# Propensity Model for Book Reading

In this project, the goal is to build a machine learning model to predict user’s book reading behavior. The dataset will be used is “GoodReads”, a public available dataset collected by UCSD. The datasets include user interaction with books. It can be accessed here:

<https://github.com/MengtingWan/goodreads>

There are multiple ways to download the data, such as:

<https://github.com/MengtingWan/goodreads/blob/master/download.ipynb>

The data file needed in this project is:

|  |  |  |
| --- | --- | --- |
| **7** | 1EFHocJIh5nknbUMcz4LnrMEJkwW3Vk6h | goodreads\_interactions\_fantasy\_paranormal.json.gz |

Data schema:

book\_id: string (mandatory column)

date\_added: string

date\_updated: string (mandatory column)

isRead: boolean

rating: long (mandatory column)

read\_at: string (mandatory column)

review\_id: string

started\_at: string

user\_id: string (mandatory column)

Scope:

1. The problem to be solved is called “Propensity Modeling”. Build a ML model that can use the first 11 months data to predict whether the user will rate a book positive or negative in December 2016.
2. More formally, given user\_id and item\_id, you need to predict if the user will be positive or negative for the item. You can assume rating >= 4 is positive label, and less than 4 is negative label. The input for the model will be (user\_id, item\_id) with other feature data and the output will be (positive or negative, confidence value). How to divide train and test set is what you define.
3. For this exercise, choose only mandatory columns (user\_id, book\_id, date\_updated, read\_at, rating)
4. Choose either date\_updated or read\_at as the timestamp column and filter data has a timestamp value in the year of 2016. Please describe the reason why you choose that timestamp column.
5. Data cleansing is needed if necessary, e.g. removing records without valid values. Besides code, please also describe how this is done.
6. The candidate has freedom to select tools, frameworks, and language.
7. By any chance if you use open source for function, please include the reference link.

The final deliverable should include:

1. A document describing the high-level architecture, analysis of the problem/solution, data preprocessing/cleansing methods, algorithm selected.
2. README file describing how to setup the development environment (tools, framework and instruction of how to run the program).
3. Code for data preprocessing, model training and validation. Proper comments are expected. Code can be in format such as Jupyter notebook.
4. Model performance report including which metric to use and how the model performs.
5. Description about the model, algorithm, and formula (short description of why you choose or created your own, why this model is proper for this problem. Any expected pros and cons of this model for the problem. If you referenced any paper or blog to get insight, please include them too).
6. A zip file with all required above.