ISE 5103 Intelligent Data Analytics

Homework 6 - Modeling Competition

Daniel Carpenter, Sonaxy Mohanty, & Zachary Knepp

October 2022

Contents

Packages

```
# Data Wrangling
library(tidyverse)
library(skimr)

# Modeling
library(MASS)
library(caret) # Modeling variants like SVM
library(earth) # Modeling with Mars
library(pls)   # Modeling with PLS
library(glmnet) # Modeling with LASSO

# Aesthetics
library(knitr)
library(cowplot)  # multiple ggplots on one plot with plot_grid()
library(scales)
library(kableExtra)
library(ggplot2)

#Hold-out Validation
library(caTools)

#Data Correlation
library(GGally)
library(regclass)

#RMSE Calculation
library(Metrics)

#p-value for OLS model
library(broom)

#ncvTest
library(car)
```

General Data Prep

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

Read Training Data

```
# Convert all character data to factor
df.train <- read.csv('Train.csv', stringsAsFactors = TRUE)

# convert the ""'s to NA
df.train[df.train == ""] <- NA

# Ensure boolean variables are numeric
df.train$adwordsClickInfo.isVideoAd <- as.numeric(df.train$adwordsClickInfo.isVideoAd)
```

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
13	70071

variable	n_missing	complete_rate	mean	sd
sessionId	0	1.00	4708198750205.95	2732187943023.75
custId	0	1.00	48874.99	27321.88
visitStartTime	0	1.00	1485110879.82	9106581.66
visitNumber	0	1.00	3.15	8.66
timeSinceLastVisit	0	1.00	256450.24	1164717.35
isMobile	0	1.00	0.23	0.42
isTrueDirect	0	1.00	0.40	0.49
adwordsClickInfo.page	68260	0.03	1.01	0.18
adwordsClickInfo.isVideoAd	68260	0.03	0.00	0.00
pageviews	8	1.00	6.30	11.69
bounces	40719	0.42	1.00	0.00
newVisits	23944	0.66	1.00	0.00
revenue	0	1.00	10.17	99.53

variable	p0	p25	p50	p75	p100
sessionId	200000120	2328800000173	4688000000146	7079450000170	9449700000194
custId	1795	25081	48673	72588	96290
visitStartTime	1470035066	1477552704	1484102061	1493099922	1501655863
visitNumber	1	1	1	2	155
timeSinceLastVisit	0	0	0	10375	30074517
isMobile	0	0	0	0	1
isTrueDirect	0	0	0	1	1
adwordsClickInfo.page	1	1	1	1	7
adwordsClickInfo.isVideoAd	0	0	0	0	0
pageviews	1	1	2	6	469
bounces	1	1	1	1	1
newVisits	1	1	1	1	1
revenue	0	0	0	0	15981

Factor Data Quality Report

Num_Factor_Variables	Total_Observations
35	70071

variable	n_missing	complete_rate	ordered	n_unique	top_counts
date	0	1.00	FALSE	366	201: 362, 201: 352, 201: 349, 201: 347
channelGrouping	0	1.00	FALSE	8	Org: 27503, Soc: 13528, Ref: 13482, Dir: 11824
browser	1	1.00	FALSE	27	Chr: 51584, Saf: 12007, Fir: 2407, Int: 1357
operatingSystem	307	1.00	FALSE	15	Mac: 23970, Win: 23707, And: 8074, iOS: 7487
deviceCategory	0	1.00	FALSE	3	des: 53986, mob: 13868, tab: 2217
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
region	38485	0.45	FALSE	309	Cal: 11254, New: 3468, Ill: 1047, Tex: 909
metro	49183	0.30	FALSE	72	San: 10072, New: 3526, Los: 1050, Chi: 1047
city	39028	0.44	FALSE	477	Mou: 4569, New: 3465, San: 2183, Sun: 1362
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
campaign	67310	0.04	FALSE	6	AW : 1229, Dat: 911, AW : 575, tes: 35
source	2	1.00	FALSE	131	goo: 29233, you: 12708, (di: 11825, mal: 10840
medium	11827	0.83	FALSE	5	org: 27503, ref: 27010, cpc: 2085, aff: 911
keyword	67412	0.04	FALSE	415	6qE: 997, 1hZ: 213, Goo: 183, (Re: 182
referralPath	43062	0.39	FALSE	383	/: 11419, /yt: 4359, /yt: 842, /an: 836
adContent	69230	0.01	FALSE	27	Goo: 449, Dis: 82, Goo: 79, Ful: 49
adwordsClickInfo.slot	68260	0.03	FALSE	2	Top: 1771, RHS: 40, emp: 0
adwordsClickInfo.gclId	68245	0.03	FALSE	1405	Cj0: 14, Cjw: 10, CIy: 9, Cj0: 9
adwordsClickInfo.adNetworkType	68260	0.03	FALSE	1	Goo: 1811, emp: 0

–>