Mercari Price Suggestion Challenge

ooooooh!

Overview

Competition URL：https://www.kaggle.com/c/mercari-price-suggestion-challenge/kernels

時程表：

- February 7, 2018 - Entry deadline. You must accept the competition rules before this date in order to compete.
- February 7, 2018 - Team Merger deadline. This is the last day participants may join or merge teams.
- February 14, 2018 - Final submission deadline. This is the last day participants are allowed to modify Kernels, submit to competition, and to select the Kernel version for Stage 2. After this date, all the Kernels versions are recorded and will be re-run based on the selections.
- February 15, 2018 - Stage 2 evaluation period starts. Sit back and watch the leaderboard.
- February 21, 2018 - Stage 2 ends. Final results announced.

Exploratory Data Analysis (EDA)

```
# 載入所需的套件
library(data.table) # Loading data
library(RColorBrewer) # wordcloud
library(wordcloud) # wordcloud
library(ggplot2) # plot
library(stringr)
```

將 train.tsv 資料載入，因為沒加 encoding 在 windows 上 category_name 會出現亂碼(mac 上還沒試過)，所以下此語法：

```
# fread 是專門針對大量資料的語言，可以快速讀取百萬筆資料，sep = '\t'因為資料是用大空格分隔。
train<-fread(file = '~/R/K/train.tsv', sep = '\t', header = TRUE,encoding = 'UTF-8')

##
Read 25.6% of 1482535 rows
Read 38.4% of 1482535 rows
Read 54.0% of 1482535 rows
```

```
Read 74.2% of 1482535 rows
Read 91.1% of 1482535 rows
Read 1482535 rows and 8 (of 8) columns from 0.315 GB file in 00:00:08
```

接下來可以進行簡單的資料觀察

```
# 觀察前5筆資料
head(train,5)
```

```
##    train_id                                 name item_condition_id
## 1:        0 MLB Cincinnati Reds T Shirt Size XL                  3
## 2:        1    Razer BlackWidow Chroma Keyboard                  3
## 3:        2                       AVA-VIV Blouse                 1
## 4:        3              Leather Horse Statues                   1
## 5:        4               24K GOLD plated rose                   1
##                                       category_name brand_name pri
ce
## 1:                                Men/Tops/T-shirts
10
## 2: Electronics/Computers & Tablets/Components & Parts      Razer
52
## 3:                        Women/Tops & Blouses/Blouse     Target
10
## 4:                Home/Home Decor/Home Decor Accents
35
## 5:                            Women/Jewelry/Necklaces
44
##    shipping
## 1:        1
## 2:        0
## 3:        1
## 4:        1
## 5:        0
##

                                  item_description
## 1:

                                No description yet
## 2: This keyboard is in great condition and works like it came out of
 the box. All of the ports are tested and work perfectly. The lights ar
e customizable via the Razer Synapse app on your PC.
## 3:                                                                A
dorable top with a hint of lace and a key hole in the back! The pale pi
nk is a 1X, and I also have a 3X available in white!
## 4:                New with tags. Leather horses. Retail for [rm] eac
h. Stand about a foot high. They are being sold as a pair. Any question
s please ask. Free shipping. Just got out of storage
## 5:
```

```
          Complete with certificate of authenticity
```

```
# 欄位探勘
str(train)
```

```
## Classes 'data.table' and 'data.frame':   1482535 obs. of  8 variable
s:
##  $ train_id         : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ name             : chr  "MLB Cincinnati Reds T Shirt Size XL" "Ra
zer BlackWidow Chroma Keyboard" "AVA-VIV Blouse" "Leather Horse Statues
" ...
##  $ item_condition_id: int  3 3 1 1 1 3 3 3 3 3 ...
##  $ category_name    : chr  "Men/Tops/T-shirts" "Electronics/Computer
s & Tablets/Components & Parts" "Women/Tops & Blouses/Blouse" "Home/Hom
e Decor/Home Decor Accents" ...
##  $ brand_name       : chr  "" "Razer" "Target" "" ...
##  $ price            : num  10 52 10 35 44 59 64 6 19 8 ...
##  $ shipping         : int  1 0 1 1 0 0 0 1 0 0 ...
##  $ item_description : chr  "No description yet" "This keyboard is in
 great condition and works like it came out of the box. All of the port
s are tested and work"| __truncated__ "Adorable top with a hint of lace
 and a key hole in the back! The pale pink is a 1X, and I also have a 3
X available in white!" "New with tags. Leather horses. Retail for [rm]
each. Stand about a foot high. They are being sold as a pair. An"| __tr
uncated__ ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
# 觀察各個欄位敘述統計
summary(train)
```

```
##     train_id           name           item_condition_id category_nam
e
##  Min.   :      0   Length:1482535     Min.   :1.000     Length:14825
35
##  1st Qu.: 370634   Class :character   1st Qu.:1.000     Class :chara
cter
##  Median : 741267   Mode  :character   Median :2.000     Mode  :chara
cter
##  Mean   : 741267                      Mean   :1.907
##  3rd Qu.:1111901                      3rd Qu.:3.000
##  Max.   :1482534                      Max.   :5.000
##   brand_name            price            shipping      item_descript
ion
##  Length:1482535     Min.   :   0.00   Min.   :0.0000   Length:148253
5
##  Class :character   1st Qu.:  10.00   1st Qu.:0.0000   Class :charac
```

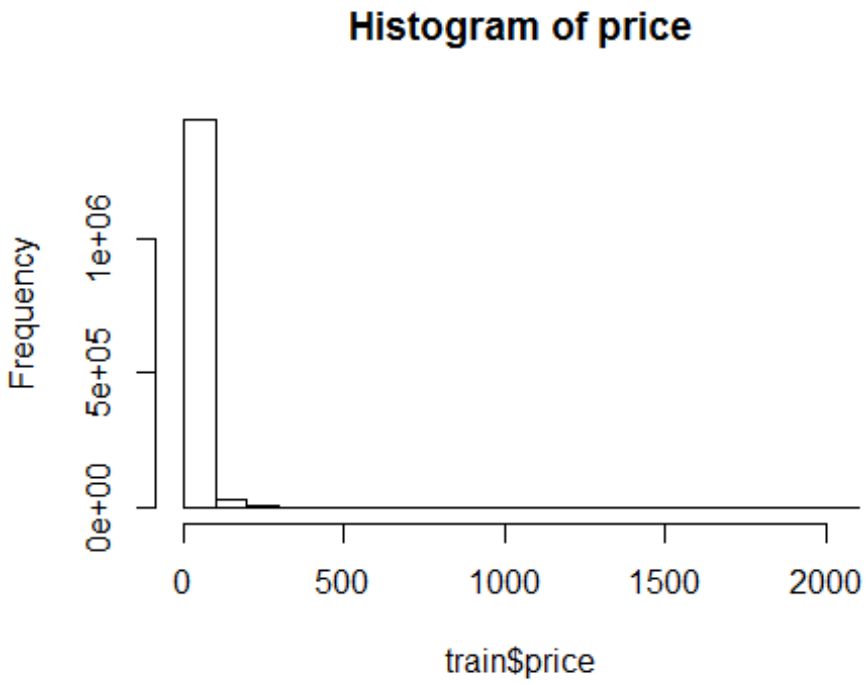
```
ter
##  Mode  :character   Median :  17.00   Median :0.0000   Mode  :charac
ter
##                     Mean   :  26.74   Mean   :0.4473
##                     3rd Qu.:  29.00   3rd Qu.:1.0000
##                     Max.   :2009.00   Max.   :1.0000
```

從提供的 Overview 可以知道欄位定義如下：

- train_id or test_id - the id of the listing
- name - the title of the listing. Note that we have cleaned the data to remove text that look like prices (e.g. $20) to avoid leakage. These removed prices are represented as [rm]
- item_condition_id - the condition of the items provided by the seller
- category_name - category of the listing
- brand_name
- price - the price that the item was sold for. This is the target variable that you will predict. The unit is USD. This column doesn't exist in test.tsv since that is what you will predict.
- shipping - 1 if shipping fee is paid by seller and 0 by buyer
- item_description - the full description of the item. Note that we have cleaned the data to remove text that look like prices (e.g. $20) to avoid leakage. These removed prices are represented as [rm]

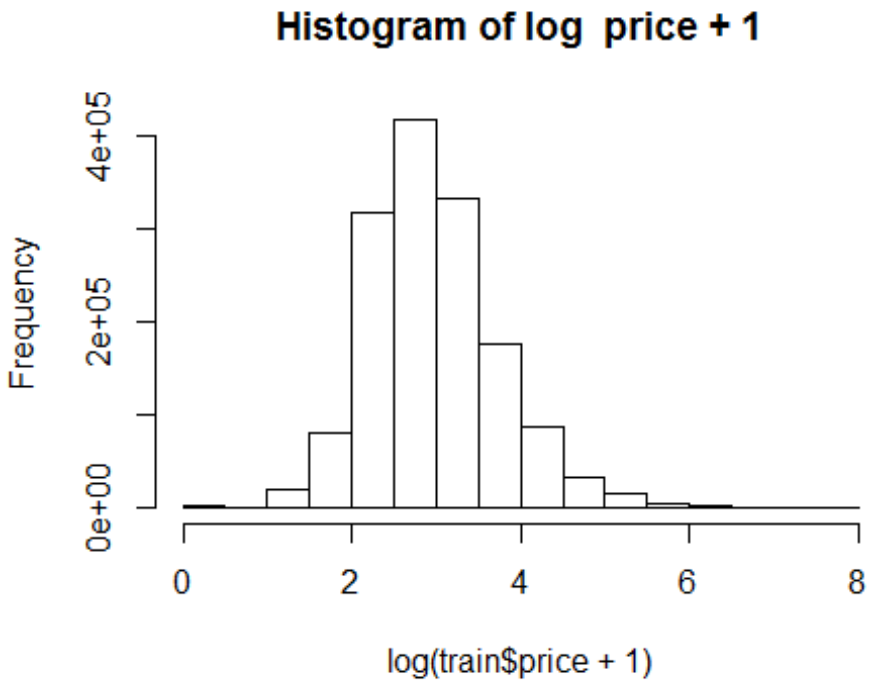
模型為從商品的其他資訊，推測應變數價格 price。首先觀察應變數資料：

```
# 觀察price 資料分布
hist(train$price, main='Histogram of price')
```



資料嚴重右偏，所以用 log(x+1)修正資料，想辦法讓資料偏向常態一點。

```
# 取log(price+1)
hist(log(train$price+1), main='Histogram of log  price + 1')
```



資料看起來正常一點了，後面會讓應變數做轉換，提高預測準確率。接下來我們來對其他欄位進行探勘。

train_id

沒意義的東西就不看了。

item_condition_id

賣方提供商品的狀況，因為沒有 know how，所以這邊只能當單純的 factor 去看它，一樣來統計看看。

```
table(train$item_condition_id)

##
##      1      2      3      4      5
## 640549 375479 432161  31962   2384
```

可以看到大部份的資料集中在 1，5 是最少的。

```
# Y 連續X 類別變數 用anova 進行變異數分析
log_price = log(train$price+1)
aov.item_condition_id <- aov(log_price~ train$item_condition_id)
summary(aov.item_condition_id)
```

```
##                               Df Sum Sq Mean Sq F value Pr(>F)
## train$item_condition_id        1      4   3.576   6.372 0.0116 *
## Residuals               1482533 832165   0.561
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value 小於 0.05，95%的信心水準下有顯著相關，但不到 0.01，這個變數對我們的價格似乎有些微影響。

category_name

這個欄位資料都是文字，並且分類用"/"符號分隔，每個間隔內容看起來是獨立的，這裡把他當成分類貼標來看，所以需要進行文字探勘將它做拆解。

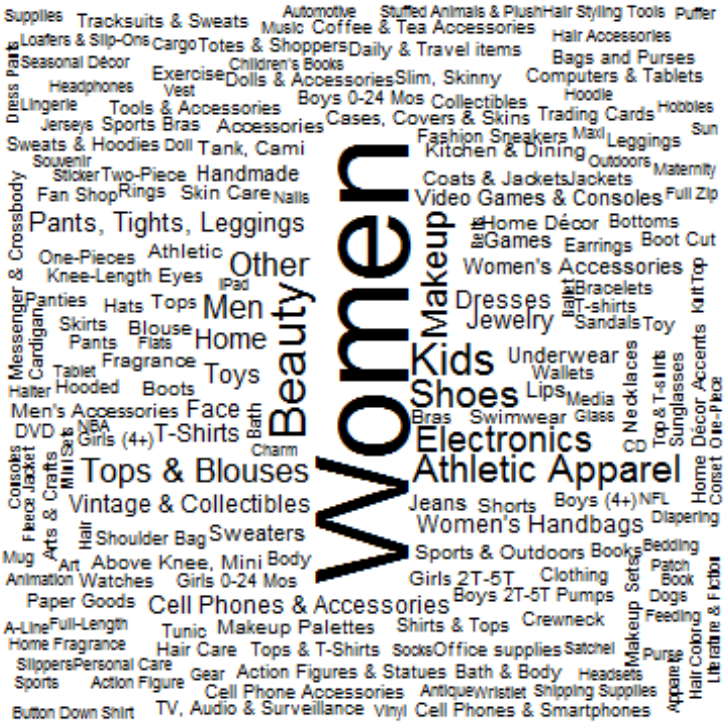
```
# 觀察此欄位內容
head(train$category_name,5)
```

```
## [1] "Men/Tops/T-shirts"
## [2] "Electronics/Computers & Tablets/Components & Parts"
## [3] "Women/Tops & Blouses/Blouse"
## [4] "Home/Home Decor/Home Decor Accents"
## [5] "Women/Jewelry/Necklaces"
```

```
# 把所有貼標轉換成清單統計結果
category_list<-unlist(strsplit(train$category_name,'/'))
category_df<-as.data.frame(table(category_list))
head(category_df[order(category_df$Freq,decreasing = T),],10)
```

```
##        category_list   Freq
## 938            Women 683360
## 85            Beauty 207828
## 492             Kids 171796
## 36  Athletic Apparel 134422
## 757            Shoes 132620
## 315      Electronics 126318
## 544           Makeup 124624
## 613            Other 115592
## 863   Tops & Blouses 107967
## 557              Men  95967
```

```
# 轉化文字雲看一下
wordcloud(category_df$category_list, category_df$Freq, min.freq =500, r
andom.order = F, ordered.colors = F)
```



可以看到女生的衣服貼標數量真的很多，比例上來看也高於其他貼標，研判據有資料參考價值，我們來看如果轉成類別會有多少元素。

```
# 計算有幾列
nrow(category_df)

## [1] 950
```

一個 950 個變數如果要轉成 factor 可能會讓演算法跑不動，到時候可能要轉化稀疏矩陣才能進行分析，我們來看一個商品最多能有幾個貼標。

```
# 計算欄位內"/"的最大值
max(str_count(train$category_name, "/"))

## [1] 4
```

可以把每個商品後面加 5 個欄位，並且根據“/”分隔，將標籤放進去，但是這樣對最後建模沒有太大幫助，最後還是決定轉為 950 個欄位的稀疏矩陣，比較好對應變數建立模型。

brand_name

```
# 統計一下品牌
brand_name_df<-as.data.frame(table(train$brand_name))
head(brand_name_df[order(brand_name_df$Freq,decreasing = T),],10)
```

```
##                   Var1   Freq
## 1                       632682
## 3417              PINK  54088
## 3111              Nike  54043
## 4581 Victoria's Secret  48036
## 2673           LuLaRoe  31024
## 271              Apple  17322
## 1648        FOREVER 21  15186
## 3121          Nintendo  15007
## 2681         Lululemon  14558
## 2885      Michael Kors  13928
```

最多的果然是空着值，畢竟大多商品不是名牌也是很正常的，在來是 PINK、Nike、Victoria's Secret，難怪女裝會這麼多不是沒有原因的，我們一樣來統計名牌的總數：

```
nrow(brand_name_df)
```

```
## [1] 4810
```

高達 4810 種名牌，看來是大大小小的名牌都進來了，我們一樣預計將之轉為稀疏矩陣處理。

shipping

從說明來看似乎跟應變數沒有太大關係，但既然他給了，一定有他的原因，我們一樣做一下多變量分析看一下。

```
table(train$shipping)
```

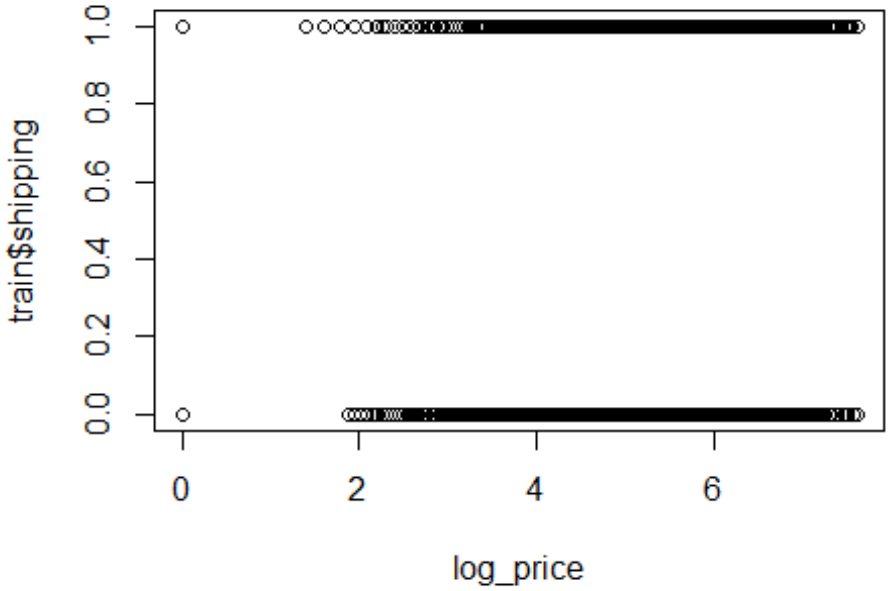
```
##
##      0      1
## 819435 663100
```

```
aov.shipping <- aov(log_price~ train$shipping)
summary(aov.shipping)
```

```
##                     Df Sum Sq Mean Sq F value Pr(>F)
## train$shipping       1  43921   43921   82607 <2e-16 ***
## Residuals      1482533 788247       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

想不到居然有高度相關！我們來看一下統計圖型。

```
plot(aov.shipping$model)
```



分布上來看確實有些差異在，此因子看來也是重要變數。

item_description

這個欄位是對商品的敘述，沒有一定的文字結構，如果用文字探勘進行單字統計，可能會非常的發散，所以我們改用字數來決定這個特徵。

```
# 抓幾筆來看一下
data.frame(item_description=train$item_description[1:5],nchar=nchar(tra
in$item_description[1:5]))

##

                                    item_description
## 1

                                  No description yet
## 2 This keyboard is in great condition and works like it came out of
the box. All of the ports are tested and work perfectly. The lights are
 customizable via the Razer Synapse app on your PC.
## 3                                                                  Ad
orable top with a hint of lace and a key hole in the back! The pale pin
k is a 1X, and I also have a 3X available in white!
## 4                New with tags. Leather horses. Retail for [rm] each.
 Stand about a foot high. They are being sold as a pair. Any questions
please ask. Free shipping. Just got out of storage
```

```
## 5

          Complete with certificate of authenticity
##   nchar
## 1    18
## 2   188
## 3   124
## 4   173
## 5    41

# 檢定字數與價格相關性
cor.test(train$price,nchar(train$item_description))

##
##  Pearson's product-moment correlation
##
## data:  train$price and nchar(train$item_description)
## t = 58.663, df = 1482500, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04651725 0.04972920
## sample estimates:
##        cor
## 0.04812335
```

P-value < 0.05 顯著相關，看來商品描述的字數長度會影響我們的商品價格，為了更精確驗證我們的假設，改根據"單字數"進行統計。

```
# 寫迴圈將單字數計算出來
nword_itdes<-NULL
for ( i in 1:length(train$item_description)){
  nword_itdes[i]<-length(unlist(strsplit(train$item_description[i],'
')))
}

# 將統計結果列出來
data.frame(item_description=train$item_description[1:5],nword=nword_itd
es[1:5])

##

                                    item_description
## 1

                                  No description yet
## 2 This keyboard is in great condition and works like it came out of
the box. All of the ports are tested and work perfectly. The lights are
 customizable via the Razer Synapse app on your PC.
## 3                                                                 Ad
orable top with a hint of lace and a key hole in the back! The pale pin
```

```
k is a 1X, and I also have a 3X available in white!
## 4                New with tags. Leather horses. Retail for [rm] each.
 Stand about a foot high. They are being sold as a pair. Any questions
please ask. Free shipping. Just got out of storage
## 5

          Complete with certificate of authenticity
##   nword
## 1     3
## 2    36
## 3    29
## 4    32
## 5     5
```

```
# 檢定單字數與價格相關性
cor.test(train$price,nword_itdes)
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$price and nword_itdes
## t = 59.197, df = 1482500, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04695493 0.05016674
## sample estimates:
##        cor
## 0.04856096
```

P-value < 0.05 顯著相關，符合我們的假設，後面會以“單字數”進行建模分析，雖然進行細部的文字探勘可以提高模型準確率，但工程浩大所以先到這裡。

name

同上面的商品敘述一樣，我們將之轉為單字數統計：

```
# 寫迴圈將單字數計算出來
nword_name<-NULL
for ( i in 1:length(train$name)){
  nword_name[i]<-length(unlist(strsplit(train$name[i],' ')))
}

# 將統計結果列出來
data.frame(name=train$name[1:5],nword=nword_name[1:5])
```

```
##                                   name nword
## 1 MLB Cincinnati Reds T Shirt Size XL     7
## 2    Razer BlackWidow Chroma Keyboard     4
## 3                      AVA-VIV Blouse     2
```

```
## 4                Leather Horse Statues     3
## 5                 24K GOLD plated rose     4
```

```
# 檢定單字數與價格相關性
cor.test(train$price,nword_name)
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$price and nword_name
## t = 42.064, df = 1482500, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03291863 0.03613419
## sample estimates:
##       cor
## 0.0345265
```

P-value < 0.05 顯著相關，看來商品名稱的長度與價格有顯著關係，後面也會以“單字數”進行建模分析。商品名稱的內容通常會影響著價格，但這樣要進行更細部的文字探勘，由於工程耗大所以我們先做到這樣。

Data processing

從前面 EDA 來看，我們有了初步建模計劃：

- name：轉為單字數。
- item_condition_id：保留原本值。
- category_name：轉為稀疏矩陣，950 個。
- brand_name：轉為稀疏矩陣，4810 個。
- shipping：保留原本值。
- item_description：轉為單字數。