Luke Murphy

Machine Learning

12/8/2022

**Machine Learning Final Paper**

1. **Executive Summary**

The company Nintendo had asked me to figure who has been their best game developer over the years in order to use that same developer to create a new video game that will do great in sales. I figured this out by using hierarchical clustering through data frames and also summaries of data using R studio to discover who this developer is in order to help Nintendo. I discovered that Nintendo EDP was the most profitable game developer for Nintendo by using the techniques listed before. The complete case study and technical information is down below.

1. **Problem**

I am a business analyst that the company Nintendo has hired to analyze their numerous numbers of bestselling games they have had over the years. Nintendo is historically known for producing some of the most well-known video games on the market. These include Mario, Pokemon,The Legend of Zelda, Animal Crossing and numerous others. According to Nintendo history, “Nintendo has sold more than 5.3 billion video games and over 800 million hardware units globally” (Nintendo,2015). There ask for me as a business analyzer is to discover which of there games have produced the most sales based on developer. Their goal for me is to help them find out which developers have produced the most for them in order to ask that same developer company to help produce a new game for them.

1. **Techniques**

So initially, I was thinking in which way I should organize the dataset of all the Nintendo games. Whether regression would be better or something like clustering to get a better visualization of all the components of the developer and sales data. I ended up choosing clustering. I loaded the dataset the company provided me into my data programmer R studio. After loading the data, I looked at some simple variable measures for the sales made out of all the games provides. I discovered that the minimum number of sales a game made was 1,000,000 dollars, the median made was 2,030,000 dollars, the average or mean 4,584,888 dollars and a max of 82,900,000 dollars. This date is extremely helpful to me in a way that it helps me get and idea of the data to help me break it down a bit. One thing to notice is as the date started reaching present time the sales amounts increased drastically for the company. So, what I did then was created a data frame of the data that didn’t include unnecessary categories including the publisher of the game, the release date, the platform or genre. So, after everything was all said and done the data frame ended up listing the developer, the game and the sales made off that game. I then edited this data frame in order to then discover only the games and developers that made games over 10,000,000 dollars. This then shrunk my data set observations from 430 to 31 in order to make it easier to process the data. After finally organizing all my data, I finally got into the cluster. I used a hierarchical cluster to display my data and used Euclidean distance to measure it. I used the developers as the x-axis label. The data is clustered in way that shows us which games by developer had similar sales. It seems the cluster shows one major outlier which is Wii sports developed by Nintendo EAD (group 2). Therefore, I will not include it in the final calculation. After created the original cluster, I created a DIANA to show the root to the leaf of the data. This was very much similar to the original cluster. Then to organize the clusters neatly I created a cluster dendrogram in order to show the actual division of the four clusters. This showed me that the bottom half of the sales data is more closely similar in sale compared to the higher sales prices. So, we can exclude that large cluster of those lower prices of which developer we should choose.

1. **Analysis**

I discussed some of the results in the approach but to continue to the main results I will finish here. So, continuing where I left off, when we calculate the total sales of the top part of the clusters, we get the following calculations. The top part of the clusters is identified as the red and green boxes on the dendrogram. The top developers in this half are Nintendo EPD, Nintendo EAD (group 2), Nintendo EAD (group 1), and Bandai Namco Studios, Sora LTD. Out of all these companies Nintendo EPD had a total of 4 of the 8 games in the top part of the sales cluster. It had a total of 134,020,000 dollars in sales.

1. **Conclusion**

After the final results and calculations, out of all the four top developers, Nintendo EPD was the best. They produced games that gained a total of 134,020,000 dollars in sales which about 7% of the total sales made of Nintendo games all time. It also represents about 54% of the total sales among the top 8 cluster of sales and developers. So therefore, I recommend Nintendo would be wisest to use Nintendo EDP as developer for there next upcoming game.

1. **Appendix**

GitHub link to data set:  <https://github.com/You-now-Who/dataset/tree/main/List%20of%20best%20selling%20nintendo%20videogames>

Nintendo. (2015). *Nintendo History*. Nintendo of Europe GmbH. https://www.nintendo.co.uk/Hardware/Nintendo-History/Nintendo-History-625945.html

Link to code: <https://github.com/Murphmania/64060_-lmurph25/blob/5c399224f1e6a34940414acda0b7d4e6ab76b436/Final/Machine%20Learning%20Final.Rmd>

Code

---

title: "Machine Learning Final Project"

output:

pdf\_document: default

html\_document: default

date: "2022-12-08"

---

```{r}

library(dbscan)

```

```{r}

library(factoextra)

```

```{r}

library(fpc)

```

```{r}

library(tidyverse)

```

```{r}

library(ISLR)

```

```{r}

mydata<-read.csv('Nintendo.csv')

```

```{r}

summary(mydata$Sales)

```

```{r}

df<-mydata[,-c(3,4,6,7)]

df<-df[df$Sales >= 10000000,]

summary(df)

```

```{r}

show(df)

```

```{r}

d<-dist(df, method = "euclidean")

hcl<-hclust(d,method = "complete")

plot(hcl,labels = df$Developer, cex = 0.55, hang = 0.2)

```

```{r}

hc\_diana<-diana(df$Sales)

hc\_diana$dc

```

```{r}

pltree(hc\_diana,labels=df$Developer,cex=0.55,hang=0.2)

```

```{r}

df<-df

d<-dist(df,method = "euclidean")

hc\_complete<-hclust(d,method = "complete")

plot(hc\_complete, cex=0.55)

rect.hclust(hc\_complete, k=4, border = 1:4)

```