BUAN 6337 PREDICTIVE ANALYTICS USING SAS

GROUP PROJECT

MAY 6, 2019

GROUP 6
Qingyuan Zhu, Minke Li, Spencer Lu, Xinyi Zhang, Jinglin Zhao

Contents

Executive Summary

Our paper is structured as follows. We first introduce three research questions we bring up in this project. Subsequently, an overview of our dataset used in the study is briefly described in tables by utilizing descriptive analysis method. We will further discuss how to clean and process out dataset so that to avoid data leakage, skewness and outlier problems.

We have three dependent variables, which are ratings, installs, and price. Linear regression model is used in this report and various variables and interaction terms are also implemented within our models. Then we try to compare different value in Mallows'Cp (forward, backward and stepwise algorithm), cross validation (forward, backward and stepwise algorithm), Lasso and Elastic Net, to see which model could provide best performances.

The results and analyses of three models are reviewed in addition to the improvement of the research for further study and conclusion will also be provided.

Research Overview

2.1. Research Questions

We have three research questions which are how to get a high rating on Google play store, how to boost installation amount and does price really matters.

We would break the first question down into three small questions, so that it would be much easier to us to find out the answer. These three small questions are finding the most relevant features, selecting a best model to predict the application rating and trying to look for effective strategies to improve the ratings.

For the second research question, how to boost installs, we explore the relationship between installation amount and the types (free or paid), and we also include the possible relative components, like size, rating, and categories into consideration.

For the third research question, does price matters, we also find two small question. These questions are: the difference of application average price in various categories, whether the size and content rating of applications are the important factors of pricing strategy, and how often should paid applications update.

2.2. Dataset Description

2.2.1. Variables

In dataset google_play_store, there are 13 features contained to describe the google store marketing with over 10,000 observations. These features include feedback from users and nature of applications, which all depend on the scraped date. Variables will be introduced as following:

Categorical variables:

- App: the name of the application
- Category: the category of the application, in the original dataset, there were 33 categories. We re-arranged them into 8 new categories (which will be described later).
- Type: describe the application is paid or free, under this feature, there are two values: Paid, Free
- Content Rating: the application is classified by age, there are 6 values: Adults only 18+, Everyone, Everyone 18+, Mature 17+, Teen and Unrated
- Update: extract first digit of current version, which indicate how many times developer has majorly updated the application, 10 represents other situation
- Last Update: the newest date to update the application
- Android Ver: minimum requirement of Android system to download the application, we use the first digit to represent the system version includes version 0 to version 8

Numerical variables:

- Rating: overall rating from users
- Reviews: number of user reviews for application
- Size: the size of application, calculated by MB
- Installs: approximately install number of application (without plus sign)
- Price: the price of paid application
- Recency: the difference between the data collected date (sep/4/2018) and newest update date

2.2.2. Data Cleaning and Preprocessing

2.2.2.1. Variables Re-arrangements

Categorical variables:

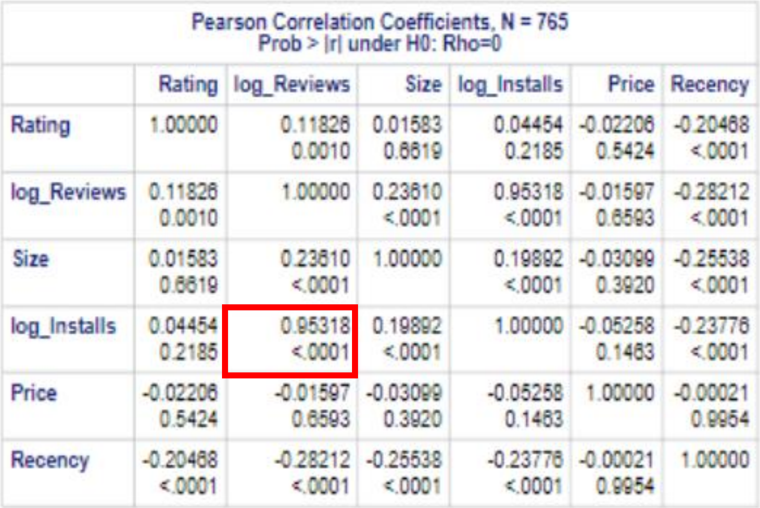
We have rearranged the Category variable in this project. Since there are 33 categories, which is not easy to interpret in model analysis, we decided to re-classify them into 8 general categories. They are Tools, Social and Entertainment, Business, Transportation, Media_Video, Game, Family and Lifestyle. Appendix 1 illustrate shares of different categories, which is much easier for us to conduct the follow analysis. And we rearranged original categories into those new ones based on daily experiences.

Numerical variables:

The distributions of variable Reviews and Installs demonstrate skewness problem. With such wide range of observations, it is hard to make correct prediction on dependent variable. Therefore, we need to use natural logarithm on these two variables in order to reduce skewness problem.

2.2.2.2. Data Leakage and Outliers

Table 1. Pearson Correlation



Pearson Correlation Coefficients, N = 765
Prob > |r| under H0: Rho=0

	Rating	log_Reviews	Size	log_Installs	Price	Recency
Rating	1.00000	0.11826 0.0010	0.01583 0.6619	0.04454 0.2185	-0.02206 0.5424	-0.20468 <.0001
log_Reviews	0.11826 0.0010	1.00000	0.23610 <.0001	0.95318 <.0001	-0.01597 0.6593	-0.28212 <.0001
Size	0.01583 0.6619	0.23610 <.0001	1.00000	0.19892 <.0001	-0.03099 0.3920	-0.25538 <.0001
log_Installs	0.04454 0.2185	0.95318 <.0001	0.19892 <.0001	1.00000	-0.05258 0.1463	-0.23776 <.0001
Price	-0.02206 0.5424	-0.01597 0.6593	-0.03099 0.3920	-0.05258 0.1463	1.00000	-0.00021 0.9954
Recency	-0.20468 <.0001	-0.28212 <.0001	-0.25538 <.0001	-0.23776 <.0001	-0.00021 0.9954	1.00000

In one of our model analyses, log_Installs is designed as dependent variable.

According to the correlation table, the correlation coefficient between log_Installs and log_Reviews is close to 1, which means they are highly correlated. If we include log_Reviews as one of the independent variables in the model to predict log_Installs, other independent variables would have smaller contribution to explanation. In this case, we targeted to research how these variables, not only log_Reviews, explain the dependent variable. That is the main reason why we need to remove log_Reviews when we predict log_Installs.

Boxplots of each category below demonstrated that there are outliers existed in Price feature. Results might be inaccuracy to fitting with these outliers. We apply Cook's D method to detect those outliers and remove them to get a new dataset. From comparison between these two boxplots, price information will be much clearer to explain the minimum and maximum value

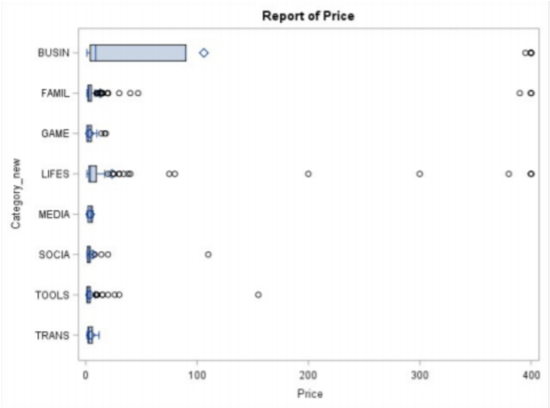


Figure 1. Box plot of Price with outliers

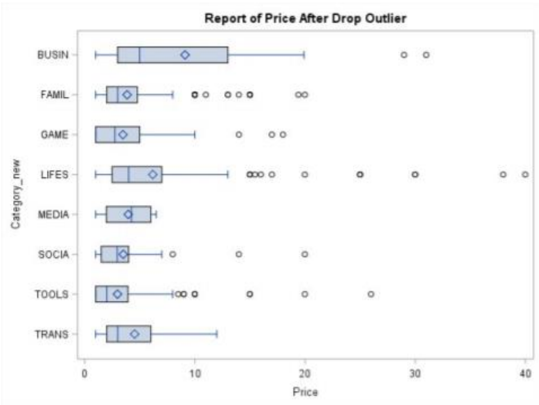


Figure 2. Box plot of Price without outliers

2.3. Data Exploration

We summarize statistic information of features including rating, size, review, installs, log(reviews) and log(installs) by each category.

Each category has similar rating pattern with the average value of 4.

However, size of each category is quite different. Game category is larger than other categories and has the largest size, while tools category is relatively smaller than others and has the smallest size.

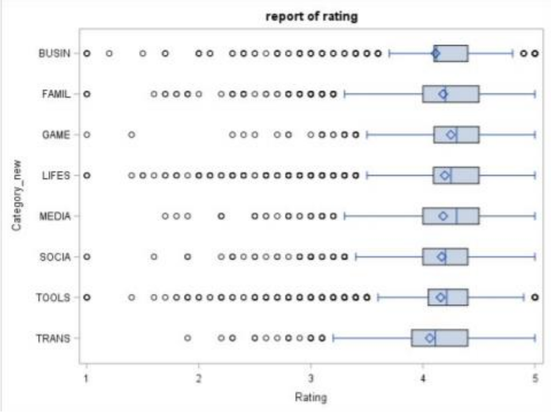


Figure 3. Box plot of Rating

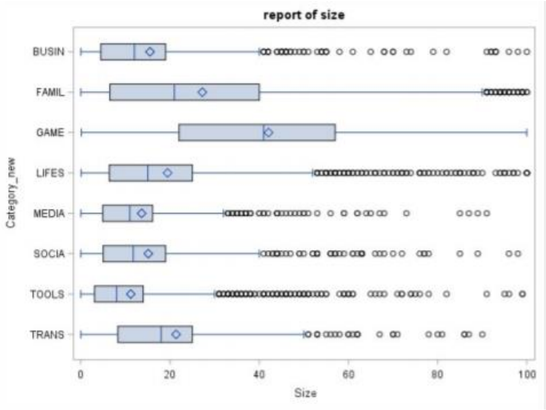
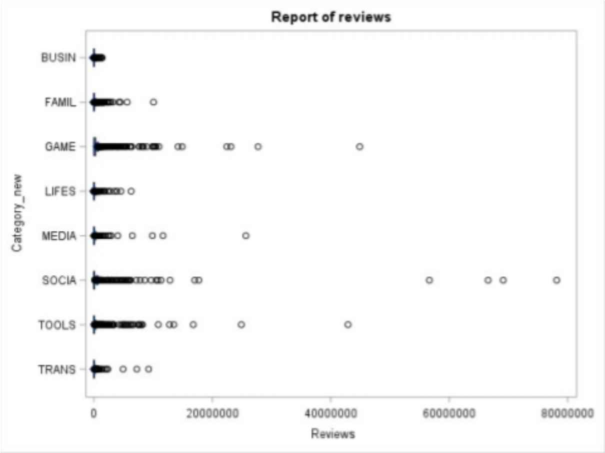
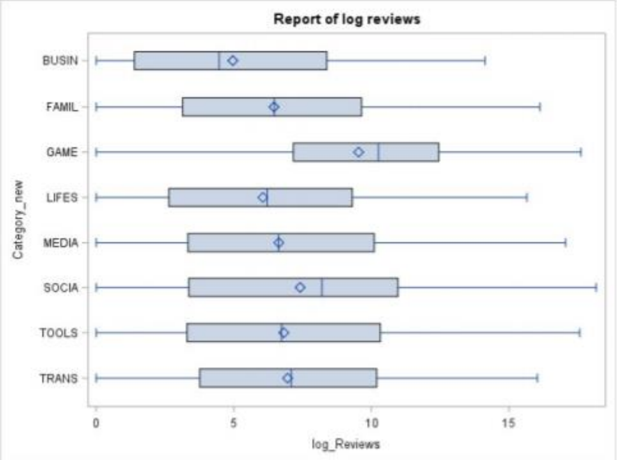
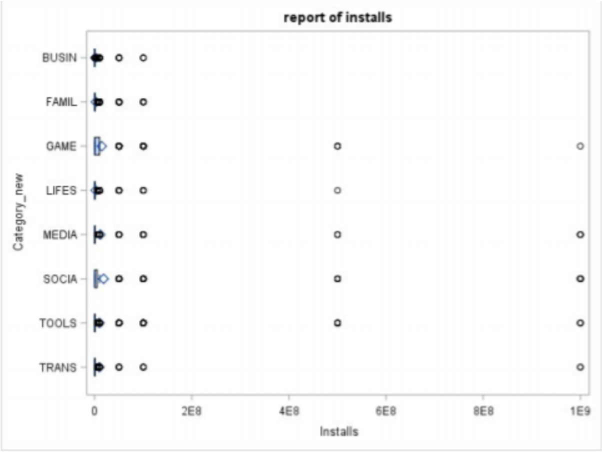


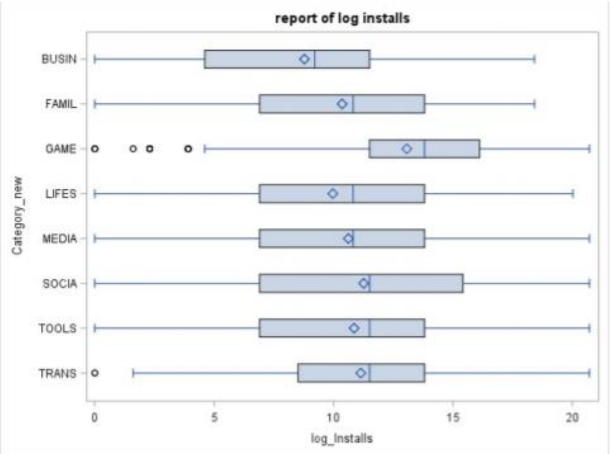
Figure 4. Box plot of Size

Both reviews and installs have large scales, if we directly use these two factors into our models, the results of the model may be seriously impacted. When we take log of reviews and installs, the data are all scaled, and the distribution plot is clear and readable.









Model Analysis

3.1. Analysis Methods

To answer our research questions, we select Rating, log (Installs) and Price as our dependent variables and apply linear regression models separately.

1) To analysis how the rating of apps is affected by other factors, we first set Rating as the dependent variable, and apply linear regression models. We also create few interaction variables, such as Category*Size and Category*log(installs), to evaluate if the sizes and installs of apps has different impact on rating across different categories.

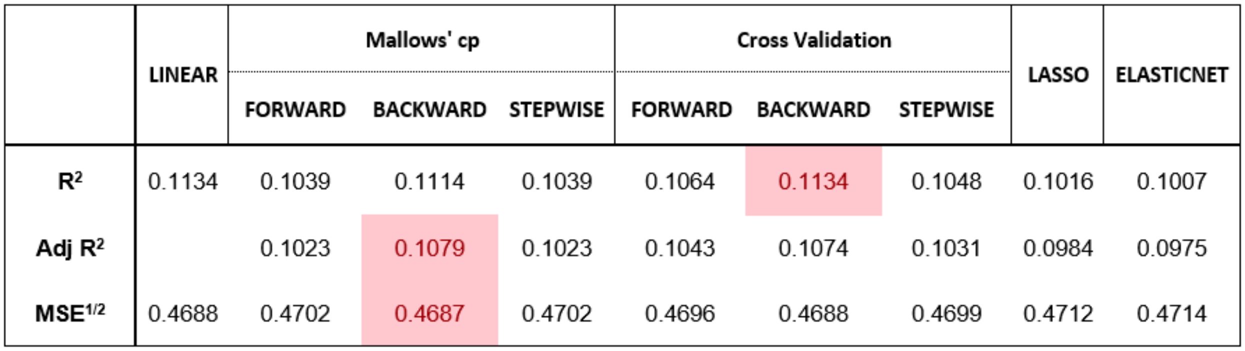
2) To analysis how the Installs of apps are affected by other factors, we then set log (Installs) as the dependent variable. As we discussed before, reviews of apps are highly correlated with installs of apps (corr >0.95), which would result in data leakage. So, we removed log(reviews) from the independent variables.

3) To analysis how should we price the applications, we explore the influential factors of existing applications. We first set Price as the dependent variable, and apply Category, Size, log(reviews), Content_Rating, and Recency as the independent variables. Moreover, the interaction terms, Size*Category and log(reviews)*Category, are also included in the model to deeply analysis the price difference in different categories.

In model selection, we apply Mallows' Cp (forward, backward and stepwise algorithm), cross validation (forward, backward and stepwise algorithm), Lasso and Elastic Net to see which model could provide best performances. The model performances for the 3 dependent variables are presented below:

Model selection - 1) Rating as dependent variable:

Table 2. Model selection-Rating as dependent variable

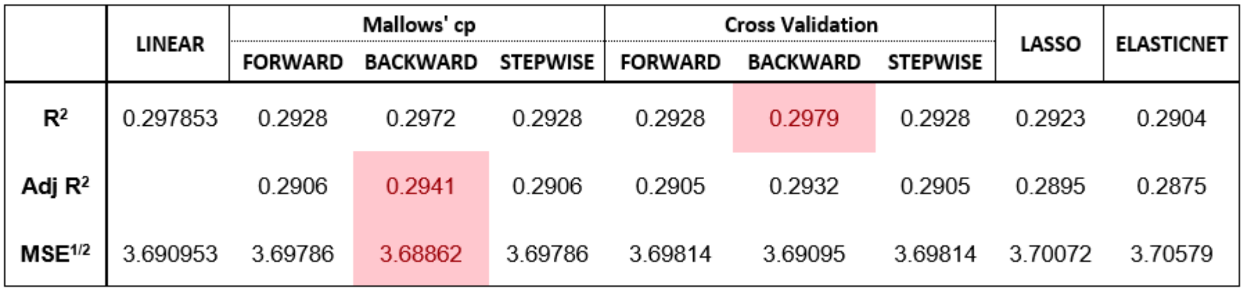


	LINEAR	Mallows' cp			Cross Validation			LASSO	ELASTICNET
		FORWARD	BACKWARD	STEPWISE	FORWARD	BACKWARD	STEPWISE		
R^2	0.1134	0.1039	0.1114	0.1039	0.1064	0.1134	0.1048	0.1016	0.1007
Adj R^2		0.1023	0.1079	0.1023	0.1043	0.1074	0.1031	0.0984	0.0975
$MSE^{1/2}$	0.4688	0.4702	0.4687	0.4702	0.4696	0.4688	0.4699	0.4712	0.4714

Based on adjusted R-square criteria, we select Mallows' Cp -Backward selection model as the best model and will use this model to interpret for App Rating.

Model selection - 2) Log(installs) as dependent variable:

Table 3. Model selection-Log(install) as dependent variable

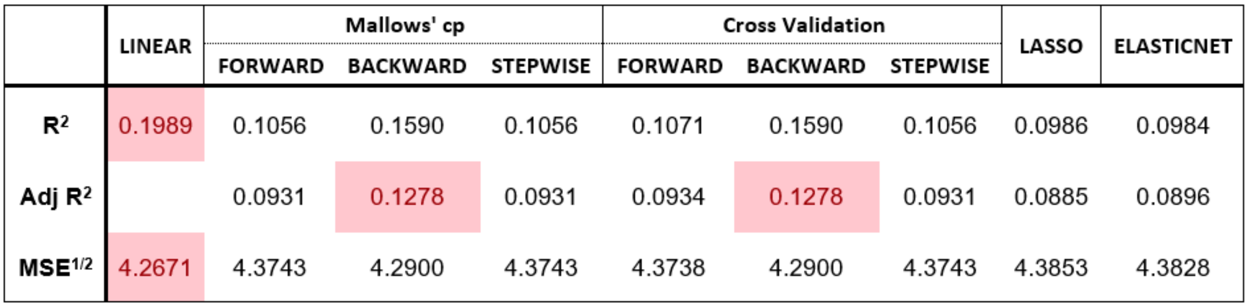


	LINEAR	Mallows' cp			Cross Validation			LASSO	ELASTICNET
		FORWARD	BACKWARD	STEPWISE	FORWARD	BACKWARD	STEPWISE		
R^2	0.297853	0.2928	0.2972	0.2928	0.2928	0.2979	0.2928	0.2923	0.2904
Adj R^2		0.2906	0.2941	0.2906	0.2905	0.2932	0.2905	0.2895	0.2875
$MSE^{1/2}$	3.690953	3.69786	3.68862	3.69786	3.69814	3.69095	3.69814	3.70072	3.70579

Based on adjusted R-square criteria, we select Mallows' Cp -Backward selection model as the best model and will use this model to interpret for Log(installs).

Model selection - 3) Price as dependent variable:

Table 4. Model selection-Price as dependent variable



	LINEAR	Mallows' cp			Cross Validation			LASSO	ELASTICNET
		FORWARD	BACKWARD	STEPWISE	FORWARD	BACKWARD	STEPWISE		
R^2	0.1989	0.1056	0.1590	0.1056	0.1071	0.1590	0.1056	0.0986	0.0984
Adj R^2		0.0931	0.1278	0.0931	0.0934	0.1278	0.0931	0.0885	0.0896
$MSE^{1/2}$	4.2671	4.3743	4.2900	4.3743	4.3738	4.2900	4.3743	4.3853	4.3828

Based on adjusted R-square criteria, we select Mallows' Cp -Backward selection model as the best model and will use this model to interpret Price.

3.2. Models

3.2.1. Ratings

According to the backward selection based on Mallow’s Cp selection criteria, the following effects are selected.

Table 5. Parameters in Rating model

Effects :	Intercept,Category_new_BUSIN,Category_new_FAMIL,Category_new_GAME, Category_new_LIFES,Category_new_MEDIA,Category_new_SOCIA, Category_new_TOOLS,Size,Size*Category_new_FAMIL, Size*Category_new_GAME,Size*Category_new_LIFES Size*Category_new_SOCIA,Size*Category_new_TOOLS,log_Installs log_Installs*Category_new_FAMIL,log_Installs*Category_new_GAME log_Installs*Category_new_MEDIA,log_Installs*Category_new_SOCIA Type_Free,Category_new_BUSIN*Type_Free,Category_new_LIFES*Type_Free, Category_new_SOCIA*Type_Free,Category_new_TOOLS*Type_Free, log_Reviews,Size*Type_Free,Content_Rating_Everyone,Content_Rating_Mature17+, Content_Rating_Everyone*Type_Free,Recency,Recency*Type_Free Android_Ver_0,update_0,update_1,update_2,update_8,Type_Free*update_1, Type_Free*update_2

Table 6. Estimated coefficient in Rating model

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.46383	0.086775	51.44	<.0001***
Category_new_BUSIN	1	-0.195453	0.111505	-1.75	0.0797*
Category_new_FAMIL	1	0.264477	0.046251	5.72	<.0001***
Category_new_GAME	1	0.109792	0.063916	1.72	0.0859*
Category_new_LIFES	1	0.182713	0.059946	3.05	0.0023***
Category_new_MEDIA	1	0.195268	0.057794	3.38	0.0007***
Category_new_SOCIA	1	0.002935	0.091088	0.03	0.9743
Category_new_TOOLS	1	0.123229	0.056417	2.18	0.029**
Size	1	-0.00093	0.001184	-0.79	0.4322
Size*Category_new_FAMIL	1	0.002742	0.000885	3.1	0.0019***
Size*Category_new_GAME	1	0.003149	0.000966	3.26	0.0011***
Size*Category_new_LIFES	1	0.001585	0.000956	1.66	0.0974*
Size*Category_new_SOCIA	1	0.001854	0.001292	1.43	0.1515
Size*Category_new_TOOLS	1	0.003139	0.001119	2.81	0.005***
log_Installs	1	-0.097779	0.004173	-23.43	<.0001***
log_Installs*Category_new_FAMIL	1	-0.020863	0.003467	-6.02	<.0001***
log_Installs*Category_new_GAME	1	-0.008706	0.004441	-1.96	0.05**
log_Installs*Category_new_MEDIA	1	-0.009524	0.004524	-2.11	0.0353**
log_Installs*Category_new_SOCIA	1	-0.011372	0.003563	-3.19	0.0014***

Type_Free	1	-0.124925	0.079541	-1.57	0.1163
Category_new_BUSIN*Type_Free	1	0.249425	0.109993	2.27	0.0234**
Category_new_LIFES*Type_Free	1	-0.093947	0.052379	-1.79	0.0729*
Category_new_SOCIA*Type_Free	1	0.164211	0.084966	1.93	0.0533*
Category_new_TOOLS*Type_Free	1	-0.099593	0.049028	-2.03	0.0422**
log_Reviews	1	0.120909	0.004233	28.56	<.0001***
Size*Type_Free	1	-0.001967	0.00097	-2.03	0.0427**
Content_Rating_Everyone	1	-0.137578	0.063267	-2.17	0.0297**
Content_Rating_Mature 17+	1	-0.096632	0.028313	-3.41	0.0006***
Content_Rating_Everyone*Type_Free	1	0.146911	0.064378	2.28	0.0225**
Recency	1	-6.99E-05	0.00003267	-2.14	0.0324**
Recency*Type_Free	1	-6.41E-05	3.5487E-05	-1.81	0.0711*
Android_Ver_0	1	0.028725	0.019487	1.47	0.1405
update_0	1	0.099509	0.037964	2.62	0.0088***
update_1	1	-0.070168	0.040539	-1.73	0.0835*
update_2	1	-0.105728	0.057109	-1.85	0.0641*
update_8	1	0.091054	0.043725	2.08	0.0373**
Type_Free*update_1	1	0.129841	0.041947	3.1	0.002***
Type_Free*update_2	1	0.142386	0.058981	2.41	0.0158**

(***: significant at 99% level; **: significant at 95% level; *: significant at 90% level)

Category:

Category_new_BUSIN	1	-0.195453	0.111505	-1.75	0.0797*
Category_new_FAMIL	1	0.264477	0.046251	5.72	<.0001***
Category_new_GAME	1	0.109792	0.063916	1.72	0.0859*
Category_new_LIFES	1	0.182713	0.059946	3.05	0.0023***
Category_new_MEDIA	1	0.195268	0.057794	3.38	0.0007***
Category_new_SOCIA	1	0.002935	0.091088	0.03	0.9743
Category_new_TOOLS	1	0.123229	0.056417	2.18	0.029**

Under 95% confident level, the Ratings are significantly different in several categories: Tools, Media_Video, Family and Lifestyle. Social_and_Entertainment, Business, and Game is not statistically significant at this level. According to the coefficients of each categories, applications in Family category are most likely to have higher ratings, followed by Media_Video, Lifestyles and Tools. However, Transportation category is not included in the model after model selection, because there is much less difference of ratings among the applications in this category.

Therefore, if the company want to develop an application in the categories of Lifestyle, Family, Tools, and Media_Video, it would be easier to get higher ratings.

Size:

Size	1	-0.00093	0.001184	-0.79	0.4322

Due to its negative coefficient, when other factors are equal, every 1MB increasing on size will decrease the rating value by 0.00093. But under 95% confident level, Size is not statistically significant, since its p-value is 0.4322, can’t reject the null hypothesis.

Size*Category:

Size*Category_new_FAMIL	1	0.002742	0.000885	3.1	0.0019***
Size*Category_new_GAME	1	0.003149	0.000966	3.26	0.0011***
Size*Category_new_LIFES	1	0.001585	0.000956	1.66	0.0974*
Size*Category_new_SOCIA	1	0.001854	0.001292	1.43	0.1515
Size*Category_new_TOOLS	1	0.003139	0.001119	2.81	0.005***

This interaction term helps us to find out the Size impact on different categories. Unfortunately, the result shows that under 95% confident level, Size impact significantly eliminated or even reversed on ratings of Family, Game, and Tools applications, because of its positive coefficient. At the same time, although the interaction terms on Lifestyle and Social_and_Entertainment are not statistically significant at this level, the coefficients also show that Size impact is reduced.

Briefly, when the company develop the applications in the categories of Family, Game, and Tools, it would better create a larger size application to gain the higher ratings; however, if the company invests on other categories, smaller size applications are preferred by users.

Log(installs):

log_Installs	1	-0.097779	0.004173	-23.43	<.0001***

Under 95% confident level, p-value of log(installs) indicates that Installs is significant difference on the Ratings. When everything is equal, log(installs) increases one unit with a 9.79% decreasing on Ratings.

This finding is different to our expectation. In common sense, an application getting higher download amounts indicates its better popularity and user preference; however, Ratings are more likely to approach a worse score with the increasement of downloads. This is probably because of the strategy the applications use which only focuses on boosting the downloads in short term by some promotion events but sacrifices the quality of the application. For figuring out the real reasons, more data should be provided.

Log(installs)*Category:

log_Installs*Category_new_FAMIL	1	-0.020863	0.003467	-6.02	<.0001***
log_Installs*Category_new_GAME	1	-0.008706	0.004441	-1.96	0.05**

log_Installs*Category_new_MEDIA	1	-0.009524	0.004524	-2.11	0.0353**
log_Installs*Category_new_SOCIA	1	-0.011372	0.003563	-3.19	0.0014***

Same as before, we also create the interaction term of log(installs) and Category to study the installation's impact on different categories. From the table above we can see, the applications in Family, Game, Media_Video, and Social_and_Entertainment are significant here, and they all strengthen the impact of log(installs), which also means in these categories, other rating improvement strategies should be much more considered, since increasing installation is not a good idea to increase Ratings at all.

As a result, it's a good hint for product managers and marketing managers, only focusing on enlarging the downloads is not a wise decision, and it's also not good for the development of applications with a lower and lower rating scores when installation increases. In the other words, if the company wants to keep a higher Ratings, it should pay more attention on other crucial components of the Ratings and applying the strategies accordingly.

Type_Free:

Type_Free	1	-0.124925	0.079541	-1.57	0.1163

Basically, compared with average rating of paid applications, the average rating of free applications will have 0.12 point less than paid ones, but under 95% confident level, Type_Free is not statistically significant.

Category*Type_Free:

Category_new_BUSIN*Type_Free	1	0.249425	0.109993	2.27	0.0234**
Category_new_LIFES*Type_Free	1	-0.093947	0.052379	-1.79	0.0729*
Category_new_SOCIA*Type_Free	1	0.164211	0.084966	1.93	0.0533*
Category_new_TOOLS*Type_Free	1	-0.099593	0.049028	-2.03	0.0422**

Compared with average rating within free apps, apps in category Business and Social_and_Entertainment will increase the rating while rating in category Lifestyles and Tools will decrease. But at 95% confident level, interactions terms of Type_Free*Lifestyles and Type_Free* Social_and_Entertainment are not statistically significant. This may indicate that people have stricter requirements for Tools and easy to be satisfied by Business applications.

Size*Type_Free:

Size*Type_Free	1	-0.001967	0.00097	-2.03	0.0427**

Under 95% confident level, p-value of Size*Type_Free indicates it is significant difference on the Ratings. When everything is equal, the size of a free application increases 1MB will cause

0.00283 decrease on Ratings, which is make sense, people don’t like their apps occupy too much limited storage on their phone, especially free applications.

Log_Reviews:

log_Reviews	1	0.120909	0.004233	28.56	<.0001***

Under 95% confident level, p-value of log(reviews) indicates that Ratings is significant difference on the log(reviews). When everything is equal, log(reviews) increases one unit with a 12.09% increasing on Ratings. Like the installs, an application getting higher reviews amounts also indicates its better popularity and user preference, hence Rating is more likely to approach a better score.

Content_Rating:

Content_Rating_Everyone	1	-0.137578	0.063267	-2.17	0.0297**
Content_Rating_Mature 17+	1	-0.096632	0.028313	-3.41	0.0006***

Under 95% confident level, Content_Rating Mature 17+ and Content_rating_Everyone has significant difference in Ratings, so that if the content rating of an application is mature 17+, the rating of this application is 0.10 less than other content rating applications. If the content rating of an application is for everyone, the rating of this application is 0.13 less than other content rating applications.

This finding tells us that users possibly have higher evaluating standards for mature 17+ and everyone applications, or in the current industry, the average quality of these two kinds applications are not satisfied the demand of the users enough.

Content_Rating_Everyone*Type_Free:

Content_Rating_Everyone*Type_Free	1	0.146911	0.064378	2.28	0.0225**

Under 95% confident level, free apps for everyone has significant difference in Ratings, so that the ratings of free applications for everyone will be 0.15 more than average rating for paid apps. Additionally, Type_Ftee reverses the impact of Content_Rating_Everyone, that is to say, free applications for all ages are easier to gain recognitions than the paid ones.

Recency:

Recency	1	-6.99E-05	0.00003267	-2.14	0.0324**

Under 95% confident level, Recency is statistically significant. Other things being equal, Recency is prolonged by 1000 days, the Rating is decreased by 0.07. Based on the Recency, we can say that the longer using the current version of applications, the better ratings can be expected.

Therefore, if the company wants to have a higher rating application, it should make application full-featured, more system stable and less bugs, which can extent the using time of one version of an application.

Recency*Type_Free:

Recency*Type_Free	1	-6.41E-05	3.5487E-05	-1.81	0.0711*

Although this interaction term is not statistically significant at 95% confident level, it tells that the ratings of free applications are greater negatively affected by Recency.

Android_Ver:

Android_Ver_0	1	0.028725	0.019487	1.47	0.1405

Others being equal, the lower requirements of the Android_Ver, the better ratings can be achieved, but this variable is also not statistically significant under the 95% confident level.

Update:

update_0	1	0.099509	0.037964	2.62	0.0088***
update_1	1	-0.070168	0.040539	-1.73	0.0835*
update_2	1	-0.105728	0.057109	-1.85	0.0641*
update_8	1	0.091054	0.043725	2.08	0.0373**

For deeply analysis of version updating, we include the categorical variable update, which points out the major updating times of an application. Under 95% confident level, update_0 and update_8 which refer to no major updating and 8 times major updating, are statistically significant, while update_1 and update_2, which refer to once or twice major updating, are not significant, and other update situations are not importance to Ratings. According to the estimates of update variables, we find that the applications are fully updated 0 or 8 times can improve the Ratings by 0.10 and 0.09, on the contrary, updating 1 or 2 times will decrease the Ratings by 0.07 and 0.11, when others keep equal.

This result makes the complement to what we find above. If an application is developed with powerful and full-featured functions, no serious bugs, and stable system at the beginning, which means it doesn’t need major update at all, this kind applications will have higher rating scores. On the other hand, if an application gets improved frequently and significantly, especially having 8 times major updating, users may give higher ratings as well. However, if an application has only one or two times updating, the Ratings will be lower than other situations. Because an application having major updating but with lower updating speed may refer to a lower developing speed, and simple functions which may make users feel board and not useful enough.

Type*Update:

Type_Free*update_1	1	0.129841	0.041947	3.1	0.002***
Type_Free*update_2	1	0.142386	0.058981	2.41	0.0158**

Under 95% confident level, Type_Free*update_1 and Type_Free*update_2 make Rating significantly different with others, but from the above analysis we can see the main effects neither Type_Free nor update (1 and 2) is significant, which means there is no overall effects of these two factors, but there is a crossover interaction. The effect of update (1 and 2) on the ratings is oppsite, depending on the value of Type_Free.

The coefficients of the all kinds of applications are as below:

Table 7. Coefficients of four kinds of applications

	free	Paid
update_1	0.059673	-0.070168
Not update_1	-0.124925	0
update_2	0.036658	-0.105728
Not update_2	-0.124925	0

From the table we can see that users' ratings are less sensitive to the free applications updating than the paid one. The reason to get this result, we guess the users may have higher expectation to the paid applications, so that they want the application is full-featured at the beginning, or updates more frequently, instead of getting improved step by step and slowly. From another side, users have higher tolerance of slower developing speed of free applications. No matter these free applications update once or twice, they both have higher rating than the average of other applications.

Therefore, if the company wants to launch a paid application, it would be better not only updating the application one or two times. Either the company should initially create a powerful application, or frequently update the application to drive users' interests, then the application can keep a respectively high rating.

3.2.2. Log Installs

According to Backward Selection based on Mallows Cp selection Criteria, the following effects are selected.

Table 8. Parameters in Log(Installs) model

Effects:	Intercept,Category_new_BUSIN,Category_new_FAMIL,Category_new_GAME Category_new,LIFES,Category_new_MEDIA,Category_new_SOCIA Category_new,TOOLS,Size,Size*Category_new_FAMIL Size*Category_new_SOCIA,Size*Category_new_TOOLS,Rating Rating*Category_new_BUSIN,Rating*Category_new_FAMIL Rating*Category_new_LIFES.Rating*Category_new_SOCIA,Type_Free Category_new_BUSIN*Type_Free,Category_new_FAMIL*Type_Free Category_new_LIFES*Type_Free,Category_new_MEDIA*Type_Free Size*Type_Free,Content_Rating_Everyone,Content_Rating_Everyone*Type_Free Recency,Recency*Type_Free,update_0,update_1,update_2,update_3,update_4 update_5,update_6,update_7,update_8,Type_Free*update_0,Type_Free*update_1 Type_Free*update_2,Type_Free*update_3,Type_Free*update_4 Type_Free*update_7, Type_Free*update_8

Table 9. Estimated Coefficient of Log(Install) model

Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	4.460641	0.848441	5.26	<.0001***
Category_new_BUSIN	1	2.718109	1.369662	1.98	0.0472**
Category_new_FAMIL	1	6.235744	0.998867	6.24	<.0001***
Category_new_GAME	1	2.1393	0.245176	8.73	<.0001***
Category_new_LIFES	1	2.951808	0.94901	3.11	0.0019***
Category_new_MEDIA	1	-2.451037	0.71083	-3.45	0.0006***
Category_new_SOCIA	1	2.880331	1.180263	2.44	0.0147**
Category_new_TOOLS	1	1.134122	0.2355	4.82	<.0001***
Size	1	0.019682	0.007313	2.69	0.0071***
Size*Category_new_FAMIL	1	0.026384	0.004532	5.82	<.0001***
Size*Category_new_SOCIA	1	0.027562	0.008528	3.23	0.0012***
Size*Category_new_TOOLS	1	-0.0319	0.00722	-4.42	<.0001***
Rating	1	0.875281	0.12338	7.09	<.0001***
Rating*Category_new_BUSIN	1	-0.705038	0.266671	-2.64	0.0082***

Rating*Category_new_FAMIL	1	-1.292638	0.215442	-6	<.0001***
Rating*Category_new_LIFES	1	-0.513164	0.200498	-2.56	0.0105**
Rating*Category_new_SOCIA	1	-0.718033	0.277719	-2.59	0.0097***
Type_Free	1	5.450601	0.646926	8.43	<.0001***
Category_new_BUSIN*Type_Free	1	-1.681634	0.84839	-1.98	0.0475**
Category_new_FAMIL*Type_Free	1	-1.522149	0.36954	-4.12	<.0001***
Category_new_LIFES*Type_Free	1	-1.628261	0.390679	-4.17	<.0001***
Category_new_MEDIA*Type_Free	1	2.408536	0.699957	3.44	0.0006***
Size*Type_Free	1	0.018761	0.007245	2.59	0.0096***
Content_Rating_Everyone	1	-0.821072	0.485727	-1.69	0.091*
Content_Rating_Everyone*Type_Free	1	0.717311	0.497117	1.44	0.1491
Recency	1	2.018E-05	0.000257	0.08	0.9375
Recency*Type_Free	1	-0.000458	0.000279	-1.64	0.1015
update_0	1	-2.578182	1.171163	-2.2	0.0277**
update_1	1	-2.148212	0.410142	-5.24	<.0001***
update_2	1	-1.007819	0.518503	-1.94	0.052*
update_3	1	-0.87284	0.635462	-1.37	0.1696
update_4	1	0.289754	0.784037	0.37	0.7117
update_5	1	-2.064508	0.221256	-9.33	<.0001***
update_6	1	-1.612613	0.26228	-6.15	<.0001***
update_7	1	0.493418	1.290028	0.38	0.7021
update_8	1	0.539661	1.443717	0.37	0.7086
Type_Free*update_0	1	-2.864084	1.211599	-2.36	0.0181**
Type_Free*update_1	1	-2.498493	0.422727	-5.91	<.0001***
Type_Free*update_2	1	-2.022606	0.534334	-3.79	0.0002***
Type_Free*update_3	1	-2.070923	0.65341	-3.17	0.0015***
Type_Free*update_4	1	-2.325625	0.804323	-2.89	0.0038***
Type_Free*update_7	1	-2.986167	1.32913	-2.25	0.0247**
Type_Free*update_8	1	-3.324689	1.486383	-2.24	0.0253**

(***: significant at 99% level; **: significant at 95% level; *: significant at 90% level)

In this model, all the important variables to the log(installs) are selected as above.

Category:

Category_new_BUSIN	1	2.718109	1.369662	1.98	0.0472**
Category_new_FAMIL	1	6.235744	0.998867	6.24	<.0001***
Category_new_GAME	1	2.1393	0.245176	8.73	<.0001***
Category_new_LIFES	1	2.951808	0.94901	3.11	0.0019***
Category_new_MEDIA	1	-2.451037	0.71083	-3.45	0.0006***
Category_new_SOCIA	1	2.880331	1.180263	2.44	0.0147**
Category_new_TOOLS	1	1.134122	0.2355	4.82	<.0001***

At 95% confidence level, the installs of applications in categories Business, Family, Game, Social Media and Tools are significantly different from average level. According to the estimated coefficients, Family category has the most impact on installs, while Lifestyle, Social media, and Business categories both have comparatively high impacts, when other things being equal. However, Media_video category has negative impact. For Transportation category, which are not included after applying model selection, the category seems not significantly affects their installs.

If companies are considering developing apps and want to obtain high installs, they probably should consider Family, Lifestyle, Social Media and Business categories.

Size:

Size	1	0.019682	0.007313	2.69	0.0071***

At 95% confidence level, the coefficient of size is statistically significant. When the size of application increases by 1 MB, while others being equal, the installs of application is estimated to increase by 1.97%.

From the results of estimation, we can conclude that with higher installs application, their size might be bigger than lower installs application.

Size*Category:

Size*Category_new_FAMIL	1	0.026384	0.004532	5.82	<.0001***
Size*Category_new_SOCIA	1	0.027562	0.008528	3.23	0.0012***
Size*Category_new_TOOLS	1	-0.0319	0.00722	-4.42	<.0001***

This interaction shows how size of apps are affecting the installs differently across different categories. At 95% confidence level, when other things being equal, and when the size of the apps are the same, apps under Family and Social Media categories is likely to have 2.6% - 2.75% higher installs, while apps under Tools category is likely to have about –3.2% lower installs, comparing to average installs.

In other words, when choosing apps under Family and Social Media categories, people tend to care less about the size of the app, while choosing apps under Tool category, people are more likely to choose smaller sizes.

Rating:

Rating	1	0.875281	0.12338	7.09	<.0001***

At 95% confidence level, the coefficient of rating is statistically significant. If the rating of application increased by 0.1, while others being equal, the installs of application is estimated to increase by 8.75%.

It is obvious to conclude that, applications with higher rating is more likely to obtain higher installs.

Rating*category:

Rating*Category_new_BUSIN	1	-0.705038	0.266671	-2.64	0.0082***
Rating*Category_new_FAMIL	1	-1.292638	0.215442	-6	<.0001***
Rating*Category_new_LIFES	1	-0.513164	0.200498	-2.56	0.0105**
Rating*Category_new_SOCIA	1	-0.718033	0.277719	-2.59	0.0097***

This interaction shows when other factors been equal, how the categorical difference of rating of the apps will affect installs. We find that at 95% confidence level, when ratings are the same, apps under Family seem to have the lowest installs, while Lifestyle, Business and Social media categories also tend to generate lower installs, comparing to average installs.

Type:

Type_Free	1	5.450601	0.646926	8.43	<.0001***

This parameter is significant at 95% confidence level, it shows how free apps and paid apps behave differently in terms of installs. When other things being equal, free apps is estimated to have 4 times more installs than the average level.

This could be a very important factor for companies to consider, that free apps would be a better choice if they want to generate a large customer base.

Category* Type:

Category_new_BUSIN*Type_Free	1	-1.681634	0.84839	-1.98	0.0475**
Category_new_FAMIL*Type_Free	1	-1.522149	0.36954	-4.12	<.0001***
Category_new_LIFES*Type_Free	1	-1.628261	0.390679	-4.17	<.0001***
Category_new_MEDIA*Type_Free	1	2.408536	0.699957	3.44	0.0006***

This interaction shows how the categorical difference of free and paid apps will affect installs of the app differently. At 95% confidence level, when other things being equal, for apps under Business, Family, and Lifestyle categories, free apps will have lower installs, comparing to average level of free apps. While free apps under Media_video category will have higher installs.

Size* Type:

Size*Type_Free	1	0.018761	0.007245	2.59	0.0096***

This interaction shows, at 95% confidence level, when other things being equal, for apps with the same size, free apps is estimated to have 1.87% more installs.

Content Rating:

Content_Rating_Everyone	1	-0.821072	0.485727	-1.69	0.091*
Content_Rating_Everyone*Type_Free	1	0.717311	0.497117	1.44	0.1491

It seems overall, the content rating level doesn’t affect the installs significantly. At 95% confidence level, the installs of apps with content rating level – Everyone is significantly different from the average level.

Update:

update_0	1	-2.578182	1.171163	-2.2	0.0277**
update_1	1	-2.148212	0.410142	-5.24	<.0001***
update_2	1	-1.007819	0.518503	-1.94	0.052*
update_3	1	-0.87284	0.635462	-1.37	0.1696
update_4	1	0.289754	0.784037	0.37	0.7117
update_5	1	-2.064508	0.221256	-9.33	<.0001***
update_6	1	-1.612613	0.26228	-6.15	<.0001***
update_7	1	0.493418	1.290028	0.38	0.7021
update_8	1	0.539661	1.443717	0.37	0.7086

The coefficients of this series of variables shows the relationship between number of major updates and installs of the apps. We can tell from the estimation results that, at 95% confidence level, the installs of applications with 0, 1, 5, and 6 times of major updates are significantly lower than the average level.

Type*Update:

Type_Free*update_0	1	-2.864084	1.211599	-2.36	0.0181**
Type_Free*update_1	1	-2.498493	0.422727	-5.91	<.0001***
Type_Free*update_2	1	-2.022606	0.534334	-3.79	0.0002***
Type_Free*update_3	1	-2.070923	0.65341	-3.17	0.0015***
Type_Free*update_4	1	-2.325625	0.804323	-2.89	0.0038***
Type_Free*update_7	1	-2.986167	1.32913	-2.25	0.0247**
Type_Free*update_8	1	-3.324689	1.486383	-2.24	0.0253**

This interaction shows how frequency of updates affect free and paid apps differently. At 95% confidence level, when the application has 0,1,2,3,4 and 7 ,8 times of major updates, free apps is estimated to have lower installs than average level.

One possible explanation is that, when paid apps are lunched, they might have already passed several strict tests to avoid serious bugs and have made more efforts to improve customer experience. But for free apps, they might focus more on entering the market early, so bugs are improvements can wait after lunch of the app. Thus, at earlier versions, paid apps would be likely to generate more installs than free apps. However, when free apps go through more major updates to improve functions and fix bugs, in general they would also be likely to generate more installs.

3.2.3. Price

According to the Backward Selection based on Mallows Cp selection Criteria, the following effects are selected.

Table 10. Parameters in Price model

Effects:	Intercept,Category_new,Size,Size*Category_new,log_Reviews, log_Revie*Category_new, Content_Rating Recency

Table 11. Estimated coefficients in Price model

Parameter	DF	Estimate	Standard	t Value	Pr > |t|
			Error		
Intercept	1	6.51938	2.96738	2.2	0.0283**
Category_new BUSIN	1	7.66387	3.36069	2.28	0.0229**
Category_new FAMIL	1	-1.6247	2.95784	-0.55	0.583
Category_new GAME	1	-1.8092	3.07185	-0.59	0.5561
Category_new LIFES	1	1.1153	2.97489	0.37	0.7078
Category_new MEDIA	1	-0.1951	3.11067	-0.06	0.95
Category_new SOCIA	1	-0.8493	3.12056	-0.27	0.7856
Category_new TOOLS	1	-1.7655	2.92626	-0.6	0.5465
Category_new TRANS	0	0	.	.	.
Size	1	-0.0272	0.04651	-0.58	0.5588
Size*Category_new BUSIN	1	0.15045	0.11617	1.3	0.1957
Size*Category_new FAMIL	1	0.01166	0.04867	0.24	0.8107
Size*Category_new GAME	1	-0.0239	0.05041	-0.47	0.636
Size*Category_new LIFES	1	0.07831	0.05003	1.57	0.118
Size*Category_new MEDIA	1	0.06177	0.07576	0.82	0.4152
Size*Category_new SOCIA	1	0.03887	0.08777	0.44	0.658
Size*Category_new TOOLS	1	-0.0103	0.05102	-0.2	0.8408
Size*Category_new TRANS	0	0	.	.	.
log_Reviews	1	0.13491	0.3832	0.35	0.7249
log_Revie*Category_n BUSIN	1	-1.2928	0.48633	-2.66	0.008***
log_Revie*Category_n FAMIL	1	0.06393	0.39459	0.16	0.8713
log_Revie*Category_n GAME	1	0.14912	0.41002	0.36	0.7162
log_Revie*Category_n LIFES	1	-0.2879	0.40612	-0.71	0.4786
log_Revie*Category_n MEDIA	1	-0.529	0.46641	-1.13	0.2571
log_Revie*Category_n SOCIA	1	-0.289	0.4482	-0.64	0.5193
log_Revie*Category_n TOOLS	1	-0.0447	0.39396	-0.11	0.9096
log_Revie*Category_n TRANS	0	0	.	.	.
Content_Rating Everyone	1	-1.4921	0.68683	-2.17	0.0302**
Content_Rating Mature 1	1	-1.4381	1.25807	-1.14	0.2534
Content_Rating Teen	0	0	.	.	.
Recency	1	-0.0006	0.0003	-1.98	0.0484**

(***: significant at 99% level; **: significant at 95% level; *: significant at 90% level)

For the dependent variable Price, only four variable shows statistically significant in our model. And we are going to dig into these four variables in the following paragraph.

Category:

From the table above, we can conclude that under 95% confident level, the only variable that shows statistically significant in this parameter is Business category. All other categories are not showing any statistically significant at this level.

Under this circumstance, we only focus on the estimated coefficients of Business category, which indicating from the table, is 7.66387. According to the estimated coefficients of each categories, it seems that applications in Business category is most likely to have higher effect on Price factor. Therefore, if company wants to develop an application in the category of Business, it should care more about the factors which may affect the Price.

Log_Review*Category:

We can observe from the above table that under 95% confident level, the only variable that shows statistically significant in this interaction term parameter is Log_review*Business category. All other interaction terms within this parameter are not showing any statistically significant at this level.

Just like what we analyze in Category parameter, we would only focus on the estimated coefficients of Log_review*Business category, which indicating from the table, is –1.2928. According to the estimated coefficients of each categories, it seems that Log_review*Business category has larger effect on Price factor. Therefore, if company wants to develop an application, it is worthy paying more attention on customers' reviews in Business category.

Content_Rating:

Content_rating is the third parameter that shows statistically significant in this model. We noticed that for content rating, there might be a slight difference between teenager and mature. In this price model, content rating from teenager is our base group. It is surprising to observe p-value of content rating from mature is not even statistically significant under 95% confident level, while that of everyone do show significant.

The estimated coefficient of variable Content_Rating from everyone is –1.4921, which can be interpreted as other things being equal, one more Content_rating would cause the price of application decreased by 1.4921. It is plausible that Content_rating has effect on price, no matter what the rating is, company would adjust their prices based on customer's rating.

Recency:

At first glance, it is reasonable to connect application's recency to the price of an application. For we define Recency as "the difference between the data collected date (sep/4/2018) and newest update date”. Statistic from about table back up our assumption. P-value of Recency is 0.0484, which under 95% confident level, is statistically significant. That means recency do have effect on price of application.

What is intriguing is the negative 0.0006 in estimated coefficient of recency. This number can be interpreted as if recency is longer, the application price would decrease. It reveals the fact that consumers do care of recency, or the update frequency, and such frequency would impact application price. Therefore, when company launch a new application, they would have a plan on when is the best time to update or modify the application.

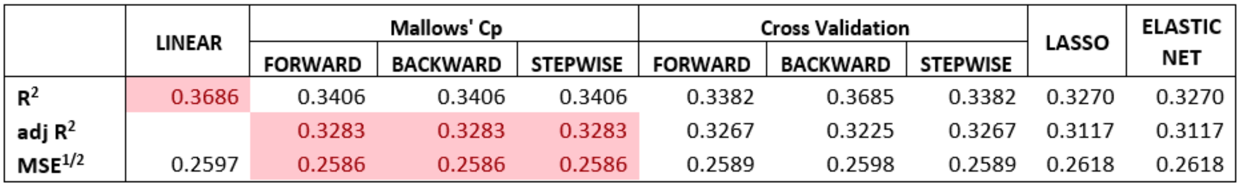
3.2.4. Sentiment Analysis of User Reviews

In 3.2.1, we analyzed how the rating of apps are affected by size, installs, category, price, content rating and frequency of updates. However, even the best model could explain only 10% of the data (adjusted R-square at 0.1 level). This inspired us that we may need to consider more sources of data, such as app reviews from users, which could be more related to rating of the apps.

In the googleplaystore_user_reviews.csv dataset, more than 64,000 reviews are generated for more than 1000 free apps with initial letters from A-H. based on this dataset we summarized the user sentiment aspects of 865 apps based on the Sentiment Score, Sentiment Polarity and Sentiment Subjectivity of app reviews.

However, only 765 of the apps are recognized after merging with the google_apps.csv dataset. We then applied a new model selection process on this new dataset and the results are the below:

Table 12. Selection process on new dataset – with sentiment analysis



	LINEAR	Mallows' Cp			Cross Validation			LASSO	ELASTIC NET
		FORWARD	BACKWARD	STEPWISE	FORWARD	BACKWARD	STEPWISE		
R^2	0.3686	0.3406	0.3406	0.3406	0.3382	0.3685	0.3382	0.3270	0.3270
adj R^2		0.3283	0.3283	0.3283	0.3267	0.3225	0.3267	0.3117	0.3117
$MSE^{1/2}$	0.2597	0.2586	0.2586	0.2586	0.2589	0.2598	0.2589	0.2618	0.2618

Based on adjusted R-square criteria, Mallows’ Cp –Forward, backward and stepwise provides the same level of best performance. We will use Forward selection model to interpret for App Rating after adding three more variables about sentiment.

The selected effects are as followed:

Table 13. Parameters of best model

Effects:	Intercept Category_new Size log_Installs log_Reviews Recency Sentiment_score Sentiment_Polarity Sentiment_Subjectivi

Table 14. Estimated coefficient in best model

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.119322	0.104798	39.31	<.0001***
Category_new BUSINESS	1	0.051311	0.052324	0.98	0.3271
Category_new FAMILY	1	0.176367	0.056047	3.15	0.0017***
Category_new GAME	1	0.167028	0.052091	3.21	0.0014***
Category_new LIFESTYLE	1	0.037956	0.044069	0.86	0.3894
Category_new MEDIA_VID	1	0.060654	0.054822	1.11	0.2689
Category_new SOCIAL_AN	1	-0.076272	0.045621	-1.67	0.095*
Category_new TOOLS	1	0.052096	0.04454	1.17	0.2425
Category_newTRANSPORT	0	0	.	.	.
Size	1	-0.001191	0.00059	-2.02	0.0437**
log_Installs	1	-0.077977	0.009923	-7.86	<.0001***
log_Reviews	1	0.100845	0.009137	11.04	<.0001***
Recency	1	-0.000205	3.292E-05	-6.22	<.0001***
Sentiment_score	1	0.090232	0.053873	1.67	0.0944*
Sentiment_Polarity	1	0.399449	0.112108	3.56	0.0004***
Sentiment_Subjectivi	1	0.264636	0.119162	2.22	0.0267**

(***: significant at 99% level; **: significant at 95% level; *: significant at 90% level)

After adding sentiment features into the regression model for Rating, adjusted R-square increases to 0.33, which implies that the new model fit the data set better. The newly added features – Sentiment_Polarity and Sentiment_Subjectivity are both significant at 95 % level, and Sentiment_score is significant at 90% level.

The coefficient of Sentiment Polarity score is 0.39, which means when other things being equal, if the average sentiment polarity scores of all reviews of the app increase by 0.1(more towards positive attitudes), the rating of the app is estimated to increase by 0.039.

The coefficient of Sentiment Subjectivity score is 0.26, which means when other things being equal, if the average subjectivity polarity scores of all reviews of the app increase by 0.1(more subjective reviews), the rating of the app is estimated to increase by 0.026.

The significance of some original features is changed. In Category, Family and Game categories are significant at 95%level, while other categories like Business, Lifestyle, Media_video and Tools have no significant impact on rating of apps. Size, log(installs), log(reviews) and Recency are still significant.

This proves that sentiment features are significantly affecting ratings of apps. Further analysis can be conducted if we could obtain more data on app reviews.

Conclusion

In this report, we create three models in order to solve our three research questions.A few model selection methods are implemented here to help us find out which one is the best model, and how we interpret the one that provide best performance. The following paragraphs are a few points we want company to pay attention to, and they are from the perspective of application's rating, the amount of installation and application's price.

Rating:

To gain a higher rating score, there are several points should keep in mind.

1) Developing applications in Family, Media_Video, Tools, and Lifestyle categories is easier to get higher ratings.

2) Creating a relatively larger size of applications in Family, Game and Tools categories will gain the higher ratings, but in other categories, the smaller the better.

3) Increasing the installs of applications may help to gain their popularity among the industry, but it may also ruin the ratings with large amounts of downloads at the same time.

4) Users are stricter to the applications which content rating is for everyone and 17+ lower ratings are usually gained comparing to other categories.

5) Prolonging the duration of version updating has positive impact on ratings; having no major updating at all and having as frequent as 8 times or no updating are better for applications to get higher ratings.

6) Free applications are much more difficult to gain higher ratings than paid ones, but paid applications should focus more on its quality.

7) Positive and subjective reviews have significant impact on improving rating scores.

Management teams should pay attention to user experience and feedbacks.

Installs:

To improve installs of app, we have the following conclusions:

1) Four categories of apps have the most installs among all: Family, Lifestyle, Social Media and Business. Category with least installs is Media_video.

2) Generally, applications with higher installs will have larger size than applications with lower installs. This is consistence in Family and Social Media categories. For particular category such as Tool, users will prefer smaller sizes to larger sizes.

3) Applications with higher ratings would generate higher installs. However, when taking categorical differences into consideration, when ratings are the same, apps under categories of Lifestyle, Business and Social Media would have lower installs comparing to average levels.

4)The installs of free applications are estimated to be 5 times of the average level. Thus free applications are more popular than paid apps, especially for Media_video category. But for Business, Family, and Lifestyle categories, the difference in installs between free and paid apps are less obvious.

5)Frequency of updates affects the installs of free and paid apps differently. At earlier versions, or when the apps have major updates less frequently, paid apps tend to generate more installs. But when the apps go through more major updates, for example, after 7 major updates, free apps tend to have more installs.

Price:

Price is the focal point in supply and demand relationship. Company do care of price because it wants to reap benefits as much as possible, while consumers would always have a second thought on it.

If company intends to maximize their revenue, some factors should be placed more weight on. Applications that can be classified as business type is highly recommended. Content_rating is also importance, and company might not need customer segmentation if time is limited, for the statistic shows only overall rating matters.

The other thing company should pay attention to the recency of application. It is of crucial importance to decide when the best time to update the application is. It might not smart to update it too frequently.

Strategies

- Free Apps

For companies want to generate large user bases, free apps should be a better choice. While free apps have more advantages in entering or expanding the market, developers and management teams should focusing more on improving quality of the apps, considering improving ratings of free apps is more difficult than paid apps.

Although users are more tolerable for frequent updates in free apps, management teams should always pay attention to user reviews and provide better functions and services to get more positive feedbacks. This is crucial for both improving ratings and stimulating installs.

For apps under Media_vedio categories, more strategies can be applied to motivate installs or subscriptions such as free trials, frequently updating more features or resources.

For apps under Family, Social Media and Game categories, larger sizes are considerable. But for apps under Tools categories, smaller sizes should be a better choice to stimulate installs.

- Paid Apps

The major advantage of paid apps is that people would be willing to pay for better quality. Thus maintaining good product quality and brand image are of crucial importance for paid apps.

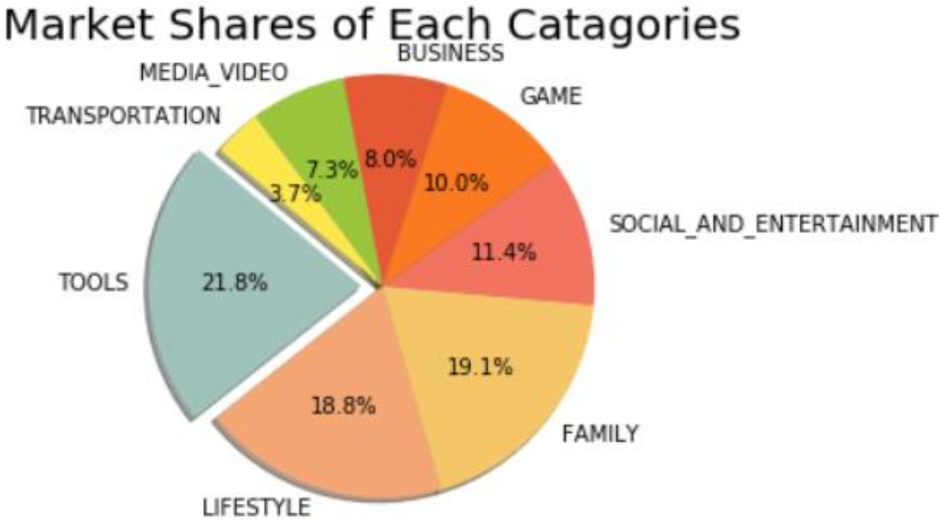
For apps under Business, Family and Lifestyle categories, users are more willing to choose paid apps. Developers and management teams may consider offering competitive prices for these as long as the quality of the app is satisfying.

Users are less tolerable for frequent updates in paid apps. Therefore, developers and management teams need to pay more attention to improve user experience and eliminate bugs before launch of the app and at early versions.

For more aggressive companies who are willing to take risks, developing games and targeting on teenagers and 17+ people has potential to be very profitable. This user group is very picky, but once accepted by them, the apps could make a huge success.

Appendix 1

Table 1. Market shares of each categories after rearrangement


Market Shares of Each Catagories
BUSINESS
MEDIA_VIDEO
GAME
TRANSPORTATION
SOCIAL_AND_ENTERTAINMENT
TOOLS
FAMILY
LIFESTYLE