IST707 – Data Analytics

My Research Project Title: Data Analysis of video games sales

My Name: Zhou, Zeyang

My Email Address: zezhou@syr.edu

## Introduction

### Background

The trend of the desire to take video games as a leisure time entertainment has been intensified in recent times. A large number of people show their huge enthusiasm in exploring the virtual world in video games. The appearance of video games can be traced back to 40 years ago. Japanese game giants, Sega and Samurai opened the global market and the popularity of them became a nice template for other video game makers who intend to succeed in the global market.

Due to Covid-19, video game market is becoming more popular than before since it changes the lifecycle of most people. Now the gaming industry is booming and, according to Global Industry Analysts, the global market for Video Games estimated at US$156.8 Billion in the year 2020, is projected to reach a revised size of US$293.2 Billion by 2027, growing at a CAGR of 9.3% over the analysis period 2020-2027.

However, there are some questions for video game manufacturers and investors such as which video games are more popular in Europe or which platform should they release their games on that can maximize their profits. As a video gamer, I would like to do data analysis about video game sales and solve these problems.

### Dataset

The dataset I will use in this study is from Kaggle (2016) which contains a list of video games with sales greater than 100,000 copies in different areas from 1980 to 2016. It was generated by a scrape of [vgchartz.com](www,vgchartz.com). The fields in the dataset include Rank - Ranking of overall sales, Name - The games name, Platform - Platform of the games release (e.g., PC, PS4, etc.), Year - Year of the game's release, Genre - Genre of the game, Publisher - Publisher of the game, NA\_Sales - Sales in North America (in millions), EU\_Sales - Sales in Europe (in millions), JP\_Sales - Sales in Japan (in millions), Other\_Sales - Sales in the rest of the world (in millions), Global\_Sales - Total worldwide sales. Its size is 2 MB and there are 16,598 records.

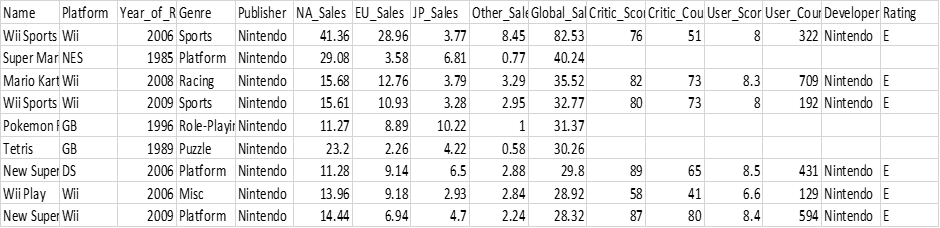


Figure.1 Raw dataset

### Objective

This final report will show a complete procedures of data mining with machine learning models including data preprocessing, descriptive analysis, visualization, and analysis for results generated by different machine learning models. Four machine learning algorithms planned to use are KNN, SVM, Naïve Bayes and Apriori.

The software R studio and weka will be used to predict the future hit games based on the past global sales or sales in other areas. In addition, Apriori will be used to detect the correlation between high sales video games and their features (e.g., genre = “action” or platform = “2600”)

## Descriptive Analysis

First, with string () function and summary () function, the data types and statistical results can be shown directly as in the figure 2 and figure 3 below.

In this case, the dataset contains 16598 records with 11 variables. Based on these two figures, I can detect outliers, process missing values, convert datatypes and preprocess them for machine learning. It seems that sales in different area can be taken as prediction results and rank, name, platform, year, genre, and publisher are the inputs to the models. Also, in the statistical results, no apparent outlier is found.

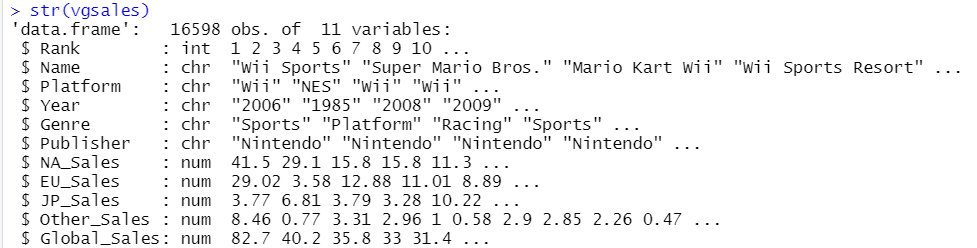


Figure.2 Data types

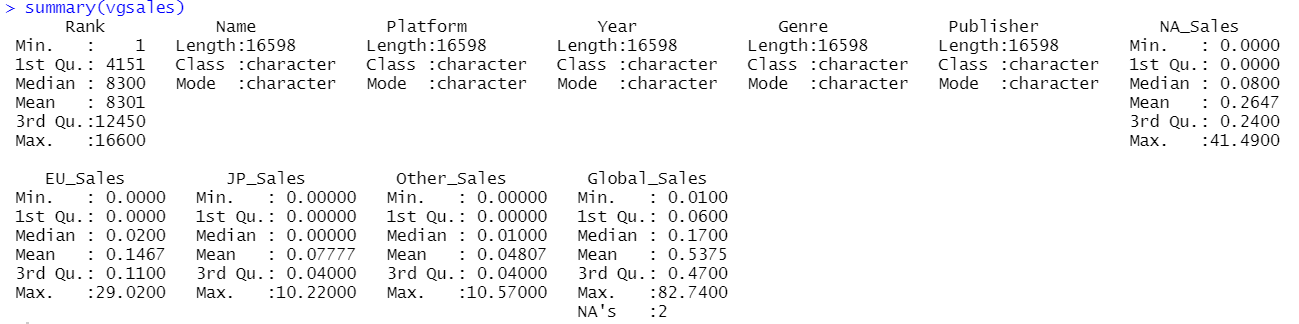


Figure.3 statistical results

In this case, the dataset contains 16598 records with 11 variables. Based on these two figures, I can detect outliers, process missing values, convert datatypes and preprocess them for machine learning. It seems that sales in different area can be taken as prediction results and rank, name, platform, year, genre, and publisher are the inputs to the models. Also, in the statistical results, no apparent outlier is found. So, the further cleaning and preparation is still required.

## Data Cleaning and Data Preparation

### Processing missing values

First, we would like to process missing values in this dataset. With locate () function, there are two missing values in Global sales column. Since Global sales is the sum of video games sales in NA, in Europe, in Japan and in other areas, I replace them with the total sales.

### Converting data type

“Year” means released year in this dataset. Its data type should be integer but not character. So, I convert its data type to the correct one.

### Outliers

There are two outliers in this dataset. They are both in the year column, too. 

By the unique () function, we can get the result above. Since this dataset contains video games sales from 1980 to 2016, video games sales in 2017 and 2020 are all outliers. So, I remove them from the dataset.

### Training set and test set

I split the raw dataset into two subsets based on the timeline because I would like to train the machine learning model to predict hit video games in the future. One is the training set with 60% of the data and the other is the test set with the rest 40% data.

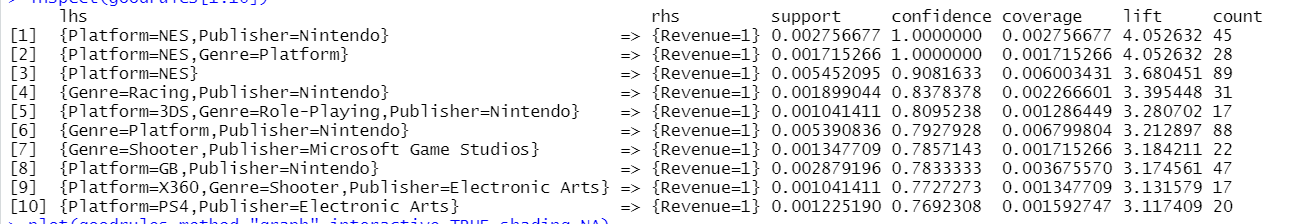
## Machine learning models and Analysis

### General preprocessing before

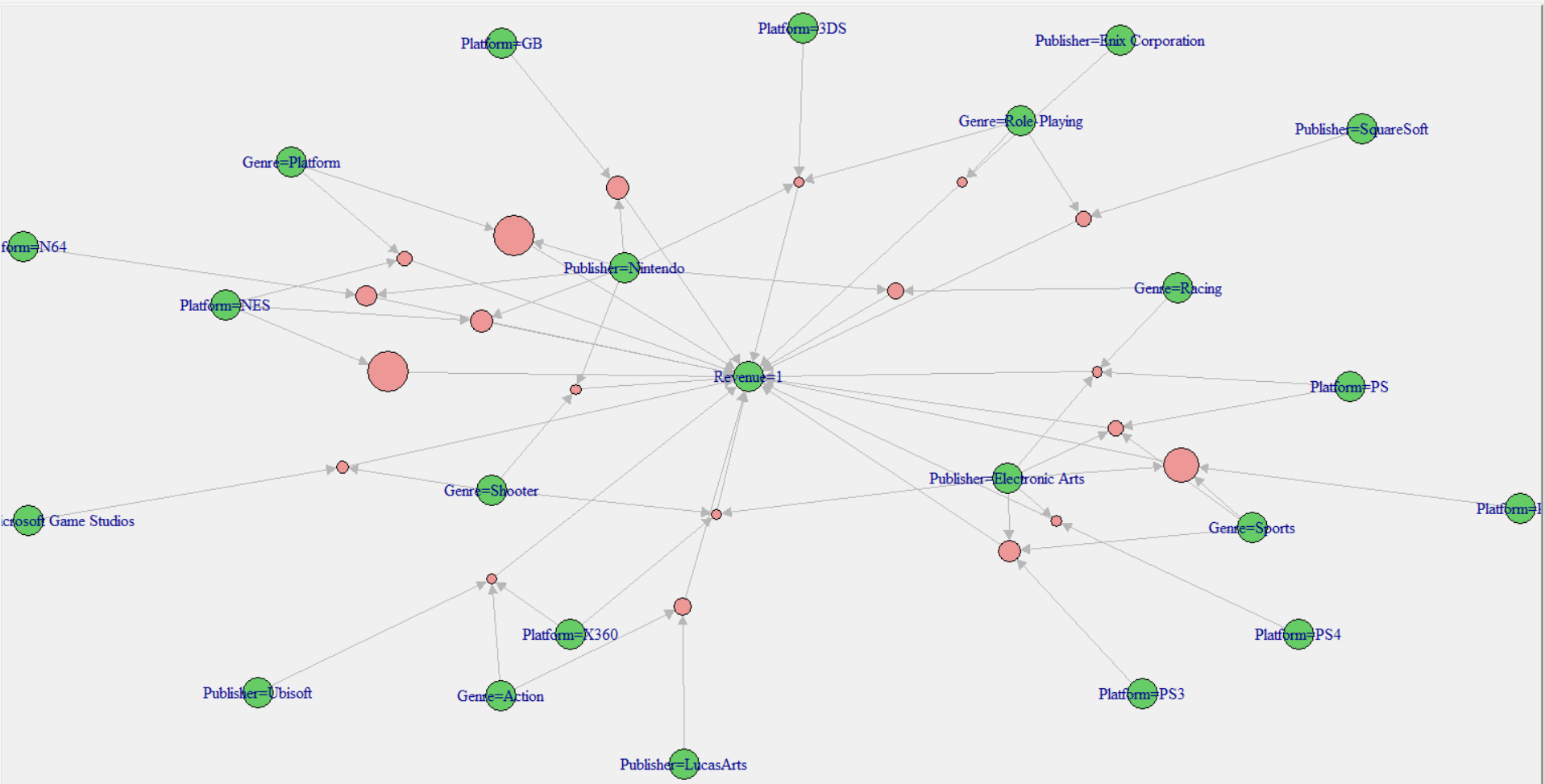
Before running machine learning algorithm in R studio, we need to preprocess it for better classification and fit the prerequisite of these algorithms. Since the values in “Global\_sales” cannot be classified directly, I add a new column called “Revenue” and its value depends on its “Global sales” that if its global sale is bigger than the medium sale, it is a high sales video game and vice versa. Now the target column is changed from “Global sales” to “Revenue” which can make it a simpler classification. (R code will be attached in the submission)

### Association rules

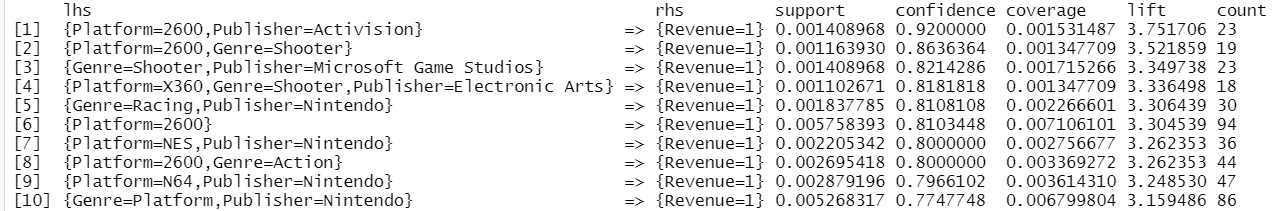
Before running machine learning algorithm in R studio, we need to decide the confidence and support of our Apriori model. To avoid too many unrelated or weak rules, I set the support to 0.01 and confidence to 0.06, the “rhs” to “Revenue = high” and just show the top 10 rules in this report. And what I want to find is the relationship between sales in different area and their features. So, I have five different datasets which remove the unrelated columns, and they are global sales, sales in NA, sales in EU, sales in JP and sales in other areas. The results are as follow:



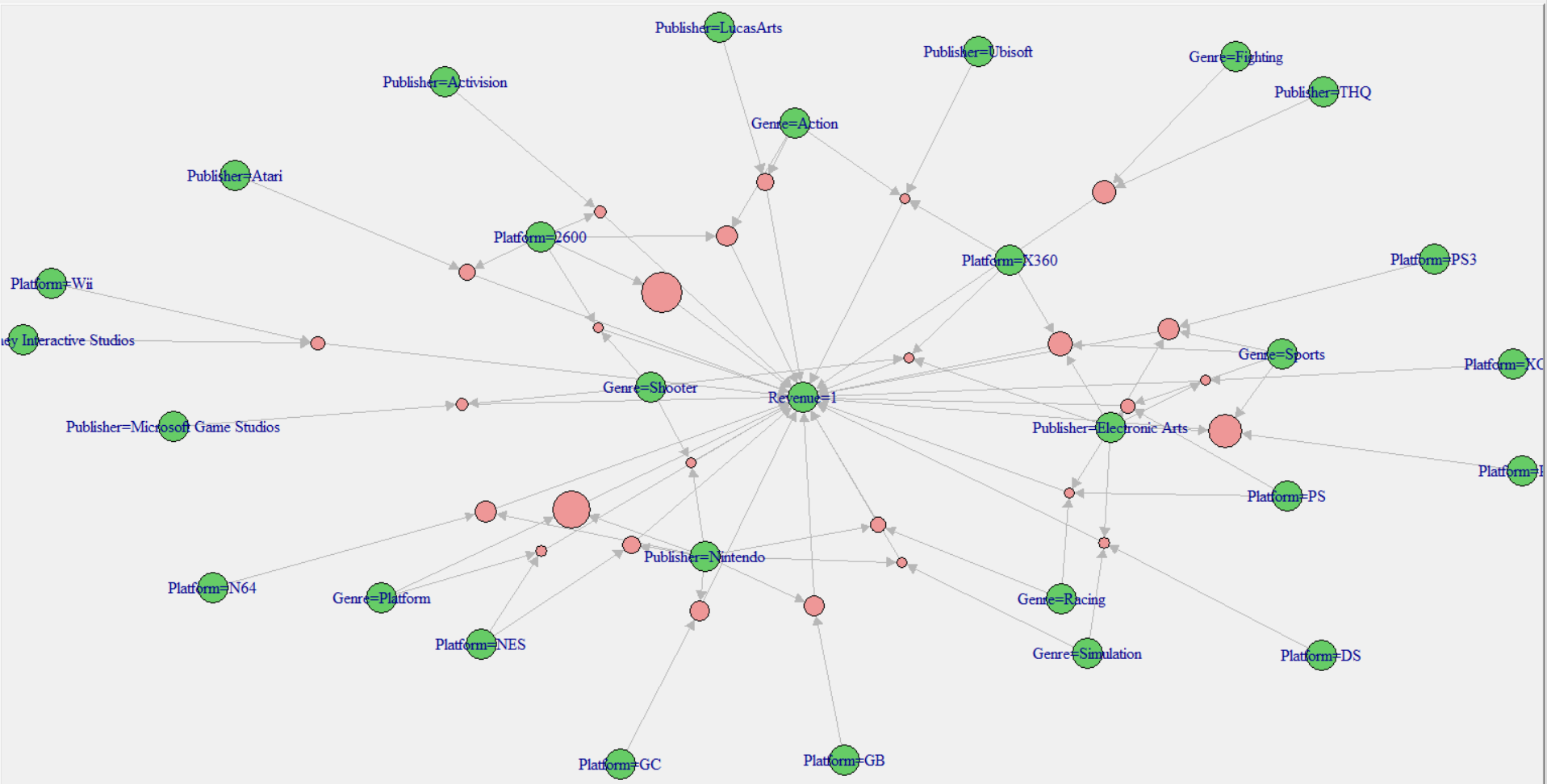
Global sales rules



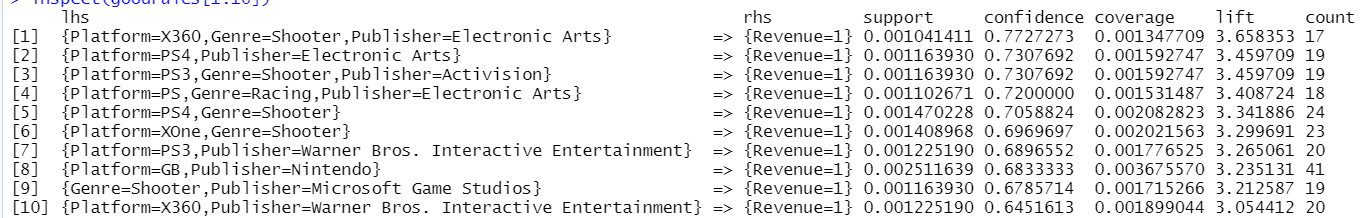
Visualization of Global sales rules



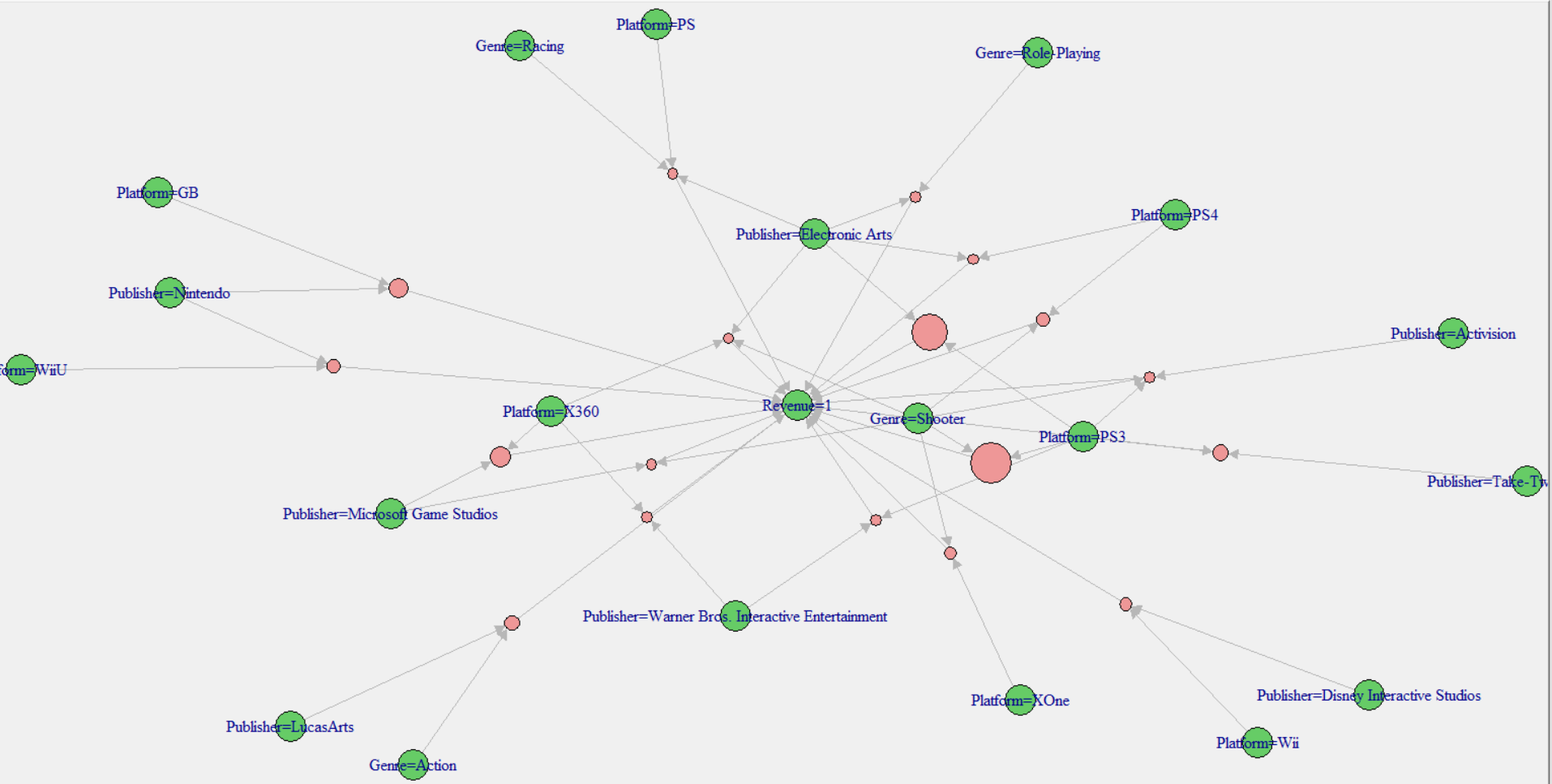
NA sales rules



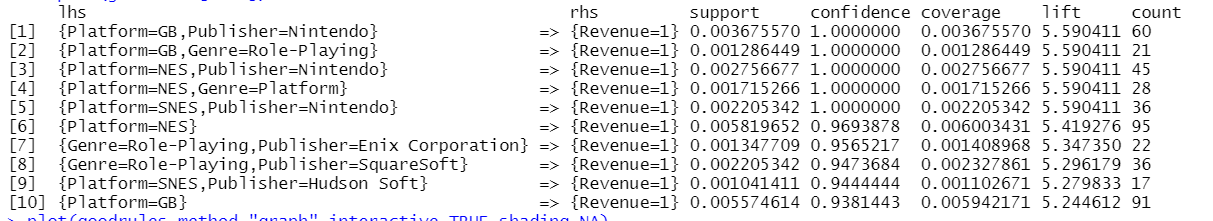
Visualization of NA sales rules



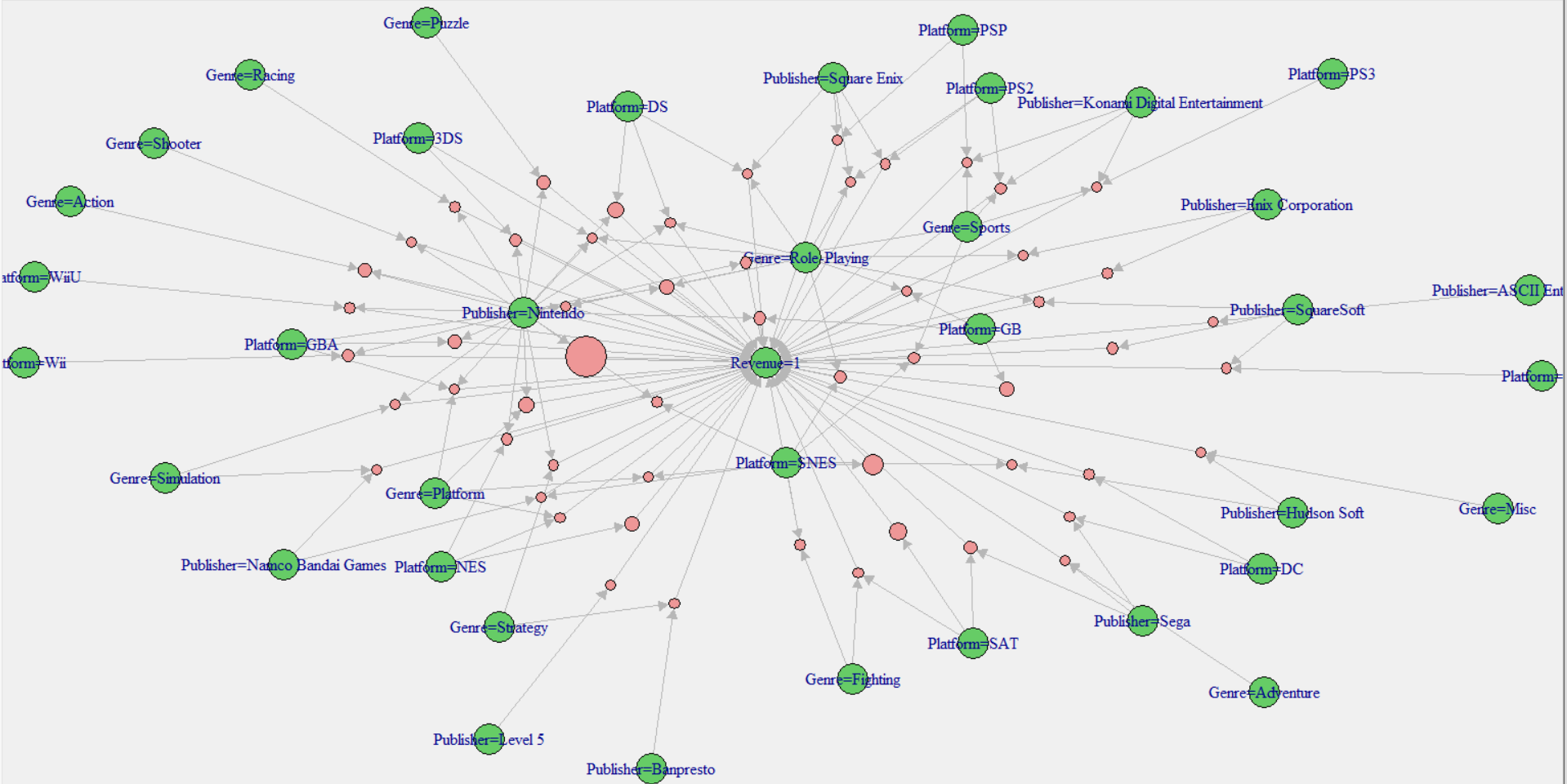
EU sales rules



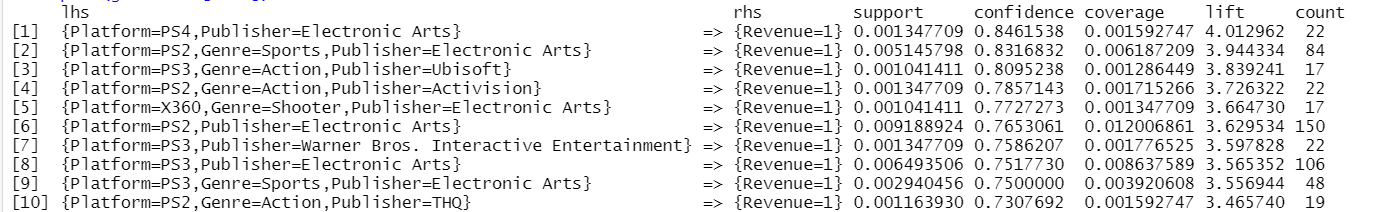
Visualization of EU sales rules



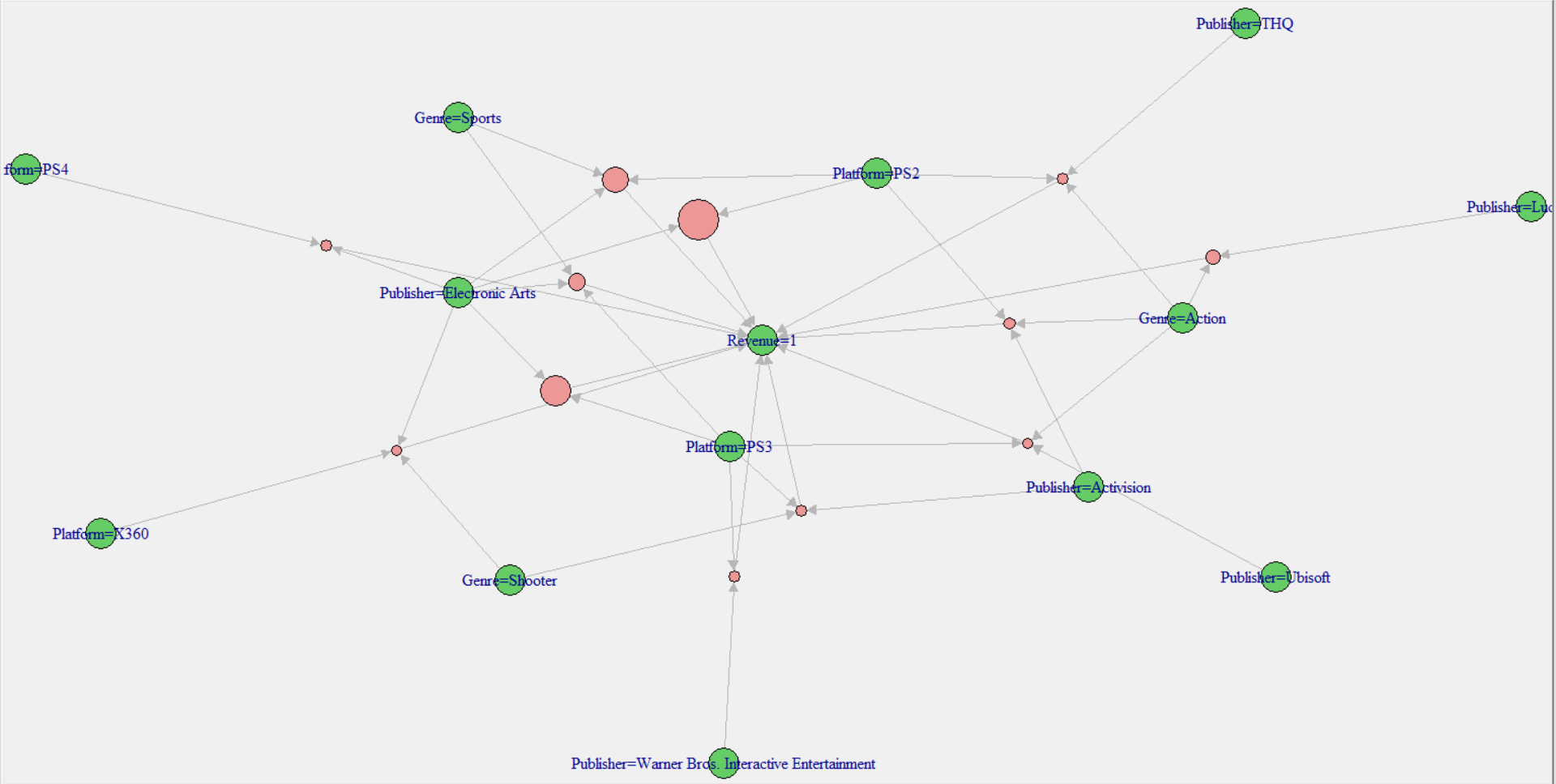
JP sales rules



Visualization of JP sales rules



Other areas sales rules

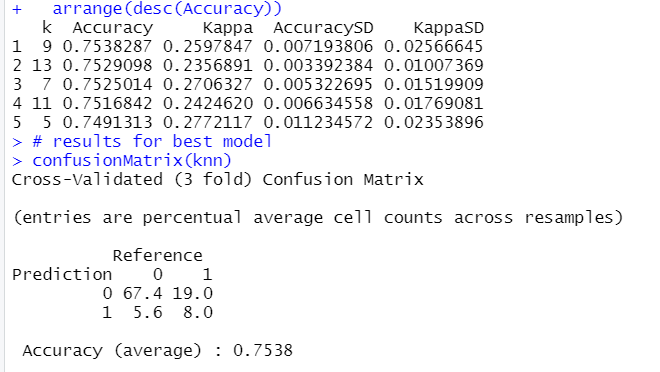
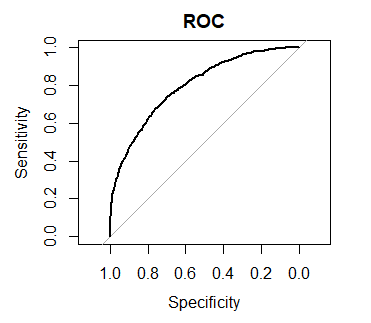


Visualization of Other areas sales rules

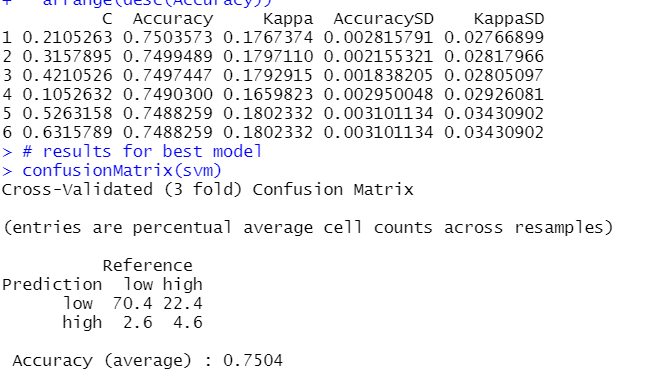
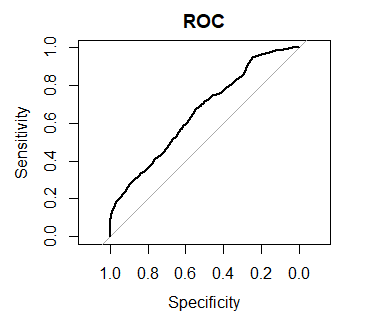
### Classification By KNN, SVM, Naïve Bayes

In this report, I select three classification machine learning algorithms, KNN, SVM with three different kernels and Naïve Bayes to do prediction works and want to find which one is the best algorithm for this dataset with the highest accuracy. Since KNN and SVM are both time consuming algorithms, I just conduct them with cross-validation to 3 folds to save time and it is enough for this small dataset. The result is shown as follow:

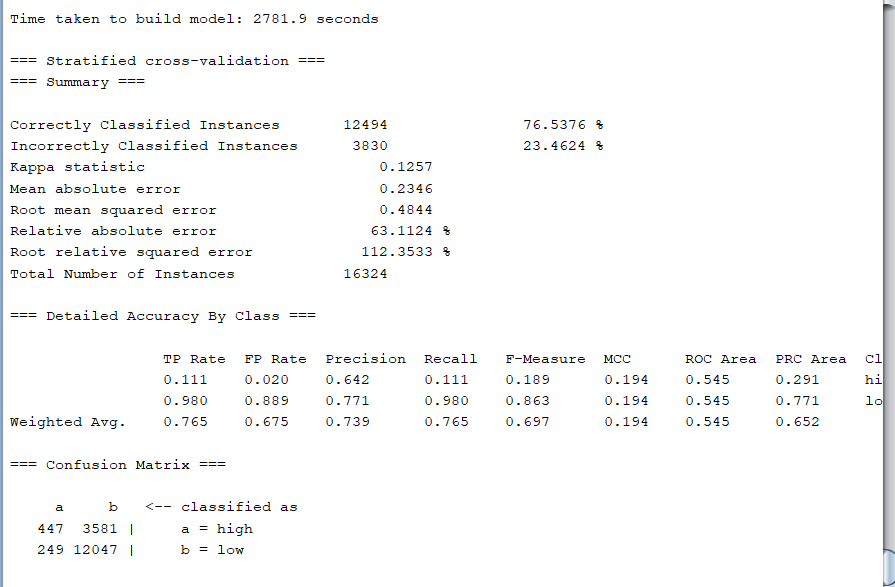
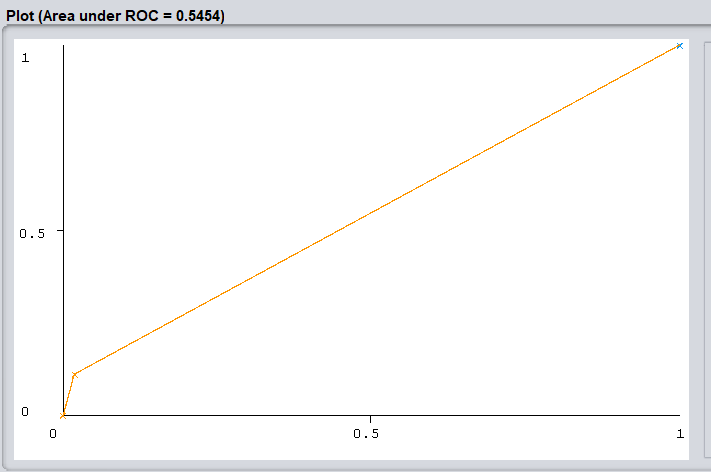
KNN:

SVM with linear kernel:

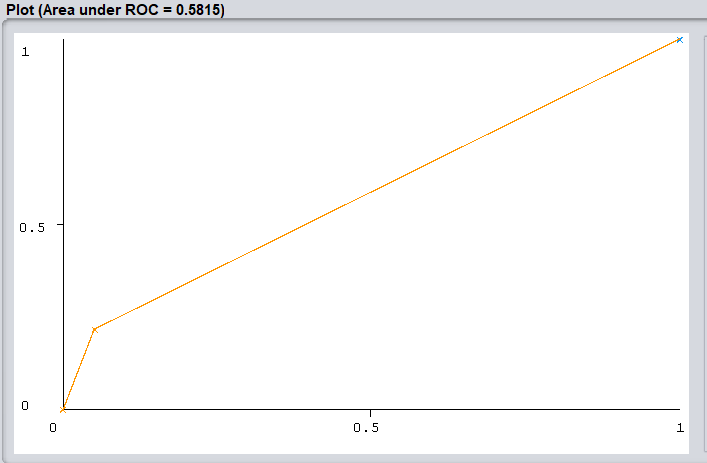
 

SVM with RBF kernel:

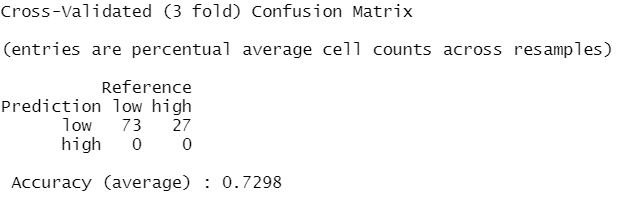
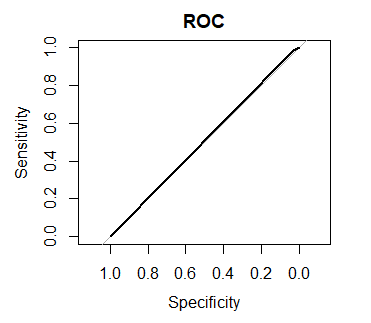
 

SVM with Poly kernel:

Table

Description automatically generated with low confidence 

Naïve Bayes:

## Conclusion

To sum up, it is time that we can compare the performance of different machine learning algorithms and extract some helpful suggestions for video game publishers and investors. Also, I would like to discuss about the issues I have in this project.

* Based on the table shown below, SVM with RBF kernel is proved to be the best classification machine learning algorithm for this study. However, it’s ROC sensitivity is just 0.5454 which means the result of it is not stable and almost random. Similarly, ROC sensitivity of SVM with different kernels all do not perform well which may be caused by the dataset. To small dataset, KNN may be a better choice.
* In association rules, we can find that although Nintendo had the top 10 video games sales in global area, the hit games in other areas like NA, EU are totally different. For example, in North America, the most popular game platform is “2600” and the publisher is “Activision”. In Europe, people prefer playing video games provided by publisher, “Electronic Arts”. For Japanese, GBA is not doubt the dominant platform. And in other areas, players seem to like PS series platforms better. Therefore, with these useful rules, video game manufacturers or investor can predict the hit video games in the future much more easily.
* It is not fast enough to use R Studio to run machine learning algorithm, I cost nearly one day on running SVM with poly kernel and the software crashed. So, I think Python IDE or other tools may perform better than R studio in data mining.
* At last, for it is my first time to complete a total data analysis report based on machine learning, there are so many problems occurred like how to choose the topic, errors in R Studio, analysis logic and so on.
* What impressed me most in this project is the strong power of machine learning. To different data analysis problems, we can select different machine learning algorithms to solve them. But it is a pity that my poor machine learning background cannot support me to behave well in this final project. So, in my future learning, I will go further study about how to apply machine learning model in solving real business problems or helping do business consulting.



Table.1 Classification accuracy



Table.2 Association rules

## References

*Video Game Sales*. (n.d.). Retrieved May 14, 2021, from <https://kaggle.com/gregorut/videogamesales>

*Explore Video Games Sales*. (n.d.). Retrieved May 14, 2021, from <https://kaggle.com/umeshnarayanappa/explore-video-games-sales>

*Video Games Market Share, Analysis and Trends by 2030*. (n.d.). Retrieved May 14, 2021, from <https://www.transparencymarketresearch.com/video-games-market.html>

Inc, G. I. A. (n.d.). *Global Video Games Market to Reach $293.2 Billion by 2027*. Retrieved May 14, 2021, from <https://www.prnewswire.com/news-releases/global-video-games-market-to-reach-293-2-billion-by-2027--301287569.html>

GregorUT. (2021). *GregorUT/vgchartzScrape* [Python]. <https://github.com/GregorUT/vgchartzScrape> (Original work published 2016)