Google Analytics Customer Revenue Prediction

Xiaofei Wu

I. Intro

1. Project overall

This report analyze a Google Merchandise Store customer dataset to predict revenue per customer through Linear mixed effect model, Linear mixed effect model with log transformation and logistic model. Before all models were build, the exploratory data analysis (EDA) is used to analyzing data sets to summarize their main characteristics, mostly with visual methods. From aspect of Root-Mean-Squared-Error (RMSE), Linear mixed effect model with log transformation leads to the least of 1.934. Then, interpretation and implication are mentioned to show the result of the analysis. In the end, limitation and future discussion are stated to look forward to an improvement in the future for this analysis.

2. Background

The 80/20 rule has proven true for many businesses–only a small percentage of customers produce most of the revenue. As such, marketing teams are challenged to make appropriate investments in promotional strategies.

3. Evaluation

Submissions are scored on the root mean squared error. RMSE is defined as:

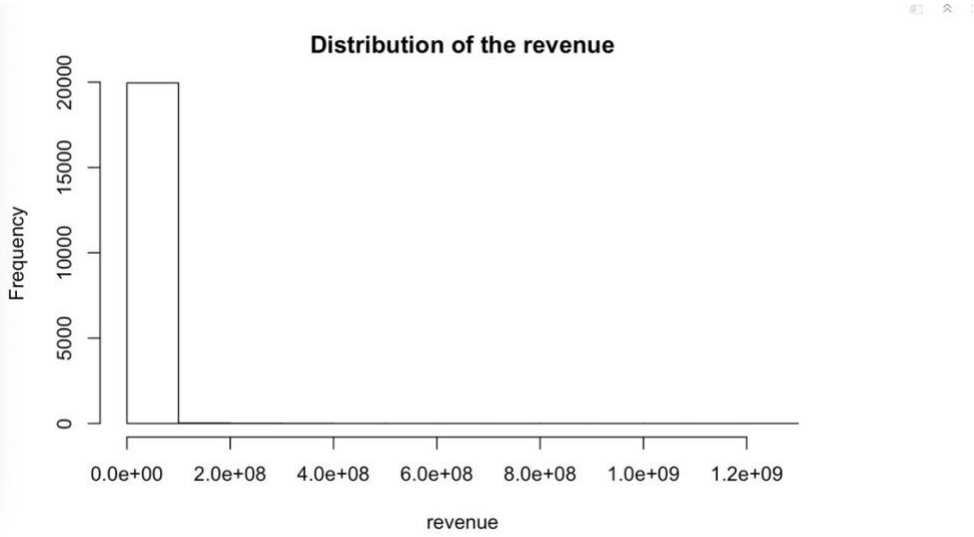
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

4. **Data source**

The data set is all from kaggle.com (https://www.kaggle.com/c/ga-customer-revenue- prediction), where it contains train dataset, test dataset and submission file. The size of all three file is more than 30GB, so I use random sampling and select 30000 rows to deal with the data (20000 rows and training data and 10000 rows as the test data). There is one column “hits” in the data which is pretty large, thus I ignore this column when importing the data. Also, there are several columns with sub-column information in json format, therefore I use “jsonlite” package in r to convert these column into normal columns. Moreover, there are constant columns like “social Engagement Type” and “visits”, and I delete them.

II. Exploratory Data Analysis(EDA)

1. Firstly group the data by outcomes, which is the outcome, and create plots.

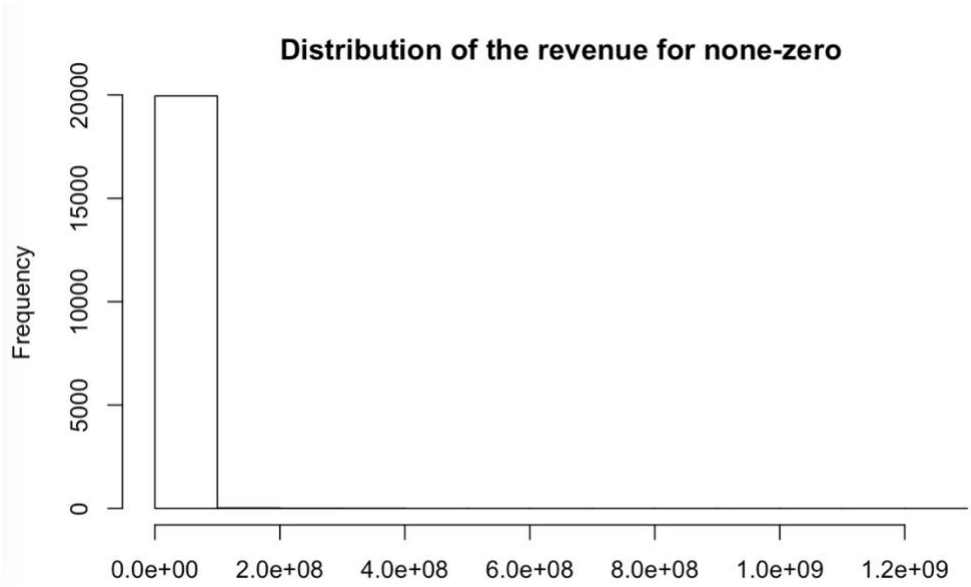


I notive most of the transactionRevenue laid arount zero, so I want to check if they are actually zero.

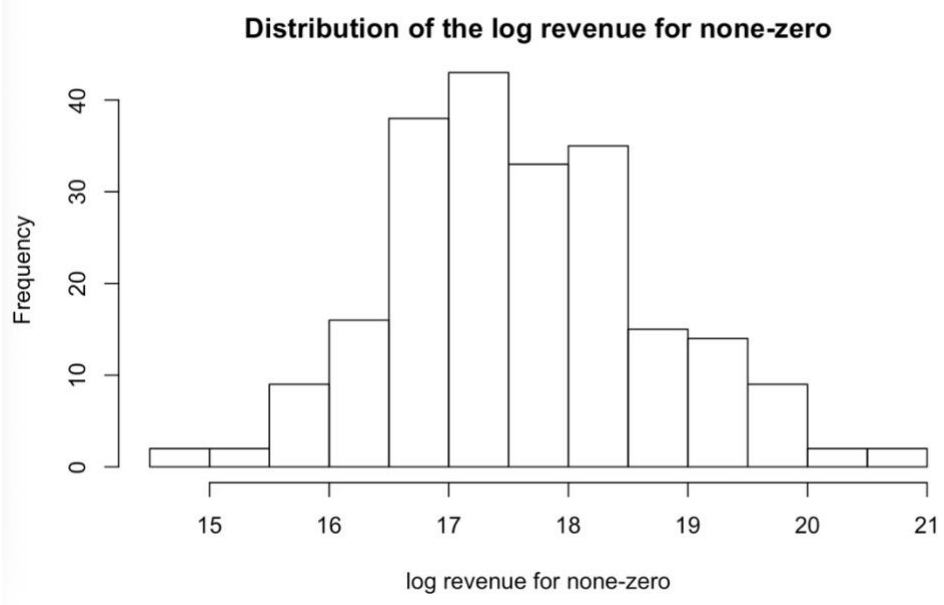
Revenue <dbl>	n <int>	percentage <chr>
0	19780	98.9%
1	220	1.1%

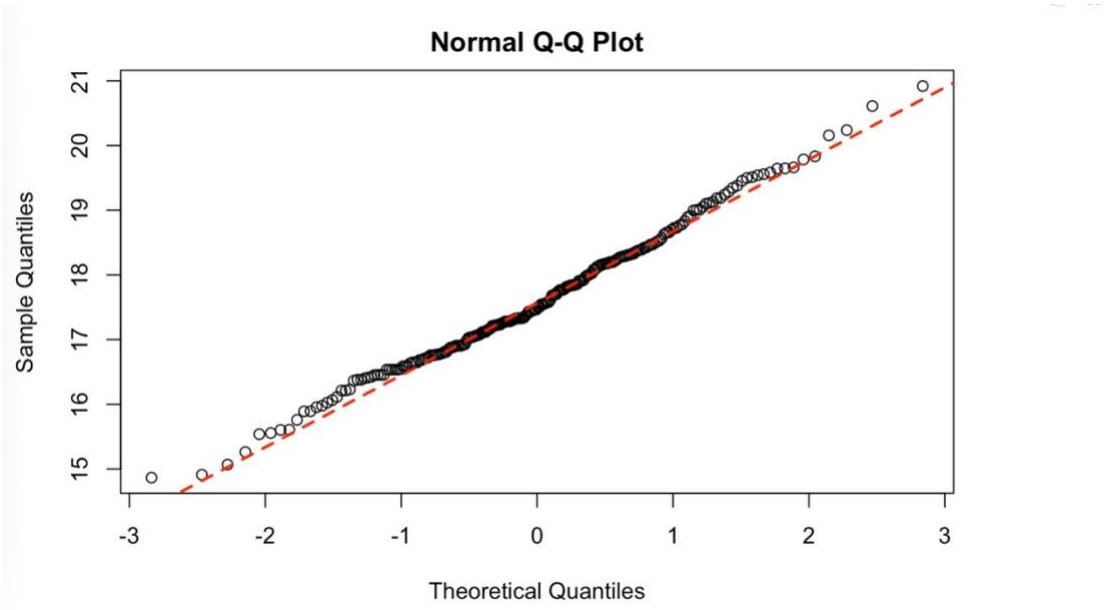
We can see that only 1.1% percentage of the data was none-zero revenue

Also most of the rows are zero revenue, we will focus on the none-zero revenue history to do the predict.



For our none-zero revenue data, the distribution of revenue is right skewed, so we see the log(revenue distribution).





For nonzero targets, they seems follow normal distribution.

Thus I created qqplot, and it indicates that the nonzero target approximately follows normal distribution.

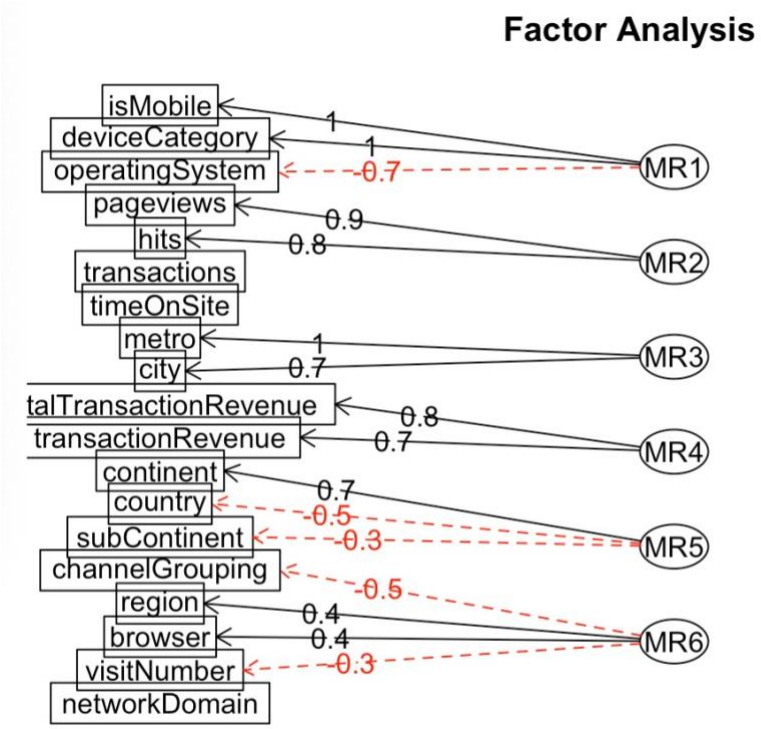
2. Create correlation table for predictors

	channelGrouping	visitNumber	browser	operatingSystem	isMobile	deviceCateg
channelGrouping	1	-0.02445196280849...	0.0225829952532298	0.143779208434578	-0.19157725387942	-0.178234679
operatingSystem	0.14377920843...	-0.04339532291292...	-0.192180391336799	1	-0.754942911342475	-0.695647062
networkDomain	0.08144609639...	-0.131192021152257	0.0614400083712701	0.116969950032838	-0.0181198452433685	-0.019773175
continent	0.02908690055...	-0.117484613249419	0.0261207435643558	0.150977839547154	-0.0537539763975498	-0.041535802
subContinent	0.02633038297...	-0.04245907600048...	0.0207415125237666	-0.00812106158189631	0.0194996397726543	0.014285914
browser	0.02258299525...	-0.09750560084831...	1	-0.192180391336799	0.322857457039855	0.329815727
transactionRevenue	0.02202047448...	0.098835344568028	-0.03123719557796...	-0.0096114352609611	-0.0419574776204648	-0.039769803
metro	0.00783789283...	0.0814859665270614	-0.015907022601383	-0.034713984378315	-0.013492858519025	0.003760565
city	0.00053338396...	0.0325281083869458	0.003209197243600...	0.0202123269840619	-0.0171070997109015	-0.001001187
isMobile	-0.1915772538...	-0.05880416665733...	0.322857457039855	-0.754942911342475	1	0.943786229
deviceCategory	-0.1782346797...	-0.05523223351005...	0.329815727490246	-0.695647062283076	0.943786229627513	1
country	-0.0402846386...	0.1226429414011	0.008419956421300...	-0.0815199118842043	-0.0147780864241942	-0.012034896
pageviews	-0.0400269511...	0.100701503008088	-0.07257460819726	-0.0152721111169108	-0.0529992709860497	-0.047807349
visitNumber	-0.0244519628...	1	-0.09750560084831...	-0.0433953229129212	-0.0588041666573323	-0.055232233
hits	-0.0222633843...	0.0735666885902391	-0.08355690181937...	-0.00774062837672993	-0.0523355729515046	-0.048525701
region	-0.0014793768...	-0.093609755463068	0.0552437619023284	0.0610866548685522	0.0498544043831791	0.060302458

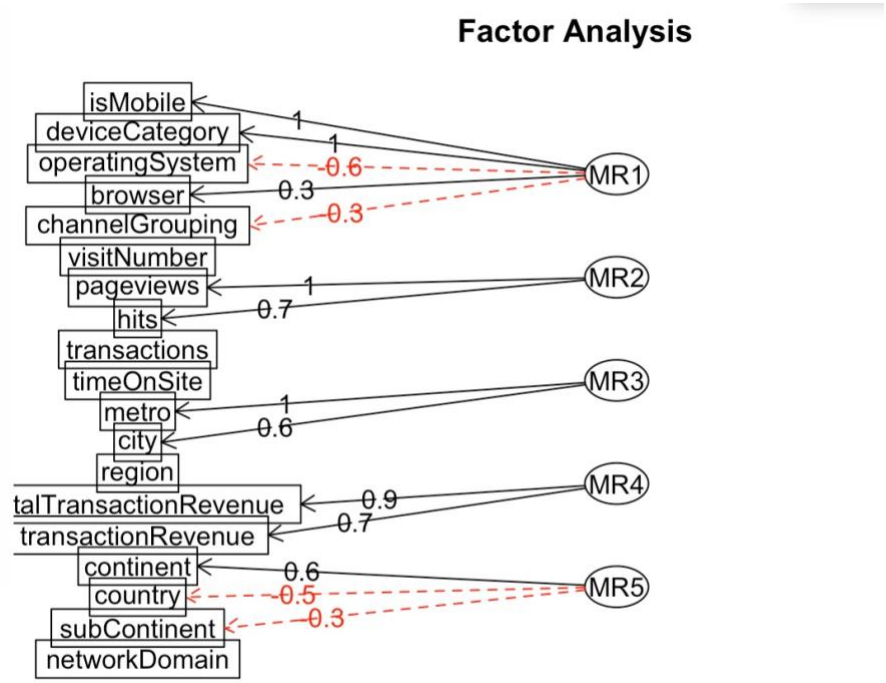
From the table, “hits” and “page views” has correlation of 1, thus I select “page views” instead of hits.

3. Try factor analysis for predictors to reduce dimensions.

Parallel analysis suggests that the number of factors = 6 and the number of components = NA.



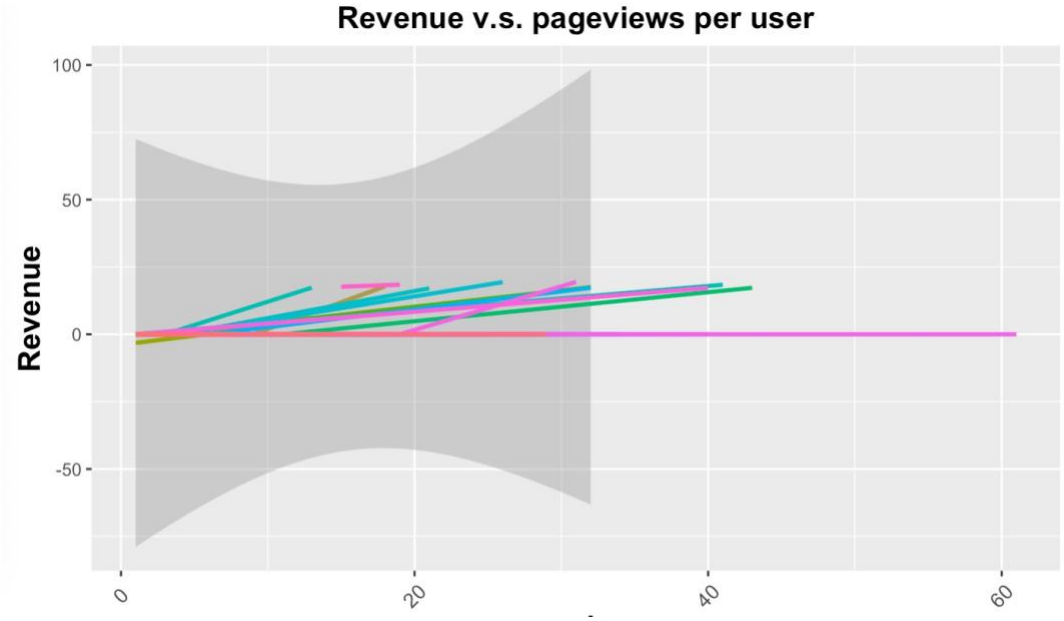
we could see that the factor analysis suggests to put all of our predictors into 5 groups, I conclude as "device", "hits", "city", "continent", "region+browser". The combination of region and browser is weird, so I also tried the number of factors = 5.



For the number of factors = 5, we could see that the factor analysis suggests to put all of our predictors into "device+browser", "hits", "city", "continent", which make more sense here.

4. Predictors Group by user

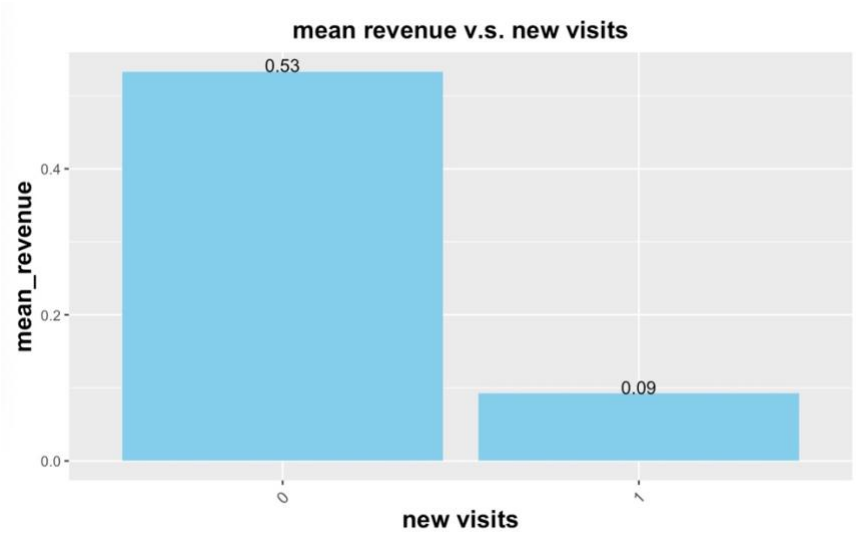
a) page views



CONCLUSION: the slope of “page views” will change a lot between different users.

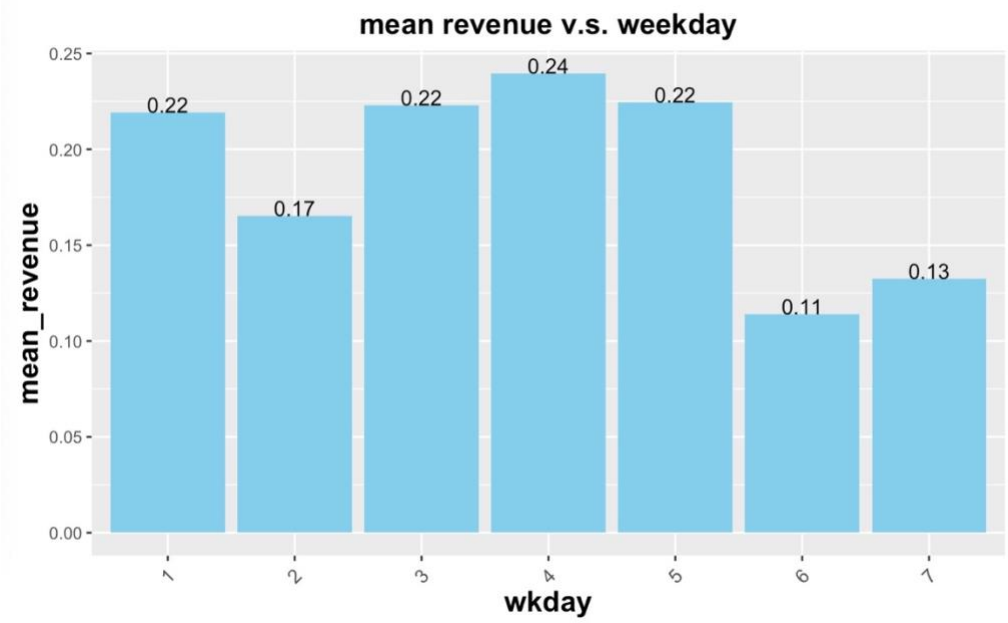
Adding (0+pageviews|fullVisitorId) in the linear mixed models.

b) new visits



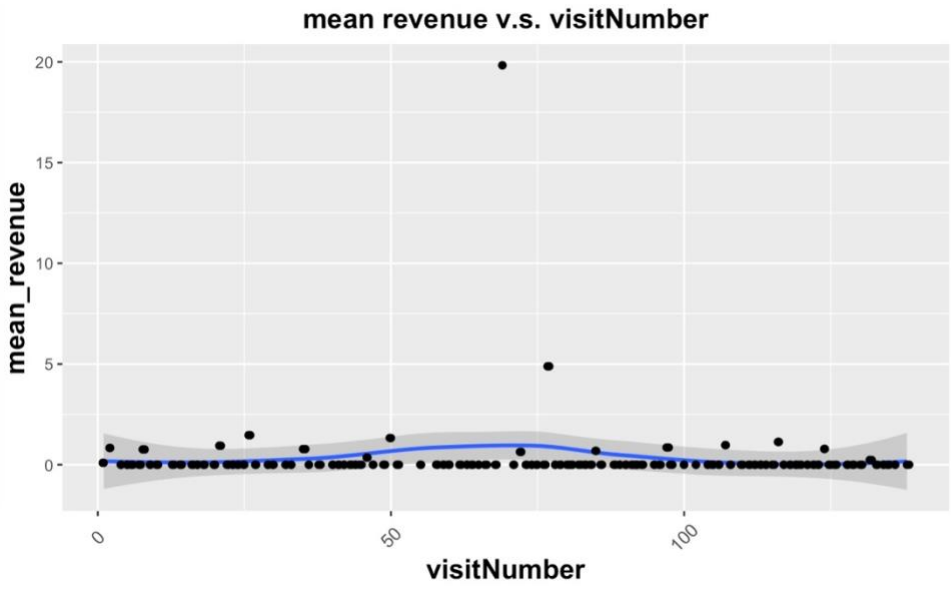
CONCLUSION: “new visits” has great impact on outcome, thus it should be included in the model.

c) dates



CONCLUSION: “wkdays” has great impact on outcome, thus it should be included in the model.

d) visit number



CONCLUSION: “visit number” may great impact on outcome, thus it may be

included in the model.

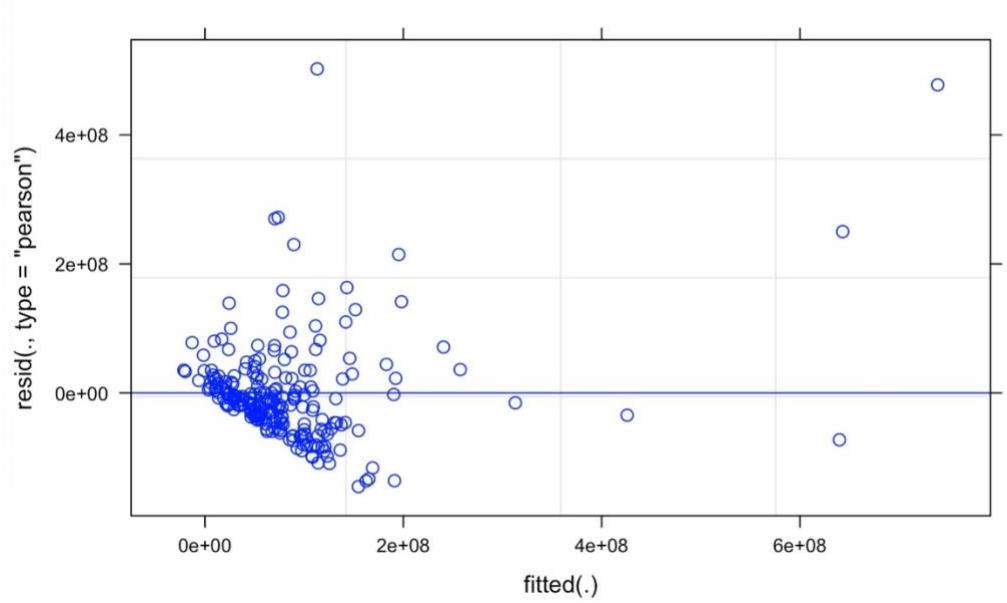
e) Similar ways to deal with “browser”, “operating system”, “is mobile” and other predictors.

5. Model

1. Fit linear mixed model group by user with possible predictors.

```
Linear mixed model fit by REML ['lmerMod']
Formula: transactionRevenue ~ (1 | fullVisitorId) + (1 | pageviews) +      factor(browser) + scale(pageviews) +
factor(newVisits) +
    scale(visitNumber) + factor(operatingSystem) + factor(isMobile) +      factor(continent) + factor(isTrueDirect) +
factor(wkday)
   Data: factor_train

REML criterion at convergence: 7938

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5200 -0.4358 -0.1221  0.2351  5.2546

Random effects:
 Groups        Name        Variance  Std.Dev.
 fullVisitorId (Intercept) 4.293e+08     20719
 pageviews     (Intercept) 1.136e+16 106560234
 Residual                  9.138e+15  95590952
Number of obs: 220, groups:  fullVisitorId, 219; pageviews, 54

Fixed effects:
                                  Estimate Std. Error t value
(Intercept)                        11039471   46877100   0.235
factor(browser)Edge               -80980558   75275012  -1.076
factor(browser)Firefox             35456762   59510695   0.596
factor(browser)Internet Explorer -56071859  107624822  -0.521
factor(browser)Opera             -49241685   90562716  -0.544
factor(browser)Safari             87040063   37703488   2.309
```

Check residuals using plot(MD1) to see if the **linear mixed model** follows assumptions (1. The errors have constant variance. 2. The errors are independent. 3. The errors are Normally distributed.**)**

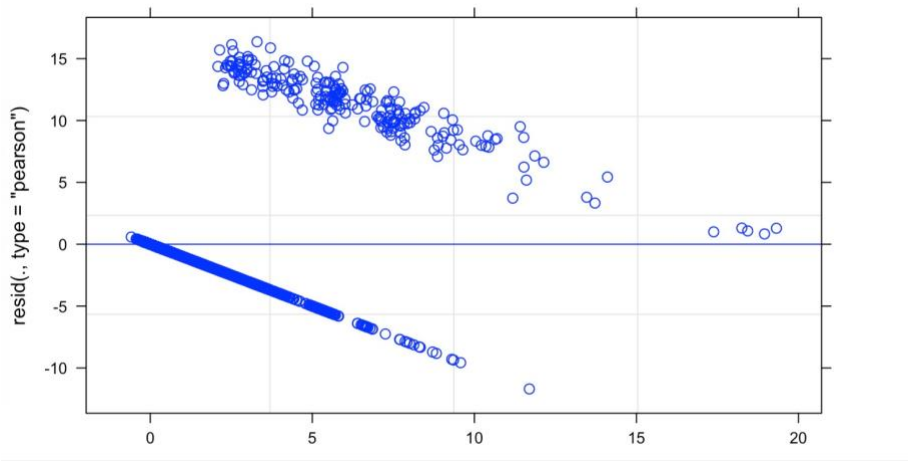


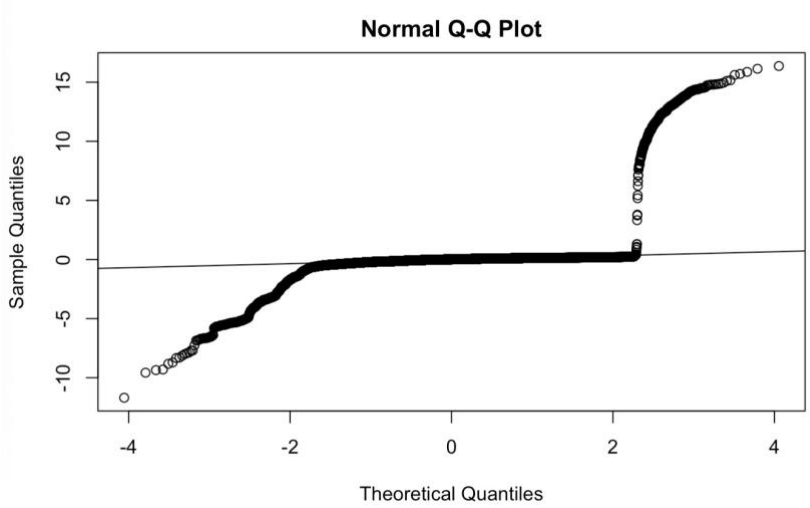
Residuals spread out, so the model needs LOG TRANSFORMATION. After

transformation, I get

```
Linear mixed model fit by REML ['lmerMod']
Formula: log(transactionRevenue + 1) ~ (1 | fullVisitorId) + (1 | pageviews) +
    factor(browser) + scale(pageviews) + factor(newVisits) +
    scale(visitNumber) + factor(operatingSystem) + factor(isMobile) +
    factor(continent) + factor(isTrueDirect) + factor(wkday)
   Data: factor_train

REML criterion at convergence: 75450.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.8911 -0.0868  0.0103  0.0667 11.0438

Random effects:
 Groups        Name        Variance Std.Dev.
 fullVisitorId (Intercept)  0.3019  0.5495
 pageviews     (Intercept) 21.8442  4.6738
 Residual                   2.1965  1.4820
Number of obs: 20000, groups:  fullVisitorId, 19797; pageviews, 76

Fixed effects:
                                    Estimate Std. Error t value
(Intercept)                         2.066263   1.818689   1.136
```

Check Residuals:

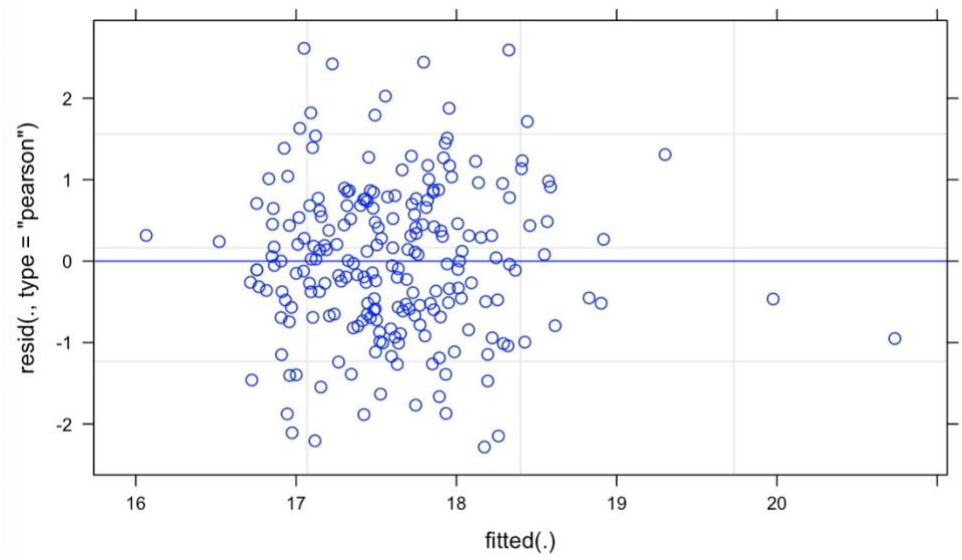


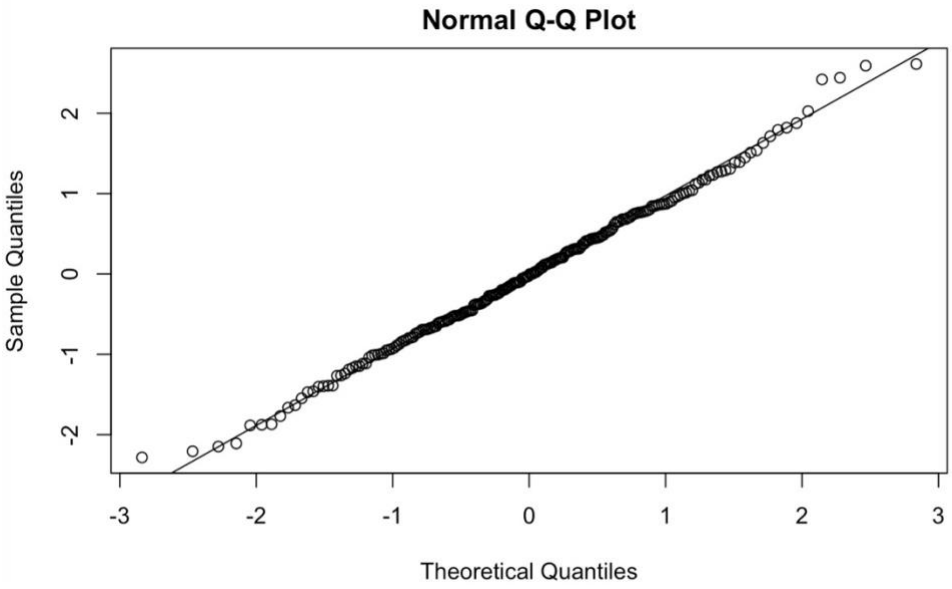


Maybe the pattern in residual plots is caused by too much zero revenue. I checked with all none zero revenue dataset and get:

```
Linear mixed model fit by REML ['lmerMod']
Formula: log(transactionRevenue + 1) ~ (1 | fullVisitorId) + (1 | pageviews) +
    factor(browser) + scale(pageviews) + factor(newVisits) +
    scale(visitNumber) + factor(operatingSystem) + factor(isMobile) +
    factor(continent) + factor(isTrueDirect) + factor(wkday)
   Data: factor_train_use

REML criterion at convergence: 613.5

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.31048 -0.62969 -0.01411  0.67149  2.64360

Random effects:
 Groups        Name        Variance Std.Dev.
 fullVisitorId (Intercept) 0.0000   0.0000
 pageviews     (Intercept) 0.0000   0.0000
 Residual                  0.9766   0.9882
Number of obs: 220, groups:  fullVisitorId, 219; pageviews, 54

Fixed effects:
                             Estimate Std. Error t value
(Intercept)                  17.50606    0.35322  49.561
factor(browser)Edge          -1.41551    0.73561  -1.924
```

Check residuals:





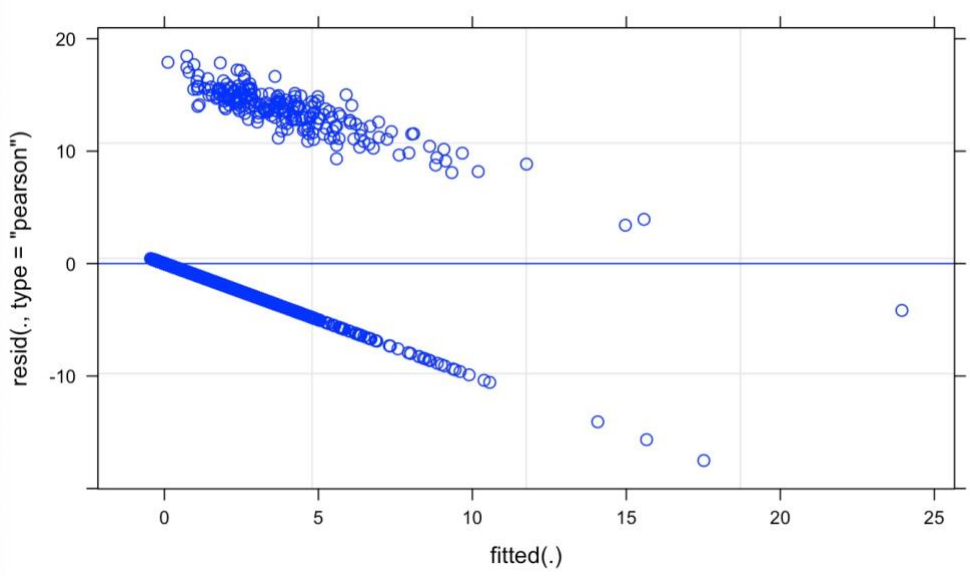
The residual looks normal.

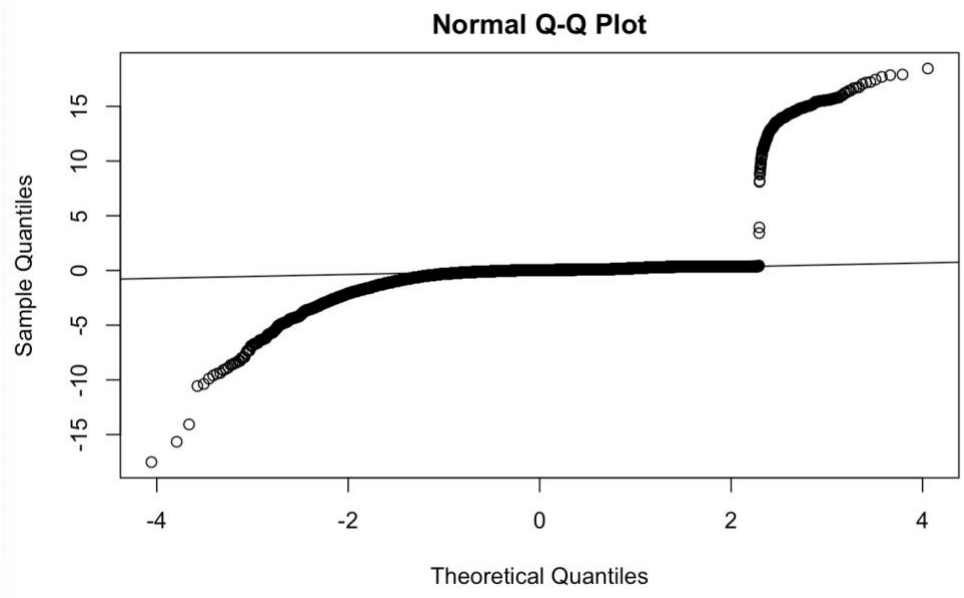
2. Fit linear mixed model group by user with possible predictors.

a) Dataset contains zero revenues with log transformation:

```
fit by REML ['lmerMod']
Formula: log(transactionRevenue + 1) ~ (1 | continent) + (0 + pageviews |
    continent) + factor(browser) + scale(pageviews) + factor(newVisits) +
    scale(visitNumber) + factor(operatingSystem) + factor(isMobile) +
    factor(continent) + factor(isTrueDirect) + factor(wkday)
   Data: factor_train

REML criterion at convergence: 76188.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.7888  -0.0829   0.0202   0.0627  11.3685

Random effects:
 Groups      Name        Variance Std.Dev.
 continent   (Intercept) 2.108509 1.45207
 continent.1 pageviews   0.004882 0.06987
 Residual                2.635691 1.62348
Number of obs: 20000, groups:  continent, 6

Fixed effects:
                                    Estimate Std. Error t value
(Intercept)                         0.453703   2.207580   0.206
```

Check residuals:

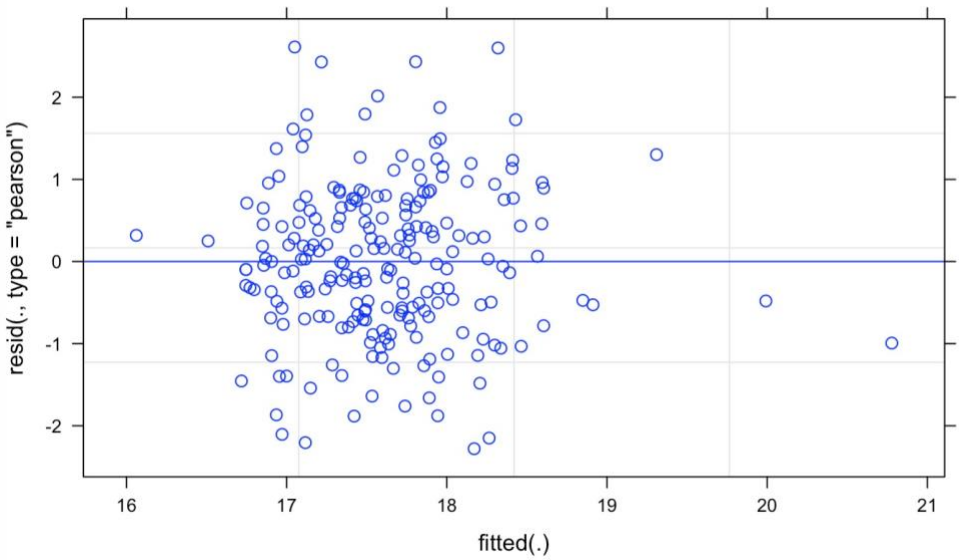


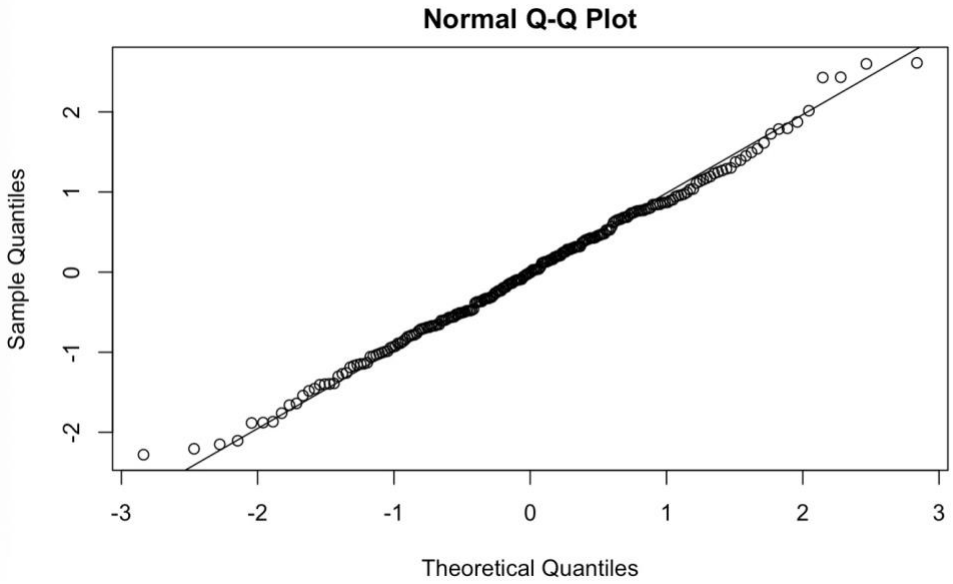


b) Dataset contain only none-zero revenue with log transformation:

```
Linear mixed model fit by REML ['lmerMod']
Formula: log(transactionRevenue + 1) ~ (1 | continent) + (0 + pageviews |
    continent) + factor(browser) + scale(pageviews) + factor(newVisits) +
    scale(visitNumber) + factor(operatingSystem) + factor(isMobile) +      factor(isTrueDirect) + factor(wkday)
   Data: factor_train_use

REML criterion at convergence: 617.3

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.31447 -0.66374 -0.00402  0.67810  2.65228

Random effects:
 Groups      Name        Variance Std.Dev.
 continent   (Intercept) 0.0000   0.0000
 continent.1 pageviews   0.0000   0.0000
 Residual                0.9699   0.9848
Number of obs: 220, groups:  continent, 3

Fixed effects:
                             Estimate Std. Error t value
(Intercept)                  17.50936    0.35199  49.744
factor(browser)Edge          -1.44214    0.73213  -1.970
```

Check residuals:





3. **Fit logistic model to predict if the transaction will produce revenue with possible predictors (with interaction).**

```
Call:
glm(formula = iftransaction ~ scale(pageviews) + factor(newVisits) +
    factor(browser) + factor(operatingSystem) + factor(isMobile) +
    factor(continent) + factor(wkday) + factor(isTrueDirect) +
    scale(visitNumber) + pageviews * browser + visitNumber *
    operatingSystem + visitNumber * isMobile + visitNumber *
    isTrueDirect + pageviews * operatingSystem + visitNumber *
    browser, family = binomial(link = "logit"), data = train_logis)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
 -8.49    0.00    0.00    0.00    8.49

Coefficients: (77 not defined because of singularities)
                                          Estimate Std. Error    z value
(Intercept)                             1.230e+15  8.027e+07   15321061
scale(pageviews)                        4.638e+15  8.563e+07   54163174
factor(newVisits)1                     -2.404e+14  2.781e+06  -86446930
factor(browser)Amazon Silk             -1.454e+15  7.503e+07  -19382555
factor(browser)Android Browser         -2.042e+15  7.854e+07  -25996930
factor(browser)Android Webview         -2.254e+15  6.988e+07  -32263145
factor(browser)BlackBerry               1.489e+15  1.071e+08   13911117
factor(browser)Chrome                   9.109e+14  6.975e+07   13059538
factor(browser)Coc Coc                 -1.033e+15  7.406e+07  -13944465

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  2421.9  on 19999  degrees of freedom
Residual deviance: 15931.3  on 19894  degrees of freedom
AIC: 16143

Number of Fisher Scoring iterations: 25
```

Used formula to predict accuracy for training data is 0.98895

sum(pre_logis_MD3_train==train_logis$iftransaction)/nrow(train_logis)

Use the same model and formula to predict accuracy for training data is 0.9887933.

These seem to be pretty high, but in fact, for the training dataset, it only predict 41 correct none-zero transaction, there are actually 220 none-zero transaction. This illusion is caused by the number whole zero transaction is so large.

6. Interpretation

1. Linear Mixed Model group by user.

Predictors	+/-	estimate
factor(browser)Edge	-	-1.41551
scale(pageviews)	+	0.43032
factor(newVisits)1	-	-0.37788
factor(operatingSystem)Chrome OS	+	0.67858
factor(wkday)4	-	-0.35928

2. Linear Mixed Model group by continent.

Predictors	+/-	estimate
scale(visitNumber)	-	0.088986
scale(pageviews)	+	0.207064
factor(wkday)2	-	-0.067022

From the tables, we can see that “page views” is the most frequent predictor that shows on the top of the list, which means it is the most important predictor to predict customer revenue. Also, all of its signs are positive, which means higher

Revenue is expected when there are more “page views”. Also, impact of “page views” to prediction of revenue may change between users and continents, thus it is the most indispensable predictor through all predictors.

Then, “visit number” is a significant predictor for revenue. Its signs are negative, thus more “visit number” may lead to lower expectation of customer revenue.

Besides, “newVisits” is a significant predictor for revenue. Its signs are negative, if th is the user’s first time visit, the expected chance to have a transaction with revenue is lower.

Next, “operating system” is another crucial predictor. Since the baseline here is "Android", thus users with “Chrome OS”, “Macintosh”, “Windows” are expected to have higher revenue than users with “Android”.

7. Discussions:

1. Assessment of the result

Linear mixed effect model with log transformation leads to the least of 1.934. Submissions are scored on the root mean squared error. It is around 600th ranking in the Kaggle leaderboard. Hopefully, the outcome will be more actionable operational changes and a better use of marketing budgets for those companies who choose to use data analysis on top of GA data.

2. Limitations

Due to time and the computing space if my computer limit, the largest column “hits”, which is also the most interesting column, has not been analyzed. Also, for different type of users, there could be several pattern exists when they shopping, such as we may figure out the costumer’s shopping habit and interest through their user information or maybe combining other shopping history dataset, then build various models depending on their types. However, due to time concerns, I will keep this thoughts for future.

3. Future directions

a) **User:** through the users’ information, based on timeline, for the most active customer, we could predict their chance to generate revenue by their historical revenues. That’s saying to treat users as fixed individual groups.

b) **Complete Dataset：** based on the Complete Dataset, there will be more complete information about every users to be explored.

c) **Factor groups:** I used some interaction in the model, however, based on the factor analysis, there should be more things happening in each factor groups. More random effect could be explored as well as more fixed interactions.

8. Acknowledgement

I would like express many thanks to Professor Masanao Yajima for helping me gaining all these modeling knowledge. Also, I would like to thank Teaching Assistant Skyler Xu for helping me out with the problems I encountered.

9. Reference

1. Notes for competition on kaggle.com: https://www.kaggle.com/sudalairajkumar/simple-exploration-baseline-ga-customer-revenue
2. Notes for competition on kaggle.com: https://www.kaggle.com/julian3833/1-quick-start-read-csv-and-flatten-json-fields