

Group 4

Kiana Dane, James Gomes,

Avery Martin, Johnny Whitaker

OBA 410, T/R 6PM

## **Video Game Sales Analysis**

### **Achievement Goals**

The goal of this analysis is to predict performance and global sales of a new video game across four markets: the United States, Japan, Europe, and the rest of the world (Other). Using a combination of machine learning tools across several different programs, our team analyzed the ability of these programs to accurately predict sales within each individual market as well as total worldwide sales. The results of this analysis will support better predictions of which games will achieve success by region, genre, and platform, which could be significant to video game developers and publishers.

### **Relevance of Analysis**

Due to the exponential growth of the video game industry, we feel that our topic of choice is very relevant when it comes to making a meaningful prediction within a dataset. The task of predicting video game sales can lead to better business decisions within the video game industry for several areas of business. Marketing, finance, and R&D all can benefit heavily from an accurate prediction for future video game sales. In 2018, the video game industry reached a record-breaking sales figure of \$43.4 billion dollars. This represents an 18% growth from 2017 which is almost unheard of. Similarly, 2017 experienced 16% growth from the year prior. This confirms just how relevant the industry and its growth is. Along with these encouraging numbers, the global video game market is expected to be worth over \$90 billion by 2020. Aside from the financial implications of video games, there is a sense of unity and belonging that comes with the industry. In the 80's, consumerism expanded as homeowners began to buy consoles and entertainment systems for the home and everyday life. As video games became more popular, a cultural shift began where people could come together and spend meaningful time with one another using video games. Fast forward to 2019 and the same holds true. Currently in the United States, 45% of gamers are women, this proves that video games are an inclusive experience. Along with this, 60% of Americans play video games daily and 64% of American households are home to at least one person who plays video games daily. Games are so much more than a way to escape the unpleasant realities of everyday life, they offer a sense of unity and belonging for so many people around the world.

### **The Data Source**

The dataset our team used in our analysis is from kaggle.com (see exhibit 1). It contains data of over 16,500 games. The unit size in this dataset is in millions of units sold. Only games that sold more than 100,000 copies are included. Variables of these data include:

- Global Rank
- Name
- Platform
- Year
- Genre
- Publisher

One of the most unique things about the dataset we chose to analyze is that it contains data from 1980 to 2014. The industry itself has only been in existence since the late 70's, so a dataset including data from the last 34 years encompasses almost the entire sales performance history of the industry.

vgsales.csv (402.59 KB)

	# Rank	A Name	A Platform	A Year	A Genre	A Publisher	# NA_Sales	# EU_Sales	# JP_Sales	# Other_...	# Global_...
1	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
4	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
5	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37
6	6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
7	7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
8	8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
9	9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
10	10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
11	11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11	1.93	2.75	24.76
12	12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
13	13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1
14	14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72
15	15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22

## Tools and Methods

To conduct our analysis, we utilized SAS and Python using the Pandas and ScikitLearn packages. In SAS, we imported the excel table and converted it into a SAS table using the Enterprise Guide then constructed our nodes in a diagram. The data was partitioned into 40% training and 60% validation with simple random sorting and a random seed of 420, then used to conduct an ensemble analysis combining the results of a LARS (Lasso Regression) model and a Decision Tree. The results of the models, particularly the Mean Squared Error, was examined to determine the strength of the models. To further test the strength of the models we conducted a prediction using the same dataset, sans target variable, to use as scoring data along with an independent prediction on *The Witcher 3: Wild Hunt (PS4)*.

In Python, specifically the Jupyter coding environment, we began by importing the necessary packages; Pandas, which is a data-structuring package allowing for the construction of matrices and tables, as well as the importation of Excel and similar csv files into the coding environment; ScikitLearn which is a comprehensive suite of machine-learning tools used to construct, run, and assess the models in the coding environment. Once the packages were imported, the dataset was imported and the categorical variables such as Publisher and Platform were given category codes for more effective analysis and easier interpretation of the results.

```
: import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge
from sklearn.tree import DecisionTreeRegressor
from sklearn.tree import export_graphviz
from sklearn.ensemble import RandomForestRegressor

vg=pd.read_csv('vgsales0.csv')
vg['Platform']=vg['Platform'].astype('category')
vg['Platform_CAT']=vg['Platform'].cat.codes
vg['Year']=vg['Year'].astype('category')
vg['Year_CAT']=vg['Year'].cat.codes
vg['Genre']=vg['Genre'].astype('category')
vg['Genre_CAT']=vg['Genre'].cat.codes
vg['Publisher']=vg['Publisher'].astype('category')
vg['Publisher_CAT']=vg['Publisher'].cat.codes
vg.head()
```

	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Platform_CAT	Year_CAT	Genre_CAT	Publisher_CAT
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74	26	26	10	359
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	11	5	4	359
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82	26	28	6	359
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00	26	29	10	359
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	5	16	7	359

The models were then constructed by first defining the necessary variables, classifying them appropriately as X (input) or Y (target) variables, and defining the range included from the data source. Then the data was divided into training and testing sets; 50/50 for the North American sales models but it was adjusted individually per model to determine the best results. Once the data was partitioned, the models were named, created and trained to the training portion of the data. Individual factors for the models, such as alpha and max\_leaf\_nodes, were adjusted individually for each model to produce the best results. The models were then run, with print functions below to display the  $R^2$  value for each model on both training and test data to identify any potential overfitting.

```

XNA,yNA=vg.iloc[:,[6,7,8,9,10,11,12,13]],vg.iloc[:,5]
XNA_train,XNA_test,yNA_train,yNA_test=train_test_split(XNA,yNA,train_size=0.5,test_size=0.5,random_state=0)
lin_NA=LinearRegression()
lin_NA.fit(XNA_train,yNA_train)
lasso_NA=Lasso(alpha=0.002)
lasso_NA.fit(XNA_train,yNA_train)
ridge_NA=Ridge(alpha=15)
ridge_NA.fit(XNA_train,yNA_train)
dt_NA=DecisionTreeRegressor(random_state=0,max_leaf_nodes=20)
dt_NA.fit(XNA_train,yNA_train)
rf_NA=RandomForestRegressor(random_state=0,n_estimators=125)
rf_NA.fit(XNA_train,yNA_train)
print('North America Sales')
print('Linear R^2 on Train Data:{:.3f}'.format(lin_NA.score(XNA_train,yNA_train)))
print('Linear R^2 on Test Data:{:.3f}'.format(lin_NA.score(XNA_test,yNA_test)))
print('Lasso R^2 on Train Data:{:.3f}'.format(lasso_NA.score(XNA_train,yNA_train)))
print('Lasso R^2 on Test Data:{:.3f}'.format(lasso_NA.score(XNA_test,yNA_test)))
print('Ridge R^2 on Train Data:{:.3f}'.format(ridge_NA.score(XNA_train,yNA_train)))
print('Ridge R^2 on Test Data:{:.3f}'.format(ridge_NA.score(XNA_test,yNA_test)))
print('DT Accuracy on Train Data:{:.3f}'.format(dt_NA.score(XNA_train,yNA_train)))
print('DT Accuracy on Test Data:{:.3f}'.format(dt_NA.score(XNA_test,yNA_test)))
print('RF Accuracy on Train Data:{:.3f}'.format(rf_NA.score(XNA_train,yNA_train)))
print('RF Accuracy on Test Data:{:.3f}'.format(rf_NA.score(XNA_test,yNA_test)))
print('')

```

The feature importances were then determined using the Random Forest models, which were determined to be the most robust, and displayed using the print function.

Finally, predictions were conducted by creating entities in the coding environment with the characteristics of specific games, removing the target variable from the data set and using the Random Forest models to predict the values. Specifically, the predictions used *The Witcher 3: Wild Hunt* on PS4, XBOX ONE, and PC to obtain values for North America Sales, Europe Sales, Japan Sales, Other Sales, and Global Sales.

```

print('The Witcher 3: Wild Hunt (PS4)')
Witcher3NA=[2,0.21,0.56,3.73,18,35,7,347]
print('NA Sales:',rf_NA.predict([Witcher3NA]))
print('Actual:0.96')
print('')
Witcher3EU=[0.96,0.21,0.56,3.73,18,35,7,347]
print('EU Sales:',rf_EU.predict([Witcher3EU]))
print('Actual:2')
print('')
Witcher3JP=[0.96,2,0.56,3.73,18,35,7,347]
print('JP Sales:',rf_JP.predict([Witcher3JP]))
print('Actual:0.21')
print('')
Witcher3Other=[0.96,2,0.21,3.73,18,35,7,347]
print('Other Sales:',rf_Other.predict([Witcher3Other]))
print('Actual:0.56')
print('')
Witcher3Global=[0.96,2,0.21,0.56,18,35,7,347]
print('Global Sales:',rf_Global.predict([Witcher3Global]))
print('Actual:3.73')
print('')

```

For both methods, once a top performing model had been determined, the various feature importances and detailed statistics could be examined for conclusions about causality and patterns in the data.

## Results

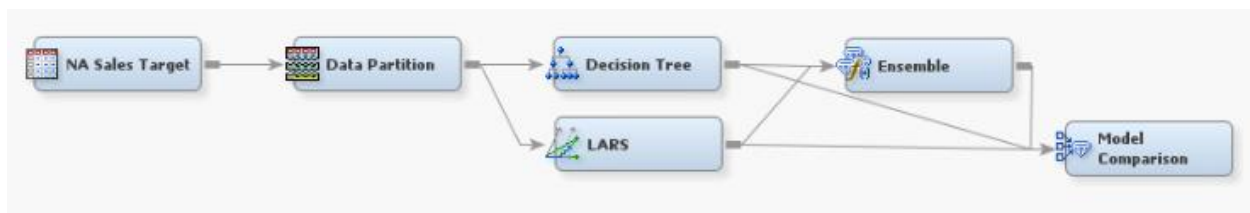
The North American region has the largest amount of video game sales in the dataset, and the majority of the industry is American. Although, we built models to predict sales in each region, for the sake of this report the focus will be on the results for the North American region and Global. This is mainly due to runtime errors in SAS, which led to the inability to run any predictions on the EU, Japan and Other region sales as target variables.

## SAS

After creating many models, it was ultimately decided that an ensemble of a decision tree and a LARS regression model would be able to make strong predictions on the sales of video games in various regions. The final model is shown below, along with a model comparison of the mean square errors (MSE) of the ensemble and the other two models that make it up.

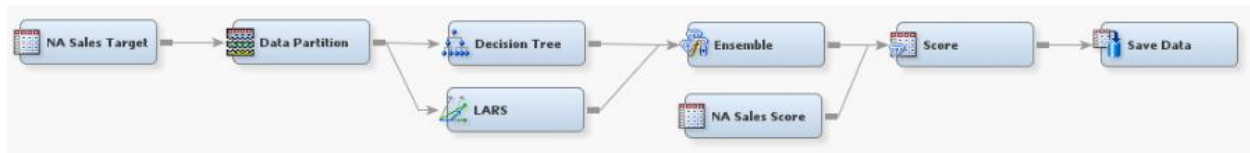
When looking at the model comparison, notice that the MSE is not the best and is larger than that of the LARS model by multiple factors of ten. The LARS regression model was not used by itself in making predictions because of how small the validation error is. It was decided that the error would be too

small to make consistent and adequate predictions on new data; therefore the LARS model would not be robust enough. In order to increase the robustness of the model it was ensembled with the decision tree.



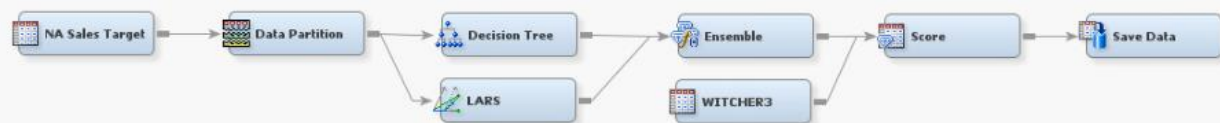
VGSales Model Validation MSE	
Model	Validation MSE
LARS	3.171E-05
Ensemble	0.1249
Tree	0.1928

With the final model, a score node and a save data node was attached to the end of the model and ran twice: once to predict the North American Sales (shown below) and another to predict the Global Sales. Surprisingly, both models predicted better, in terms of MSE, than SAS had shown for the validation scores. This could have been due to the random seed used when training and validating the models and could highlight a possible, slight instability in the model.



VGSales Model MSE			
Region	Validation MSE	Score MSE	Difference
North America	0.0485	0.0348	(0.0137)
Global	0.1249	0.0833	(0.0416)

After predicting the North American and Global sales for all the video games in the original data set, the model was used to predict the sales for a specific video game: *The Witcher 3: Wild Hunt (PS4)*. Both predictions made on *The Witcher 3* were relatively close to the actual sales numbers with North America predictions off by 0.22 (220,000 units) and Global by 0.13 (130,000 units). Though, the North America prediction is worse than the Global prediction due to the actual number of sales being smaller but the prediction being off by a larger amount, it is still important to note that the prediction is still decent. However, what is more important to note is how close the Global prediction is to the actual amount. In terms of millions of units sold, being off by only 130,000 unit is a good prediction.



<i>The Witcher 3: Wild Hunt (PS4)</i>			
Region	Actual	Prediciton	Difference
North America	0.96	1.18	0.22
Global	3.73	3.86	0.13

### Python

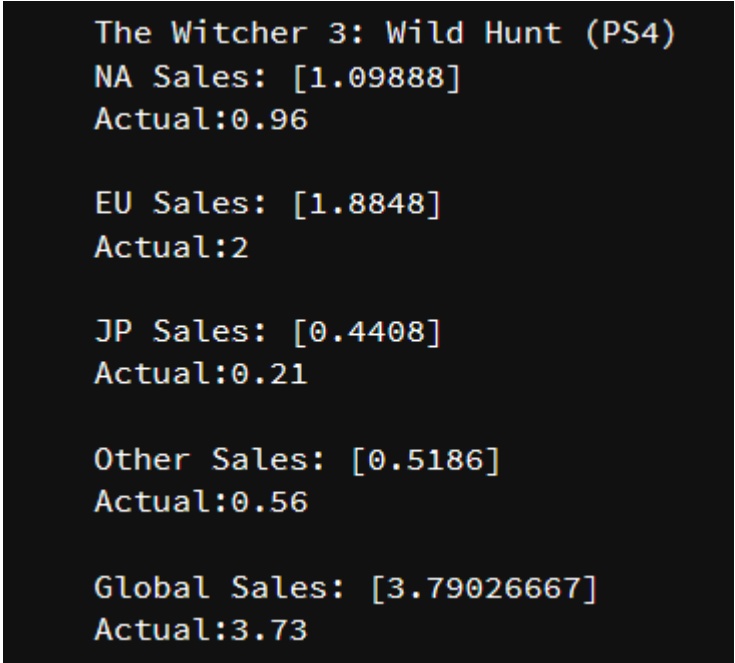
After creating, running and scoring quite a few models, it was eventually decided that the random forest model was the most robust model and would be used in any subsequent predictions. This is due to some of the models performing so well, as seen in the scores for North American sales shown below, that there is cause for concern that the regression models are overfitted. This concern comes from how close the models' scores are to one and also how close the train and test scores are to one another. Ultimately, we are at least 85% confident in predictions made by the models in Python, which is a less than the score of the random forest however the decrease in confidence is due to the possibility of overfitting.

The feature importances of the random forest highlight an interesting importance. It is not unexpected that the highest importance would be that of the Global Sales and other region sales, however the interesting importance is that of the Year. What is interesting about looking at the year the video game was released is that it can highlight the idea of a video game increasing in popularity over time. With that extra time and increase in popularity, one could infer that it would play and impact the overall sales of the video game.

	TRAIN/TEST			
<b>North America Sales</b>	0.5/0.5			NA Importance
Linear R^2 on Train Data	1			Global_Sales 0.877840
Linear R^2 on Test Data	1			EU_Sales 0.054657
Lasso R^2 on Train Data	0.999			JP_Sales 0.027775
Lasso R^2 on Test Data:	0.998			Year_CAT 0.014816
Ridge R^2 on Train Data:	0.999			Other_Sales 0.013778
Ridge R^2 on Test Data:	0.999			Platform_CAT 0.005195
DT Accuracy on Train Data:	0.952			Genre_CAT 0.004611
DT Accuracy on Test Data:	0.811			Publisher_CAT 0.001329
RF Accuracy on Train Data:	0.972			
RF Accuracy on Test Data:	0.884			

To test our model using a prediction, we removed all three *The Witcher 3: Wild Hunt* records (PS4, XBOX ONE, and PC), predictions on the sales figures were made for the PS4 version. Relatively, all the predictions are really close to their actual values, except for the Japan sales prediction. The prediction may only be about 0.2 above the actual sales figure, however relative to the actual sales (0.21 million units) this is a terrible prediction. With an actual error of over 100%, there seems to be something that is not quite working right with the Japan Sales model, at least when predicting *The Witcher 3: Wild Hunt (PS4)* sales.

It's important to also look at how accurate the Global Sales prediction was. The Global Sales prediction shows how robust the model is as it predicted *The Witcher 3: Wild Hunt (PS4)* to sell 0.06 million (60,000) or about 1.6% more units than it actually did. Ideally, a business would likely want that forecast to be less than the actual amount, due in part to the conservative nature of business, but in terms of millions of units, being off by 60,000 is not very much and constitutes a very close prediction.



```
The Witcher 3: Wild Hunt (PS4)
NA Sales: [1.09888]
Actual:0.96

EU Sales: [1.8848]
Actual:2

JP Sales: [0.4408]
Actual:0.21

Other Sales: [0.5186]
Actual:0.56

Global Sales: [3.79026667]
Actual:3.73
```

## Conclusion

The video game industry is growing. Within the next few years, it will achieve annual sales of \$50 billion. With such a rapidly growing market, sales performance forecasting will become increasingly important for game developers and publishers as they continue to learn the complexities of consumer preferences across many different markets. Because the industry is so young, it enables data scientists to perform analysis of the performance of nearly every modern video game. This cohesive collection of data from the last 34 years also enables researchers to make predictions about how factors like release date, genre, and platform significantly influence performance of new games. Our team's analysis of this dataset led us to the conclusion that the model we created was able to predict global sales extremely well, while its ability to predict sales within individual markets was limited due to complex circumstances. With more time, however, our team would take the opportunities to further examine



these complexities in the data in order to provide a better analysis. For example, SAS could not complete some regional analysis models to the end and showed us run time errors. Additionally, Japanese consumers typically prefer Sony Playstation games, so low Xbox game sales skewed the prediction for that market. Our dataset was comprehensive and allowed us to achieve meaningful predictions while proving that there is still much more information to be derived from it.