Analysis of the Popular Video Game Sales features

Ki Min Lee

April 7th, 2020

I. Introduction

The video game industry has experienced dramatic changes over time. Currently a number of major players including Sony, Nintendo, EA, Microsoft and etc. dominate the revenue of the game industry. This study shows the trend of the game sales over time and discusses descriptive models for game sales of global, north american and japanese game market in an attempt to provide information for the video game sales business.

II. Data acquisition

The raw data can be found in Kaggle dataset (link: https://www.kaggle.com/gregorut/videogamesales/data#). The dataset contains the fields include ranking of overall sales (top 16600), the game name, platform of the games release, release year, game genre, game publisher, sales in North America, Europe, Japan, the rest of the world in millions, and total worldwide sales.
The data does not include enough number of observations after the year of 2017.

III. Descriptive Statistics

The original dataset has 16,598 observations.

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16598.000000	16327.000000	16598.000000	16598.000000	16598.000000	16598.000000	16598.000000
mean	8300.605254	2006.406443	0.264667	0.146652	0.077782	0.048063	0.537441
std	4791.853933	5.828981	0.816683	0.505351	0.309291	0.188588	1.555028
min	1.000000	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	4151.250000	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	8300.500000	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	12449.750000	2010.000000	0.240000	0.110000	0.040000	0.040000	0.470000

max	16600.000000	2020.000000	41.490000	29.020000	10.220000	10.570000	82.740000

Figure 1. Descriptive Statistics Table

The descriptive statistics of the raw data shows that sales in North America counts the most in the market share, therefore it is important to know the market trend of North America for the game companies to succeed.

IV. Data Processing

As the main focus of this study is to describe the recent game market to provide insight for video game sales, the data prior to the year 2011 is opted out. As the raw data does not contain enough records for analysis after the year 2016, only data between the year 2011 to 2015 are taken out. 3538 observations from the raw data are obtained after the extraction for years.

Also, the data was extracted for the five most popular game genres including Sports, Action, Shooter, Role-Playing, and Platform only.

In addition, dummy variables for descriptive modeling were added based on the exploratory data analysis.

V. Exploratory Data Analysis

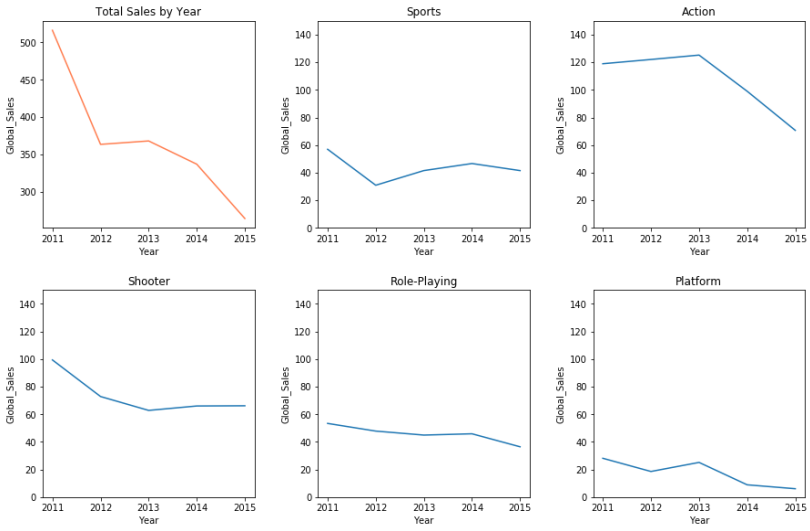


Figure 2. Global Sales by Genre and Year

In order to concentrate on features shared among games with high sales, the data is extracted for year 2011 and 2015, and five game genres including Sports, Action, Shooter, Role-Playing, and Platform.

As the total game sales significantly declined between 2011 and 2012 and continues to drop until 2015, sales of all game genres decreased throughout the years but sports.

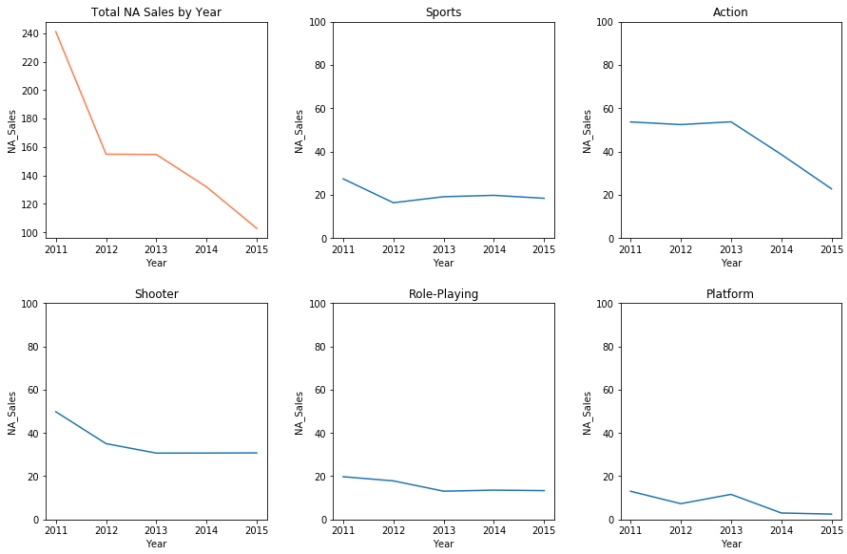


Figure 3. NA Sales by Genre and Year

While Action and Platform game sales experienced a significant drop, other genres are relatively staying flat in sale over the years.

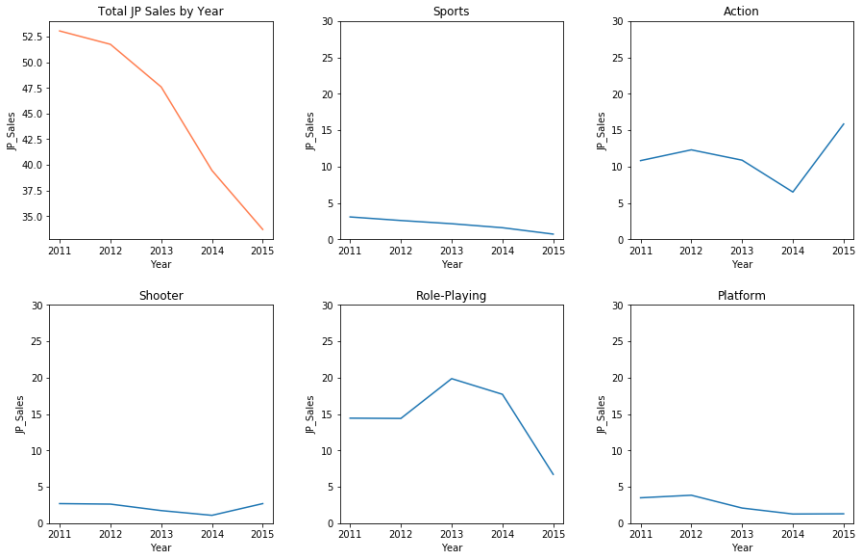


Figure 4. JP Sales by Genre and Year

In Japanese game market, Role-Playing has been the most popular game genre for over all time. However, the recent sales data shows that it is not as popular as before 2014 and sales of action games is replacing the place.

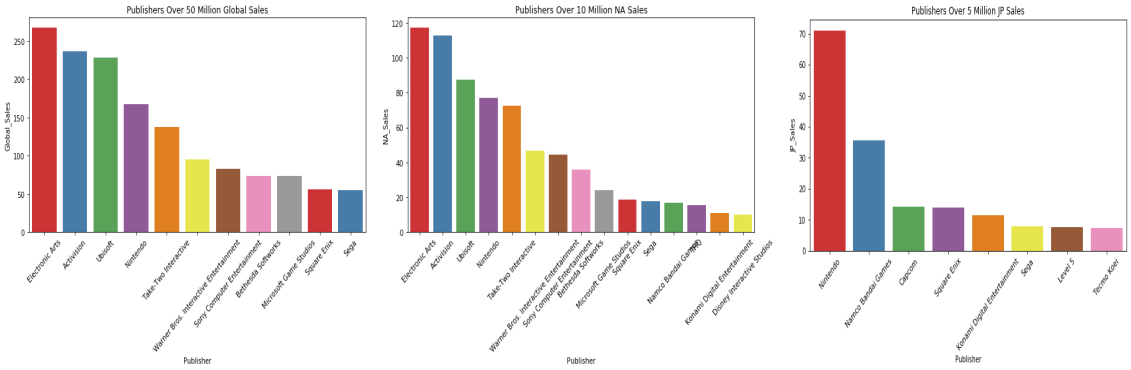
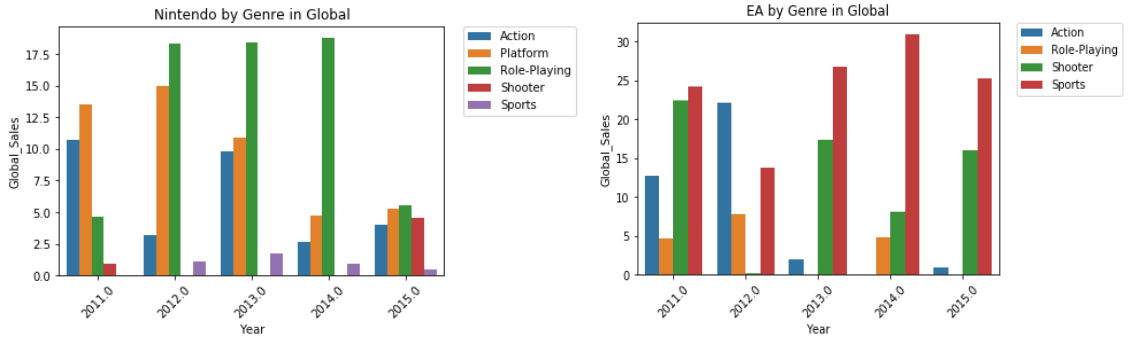


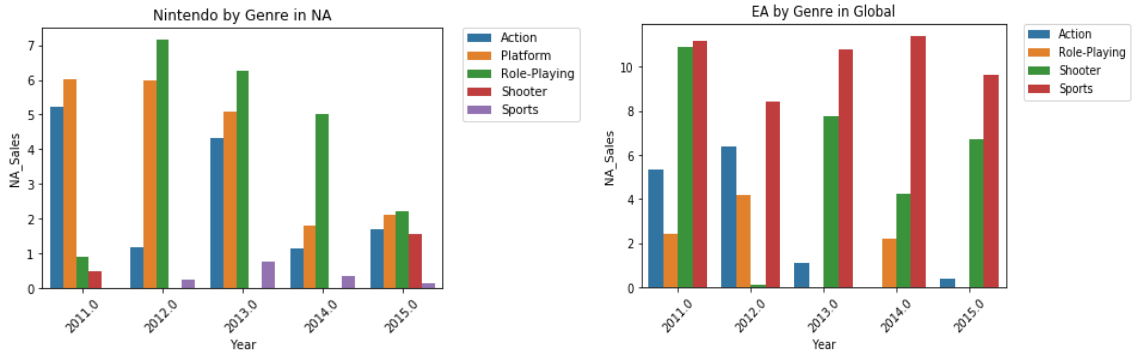
Figure 5. JP Sales by Genre and Year

As the graph shows in the above, Electronic Arts and Activision are dominating the market followed by Ubisoft and Nintendo. In contrast, Nintendo monopolizes the video game market of Japan.

Sales in Global



Sales in North America



Sales in Japan

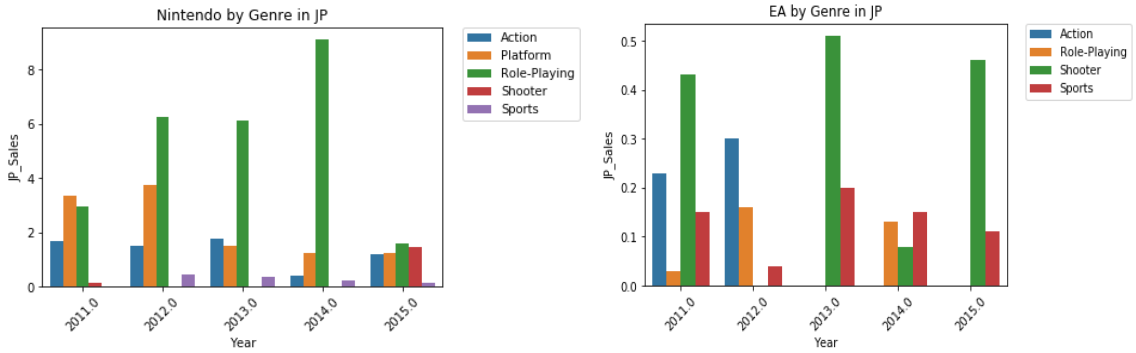


Figure 6. Nintendo vs. Electronic Arts, Sales by Genre over years

As it's shown in the figure 6, The top three genres of game sales for Nintendo are Role-Playing, Platform, and Action games. On the other hand, the most Electronic Arts (EA) game sales are from Sports, Shooter, and Action. The japanese game market behaves differently from other markets. Nintendo sold Role-Playing games twice more than Platform games in Japan. In contrast, EA has most of the game sales from the Shooter genre and also, the absolute number of EA game sales is significantly lower than Nintendo in Japan.

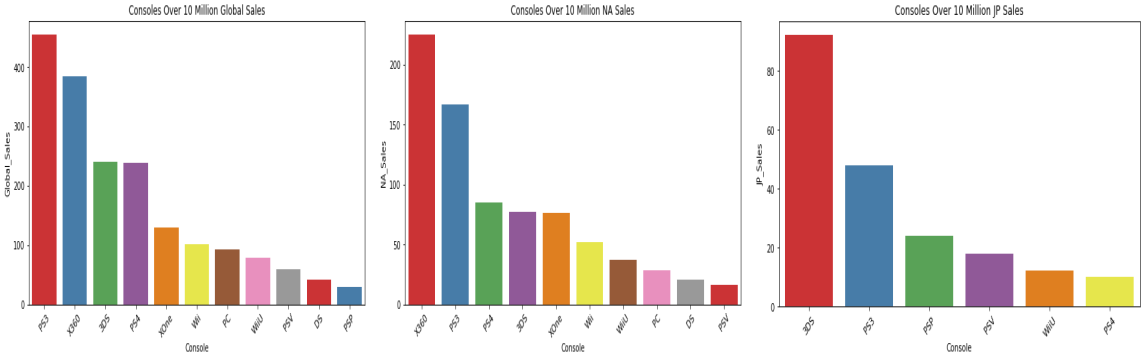


Figure 7. Game Sales by Consoles in Global, NA, and JP

In Japan, games published in 3DS records the highest sales. XBox 360 is the most popular platform with top game sales in North America and Playstation 3 has the top game sales in the global game market.

VI. Descriptive Modeling

Two descriptive methods, logistic regression and Ordinary Least Squares regression (OLS) were employed to construct models in an attempt to describe the global, north america, and japan game market.

For the logistic models, dummy variables are created for the games having the cumulative sales higher than mean values between 2011 to 2015 (Global sales > 0.523, NA sales > 0.222, JP sales > 0.064).

1. Global Sales Model

Based on the information from the exploratory data analysis, EA, Shooter, and PS3 dummy variables are included for the global market model.

OLS Regression Results

Dep. Variable:	Global_Sales	R-squared (uncentered):	0.181
Model:	OLS	Adj. R-squared (uncentered):	0.180
Method:	Least Squares	F-statistic:	137.0
Date:	Tue, 07 Apr 2020	Prob (F-statistic):	7.08e-106
Time:	00:50:13	Log-Likelihood:	-4239.1
No. Observations:	2476	AIC:	8486.
Df Residuals:	2472	BIC:	8509.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Shooter	0.7441	0.097	7.663	0.000	0.554	0.934
PS3	0.1880	0.069	2.733	0.006	0.053	0.323
EA	0.6631	0.106	6.272	0.000	0.456	0.870
Year	0.0002	1.57e-05	13.080	0.000	0.000	0.000

Omnibus:	3073.929	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	423549.533
Skew:	6.676	Prob(JB):	0.00
Kurtosis:	65.667	Cond. No.	8.03e+03

Figure 7. Global Sales OLS Model Report: Mean Absolute Error: 0.5028919325862423, Mean Squared Error: 0.9290539170663538, Root Mean Squared Error: 0.9638744301341092

Logit Regression Results

Dep. Variable:	GlobSales_High	No. Observations:	2476
Model:	Logit	Df Residuals:	2472
Method:	MLE	Df Model:	3
Date:	Tue, 07 Apr 2020	Pseudo R-squ.:	0.07105
Time:	00:53:50	Log-Likelihood:	-1244.3
converged:	True	LL-Null:	-1339.5
Covariance Type:	nonrobust	LLR p-value:	5.117e-41

	coef	std err	z	P>|z|	[0.025	0.975]
Shooter	1.1163	0.153	7.285	0.000	0.816	1.417
PS3	0.6578	0.116	5.655	0.000	0.430	0.886
EA	1.5851	0.166	9.571	0.000	1.260	1.910
intercept	-1.6097	0.062	-25.804	0.000	-1.732	-1.487

train data confusion matrix:

predicted	0	1
actual		
0	1860	43
1	498	75

AIC: 2496.6203118830117

BIC: 2519.8779104182295

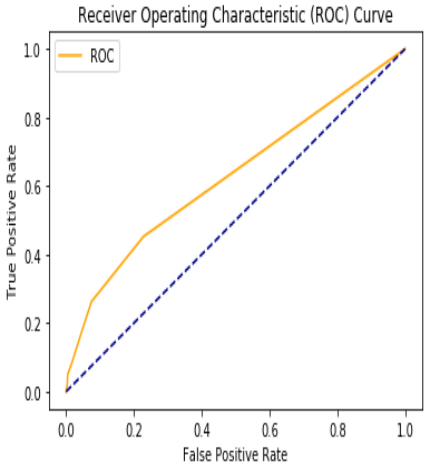


Figure 8. Global Sales Logit Model Report and ROC curve: AUC score: 0.626223360468307

Based on the information from the exploratory data analysis, EA, Shooter, and PS3 dummy variables are included for the global market model. The R-squared value of the OLS model is 0.18 which tells that the model can explain only 18% of the sample. It also has a very high error rate.

In the logistic model, the very high AIC and BIC scores and very low R-squared indicate that the model is not robust to describe the games with sales over sales mean.

2. NA Sales Model

OLS Regression Results

Dep. Variable:	NA_Sales	R-squared (uncentered):	0.180
Model:	OLS	Adj. R-squared (uncentered):	0.179
Method:	Least Squares	F-statistic:	136.1
Date:	Tue, 07 Apr 2020	Prob (F-statistic):	3.15e-105
Time:	01:03:55	Log-Likelihood:	-2297.2
No. Observations:	2476	AIC:	4602.
Df Residuals:	2472	BIC:	4626.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Shooter	0.3713	0.045	8.334	0.000	0.284	0.459
X360	0.2375	0.036	6.678	0.000	0.168	0.307
EA	0.2491	0.048	5.150	0.000	0.154	0.344
Year	7.691e-05	6.92e-06	11.109	0.000	6.33e-05	9.05e-05

Omnibus:	3051.730	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	411829.405
Skew:	6.594	Prob(JB):	0.00
Kurtosis:	64.790	Cond. No.	8.03e+03

Figure 9.NA Sales OLS Model Report: Mean Absolute Error: 0.22917447944088568, Mean Squared Error: 0.2265415900666995,Root Mean Squared Error: 0.4759638537396506

Logit Regression Results

Dep. Variable:	NASales_High	No. Observations:	2476
Model:	Logit	Df Residuals:	2472
Method:	MLE	Df Model:	3
Date:	Tue, 07 Apr 2020	Pseudo R-squ.:	0.07575
Time:	01:04:16	Log-Likelihood:	-1226.8
converged:	True	LL-Null:	-1327.4
Covariance Type:	nonrobust	LLR p-value:	2.441e-43

	coef	std err	z	P>|z|	[0.025	0.975]
Shooter	1.0448	0.155	6.744	0.000	0.741	1.348
X360	0.9520	0.126	7.532	0.000	0.704	1.200
EA	1.4346	0.166	8.629	0.000	1.109	1.760
intercept	-1.6337	0.061	-26.798	0.000	-1.753	-1.514

train data confusion matrix:

predicted	0	1
actual		
0	1866	47
1	494	69

AIC: 2461.640981203239

BIC: 2484.8985797384566

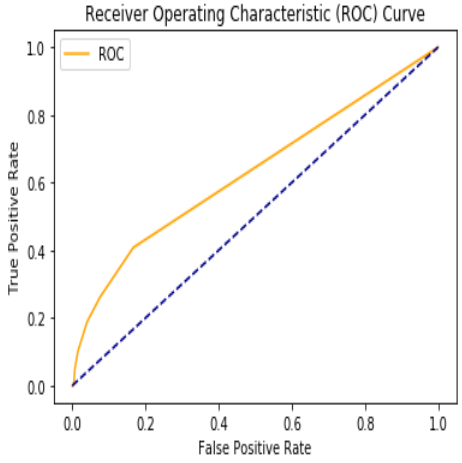


Figure 10. NASales Logit Model Report and ROC curve: AUC score: 0.6293507326405495

The NA Sales model includes Shooter, XBox 360, and EA variables. The NA models show very similar results to the global sales models with a slightly lower error rate of approximately 0.48.

3. JP Sales Model

OLS Regression Results

Dep. Variable:	JP_Sales	R-squared (uncentered):	0.271
Model:	OLS	Adj. R-squared (uncentered):	0.270
Method:	Least Squares	F-statistic:	229.6
Date:	Tue, 07 Apr 2020	Prob (F-statistic):	8.78e-168
Time:	00:50:12	Log-Likelihood:	257.23
No. Observations:	2476	AIC:	-506.5
Df Residuals:	2472	BIC:	-483.2
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
RPG	0.1264	0.014	9.323	0.000	0.100	0.153
Nintendo	0.5027	0.025	20.282	0.000	0.454	0.551
3DS	0.1053	0.013	7.828	0.000	0.079	0.132
Year	9.762e-06	2.46e-06	3.963	0.000	4.93e-06	1.46e-05

Omnibus:	3710.192	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1565003.155
Skew:	9.051	Prob(JB):	0.00
Kurtosis:	124.828	Cond. No.	1.15e+04

Figure 11.JP Sales OLS Model Report: Mean Absolute Error: 0.22917447944088568, Mean Squared Error: 0.2265415900666995, Root Mean Squared Error: 0.4759638537396506

Logit Regression Results

Dep. Variable:	JPSales_High	No. Observations:	2476
Model:	Logit	Df Residuals:	2472
Method:	MLE	Df Model:	3
Date:	Tue, 07 Apr 2020	Pseudo R-squ.:	0.1263
Time:	01:21:37	Log-Likelihood:	-1091.9
converged:	True	LL-Null:	-1249.7
Covariance Type:	nonrobust	LLR p-value:	3.902e-68

	coef	std err	z	P>|z|	[0.025	0.975]
RPG	1.4494	0.137	10.592	0.000	1.181	1.718
Nintendo	2.7697	0.301	9.203	0.000	2.180	3.360
3DS	1.0167	0.141	7.208	0.000	0.740	1.293
intercept	-1.8946	0.067	-28.315	0.000	-2.026	-1.763

train data confusion matrix:

predicted	0	1
actual		
0	1941	32
1	408	95

AIC: 2191.7279015609074
BIC: 2214.985500096125

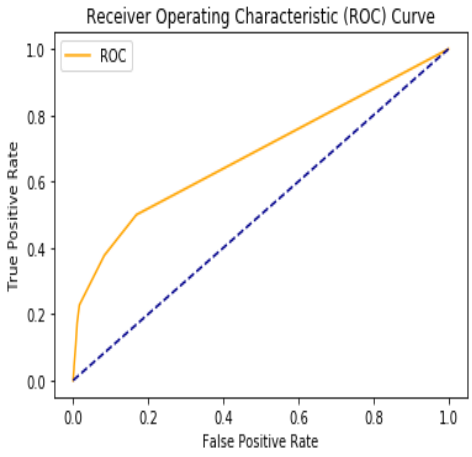


Figure 12. JP Sales Logit Model Report and ROC curve: AUC score: 0.6832163679550852

Independent variables including 3DS, RPG, and Nintendo are incorporated into the model for the JP market.

The models perform better as the OLS model has higher R-squared value with 0.271 although the error rate is not very improved compared to the models for the NA and global market. Meanwhile, the logit model for the JP market has lower AIC and BIC scores as well as higher AUC value. This tells that the models for JP Sales explain the market better in comparison to the models for North America and global game sales.

VII. Conclusion and Future Directions

The objective of this study is to determine the features shared among games with high sales.

The exploratory data analysis in this study revealed some information of the game sales. For example, in North America, EA and Activision are the two most popular game publishers and Action and Shooter games are sold much higher than other types of games. Also, XBox 360 is slightly more popular in North America while PS3 games are sold better in global sales. Japanese market shows unique features as Nintendo is almost monopolizing the market with Role-Playing games and 3DS is the most popular console in the market. However, given the raw data includes a limited number of observations and variables, the models created are not robust to perform the task, although the models for the Japanese game sales performs slightly better than models for other regions.

The models in this study interact only with publisher, year, console, and genre. However, there are many other features such as studio, promotion, and review score. The models of this study miss lots of information which may have led to large error rates, but if more information is given, could produce notable improvements to the models.