Estimating Video Game Success using Machine Learning

Research Proposal · August 2019

DOI: 10.13140/RG.2.2.14389.01767

1 author:

Aashish Prasad
National College of Ireland
8 PUBLICATIONS **1** CITATION

SEE PROFILE

National College of Ireland

Estimating Video Game Success using Machine Learning

MSc Research Project
Data Analytics

Aashish Prasad
Student ID: x17170826

School of Computing
National College of Ireland

Supervisor: Sachin Sharma

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Aashish Prasad
Student ID:	x17170826
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Sachin Sharma
Submission Due Date:	02/08/2019
Project Title:	Estimating Video Game Success using Machine Learning
Word Count:	5465
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	1st August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Estimating Video Game Success using Machine Learning

Aashish Prasad
x17170826

Abstract

Video games are one of the most popular source of entertainment within our society. With the evolution of science and technology the demand of better quality contents with respect to improved graphics and more real life experience have increased for video games. Such demand from users has led to high production cost, ultimately leading to higher business risk for the development companies. In order to minimize this risk this research proposes a machine learning approach towards estimating video game success before the initial release. In addition to the descriptive features the research uses game plots which plays an important role to keep the player engaged and reduce churn rate. Hence, in this research key sentiments extracted from the game plots along with descriptive features will be used to classify video game success level as Low, Medium, High and Very High. The implementation involves use of Word2vec based techniques for Natural Language Processing (NLP). The classification algorithms for the proposed research are Support Vector Machine (SVM) , Artificial Neural Network (ANN), K-Nearest Neighbor (KNN) and Random Forest (RF). These models will evaluated using Confusion Matrix with Accuracy, Precision and F-Score.

Contents

1 Introduction

Video Games have gained huge popularity among most age groups in recent decades and becoming part of our daily life. The emergence of a vast number of electronic devices, such as personal computers, mobile devices and gaming consoles at low-cost in has promoted accessibility of video games through the mass population. Also the easy availability of internet over 4G and 5G network has created accessibility to games over mobile devices. With the growth rate of 6.77 percent, the global gaming market is expected to rise at 158.33 USD by the year 2023, Zhang et al. (2019).

Realizing the market potential organizations from all around world have stepped into this industry with few big names like EA Games and Ubisoft. Despite the rapid market growth it is getting difficult to retain significant profit due to high budget as a result of growing size of development team and high quality contents, Bailey and Miyata (2019). With time game developers have realized certain patterns followed by users of human behaviours towards video games that can be used to generate maximum investment returns, Ahmad et al. (2017). Still with the increasing dynamic market for video game industry it is difficult to estimate the chances of video game success. With large investments into game development, the success of the games sometimes become a crucial factor for the survival of business organizations. In order to minimize business risk and increase the chances of video game success, machine learning algorithms can be used as a reliable tool.

Determining success of a video game using monetary value is difficult when considering multiple games from different publishers with a variety of game features. For example, sale of 10,000 copies for a low budget game could be a huge success. Whereas the same sales figure for a high-end game involving heavy production cost could be a nightmare for the organization. Also, success based on sale of game can occur at different point of time for individual games. It is generally seen that games that are a sequel to a previous release may gain popularity faster than games released with new title. This does not mean that the later could not reach the same level of popularity in late time period.

To deal with this challenge, in this project success of a video games will be measured by the overall user ratings, assuming the fact that higher user rating is directly proportional to greater sales in long run. Apart from the above descriptive features, there are also other factors which can impact success of a video games. Such as, social media responses and influencers, example bloggers. But a game needs to obtain minimum level of popularity to get such social attention, Trněnỳ (2017). One **influencing factor** that has been never

taken into consideration while analyzing success is **story-line**, which plays a very crucial role in determining success in long run. A research by Bormann and Greitemeyer (2015) showed the positive influence of in game story telling on players. A good story keeps the player interest up and thus reduces the chance of player churn. In a video game users always prefer content quality or content quantity, Bailey and Miyata (2019). It is a similar case we have seen in movie success at box office. A movie may have a very good cast but if the story does not appeal to the intended audience it is ought to fail. Similar criteria applies to video games but it is generally seen in the long run in form of user experience or ratings. Thus, this factor shall be considered in this research along with the descriptive features.

In addition, the game developers may launch Downloadable Content (DLC) post release inorder to keep the game running and players sometimes have to pay extra charges to avail these premium features. Hence, such DLCs generate revenue even after initial sale of the game copy, Ahmad et al. (2017). As stated above, players need to be kept engaged with interesting plot which will directly or indirectly increase revenue from DLCs.

This paper proposes predictive analysis using various descriptive features of video games along with sentiments derived for specific video game story/plot. The two widely used approaches sentiments analysis are lexicon based and machine learning. Unlike machine learning approach, Lexicon based classification does not require any pre-processing and training of classifier. Whereas, machine learning sentiment analysis also known as supervised learning approach has been seen more accurate in previous researches, Dhaoui et al. (2017) Ge et al. (2018). Hence, the proposed methodology uses sentiment classification of video game story using word2vec vectorization technique as natural language processing combined with deep neural network. The output for which shall be merged with the descriptive data to form a feature rich dataset of over 10,000 games. Using label encoder all the textual data will be converted into machine readable or numeric form before passing into the machine learning algorithms; Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbour (KNN), Random Forest (RF). To evaluate the proposed models Confusion matrix, Precision, Accuracy and F score will be used.

The research questions undertaken in this research is

How video game plots can be effectively utilized to improve the classification accuracy while predicting its success rate using machine learning algorithms when measured based on user rating?

2 Related Work

2.1 Research on video games using machine learning

Many games do offer purchase of virtual products or items within the game to enhance player progress. To improve such in-game purchase a recommendation system was developed by Bertens et al. (2018). Unlike traditional recommendation systems the principle of this system was based on predicting rating of an item by the player. The paper compares performance of an Ensemble based model using randomized tree (ERT) with Deep Neural Network (DNN). The data used for study was taken from a Japanese card game containing player information and eight different types of items for in-game purchase. According to the author both ensemble approach and neural network are reliable

algorithms. The main advantage of an ensemble approach is that the models can be trained on parallel processors ensuring less execution time with better accuracy. The results for the models were evaluated into three scenarios namely item purchased on next purchase date, next purchase and purchased within the given window. Both the DNN and ERT performed similar with accuracy of 81 and 74 percent for the first two categories with slight improvement in the ERT for the third category with 91 percent.

Previous research on predicting video game success were mostly based on estimating sales figures. However, determining success based on sales is difficult when comparing multiple games developed with different motivation. The expectation of the development team or the publisher in terms of monetary value differs. In order to find an alternate to this a research on video game success was accomplished by Trněnỳ (2017). The success criteria was estimated based on number of active players or game owners reached after two months of game release date. The data for the research involved steam game details collected from steam charts and steam spy which included descriptive features such as price, genre, developer, publisher, etc. The results showed 95 percent precision and 75 percent recall rate with random forest. However, the research had few drawbacks. The number of owners is not a reliable source to determine success as it does not ensure player stability to continue in the game. Also, the game sales continue ever after several months of its release and the actual revenue could vary from the initial sold copies. So even though the results showed higher accuracy the approach cannot guarantee reliable results when tested into actual scenario which could be much more dynamic in nature.

A research on determining video game completion rates was conducted by Bailey and Miyata (2019) on a sample data of 725 games collected from steam service. The results of correlation analysis among completion rate and different descriptive factors indicted significant correlation with user ratings, price, genre, publisher and least with release date. Furthermore, regression analyses on various factors showed user ratings and genre proved to be signification predictor in estimating completion rate.

In this research, video game success will be predicted using similar descriptive features used by Trněnỳ (2017). Along with this game story shall be taken into consideration, with the previously used descriptive features which plays an important role by reducing chances of player churn. Also, the success criteria will be determined based on overall ratings by the users which is more reliable metric to analyze success level categorized into four classes as Low, Medium, High and Very High. This differs from previous approach that used number of active players after two months of game release as mentioned in the above paragraph.

2.2 Study on Sentiment Classification of Text and Feature Extraction Techniques

Sentiment analysis is widely used for Natural Language Processing (NLP) to study sentiments in textual data. The two types of sentiment classification used are lexicon based using weighted words and machine learning approach. According to Giatsoglou et al. (2017), Lexicon based analysis works on overall sentiments of the documents, however they do not consider the context in which the text has been written. In contrast word embedding based analysis used in machine learning such as word2vec converts textual data into vectors and successfully capture semantic between texts. The results of which are beneficial towards sentiment classification. Documents containing moview review in English and Greek language was used for the experiment. The proposed methodology

converts textual documents into numeric vectors using word2vec algorithm. A hybrid vectorization process was used that combined lexicons or words representing emotions with word2vec. The experiment showed word2vec trained models resulting in improved accuracy. Similar research on the above two approaches was accomplished by Dhaoui et al. (2017) on social media conversations using a sample data of 850 user comments from 83 Facebook pages. The experiment was conducted using RTextTools package available in R language for machine learning. Using evaluation metrics such as True Positive Rate (TPR), False Positive Rate (FPR), Precision and Recall the model was evaluated. The result performance were similar for both approaches with 74% when only positive and negative polarity were considered. However, there was substantial difference in their classification ensembles.

To improve sentiment analysis a feature based on word2vec is proposed by Alshari et al. (2017). The proposed method was tested on a internet movie review dataset and the results were evaluated by the author using Support Vector Machine (SVM) and Logistic Regression classifiers. The feature set consists of clustering of terms by opinion words from sentiment lexical based dictionary. To redistribute the terms based on polarity based on space transformation was done on negative term vectors. Based on these set of clusters document vectors were produced in form of small matrix. The proposed method reduced the complexity of classifier with accuracy level of 93% and 86% for logistic regression and svm respectively. Ge et al. (2018) says sentiment analysis can be viewed as a quantitative information based on emotions. The author has used word2vec for feature extraction for sentiment analysis the results for which were later fed into machine learning classifiers. To determine important structural elements in the data Principal Component Analysis (PCA) method was used. The data was split into train and test sets using sklearn library and applied to the Logistic Regression, SVM and Random Forest (RF) classifiers. Performance metrics were evaluated using TP and FP rates. Although, Logistic regression showed best performance among the three classifiers, the accuracy score was just around 60%.

According to Yang and Xia (2016) sentiment analysis could be termed as a classification problem and the solution could be defined into multiple classes. The author has used word embedding (word2vec) in the process of sentiment classification of Chinese documents. To differentiate between text word segmentation was applied, after which the texts were vectorized into their numeric format using Gensim package in python programming language. Further, the vectors were used as input for Convolutional Neural Network (CNN). The model was evaluated using Precision, Recall and F1 score. An accuracy of 92% was achieved by the proposed model while the baseline SVM and Naive Bayes scored level of 73% and 67%. The feature vector technique and architecture of neural network was the main reason for such high accuracy.

Xiao et al. (2018) proposed a text classification model to classify patent texts based on word2vec and long short term memory (LSTM) algorithm. The author has used word2vec to solve the problem of dimension that occurs in traditional method by converting one hot encoder into continuous values, thus avoiding over-fitting. In the experiment word2vec is used to train Chinese textual data from Wikipedia. Initially, the texts were converted into numeric vectors and the result for which were used as an input for LSTM model for text classification. The literature states that text classification model based on combination of word2vec and LSTM has much higher accuracy when compared to individual LSTM model as word2vec reduces over-fitting of data. The experiment result showed that the combination outperformed K Nearest Neighbor (KNN) and Convectional Neural Network

(CNN) with an accuracy of 93%.

In this research, word2vec will be used with neural network to derive sentiments from video game story/plot and to convert textual data present in the descriptive video game dataset into vectors for the machine learning models.

2.3 Algorithms used for Multi-Class Classification

Algorithms could be used to classify data into two or more classes. Thaseen and Kumar (2017) have used Support Vector Machine (SVM) as a multi class classifier for intrusion detection model. A chi-square feature selection technique is used with SVM for higher accuracy with low false positive rate (FPR). The proposed methodology employs normalization of dataset during preprocessing stage followed by feature selection to remove low ranking attributes. The data is then divided into the validation, training and test set. Cross validation is performed on the validation set using K-fold validation technique to compute the performance measure. The final optimal parameters are used in the training and test set for the SVM model to classifies into label intrusion classes as Normal, Probe, DoS, U2R and R2L. The model showed improved accuracy with 98% , reducing the number of false alarms achieved by previous approach. Liu et al. (2017) says SVM works on structural risk minimization principle in which it splits feature vectors using hyperplane with maximum margin. As SVM was created as a binary classifier, one verses one strategy need sto be applied for multi-class classification. SVM has been used for multi-class sentiment classification of online texts consisting of three different public datasets, which is further divided into 12 subsets. Each dataset was labeled into 3,4 or 5 classes with range of positive and negative sentiments as 1 to 5. Bag-of-words (BOW) has been used as feature selection technique and the selected features were ran five machine learning algorithms including SVM. These are Decision tree, Naive Bayes, SVM, Radial basic function neural network (RBFNN) and K-Nearest Neighbor (KNN). On the basis of accuracy SVM performed the best among all with highest accuracy of 82%, while KNN had the fastest execution time.

Zhang, Zhang, Liu, Li, Yang and Tian (2017) used multi-class SVM to predict weather condition by classifying it into three categories as sunny, foggy and cloudy. The data from Photovoltaic (PV) power station situated in China is used to train the proposed model which was collected for a month. The model is a replacement towards traditional measurement and monitoring system (physical instruments) to determine weather conditions in near future. PCA analysis was conducted to find the key correlation between attributes int the collected data and further used as an input to SVM. Using radial basic function kernel the multi-class SVM is implemented resulting in 88% accuracy which can save money on physical instruments.

Jaiswal and Banka (2017) used Artificial Neural Network (ANN) to classify epileptic electroencephalogram signals. Epileptic a brain disorder which needs to be classified into seizure and non seizure signals using neural signals captured by EGG from human brain. Local Descriptive Patter (LNDP) and One-dimensional Local Gradient Pattern (1D-LGP) are the two feature extraction techniques used in the experiment. The data was transformed into 8 bit code using the above feature extraction technique. The transformed data is used to train the four different classifiers, namely KNN, Decision Tree (DT), SVM and ANN. For evaluation 10 fold cross validation was used considering sensitivity, specificity and accuracy. During the experiment it was observed that KNN took the least time while the ANN took the most but it performed well in terms on classification

accuracy with 99%. In another research involving human brain ANN was used by Fuad et al. (2019) to classify Brain Balance Index (BBI) into 5 different groups as index1 (unbalanced brain), index2 (less balanced brain), index3 (moderately balanced brain), index4 (balanced brain) and index5 (highly balanced brain). The frequency data collected from EEG known as Alpha, Beta, Theta and Delta grouped by their respective frequency values were used as an input for the classification model. The ANN model used in this research had sigmoid activation function. Using confusion matrix the sensitivity, specificity and accuracy of the proposed model was evaluated. The results for sensitivity was within a range of 87% to 92% while the specificity was calculated between 94% and 98%. However, the overall accuracy achieved by ANN was 88%.

Based on text-speech a classification model was build by Klumpp et al. (2018) to differentiate patients suffering from Alzheimer disease. According to the author, Alzheimer can be detected based on speech signal or spoken texts of patients. Hence, speech data of healthy and non-healthy individuals were collected. The proposed model was build using bag of words technique for feature vectors. The model consists of one input layer, one hidden layer and one output layer with two node with rectified linear unit (ReLU) activation function. The accuracy achieved by the classifier was 84%. In contrast, to predict academic success level of students as low, medium, high Amrieh et al. (2016) used ANN classification with Decision tree, Naive Bayes. Data from an online educational website was used consisting of basic student details along with behavioural features such as discussion groups, visited resources, interaction in class. The cleaned data was processed to find the best features using a filter methodology and was used as input for classifiers. The results were further passed towards ensemble methods of bagging, boosting and random forest. Performance evaluation was done using TPR, FPR, Precision, Recall rate and accuracy. The results showed improvement with the bagging method for Decision tree and Naive Bayes, however, the accuracy level reminded the same for ANN with 79% which is highest value achieved within the entire experiment.

K Nearest Neighbour (KNN) is another popular classification algorithm with simple implementation and high accuracy. Zhang, Li, Zong, Zhu and Wang (2017) proposed two KNN classification algorthims, kTree and k*Tree. kTree was used to find the most optimal k value for testing data during training stage for the classifier. The k value learned during the training stage is then applied on testing set by KNN classification. This method of using different k values results in higher accuracy than the tradition fixed k value method with less running cost. Further, this kTree method is improved using K*Tree method in which extra sorting on training data was done on leaf nodes of kTree. Both the approaches resulted in improved accuracy of the classifier with 77%.

Krithika and Selvarani (2017) used KNN for agriculture application to classify variour types of leaf diseases. Texture feature from leaf images were extracted and converted into numeric values on which KNN classification was performed to identify leaf disease. Furthermore, Islam et al. (2018) implemented KNN for analysing emotions on social media platform. People express their emotions such as fear, happiness, joy, surprise with their comments on various social media platforms. Based on such data, a KNN classifier was build to detect depression among users. Similarly, based on personality test of individuals a KNN classification was proposed by Bhannarai and Doungsa-ard (2016) who are fit to work under agile process using the big five personality traits, namely; Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The model was tested with different k value, with highest accuracy of 65% when k=1 and a standard deviation of 9.99.

In order to maximize security over cloud infrastructure a KNN classification is used by Sarma et al. (2017). Cloud facility can be accessed by multiple users and are prone to unauthorized access. Some files with critical information needs extra security protection. To identify such files KNN classifier was used. The user files were categorized as Confidential, Strictly Confidential, Public and Open. The files categorized as Confidential and Strictly Confidential require high security. As per the author, KNN algorithm is preferable due to its high performing capability with multi-class classification problems.

According to, More and Rana (2017) with the rapid growth of data, imbalanced datasets are difficult to analyse. Due to this, applying classification algorithms on such datasets lead to biased output. Random Forest (RF) is an ensemble approach that combines multiple classifiers or decision trees to generate classification output. Due to the use of dimensional reduction and capability to deal with missing values, RF has been seen as a high performing classifier. When tested on a sample dataset, 86% of the data was correctly classified by RF. In a research on imbalanced data by Arafat et al. (2017) it is said that most traditional algorithms ignores minority class instances when faced with imbalanced dataset. The author has used a clustering based RF for multi-class classification. The data is divided into majority and minority groups and clustering is performed to group them in several small clusters until balanced data is achieved. The RF is trained using the balanced data and further the classifier uses majority voting principle to classify test data. The results showed better performance of RF with over 90% in most cases when compared with popular AdaBoost, RUSBoost and SOMTEBoost algorithms.

Wang et al. (2015) used RF to build a multi-class object recognition application for robot programming. Using features such as colour histogram, binary pattern, aspect ration, etc the RF classified 20 types of objects. The accuracy of model was compared with KNN, SVM and Decision Tree (DT) was highest among all with 99%. While SVM score 92%, KNN and DT scored 81% and 88% respectively. In another study conducted by Chaudhary et al. (2016)involving multi-class classification RF was optimized for better results. As per the author, most crops are destroyed due to diseases resulting in heavy financial losses. The data used in this research consists of climate and crop features labeled by various types of diseases. Using features with high correlation values and balancing the dataset using gain ratio the performance of RF is improved. The results were evaluated using accuracy, sensitivity and specificity values. 97% overall accuracy was obtained by RF algorithm.

In this research, a video game descriptive dataset along with sentiments from game plots with over 10,000 rows and 11 features will be used. SVM works on risk maximization principle Liu et al. (2017) and have shown high accuracy for classification tasks Zhang, Zhang, Liu, Li, Yang and Tian (2017). While ANN has also shown high accuracy and is good at dealing with complex classification problems using due to the structural concept of neural network. KNN is known classification algorithm with good results obtained in previous researches Zhang, Li, Zong, Zhu and Wang (2017). Lastly Random Forest is best in handling imbalanced dataset and have shown high accuracy because of its ensemble working principle More and Rana (2017). All the above four algorithms are well suited for classification and will be used in this research to estimate video game success.

2.4 Table of Comparison

Title	Description
Improvement of sentiment analysis based on clustering of word2vec features	Word2vec was used for feature extraction and sentiment classification with logistic regression, SVM and random forest. Clustering of vocabulary on the basis of opinion words was used to improve word2vec performance.
Mining educational data to predict students academic performance using ensemble methods	Using behavioural features student performance was estimated. Accuracy of 79% achieved by ANN model.
Improving video game project scope decisions with data: An analysis of achievements and game completion rates	Research on game completion rate. The study shows high correlation between game completion rate and descriptive features such as user rating, price, genre and publisher.
A machine-learning item recommendation system for video games	Built a recommendation system based on principle of predicting user ratings for in-game items using random forest and deep neural network.
Social media sentiment analysis: lexicon versus machine learning	Comparative study on sentiment analysis using lexicon and machine learning approach. Used true positive, false positive, precision and recall for evaluation.
ANN classification for the analysis of 3d EEG data in BBI	Sigmoid activation function used for ANN classification and evaluated using confusion matrix.
Local pattern transformation based feature extraction techniques for classification of Epileptic EEG signals	Created ANN and KNN classification models. Cross validation was done using K-fold technique with sensitivity, specificity and accuracy. Results showed KNN to be the fastest while ANN scored highest accuracy with 99%.
ANN-based Alzheimers disease classification from bag of words	ANN model classification model created using rectified linear unit activation function and multiple processing layers.
Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms	Compared SVM, neural network, KNN, decision tree and naive bayes classifiers on three different datasets. SVM performed the best among all with 82% accuracy score.

Title	Description
Intrusion detection model using fusion of chi-square feature selection and multi class svm	SVM was used as a multi-class classifier for intrusion detection. Result showed accuracy of 98% with reduced false positives.
Machine learning for predicting success of video games	Predictive models for video game success were created based on descriptive game features and sales value as success criteria.
Research on patent text classification based on word2vec and lstm	Word2vec was used for text classification with lstm algorithm. The combination resulted in improved accuracy with 93%.
A convolutional neural network method for chinese document sentiment analyzing	Used word2vec embedding for sentiment classification with convolutional neural network and achieved an accuracy of 93%.
Weather prediction with multiclass support vector machines in the fault detection of photovoltaic system	Radial basic function kernel used for SVM classifier for weather prediction into three classes as sunny, foggy and cloudy.
Research on Sentiment Analysis of Multiple Classifiers Based on Word2vec	Feature extraction and dimensionality reduction was accomplished using Word2vec and tested on multiple classifiers.
Sentiment Analysis leveraging Emotions and Word Embeddings	Used Word2vec for sentiment classification on textual data resulting in better accuracy.
Efficient knn classification with different numbers of nearest neighbors	Proposed Ktree, an extension of KNN for classification to find most optimal k value improving overall accuracy with 77%.
KNN file classification for securing cloud infrastructure	The research used 109 samples to classify into groups based on confidentiality using KNN inorder to increase cloud security.
Detecting Depression Using K-Nearest Neighbors (KNN) Classification Technique	KNN model used for classifying human emotions.
Agile person identification through personality test and kNN classification technique	Used KNN to classify individual personality with various k values. Highest accuracy of 65% achieved when k=1.
Review of random forest classification techniques to resolve data imbalance	Paper used random forest as it has dimensional reduction feature and high accuracy because of ensemble approach.
Multi-class assembly parts recognition using composite feature and random forest for robot programming by demonstration	Created KNN, SVM, DT and RF classifiers for object recognition program. RF achieved highest accuracy with 99%.

Title	Description
Cluster-based under-sampling with random forest for multi-class imbalanced classification	Used random forest on imbalanced dataset for multi-class classification. Results showed better performance of RF with 90% accuracy.
An improved random forest classifier for multi-class classification	Improved performance of random forest by balancing dataset using better feature selection and gain ratio.

3 Proposed Methodology & Implementation

In this research, we plan to estimate the success of level of video games using machine learning models. In order to achieve this goal, the proposed research will be based on Cross Industry Standard for Data Mining (CRISP-DM) methodology that follows a structured pattern for scientific experiment. Figure 1 shows the steps for the proposed methodology.

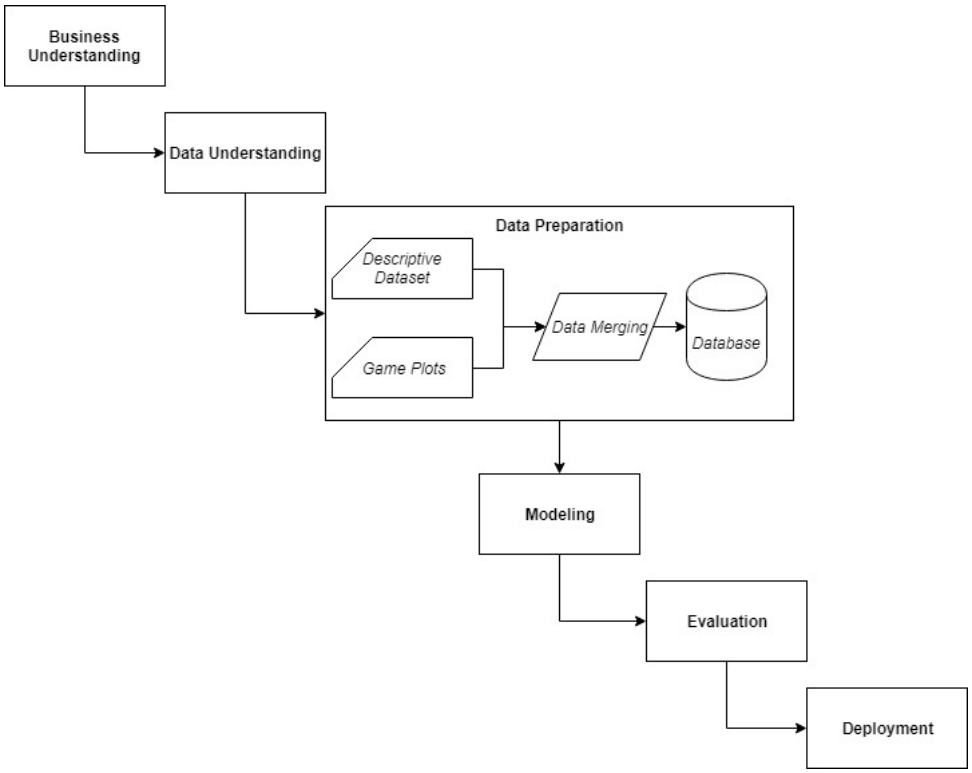

Figure 1: CRISP-DM

3.1 Business Understanding

Video games industry is one of the fastest growing industry. With the increasing popularity of games, the demand for high quality contents has also increased. This requirement directly affects the development time and cost with increased development team size and resources. As the business risk is high the failure can impact the organization heavily. With more number of competitors and dynamic choices of players, it is difficult to determine success or failure of a video game. To solve this problem this research proposes a machine learning approach to estimate success level of video games in the market. Estimation of video game success will be based on descriptive features such and name, price, genre, platform, etc and features derived from specific game plots. Based on the above factors, the success level will be determined as Poor, Average, Good and Excellent.

3.2 Data Understanding

The data in this research will be acquired from two different sources. The first source is Kaggle, which contains descriptive features as Name, Genre, Publisher, Age rating, User

Score, Critic Score, Sales figures, etc. The url for this dataset is `https://www.kaggle.com/kendallgillies/video-game-sales-and-ratings/version/1`. There are 17,000 rows in the dataset containing details about video games. As this research is on estimating video game success during pre-release phase, specific columns will be selected as not all values mentioned in the dataset would have been known before initial release; for instance, critic score and sales figures can only be occur after launch of the game. The target variable for the proposed models will be user score, which will be encoded as labels explained in the next stage of Data Preparation.

The second source of data is Wikipedia which contains video game story/plots which will be scrapped from the website.

3.3 Data Preparation

In order to transform the data into machine readable format and obtain high accuracy the data collected from both the sources must be brought down to the same level using Natural Language Processing (NLP). This step can be divided into three stages; 1) Processing Game Plots (texts) and extracting required sentiments; 2) Cleaning Video game dataset; 3) Data Merging.

In the first step, the text data of game plots will be tokenized and all the stops words shall be removed. Using stemming method the words will be transformed into its roots. These functions for tokenization and stemming are provided in NLTK python library. Next, word vectorization will be performed using word2vec that will convert the texts into numeric vectors and assign weights to each of the word. After the above preprocessing, the vectorized data will be passed into LSTM model for sentiment analysis. The output for the analysis will give values for different human emotions as happy, sad, fear, anger, surprise and disguise in the text.

In the second step, the video game dataset will be cleaned which will include removal of duplicate entries. Using data encoding, the target column, 'user score' will be converted into groups as Low for user score 4 and below; Medium for user score above 4 and below 6; High for user score between 6 and 8; Very High for user score above 8. As the machine learning models understands only numeric values, the above categories will be detonated in numbers as '1' for Low, '2' for Medium, '3' for High, '4' for Very High. Further, the all the textual data in the dataset such as name, genre, publisher must be converted to numbers for machine learning models hence word2vec will be used for this task.

Once data from both the sources are transformed, in the last step they will be merged to form a larger feature rich dataset. In order to standardize the variables, scaling function found in sklearn python package will be applied on this final dataset.

3.4 Modeling

The models to be used to determine video game success level are SVM, ANN, KNN and RF. These are the most preferred algorithms for classification algorithms as stated in literature review in section 2.

3.4.1 Support Vector Machine

SVM is a supervised learning model which is known for its high accuracy using kernel trick. Data points are separated by line(s) known as hyperplane and the classification is

achieved. The Scikit-Learn library in python will be used which contains functions for building svm classifier.[1]

3.4.2 Artificial Neural Network

ANN is a machine learning algorithm inspired by the work of human brain or the neural network. There are multiples connected nodes known as neurons that performs non-linear functions.[2] Python provides tensorflow library which can be used to build this model.

3.4.3 K-Nearest Neighbour

KNN is widely used to solve classification and regression problems. It works on the assumption that similar things are present at close proximity. [3]

3.4.4 Random Forest

Random Forest is an ensemble learning model used for classification tasks. It is a collection of multiple Decision Trees.[4]

3.5 Evaluation

To evaluate the classification algorithms confusion matrix will be used which provides number of correctly and incorrectly classified values as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Further, using these values, Accuracy, Precision and F-score will be calculated.

The accuracy defines the overall performance of the model, but to better understand the number of correctly classified cases for each class, precision and recall are necessary evaluation factors. This can be explained with the following example; **Scenario 1** the model incorrectly classifies a video game as 'low success', when it is actually comes under 'high' or 'very high success'. **Scenario 2** the model incorrectly classifies a video game as 'high success', when it is actually comes under 'low' or 'medium success'. Among the above two cases, **Scenario 2** is more harmful in comparison to **Scenario 1** as it can lead to a direct loss for the organization. Hence, the correct metric for the models in this research is 'Precision' as it uses False Positives (FP) in calculation.

The evaluation cases are divided into two experiments as mentioned in 3.5.1 and 3.5.2.

3.5.1 Evaluation based on Descriptive features

In this evaluation phase only descriptive features of the video game dataset shall be considered as independent variables. That are Name of the game, Platform, Genre, Publisher and Age Ratings to estimate the video game success.

[1] https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

[2] https://en.wikipedia.org/wiki/Artificial_neural_network

[3] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[4] https://en.wikipedia.org_wiki_Random_forest

3.5.2 Evaluation based on Descriptive features and Game Plots

In contrast with the above case, in this evaluation phase, all the features shall be considered including sentiments derived from game plots.

3.6 Deployment

Since this project is for academic research, at this stage a detailed report will be prepared which will contain project planning, implementation and evaluation of the results. Moreover, the challenges and techniques used to overcome will also be discussed in this report.

4 Proposed Project Timeline

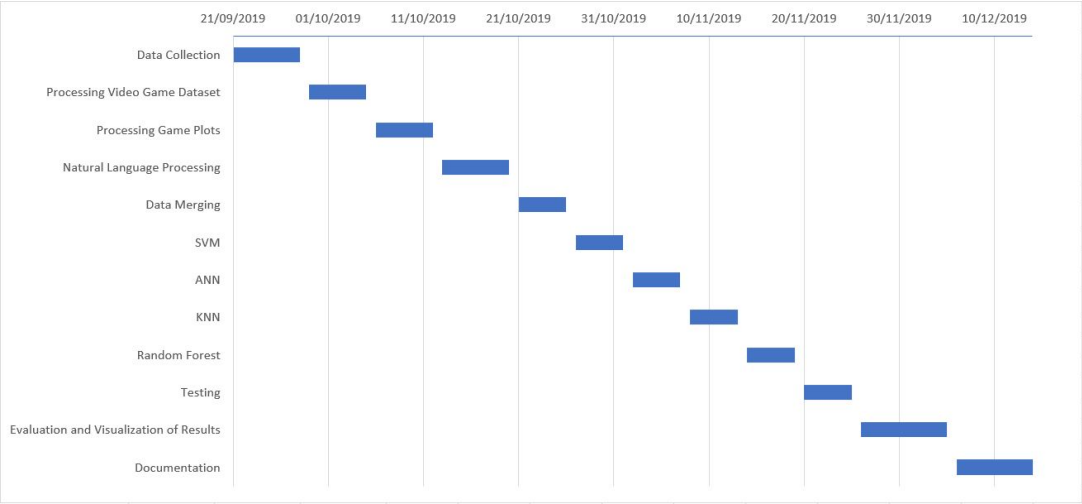

Figure 2: Gantt Chart

5 Requirements & Specification

5.1 Hardware

The research experiments will be performed on system manufactured by Dell with following hardware specifications. CPU: Intel Core i7-8750H, RAM: 16 GB, Graphics: Nvidia GeForce GTX 1610, Storage: 1TB HDD/ 256 GB SSD.

5.2 Software

The proposed methodology shall be implemented using python programming language on Jupyter Notebook development environment. Hence, as a prerequisite latest version of Python and development environment for Jupyter Notebook must be installed and configured for the research experiments.

5.3 Ethical

The source of Data used for the proposed research is Kaggle and Wikipedia, which are open source platforms. So, this avoids any conflict of interest regarding ethical violations as per GDPR regulations.

6 Conclusion

This research proposal presents a machine leaning approach to estimate video game success before initial release using sentiments derived from game plots with other descriptive game features. This information can be used to accurately plan a video game launch and minimize involved business risk for game development organizations. The extracted data of game plots will be processed using word2vec and NLP techniques in order to find the hidden sentiments and is combined with the descriptive dataset. The large dataset shall be processed using label encoder and word2vec before passing it to the proposed classifiers and predict the level of video game success. The evaluation will be based on comparison of results using accuracy, precision and f-score, in case when only descriptive features are used and the combination of game plot sentiments along with the descriptive features.

References

Ahmad, N. B., Barakji, S. A. R., Shahada, T. M. A. and Anabtawi, Z. A. (2017). How to launch a successful video game: A framework, *Entertainment computing* **23**: 1–11.
URL: *https://www.sciencedirect.com/science/article/pii/S1875952117300861*

Alshari, E. M., Azman, A., Doraisamy, S., Mustapha, N. and Alkeshr, M. (2017). Improvement of sentiment analysis based on clustering of word2vec features, *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*, IEEE, pp. 123–126.
URL: *https://ieeexplore.ieee.org/document/8049699*

Amrieh, E. A., Hamtini, T. and Aljarah, I. (2016). Mining educational data to predict students academic performance using ensemble methods, *International Journal of Database Theory and Application* **9**(8): 119–136.
URL: *http://evo-ml.com/ibrahim/publications/13.pdf*

Arafat, M. Y., Hoque, S. and Farid, D. M. (2017). Cluster-based under-sampling with random forest for multi-class imbalanced classification, *2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, IEEE, pp. 1–6.
URL: *https://ieeexplore.ieee.org/document/8294105?arnumber=8294105*

Bailey, E. and Miyata, K. (2019). Improving video game project scope decisions with data: An analysis of achievements and game completion rates, *Entertainment Computing* **31**: 100299.
URL: *https://www.sciencedirect.com/science/article/pii/S1875952118300181*

Bertens, P., Guitart, A., Chen, P. P. and Periáñez, Á. (2018). A machine-learning item recommendation system for video games, *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, IEEE, pp. 1–4.
URL: *https://ieeexplore.ieee.org/abstract/document/8490456*

Bhannarai, R. and Doungsa-ard, C. (2016). Agile person identification through personality test and knn classification technique, *2016 2nd International Conference on Science in Information Technology (ICSITech)*, IEEE, pp. 215–219.
URL: *https://ieeexplore.ieee.org/document/7852636?arnumber=7852636*

Bormann, D. and Greitemeyer, T. (2015). Immersed in virtual worlds and minds: effects of in-game storytelling on immersion, need satisfaction, and affective theory of mind, *Social Psychological and Personality Science* **6**(6): 646–652.
URL: *https://journals.sagepub.com/doi/abs/10.1177/1948550615578177*

Chaudhary, A., Kolhe, S. and Kamal, R. (2016). An improved random forest classifier for multi-class classification, *Information Processing in Agriculture* **3**(4): 215–222.
URL: *https://www.sciencedirect.com/science/article/pii/S2214317316300099*

Dhaoui, C., Webster, C. M. and Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning, *Journal of Consumer Marketing* **34**(6): 480–488.
URL: *https://www.emerald.com/insight/content/doi/10.1108/JCM-03-2017-2141/full/pdf?title=social-media-sentiment-analysis-lexicon-versus-machine-learning*

Fuad, N., Taib, M., Jailani, R. and Marwan, M. (2019). Ann classification for the analysis of 3d eeg data in bbi, *Advances in Computing and Intelligent System* **1**(1).
URL: *https://fazpublishing.com/acis/index.php/acis/article/view/1*

Ge, X., Jin, X. and Xu, Y. (2018). Research on sentiment analysis of multiple classifiers based on word2vec, *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Vol. 2, IEEE, pp. 230–234.
URL: *https://ieeexplore.ieee.org/document/8530220*

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G. and Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings, *Expert Systems with Applications* **69**: 214–224.
URL: *https://www.sciencedirect.com/science/article/pii/S095741741630584X*

Islam, M. R., Kamal, A. R. M., Sultana, N., Islam, R., Moni, M. A. et al. (2018). Detecting depression using k-nearest neighbors (knn) classification technique, *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, IEEE, pp. 1–4.
URL: *https://ieeexplore.ieee.org/document/8465641?arnumber=8465641*

Jaiswal, A. K. and Banka, H. (2017). Local pattern transformation based feature extraction techniques for classification of epileptic eeg signals, *Biomedical Signal Processing and Control* **34**: 81–92.
URL: *https://www.sciencedirect.com/science/article/pii/S174680941730006X*

Klumpp, P., Fritsch, J. and Noeth, E. (2018). Ann-based alzheimer's disease classification from bag of words, *Speech Communication; 13th ITG-Symposium*, VDE, pp. 1–4.
URL: *https://ieeexplore.ieee.org/abstract/document/8578051*

Krithika, N. and Selvarani, A. G. (2017). An individual grape leaf disease identification using leaf skeletons and knn classification, *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, IEEE, pp. 1–5.
URL: *https://ieeexplore.ieee.org/document/8275951?arnumber=8275951*

Liu, Y., Bi, J.-W. and Fan, Z.-P. (2017). Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms, *Expert Systems with Applications* **80**: 323–339.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417417301951*

More, A. and Rana, D. P. (2017). Review of random forest classification techniques to resolve data imbalance, *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, IEEE, pp. 72–78.
URL: *https://ieeexplore.ieee.org/document/8122151?arnumber=8122151*

Sarma, M. S., Srinivas, Y., Abhiram, M., Prasanthi, M. S. and Ramya, M. L. (2017). Knn file classification for securing cloud infrastructure, *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, IEEE, pp. 5–9.
URL: *https://ieeexplore.ieee.org/document/8256548?arnumber=8256548*

Thaseen, I. S. and Kumar, C. A. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class svm, *Journal of King Saud University-Computer and Information Sciences* **29**(4): 462–472.
URL: *https://www.sciencedirect.com/science/article/pii/S1319157816300076*

Trněnỳ, M. (2017). Machine learning for predicting success of video games.
URL: *https://is.muni.cz/th/k2c5b/diploma*$_t$*hesis*$_t$*rneny.pdf*

Wang, Y., Xiong, R., Wang, J. and Zhang, J. (2015). Multi-class assembly parts recognition using composite feature and random forest for robot programming by demonstration, *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, pp. 698–703.
URL: *https://ieeexplore.ieee.org/document/7418850?arnumber=7418850*

Xiao, L., Wang, G. and Zuo, Y. (2018). Research on patent text classification based on word2vec and lstm, *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, Vol. 1, IEEE, pp. 71–74.
URL: *https://ieeexplore.ieee.org/abstract/document/8695493*

Yang, S. and Xia, Z. (2016). A convolutional neural network method for chinese document sentiment analyzing, *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, IEEE, pp. 308–312.
URL: *https://ieeexplore.ieee.org/document/7924714*

Zhang, S., Li, X., Zong, M., Zhu, X. and Wang, R. (2017). Efficient knn classification with different numbers of nearest neighbors, *IEEE transactions on neural networks and learning systems* **29**(5): 1774–1785.
URL: *https://ieeexplore.ieee.org/document/7898482?arnumber=7898482*

Zhang, W., Zhang, H., Liu, J., Li, K., Yang, D. and Tian, H. (2017). Weather prediction with multiclass support vector machines in the fault detection of photovoltaic system, *IEEE/CAA Journal of Automatica Sinica* **4**(3): 520–525.
URL: *https://ieeexplore.ieee.org/abstract/document/7974898*

Zhang, X., Chen, H., Zhao, Y., Ma, Z., Xu, Y., Huang, H., Yin, H. and Wu, D. O. (2019). -improving cloud gaming experience through mobile edge computing, *IEEE Wireless Communications* .
URL: *https://ieeexplore.ieee.org/abstract/document/8685768*