APPLIED DATA SCIENCE

16971-A

FINAL PROJECT REPORT

DETERMINING POTENTIAL READERS FOR A TARGET BOOK

TEAM G:

Isha Bandi

Maitri Vasa

Rupika Nilakant

Table of Contents

1. Executive Summary

The project aims to find potential readers for a specific book. It presents stakeholders as publishing houses or bookstores. It takes into account the preferences of the users based on the books that they have read and reviewed, defined here on as user-book interaction. The user is defined as the main agent. In our problem statement, given user-book interaction data, potential users are ascertained for a target book. The supplemental objective here is to understand which users are more likely to buy the "new" book, given their likes, dislikes, past reading history, and their reviews/ratings on those books. The project's aim is to use a similarity score between the "target" book and the books read/liked by the users and present a group of users who are the target readers for that specific book.

In a broader scope, this tool will help the publishing houses/vendors to identify a set of users towards whom the marketing communications must be pushed out. Customized and targeted marketing strategies can be prepared, and thus, using the data science principles, the hope is to market the book to the correct crowd so as to utilize the marketing budget efficiently.

The project uses user-book interaction data from Goodreads. Due to the large amount of data, the data was scoped down to accommodate for the limited hardware resources. Feature Engineering included multiple aggregated statistics about the user's reading habits and similarity of the past read books. Multiple classification models were analyzed and Random Forests performed the best with an AUC of 0.95.

2. Motivation

Declining Book Industry

- Book publishing is a massive industry in the U.S., with revenue that is projected to reach nearly 44 billion U.S. dollars in 2020[1].
- In 2015, about 2.7 billion books were sold; a number that has remained fairly consistent in the last few years.
- With over 3 million books in print in the U.S. every year, more than a hundred thousand of them are new titles. Yet, only a tiny fraction attract considerable readership. For example, less than 500 books make it to the New York Times bestseller lists and only a handful of authors stay on the list for ten or more weeks.
- These demanding odds reflect the challenges of capturing an audience in today's highly competitive world.

Customer Needs

- Predicting a customer's potential next purchase is a powerful advantage that companies aim to have. It enables companies to easily forecast stock replenishment or create the right

[1] U.S. book industry/market—statistics & facts. Statista. https://www.statista.com/topics/1177/book-market/. Accessed 2015-09-29

marketing-mix, ranging from promotions to a truly personalized experience for their customers, with the final intent of influencing positively (as a facilitator) their purchasing behavior.
- Most of the existing software used in companies for cross-selling products to their customers use association rules of increasing numbers and complexities and therefore making it extremely difficult for the marketer or data analyst to predict the next-item that each customer will buy.
- Leveraging this predictive power allows companies to understand their customers at a truly user-level or comprehend future market influences. It provides the capability to create a personalized promotion or an experience for customers or easily forecasts stock replenishment required in stores.

3. Objective and Goals

The project intends to be an analytic solution to determine potential readers from a database for marketing a target book. It aims to find the target audience given their reading habits and the similarity of the current target book with the books the user has read in the past.
Output: Classify whether a particular user shall be interested in buying a book.

4. Stakeholders

Major Stakeholders include the Publishing and Printing Industry. The marketing departments of the firm can benefit from having a list of ‘most-likely-to-buy’ users.

5. Use Case Scenario

The project aims to create a foundation for an interactive tool that would automate the process of identifying potential readers for the target book. The user, ideally the marketing department employee, shall upload the details of the book that needs to be marketed. The tool will perform internal calculations and modeling tasks to predict whether a reader shall be interested in the new target book. The tool shall deliver a list of user id and their contact information (sensitive secure user information) and push out marketing communications aggressively.

6. Dataset

The dataset that was scraped from Goodreads was available at UCSD Book Graph. The statistics were:

- 2,360,655 books (1,521,962 works, 400,390 book series, 829,529 authors)
- 876,145 users; 229,154,523 user-book interactions in users' shelves (include 112,310,716 reads and 104,713,520 ratings)
- interactions per user: 261.55; interactions per book: 97.07

After multiple attempts to parse through the data on local machines, it was decided that onbly the Comics and Graphics Genre will be analysed for the purpose of the project. The following are the statistics of the Comics_Graphs Data:

- Number of Books: *89,411* books
- Number of User-Interaction Data:*7,366,386* interactions (*4,774,055* reads, *4,523,249* ratings)
- Number of Users: *342,415* users

After exploratory analysis, pre-processing and aggregation: the number of user-interaction data in the final dataset is 144360

7. Project Plan: Flowchart

Figure 1 depicts the internal working of the project:

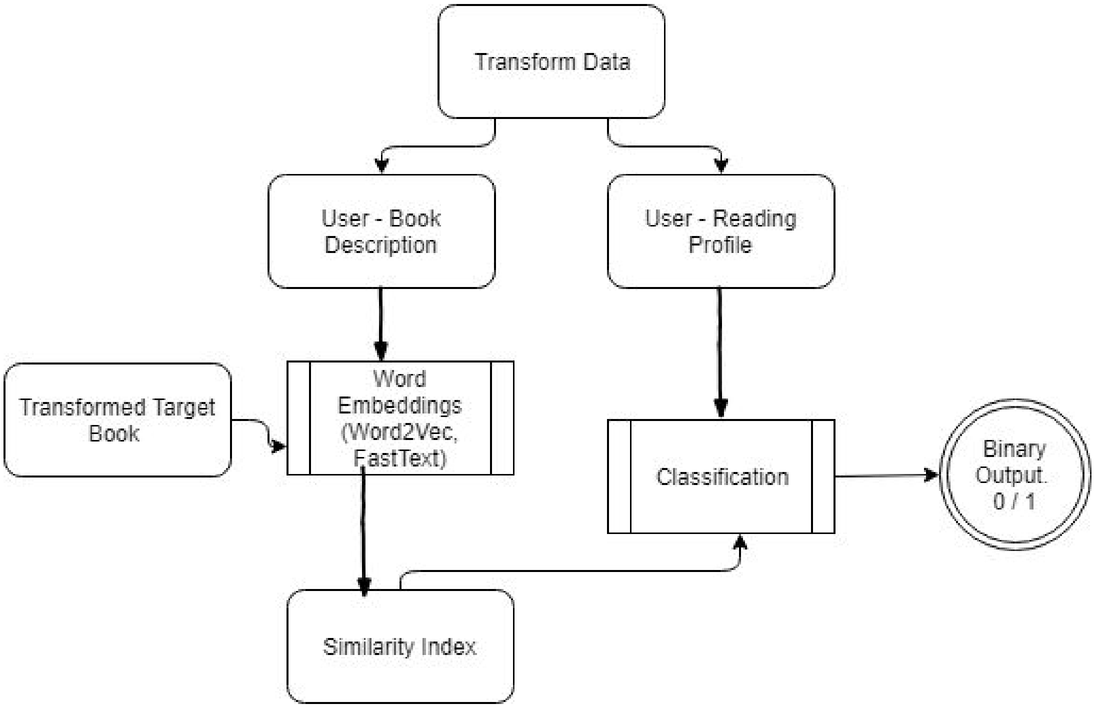


Figure 1: Flowchart depicting the internal working of the project

Further explanation of the flowchart is encapsulated below.

8. Exploratory Analysis

The initial exploratory analysis uncovered that the timestamps when the user read each book were missing for most of the rows in the user data. Hence, the change in the user's change of interest over each time could not be tracked.

Pre-processing

1. User - Interaction Data: Relevant field extracted.
    a. User_id
    b. Book_id
    c. Ratings given to book_id by user_id

2. Books Data
    a. Book_id
    b. ISBN
    c. Description
    d. Title
    e. Series: if the book belongs to a series.
    f. isFirst: If the book is first in the series.

3. Dealing with missing values and biases in Books and User - Description:
    a. Removed books with the missing description field
    b. Removed books if they were part of a series and are not the first book of the series.

Selection of Target Book

To maintain a uniform number of users in both test and train split and to maintain the balance in the date, the most read book in the dataset was selected as the target book. Figure 2 shows the top 10 books from the dataset.

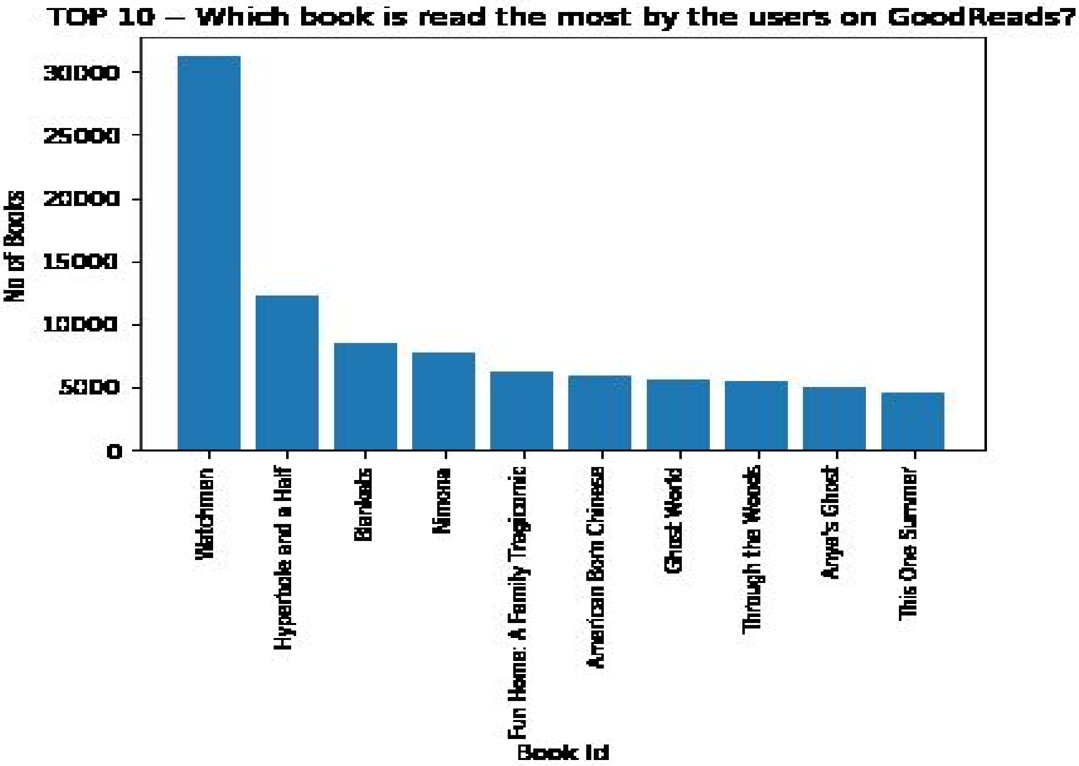


Figure 2: Most read books by users on Goodreads

9. Feature Engineering

The following are the features and output class of our final dataset after aggregating measures and performing Natural Language Processing techniques to find the similarities between the target book and the past books read by the user.

User Id	Total number of books read in Comics	Average Rating of books in Comics	The proportion of Similar books	Average similarity with target book	Target Class

*Aggregated Feature*s:

1. For each user, the number of books read.
2. Average Rating given to the books ready by the user.

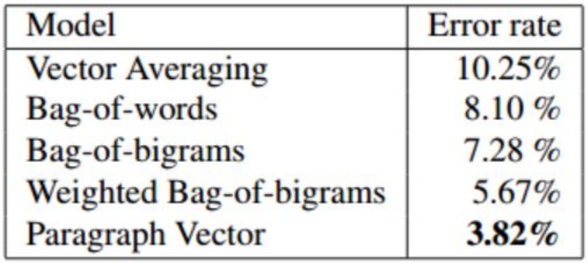
Natural Language Processing

1. Cleaning of Data:
   a. Removal of HTML tags.
   b. Removal of non-english alphabets

2. Stopword Removal
   A stop word is a commonly used word (such as “the”, “a”, “an”, “in”). These words are removed as they do not add value to the context of the document. This was done using NLTK library.

3. Lemmatization
   Breaking down similar extensions of one word to root form. We used WordNet lemmatizer to achieve the same. At a higher level, WordNet resembles something like a thesaurus and seems to work better as a lemmatizer.

4. Doc2Vec:
   Doc2vec is an NLP tool for representing documents as a vector and is a generalizing of the word2vec method. Doc2vec seems to perform better than other methodologies of aggregating word vectors to represent documents, as found in this study by Tom Mikilov (here)



Model	Error rate
Vector Averaging	10.25%
Bag-of-words	8.10 %
Bag-of-bigrams	7.28 %
Weighted Bag-of-bigrams	5.67%
Paragraph Vector	**3.82%**

Figure 3: Error rates for Doc2Vec models

5. Cosine Similarity
   Cosine similarity is an simple way to find similarity of the word documents. Once, the words are represented numerical vectors by Doc2Vec model. Cosine Similarity finds the angle between these vectors and gives a number between 0 to 1. 0 being highly dissimilar and 1 being exactly similar.

6. Deriving Proportion and Average Similarity after Thresholding

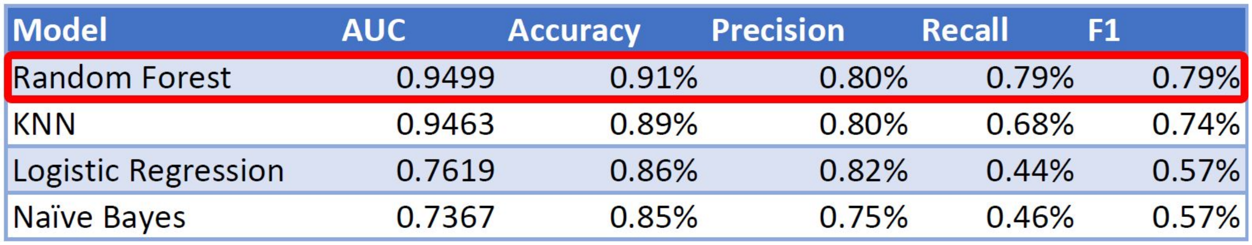
a. Threshold:
   A threshold similarity of 0.9 was imposed. Books above this threshold were considered similar (highly similar in context/type) to the target book.

b. Proportion:
   The proportion is derived by dividing the number of books similar to the target book read by the user with the total number of books read by the user, per user.

c. Average Similarity:
   The average similarity of all the books read by the user. This was done to take into account also those books which were not very similar to the books read by the user.

10. Modeling and Performance

Figure 4 shows the different models and their performance using cross-validation:



Model	AUC	Accuracy	Precision	Recall	F1
Random Forest	0.9499	0.91%	0.80%	0.79%	0.79%
KNN	0.9463	0.89%	0.80%	0.68%	0.74%
Logistic Regression	0.7619	0.86%	0.82%	0.44%	0.57%
Naïve Bayes	0.7367	0.85%	0.75%	0.46%	0.57%

Figure 4: Performances of all the models

Figure 5 depicts the Confusion Matrices of the different models:

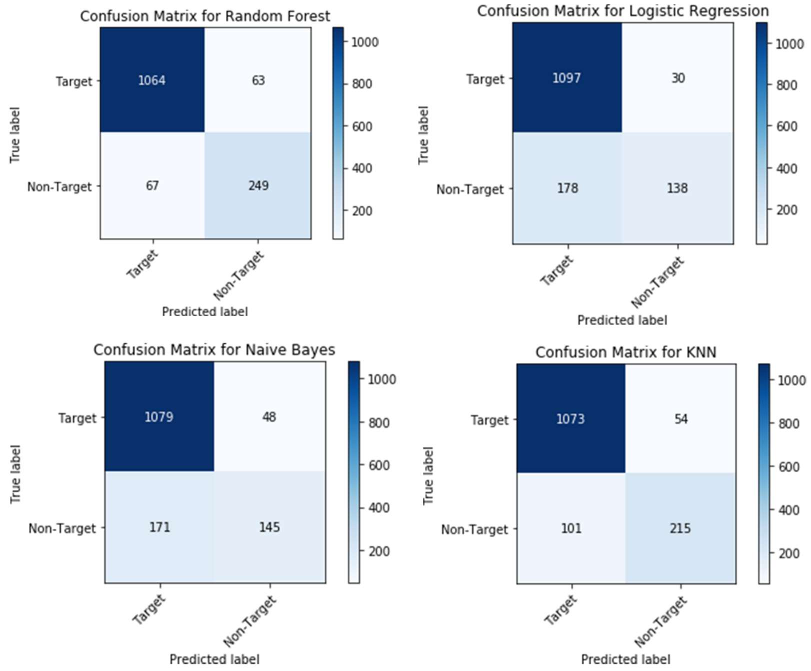


Figure 5: Confusion Matrices for all models

10.1 Hyperparameter Testing for KNN

With KNN, different values of k were applied, with Figure 6 depicting the ROC curves for them.

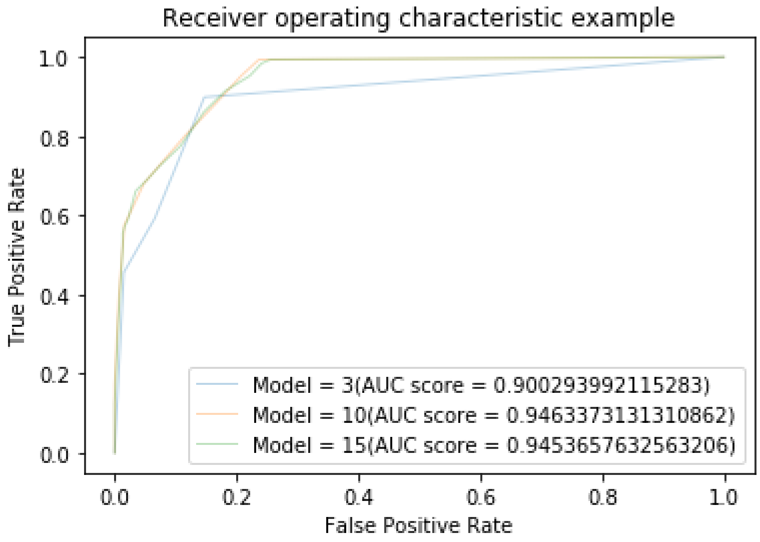


Figure 6: ROC for the different KNN models

Figure 6 shows that the best results are when **k=10, with AUC = 0.946**

11. Results and Interpretation

Figure 7 presents a comparison of the ROC curves:

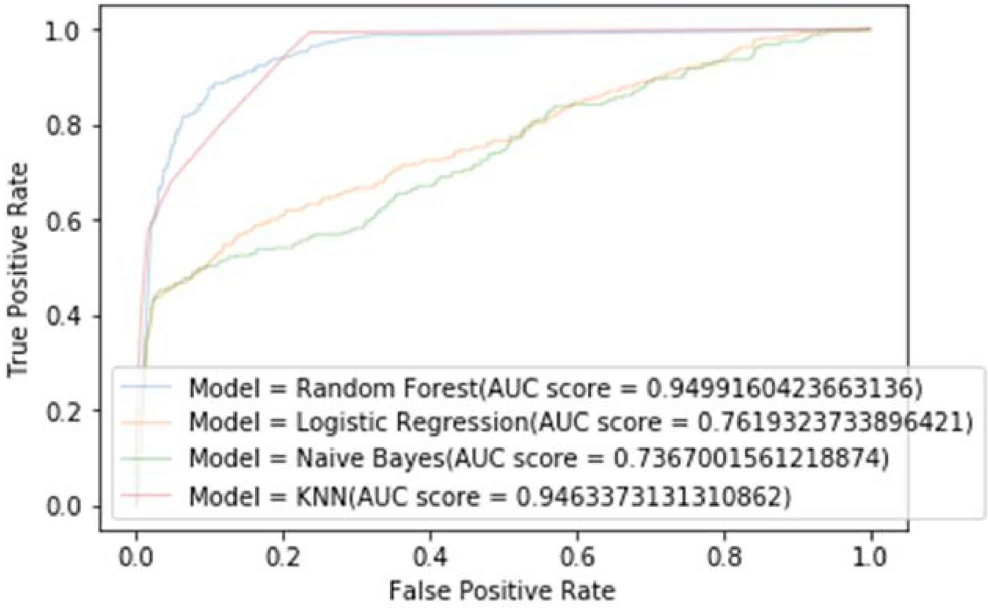


Figure 7: Comparing the different ROC Curves

From the ROC curves plotted for the four models, we can see that the Random Forest Classifier gives the most true positives for a fixed false positive. K- Nearest Neighbour is a close second with a comparable AUC score.

Feature Importance**:**

The use of forests of trees to evaluate the importance of features on our classification task:

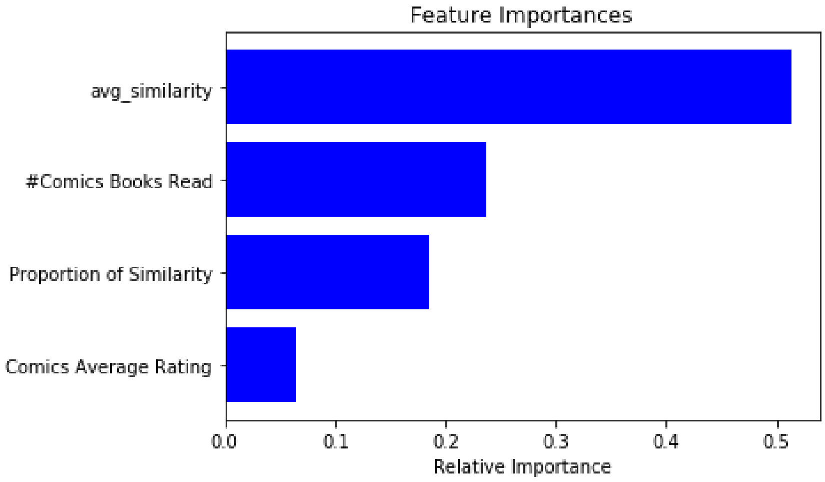


Figure 8: Relative importance of features

Figure 8 reveals that, as expected, average similarity has the highest feature importance in predicting the willingness of the user to read the book. The model also shows that it matters how many books the users have read in order to get accurate results.

12. Stakeholder Utility

The model will provide a binary output for each user about the willingness to read the book

13. Challenges

Following were the challenges:

1. A large amount of data, parsing, and processing was a major hurdle.
2. Lack of demographic data
3. Too many missing values in certain columns, as a result, had to drop them.

14. Risks and Mitigations

1. **Overfitting tends to be a major risk in any data science project**
   Necessary hyperparameter tuning for KNN has been done, due to the large amount to dimensionality ratio. The model has a good test accuracy as well as an AUC score, hence this risk has been mitigated, along with a 10 fold cross-validation
2. **Books that are in a series are removed except the first book, i.e. sequels are removed**
   Books that are a sequel to a previously published book have more tendency to be read and reviewed. For the scope of this project, this bias has been removed and hence this bias affecting the accuracy has been mitigated

15. Future Work

For the continuation of work, the popularity of the author can be taken into account. If the demographic of the user data is available, age groups can be targeted for marketing. It would also be useful to include other genres that see a consistent decline, for example, poetry, in order to boost sales. It would also be interesting to obtain and use timestamps and temporal information to formulate analysis about the trends in reading.