**Data Analytics Capstone Topic Approval Form**

**Student Name:** Jonathon Fryman

**Student ID:** 000974544

**Capstone Project Name:** Book Recommendation System – Comparing recommendations with historical user information

**Project Topic**: The use of recommendation systems in today’s online marketplace is an important business tool that can assist a business in pairing customers with products and services that best align with historical purchasing or usage behaviors. The goal of this project is to determine if a developed recommendation system model for a catalog of books is capable of recommending titles that correlate with existing ‘to\_read’ classifiers users have set regarding future reading interests.

**This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** Can a book recommendation model be developed that provides recommendations closely aligned with the titles existing users with similar reading interests have added to their ‘to\_read’ classification reflecting interest in reading later?

**Hypothesis**:   
**Null hypothesis**-. The recommendation system is not able to sufficiently generate recommendations that align with titles in the user’s ‘to\_read’ list  
  
**Alternate Hypothesis**-. The recommendation system is able to sufficiently recommend titles that are included in a user’s ‘to\_read’ list

**Context:** The books dataset contains 10,000 titles. Users can rate their previously read titles, add tags to the titles that classify the specific title, and add titles to a ‘to\_read’ list expressing interest in reading at a later date.  
  
The existing ‘to\_read’ dataset has an existing catalog of 912,705 occurrences of a user adding on of the 10,000 books with the ‘to\_read’ classifier. This catalog spans 53,424 individual user\_id that correlate with the individual users reading preference. It would be beneficial to determine if a model could be developed that would sufficiently be capable of recommending titles aligned with what existing users have specified having future interest.

**Data:** The dataset to be used for this analysis is publicly available at https://www.kaggle.com/datasets/zygmunt/goodbooks-10k.

The existing dataset to be used for this analysis contains 10,000 unique ‘book\_id’ that associate to specific titles. The full dataset to be used contains five unique .csv files. The files contained within the data set are ‘books.csv’, ‘ratings.csv’, ‘book\_tags.csv’, ‘tags.csv’, and ‘to\_read.csv’.  
  
The ‘books.csv’ file contains information on the specific title, such as ‘book\_id ‘, ‘original\_title’, ‘authors’, and ‘average\_rating’. The ‘ratings.csv’ file contains ratings associated with a specific ‘book\_id’ from a specific ‘user\_id’.   
  
The ’rating.csv’ file contains information records that associate a specific book, user, and the rating that individual user assigned for the title.  
  
The ‘book\_tags.csv’ file contains information that includes a specific ‘book\_id’, the associated ‘tag\_id’, and a ‘count’ of how many times that specific ‘book\_id’ has been associated with the given ‘tag\_id’.  
  
The ‘tags.csv’ file associates a ‘tag\_name’ with a ‘tag\_id’ that correlates to the ‘tag\_id’ used in the previously mentioned ‘book\_tags.csv’ file.   
  
Finally, the ‘to\_read.csv’ file contains a pairing between a ‘user\_id’ and a ‘book\_id’ that the given user had previously marked as having interest to be read later.  
  
The data being used for this analysis is provided under a [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/) license that allows users to share and adapt the data with proper credit given to the original data provider.

**Data Gathering:** The data-gathering methodology to be used for this analysis is Transactional Data Tracking. This methodology will be used with existing user reading history, historical ratings for titles within the Books data set, and the existing collection of titles associated with the user bases ‘to\_read’ classifier.

**Data Analytics Tools and Techniques**: Regression analysis, which is a predictive modeling technique, will be utilized for this project’s data-analysis.

**Justification of Tools/Techniques:** Regression analysis is appropriate for this analysis as the primary objective is to identify correlation between a dependent variable (a recommended book) and a set of independent variables( a user’s reading/rating history, the type/genre of books previously enjoyed, etc.).

The planned tools to be used for this analysis include the Python programming language, Jupyter Notebooks, and Tableau.

Jupyter Notebooks will be utilized for any required data cleaning, data analysis, data visualization, and model development for use with generating predictions and comparing against the existing ‘to\_read’ classification set. This will be achieved using the Python programming language and its many available packages and libraries that facilitate data analysis and model development.

Additionally, Tableau is planned to be utilized to create additional data visualizations for use within the written report associated with this project.

**Project Outcomes**: The anticipated project outcomes for this analysis would be to develop a model that can capably provide reading recommendations to users that closely align with titles the historical userbase has previously added with a ‘to\_read’ classification. This will ensure that the provided recommendations align with associated reading interests a user has previously identified.   
  
A model that can capably provide recommendations that correlate with the ‘to\_read’ classification will be able to relevantly provide suggestions to current/future users labeled as having similar reading interests that have not purchased a title or added it with their own ‘to\_read’ classifier.   
  
Providing suggestions of this type can increase future sales and generate interest in titles that a subset of users may have otherwise been unaware.

**Projected Project End Date**: 03/30/2023

**Sources**:

1. Foxtrot. (2017, September 2). *Goodbooks-10K*. Kaggle. Retrieved February 12, 2023, from https://www.kaggle.com/datasets/zygmunt/goodbooks-10k
2. Zygmuntz. (n.d.). *Zygmuntz/goodbooks-10K: Ten Thousand books, six million ratings*. GitHub. Retrieved February 12, 2023, from https://github.com/zygmuntz/goodbooks-10k
3. *Creative Commons License Deed*. Creative Commons - Attribution-ShareAlike 4.0 International - CC BY-SA 4.0. (n.d.). Retrieved February 12, 2023, from https://creativecommons.org/licenses/by-sa/4.0/
4. atharvakarpe. (2022, May 18). *Difference between null and alternate hypothesis*. GeeksforGeeks. Retrieved February 12, 2023, from https://www.geeksforgeeks.org/difference-between-null-and-alternate-hypothesis/

**Course Instructor Signature/Date:**

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.