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

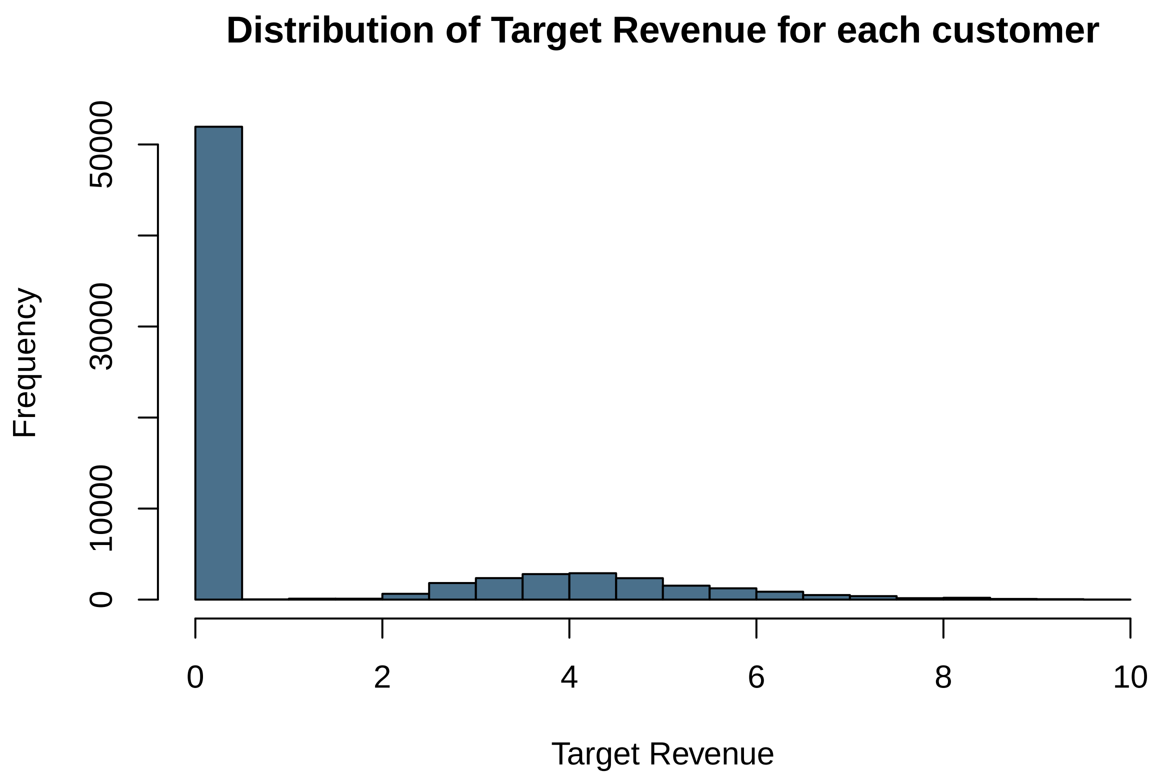
variable	n_missing	complete_rate	n_unique	top_counts
sessionId	0	1.00	70071	200: 1, 400: 1, 600: 1, 700: 1
custId	0	1.00	47249	234: 155, 558: 135, 455: 129, 818: 115
channelGrouping	0	1.00	8	Org: 27503, Soc: 13528, Ref: 13482, Dir: 11824
deviceCategory	0	1.00	3	des: 53986, mob: 13868, tab: 2217
isTrueDirect	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

- Need to predict a transformation of the `aggregate customer-level sales` value based on the natural log

Analysis 1:

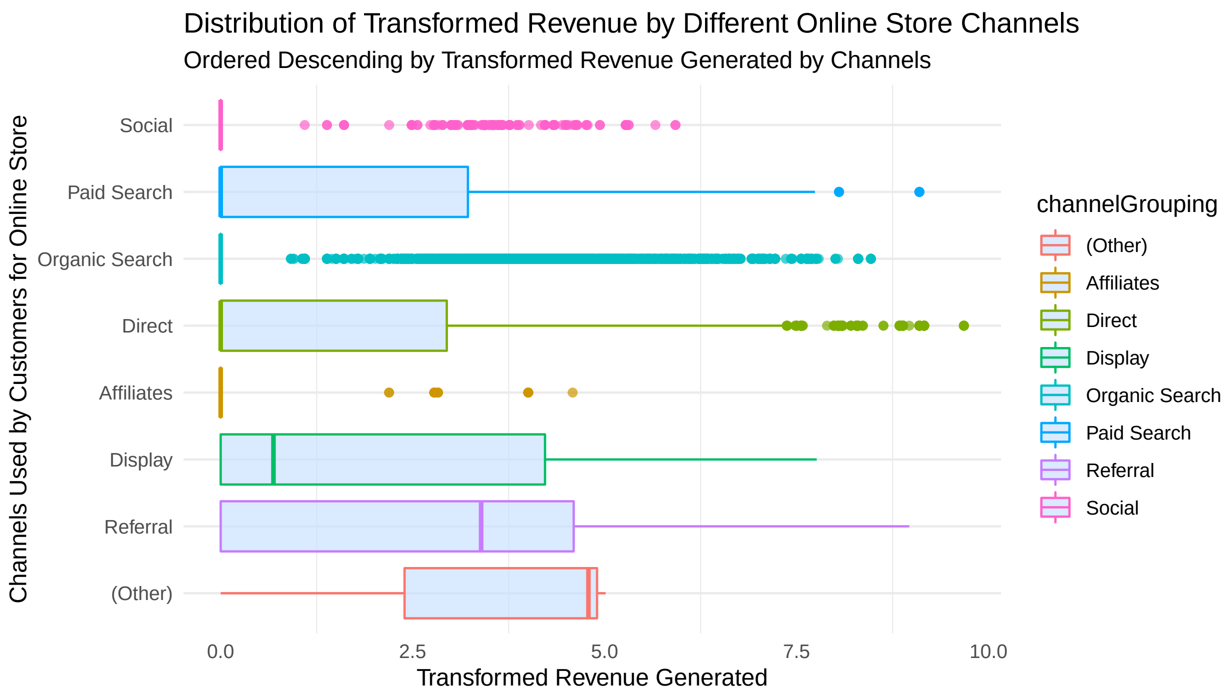
- Checking the distribution of the transformation of the aggregrate customer-level sales value based on the natural log:



- We can see that the transformed revenue doesn't look like a normal distribution with a spike at 0 revenue which means it can be an outlier.

- From the dataset, we can also see two sets of cutomers - one set who visited the site once and another who visited multiple times.

- The histogram here doesn't take into account the above fact and therefore the frequency of the target revenue is compromised.



Analysis 2:

- The relevance behind this plot is to analyze whether the revenue generated is dependent on the channels via which the user came to the online store.

- Social media definitely plays an important role to attract customers to ttry online shopping.

(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 `region` is `Osaka Prefecture` and `city` is `Osaka` some location details are `NULL`

- Implication: the other fields can be manually set to correct values based on region and city criteria
- So, set `location related` null fields to `know` description for the above `region` and `city` condition

See that when `continent` is `null`, then other `location` related fields are also null

- Implication: these other fields depend on the `continent` variable
- So, set `location related` null fields to `Unknow` description

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

- Implication: these other fields depend on the `medium` variable
- So, set these null fields to `None` description, since a null value indicates the user did not has `no traffic source`

See that when `campaign` is `null`, then some `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

Similar approach is done to impute the rest `NAs` in the categorical variables of the data set

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

```
## [1] "There are 318 rows with nulls"
```

```
## [1] "That equates to 0.5% rows with nulls"
```

```
## [1] "Total Rows Remaining: 69753"
```

- We are going to factor collapse factor columns with more than 4 columns
- So there will be 5 of the original, and 1 containing 'other'

```
## [1] "Before cleaning, there are 24 factor columns with more than 4 unique values"
```

```
## [1] "After cleaning, there are 2 columns with more than 5 unique values (omitting NA's)"
```

Group by Customer

Get list of customers who visited once and twice

Group by customer & Sum up all numeric data

- Filter to only the customers who visited twice
- Get the unique visits and choose the first visit
- THis is just an assumption! Not the best, but we have to make a choice.
- Append unique customers to non-unique customers (that are now unique)
- Note not using all columns, only columns NOT specific to the model

```
## [1] 46967
```

```
## [1] 46967
```

```
## [1] 28
```

```
## [1] 28
```

Create `targetRevenue` Variable

```
df.train.clean.cust <- df.train.clean.cust %>%
  mutate(targetVariable = log(revenue + 1)) %>%
  dplyr::select(-revenue)
```

Then create dataset without the `custID` field called `df.train.clean.noCust`

(a, iii) - Modeling

OLS Model

Fit the Model

- Initially created a model with all variables, then used `stepAIC()` to identify important variables
- Implemented in the OLS model to realize a better fit model.

```
# The OLS model
# See RMD for stepAIC function that generated these relevant variables for the model
ols <- lm(targetVariable ~ operatingSystem + country + metro + city + networkDomain +
            source + keyword + isTrueDirect + referralPath + bounces +
            newVisits + pageviews,
          data = df.train.clean.noCust)
```

View and Interpret Results

Model	Notes	Hyperparameters	RMSE	Rsquared
OLS	lm	N/A	0.93	0.5

- Comparing the OLS model with various other robust models to see how better these robust models perform as compared to the OLS model based on the RMSE and R^2.

Model 2: PCR Model

Fit the Model

- Based on model testing, highest R^2 is around 68 number of components.
- Fits data much better than the former model.

View and Interpret Results

Model	Notes	Hyperparameters	RMSE	Rsquared
PCR	pcr	ncomp = 36	0.94	0.49

- 28 components explain 100% variance in the data set, but 15 components are enough to justify more than 75% of data variance.
- We will see if MARS and ELasticNet models outperform the PCR model.

Model 3: MARS

Fit the Model

- Use MARS model from earth package.
- Fits data similarly to the former models.

View and Interpret Results

Model	Notes	Hyperparameters	RMSE	Rsquared
MARS	caret and earth	Degree = 1 , nprune = 8	0.76	0.66

- See that the model overall performs well, and in fact performs better as compared to the PCR model (in terms of RMSE and R^2).

Model 4: Elastic Net Model

Fit the Model

View and Interpret Results

Model	Notes	Hyperparameters	R
Elastic Net	caret and elasticnet	Alpha = 0.3 , Lambda = 0.000381198688071757	

- The Elastic Net Model performs similar to PCR model.
- Thus, MARS model outperformed all the rest models based on the `RMSE` and R^2 measures.
- We will now predict the test set with the MARS model that we created.

(a, iv) - Debrief

Summary Table

Model	Notes	Hyperparameters	RMSE	Rsquared
OLS	lm	N/A	0.93	0.50
PCR	pcr	ncomp = 36	0.94	0.49
MARS	caret and earth	Degree = 1 , nprune = 8	0.76	0.66
Elastic Net	caret and elasticnet	Alpha = 0.3 , Lambda = 0.000381198688071757	0.94	0.49

Interpretations of Debrief

- For MARS model, we used `caret` and `earth` method

- The hyperparameters that worked best for the model is`Degree = 1` , `nprune = 8`

- We first imputed all the `NAs` we had in the data set and got the final dataset of `69753` rows, i.e., we only dropped about 0.5% rows with nulls in order to capture the sanctity of the dataset and performed transformation of revenue on this dataset

- Since after imputation, there were a large number of dummy categorical values, so initially when the model were tested it took a lot of time and for some cases the model didn't even fit

- So, factor collapsing was performed on these factor columns which gave us 5 of the `original` values and 1 `other`

- Same steps were performed on the test data set

- After that even, the `RMSE` and R^2 values were not upto the mark for the models
- So, came up with the idea of grouping of customers, i.e., grouping based on which customers visited once or twice

- Then the transformation of revenue, `targetRevenue` was done on this grouped data set, and then `custId` column was removed from the dataset to avoid overfitting of the model

- Then the performance of the models boosted up

Apply to Test Data

- Need to clean test data like we did in the train
- Note all comments for the main model apply here
- Then apply the models to this dataset
- Outputs a CSV with predicted customer log revenue
- For general data preparation, please see conceptual steps below. See `.rmd` file for detailed code.