Estimating Video Game Ratings

- using Machine Learning

MSc Research Project

Data Analytics

Aashish Prasad

Student ID: x17170826

School of Computing

National College of Ireland

Supervisor: Paul Laird

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Aashish Prasad
Student ID:	x17170826
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Paul Laird
Submission Due Date:	12/12/2019
Project Title:	Estimating Video Game Ratings using Machine Learning
Word Count:	4554
Page Count:	17

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:	28th January 2020

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):	

Contents

Estimating Video Game Ratings using Machine Learning

Aashish Prasad

x17170826

Abstract

Playing video games have emerged as a popular entertainment activity in our society, without any barricade on the age group. It can be seen as an outcome of expanding digital platforms developed in the recent decade with each competing to offer better virtual experience than the other. This competition often puts video game developers to push their limits towards achieving the best user experience over various platforms by exploring innovative ideas and technology in building a state of art video game. The result of which directly impacts budget with high production costs for a video game, that is directly proportional to higher business risks. This research focuses on the minimization of this business risk using the machine learning approach. In this research we explore the sentiments from video game plots, which reduces the churn rate by keeping the players engaged; along with various descriptive features to predict video game ratings. The research involves the creation of a video game dataset on which regression-based prediction models, Random Forest, Support Vector Machine and Extreme Gradient Boosting has been implemented. Also, feature importance using the Boruta package has been explored to improve model performance which was evaluated using Root Mean Square Error and R-Squared as an evaluation metric.

1 Introduction

Video Gaming has achieved significant popularity among different age groups with the widespread of digital technology in the recent past. With the accessibility of devices at low cost; modern games are currently running on the fingertips of mass individuals. As a result, video games have paved its way from stationary personal computers to portable devices such as smartphones, tablets, gaming consoles, etc. On a global scale by the year 2023, the gaming market is expected to reach a high of 158.33 US Dollars, Zhang et al. (2019).

The scope for this massive market growth has attracted a mix of small and large organizations with names such as Ubisoft, EA Sports and Gameloft into the video game industry. Increasing competition among developers has set new standards within the games in terms of quality products which comes with high development costs. This includes large development teams and quality infrastructure, necessary to meet the set standards. All this makes it difficult to obtain a significant profit margin despite rapid market growth, Bailey and Miyata (2019). With experience in the gaming industry, developers have found certain human behavioural patterns that can contribute to maximizing investment returns, Ahmad et al. (2017). The success of a video game could

1

sometimes be a crucial factor for the survival of the organization when such huge investments are involved in game development. The machine learning approach could be used as a trustworthy tool to maximise the chance of video game success by minimizing the business risk involved.

When considering multiple games from various publishers, it is difficult to estimate the success of a video game based on monetary gain. Sales of ten thousand copies for a game with low production cost could be a huge success whereas the same count for a game with the high cost involved could be a heavy loss for the organization. Success based on sales figures could occur at different point of time for various games. A game with a previous sequel may achieve popularity much faster in the market than a game with a fresh title. This does not change the fact that the new title could achieve similar popularity during the later part of the time-frame.

Considering the above factors, this project estimates video game success by the overall game rating, which unlike sales figures remains constant after a certain period of time from the date of the initial release. Also, a higher rating is an indication of a higher sales figure. It is well known that descriptive features such as game genre and publisher affect video game success. Apart from these, social media reviews/responses and influencers/bloggers do also contribute towards video game success or failure. But to make the above factors work, a game need to grab a certain level of popularity for social media attention, Trněnỳ (2017). Another important factor influencing games are the game-story that which has never been considered, which in the long-run has a crucial role in a game's success. In-game storytelling has been seen to influence users positively in research by Bormann and Greitemeyer (2015). A good story has a positive impact on keeping up players interests, thus decreasing the chances of player churn. A player always chooses content quality over content quantity in a video game, Bailey and Miyata (2019). It is similar to the success of a movie at the box office. Although a movie may have a very good cast, it may fail if the intended audience does not find the story much appealing to their interests. Such criteria do apply for a video games success which can be seen generally in the long run as video game ratings. Hence, analysis of game story has been included along with the descriptive game features in this research project.

This research has been accomplished to find the success level of video games based on descriptive features and the game plot. For this purpose, three machine learning models that are Random Forest, Support Vector Machine and Extreme Gradient Boosting have been implemented to predict game ratings. Natural language processing techniques have been used to derive sentiment scores from video game plots. In addition, the data encoding has been used to transform categorical data and followed by feature selection using Boruta. Evaluation metrics used for the models are RMSE and R-Squared.
The research has undertaken the following question.

How video game plots can be used to improve prediction accuracy while predicting game ratings using machine learning algorithms along with descriptive game features?

2 Related Work

2.1 Study on Video Games

Various researches have used sales figures as an estimator of video game success. Using this criterion for deciding success is difficult when multiple publishers with different

monetary expectation are involved. In research by Trněnỳ (2017), an alternate metric to determine game success has been used in order to resolve this challenge. The author has used the number of players after two months of initial release to estimate success. The data from steam charts and steam spy was collected for the propose of research which included descriptive details such as game name, price, developer, publisher, genre, etc. If analysed, the number of players may be a better estimator than sales figures, but it is also notable that it does not ensure continuity of players in a game. Also, a game may have a higher number of players as a result of popular production house during initial months of release but subsequently, lose this count in the later months. Similar research involving steam data by Bailey and Miyata (2019) on game completion rate showed a correlation between completion rate with price, genre, ratings and least with release data. While the regression analysis resulted in a significant correlation with ratings and genre.

2.2 Study based on Predicting Ratings

Some video games offer purchase option within the game as an opportunity to enhance player progress within a short time-frame. Research by Bertens et al. (2018) focuses on improving such in-game purchase with the help of a recommendation system that works on the principle of predicting ratings for an item by the player. The rating prediction was done for eight different in-game items and the higher values were considered for recommendation. In research by Zhao et al. (2016) user-service rating was predicted by analysing a user's social rating behaviour on Yelp and movie dataset. The concept of user's rating schedule has been proposed to determine user rating behaviour. Four factors of user interest as item topic, similarity, interpersonal rating behaviour and rating diffusion has been fused into the matrix factorisation framework for rating prediction. The prediction models evaluated using cross-validation technique and Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Similar metrics were used to evaluate rating prediction using sentiment analysis on textual data by Lei et al. (2016). The user's sentiments were calculated for specific product and factors affecting the sentiments such as product popularity has been used to improve the predictive capability of the recommendation system. The textual data was pre-processed to remove stop words and unwanted characters and fed into matrix factorisation based model which was further evaluated using RMSE value.

Prediction of movie rating before the initial release has been addressed in research by Ning et al. (2018). A regression model has been proposed based on Convolutional Neural Network for predictive analysis using pre-release descriptive information such as genes, director, cast and plot information. As part of pre-processing, the author has used the parameter tuning technique to determine parameters that were optimal for the designed models. In another research-based involving movie dataset a prediction model based on XGBoost was implemented to predict user rating given to specific movie title, Viard and Fournier-S'niehotta (2019). For the evaluation of results, RMSE and MAE metrics were used. Similarly for predicting travel time for taxi XGBoost regression model was used by Kankanamge et al. (2019). In both the researches, the XGBoost showed good accuracy when used for regression.

2.3 Analysing Textual Data with Natural Language Processing

Sentiment Analysis on textual data is generally observed in researches where opinion or polarity towards a specific topic of interest is concerned. When considering video game plots, the flow of the story has an influencing effect on the players, Bormann and Greitemeyer (2015). Lexicon based sentiment analysis on textual data, when combined with pre-processing, has shown improvement in past research. When NRC sentiments were applied on a Persian language data after removal of punctuation, special characters and removal of non-Persian words, there was a significant improvement in the sentiment scores. Basiri and Kabiri (2017). In Bushi and Zaïane (2019) research, for the purpose of sentiment extraction from the web documents various pre-processing steps were followed. At first, the unnecessary elements in the Document Object Model (DOM) were eliminated. Followed by tokenization, removal of stopwords and punctuation were implemented. Similarly, tokenization and stopword removal technique along with stemming was used for processing Arabic texts while finding the hidden sentiments by Kanan et al. (2019) and Marie-Sainte et al. (2018).

There are various techniques for feature selection used in previous researches. Vector representation of words is one of the popular method used for this purpose where words are represented in binary format[1]. As discussed in the previous section Ning et al. (2018) performed research on predicting movie ratings using descriptive information. For which one-hot encoding was used to transform data. The categorical data was set as '1' for the specific column values where it actually resided and the rest were set to '0'.

The contribution of various variables may differ in predicting the dependent variable. Such selection methods for variable selection based on the random forest selection process is compared by Speiser et al. (2019). Boruta package provided in R language offers high computational efficiency and is preferable for datasets with a large number of predictors. To improve the conservation of energy consumption Boruto package was used, Nagpal et al. (2017). It is a wrapper method for feature selection that selects and plots the graphs representing variable importance.

3 Methodology

A research-based in data mining involves various processes and decision-making tasks. Hence, a well-structured set of steps or methodology is required to achieve the research objectives in an organised manner following the documentation of the research accomplished. The methodology chosen for the researcher is Cross Industry Standard for Data Mining (CRISP-DM), Wirth and Hipp (2000).

3.1 Business Understanding

Among various entertainment industries, the video game industry has seen a very high growth rate in terms of popularity along with the high rate of returns on investments. With a number of competitors within the market, the huge growth in the recent decade has developed high demanding features within the games such as high-end graphics and ultimate user experience. Meeting such demands increase the production cost with the expansion of resources. These resources may include a large development team involved

---

[1]https://towardsdatascience.com/introduction-to-natural-language-processing-for-noobs-8f47d0a27fcc

and expensive licensed tools and technology. Such huge investments are directly proportional to high business risks. Here in this research, to minimize the risk involved, estimation of video game success based on game ratings has been undertaken.

3.2 Data Understanding

The data for this research has been taken from two sources. The first source of data is `https://ie.ign.com/` which contains descriptive information about video games. This includes game name, month, year, publisher, developer, age rating and game rating. The total number of rows fetched from the website are 8,529.

The second source of data used in this research is from `https://www.wikipedia.org/`. The game plots were extracted from this source for those games that were initially extracted from the first source. This textual data was transformed into numeric values as the sentiment score for further analysis. The resulting columns for the dataset created are as follows.

game_name	developer	publisher	genre	platform	release_date	age_rating	game_rating	game_text
A Boy and His Blob	WayForward Techn...	Majesco	Platformer	Wii, Xbox One, Pla...	released	"E"	7.6	The planet Blobolonia
A Boy and His Blob: ...	Absolute Entertain...	Jaleco	Puzzle	NES, Wii	June 1, 1989	"K-A"	4.5	Since the original relea
A Bug's Life	Traveller's Tales	Sony Computer Ente...	Platformer, Ac...	PlayStation, Game...	November 18, 1998	"E"	4.0	Ant Island is a colony
A Collection of Intell...	Activision	Activision	Action	PlayStation	October 30, 1999	"E"	4.0	Gameplay focuses on
A Game of Thrones: ...	Cyanide	Focus Home Interac...	Strategy	PC	September 29, 2011	"T"	6.0	Death Wish is a challer
A Hat in Time	Gears for Breakfast	Gears for Breakfast	Platformer	PC, Macintosh, Xb...	December 5, 2017	"T"	8.0	In the game, the playe
A Kingdom for Kefli...	NinjaBee	NinjaBee	Simulation	PC, Xbox 360	TBA	"E"	8.2	The Arts & Letters edit
Brain Age Express: A...	Nintendo	Nintendo	Educational	Nintendo DSi	August 10, 2009	"E"	7.9	The Arts & Letters edit
Brain Age Express: ...	Nintendo	Nintendo	Educational	Nintendo DSi	April 5, 2009	"E"	7.9	A new Brain Age title f
Brain Age Express: S...	Nintendo	Nintendo	Educational	Nintendo DSi	August 17, 2009	"E"	8.0	In the game, the playe
A Plague Tale: Innoc...	Asobo Studio	Focus Home Interac...	Action	PC, Xbox One, Pla...	May 14, 2019	"M"	7.0	Lost in Shadow is a 2D
Lost in Shadow	Hudson Soft	Hudson Soft	Platformer	Wii, Wii U	January 4, 2011	"E10+"	8.5	In the year 2055, time t

Figure 1: Video Game Dataset

1. **game_name:** It describes the name of the game and are unique values.

2. **release_date:** It represents the initial release date of the game.

3. **publisher:** It contains the name of the publisher for specific games.

4. **developer:** It contains the name of the developer for specific games.

5. **genre:** It contains various genres for specific games.

6. **platform:** It describes the platforms on which the video game is available.

7. **age_rating:** It contains the age rating as provided by the Entertainment Software Rating Board (ESRB). These are categorical in nature.

8. **game_rating:** It contains the video game ratings, ranging between 1 to 10 provided by `https://ie.ign.com/`. These are continuous variables.

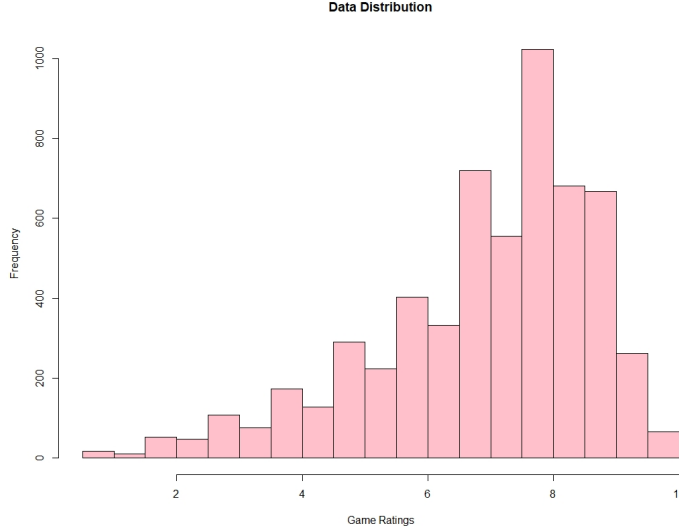9. **game_text:** It contains the game plots scraped from Wikipedia page.

Figure 2: Distribution of Dependent Variable (Game Ratings)

The distribution of the dependent variable, 'game_rating' can bee seen in Figure 2 with most of the rating values between 7.0 to 9.0.

3.3 Data Preparation

It is important to eliminate inconsistency and transform the data into numeric values for the machine learning models. The collected data consists of numeric as well as categorical values.

The 'date of release' for the game was split to form month and year columns. While the value for 'developer', 'publisher','genre' and 'platform' columns was set as 'other' where the count was less than a minimum threshold. This was done to eliminate outliers for these columns. All the columns containing categorical values that are developer, publisher, genre, platforms, month and age ratings were transformed into a binary format using one-hot encoding. Further, all the rows with null values were removed using na.omit() function in R.

Based on the 'game name' column, Wikipedia page URL for each video game was extracted using 'wikipediaapi' library in Python. Further, the game plots were scraped using these url(s). The scraped game plots from were processed using various natural language processing libraries provided under NLTK package in Python. Firstly, any HTML content or punctuation in the texts was removed. Then they were tokenized for the removal of stopwords. The resultant output was further processed to their root form. Using R, NRC sentiment analysis was performed over the processed texts.

Finally, data from both the sources were merged to form a feature-rich video game dataset with 4,615 relevant entries. For further experiments, this dataset was replicated to form three more datasets. Firstly, as 'Dataset_2' with 'year' value greater than or equal to '2000'. Second, as 'Dataset_3'with rows eliminated containing publisher as 'other' and third, as 'Dataset_4' with rows containing publisher as 'other' removed. Each dataset containing features of the previous dataset as mentioned in the above order. The below Table 1 describes the four datasets.

6

Dataset	Number of Rows	Description
Dataset_1	4,615	All rows with valid entries.
Dataset_2	3,840	'Year' greater than equal to 2000.
Dataset_3	2,987	'Year' greater than equal to 2000 and publisher as 'other' removed.
Dataset_4	1,372	'Year' greater than equal to 2000, publisher and developer as 'other' removed.

Table 1: Dataset Description

3.4 Modeling

3.4.1 Random Forest

Random Forest is a combination of multiple decision trees that follows the working process of an ensemble model. The model uses a majority vote principle to decide the final result while performing prediction. The use of a combination of multiple trees is an advantage over the decision tree. It is widely used for regression problems and the use of bagging method while training the data makes it more efficient.

3.4.2 Support Vector Machine

Support Vector Machine (SVM) is widely known as a supervised machine learning algorithm that works on the principle of a hyperplane. It is generally seen to be used to solve classification problem using the kernel trick, but can also handle the regression problem. Also, the model is suitable to work well with non-linear data.

3.4.3 Extreme Gradient Boosting

Extreme Gradient Boosting, popularly known as XGBoost is an ensemble-based algorithm that works on a gradient boosting framework. An optimized gradient boosting model with data pruning and parallel processing capability. It can also handle missing values and is good at dealing with the problem of overfitting. It is known for its high computational ability and is less time-consuming in comparison to conventional tree-based algorithms such as decision tree and random forest.

4 Design Specification

As shown in Figure 3 the research begins with the data collection using web scraping technology followed by further processing, modeling and evaluation of results.
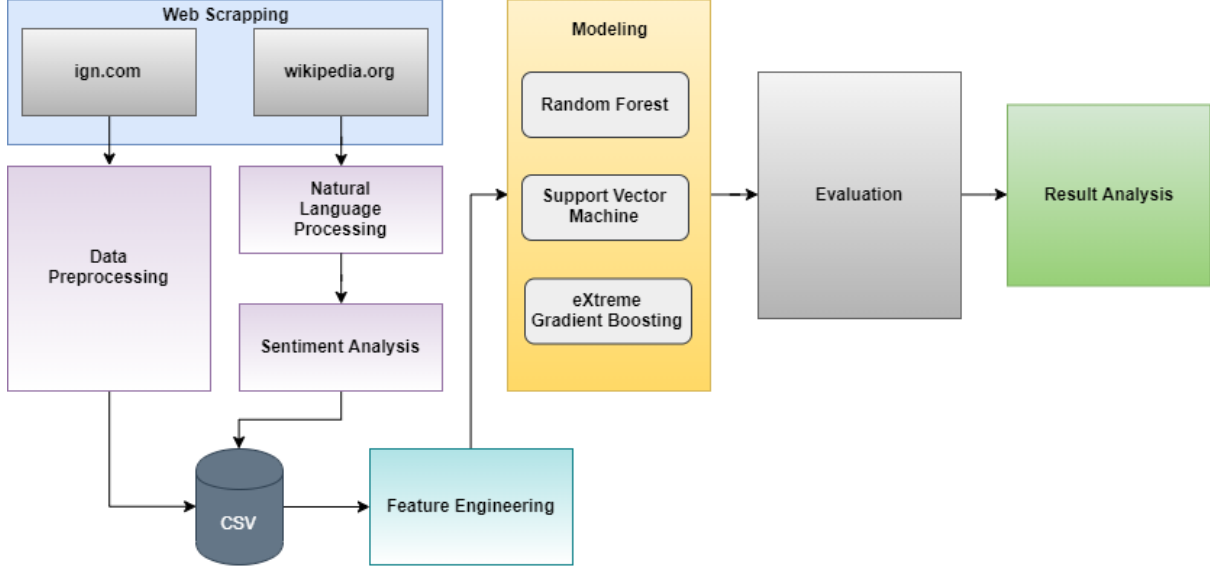

Figure 3: Research Architecture

5 Implementation

In this section, the steps followed in order to achieve the research objective as discussed in section 1 has been briefly elaborated. Firstly, a video game dataset was created using web scraping technique using the selector gadget tool. Following which the pre-processing tasks were undertaken for the machine learning models. Lastly, the prediction models were implemented.

5.1 Data Collection & Pre-Processing

As discussed in section 3, the data for the research has been collected from two sources to form a large dataset. For this purpose, the 'html_nodes' function for web scraping found in the 'rvest' library in the R programming language. The video game data was extracted from `https://ie.ign.com/` consisting of 8,529 rows into eight columns as game name, release_date, developer, publisher, platform, genre, age rating and game rating.

Based on the column 'game name', using Wikipedia-api available in python, video page URL for each video game was extracted. For the games that do not have any Wikipedia page, the entire rows were omitted later in R. Each URL was used for web scrapping the 'game plot' section on for the associated video game.

Further in this stage, the columns within the dataset were manipulated for the machine learning models. Firstly, the column 'date of release' was split into three different columns as date, month and year using 'separate' function while only the month and year columns were retained into the dataset.

5.1.1 Data Imputation

The missing data for the columns developer and publisher were imputed with the developer and publisher information available on Wikipedia page for the respective video game. To accomplish this web scrapping was done using the Wikipedia page URLs that was generated in the earlier stage of data collection. The columns month, year and age ratings were not imputed as the page structure for the Wikipedia differed for these columns.

### 5.1.2 Feature Selection and Transformation

1. **Removal of special symbols:** All the special symbols like quotes and commas were removed from the game name and age ratings column in R.

2. **Tokenization of Words:** The sentences in the text were broken down into words for further processing which involved stopwords removal and stemming.

3. **Removal of Stopwords:** Game plots were processed to remove stopwords fro the texts such as, 'the', 'is' , 'an' using NLTK package available in Python.

4. **Stemming the words:** Further, the words were brought down to their root form such as 'fighting' to 'fight'. This was achieved using the "nltk.stem' provide in Python.

5. **Sentiment Analysis:** The textual data or the game plots extracted from wikipedia needed to be processed into the numeric format for further analysis. This was achieved by applying NRC sentiment analysis using 'syuzhet' library in R. This resulted in 10 columns as anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative and positive.

6. **One-Hot Encoding:** The dataset consists of categorical values or repeated values for the columns developer, publisher, month, age ratings, platforms and genes. These categorical values were converted into the binary format as '0' and '1' using one-hot encoding technique. The one-hot method provided in 'mltools' R library was used for this purpose. Two new columns were added as 'otherPlatform' and 'otherGenre' indicating the presence of specific platform and genre with a total count less than 50 in their respective column. At the end of this step, the total number of columns increased to 239 in the dataset.

7. **Removal of Irrelevant rows and Missing Data:** The rows that contained value as '0' for all publisher and genre columns were removed. Also, the rows containing missing values or NA(s) were omitted in R. At the end of this step, the total number of relevant rows were 4,950.

8. **Feature Importance with Boruta:** The Boruta Package in R was used to determine the important variables and reduced the number of additional variables to improve predictability for the models.
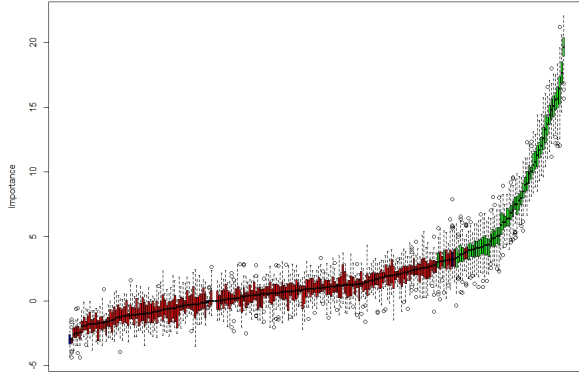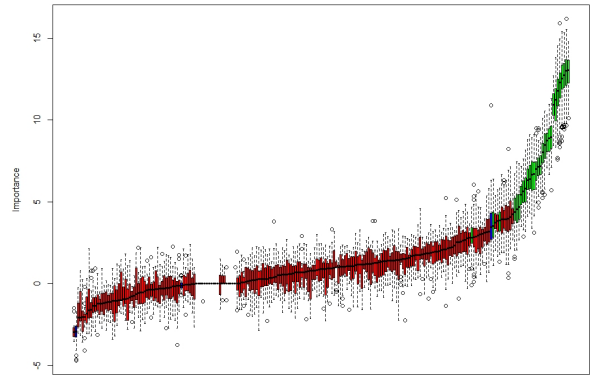
(a) Boruta Plot for Dataset 1
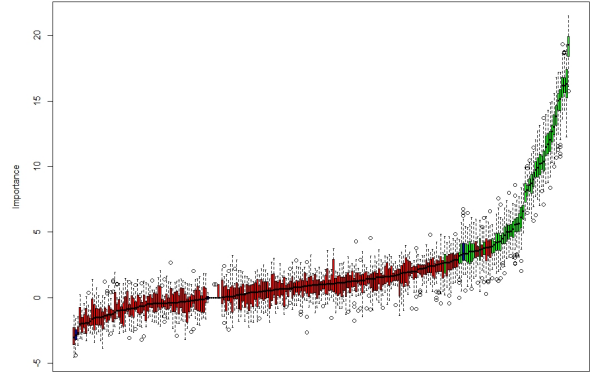
(b) Boruta Plot for Dataset 2
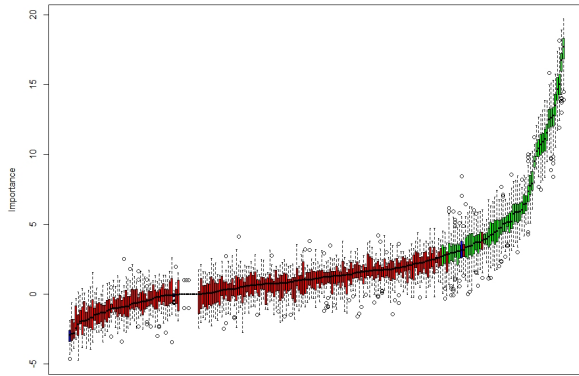
(c) Boruta Plot for Dataset 3

(d) Boruta Plot for Dataset 4

Figure 4: Boruta Output

The figure 4, shows the plots for Boruta feature importance. For all the datasets most of the columns with less importance were rejected. For Dataset 1 represented by Plot 4a 54 columns were selected, where as for Dataset 2 represented by Plot 4b, 51 columns were marked as important. Similarly for Dataset 3, Plot 4c, 55 columns and for Dataset 4, Plot 4d, 28 columns were confirmed as important.

5.2   Modeling

As observed in Section 3.2, the distribution of data is left-skewed, Figure 2. Hence, for predicting game ratings, three non-linear regression models have been implemented. Each of the four datasets was split into train and test set. The ratio for which is 80% and 20% respectively.

The first is Random Forest(RF) which was implemented using the 'randomForest' package in R. The value of mtry was set as '6' to maximized the model performance.

The second model is Support Vector Machine(SVM) found in the library 'e1071' of R. The model is trained based on 'radial' kernel with type as 'eps-regression'. Other kernels such as polynomial and linear were also tested but the highest R-Squared value was achieved using the radial kernel.

The third model is XGBoost found in 'xgboost' R package. The model has been developed with hyperparameters tuning.

6   Evaluation

In this research, for evaluating the results, Root Mean Square Error (RMSE) and R-Square metric has been used. RMSE describes the best fit of the data over the regression line. It is the standard deviation of the residual or the wrongly predicted values from the regression line.The formula to calculate RMSE value is given below.

$$\text{RMSE} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i-f_i}{\sigma_i}\right)^2}$$

Whereas the R-squared represents how much of the proportion of variance for the dependent variable is explained by the independent variables.

Each of the three implemented models, Random Forest, Support Vector Machine and XGBoost were trained and tested over the four datasets described in table 1. The experiment conducted involved training the dataset over Boruta Package results, that was used to identify important columns with the dataset.

6.1   Experiment : Dataset 1

Model	Without Feature Selection		With Feature Selection	
	RMSE	R-Squared	RMSE	R-Squared
RF	1.47	0.18	1.47	0.16
SVM	1.50	0.15	1.49	0.16
XGBoost	1.47	0.16	1.5	0.12

Table 2: Results for Dataset 1

In Table 2, the R-Squared value for both the Random Forest and XGBoot can be seen to decrease after selecting the important columns chosen by the Boruta algorithm.

6.2   Experiment : Dataset 2

Model	Without Feature Selection		With Feature Selection	
	RMSE	R-Squared	RMSE	R-Squared
RF	1.46	0.20	1.51	0.11
SVM	1.56	0.10	1.60	0.05
XGBoost	1.49	0.14	1.57	0.04

Table 3: Results for Dataset 2

Similar result can be observed in the above Table 3. Among the three models Random Forest can be seen with highest R-Squared value with 0.20.

6.3   Experiment : Dataset 3

Model	Without Feature Selection		With Feature Selection	
	RMSE	R-Squared	RMSE	R-Squared
RF	1.39	0.17	1.36	0.20
SVM	1.44	0.11	1.34	0.21
XGBoost	1.34	0.21	1.41	0.13

Table 4: Results for Dataset 3

For Dataset 3, the feature selection resulted in slight increase in the model performance for Random Forest and SVM. The R-Squared value can be seen to increase by 0.10. While XGBoost before feature selection showed similar results as SVM after feature selection.

6.4   Experiment : Dataset 4

Model	Without Feature Selection		With Feature Selection	
	RMSE	R-Squared	RMSE	R-Squared
RF	1.27	0.17	1.30	0.13
SVM	1.40	0.06	1.32	0.13
XGBoost	1.32	0.09	1.49	0.14

Table 5: Results for Dataset 4

With more filtration on data in Dataset 4, the overall model performance can be seen decreasing further although the RMSE values showed improvement when compared to previous datasets. Random Forest without feature selection showed the best RSME score across all the experiment cases.

6.5   Discussion

(a) Actual vs Predicted: RF (Dataset 1)   (b) Actual vs Predicted: RF (Dataset 2)

(c) Actual vs Predicted: XGBoost (Dataset 3)   (d) Actual vs Predicted: RF (Dataset 4)
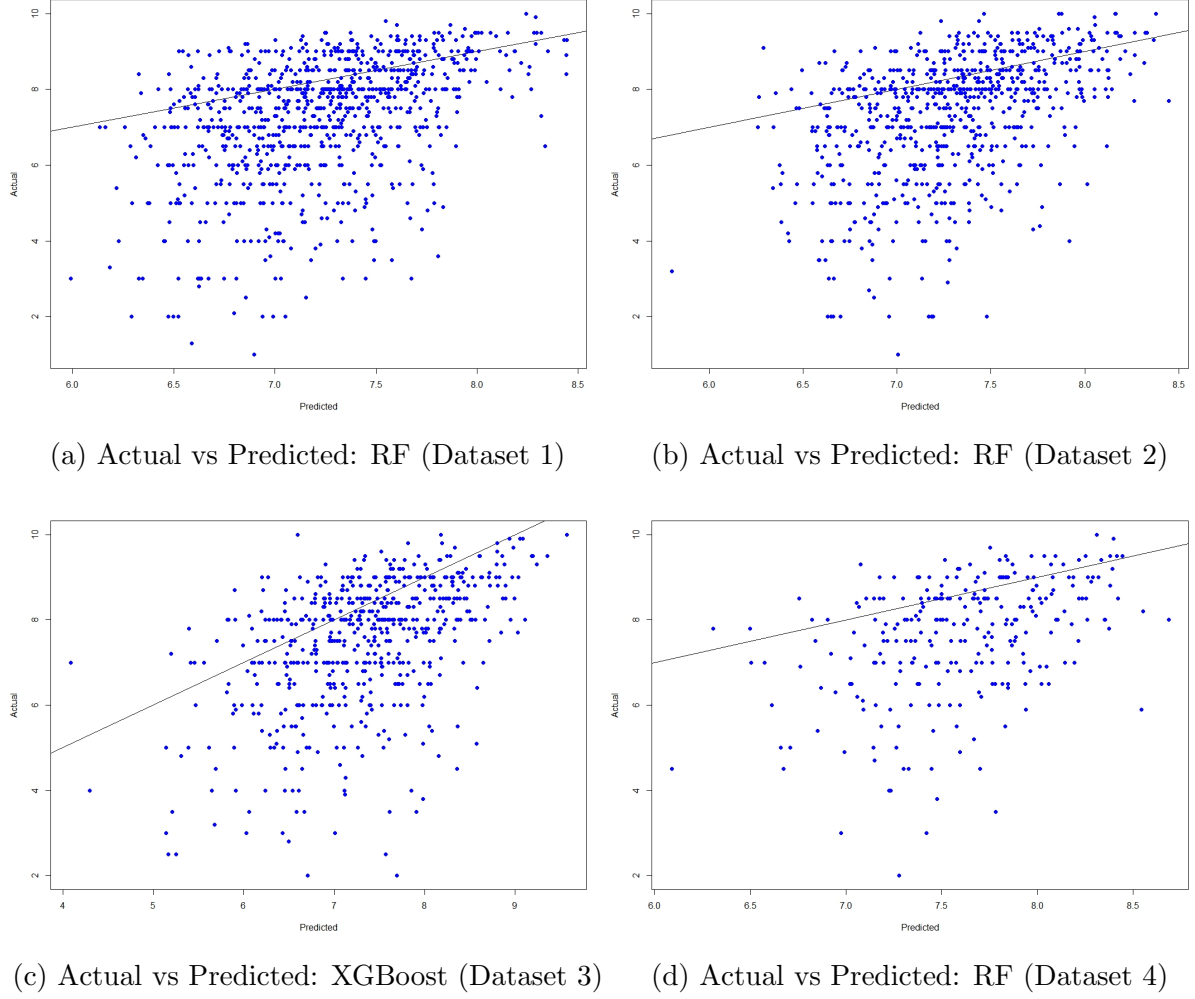
Figure 5: Actual vs Predicted

The research has addressed the prediction of video game ratings using descriptive features along with game plots. The dataset for the research has been built from scratch using web scraping technique. In order to make the data into a machine-readable format, various pre-processing tasks including natural language processing has been implemented. Further, to improve accuracy, the Boruta package is used for feature importance. The experiment result showed this approach of feature selection to be unsuccessful in most of the dataset cases. The accuracy of the model decreased after feature selection which is similar to what was observed in research by Masduki and Ramli (2016). This is a special case, which shows that the variables rejected by Boruta, had a certain level of dependency over the variables marked as important. In another research, model accuracy decreased when unnecessary features were selected, Ahmadi et al. (2016). It is also important to note that for some cases like Table 4 in Section 6 there was a significant improvement in the model performance which can be observed in RMSE and R-Squared values with Boruta feature selection for SVM. Although, there is an inconsistency observed among RMSE and R-Squared results for the three models across the different versions of the dataset.

13

The actual data points compared to predicted values can be observed in Figure 5. Although the results are not good enough, all the four plots can be seen to have some degree of positive relationships. The highest accuracy achieved is 21% using XGBoost (without feature importance) and Random Forest (with feature importance) for Dataset 3 as observed in Experiment 6.3.

7   Conclusion and Future Work

The research tried to predict video game ratings using descriptive data and game plot using machine learning models. For the prediction algorithm, the categorical data has been processed using one-hot encoding while the textual data or the game plots were processed using natural language processing with NRC sentiment analysis. The dataset created was used for four different experiment cases, each case was tested with and without feature importance selection of independent variables. Regression models of Random Forest, Support Vector Machine and Extreme Gradient Boosting were evaluated using RMSE and R-Squared values. Using this approach 21% of the game ratings were correctly predicted. Unfortunately, the models did not achieve high accuracy rate as expected. This suggests that game plots do not contribute significantly towards the prediction of game ratings.

There were few limitations observed during the research which has contributed to the low predictability of the chosen models. Firstly the data collected from `https://ie.ign.com/` consisted of 8,529 rows, from which only those game records were considered which has an existing Wikipedia page. Also, due to the difference in page structure across Wikipedia, many missing values could not be successfully imputed. Such rows with NA or missing values were later eliminated during data pre-processing stage. As a result, the total number of valid rows for the experiment dropped to 4,615.

For future work, with more availability of time, the research could be expanded using other feature selection processes. Also, more important features in the dataset could be included that affects the video game ratings. As the Wikipedia page structure are not uniform across all the web pages, there is a possibility of additional texts included from other page section, such as 'gameplay' for some games. Further study may include finding solutions to such technical challenges and overcome limitations observed as mentioned above by incorporating video game data from other sources.

8   Acknowledgement

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*

Ahmadi, M., Ulyanov, D., Semenov, S., Trofimov, M. and Giacinto, G. (2016). Novel feature extraction, selection and fusion for effective malware family classification, *Proceedings of the sixth ACM conference on data and application security and privacy*, ACM, pp. 183–194.
URL: *https://dl.acm.org/citation.cfm?id=2857713*

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*

Basiri, M. E. and Kabiri, A. (2017). Translation is not enough: comparing lexicon-based methods for sentiment analysis in persian, *2017 International Symposium on Computer Science and Software Engineering Conference (CSSE)*, IEEE, pp. 36–41.
URL: *https://ieeexplore.ieee.org/document/8320114?arnumber=8320114*

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*

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*

Bushi, S. S. and Zaïane, O. R. (2019). Apnea: Intelligent ad-bidding using sentiment analysis, *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, pp. 76–83.
URL: *https://ieeexplore.ieee.org/document/8909592?arnumber=8909592*

Kanan, T., Sadaqa, O., Aldajeh, A., Alshwabka, H., AlZu'bi, S., Elbes, M., Hawashin, B., Alia, M. A. et al. (2019). A review of natural language processing and machine learning tools used to analyze arabic social media, *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, IEEE, pp. 622–628.
URL: *https://ieeexplore.ieee.org/document/8717369?arnumber=8717369*

Kankanamge, K. D., Witharanage, Y. R., Withanage, C. S., Hansini, M., Lakmal, D. and Thayasivam, U. (2019). Taxi trip travel time prediction with isolated xgboost regression, *2019 Moratuwa Engineering Research Conference (MERCon)*, IEEE, pp. 54–59.
URL: *https://ieeexplore.ieee.org/document/8818915?arnumber=8818915*

Lei, X., Qian, X. and Zhao, G. (2016). Rating prediction based on social sentiment from textual reviews, *IEEE Transactions on Multimedia* **18**(9): 1910–1921.
URL: *https://ieeexplore.ieee.org/abstract/document/7484319*

Marie-Sainte, S. L., Alalyani, N., Alotaibi, S., Ghouzali, S. and Abunadi, I. (2018). Arabic natural language processing and machine learning-based systems, *IEEE Access*

7: 7011–7020.
URL: *https://ieeexplore.ieee.org/document/8594541?arnumber=8594541*

Masduki, B. W. and Ramli, K. (2016). Improving intrusion detection system detection accuracy and reducing learning time by combining selected features selection and parameters optimization, *2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, IEEE, pp. 397–402.
URL: *https://ieeexplore.ieee.org/abstract/document/7893606*

Nagpal, D., Srivastava, R., Mehrotra, D. et al. (2017). Feature selection approach for reducing the power consumption for a greener environment, *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, IEEE, pp. 1–5.
URL: *https://ieeexplore.ieee.org/abstract/document/8343509*

Ning, X., Yac, L., Wang, X., Benatallah, B., Dong, M. and Zhang, S. (2018). Rating prediction via generative convolutional neural networks based regression, *Pattern Recognition Letters* .
URL: *https://www.sciencedirect.com/science/article/pii/S0167865518303325*

Speiser, J. L., Miller, M. E., Tooze, J. and Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling, *Expert Systems with Applications* .
URL: *https://www.sciencedirect.com/science/article/pii/S0957417419303574*

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

Viard, T. and Fournier-S'niehotta, R. (2019). Augmenting content-based rating prediction with link stream features, *Computer Networks* **150**: 127–133.
URL: *https://www.sciencedirect.com/science/article/pii/S1389128618313215?*

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Citeseer, pp. 29–39.
URL: *https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133rep=rep1type=pdf*

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*

Zhao, G., Qian, X. and Xie, X. (2016). User-service rating prediction by exploring social users' rating behaviors, *IEEE transactions on multimedia* **18**(3): 496–506.
URL: *https://ieeexplore.ieee.org/abstract/document/7373661*

Appendices

Technical implementation related to this academic research is accessible on Github, `https://github.com/aashishprasad/Research-Project`

2. Your Use of Content

You acknowledge that the Services contain information, software, articles, reviews, content, photographs, audio and video clips, graphics, links, logos, trademarks, the "look and feel" of our websites, applications and software and other material (collectively, the **"Content"**) that are protected by copyright, trademark and/or other proprietary rights of Ziff Davis or third parties. All Content on the Services is protected by applicable copyright laws. You agree to comply with all copyright laws and any copyright notices, information, or restrictions contained in any Content available on or accessed through the Services. Users of the Services may use the Content only for their personal, noncommercial use. Businesses, organizations or other legal entities may not become Members, and are not permitted to use the Services for any purpose, including but not limited to collecting usernames and/or email addresses of Members by electronic or other means for the purpose of sending unsolicited email and unauthorized framing of, or linking to, the Services or users on the Services. **Please refer to our Linking Policy for more information on linking and use of Content.**

Except as expressly permitted, you may not (i) modify, publish, transmit, reproduce, create derivative works from, distribute, perform, display, adapt, aggregate, sell, transfer or in any way exploit any of the Content, code or Material (defined below), in whole or in part, or (ii) use any robot, spider, site search and/or retrieval application, or other device to scrape, extract, retrieve or index any portion of the Services. Content consisting of downloadable or web-based software may not be reverse engineered unless specifically authorized by the owner of the software's patent and/or copyright. You also agree not to circumvent, disable or otherwise interfere with any security related features of the Services or the Content, including features that prevent or restrict use or copying, or that enforce limitations on use.

You may only post on the Services content owned by you (such as your original statements or video clips), content for which you have received express written permission from the owner and content in the public domain (collectively, the **"Material"**). You assume all risk and responsibility for determining whether any Material is in the public domain. You hereby grant, transfer and assign to Ziff Davis and its affiliates, successors, assigns and licensees (collectively, **"Licensee"**) a fully-paid, royalty-free, irrevocable, perpetual, worldwide right and license to publish, distribute, reproduce, transmit, use, translate, display, perform, modify, revise, create derivative works of and archive the Material, in any form or media now known or hereafter developed (including without limitation in print, magnetic or electronic form), including, but not limited to, as part of the Services or in support of the Services through advertising or marketing, on any number of occasions in any form, and to sublicense third parties (including other users of the Services) to do any of the foregoing with further right of sublicense (the **"License"**), without payment or compensation to you and without seeking further permission from you. You represent and warrant that you are authorized to grant all rights set forth in the preceding sentence and that the exercise by Licensee of Licensee's rights under the License shall not violate any laws, defame or libel any person, invade any person's right of privacy or publicity or otherwise violate, misappropriate or infringe the rights of any person (including but not limited to any copyright or moral right). You agree that you are solely responsible for the Material that you post on the Services or transmit to others and agree that you will not hold Company responsible or liable for any materials or content you access or receive from other users of the Services.

You may download, copy and make any personal, non-commercial use of the Content and use that is expressly permitted by the United States Copyright Act of 1976, as amended (**"Copyright Act"**) and not prohibited by any section of the Copyright Act or by any other applicable law, rule or regulation; provided, however, that you maintain all copyright and other notices contained in such Content; and provided further that you shall not store electronically any significant portion of any Content.

To the extent that you provide us with any suggestions, feedback or other information relating to our business or the Services (including, but not limited to, suggested new products or services or improvements to existing products and services), such information is provided to us on a non-confidential and unrestricted basis, and you hereby grant to the Licensee (as defined above) a non-exclusive, worldwide, perpetual, royalty-free, fully transferrable and sublicensable right and license to reproduce, display, distribute, use and fully exploit such suggestions, feedback and information.

For questions regarding use of Content please email **Brand_Licensing@ziffdavis.com**.

Figure 6: IGN Usage of Content

Figure 7: Wikipedia Terms of Use