Uber Take Home

Jeffrey Uslan

November 1, 2015

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(lubridate)
library(rpart)
library(randomForest)
```

```
## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(zoo)
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(caret)
```

```
## Loading required package: lattice
## Loading required package: ggplot2
```

```r
library(pander)
```

```r
 load(file="uber_test_data.rda")
```

```r
data=tbl_df(uber_unlist)
data$trips_in_first_30_days=as.numeric(as.character(data$trips_in_first_30_days))
data$signup_date=as.Date(data$signup_date)
data$avg_rating_of_driver=as.numeric(as.character(data$avg_rating_of_driver))
data$avg_surge=as.numeric(as.character(data$avg_surge))
data$last_trip_date=as.Date(data$last_trip_date)
data$surge_pct=as.numeric(as.character(data$surge_pct))
data$weekday_pct=as.numeric(as.character(data$weekday_pct))
data$avg_dist=as.numeric(as.character(data$avg_dist))
data$avg_rating_by_driver=as.numeric(as.character(data$avg_rating_by_driver))
```

```r
numeric_data_inds=sapply(data,is.numeric)
cat_data_inds=!sapply(data,is.numeric)

#tabulations of categorical data
pander(data %>% group_by(city) %>% summarise(Count=n()))
```

city	Count
King's Landing	10130
Astapor	16534
Winterfell	23336

```r
pander(data %>% group_by(phone) %>% summarise(Count=n()))
```

phone	Count
iPhone	34582
Android	15022
NA	396

```r
pander(data %>% group_by(uber_black_user) %>% summarise(Count=n()))
```

uber_black_user	Count
TRUE	18854
FALSE	31146

```
#generating retained variable
data$retained=0
data$retained[which(data$trips_in_first_30_days>0)]=1
mean(data$retained,na.rm=TRUE)
```

```
## [1] 0.6922
```

```
data$retained=as.factor(data$retained)
data %>% group_by(retained) %>% summarise(Count=n())
```

```
## Source: local data frame [2 x 2]
##
##   retained Count
## 1        0 15390
## 2        1 34610
```

numeric exploration

```
summary(data[,numeric_data_inds])
```

```
##  trips_in_first_30_days avg_rating_of_driver   avg_surge
##  Min.   :  0.000        Min.   :1.000        Min.   :1.000
##  1st Qu.:  0.000        1st Qu.:4.300        1st Qu.:1.000
##  Median :  1.000        Median :4.900        Median :1.000
##  Mean   :  2.278        Mean   :4.602        Mean   :1.075
##  3rd Qu.:  3.000        3rd Qu.:5.000        3rd Qu.:1.050
##  Max.   :125.000        Max.   :5.000        Max.   :8.000
##                         NA's   :8122
##    surge_pct       weekday_pct        avg_dist       avg_rating_by_driver
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :1.000
##  1st Qu.:  0.00   1st Qu.: 33.30   1st Qu.:  2.420   1st Qu.:4.700
##  Median :  0.00   Median : 66.70   Median :  3.880   Median :5.000
##  Mean   :  8.85   Mean   : 60.93   Mean   :  5.797   Mean   :4.778
##  3rd Qu.:  8.60   3rd Qu.:100.00   3rd Qu.:  6.940   3rd Qu.:5.000
##  Max.   :100.00   Max.   :100.00   Max.   :160.960   Max.   :5.000
##                                                      NA's   :201
```

covariates

```
covariates=c("city","phone","uber_black_user","avg_rating_of_driver",
             "avg_surge","surge_pct","weekday_pct","avg_dist","avg_rating_by_driver")
```