**CS 839 Data Science Project Stage 2 Report**

Sreejita Dutta, Deepanshu Gera, Rahul Jayan

1. **Data Description**

We chose books as the entity. We scraped this data from two web sources Walmart.com and amazon.com. The scraped table consists of the following attributes:

1. Name/Title of the book
2. Sale Price
3. Category/Genre
4. Author
5. ISBN
6. Pages
7. Publisher
8. Language
9. Dimension of the book
10. Weight of the book
11. User Rating

Though we have retained the attribute ISBN in the dataset, we have not used that attribute for the entity matching.

amazon\_books.csv consists of **3212 tuples** and walmart\_books.csv consists of **3114 tuples**.

1. **Blocker**

We have used all three blockers for our blocking stage.

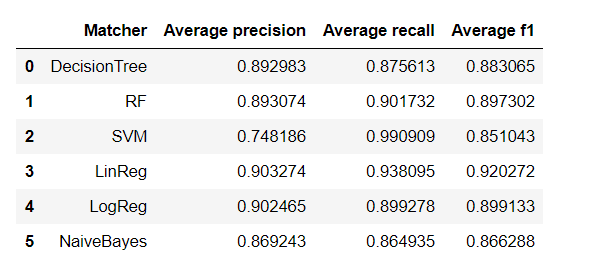
1. Black Box Blocker with specific rules for name similarity
2. Word level Overlap Blocker on attribute ‘Author’ with overlap size 2
3. Word level Overlap Blocker on attribute ‘Name’ with overlap size 2
4. Rule based blocker with absolute norm compute on attribute ‘Pages’
5. Rule based blocker with absolute norm compute on attribute ‘Sales\_Price’

We obtained **803** tuple pairs after this blocking stage. We labelled **300** of these pairs for our learning stage

1. **Results**

The results obtained on performing cross validation for the first time on set I is compiled in the table given below.

First Time:

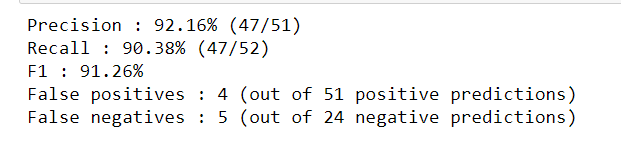


After this iteration we selected Random Forest as the matcher since Magellan provides debugger support for only Decision Tree and Random Forest matchers. Even though Linear Regression and Logistic regression matchers had already reached 90% precision, we suspected that their precision might suffer a lot more once test set is applied.

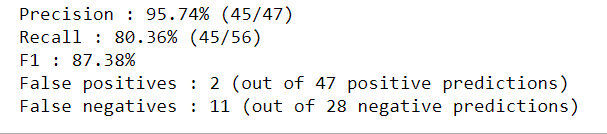
Iteration 1 RF matcher

We decided to debug RF matcher for this iteration. On debugging we realized that the most common issues for false positives and negatives were incorrect labelling. Using the debugger we corrected many labels. This was the result after running RF matcher after splitting H into U and V.

The results after debugging on different splits of the training data is given below.

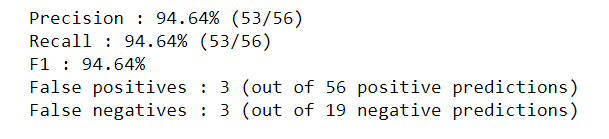


Iteration 2: RF matcher

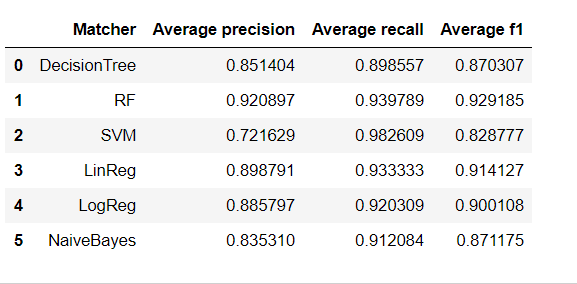


The only thing we checked is if the labels are proper or not. Since we were imputing the missing value fields with 0 certain entries were infact misclassified.

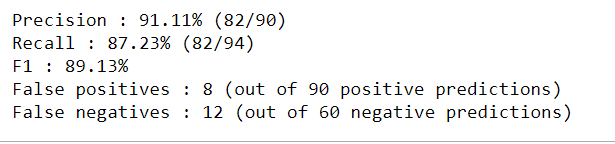
Iteration 3: RF matcher:



Once we achieved this level of precision we ran matcher on train dataset and performed CV for one more time. This was the result we obtained.

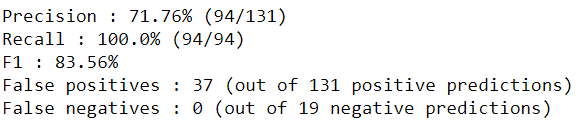


From these results we selected Random Forests matcher as the best one and we ran it on test dataset J. The results obtained on test dataset J is given below.

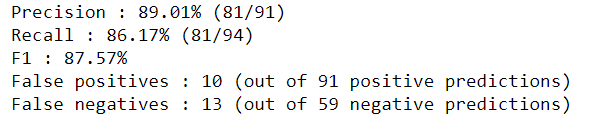


Result on test dataset for other 5 classifiers are given below

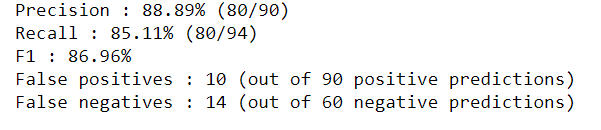
SVM



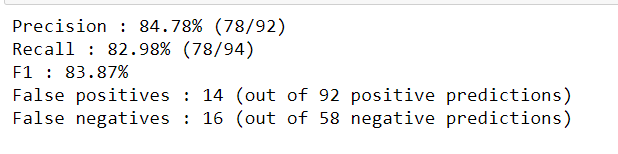
Decision Tree



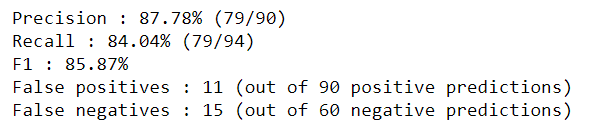
Logistic Regression



Naïve Bayes



Linear Regression



1. **Time Estimates**

The time estimates are given below.

1. Blocking:
2. Labelling:
3. Finding best matcher:
4. **Discussion**

Magellan Good points