are Content-Based and Collaborative Filtering. Collaborative filtering is a technique used by websites like Amazon, YouTube, and Netflix. It filters out items that a user might like on the basis of reactions of similar users. There are two categories of collaborative filtering algorithms: memory based and model based.

Model based approach involves building machine learning algorithms to predict user's ratings. They involve dimensionality reduction methods that reduce high dimensional matrix containing abundant number of missing values with a much smaller matrix in lower-dimensional space. The goal of this section is to compare SVD and NMF algorithms, try different configurations of parameters and explore obtained results. Model based approach involves building machine learning algorithms to predict user's ratings. They involve dimensionality reduction methods that reduce high dimensional matrix containing abundant number of missing values with a much smaller matrix in lower-dimensional space. The goal of this section is to compare SVD and NMF algorithms, try different configurations of parameters and explore obtained results This analysis will focus on book recommendations based on Book-Crossing dataset. To reduce the dimensionality of the dataset and avoid running into memory error we will focus on users with at least 3 ratings and top 10using svd we get result

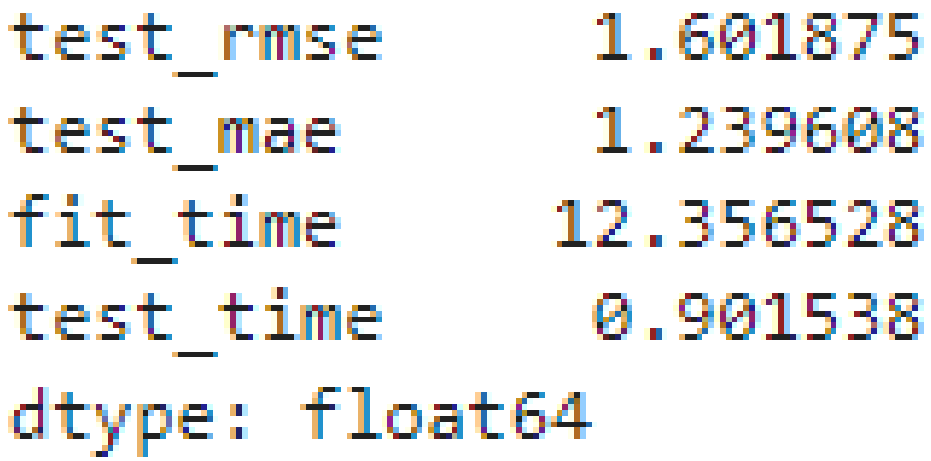

Figure 2.4: single value decompositation result

using negative matrix factorization result

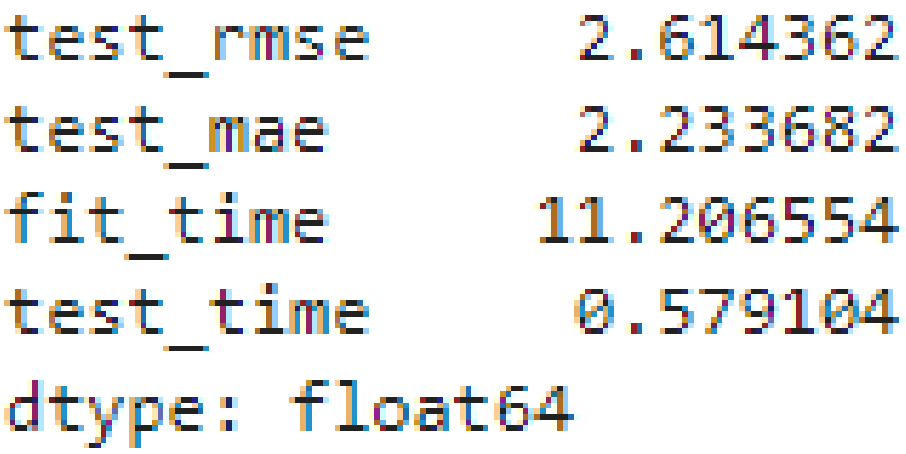

Figure 2.5: nvf result

Optimisation of SVD algorithm

Grid Search Cross Validation computes accuracy metrics for an algorithm on various combinations of parameters, over a cross-validation procedure. It's useful for finding the best configuration of parameters.

It is used to find the best setting of parameters: $\mathrm{n}_{factors} - the number of factors n_{epochs} - the number of iteration of the SGD procedure lr_{all} - the learning rate for all parameters reg_{all} - the regularizat$

Distribution of actual and predicted ratings in the test set According to the distribution of actual ratings of books in the test set, the biggest part of users give positive scores - between 7 and 10. The mode equals 8 but count of ratings 7, 9, 10 is also noticeable. The distribution of predicted ratings in the test set is visibly different. One more time, 8 is a mode but scores 7, 9 and 10 are clearly less frequent. Absolute error of predicted ratings The distribution of absolute

errors is right-skewed, showing that the majority of errors is small: between 0 and 1. There is a long tail that indicates that there are several observations for which the absolute error was close to 10.

How good/bad the model is with predicting certain scores? As expected from the above charts, the model deals very well with predicting score = 8 (the most frequent value). The further the rating from score = 8, the higher the absolute error. The biggest errors happen to observations with scores 1 or 2 which indicates that probably the model is predicting high ratings for those observations.

2.3 knn algorithm

kNN is a machine learning algorithm to find clusters of similar users based on common book ratings, and make predictions using the average rating of top-k nearest neighbors. For example, we first present ratings in a matrix, with the matrix having one row for each item (book) and one column for each user, like so: then find the k item that have the most similar user engagement vectors. In this case, Nearest Neighbors of item id 5= [7, 4, 8, . . .].convert our table to a 2D matrix, and fill the missing values with zeros (since we will calculate distances between rating vectors). then transform the values(ratings) of the matrix dataframe into a scipy sparse matrix for more efficient calculation.i use unsupervised algorithms with sklearn.neighbors. The algorithm we use to compute the nearest neighbors is “brute”, and we specify “metric=cosine” so that the algorithm will calculate the cosine similarity between rating vectors. Finally, fit the model.In this step, the kNN algorithm measures distance to determine the “closeness” of instances. It then classifies an instance by finding its nearest neighbors, and picks the most popular class among the neighbors.

```
Recommendations for Wish You Well:

1: Dark Lady, with distance of 0.8869215460827691:
2: Gone For Good, with distance of 0.8983158901902949:
3: Plum Island, with distance of 0.9013193486412017:
4: The Cottage, with distance of 0.9052598738590031:
5: Middle of Nowhere, with distance of 0.9064056942266255:
```

Figure 2.7: knn prediction on distance

Chapter 3

Results and Discussion

3.1 Experimental analysis and Performance Measures

3.1.1 Performance measure

The books dataset which initially has 2080 features in it has been reduced to 200 features after using Truncated SVD in feature extraction. This reduced dataset has been used to build the models. By calculating the coefficient of silhouette, the quality of clustering of different trained models can be compared.After building the model using the training data for clustering, the next step is to measure the performance of the model. To evaluate the efficacy of the model, silhouette score is calculated

3.1.2 Root Mean Square

Root mean squared error (RMSE) is the square root of the mean of the square of all of the error. The use of RMSE is very common, and it is considered an excellent general-purpose error metric for numerical predictions. The accuracy calculated by taking the RMSE value is known as Fundamental vertical accuracy whose value is computed by 1.96 * RMSE. Root mean square error can be expressed as Where N is the number of data points, y(i) is the ith measurement, and y (i) is its corresponding prediction. By squaring errors and calculating a mean, RMSE can be heavily affected by a few predictions which are much worse than the rest. If this is undesirable, using the absolute value of residuals and/or calculating median can give a better idea of how a model performs on most predictions, without extra influence from unusually poor predictions. Cosine Similarity Matrix Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, the higher the cosine similarity. You have to compute the cosine similarity matrix which contains the pairwise cosine similarity score for every pair of sentences (vectorized using tf-idf). Remember, the value corresponding to the ith row and jth column of a similarity matrix denotes the similarity score for the ith and jth vector. Use $\text{linear}_kernel()and pass tfidf_matrix to compute the cosine similarity matrix cosine_sim.$

3.1.3 Performance Analysis and Model Comparison

Out of all the trained models we need to choose the best model. We need to analyze the performance of each model and then compare the silhouette scores of the two trained modelssilhouette score for K- Means clustering model. This coefficient of silhouette of this model is found

out to be 0.02688953262460168, which is positive that means the clustering is good and appropriate.This coefficient of determination of this model is found out to be 0.025855911136862, which is also positive but less than that of K-Means clustering model. By the above data and calculations, it is evident that K-Means Clustering model is more efficient for clustering. Then we have analyzed each cluster

```
The cluster 0 mean is: 3.7488141202426917
book_title
Fifty Shades of Grey                    4.866667
The Clan of the Cave Bear               4.611111
Ready Player One                        4.583333
The Notebook                            4.526316
Lolita                                  4.517241
Harry Potter and the Goblet of Fire     4.500000
The Three Musketeers                    4.485714
A Million Little Pieces                 4.466667
The Art of Racing in the Rain           4.454545
2001: A Space Odyssey                   4.454545
dtype: float64
```

Figure 3.1: analyzing cluster using model

I have performed our clustering on a dataset that included 200 composite features. It is difficult to create a visualization that effectively illustrates all of these features. Therefore, I have selected the two top features, which played the most significant role in the clustering, and created a scatterplot that illustrates the clusters across those features.

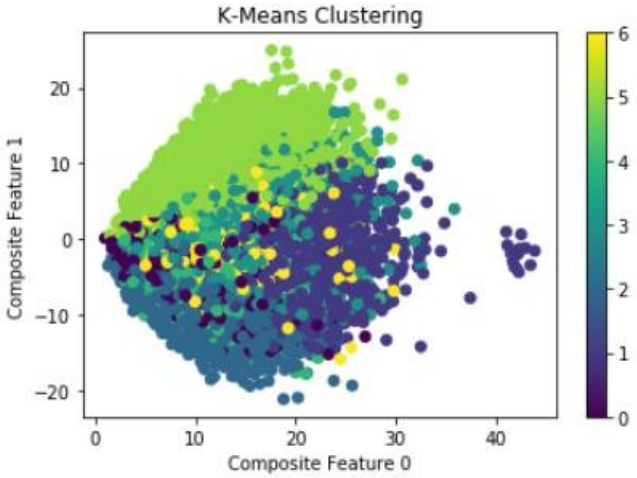

Figure 3.2: Visualizing cluster across two composite feature

indicates that the clustering model was effective in grouping users based on their values for composite features 0 and 1. The scatterplot also indicates that a few users had outlier values for composite feature 0.

3.1.4 Results and Discussion

The best fit model for our system has been found out through above mentioned model's comparison. The below table shows silhouette coefficient for two models.

	K-Means	Gaussian Mixture
Silhouette score	0.02688953262460168	0.025855911136862

Figure 3.3: Comparing Silhouette scores of two models

In the above table, we are checking the silhouette coefficient of two models. After checking the two scores, we consider the model using K-Means method with cluster count of 7 because of higher score. Then, this model is used to cluster the users. indicates that users in the test set, on average, rated their clusters' favorite books higher than a random set of 10 books by 0.47 stars, or nearly half a star. RMSE (Root mean square error) measures the error caused by the deviation between the sample values and predicted values by the model. The accuracy calculated by taking the RMSE value is known as Fundamental vertical accuracy whose value is computed by 1.96 * RMSE

RMSE	0.5957791790493179
Accuracy	1.167727190936663

Figure 3.4: RMSE and Accuracy for the obtained values

3.1.5 Result

final result of this project using knn algorithm is given below.it recommend book upon how many user read tis book and give rating using distance calculation.

```
Index(['Harry Potter and the Chamber of Secrets (Book 2)',
       'Harry Potter and the Prisoner of Azkaban (Book 3)',
       'Harry Potter and the Goblet of Fire (Book 4)',
       'Harry Potter and the Sorcerer's Stone (Book 1)', 'Exclusive',
       'The Cradle Will Fall'],
      dtype='object', name='Book-Title')
```

Figure 3.5: Recommended book

Chapter 4

Summary and Conclusion

In this project, I have recommended the books for a user using the model trained using K-Means Clustering which is a Collaborative Filtering Technique. I have also compared different models built using different methods and identified the best model and justifies why it has chosen that model. We have used the books dataset that is available in the kaggle website which consists of more than 3000 books. The models are built using the reduced features which is done by Truncated SVD. Based on those features the author built a model that gives a positive Silhouette score. The model that is suggested by this paper is useful for book readers. The system we have developed can make recommendations for

1. To calculate nearest neighbours.
2. To do matrix Factorization.
3. To create a pivot table.
4. To implement knn algorithm.
5. Build model using python language.

Bibliography