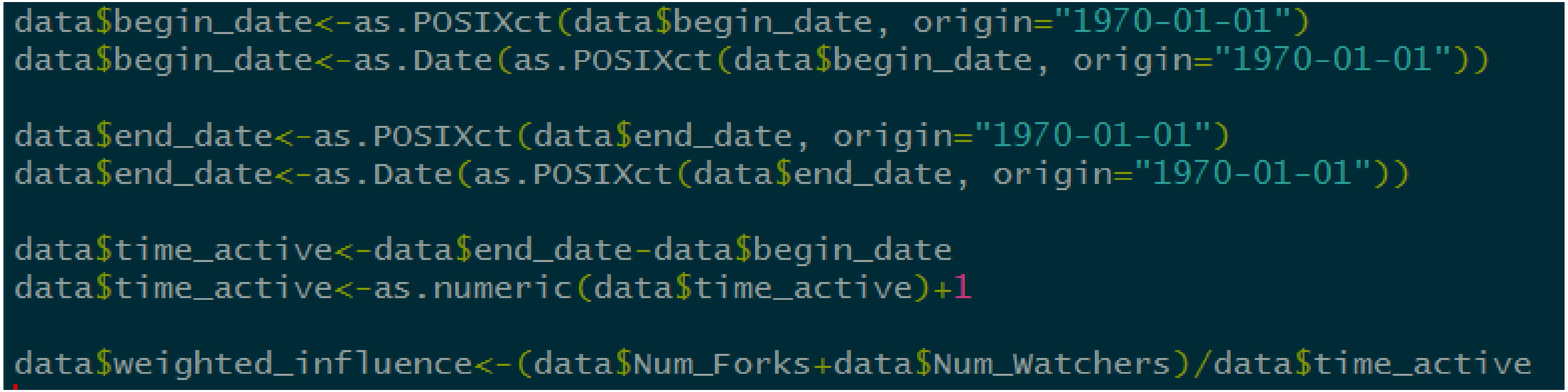
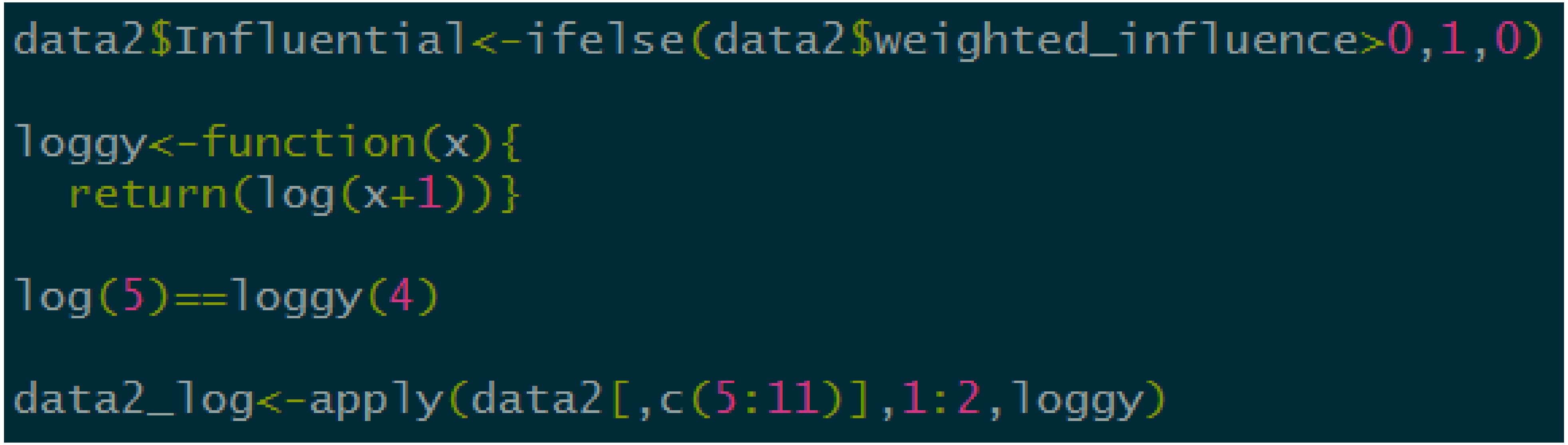
# Project 2c

# New Variable



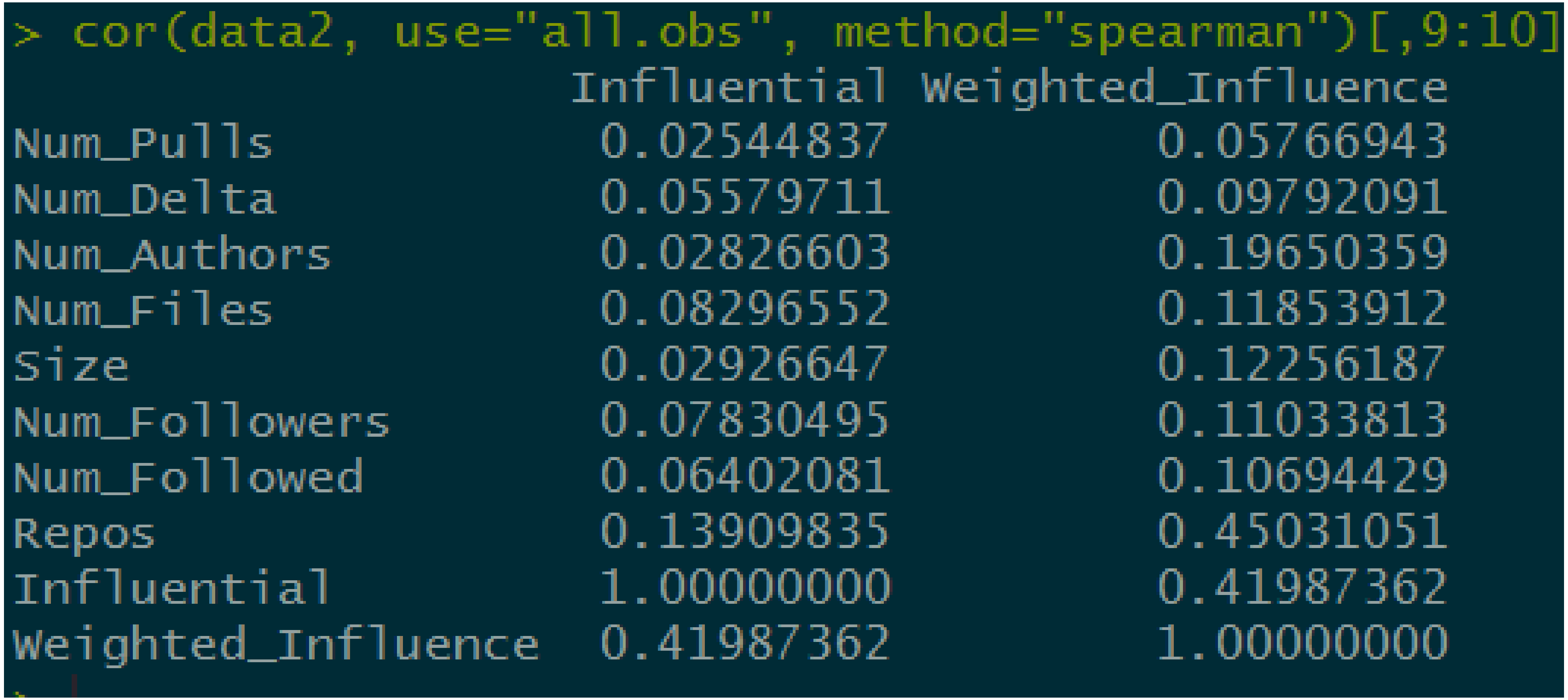
```
data$begin_date<-as.POSIXct(data$begin_date, origin="1970-01-01")
data$begin_date<-as.Date(as.POSIXct(data$begin_date, origin="1970-01-01"))

data$end_date<-as.POSIXct(data$end_date, origin="1970-01-01")
data$end_date<-as.Date(as.POSIXct(data$end_date, origin="1970-01-01"))

data$time_active<-data$end_date-data$begin_date
data$time_active<-as.numeric(data$time_active)+1

data$weighted_influence<-(data$Num_Forks+data$Num_Watchers)/data$time_active
```

# Log *most of* the Data



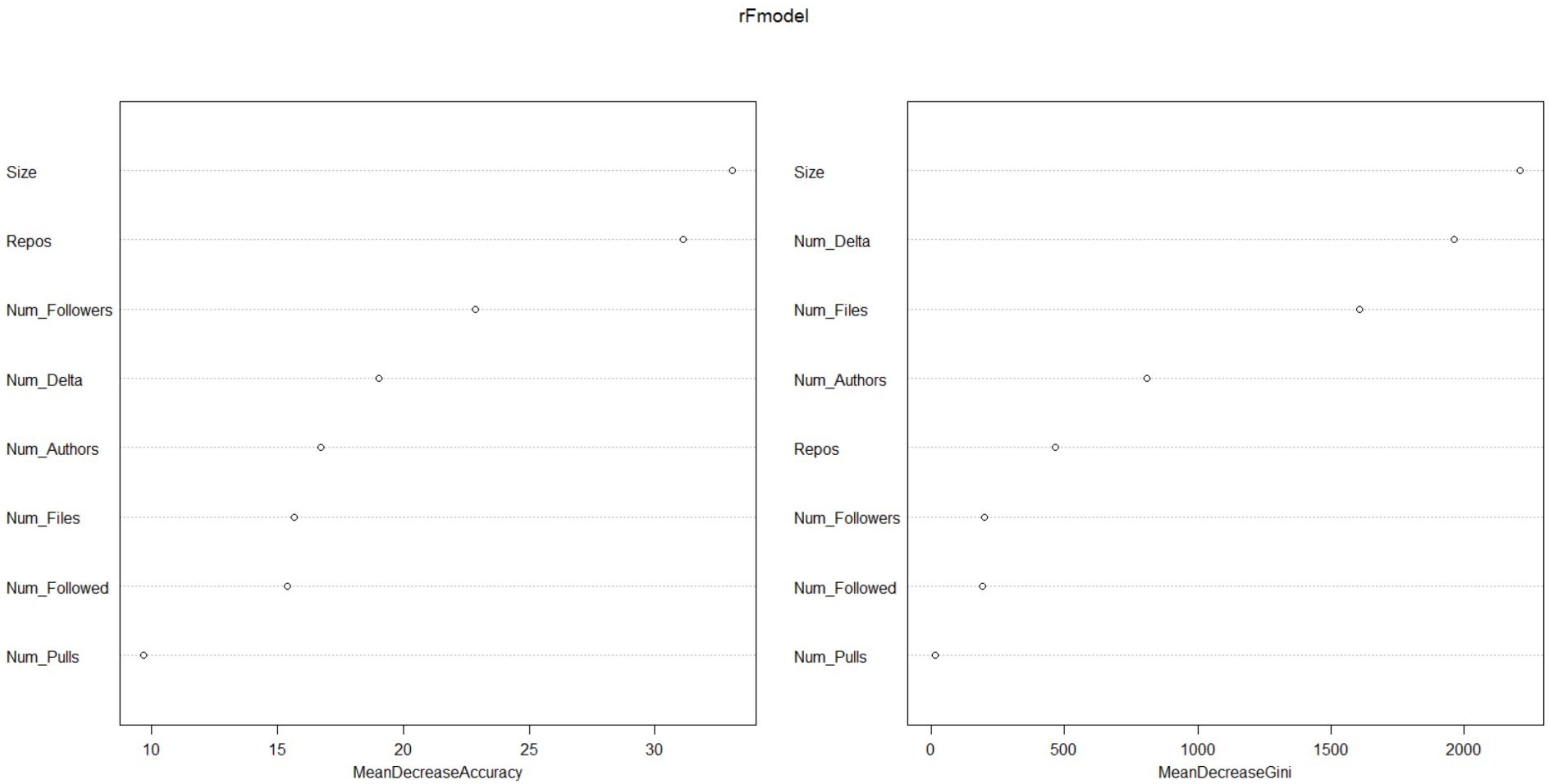
```
data2$Influential<-ifelse(data2$weighted_influence>0,1,0)

loggy<-function(x){
  return(log(x+1))}

log(5)==loggy(4)

data2_log<-apply(data2[,c(5:11)],1:2,loggy)
```

# All Data -> User Level

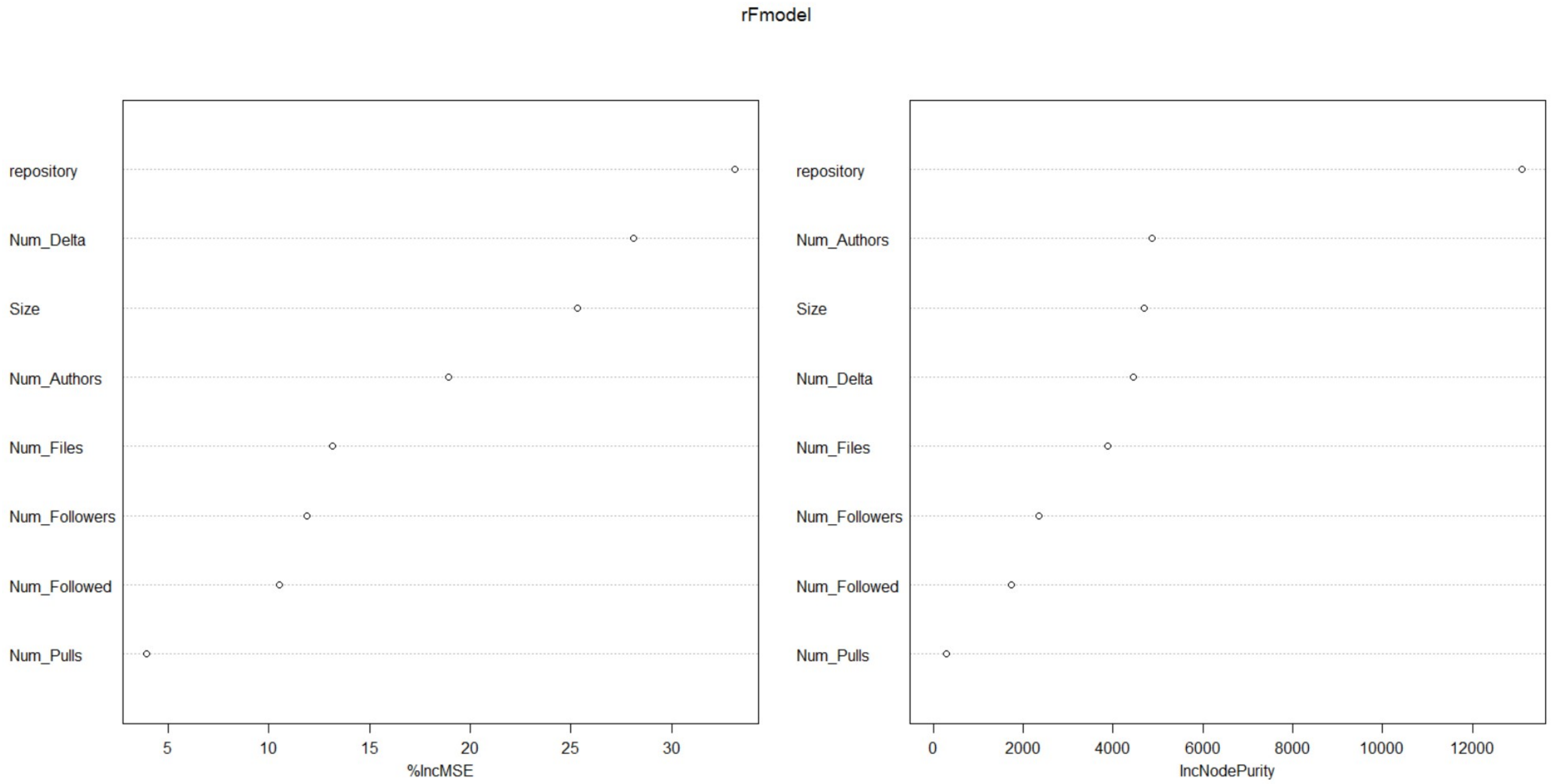


```
> cor(data2, use="all.obs", method="spearman")[,9:10]
                   Influential Weighted_Influence
Num_Pulls           0.02544837         0.05766943
Num_Delta           0.05579711         0.09792091
Num_Authors         0.02826603         0.19650359
Num_Files           0.08296552         0.11853912
Size                0.02926647         0.12256187
Num_Followers       0.07830495         0.11033813
Num_Followed        0.06402081         0.10694429
Repos               0.13909835         0.45031051
Influential         1.00000000         0.41987362
Weighted_Influence  0.41987362         1.00000000
>
```

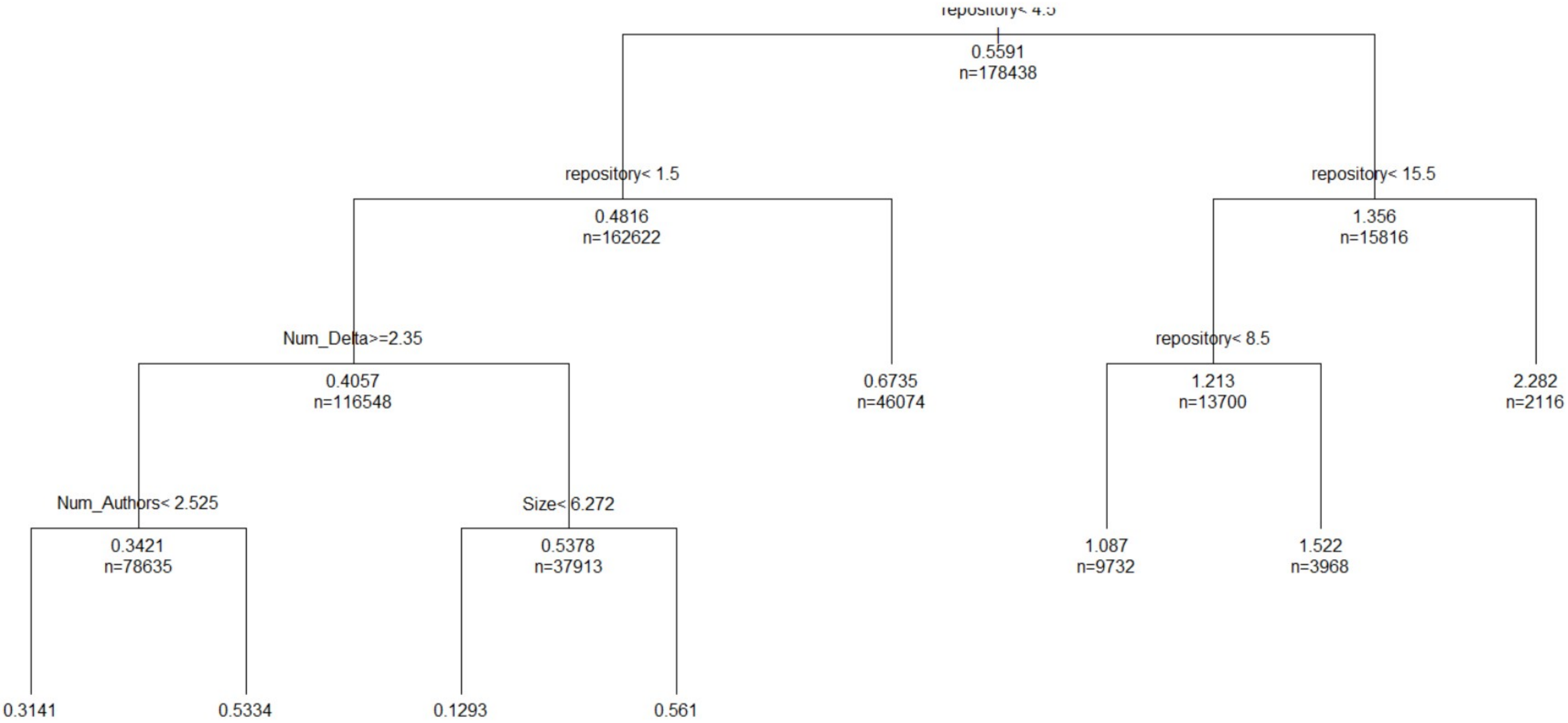
# Random Forest



# Random Forest



## Classification Tree for Influence


repository< 4.5
0.5591
n=178438
repository< 1.5
0.4816
n=162622
repository< 15.5
1.356
n=15816
Num_Delta>=2.35
0.4057
n=116548
0.6735
n=46074
repository< 8.5
1.213
n=13700
2.282
n=2116
Num_Authors< 2.525
0.3421
n=78635
Size< 6.272
0.5378
n=37913
1.087
n=9732
1.522
n=3968
0.3141
0.5334
0.1293
0.561