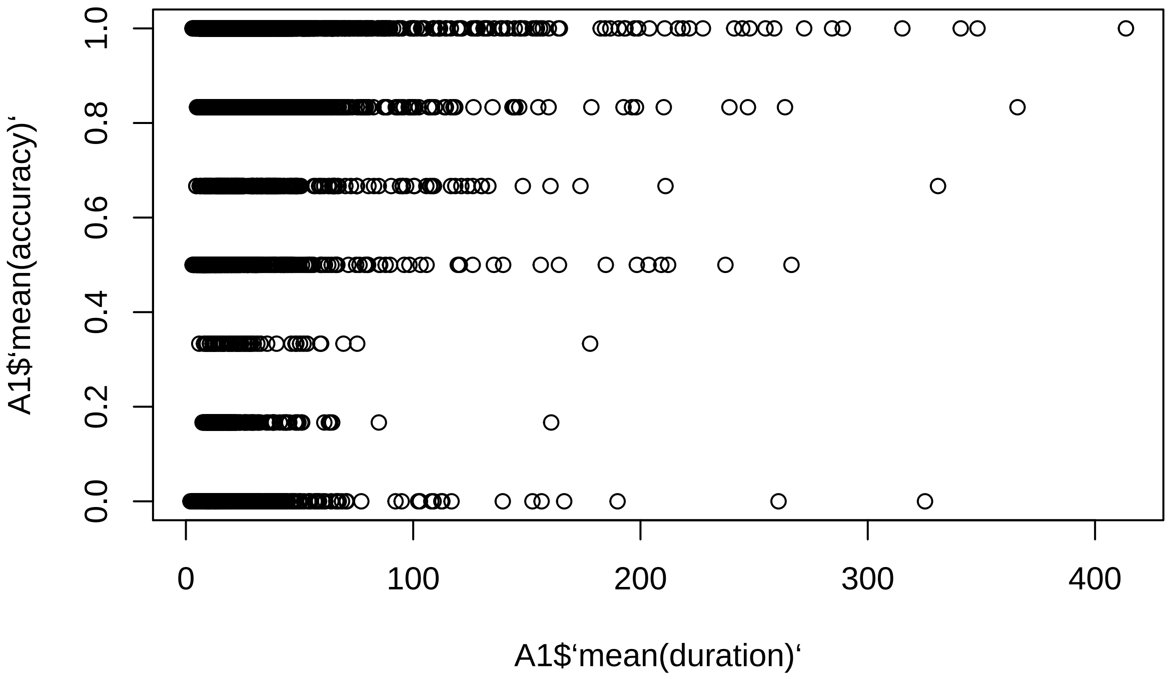
Random Forest of Binary Classification

```
library(data.table)
library(ggplot2)
library(dplyr)
library(knitr)
require(bit64)
library(randomForest)
```

```
Accuracy = fread('https://raw.githubusercontent.com/cszys888/BEGGER-DATA---Team-1/master/CloudFactory_Da
colnames(Accuracy)[4] = "keytype"
colnames(Accuracy)[5] = "mousemove"
colnames(Accuracy)[6] = "mouseclick"
colnames(Accuracy)[7] = "duration"
```

```
dt1 = Accuracy
A1 = dt1 %>%
  group_by(task_id)%>%
  summarise(mean(accuracy), mean(duration))
plot(A1$`mean(duration)`,A1$`mean(accuracy)`)
```



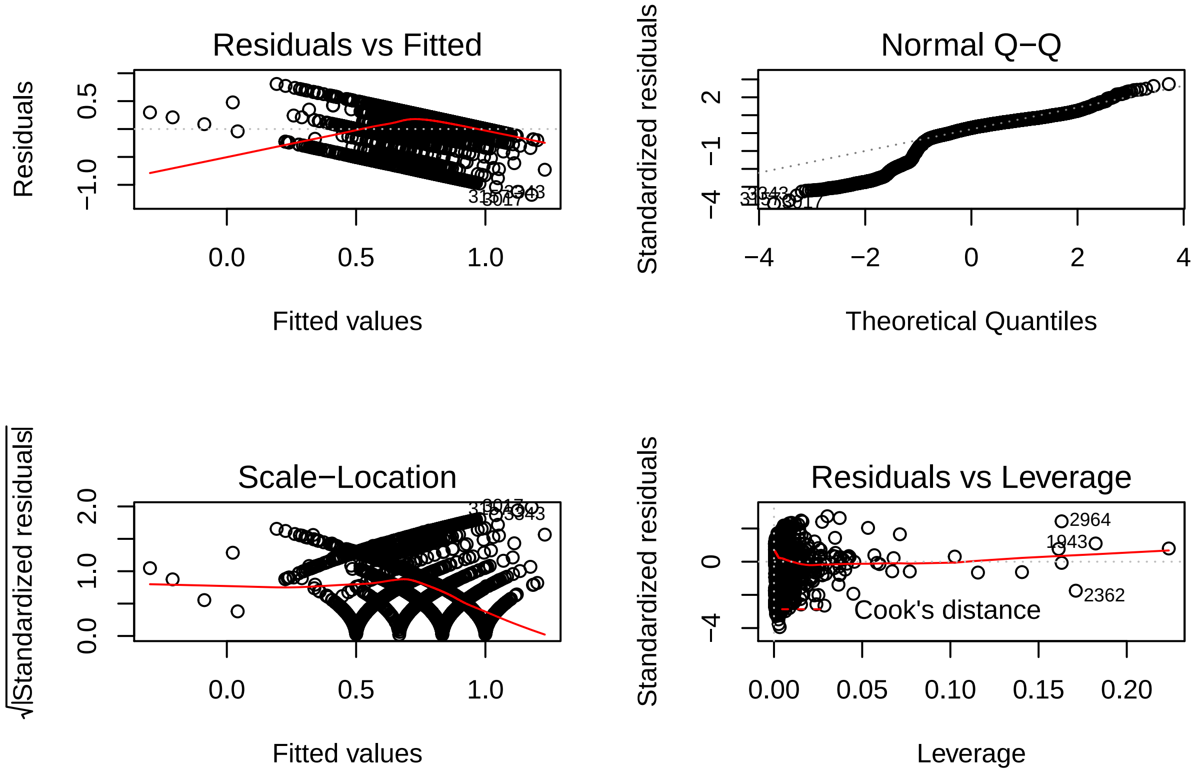
```
dt2 = Accuracy
dt2$keytype[is.na(dt2$keytype)] = 0
dt2$mousemove[!is.na(dt2$mousemove)] = "Yes"
dt2$mousemove[is.na(dt2$mousemove)] = "No"
dt2$mouseclick[!is.na(dt2$mouseclick)] = "Yes"
dt2$mouseclick[is.na(dt2$mouseclick)] = "No"

dt2_trans = dt2 %>%
  group_by(task_id) %>%
  summarise(duration = duration[1],
            count_mousemove = sum(mousemove == "Yes")/n(),
            count_mouseclick = sum(mouseclick == "Yes")/n(),
            key1 = sum(keytype == 1)/n(),
            key2 = sum(keytype == 2)/n(),
            key3 = sum(keytype == 3)/n(),
            key4 = sum(keytype == 4)/n(),
            key5 = sum(keytype == 5)/n(),
            key6 = sum(keytype == 6)/n(),
            key7 = sum(keytype == 7)/n(),
            key8 = sum(keytype == 8)/n(),
            key9 = sum(keytype == 9)/n(),
            key10 = sum(keytype == 10)/n(),
            key11 = sum(keytype == 11)/n(),
            key12 = sum(keytype == 12)/n(),
            accuracy = accuracy[1],
            worker_id = worker_id[1])

#linear regression
lm_dt2 = lm(data = dt2_trans, accuracy~(.-accuracy - task_id - worker_id))
summary(lm_dt2)
```

```
## 
## Call:
## lm(formula = accuracy ~ (. - accuracy - task_id - worker_id), 
##     data = dt2_trans)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.17893 -0.05819  0.08730  0.18697  0.80736 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       0.8840383  0.8252281   1.071 0.284102    
## duration          0.0009351  0.0001473   6.351 2.34e-10 ***
## count_mousemove  -0.2075572  0.8182629  -0.254 0.799772    
## count_mouseclick -0.5773009  0.8239354  -0.701 0.483546    
## key1             -0.1671828  0.8262144  -0.202 0.839653    
## key2             -6.7831613  2.1475152  -3.159 0.001595 ** 
## key3             -0.3075204  0.8316144  -0.370 0.711557    
## key4              3.0234253  2.8583493   1.058 0.290220    
## key5             -3.4271253  0.9671257  -3.544 0.000398 ***
## key6             -0.3372645  0.8280136  -0.407 0.683792    
## key7             -0.5672852  0.8270160  -0.686 0.492781    
## key8              0.5745628  0.8264554   0.695 0.486954    
## key9             -0.8436830  0.8472879  -0.996 0.319422    
## key10             0.8369146  0.8265302   1.013 0.311318    
## key11            -0.8443699  0.8584546  -0.984 0.325363    
## key12            -0.1460888  0.8268063  -0.177 0.859759    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2995 on 4984 degrees of freedom
## Multiple R-squared:  0.1888, Adjusted R-squared:  0.1864 
## F-statistic: 77.35 on 15 and 4984 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(lm_dt2)
```



```r
#logistic regression
glm_dt2 = glm(data = dt2_trans, accuracy~(.-accuracy - task_id - worker_id), family = binomial(link = "
```

```
## Warning: non-integer #successes in a binomial glm!
```

```r
summary(glm_dt2)
```

```
##
## Call:
## glm(formula = accuracy ~ (. - accuracy - task_id - worker_id),
##     family = binomial(link = "logit"), data = dt2_trans)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8596  -0.1339   0.3684   0.6045   2.2017
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.924591   6.475898   0.143  0.88647
## duration             0.006322   0.001529   4.135 3.54e-05 ***
## count_mousemove     -0.503182   6.429765  -0.078  0.93762
## count_mouseclick    -2.258609   6.460108  -0.350  0.72662
## key1                 0.094647   6.483291   0.015  0.98835
## key2               -33.877352  15.925825  -2.127  0.03340 *
## key3                -1.212950   6.533234  -0.186  0.85271
## key4                25.229006  30.317202   0.832  0.40531
## key5               -22.227907   8.448975  -2.631  0.00852 **
## key6              -0.991356   6.496494  -0.153  0.87871
## key7              -1.903432   6.489760  -0.293  0.76929
## key8               4.005656   6.486761   0.618  0.53690
## key9              -2.182509   6.631441  -0.329  0.74207
## key10              6.600020   6.491571   1.017  0.30929
## key11             -4.678210   6.733323  -0.695  0.48719
## key12             -0.004596   6.496866  -0.001  0.99944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3371.8  on 4999  degrees of freedom
## Residual deviance: 2784.6  on 4984  degrees of freedom
## AIC: 4415.4
##
## Number of Fisher Scoring iterations: 5
```

```
#randomforest 7 classification
dt7_trans = dt2_trans
dt7_trans_ = dt7_trans %>%
  select(-task_id, -worker_id)
dt7_trans_$accuracy = as.factor(dt7_trans_$accuracy)

n = names(dt7_trans_)
f = as.formula(paste("accuracy~", paste(n[!n %in% "accuracy"], collapse = "+")))
dt7_rf = randomForest(data = dt7_trans_,
                      f, importance = TRUE)
dt7_rf_predict = predict(dt7_rf)
table7 = table(dt7_trans_$accuracy, dt7_rf_predict)
kable(table7)
```

	0	0.166666666666667	0.333333333333333	0.5	0.666666666666667	0.833333333333333	
0	100	2	0	93	0	7	27
0.166666666666667	1	50	0	2	1	22	14
0.333333333333333	1	2	2	4	3	24	1
0.5	18	1	3	358	5	20	7
0.666666666666667	4	3	0	4	8	53	15
0.833333333333333	3	7	1	8	7	207	69
1	57	22	0	36	4	99	240

```
accurate7 = sum(diag(table7))/5000
paste0("The accuracy of prediction of 7-type classification is ",accurate7)
```

```
## [1] "The accuracy of prediction of 7-type classification is 0.6266"
```

```
#randomforest binary classification
dt2_trans_ = dt2_trans %>%
  select(-task_id, -worker_id) %>%
  mutate(accuracy = (accuracy == 1))
dt2_trans_$accuracy = as.factor(dt2_trans_$accuracy)

n = names(dt2_trans_)
f = as.formula(paste("accuracy~", paste(n[!n %in% "accuracy"], collapse = "+")))
```

```
dt2_rf = randomForest(data = dt2_trans_,
                      f, importance = TRUE)
dt2_rf_predict = predict(dt2_rf)
table2 = table(dt2_trans_$accuracy, dt2_rf_predict)
kable(table2)
```

	FALSE	TRUE
FALSE	1502	872
TRUE	582	2044

```
accurate2 = sum(diag(table2))/5000
paste0("The accuracy of prediction of binary classification is ",accurate2)
```

```
## [1] "The accuracy of prediction of binary classification is 0.7092"
```

```
paste0("The percent information gain (PIG) of this model is 12.52%")
```

```
## [1] "The percent information gain (PIG) of this model is 12.52%"
```