

SHAO

shao

Friday, May 15, 2015

Preliminary Exploration

Load the data, and take a look.

```
train <- read.csv("../data/train.csv")
test <- read.csv("../data/test.csv")
dim(train)
```

```
## [1] 10506    12
```

```
dim(test)
```

```
## [1] 116293    11
```

```
head(train)
```

```
##           Date                               Address
## 1 2007-05-29 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 2 2007-05-29 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 3 2007-05-29 6200 North Mandell Avenue, Chicago, IL 60646, USA
## 4 2007-05-29 7900 West Foster Avenue, Chicago, IL 60656, USA
## 5 2007-05-29 7900 West Foster Avenue, Chicago, IL 60656, USA
## 6 2007-05-29 1500 West Webster Avenue, Chicago, IL 60614, USA
##           Species Block           Street Trap
## 1 CULEX PIPIENS/RESTUANS    41  N OAK PARK AVE T002
## 2           CULEX RESTUANS    41  N OAK PARK AVE T002
## 3           CULEX RESTUANS    62  N MANDELL AVE T007
## 4 CULEX PIPIENS/RESTUANS    79   W FOSTER AVE T015
## 5           CULEX RESTUANS    79   W FOSTER AVE T015
## 6           CULEX RESTUANS    15  W WEBSTER AVE T045
##           AddressNumberAndStreet Latitude Longitude AddressAccuracy
## 1 4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099           9
## 2 4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099           9
## 3 6200  N MANDELL AVE, Chicago, IL 41.99499 -87.76928           9
## 4 7900  W FOSTER AVE, Chicago, IL 41.97409 -87.82481           8
## 5 7900  W FOSTER AVE, Chicago, IL 41.97409 -87.82481           8
## 6 1500  W WEBSTER AVE, Chicago, IL 41.92160 -87.66645           8
##           NumMosquitos WnvPresent
## 1                1          0
## 2                1          0
## 3                1          0
## 4                1          0
## 5                4          0
## 6                2          0
```

```
head(test, 10)
```

```
##      Id      Date      Address
## 1    1 2008-06-11 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 2    2 2008-06-11 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 3    3 2008-06-11 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 4    4 2008-06-11 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 5    5 2008-06-11 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 6    6 2008-06-11 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 7    7 2008-06-11 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 8    8 2008-06-11 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 9    9 2008-06-11 6200 North Mandell Avenue, Chicago, IL 60646, USA
## 10 10 2008-06-11 6200 North Mandell Avenue, Chicago, IL 60646, USA
##      Species Block      Street Trap
## 1  CULEX PIPIENS/RESTUANS  41  N OAK PARK AVE T002
## 2      CULEX RESTUANS  41  N OAK PARK AVE T002
## 3      CULEX PIPIENS  41  N OAK PARK AVE T002
## 4      CULEX SALINARIUS  41  N OAK PARK AVE T002
## 5      CULEX TERRITANS  41  N OAK PARK AVE T002
## 6      CULEX TARSALIS  41  N OAK PARK AVE T002
## 7      UNSPECIFIED CULEX  41  N OAK PARK AVE T002
## 8      CULEX ERRATICUS  41  N OAK PARK AVE T002
## 9  CULEX PIPIENS/RESTUANS  62  N MANDELL AVE T007
## 10      CULEX RESTUANS  62  N MANDELL AVE T007
##      AddressNumberAndStreet Latitude Longitude AddressAccuracy
## 1  4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099      9
## 2  4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099      9
## 3  4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099      9
## 4  4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099      9
## 5  4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099      9
## 6  4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099      9
## 7  4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099      9
## 8  4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099      9
## 9   6200  N MANDELL AVE, Chicago, IL 41.99499 -87.76928      9
## 10  6200  N MANDELL AVE, Chicago, IL 41.99499 -87.76928      9
```

```
str(train, strict.width="cut")
```

```
## 'data.frame':  10506 obs. of  12 variables:
## $ Date      : Factor w/ 95 levels "2007-05-29","2007-06-05",...
## $ Address   : Factor w/ 138 levels "1000 East 67th Street, "...
## $ Species   : Factor w/ 7 levels "CULEX ERRATICUS",...: 3 4 4..
## $ Block     : int  41 41 62 79 79 15 25 11 11 11 ...
## $ Street    : Factor w/ 128 levels " W ARMITAGE AVENUE",...:
## $ Trap      : Factor w/ 136 levels "T001","T002",...: 2 2 7 1..
## $ AddressNumberAndStreet: Factor w/ 138 levels "1000 E 67TH ST, Chicag"..
## $ Latitude  : num  42 42 42 42 42 ...
## $ Longitude : num  -87.8 -87.8 -87.8 -87.8 -87.8 ...
## $ AddressAccuracy : int  9 9 9 8 8 8 8 8 8 8 ...
## $ NumMosquitos : int  1 1 1 1 4 2 1 1 2 1 ...
## $ WnvPresent  : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
str(test, strict.width="cut")
```

```
## 'data.frame': 116293 obs. of 11 variables:
## $ Id : int 1 2 3 4 5 6 7 8 9 10 ...
## $ Date : Factor w/ 95 levels "2008-06-11","2008-06-17",...
## $ Address : Factor w/ 151 levels "1000 East 67th Street, "..
## $ Species : Factor w/ 8 levels "CULEX ERRATICUS",...: 3 4 2..
## $ Block : int 41 41 41 41 41 41 41 41 62 62 ...
## $ Street : Factor w/ 139 levels " W ARMITAGE AVENUE",...:..
## $ Trap : Factor w/ 149 levels "T001","T002",...: 2 2 2 2..
## $ AddressNumberAndStreet: Factor w/ 151 levels "1000 E 67TH ST, Chicag"..
## $ Latitude : num 42 42 42 42 42 ...
## $ Longitude : num -87.8 -87.8 -87.8 -87.8 -87.8 ...
## $ AddressAccuracy : int 9 9 9 9 9 9 9 9 9 9 ...
```

And there is no missing value.

```
any(is.na(train) == TRUE)
```

```
## [1] FALSE
```

```
any(is.na(test) == TRUE)
```

```
## [1] FALSE
```

Feature Engineering

Date

```
train$Date <- as.Date(train$Date)
test$Date <- as.Date(test$Date)
```

Naturally, we create four features Year, Month, Week and Weekday based on Date.

```
library(lubridate)
## train set
train$Year <- as.integer(year(train$Date))
train$Month <- factor(months(train$Date),
                      labels=c("May", "June", "July", "August", "September", "October"))
train$Week <- week(train$Date)
train$Weekday <- factor(weekdays(train$Date),
                        labels=c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"))
## test set
test$Year <- as.integer(year(test$Date))
test$Month <- factor(months(test$Date),
                     labels=c("June", "July", "August", "September", "October"))
test$Week <- week(test$Date)
test$Weekday <- factor(weekdays(test$Date),
```

```

labels=c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"))
## take a look
head(train[,c("Date", "Year", "Month", "Week", "Weekday")])

```

```

##      Date Year  Month Week Weekday
## 1 2007-05-29 2007 August   22 Thursday
## 2 2007-05-29 2007 August   22 Thursday
## 3 2007-05-29 2007 August   22 Thursday
## 4 2007-05-29 2007 August   22 Thursday
## 5 2007-05-29 2007 August   22 Thursday
## 6 2007-05-29 2007 August   22 Thursday

```

```

head(test[,c("Date", "Year", "Month", "Week", "Weekday")])

```

```

##      Date Year  Month Week Weekday
## 1 2008-06-11 2008 August   24 Friday
## 2 2008-06-11 2008 August   24 Friday
## 3 2008-06-11 2008 August   24 Friday
## 4 2008-06-11 2008 August   24 Friday
## 5 2008-06-11 2008 August   24 Friday
## 6 2008-06-11 2008 August   24 Friday

```

```

str(train[,c("Date", "Year", "Month", "Week", "Weekday")])

```

```

## 'data.frame':  10506 obs. of  5 variables:
## $ Date   : Date, format: "2007-05-29" "2007-05-29" ...
## $ Year    : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
## $ Month   : Factor w/ 6 levels "May","June","July",...: 4 4 4 4 4 4 4 4 4 4 4 ...
## $ Week    : num  22 22 22 22 22 22 22 22 22 22 22 ...
## $ Weekday : Factor w/ 5 levels "Monday","Tuesday",...: 4 4 4 4 4 4 4 4 4 4 4 ...

```

```

str(test[,c("Date", "Year", "Month", "Week", "Weekday")])

```

```

## 'data.frame':  116293 obs. of  5 variables:
## $ Date   : Date, format: "2008-06-11" "2008-06-11" ...
## $ Year    : int  2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...
## $ Month   : Factor w/ 5 levels "June","July",...: 3 3 3 3 3 3 3 3 3 3 3 ...
## $ Week    : num  24 24 24 24 24 24 24 24 24 24 24 ...
## $ Weekday : Factor w/ 5 levels "Monday","Tuesday",...: 5 5 5 5 5 5 5 5 5 5 5 ...

```

Species

As a factor, Species has 7 levels in train set, but 8 levels in test set. The extra one in test set is “UNSPECIFIED CULEX”.

```

table(train$Species)

```

```

##
##      CULEX ERRATICUS      CULEX PIPIENS CULEX PIPIENS/RESTUANS

```

```
##           1           2699           4752
##      CULEX RESTUANS      CULEX SALINARIUS      CULEX TARSALIS
##           2740           86           6
##      CULEX TERRITANS
##           222
```

```
table(test$Species)
```

```
##
##      CULEX ERRATICUS      CULEX PIPIENS CULEX PIPIENS/RESTUANS
##           14345           14521           15359
##      CULEX RESTUANS      CULEX SALINARIUS      CULEX TARSALIS
##           14670           14355           14347
##      CULEX TERRITANS      UNSPECIFIED CULEX
##           14351           14345
```

The “UNSPECIFIED CULEX” does not appear in the train set. We need a mechanism to specify it to another level, which is one of the 7 levels in train set.

NumMosquitos

These train results are organized in such a way that when the number of mosquitos exceed 50, they are split into another record (another row in the dataset), such that the number of mosquitos are capped at 50. For the test set, it is the same, while NumMosquitos does not appear as a predictor.

This section we combine the same record into one row, instead of several rows since its NumMosquitos > 50.

```
library(plyr)
```

```
##
## Attaching package: 'plyr'
##
## The following object is masked from 'package:lubridate':
##
##      here
```

```
train <- ddply(train,
               .(Date, Address, Species, Block, Street, Trap,
                 AddressNumberAndStreet, Latitude, Longitude,
                 AddressAccuracy, Year, Month, Week, Weekday),
               summarize,
               #      WnvPresent = sum(WnvPresent*NumMosquitos),
               WnvPresent = as.integer(as.logical(sum(WnvPresent))),
               NumMosquitos = sum(NumMosquitos)
               )
## show
head(train[train$NumMosquitos > 50,])
```

```
##      Date      Address
## 292 2007-07-11 2200 West 113th Street, Chicago, IL 60643, USA
## 312 2007-07-11 7200 North Oketo Avenue, Chicago, IL 60631, USA
## 342 2007-07-11 3500 West 116th Street, Chicago, IL 60655, USA
```

```
## 511 2007-07-18      South Stony Island Avenue, Chicago, IL, USA
## 512 2007-07-18      South Stony Island Avenue, Chicago, IL, USA
## 524 2007-07-18      3700 118th Street, Chicago, IL 60617, USA
##              Species Block              Street Trap
## 292 CULEX PIPIENS/RESTUANS      22      W 113TH ST T086
## 312 CULEX PIPIENS/RESTUANS      72      N OKETO AVE T016
## 342 CULEX PIPIENS/RESTUANS      35      W 116TH ST T158
## 511 CULEX PIPIENS/RESTUANS      10  S STONY ISLAND AVE T138
## 512      CULEX RESTUANS      10  S STONY ISLAND AVE T138
## 524 CULEX PIPIENS/RESTUANS      37      E 118TH ST T212
##              AddressNumberAndStreet Latitude Longitude
## 292      2200  W 113TH ST, Chicago, IL 41.68832 -87.67671
## 312      7200  N OKETO AVE, Chicago, IL 42.01160 -87.81151
## 342      3500  W 116TH ST, Chicago, IL 41.68259 -87.70797
## 511 1000  S STONY ISLAND AVE, Chicago, IL 41.72646 -87.58541
## 512 1000  S STONY ISLAND AVE, Chicago, IL 41.72646 -87.58541
## 524      3700  E 118TH ST, Chicago, IL 41.68095 -87.53520
##      AddressAccuracy Year Month Week Weekday WnvPresent NumMosquitos
## 292      8 2007   June   28  Friday      0      143
## 312      8 2007   June   28  Friday      0      51
## 342      9 2007   June   28  Friday      0     189
## 511      5 2007   June   29  Friday      0     100
## 512      5 2007   June   29  Friday      0      76
## 524      8 2007   June   29  Friday      0     102
```

Trap

First, check out the names of all traps.

```
unique(train$Trap)
```

```
##      [1] T002 T007 T015 T045 T046 T048 T049 T050 T054 T086 T091
##      [12] T094 T096 T129 T143 T148 T153 T159 T009 T011 T016 T019
##      [23] T025 T028 T031 T033 T089 T090 T092 T135 T141 T142 T145
##      [34] T146 T147 T149 T150 T151 T152 T154 T158 T162 T218 T220
##      [45] T001 T003 T006 T008 T012 T034 T037 T040 T043 T047 T051
##      [56] T085 T088 T161 T219 T013 T014 T018 T030 T084 T144 T160
##      [67] T005 T017 T044 T095 T004 T035 T036 T039 T060 T061 T062
##      [78] T065 T066 T067 T069 T070 T071 T073 T074 T075 T076 T077
##      [89] T079 T080 T081 T082 T083 T114 T155 T063 T115 T138 T200
##     [100] T206 T209 T212 T215 T107 T128 T072 T078 T097 T099 T100
##     [111] T102 T103 T027 T156 T157 T221 T900 T903 T222 T223 T225
##     [122] T227 T224 T226 T229 T230 T228 T232 T231 T235 T233 T236
##     [133] T237 T238 T094B T054C
## 136 Levels: T001 T002 T003 T004 T005 T006 T007 T008 T009 T011 T012 ... T903
```

```
unique(test$Trap)
```

```
##      [1] T002 T007 T015 T045 T046 T048 T049 T050 T054 T086 T091
##      [12] T094 T096 T129 T143 T148 T153 T159 T009 T011 T016 T019
##      [23] T025 T028 T031 T033 T089 T090 T092 T135 T141 T142 T145
##      [34] T146 T147 T149 T150 T151 T152 T154 T158 T162 T218 T220
```

```
## [45] T001 T003 T006 T008 T012 T034 T037 T040 T043 T047 T051
## [56] T085 T088 T161 T219 T013 T014 T018 T030 T084 T144 T160
## [67] T005 T017 T044 T095 T004 T035 T036 T039 T060 T061 T062
## [78] T065 T066 T067 T069 T070 T071 T073 T074 T075 T076 T077
## [89] T079 T080 T081 T082 T083 T114 T155 T063 T115 T138 T200
## [100] T206 T209 T212 T215 T107 T128 T072 T078 T097 T099 T100
## [111] T102 T103 T027 T156 T157 T221 T900 T903 T090A T090B T090C
## [122] T200A T128A T200B T218A T218C T218B T222 T223 T225 T227 T224
## [133] T226 T229 T230 T228 T231 T232 T002A T002B T233 T234 T235
## [144] T236 T237 T238 T065A T094B T054C
## 149 Levels: T001 T002 T002A T002B T003 T004 T005 T006 T007 T008 ... T903
```

Through the observation, we know that initial “T” is shared by all traps’ names. The single letter after the 3 digits means this trap is a satellite trap of the main trap with the same 3 digits in the name.

Two new features can be naturally generated from this Trap feature:

- The main trap number (TrapNumber): the 3 digits;
- Main or satellite trap (TrapMS): the single letter after digits in Trap. If it is the main trap, there is no single letter. We label it “M”.

```
## train set
train$TrapNumber <- as.integer(substr(as.character(train$Trap), 2, 4))
train$TrapMS <- as.factor(substr(as.character(train$Trap), 5, 5))
levels(train$TrapMS) <- c("M", "B", "C")
## test set
test$TrapNumber <- as.integer(substr(as.character(test$Trap), 2, 4))
test$TrapMS <- as.factor(substr(as.character(test$Trap), 5, 5))
levels(test$TrapMS) <- c("M", "A", "B", "C")
## show
head(train[,c("Trap", "TrapNumber", "TrapMS")])
```

```
##   Trap TrapNumber TrapMS
## 1 T002          2      M
## 2 T002          2      M
## 3 T007          7      M
## 4 T015         15      M
## 5 T015         15      M
## 6 T045         45      M
```

```
head(test[,c("Trap", "TrapNumber", "TrapMS")])
```

```
##   Trap TrapNumber TrapMS
## 1 T002          2      M
## 2 T002          2      M
## 3 T002          2      M
## 4 T002          2      M
## 5 T002          2      M
## 6 T002          2      M
```

Save this data set

```
## write.csv(train, "data/train3.csv", row.names=F)
## write.csv(test, "data/test2.csv", row.names=F)
```

Data Exploration

The distribution of blocks

It seems like the distribution of blocks does not have a very regular sequence.

```
library(ggplot2)
library(ggmap)
mapdata <- readRDS("../data/mapdata_copyright_openstreetmap_contributors.rds")
ggmap(mapdata) + geom_point(data=train, aes(x=Longitude,y=Latitude, size=Block))
```

