Record Linkage

Arwa Ashi

09 March, 2022

Contents

Table 1: Example: First Rows of 'RLdata500' Dataset

fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd
CARSTEN	NA	MEIER	NA	1949	7	22
GERD	NA	BAUER	NA	1968	7	27
ROBERT	NA	HARTMANN	NA	1930	4	30
STEFAN	NA	WOLFF	NA	1957	9	2
RALF	NA	KRUEGER	NA	1966	1	13
JUERGEN	NA	FRANKE	NA	1929	7	4

1 Introduction

Digital transformation after COVID 19, increase the data collecting for public and privet sector. If the data linked in a proper way, that would improve the provide service and client experience. This report represent a several data linkage methodologies. There are three stages: Pre-linkage (see data preparation section), Linkage (see methodology section), and Post-linkage. The report will be divided into data, data exploration, data preparation, methodology, result, and conclusion sections.

2 Data

Finding an available dataset for data linkage ptoject is not easy. Based on that the 'RLdata500' and 'RLdata10000' datasets under the package 'RecordLinkage' are considered.

3 Data Exploration

the data contain the first name, last name and date of birth for individuals. Notice that the data field can be different slightly, for example two records refer to the same entity i.e. peter can have a slight change in his last name or his date of birth. The 'RLdata500' and 'RLdata10000' have 7 columns for each, and 500 and 10000 rows consequently. The first name as 'fname_c1' and last name as 'lname_c1' are separated into two columns and date of birth is separated into several columns for year as 'by', month as 'bm', and day as 'bd', See Table 1.

The dataset assumed that it marge individual information from different databases. Based on that the total duplication in 'RLdata500' and 'RLdata10000' are and consequently.

4 Data Preparation

First step is pre-linkage stage which is to prepare the data for evaluation by generate the features that will be used in the models. In order to do that, a 'compare.depdup' function under the 'RecordLinkage' package is used to generate the feature. The generated feature (pair) compare two ids in each row. See Table 2. The number 1 and 0 mean perfect match or no match consequently. If the number is less than 0, then it means that it is a float number for a string comparison. The final column indicates if there is a match or not.

5 Methodology

The linkage stage that has the objective of matching the records in each 'RLdata500' and 'RLdata10000' datasets with no common unique identifiers and deduplicating with a dataset. There will be two methodologies: a probabilistic method and machine learning method.

The preprocessing stage was done in data preparation section by developing link keys by using blocking as 'blockfld' function under 'compare.dedup' function.

Table 2: Example: Generated Pairs of 'RLdata500' Dataset

id1	id2	fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd	is_match
1	174	1	NA	0.1428571	NA	0	0	0	NA
1	204	1	NA	0.0000000	NA	0	0	0	NA
2	7	1	NA	0.3750000	NA	0	0	0	NA
2	43	1	NA	0.8333333	NA	1	1	1	NA
2	169	1	NA	0.0000000	NA	0	0	0	NA
4	19	1	NA	0.1428571	NA	0	0	0	NA

Table 3: Example: Initial Matched. of 'RLdata500' Dataset

id	fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd	Weight
313	URSULA	BIRGIT	MUELLRR	NA	1940	6	15	
457	URSULA	BIRGIT	MUELLER	NA	1940	6	15	35.608887
467	ULRIKE	NICOLE	BECKRR	NA	1982	8	4	
472	ULRIKE	NICOLE	BECKER	NA	1982	8	4	35.568245

5.1 Probabilistic Method

There are a deterministic method that is a direct match by comparing everything needs to match, and a probabilistic method that is to estimate a probability or liklihood for two records. The focusing here is by using a probabilistic matching. For the classification, a Fellegi-Sunter Model is considered. Define a cut off for string comparing at 80% by using EM algorithm as 'emWeights' function in 'RecordLinkage' package. For a summary of weights for 'RLdata500', see the following:

```
##
## Deduplication Data Set
##
## 500 records
## 1221 record pairs
##
## 0 matches
## 0 non-matches
## 1221 pairs with unknown status
##
##
## Weight distribution:
##
## [-15,-10]  (-10,-5]    (-5,0]     (0,5]    (5,10]   (10,15]   (15,20]   (20,25]
##      1006       119        34        10         0        18        12         4
##   (25,30]   (30,35]   (35,40]
##        16         0         2
```

See Table 3 for initial matched.

The initial matches is used as base to determine threshold. For 'RLdata500', the threshold is 30. See table 4 for final pairs.

Table 4: Example: Final Matched. of ’RLdata500’ Dataset

id	fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd	Weight
48	WERNER	NA	KOERTIG	NA	1965	11	28	
238	WERNIER	NA	KOERTIG	NA	1965	11	28	29.850470
68	PETEVR	NA	FUCHS	NA	1972	9	12	
190	PETER	NA	FUCHS	NA	1972	9	12	29.850470
85	THORSKTEN	NA	MARTIN	NA	1995	11	15	
187	THORSTEN	NA	MARTIN	NA	1995	11	15	29.850470
158	PETER	NA	BECKER	NA	1960	9	5	
229	PETERS	NA	BECKER	NA	1960	9	5	29.850470
177	JOHANNNES	NA	SCHULZ	NA	1974	1	17	
207	JOHANNES	NA	SCHULZ	NA	1974	1	17	29.850470
209	ROLBF	NA	NEUMANN	NA	1967	3	29	
227	ROLF	NA	NEUMANN	NA	1967	3	29	29.850470
265	MARIANNFE	NA	MOELLER	NA	1961	9	17	
456	MARIANNE	NA	MOELLER	NA	1961	9	17	29.850470
266	KARIN	NA	HORN	NA	2002	6	4	
437	KARINW	NA	HORN	NA	2002	6	4	29.850470
298	SONJA	NA	FISCHER	NA	1989	7	17	
464	SONJAD	NA	FISCHER	NA	1989	7	17	29.850470
310	MONIKA	NA	SCHNEIDER	NA	1937	6	2	
432	MONIYKA	NA	SCHNEIDER	NA	1937	6	2	29.850470
377	SABAINE	NA	OTTO	NA	1940	7	23	
448	SABINE	NA	OTTO	NA	1940	7	23	29.850470
391	GABRIELE	NA	BECKER	NA	1990	3	27	
496	GABRIHELE	NA	BECKER	NA	1990	3	27	29.850470
395	GISOELA	NA	BECK	NA	2003	4	16	
404	GISELA	NA	BECK	NA	2003	4	16	29.850470
402	CHRISTA	NA	SCHWARZ	NA	1965	7	13	
462	CHRISTAH	NA	SCHWARZ	NA	1965	7	13	29.850470
37	HARTMHUT	NA	HOFFMSNN	NA	1929	12	29	
72	HARTMUT	NA	HOFFMANN	NA	1929	12	29	29.657824
290	HELGA	ELFRIEDE	BERGER	NA	1989	1	18	
466	HELGA	ELFRIEDE	BERGER	NA	1989	1	28	28.768566

5.2 Machine Learning Method

For the machine learning approach ((logistic regression)), a 'reclin2' packages is considered for preparing the data for the algorithm. First, creating a pari by blocking fields by using 'pair_blocking' function in 'reclin2' package. Second, comparing the pairs to get comparing score for each feature by using 'compare_pairs' function in 'reclin2' package. Third, preparing the binary parameters 'TRUE' and 'FALS' by using 'compare_vars' function in 'reclin2' package. Fourth, using 'glm' function with a family = binomial() for the logistic regression.fifth, predict the matching probability. Sixth, selecting a matching probability that is greater than 50%. Seventh, generating a FALSE TRUE table for the result evaluation. Finally, generate the Final matching pairs. See table 5.

```
## 
##         TRUE
##   TRUE  600
```

Table 5: Example: Final Matched. of 'RLdata500' Dataset

.y	.x	fname_c1.x	fname_c2.x	lname_c1.x	lname_c2.x	by.x	bm.x	bd.x	id.x	fname_c1.y	fname_c2.y	lname_c1.y	lname_c2.y	by.y	bm.y	bd.y	id.y
1	1	CARSTEN	NA	MEIER	NA	1949	7	22	34	CARSTEN	NA	MEIER	NA	1949	7	22	34
2	2	GERD	NA	BAUER	NA	1968	7	27	51	GERD	NA	BAUER	NA	1968	7	27	51
2	43	GERD	NA	BAUERH	NA	1968	7	27	51	GERD	NA	BAUER	NA	1968	7	27	51
3	3	ROBERT	NA	HARTMANN	NA	1930	4	30	115	ROBERT	NA	HARTMANN	NA	1930	4	30	115
4	4	STEFAN	NA	WOLFF	NA	1957	9	2	189	STEFAN	NA	WOLFF	NA	1957	9	2	189
5	5	RALF	NA	KRUEGER	NA	1966	1	13	72	RALF	NA	KRUEGER	NA	1966	1	13	72
6	6	JUERGEN	NA	FRANKE	NA	1929	7	4	142	JUERGEN	NA	FRANKE	NA	1929	7	4	142
7	7	GERD	NA	SCHAEFER	NA	1967	8	1	162	GERD	NA	SCHAEFER	NA	1967	8	1	162
8	8	UWE	NA	MEIER	NA	1942	9	20	48	UWE	NA	MEIER	NA	1942	9	20	48
9	9	DANIEL	NA	SCHMIDT	NA	1978	3	4	133	DANIEL	NA	SCHMIDT	NA	1978	3	4	133
10	10	MICHAEL	NA	HAHN	NA	1971	2	27	190	MICHAEL	NA	HAHN	NA	1971	2	27	190
11	11	PETER	NA	JUNG	NA	1988	1	9	178	PETER	NA	JUNG	NA	1988	1	9	178
12	12	MANFRED	NA	HOFFMANN	NA	1933	8	25	217	MANFRED	NA	HOFFMANN	NA	1933	8	25	217
13	13	MICHAEL	NA	FISCHER	NA	1951	7	1	175	MICHAEL	NA	FISCHER	NA	1951	7	1	175
14	14	MANFRED	NA	BECKER	NA	1973	7	27	197	MANFRED	NA	BECKER	NA	1973	7	27	197
15	15	WALTER	NA	SCHNEIDER	NA	1953	8	26	44	WALTER	NA	SCHNEIDER	NA	1953	8	26	44
16	16	MARTIN	NA	SCHROEDER	NA	1988	2	3	84	MARTIN	NA	SCHROEDER	NA	1988	2	3	84
17	17	ALEXANDER	NA	MUELLER	NA	1974	9	9	35	ALEXANDER	NA	MUELLER	NA	1974	9	9	35
18	18	HANS	NA	SCHAEFER	NA	2003	6	22	88	HANS	NA	SCHAEFER	NA	2003	6	22	88
19	19	STEFAN	NA	MUELLER	NA	1949	8	13	77	STEFAN	NA	MUELLER	NA	1949	8	13	77
20	20	GERHARD	NA	SCHAEFER	NA	1964	4	29	91	GERHARD	NA	SCHAEFER	NA	1964	4	29	91
21	21	DENNIS	NA	SCHAEFER	NA	1956	4	11	90	DENNIS	NA	SCHAEFER	NA	1956	4	11	90
22	22	THORSTEN	NA	KLEIN	NA	1966	9	20	171	THORSTEN	NA	KLEIN	NA	1966	9	20	171
23	23	PETER	NA	BRANDT	NA	1997	4	1	61	PETER	NA	BRANDT	NA	1997	4	1	61
24	24	WALTER	NA	FISCHER	NA	1997	4	7	26	WALTER	NA	FISCHER	NA	1997	4	7	26
25	25	MATTHIAS	NA	HAAS	NA	1955	7	8	33	MATTHIAS	NA	HAAS	NA	1955	7	8	33
25	107	MATTHIAS	NA	HAAS	NA	1955	8	8	33	MATTHIAS	NA	HAAS	NA	1955	7	8	33
26	26	WOLFGANG	NA	WOLF	NA	1961	12	28	38	WOLFGANG	NA	WOLF	NA	1961	12	28	38
27	27	BENJAMIN	NA	MUELLER	NA	1997	4	19	195	BENJAMIN	NA	MUELLER	NA	1997	4	19	195
28	28	JAN	JUERGEN	SCHAEFER	NA	1946	5	25	199	JAN	JUERGEN	SCHAEFER	NA	1946	5	25	199
29	29	PETER	NA	WINTER	NA	1958	8	18	216	PETER	NA	WINTER	NA	1958	8	18	216
30	30	SVEN	NA	BRAUN	NA	1994	1	22	105	SVEN	NA	BRAUN	NA	1994	1	22	105
31	31	WOLFGANG	NA	LEHMANN	NA	1971	1	3	71	WOLFGANG	NA	LEHMANN	NA	1971	1	3	71
32	32	CHRISTIAN	JENS	SCHULZ	NA	2008	5	18	46	CHRISTIAN	JENS	SCHULZ	NA	2008	5	18	46
33	33	ANDREAS	FRANK	SCHUMACHER	NA	1928	7	7	118	ANDREAS	FRANK	SCHUMACHER	NA	1928	7	7	118
34	34	HEINZ	NA	BOEHM	NA	1938	12	20	27	HEINZ	NA	BOEHM	NA	1938	12	20	27
34	111	HEINZ	NA	BOEHMR	NA	1938	12	20	27	HEINZ	NA	BOEHM	NA	1938	12	20	27
35	35	HARALD	NA	BECKER	NA	1951	2	24	220	HARALD	NA	BECKER	NA	1951	2	24	220
36	36	KURT	NA	SCHMIDT	NA	1998	4	27	59	KURT	NA	SCHMIDT	NA	1998	4	27	59
37	37	HARTMHUT	NA	HOFFMSNN	NA	1929	12	29	139	HARTMHUT	NA	HOFFMSNN	NA	1929	12	29	139
37	72	HARTMUT	NA	HOFFMANN	NA	1929	12	29	139	HARTMHUT	NA	HOFFMSNN	NA	1929	12	29	139
38	38	HORST	NA	ENGEL	NA	1999	2	27	224	HORST	NA	ENGEL	NA	1999	2	27	224
39	39	DIETER	NA	NEUMANN	NA	1994	6	20	211	DIETER	NA	NEUMANN	NA	1994	6	20	211
40	40	WOLFGANG	NA	SCHMITT	NA	1986	11	29	47	WOLFGANG	NA	SCHMITT	NA	1986	11	29	47
41	41	SVEN	NA	SCHUMACHER	NA	2008	2	27	163	SVEN	NA	SCHUMACHER	NA	2008	2	27	163
42	42	DIETER	NA	WEISS	NA	1945	1	10	19	DIETER	NA	WEISS	NA	1945	1	10	19
43	2	GERD	NA	BAUER	NA	1968	7	27	51	GERD	NA	BAUERH	NA	1968	7	27	51
43	43	GERD	NA	BAUERH	NA	1968	7	27	51	GERD	NA	BAUERH	NA	1968	7	27	51

6 Result

Table 4 and 5 have the matching pairs' results by using a probabilistic methodology and logistic regression consequently.

7 Conclusion

The probabilistic and machine learning approaches for records linkages are working and matched the records. For the future work, a big dataset would be considered that have more features to have the ability to evaluate the performance of each approach and consider more approaches and assumptions.