EODP Presentation

How can different attributes from both books and users influence book ratings?

Kerui Huang
Peter Lu
Dylan Tran

Overview

- Brief introduction to Data and relevant background information
- Preprocessing
- Feature selection
- Model justification
- K Nearest Neighbours
- Decision Trees

Preprocessing

Methods:

- **RegEx** (Handled simple string cleaning)
- **Normalising to NFKD standard** (Handled diacritic marks)
- **Fuzzy Match** (Handled misspelt entries)
- **KNN imputation** (Handled Nan entries)

Common data irregularities:

	User-ID	User-City	User-State	User-Country	User-Age
0	8	timmins	ontario	canada"	NaN
1	9	germantown	tennessee	usa"	NaN
2	16	albuquerque	new mexico	usa"	NaN
3	17	chesapeake	virginia	usa"	NaN
4	19	weston		NaN	14"
5	26	bellevue	washington	usa"	NaN
6	32	portland	oregon	usa"	NaN
7	39	cary	north carolina	usa"	NaN
8	42	appleton	wisconsin	usa"	17
9	44	black mountain	north carolina	usa"	51
10	51	renton	washington	usa"	34
11	53	tacoma	washington	usa"	NaN
12	56	cheyenne	wyoming	usa"	24
13	69	vancouver	british columbia	canada"	NaN
14	73	wentzville	missouri	usa"	NaN
15	75	long beach	california	usa"	37
16	78	oakland	california	usa"	18
17	81	santa cruz	california	usa"	NaN
18	83	eugene	oregon	usa"	NaN
19	85	london	england	united kingdom"	41


```
Number of matches to make: 23
NEWYORK NEWPORT progress: 1
MONTREAL MOUNTPEARL progress: 2
LISBOA LISBON progress: 3
MÜNCHEN MÜNCHBERG progress: 4
STPAUL STPAULI progress: 5
WIEN WIES progress: 6
MILANO MIDLAND progress: 7
STCHARLES SAINTCHARLES progress: 8
FRANKFURT FRANKFORT progress: 9
PLOIESTI PLOIEŞTI progress: 10
NYC nan progress: 10
ACORUÑA CORUNNA progress: 11
ISTANBUL nan progress: 11
DUESSELDORF DÜSSELDORF progress: 12
VITORIA VICTORIA progress: 13
ZURICH AURICH progress: 14
GENOVA GÉNOVA progress: 15
HONGKONG HONGTANG progress: 16
BRICK BERWICK progress: 17
FIRENZE FORENZA progress: 18
SAOPAULO SANPABLO progress: 19
ALLSTON GALSTON progress: 20
MILILANI MAILANI progress: 21
Current unmatched series
...
```

Feature Selection

Performed mutual information analysis on each feature against the class label of Book-Rating

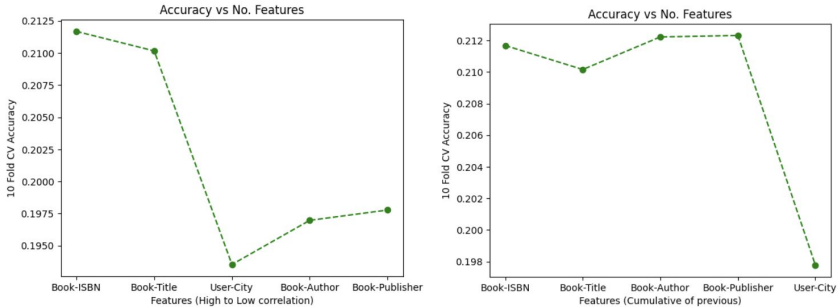
- Sort features with NMI
- Found top 5 features: ISBN, Book-Title, User-City, Book-Author and Book-Publisher, respectively.

However, during testing of the models, we found that the User-City feature drastically decreased performanced, and thus we removed it.

Figure 1. MI Correlations

```
User-ID                0.418
ISBN                   0.206
Book-Title             0.185
User-City              0.113
Book-Author            0.090
Book-Publisher         0.026
User-State             0.015
User-Country           0.007
User-Age               0.005
Year-Of-Publication    0.004
```

Figure 2. Feature vs Accuracy



Model Justification

Attempt to find correlation with Pearson correlation and Normalised Mutual Information

- Pearson correlation generally quite low (Figure 1)
- Normalised Mutual information relatively higher.

We decided to use KNN and Decision Trees

Figure 1. Pearson Correlation

	City	Author	ISBN	Title	Publisher	Age	Ratings
City	1.000000	-0.001908	0.001005	-0.002609	0.007628	0.002093	0.011552
Author	-0.001908	1.000000	0.013733	-0.009224	0.059569	-0.015715	-0.012048
ISBN	0.001005	0.013733	1.000000	0.002909	-0.005317	0.004842	-0.006982
Title	-0.002609	-0.009224	0.002909	1.000000	-0.017696	-0.000350	-0.016328
Publisher	0.007628	0.059569	-0.005317	-0.017696	1.000000	0.013759	-0.025908
Age	0.002093	-0.015715	0.004842	-0.000350	0.013759	1.000000	-0.014507
Ratings	0.011552	-0.012048	-0.006982	-0.016328	-0.025908	-0.014507	1.000000

Figure 2. Mutual Information against Book-Rating

```
User-City 0.113
User-State 0.015
User-Country 0.007
User-Age 0.005
User-Age-Binned 0.001
User-ID 0.418

ISBN 0.206
Book-Title 0.185
Book-Author 0.09
Year-Of-Publication 0.004
Book-Publisher 0.026
```

K Nearest Neighbours

- Selecting K Hyperparameter

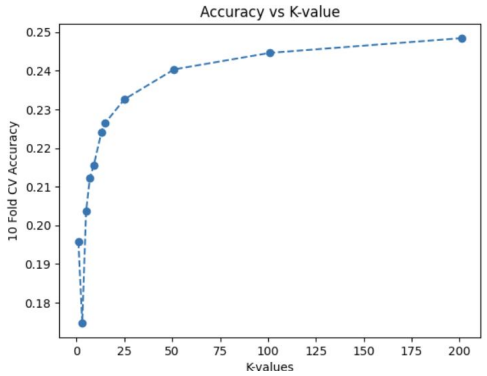


Figure 1. K selection for default category KNN model

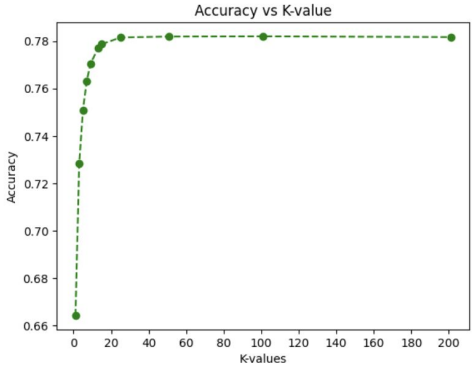


Figure 2. K selection for grouped category KNN model

- Model Evaluation

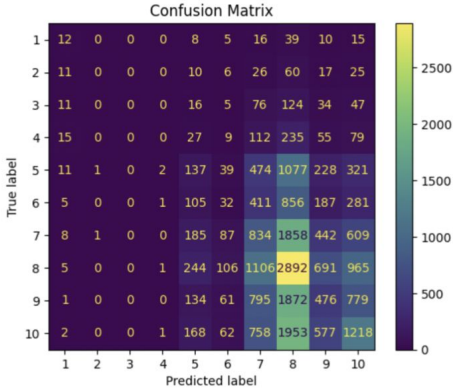


Figure 3. Confusion matrix for default category KNN model

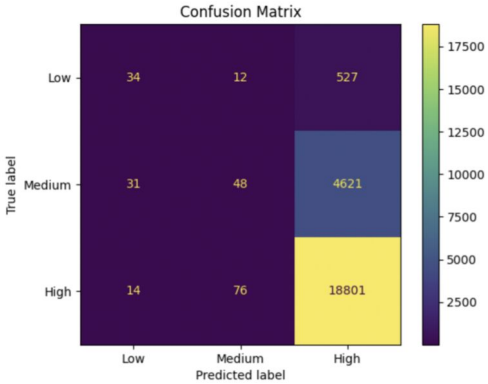


Figure 4. Confusion matrix for default category KNN model

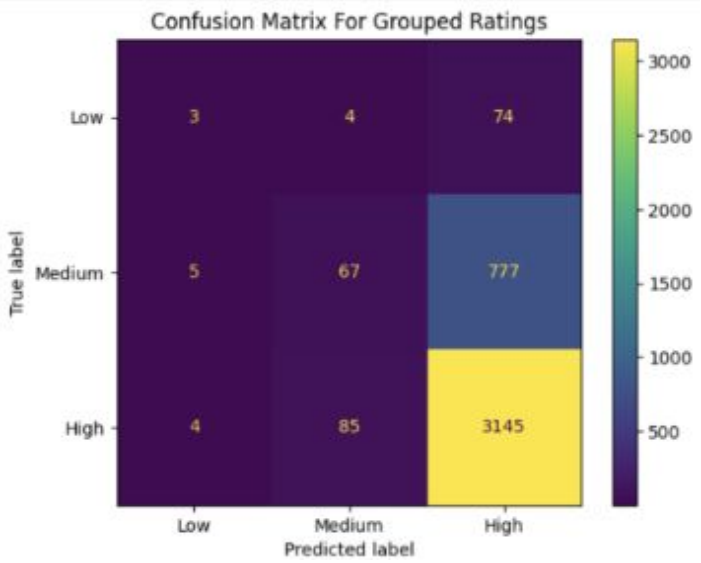
Metrics:

Default category model
accuracy: 23.2%
base line: 45.4%
Grouped category model
accuracy: 78.1%
Baseline: 99.1%

Decision Trees

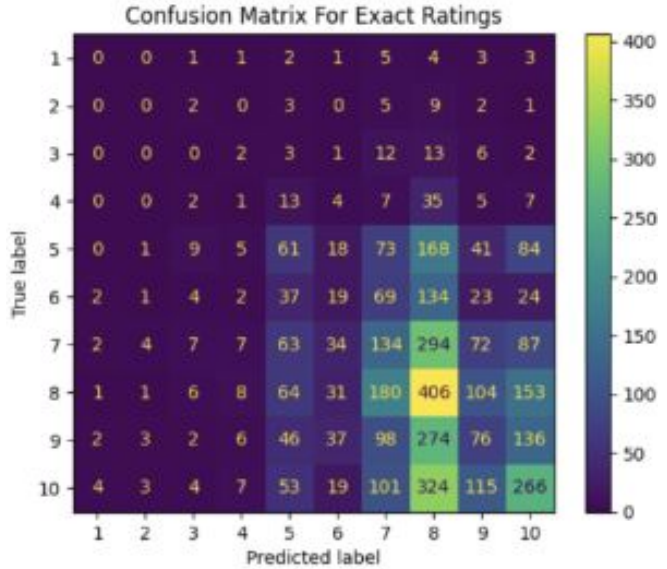
Confusion Matrix of Grouped DT Model

- Accuracy = 0.77
- Recall = 0.77
- Precision = 0.70
- F1 = 0.70

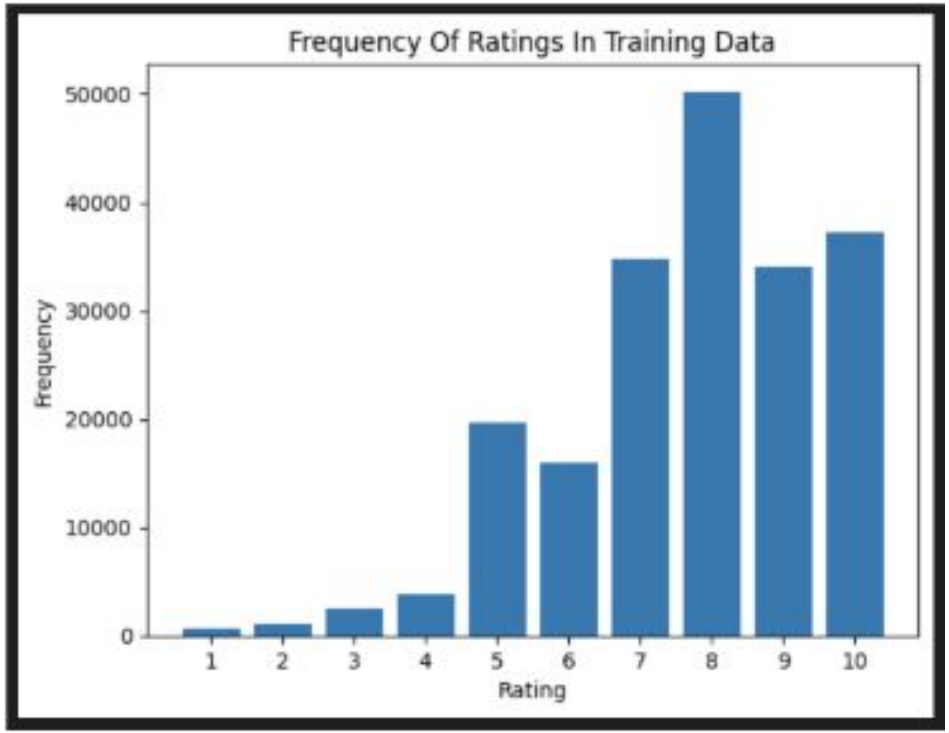


Confusion Matrix of Exact DT Model

- Accuracy = 0.23
- Recall = 0.23
- Precision = 0.22
- F1 = 0.22



Limitations, Improvements and Conclusions



Thank You !!