

ISE 5103 Intelligent Data Analytics

Homework 6 - Modeling Competition

Daniel Carpenter, Sonaxy Mohanty, & Zachary Knepp

October 2022

Contents

General Data Prep	2
Read Training Data	2
Create <code>numeric</code> and <code>factor data frames</code>	2
2 (i) - Data Understanding	2
Numeric Data Quality Report	2
Factor Data Quality Report	3

General Data Prep

For general data preparation, please see conceptual steps below. See `.rmd` file for detailed code.

Read Training Data

Clean data to ensure each read variable has the correct data type (factor, numeric, Date, etc.)

Create numeric and factor data frames

Make data set of `numeric` variables called `df.train.numeric`

Make data set of `factor` variables called `df.train.factor`

2 (i) - Data Understanding

Create a data quality report of `numeric` and `factor` data

Numeric Data Quality Report

Num_Numeric_Variables	Total_Observations
4	70071

variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
visitNumber	0	1	3.1	8.7	1	1	1	2	155
timeSinceLastVisit	0	1	256450.2	1164717.4	0	0	0	10375	30074517
revenue	0	1	10.2	99.5	0	0	0	0	15981
pageviews	8	1	6.3	11.7	1	1	2	6	469

Factor Data Quality Report

- Appears that few people use website from the ads, which cause many null values.

Num_Factor_Variables	Total_Observations
35	70071

variable	n_missing	complete_rate	ordered	n_unique	top_counts
sessionId	0	1.00	FALSE	70071	200: 1, 400: 1, 600: 1, 700: 1
custId	0	1.00	FALSE	47249	234: 155, 558: 135, 455: 129, 818: 115
channelGrouping	0	1.00	FALSE	8	Org: 27503, Soc: 13528, Ref: 13482, Dir: 11824
isMobile	0	1.00	FALSE	2	0: 53993, 1: 16078
deviceCategory	0	1.00	FALSE	3	des: 53986, mob: 13868, tab: 2217
isTrueDirect	0	1.00	FALSE	2	0: 42026, 1: 28045
bounces	0	1.00	FALSE	2	0: 40719, 1: 29352
newVisits	0	1.00	FALSE	2	1: 46127, 0: 23944
browser	1	1.00	FALSE	27	Chr: 51584, Saf: 12007, Fir: 2407, Int: 1357
source	2	1.00	FALSE	131	goo: 29233, you: 12708, (di: 11825, mal: 10840
continent	85	1.00	FALSE	5	Ame: 42508, Asi: 13697, Eur: 11992, Oce: 901
subContinent	85	1.00	FALSE	22	Nor: 38860, Sou: 4823, Nor: 3601, Wes: 3563
country	85	1.00	FALSE	176	Uni: 36941, Ind: 3044, Uni: 2330, Can: 1918
operatingSystem	307	1.00	FALSE	15	Mac: 23970, Win: 23707, And: 8074, iOS: 7487
medium	11827	0.83	FALSE	5	org: 27503, ref: 27010, cpc: 2085, aff: 911
networkDomain	33448	0.52	FALSE	5014	com: 2890, ver: 1372, rr.: 1319, com: 1247
topLevelDomain	33448	0.52	FALSE	183	net: 15027, com: 6297, tr: 874, in: 868
region	38485	0.45	FALSE	309	Cal: 11254, New: 3468, Ill: 1047, Tex: 909
city	39028	0.44	FALSE	477	Mou: 4569, New: 3465, San: 2183, Sun: 1362
referralPath	43062	0.39	FALSE	383	/: 11419, /yt: 4359, /yt: 842, /an: 836
metro	49183	0.30	FALSE	72	San: 10072, New: 3526, Los: 1050, Chi: 1047
campaign	67310	0.04	FALSE	6	AW : 1229, Dat: 911, AW : 575, tes: 35
keyword	67412	0.04	FALSE	415	6qE: 997, 1hZ: 213, Goo: 183, (Re: 182
adwordsClickInfo.gclid	68245	0.03	FALSE	1405	Cj0: 14, Cjw: 10, ClY: 9, Cj0: 9
adwordsClickInfo.page	68260	0.03	FALSE	5	1: 1806, 2: 2, 3: 1, 5: 1
adwordsClickInfo.slot	68260	0.03	FALSE	2	Top: 1771, RHS: 40, emp: 0
adwordsClickInfo.adNetworkType	68260	0.03	FALSE	1	Goo: 1811, emp: 0
adwordsClickInfo.isVideoAd	68260	0.03	FALSE	1	0: 1811
adContent	69230	0.01	FALSE	27	Goo: 449, Dis: 82, Goo: 79, Ful: 49