R Notebook: random forest on dataset with classes - full modeal

```
library(caret)
```

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

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

```
library(ggplot2)
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::combine()       masks randomForest::combine()
## x dplyr::filter()        masks stats::filter()
## x dplyr::lag()           masks stats::lag()
## x purrr::lift()          masks caret::lift()
## x randomForest::margin() masks ggplot2::margin()
```

```
setwd("~/CSP571ProjectGroup")
df <- read_csv("df_with_class.csv")
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 114660 Columns: 17
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): number, incident_state, sys_updated_by, contact_type, category, su...
## dbl  (6): ...1, reassignment_count, reopen_count, sys_mod_count, problem_id,...
## lgl  (1): made_sla
```

```
##
## 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.
```

```
df %>% group_by(class) %>% summarise(n = n())
```

```
## # A tibble: 19 x 2
##    class       n
##    <chr>   <int>
##  1 >28days  5807
##  2 0mins   27618
##  3 10days   3499
##  4 10mins   2326
##  5 14days   5356
##  6 1day     9859
##  7 1hr      2934
##  8 28days   6425
##  9 2days    8289
## 10 30mins   4331
## 11 3days    5244
## 12 3hr      5316
## 13 4days    4079
## 14 5days    3609
## 15 5mins    5890
## 16 6days    3551
## 17 6hr      3928
## 18 7days    3617
## 19 8days    2982
```

```
df = subset(df, select=-c(sys_updated_by, number, subcategory, resolved_updated_diff))
df = subset(df, select=-1)
df$class = as.factor(df$class)
```

```
# try on the full data set (df_small from previous version of the code)
df_small = df
train_idx = createDataPartition(df_small$class, p = 0.8, list=F)
df_train = df_small[train_idx,]
df_test = df_small[-train_idx,]
# try a randomforest model
rf_mod = randomForest(class~., data=df_train, ntree=1000)
df_test_X = subset(df_test, select=-c(class))
df_test_y = df_test$class
pred = predict(rf_mod, df_test)


columns = c("pred", "actual")
acc_df = data.frame(matrix(nrow=nrow(df_test), ncol = length(columns)))
colnames(acc_df) = columns
acc_df$pred = c(pred)
acc_df$actual = c(df_test_y)
#acc_df
acc = nrow(acc_df[acc_df$pred == acc_df$actual,])
acc/nrow(df_test)
```

```
## [1] 0.3096017
```

```
table(pred, df_test$class)
```

```
##
```

```
## pred      >28days 0mins 10days 10mins 14days 1day  1hr 28days 2days 30mins
##   >28days     164    44     14      7     23   42   12     61    51     15
##   0mins       111  4894     63     33     88  171   59    107   160     74
##   10days        0     1      2      0      0    0    0      0     0      1
##   10mins        0     0      0      0      0    0    0      0     0      0
##   14days        0     3      0      0      8    3    0      2     3      0
##   1day        629   458    463    130    710 1368  276    807  1041    317
##   1hr           0     0      0      0      0    0    1      0     0      0
##   28days       11     5      8      0     11   14    6     50     9      6
##   2days        19    44     31      5     47   69   12     35   113     11
##   30mins      188    21    107    244    151  259  181    179   225    368
##   3days         1     1      0      0      0    0    1      1     2      2
##   3hr           0     1      1      0      1    0    0      0     1      0
##   4days         1     2      0      0      1    1    1      1     0      0
##   5days         0     0      2      0      0    1    0      0     1      0
##   5mins        37    49      7     46     31   41   36     40    48     72
##   6days         0     0      0      0      0    0    1      1     1      0
##   6hr           0     0      0      0      0    1    0      0     0      0
##   7days         0     0      1      0      0    0    0      0     1      0
##   8days         0     0      0      0      0    1    0      1     1      0
##
## pred      3days  3hr 4days 5days 5mins 6days  6hr 7days 8days
##   >28days    23   21    11    18    36    17   16    20    12
##   0mins      88  109    74    61   216    53   79    59    30
##   10days      0    0     1     1     0     0    0     1     0
##   10mins      0    0     0     0     0     0    0     0     0
##   14days      5    2     3     0     0     1    1     0     4
##   1day      699  621   546   481   508   492  505   501   431
##   1hr         0    0     0     0     0     0    0     0     0
##   28days      8   11     9     6     4     7    9     8     2
##   2days      44   48    37    35    26    36   29    33    21
##   30mins    153  202   106    93   275    86  118    79    74
##   3days       4    2     1     0     0     0    1     0     2
##   3hr         0    1     0     0     1     1    1     1     0
##   4days       1    0     5     2     2     1    2     1     1
##   5days       0    1     1     4     2     1    0     0     3
##   5mins      23   44    21    20   107    14   22    15    14
##   6days       0    0     0     0     0     0    1     0     0
##   6hr         0    1     0     0     1     0    1     0     0
##   7days       0    0     0     0     0     0    0     5     0
##   8days       0    0     0     0     0     1    0     0     2
```

```
summary(rf_mod)
```

```
##                 Length  Class  Mode
## call                 4 -none- call
## type                 1 -none- character
## predicted        91737 factor numeric
## err.rate         20000 -none- numeric
## confusion          380 -none- numeric
## votes           1743003 matrix numeric
## oob.times        91737 -none- numeric
## classes             19 -none- character
## importance          11 -none- numeric
## importanceSD         0 -none- NULL
```

```
## localImportance       0 -none- NULL
## proximity             0 -none- NULL
## ntree                 1 -none- numeric
## mtry                  1 -none- numeric
## forest               14 -none- list
## y                 91737 factor numeric
## test                  0 -none- NULL
## inbag                 0 -none- NULL
## terms                 3 terms  call
```