Project 1: DMP system

Mukukenova Victoria
Prutko Alexander

NewProLab, 2020
Big Data 13.0

Contents

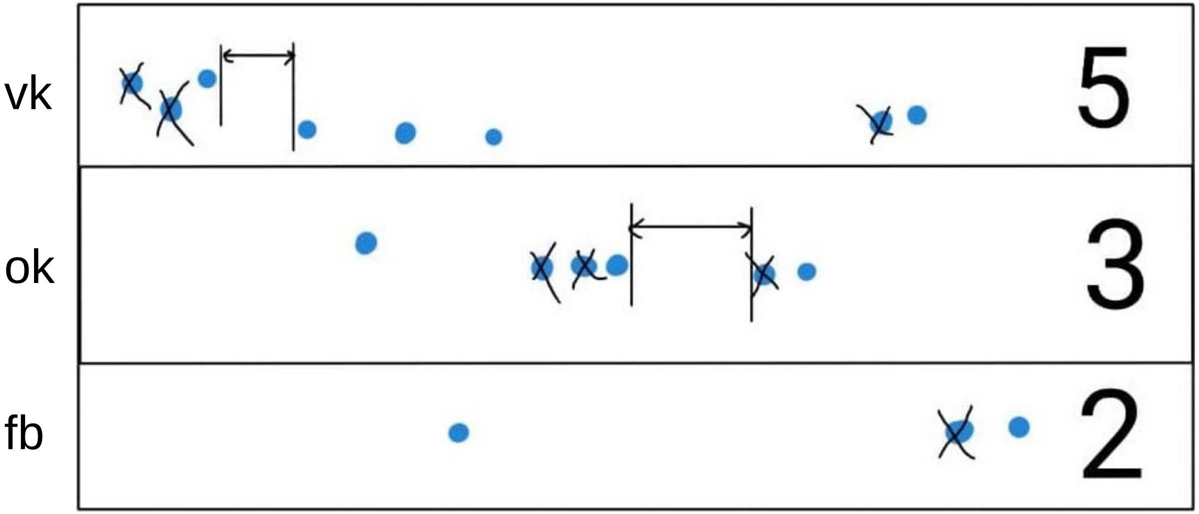
- Features
  - Feature cleaning
  - Feature engineering
- Model / Conclusion

Features: feature cleaning

- url → domain
- unique visitor counts
- remove rare domains
- remove common domains (tf-idf stop words)

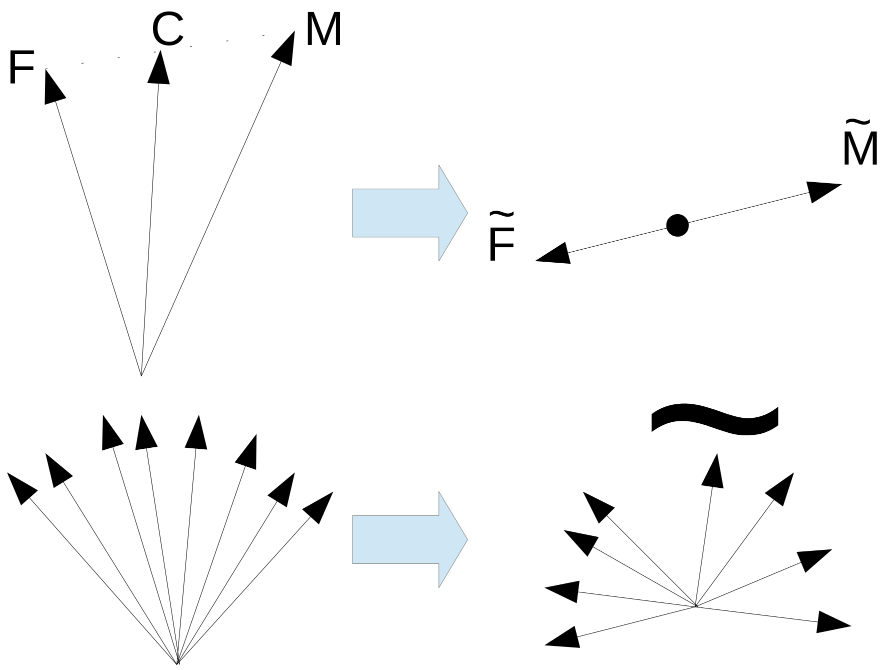
Features: timestamps

- timestamp → [day of week, weekend, day, hour] visit counts
- timestamp → visit sessions

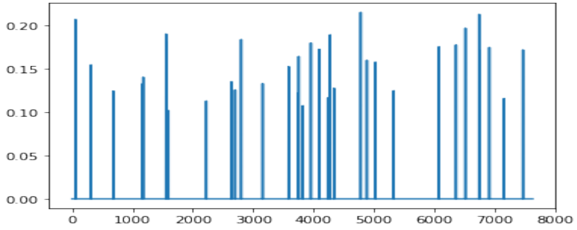


Features: tf-idf and vectors

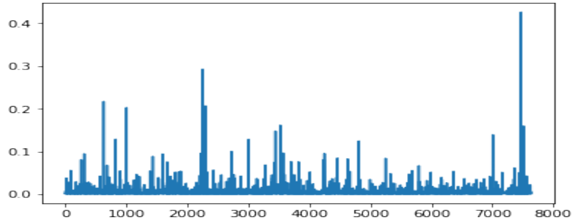
- domain string → tf-idf vector
- mean {Female, Male} vector
- mean Common vector
- subtraction of Common vector
- cosine similarity between object vectors and F,M-vectors
- common domains from common vector → stop words
- the same with age vectors and category vectors



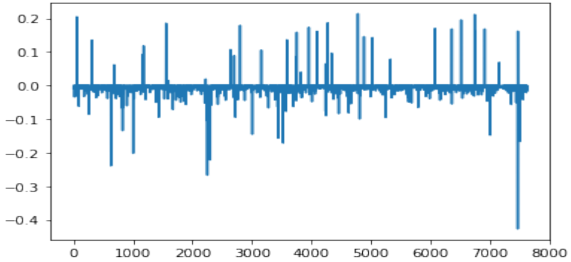
Features: cosine similarity



vector



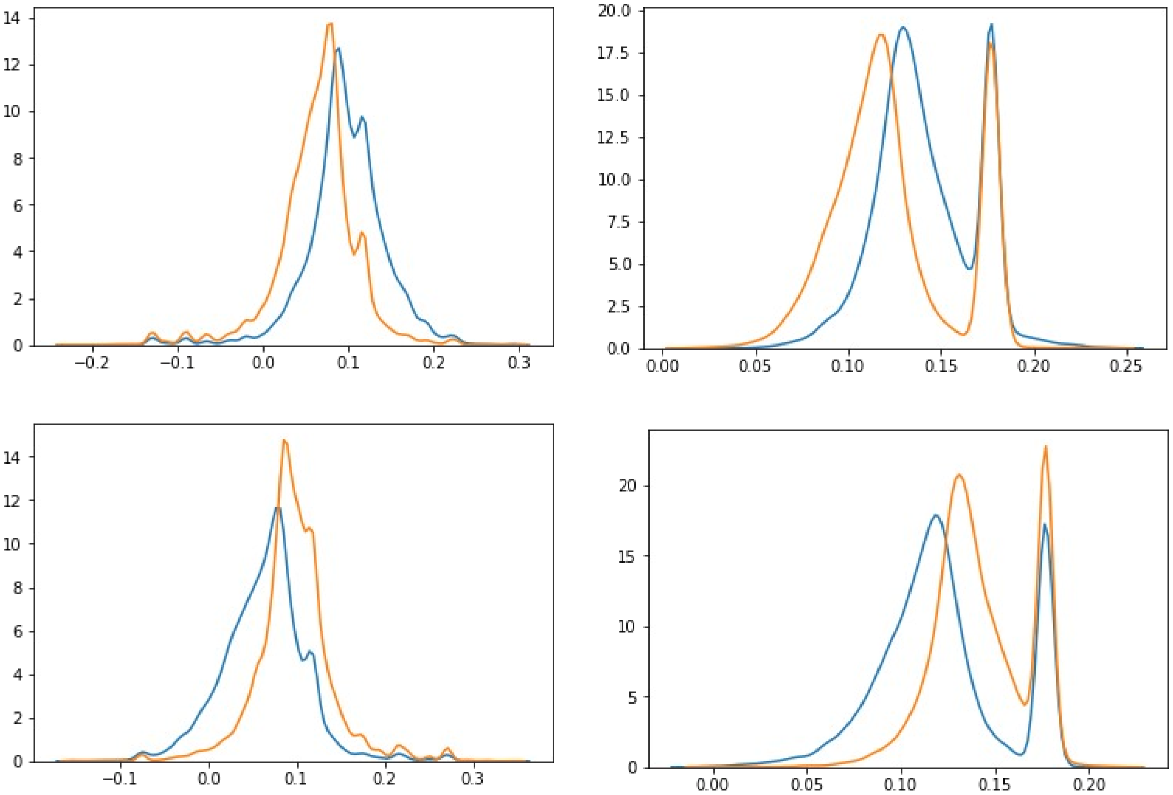
common vector



vector after subtraction

cos sim to male

cos sim to female



all domains

all domains - stop words

Model / Conclusion

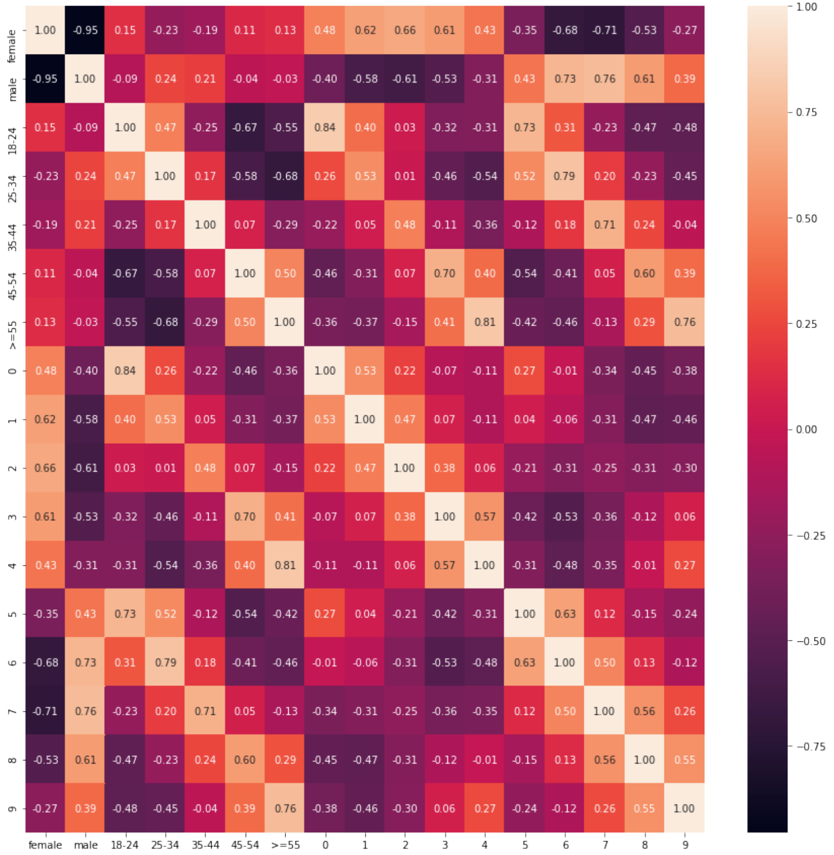
- CatBoost
  - loss_function: MultiClassOneVsAll
  - max_depth: 2
  - subsample: 0.7
  - colsample_bylevel: 0.6
  - learning_rate: 0.05
  - n_estimators: 200

```
           Train                            Test
F0 08.9235%  189 / 2118      F0 07.9427%   61 / 768
F1 63.6293% 3254 / 5114      F1 57.4240%  963 / 1677
F2 21.1250%  676 / 3200      F2 16.7134%  179 / 1071
F3 17.6471%  351 / 1989      F3 12.6645%   77 / 608
F4 00.0000%    0 / 673       F4 00.0000%    0 / 222
M0 00.0000%    0 / 1507      M0 00.0000%    0 / 505
M1 59.3100% 3851 / 6493      M1 57.6622% 1253 / 2173
M2 26.8127% 1028 / 3834      M2 24.7809%  311 / 1255
M3 00.0000%    0 / 1592      M3 00.0000%    0 / 555
M4 00.0000%    0 / 583       M4 00.0000%    0 / 201
-----------------------      ----------------------
-- 34.4943% 9349 / 27103     -- 31.4776% 2844 / 9035
```

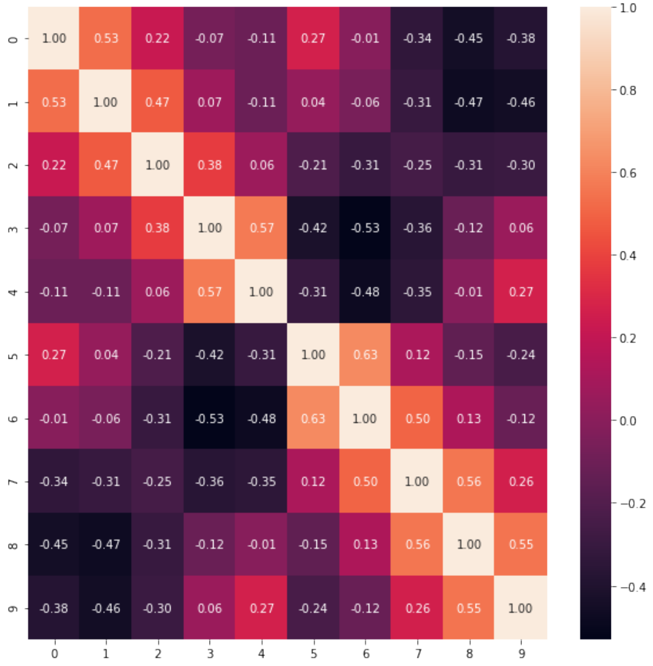
Questions?

No questions you have

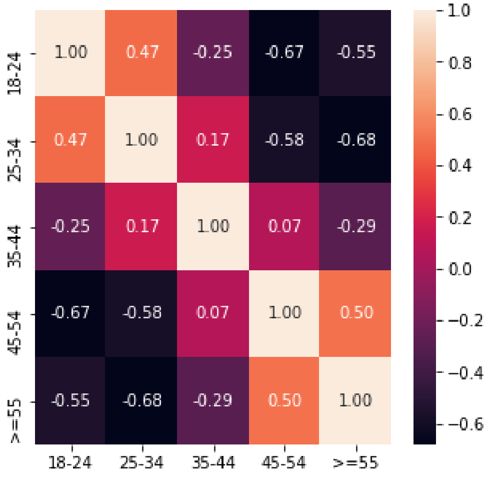
Features: cosine similarity



	female	male	18-24	25-34	35-44	45-54	>=55	0	1	2	3	4	5	6	7	8	9
female	1.00	-0.95	0.15	-0.23	-0.19	0.11	0.13	0.48	0.62	0.66	0.61	0.43	-0.35	-0.68	-0.71	-0.53	-0.27
male	-0.95	1.00	-0.09	0.24	0.21	-0.04	-0.03	-0.40	-0.58	-0.61	-0.53	-0.31	0.43	0.73	0.76	0.61	0.39
18-24	0.15	-0.09	1.00	0.47	-0.25	-0.67	-0.55	0.84	0.40	0.03	-0.32	-0.31	0.73	0.31	-0.23	-0.47	-0.48
25-34	-0.23	0.24	0.47	1.00	0.17	-0.58	-0.68	0.26	0.53	0.01	-0.46	-0.54	0.52	0.79	0.20	-0.23	-0.45
35-44	-0.19	0.21	-0.25	0.17	1.00	0.07	-0.29	-0.22	0.05	0.48	-0.11	-0.36	-0.12	0.18	0.71	0.24	-0.04
45-54	0.11	-0.04	-0.67	-0.58	0.07	1.00	0.50	-0.46	-0.31	0.07	0.70	0.40	-0.54	-0.41	0.05	0.60	0.39
>=55	0.13	-0.03	-0.55	-0.68	-0.29	0.50	1.00	-0.36	-0.37	-0.15	0.41	0.81	-0.42	-0.46	-0.13	0.29	0.76
0	0.48	-0.40	0.84	0.26	-0.22	-0.46	-0.36	1.00	0.53	0.22	-0.07	-0.11	0.27	-0.01	-0.34	-0.45	-0.38
1	0.62	-0.58	0.40	0.53	0.05	-0.31	-0.37	0.53	1.00	0.47	0.07	-0.11	0.04	-0.06	-0.31	-0.47	-0.46
2	0.66	-0.61	0.03	0.01	0.48	0.07	-0.15	0.22	0.47	1.00	0.38	0.06	-0.21	-0.31	-0.25	-0.31	-0.30
3	0.61	-0.53	-0.32	-0.46	-0.11	0.70	0.41	-0.07	0.07	0.38	1.00	0.57	-0.42	-0.53	-0.36	-0.12	0.06
4	0.43	-0.31	-0.31	-0.54	-0.36	0.40	0.81	-0.11	-0.11	0.06	0.57	1.00	-0.31	-0.48	-0.35	-0.01	0.27
5	-0.35	0.43	0.73	0.52	-0.12	-0.54	-0.42	0.27	0.04	-0.21	-0.42	-0.31	1.00	0.63	0.12	-0.15	-0.24
6	-0.68	0.73	0.31	0.79	0.18	-0.41	-0.46	-0.01	-0.06	-0.31	-0.53	-0.48	0.63	1.00	0.50	0.13	-0.12
7	-0.71	0.76	-0.23	0.20	0.71	0.05	-0.13	-0.34	-0.31	-0.25	-0.36	-0.35	0.12	0.50	1.00	0.56	0.26
8	-0.53	0.61	-0.47	-0.23	0.24	0.60	0.29	-0.45	-0.47	-0.31	-0.12	-0.01	-0.15	0.13	0.56	1.00	0.55
9	-0.27	0.39	-0.48	-0.45	-0.04	0.39	0.76	-0.38	-0.46	-0.30	0.06	0.27	-0.24	-0.12	0.26	0.55	1.00



	0	1	2	3	4	5	6	7	8	9
0	1.00	0.53	0.22	-0.07	-0.11	0.27	-0.01	-0.34	-0.45	-0.38
1	0.53	1.00	0.47	0.07	-0.11	0.04	-0.06	-0.31	-0.47	-0.46
2	0.22	0.47	1.00	0.38	0.06	-0.21	-0.31	-0.25	-0.31	-0.30
3	-0.07	0.07	0.38	1.00	0.57	-0.42	-0.53	-0.36	-0.12	0.06
4	-0.11	-0.11	0.06	0.57	1.00	-0.31	-0.48	-0.35	-0.01	0.27
5	0.27	0.04	-0.21	-0.42	-0.31	1.00	0.63	0.12	-0.15	-0.24
6	-0.01	-0.06	-0.31	-0.53	-0.48	0.63	1.00	0.50	0.13	-0.12
7	-0.34	-0.31	-0.25	-0.36	-0.35	0.12	0.50	1.00	0.56	0.26
8	-0.45	-0.47	-0.31	-0.12	-0.01	-0.15	0.13	0.56	1.00	0.55
9	-0.38	-0.46	-0.30	0.06	0.27	-0.24	-0.12	0.26	0.55	1.00



	18-24	25-34	35-44	45-54	>=55
18-24	1.00	0.47	-0.25	-0.67	-0.55
25-34	0.47	1.00	0.17	-0.58	-0.68
35-44	-0.25	0.17	1.00	0.07	-0.29
45-54	-0.67	-0.58	0.07	1.00	0.50
>=55	-0.55	-0.68	-0.29	0.50	1.00