Predicting Video Game Sales

Andrew Pike

02/01/2021

Introduction:

The video game industry is estimated to be worth $159.3 billion USD in 2020, a 10% increase over the previous year (WePC, 2020). With hundreds of games being produced a year, what drives games sales? Is there a way to gather some data-based information that could help drive development decisions? The goal of this project is to take a dataset which details video game sales over the past 35 years and extract useful insights about parameters that drive global video game sales. This report will detail steps taken in applying machine learning to global video game sales to try and gain information about key drivers for global sales. The dataset that will be used and explored was first pulled from Kaggle.com's data page. The dataset was posted by Rush Kirubi (Kirubi, 2016) and was produced via a web-scrape from Metacritic. Metacritic is a site where users and critics can submit ratings for movies, tv, music and video games (Metacritic, 2020). The dataset, as downloaded, includes just over 16700 rows and 16 columns.

```
##                     Name Platform Year_of_Release    Genre Publisher
## 1            Wii Sports      Wii            2006   Sports  Nintendo
## 2        Mario Kart Wii      Wii            2008   Racing  Nintendo
## 3     Wii Sports Resort      Wii            2009   Sports  Nintendo
## 4 New Super Mario Bros.       DS            2006 Platform  Nintendo
## 5              Wii Play      Wii            2006     Misc  Nintendo
##   Critic_Score Critic_Count User_Score User_Count Developer Rating Global_Sales
## 1           76           51         79        322  Nintendo      E     82530000
## 2           82           73         82        709  Nintendo      E     35520000
## 3           80           73         79        192  Nintendo      E     32770000
## 4           89           65         84        431  Nintendo      E     29800000
## 5           58           41         65        129  Nintendo      E     28920000
##   SystemType       Date Salesfact GamesReleased
## 1    Console 2006-01-01  1109.473           294
## 2    Console 2008-01-01  1109.473           294
## 3    Console 2009-01-01  1109.473           294
## 4   Handheld 2006-01-01  1109.473           294
## 5    Console 2006-01-01  1109.473           294
```

Five of the columns are sales related columns, for this analysis the focus will be Global Sales. Sales are reported in millions but were converted to a sales in units number (sales * 10^6). Four of the columns include information about public and critical response to the game. The columns are critic score/count and user score/count. These fields represent the mean score assigned to a game by critics as well as the count of critics that led to that score and the average user score and the number of users that led to that score. Critic/User scores are based on a range from 0-100 and both have a mean of approximately 70. Other Information about the video games include, publisher, developer, title, year the game was released, game platform, game genre and game ESRB rating. The ESRB is the Entertainment Software Rating Board, they provide consumer ratings for users to make appropriate gaming decisions for their families (ESRB, 2020). The ratings are

based on in-game violence, explicit language, blood/gore etc., there are 9 total possible ratings. Platform describes the gaming system that the game was released on. There are 31 different platforms listed in this parameter and each game has only one platform listed. Publisher and developer are self explanatory, there are 572 unique values for publisher and 1697 unique values for developer. Genre describes the type of game and includes 13 different unique values. The table is a long and skinny format and games are repeated when they are released on different platforms. The global sales are unique to the platform the game was released on. Three extra parameters were parsed out and created from the dataset to include in model training and testing. The first additional parameter is a grouping of platform to try and reduce the dimensionality of the platform type. The existing levels were grouped into one of three different broader categories, Handheld, PC or Console. In the broader gaming world users are often divided between these categories so they will be used to see if they have predictive power. The second additional field added to the dataset is a sales factor. Sales factor is a measure of a publisher's success above the norm. Theoretically, publishers who tend to perform above the mean sales will release games that sell better than new publishers with few releases. The publishers with high sales numbers likely have more money for advertising and may be releasing games in well known and beloved game series that will sell above average based on fans of previous versions. The final additional parameter is a simple sum of games released per publisher. The logic is similar to the sales factor, publishers who have sold more games than others may be better established stable publishers that are likely to sell more copies of the new game.

Two different model types were tested with various combinations of parameter inputs. The approach to training and testing the model involved stepwise inclusion of parameters based on logical relationships. To prevent over-training, 5-fold cross validation was used to train and assess the model. To evaluate model performance R^2 and RMSE was inspected and parameters were included or discarded based on the returned value. The final model is a gradient boosted model that includes platform, genre, user count, critic score, year of release, rating, and sales factor. The model with the lowest RMSE value was ultimately chosen as the final model. It was then tested against validation data set and returned an RMSE value of 957 722.7 copies.

Methods/Analysis:

The following libraries were installed (when necessary) and loaded into the working environment.

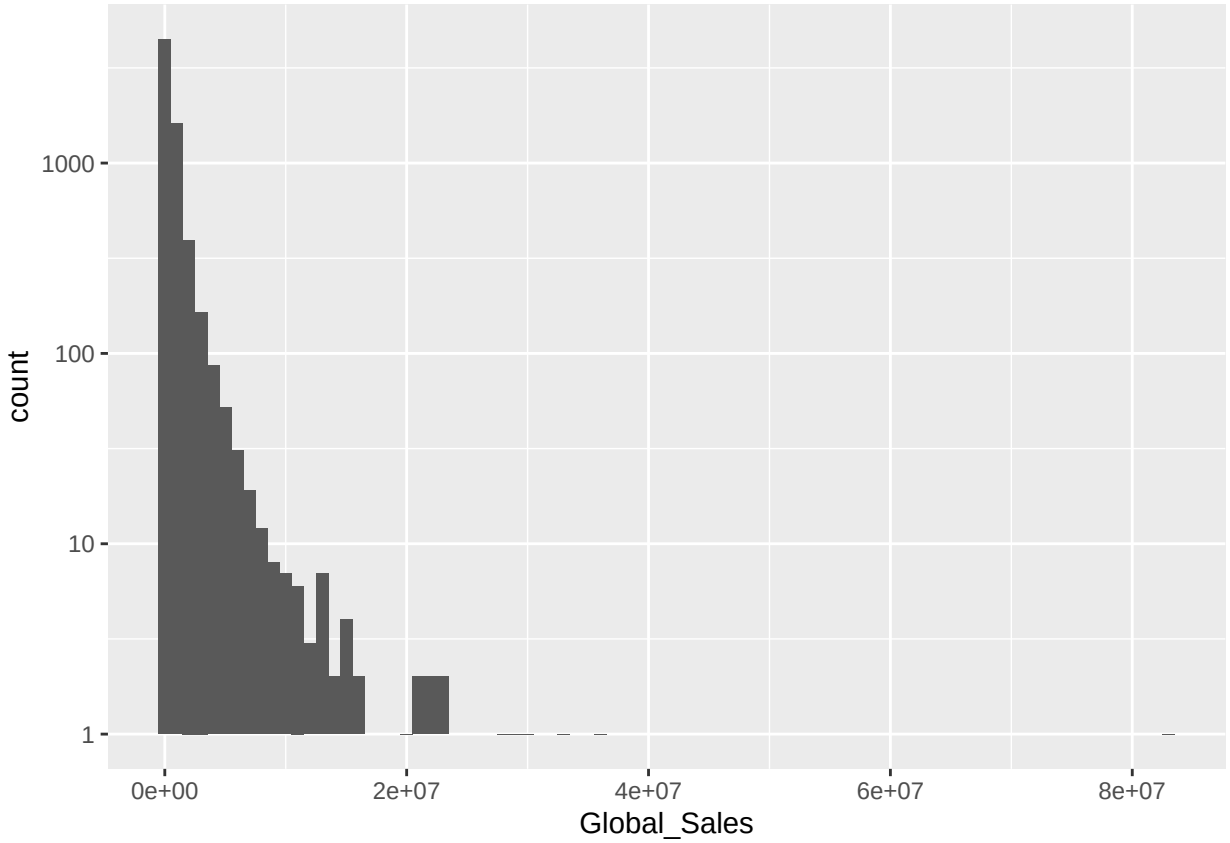
```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(dplyr)) install.packages("dplyr", repos = "http://cran.us.r-project.org")
if(!require(corrplot)) install.packages("corrplot", repos = "http://cran.us.r-project.org")
if(!require(ggplot2)) install.packages("ggplot2", repos = "http://cran.us.r-project.org")
if(!require(readr)) install.packages("readr", repos = "http://cran.us.r-project.org")
if(!require(Rborist)) install.packages("Rborist", repos = "http://cran.us.r-project.org")
if(!require(gbm)) install.packages("gbm", repos = "http://cran.us.r-project.org")
if(!require(knitr)) install.packages("knitr", repos = "http://cran.us.r-project.org")

#load libraries
library(tidyverse)
library(caret)
library(data.table)
library(dplyr)
library(corrplot)
library(ggplot2)
library(readr)
library(Rborist)
library(gbm)
library(knitr)
```

The data was downloaded from Kaggle originally, but it requires authentication to access so it was then hosted in GitHub to make loading the dataset easier and allow script level access. The data is downloaded from the github repository and read into memory. The source for the dataset indicates that there are multiple NA's in the dataset (Kirubi, 2016), so the next step was to see if the NA's are temporally dependent. Its possible that the NA's exist only in older games. Unfortunately, the NA's are independent of time and littered throughout the dataset. A "clean" version of the dataset was created that removed all rows with NAs and all sales fields except the Global sales field.

Sales numbers were converted to numerical sales (instead of sales in millions) and central tendencies and distribution were explored. Global sales show a parabolic distribution with most sales below 1 million units and very few games with over 2 million in sales

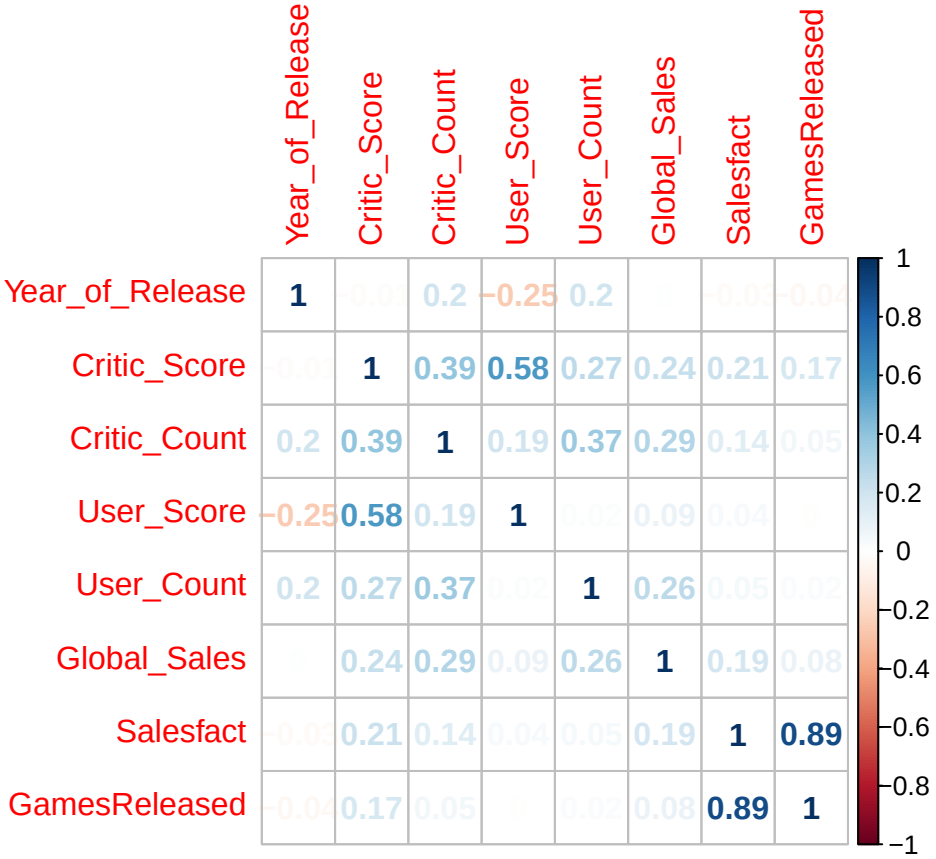
	Sales
MEAN	767048.6
SD	1940316.7
MIN	10000.0
MAX	82530000.0

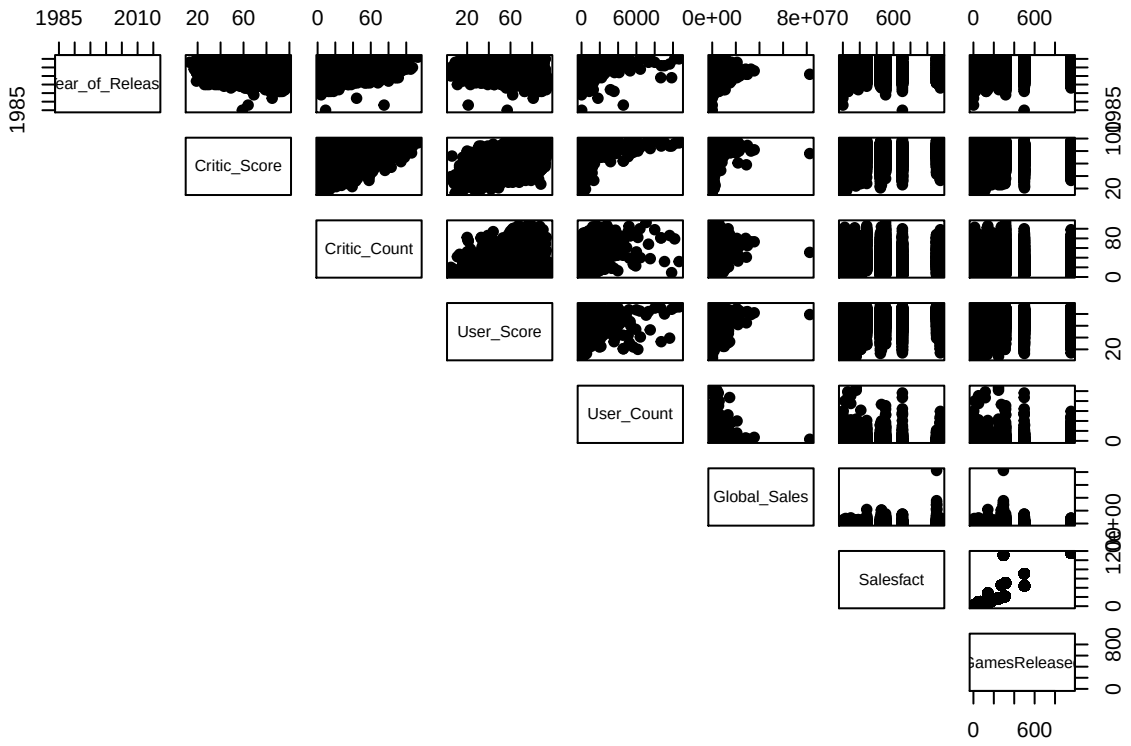


All games were assigned a new system type based on grouping the games into one of three categories, PC, Handheld or Console. Year of release and user score were converted to numeric values (from factors). A date type field was created from the numeric year of release field. Sales factor was calculated as total sales per publisher divided by the mean sales for all games. A count of games released for each publisher was added to each game in the dataset.

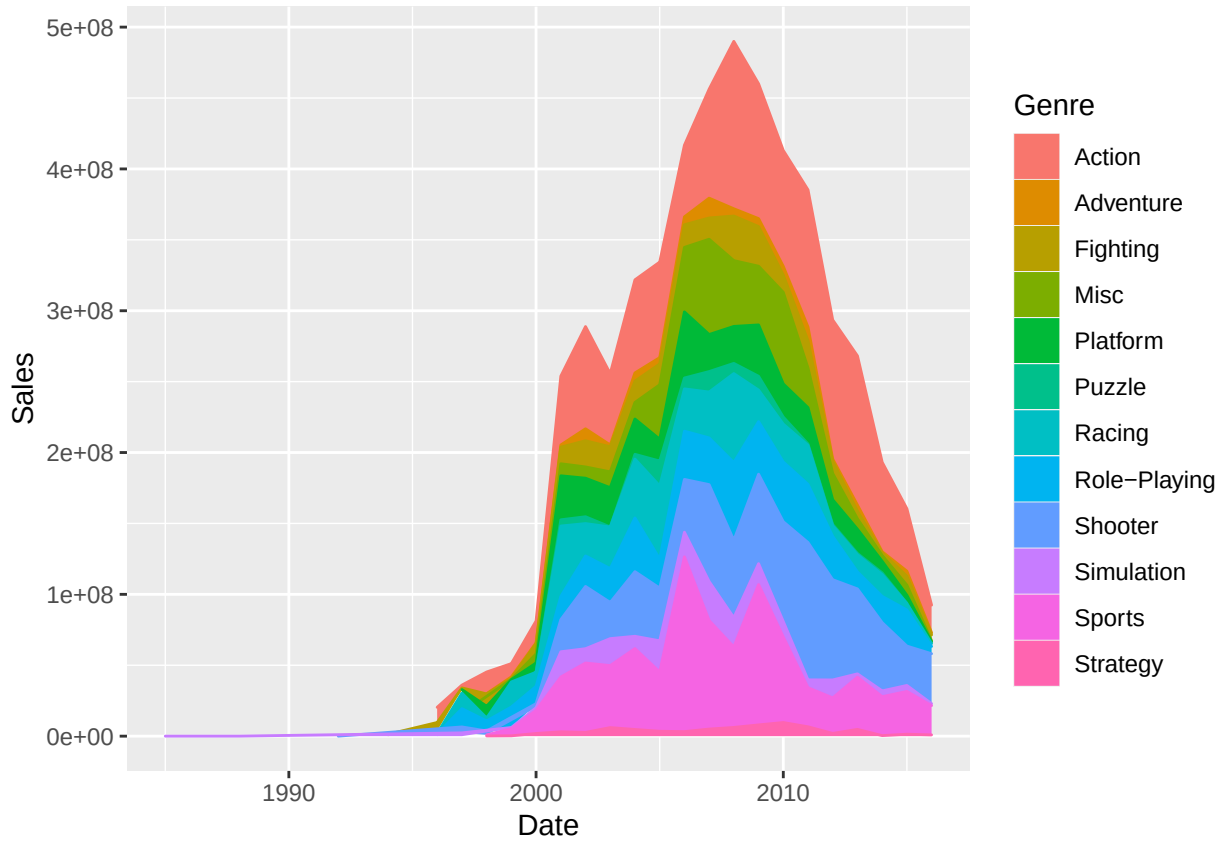
```
##                       Name Platform Year_of_Release    Genre Publisher
## 1               Wii Sports      Wii            2006   Sports  Nintendo
## 2            Mario Kart Wii      Wii            2008   Racing  Nintendo
## 3         Wii Sports Resort      Wii            2009   Sports  Nintendo
## 4  New Super Mario Bros.        DS            2006 Platform  Nintendo
## 5                 Wii Play      Wii            2006     Misc  Nintendo
## 6 New Super Mario Bros. Wii      Wii            2009 Platform  Nintendo
##   Critic_Score Critic_Count User_Score User_Count Developer Rating Global_Sales
## 1           76           51         79        322  Nintendo      E     82530000
## 2           82           73         82        709  Nintendo      E     35520000
## 3           80           73         79        192  Nintendo      E     32770000
## 4           89           65         84        431  Nintendo      E     29800000
## 5           58           41         65        129  Nintendo      E     28920000
## 6           87           80         83        594  Nintendo      E     28320000
##   SystemType       Date Salesfact GamesReleased
## 1    Console 2006-01-01  1109.473           294
## 2    Console 2008-01-01  1109.473           294
## 3    Console 2009-01-01  1109.473           294
## 4   Handheld 2006-01-01  1109.473           294
## 5    Console 2006-01-01  1109.473           294
## 6    Console 2009-01-01  1109.473           294
```

General trends and relationships between prediction parameters were explore.





The final cleaned dataset that has all the NA values dropped includes a total of 6984 rows. The data was partitioned into a training set and a validation set using a 90/10 split that leaves 691 rows in the validation set. The 90/10 split ratio was selected to try and keep the number of observations in the training set high to prevent poor representation during cross validation. Correlations between predictive parameters and Global sales were found to be generally low (<0.3). Interestingly, user count shows a significantly higher correlation to global sales than user score. This may indicate that the more sales a game has the more likely multiple users are willing to rate that game. On the other hand, critics who may be paid to rate video games, are more likely to rate games independent of game sales. The sales factor parameter that was created shows a correlation of 0.19 which is more significant than the games released parameter. There are a couple pairs of highly cross correlated parameters. Sales factor and games released have a Pearson's r score of 0.89 indicating a strong relationship. Sales factor is based on publisher sales, since a unit of sale would be 1 game it follows that these fields should show correlation. User score and critic score also show a relatively high correlation value (0.58). Not a surprising result, if a game is critically reviewed as good or bad it would be surprising for users to submit ratings that were wildly different. When training the final model, combinations of highly correlated variables were avoided. A look at a plot showing game sales per Genre over time shows a trend towards a temporal and genre-specific trend. Action, Misc, Shooter and sports games seem to sell the most globally. There is an obvious peak in sales in the late 2000's and a slight decline following the peak and indicates that global sales are temporally dependent.



The caret package was used to train the models using the Gradient Boosting Machine and the Random Forest methods. To prevent overtraining 5-fold cross validation was used on all the models tested. Models were iteratively trained with different parameters to test and find the model that returned the lowest RMSE. Parameters were added and removed based on model performance. When the model with the lowest RMSE was finally selected. It was then used to predict Global sales for the validation (final hold out) dataset and the resulting RMSE was evaluated.

Results:

Multiple different parameter combinations and model methods were tested resulting in a range of (cross validated) RMSE values from 1 804 205 – 1 553 156. The R^2 values for the cross validated models ranged from 0.1522 – 0.2981. GBM models performed better than random forest models with the same parameters. Interestingly, a random forest model returned the highest R^2 value of any model with 0.2981. However, that same model also returned the third highest RMSE score at 1704226, so it was discarded. The final model returned a cross validated RMSE value of 1 553 156 and an R^2 of 0.2338. This final model represents a 14% improvement over the worst model tested.

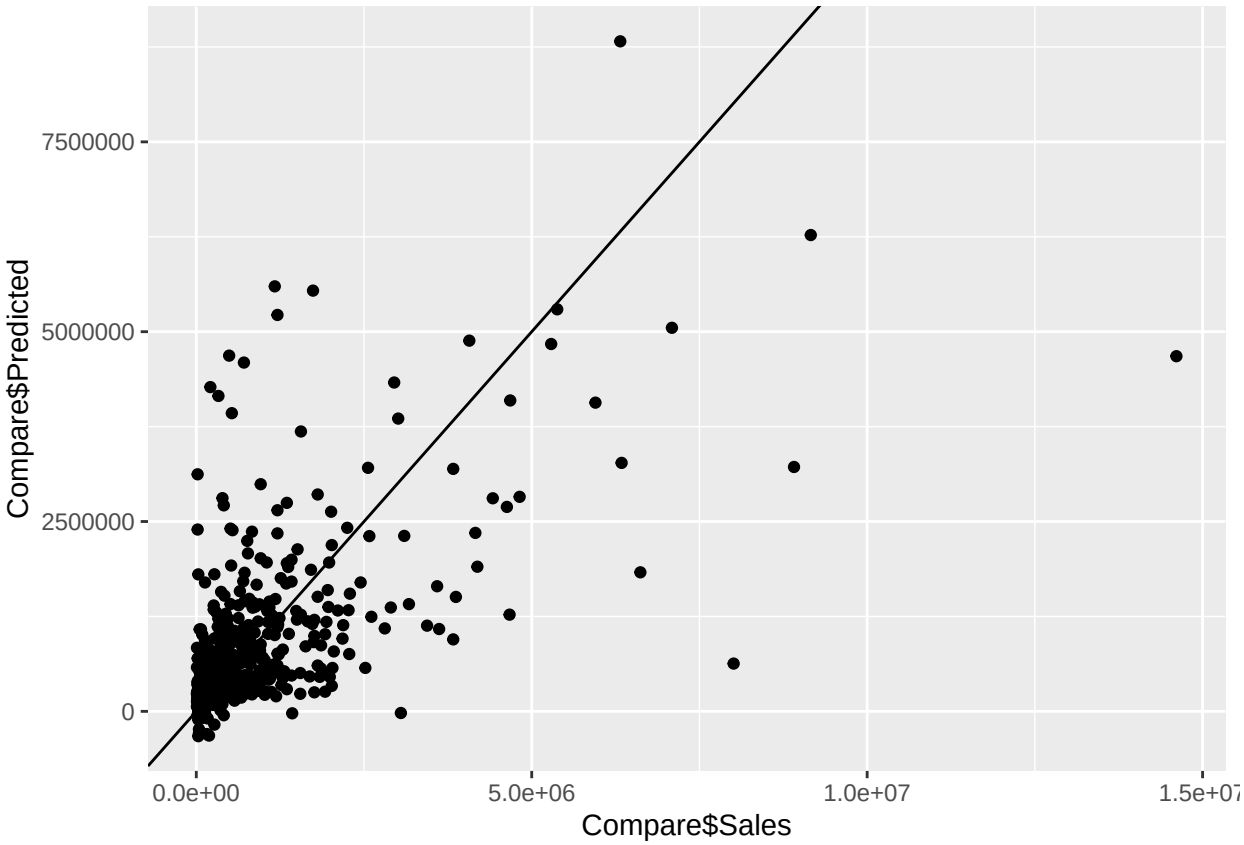
	RMSE	R_2
GBM_SYS_GEN_UC_CS	1722865	0.2153344
GBM_SYS_GEN_US_CS	1804205	0.1522106
GBM_SYS_GEN_UC_CC	1736055	0.1859624
GBM_PLT_GEN_UC_CS	1649025	0.2227138
GBM_PLT_GEN_UC_CS_YR	1643237	0.2228376
RF_PLT_GEN_UC_CS_YR	1704226	0.2981753
GBM_PLT_GEN_UC_CS_YR_RT	1591295	0.2273773

	RMSE	R_2
GBM_PLT_GEN_UC_CS_YR_GR_SF	1553541	0.2338737
GBM_PLT_GEN_UC_CS_YR_RT_SF	1553156	0.2338737

The parameters that performed the best in the 5-fold cross validation GBM model were Platform, Genre, User Count, Critic Score, Rating, Year of Release, and Sales Factor. System type was meant to replace platform and capture the similar variability; however, the model testing showed that using system type over platform tended to return higher RMSE and lower R^2 values. User count and critic score were significant variables in the final model as expected from their correlation to global sales. Rating was also a key indicator in sales with the everyone rating being most important Genre predictor for sales. Theoretically, games that are rated for everyone (E) have a larger intended audience and have higher sales potential. Sales factor was an important variable in model prediction and provides evidence to support the theory that publishers that are historically successful will continue to sell above the industry average. This parameter is possibly capturing the success of games that are released in a series of games. If the first game in a series sells well, the follow up game is likely to sell well based on the success of the first game alone.

```
## gbm variable importance
##
##   only 20 most important variables shown (out of 54)
##
##                   Overall
## User_Count      100.00000
## Critic_Score     25.90168
## RatingE          22.85801
## PlatformWii      19.50295
## Year_of_Release  15.47331
## Salesfact        14.41111
## PlatformPC       12.90053
## GenreSports       8.60689
## GenreMisc         1.66880
## RatingM           1.27620
## PlatformX360      1.15503
## GenrePlatform     0.98725
## PlatformDS        0.87424
## PlatformPS2       0.78921
## GenreAction       0.63834
## GenreRacing       0.52148
## PlatformPS        0.42555
## PlatformPS3       0.35112
## GenreShooter      0.12035
## PlatformWiiU      0.08722
```

When validated against the final hold out dataset the model performed better than in the cross validated samples and returned an RMSE value of 957 722. The mean sales for a game is around 770 000 copies but the standard deviation is almost 2 000 000. Looking at a plot of the predicted values vs true sales numbers shows that predictions are not bad when sales are low (up to about 1.5 million). Beyond that value, dots become much more dispersed and the model is more likely to over or under-estimate sales.



Looking at the poorly predicted games, we see some interesting trends.

	Name	Year	Platform	Sales	Predicted	Diff	DiffA	no0
1	Call of Duty: Black Ops	2010	X360	14610000	4677279.40	9932721	99327214677279.4	
4	The Sims 3	2009	PC	8010000	629225.64	7380774	7380774629225.6	
3	Mario Party DS	2007	DS	8910000	3218428.69	5691571	56915713218428.7	
6	Big Brain Academy	2005	DS	6620000	1831092.58	4788907	47889071831092.6	
111	Grand Theft Auto V	2015	PC	1170000	5596914.14	-4426914	44269145596914.1	
249	Fire Emblem: Radiant Dawn	2007	Wii	490000	4684796.02	-4194796	41947964684796.0	
421	Tony Hawk's Pro Skater 5	2015	PS4	210000	4270349.79	-4060350	40603504270349.8	
106	Boom Blox	2008	Wii	1210000	5220457.71	-4010458	40104585220457.7	
182	Silent Hill 3	2003	PS2	710000	4593457.34	-3883457	38834574593457.3	
320	SSX Blur	2007	Wii	330000	4153834.73	-3823835	38238354153834.7	
68	The Order: 1886	2015	PS4	1740000	5540679.75	-3800680	38006805540679.8	
230	Excite Truck	2006	Wii	530000	3926720.75	-3396721	33967213926720.7	
14	The Lord of the Rings: The Two Towers	2002	PS2	4670000	1273737.93	3396262	33962621273737.9	

	Name	Year	Platform	Sales	Predicted	Diff	DiffA	no0
671	Left 4 Dead	2008	PC	20000	3121852.77	-3101853	31018533121852.8	
28	The Sims 4	2014	PC	3050000	-21376.84	3071377	3071377	0.0
7	Halo 3: ODST	2009	X360	6340000	3272119.59	3067880	30678803272119.6	
2	Animal Crossing: New Leaf	2012	3DS	9160000	6272701.02	2887299	28872996272701.0	
22	Flash Focus: Vision Training in Minutes a Day	2007	DS	3830000	947654.79	2882345	2882345947654.8	
23	Guitar Hero: World Tour	2008	Wii	3620000	1084312.46	2535688	25356881084312.5	
8	Red Dead Redemption	2010	X360	6320000	8823545.60	-2503546	25035468823545.6	
289	Mass Effect 2	2010	PC	390000	2807646.53	-2417647	24176472807646.5	
676	Darksiders	2010	PS3	20000	2394500.25	-2374500	23745002394500.2	
20	MotorStorm	2006	PS3	3870000	1507236.43	2362764	23627641507236.4	
25	Tekken 4	2002	PS2	3440000	1128755.83	2311244	23112441128755.8	
278	Animal Crossing: Amiibo Festival	2015	WiiU	410000	2712626.46	-2302626	23026262712626.5	
17	Gran Turismo 5 Prologue	2007	PS3	4190000	1904587.52	2285412	22854121904587.5	
75	Pokemon Mystery Dungeon: Explorers of Sky	2009	DS	1560000	3685916.22	-2125916	21259163685916.2	
5	Mario Kart 8	2014	WiiU	7090000	5051268.28	2038732	20387325051268.3	
133	Far Cry 3	2012	PC	960000	2991518.40	-2031518	20315182991518.4	

The games on this list can easily be placed into two categories, a huge success or a bust. In other words, a number of these games wildly outperformed expectations including Call of Duty: Black Ops and The Sims 3. Some games wildly under-performed like Tony Haw's Pro Skater 5 which was a flop and failed to reach similar numbers to previous versions (Wikipedia, 2021). This is an example of a game getting a boost in prediction from success in previous series but failing to reach the same level of success. The Sales factor parameter would likely be the culprit in these scenarios. There are only 29 observations out of 691 total that have a residual value above 2 000 000. However, the magnitude of these residuals ranges from 2 031 518 to 9 932 721. If these outliers are removed from the RMSE calculation the RMSE drops to 528 373. A potential problem with the model is that it does not force positive sales predictions. Negative sales numbers are impossible, and 21 games receive negative sales values. A quick fix would be to coerce the negative predictions to the lowest reported sales value (10 000). When negative values are forced to 0 the RMSE only changes slightly and drops about 1000 copies so overall not a major source of error.

Conclusion:

Video game sales are highly variable and extremely complicated. Using a dataset containing ~17000 video games with ratings. After cleaning up the data, some feature engineering the final training dataset included ~6900 rows. A Gradient Boosted Machine model was trained on a training dataset and cross validated for RMSE and R^2 values. The final model predicted global sales for the validation data set with an RMSE of ~950 000. This is a large discrepancy relative to the mean sales number of ~770 000. A difference in predicted to actual sales of close to a million could be the difference between success and failure for a smaller company. The main source of error in the final model is the inability of the algorithm to predict outlier's success or failures in the dataset. Some video games wildly outperform expectations while others fail to reach the level of previous games in a series. Although there are only 26 (depending on where you apply the cut-off) outliers, they come close to doubling the output RMSE. With these outlier values removed the

final RMSE is only ~550 000. Given the large swing caused by the outliers, and the huge variety in the dataset (Global Sales standard deviation is almost 2 000 000 units) the final model does an adequate job of predicting sales for the average video game.

The final model indicates that high sales values are mainly driven by the number of users that rate the game, critic score, a rating of E or M, being released on a major platform (Wii, Xbox, PS etc), specific Genres and a Sales factor. What insights are most important to a publisher or developer? A developer (if they are not the publisher as well) will sell the most copies if they can link up with a publisher with a history of success (high historical sales numbers). The best sold video games were rated Mature or Everyone. Games released on major platforms like Xbox, Wii, Play Station (and all their versions) tend to sell more copies than smaller platforms like some handhelds. Action, Sports, Misc, Racing, and Shooting games tend to sell more copies than other genre types.

Further improvements to the model would include forcing the model to predict positive sales numbers. For a more realistic approach the minimum sales value that can be predicted could be set to the minimum sales number (10 000). That improvement would give us a more realistic prediction range and likely improve prediction RMSE. There are improvements that could be made through more feature engineering. Creating a parameter that captures inclusion in a gaming series could potentially improve upon and replace the sales factor parameter. Further data clean-up to extend the dataset would be the most helpful addition. Increasing the training and test set size could help improve model training and ensure confidence in model performance reporting.

References:

1. Video Game Industry Statistics In 2020, WePC, retrieved January 2nd 2021, https://www.wepc.com/news/video-game-statistics/
2. Video Game Sales With Ratings, Rush Kirubi, 2016, retrieved December 27th 2020, https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings
3. Metacritic, Metacritic, retrieved January 2nd 2021, https://www.metacritic.com/
4. ESRB Home, ESRB, retrieved January 2nd 2021, https://www.esrb.org/
5. Tony Hawk's (Series), Wikipedia, retrieved January 2nd 2021, https://en.wikipedia.org/wiki/Tony_Hawk%27s_(series)