

Expedia Training Data-set Analysis

SarahLynne Palomo

21/04/2021

Exploring srch_id

Find the unique srch_id's and count how many of them are in the submission_sample data-set:

```
# Load Training data-set
tr <- read.csv("training_set_VU_DM.csv")

# Find unique search id's
uniq_srch <- unique(tr$srch_id)
num_uniq_ids <- nrow(as.matrix(uniq_srch))
head(uniq_srch, 10)

## [1] 1 4 6 8 11 12 17 21 25 28

tail(uniq_srch, 10)

## [1] 332765 332768 332772 332774 332776 332777 332781 332782 332784 332785

last_srch_id <- tail(uniq_srch, n=1) # Highest srch_id in ascending list
perc_missing_srch <- (1 - (num_uniq_ids/last_srch_id)) * 100
perc_missing_srch

## [1] 39.96274
```

Here we can see that a good portion of srch_id's are missing from the training data-set sequence (40%)

Exploring NULL values

```
#which(is.null(tr$srch_id)) # <-- this does not work!
#which(tr$srch_id == "NULL") # Not a real NULL value; it's only a string
#which(tr$comp3_rate == "NULL") # testing function on a column that contains "NULL"

# Create a new table for analyzing NULL stats
df_null_ratios <- data.frame(matrix(ncol = 3, nrow = 54))
names(df_null_ratios) <- list("Column_Name", "Number_of_Nulls", "Percentage_of_Rows")

tr_row <- nrow(tr)
tr_col <- ncol(tr)
i <- 0

for (i in 1:tr_col) {
  # Populate column names from the training set into new table
  df_null_ratios[i, 1] <- names(tr)[i]

  # Find the rows in the training set that contain string "NULL"
```

```

null_rows <- which(tr[,i] == "NULL")

# Populate "NULL" counts into new table
num_nulls <- nrow(as.matrix(null_rows))
df_null_ratios[i, 2] <- num_nulls

# Populate percentage of counts into new table
null_perc <- (num_nulls / tr_row) * 100
df_null_ratios[i, 3] <- null_perc
}

```

```
df_null_ratios
```

| ## | Column_Name | Number_of_Nulls | Percentage_of_Rows |
|-------|-----------------------------|-----------------|--------------------|
| ## 1 | srch_id | 0 | 0.00000 |
| ## 2 | date_time | 0 | 0.00000 |
| ## 3 | site_id | 0 | 0.00000 |
| ## 4 | visitor_location_country_id | 0 | 0.00000 |
| ## 5 | visitor_hist_starrating | 4706481 | 94.92036 |
| ## 6 | visitor_hist_adr_usd | 4705359 | 94.89774 |
| ## 7 | prop_country_id | 0 | 0.00000 |
| ## 8 | prop_id | 0 | 0.00000 |
| ## 9 | prop_starrating | 0 | 0.00000 |
| ## 10 | prop_review_score | 0 | 0.00000 |
| ## 11 | prop_brand_bool | 0 | 0.00000 |
| ## 12 | prop_location_score1 | 0 | 0.00000 |
| ## 13 | prop_location_score2 | 1090348 | 21.99015 |
| ## 14 | prop_log_historical_price | 0 | 0.00000 |
| ## 15 | position | 0 | 0.00000 |
| ## 16 | price_usd | 0 | 0.00000 |
| ## 17 | promotion_flag | 0 | 0.00000 |
| ## 18 | srch_destination_id | 0 | 0.00000 |
| ## 19 | srch_length_of_stay | 0 | 0.00000 |
| ## 20 | srch_booking_window | 0 | 0.00000 |
| ## 21 | srch_adults_count | 0 | 0.00000 |
| ## 22 | srch_children_count | 0 | 0.00000 |
| ## 23 | srch_room_count | 0 | 0.00000 |
| ## 24 | srch_saturday_night_bool | 0 | 0.00000 |
| ## 25 | srch_query_affinity_score | 4640941 | 93.59855 |
| ## 26 | orig_destination_distance | 1607782 | 32.42577 |
| ## 27 | random_bool | 0 | 0.00000 |
| ## 28 | comp1_rate | 4838417 | 97.58125 |
| ## 29 | comp1_inv | 4828788 | 97.38705 |
| ## 30 | comp1_rate_percent_diff | 4863908 | 98.09535 |
| ## 31 | comp2_rate | 2933675 | 59.16639 |
| ## 32 | comp2_inv | 2828078 | 57.03671 |
| ## 33 | comp2_rate_percent_diff | 4402109 | 88.78179 |
| ## 34 | comp3_rate | 3424059 | 69.05646 |
| ## 35 | comp3_inv | 3307357 | 66.70281 |
| ## 36 | comp3_rate_percent_diff | 4485550 | 90.46462 |
| ## 37 | comp4_rate | 4650969 | 93.80080 |
| ## 38 | comp4_inv | 4614684 | 93.06900 |
| ## 39 | comp4_rate_percent_diff | 4827261 | 97.35626 |
| ## 40 | comp5_rate | 2735974 | 55.17916 |

```
## 41          comp5_inv      2598327      52.40309
## 42  comp5_rate_percent_diff  4117248      83.03671
## 43          comp6_rate      4718190      95.15651
## 44          comp6_inv      4697371      94.73663
## 45  comp6_rate_percent_diff  4862173      98.06036
## 46          comp7_rate      4642999      93.64006
## 47          comp7_inv      4601925      92.81168
## 48  comp7_rate_percent_diff  4819832      97.20643
## 49          comp8_rate      3041693      61.34490
## 50          comp8_inv      2970844      59.91602
## 51  comp8_rate_percent_diff  4343617      87.60212
## 52          click_bool           0      0.00000
## 53  gross_bookings_usd      4819957      97.20895
## 54          booking_bool           0      0.00000
```

Check the fields with more than 90% NULL values against the Assignment description of the data-set for significance!

Exploring Date/Time

Dates are not in sequential order. It may be necessary to split *date_time* field into Date and Time.

```
min(tr$date_time)
```

```
## [1] "2012-11-01 00:08:29"
```

```
max(tr$date_time)
```

```
## [1] "2013-06-30 23:58:24"
```