Rotman Commerce, University of Toronto
RSM 456 Big Data and Marketing Analytics
Professor Gerhard Trippen
5 April 2019

Big Data Project

Final Report

Jiarou Yu 1002059309
Sihan Yang 1002113517
Xinyue Wang 1002214409
Wenjie Hao 1002183059
Yijia Liu 1002250734
Leyin Zheng 1002317824
Sheng Qin 1002207836

1. Objective

Our project takes the perspective of the management team at Nintendo, a Japanese-based major game developer and publisher, to determine the ideal type of video games that would generate the highest sales and best user ratings for the company, taking global preferences into account. Our analysis is based on the *Video Game Sales with Ratings* Kaggle dataset with 16720 entries including column names.[1]

Our dataset is comprehensive with 16 columns of information for each game. We filter-in and highlight the following variables that are particularly useful for our data analysis: categorical variables such as platform (on which the game is running), genre (game's category), developer (name) and game ratings. And quantitative fields like global video game sales, user score, and critic scores.

2. Data manipulation

1)Missing data

After carefully examining the dataset, we identified that some data contain missing values may hinder our data analysis. Out of 16719 rows of data in our original dataset, we now have 16415 rows after cleaning the data. There were N/A values in "Platform", "Publisher", "Sales", "Year of Release", "Genre", or "Game Name" columns, thus we decided to delete those rows with N/A values in these columns given that the "Platform", "Genre", "Rating", "Critic Score" and "User Score" columns are key independent variables that drive revenue and "Game Names" are important in identifying the specific games. Game rating is an important factor when analyzing sales figures. For instance, giving games a rating of M could potentially reduce revenue as they are not suitable for customers under 17. For the "Rating" column, we want to keep them for further analysis. Since they are categorical and qualitative data, we fill the missing data in this

[1] This dataset was combined on Kaggle from the *Video Game Sales* dataset from VGChartz and corresponding ratings from Metacritics. VGChartz is a video game sales tracking website that provides weekly sales figures by region; Metacritics is a website that rates and aggregates reviews of media products such as films and video games.

column with “Others”. For our regression analysis with the quantitative variable, we deleted any rows that missed any value or had “tbd” in “user score”, “critic score” and “global sales”.

2) Outliers

We check for negative values in global sales and the extreme high or low value in the global sales using statistical interquartile range. We found some games with high sales but decided to keep them for now. We also check the figure of regional sales versus global sales provided. By adding the sales of four regions and dividing the sum of the sales into four regions by global sales, we get the result that most ratios are one while the rest of the ratios can be rounded to one if keeping three decimals. Therefore, we think the slight difference between the sum of regional sales and global sales is due to a rounding error. There is no need to clean these data.

3) Data Cleaning

Our objective is to determine the ideal type of video games that will generate the highest sales for Nintendo globally. We want to determine the ideal type of video games, the columns related are “Name”, “Platform”, “Year of release”, “Genre”, “Publisher” and “Rating”, “Critic Score”, “User Score”. Based on the above columns, our team will be able to develop the type of video game that can gain the highest sales for Nintendo. As for rows, we want all the observations in “Sales” columns to be non-negative since the unit is millions of sales units; However, there are no such rows in the database after the data cleaning, so we do not need to filter any rows. After filtering, we have in total of 11 columns and 16,415 rows.

- **Grouping:**

We group respectively by “Platform”, “Publisher”, “Rating” and “Genre” since they are the independent variables of our research such that we could set up comparisons between the sub-results of each variable (such as in platform, look at Wii versus PSP). The dependent variable is the impact on sales and the dataset is already broken down regional (NA, EU, JP and Other) and global sales. So there is no grouping code needed here.

- **Transforming:**

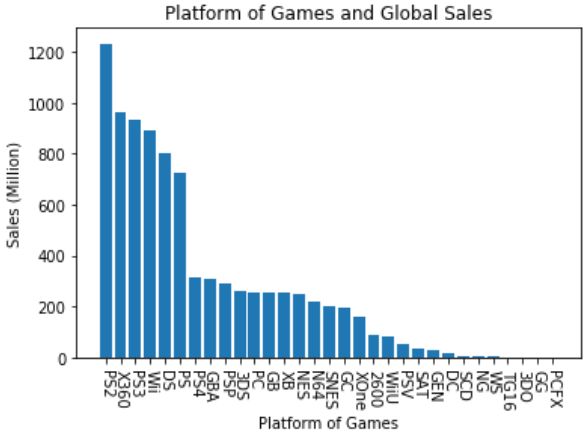
We find the column "Year of Release's" datatype is floating-point numbers. We need to change the data type from floating-point numbers to integers since years are supposed to be integers without decimal.

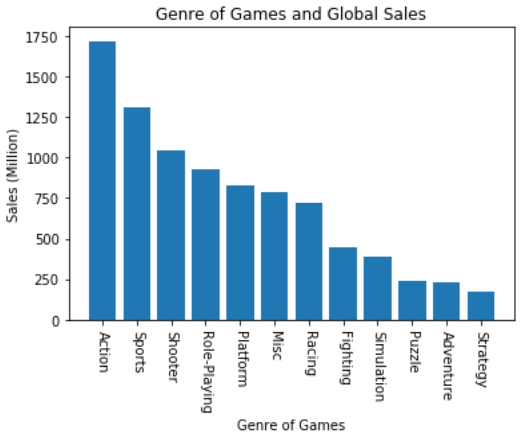
- **Sorting:**

For now, there does not seem to be an explicit need to sort by any variable because we have a multi-regression model's intuition. But for now, we sorted by descending orders of "global sales" to help with understanding the most significant results first.

After checking for cleaning the missing data, checking for outliers, grouping, transforming, and sorting the data, the dataset is ready for visualization and machine learning.

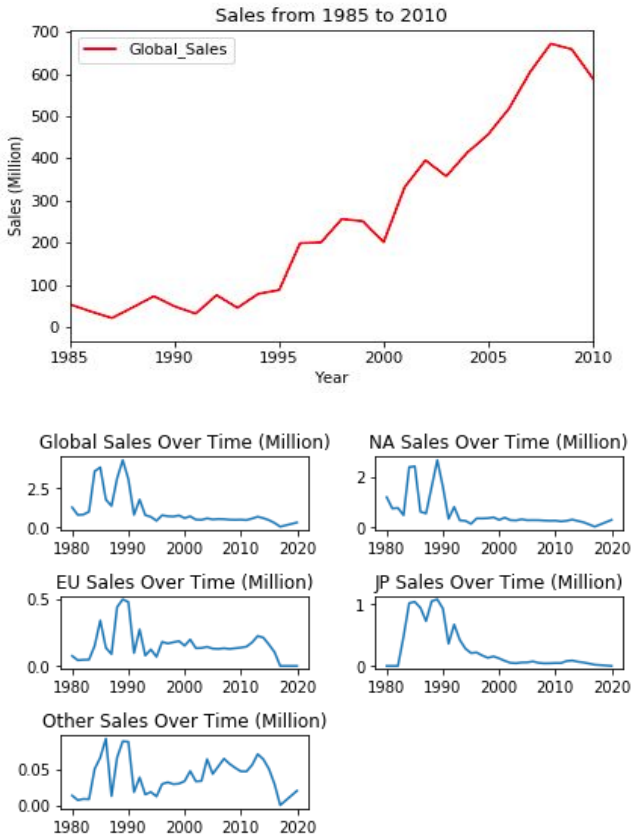
3. Data Visualization





According to the bar chart "Platform of Games and Global Sales", the top six platforms are PS2, X360, PS3, Wii, DS, and PS, respectively. Among the six platforms, PS2, PS3 and PS platforms belong to PlayStation home consoles, a gaming brand owned by Sony Interactive Entertainment in Japan. X360 represents Xbox X360, a home video game console launched by Microsoft. DS is a dual-screen handheld game console and Wii is a home video game console. Both platforms are developed by Nintendo. We can find that Sony, Microsoft, and Nintendo dominated the video games console market.

When it comes to the genre of games, the most popular types are the “action game” and “sports game”. Action game requires hand-eye coordination and real-time reaction. One of the most popular action games in recent years is Player Unknown Battleground. The second and third popular genre of games is Sports and Shooter, with sales of around 1300 million and 1100 million, respectively.



In this line diagram, we wanted to visualize the changes in overall video game market size over time by plotting “year of release” and the sum of “global sales” in each year. This chart tells mostly an expansionary tale from 1985 to 2008, with the most accelerated sales growth periods between 2005 and 2008. This is primarily driven by economic prosperity, massive mobile phone adoptions and lower game development costs as we discussed in our research proposal.

In the subplots, we compare the average of sales (units) in each year in each of the 4 distinct regions and global sales. These insights are drawn from this diagram:

1) The relative average market sizes (y-axis):

The peak average sales units on a global scale are around 3M in the 1980s. It is interesting to note that NA (North America) has a peak of around 2.5M in that same time period: the largest sub-market for games. Ranking second is the Japanese market with a peak around 1M sales in the later years of the 1980s and 1990s. Finally, the EU market is also significant at 0.5M at around 1990 and all other parts total to only 0.05M during the same time period (not as significant)

2) Time series analysis (x-axis):

From the global average sales chart, we can see that on average, fewer game units are sold over time. In particular, the largest markets like NA and JP have both shown significant decreases in average sales units since the 1990s and 2000s to a stable level close to 0M (rough estimate). From a business point of view, this implies current game designers face an extremely competitive and diversified market than the 1980-1990s. For the EU, the decreasing trend in average sales is also apparent but not as quantitatively significant (drop from 0.5M to 0 in EU versus 2M to 0 (NA) and 1M to 0 (JP). Finally, for "other sales regions", the average sales unit trend has been increasing over time. However, the "other regions" market is less than 4% of the global market so its positive trend does not materially influence the global trend.

From the visualization, we find that the total sales keep increasing in the recent years while the average sales indicates a downward trend, which is due to the extremely competitive market. The video game console market is dominated by Microsoft, Sony, and Nintendo, and action game, sports game, and Shooter game are the most popular genre.

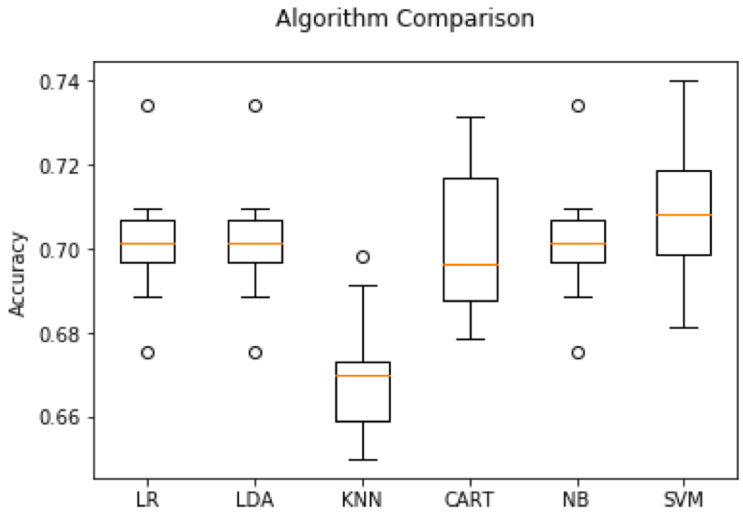
4. Machine Learning: Applying Algorithms

- **Classification**

We use classification to predict whether the game is popular or not. The target variables are unpopular game (labeled as '0') with sale less than 0.47 million and popular game (labeled as

‘1’) with sales more than 0.47 million. 0.47 million is the top 25% percentile of the global sales performance. The features/independent variables are “genre”, “platform”, “rating”—the three most important game design factors. We believe this is an important investigation given Nintendo’s corporate strategy to always lead in the market by producing the most popular games.

At first, we split the dataset and separate the data into training data and test data, we chose 30% of the data as our test data. Then we have run in of total 6 models, they are Logistic regression(LR), Linear discriminant analysis (LDA), The k-nearest neighbors algorithm(k-NN), Decision tree learning(CART), Naive Bayes classifiers(NB) and Support Vector Machine(SVM). We evaluated the models in turn and compare them in order to choose the one with the highest accuracy. We also used a Box plot to show the accuracy of each algorithm. From the box plot, we have chosen the SVM model since it has the highest mean accuracy (71%) among all the models, and It also showed no outliers comparing with the other models.



By using SVM as our classification model, we made predictions that 316 games are best-selling games/popular games with global sales greater than 0.47 million, which are among the top 25% of the sales performance.

```
This is confusion matrix for the predictions from SVM model we chose for classification:
[[1959  152]
 [ 710  164]]

This is classification report for the predictions from SVM model we chose for classification:
              precision    recall  f1-score   support

           0       0.73      0.93      0.82      2111
           1       0.52      0.19      0.28       874

   micro avg       0.71      0.71      0.71      2985
   macro avg       0.63      0.56      0.55      2985
weighted avg       0.67      0.71      0.66      2985
```

	Class Predicted by your model	
	0	1
0 [[	1959	152]
1 [	710	164]]

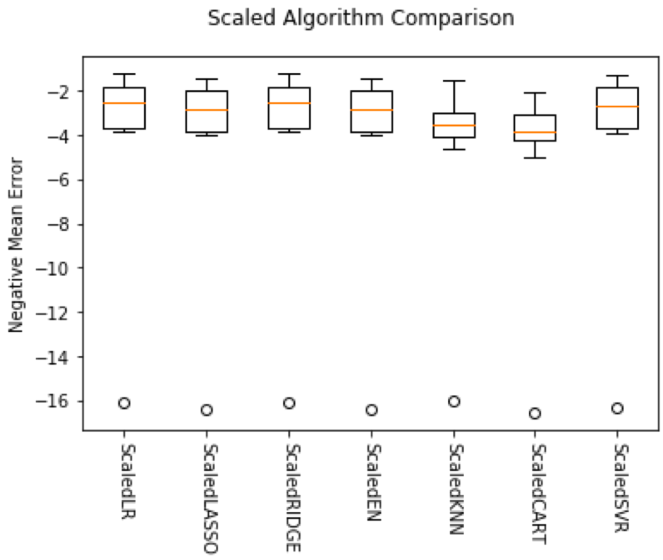
We used accuracy score along with confusion matrix and classification report in predictive analytics. Firstly, we use classification accuracy as the start point, which is the correct predictions made from all predictions made. The accuracy scores for the predictions from the SVM model we chose for classification is around 0.71, which is relatively higher than the rest 5 models.

While the accuracy score alone cannot offer sufficient information when evaluating the model. Thus, we use the confusion matrix and the classification report to present the prediction. Coming to the confusion matrix, as a binary classification, we set two classes, “unpopular games” (label 0) and “popular games” (label 1). There are 2111 games in unpopular class (label 0), where the SVM classifier successfully identifies 1959 of 2111 games, given the recall of 1959/2111=0.93. There are 874 games in popular class (label 1). Among them, 164 games are marked correctly. SVM predicted the rest 710 games in class 0 (as unpopular games), given the recall of 164/874=0.19. In addition, by looking at first column in the table above, 1959 games are marked correctly in class 0, the rest 710 are incorrect, which given the precision of 1959/ (1959+710)=0.73. Similarly, the precision in class 1 is 164/(164+152)=0.52. “The f1 score is the harmonic mean of recall and precision, with a higher score as a better model.” (Andrew Long, 2018) The f1 score is higher (0.82) in class 0, and much lower (0.28) in class 1.

- **Regression**

We use regression to predict the global sale of the game. The target/dependent variable is global sales. The features/independent variables are “user score” and “critic score”. These are two quantitative data sources that measure both the consumers (users) and experts (critic)’s receptiveness of the game.

At first, we split the dataset and separate the data into training data and test data, we select 30% data for the test set. After that, we have run 7 models in total, they are Linear regression(LR), Lasso Regression(LASSO), Ridge regression(RIDGE), Elastic Net(EN), The k-nearest neighbors algorithm (k-NN), Decision tree learning(CART), and Support Vector Regression model(SVR). We also standardized the data using the pipeline method. After normalizing the data, we evaluated the models in turn and compare them in order to choose the one with the highest accuracy. We used a box plot to show the accuracy of each Algorithm. The model with the highest negative mean error (negative MSE) (smallest absolute value) is the most accurate. By comparing the seven models, we chose to use Support Vector Regression model(SVR) with MSE of 2.84 to predict global sales of the game.



We used scaler to rescale the data to run regression models. And we also used tuning hyper parameter to improve our results.

The estimates for the game sales in X_test set predicted by the SVR model are [0.65195131 0.64603629 0.76360886 ... 0.56200845 0.7228997 0.51766172]. From the predictions of global sales, we can tell that New Super Mario Bros will have the highest global sales compared to other video games.

As a result**,** for the classification model, the SVM cannot make effectively (accurate) predictions of the best-selling games by just using "genre", "rating", and "platform", but compared with other models such as Logistic Regression Model, and Linear Discriminant Analysis Model etc. (in our project), it is a much better classifier. (Based on the accuracy and confusion matrix, as well as the classification report).

For the regression model, the SVR model we chose still has higher MSE (2.84) even it is smaller than the rest of the regression models, which suggests that we might need more dimensions for our predictions to be more accurate. We used "user score" and "critic score" to predict the global sales of a game. Based on our large MSE, we think that these two independent variables alone do not have enough predicting power. We brainstormed about adding in more independent variables to potentially make our predictions more accurate, thus MSE would decrease.

5. Analyses and Insights

Throughout this project, we used information and results obtained from our data analytics to approach our objective and assist with the decision-making process. Both our research and data visualization show a steady increase in global video game market size and a decrease in average sales per game, indicating the growing challenges and intensifying competition that we face as game developers. This current situation reinforces the importance of our decision-making process.

We conducted analysis on sales performance of various geographical regions in order to better understand different regional preferences and tastes regarding video games. However, as we

made our final decision, we mainly focused on the current and predicted *global* sales, because Nintendo is a world-wide industry leader and the new game will be introduced to the global market.

We took into consideration multiple important factors regarding the characteristics of the game we plan to develop. We wanted to obtain as much information as possible by choosing a dataset that includes both qualitative and quantitative aspects of individual games. Among all independent variables, we found that platform and genre have the most impact on global sales and decided to focus on these two variables. From data visualization, we found that two of our own platforms, namely the Wii and DS consoles, are listed among the top 6 platforms with the highest global sales.

6. Conclusion

Therefore, we decided to publish our new game on the Wii console as it has higher sales compared to DS. Then we used results from data visualization again to decide which genre to develop. We compared the top 2 genres that currently generate most sales, namely action and sports games, as well as their future potentials in sales performance. Interestingly, data visualization shows that current sales of action games exceeds that of sports games, while regression analysis predicts higher sales for sports games. The action genre represents around 1740 million sales units while the sports genre has sales of around 1300 million. Although action games has a higher sales than sports games for now, we also need to take critic scores and user scores into consideration using our regression algorithm. Since our research proposal showed that gamer feedback and involvement (represented by user and critic score variables) have prolonged effects on sales and branding, we decided to weigh more on the regression result.

Overall, from our market overview, data visualization and regression analysis, Nintendo's new best-selling game will be **a sports game published on the Wii platform**.