SmartFly: Exploratory Data Analysis

Cindy Lamm

January 12, 2015

First load variable names and types of historic data (prepared in an additional csv file):

```
nameTypeDataFile  <- "resources/raw_variables.csv"
variableNames <- read.csv(nameTypeDataFile, header=TRUE, stringsAsFactors=FALSE)
variableNames

##                        name      type
## 1                        id character
## 2                      year    factor
## 3                     month    factor
## 4              day_of_month    factor
## 5               day_of_week    factor
## 6  scheduled_departure_time    factor
## 7    scheduled_arrival_time    factor
## 8                   airline    factor
## 9             flight_number    factor
## 10              tail_number    factor
## 11              plane_model    factor
## 12       seat_configuration    factor
## 13          departure_delay   numeric
## 14           origin_airport    factor
## 15      destination_airport    factor
## 16       distance_travelled   numeric
## 17             taxi_time_in   numeric
## 18            taxi_time_out   numeric
## 19                cancelled   integer
## 20        cancellation_code    factor
```

Then load historic data into R:

```
historicDataFile <- "../../data/smartfly_historic.csv"
#historicDataFile <- "../../data/mod4000.csv"
trainDataTyped <- read.csv(historicDataFile, header=FALSE, stringsAsFactors=FALSE,
                           col.names=variableNames$name, colClasses=variableNames$type)
# convert integer to logical
trainDataTyped$cancelled <- as.logical(trainDataTyped$cancelled)
```

Checkout first 10 historic data rows

```
head(trainDataTyped)

##                      id year month day_of_month day_of_week scheduled_departure_time
## 1 4982598272866526024 2013     8           11           7                     1015
## 2 5074130684343212714 2013     8           17           6                     1015
## 3 8872634703988349126 2013     8           18           7                     1015
## 4 1147433994031419585 2013     8           24           6                     1015
## 5  739211944918463275 2013     8           25           7                     1015
## 6 7526342364355579297 2013     8           31           6                     1015
##   scheduled_arrival_time airline flight_number tail_number plane_model seat_configuration
## 1                   1132      US           923      N728UW         757        Three Class
## 2                   1132      US           923      N746UW         757           Standard
## 3                   1132      US           923      N706UW         787                 V1
## 4                   1132      US           923      N707UW         747                 V2
## 5                   1132      US           923      N758UW         787                 V1
## 6                   1132      US           923      N702UW         747                 V2
##   departure_delay origin_airport destination_airport distance_travelled taxi_time_in
## 1              -5            BWI                 CLT                361            9
## 2               5            BWI                 CLT                361            7
## 3              -4            BWI                 CLT                361            6
## 4              -6            BWI                 CLT                361           15
## 5              -3            BWI                 CLT                361            7
## 6              -8            BWI                 CLT                361            5
##   taxi_time_out cancelled cancellation_code
## 1            11     FALSE              <NA>
## 2             7     FALSE              <NA>
## 3             9     FALSE              <NA>
## 4            11     FALSE              <NA>
## 5            12     FALSE              <NA>
## 6            15     FALSE              <NA>
```

and a summary of the historic data:

```
summary(trainDataTyped)
```

```
##       id              year             month          day_of_month     day_of_week
##  Length:7374365     2013:2185499   8      :1023748   13     : 252615   1:1079862
##  Class :character   2014:5188866   9      : 957710   6      : 252560   2:1063516
##  Mode  :character                  10     : 782952   3      : 252160   3:1069847
##                                    3      : 559342   17     : 251944   4:1096825
##                                    7      : 558568   16     : 250869   5:1096417
##                                    1      : 552109   2      : 250647   6: 935465
##                                    (Other):2939936   (Other):5863570   7:1032433
##  scheduled_departure_time scheduled_arrival_time    airline        flight_number
##  700    : 105996          1810   :  21315        WN     :1171236   192    :   5702
##  800    :  74502          1715   :  21191        AA     : 960866   64     :   5639
##  600    :  66567          1215   :  21074        DL     : 825543   706    :   5409
##  900    :  65778          1615   :  21048        UA     : 686409   186    :   5373
##  630    :  60479          1605   :  20639        NW     : 619091   751    :   5209
##  1700   :  56619          1630   :  20359        US     : 529032   340    :   5060
##  (Other):6944424          (Other):7248739        (Other):2582188   (Other):7341973
##   tail_number      plane_model      seat_configuration  departure_delay
##         :  42213   737 :2317735   Standard   :2130560   Min.   :-1410.00
##  0      :  17138   747 :1579936   Three Class: 779700   1st Qu.:   -4.00
##  000000 :  10157   757 : 999512   Two Class  : 779964   Median :    0.00
##  N183UW :   4694   777 : 634170   V1         :1430984   Mean   :    4.87
##  N80    :   4290   787 : 633182   V2         :1105044   3rd Qu.:    2.00
##  N96    :   4269   A320:1209830   V3         :1148113   Max.   : 2119.00
##  (Other):7291604                                        NA's   :104127
##  origin_airport    destination_airport distance_travelled  taxi_time_in
##  ORD    : 431004   ORD    : 431004     Min.   :  11       Min.   :   0.000
##  ATL    : 389963   ATL    : 389886     1st Qu.: 308       1st Qu.:   4.000
##  DFW    : 382123   DFW    : 382349     Median : 569       Median :   5.000
##  LAX    : 255642   LAX    : 255786     Mean   : 726       Mean   :   6.808
##  PHX    : 209831   PHX    : 209839     3rd Qu.: 964       3rd Qu.:   7.000
##  IAH    : 195923   IAH    : 195926     Max.   :4962       Max.   :1495.000
##  (Other):5509879   (Other):5509575
##  taxi_time_out     cancelled       cancellation_code
##  Min.   :   0.00   Mode :logical       :2484977
##  1st Qu.:  10.00   FALSE:7270238   A    :  14587
##  Median :  13.00   TRUE :104127    B    :   8072
##  Mean   :  15.05   NA's :0         C    :   8309
##  3rd Qu.:  18.00                   D    :    179
##  Max.   :1439.00                   NA's:4858241
##
```

What are the factor levels for the different variables:

```
all_levels <- sapply(1:20, FUN=function(x) {levels(trainDataTyped[,x])})
names(all_levels) <- variableNames$name
number_of_levels <- unlist(lapply(all_levels, FUN=length))
number_of_levels
```

```
##                       id                     year                    month
##                        0                        2                       12
##             day_of_month              day_of_week scheduled_departure_time
##                       31                        7                     1190
##   scheduled_arrival_time                  airline            flight_number
##                     1323                       17                     6889
##              tail_number              plane_model       seat_configuration
##                     5036                        6                        6
##          departure_delay           origin_airport      destination_airport
##                        0                      279                      279
##       distance_travelled             taxi_time_in            taxi_time_out
##                        0                        0                        0
##                cancelled        cancellation_code
##                        0                        5
```

```
lapply(all_levels, FUN=head, n=10)
```

```
## $id
## NULL
##
## $year
## [1] "2013" "2014"
##
## $month
##  [1] "1"  "10" "11" "12" "2"  "3"  "4"  "5"  "6"  "7"
##
## $day_of_month
##  [1] "1"  "10" "11" "12" "13" "14" "15" "16" "17" "18"
##
## $day_of_week
## [1] "1" "2" "3" "4" "5" "6" "7"
##
## $scheduled_departure_time
##  [1] "0"    "10"   "100"  "1000" "1001" "1002" "1003" "1004" "1005" "1006"
##
## $scheduled_arrival_time
##  [1] "0"    "1"    "10"   "100"  "1000" "1001" "1002" "1003" "1004" "1005"
##
## $airline
##  [1] "AA" "AS" "B6" "CO" "DH" "DL" "EV" "FL" "HP" "MQ"
##
## $flight_number
##  [1] "1"    "10"   "100"  "1000" "1001" "1002" "1003" "1004" "1005" "1006"
##
## $tail_number
##  [1] ""        "0"      "000000" "N050AA" "N051AA" "N052AA" "N054AA" "N055AA" "N056AA"
## [10] "N057AA"
##
## $plane_model
## [1] "737"  "747"  "757"  "777"  "787"  "A320"
##
## $seat_configuration
## [1] "Standard"    "Three Class" "Two Class"   "V1"          "V2"          "V3"
##
## $departure_delay
## NULL
##
## $origin_airport
##  [1] "ABE" "ABI" "ABQ" "ABY" "ACK" "ACT" "ACV" "ACY" "ADK" "ADQ"
##
## $destination_airport
##  [1] "ABE" "ABI" "ABQ" "ABY" "ACK" "ACT" "ACV" "ACY" "ADK" "ADQ"
##
## $distance_travelled
## NULL
##
## $taxi_time_in
## NULL
##
## $taxi_time_out
## NULL
##
## $cancelled
## NULL
##
## $cancellation_code
## [1] ""  "A" "B" "C" "D"
```

(Note: The number of levels matters if we would want to create a dummy variable for each level. With lots of levels the number of variables would be HUGE and so would be the sparsity of the design matrix.)

Conclusions:

- `tail_number` and `cancellation_code` contain spaces that need to be converted to `NA`
- `scheduled_departure_time` (and probably `scheduled_arrival_time` as well) need to be prefixed

with 0 for the 3-character strings (assuming "900" means "0900" and thus a time of 09h00)

- `departure_delay` needs to be converted to a binary factor (based on definition that only positive delay counts as "delayed")

Create binary target variable `is_delayed`:

```
trainDataTyped$is_delayed <- factor(trainDataTyped$departure_delay > 0
                                    & trainDataTyped$cancelled==FALSE,
                                    labels= c("delayed", "on_time"))
summary(trainDataTyped$is_delayed)

## delayed on_time
## 5263866 2110499
```

Look at correlations between continuous variables:

```
cor(trainDataTyped$departure_delay, trainDataTyped$taxi_time_in, use="pairwise.complete.obs")

## [1] 0.03345877

cor(trainDataTyped$departure_delay, trainDataTyped$taxi_time_out, use="pairwise.complete.obs")

## [1] 0.06387488

cor(trainDataTyped$departure_delay, trainDataTyped$distance_travelled, use="pairwise.complete.obs")

## [1] -0.0007718446
```

Look at some dependency between the binary target variable and other factor variables using the Chi-Square test of independence. The null hypothesis is that the two variables are independent, which we reject if the p-value is smaller than $\alpha = 0.001(chosensosmallduetolargesamplesize)$:

```
suitable <- intersect(which(variableNames$type=="factor"),
                      which(number_of_levels < 20))
test_independence <- function(col, alpha){
  result <- suppressWarnings(
      chisq.test(table(trainDataTyped$is_delayed, trainDataTyped[,col]))
      )
  return(result$p.value < alpha)
}
dependent_with_target <- sapply(suitable, FUN=test_independence, alpha=ALPHA)
names(dependent_with_target) <- variableNames$name[suitable]
dependent_with_target

##               year              month        day_of_week            airline
##               TRUE               TRUE               TRUE               TRUE
##        plane_model seat_configuration  cancellation_code
##               TRUE               TRUE               TRUE
```

Save data frame for next step:

```
save(trainDataTyped, file="trainDataTyped.Rdata")
```