HW4_Feature Engineering

Jasmine Zhang

2023-04-29

Feature engineering

Building on the previous model, we will create other feature variables and transformations to improve our model.

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
ord <- read_csv("~/Downloads/orders.csv")
```

```
## Rows: 353687 Columns: 6

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): orddate
## dbl (5): id, ordnum, category, qty, price
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
ord
```

```
## # A tibble: 353,687 x 6
##       id orddate   ordnum category   qty price
##    <dbl> <chr>      <dbl>    <dbl> <dbl> <dbl>
##  1   957 10FEB2008  38650       35     1  5.01
##  2   957 10FEB2008  38650       35     1 20.4
##   3   957 10FEB2008  38650       19      1 20.4
##   4   957 15MAR2008  48972       40      1 25.5
##   5   957 22NOV2008 150011       40      1 14.3
##   6   957 22NOV2008 150011       40      1  8.59
##   7   957 03OCT2009 286151       19      1 15.3
##   8   957 04APR2010 376779       14      1 12.8
##   9   957 04APR2010 376779       14      1  5.09
##  10   957 04APR2010 376779       35      1  6.54
## # i 353,677 more rows
```

Order per category

```
#Total orders per category
ord_per_cat <- ord %>%
  group_by(id, category) %>%
  summarise(ord_per_cat=n(), .groups = "drop")
ord_per_cat
```

```
## # A tibble: 108,052 x 3
##       id category ord_per_cat
##    <dbl>    <dbl>       <int>
##  1   957        1           1
##  2   957        5           2
##  3   957       14           4
##  4   957       19           4
##  5   957       20           4
##  6   957       26           1
##  7   957       35           7
##  8   957       37           4
##  9   957       40           4
## 10   957       41           2
## # i 108,042 more rows
```

```
#Transform to wide format for total orders per category
ord_per_cat_wide = spread(ord_per_cat, category, ord_per_cat, fill = 0)
ord_per_cat_wide
```

```
## # A tibble: 16,781 x 31
##       id   `1`   `3`   `5`   `6`   `7`   `8`   `9`  `10`  `12`  `14`  `17`  `19`
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1   957     1     0     2     0     0     0     0     0     0     4     0     4
##  2  2062     3     0     2     1     0     4     0     0     0     1     0     1
##  3  2232     0     0     0     0     0     0     0     1     7    12     3    17
##  4  2623     1     0     1     1     0     0     0     0     0     2     0     0
##  5  3000     0     0     1     0     0     0     0     1     0     0     0     2
##  6  3689     0     0     0     0     0     1     0     0     0     0     0     3
##  7  4251     1     0     0     0     0     0     0     0     0    32     0     0
##  8  4642     0     0     0     0     0     0     0     0     0     0     0     2
##  9  5002     2     0     0     0     0     0     0     0     3    12     0     5
## 10  6084    12    11     5     4     2     9     0     0     3    28     0    44
## # i 16,771 more rows
## # i 18 more variables: ‘20‘ <dbl>, ‘21‘ <dbl>, ‘22‘ <dbl>, ‘23‘ <dbl>,
## #   ‘26‘ <dbl>, ‘27‘ <dbl>, ‘30‘ <dbl>, ‘31‘ <dbl>, ‘35‘ <dbl>, ‘36‘ <dbl>,
## #   ‘37‘ <dbl>, ‘38‘ <dbl>, ‘39‘ <dbl>, ‘40‘ <dbl>, ‘41‘ <dbl>, ‘44‘ <dbl>,
## #   ‘50‘ <dbl>, ‘99‘ <dbl>
```

Average order value