Manga Recommender

TABLE OF CONTENTS

01

Problem Statement

Problem Statement

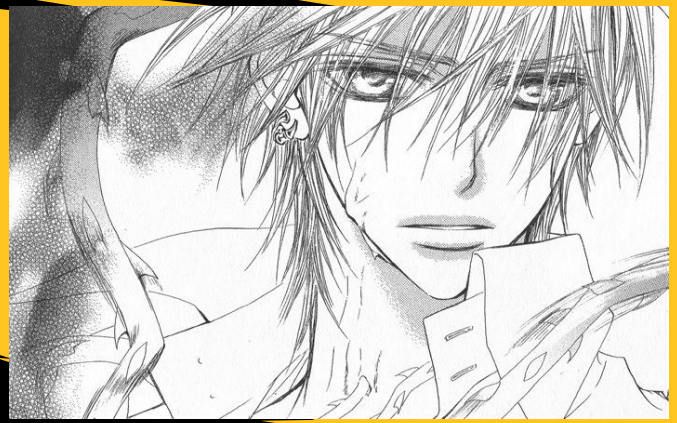
You are a data scientist working in a firm who has been engaged by a client who specialises in selling manga. In order to establish their presence online, they plan to engage your firm to help set up an online store. However, the owner is skeptical about the effectiveness of a recommender system and is unwilling to include it as part of the package. As instructed by your manager, you are tasked to create a manga recommender to showcase to the client that a recommender system can improve its sales online and should be included into the package.

02

What's Manga?

What's Manga?

Manga = Japan Originated Comic
Comes from the Kranji "漫画"
Usually coloured in black and white



03

Methodology

Methodology

Web Scraping

On myanimelist.net

Cleaning

Remove rows without scores
Fill in missing values for themes and demographic

EDA

Looking for useful information

Features Engineering

Creating new features:-
Manga length
Clusters

Modeling

Basic Similarity Metrics
Item Based Collaborative Filtering
LightFM

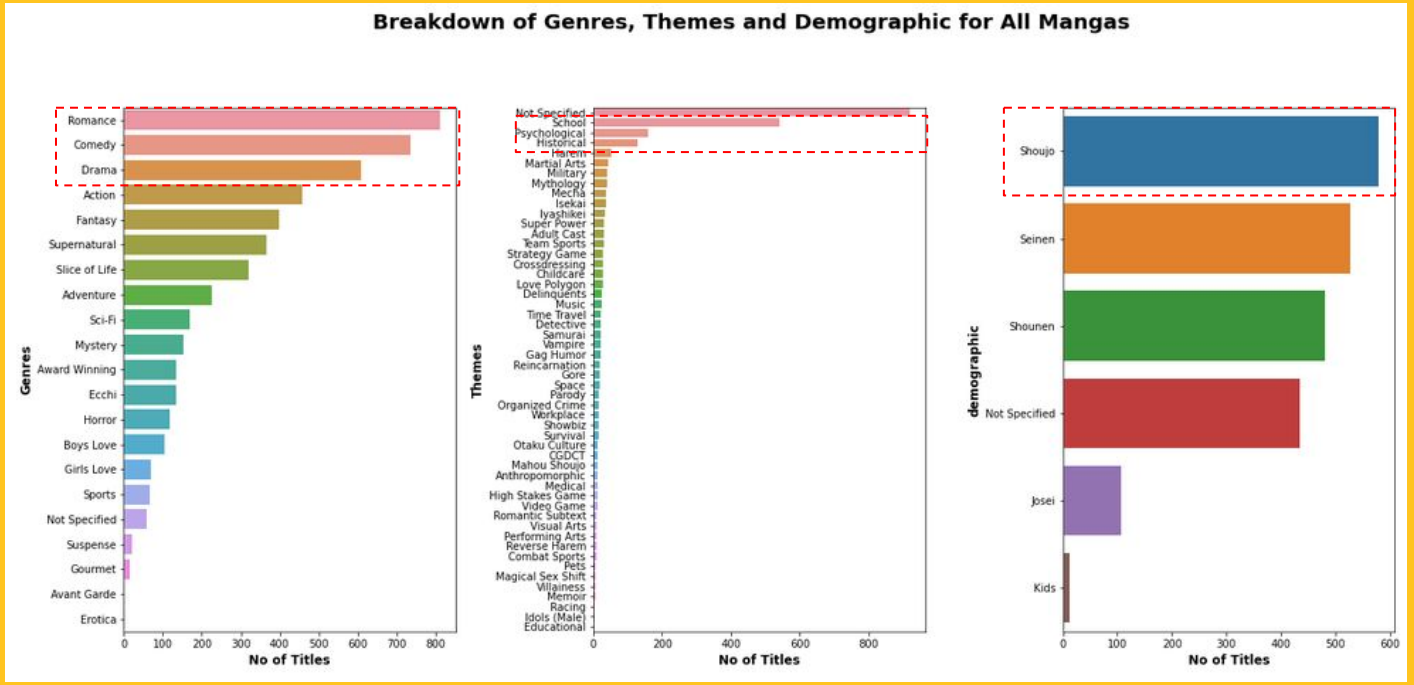
Final Model

What is behind the hood?

04

EDA

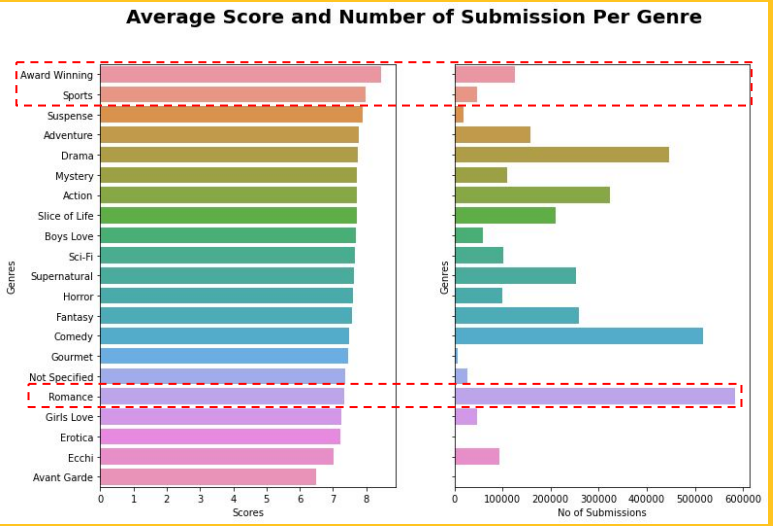
EDA - Types of Manga

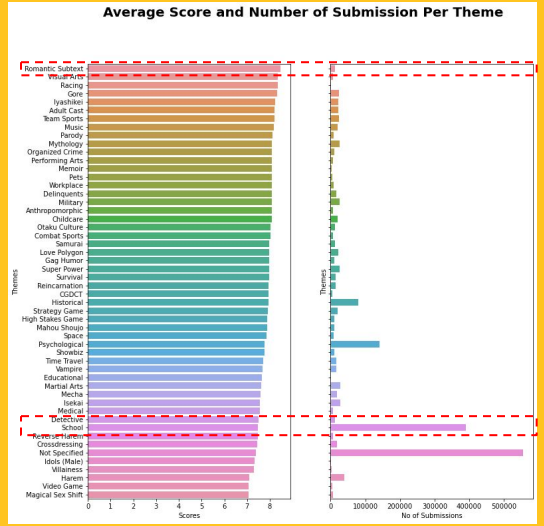


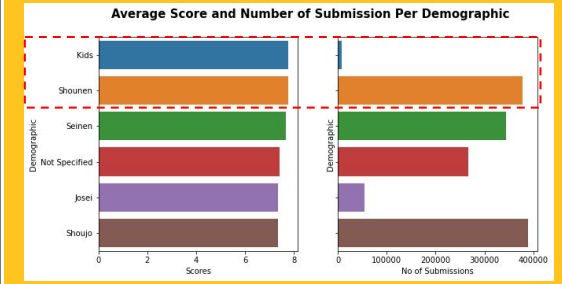
Term	Meaning
Shoujo	Adolescent girls and young adult women
Seinen	Young adult men
Shounen	Young teen male
Josei	Adult women
Kids	Young children

- Romance, comedy and drama are the most common genres
- School, psychological and historical are the most common themes
- Shoujo is the most targeted audience

EDA - Ratings







- Award winning and sports are the highest rated genres
  - Romance and comedy has the most readers
- Romantic subtext has the highest rated themes
  - School themes has the most readers
- Kids and shounen manga have highest rating
  - Kids manga has the lowest number of readers

05

Modeling

Metrics: How Do Grade Our Models

SUCCESS: At least 1 title which the reader read is in the top 10 recommendations

FAILURE: None of the titles which the reader read is in the top 10 recommendations

1.5584%

Null Mode (Train Dataset) Score

8.2798%

Null Mode (Test Dataset) Score

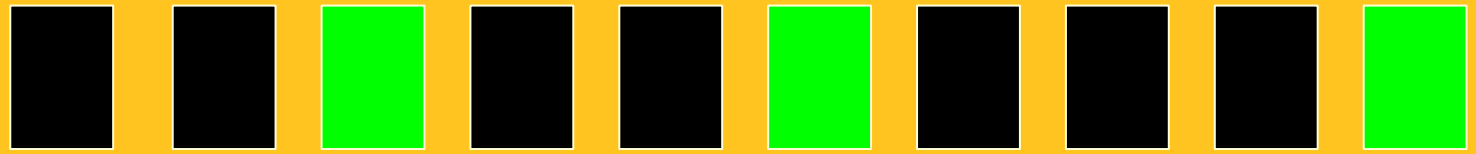
Hypergeometric Distribution

- Discrete probability distribution
- Describes the probability of k successes in n draws
  - Random draws for which the object drawn has a specified feature
- Without replacement
- From a finite population of size N that contains exactly K objects with that feature
- Where each draw is either a success or a failure

Back

Hypergeometric Distribution

Objective: Choose 3 books from the 10 books below

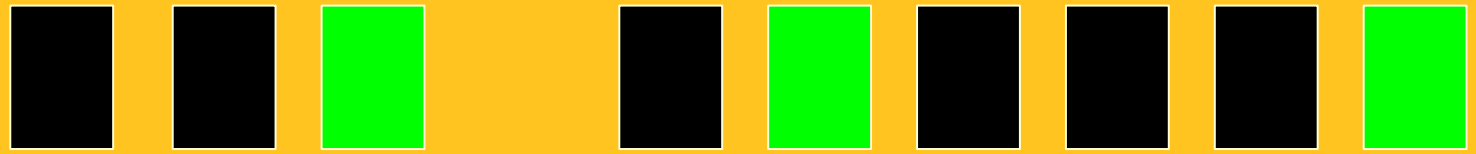


p(SUCESS) = 0.3

Back

Hypergeometric Distribution

Objective: Choose 3 books from the 10 books below



p(SUCESS) = 3/9 = 0.333

Back

Modeling - What Did We Use?

- Basic Similarity Metrics (BSM)
  - Used genres, themes, demographic as features
- Item-Based Collaborative Filtering (CFSM)
- LightFM

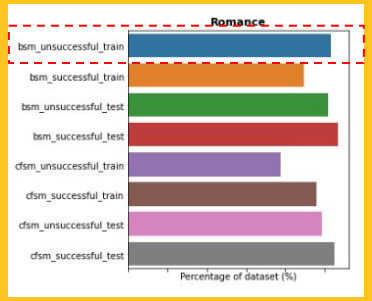
Modeling Results

Model	Train Accuracy	Test Accuracy	Train Precision@k	Test Precision@k
Null Model	0.015584	0.082798	-	-
Basic Similarity Matrix	0.823500	0.319338	0.13937	0.041654
Collaborative Filtering Similarity Matrix	0.519700	0.766412	0.16218	0.171908
LightFM – Basic Model	0.266600	0.776845	0.132022	0.182901

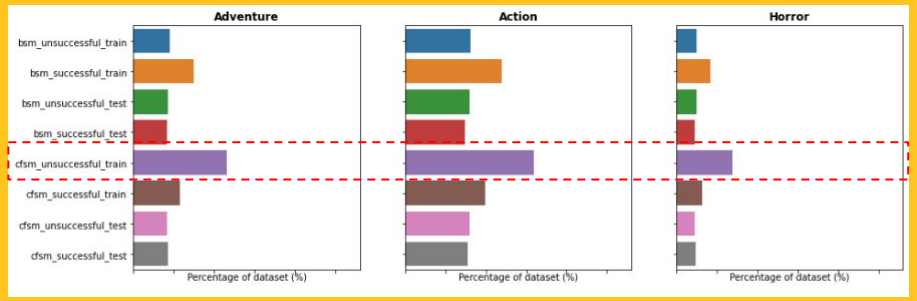
While LightFM had a better test accuracy than collaborative filtering, LightFM was not used in the final production model due to its lower train accuracy.

Why It Fails? - Genres

Basic Similarity Metrics (BSM)

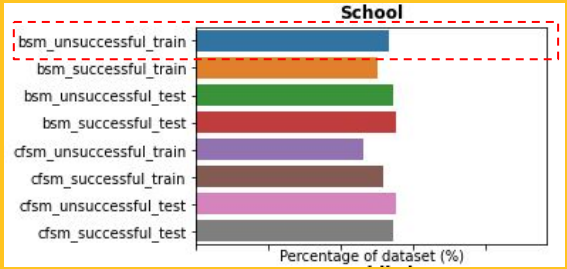


Item-Based Collaborative Filtering (CFSM)

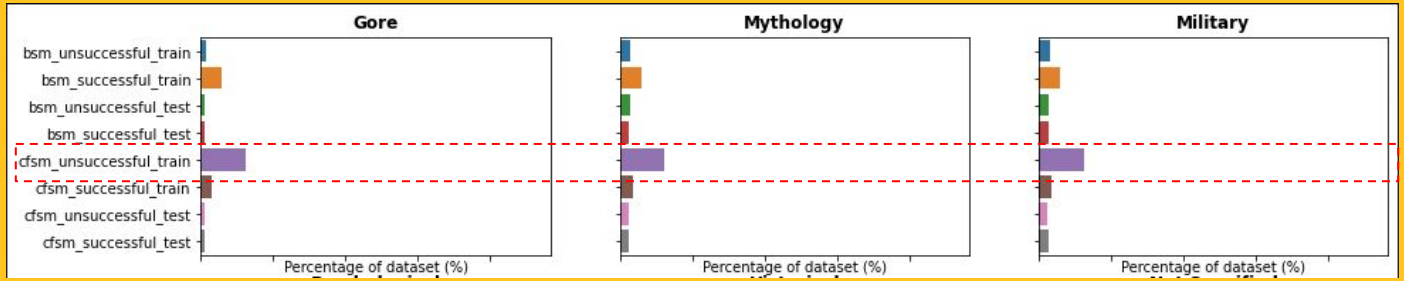


Why It Fails? - Themes

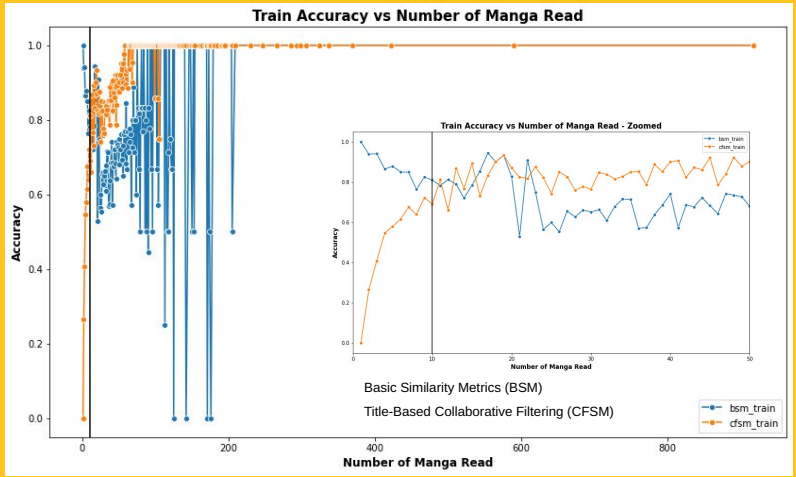
Basic Similarity Metrics (BSM)

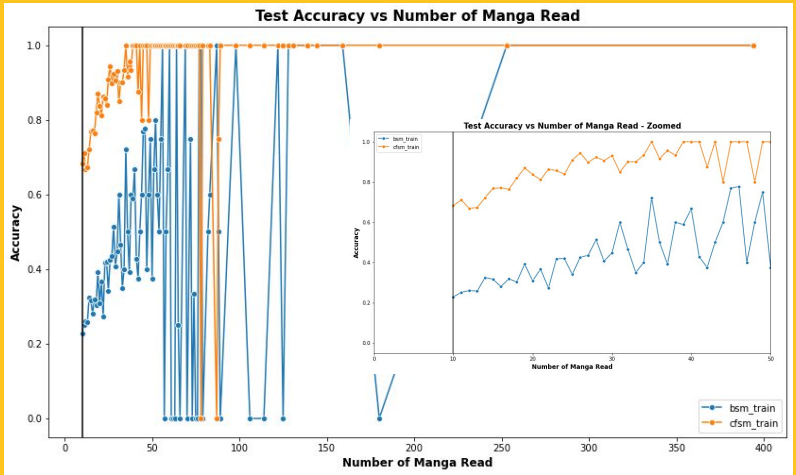


Item-Based Collaborative Filtering (CFSM)



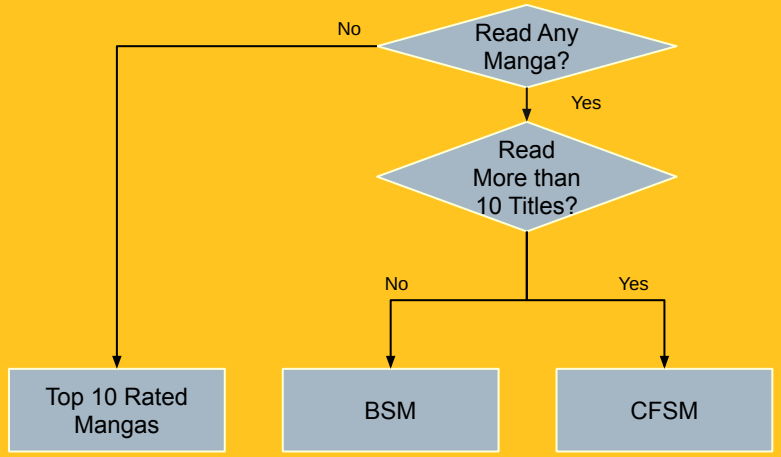
Hybrid Recommender - BSM & CFSM



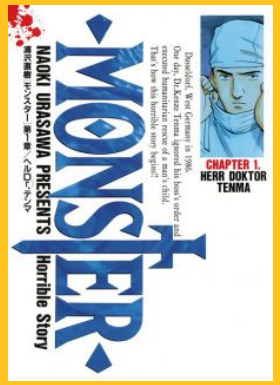


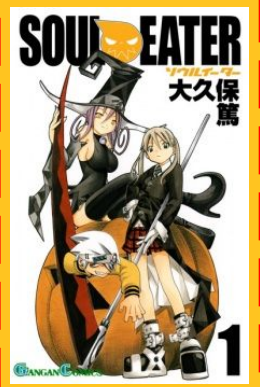
	BSM_Train	BSM_Test	CFSM_Train	CFSM_Test
light_reader	0.864972	0.298441	0.401923	0.749722
moderate_reader	0.660944	0.52988	0.857143	0.936255
heavy_readers	0.75964	0.574713	0.96144	0.965517

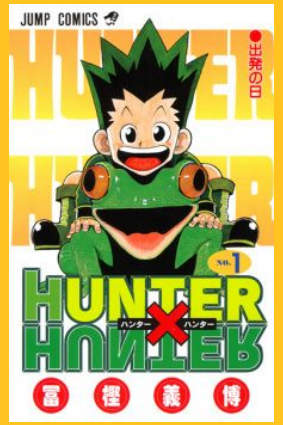
Hybrid Recommender - BSM & CFSM



Recommendations (Own List)







Business Case

Assumption: 10% of the reader eventually purchased the recommended book
Customer Per Month: 900,000
Price Per Book: $10.00
Model Recommendation Accuracy: 80%

Total Revenue Per Month:

With recommender system: $720,000

Without recommender system: $ 72,000

06

Conclusion

Possible Improvements

- Look into other recommender system
- Further optimizing the LightFM model by changing default parameters
- Combining different other types of recommender system
- Using neural net to decide the weightage of each model
- Input user preference when recommending
- Having more titles in the dataset

CONCLUSIONS

- Hybrid recommender is necessary as no one model can fit all cases
  - Basic similarity works well when the number of titles read is low
  - Collaborative filtering works better when the number of titles read by readers are high
- Having an recommender system greatly helps increase sales as it brings user to items that they would be interested in purchasing

End