ISE 5103 Intelligent Data Analytics

Homework 6 - Modeling Competition

Daniel Carpenter, Sonaxy Mohanty, & Zachary Knepp

October 2022

Contents

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` *base* `data frames`

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

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

`(a, i)` - Data Understanding

Create a data quality report of `numeric` and `factor` data
Created function called `dataQualityReport()` to create factor and numeric QA report

Numeric Data Quality Report

- `pageviews` has some null values, but there are an insignificant amount, so we will just drop those rows.

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

- Location data unknown, so add an `Unknown` label for `null` values
- Appears that few people use website from the ads, which cause many null values. See more details below.

Num_Factor_Variables	Total_Observations
28	70071

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

Exploratory Analysis

Analysis 1: TODO

Analysis 2: TODO

(a, ii) - Data Preparation

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

Clean up Null Data

See that when `campaign` is `null`, then other `ad` related fields are (mostly) null

- Implication: these other fields depend on the `campaign` variable
- So, set `adwordsClickInfo.page` null fields to `None` description, since a null value indicates the user did not come using an advertisement

Now, we are going to do a similar tactic for the location data.

- There is a substantial amount of unique location data, which will not work well within a model.
- So, then location data is null, then we flag as `Unknown`

Similar to the location data, if the `referralPath` or `medium` is null, label as `None`. We would likely factor lump this data anyways.

Now we have very few null values rows. Let's simply remove them. See below for how many.

```
## [1] "There are 2497 rows with nulls"
```

```
## [1] "That equates to 3.6% rows with nulls"
```

```
## [1] "Total Rows Remaining: 67574"
```

Factor Lump the Factor Data

Overview: Create `Other` Bin for Columns over `5` Unique Values

- Applied to any `factor` column with over 5 unique values
- Applies `fct_lump()` function to columns via dynamic `dplyr` logic

```
## [1] "Before cleaning, there are 18 factor columns with more than 5 unique values"
```

(a, iii) - Modeling

OLS Model

Model Setup

Fit the Model

View and Interpret Results

Model 2: PCR Mixed with SVM

Model Setup

- Uses `numeric` data for Principal Component Analysis
  - Data includes outliers
  - Chose number of PC's that explain 70% of the variation. This is just a general judgement call to keep the number of principal components low.
- Then appends the `factor` columns *without `NULL` values* and `Revenue` to the data
- Finally, uses `stepAIC()` to best model data
- See interpretation at end

```
# Get cleaned `numeric` and `factor` `data frames`
# After cleaning, two data sets that contain..

## Numeric data
df.train.clean.numeric <- df.train.clean %>% select_if(is.numeric)

## Factors
df.train.clean.factor  <- df.train.clean %>% dplyr::select(where(is.factor))
```

```
# Perform PCA
# Principal component analysis on numeric data
pc.train <- prcomp(df.train.clean.numeric %>% dplyr::select(-revenue), # do not include response var
                   center = TRUE, # Mean centered
                   scale  = TRUE  # Z-Score standardized
                   )
# See first 10 cumulative proportions
pc.train.summary <- summary(pc.train)
pc.train.summary
```

```
## Importance of components:
##                          PC1   PC2   PC3
## Standard deviation     1.048 0.989 0.961
## Proportion of Variance 0.366 0.326 0.308
## Cumulative Proportion  0.366 0.692 1.000
```

```
# Now we choose number of PC's that explain 75% of the variation
# Note this threshold is just a judgement call. No significance behind 75%
cumPropThreshold = 0.70 # The threshold
numPCs <- sum(pc.train.summary$importance['Cumulative Proportion', ] < cumPropThreshold)
paste0('There are ', numPCs, ' principal components that explain up to ', cumPropThreshold*100,
       '% of the variation in the data')
```

```
## [1] "There are 2 principal components that explain up to 70% of the variation in the data"
```

```
chosenPCs <- as.data.frame(pc.train$x[, 1:numPCs])
```

Join on the `factor` data and `revenue`

```
df.svm <- cbind(revenue = df.train.clean.numeric$revenue,
                chosenPCs,
                df.train.clean.factor) %>%

  dplyr::select(-c(
    # Remove ID cols since will overfit model
    sessionId, custId,

    # Remove ad content for the time being
    starts_with('ad')
    ) )
```

Model controls

```
ctrl <- trainControl(method  = "repeatedcv",
                     number  = 5, # 5 fold cross validation
                     repeats = 1  # 2 repeats
                     )
```

Fit the Model

- SVM model containing:
  - Principal components explaining 70% of variation in `numeric` data
  - Non-null `factor` data
  - *Predicted variable:* `revenue`

```
# Train and tune the SVM

fit.svm <- train(data = df.svm[1:100, ], # TODO - Only showing 1000 records
                 revenue ~ .,
                 method     = "svmRadial",         # Radial kernel
                 tuneLength = 9,                   # 9 values of the cost function
                 preProc    = c("center","scale"), # Center and scale data
                 trControl  = ctrl)
```

View and Interpret Results

```
## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 1
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.0104248452673432
##
## Number of Support Vectors : 35
##
## Objective Function Value : -153
## Training error : 122.830676
```

Model	Notes	Hyperparameters	RMSE	Rsquared
SVM	caret and svmRadial	C = 1 , Epsilon = 0.1	6.9	0.09

Model 3:

Model Setup

Fit the Model

View and Interpret Results

Model 4:

Model Setup

Fit the Model

View and Interpret Results

`(a, iv)` - Debrief

Summary Table

Model	Notes	Hyperparameters	RMSE	Rsquared
SVM	caret and svmRadial	C = 1 , Epsilon = 0.1	6.9	0.09