Data Analysis and Prediction of Video Game Sales

Haoyang Yu, hy2581, Meng Ren, mr3847

Introduction:

Video game industry has been sky-rocket booming for years and has become one of the fastest growing industry. In 2017, PC games worth over 28 billion dollars and it has been estimated that PC games sales would reach 33 billion dollars in 2020. Over the global market, the whole sales of video games from all different platforms had hit 108.9 billion dollars. Considering the rapid growth of sale volume and a huge amount of industry value, the markets are huger for sales data analysis and future prediction.

Dataset description:

The information in the original dataset has been distributed into rows and columns. Each row represents a different video game and each column represent a different feature of a specific video game. The columns include 'Name', 'Platform', 'Year_of_Release', 'Genre', 'Publisher', 'Developer', 'NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales', 'Global_Sales'. 'Name' is the name of each video games. 'Platform' is the hardware system which video games are running on (i.e. PC, PS4, Xbox, etc.). 'Year_of_Release' is the year that the video games are released to the market. 'NA_Sales' is the sales amount in North America. 'EU_Sales' is the sales number for Europe. 'JP_Sales' is the sales amount of Japan. 'Other_Sales' is the sum of the sales amount of the area except North America, Europe, and Japan. 'Global_Sales' is the whole sales number of a game all over the world.

Data Loading:

We used the 'pandas' library to read and process dataset. The origin 'csv' file would be read and load into 'table' liked dataset. The type of the loaded dataset in Python is 'dataframe', which is a 2-dimension format.

Data cleaning:

We found that some rows in the dataset should be deleted after stared processing data, so we will submit a new dataset that has been cleaned instead of the dataset we submitted before. The data have been cleaned is that some game in 1993 has not classifier to any genre. We use 'notepad++' to clean data. The data should be further cleaned based on different questions below.

Project description and conclusion:

1. Data visualization:

a) Do the platforms with higher average sales on games have higher average user score?

Data cleaning:

- Drop NaN value in the columns 'Platform', 'Global_Sale' and 'User_Score'v.

The sales of video games from different platform:

- Group the dataset by different platforms and calculate the mean value of global sales for each platform, then rank the result in descending order.
- Calculate the total global sales for different platform.
- Plot a bar chart for the average global sales of different platforms in descending order.

The user score of video games from different platform:

- Drop the rows whose user score is recorded as 'tbd'.
- Change the type of data in the column 'User_Score' from string to float.
- Calculate mean user scores for different platforms and drop the NaN values.
- plot a bar chart for the average user scores of different platform in descending order.

The relationship between global sales and user score:

- Create a new dataset that contained 'Platform','Global_Sales' and 'User_Score'.
- Calculate the correlation coefficient for global sales and user score.

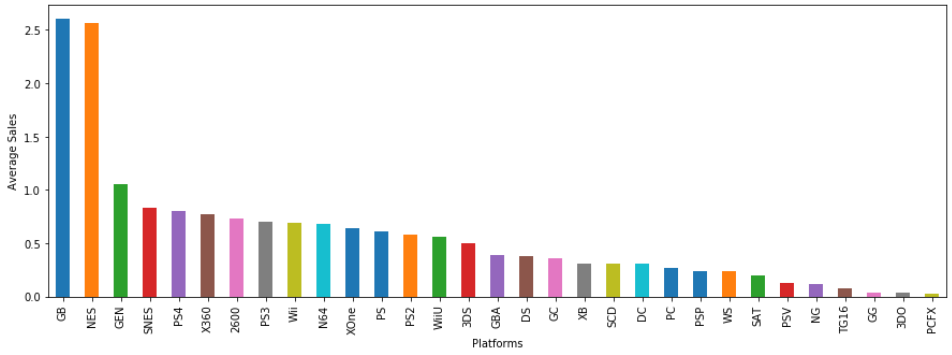


Fig.1: Bar chart for the average global sales of every game published by different platform in descending order.

The average sales of every game published by ‘GB’ and ‘NES’ are higher than any other platforms, about 2.6, while sales of other platform are all below 1.1. The bar chart contains 31 platforms.

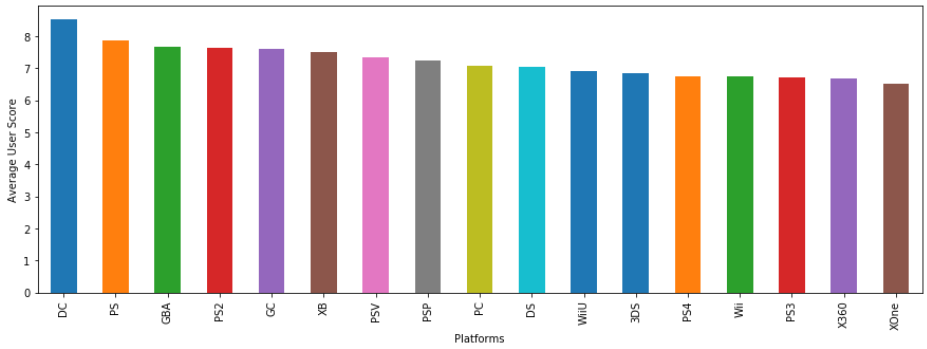


Fig.2: Bar chart for the average user scores for different platform in descending order.

The average user score of the games on different platform are close to each other. ‘DC’ platform has the highest average user score, while ‘XOne’ platform has the lowest average user score.

We selected the rows that the platform is ‘DC’ and found that most of the video games on the ‘DC’ platform were released from 1998 to 2000. Furthermore, more than half of games whose platform is DC are containing ‘NaN’ value of user score. So the statistical result of user score might be biased due to the fragmentary information.

Conclusion: We find that DC platform has the highest average user score, while GB platform have the highest average sales. The platform with higher average sales on games do not have higher average users score.

Global sales and user score have a positive correlation with each other, the value of correlation coefficient is extremely low, about 0.088139, showed that the two variables had a relatively low relationship with each other.

b) The percentage of sales of each genre over years.

- group the data by the release year and different types of game.
- Calculate the proportion of different game genres for global sales of each year.
- Make a stack bar chart of the proportion of different game genre alone years.

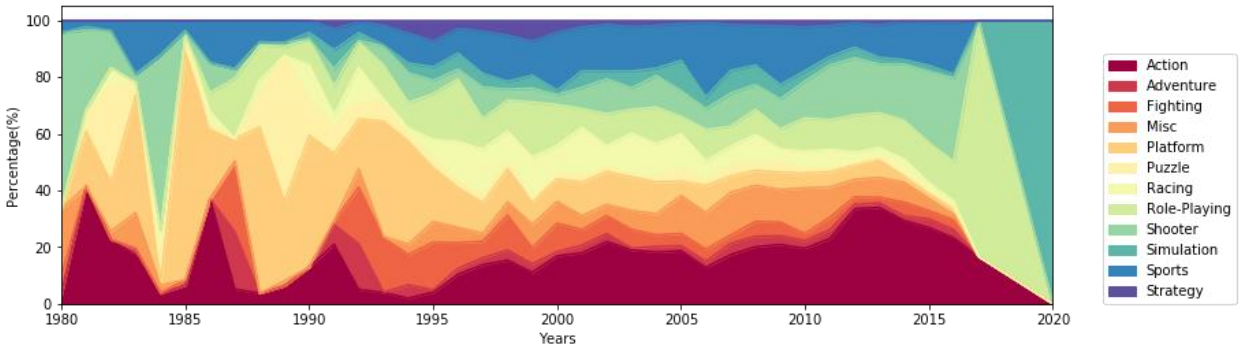


Fig. 3: Stack bar chart of the proportion of different game genre alone years.

The data collected between 1995 to 2016 is intact compare with other time periods alone the dataset. In this period, we can see the change of percentage is relatively smooth. In 1995, the ‘platform’ genre has the highest percentage of the market and has decreased in following years. The percentage of action videogames increased, while the percentage of fighting and racing video games decreased during this time. As contrary, the change of the percentage of other time period fluctuate wildly, some genres just disappear in those years. In 2020, there is just one game being recorded and it will be published in the future, that is a simulation video game.

c) Find the best sales platform based on the 7^{th} generation and 8^{th} generation.

- Create a new dataset that contained the rows that their platform is Wii or PS3 or X360.
- Group the data by platforms in the 7^{th} generation and release year (same as the 8^{th} generation).

- Calculate the total global sales of 3 different platforms of each year and make a stacked bar plot for 7th generation (same as the 8th generation).
- Plot pie chart of the sum of global sales for 7th generation and 8th generation.

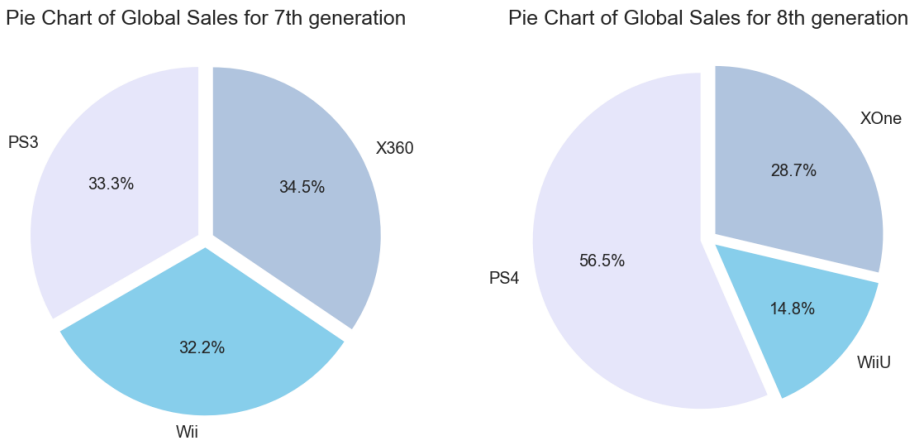


Fig.4: Pie chart of the sum of global sales for 7th generation and 8th generation.

The pie chart showed the proportion of global sales for each year of three platforms. For the 7th generation, the area of the three platforms were similar to each other. However, for the 8th generation, the area of 'PS4' took up more than a half of the pie chart, and the areas of 'XOne' is as twice large as the area of 'WiiU'.

Conclusion: The consoles for video games have been divided into 8 different generations by their released and prevalent periods. The 8th generation is the current video game consoles generation and is the descendant of the 7th generation. The 7th generation included Nintendo's 'Wii', Microsoft's 'Xbox 360' and Sony's 'PS3'. 8th generation included Nintendo's 'Wii U', Microsoft's 'Xbox One' and Sony's 'PS4'.

The best sales platform based on the 7th is X360, the marketing segmentation of 3 platforms is relatively equal. The best sales platform based on the 8th is PS4, the proportion of it is 56.5%, almost two times of the second sales platform, XOne. The marketing segmentation is not equal to 3 platforms.

However, with the innovation of video game developed pattern, the amount of exclusive games is decreasing. Most current video games could be played on different platforms. Thus, the sales data for different platforms might not be able to reflect the actual sales.

2. Do critic score and users score related to each other? Which region of sales has the most contribution to global sales?

- Data cleaning: Clean data by drop all NaN because we will use the whole dataset here.
- Convert data type in user score from string to float.
- Plot a correlation map, get the correlation coefficients between different regions of sales with global sales. The correlation coefficient between user score and critic score is 0.58.

- Plot a cross-plots and plot the scatter and regression line between critic score and users score by using the linear regression model in ‘sklearn’.
- Calculate r2 score of the regression model, the result is 0.3368. The score is fine, critic score and user score do not have high relationship with each other.
- The slope of the regression line is 0.0603, the intercept is 2.9515.
- Plot the scatter and regression line of predict user score and true user score. The scatter plot is spindle-shaped, show a linear relationship with each other.
- Plots the residual scatter of regression line.
- Use the bootstrapping method to find the confidence interval of the slope.

Conclusion:

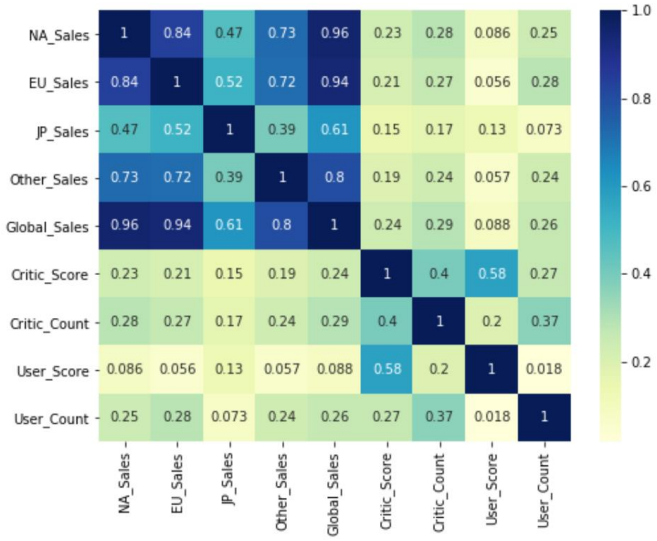
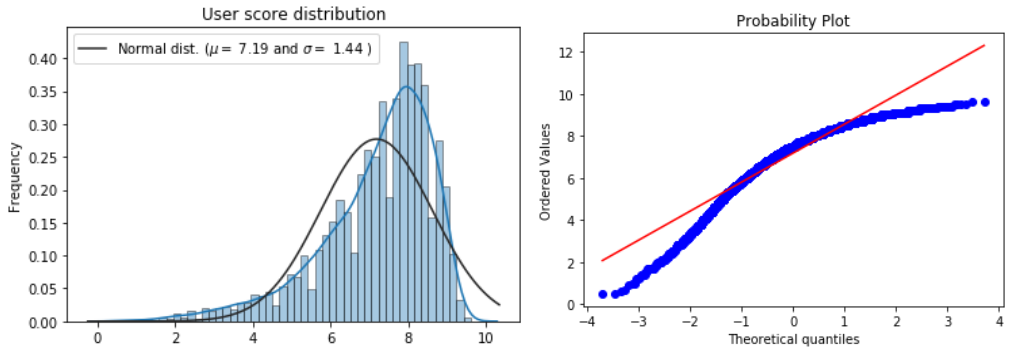


Fig.5: Correlation map.

The region that has the most contribution for global sale score is NA sales based on 0.96 correlation value with global sales. However, since the global sale is the sum of NA, EU, JP, and other sales, we should focus on the correlations of other features with global sale. From the map, critic count has the highest correlation with global sales.



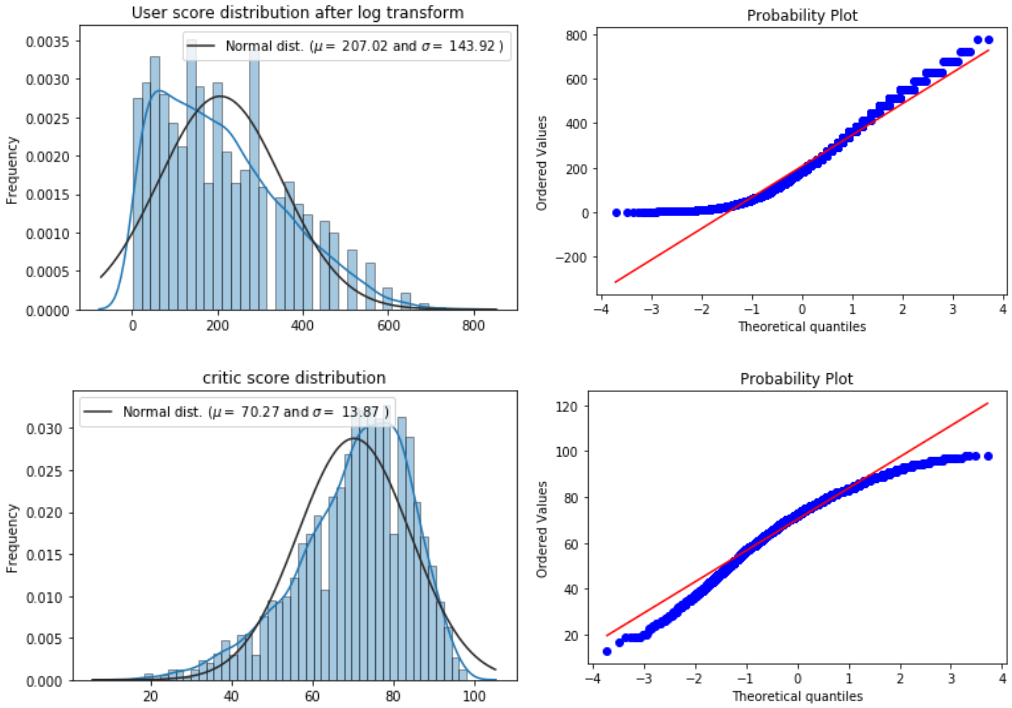


Fig.6: Distribution and probability plot for user score, user score after log transform, critic score.

Both user score and critic score show left-skewed on distribution, the same conclusion as probability plot. For linear regression, we should try to adjust data distribution to normal distribution, so we try np.exp2(2**X) function. But the result distribution becomes right-skewed. The r2 score based on transferred user score is lower than the original score, so we will use the original user score.

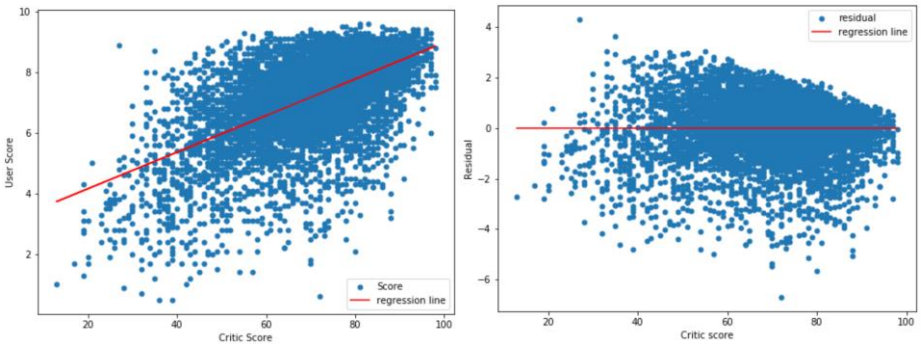


Fig.7 Fig.8

Fig.7: The scatter and regression line between critic score and users score. Fig.8: The residual scatter of regression line.

The user score and critic score do have a correlation with each other, even though it is not high enough. But the regression of two variable is not perfect based on low r2 score, 0.3378.

The residual plot appears to be centered around the horizontal line at level 0 with no upward or downward trend. Residual plots help us make visual assessments of the quality of a linear

regression analysis. This plot indicates that linear regression was a reasonable method of estimation: the shape of residual scatter plot is similar with the original scatter plot, and the points of scatter plot distribution is roughly even below and above 0.

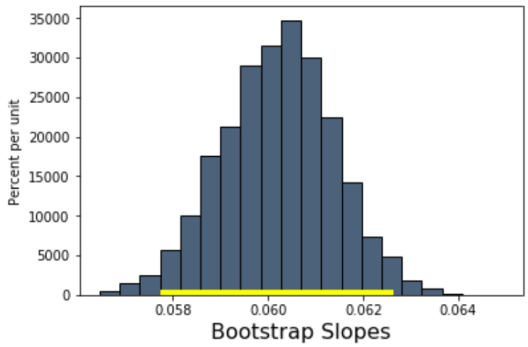


Fig.9: the confidence interval of slope.

Bootstrapping methods are useful when processing large samples, we can bootstrap different scatter plots in order to have a better estimate of the true slope. After bootstrap the scatter plot a large number of times, we draw a regression line through each bootstrapped plot. Each of those lines has a slope. We can simply collect all the slopes and draw their empirical histogram. Then we can print the interval consisting of the "middle 95%" of the slopes. Approximate 95% confidence interval for the true slope:0.058 to 0.063.

3. Predict the global sale for the 7th generation, the 8th generation and PC platform by critic score, platform, genre, publisher, rating and developer, using decision tree classifier, random forest classifier and random forest regression, gradient boosting regression and XGBoost regression methods:

Data processing:

- Create a new dataset that will be analyzed: select the columns that only contain features: global sale, critic score, platform, genre, publisher, rating and developer, select the data of the 7th generation, the 8th generation, and PC platform;
- Drop the rows that contain 'NaN' values of this dataset;
- Convert data type in user score from string to float.
- Find outlier values by plot scatter between critic score and global sale, drop this outlier;

Decision tree classifier:

- Create the label for the x dataset: since user score is continuous datatype, we divide user score into 4 categories. The principle of category process is trying to control the amounts of samples in every category is close to each other. Processing the critic score in the same steps.

Categories	Score range	Percentage of samples in each category	
		User score	Critic score
A	X >= 8.5	17%	15%
B	7.5 <= X < 8.5	34%	31%
C	6.5 <= X < 7.5	23%	25%
D	X <= 6.5	26%	29%

Table 1: categories of user score and critic score.

- Create the label for the y dataset: we divide global sales into 3 categories.

Categories	Global sales range (million)	Percentage of samples in each category
High	X >=1	21%
Medium	0.2 =< X < 1	41%
Low	X < 0.2	38%

Table 2: Categories of global sales

- Encoder the categorical data type, transform category data into numbers to train decision tree model;
- Split the dataset into train set and test set as 80% and 20%;
- Use random forest classifier to learn features;
- Use the test set to predict the global sale range, calculate the accuracy of prediction;
- Plot a confusion matrix.

Random forest classifier:

- Repeat the steps of decision tree classifier, get the accuracy of prediction and confusion matrix.

Random forest regression:

- Convert the categorical data type into dummy variables to train the random forest regression model;
- Since the distribution of global sale is right-skewed, we take the logarithm of global sale;

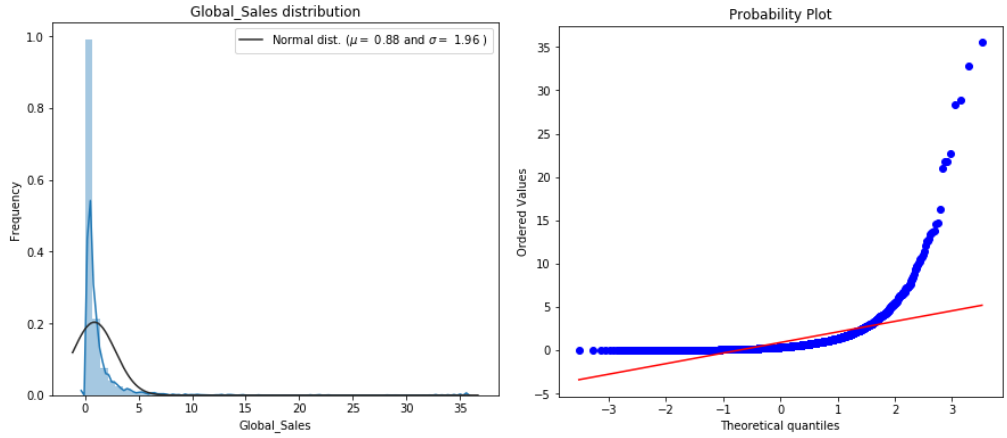


Fig.10: Global sales distribution and probability plot. Global sales distribution is right-skewed.

- Use model selection function: GridSearchCV to find the best parameters for random forest regressor include n_estmators, max_features, max_depth;
- Use the best parameters to predict the output, get predict-true scatter plot and score of prediction.

Gradient boosting regression:

- Repeat the steps of random forest regression, get predict-true scatter plot and score of prediction.

XGBoost regression:

- Repeat the steps of random forest regression, get predict-true scatter plot and score of prediction.

Conclusion:

Decision tree classifier	Accuracy	52.41%
Random forest classifier		55.91%
Random forest regression	R2 score	0.6554
Gradient boosting regression		0.6597
XGBoost regression		0.6851

Table 3: Fit result of 5 models.

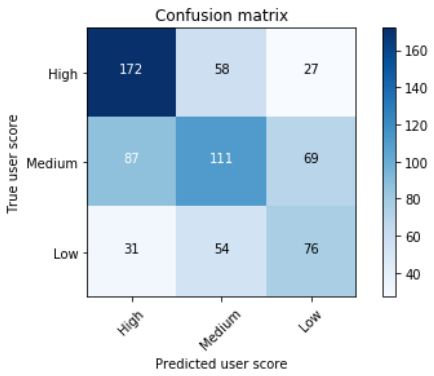


Fig.11

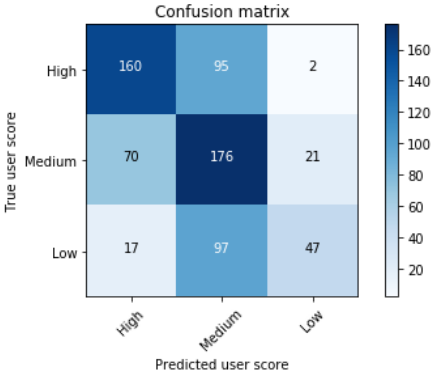


Fig.12

Fig.11: confusion matrix of decision tree classifier. Fig.12: confusion matrix of random forest classifier

The accuracy of the random forest classifier is higher than the decision tree classifier. Because random forest classifier can process plenty of different decision tree algorithm, and vote the best decision to produce an output. But the accuracy is not good enough based on classification by 3 categories, the baseline of accuracy is 41%.

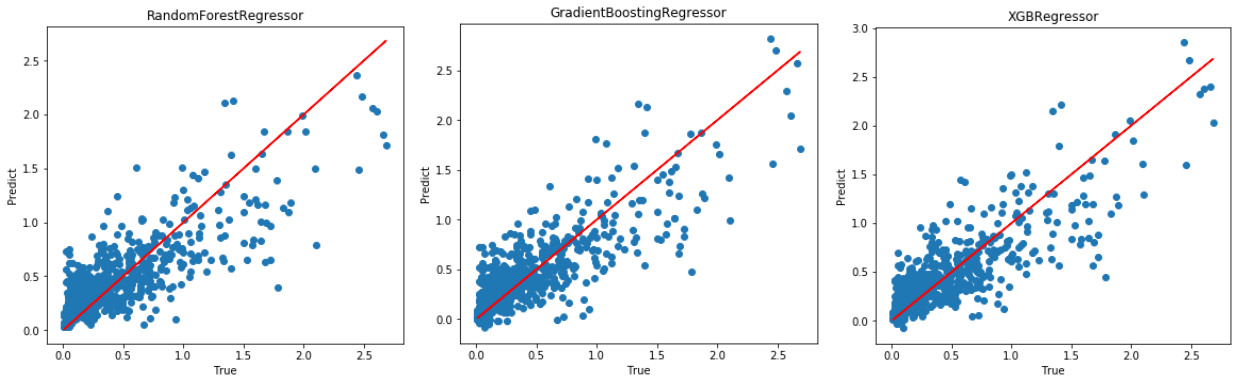


Fig.13: True and predict global sales value based on different regression model.

We try to use different regression models. The r2 score of different models: Random forest regression < Gradient boosting regression < XGBoost regression. The scatter plots of predict and true value of global sales based on different models looked similar to each other. Predict and true value show a good linear relationship with each other.

Deficiency and future work:

1. The sales number of different platforms show various characteristics. As the platforms developed over years, most platforms were weed out. Currently, most competitive platforms are PS4, XboxOne, Wii U and PC. However, due to the decreasing amount of exclusive games and the increasing number of free-installation games, this dataset of video games sales information might not reflect the actual performance of different platforms.

2. The sales number of North America made the most contribution to global sales in this dataset. Whereas, according to the report by 'Newzoo', the sales number of video games in China had exceeded the US in 2017. Furthermore, 'in-app purchase' has been the most profitable strategy of video game developer and publisher, dataset about free-installation games should be considered to perform a further study.

3. The critic scores in this dataset were provided by 'Metacritic', which is not the most authoritative video game comment organization. The critic score from 'IGN' might have a more impact on user score and sales number of a video game.

4. Regression model has a better performance of the prediction of global sales. But the huge amount of missing values in the dataset, nearly half of video games were drop out from the original dataset, which might lead to the inaccurate of the prediction.