STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Data Mining of Video Game Sales

Xin Lin, Fanshu Li, Jingmiao Shen
Instructor: Yifan Hu

Content

1

Introduction

Video Games

“

Background

- Video game industry shows a tremendous growth in recent years
- The video games market in the United States, worth over 20 billion annually *(Charfield, 2010)*

Project Goal

- Predicting video game global sales
- More than one million or not

Data Source & Pre-processing

Data Source

★ *VG Chartz*

★ *Metacritic*

Data Interpretation

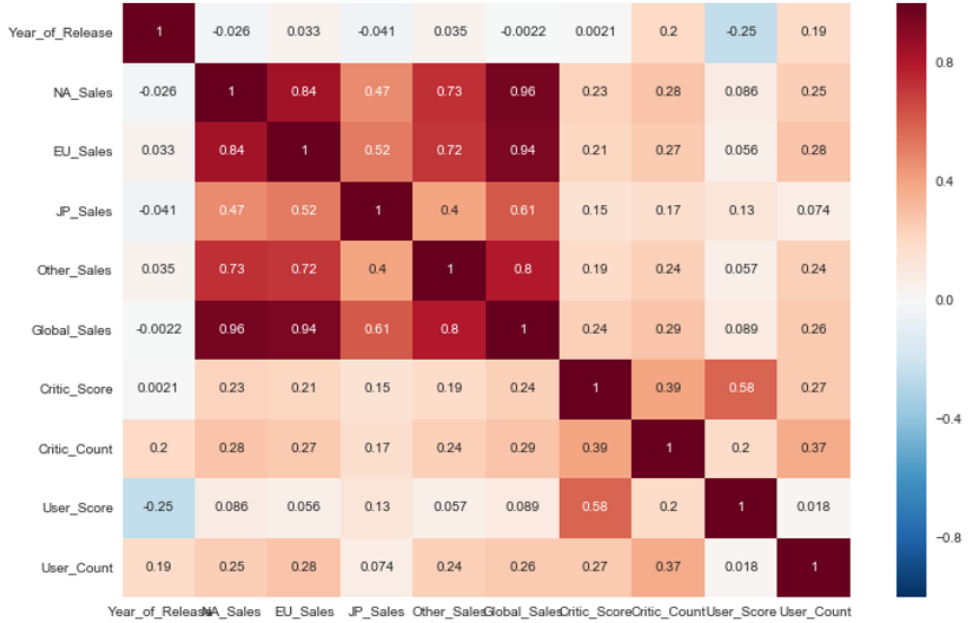
Variables	Explanation
Names	Lists of video games name
Platform	Lists of platforms of each video game such as GB, Wii
Year_of_Release	Year that video games are released (1980 – 2017)
Genre	Lists of genre for each video game such as sports, racing
NA_Sales; EU_Sales, JP_Sales Other_Sales, Global_Sales	Sales in North America, Europe, Japan, in the rest of the world, and Total worldwide sales (in millions)
Critic_Score	Score based on critic reviews (0-100)
Critic_Count	Number of critic reviews
User_Score	Score based on user reviews (0-10)
User_Count	Number of user reviews
Developer & Publisher	Company builds a product from design to implement ; Company sells product and getting returns
Rating	Age appropriateness suggested by The Entertainment Software Rating Board (ESRB) rating

Data Cleaning

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count	Developer	Rating
0	Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	8	322.0	Nintendo	E
1	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	NaN	NaN	NaN	NaN
2	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0	8.3	709.0	Nintendo	E
3	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	8	192.0	Nintendo	E
4	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	NaN	NaN	NaN	NaN

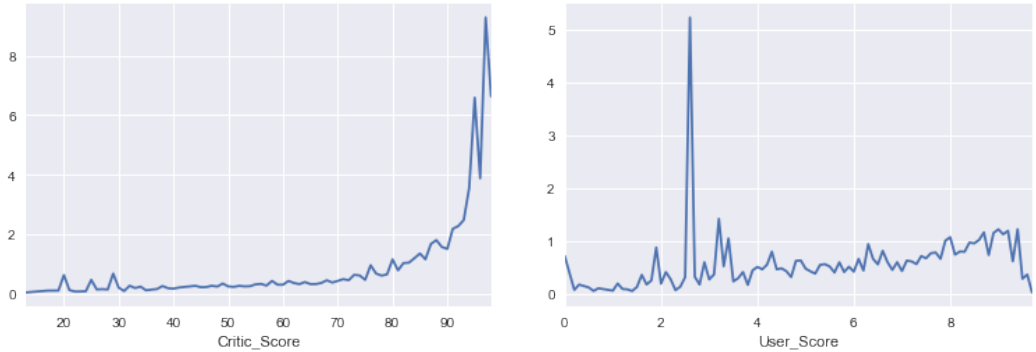
- Categories AO (adult only), EC (early childhood) and RP (rating pending) in rating variable are so rare that we decided to drop them
- Convert type of Year_of_Release from “object” to “Int” and drop NaN values

Feature Selection



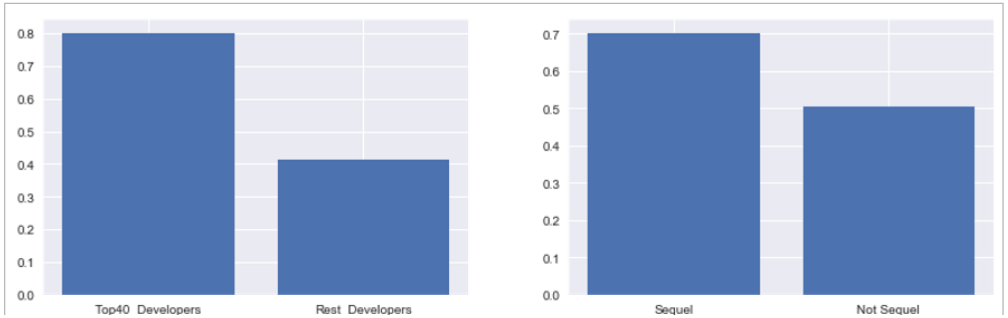
Coefficient Matrix of Sales

Feature Selection



Critic_Score shows high correlation with Average Sales while User_Score seems uncorrelated with Average Sales

Feature Selection



Top 40 Developer's game average sales is almost twice over the rest of developer's sales.

Sequel would influence the sale of a specific video game. Add this new feature for our prediction.

Features

	Platform	Year_of_Release	Genre	Publisher	Critic_Score	Critic_Count	Developer	Rating	Sequel	Million	Top_Developer
0	Wii	2006	Sports	Nintendo	76.0	51.0	Nintendo	E	0	1	1
2	Wii	2008	Racing	Nintendo	82.0	73.0	Nintendo	E	0	1	1
3	Wii	2009	Sports	Nintendo	80.0	73.0	Nintendo	E	1	1	1
6	DS	2006	Platform	Nintendo	89.0	65.0	Nintendo	E	0	1	1
7	Wii	2006	Misc	Nintendo	58.0	41.0	Nintendo	E	0	1	1
8	Wii	2009	Platform	Nintendo	87.0	80.0	Nintendo	E	0	1	1
10	DS	2005	Simulation	Nintendo	NaN	NaN	NaN	NaN	0	1	0
11	DS	2005	Racing	Nintendo	91.0	64.0	Nintendo	E	0	1	1
13	Wii	2007	Sports	Nintendo	80.0	63.0	Nintendo	E	0	1	1
14	X360	2010	Misc	Microsoft Game Studios	61.0	45.0	Good Science Studio	E	0	1	0
15	Wii	2009	Sports	Nintendo	80.0	33.0	Nintendo	E	1	1	1

3

Data Analysis

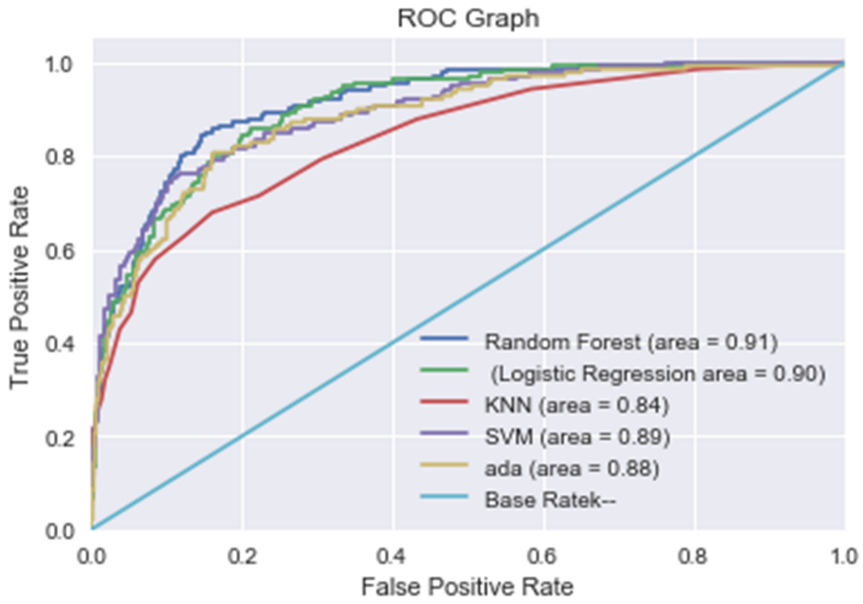
Classification Models

Classification Model

- Random Forest Classifier
- Logistic Regression
- K-Nearest Neighbor
- Support Vector Machine
- AdaBoost
- Base Rate

90% as training data
10% as testing data

Results Comparison



Best performance

Random Forest (0.91)

4

Conclusion & Discussion

Real World Impact

Conclusion

Random Forest model to fit our new data set

Randomly pick some video games to conduct prediction

Name	Platform	Year_of_Release	Genre	Publisher	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count
Horizon: Zero Dawn	PS4	2017	Action	Sony Computer Entertainment	3.48	89	115	8.3	5100
Crash Bandicoot N. Sane Trilogy	PS4	2017	Platform	Activision	2.22	80	83	8.8	711
Injustice 2	XOne	2017	Fighting	Warner Bros. Interactive Entertainment	0.46	89	23	7.6	110
Halo Wars 2	XOne	2017	Strategy	Microsoft Game Studios	0.33	79	79	5.8	393
LEGO Worlds	XOne	2017	Misc	Warner Bros. Interactive Entertainment	0.29	69	20	6.3	25
Sniper Elite 4	PS4	2017	Shooter	Rebellion Developments	0.35	77	55	6.8	119
Halo Wars 2	XOne	2017	Strategy	Microsoft Game Studios	0.33	79	79	5.8	393
The Legend of Zelda: Breath of the Wild	WiiU	2017	Action	Nintendo	1.14	96	13	8.1	1731
Tom Clancy's Ghost Recon Wildlands	XOne	2017	Shooter	Ubisoft	1.13	76	27	6.6	166

Testing Results

Name	Million	Predict
Horizon: Zero Dawn	1	1
Crash Bandicoot N. Sane Trilogy	1	1
Injustice 2	0	0
Halo Wars 2	0	0
LEGO Worlds	0	0
Sniper Elite 4	0	0
Halo Wars 2	0	0
The Legend of Zelda: Breath of the Wild	1	1
Tom Clancy's Ghost Recon Wildlands	1	1

Testing results show 100% prediction accuracy

Real World Impact

1. For game shop owner

2. For video game industry

3. For any product sale

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Thanks for listening.