IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

Predictive Analysis on Commercial Success of Game

Sheikh Arafat Rahman Shovo[1], ShafiulAlam[2], Shaikh Shariful Habib[3]

Student, CSE Department, City University, Dhaka, Bangladesh[1]

Student, CSE Department, City University, Dhaka, Bangladesh[2]

Assistant Professor, CSE Department, City University, Dhaka, Bangladesh[3]

Abstract: The main goal of the thesis, titled "Predictive Analysis on the Commercial Success of Game," is to predict whether a game will be successful or not after being published by analyzing previous game sale data. Whether a game is considered a hit or not is determined by the number of game copies sold. Previous game sale data were collected from the website Kaggle, Meta critics and Gamespot. The attribute global sale was correlated with all other attributes from the dataset. This enabled the identification of trend lines and determination of the attributes that have the most impact on global sales. Prediction was carried out using Random Forest Classifier and Logistic Regression. One Hot Encoding was used as LR and RFC cannot work with string value.

Keywords: Video Game, Data Science, Data Analysis, Machine learning, Data visualization, LR, RFC, One HOT Encoding

I. INTRODUCTION

Predictive analysis on the commercial success of a game typically involves analyzing various features of the game, as well as external factors such as market trends and user demographics, to predict the likelihood of the game's success in terms of sales, user adoption, and other metrics. Some common features of a game that may be analyzed include Genre, Platform, Rating , Publisher etc. In addition to analyzing these features of the game, predictive analysis may also involve looking at external factors such as market trends, user demographics, and competing games to predict the game's potential success.

II. PROBLEM STATEMENT

The gaming industry is rapidly growing, with thousands of games being released each passing year, but not all of them achieve success.

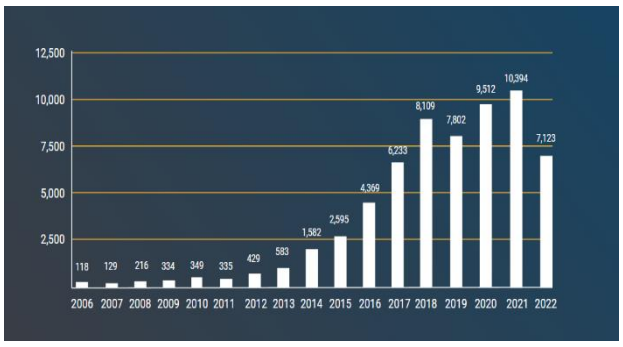

Fig. 1 The number of games released from 2006 to 2022.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

A study by EEDAR (Electronic Entertainment Design and Research) found that of the 7,000 games released on Steam between 2006 and 2016, only around 10% of games generated more than $50,000 in revenue. Another study by Statista found that in 2020, only 12.3% of video games released on the App Store were successful, defined as those that generated more than $1 million in revenue.

III. OBJECTIVE

1. Analyze sales data from previously released video games to identify clusters and determine industry trends.
2. Use the identified clusters to create a trend line for successful games. Develop a prediction model to forecast the probability of a game being successful.

IV. RELATED WORK

Author [1] applied machine learning techniques to predict the success of video games. It uses a dataset of over 6000 video games with various features like genre, platform, release date, and user ratings.The thesis is that it only focuses on a specific dataset and may not be applicable to other datasets. The accuracy of the model may also be affected by factors that are not included in the dataset, such as marketing strategies, game quality, and player preferences.

The authors[2]focused on game and console sales in Europe from March 12, 2005 to December 31, 2011. The authors used data about 2,450 games. Limitation of the thesis is its focus on the European market, which may not be representative of the global market. Another limitation is that the model does not take into account external factors such as economic trends, social changes, or technological advancements, which could impact video game sales. Additionally, the thesis uses data from a specific time period (2008-2011), which may not be applicable to current market trends.

The authors[3] aim of this thesis was to collect gaming forum posts and use this data to predict sales of video games. Limitation of this thesis is its reliance on internet message board discussions as the primary source of data.

The authors[4] proposes a predictive model, called the Game Prophet, to forecast the commercial success of video games using data analytics and machine learning techniques. Model's accuracy and applicability may depend on the quality and quantity of input data, which can be limited or biased. Moreover, the model's predictions may not account for unexpected market changes or subjective factors, such as user preferences and reviews, which can influence game success.IEEE Conference on Computational Intelligence and Games1stands out as a prominent source of papers dealing with applying machinelearning methods in video games development.

The authors[5] processed data about at what times players were playing and what they were doing within the game. While the study provides valuable insights into predicting player retention in sandbox games, there are some limitations to consider. The study was conducted on a specific game, and the results may not be generalizable to other games or datasets. Secondly, the study only considers a 7-day retention period, which may not be sufficient for longer-term predictions.

The authors[6]used very detailed data about player activities in a major title, Destiny. The data span across 17 months andincluded the activities of 10,000 players. They focused on the useof multinomial Hidden Markov Model which returned the highestprecision of 92 % with a relatively low recall of 43 % compared to othermodels.
Limitation of this study is that the dataset is limited to a single game, and the results may not generalize to other games. The study also did not consider external factors that may influence player churn, such as changes in the game's environment or the release of competing games.

The authors[7]describes learning of how players experience one gameand making predictions about theirexperience in another game. The authors used two methods for the task of automatically mapping features between games, referred to as

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

”supervised feature mapping” and ”unsupervised transfer learning”. Both methods produced accuracies above 58 % and 55 %, respectively, achieving 83 %accuracy on one of the subtasks (predicting challenge). These results were comparable with manual mappings created by experts. Limitation of the paper is that it only considers a limited number of games, and it is unclear how well the proposed approach would generalize to a larger and more diverse set of games.

V. METHODOLOGY

Step 1: Two game sale datasets were collected from Kaggle, and one was manually collected from Metacritic and Gamespot.
Step 2: These datasets were merged
Step 3: The most influential factors on global sale were identified by relating other attributes with global sale
Step 4: The rest of columns were dropped
Step 5: Any remaining row with null values were also dropped.
Step 6: String variables were represented as binary vectors by applying One Hot Encoding .
Step 7: The dataset was then split into train and test data
Step 8: Prediction models were created using the LR and RFC algorithms
Step 9: The models were trained using the train data, and their accuracy was tested using the test data.

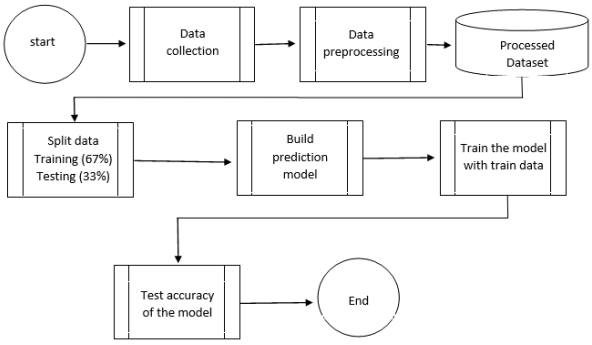

Fig. 2 Workflow diagram

VI. MATERIALS

1. USED Algorithms

1.1 Logistic Regression

The logistic regression model uses the logistic function to model the relationship between the independent variables and the probability of the dependent variable taking on the value of 1. The logistic function is an S-shaped curve that maps any real-valued number to a value between 0 and 1. The logistic function is defined as:

$p(x) = 1 / (1 + e^{-z})$

where p(x) is the probability of the dependent variable taking on the value of 1, e is the mathematical constant approximately equal to 2.71828, and z is the linear combination of the independent variables.

The logistic regression model estimates the parameters of the logistic function using a maximum likelihood estimation method. The resulting model can then be used to predict the probability of the dependent variable taking on the value of 1 for a given set of values of the independent variables. Logistic regression can be extended to handle multiple independent variables and interactions between them. It can also be used for multi-class classification problems, where the dependent variable can take on more than two possible values.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

Logistic regression is widely used in many fields, including medicine, social sciences, marketing, and finance, to predict the likelihood of an event occurring based on a set of predictor variables.

1.2 Random Forest Classifier

Overall, random forest classifier is a powerful machine learning algorithm that can be used for a wide range of classification and regression tasks. Its ability to combine multiple decision trees and introduce randomness makes it a popular choice among data scientists and machine learning practitioners.

1.3 One Hot Encoding

One hot encoding is a technique used in data processing and machine learning to convert categorical data into numerical data that can be used in algorithms.

2.USED Language:

Python

VII. DATASET DESCRIPTION

Video Game Sales with Ratings 2.0[8] contains 17417 game sale data (1.36Mb). PCGame Sales [9] contains 176 game sale data(15Kb). These datasetswere collected from the Kaggle website.
Dataset[10] is collected manually from critics score and game spot contains 108 game sale data (5Kb).

VIII. DATA EXPLORATION AND ANALYSIS

1. Median sales (in millions of units) vs. critic scores

The following three heatmaps show how game sales vary according to critic scores, which are split into six scoring groups. Additionally, each heatmap segments the data further by one of the following features: genre, ESRB rating, publisher.

Under each heatmap, we identify the categories where games sell best. This is done for okay, good, and great games, as defined by games with scores in the 70s, 80s, and 90s, respectively.

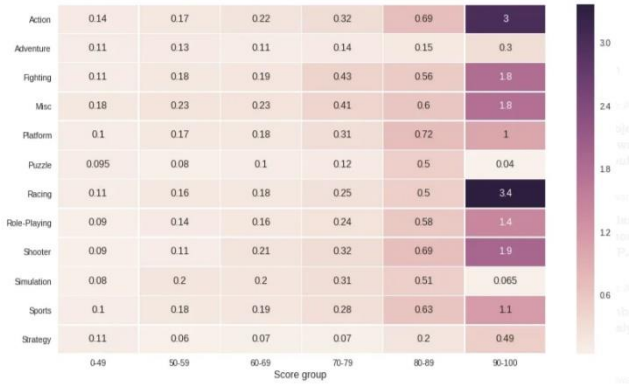

Fig. 3 Genre vs critics score by median sale

- Genres where great games sell best: Racing, Action
- Genres where good games sell best: Action, Shooter
- Genres where okay games sell best: Fighting

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

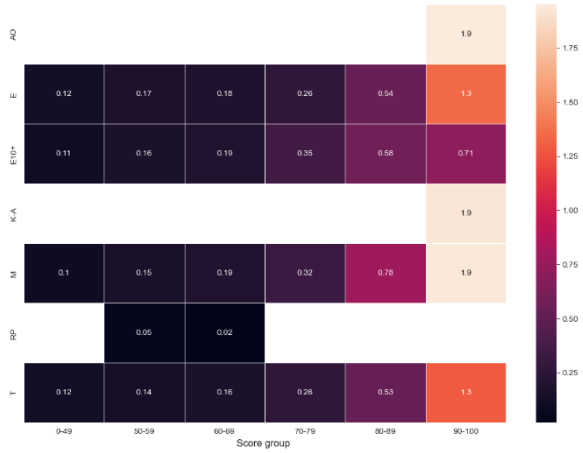

Fig. 4 ESRB_RAting vs critics Score by median sale

- ESRB_RAting where great games sell best: AO, M,K-A
- ESRB_RAting where good games sell best: E,T
- ESRB_RAting where okay games sell best: E10+
-

Fig. 5 Publisher vs critics score by median sale

- Publisher where great games sell best: Warner Bros Interactive Entertainment
- Publisher where good games sell best: Sony Computer Entertainment, Nintendo
- Publisher where okay games sell best: Electronic Art

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

2.Top values in the dataset

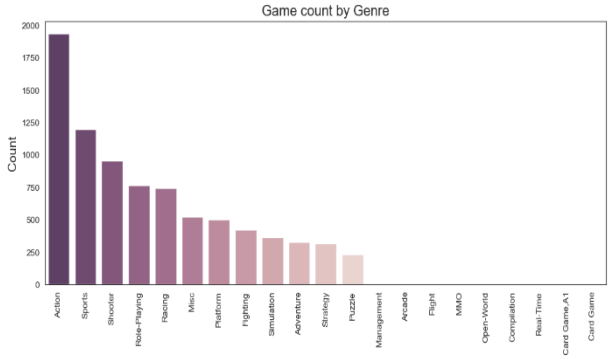

Fig. 6 Game count by Genre

Genres with most games in dataset:

- Action
- Sports

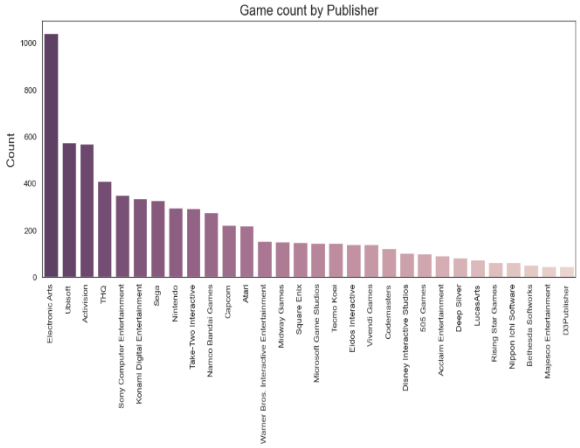

Fig. 7 Game count by Publisher

Publishers with most games in dataset:

- Electronic Art
- Ubisoft
- Activision

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

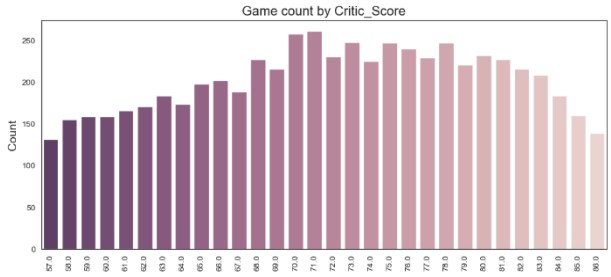

Fig. 8 Game count by Critic Score

Critic Scores with most games in dataset are between 65 to 78

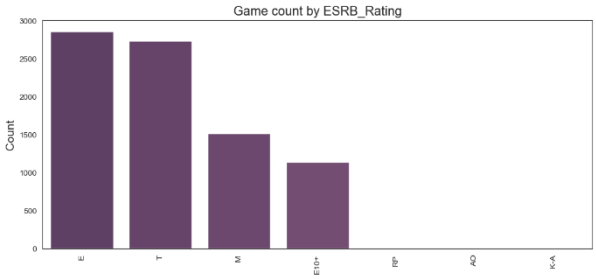

Fig. 9 Game count by ESRB Rating

ESRB Ratings with most games in dataset are :

- E
- T

3. Dataset correlations

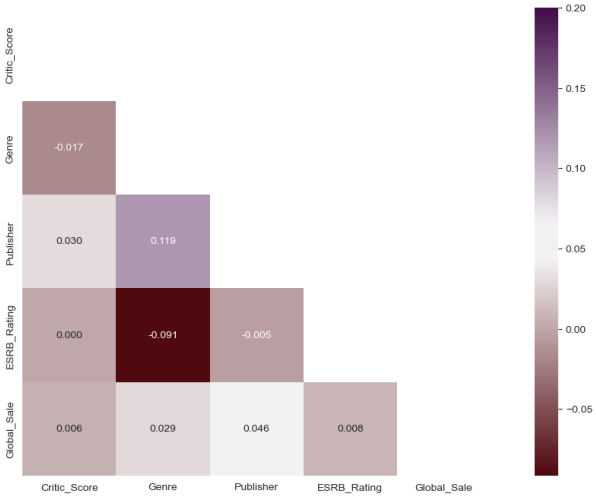

Fig. 10 Dataset correlations for numeric and categorical variables

Strongest correlations:

- global sales to Publisher and ESRB rating
- Publisher to Genre
- **Critic scores to global sales**

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

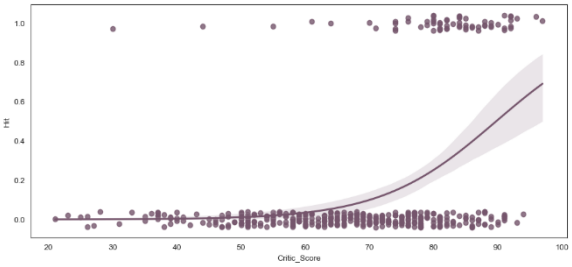

Fig. 11 Critic scores vs global sales after defining hits as those with sales above 1 million units

As expected Critics Score , ESRB rating , publisher, Genre are important feature for global sale and we will take year of release science each years trend impacts
global sale

IX. IMPLEMENTATION AND REPORT

1. Logistic Regression

1.1 Classification Report

	precision	recall	f1-score	support
0	0.86	0.98	0.91	2280
1	0.60	0.18	0.28	448
accuracy			0.85	2728
macro avg	0.73	0.58	0.59	2728
weighted avg	0.82	0.85	0.81	2728

Table. 1 Accuracy report of LR

Model made by Logistic Regression has prediction accuracy of 85%and loss of 35%.

1.2 Confusion matrix

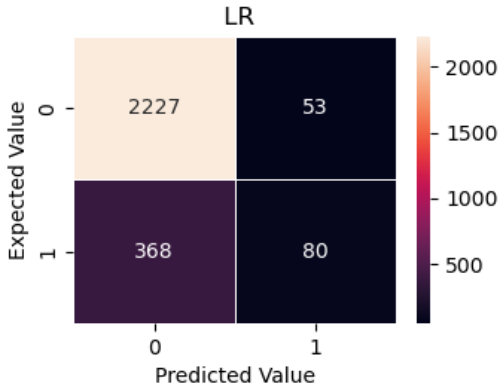

Fig. 12 Confusion matrix of LR

For LR :
The number of game will be hit predicted successfully(TP)=80
The number of game will be hit predicted unsuccessfully(FP)=53
The number of game will not be hit predicted successfully(TN)=2227

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

The number of game will not be hit predicted unsuccessfully(FN)=368

1.3 ROC Curve

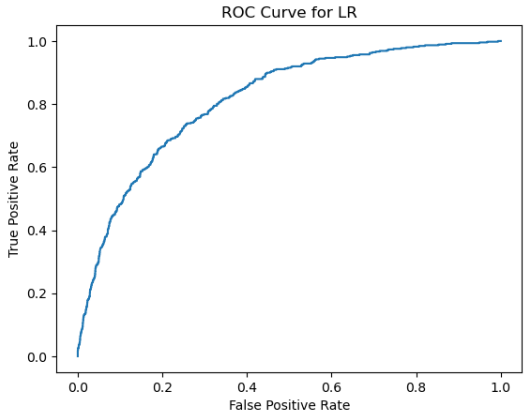

Fig . 13 ROC curve of LR

For logistic regression True positive rate is higher than False positive rate

2. Random Forest Classifier

2.1 Classification Report

	precision	recall	f1-score	support
0	0.88	0.93	0.90	2280
1	0.50	0.38	0.43	448
accuracy			0.84	2728
macro avg	0.69	0.65	0.67	2728
weighted avg	0.82	0.84	0.83	2728

Table. 2 Accuracy Report of RFC

Model made by Random Forest Classification has prediction accuracy of 84% and loss of 47%.

2.2 Confusion matrix

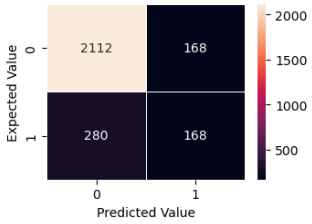

Fig. 14 Confusion Matrix of RFC

For RFC:

The number of game will be hit predicted successfully(TP)=168

The number of game will be hit predicted unsuccessfully(FP)=168

The number of game will not be hit predicted successfully(TN)=2112

The number of game will not be hit predicted unsuccessfully(FN)=280

2.3 ROC Curve

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

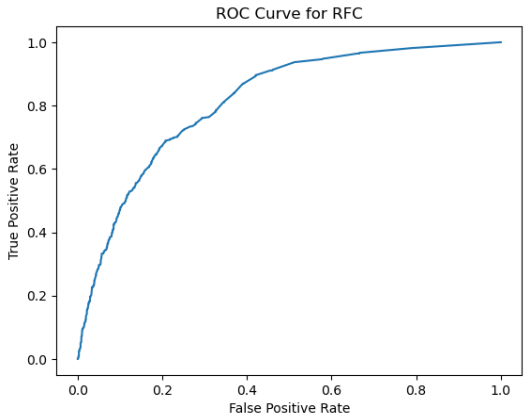

Fig . 15 ROC curve of RFC

For RFC True positive rate is higher than False positive rate but its slightly worse than LR.

X. RESULTS AND DISCUSSION

1 Result

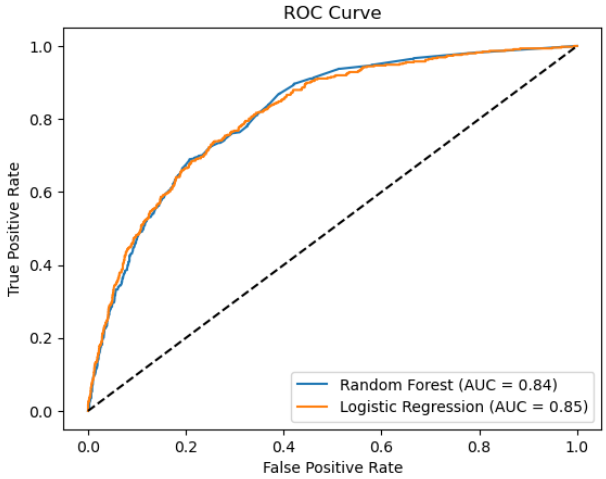

Fig. 16 Compressive ROC curve of LR and RFC

Model	Precession		Recall		F1 Score		Accuracy	Loss
	0	1	0	1	0	1		
LR	0.86	0.60	0.98	0.18	0.91	0.28	0.85	0.35
RFC	0.88	0.50	0.93	0.38	0.90	0.43	0.84	0.48

Table. 3 Performance comparison of LR and RFC

Analyzing the above results we can conclude that LR model provides us higher accuracy (85%) and loss (35%) compared to RFC with accuracy (85%) and loss (35%).

2. Discussion

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10536

Sources such as Kaggle, Gamespot, and Metacritic were mainly used to collect the data for this study. Accurate game sales data is not always released by publishers in order to maintain their stock prices, and other important features such as marketing costs, developer information, or game mechanics are often kept secret to protect their company's privacy. If provided by the publisher, these features can improve the accuracy of our model.

XI. CONCLUSION

An approach has been proposed to predict the probability of a game's success prior to its release by analyzing previous game sales data. More data sets with additional features can be included in the proposed models to enhance their accuracy. By using this model, the game's publisher will have a means to predict its success after release with a high degree of accuracy. This will assist in reducing the number of unsuccessful games.

References:

[1] Trneny, M. (2017). Machine learning for predicting success of video games. Masaryk University, Faculty of Informatics.

[2] BEAUJON, Walter Steven. Predicting Video Game Sales in the European Market. 2012. Available also from: https://www.few.vu.nl/nl/Images/werkstuk-beaujon_tcm243-264134.pdf.

[3] EHRENFELD, Steven Emil. Predicting Video Game Sales Using an Analysis of Internet Message Board Discussions. 2011. Available also from:https://sdsu-dspace.calstate.edu/bitstream/handle/10211.10/1073/Ehrenfeld_Steven.pdf. Master's thesis. San Diego StateUniversity

[4] GHATTAMANENI, Sriram; KOMARRAJU, Agastya kumar. The Game Prophet: Predicting the success of Video Games. 2012. Available also from: https://cepd.okstate.edu/files/Analytics_Ghatta.pdf.

[5] SIFA, Rafet; SRIKANTHY, Sridev; DRACHENZ, Anders; OJEDA,Cesar; BAUCKHAGE, Christian. Predicting Retention in SandboxGames with Tensor Factorization-based Representation Learning.In: NOMIKOS, Petros M. (ed.). 2016 IEEE Conference on Computational Intelligence and Games. 2016, p. 142.

[6] TAMASSIA, Marco; RAFFEY, William; SIFAZ, Rafet; DRACHENX,Anders; ZAMBETTA, Fabio; HITCHENS, Michael. Predicting PlayerChurn in Destiny: A Hidden Markov Models Approach to Predicting Player Departure in a Major Online Game. In: NOMIKOS, Petros M. (ed.). 2016 IEEE Conference on Computational Intelligence andGames. 2016, p. 325.

[7] SHAKER, Noor; ABOU-ZLEIKHA, Mohamed. Transfer Learning forCross-Game Prediction of Player Experience. In: NOMIKOS, Petros M. (ed.). 2016 IEEE Conference on Computational Intelligence andGames. 2016, p. 209.

[8] Video Game Sales with Ratings:https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

[9] PC Game Sales :https://www.kaggle.com/datasets/khaiid/most-selling-pc-games

[10] Manual Dataset :https://drive.google.com/file/d/1tvFanh-sjT7CiIkErtRh-ZNhdh5LFG1Z/view?usp=share_link

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License