Analyzing the Gaming Industry

From Pixels to Profits

PLAYALYTICS

Ang Mei Chi

Lew Kuan Teng Roy

Liu Wudi

Michael Wong Wait Kit

Ong Wee Yang

Business
Understanding

- Background Intro
- Project Objectives

Business Understanding

Business Understanding

Data Source Preprocessing

Modelling
Text Analytics

- Predictive Modelling
- Text Analytics

Business Understanding

Data Source Preprocessing

Modelling
Text Analytics

Model Evaluation

- Assessment
- Interpretability

Business Understanding

Data Source Preprocessing

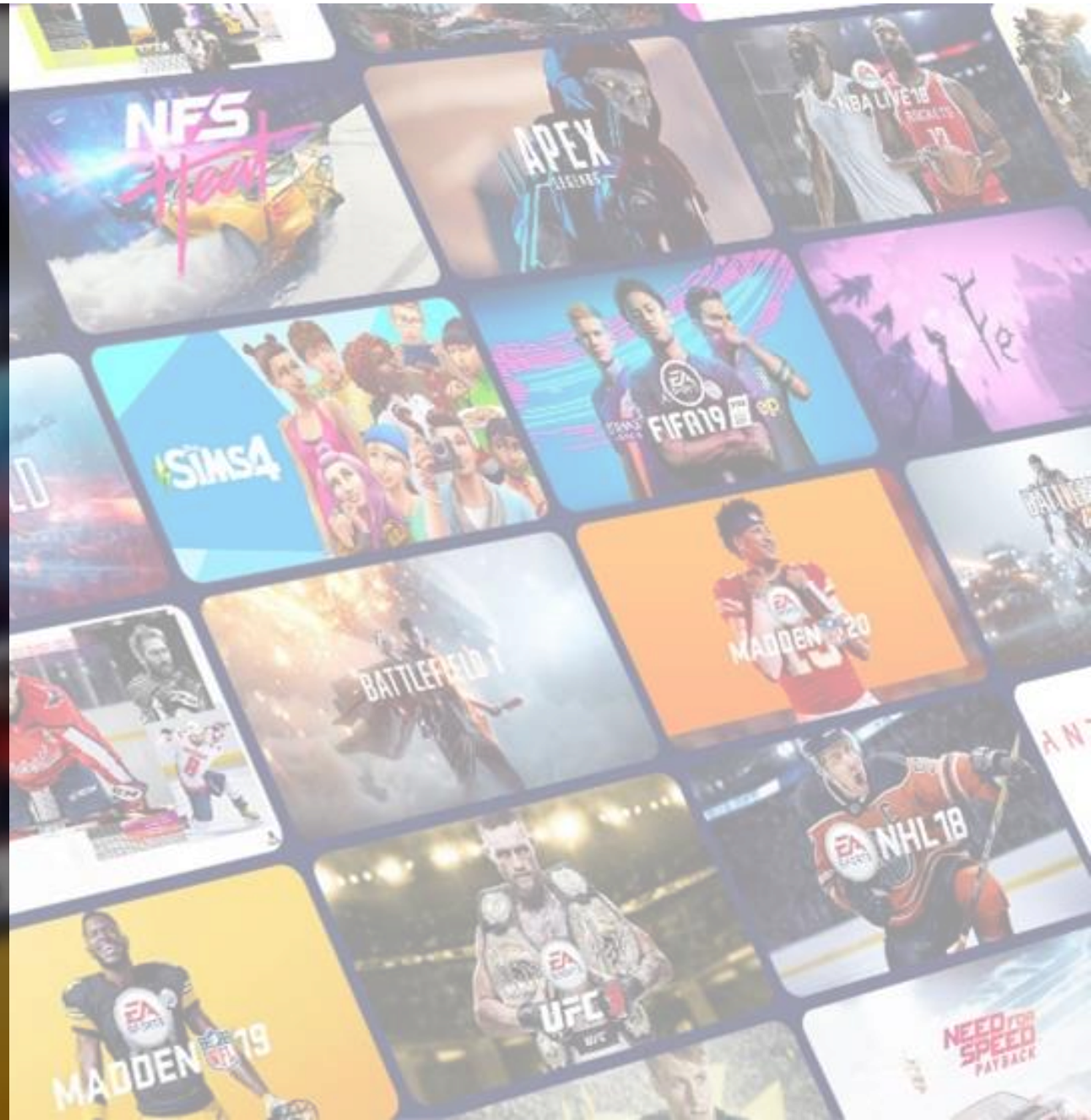
Modelling
Text Analytics

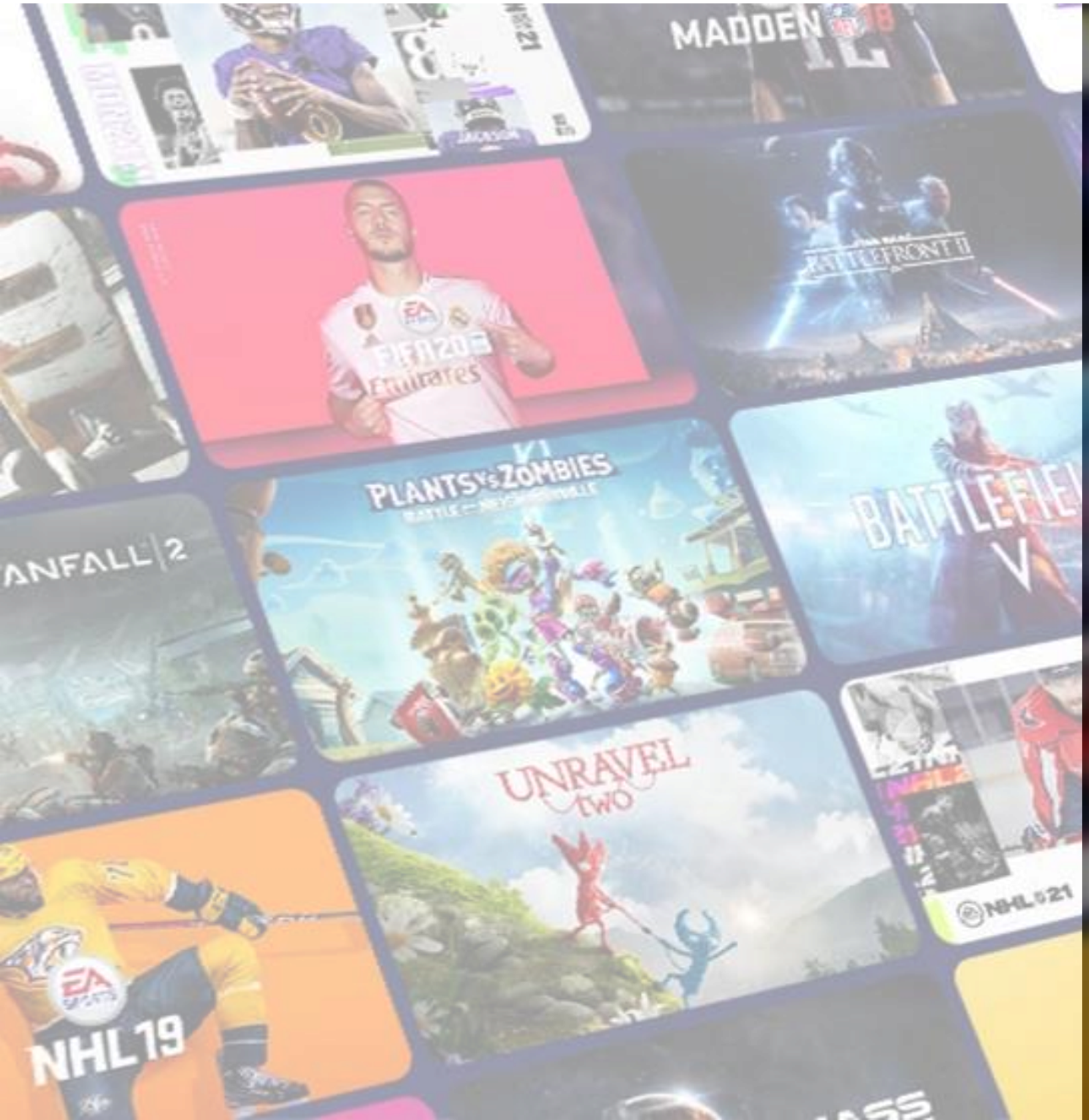
Model Evaluation

Retrospective

- Limitations
- Learning Points

Which **variable(s)** effects game sales the most?





Historical data

Guiding patterns and insights to better business decisions

Project Objectives

Key Revenue Indicators

Sales Revenue Prediction Model

Project Objectives

Project Objectives

- Key Revenue Indicators
- Improvements Game Features

Sales Revenue Prediction Model

More **informed** business decisions

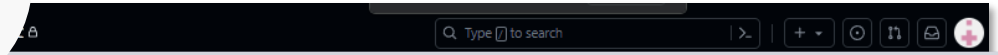
Text Analysis of Players' Reviews

Michael Wong push — fbe87df · 3 weeks ago — History

Preview | Code | Blame — 65 lines (65 loc) · 130 KB — Code 55% faster with GitHub Copilot — Raw

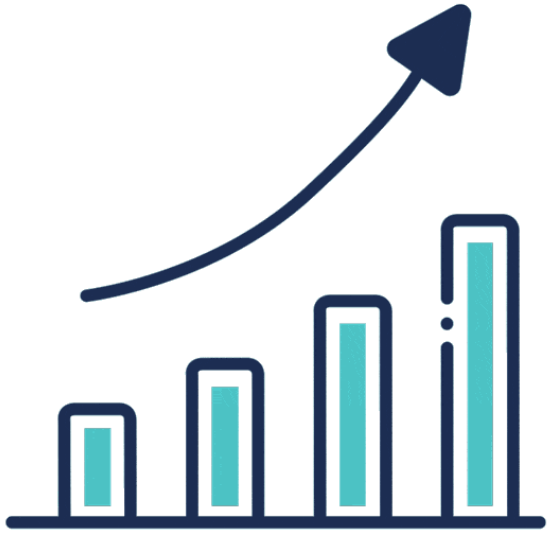
Search this file

	title	developer	publisher	tags
2	Spooky's Jump Scare Mansion	Lag Studios	Lag Studios	['Horror', 'Free to Play', 'Cute', 'First-Person', 'Singleplayer', 'Psychologica
3	Sakura Clicker	Winged Cloud	Winged Cloud	['Nudity', 'Anime', 'Free to Play', 'Mature', 'Sexual Content', 'Clicker', 'Fem
4	WARMODE	WARTEAM	WARTEAM	['Early Access', 'Free to Play', 'FPS', 'Multiplayer', 'Shooter', 'Action', 'First-
5	Fractured Space	Edge Case Games Ltd.	Edge Case Games Ltd.	['Space', 'Multiplayer', 'Free to Play', 'PvP', 'MOBA', 'Action', 'Strategy', 'Sc
6	Counter-Strike: Global Offensive	Valve, Hidden Path Entertainment	Valve	['FPS', 'Multiplayer', 'Shooter', 'Action', 'Team-Based', 'Competitive', 'Tacti
7	Dota 2	Valve	Valve	['Free to Play', 'MOBA', 'Strategy', 'Multiplayer', 'Team-Based', 'Action', 'e-
8	Path of Exile	Grinding Gear Games	Grinding Gear Games	['Free to Play', 'Action RPG', 'Hack and Slash', 'RPG', 'Multiplayer', 'Massiv
9	War Thunder	Gaijin Entertainment	Gaijin Entertainment	['Free to Play', 'World War II', 'Multiplayer', 'Simulation', 'Flight', 'War', 'Ta
10	Ring of Elysium	Aurora Studio	Tencent Games	['Early Access', 'Free to Play', 'Battle Royale', 'Shooter', 'Survival', 'Multipla
11	Team Fortress 2	Valve	Valve	['Free to Play', 'Multiplayer', 'FPS', 'Action', 'Shooter', 'Class-Based', 'Team
12	World of Warships	Wargaming Group Limited	Wargaming Group Limited	['Free to Play', 'Naval', 'World War II', 'Multiplayer', 'PvP', 'Action', 'Online
13	Star Trek Online	Cryptic Studios	Perfect World Entertainment	['Free to Play', 'Sci-fi', 'Massively Multiplayer', 'Space', 'RPG', 'MMORPG',
14	Yu-Gi-Oh! Duel Links	Konami Digital Entertainment	Konami Digital Entertainment	['Free to Play', 'Card Game', 'Anime', 'Multiplayer', 'Strategy', 'Trading Car
15	Paladins®	Evil Mojo Games	Hi-Rez Studios	['Free to Play', 'Multiplayer', 'FPS', 'Shooter', 'Action', 'Team-Based', 'First-
16	SMITE®	Titan Forge Games	Hi-Rez Studios	['Free to Play', 'MOBA', 'Multiplayer', 'Action', 'Third Person', 'Mythology'
17	Brawlhalla	Blue Mammoth Games	Ubisoft	['Free to Play', 'Multiplayer', 'Fighting', '2D', '2D Fighter', 'Action', 'Local M
18	World of Tanks Blitz	Wargaming Group Limited	Wargaming Group Limited	['Tanks', 'Free to Play', 'Multiplayer', 'Action', 'World War II', 'War', 'Massiv
19	Shadowverse CCG	Cygames, Inc.	Cygames, Inc., Cygames Korea Inc.	['Free to Play', 'Card Game', 'Anime', 'Strategy', 'Trading Card Game', 'Fan
20	DCS World Steam Edition	Eagle Dynamics SA	The Fighter Collection	['Simulation', 'Flight', 'Free to Play', 'Military', 'Multiplayer', 'Realistic', 'Wa
21	Heroes & Generals	RETO MOTO	RETO MOTO	['Free to Play', 'World War II', 'Multiplayer', 'FPS', 'War', 'Shooter', 'Action'
22	The Elder Scrolls®: Legends™	Sparkypants Studios, LLC	Bethesda Softworks	['Free to Play', 'Card Game', 'Multiplayer', 'Strategy', 'Trading Card Game'
23	Tree of Savior (English Ver.)	IMCGAMES Co.,Ltd.	IMCGAMES Co.,Ltd.	['MMORPG', 'Massively Multiplayer', 'Free to Play', 'RPG', 'Anime', 'Fantas
24	Trove	Trion Worlds	Trion Worlds	['Free to Play', 'Open World', 'Multiplayer', 'Adventure', 'Massively Multip
25	Neverwinter	Cryptic Studios	Perfect World Entertainment	['Free to Play', 'MMORPG', 'RPG', 'Massively Multiplayer', 'Adventure', 'Ac



	platform	genre	publisher	vg_score	critic_score	user_score	total_shipped	Production Cost	release_year	Developer_metac	Genre_metacritic_game_info	Metascore_metacritic_game_ir	Avg_Userscore_metacritic_game_	No. of players_metacritic_game_info	releasedate_qua
...t A.D.	PC	Simulation	Sony Computer Entertainment	8	7.7	5	1	6.36	2006	RelatedDesigns	Strategy;Real-Time;Historic;General;Historic	79	8.4	1 to 4	26/10/2006
2002 FIFA World Cup	PC	Sports	Sony Computer Entertainment	2	6	7	277	6.06	2002	EASports	Sports;Traditional;Team;Soccer;Sim;Sim	80	8	Mutliplayer	26/4/2002
2010 FIFA World Cup South Afric	PS3	Sports	Ubisoft	7	8.2	9	53	4.24	2010	EACanada	Sports;Traditional;Team;Soccer;Sim;Sim	82	7.8	1	27/4/2010
ABZU	PC	Adventure	Activision	9	8	9	0.58	3.75	2016	GiantSquid	Action Adventure;Adventure;General;3D;Third-Person	83	6.9	1	2/8/2016
ABZU	PS4	Adventure	Electronic Arts	8	7	7	117	3.24	2017	GiantSquid	Fantasy;Action Adventure;Adventure;General;3D;Third-Person	78	7.6	1	2/8/2016
Aggressive Inline	PS2	Sports	Activision	6	8.7	10	117	3.21	2002	Z-Axis,Ltd.	Sports;Alternative;Skating	85	8.7	1 to 2	28/5/2002
Alice in Wonderland	DS	Adventure	Electronic Arts	8	7	3	59	0.03	2010	EtrangesLibellules	Action;Action Adventure;General;General;Fantasy	78	tbd	1	2/3/2010
Alien Isolation	PS4	Adventure	Frontier Developments	10	4	8	189	2.61	2014	CreativeAssembly	Action Adventure;Sci-Fi;General;Survival	79	8.2	1	6/10/2014
Alien Hominid	PS2	Shooter	Frontier Developments	3	4	8	296	2.58	2004	TheBehemoth	Action;Shooter;Scrolling	78	8.3	1	21/11/2004
Aliens versus Predator 2	PC	Shooter	Sony Computer Entertainment	5	8.2	2	275	2.07	2001	MonolithProduction	Action;Shooter;Shooter;First-Person;Sci-Fi;Sci-Fi;Arcade	85	8.7	Mutliplayer	22/10/2001
Allegiance	PC	Simulation	Activision	1	10	10	231	2.81	2000	MicrosoftGameStuc	Simulation;Sci-Fi;Space;Small Spaceship;Small Spaceship;Combat	86	8.3	1	16/3/2000
Alone in the Dark The New Nightm	PS	Adventure	Electronic Arts	3	7.2	4	162	5.65	2001	Darkworks	Action Adventure;Horror	77	8.2	1	8/5/2001
Animal Crossing Wild World	DS	Simulation	Amazon Game Studios	6	8.5	4	11.75	5.09	2005	Nintendo	Simulation;Miscellaneous;Virtual Life;Virtual;Virtual Life	86	8.6	1 to 4	23/11/2005
Anno 1701 Dawn of Discovery	DS	Simulation	Ubisoft	9	8	5	221	4.05	2008	SunflowersInteractiv	Simulation;General;General	78	7.4	1 to 4	8/6/2007
Apollo Justice Ace Attorney	DS	Adventure	Ubisoft	7.9	8	10	23	5.95	2008	Capcom	Adventure;General;General;Visual Novel	78	8.2	1	12/4/2007
Aqua Aqua	PS2	Puzzle	Frontier Developments	10	4	2	239	4.23	2001	ZedTwoLimited	Miscellaneous;Puzzle;Puzzle;General	79	tbd	1 to 2	2/11/2000
Aquaria	PC	Adventure	Activision	4	4	10	0.16	2.24	2007	BitBlot	Action Adventure;Fantasy;General;Fantasy	82	8.2	1	7/12/2007
ArmA II	PC	Shooter	Electronic Arts	3	8	1	214	4.22	2009	BohemiaInteractive	Action;Shooter;Shooter;First-Person;Modern;Modern;Arcade	77	7.5	1	17/6/2009
Armed and Dangerous	PC	Shooter	Electronic Arts	9	8	9	112	2.78	2003	PlanetMoonStudios	Action;Shooter;Shooter;Third-Person;Sci-Fi;Sci-Fi;Arcade	78	7.8	1	2/12/2003
Armored Core 2	PS2	Simulation	Sony Computer Entertainment	1	10	9	243	6.26	2000	FromSoftware	Simulation;Sci-Fi;Mech	78	8.6	1 to 2	3/8/2000
ATV Offroad Fury	PS2	Racing	Activision	7	8.4	2	208	3.34	2001	RainbowStudios	Driving;Racing;Rally / Offroad	82	8.3	1 to 2	6/2/2001
ATV Offroad Fury															

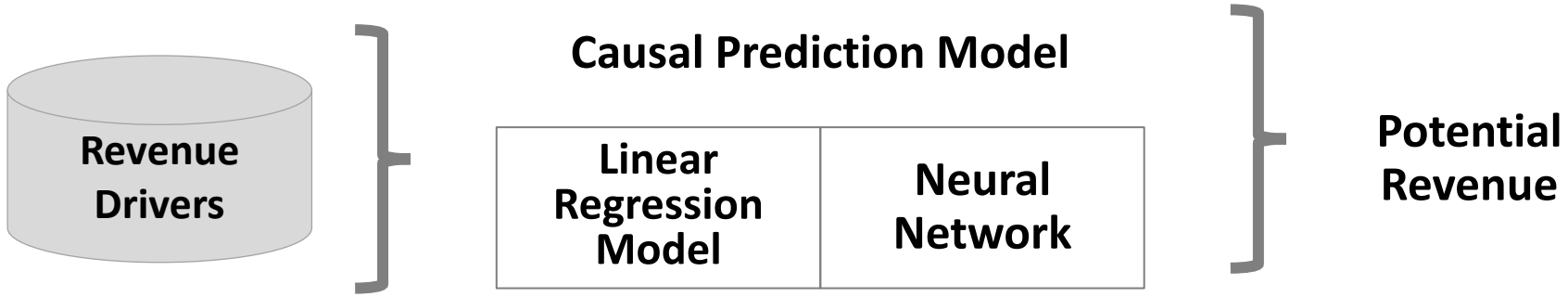
Title	Platform	Userscore	Comment
Halo 3	Xbox360	5.0	This game could have been a lot better. The campaign was way too short and it didn't really explain much of the story. The multiplayer is just downright bad. Every time I play it, i remember why i stopped. The game is fun if u have a bunch of friends who play and you guys can screw around of forge. This is to Oscar G. : yes a boring history..even though this game takes place in 500 This game could have been a lot better. The campaign was way too short and it didn't really explain much of the story. The multiplayer is just downright bad. Every time I play it, i remember why i stopped. The game is fun if u have a bunch of friends who play and you guys can screw around of forge. This is to Oscar G. : yes a boring history..even though this game takes place in 500 years. And of course the gameplay didn't change. What are you going to be killing this game? Zombies? Overall, the campaign was fun for the 2 hours and the multiplayer is way too tournament based.
The Legend of Zelda: Breath of the Wild	WiiU	10.0	This game is a masterpiece, it would definitely feature in my Top 10 favourite games of all times. Simple controls, beautiful environments, great sense of freedom and your discoveries really feel like the result of your own work.
FIFA 16	PlayStation4	6.0	Every year the same.If u r fan its ok ;) the line up is good but the glitches ,pace and pass system is nt what we want again and again :(i would prefer every 2 years a new game and a live team that keeps updating the game over the years until the next one release.
Middle-earth: Shadow of War	XboxOne	10.0	Loll'm playing on nemisis difficulty, I never paid a cent in, games brilliant and I'm super powerful now.Microtransactions didn't ruin nothing.I earned 6k mithril last night 5hr session.If you are putting money into this game you probably aren't any good at it.
Counter-Strike: Global Offensive	PC	10.0	One of the best of the CS series. Sure the first CS is the leader of them all, but CS:GO just puts the series onto a whole new level. It proves that Source can give amazing graphics as well as incredible gameplay, and it shows that CS will never die out. I think this is an amazing release made by VALVe and it's worth every penny.



Predictive Modelling

Revenue Prediction

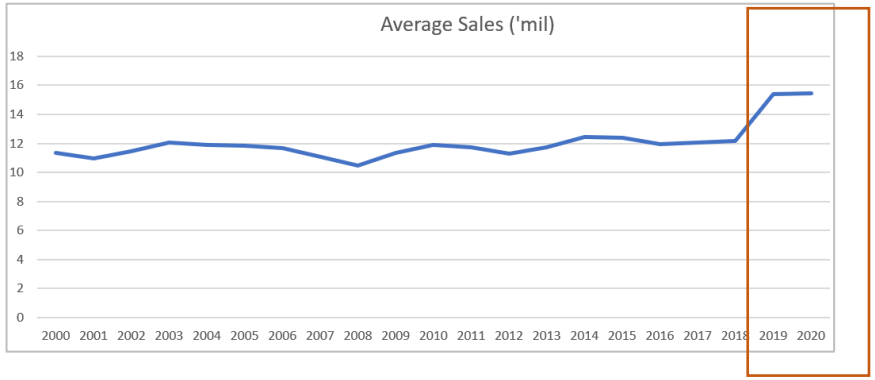
1. Platform
2. Genre
3. Publisher
4. Developer
5. No. of players
6. VG Score
7. Critic Score
8. User Score
9. Meta Score
10. Production Cost
11. Release Date



Data Pre-processing

Apply Exclusion Rules

- There is a spike in the average sales from 2019 onwards due to Covid-19
- Remove those data points to eliminate Covid-19 impact



Remove Missing Values

- 8 missing values identified
- Remove those data points due to not missing completely at random based on MCAR test

Addressed Data Inconsistencies

- Standardized the description for Game Developer
- E.g. EA LA; ElectronicArts .. Standardized to Electronic Arts

Group by Top 10 Developer

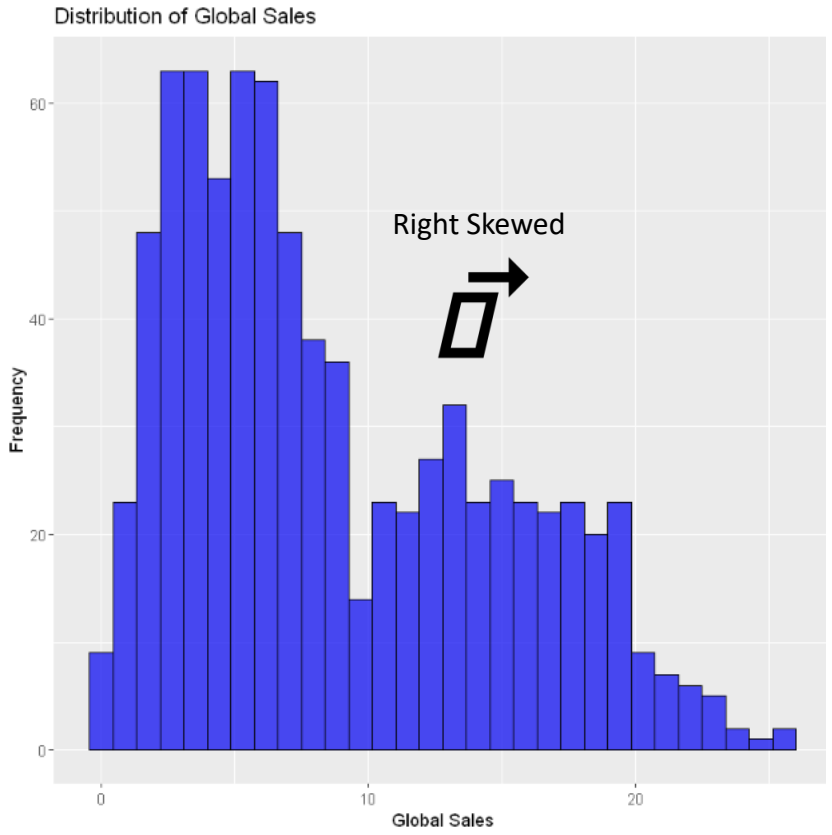
- 283 Developers had been grouped into Top 10 Developers & Others

Derived Time Dimension

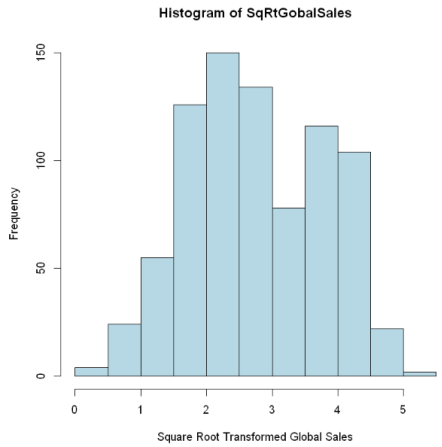
- Release Year; Release Quarter; Release Month; Release Weekday; Release Weekend; Release End of Month

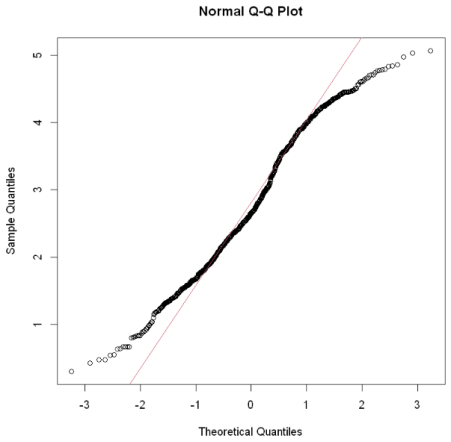
Data Pre-processing

Addressed Non-Normality



Square Root Transformation





Linear Regression

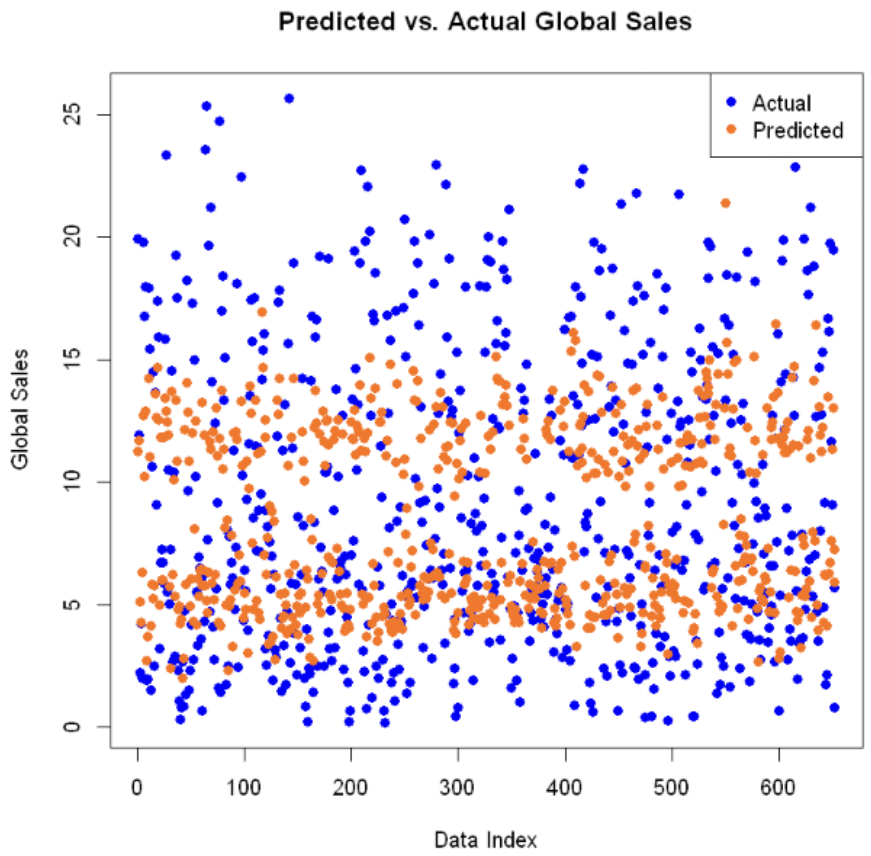
- Observed consistently **poor R-square** results across all models, ranging from **30% to 42%**

	Dependent Variable			
	Response Variable			
	(1)	(2)	(3)	(4)
PlatformPC	1.710***	1.365***	1.614*	1.506*
	(0.173)	(0.165)	(0.825)	(0.854)
PlatformPS	0.900***	0.545*	1.546*	1.823**
	(0.284)	(0.281)	(0.846)	(0.863)
GenreAdventure	0.608***	0.550***	0.801	0.930
	(0.180)	(0.184)	(0.631)	(0.651)
GenreSports	0.078	-0.001	1.639	1.467
	(0.165)	(0.172)	(1.010)	(1.044)
PublisherActivision	0.228**	0.182*		0.192*
	(0.108)	(0.110)		(0.109)
PublisherIBM	0.310**	0.272*		0.363**
	(0.155)	(0.156)		(0.152)
CriticScore	-0.015	-0.011		
	(0.012)	(0.013)		
UserScore	-0.022*	-0.017		
	(0.012)	(0.012)		
ProductionCost	-0.001	-0.0005		
	(0.019)	(0.020)		
ReleaseMonth	0.325*	0.287		
	(0.195)	(0.198)		
ReleaseQuarter	-0.077	-0.082		
	(0.130)	(0.131)		
ReleaseWeekendWeekend	0.370	0.408*	0.410**	0.456**
	(0.237)	(0.239)	(0.205)	(0.211)
ReleaseWeek	-0.067	-0.057		
	(0.043)	(0.044)		
MetaScore	0.021**	0.025***		
	(0.008)	(0.008)		
NoOfPlayers1	-0.347	-0.360		
	(0.856)	(0.865)		
TopDeveloper_SumNamco	1.046***		1.190***	
	(0.354)		(0.360)	
TopDeveloper_SumNintendo	2.566***		2.389***	
	(0.446)		(0.451)	
TopDeveloper_CountNamco		1.236***		
		(0.388)		
TopDeveloper_CountSportsInteractive		1.019**		
		(0.411)		
PlatformPS3:GenrePuzzle			2.558**	2.306**
			(1.073)	(1.106)
PlatformPS:GenreSports			-2.187*	-2.237*
			(1.219)	(1.255)
Constant	-11.673	-15.346	0.666	1.261**
	(19.707)	(19.885)	(0.629)	(0.603)
Observations	652	652	652	652
R2	0.403	0.391	0.419	0.382
Adjusted R2	0.357	0.344	0.368	0.331
Residual Std. Error	0.832 (df = 605)	0.840 (df = 605)	0.825 (df = 599)	0.848 (df = 602)
F Statistic	8.873*** (df = 46; 605)	8.431*** (df = 46; 605)	8.298*** (df = 52; 599)	7.584*** (df = 49; 602)
Note:				*p<0.1; **p<0.05; ***p<0.01

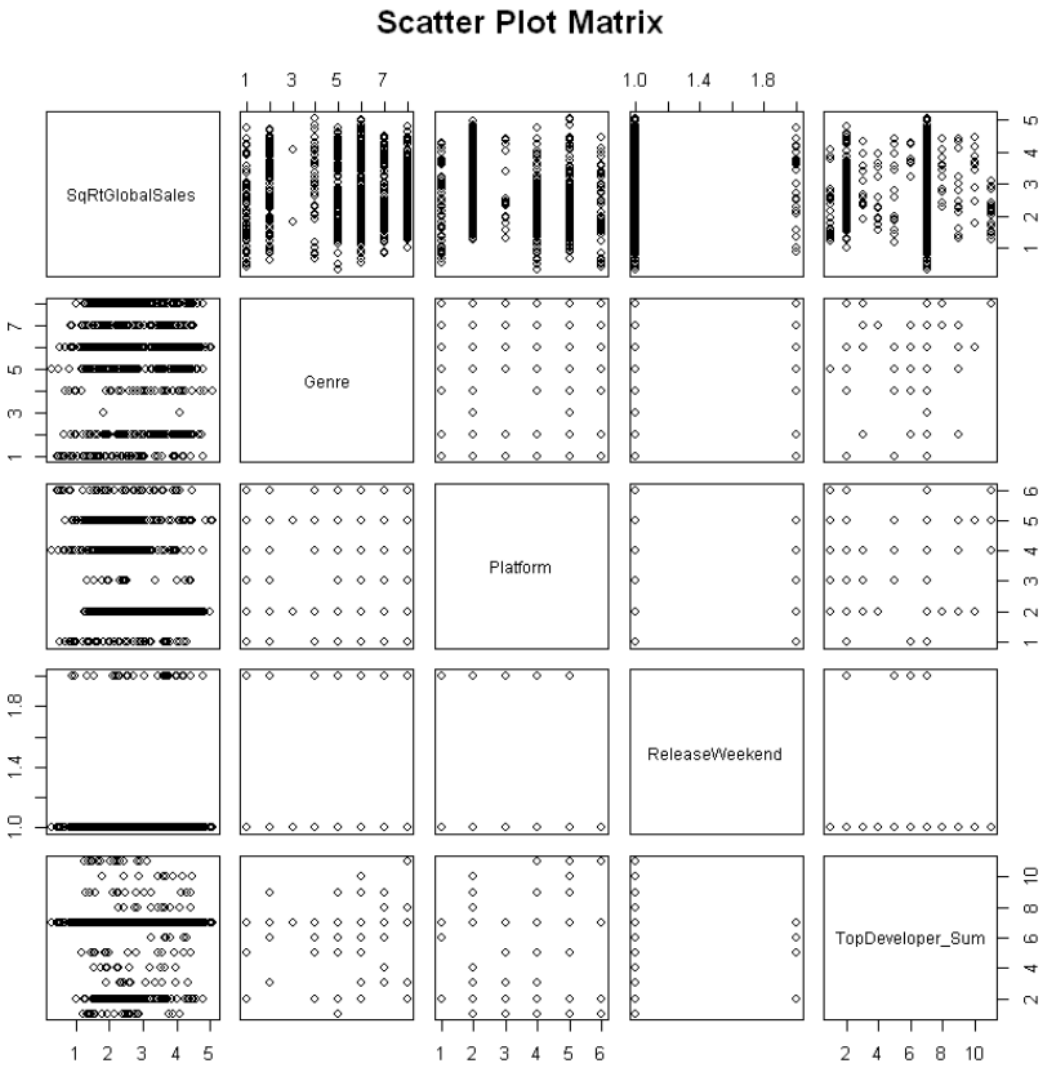
Note: Please note that this simplified version of the stargazer output is provided for presentation purposes. For comprehensive and detailed information, kindly refer to the project documentation

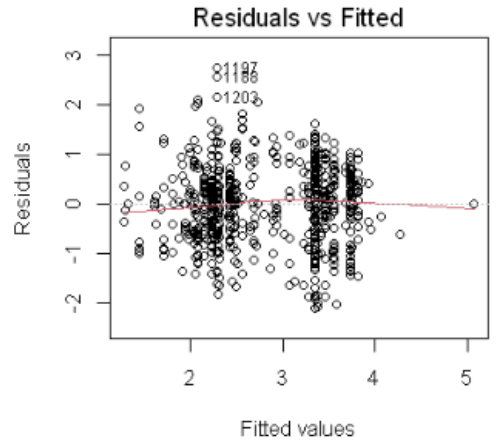
Actual vs Predicted Global Sales

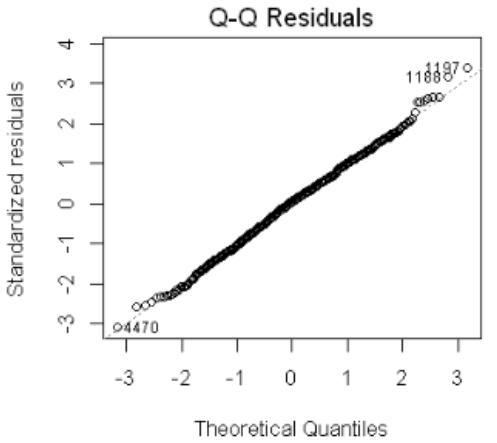
Model 3

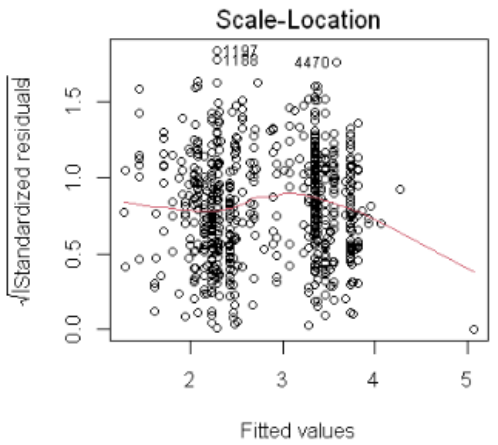


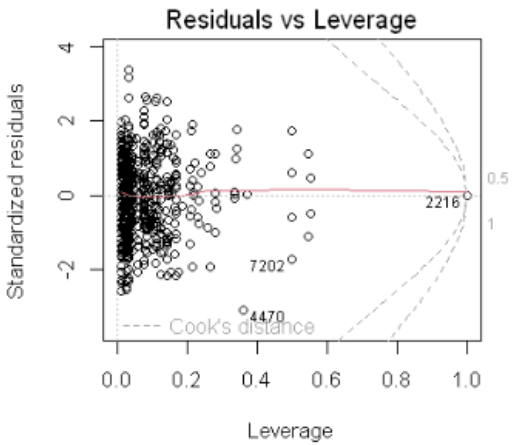
	Training Dataset (80%)	Testing Dataset (20%)
Mean Absolute Deviation	3.55	3.60
Mean Squared Error	21.56	22.30
Standard Error	4.64	4.72


Scatter Plot Matrix
SqRtGlobalSales
Genre
Platform
ReleaseWeekend
TopDeveloper_Sum


Residuals vs Fitted
Residuals
Fitted values
1197
1188
1202

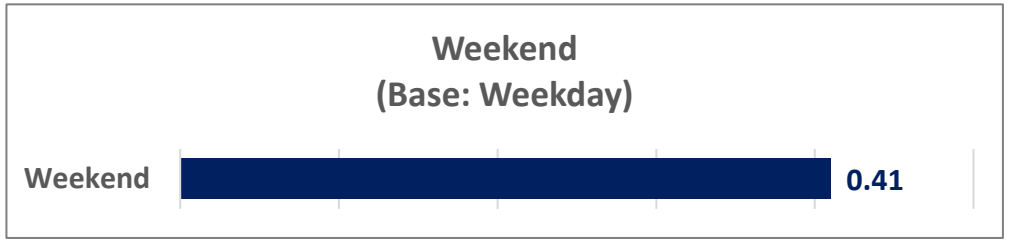

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
1197
1188
4470

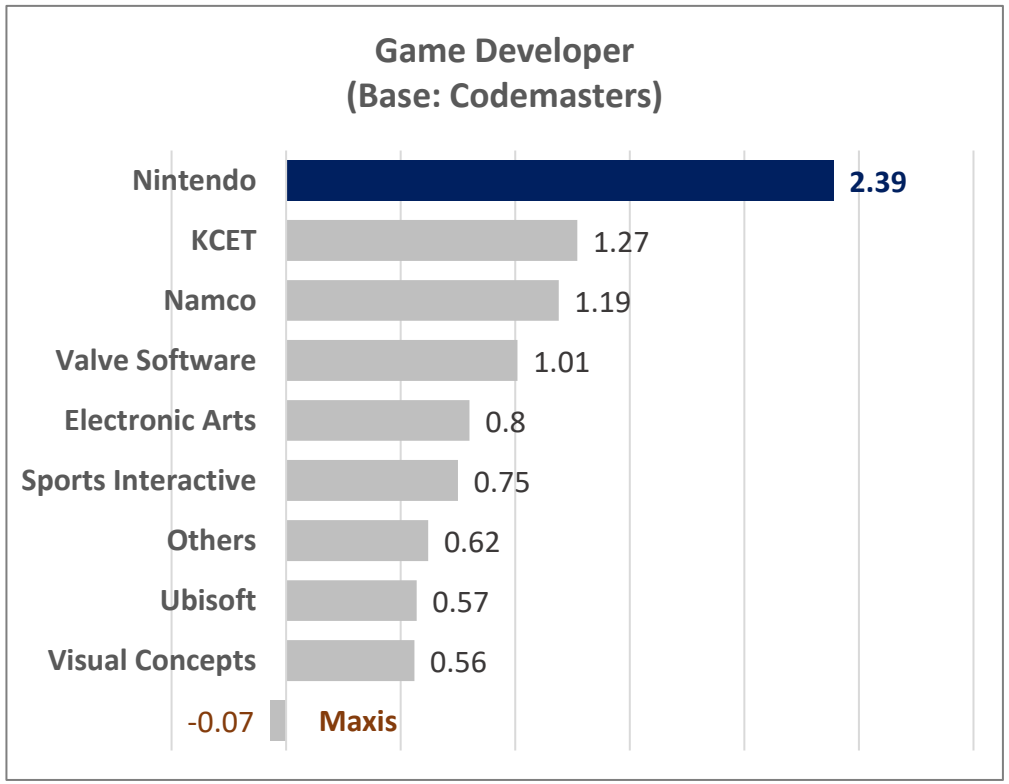

Scale-Location
√|Standardized residuals|
Fitted values
1197
1188
4470

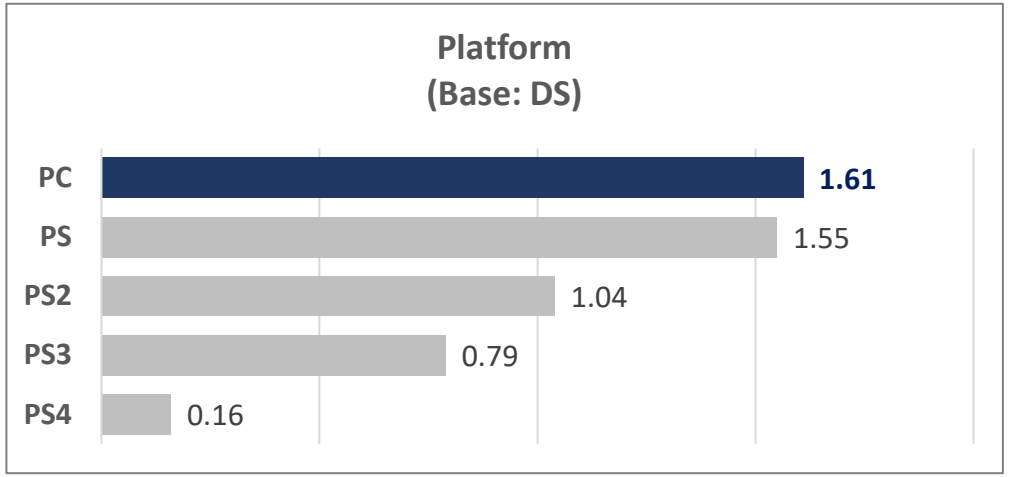

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
2216
7202
4470

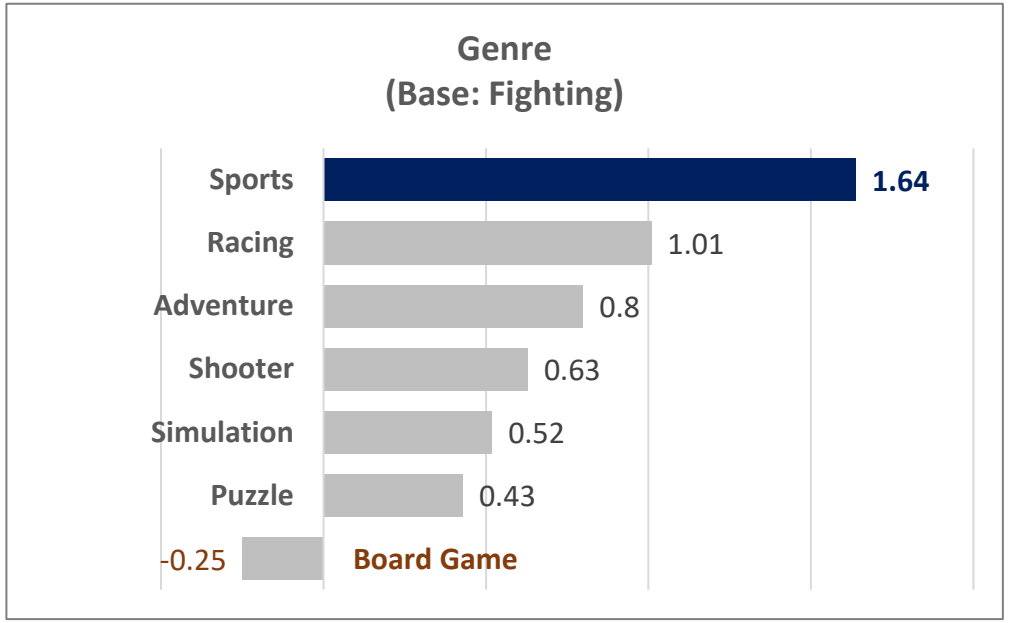
Key Revenue Contributors

Net impact on Revenue ($'mil)









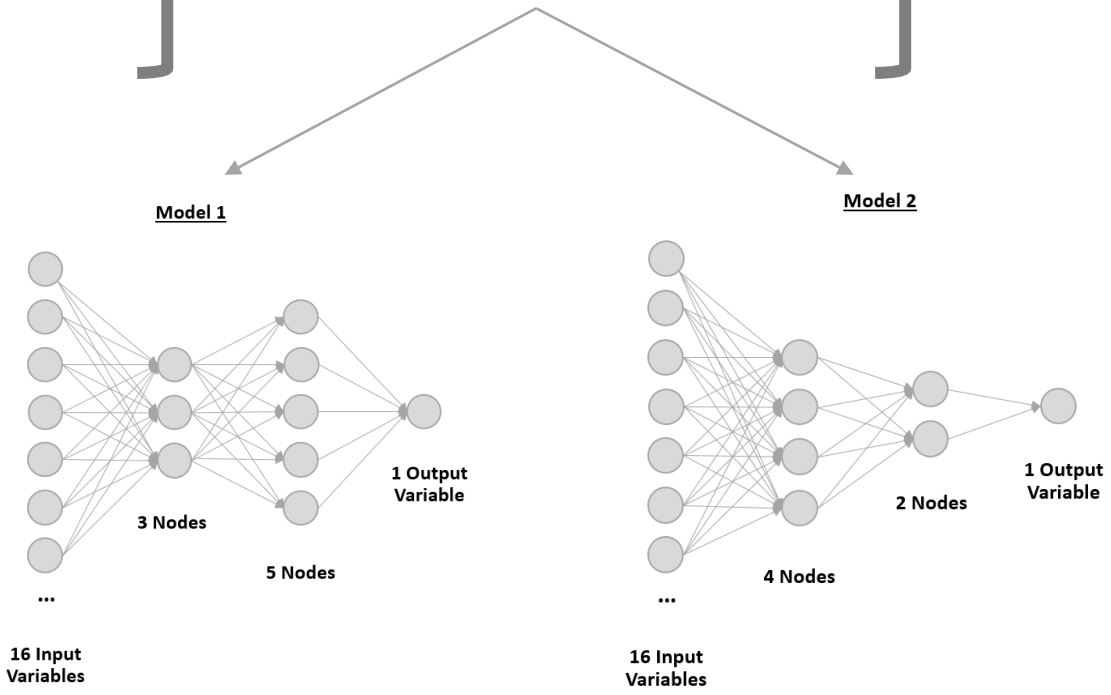
Neural Network

Revenue Dataset

Normalized all the input variables

Neural Network Model

Evaluate the Models



Architecture:

- Activation Function: Relu
- Loss Function: Mean squared error

Training Configuration:

- Epochs: 100

Models Comparison

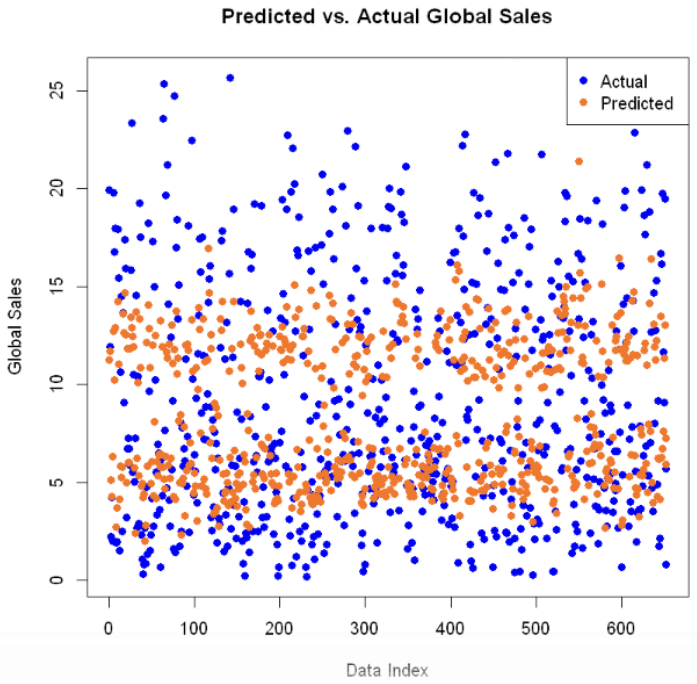
	Linear Regression [Model 3]		Neural Network [Model 1]		Neural Network [Model 2]	
	Training	Testing	Training	Testing	Training	Testing
Mean Absolute Deviation	3.67	3.66	5.11	4.91	5.10	4.98
Mean Squared Error	22.87	22.52	37.03	32.84	36.97	33.98
Standard Error	4.78	4.75	6.09	5.73	6.08	5.83

Linear Regression Outperforms Neural Network:

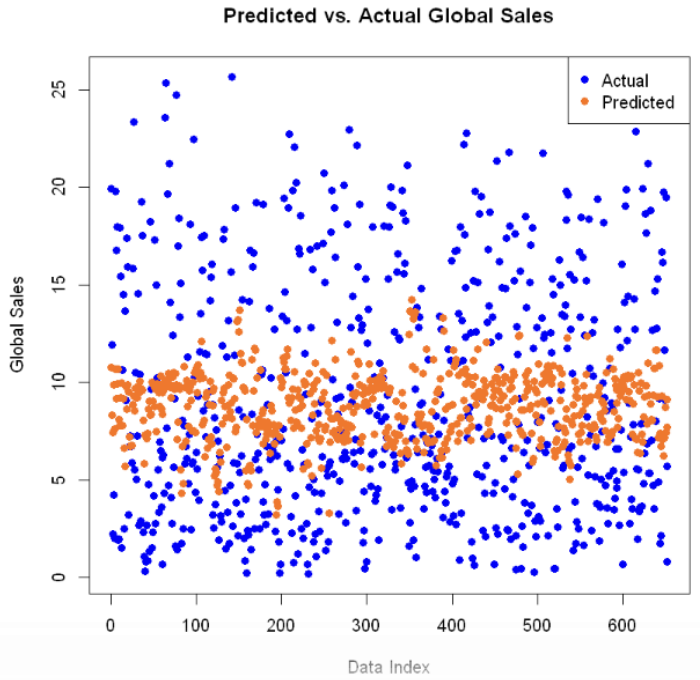
- Highly likely attributed to the **limited dataset size**, which consists of around **815 data points**.
- Limited dataset **restricted the complexity of the neural network model**, allowing only a **simple architecture with few nodes and layers to be set up**.

Models Comparison

Linear Regression [Model 3]



Neural Network [Model 2]



- **Both Models Underfitting**: Indicating limitations in capturing the complexity of the relationship within the data
- **Consider Categorical Prediction Models**: These models aim to predict the hit or miss category rather than attempting to predict a precise point estimate of revenue

Classification model

“Hit” or “Miss”

Model variables

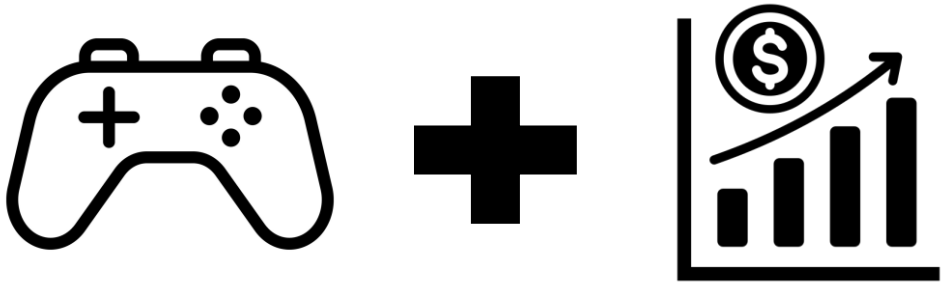
Input variables

- **Numerical:**
  Critic score, User score, Production Cost
- **Categorical:**
  Genre, platform, publisher, number of players, Console, Developer
- **Output from Text Mining TDM:**
  Features and improvements in beta players' user comments

Output variables

- Game's Hit or Miss (Based on Net profit)

Data Preparation



Inner Join game sales dataset with user review dataset

Derived “Game hit/miss” column

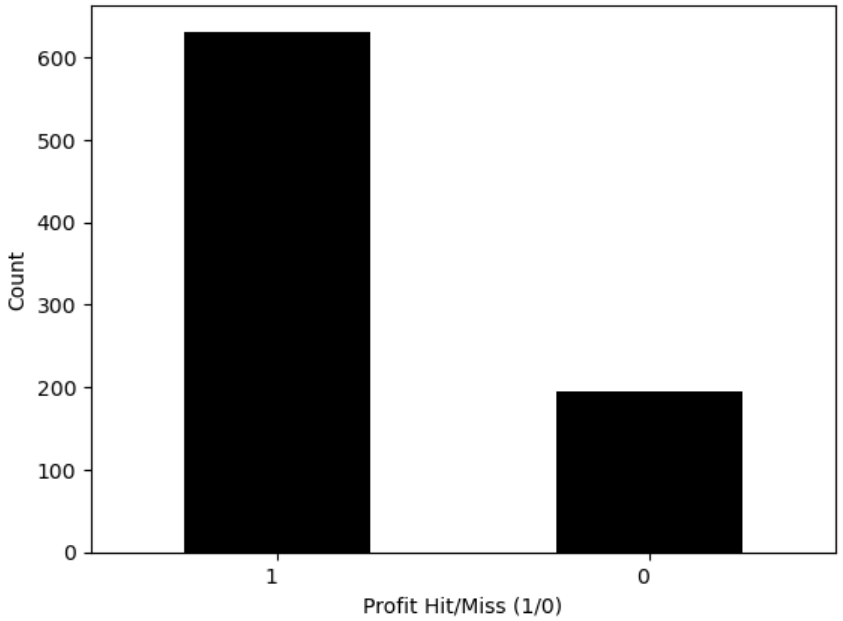
platform	genre	publisher	Developer_metacritic_game_info
PC	Simulation	Sony Computer Entertainment	RelatedDesigns
PS3	Sports	Ubisoft	EACanada
PC	Adventure	Activision	GiantSquid
PS2	Sports	Activision	Z-Axis,Ltd.
DS	Adventure	Electronic Arts	EtrangesLibellules
PS	Adventure	Electronic Arts	Darkworks
DS	Simulation	Amazon Game Studios	Nintendo
DS	Simulation	Ubisoft	SunflowersInteractive
DS	Adventure	Ubisoft	Capcom
PS2	Puzzle	Frontier Developments	ZedTwoLimited
PC	Adventure	Activision	BitBlot

genre_Board Game	genre_Fighting	genre_Puzzle	genre_Racing	genre_Shooter	genre_Simulation	genre_Sports	publisher_Amazon Game Studios
0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0

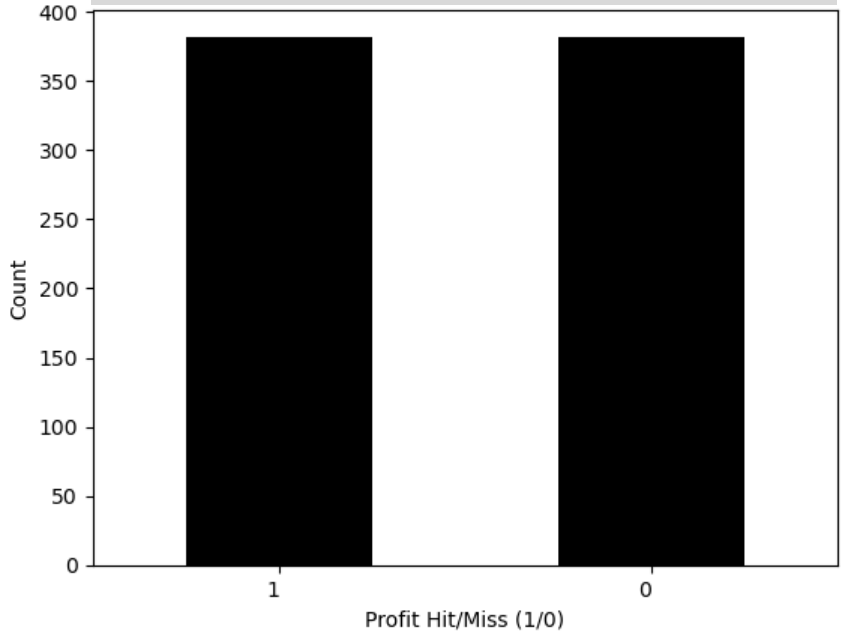
Encode categorical columns

Data Imbalance Handling

Original dataset Distribution



Distribution of class after SMOTE

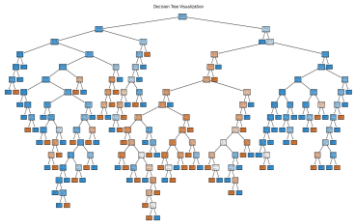


- Dataset exhibits imbalance with the profit hit/miss
  - Class 1: Profit 76%
  - Class 2: Loss 24%
- Applied **SMOTE** to training set to generate synthetic samples for minority class
- Applied train test split at 80:20 rule

Parameters & Tuning Factor

Random Forest

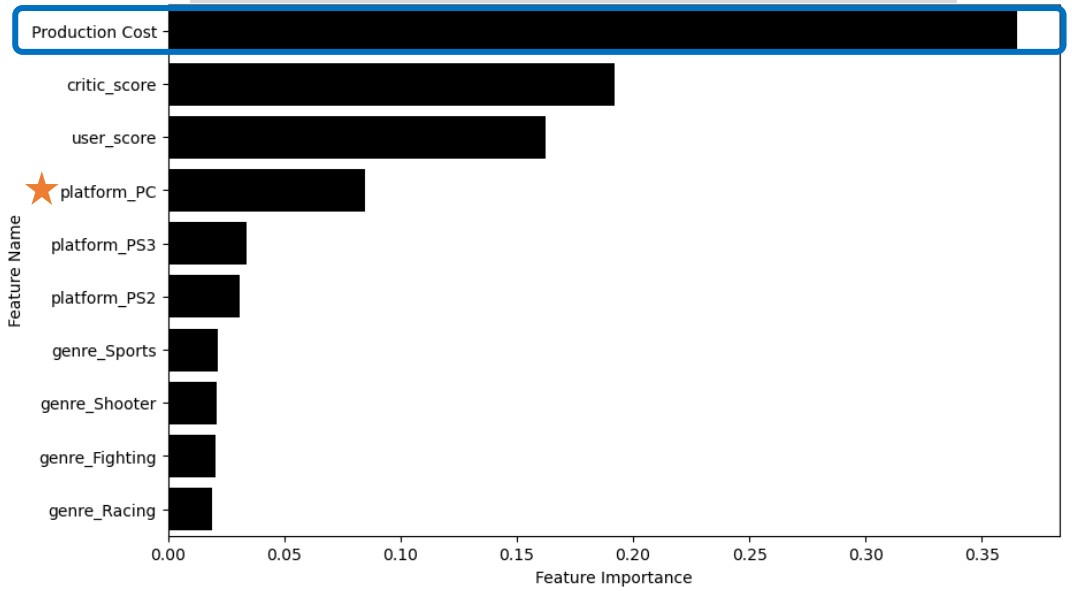
- Hyperparameter Tuning: Grid search
- Features selected: Genre, Platform, critic score, user score, Production cost



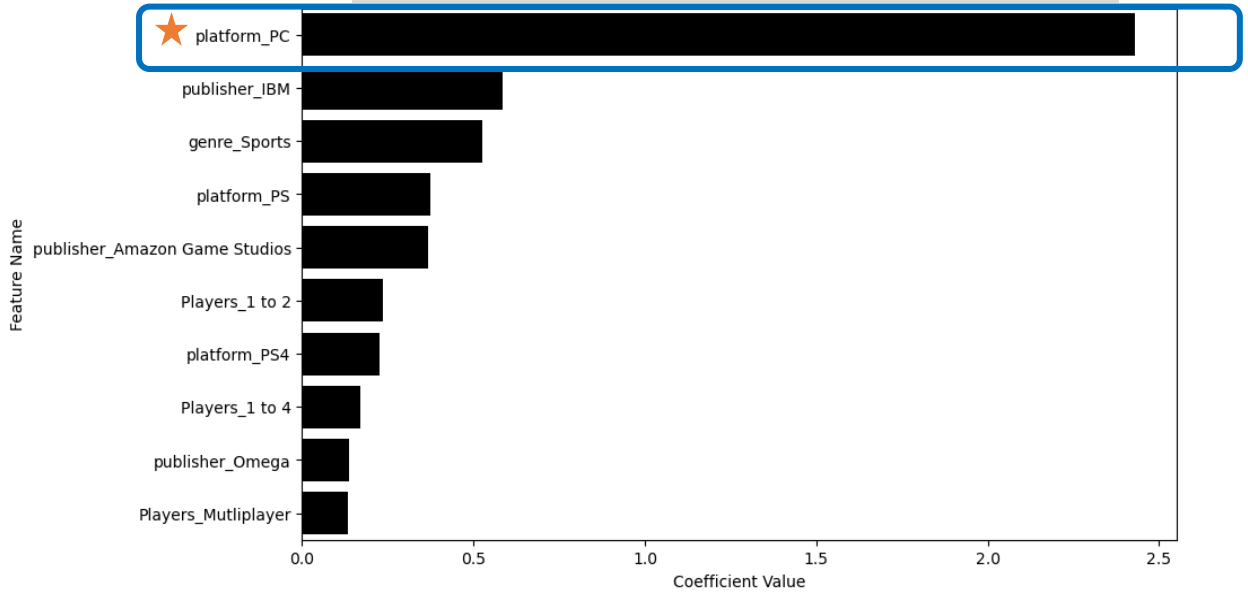
Logistic regression

- Maximum number of iterations: 1000
- Features selected: Platform, No. of players, critic score, user score, Production cost

RF Feature Importance

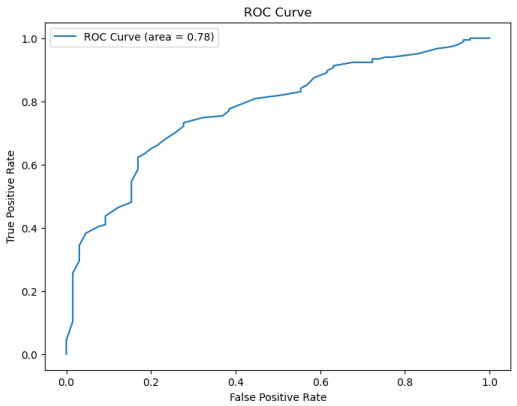


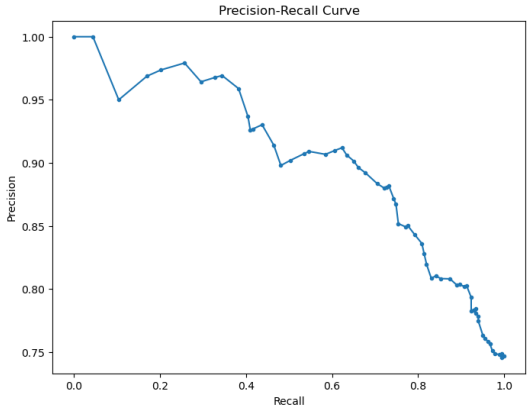
LogReg Coefficient Value



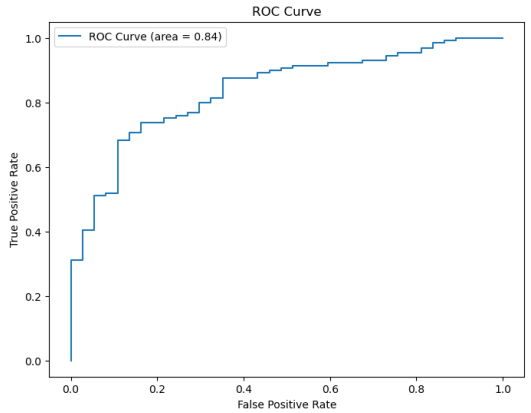
Model Performance

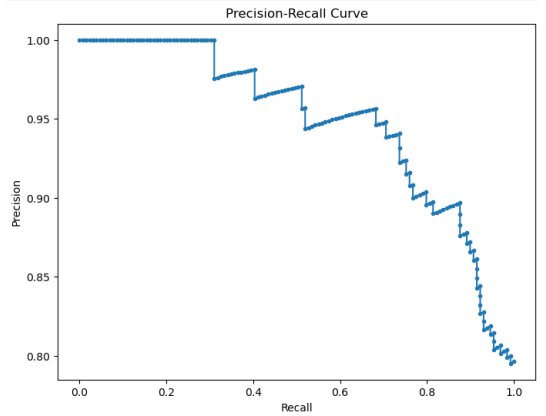
Random Forest





Logistic regression





	Random Forest	Logistic regression	Neural Network	SVM
Accuracy	0.76	0.81	0.72	0.73
ROC AUC	0.78	0.84	0.56	0.71
Cross Entropy	0.48	0.43	9.57	0.46

- LogReg outperforms RF in accuracy and ROC AUC.
- Lower cross-entropy suggests a higher level of confidence in its predictions.

Note: Scores are based on test dataset.

Next steps

Further refine Logistic Regression model.

Evaluate the interpretability of the model and ensure it aligns with business needs.

Leveraging domain knowledge in the gaming industry dynamics

Adopt an iterative approach to model improvement

Text Analytics

Text Analytics

1. “How would you rate the stated game, played via this stated platform, on a Likert scale of 0 to 10?”
2. “Please justify why you gave this rating score.”

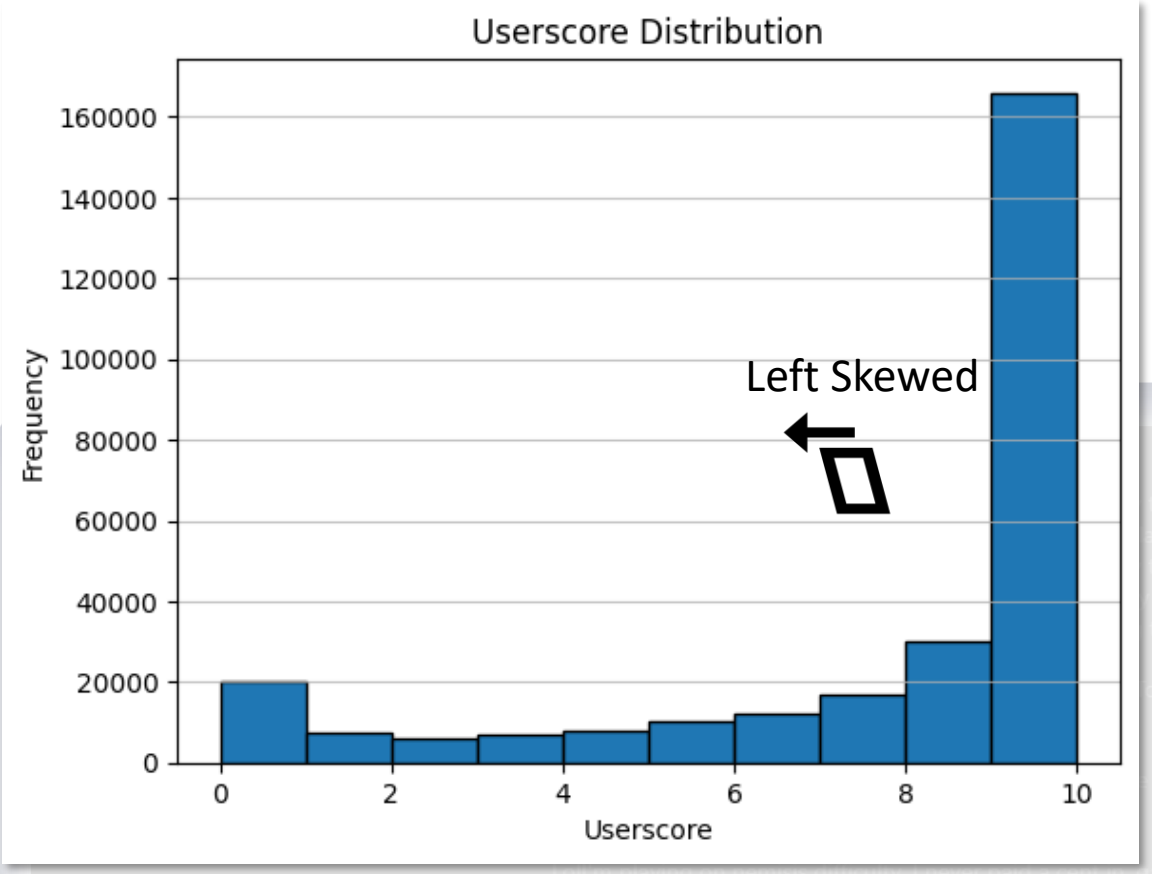
Objective:

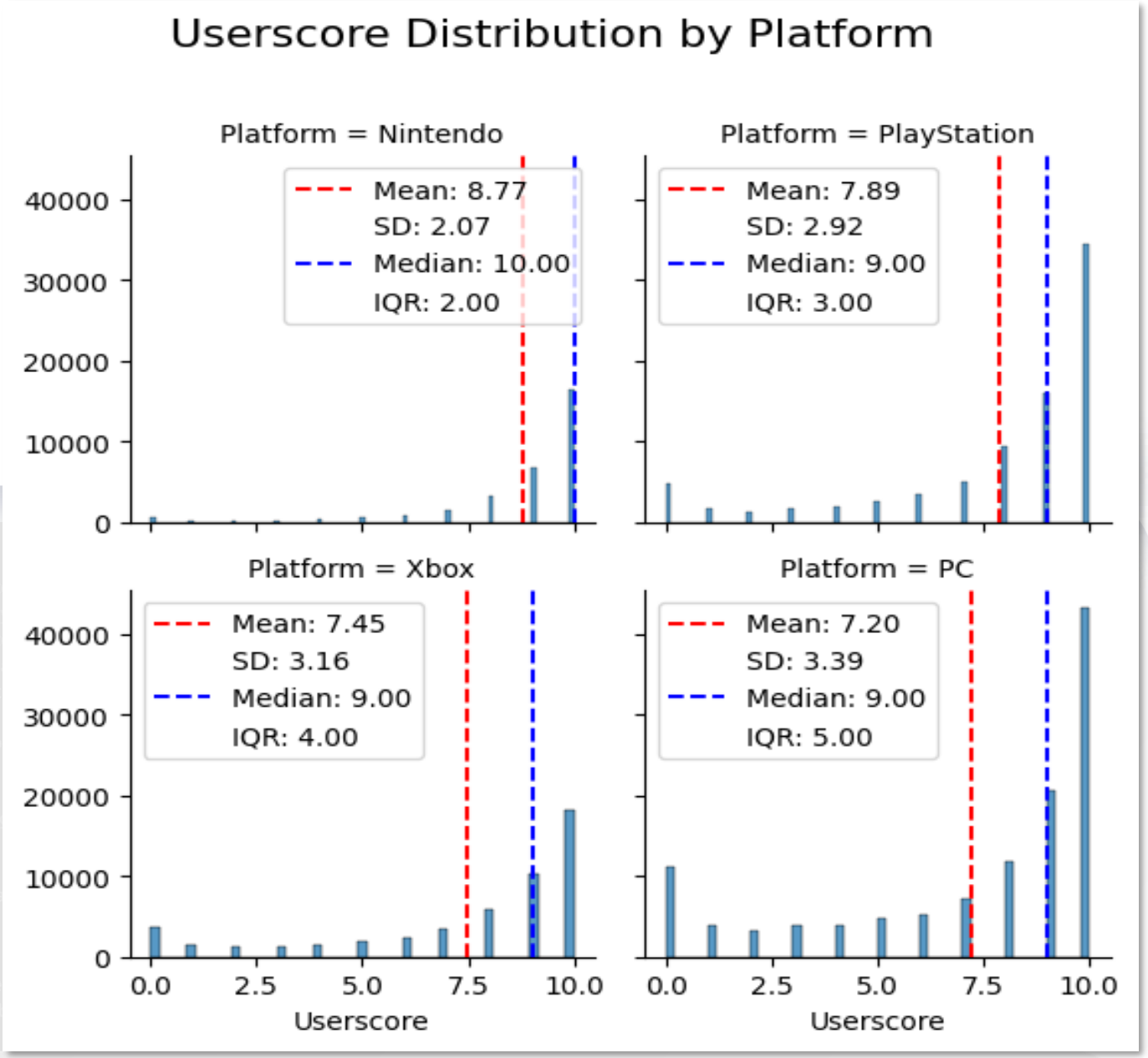
To use these two questions to find out 3 to 5 key **‘features’** and main ‘**improvements**’ that players look for in Games per Gaming Platform

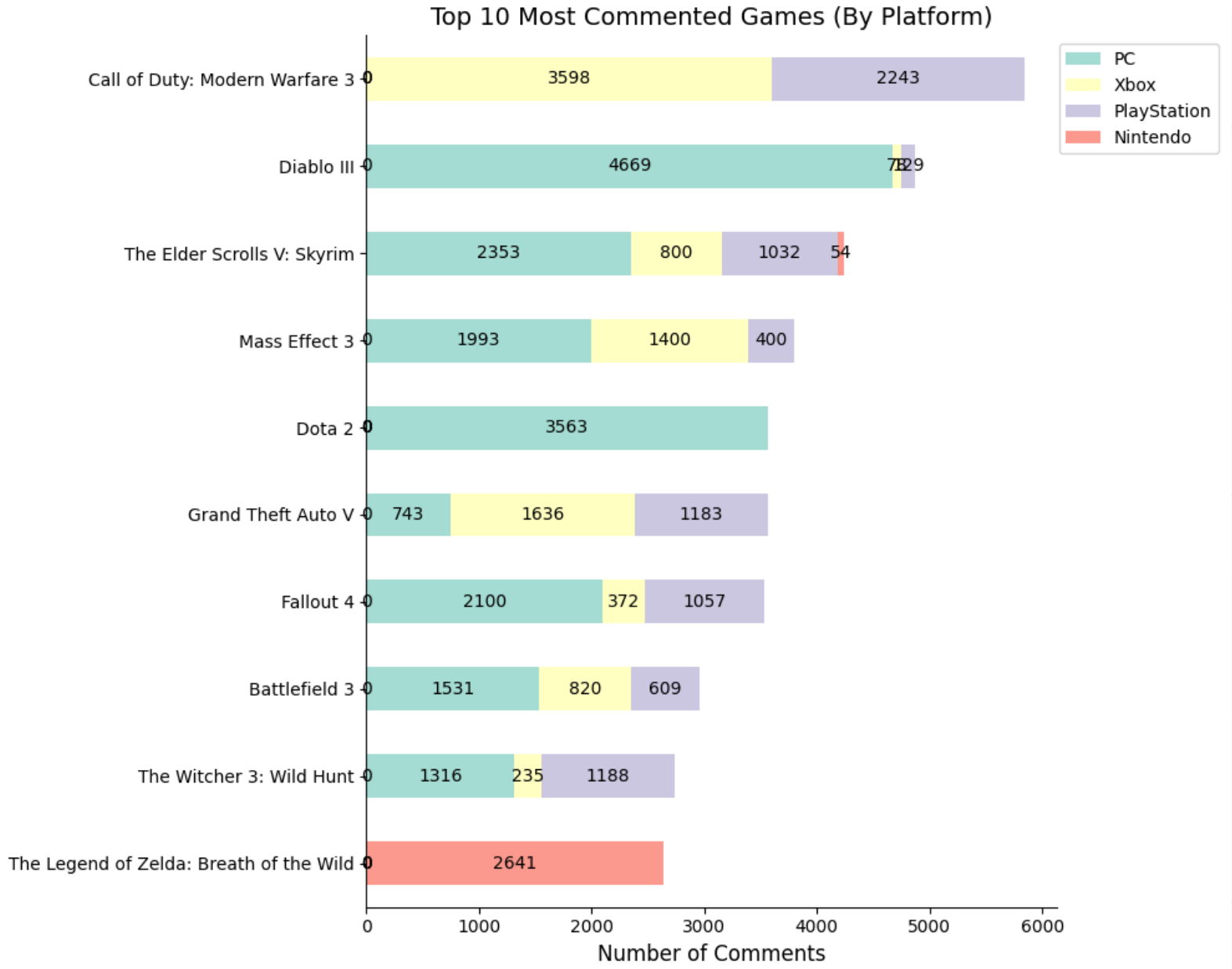
1

2

Title	Platform	Userscore	Comment
Halo 3	Xbox360	5.0	This game could have been a lot better. The campaign was way too short and it didn't really explain much of the story. The multiplayer is just downright bad. Every time I play it, i remember why i stopped. The game is fun if u have a bunch of friends who play and you guys can screw around of forge. This is to Oscar G. : yes a boring history..even though this game takes place in 500 This game could have been a lot better. The campaign was way too short and it didn't really explain much of the story. The multiplayer is just downright bad. Every time I play it, i remember why i stopped. The game is fun if u have a bunch of friends who play and you guys can screw around of forge. This is to Oscar G. : yes a boring history..even though this game takes place in 500 years. And of course the gameplay didn't change. What are you going to be killing this game? Zombies? Overall, the campaign was fun for the 2 hours and the multiplayer is way too tournament based.
The Legend of Zelda: Breath of the Wild	WiiU	10.0	This game is a masterpiece, it would definitely feature in my Top 10 favourite games of all times. Simple controls, beautiful environments, great sense of freedom and your discoveries really feel like the result of your own work.
FIFA 16	PlayStation4	6.0	Every year the same.If u r fan its ok ;) the line up is good but the glitches ,pace and pass system is nt what we want again and again :(i would prefer every 2 years a new game and a live team that keeps updating the game over the years until the next one release.
Middle-earth: Shadow of War	XboxOne	10.0	Loll'm playing on nemisis difficulty, I never paid a cent in, games brilliant and I'm super powerful now.Microtransactions didn't ruin nothing.I earned 6k mithril last night 5hr session.If you are putting money into this game you probably aren't any good at it.
Counter-Strike: Global Offensive	PC	10.0	One of the best of the CS series. Sure the first CS is the leader of them all, but CS:GO just puts the series onto a whole new level. It proves that Source can give amazing graphics as well as incredible gameplay, and it shows that CS will never die out. I think this is an amazing release made by VALVe and it's worth every penny.


Userscore Distribution
Frequency
Userscore
Left Skewed
0
20000
40000
60000
80000
100000
120000
140000
160000
0
2
4
6
8
10


Userscore Distribution by Platform
Platform = Nintendo
Mean: 8.77
SD: 2.07
Median: 10.00
IQR: 2.00
Platform = PlayStation
Mean: 7.89
SD: 2.92
Median: 9.00
IQR: 3.00
Platform = Xbox
Mean: 7.45
SD: 3.16
Median: 9.00
IQR: 4.00
Platform = PC
Mean: 7.20
SD: 3.39
Median: 9.00
IQR: 5.00
Userscore
Userscore
0
10000
20000
30000
40000
0.0
2.5
5.0
7.5
10.0


Top 10 Most Commented Games (By Platform)
PC
Xbox
PlayStation
Nintendo
Call of Duty: Modern Warfare 3
Diablo III
The Elder Scrolls V: Skyrim
Mass Effect 3
Dota 2
Grand Theft Auto V
Fallout 4
Battlefield 3
The Witcher 3: Wild Hunt
The Legend of Zelda: Breath of the Wild
0
3598
2243
4669
729
2353
800
1032
54
1993
1400
400
3563
743
1636
1183
2100
372
1057
1531
820
609
1316
235
1188
2641
0
1000
2000
3000
4000
5000
6000
Number of Comments

Q&A method

Topic Modelling method

POS Tagging and TF-IDF method (chosen)

1. *'What are the best features of the game?'*
2. *'What are the main improvements needed for the game?'*

Plan:
Test if the sentiments drawn from both questions are different. If yes, it means that Q&A method is able to discern key features vs. main improvements well and may be used. Else, cannot be used.

Q&A method

Topic Modelling method

POS Tagging and TF-IDF method

1. *'What are the best features of the game?'*
2. *'What are the main improvements needed for the game?'*

Plan:
Test if the sentiments drawn from both questions are different. If yes, it means that Q&A method is able to discern key features vs. main improvements well and may be used. Else, cannot be used.

Key Features

Improvement Needed

Q&A method

Topic Modelling method

POS Tagging and TF-IDF method

Userscore

Q&A method

Topic Modelling method

POS Tagging and TF-IDF method

Score = 10

Score between 0-2

Plan:

If the words in each group points to a consistent topic, we may use that topic as the ‘features’ (for high score) or ‘improvements’ (for low score). Else, cannot be used.

```
High Score:
[(0, '0.002*"gamecube" + 0.002*"metroid" + 0.002*"owns" + 0.002*"decade" + 0.002*"fifa" + 0.002*"1.6" + 0.001*"soccer" + 0.001*"sexy" + 0.001*"pe" + 0.001*"в"'),
(1, '0.005*"de" + 0.004*"e" + 0.004*"sweet" + 0.004*"que" + 0.003*"un" + 0.002*"la" + 0.002*"juego" + 0.002*"comment" + 0.002*"en" + 0.002*"jogo"'),
(2, '0.006*"best" + 0.005*"ever" + 0.004*"played" + 0.004*"good" + 0.004*"story" + 0.004*"like" + 0.003*"graphic" + 0.003*"amazing" + 0.003*"really" + 0.003*"get"')]
Low Score:
[(0, '0.004*"volvo" + 0.002*"в" + 0.002*"hat" + 0.002*"не" + 0.001*"как" + 0.001*"wer" + 0.001*"а" + 0.001*"что" + 0.001*"halloween" + 0.001*"микротранзакции"'),
(1, '0.003*"like" + 0.003*"good" + 0.002*"money" + 0.002*"get" + 0.002*"bad" + 0.002*"give" + 0.002*"even" + 0.002*"player" + 0.002*"story" + 0.002*"boring"'),
(2, '0.003*"pokemon" + 0.002*"uplay" + 0.002*"blade" + 0.001*"batmobile" + 0.001*"blacklist" + 0.001*"pet" + 0.000*"washing" + 0.000*"batman" + 0.000*"smashing" + 0.000*"spirit"')]
```

Q&A method

Topic Modelling method

POS Tagging and TF-IDF method

Q&A method

Topic Modelling method

POS Tagging and TF-IDF method

```
filter_list = ["would", "could", "left", "right", "a.m.", "p.m.", "'s", "! ! !", "...", ":", ";", "n't",
               "game", "games", "play", "fun", "much", "one", "great", "perfect", "time", "year", "lot", "thing", "etc",
               "hour", "hours", "way", "ways", "everything", "anything", "thing", "things", "review", "reviews", "year", "years",
               "feel", "feels", "thing", "nothing", "problem", "end", "begin", "kind", "piece", "work", "call", "anyone",
               "minute", "minutes", "waste", "crap", "garbage", "masterpiece"]

def preprocess_text(tokens, needtokenizeBoolean = True, grams = False, ngramsNumber = 2, furtherPreProcessNgrams = False):

    if needtokenizeBoolean:
        tokens = nltk.word_tokenize(tokens)
        if grams:
            tokens = list(ngrams(tokens, ngramsNumber))
    if grams:
        tokens = [' '.join(gram) for gram in tokens]
        if furtherPreProcessNgrams == False:
            return tokens

    tokens = [t.lower() for t in tokens]
    tokens = [t for t in tokens if t not in stopwords.words('english') + filter_list]
    tokens = [t for t in tokens if t not in string.punctuation]
    tokens = [t for t in tokens if not t.isnumeric()]
    # tokens = [SnowballStemmer('english').stem(t) for t in tokens]
    tokens = [nltk.WordNetLemmatizer().lemmatize(w) for w in tokens]

    return tokens
```

```
for df in userCommentsTESTExtreme_list:
    for platform in platformCondensed_list:
        if df.scoreBin.max() == 'High':
            features_improvements = 'Key Features'
        elif df.scoreBin.max() == 'Low':
            features_improvements = 'Improvements Needed'
        try:
            wc(df = df[df.platformCondensed == platform], columnName = 'Comment',
            preProcessingFunctionBoolean = True, vectorizerMinDf = 2, vectorizerMaxDf = 0.7,
            countVectorizerBinary = True, tfidfVectorizerBoolean = True,
            ngrams = False, ngramsNumber = 3, furtherPreProcessNgrams = False,
            nounTaggingBoolean = True, universalNounTagsetBoolean = False,
            top = 20, features_improvements = features_improvements, platform = platform, titleFontSize = 20)
        except ValueError:
            print(f"\n\nNo Word Cloud for '{features_improvements} ({platform})' due to insufficient sample size (No. of comments = 0 or < vectorizerMinDf).\n\n")
```

```
def wc(df, columnName, preProcessingFunctionBoolean = True, tfidfVectorizerBoolean = True, vectorizerMinDf = 2, vectorizerMaxDf = 0.7, countVectorizerBinary = True,
       nounTaggingBoolean = False, universalNounTagsetBoolean = False, ngrams = False, ngramsNumber = 2, furtherPreProcessNgrams = False,
       top = 10, features_improvements = 'Word Cloud', platform = 'All', titleFontSize = 20):

    if preProcessingFunctionBoolean == True & ngrams == True:
        if tfidfVectorizerBoolean == True:
            if furtherPreProcessNgrams == True:
                tdm = TfidfVectorizer(tokenizer = lambda text: preprocess_text(text, grams = True, ngramsNumber=ngramsNumber, furtherPreProcessNgrams = True),
                                      min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
            else:
                tdm = TfidfVectorizer(tokenizer = lambda text: preprocess_text(text, grams = True, ngramsNumber=ngramsNumber),
                                      min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
        else:
            if furtherPreProcessNgrams == True:
                tdm = CountVectorizer(binary = countVectorizerBinary, tokenizer = lambda text: preprocess_text(text, grams = True, ngramsNumber=ngramsNumber, furtherPreProcessNgrams = True),
                                      min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
            else:
                tdm = CountVectorizer(binary = countVectorizerBinary, tokenizer = lambda text: preprocess_text(text, grams = True, ngramsNumber=ngramsNumber),
                                      min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
        tdmMatrix = tdm.fit_transform(df[columnName])

    if nounTaggingBoolean == True and ngrams == False:
        noun_list = []
        for value in df[columnName]:
            if universalNounTagsetBoolean == True:
                tagged_value = pos_tag(word_tokenize(value), tagset='universal')
                noun = [word for word, tag in tagged_value if tag == 'NOUN']
            else:
                tagged_value = pos_tag(word_tokenize(value))
                noun = [word for word, tag in tagged_value if tag == 'NN' or tag == 'NNS']
            noun_list.append(noun)
        df['Text_NounOnly'] = noun_list
        if preProcessingFunctionBoolean == True:
            df['Text_NounOnly'] = df['Text_NounOnly'].apply(lambda x: preprocess_text(tokens = x, needtokenizeBoolean = False))
            df['Text_NounOnly'] = df['Text_NounOnly'].apply(lambda x: ', '.join(x))
        else:
            df['Text_NounOnly'] = df['Text_NounOnly'].apply(lambda x: ', '.join(x))
        if tfidfVectorizerBoolean == True:
            tdm = TfidfVectorizer(min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
        else:
            tdm = CountVectorizer(binary = countVectorizerBinary, min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
        tdmMatrix = tdm.fit_transform(df['Text_NounOnly'])

    if nounTaggingBoolean != True and ngrams == False:
        if tfidfVectorizerBoolean == True:
            if preProcessingFunctionBoolean == True:
                tdm = TfidfVectorizer(tokenizer = preprocess_text, min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
            else:
                tdm = TfidfVectorizer(min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
        else:
            if preProcessingFunctionBoolean == True:
                tdm = CountVectorizer(binary = countVectorizerBinary, tokenizer = preprocess_text, min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
            else:
                tdm = CountVectorizer(binary = countVectorizerBinary, min_df = vectorizerMinDf, max_df = vectorizerMaxDf)
        tdmMatrix = tdm.fit_transform(df[columnName])

    array = tdmMatrix.toarray()
    feature_names = tdm.get_feature_names_out()
    word = dict(zip(feature_names, array.sum(axis=0)))
    fd = FreqDist(word)
    wc = WordCloud(background_color="white").generate_from_frequencies(fd)
    plt.figure()
    plt.suptitle(f"{features_improvements} ({platform})", fontsize = titleFontSize, x = 0.5, y = 0.85, fontweight = 'bold', fontname = 'Calibri')
    plt.imshow(wc, interpolation='bilinear')
    plt.axis("off")
    displayList = []
    for x,y in fd.most_common(top):
        displayList.append((x, round(y, 2)))

    plt.show()
    print(displayList)
```

Q&A method

Topic Modelling method

POS Tagging and TF-IDF method

Chosen plan:

Preprocessing
- capitalization
- stop words
- custom filter list
- punctuation
- numeric removals

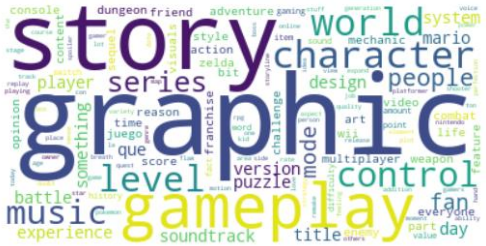
Unigram
Lemmatization
POS Tagging for Nouns
TF-IDF

Q&A method

Topic Modelling method

POS Tagging and TF-IDF method

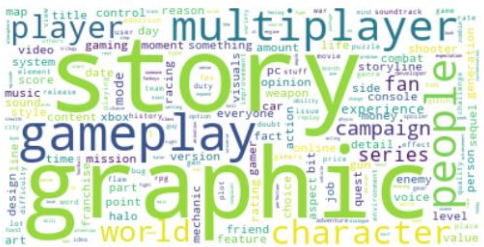
Key Features (Nintendo)



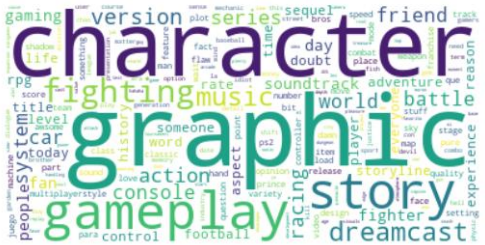
Key Features (PlayStation)

Key Features (PC)

Key Features (Xbox)



Key Features (Others)

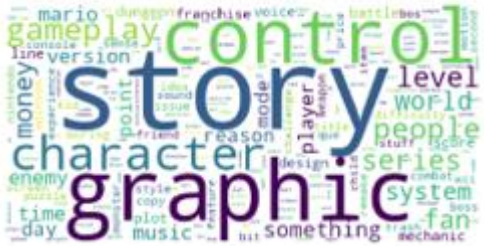


Key Features (All platforms)

Similarities:
Graphic, Story (storyline),
Character (character development)

Differences:
Music in *Nintendo,*
Multiplayer in *XBox*

Improvements Needed (Nintendo)



Improvements Needed (PlayStation)

Improvements Needed (PC)

Improvements Needed (Xbox)

Improvements Needed (Others)

Improvements Needed (All platforms)

Similarities:
Money, Graphic, Story (storyline)

Differences:
Level (levelling difficulty) in *Nintendo,*
Map (limited world map) in *Xbox,*
Camera (POV camera angle) in *Others*

Applying back to business objective

- Information from Word Cloud:
  - Better advisory to clients
    - Recommend must-have key features and pitfalls (improvements) to avoid based on gaming platform type
    - Appeal to gamers

Conclusion

Model

1. Logistic Regression ***

- ROC and AUC
- Business Use Cases (Sales Revenue vs Hit/Miss)
- Availability and Relevancy of Data

2. POS Tagging

- Term Frequency Distribution
- Critical Game attributes
  - Character
  - Gameplay
  - Story

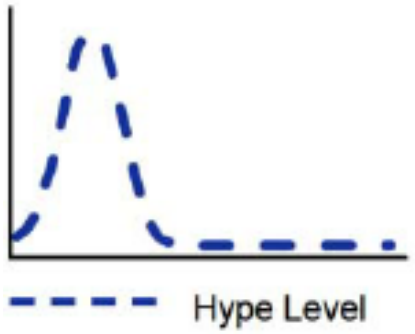
Retrospective

1. Imbalance Dataset

- SMOTE, under sampling etc

2. Sales Revenue Data

- Assumption games sales tapering (older vs newer games)
- Gartner Hype Cycle



Limitations

1. Gaming Industry
    - Dynamic and fast product sale lifecycles
    - Changing consumer preferences

2. Respondent Language
    - Languages other than English in review comments

3. Limited usable data size
    - Not adequate for a NN model

Possible Improvements

1. Enhancing Sentiment Analysis

- **Positive** vs **Negative** Categorization
- Eliciting useful response thru targeted questions

2. Language Handling

- Respondent Demographics (российский & español)
- Google Translate API

3. Potential Model Improvements

- Gradient Boosting / Ensemble model