Foundation of Data Science

Professor Rumi Chunara

Course: CS-GY 6053 Fall 2017

Final Project Report

Project Name:

Game Culture - An approach to detect popular games

Team Members:
Name: Arhant Jain Net Id: aj1973
Name: Harsh Yadav Net Id: hy1217

Problem Motivated with appropriate Background:

Nowadays, people are very eager about games and take a lot of interest in them. People belonging to all age groups show enthusiasm and willingness to buy games. Not only games, but the platform, it has been made for is also becoming popular. Different platforms provide different features which increases overall benchmark score, a term and an important feature used by gamers, to decide overall performance of games. Games have become an important part of people's life and determines the social behavior of a person. Thus, video games are a great influence.

Over the period, games have changed their concept. Games have developed into new generation of concepts, example call of duty 1 and then call of duty 2. Considering games as movies with their own storyline and their own motions, providing an interactive and almost perfect behavior and characteristics of human imagination. For gamers it is a big judging factor. Some play games because of their storyline, example war/action or for its surreal features and animations. Some play because for them it's like an addiction. Whatever might be the reason, every rating counts and plays an important role in deciding the overall sales and popularity of a game.

It is important to find features and correlations between features, so to maximize game sales we need to see for which platform it is made for. It is essential that manufacturers maintain proper software cycle for the games that are sold in the largest quantity for a platform. Not only that, they should provide adequate after sales patches for any game on that platform making considerable amount of profits.

Today we have video games for almost everyone and it has been found that it helps to develop our cognitive part of the brain. The researchers discovered that those who played StarCraft were quicker and more accurate in performing cognitive flexibility tasks, than those who played The Sims. It has been written in one of the research papers that "Creative problem solving and 'thinking outside the box' require cognitive flexibility. Perhaps in contrast to the repetitive nature of work in past centuries, the modern knowledge economy places a premium on cognitive flexibility". It is their passion, their hobby and for some their career (game critics and testers).

So, it becomes pivotal for us to determine which video game holds certain importance in their niche category and thus, we can determine which games are popular in different parts of the world, their genre, what are their critic scores and sales around the world.

Target and predictor variables:

As the video game industry has a broad scope to discover, so we considered two problems that need to be determined, in-order to facilitate the growth of this industry.

First, we considered Ratings of different video games as our target variable which represents what games are most played by which age group. So, our **Ratings** (AGE GROUPS for VIDEO GAMES) were divided into categories of:

E: Everyone
AO: Adult Only
E10+: Everyone above 10 years of age
EC: Early Childhood, K.A: Kids to Adults
M: Mature
RP: Rating Pending
T: Teenagers

Second, we considered Sales across the world as our target variable, which represent how much a game is sold around the world (in millions).

To predict the Ratings of a game, we cleaned our dataset in many ways and finally came up to consider the following as our predictors for the Decision trees algorithm based on their feature importance:

Name: Name of a video game
Platform: The platform on which the game a released on
Year of Release: Year when the game was first released
Genre: the genre the game belongs to
Publisher: Company that published the game
Critic Count: How many critics have given their scores for a game

Then further for our analysis based on Random forest and Gradient boost algorithms, we again cleaned our dataset and this time we considered:

Name: Name of a video game
Platform: The platform on which the game a released on
Year of Release: Year when the game was first released
Genre: the genre the game belongs to
Publisher: Company that published the game
Critic Count: How many critics have given their scores for a game
User Score: Score given by users based on game performance from 1 to 10
User Count: How many users gave the scores for a game.
Critic Score: Score given by Critics based on game performance from 1 to 100

NA Sales: Sales in North America in millions
EU Sales: Sales in Europe in millions
Other Sales: Sales in Other parts of the world excluding North America, Europe and Japan in millions
Global Sales: Sales all around the world in millions
JP Sales: Sales in Japan in millions

Secondly, for predicting our second target variable that is the Sales across different parts of the world we did our estimation by converting the continuous Global Sales data into categories (from 1 to 3 million, 3 to 7 million and so on) and then predicted the sales on our test data based on Decision Trees and Random Forest models. Further analysis is done based on the visualizations that we created by plotting Genre, Platform, User Scores, and Critic Scores for determining the sales across different parts of the world.

Clear Problem Statement:

(a) To train a machine learning model to predict the Ratings (Age Group) of a game based on parameters like game Platform, Genre, Year of Release, Platform, Name of the game, Critic score, Critic Count, User score, user count and sales across different parts of the world.
(b) Which genre of video games will be most influential, which depends on predicting a Video Game's Global Sales? The criteria for more influential is more sales in different parts of the world. So, we are predicting the Global sales of a Video game.
(c) Which video game publisher is most popular and will continue to be popular in the coming years depending on the platform it produces game on, genre of games and sales in all continents? The criteria for becoming more famous is more Global Sales.

Both parts (b) and (c) depend on Global Sales prediction values and are considered as second part in our Evaluation approach as we can determine the sales of a video game by knowing its publisher, genre and other predictor values.

Type of model motivated:

We have used Decision Trees algorithm, for making the predictions of Ratings and Global Sales of a Video game. We used this model as it is much faster to train as compared to the simple neural networks for comparable performance and has less time complexity. Apart from this it is robust as compared to other algorithms.

Then we performed feature importance and selected only the features that are most important for getting better prediction results. As our results were still not quite accurate, so we changed the number of leaves and splits of the decision tree and evaluated the results. For further improvement we considered k fold cross validation for varied values of tree depth and thus, we were able to see an improvement in our results.

Secondly, we used Random forest algorithm for our further analysis to predict the Ratings and Global Sales. We evaluated the accuracy and other measures like precision, recall and F1 score for various values of n_estimators for the random forest algorithm. We chose to train on this algorithm as its nicely handles the missing values in the dataset (which in our case are too many), apart from this it does not overfit the model and provides wrong results and performs well on classification data.

Third, we used the XG Boost algorithm for even better performance than the Random forest algorithm. Even though the results obtained using the Random forest algorithm were quite satisfactory, but still we used this algorithm, to learn something beyond the scope of the class and improve our results for predicting the Ratings of a video games. XG Boost is an ensemble technique and uses Bagging and boosting for making predictions. We trained the XG Boost model based on 2 parameters namely tree depth and n_estimators, so we iterated over a range of values for these parameters and plotted the accuracy for each result.

Next, for making predictions on the sales we again used Decision Trees as it is much faster to train as compared to the simple neural networks for comparable performance and has less time complexity and we obtained good results. Then we also, used Random forest due to lot of Null values in our dataset, as it performs good with Null values. Again, we obtained good results of nearly 98.7 percent.

Additionally, we created a lot of visualizations and made inferences for sales from them based on Genre, Platform, Publisher, Critic Scores, and User scores. We chose to make solid inferences from visualizations as we wanted to learn about creating effective visualizations and how could they help us in interpreting results (as shown by TA Josua). Also, it’s a great method to make inferenced from.

Evaluation approach:

As, we had three different problems that need to be solved for the scope we defined for our problem.

For the first part, to find the Ratings (Age Distribution) of individual video games we first split the dataset into training and testing data using a randomizer function.

- Then cleaned the dataset and removed the rows that had NULL values in the essential columns for video game such as Name, Publisher, User score, Developer, and ratings as if would make no sense to make predictions for a video game whose Name, Developer, Publisher, and Ratings are previously not known, and this would produce an error as our model will get trained for NULL Ratings or it will create a new category called NULL, which will not be correct. Next, correlation has been plotted in the form of heatmap where the blue color represents positive correlation and red represents negative correlation between the variables. As we are considering Rating as our target variable, we can see in the notebook that only Genre, Critic Count and Year of Release are the three variables that are positively correlated to it, few of the variables are negatively correlated and most variables are independent as represented on the color scale.

- Then we tried to predict with the simple Decision Tree algorithm by training on 70 percent data considering all the columns present in our data frame except the sales parameters. We choose the decision trees algorithm as it is much faster to train as compared to the simple neural networks for comparable performance and has less time complexity. Apart from this it is robust as compared to other algorithms. So, we decided to perform our analysis based on this algorithm. But initially we got an accuracy of nearly 60 percent, with a precision score of 32 percent, recall score of 32.1 percent and a recall of 32.1 percent. So, the model is not good enough to predict values of Ratings accurately. So, next we performed the feature importance, to see what all features are important for training our Decision Tree classifier and we saw that the most important features were Name, Platform, Year of Release, Genre, Publisher, and Critic Count. Then we trained our Decision Tree algorithm based on these 6 parameters and got an accuracy of 65 percent, precision of 34.8 percent, recall of 35 percent and F-1 score of 34.9 percent, which was an improvement over our previous version, but still not good enough to predict correct values of Ratings.
- So, next we considered to change parameters like tree splits and leaves and tried to evaluate if there are any improvement on the training of our Decision tree algorithm. We plotted the accuracy for different values of these parameters and found out that the maximum accuracy of 65 percent would occur at a value of 30 for leaf and 25 for splits in the decision tree, after which the accuracy fell-down drastically with no other peak. Therefore, we did not receive any improvement in our results and then as a final step in the decision tree algorithm, we decided to perform a 10-fold cross validation with different values of tree depth to improve the result, but we got the best accuracy of only 62.1 percent with the value of tree depth as 15. We had plotted this result for many other values of tree depth, but the ones that are shown in our python notebook are best ones received so far. So, by this we summarized our approach of using the Decision tree algorithm and predicting results.
- As, the next step for predicting the Ratings, we first cleaned our dataset using a new clean function where we removed the rows that had NULL values for video game Name, Publisher, User score, Developer, and ratings as described above.
- For our further analysis we used the Random forest algorithm to evaluate the value of Ratings on our test data and see of its an improvement over the other algorithm that we

have used so far. We evaluated the accuracy and other measures for various values of n_estimators for the random forest algorithm and the best accuracy, recall, precision and F1 score was received at n_estimators value equal to 500. For this case we got the accuracy of 70 percent, precision of 54.7 percent, recall of 50.4 percent and F1 score of 51.8 percent, which are the best results so far. So, we can easily see that this algorithm is the best model that we have trained so far and is quite accurate in predicting the values of Ratings of video games on the test data. We chose to train on this algorithm as its nicely handles the missing values in the dataset (which in our case are too many), apart from this it does not overfit the model and provides wrong results and performs well on classification data.

- Next, we tried to train our model based on the XG Boost algorithm. It's an implementation of gradient boosted decision trees designed for speed and it performs better than the other algorithms. We choose this algorithm to have better results for predicting the values of Ratings. We trained the XG Boost model based on 2 parameters namely tree depth and n_estimators, so we iterated over a range of values for these parameters and plotted the accuracy for each result. We saw that the accuracy became constant after a value of tree depth, which was 50 for tree depth and 1500 for n_estimators. We had also, tried for a longer range of tree depth and n_estimators, but the results shown in the notebook are the best ones we received. After, training this model on these parameters we were able to get an accuracy of 76 percent, precision of 59percent, recall of 58 percent and F1 score of 58 percent. These results are the best ones received so far and we can train good model which could quite precisely and accurately predict the ratings of a game based on parameters like Game, Genre, Platform, its sales in different parts of the world, publisher, etc.

The second part that we considered was to predict the sales of the video games across the world.

- So, for performing this analysis, we first created a clean function which had the same functionality of the above defined clean function with an addition of converting our Global sales column value from continuous value to categorical by defining ranges like 1 to 3 million, 3 to 7 million and so on. We did this, as we just wanted to predict the range of sales of a game world-wide. We considered the Global sales feature, as it is the sum of all other sales and summarizes them.
- Then we implemented Decision Trees and Random Forest algorithm on this cleaned data as described in the notebook and obtained good results in both cases. By we had the best results for Random Forest which had an accuracy of 98.7 percent, Precision of 79.7 percent, Recall of 81.1 percent and F1 score of 79.3 percent.
- Further, we considered to use different visualizations for making inferences about sales, we were able to draw strong inference about what all parameters affect the sales of video games in different parts of the world. The observations that we saw are that *firstly*, Xbox 360, Wii and PS are the bestselling platforms across all the continents except Japan where

SNES, DS and 3DS are the bestselling platforms, but these are don't have good sales in other parts of the world.

- Secondly, we can see that if a game a Role-playing game produced in Japan on a SNES, DS or 3DS gaming platform then they are substantially famous than rest of the gaming genre.
- Third, an Action, Shooter or Sports video game in North America and Europe on an Xbox 360 or Wii platform are highly successful as compared to any other genre or gaming platform.
- Fourth, people in Europe are least interested in strategy and puzzle, role playing and adventure games.
- Fifth, based on the critic scores we can say that PS, Xbox 360, and Wii are among the best gaming platforms.
- Sixth, based on the Gaming Publishers we can say that Nintendo is the best-selling gaming publisher followed by Electronic Arts, but Electronic Arts has negligible sales in Europe.
- Seventh, Sports, Role Playing and Strategy games have the best critic scores as compared to the other game Genre.

<u>Assumptions/Limitations discussed</u>:

The only limitation that we had was we had a small chunk of data that gave results of sales of video games on Japan and we had a lot of missing values in our dataset, which we filtered out and replaced my mean and median values.

<u>Problem in Scope of class</u>:

1. We were quite accurately able to predict the Ratings of a new video game based on its parameters like its Publisher, Genre, Developer, Platform the game is made available for, its sales in different continents, its critic score, critic counts, user score and user counts.
2. We were very accurately able to predict the Global sales of a video game based on its parameters like Developer, Publisher, Genre, Producer, Critic Scores, etc.
3. We can infer from our data that Xbox 360, Wii, and PS are the bestselling platforms across all the continents except Japan where SNES, DS, NES and 3DS which are Nintendo based gaming platforms are the bestselling platforms, but they don't have good sales in other parts of the world.
4. We came to know that if a Role-playing game is produced in Japan on a SNES, DS, NES or 3DS gaming platform then it will most likely be successful as compared to other gaming platforms.
5. If a publisher produces an Action, Shooter or Sports video game in North America and Europe on an Xbox 360 or Wii platform it will be highly successful as compared to any other genre or gaming platform.

6. People in Europe are least interested in strategy and puzzle, role playing and adventure games and if they are produced on gaming platforms like Xbox 360, PS, Wii then they might be successful.
7. Based on the critic scores we can say that PS, Xbox 360 and Wii are among the best gaming platforms.
8. Based on the Gaming Publishers we can say that Electronic Arts is the bestselling gaming publisher followed by Nintendo, but it has negligible sales in Europe. So, we can infer that for Electronic Arts to gain popularity in Europe and increase their sales, they can release their games on Gaming Platforms like PS2 or Xbox 360 and make Action or Shooter games.
9. Sports, Role Playing and Strategy games have the best critic scores as compared to the other game Genre.

What did you change from your original proposal and why:

We did not change much from what we had proposed, but we came up with better results from what we had proposed before. We had proposed that we will use K means modeling to find which video game genre will be more popular but instead we used Decision trees and Random forest as they were more accurate for our case in Predicting Sales.

We also, proposed to use Decision Tree/ Random Forest modeling to find which game publisher will be more popular in the coming years based on the features in dataset (Popularity is the Ratings of that Game). So, for this we used other algorithms including Decision trees and Random forest and got the best results using the XG Boost algorithm. So, the criteria for becoming more famous was to determine the Ratings of the game in different parts of the world.

Apart from this, we also inferred from our visualizations that which Platform will be the best selling in the future depending on the current sales of video games in different parts of the world. We implemented this as an additional part of our project.

Team Evaluation:

It was never a big task for us to meet, discuss and work on this project. It was really an enlightening experience, working on this project and most of all learning so much about the value of DATA.

Arhant Jain 5/10
Harsh Yadav: 5/10