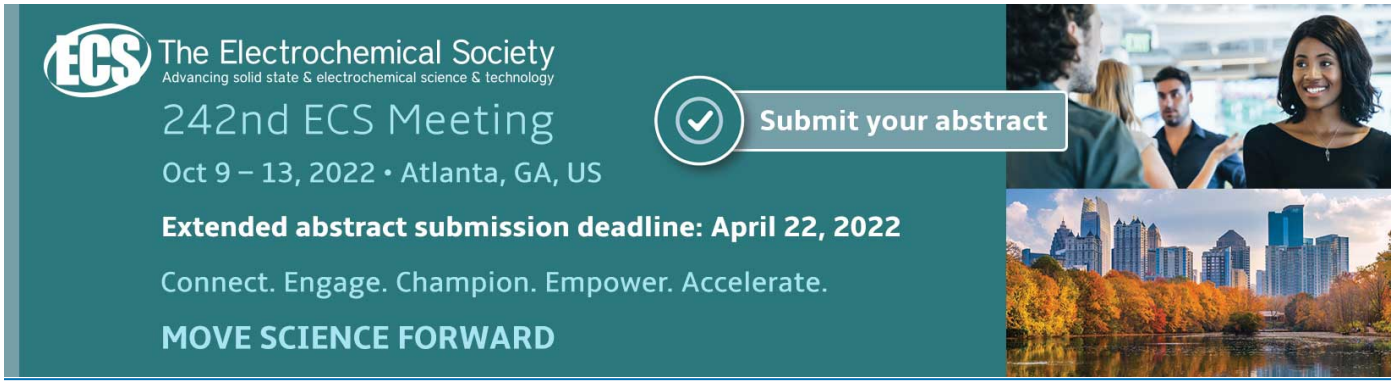
Empirical Analysis on Sales of Video Games: A Data Mining Approach



IOP Conf. Series: Journal of Physics: Conf. Series **1049** (2018) 012086 doi:10.1088/1742-6596/1049/1/012086

Empirical Analysis on Sales of Video Games: A Data Mining Approach

Amar Aziz[1], Shuhaida Ismail[2], Muhammad Fakri Othman[1], Aida Mustapha[1]

[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia.

[2]Faculty of Science and Technology, Universiti Tun Hussein Onn Malaysia, 84600 Pagoh, Muar, Johor, Malaysia.

E-mail: shuhaida@uthm.edu.my, fakri@uthm.edu.my, aidam@uthm.edu.my

Abstract. This paper studies factors that make the sales of video games becomes a blockbuster. The dataset used is collected from an online database maintained by VGChartz.com. Using the dataset, the Rapid Miner tool is used to select the features or factors and produce efficient estimation of the data. The techniques used in this project included the k-Nearest Neighbour (k-NN), Random Forest and Decision Tree. The factors and differences in the results are deliberated and discussed.

1. Introduction

In technology world, people from various age ranging from small kids to adults play the video games. This has lead to the spike in the sales of video games nowadays. Video games are released by major publishers across many popular hardware platforms. It provides the only real experience of interactive entertainment offered by the modern technologies. It also provides a rapidly growing form of entertainment and is being used for educational as well as business purposes. In the previous decade, several major video gaming releases have raised the bar of conventional entertainment goods in terms of revenues earned. There are several types of blockbuster video games on sale today such as Grand Theft Auto IV' by Rock Star Games and Call of Duty' Series by Activision. These types of video games have produced a series of annual records for revenues over the course of a three-year period.

Predictive modelling has long been the goal of many individuals and organizations. This science has many techniques, with simulation and machine learning at its heart. Aside from the potential of simulations, machine learning techniques are known for their ability to uncover hidden data trends. While the choice of algorithms used in each study may differ, they all had one common similarity, they give choices that human track experts made and are able to use the data to create arbitrage opportunities [7]. For example, [5] used Back-Propagation Neural network with Principal Component Analysis to predict the weekly video games sales. Other than that, transformation of classification trees into a decision-analytic model has also been used in solving the value-maximizing game development policy. The results showed that the compact predictive models created by data mining algorithms can help to make decision-analytic feed-forward control feasible, even for large, complex problems [6].

The main purpose of this study is to find out the contributing factors that lead video game sales becoming blockbuster. The collected dataset consisting of approximately 1,800 observations relating to individual video games titles released across a various platforms. This study uses a unique dataset of video game title to estimate the effect of dependent variables that used to empirical the lifetime unit sales of games released in the US in September 2010, which are collected from online database maintained by VGChartz.com. Then, the Rapid Miner tool is used to produce efficient estimation of the data. Aside from finding the factors that lead video game sales becomes blockbuster, the experiment also estimate the effect sales of video games becoming blockbuster.

The remaining of this paper proceeds as follows. Section 2 presents the methodology of the project including the machine learning algorithms and the dataset, Section 3 presents the experimental setup, Section 4 presents and discusses the results, and finally Section 5 concludes the work.

2. Methodology

In this research, a benchmark dataset which was originally collected and analyzed by Dr. Joe Cox is used. The dataset was collected from two Portuguese secondary schools during 2005-2006 by using two different methods which are mark reports and questionnaire. Decision Tree is used to predict and to find the correlation between features and as before for pre-processing the dataset. Data cleaning is performed in order to remove noise and to correct the inconsistencies exits in the data. The next step is merging two different data sets by using data integration technique. Classification is performed to analyse the statistical value between pair or two items. For this research, techniques such as k-NN [2], Random Forest [3] and Naïve Bayes [4] will be used to identify the difference results and outputs. The different and descriptions of techniques used during this project were documented in following section. Figure 1 shows the steps taken in the methodology to perform the experiment.

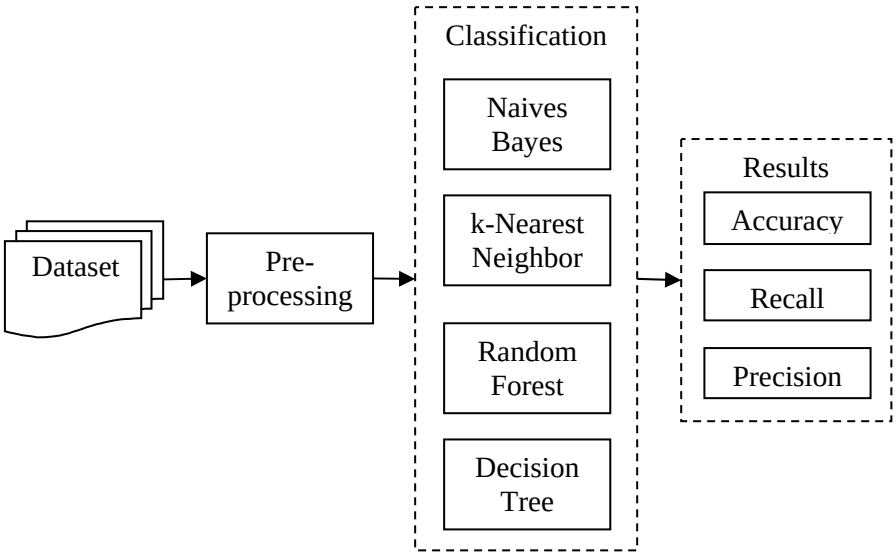


Figure 1: Methodology

A. Machine Learning Algorithms

There are four machine learning algorithms used in the experiments, which are the naïve Bayes, k-Nearest Neighbor, Random Forest, and Decision Tree.

- Naïve Bayes: This technique is used to create a Bayesian model that predicts the value of a target attribute (often called class or label) based on several input attributes of the dataset. The naive Bayes technique was used by previous work in [5].

- k-Nearest Neighbor: By comparing a given test example with training examples that are similar, the k-Nearest Neighbour algorithm is learning based on the analogy. The k-nearest neighbour algorithm is also one of the simplest of all machine learning algorithms.
- Random Forest: The Random Forest operator is used to generate a set of random trees. The random trees are generated in exactly the same way as the Random Tree operator generates a tree. The resulting forest model contains a specified number of random tree models. The number of trees parameter specifies the required number of trees.
- Decision Tree: The technique used to create a classification model that predicts the value of a target attribute (often called class or label) based on several input attributes of the dataset.

B. Dataset

Rapid Miner studio is used as the tool to import the data set and the attribute and type for the data is assigned before into the process. The attribute and the data type are shown in Figure 2. The data is reviewed and normalization is performed accordingly. Normalization is a pre-processing technique that used to rescale attribute values to fit in a specific range. Data normalization is very important especially when dealing with attributes that have different units and scales. For example, some data mining techniques use the Euclidean distance. Therefore, all attributes should have the same scale for a fair comparison between them. In other words, normalization is a technique used to level the playing field when looking at attributes that widely vary in size as a result of the units selected for representation. This operator performed normalization of selected attributes. Four normalization methods are provided. These methods are explained in the parameters.

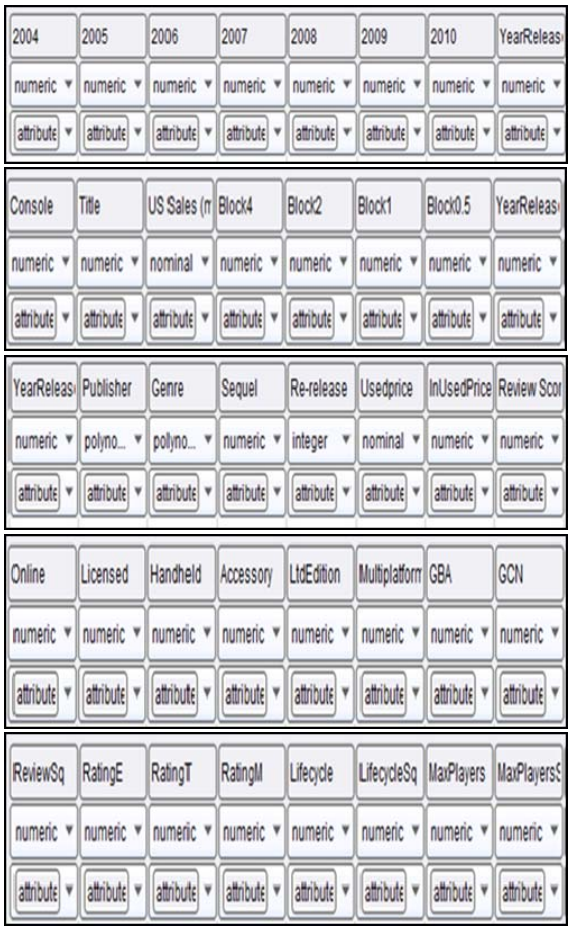


Figure 2: Attributes of the Dataset

3. Experiments

This study uses a unique dataset of video game title to estimate the effect of dependent variable that used to empirical the lifetime unit sales of games released in the US in September 2010, which are collected from online database maintained by VGChartz.com. RapidMiner Studio Basic is use as a tool to run the experiment. RapidMiner Studio Basic is a code-free environment for designing advanced analytic processes with machine learning, data mining, text mining, predictive analytics and business analytics [1].

A. Data Pre-Processing

Data pre-processing technique that has been used in this experiment is normalization. Normalization technique is generally used to rescale the attribute values in the dataset as shown in Figure 1. Other than normalization, data cleaning and data integration are also use in order to remove noise, correct the inconsistencies and merge two different data sets respectively.

B. Data Transformation

For our work, cross-validation technique is used to estimate the statistical performance of the learning operator and to estimate the accurate of model performance in training and testing phase.

C. Operators Parameters

Based on the dataset, suitable operators are chosen in order to produce the result and to fix error occurs during the experiment. Table 1 shows the details and the parameter of the operators used in this work.

D. Assessment Criteria

The results of each technique, which are Naïve Bayes, Decision Tree, K-NN and Random Forest were classified into three parameters which are accuracy, recall and precision. Assessments criteria accuracy is calculated by taking the percentage of correct predictions over the total number of examples, while the weighted recall is the result of an average of recall of every class. The weighted precision on the other hand, is calculated by taking the average of precision of every class.

Table 1: Operators and Parameter

No	Type	Parameter
1	Frequency-based Discretization	Filter type: regular expression Regular expression: Usedprice Number of bin: 2 Range name type: long
2	Decision Tree	Criterion: gain ratio Maximal depth: 20 Confidence: 0.25
3	k-Nearest Neighbor (k-NN)	k: 1 Measure types: Mixed Measures Mixed Measure: Mixed Euclidean Distance
4	Random Forest	Number of Tree: 10 Criterion: gain ratio Maximal depth: 20 Confidence: 0.25
5	Naïve Bayes	Laplace Correction: yes
6	Performance Classification	Main criterion: First Accuracy: Yes Weighted mean recall: Yes Weighted mean precision: Yes

4. Results and Discussion

The accuracy, recall and precision of the data were gathered by based on four machine learning algorithms; Naïve Bayes, Decision Tree, k-NN and Random Forest. The result of the data by using these techniques were documented in the following section. Frequency-based discretization was used to convert the selected numerical attributes into nominal attributes by changing the numerical attributes into a user-specified number of bins. Table 2 shows the final results for Naïve Bayes, Decision Tree, K-NN and Random Forest.

Table 2: Results for Naïve Bayes, Decision Tree, k-NN and Random Forest

Techniques	Accuracy	Recall	Precision
Naïve Bayes	81.58%	43.00%	44.12%
Decision Tree	99.55%	86.61%	86.20%
K-NN	24.86%	15.07%	13.97%
Random Forest	26.89%	3.45%	0.93%

The results in Table 2 showed that the percentage of accuracy for all method using Decision Tree produced the same percentage, which is 99.55% and the percentages of recall and precision for all method produced 86.61% and 86.20% respectively. Meanwhile the highest accuracy, recall and precision is obtained using z-Transformation method for k-NN technique with the values of 24.86%, 15.07% and 13.97% respectively. The highest accuracy, recall and precision using Random Forest technique are recorded at 26.89%, 3.45%and 0.93%, respectively. Besides, the highest accuracy, recall and precision percentage using Naïve Bayes technique is the Interquartile Transformation which produced 81.58%, 43.00% and 44.12% respectively. The experiments also showed that the most contributing factors to the blockbuster game sales include the year that the game is released, the genre, the price, and the review score.

5. Conclusion

In conclusion, from dataset there has been several technique perform to compare the result to find the outcome of what makes blockbuster video game. The techniques are Decision Tree, K-NN Result and Random Forest. The results are then compared between three criteria, which are accuracy, recall and precision. The results showed that Naïve Bayes technique using Interquartile Transformation have much closer accuracy, recall and precision compared to decision tree technique. It is proven that this method is more suitable to calculate the accuracy, recall and precision of the work compared to others.

Acknowledgement

The research publication is supported by Universiti Tun Hussein Onn Malaysia (UTHM) under Short Term Grant (STG), Vot.U367 and is also seconded by Gates IT Solution Sdn. Bhd.

References

[1] RapidMiner Studio 6.5. (2016). Retrieved on 27 November 2016 from http://docs.rapidminer.com
[2] K-NN (k-Nearest Neighbor). (2016). Retrieved on 27 November 2016 from http://docs.rapidminer.com/studio/operators/modelling/predictive/lazy/k_nn.html.
[3] Random Forest. (2016). Retrieved on 27 November from http:// docs.rapidminer.com/studio/operators/modelling/predictive/trees/parallel_random_forest.html
[4] Naïve Bayes. (2016). Retrieved on 27 November 2016 from http://docs.rapidminer. com/studio/operators/modelling/predictive/bayesian/naive_bayes.html
[5] Marcoux, J., & Selouani, S. A. (2009). A hybrid subspace-connectionist data mining approach for sales forecasting in the video game industry. In Computer Science and Information Engineering, 2009 WRI World Congress on (Vol. 5, pp. 666-670). IEEE. Brydon, M., & Gemino, A. (2008). Classification trees and decision-analytic feedforward control: a case study from the video game industry. Data Mining and Knowledge Discovery, 17(2), 317-342.

[6] Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). Predictive modelling for sports and gaming. In Sports Data Mining (pp. 55-63). Springer US.
[7] Cox, J. (2014). What makes a blockbuster video game? An empirical analysis of US sales data. Managerial and Decision Economics, 35(3):189-198.