CDS513: Predictive Business Analytics
Academic Session: Semester 2, 2023/2024
School of Computer Sciences, USM, Penang

ASSIGNMENT 1

Task Type

This assignment is weighted as **10%** of the assessed work for the course. The assignment is an **individual** work.

COURSEWORK SUBMISSION GENERAL INFORMATION

Academic Integrity Statement

You must adhere to the university regulations on academic conduct. Formal inquiry proceedings will be instigated if there is any suspicion of plagiarism or any other form of misconduct in your work. Students must **NOT** collude with other groups of students or plagiarize their work. We practice zero tolerance towards plagiarism, and we use Turnitin to evaluate the similarity index. Your similarity index score must not exceed 20%.

Your tasks must be your own work. Unless the use of Artificial Intelligence (AI) is permitted in your assessment task, using AI to complete your assignment is a form of plagiarism.

Nature of the submission required

A softcopy of your assignment in PDF version should be submitted to lecturer through eLearn@USM. The zip package must be named according to the following notation: CDS513_ Assignment1_MatricNo_TopicNo. For example, the zip package is named as ***CDS513_Assignment1_111111/111222_T01.*** (Please refer to submission arrangement).

Diagrams may be used where they are helpful to support your arguments or description. If they are not your own work, the source must be referenced. Please help us to handle and mark your work efficiently.

Documentation guidelines

Student is required to submit a **SOFTCOPY** of the report and ensure that it use the following formatted styles: 1) Font type: **ARIAL**, 2) Font size: **11 pt**., 3) Line spacing: **Single spacing** and 4) Page layouts: **Justify**. Please make sure you have proper format alignment for all paragraphs, following standard writing style and use **American Psychological Association (APA)** for citation. Please include a **HEADER** with the following information: **MatrixID, Student name, Course code and Assignment type**. Please also include a proper cover page **(Appendix-B)**. Also include page number and list of references, which is shown in the last page.

CDS513: Predictive Business Analytics
Academic Session: Semester 2, 2023/2024
School of Computer Sciences, USM, Penang

ASSIGNMENT 1

Submission arrangement

1. Cover page (Appendix-B)
2. Table of Content
3. Main Report (Refer: Report Format)
4. Reference List or Bibliography List (whichever applicable)
5. Marking Rubric

Assignment Description

The assignment requires the students to choose a problem or data set from **Appendix A** and propose a solution using **Recommender systems (RS).** Provide background information about the chosen problem domain and describe the specific problem that will be addressed through the proposed recommender systems.

Name of Dataset Selection: See *Appendix A*.

Note: *You may come out with suitable scenarios and assumptions as deemed necessary. For a topic with more than one dataset, you are required to understand the relationship between the files and perform necessary action to come out with relevant information.*

Background

Recommendation systems have proven to be game-changers in the digital content industry. From streaming services like Netflix and Spotify to e-commerce giants like Amazon, personalized recommendations drive user engagement and sales. The backbone of any recommendation system is the data it relies on. A recommender system is a subclass of systems that seeks to predict the `"rating"` or `"preference"` a user would give to an item. It estimates a utility function that automatically predicts how a user will like an item, based on past behavior, relation to other users, item similarity and context.

Task 1: Perform the Recommender Systems: (35%)

Perform all three (3) recommender systems. Your recommender systems should be based on the topic chosen. Accomplish the necessary analysis have taught and discussed in the class. You are also strongly encouraged to do research to add value to your analysis. The analysis should reflect the real world's scenario application.

a. The Collaborative Filtering System:
   i. Model-based (as in the lab) OR
      Memory-based (similarity measure)
      ➢ User-to-User CF
      ➢ Item-to-Item CF

b. The Content-based System.

c. Calculate AUC, precision and mean average precision for all the recommender systems. (You may opt for suitable calculation type)

CDS513: Predictive Business Analytics
Academic Session: Semester 2, 2023/2024
School of Computer Sciences, USM, Penang

ASSIGNMENT 1

Task 2: Performance Analysis and Application Research (RS): (25%)

a. Select certain items for focus and analyze the pattern and impact. (not limited to:)
   i. *Item Selection:* Choose a subset of items from the dataset to focus on.
   ii. *Data Preparation*: Gather relevant data related to the selected items, including user ratings, item features, and any other metadata available.
   iii. Pattern Analysis: rating, correlation, temporal etc.,

b. Do precise application research on how to visualize the results. Do not forget to cite the reference/s.

Note: You may repeat the experiments with different options to get the optimal recommender systems.

<u>Report Format:</u> (40%)

Student is required to submit a **SOFTCOPY** of the report and ensure that it use the following formatted styles: 1) Font type: ARIAL, 2) Font size: 11 pt., 3) Line spacing: Single spacing and 4) Page layouts: **Justify.** Please make sure you have proper format alignment for all paragraphs, following standard writing style and use **American Psychological Association (APA)** for citation. Please include a **HEADER** with the following information: **MatrixID, Student name, Course code and Assignment type**. Please also include a proper cover page **(Appendix-B).** Also include page number and list of references, which is shown in the last page

Table of Contents

1.0 Project Background
- Background of the problem domain
- Description of the problem

2.0 Data Understanding & Integration*

3.0 Recommender Systems

4.0 Discussion & Analysis

5.0 Conclusion

Note:

**Data Understanding & Integration may include:*
- *Data description and visualization, understanding of the data distribution of the data etc.*
- *Integration – process of split or merging the data from original dataset, as well as the assumption regarding the dataset you deal with.*

- Summit your assignment report (minimum 5 pages, maximum of 20 pages, PDF document file).
- Your report should include the snap short of the process and output.

CDS513: Predictive Business Analytics
Academic Session: Semester 2, 2023/2024
School of Computer Sciences, USM, Penang

ASSIGNMENT 1

Report Submission Instruction

- Submit soft copy to eLearn@USM.
- The zip package must be named according to the following notation: CDS513_ Assignment1_*MatricNo_TopicNo*. For example, the zip package is named as **CDS513_Assignment1_*111111/111222_T01***.

ASSIGNMENT 1

Assignment Evaluation

This assignment will be graded **based on the marking scheme**

> **IMPORTANT:** Students who copied or plagiarized other's work or let their work be copied or plagiarized will be given an F grade. The student may be barred from sitting for final exam and reported to the university's disciplinary board.

Grading Rubric – Assignment 1

Course Learning Outcome (CLO):

- CLO2 Design strategies relevant to predictive business analytics using appropriate technologies and tools.
- CLO3 Assess the role of predictive business analytics in enhancing business performance.
- CLO5 Evaluate predictive business analytics using recent tools.

CDS513: Predictive Business Analytics Marking Rubric Assignment-1 – Weighted (10%)						
	Report: Table of Contents					
Report Component	(Poor)	(Average)	(Good)	(Excellent)	Weighted	Marks Obtained
Intro & Problem Background	Introduction and problem background are **poorly** explained.	Introduction and problem background are **fairly** explained.	Introduction and problem background are **adequately** explained.	Introduction and problem background are **clearly** explained.	10%	
Data Understanding & Integration	Data Understanding & Integration are **poorly** explained.	Data Understanding & Integration are **fairly** explained.	Data Understanding & Integration are **adequately** explained.	Data Understanding & Integration are **clearly** explained.	15%	
Discussion & Analysis	The results are **poorly** discussed and tailored to the problem addressed. Insights from the analysis are discussed and **poorly** explained.	The results are **fairly** discussed and tailored to the problem addressed. Insights from the analysis are discussed and **minimally** explained.	The results are **adequately** discussed and tailored to the problem addressed. Insights from the analysis are discussed and **adequately** explained.	The results are **clearly** discussed and tailored to the problem addressed. Insights from the analysis are discussed and well-explained.	15%	
					Report (40%)	

ASSIGNMENT 1

	Task 1: Performing & Task 2 Performance Analysis of Recommendation systems					Marks Obtained
Recommender Systems	RS is **poorly** performed.	RS is **fairly** performed.	RS is **adequately** performed.	RS is **successfully** performed.	35%	
Analysis of RS	Analysis of RS is **poorly** analysed.	Analysis of RS is **fairly** analysed.	Analysis of RS is **adequately** analysed.	Analysis of RS is **clearly** analysed.	20%	
Application of RS	The application of RS is **poorly** written and justified.	The application of RS is **fairly** written and justified.	The application of RS is a**dequately** written and justified.	The application of RS is **clearly** written and justified.	5%	
					Task1 & Task 2 (60%)	
					Overall Report (40% & Tasks (60%)	
					Weighted (10%)	
Comments:						

ASSIGNMENT 1

Appendix A – Dataset and Topic Selection

No.	Name of Dataset	Description	Sample Size	Recommender
1.	Steam Video Games	Steam is the world's most popular PC Gaming hub, with over 6,000 games and a community of millions of gamers. With a massive collection that includes everything from class AAA blockbusters to small indie titles, great discovery tools are a highly valuable asset for Steam. This dataset is a list of user behaviors, with columns: user-id, game-title, behavior-name, value. The behaviors included are 'purchase' and 'play'. The value indicates the degree to which the behavior was performed - in the case of 'purchase' the value is always 1, and in the case of 'play' the value represents the number of hours the user has played the game.	6000	X
		https://www.kaggle.com/datasets/tamber/steam-video-games/data		
2.	The Movies Dataset	These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been	45,000	X

		obtained from the official GroupLens website.		
		https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset		
3.	Book-Crossing Dataset	The Book-Crossing (BX) dataset was collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. It contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.	278,858	X
		https://www.kaggle.com/datasets/syedjaferk/book-crossing-dataset		
4.	Walmart Recruiting: Trip Type Classification	This dataset is used to classify shopping trips. Potential for market basket and recommender.	4521127	X
		https://www.kaggle.com/c/walmart-recruiting-trip-type-classification/data		
5.	Rating Disposition 2023	The ratings.csv file contains approximately 3,908,657 anonymous ratings of 68,044 movies made by 6,724 Movie Lens users who have logged in to the website over 12 times from 2019 to 2020 and rated over 20 movies since their registration.	3908657	X
		https://grouplens.org/datasets/rating-disposition-2023/		
6.	Restaurant Data with Consumer Ratings	This dataset was used for a study where the task was to generate a top-n list of restaurants according to the consumer preferences and find the significant features.	>1100	X
		https://www.kaggle.com/datasets/uciml/restaurant-data-with-consumer-ratings		

7.	Goodbooks-10k	This dataset contains ratings for ten thousand popular books. As to the source, let's say that these ratings were found on the internet. Generally, there are 100 reviews for each book, although some have less - fewer - ratings. Ratings go fromone to five	10000	X
		https://www.kaggle.com/datasets/zygmunt/goodbooks-10k		
8.	Retail Rocket	The dataset consists of three files: a file with behaviour data (events.csv), a file with item properties (item_properties.csv) and a file, which describes category tree (category_tree.csv). The data has been collected from a real-world ecommerce website. The behaviour data, i.e. events like clicks, add to carts, transactions, represent interactions that were collected over a period of 4.5 months. A visitor can make three types of events, namely “view”, “addtocart” or “transaction”. In total there are 2 756 101 events including 2 664 312 views, 69 332 add to carts and 22 457 transactions produced by 1 407 580 unique visitors. For about 90% of events corresponding properties can be found in the “item_properties.csv” file.	2756101	X
		https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset		

9.	Anime Recommendations Database	This data set contains information on user preference data from 73,516 users on 12,294 anime. Each user is able to add anime to their completed list and give it a rating and this data set is a compilation of those ratings.	12294	X
		https://www.kaggle.com/CooperUnion/anime-recommendations-database		
10.	Nursing Home Rating	The dataset ha the general information on currently active nursing homes, including number of certified beds, quality measure scores, staffing and other information used in the Five-Star Rating System. Data are presented as one row per nursing home.	14878	X
		https://data.cms.gov/provider-data/dataset/4pq5-n9py		
11.	H&M Personalized Fashion dataset	H&M Group is a family of brands and businesses with 53 online markets and approximately 4,850 stores. Our online store offers shoppers an extensive selection of products to browse through. But with too many choices, customers might not quickly find what interests them or what they are looking for, and ultimately, they might not make a purchase. To enhance the shopping experience, product recommendations are key. More importantly, helping customers make the right choices also has positive implications for sustainability, as it reduces returns, and thereby minimizes emissions from transportation.	>100000	X
		https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/discussion/306078		

12.	Food dataset	This data records interactions with Entree Chicago restaurant recommendation system (originally http://infolab.cs.uchicago.edu/entree) from September, 1996 to April, 1999. The data is organized into files roughly spanning a quarter year -- with Q3 1996 and Q2 1999 each only containing one month	50672	X
		http://archive.ics.uci.edu/dataset/123/entree+chicago+recommendation+data		
13.	Amazon Product Reviews	This is a large-scale Amazon Reviews dataset collected in 2023. This dataset contains 48.19 million items, and 571.54 million reviews from 54.51 million users	>48M	X
		https://jmcauley.ucsd.edu/data/amazon/ https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews		
14.	IMDB data from2006 to 2016	Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascrore. It consists of 1,000 most popular movies on IMDB in the last10 years.	1000	X
		https://www.kaggle.com/PromptCloudHQ/imdb-data		

Appendix -B: Assignment Cover page

ASSIGNMENT 1

Course/Semester	**Course Code and Title**
MASTER OF SCIENCE (DATA SCIENCE AND ANALYTICS) **SEMESTER II, ACADEMIC SESSION 2023/2024/COURSEWORK MODE**	**CDS513 Predictive Business Analytics**
Student's name / MatrixID:	**Lecturer's name**
	Prof. Madya Dr. J. Joshua Thomas (JJT)

Date issued	**Submission Deadline**	**Indicative Weighting**
15th April 2024	6th May 2024	20%
Assignment [1] title	Recommendation Systems	

This assessment assesses the following course learning outcomes

# as in Course Guide	**Learning Outcomes**
CLO1	-
CLO2	Design strategies relevant to predictive business analytics using appropriate technologies and tools.
CLO3	Assess the role of predictive business analytics in enhancing business performance.
CLO4	-
CLO5	Evaluate predictive business analytics using recent tools

Student's declaration

I certify that the work submitted for this assignment is my own and research sources are fully acknowledged.

Student's signature: Submission Date: