Use SAS Enterprise Miner Workstation 15.1 to Do Predictive Analysis for Mobile Strategy Games Industry

Chun Du, Oklahoma State University, Stillwater, OK

As a data analyst and agricultural economist, graduated from Oklahoma State University in spring 2020 from M.S. in Agricultural Economics and a 1st year master student of Business Analytics and Data Science program. Interested in using the big-data methods, statistical methods to analyze data, define problems, find the value behind the data, and estimate and evaluate the best situation. In my last master program, devoted to do research Chinese preference and nutritional knowledge for pecans.

ABSTRACT

This study aims to predict mobile strategy games' ratings, to find the relationships between the game's factors and games ratings, and help game developers, game players, and game companies to define a successful game. The rating of games is divided into two groups: group 1 - rating is below 4 (bad performance), group 2 - rating is 4 and above 4 (good performance). The overview of study plan is dividing the sample into 70/30 training and validation. Logistic regression, decision tree and neural network will be used to build predictive models. By conducting text cluster analysis and topic analysis to analyze games descriptions and find the traits and categories of strategy mobile game with rating above 4. The final predictive model, neural network is the winner models with sensitivity of predicting game rating of 97.91% and total 25.87% misclassification rate.

INTRODUCTION

Games have always been the most popular category in the iTunes App Store. Whatever way you care to slice it – number of active apps, number of downloads, time spent, and revenue generated (Simon, 2014). In this modern time and society, the mobile games industry is worth billions of dollars. Companies are spending vast amounts of money on the development and marketing of these games to an equally large market. One of the most popular and favorite genre of mobile games is Strategy Games (Prasetya, 2019). A new Sensor Tower report reveals that gaming made up approximately $29.6 billion of mobile revenue during 2019's first half, or approximately 75 percent of $39.7 billion total (Ana, 2019). In 2020, smartphone games generated approximately 74.9 billion U.S. dollars in annual revenue, accounting for 43 percent of the global gaming market during the measured period (J.Clement, 2021). According to Statista Research Department's January of 2021 reports, smartphone games generated $63.6 billion in global revenue in 2020, while console gaming was the third largest segment on the market with a revenue of only $15.4 billion at end of 2019. Therefore, the most intuitive way to define a successful game is ratings. I work on creating a better fit model to look at the different factors that go into making a popular game.

Previous studies have found that the rating of a strategy game is closely related to its size and price, the age of the user. The bigger the size the smaller the ratings. But it's kind different for Paid Apps, as the big sized apps still get a decent ratings. The most of the Mobile Strategy Games in the App Store have the 4+ rating. This might means that Mobile Strategy Games are really popular for kids with the age of 4 and above (Prasetya, 2019). Where games above $8.99 always scores 4.0 and above. While games below $5.99 shows a rating range of 1.5 to 5. A very small portion of games (less than 100 titles) are above 1GB, in which the minimum score for the game is 3-3.5. This might be due to the user sentiment who gives credit to the huge game content and possibly better game graphics. We can see that games that lies between 1.2GB and 1.7GB gets score >4 (Hkhoi, 2019). However, those previous studies have looked at the impact of each factor on the rating independently, but ignoring the internal relationship among the size of a game's APP and the price of the game, the age of the game, the availability of in-app purchases and the time of release, as well as their combined impact on the rating. Therefore, the interest of this study is by using game price data, game size data, language, game genre, and age rating of app to predict the ratings of strategy games. I decided to use Latent Dirichlet Allocation (LDA) to create topics for games which were rated above 4+ and explore what topics could bring a popular games.

DATA PREPARATION

In this study, 2004-2019 strategy game on the Apple App Store dataset is an open data set uploaded from Kaggle, where 17,007 strategy games collected on the 3rd of August 2019, using the iTunes API and the App Store sitemap. I created target variable GAME_RATING and divided the Ratings into 3 groups: Group 1 - rating is below 4 (bad performance); Group 2 - rating is 4 and above 4 (good performance). Unfortunately, there are a large number of missing values for my target variable. Since the total amount of observation samples is very large, and the target variable is required effective value during this studying how the dependent variable is affected by other factors, I removed the rows where there are missing values in target variable. Finally I got 7561 sample sizes. That's still a pretty impressive data set. In addition to considering the language type of the game software to effect rating scores, I have created NEW_LANGUAGES variable: Group 1 - only one language, Group 2 -2 or 3 types of languages, and Group 3 -4 anOrigniald 4 above. It was generated based on the original LANGUAGES variable. The last step to prepare data is dealing with PRIMARY_GENRE variable. Considering this study is for strategy games, I created a new dummy variable for the main app category, PRIMARY_GENRE2: If the main genre is Game, it goes to 1; if not, it goes to 0. The detailed variable information collected in this study is described in the Table 1.

Variable Name	Model Rule	Measurement Level	Description
URL	Rejected	Nominal	The link to the app through the App Store.
ID	ID	Nominal	The ID of the app in the App Store.
NAME	Rejected	Nominal	The name of the app.
SUBTITLE	Rejected	Nominal	The secondary text under the name.
ICON_ URL	Rejected	Nominal	URL to the app’s icon image.
AVERAGE_USER_RATING	Rejected	Interval	The average user rating of the app, rounded to nearest, 5.
GAME_RATING	Target	Ordinal	The new rating group. Group 1 - rating is below4 (bad performance), group 2 -rating is 4 and above 4 (good performance).
USER_RATING_COUNT	Rejected	Interval	The numbers of user rating the app have obtained internationally.
PRICE	Input	Interval	The price of the apps in the App Store (USD).
IN_APP_PURCHASE	Rejected	Interval	Prices of available in app purchases.
DESCRIPTION	Text	Nominal	A quick description of the app.
DEVELOPER	Rejected	Nominal	The team that develops the app.
AGE_RATING	Input	Interval	The age ratings of the app.
LANGUAGES	Rejected	Nominal	The languages the apps use.
NEW_LANGUAGES	Input	Ordinal	The new language group: group 1 - only one language, group 2 - 2 or 3 types of languages, and group 3 - 4 and 4.
SIZE	Input	Interval	The size of the apps (bytes).
PRIMARY_GENRE	Rejected	Nominal	The main genre of the app.
PRIMARY_GENRE2	Input	Ordinal	Whether the main genre is Game. 0 is No; 1 is Yes.
GENRES	Rejected	Nominal	Genres of the app.
ORIGINAL_RELEASE_DATE	Rejected	Nominal	When the app was released.
CURRENT_VERSION_RELEASE_DATE	Rejected	Nominal	When the app was last updated.

Table 1. The model rules and description of collected variable in predictive modeling.

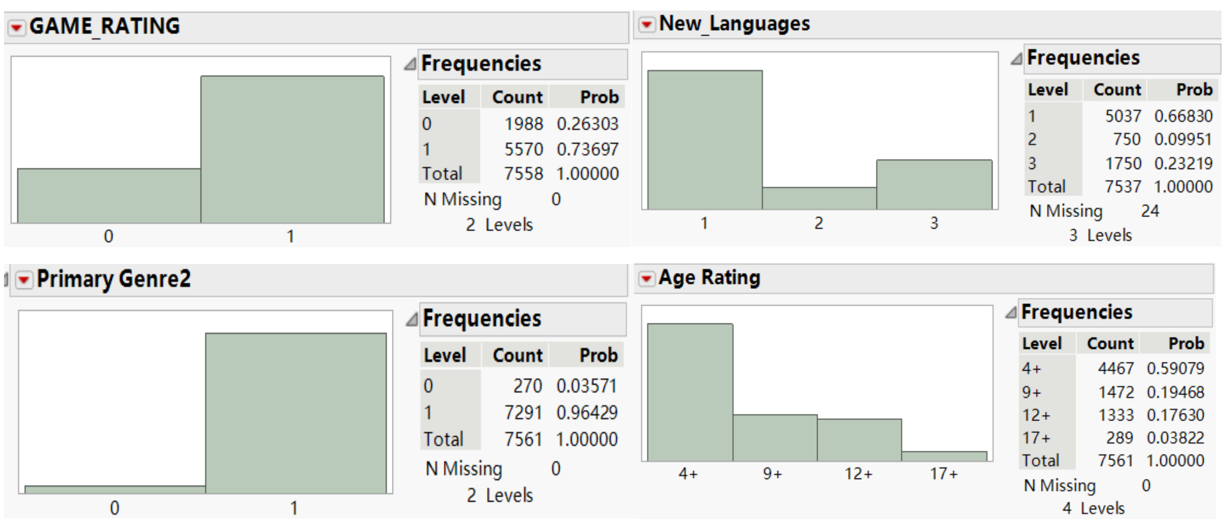
METHODOLOGY

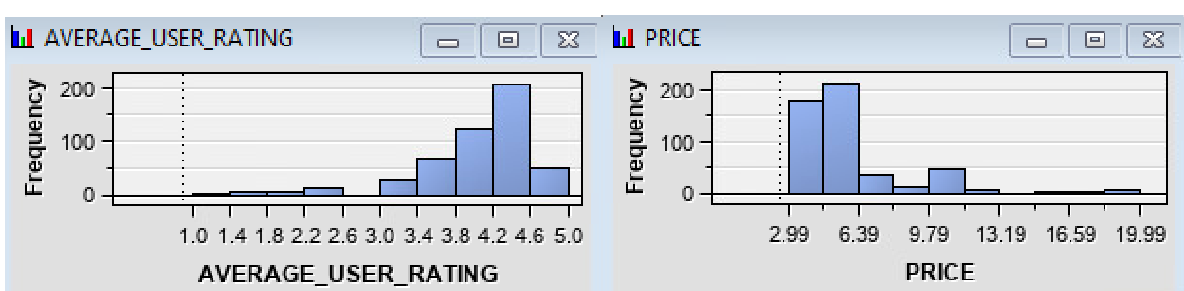
The analysis starts with understanding how the factors influencing game's rating by applying one-way analysis and bivariate analysis through SAS JMP 14 to look at the significance between factors and target variables. After descriptive statistics analysis, SAS Enterprise Miner Workstation 14.3 is used to do sampling, data transformation, 70/30 training and validation partition, and imputation. Then I built predictive models through logistic regression, decision tree, and neural network modeling based on the best misclassification rate of validation data. Finally, Expectation-Maximization Cluster Algorithms and Hierarchical Cluster Algorithms were selected since there is one text variables within the data, which is significant to be explored how descriptions of game have effects on players' rating. Text mining and cluster analysis were the descriptions for games which were rated above 4+. By analyzing these descriptions, it's easy to explore any trends within the descriptions among 4+ games. What words are common among these descriptions?

DESCRIPTIVE STATISTICS AND ANALYSIS

INSIGHT FROM DESCRIBTIVE STATISTICS

The table 2 below describes summary statistics for the five numeric variables in my data set, PRICE, SIZE, USER_RATING_COUNT, AGE_RATING, and AVERAGE_USER_RATING. Histograms for four categorized variables are provided below too. The mean for PRICE or AGE_RATING in years is slightly greater than the median so the data has a longer right-hand tail. Visually, GAME_RATING seems left skewed. The maximum value of PRICE appears to be an outlier, 139.99, which is much higher than mean and median. Therefore, the variable has a large standard deviation and creates the perception of high variation in price. Likewise, we can see the effect of the outlier in the histogram as the data does not seem appropriately distributed.





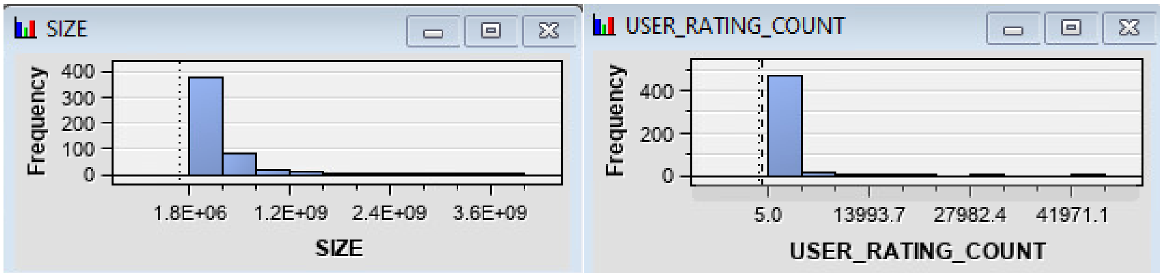


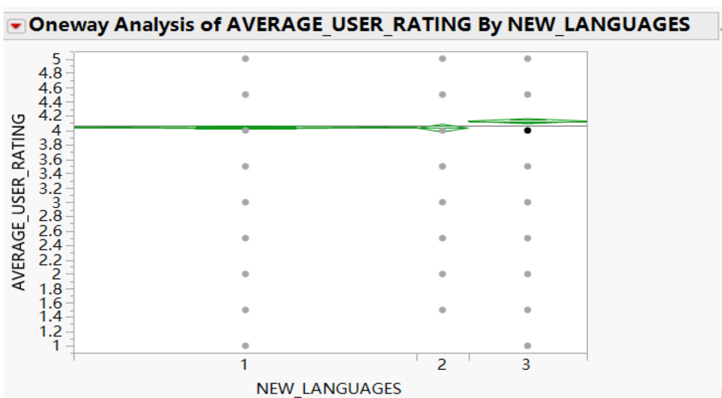
Figure 1. Histograms of Three Categorized Variables and Five Numerical Variables in Data Set

Variable	Mean	Median	Standard Deviation	Maximum	Minimum
PRICE	0.5713	0	2.416	139.99	0
SIZE	151467919	79646720	255038050	4008891040	215840
USER_RATING_COUNT	3306.53	46	42322.561	3032734	5
AGE_RATING	6.8807	4	3.7833	17	4
AVERAGE_USER_RATING	4.0609	4.5	0.7514	5	1

Table 2. Summary Statistics for the Five Numeric Variables in Data Set

INSIGHT FROM ONEWAY ANOVA ANALYSIS

Through the One-way ANOVA analysis, the mean AVERAGE_USER_RATING of Group 3 in NEW_LANGUAGES is significant different from other groups, and the mean AVERAGE_USER_RATING of Game type is significant different from Non-Game type. From the results of One-way ANOVA (Figure 2), the mean of AVERAGE_USER_RATING is 4.0613 and is statistically significant different from three language groups, which are ranging from 4.03267 to 4.12914. The game which has only one language is approximately 0.08 of average rating lower than game with 4 and 4 above languages. It might indicate the game app with more languages could attracts more different types of race to rate, which is a potential advantage. The mean of PRIMARY_GENRE2 is 4.0609 and is statistically significant different from whether it is a game type, which are ranging from 3.9259 to 4.0659. If the definition of one mobile strategy game is 'game' type, it is approximately 0.14 of average rating higher than it is not "game" type. It might indicate pure game type would have more potential advantage than other types, such as educational game, cooking game.



Oneway Anova

Summary of Fit

Rsquare	0.00248
Adj Rsquare	0.002215
Root Mean Square Error	0.749776
Mean of Response	4.061364
Observations (or Sum Wgts)	7537

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
NEW_LANGUAGES	2	10.5286	5.26428	9.3643	<.0001*
Error	7534	4235.3406	0.56216		
C. Total	7536	4245.8692			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	5037	4.04209	0.01056	4.0214	4.0628
2	750	4.03267	0.02738	3.9790	4.0863
3	1750	4.12914	0.01792	4.0940	4.1643

Std Error uses a pooled estimate of error variance

Missing Rows 24

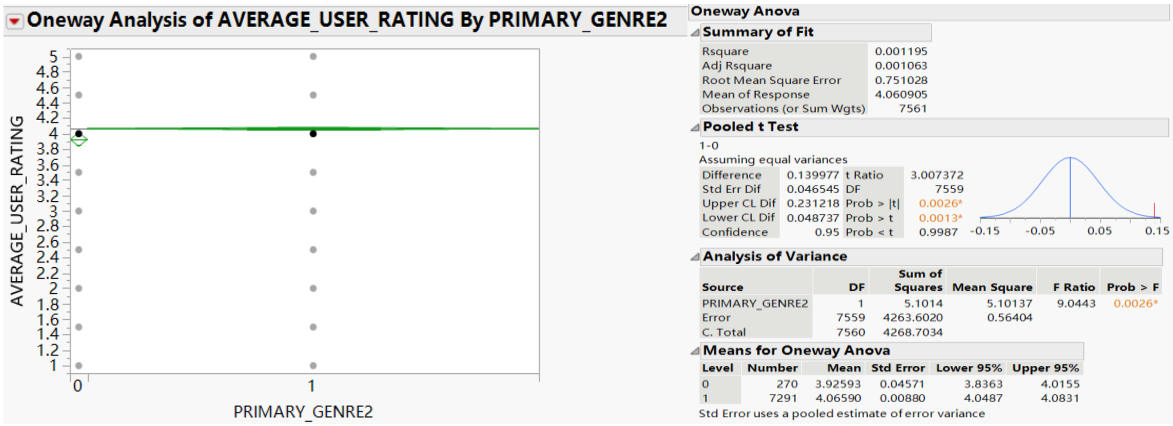
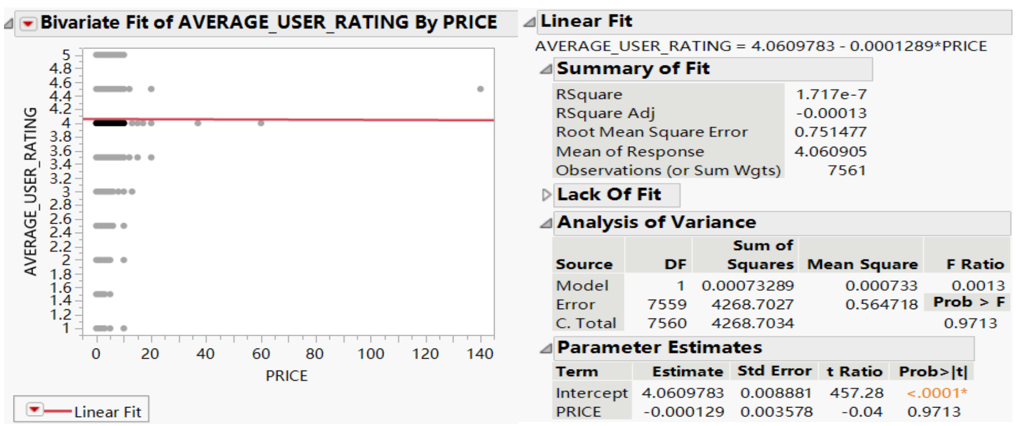
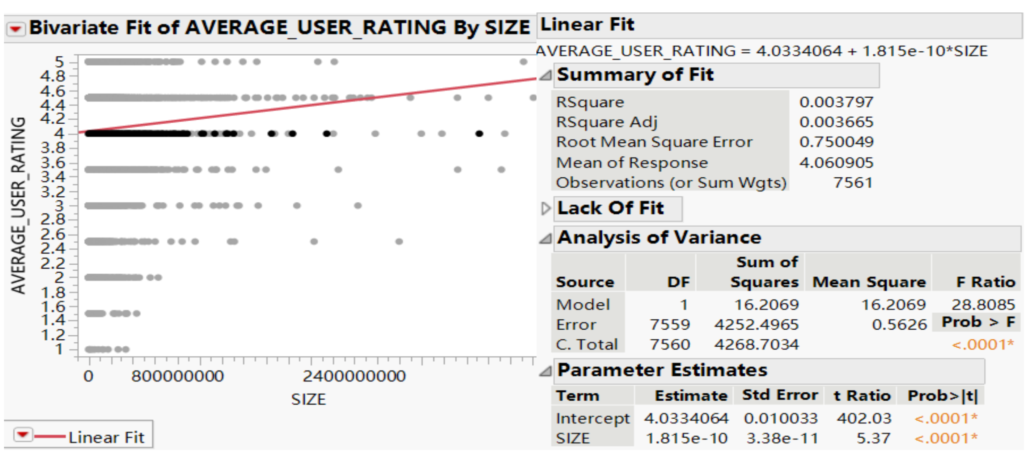


Figure 2. Oneway-ANOVA Analysis of Two Variables: NEW_LANGUAGES and PRIMARY_GENRE2

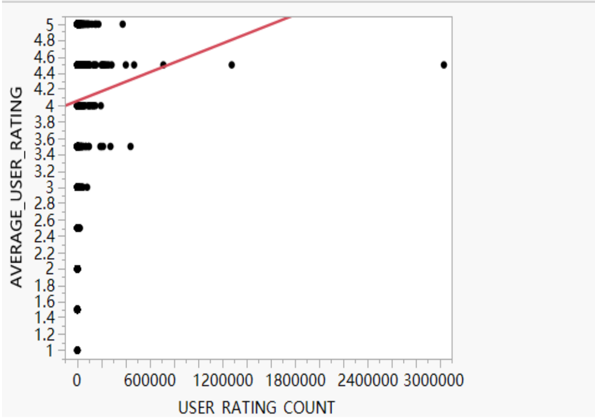
INSIGHT FROM BIVARIATE FIT ANALYSIS

To find the relationship among average rating and price, size, user rating count, and age rating, I use bivariate analysis. Based on the results of bivariate analysis (Figure 3), the Size, USER_RATING_COUNT, and AGE_RATING are significantly important variables of AVERAGE_USER_RATING at significant level of 0.1. However, the PRICE is not significantly with p-value of 0.97. The reason I suspected is there are high effect from extreme values. There are only three extreme values in price: 139.99, 59.99, and 36.99. Those three sample sizes won't be considered in final predictive models.





Bivariate Fit of AVERAGE_USER_RATING By USER_RATING_COUNT



Linear Fit

AVERAGE_USER_RATING = 4.0589659 + 5.8635e-7*USER_RATING_COUNT

Summary of Fit

RSquare	0.001091
RSquare Adj	0.000958
Root Mean Square Error	0.751067
Mean of Response	4.060905
Observations (or Sum Wgts)	7561

Lack Of Fit

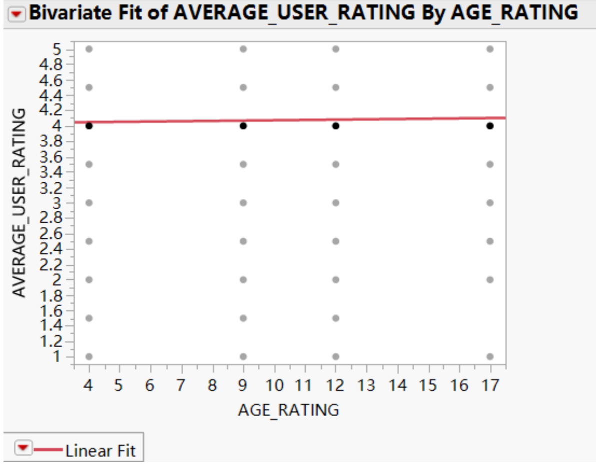
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	4.6556	4.65560	8.2531
Error	7559	4264.0478	0.56410	Prob > F
C. Total	7560	4268.7034		0.0041*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.0589659	0.008664	468.49	<.0001*
USER_RATING_COUNT	5.8635e-7	2.041e-7	2.87	0.0041*

Bivariate Fit of AVERAGE_USER_RATING By AGE_RATING



Linear Fit

AVERAGE_USER_RATING = 4.0331528 + 0.0040333*AGE_RATING

Summary of Fit

RSquare	0.000412
RSquare Adj	0.00028
Root Mean Square Error	0.751322
Mean of Response	4.060905
Observations (or Sum Wgts)	7561

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1.7603	1.76027	3.1184
Error	7559	4266.9431	0.56449	Prob > F
C. Total	7560	4268.7034		0.0775

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.0331528	0.017934	224.89	<.0001*
AGE_RATING	0.0040333	0.002284	1.77	0.0775

Figure 3. Bivariate Fit Analysis of Four Variables: Price, Size, USER_RATING_COUNT, and AGE_RATING

PREDICTIVE MODELING

By using SAS Enterprise Miner Workstation 15.1, I created work flow diagram as below Figure 4. The excel dataset was exported into uniform SAS dataset by JMP. In the third transform variable node, it was add to transform the skewed interval variable PRICE, SIZE, USER_RATING_COUNT, AGE_RATING, and AVERAGE_USER_RATING. The algorithms of transformation to each variables are showed in Figure 5. The fourth node is data partition node, which is used to split the data into 70% for training data set and 30% for validation data set. Imputation node is used for imputing missing value. Fortunately, the results of imputing showed that there is no missing value.

Finally, I created 4 predictive models:

a) Log Regression

b) Decision Tree

c) Neural Network

d) Neural Network base on Decision Tree.

I have tried to select optimizing regression model with stepwise selection and Validation Error as selection Criterion. But there are only 9 inputs within modeling processing, so there is no difference between optimization and default setting.

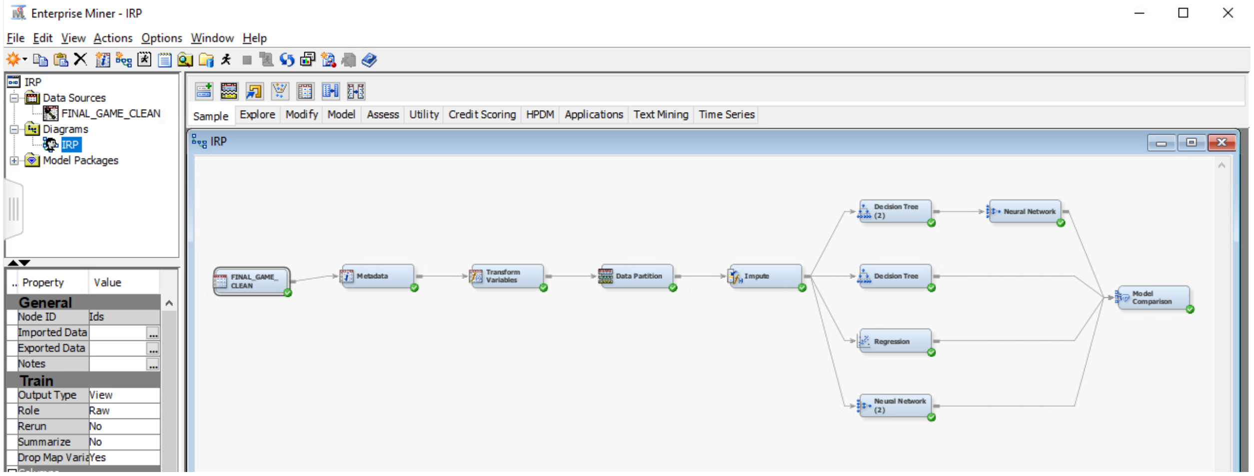


Figure 4. The Workflow Diagram of Four Predictive Models in SAS Enterprise Miner.

Computed	EXP_AVERAGE_USER_RATING	exp(max(AVERAGE_USER_RATING-1, 0.0)/4)
Computed	LOG_AGE_RATING	log(max(AGE_RATING-4, 0.0)/13 + 1)
Computed	LOG_PRICE	log(max(PRICE-0, 0.0)/19.99 + 1)
Computed	LOG_SIZE	log(max(SIZE-215840, 0.0)/4005375200 + 1)
Computed	LOG_USER_RATING_COUNT	log(max(USER_RATING_COUNT-5, 0.0)/303272...

Figure 5. Transforming Algorithms of Five Interval Variables

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate and False Positive Rate. AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0, 0) to (1, 1) (Google Developer). ROC and AUC was used measure how well predictions are ranked and the quality of the model's predictions. The bigger value of AUC are, the better the classier performs. The results shows the Neural Network base on Decision Tree model has the largest Valid ROC index, 0.687.

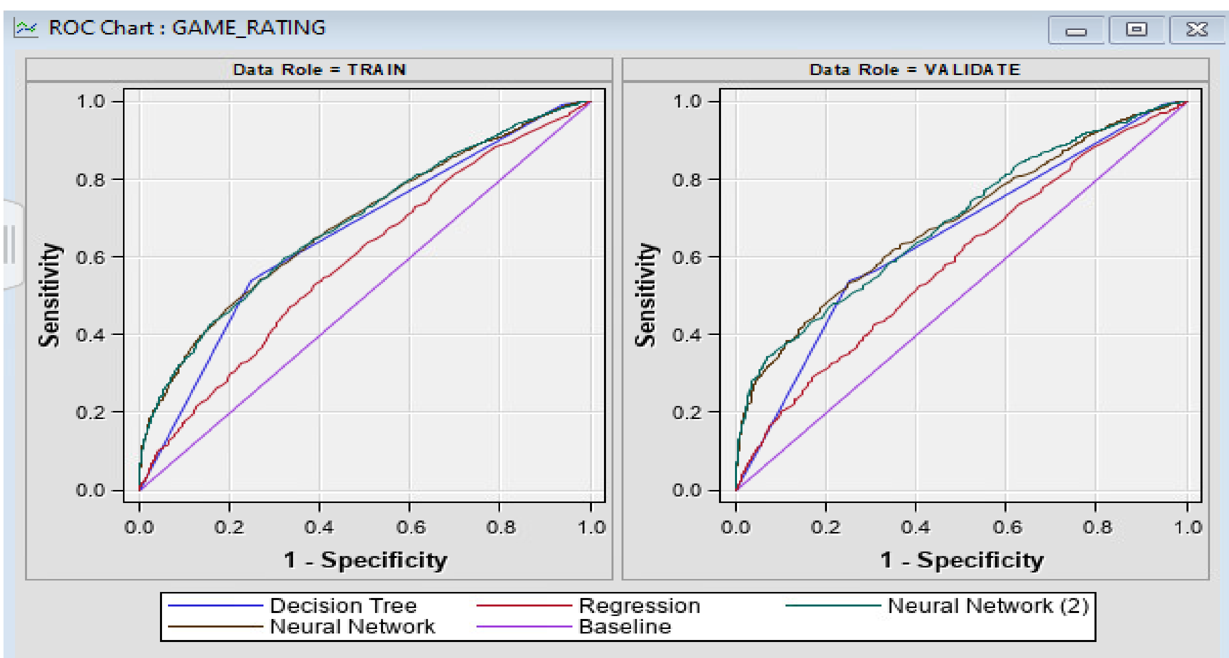


Figure 6. Receiver Operating Characteristic Curve (ROC) Charts of Train Data and Validate Data

The Fit Statistics result from model comparison node shows the four predictive models' valid misclassification rate of 25.65 % to 26.31%. The decision tree model provides the lowest valid and train misclassification: 25.65% and 25.37%. The train average squared error of predictive models are range from 0.17811 to 0.1909. The valid average squared error of predictive models are range from 0.17664 to 0.19104. In general, the differences among those statistics number are slight.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree	Decision Tree	0.25650	0.17848	0.25373	0.18010
	Neural2	Neural Network (2)	0.25870	0.17811	0.26016	0.17664
	Neural	Neural Network	0.26179	0.17866	0.25865	0.17760
	Reg	Regression	0.26311	0.19090	0.26300	0.19104

Figure 7. Misclassification Rate of Four Predictive Models

Error! Reference source not found. and Table 3 show the precision and sensitivity of predicting whether a strategy mobile phone got a 4-star or 4-star above rating with regression model, decision tree, neural networks, and neural networks based on decision tree. Those four models have good performances on precision of approximately 75%, and sensitivity of approximately 98%. The differences among those parameters are much slight. According to the valid misclassification rate, the winner model would be decision tree. But based on other fit statistics parameters, neural network's performance of predicting game rating is the best, with the least ASE and RMSE, the highest ROC index and precision, and the misclassification rate of this model is in the second best rank.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg	Regression	TRAIN	GAME_RATING		0	0	1391	3898
Reg	Regression	VALIDATE	GAME_RATING		0	0	597	1672
Tree	Decision Tree	TRAIN	GAME_RATING		40	89	1302	3858
Tree	Decision Tree	VALIDATE	GAME_RATING		17	32	565	1655
Neural	Neural Network	TRAIN	GAME_RATING		65	88	1303	3833
Neural	Neural Network	VALIDATE	GAME_RATING		35	38	559	1637
Neural2	Neural Network (2)	TRAIN	GAME_RATING		67	82	1309	3831
Neural2	Neural Network (2)	VALIDATE	GAME_RATING		32	42	555	1640

Figure 8. Event Classification Table

	ASE	RMSE	Misclassification Rate	ROC Index	Precision	Sensitivity
Log Regression	0.191041	0.437082	0.263112	0.586	0.7369	**1**
Decision Tree	0.180096	-	**0.256501**	0.647	0.7455	0.9898
Neural Network	**0.17664**	**0.42029**	0.258704	**0.687**	**0.7472**	0.9791
Neural Network base on Decision Tree.	0.177604	0.421431	0.261789	0.685	0.7454	0.9809

Table 3. Summary of Fit Statistics Comparison

TEXT ANALYTICS

As the diagram in Figure 9 shown below, in the first text parsing node I chose to ignore the parts of speech in abbr, aux, conj, det, interj, num, part, pref, prep, pron, and prop, with Max SVD Dimension to 50 and cluster algorithm to Expectation-Maximization. In the second text parsing node, I changed the cluster algorithm to Hierarchical with other parameters in default setting. Singular Value Decomposition (SVD) is used to reduce dimensionality by converting the term frequency matrix into a lower dimensional form. Smaller values of k (2 to 50) are thought to generate better results for text clustering using short text (OSU BAN 5743 Tutorial).

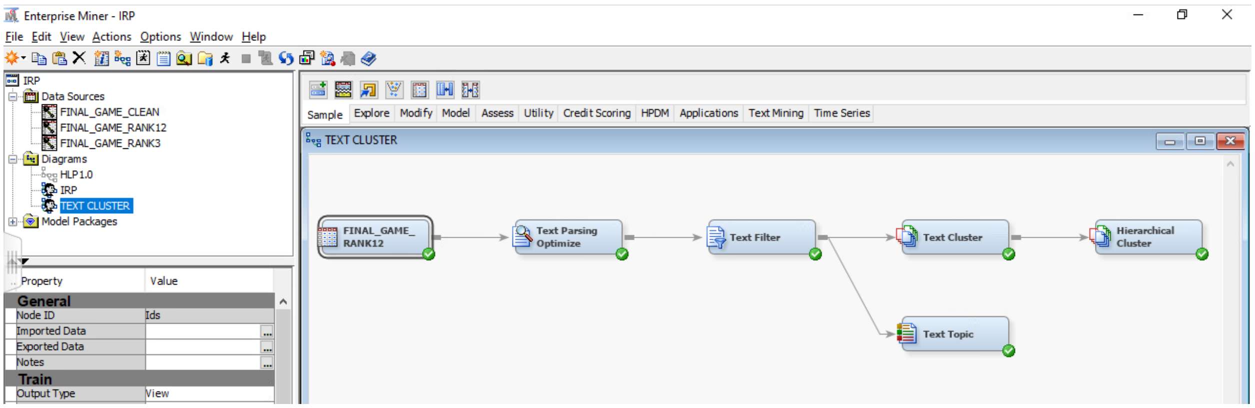


Figure 9. The Workflow Diagram of Text Cluster Analysis and Topic Analysis in SAS Enterprise Miner.

INSIGHT FROM TEXT CLUSTER ANALYSIS

Cluster analysis could provide big idea to what kind of game's distribution is popular. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar amongst themselves and dissimilar compared to objects of other groups (Abbas, 2018). Cluster analysis is the generic name for a wide variety of procedures that can be used to create a classification of entities/objects. The expectation-maximization algorithm is an approach for performing maximum likelihood estimation in the presence of latent variables. It does this by first estimating the values for the latent variables, then optimizing the model, then repeating these two steps until convergence (Brownlee, 2019). Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence. In contrast, hierarchical algorithms combine or divide existing groups, creating a hierarchical structure that reflects the order in which groups are merged or divided (Abbas, 2018). Comparison is shown in Table 4. It's interesting that upgrade and game player social, and free charge are main differences among those two target groups.

Strategy Mobile Game with Rating below 4	**Strategy Mobile Game with Rating above 4**
Expectation-Maximization algorithm	
Business, Management, Money, Worktime, Build (22%)	Battle, Strategy, World, Enemy, Upgrade, Unique (46%)
Challenge, Puzzle, Score Addictive, Simple (20%)	Game play, Level, Challenge, Time, Puzzle, Friend Game (41%)
Battle, Strategy, War, Army, Unit (14%)	Purchase, Manage, End money, Charge, Upgrade Free (13%)
Player, Board, Move, Opponent, Piece (11%)	
Hierarchical algorithm	
Play Games, Board, Move, Score (22%)	Game Friends, Challenge Players, Board Challenging (40%)
Addictive, Money, fun, simple (20%)	Battle, Strategy, World, Fight, Upgrade, Unique (36%)
Build, Fight Powerful Resources, and Conquer Epic (19%)	Money, Free, Earn Purchases Coins, Unlock (23%)
Enemies, Defense, Attract, Weapons, Destroy (12%)	

Table 4. Matrix Comparison of Top Descriptive Terms in terms of Expectation-Maximization Cluster Algorithm and Hierarchical Cluster Algorithm for Game Rating below Four Stars verses Game Rating above Four Stars

INSIGHT FROM TEXT TOPIC ANALYSIS

Topic modeling is a collective term for a family of computational algorithms that are used to model text in a collection of documents as arising from a much smaller set of topics (Isoaho, et al., 2019). Topic analysis is a Natural Language Processing (NLP) technique that allows us to automatically extract meaning from texts by identifying recurrent themes or topics. It's easy to find patterns and unlock semantic structures within texts to extract insights and help make data-driven decisions (MonkeyLearn). I my study, I generated 10 topics for each target group.

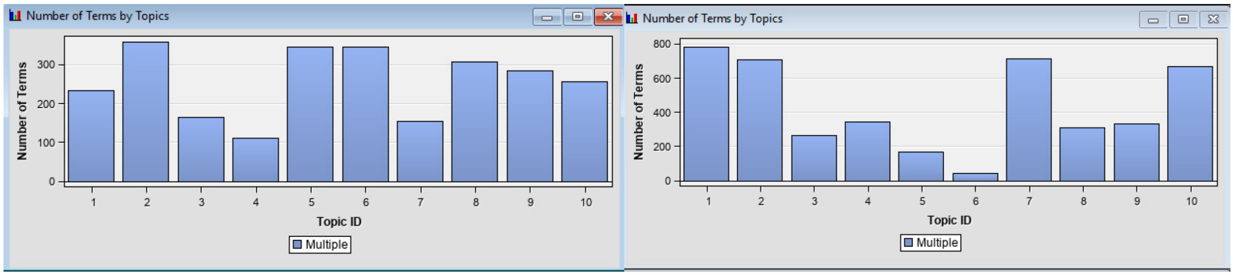


Figure 10. The Frequency Distribution for Number of Terms of 10 Topics (Target =0: Left, Target = 1: Right)

Due to the unbalance sampling, the first group of target has 5,570 sample sizes and the second group has 1,888 sample sizes. So the topic frequency of the second group is much higher than the first group. The frequency of topics in Table 5 could be used for vertical comparison within the group. The results showed that mobile strategy games with higher ratings are tend to card games and kitchen games. Survival games are not as so much popular, but war games have a strong advantage in both groups. It is worth to mention that in the attacking-tower game, the high score game uniformly appeared the word of zombie.

Strategy Mobile Game with Rating below 4	Strategy Mobile Game with Rating above 4
8 Significant Topics	
Battle, Mission, Campaign, Tank (359)	Board , Chess, Card (779)
Tower, Defense (345)	customer, Cook, Restaurant, Serve (711)
Business, Customer, Money, Management (344)	Battle, Army, Hero, War (707)
Hero, Dragon, Tower (308)	Tower, Defense, Zombie (666)
Puzzle, Room, Escape (283)	Evolution, Clicker, Mutation (344)
Card, Deck (255)	Puzzle, Door, Floor (311)
Chess, Card, Board (233)	Subscription, Current, Renewal (265)
Survival, Simulator, Animal, Wild (164)	Scene, collector, scramble (170)

Table 5. Comparison of Descriptive Terms in terms of Text Topic for Game Rating below Four Stars verses Game Rating above Four Stars

CONCLUSION

This study built four predictive models to predict strategy mobile game's rating, whether it has 4 and 4 above rating score. Neural Network is the winner models with sensitivity of predicting game rating of 97.91% and total 25.87% misclassification rate. In decision tree, USER_RATING_COUNT (100% importance) and SIZE (91.62%) are two most important variables to predict rating. However, in bivariate analysis, AGE_RATING is significantly important variables for predicting AVERAGE_USER_RATING. PRICE has no significance to predicting AVERAGE_USER_RATING. Based on the results of text analysis, I think there are several aspects that define a successful game. The first one is promotion of users. The more people review the game, the higher the rating would be. Secondly, the effect of price on a good game in this study is small, but players will be attracted by terms such as free upgrades in the description. So game developers could consider pricing as appropriate. Thirdly, a game having more game languages would be good potential power to make better game. Fourthly, there are a lot of strategy war games that make people tired. Therefore card games, aesthetic games, zombie games, and kitchen games has potential advantage in the market. Finally, good games create social circles that allow players to make friends within the game instead of acting alone.

REFERENCES

Abbas, O. A. 2008. Comparisons between data clustering algorithms. *International Arab Journal of Information Technology (IAJIT)*, *5*(3).

Ardhana, Prasetya. 2019. "Ardhana App Store Strategy Game EDA." RPubs. Accessed August 12, 2019. https://www.rpubs.com/chinesefoooood/MobileStratGames

Brownlee, Jason. 2019. "A Gentle Introduction to Expectation-Maximization (EM Algorithm)." Machine Learning Mastery. Accessed Nov. 1, 2019. https://machinelearningmastery.com/expectation-maximization-em-algorithm/

Clement J. 2021. "Gaming revenue worldwide 2020, by segment." Statista. Accessed Mar. 17, 2021. https://www.statista.com/statistics/292751/mobile-gaming-revenue-worldwide-device/

Hill, Simon. 2014. "Games rule the iTunes App Store: Most popular genres revealed." Gamasutra. Accessed Dec. 16, 2014. https://www.gamasutra.com/blogs/SimonHill/20141216/232458/Games_rule_the_iTunes_App_Store_Most_popular_genres_revealed.php

Isoaho, K., Gritsenko, D., & Mäkelä, E. 2021. "Topic modeling and text analysis for qualitative policy research." *Policy Studies Journal*, *49*(1), 300-324.

"Mobile Game Analysis." Github. Accessed Oct. 11, 2019. https://github.com/hkhoi/Mobile_Game_Analysis

"Topic Analysis." N.D. https://monkeylearn.com/topic-analysis/

Valens, Ana. 2019. "Report: Mobile gaming makes up nearly 75 percent of global mobile revenue during first half of 2019." GameDaily.biz. Accessed July 9, 2019. https://gamedaily.biz/article/1020/report-mobile-gaming-makes-up-nearly-75-percent-of-global-mobile-revenue-during-first-half-of-2019

ACKNOWLEDGMENTS

Thank the supports and advices from Mr. Gao, Dr. Chakraborty, Dr. McGaugh and Dr. Chabinak.

RECOMMENDED READING

- *JMP® Statistical Discovery™ Basic Analysis*
- *SAS® Enterprise Miner™*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chun Du
Oklahoma State University
405 762 2391
chund@okstate.edu
www.linkedin.com/in/chundu