Visualizing and Predicting Success of Video Games Sales

Devesh D R
Vellore Institute of Technology
Vellore Campus, Tiruvalam Rd,
Katpadi, Vellore, Tamil Nadu, India - 632014.
deveshd.r2019@vitstudent.ac.in

Mukunth BS
Vellore Institute of Technology
Vellore Campus, Tiruvalam Rd,
Katpadi, Vellore, Tamil Nadu, India - 632014.
mukunthb.s2019@vitstudent.ac.in

***Abstract*— This paper examines sales performed by various video games over the years. Various details of the video games are taken and used to determine what features of a game are most likely to succeed. Machine learning algorithms are used to analyse this data to make predictions on the success of video games and python is used to show the trends and details of both how sales occurred over the years and also visualize how our model predicts using various charts and graphs.**

***Keywords*— Video Games Sales, Random Forest Classifier, Logistic Regression, Data Visualization.**

I. INTRODUCTION

The immense increase in popularity of computers has helped video games become one of the biggest and most profitable markets to exist. In 2018, 2.3 billion gamers around the world are expected to spend $137.9 billion on games Over the last few decades, the video game business has evolved at a breakneck pace. It's difficult to deny that successful video games have a financial impact on video game companies. World of Warcraft, for example, has over 11 million players/subscribers worldwide and generates over $1 billion in annual income for Blizzard Entertainment. Microsoft's Xbox Live gaming program has a total of 20 million subscribers, with 39 million consoles and ten million non-gaming accounts used purely for social networking purposes. The video game market has increased from a few small companies publishing small games to big corporations of multiple people working on games for months. In such a tedious market where a huge number of people work on a single product, failure by the product, not selling, gives huge losses to the company. In an aim to rectify that visualization of previous games success and what details about the games made it a success could make a huge impact on the decision on making such games and change the outcome of whether or not the game would succeed commercially. Since video games are no longer developed by a handful of people but by multi-million-dollar corporations with several hundred people such visualizations could help in making the employee’s work succeed and be lucrative for the company.

One of the main goals of this research is to use machine learning algorithms to discover trending sales. Sales forecasting is an important aspect of every company's operation. It gives timely data that may be used to make informed business decisions. Sales forecasting is a critical tool for new business initiatives and other endeavours. The client and market demand are determined by two different aspects: sales and market predictions. Sales forecasting gives useful information for making strategic company decisions to prevent huge losses.

II. RELATED WORK

Analysis of the video games sales dataset (TM Geethanjali. et al, 2020) based on only one region (NA_Sales) using a linear regression model to find the top features influencing the sales, resulted in three consoles (GB, NES, Wii) being the top influencing factors.

Extensive analysis of the dataset using different models (Amar Aziz. et al, 2018) like Naive Bayes, Decision Trees, and K-NN resulted in Decision Tree model being the most accurate one, yielding the influencing factors for a game’s success as the year, genre and the review scores.

Another analysis study (Alice Yufa. et al, 2019) ran statistical tests through the dataset, adding to the conclusion that the success of a game might be a user’s ‘personal taste’ and not depending on various factors such as genre, platform etc.

A different study (Bodduru Keerthana. et al, 2019) on this same dataset compared four different models such as Linear Regression, Support Vector Regression, Random Forest Classifier and Decision Trees, with Random Forest having the most accuracy of all the four models.

Electronic games are starting to incorporate in-game telemetry that collects data about player, team, and community performance on a massive scale, and as data begins to accumulate, so does the demand for effectively analysing this data. Playing video games for many years has led to a large volume of gaming data that consist of gamer's likings and their playing behaviour. Such data can be used by game creators to extract knowledge for enhancing games. Most of the video gaming business organizations highly depend on a knowledge base and demand prediction of sales trends. However, no studies are conducted to work out the variables that inspire industrial sales predict involvement in and contribution to the sales prediction method. Machine learning techniques are very effective tools in extracting hidden knowledge from an enormous dataset to enhance accuracy and efficiency in predictions.

Random Forest is a supervised machine learning algorithm that randomly creates a forest with several trees.

Modified RF approach: Using disjoint partitions of training dataset to train individual base decision trees, helps in creating diversity in base decision trees. Also, different subsets of attributes are used at each node of the decision tree to increase diversity. This approach generates an RF classifier which is trained efficiently and gives a better classification accuracy compared to the original Random Forest approach. It reduces learning time notably while achieving comparable accuracy as that of original Random Forest.

RF Classifiers for Recommendations: Calculating scores for each node pair according to different measures called predictors, the resulting scores can be interpreted as indicative of the likelihood of future linkage for the given node pair. To determine the relative merit of each predictor, a random forest classifier is trained on older data. The same classifier can then generate predictions for newer data. This method proved to yield accurate recommendations.

LR approach: Logistical Regression method is used to regulate the impact of numerous autonomous variables which are conferred at the same time. This method also predicts any one of the two independent categories of variables. Logistic regression designs the best-fitting function with the help of the maximum likelihood method in order to maximize the probability of classifying the recognized data into the proper division. It is a specific category of regression and it is used in the best way to predict the binary and categorical output. Studies were undertaken to evaluate the performance of ML algorithms and to compare them with the traditional regression technique for predictions. Several studies have found LR to perform very well for prediction with predictive performance, at least as good as ML techniques. It has already been suggested that simple methods typically yield performance almost as good as more sophisticated methods. Therefore, LR classifiers show an acceptable level of sensitivity, accuracy and specificity rates based on the feature sets chosen.

III. PROBLEM STATEMENT

The sales of video games can only succeed if executive decisions can be made accurately. Unlike other industries, the entire company works on one product to release and there is a liability for that product to be a failure. Using visualization methods and machine learning models can greatly help in making better decisions in the direction in which the video game needs to be developed and output a good game. Poor decisions and planning not only harm the sales of the company that quarter but also the reputation leading to future losses. Visualization methods help in seeing what made a game succeed in the past and machine learning models help us predict what decisions help in future games and can be taken. In this research project, we aim to find suitable visualizations to find past trends and learn valuable information unavailable from raw tabulated data along with employing efficient and effective visualization methods as to which aspects of the video game market are most profitable.

IV. PROPOSED METHOD

Machine learning models are mathematical models which use input data and learn correlations using various mathematical and statistical models and use the input numbers to generate a model which returns the result or value which it was meant to calculate based on past features. They can be unsupervised or supervised. Unsupervised algorithms do not have values that are to be predicted and generally group the data based on the features. Supervised algorithms on the other hand already have the results and are generally based on past data to more accurately predict future data. For this research project, we will be using supervised machine learning as we already have details of past success as seen in the workflow diagram in Fig 1.

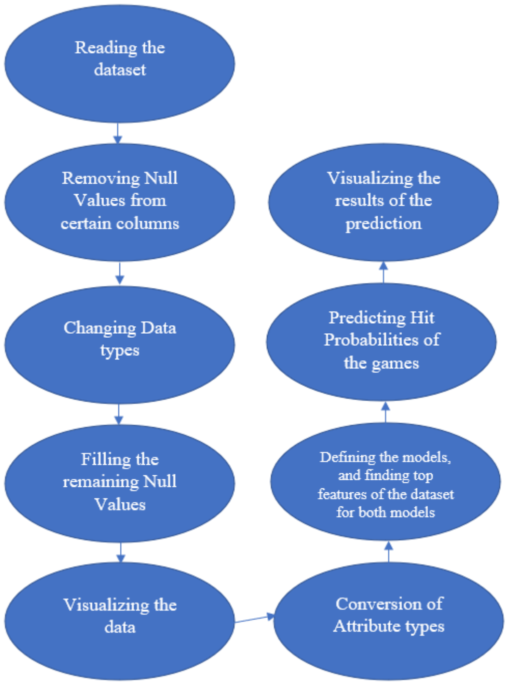

Reading the dataset
Removing Null Values from certain columns
Changing Data types
Filling the remaining Null Values
Visualizing the data
Conversion of Attribute types
Defining the models, and finding top features of the dataset for both models
Predicting Hit Probabilities of the games
Visualizing the results of the prediction

Fig 1: Workflow of visualizing and predicting video game sales

a) Random Forest Classifier

Random Forest is a method used for classification, regression, and other tasks by constructing trees based on features or aspects of the data and finally averaging them to a suitable predictor based on the information learned by the model as seen in Fig 2. Random Forest Classifier results in increased performance. When compared to decision-making, this is a suitable candidate for prediction purposes.

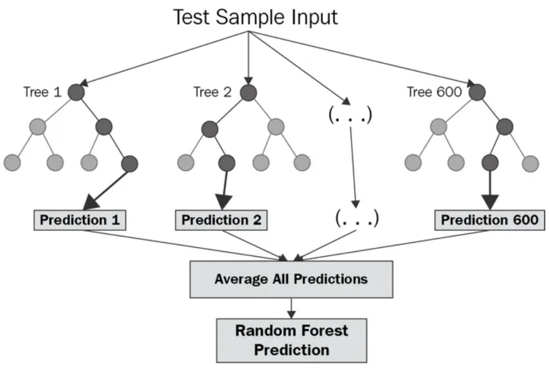


Fig 2: Random Forest Classifier Example

b) Logistic Regression

Logistic regression is a statistical model which uses a logistic function as seen in Fig 3 to generate a model which has parameters based on input to generate a result or output comparable to the data set as input. The logistic regression model itself simply models the probability of output in terms of input and can be modified into a classifier for prediction purposes.

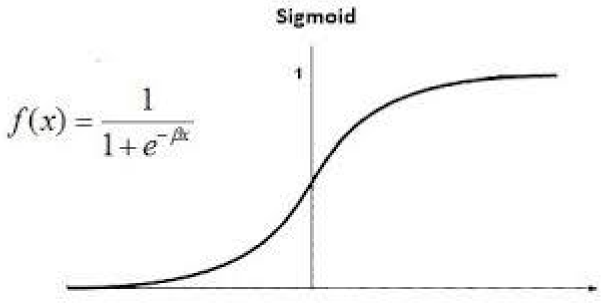


Fig 3: Function used in logistic regression

V. METHODOLOGY

The research project is visualized and implemented using python as it has access to various libraries making it easy for implementing visualizations and machine learning models. We begin by importing the required modules. The dataset is then read from a CSV file.

a) Dataset

The dataset consists of various video games over the years and various details about the game which can be used as features. The dataset consists of details about sales of video games taken from the website VGChartz which contains almost all games that were released with their corresponding platform, year, publisher, and other details along with their user score, critic score, and the number of critics and users used to calculate the score from Metacritic. The columns in the dataset are the names of the games and the years in which they were released. The next attribute is the platform or the console it was meant for such as the PC, PlayStation, Gameboy, Xbox, and so on. We then have the Genre of the game such as Action, Shooter, Sports, Racing, and other miscellaneous genres. We then have the publisher or the company which made the game. We then have the critic scores and number of critics who gave that review similarly we have the user scores and the number of users who gave that review. We then have the rating of the game which can have various categorical values such as Early childhood, Everyone 10+, Teen, Mature, and so on. We then have sales of the game region-wise (North America, Europe, Japan, the Rest of the world, Global).

b) Implementation

The dataset consisted of various null values such as for the year of release and publisher, critic and user score, and count and rating. The null values in the year of release and publisher proved to be small and respective rows could be removed. The null values in the other column however proved to be too significant to just drop so we decided to fill it with the mode value. Along with this, we pre-process the dataset to data types suitable for manipulation, and finally, we have no null values in the data frame. Using this we perform data visualization to understand the data.

c) Visualization

We compare the various features of the dataset concerning the year or time of the release of the game. The features compared here are the global sales for the genre and platform.

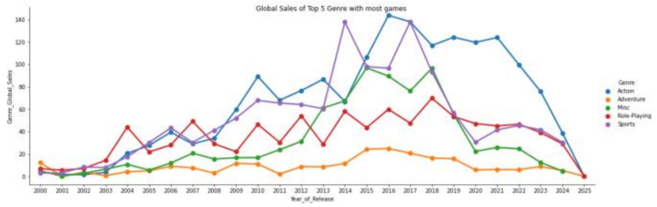


Fig 4: Time chart of global sales of top 5 genres with most games

Fig 4 shows certain features and when they peaked. For example, the action genre was popular in 2012 and peaked while the strategy games were really popular in 2010.

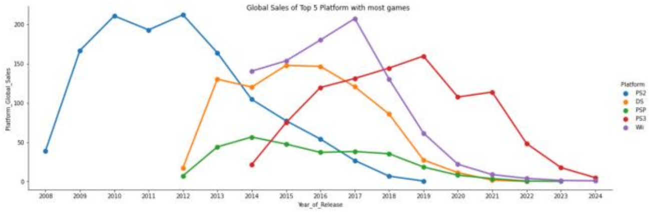


Fig 5: Time chart of global sales of top 5 platforms with most games

Fig 5 plots a time chart according to platform. We find certain trends in the consoles over the years. It is found that the sales peak when the console is first released and drop when the next console in the series is released. Examples of these are sales in ps3 sales dropping when ps4 is released and the game boy console games peaking in 2005 and dropping by 2010. However, we also find that PC game sales have been steadily increasing.

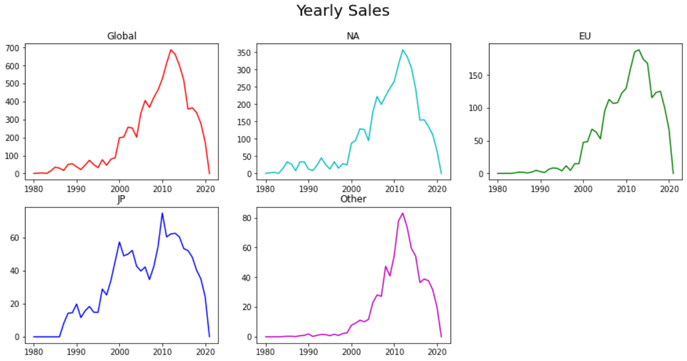


Fig 6: Time chart of global and regional sales

Fig 6 shows the plot of the global and region-wise sales for all games. It is found that there has been a steady increase from the 1980s and have reached the peak at around 2013. It is also found that the curves of North America and Europe closely resemble the global sales showing that they contribute the most and the global trends generally follow these two regions.

Fig 7 plots bar charts to understand the best performers in each category.

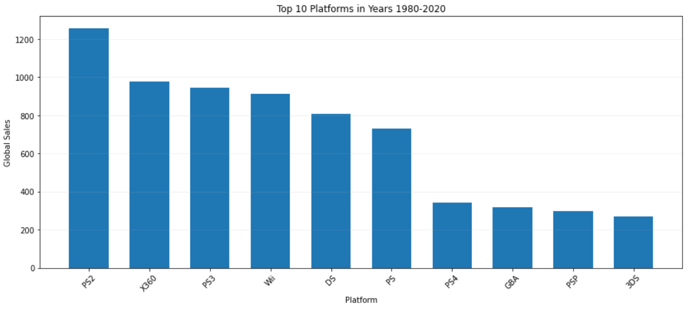


Fig 7 : Top platforms with most sales over the years

It is found that the PS2 and the X360 have had the most success in selling video games which also corresponds to the peak in sales as that was the time both the consoles were released.

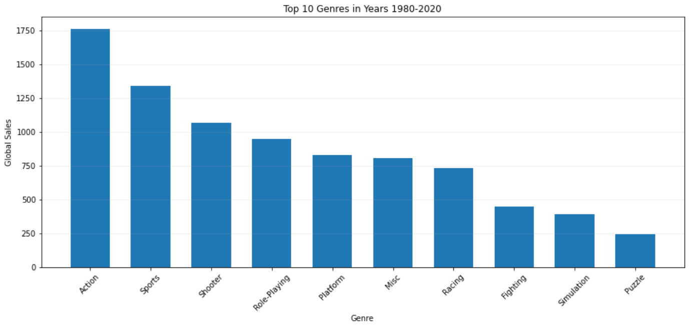


Fig 8 : Top selling genres over the year

From Fig 8 it is found that most sales over the years have happened in the Action genre of games followed by sports and shooter. We can safely say that these genres tend to sell better over the past few years.

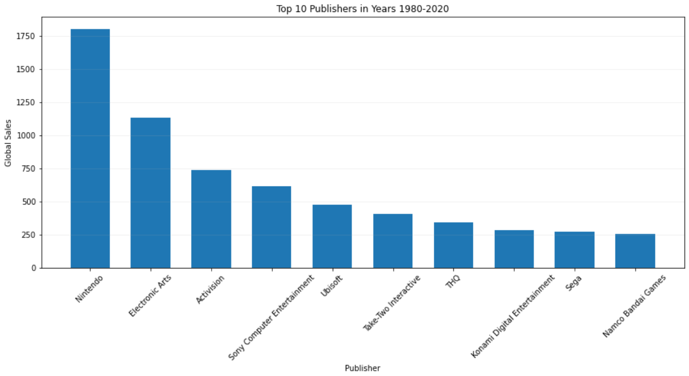


Fig 9 : Top performing Publishers over the years

Fig 9 plots the publishers. It is seen that Nintendo, a company that makes action-adventure games, is the highest which also agrees with our previous data. This is followed by Electronic Art which makes sports games again corresponding to our previous data.

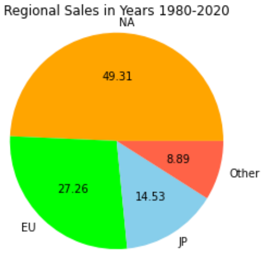


Fig 10 : Composition of Global Sales according to region

Fig 10 is a pie chart to show the percentage of parts of a whole. In this case, we have global sales which are made up of sales in each of the individual regions. We find that the most profitable market for video games is North America generating the highest revenue followed closely by the EU.

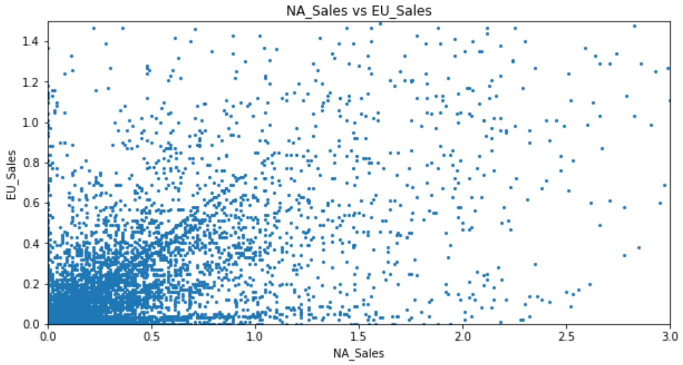


Fig 11: Sales in North America vs Sales in Europe

Fig 11 shows scatter plots where each dot represents a game and its position with respect to the x and y-axis represents the sales in each region. Here we see most of the games are linear showing that most games tend to do equally well in North America and Europe.

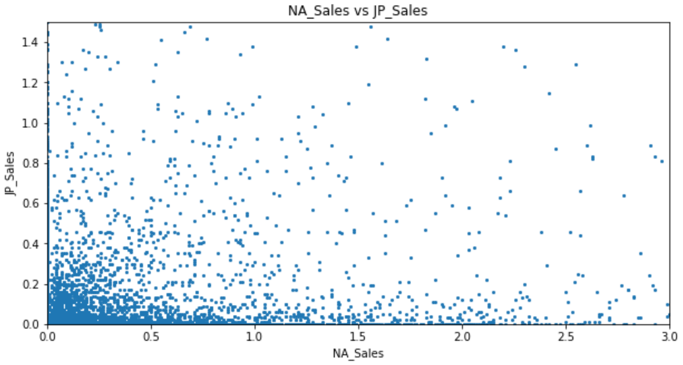


Fig 12 : Sales in North America vs Sales in Japan

Fig 12 compares North America to Japan. When comparing North America to Japan or the rest of the world however it leans towards North America showing that even if it does well in NA it may not do well in Japan and the rest of the world. Also, we see that most of the games have very few sales, and only rarely do games tend to cross 3 million and most have less than a million sales.

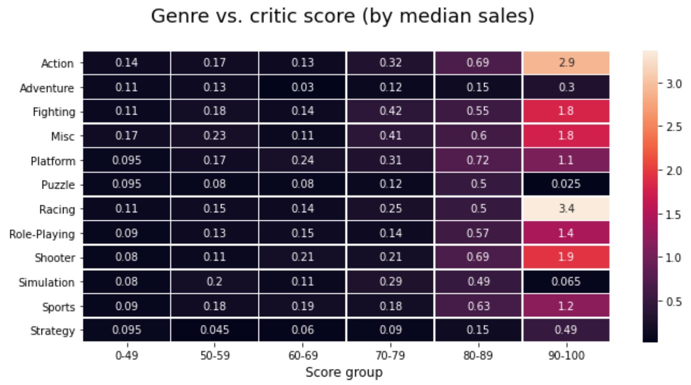


Fig 12 : Correlation of Genre with critic score

Fig 12 shows a heat map to find the correlation between datasets. Here we use the genre and critic score. It is found that critics tend to be most partial to action games and racing games which could attribute their success.

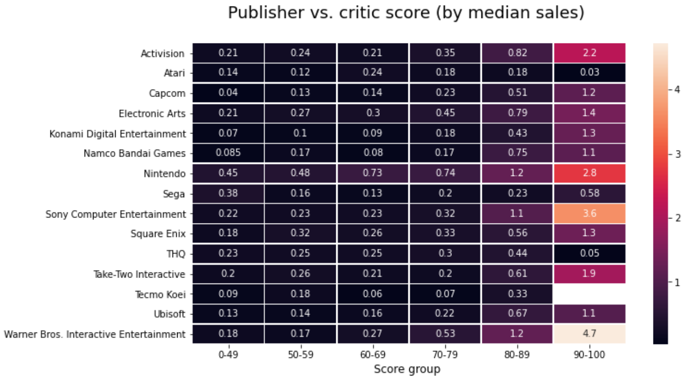


Fig 13: Correlation between Publisher and critic score

From Fig 13 it is seen which publishers do critics tend to favour the most. Here we find that the Warner brothers are favoured among the critics along with Sony which releases many PlayStation hits.

d) Model Prediction

After visualizing the data, we begin prediction by pre-processing the input to suitably be used by the model. We convert critical scores from quantitative categorical data to qualitative and then convert them and group them and convert 'Platform', 'Genre', 'Publisher', 'Rating' from categorical to a numerical value denoting the category. We also plot a general correlation between all columns.

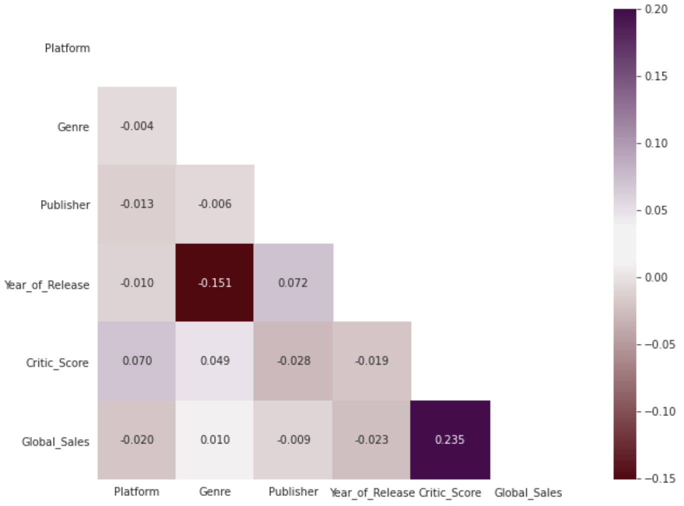
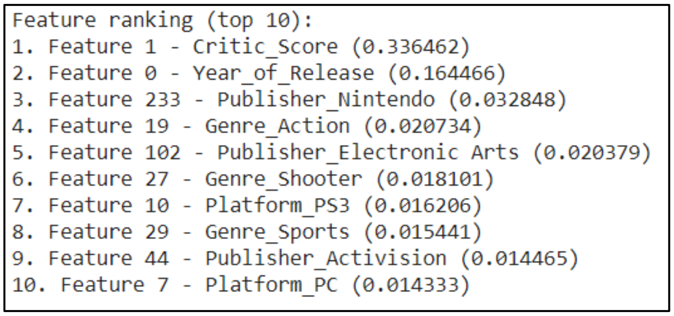


Fig 14 : Correlation between all features

From Fig 14 it is found that the Global sales tend to follow critical scores closely and that genre is in no way related to when the game was published.

To classify according to whether the game will be a hit we keep a threshold of 1 million. If a game gets more global sales than the threshold it is deemed as a success otherwise it is not a hit. Denoting this hit the input data is converted to one-hot encoding and entered into the random forest classifier and the logistic regression model. The features that are most useful for prediction are shown in Fig 15.



```
Feature ranking (top 10):
1. Feature 1 - Critic_Score (0.336462)
2. Feature 0 - Year_of_Release (0.164466)
3. Feature 233 - Publisher_Nintendo (0.032848)
4. Feature 19 - Genre_Action (0.020734)
5. Feature 102 - Publisher_Electronic Arts (0.020379)
6. Feature 27 - Genre_Shooter (0.018101)
7. Feature 10 - Platform_PS3 (0.016206)
8. Feature 29 - Genre_Sports (0.015441)
9. Feature 44 - Publisher_Activision (0.014465)
10. Feature 7 - Platform_PC (0.014333)
```

Fig 15 : Feature ranking with their weightage

From Fig 16 it is seen that the critic score strongly correlates to the global sales success which is true as people tend to see such reviews before buying a game. Other features which affect output are features such as the year when it was released or the publisher being Nintendo or EA or the game being an action genre.

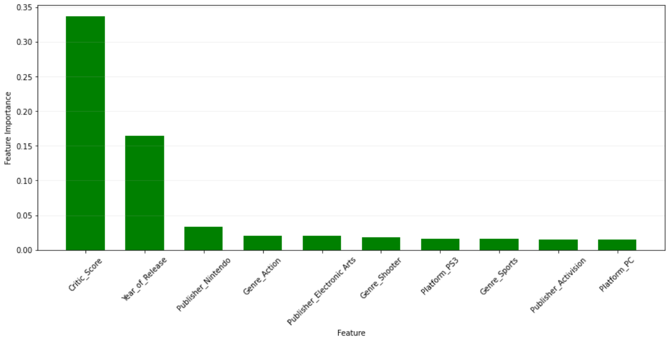
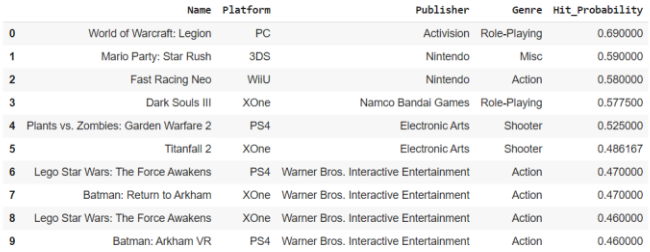


Fig 16 : Importance of features

For this project, a random forest classifier is selected as the feature weightage seemed more accurately co-relate with previous data even though both the accuracies are similar.

Using this trained model, the probability of the game being a success for the games released during 2020 in our dataset is generated.



	Name	Platform	Publisher	Genre	Hit_Probability
0	World of Warcraft: Legion	PC	Activision	Role-Playing	0.690000
1	Mario Party: Star Rush	3DS	Nintendo	Misc	0.590000
2	Fast Racing Neo	WiiU	Nintendo	Action	0.580000
3	Dark Souls III	XOne	Namco Bandai Games	Role-Playing	0.577500
4	Plants vs. Zombies: Garden Warfare 2	PS4	Electronic Arts	Shooter	0.525000
5	Titanfall 2	XOne	Electronic Arts	Shooter	0.486167
6	Lego Star Wars: The Force Awakens	PS4	Warner Bros. Interactive Entertainment	Action	0.470000
7	Batman: Return to Arkham	XOne	Warner Bros. Interactive Entertainment	Action	0.470000
8	Lego Star Wars: The Force Awakens	XOne	Warner Bros. Interactive Entertainment	Action	0.460000
9	Batman: Arkham VR	PS4	Warner Bros. Interactive Entertainment	Action	0.460000

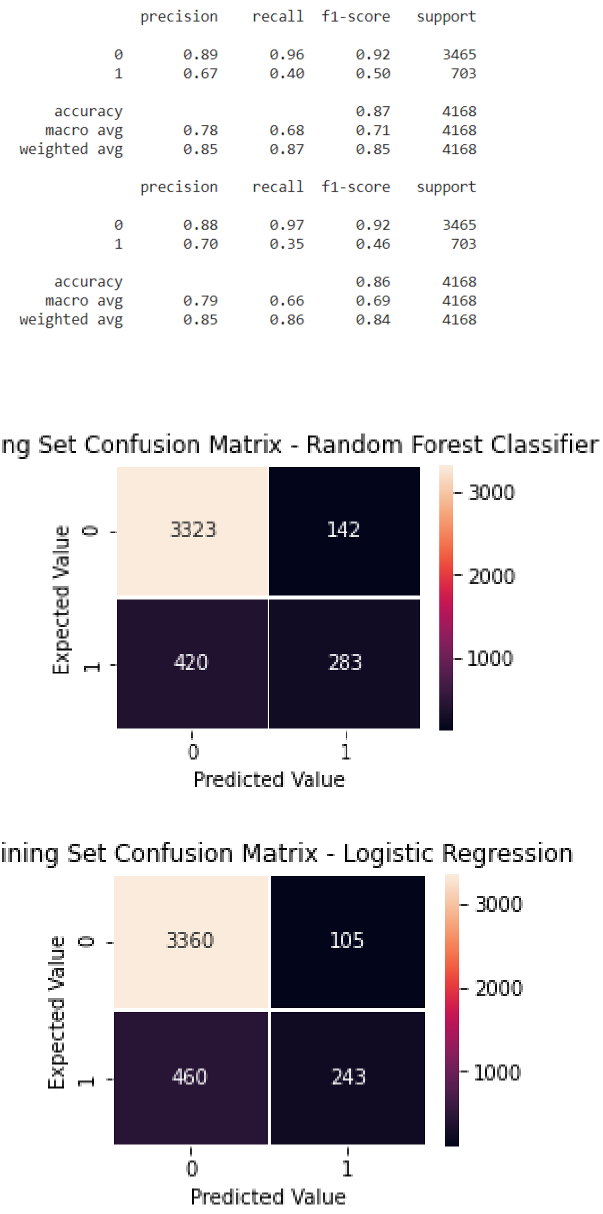
Fig 17 : Results of model as Hit Probability

From Fig 17 it is seen that these games sold well during the year and the data features such as the genre and publisher other such features strongly relate to previous data plotted and feature weights.

For added functionality, a function that takes user input and uses the model to generate the hit probability is also written for easy checking and practical use of the model as seen in Fig 18.

VI. RESULTS

For both logistic regression and random forest classifier accuracy of about 86% is achieved with the confusion matrix of both having precision recall of games not having a success being predicted more accurately than a game being a success. The corresponding F1 score is calculated from both precision and recall values being



```
              precision    recall  f1-score   support

           0       0.89      0.96      0.92      3465
           1       0.67      0.40      0.50       703

    accuracy                           0.87      4168
   macro avg       0.78      0.68      0.71      4168
weighted avg       0.85      0.87      0.85      4168
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      3465
           1       0.70      0.35      0.46       703

    accuracy                           0.86      4168
   macro avg       0.79      0.66      0.69      4168
weighted avg       0.85      0.86      0.84      4168
```

Fig 19 :Confusion Matrix and accuracy measure of both models

Although the accuracy of both models is similar, the feature ranking and weighting of random forest agrees with previous data more and is used as the primary model.

VII. REPORT

Using generated hit probabilities for 2020 games charts and graphs are plotted to better understand the predicted data.

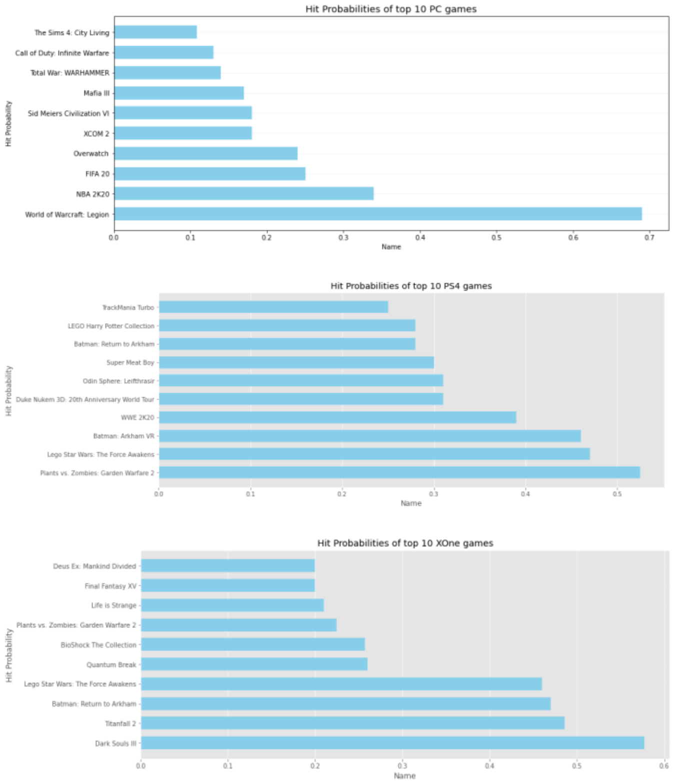
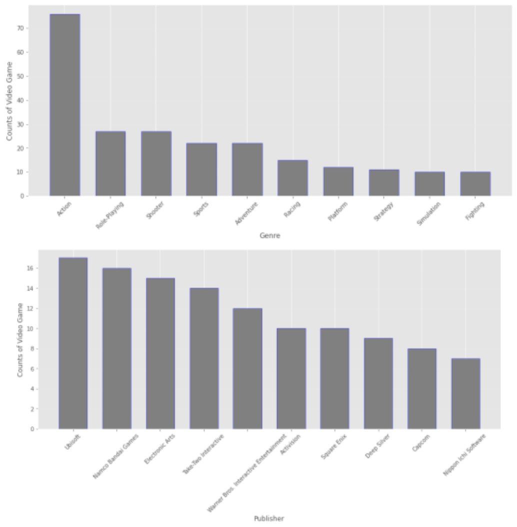


Fig 20 : Successful games in top platforms predicted by model

Fig 20 shows the most popular games with how likely they are going to be a hit in the most famous consoles like pc, ps4, and x360. It is found that World of Warcraft Legion, Plants vs Zombies, and Dark Souls 3 are most likely to succeed according to our model and they are all games that did well in 2020.



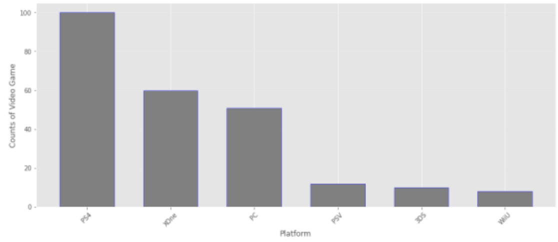


Fig 21 :Best performing Genre Publisher platform predicted by model

From the predicted model, it is found that like previously Action games are going to do well in the market. Ubisoft is going to be the most profitable while most people are going to buy games for ps4, Xone, and PC as seen in Fig 21.

VIII. CONCLUSION

This project visualized the dataset to understand trends and details of the sales that occur in video games sales that happened over the years. Features like which console is going to have the most sales, which company will most likely release a game that will be a success, and also what type of genre should a company make for the most profit should they venture into this market. This project gives a comprehensive view of the raw sales data which can be used to make an informed decision. While we have made models, which use most of the features, the rating attribute is less utilized due to the lack of data and could be recorded and made use of to make the model better and more accurate. Along with the rating, the age demographic would also give valuable insight to predicting a game's success.

IX. ACKNOWLEDGMENT

We would like to express our gratitude to our guide, Prof. Sivasankari S for showing her support, guidance, and encouragement for taking up this project. We would like to thank our university, Vellore Institute of Technology, for giving us this opportunity. We would like to thank our family and friends for supporting us through this project.

X. REFERENCES

[1] Babb, J. and Terry, N., 2013. Comparing video game sales by gaming platform. *Southwestern Economic Review*, *40*, pp.25-46.

[2] Aziz, A., Ismail, S., Othman, M.F. and Mustapha, A., 2018, July. Empirical Analysis on Sales of Video Games: A Data Mining Approach. In *Journal of*

Physics: Conference Series (Vol. 1049, No. 1, p. 012086). IOP Publishing.

[3] Bodduru Keerthana, Dr. K.Venkata Rao, 2019, June. Sales Prediction On Video Games Using Machine Learning. In *JETIR June 2019, Volume 6, Issue 6.*

[4] Alice Yufa, Jonathan L. Yu, Henry Chan, Paul D. Berger, 2019. Predicting Global Video-Game Sales. *Journal of Research in Business and Management Volume 7 ~ Issue 3 (2019)* pp: 60-64

[5] TM Geethanjali, Ranjan D, Swaraj HY, Thejaskumar MV, Chandana HP, 2020, May. Video Games Sales Analysis: A Data Science Approach. In *2020 IJCRT Volume 8, Issue 5 May 2020.*

[6] Bowman, B., Elmqvist, N. and Jankun-Kelly, T.J., 2012. Toward visualization for games: Theory, design space, and patterns. *IEEE transactions on visualization and computer graphics*, *18*(11), pp.1956-1968.

[7] Alyssa, D. and Yong, T., 2019. Visualizing the International Video Game Industry. *innovation*, p.9.

[8] Sinha, V.Y.K.P.K. and Kulkarni, V.Y., 2013. Efficient learning of random forest classifier using disjoint partitioning approach. In *Proceedings of the World Congress on Engineering* (Vol. 2).

[9] Guns, R. and Rousseau, R., 2014. Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, *101*(2), pp.1461-1473.

[10] Thabtah, F., Abdelhamid, N. and Peebles, D., 2019. A machine learning autism classification based on logistic regression analysis. *Health information science and systems*, *7*(1), pp.1-11.

[11] S. Celine, M. Maria Dominic, M. Savitha Devi, 2020, January. Logistic Regression for Employability Prediction. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-9 Issue-3.*

PLAGIARISM SCAN REPORT

Words	956	Date	May 29,2021
Characters	6339	Excluded URL	

5% Plagiarism	95% Unique	2 Plagiarized Sentences	39 Unique Sentences

Content Checked For Plagiarism

Visualizing and Predicting Success of Video Games Sales

Devesh D R
Vellore Institute of Technology
Vellore Campus, Tiruvalam Rd,
Katpadi, Vellore, Tamil Nadu, India - 632014.
deveshd.r2019@vitstudent.ac.in
Mukunth BS
Vellore Institute of Technology
Vellore Campus, Tiruvalam Rd,
Katpadi, Vellore, Tamil Nadu, India - 632014.
mukunthb.s2019@vitstudent.ac.in

Abstract— This paper examines sales performed by various video games over the years. Various details of the video games are taken and used to determine what features of a game are most likely to succeed. Machine learning algorithms are used to analyse this data to make predictions on the success of video games and python is used to show the trends and details of both how sales occurred over the years and also visualize how our model predicts using various charts and graphs.

Keywords— Video Games Sales, Random Forest Classifier, Logistic Regression, Data Visualization.

I. Introduction

The immense increase in popularity of computers has helped video games become one of the biggest and most profitable markets to exist. In 2018, 2.3 billion gamers around the world are expected to spend \$137.9 billion on games Over the last few decades, the video game business has evolved at a breakneck pace. It's difficult to deny that successful video games have a financial impact on video game companies. World of Warcraft, for example, has over 11 million players/subscribers worldwide and generates over \$1 billion in annual income for Blizzard Entertainment. Microsoft's Xbox Live gaming program has a total of 20 million subscribers, with 39 million consoles and ten million non-gaming accounts used purely for social networking purposes. The video game market has increased from a few small companies publishing small games to big corporations of multiple people working on games for months. In such a tedious market where a huge number of people work on a single product, failure by the product, not selling, gives huge losses to the company. In an aim to rectify that visualization of previous games success and what details about the games made it a success could make a huge impact on the decision on making such games and change the outcome of whether or not the game would succeed commercially. Since video games are no longer developed by a handful of people but by multi-million-dollar corporations with several hundred people such visualizations could help in making the employee’s work succeed and be lucrative for the company.

One of the main goals of this research is to use machine learning algorithms to discover trending sales. Sales forecasting is an important aspect of every company's operation. It gives timely data that may be used to make informed business decisions. Sales forecasting is a critical tool for new business initiatives and other endeavours. The client and market demand are determined by two different aspects: sales and market predictions. Sales forecasting gives useful information for making strategic company decisions to prevent huge losses.

II. Related Work

Analysis of the video games sales dataset (TM Geethanjali. et al, 2020) based on only one region (NA_Sales) using a linear regression model to find the top features influencing the sales, resulted in three consoles (GB, NES, Wii) being the top influencing factors.

Extensive analysis of the dataset using different models (Amar Aziz. et al, 2018) like Naive Bayes, Decision Trees, and K-NN resulted in Decision Tree model being the most accurate one, yielding the influencing factors for a game's success as the year, genre and the review scores.

Another analysis study (Alice Yufa. et al, 2019) ran statistical tests through the dataset, adding to the conclusion that the success of a game might be an user's 'personal taste' and not depending on various factors such as genre, platform etc.

A different study (Bodduru Keerthana. et al, 2019) on this same dataset compared four different models such as Linear Regression, Support Vector Regression, Random Forest Classifier and Decision Trees, with Random Forest having the most accuracy of all the four models.

III. Literature Survey

Electronic games are starting to incorporate in-game telemetry that collects data about player, team, and community performance on a massive scale, and as data begins to accumulate, so does the demand for effectively analyzing this data. Playing video games for many years has led to a large volume of gaming data that consist of gamer's likings and their playing behaviour. Such data can be used by game creators to extract knowledge for enhancing games. Most of the video gaming business organizations highly depend on a knowledge base and demand prediction of sales trends. However, no studies are conducted to work out the variables that inspire industrial sales predict involvement in and contribution to the sales prediction method. Machine learning techniques are very effective tools in extracting hidden knowledge from an enormous dataset to enhance accuracy and efficiency in predictions.

Random Forest is a supervised machine learning algorithm that randomly creates a forest with several trees.

Modified RF approach: Using disjoint partitions of training dataset to train individual base decision trees, helps in creating diversity in base decision trees. Also different subsets of attributes are used at each node of the decision tree to increase diversity. This approach generates an RF classifier which is trained efficiently and gives a better classification accuracy compared to the original Random Forest approach. It reduces learning time notably while achieving comparable accuracy as that of original Random Forest.

RF Classifiers for Recommendations: Calculating scores for each node pair according to different measures called predictors, the resulting scores can be interpreted as indicative of the likelihood of future linkage for the given node pair. To determine the relative merit of each predictor, a random forest classifier is trained on older data. The same classifier can then generate predictions for newer data. This method proved to yield accurate recommendations.

Sources	Similarity
JETIR1907H50.pdf - © 2019 JETIR June 2019 Volume 6 ... https://www.coursehero.com/file/93829983/JETIR1907H50pdf/	4%
Recommending research collaborations using link prediction ... · The same classifier can then generate predictions for newer data. The top predictions are treated as recommendations for future collaboration. We apply the technique to research collaborations between cities in Africa, the Middle East and South-Asia, focusing on the topics of malaria and tuberculosis. Results show that the method yields accurate recommendations. Moreover, the method can be used to determine the relative strengths of each predictor. https://link.springer.com/article/10.1007/s11192-013-1228-9	3%

PLAGIARISM SCAN REPORT

Words	981	Date	May 29,2021
Characters	6062	Excluded URL	

0% Plagiarism	100% Unique	0 Plagiarized Sentences	43 Unique Sentences

Content Checked For Plagiarism

LR approach: Logistical Regression method is used to regulate the impact of numerous autonomous variables which are conferred at the same time. This method also predicts any one of the two independent categories of variables. Logistic regression designs the best-fitting function with the help of the maximum likelihood method in order to maximize the probability of classifying the recognized data into the proper division. It is a specific category of regression and it is used in the best way to predict the binary and categorical output. Studies were undertaken to evaluate the performance of ML algorithms and to compare them with the traditional regression technique for predictions. Several studies have found LR to perform very well for prediction with predictive performance, at least as good as ML techniques. It has already been suggested that simple methods typically yield performance almost as good as more sophisticated methods. Therefore, LR classifiers show an acceptable level of sensitivity, accuracy and specificity rates based on the feature sets chosen.

IV. Problem Statement

The sales of video games can only succeed if executive decisions can be made accurately. Unlike other industries, the entire company works on one product to release and there is a liability for that product to be a failure. Using visualization methods and machine learning models can greatly help in making better decisions in the direction in which the video game needs to be developed and output a good game. Poor decisions and planning not only harm the sales of the company that quarter but also the reputation leading to future losses. Visualization methods help in seeing what made a game succeed in the past and machine learning models help us predict what decisions help in future games and can be taken. In this research project, we aim to find suitable visualizations to find past trends and learn valuable information unavailable from raw tabulated data along with employing efficient and effective visualization methods as to which aspects of the video game market are most profitable.

V. Proposed Method

Machine learning models are mathematical models which use input data and learn correlations using various mathematical and statistical models and use the input numbers to generate a model which returns the result or value which it was meant to calculate based on past features. They can be unsupervised or supervised. unsupervised algorithms do not have values that are to be predicted and generally group the data based on the features. Supervised algorithms on the other hand already have the results and are generally based on past data to more accurately predict future data. For this research project, we will be using supervised machine learning as we already have details of past success as seen in the workflow diagram in Fig 1.

Fig 1: Workflow of entire project

a) Random Forest Classifier

Random Forest is a method used for classification, regression, and other tasks by constructing trees based on features or aspects of the data and finally averaging them to a suitable predictor based on the information learned by the model as seen in Fig 2. Random Forest Classifier results in increased performance. When compared to decision-making, this is a suitable candidate for prediction purposes.

Fig 2: Random Forest Classifier Example

b) Logistic Regression

Logistic regression is a statistical model which uses a logistic function as seen in Fig 3 to generate a model which has parameters based on input to generate a result or output comparable to the data set as input. The logistic regression model itself simply models the probability of output in terms of input and can be modified into a classifier for prediction purposes.

Fig 3: Function used in logistic regression

VI. Methodology

The research project is visualized and implemented using python as it has access to various libraries making it easy for implementing visualizations and machine learning models. We begin by importing the required modules. The dataset is then read from a CSV file.

a) Dataset

The dataset consists of various video games over the years and various details about the game which can be used as features. The dataset consists of details about sales of video games taken from the website VGChartz which contains almost all games that were released with their corresponding platform, year, publisher, and other details along with their user score, critic score, and the number of critics and users used to calculate the score from Metacritic. The columns in the dataset are the names of the games and the years in which they were released. The next attribute is the platform or the console it was meant for such as the PC, PlayStation, Gameboy, Xbox, and so on. We then have the Genre of the game such as Action, Shooter, Sports, Racing, and other miscellaneous genres. We then have the publisher or the company which made the game. We then have the critic scores and number of critics who gave that review similarly we have the user scores and the number of users who gave that review. We then have the rating of the game which can have various categorical values such as Early childhood, Everyone 10+, Teen, Mature, and so on. We then have sales of the game region-wise (North America, Europe, Japan, the Rest of the world, Global).

b) Implementation

The dataset consisted of various null values such as for the year of release and publisher, critic and user score, and count and rating. The null values in the year of release and publisher proved to be small and respective rows could be removed. The null values in the other column however proved to be too significant to just drop so we decided to fill it with the mode value. Along with this, we preprocess the dataset to data types suitable for manipulation, and finally, we have no null values in the data frame. Using this we perform data visualization to understand the data.

Sources	Similarity

SmallSEOTools

PLAGIARISM SCAN REPORT

Words	445	Date	May 29,2021
Characters	2875	Excluded URL	

0% Plagiarism	100% Unique	0 Plagiarized Sentences	25 Unique Sentences

Content Checked For Plagiarism

What all have this pandemic not done to all of us!? The way every individual looks at it differs. Every sector has been hit down economically and whatnot. Mainly second wave(especially for some countries) of coronavirus is still worse than the first. The way countries have been dealing with vaccinations is at the next level.
Anyhow, we are mainly going to focus on the impact on the sector of "GLOBAL- EDUCATION". So many of the students, teachers, working employees, and their families have been affected by this. 2nd year in a row all the schools, institutions, colleges, and everything related to the educational system had to be forced to stay closed. All the dreams of students to pass or attend exams that shape their future were not been conducted or considered.
It's a sort of situation where no one is to be blamed. It's not a mistake of the education system or the Government or people, these are the circumstances where everyone needs to understand and act according to it.
There are cases where parents are forced to pay the full fees even for online education. Due to this many parents couldn't afford to pay the fees and students were denied to attend classes and lost an academic year.
Many libraries and transportation systems, stationery shops, and many are affected by these lockdowns. Unable to survive many schools are closed permanently. Many students as well as teachers have problems, accessing the internet and even teachers are not able to cope up with online teaching.
Universities are unable to allocate seats for students. From this impact, many poor students were not able to get into a good institution, for their capability or skill. Many students lost their golden opportunities to go and study abroad in one of the best universities and vice versa. From this many illiterate, Family members are doing their daughter’s marriage at a very young age without worrying about their goals and getting their life into trouble.
Also, children are made to learn through, online classes and smartphones. Thereby this has led to more use for other reasons also and creating health issues and more use of technology at an early stage. The lack of teacher-tostudent interaction has also led students to feel less passionate about the
integrity of their work. This turned out students to do half-completed assignments, get the answers from their friends in class, or simply sit instead
of concentrating on what is taught because of all this education has become less important due to COVID. So, COVID has had its impact on global

education. Hoping soon everyone will be free from this and our education goes back to normal.

Sources	Similarity

SmallSEOTools

PLAGIARISM SCAN REPORT

Words	534	Date	May 29,2021
Characters	3172	Excluded URL	

0% Plagiarism	100% Unique	0 Plagiarized Sentences	25 Unique Sentences

Content Checked For Plagiarism

Fig 16 : Importance of features
For this project, a random forest classifier is selected as the feature weightage seemed more accurately co-relate with previous data even though both the accuracies are similar.
Using this trained model, the probability of the game being a success for the games released during 2020 in our dataset is generated.
Fig 17 : Results of model as Hit Probability
From Fig 17 it is seen that these games sold well during the year and the data features such as the genre and publisher other such features strongly relate to previous data plotted and feature weights.
For added functionality, a function that takes user input and uses the model to generate the hit probability is also written for easy checking and practical use of the model as seen in Fig 18.
Fig 18 : Results of returning probability of user input features
I. Result
For both logistic regression and random forest classifier accuracy of about 86% is achieved with the confusion matrix of both having precision recall of games not having a success being predicted more accurately than a game being a success. The corresponding F1 score is calculated from both these precision and recall values.
being
Fig 19 :Confusion Matrix and accuracy measure of both models
Although the accuracy of both models is similar, the feature ranking and weighting of random forest agrees with previous data more and is used as the primary model.
II. Report
Using generated hit probabilities for 2020 games charts and graphs are plotted to better understand the predicted data.
Fig 20 : Successful games in top platforms predicted by model
Fig 20 shows the most popular games with how likely they are going to be a hit in the most famous consoles like pc, ps4, and x360. It is found that World of Warcraft Legion, Plants vs Zombies, and Dark Souls 3 are most likely to succeed according to our model and they are all games that did well in 2020.
Fig 21 :Best performing Genre Publisher platform predicted by model
From the predicted model, it is found that like previously Action games are going to do well in the market. Ubisoft is going to be the most profitable while most people are going to buy games for ps4, Xone, and PC as seen in Fig 21.
III. Conclusion
This project visualized the dataset to understand trends and details of the sales that occur in video games sales that happened over the years. Features like which console is going to have the most sales, which company will most likely release a game that will be a success, and also what type of genre should a company make for the most profit should they venture into this market. This project gives a comprehensive view of the raw sales data which can be used to make an informed decision. While we have made models, which use most of the features, the rating attribute is less utilized due to the lack of data and could be recorded and made use of to make the model better and more accurate. Along with the rating, the age demographic would also give valuable insight to predicting a game's success.

Sources	Similarity