어플리케이션 성공 요인 분석

1 팀 주제분석

이재일 김정민 양지연 임소영 한대룡

1. 서론

최근 우리나라의 스마트폰 사용자가 5천만 명에 육박하면서 스마트폰은 이제 현대인의 생활 필수품이 되었다. 이에 스마트폰 어플리케이션 시장도 나날이 활성화되는 추세이다. 2016년 대비 2017년 구글 플레이스토어의 총 수입은 34.2% 증가한 20억 1천만 달러였고, 앱스토어의 2017년 총 수입은 2016년 대비 34.7% 증가한 38억 5천만 달러로 앱 시장의 규모가 굉장히 빠른 속도로 성장하고 있다는 것을 알 수 있다. 현재 앱스토어에는 20개 이상의 카테고리가 존재하고, 그 안에서도 하위 카테고리가 나뉘어진다. 그리고 각각의 카테고리에서는 끊임없이 새로운 앱이 등재되고 있다. 그러나 앱스토어의 경우 하위 73% 앱은 최대 월 5000 달러밖에 벌지 못하고 심지어 하위 35%의 앱은 월 수입이 100달러도 되지 않았다. 반면 상위 1.6%의 앱은 월 50만 달러 이상을 벌어들이고 있는 것을 보아 어플리케이션 시장 내 수입 불균형이 아주 심각한 것을 확인할 수 있었다.

때문에 이러한 시장 불균형에 대한 원인을 찾기 위해 상위 어플의 인기 요인을 통계적인 분석으로 알아낼 수 있다면 앱 개발자들을 위한 지표가 될 수 있을 것이다. 따라서 본 연구를 통해 앱스토어에 등재된 표면적인 어플의 정보와 어플 순위와의 상관관계를 나타내는 모델을 적합하여 해석하고자 하였다.

2. 연구대상과 분석목표

2.1. 데이터 소개

분석에 사용할 데이터는 2013년 앱스토어 내의 앱 정보를 크롤링한 데이터이다. 앱스토어 내에는 아래처럼 23개의 상이한 카테고리가 존재하므로 카테고리별 1위부터 300위까지의 어플만 추려 6900개의 obs를 얻었다.

Category			
Food&Drink	Health&Fitness	NewsStand	SocialNetworking
Books	Business	Catalogs	Education
Entertainment	Finance	Games	LifeStyle
Medical	Music	Navigation	News
Photo&Video	Productivity	References	Sports
Travel	Utilities	Weather	

Table 1 Category 변수 종류

다음은 각 변수와 그에 대한 간략한 설명이다.

변수 명	변수 설명
AppID	앱ID
Rank	순위
Name	이름
ReleaseData	앱스토어에 등재된 일자
Category	앱이 속한 카테고리
Price	앱의 가격
Seller	판매자
Developer	개발자
Screenshot	앱스토어에 업로드 된 스크린샷 수
Size	앱의 용량
Description	앱에 대한 요약
StarsAllVersion	모든 버전의 총 별점
RatingsAllVersion	모든 버전의 총 리뷰 수
StarsCurrentVersion	현재 버전의 별점
RatingsCurrentVersion	현재 버전의 리뷰 수
Version	앱의 버전
TimeStamp	크롤링한 날짜
WhatsNew	최근 업데이트 내용
UpdatedData	최근 업데이트 날짜

Table 2 전처리 전 변수 설명

여기서 Rank는 분석의 반응변수가 되며, 나머지 변수들은 수정과 삭제를 거쳐 다양한 설명변수로 활용하였다.

2.2. 분석목표

앞서 언급했듯 분석의 목적은 앱의 Rank에 영향을 미치는 변수를 살펴보는 것이다. 따라서 모델을 더 정교화 해줄 수 있는 파생변수를 만들고 유의미한 변수를 찾아 해석을 시도한다. 또한 텍스트마이닝과 동시에 클러스터링을 진행하여 Description 변수의 직관적인 해석으로 좀 더 뚜렷한 결과를 얻을 수 있을 것을 기대한다. 결과적으로 앱스토어에 등재된 여러가지 정보 중 어떤 변수가 얼마나 영향을 미치는지 알아내고 해석할 수 있다면 앱스토어에 포함할 앱의 다양한 정보를 조정하여 앱의 수입을 높이는 데에 도움을 줄 수 있다.

본 연구의 목표는 앱의 매출에 영향을 미치는 변수를 파악하는 것이기 때문에 반응변수를 앱 매출로 하고자 했지만 앱스토어 규정상 매출은 공개하지 않아 어려움을 겪었다. 또한, 해당 앱의 총 매출에 따라 Rank가 결정되기 때문에 Rank는 등간 척도가 아닌 서열 척도라고 할 수 있으며(ex. 순위 한 단위간 매출액의 차이가 상이함) 서열척도의 경우 원칙적으로 사칙연산이 불가능해 Rank 값을 반응변수로 사용할 수 없었다. 따라서 주어진 정보인 앱의 Rank에 -log를 취해 앱의 매출을 추정하였다. 아래의 그래프를 보면 -log를 취한 그래프의 형태가 실제 App Store Revenue의 형태와 유사함을 알 수 있다.

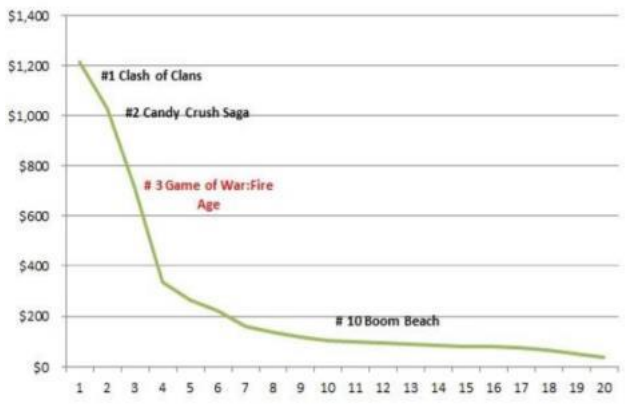

Figure 1 App Store의 Revenue

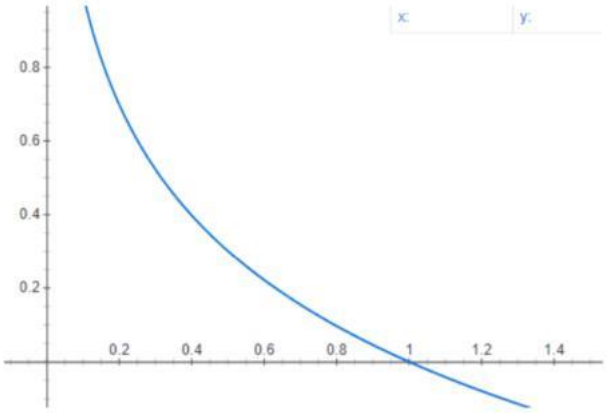

Figure 2 – logX의 그래프(우)

이후 출시 분기, 카테고리, 가격, 판매사 등급, 스크린샷 개수, 앱 용량, 총 리뷰 수, 현재 버전 리뷰 수, 전체 평점, 현재 버전 평점, 출시일로부터 지난 날짜, 마지막 업데이트로부터 지난 날짜, 유/무료 여부를 설명 변수로 활용하였다.

3. 연구과정

3.1. 선형회귀모델

반응변수인 앱 매출에 대한 설명변수들의 영향력을 확인하기 위해 해석이 용이한 선형 회귀 모델을 사용하였다. 우선적으로 불필요한 설명변수들을 제거하기 위해 Hybrid stepwise selection 기법을 이용한 결과 가격, 판매사 등급, 전체 평점, 총 리뷰 수, 출시 분기, 출시일로부터 지난 날짜, 마지막 업데이트로부터 지난 날짜가 유의미한 변수로 선택되었다. F-test의 P-value는 굉장히 작아 모델 자체는 유의미하나 수정된 결정계수는 0.3146로 활용한 설명변수들이 반응변수를 충분히 설명하지 못하는 것으로 확인되었다.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -5.6248207  0.0690683 -81.439  < 2e-16 ***
CategoryBusiness           0.0118954  0.0667331   0.178 0.858529
CategoryCatalogs           0.6049171  0.0683784   8.847  < 2e-16 ***
CategoryEducation         -0.1951052  0.0661265  -2.950 0.003184 **
CategoryEntertainment     -0.4170689  0.0671742  -6.209 5.67e-10 ***
CategoryFinance            0.2015523  0.0673814   2.991 0.002789 **
CategoryFood & Drink       0.3964066  0.0662383   5.985 2.29e-09 ***
CategoryGames             -1.4930322  0.0734255 -20.334  < 2e-16 ***
CategoryHealth & Fitness  -0.2148734  0.0661493  -3.248 0.001167 **
CategoryLifestyle         -0.0127666  0.0662066  -0.193 0.847098
CategoryMedical            0.2048978  0.0675534   3.033 0.002430 **
CategoryMusic             -0.2145932  0.0666686  -3.219 0.001294 **
CategoryNavigation         0.0582431  0.0673110   0.865 0.386915
CategoryNews               0.3096938  0.0671504   4.612 4.07e-06 ***
CategoryPhoto & Video     -0.3318680  0.0672528  -4.935 8.23e-07 ***
CategoryProductivity      -0.2765417  0.0674663  -4.099 4.20e-05 ***
CategoryReference          0.0748107  0.0661754   1.130 0.258311
CategorySocial Networking -0.3574461  0.0676090  -5.287 1.28e-07 ***
CategorySports             0.2422279  0.0665056   3.642 0.000272 ***
CategoryTravel             0.3469249  0.0660068   5.256 1.52e-07 ***
CategoryUtilities         -0.1579344  0.0677910  -2.330 0.019851 *
CategoryWeather            0.2782707  0.0669527   4.156 3.28e-05 ***
Price                      0.0223701  0.0015318  14.604  < 2e-16 ***
Seller1                    0.2484553  0.0215382  11.536  < 2e-16 ***
Size                      -0.0121997  0.0075994  -1.605 0.108465
StarsAllVersions          -0.0517666  0.0105558  -4.904 9.62e-07 ***
RatingsAllVersions         0.2790358  0.0084539  33.007  < 2e-16 ***
RatingsCurrentVersion      0.0233452  0.0084229   2.772 0.005594 **
free1                      0.0843269  0.0255221   3.304 0.000958 ***
release_quarter2nd        -0.1304111  0.0296344  -4.401 1.10e-05 ***
release_quarter3rd        -0.1355625  0.0287111  -4.722 2.39e-06 ***
release_quarter4th        -0.0989481  0.0269887  -3.666 0.000248 ***
mondiff                   -0.0180924  0.0009807 -18.448  < 2e-16 ***
updatemondiff             -0.0090054  0.0015135  -5.950 2.82e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7979 on 6488 degrees of freedom
Multiple R-squared:  0.3181,    Adjusted R-squared:  0.3146
F-statistic: 91.71 on 33 and 6488 DF,  p-value: < 2.2e-16
```

Figure 3 선형회귀모델 결과

변수들의 VIF를 확인해본 결과 10 이상인 값이 발견되지 않아 다중 공선성 문제는 없음을 확인하였다.

```
> vif(lm1)
                          GVIF Df GVIF^(1/(2*Df))
Category              3.190472 21        1.028008
Price                 1.477378  1        1.215474
Seller                1.096124  1        1.046959
Screenshot            1.102930  1        1.050205
Size                  1.362527  1        1.167273
StarsAllVersions      3.650585  1        1.910650
RatingsAllVersions    5.009064  1        2.238094
StarsCurrentVersion   3.590810  1        1.894943
RatingsCurrentVersion 2.919346  1        1.708609
free                  1.345708  1        1.160046
release_quarter       1.057077  3        1.009294
mondiff               2.265266  1        1.505080
updatemondiff         1.355728  1        1.164358
```

Figure 4 다중 공선성 확인

이후 회귀분석의 기본 가정인 오차항의 정규성, 등분산성, 독립성을 검증하였다. 이 때 해당 데이터는 동일한 시간에 대한 데이터인 횡단면 데이터이므로 독립성 가정은 의미가 없었으나 잔차 plot을 확인한 결과 정규성과 등분산성을 만족하지 못해 회귀 분석 모델을 사용할 수 없음을 확인하였다.

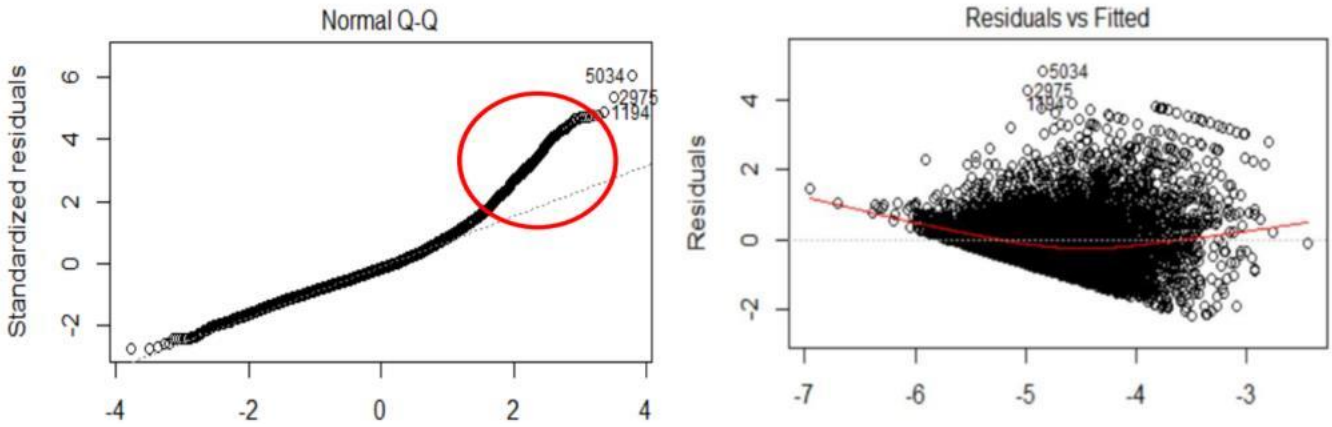

Figure 5 회귀분석 잔차 plot

3.2. 순서형 다항 로짓 모델

반응변수인 Rank 변수가 순서척도를 지니고 있어 1위부터 300위 까지의 랭킹을 적절하게 범주화 한다면 순서형 다항 로짓 모델을 사용하기 적합하다고 판단했다. 또한 앱의 매출이 상위권 랭킹으로 갈수록 급격하게 증가하는 점을 감안해 1위부터 300위 까지의 랭킹을 4개의 범주인 1~50, 51~100, 101~200, 201~300으로 범주화 하여 분석을 진행하였다. 그 후, 모델이 유의한지를 판단하기 위해 적합도 검정을 실시한 결과 P-value가 0.05보다 작아 모델이 유의함을 확인하였다. 그러나 순서형 다항 로짓 모델의 기본 가정인 비례 오즈 가정을 만족하지 않아 해당 모델을 사용할 수 없다는 결론을 내렸다.

```
> nominal_test(multilogit.fit) #비례오즈가정 만족 x
Tests of nominal effects

formula: Newrank ~ Category + Price + Seller + Screenshot + Size + Sta
rsAllVersions + RatingsAllVersions + StarsCurrentVersion + RatingsCurr
entVersion + free + release_quarter + mondiff + updatemondiff
                      Df  logLik    AIC    LRT  Pr(>Chi)
<none>                   -7601.3 15278
Category              42 -7598.0 15356  6.598 1.0000000
Price                  2 -7596.9 15274  8.721 0.0127707 *
Seller                 2 -7593.6 15267 15.244 0.0004896 ***
Screenshot             2 -7600.6 15281  1.267 0.5308562
Size                   2 -7600.4 15281  1.683 0.4310036
StarsAllVersions       2 -7596.7 15274  9.044 0.0108689 *
RatingsAllVersions     2 -7570.6 15221 61.318 4.842e-14 ***
StarsCurrentVersion    2 -7599.4 15279  3.773 0.1515659
RatingsCurrentVersion  2 -7586.9 15254 28.654 5.995e-07 ***
free                   2 -7597.2 15274  8.090 0.0175087 *
release_quarter        6 -7599.4 15287  3.700 0.7172292
mondiff                2 -7600.5 15281  1.633 0.4418971
updatemondiff          2 -7594.0 15268 14.491 0.0007134 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6 비례오즈가정 확인

위의 검정에서 가격, 판매사, 총 평점, 총 리뷰 수, 현재 버전 리뷰 수, 유/무료 여부, 마지막 업데이트로부터 지난 날짜 변수가 비례 오즈 가정을 만족하지 못해 모델이 사용 불가능함을 확인하였다.

3.3. 명목형 다항 로짓 모델

명목형 다항 로짓 모델은 데이터가 비례 오즈 가정을 만족할 필요가 없으며 반응변수에 순서척도가 없는 경우 사용되는 다항 로짓 모델이다. 연구에 사용된 데이터의 경우 반응변수에 순서척도가 있지만 순서형 비례 오즈 가정을 만족하지 못하므로 정보를 일부 손실하면서 사용하였다. 명목형 다항 로짓 모델도 순서형과 마찬가지로 반응변수를 Rank에 따라 4개의 범주로 나눈 뒤 분석을 진행하였다. 유의미한 변수를 확인한 결과는 다음과 같다.

```
> p<0.05 #유의한 변수 확인
     (Intercept) CategoryBusiness CategoryCatalogs CategoryEducation CategoryEntertainment CategoryFinance CategoryFood & Drink
grp2        TRUE            FALSE             TRUE             FALSE                 TRUE           FALSE                TRUE
grp3        TRUE            FALSE             TRUE             FALSE                 TRUE            TRUE                TRUE
grp4        TRUE            FALSE             TRUE             FALSE                 TRUE            TRUE                TRUE
     CategoryGames CategoryHealth & Fitness CategoryLifestyle CategoryMedical CategoryMusic CategoryNavigation CategoryNews
grp2          TRUE                    FALSE             FALSE            TRUE          FALSE              FALSE        FALSE
grp3          TRUE                    FALSE             FALSE            TRUE          FALSE              FALSE         TRUE
grp4          TRUE                     TRUE             FALSE            TRUE          FALSE              FALSE         TRUE
     CategoryPhoto & Video CategoryProductivity CategoryReference CategorySocial Networking CategorySports CategoryTravel
grp2                  TRUE                 FALSE             FALSE                      TRUE          FALSE           TRUE
grp3                  TRUE                  TRUE             FALSE                      TRUE           TRUE           TRUE
grp4                  TRUE                  TRUE             FALSE                      TRUE           TRUE           TRUE
     CategoryUtilities CategoryWeather Price Seller1 Screenshot  Size StarsAllVersions RatingsAllVersions StarsCurrentVersion
grp2             FALSE           FALSE  TRUE   FALSE      FALSE FALSE            FALSE                TRUE                TRUE
grp3             FALSE            TRUE  TRUE    TRUE      FALSE FALSE            FALSE                TRUE                TRUE
grp4             FALSE            TRUE  TRUE    TRUE      FALSE  TRUE            FALSE                TRUE               FALSE
     RatingsCurrentVersion free1 release_quarter2nd release_quarter3rd release_quarter4th mondiff updatemondiff
grp2                 FALSE FALSE                FALSE                FALSE                TRUE    TRUE          TRUE
grp3                  TRUE FALSE                 TRUE                 TRUE                TRUE    TRUE          TRUE
grp4                  TRUE FALSE                 TRUE                 TRUE                TRUE    TRUE          TRUE
```

Figure 7 각 변수의 P-value

분석 결과를 보면 반응 범주마다 유의미한 변수가 다름을 알 수 있다. 이는 비례 오즈 가정이 없기 때문에 각 반응 범주마다 절편과 계수가 모두 다르기 때문이다. 명목형 다항 로짓 모델의 최종 분석 결과는 다음과 같다.

```
> exp(coef(nnet)) #상관계수 보기
     (Intercept) CategoryBusiness CategoryCatalogs CategoryEducation CategoryEntertainment CategoryFinance CategoryFood & Drink
grp2    6.290255        0.9469185       0.32059002         1.266919              2.271062       0.7360176           0.4607566
grp3   54.978554        0.8568704       0.14128474         1.543615              3.440199       0.5475706           0.2695447
grp4  151.084029        0.8189406       0.07405121         1.758131              4.340513       0.4045757           0.1554841
     CategoryGames CategoryHealth & Fitness CategoryLifestyle CategoryMedical CategoryMusic CategoryNavigation CategoryNews
grp2      13.85965                1.304886         0.9197500       0.5225767      1.486446          0.8403286    0.7502055
grp3      71.91666                1.605163         0.8303199       0.4066292      1.717439          0.7801309    0.4688128
grp4     286.93473                1.837603         0.7436371       0.3315421      1.794112          0.6634539    0.2876967
     CategoryPhoto & Video CategoryProductivity CategoryReference CategorySocial Networking CategorySports CategoryTravel
grp2             1.869765              1.732773          0.8425449                 2.285328       0.5664823      0.4819514
grp3             2.640264              2.144192          0.7207101                 3.159391       0.3801448      0.3070949
grp4             3.059987              2.419802          0.6231329                 3.694878       0.2749201      0.1955767
     CategoryUtilities CategoryWeather     Price   Seller1 Screenshot     Size StarsAllVersions RatingsAllVersions StarsCurrentVersion
grp2          1.476893       0.7480366 0.9667029 0.8818331  1.0060040 1.028366        0.9393738          0.5966089           1.222828
grp3          1.561878       0.4983472 0.9465189 0.6581983  0.9828282 1.037604        0.9332649          0.4433945           1.183871
grp4          1.503728       0.2979297 0.9123846 0.3257649  0.9716538 1.078203        1.0405529          0.3369761           1.105874
     RatingsCurrentVersion     free1 release_quarter2nd release_quarter3rd release_quarter4th  mondiff updatemondiff
grp2             0.9582447 0.9796505          1.293672          1.162274          1.314139 1.031977      1.052453
grp3             0.9334856 0.9453442          1.497430          1.441160          1.406129 1.054859      1.065288
grp4             0.9200183 0.8066820          1.666163          1.652459          1.477318 1.075719      1.079335
```

Figure 8 각 변수의 상관계수

3.4. 로지스틱 회귀 모델

명목형 다항 로짓 모델은 여러 개의 범주로 이뤄진 반응변수를 분류할 수 있다는 장점이 있지만 하나의 기준 범주와 나머지 반응 범주를 비교한다는 점에서 그 해석이 직관적이지 못하다. 예를 들어 기준 범주를 201~300위로 두고 비교하면 최상위권인 1~50위와 최하위인 201~300위를 비교하게 된다. 이러한 이유로 반응 변수가 이항 변수일 때 가장 흔히 사용되는 로지스틱 회귀 분석 모델을 이용한 분석을 진행하였다. 분석의 목적이 앱 매출 상승에 영향을 미치는 요인 파악이고 랭킹이 최상위권으로 갈수록 매출이 기하급수적으로 증가하므로 상위권과 하위권을 150:150이 아닌 50:250으로 나누었다. 분석 결과는 아래와 같다.

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -4.672464   0.270051 -17.302  < 2e-16 ***
CategoryBusiness           0.171741   0.253141   0.678 0.497494
CategoryCatalogs           1.838084   0.266246   6.904 5.07e-12 ***
CategoryEducation         -0.411103   0.253924  -1.619 0.105447
CategoryEntertainment     -1.155611   0.261058  -4.427 9.57e-06 ***
CategoryFinance            0.629379   0.257607   2.443 0.014559 *
CategoryFood & Drink       1.267051   0.259127   4.890 1.01e-06 ***
CategoryGames             -4.081618   0.288288 -14.158  < 2e-16 ***
CategoryHealth & Fitness  -0.431730   0.250548  -1.723 0.084863 .
CategoryLifestyle          0.189316   0.257764   0.734 0.462672
CategoryMedical            0.918447   0.260339   3.528 0.000419 ***
CategoryMusic             -0.508309   0.261578  -1.943 0.051987 .
CategoryNavigation         0.249218   0.270135   0.923 0.356231
CategoryNews               0.744222   0.264598   2.813 0.004913 **
CategoryPhoto & Video     -0.892401   0.257082  -3.471 0.000518 ***
CategoryProductivity      -0.695021   0.257343  -2.701 0.006918 **
CategoryReference          0.302601   0.256460   1.180 0.238035
CategorySocial Networking -1.039622   0.260597  -3.989 6.62e-05 ***
CategorySports             0.958398   0.256340   3.739 0.000185 ***
CategoryTravel             1.156856   0.253725   4.559 5.13e-06 ***
CategoryUtilities         -0.395381   0.260585  -1.517 0.129196
CategoryWeather            0.684051   0.265376   2.578 0.009947 **
Price                      0.053424   0.005427   9.845  < 2e-16 ***
Seller1                    0.484340   0.080535   6.014 1.81e-09 ***
RatingsAllVersions         0.788344   0.034206  23.047  < 2e-16 ***
StarsCurrentVersion       -0.137082   0.042850  -3.199 0.001379 **
RatingsCurrentVersion      0.062358   0.030041   2.076 0.037911 *
release_quarter2nd        -0.387668   0.115645  -3.352 0.000802 ***
release_quarter3rd        -0.326432   0.112544  -2.900 0.003726 **
release_quarter4th        -0.331523   0.105504  -3.142 0.001676 **
mondiff                   -0.051484   0.003793 -13.572  < 2e-16 ***
updatemondiff             -0.062427   0.010817  -5.771 7.87e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5864.2  on 6521  degrees of freedom
Residual deviance: 4359.7  on 6490  degrees of freedom
AIC: 4423.7

Number of Fisher Scoring iterations: 6
```

Figure 9 로지스틱 회귀 모델 결과

대부분의 앱 카테고리, 가격, 판매사 등급, 총 리뷰 개수, 현재 버전 평점, 출시일로부터 지난 날짜, 업데이트 여부가 유의미한 변수로 확인되었다. 해당 결과에서 카테고리 변수의 계수는 해석에 주의가 필요하다. 어떠한 카테고리 변수의 계수가 음수일 경우 동일한 App으로 기준 카테고리(Book 카테고리)일 경우에 비해 낮은 매출을 기록할 것이라고 해석이 가능하다. 구체적인 예시를 들자면 'Game 카테고리는 계수가 -4.18로 기준 카테고리에 비해 높은 매출을 기록하기 힘들다' 라고 할 수 있으며 일반화하자면 카테고리 변수의 계수가 낮을수록 인기 있는 카테고리라고 할 수 있다.

```
> lrtest(glm.fit3) #모델 유의함
Likelihood ratio test

Model 1: Rank ~ Category + Price + Seller + Screenshot + Size + StarsAllVersions +
    RatingsAllVersions + StarsCurrentVersion + RatingsCurrentVersion +
    free + release_quarter + mondiff + updatemondiff
Model 2: Rank ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  36 -2178.5
2   1 -2932.1 -35 1507.2  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10 Likelihood ratio test

적합도 검정을 실시한 결과 P-value 또한 0.05 미만으로 모델이 유의하다고 판단하였다.

3.5. 텍스트마이닝 & 클러스터링

이전의 분석에서는 어플의 성공 요인으로 스크린 샷의 개수, 별점, 가격 등을 변수로 설정하고 분석을 진행하였다. 하지만 이러한 변수들은 각각의 앱 자체의 기능과 특징을 설명해주지 않기 때문에 분석의 결과를 향상시키기 위해 원래 데이터에서 사용하지 않았던 변수인 'Description' 변수를 사용하였다. 이 때, 'Description' 변수는 해당 앱에 대한 소개가 텍스트 형식으로 나와있어 텍스트 마이닝을 진행하였다. 텍스트 마이닝이란 자연어처리 기술을 바탕으로 비정형 텍스트에서 특징이나 패턴을 도출해 의미 있는 정보를 창출하고 활용하는 과정이다. 본 연구에서는 텍스트 마이닝을 이용해 앱의 description에서 키워드의 빈도수를 파악하고, 이를 클러스터링 기법을 사용해 어플의 기능과 특징을 나타내는 새로운 변수를 생성하였다.

모든 카테고리의 앱을 포함한 원래 데이터에서 텍스트 마이닝을 진행하고 이를 워드 클라우드로 시각화해본 결과, [그림]과 같은 결과가 나타났다. 이를 보면, 'weather', 'map', 'friend'순으로 빈도수가 높게 나타나는 것을 알 수 있는데, 이를 이상하게 여겨 'weather' 카테고리 내에서만 다시 워드 클라우드로 시각화하고 출현 빈도를 구해본 결과, 전체 300개의 obs 중 'weather'라는 단어가 296회 출현하였고, 전체 카테고리의 6522개의 obs 중 'weather'라는 단어는 331회 출현하였음을 알 수 있었다.

Figure 11 'Weather' 카테고리의 워드클라우드

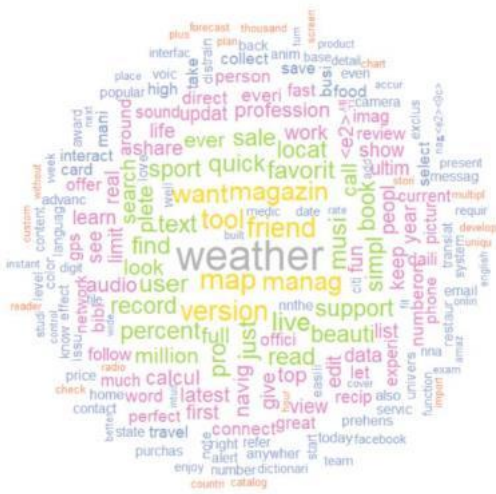

Figure 12 전체 카테고리의 워드클라우드

'weather' 카테고리에 높은 빈도수로 등장하는 단어가 전체 결과 값에 큰 영향을 미치는 것을 발견함으로써, 특정 카테고리가 전체 결과에 대해 영향력을 미친다고 판단해, 전체 카테고리에 대한 텍스트 마이닝은 카테고리 별 상이한 정보나 패턴을 모두 고려하기 어렵기 때문에 카테고리 별로 텍스트 마이닝을 진행하였다.

본 연구에서는 카테고리 별로 텍스트 마이닝을 상위 5개의 카테고리에 대해서만 분석을 진행하였다. 여기서 상위 5개의 카테고리는 기존의 로지스틱 회귀분석에서의 결과에서 'Category' 변수 내의 유의한 카테고리를 estimate값으로 오름차순으로 나열하고, 이 중 estimate값이 가장 작은 순서대로 5개를 선정하였다. 로지스틱 회귀분석에서의 반응변수 Y는 sales값으로, 'Category' 변수 내의 모든 범주들은 sales를 같은 값으로 설정하게 된다. 여기서 'Category' 변수의 범주 내에서 Book 카테고리가 기준 범주로 설정되었는데, estimate값이 Book보다 작은 값을 가진다는 것은 해당 카테고리가 Book 카테고리보다 높은 Sales를 만족하기가 어렵다는 것을 의미한다. 이를 실제 연도별 앱의 인기도를 측정한 순위와 비교해 보았을 때, 상위권의 순위가 상당히 유사함을 확인할 수 있었다.

분석의 흐름은 카테고리별로 같기 때문에, 가장 순위가 높은 'Games' 카테고리에서의 분석과정을 서술하겠다. 분석은 크게 'Description'내의 텍스트를 정제해 DTM(Document Term Matrix)을 만들고, 이를 클러스터링을 통해 그룹화한 것을 새로운 변수로 설정하고, 이를 기존의 변수들과 함께 로지스틱 모형에 적합하는 방식으로 진행되었다.

먼저, 앱 별로 Description의 길이가 다르기 때문에 텍스트 내의 단어 개수를 비교하여 앞에서부터 50단어가 되게 길이를 조정해주었다. 이는 평균적으로 두 개의 문장에 있는 단어의 개수와 비슷한데, 실제 사용자들이 앱을 다운로드하기 전에 처음 한두 문장의 앱 설명만을 읽는다는 것에 기인한 것이기도 하다. 텍스트 마이닝을 시작하기 위해, 우선 'Description' 변수의 텍스트를 텍스트 처리를 해주기 위한 Corpus라는 벡터에 저장을 시켰다. 다음은 Parsing 과정으로, 문장의 구조를 알아내는 작업이다. Parsing을 진행하기 위해서는 먼저 텍스트의 모든 단어들을 소문자로 바꿔줄 필요가 있어서, 이를 진행하였다.

```
[1] Lead your clan to victory! Clash of Clans is an epic combat strategy game. Build your village, train your tr
attle with thousands of other players online!\n\nPLEASE NOTE! Clash of Clans is completely free to play, however
 items can also be purchased for real money. If you
```

Figure 13 기존의 텍스트

```
[1] lead your clan to victory! clash of clans is an epic combat strategy game. build your village, train your tr
attle with thousands of other players online!\n\nplease note! clash of clans is completely free to play, however
 items can also be purchased for real money. if you
```

Figure 14 소문자화된 텍스트

다음으로, 불필요하거나 무의미한 단어를 지워주는 과정이다. 이 과정에서는 진행할 데이터마다 지워줄 단어가 다른데, 우선 해당 데이터가 앱 데이터이기 때문에 이와 관련된 'mobile', 'app', 'iphone' 과 같은 단어들을 삭제하였다. 그리고 해당 카테고리는 게임 카테고리이기 때문에, 이와 관련된 'game', 'play', 'fun' 과 같은 단어들을 추가로 삭제해주었다. 또, 이 Parsing 과정에서 추가적으로 불필요한 문법이나, 숫자, 문장부호, 특수문자, 그리고 공백까지 제거해주었다.

```
[1] lead clan victory clash clans epic bat strategy village train troops battle thousands ers online please
ns pletely however items also purchased real money
```

Figure 15 Parsing과정 후의 텍스트

다음은 Stemming 과정으로, 과거형, 복수형, 진행형 같이 되어있는 단어들을 어원으로 만들어주었다. Stemming이 끝난 텍스트를 보게 되면 완전한 어원이 아니라 의미를 파악할 수 있을 만한 정도의 어원으로 바뀐 걸 확인할 수 있는데, 이처럼 의미 파악을 하기 위해서는 직접 확인해봐야 한다.

```
[1] lead clan victori clash clan epic bat strategi villag train troop battl thousand
howev item also purchas real money
```

Figure 16 Stemming과정 후의 텍스트

이렇게 Stemming을 거친 후에는, 단어와 텍스트의 관계를 보기 위해 Corpus 형태로 되어있는 텍스트를 DTM, 즉 행렬 형태로 변환시켜 주었다. 여기서 행렬 내의 각각의 값은 해당 Description에서 그 단어의 빈도수를 나타낸다. 그 후, 이 DTM 내에서 빈도수가 작거나 중요도가 떨어진다고 생각하는 단어들을 제거해주는 작업을 진행하였다.

	Terms									
Docs	addict	adventur	battl	creat	experi	join	million	monster	slot	word
120	0	0	2	0	0	0	0	0	0	0
122	0	0	0	0	0	0	0	6	0	0
159	0	1	1	0	0	0	0	1	0	0
176	0	0	0	0	0	0	0	0	0	0
179	0	0	0	0	0	0	0	0	0	0
189	0	1	0	0	0	0	0	0	0	0
243	0	0	0	0	0	0	1	0	0	0
25	0	0	0	0	0	0	0	0	0	0
250	0	0	0	0	0	0	0	0	0	0

Figure 17 게임 카테고리의 DTM

그 후, 단어 별 빈도수를 확인하기 위해 워드 클라우드로 시각화를 해주었다. 다음으로, 빈도수가 높은 단어들을 키워드라고 파악하고, 이 키워드들을 대상으로 군집분석을 실시하였다. 군집분석의 방법으로는 덴드로그램을 통해 이해와 해석이 용이한 계층적 군집분석을 사용하였다. 계층적 군집분석의 평균연결법, 완전연결법, 와드연결법 등을 적용해본 결과, 가장 효율적으로 분할되는 와드연결법을 사용하였다.

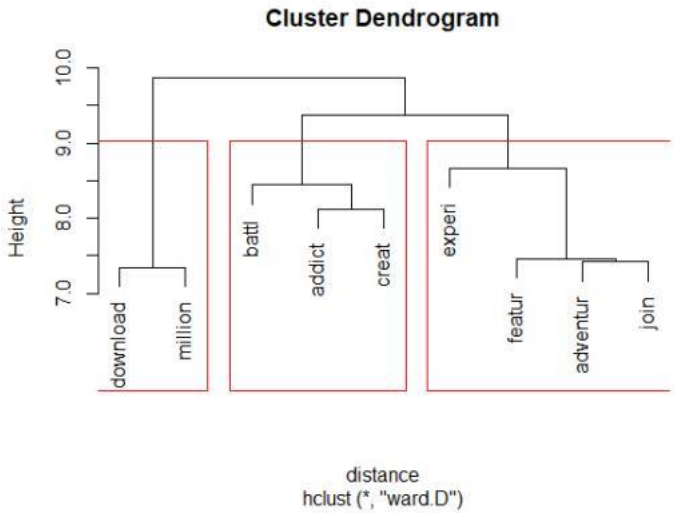

Figure 18 'Games' 카테고리의 워드클라우드 **Figure 19 'Games' 군집분석 결과 - 와드연결법**

군집분석에서 나온 결과를 바탕으로 군집 별로 의미를 파악하기 위해 해당 단어가 있는 텍스트를 찾고, 동일 군집 내에 있는 단어들 간의 연관성을 바탕으로 Labeling을 해주었다.

	Cluster1	Cluster2	Cluster3
N	3	4	2
Terms	Battle addict create	Adventure join feature experience	Download million
Example	"Battle legendary monsters and lead your kingdom to victory! Join your friends to create an army and defeat epic bosses." "Knights & Dragons is an addictive combination of strategic combat and fantasy"	"Join millions of players worldwide in this exciting city-building adventure game. "TONS OF FREE FUN FEATURES:"	"#1 FREE GAME in countries on the AppStore with over 3 Million Downloads!!"
Label	Battle & create	adventure	worldwide

Table 3 군집 분석 – 의미 파악

아래의 표는 분석을 진행한 상위 5개 카테고리의 각 클러스터에 이름을 부여한 클러스터 label이다.

Category	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Game	Battle & Create	Adventure	Worldwide	
Entertainment	Video	Game	Photo	
Social Networking	Communicate	Worldwide	New friends	
Photo & Video	Filter	Share	Performance	Using
Productivity	Text	Easy, Powerful	(?)	

Table 4 카테고리별 Cluster Label

다음으로, 각 앱의 Description마다 해당 군집에 속하는 단어가 있을 때마다 해당 군집 별로 1점씩 부여하는 과정을 진행하고, 이들 군집을 기존 game 카테고리의 데이터와 결합해주었다.

텍스트마이닝과 클러스터링으로 생성한 변수를 기존 데이터와 결합해 최종 모델인 로지스틱 회귀 모델에 적합하고, hybrid subset selection을 적용한 결과, 총 15개의 설명변수 중 10개의 변수가 유의한 수준으로 나왔으며 그림 11에서 볼 수 있듯이 새로 추가한 Cluster1과 Cluster3는 앱의 성공 요인임을 알 수 있다. 또한, ROC커브를 그린 후 AUC를 구해보니 0.84의 값을 가졌고, 이는 모델의 분류 성능이 뛰어남을 증명해주는 지표이다.

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.64404    1.80136  -0.358 0.720695
Price                 0.22957    0.11422   2.010 0.044442 *
Seller1               0.72795    0.29964   2.429 0.015122 *
Size                 -0.35879    0.15968  -2.247 0.024645 *
StarsAllversions     -1.20081    0.42820  -2.804 0.005042 **
RatingsAllversions    0.50590    0.13413   3.772 0.000162 ***
RatingsCurrentVersion 0.32026    0.11305   2.833 0.004612 **
mondiff              -0.07846    0.02143  -3.661 0.000251 ***
updatemondiff        -0.56665    0.15909  -3.562 0.000368 ***
Cluster1              0.38883    0.19533   1.991 0.046522 *
Cluster3              0.52939    0.25581   2.069 0.038505 *
```

Figure 20 최종 모델 summary (Game Category)

결과의 해석은 다음과 같다. 특정 변수 X1이 한 단위 증가할 때 100위 안에 들 오즈가 해당 변수의 상관계수인 β1배만큼 증가한다. 여기서 몇몇 변수에는 상관계수에 추가로 0.01을 곱하여 계산을 하였는데 이는 데이터 전처리 과정에서 데이터의 극단적 값을 완화시키기 위해 log값을 취해줬기 때문이다.

변수	상관계수		변수	상관계수	
Price	$1.258(e^{0.22957})$	▲	RatingsCurrentVersion	$1.003(e^{0.32026*\ 0.01})$	▲
Seller	$2.07(e^{0.72795})$	▲	mondiff	$0.924(e^{-\ 0.07846})$	▼
Size	$0.996(e^{-\ 0.35879*\ 0.01})$	▼	updatemondiff	$0.567e^{-\ 0.56665})$	▼
StarsAllVersions	$0.3(e^{-\ 1.20081})$	▼	Cluster1(Battle & Create)	$1.475(e^{0.38883})$	▲
RaitingsAllVersions	$1.1005(e^{0.50590*\ 0.01})$	▲	Cluster3(Worldwide)	$1.698(e^{0.52939})$	▲

Table 5 변수 결과값 해석

4. 결론 및 한계·의의

4.1. 결론

텍스트 마이닝과 클러스터링을 통해 추출해낸 클러스터 변수들까지 설명 변수로 모두 포함시켜 앞선 모델들 중 최종 모델로 선정한 로지스틱 회귀 모델을 적합시켰다. 아래 표는 각 상위 5개 카테고리에 대해 유의하다고 나온 변수들을 모두 정리하여 합친 것이다. 변수가 모델 안에서 유의하다고 나왔다면 그 변수가 반응 변수에 어느 방향으로 영향을 미치는지를 총체적으로 한눈에 파악하기 위함이다. 표에서 +는 설명 변수가 반응 변수에 대해 양의 영향을, -는 음의 영향을 준다는 의미이다. 살펴보면 확연한 경향성을 보이는 변수들은 순위권 기업, 리뷰 수, 출시 개월 수, 업데이트 나이이다. 각각에 대해 조금 더 자세한 설명을 덧붙이자면 먼저 순위권 기업이 앱을 만들 때 순위권이 아닌 기업보다 높은 순위로 가는데

긍정적인 영향을 미친다. 마찬가지로 리뷰 수가 많을수록 긍정적 영향을 미치며, 출시 개월 수가 늘어날수록, 즉 앱이 출시된 지 시간이 오래 될수록 부정적인 영향을, 업데이트를 한 후 시간이 오래 될수록 부정적인 영향을 미침을 알 수 있다.

Category	가격	순위권 기업	용량	평점 (All)	리뷰수 (All)	무료	출시 개월 수	업데이트 나이	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Game	++	+	-	-	+		--	--	+		+	
Entertainment		+			+		-	-		-	-	
Social Networking					+	+	-	-			+	
Photo & Video		++	-		+	-	-	-				
Productivity	+	+			++		-				-	

Table 6 카테고리 별 설명 변수의 영향력

설명 변수 해석에 대해 좀 더 자세히 짚고 넘어가 보자면 직관적인 해석 및 활용이 가능한 변수는 8개 중 5개로 가격, 용량, 업데이트 빈도, 무료/유료, 출시 분기이다. 이러한 변수들은 '각각을 조정함으로 순위 상승을 기대할 수 있다'는 방식으로 직접적인 해석과 적용이 가능하다.

직관적인 해석 및 활용이 어려운 변수 3개는 별점, 리뷰 수, 앱의 나이 변수들이다. 이러한 변수들은 물론 직접적인 해석은 어렵지만 다른 의미로 우회하여 분석이 가능하다.

가장 먼저 별점 변수의 경우 반응 변수와의 논리적인 전개 순서가 역전되어 있다고 생각할 수 있다. 순위가 높고 이미 잘 만들어진 앱이라 별점이 높은 것일 수 있기 때문이다. 하지만 별점 변수는 변수 자체의 의미가 아닌 이에 영향을 미치는 요인들을 찾아 간접적으로 분석하는 것이 가능하다. 몇 가지 예를 들자면 유저들의 리뷰에 대한 빠른 피드백, 앱에 대한 만족도가 높은 시점 공략, 주기적인 버그 확인 및 제거 등의 방법을 통해 높은 별점을 유도하고 간접적으로 매출을 향상시킬 수 있다. 이는 앱을 개발하는 시점보다는 운영하는 시점에 활용될 수 있는 정보이다.

두 번째로 리뷰 수라는 변수 또한 별점과 마찬가지로 반응 변수와의 논리적인 전개 순서가 역전되어 있다고 생각할 수 있다. 순위가 높고 이용하는 유저가 많기 때문에 리뷰 수가 높은 것일 수 있기 때문이다. 하지만 역으로 리뷰가 많은 앱이 유저를 끌어들이기도 한다. 따라서 리뷰 작성의 비율을 극대화시키는 것이 높은 순위에 긍정적인 영향을 미칠 수 있다는 점을 생각하면 리뷰 작성에 대한 요인을 소비자들에게 제공하는 방식으로 순위를 높일 방법을 찾아낼 수 있다.

마지막으로 앱의 나이라는 변수는 앱이 출시된 후 시기에 따라 결정되는 변수로 조정이 불가능해 주제 분석의 취지와는 어긋나 보인다. 하지만 변수의 분석 결과 앱의 나이가 많아질수록 부정적인 영향을 끼치는 것을 확인할 수 있었는데, 이는 앱의 인기 수명과 연관성이 있다. 여러 앱을 보유한 제작사의 경우 최근 출시한 앱에 더 많은 투자를 하거나 앱의 라이프 사이클을 짧게 가져간다는 방식을 이용할 수 있는 것이다.

변수	전략
가격	가격을 조정함으로써 순위 상승 기대
용량	용량을 조정함으로써 순위 상승 기대
업데이트 빈도	업데이트 빈도를 조정함으로써 순위 상승 기대
무료/유료	뮤료/유료 여부를 조정함으로써 순위 상승 기대
출시 분기	출시 분기를 조정함으로써 순위 상승 기대

Table 7 직관적 해석이 가능한 변수

변수	전략
별점	1. 유저들의 리뷰에 대해 빠르게 피드백
	2. APP의 만족도가 높은 시점 공략
	3. 주기적인 버그 확인 및 제거
리뷰 수	1. 리뷰 작성에 대한 요인 제공(아이템 선물 등)
어플 나이	1. 최근 출시한 어플에 더욱 투자
	2. 어플의 라이프 사이클을 짧게 가져감

Table 8 직관적 해석이 불가능한 변수

4.2. 의의와 한계

본 연구의 의의는 크게 두 가지 측면으로 나누어 생각해 볼 수 있다. 첫번째로 비즈니스적인 측면에서 보면 앞으로 우리가 실제로 참여하게 될지도 모르는 비전 있는 분야에 대해서 연구할 수 있었다는 점이다. 어플리케이션 사업은 지속적으로 다양한 분야로 확장되고 있으며 비전문가도 참여, 성공할 수 있는 분야이다. 본 연구를 통해 어플리케이션 개발 및 운영에 직접적으로 도움이 되는 지표들을 찾아냈다는 점에 의의를 둘 수 있다.

두번째로는 다양한 통계분석 기법들을 하나의 큰 흐름에 따라 활용했다는 점이다. 가장 먼저 로지스틱 회귀, 다항 로짓 회귀와 같은 범주형 자료 분석 기법에서 시작하여 비정형 데이터를 다루는 텍스트 마이닝 그리고 텍스트에서 의미를 추출하기 위한 클러스터링을 활용하였다. 필요에 따라 다양한 분석 기법을 활용하여 데이터에서 의미를 추출해냈다는 점에 큰 의의를 둘 수 있다.

본 연구가 가지는 한계는 대체적으로 데이터 자체가 지닌 한계와 밀접하게 관련되어 있다. 한계는 두 가지로 설명할 수 있다. 먼저 연구에 사용한 데이터는 2018년이 아닌 2013년에 조사된 데이터이다. 빠르게 변화하는 앱 시장을 적절히 분석하기에는 5년이라는 작지 않은 간극이 있기 때문에 현실성을 부여하기에는 다소 부족함이 있다. 만약 크롤링을 통해 최신 데이터를 사용했더라면 좀 더 시의성 있는 분석을 해냈을 것이다.

두 번째로는 Rank 변수 자체를 반응 변수로 사용할 수 없었다는 점이다. 쉬운 해석을 위해 첫 번째로 선택한 회귀 분석 모델을 돌리는 과정에서 앱의 수입 요인 분석을 하고자 하는 목적 때문에 반응 변수를 -log를 취한 rank 값으로 대체하였는데, 이로 인해 모델은 기본 가정을 만족하는 것이 어려웠을 뿐만 아니라 정보의 손실이 발생할 수밖에 없었다. 최종 모델이었던 로지스틱 회귀 분석에서도 마찬가지로 모델을 돌리기 위해 임의로 범위를 나누어 반응 변수를 범주형 변수 형태로 바꿀 수밖에 없었다. 이는 300개의 순위를 단순히 두 가지 범주로 줄여 표현할 수밖에 없었기 때문에 정보 손실이 매우 컸다. 만약 순위 자체를 y변수로 갖는 모델을 사용한다면 기존의 분석과 다르게 정보의 손실 없이 더 정확한 분석이 가능할 것이다.