Model Selection

Yuqing Yang

11/12/2023

Please refer to facebook-metrics.Rmd in Ryan branch for Matrix of correlation and EDA

This Rmd is used for model selection

```r
train_df <- model_df %>% filter(obs_type == 'Training') %>% select(c(Lifetime.Post.Consumers, Category,
# transform on numerical variables only
transform <- train_df %>% select(c(Lifetime.Post.Consumers, Page.total.likes))

boxcox_result <- preProcess(transform, method = "BoxCox")
boxcox_result
```

```
## Created from 342 samples and 2 variables
##
## Pre-processing:
##   - Box-Cox transformation (2)
##   - ignored (0)
##
## Lambda estimates for Box-Cox transformation:
## 0.1, 2
```

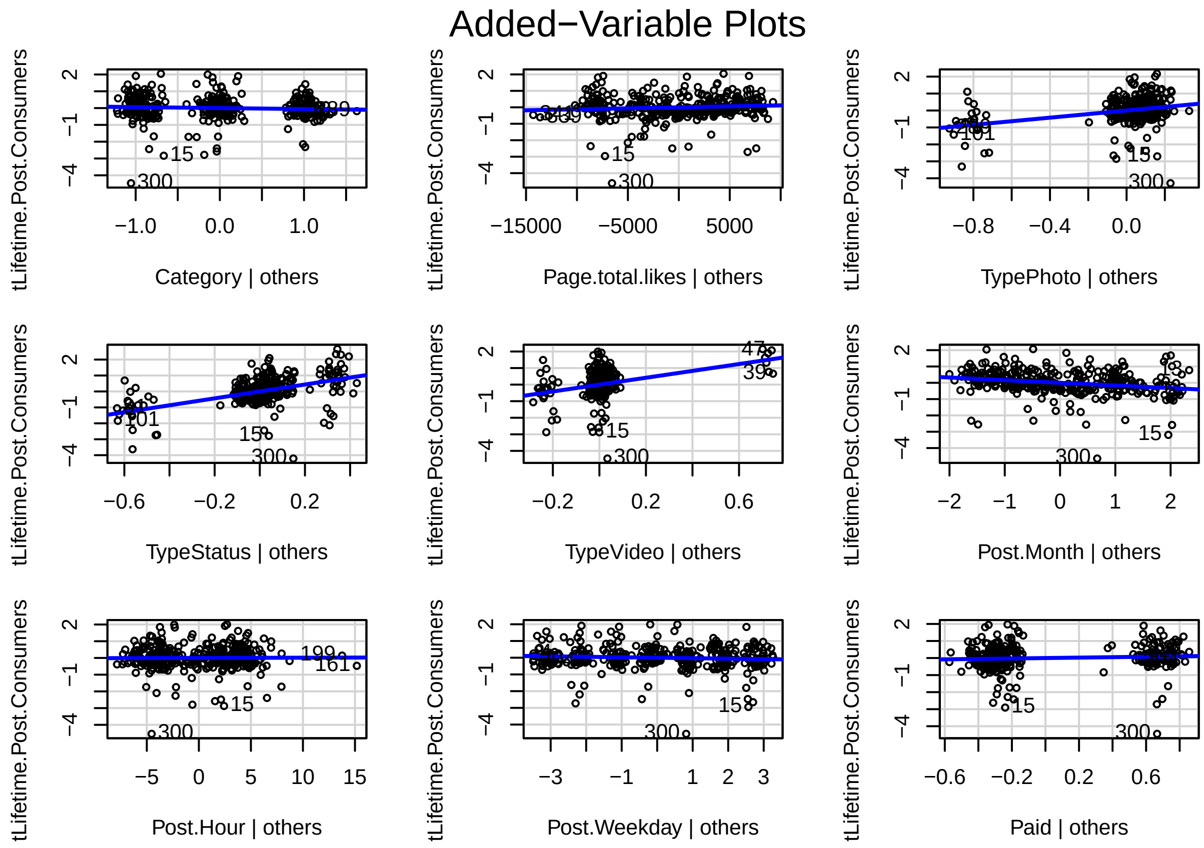
```r
# only log transform on `Lifetime.Post.Consumers`
# `Page.total.likes**2` does not normalize distribution according to histogram...
# ... Keep Page.total.likes as original form currently
t_train_df<-train_df %>% mutate(tLifetime.Post.Consumers=log(Lifetime.Post.Consumers))
```

```r
#fit the full model with all predictors
full_model<-lm(
    tLifetime.Post.Consumers ~ Category + Page.total.likes + Type +
      Post.Month + Post.Hour + Post.Weekday + Paid, data = t_train_df)
#examine full model regression
summary(full_model)
```

```
##
## Call:
## lm(formula = tLifetime.Post.Consumers ~ Category + Page.total.likes +
##     Type + Post.Month + Post.Hour + Post.Weekday + Paid, data = t_train_df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5329 -0.3245 -0.0187  0.3660  2.0007
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.041e+00  6.683e-01   7.543 4.48e-13 ***
## Category         -5.883e-02  4.979e-02  -1.182    0.238
## Page.total.likes  1.207e-05  7.436e-06   1.623    0.106
## TypePhoto         1.056e+00  1.736e-01   6.080 3.30e-09 ***
## TypeStatus        2.172e+00  2.100e-01  10.341  < 2e-16 ***
## TypeVideo         2.068e+00  3.270e-01   6.325 8.18e-10 ***
## Post.Month       -1.620e-01  3.675e-02  -4.407 1.42e-05 ***
## Post.Hour         1.511e-03  9.717e-03   0.155    0.877
## Post.Weekday     -2.799e-02  2.000e-02  -1.400    0.163
## Paid              1.272e-01  8.739e-02   1.455    0.147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7363 on 331 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3655, Adjusted R-squared:  0.3483
## F-statistic: 21.19 on 9 and 331 DF,  p-value: < 2.2e-16
```

```
# added variable plots
avPlots(full_model)
```



Added−Variable Plots

```r
# variance inflation factors
vif(full_model)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## Category         1.116446  1        1.056620
## Page.total.likes 9.241635  1        3.040006
## Type             1.235242  3        1.035838
## Post.Month       9.262833  1        3.043490
## Post.Hour        1.093990  1        1.045940
## Post.Weekday     1.011339  1        1.005653
## Paid             1.034226  1        1.016969
```

```r
stepwise_model <- step(full_model, direction = "both")
```

```
## Start:  AIC=-198.89
## tLifetime.Post.Consumers ~ Category + Page.total.likes + Type +
##     Post.Month + Post.Hour + Post.Weekday + Paid
##
##                    Df Sum of Sq    RSS      AIC
## - Post.Hour         1     0.013 179.48 -200.860
## - Category          1     0.757 180.22 -199.450
## <none>                          179.47 -198.885
## - Post.Weekday      1     1.062 180.53 -198.873
## - Paid              1     1.148 180.62 -198.710
## - Page.total.likes  1     1.428 180.90 -198.182
## - Post.Month        1    10.531 190.00 -181.441
## - Type              3    69.904 249.37  -92.713
##
## Step:  AIC=-200.86
## tLifetime.Post.Consumers ~ Category + Page.total.likes + Type +
##     Post.Month + Post.Weekday + Paid
##
##                    Df Sum of Sq    RSS      AIC
## - Category          1     0.786 180.27 -201.370
## - Post.Weekday      1     1.051 180.53 -200.869
## <none>                          179.48 -200.860
## - Paid              1     1.136 180.62 -200.709
## - Page.total.likes  1     1.448 180.93 -200.120
## + Post.Hour         1     0.013 179.47 -198.885
## - Post.Month        1    10.778 190.26 -182.974
## - Type              3    70.378 249.86  -94.047
##
## Step:  AIC=-201.37
## tLifetime.Post.Consumers ~ Page.total.likes + Type + Post.Month +
##     Post.Weekday + Paid
##
##                    Df Sum of Sq    RSS     AIC
## - Post.Weekday      1     0.963 181.23 -201.55
## <none>                          180.27 -201.37
## - Paid              1     1.135 181.40 -201.23
## + Category          1     0.786 179.48 -200.86
## - Page.total.likes  1     1.339 181.60 -200.85
## + Post.Hour         1     0.042 180.22 -199.45
## - Post.Month        1    10.379 190.65 -184.28
## - Type              3    69.612 249.88  -96.02
##
## Step:  AIC=-201.55
## tLifetime.Post.Consumers ~ Page.total.likes + Type + Post.Month +
##     Paid
##
##                    Df Sum of Sq    RSS      AIC
## <none>                          181.23 -201.553
## - Paid              1     1.090 182.32 -201.509
## + Post.Weekday      1     0.963 180.27 -201.370
## - Page.total.likes  1     1.386 182.62 -200.955
## + Category          1     0.698 180.53 -200.869
## + Post.Hour         1     0.018 181.21 -199.586
## - Post.Month        1    10.555 191.78 -184.249
## - Type              3    69.630 250.86  -96.683
```

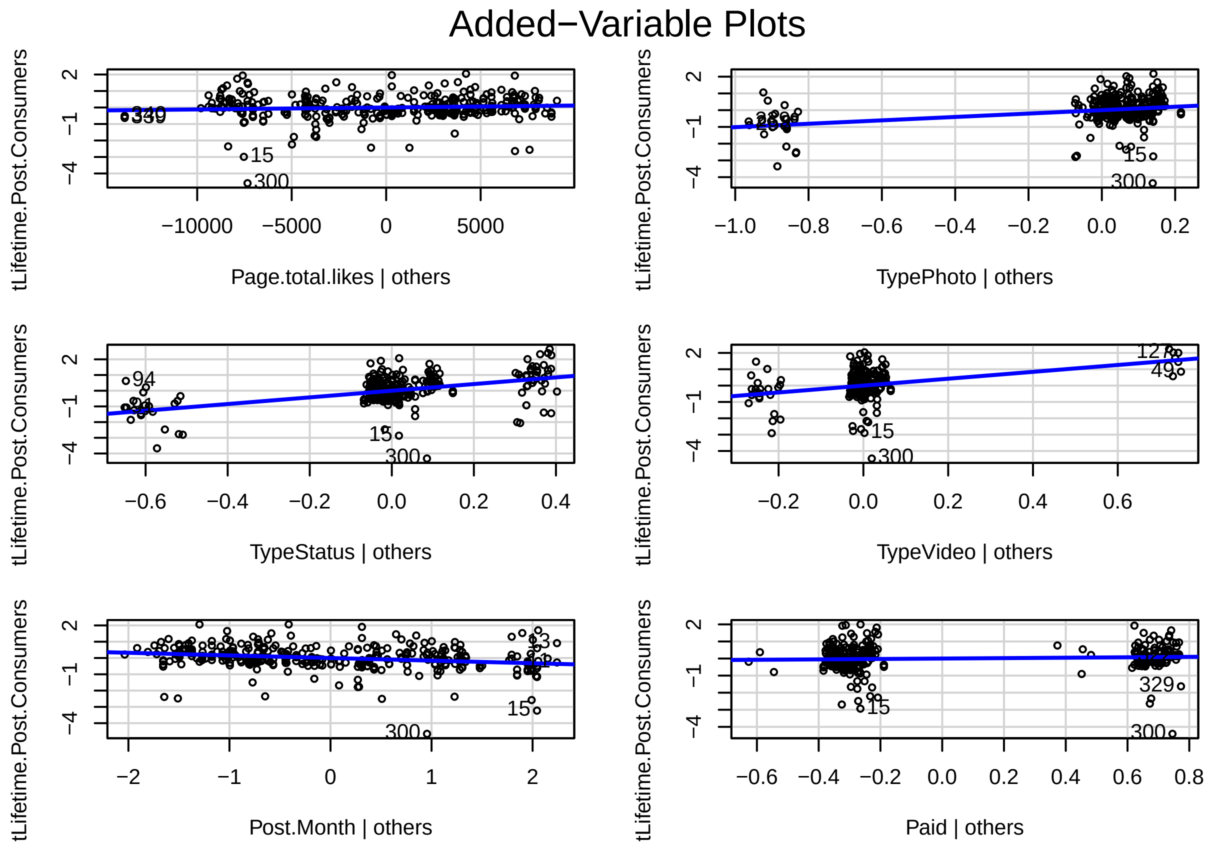
```
#result shows that the lowest AIC model is...
#...tLifetime.Post.Consumers ~ Page.total.likes + Type + Post.Month + ...
#...Paid
```

```
reduced_model<-lm(tLifetime.Post.Consumers~Page.total.likes + Type + Post.Month+Paid, data=t_train_df)
summary(reduced_model)
```

```
##
```

```
## Call:
## lm(formula = tLifetime.Post.Consumers ~ Page.total.likes + Type +
##     Post.Month + Paid, data = t_train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5027 -0.3347 -0.0351  0.3617  2.0301
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.879e+00  6.597e-01   7.397 1.14e-12 ***
## Page.total.likes 1.185e-05  7.414e-06   1.598    0.111
## TypePhoto        1.015e+00  1.673e-01   6.068 3.51e-09 ***
## TypeStatus       2.120e+00  2.045e-01  10.368  < 2e-16 ***
## TypeVideo        2.101e+00  3.245e-01   6.475 3.39e-10 ***
## Post.Month      -1.604e-01  3.636e-02  -4.411 1.39e-05 ***
## Paid             1.229e-01  8.670e-02   1.417    0.157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7366 on 334 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3593, Adjusted R-squared:  0.3478
## F-statistic: 31.21 on 6 and 334 DF,  p-value: < 2.2e-16
```

```
# added variable plots
avPlots(reduced_model)
```



```
# variance inflation factors
vif(reduced_model)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## Page.total.likes 9.181151  1        3.030041
## Type             1.110272  3        1.017587
## Post.Month       9.061147  1        3.010174
## Paid             1.017145  1        1.008536
```

```
anova(full_model, reduced_model)
```

```
## Analysis of Variance Table
##
## Model 1: tLifetime.Post.Consumers ~ Category + Page.total.likes + Type +
##     Post.Month + Post.Hour + Post.Weekday + Paid
## Model 2: tLifetime.Post.Consumers ~ Page.total.likes + Type + Post.Month +
##     Paid
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    331 179.47
## 2    334 181.23 -3   -1.7624 1.0835 0.3562
```

```
#Insignificant p-value of F-test indicates excluding these predictors may not affect the model fit...
#... The reduced model may be better fit than full model given lower AIC and insignificant Partial... #
```

```r
#fit the full model with all predictors with transformation on Page.total.likes
full_model_2<-lm(
    tLifetime.Post.Consumers ~ Category +Page.total.likes+ I(Page.total.likes^2) + Type +
      Post.Month + Post.Hour + Post.Weekday + Paid, data = t_train_df)
stepwise_model_2 <- step(full_model_2, direction = "both")
```

```
## Start:  AIC=-197.97
## tLifetime.Post.Consumers ~ Category + Page.total.likes + I(Page.total.likes^2) +
##     Type + Post.Month + Post.Hour + Post.Weekday + Paid
##
##                         Df Sum of Sq    RSS      AIC
## - Post.Hour              1     0.020 178.92 -199.931
## - Page.total.likes       1     0.346 179.24 -199.311
## - I(Page.total.likes^2)  1     0.570 179.47 -198.885
## - Category               1     0.778 179.68 -198.490
## <none>                               178.90 -197.969
## - Post.Weekday           1     1.124 180.02 -197.833
## - Paid                   1     1.173 180.07 -197.741
## - Post.Month             1     7.070 185.97 -186.752
## - Type                   3    68.834 247.73  -92.962
##
## Step:  AIC=-199.93
## tLifetime.Post.Consumers ~ Category + Page.total.likes + I(Page.total.likes^2) +
##     Type + Post.Month + Post.Weekday + Paid
##
##                         Df Sum of Sq    RSS      AIC
## - Page.total.likes       1     0.339 179.26 -201.285
## - I(Page.total.likes^2)  1     0.563 179.48 -200.860
## - Category               1     0.812 179.73 -200.387
## <none>                               178.92 -199.931
## - Post.Weekday           1     1.109 180.03 -199.825
## - Paid                   1     1.153 180.07 -199.741
## + Post.Hour              1     0.020 178.90 -197.969
## - Post.Month             1     7.125 186.04 -188.614
## - Type                   3    69.351 248.27  -94.224
##
## Step:  AIC=-201.29
## tLifetime.Post.Consumers ~ Category + I(Page.total.likes^2) +
##     Type + Post.Month + Post.Weekday + Paid
##
##                         Df Sum of Sq    RSS      AIC
## - Category               1     0.799 180.06 -201.768
## <none>                               179.26 -201.285
## - Post.Weekday           1     1.062 180.32 -201.271
## - Paid                   1     1.125 180.38 -201.153
## - I(Page.total.likes^2)  1     1.672 180.93 -200.120
## + Page.total.likes       1     0.339 178.92 -199.931
## + Post.Hour              1     0.014 179.24 -199.311
## - Post.Month             1     9.688 188.94 -185.336
## - Type                   3    69.844 249.10  -95.083
##
## Step:  AIC=-201.77
## tLifetime.Post.Consumers ~ I(Page.total.likes^2) + Type + Post.Month +
##     Post.Weekday + Paid
##
##                          Df Sum of Sq    RSS      AIC
## - Post.Weekday            1     0.973 181.03 -201.931
## <none>                                180.06 -201.768
## - Paid                    1     1.124 181.18 -201.646
## + Category                1     0.799 179.26 -201.285
## - I(Page.total.likes^2)   1     1.549 181.60 -200.848
## + Page.total.likes        1     0.327 179.73 -200.387
## + Post.Hour               1     0.043 180.01 -199.849
## - Post.Month              1     9.307 189.36 -186.582
## - Type                    3    69.059 249.12  -97.063
##
## Step:  AIC=-201.93
## tLifetime.Post.Consumers ~ I(Page.total.likes^2) + Type + Post.Month +
##     Paid
##
##                          Df Sum of Sq    RSS      AIC
## <none>                                181.03 -201.931
## - Paid                    1     1.080 182.11 -201.903
## + Post.Weekday            1     0.973 180.06 -201.768
## + Category                1     0.710 180.32 -201.271
## - I(Page.total.likes^2)   1     1.587 182.62 -200.955
## + Page.total.likes        1     0.283 180.75 -200.465
## + Post.Hour               1     0.018 181.01 -199.966
## - Post.Month              1     9.439 190.47 -186.600
## - Type                    3    69.088 250.12  -97.695
```

```
reduced_model_2<-lm(tLifetime.Post.Consumers ~ I(Page.total.likes^2) + Type + Post.Month +
    Paid, data=t_train_df)

#Perform 4 steps by removing some variables
#The final model includes "I(Page.total.likes^2)", "Type","Post.Month","Paid"
#The final LOWEST AIC is -201.93, which is smaller than the first stepwide_model
#Therefore, the reduced_model_2 may be a better fit than reduced_model
```

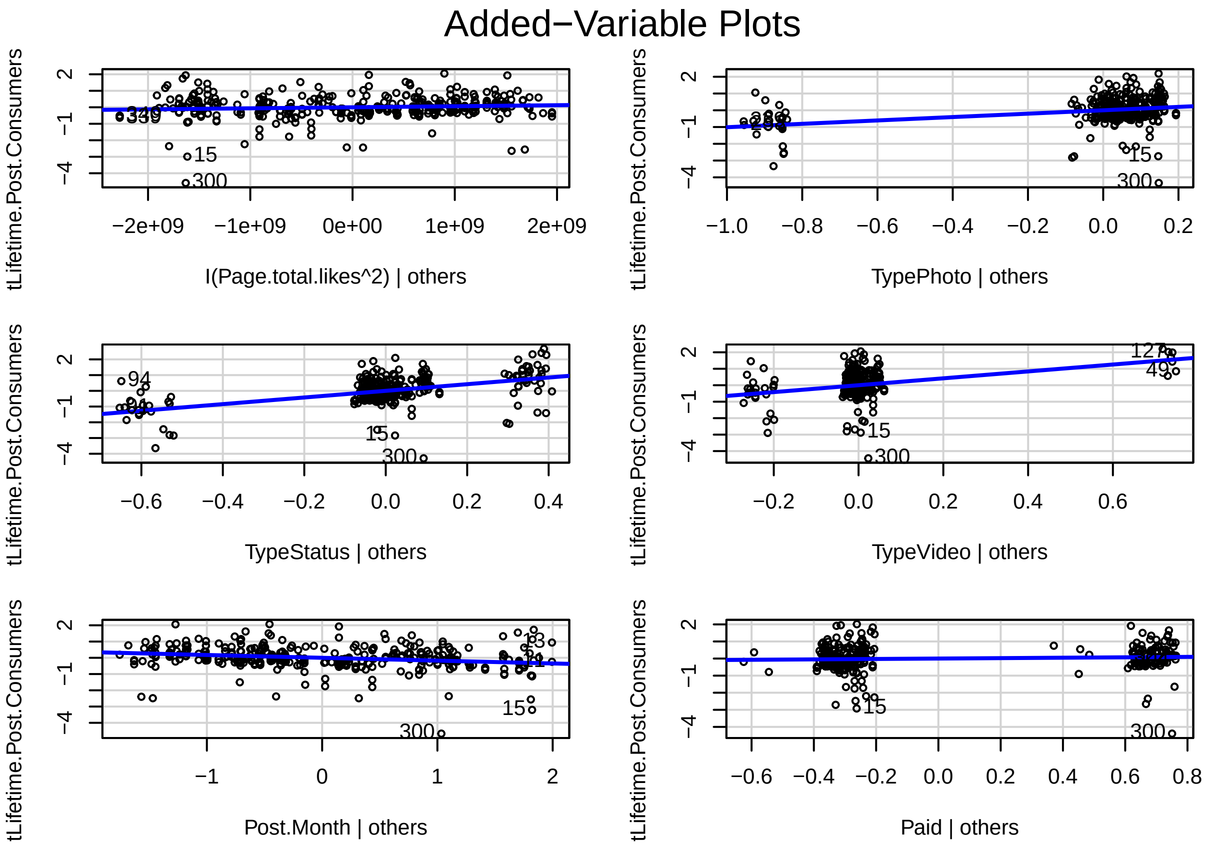
```
#examine reduced_model_2
summary(reduced_model_2)
```

```
##
## Call:
## lm(formula = tLifetime.Post.Consumers ~ I(Page.total.likes^2) +
##     Type + Post.Month + Paid, data = t_train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4868 -0.3297 -0.0396  0.3557  2.0430
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.465e+00  3.075e-01  17.772  < 2e-16 ***
## I(Page.total.likes^2)  6.302e-11  3.683e-11   1.711    0.088 .
## TypePhoto              1.010e+00  1.674e-01   6.033 4.27e-09 ***
## TypeStatus             2.113e+00  2.046e-01  10.327  < 2e-16 ***
## TypeVideo              2.093e+00  3.246e-01   6.449 3.95e-10 ***
## Post.Month            -1.738e-01  4.164e-02  -4.173 3.84e-05 ***
## Paid                   1.223e-01  8.663e-02   1.411    0.159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7362 on 334 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:   0.36,  Adjusted R-squared:  0.3485
## F-statistic: 31.31 on 6 and 334 DF,  p-value: < 2.2e-16
```

```
vif(reduced_model_2)
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## I(Page.total.likes^2) 12.043080  1        3.470314
## Type                   1.113052  3        1.018011
## Post.Month            11.895482  1        3.448983
## Paid                   1.016732  1        1.008331
```

```
avPlots(reduced_model_2)
```



```
anova(full_model, reduced_model_2)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: tLifetime.Post.Consumers ~ Category + Page.total.likes + Type +
##     Post.Month + Post.Hour + Post.Weekday + Paid
## Model 2: tLifetime.Post.Consumers ~ I(Page.total.likes^2) + Type + Post.Month +
##     Paid
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1    331 179.47
## 2    334 181.03 -3   -1.5615 0.96 0.4118
```

```
#Insignificant p-value of F-test indicates excluding these predictors may not affect the model fit...
#... The reduced model 2 may be better fit than full model given lower AIC and insignificant Partial...
```

```
sprintf(paste("BIC of full model", BIC(full_model)))
```

```
## [1] "BIC of full model 812.981653521479"
```

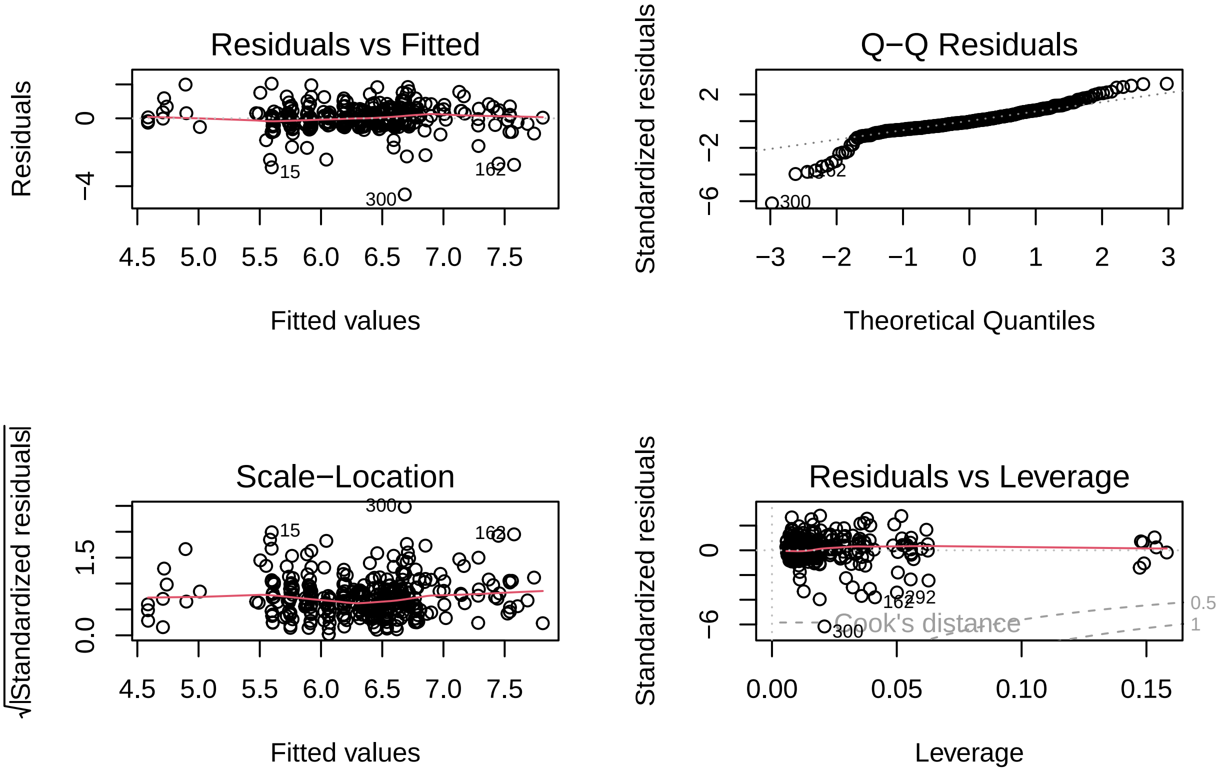
```
sprintf(paste("BIC of 1st reduced model", BIC(reduced_model)))
```

```
## [1] "BIC of 1st reduced model 798.818348084141"
```

```
sprintf(paste("BIC of 2ed reduced model", BIC(reduced_model_2)))
```

```
## [1] "BIC of 2ed reduced model 798.44009791478"
```

```
par(mfrow=c(2,2))
plot(reduced_model_2)
```



Conclusion: ## lm(tLifetime.Post.Consumers ~ I(Page.total.likes^2) +Type + Post.Month + Paid, data = t_train_df)