VIDEO GAME SALES ANALYSIS

V.SARALA, DOREPALLI AKHILA
1. Assistant Professor in DEPARTMENT OF MASTER OF COMPUTER SCIENCE, BHIMAVARAM-534202.
E Mail Id: vedalasarala21@gmail.com
2.PG STUDENT OF MCA Dept , D.N.R. COLLEGE, P.G. COURSES (AUTONOMOUS), BHIMAVARAM-534202.
E Mail Id: akhiladorepalli@gmail.com

ABSTRACT

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 depends on a knowledge base and demand prediction of sales trend. 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. In this paper, we briefly analysed the concept of gaming sales data and sales predictions. The various machine learning techniques and measures used for this sales prediction. On the basis of a performance evaluation, best suited predictive models like linear regression, support vector regression, random forest and decision trees etc. are used for the sales trend predictions. The results summarized in performance measures are root mean square error, r-square, and mean absolute error. The studies found that the best fit model is Random forest algorithm, which shows maximum accuracy in future sales prediction.

1. INTRODUCTION

Video game industry needs accurate sales in an exponential market growth. In the last 10 years in the United States the revenue coming from computer and video games increased imposingly. So we have to predict the buying nature of several video game followers by using historical sales data. This study involves extracting the video game sales data and analysing which game has more sales globally when compared to other countries [1]. With this we used machine learning techniques which predict the sales of video game in the market. This approach is useful to several industries which are interested in predicting the sales data [9]. In this paper, we are concerned with predicting the sales of a video game. For this we have used historical time series sales data. Our dataset consists of 11 variables and 500 samples with a combination of categorical and numeric variables. We need to perform data pre-processing on dataset to check whether the data is properly loaded or not, is there any missing values or NA values etc. Out of all these variables few variables are unused so drop those variables. Now find correlation between variables to know the input variable and target variable for applying machine learning algorithms. After applying correlation matrix, we came to know the target variable and input variables. Before applying machine learning algorithms we have to split the dataset into training and testing sets. Finally to obtain better performance, we have to apply possible machine learning algorithms which give us best result

Machine learning algorithms are classified into three categories: supervised learning, unsupervised learning, reinforcement learning [6]. In supervised learning we have input variables and output variables and we apply machine learning technique to

learn mapping function from input to target variable. Supervised learning has two categories: classification and regression. In un-supervised learning we have only input variables and no target variables. It has its own way to discover the structure in the data. In this project we have used supervised learning algorithms they are linear regression, support vector regression, random forest, and decision tree. We also use performance measures such as root mean square error, r-square, mean absolute error. One of the major objective of this research work is to find the trending sales by using machine learning algorithms. Sales prediction is an essential part of business organizations. It provides relevant information that can be used to make strategic business decisions [2]. Sales prediction is very important tool for upcoming business ventures etc.

2. LITERATURE SURVEY AND RELATED WORK

Julie Marcous and Sid-Ahmed Selouani, the authors proposed, "A hybrid subspace-connectionist data mining approach for sales forecasting in video game sales industry" [1], and this paper addresses the issue of sales forecasting using an approach based on connectionist and subspace decomposition methods. Back propagation algorithm is used to predict weekly sales of a video game. For this purpose optimal topology and time-series neural network is implemented. The performance of this system is evaluated and compared with base line reference sales.

Hycinta Andrat and Nazneen Ansari, the authors proposed, "Integrating data mining with computer games" [2], this paper address the information about mining computer game data is new data mining approach that can help in developing games a per a gamers requirements. For this purpose the data mining techniques are applied such as association, classification and clustering for improving game design, game marketing, and game stickiness monitoring, respectively, to enrich game quality

David Buckley, Ke Chen and Joshua Knowles, the authors proposed, "Predicting skill from game play input to a first person shooter" [3], this paper explores how game play input recorded in a first person shooter can predict a player's ability. For this purpose random forest methodology is used to predict player's skill without using game specific features

Jing Zhang and Juan Li, the authors proposed, "Retail Commodity Sale Forecast Model Based on Data Mining" [4], this paper address the information about retail commodity sales forecast, people done more in particular aspect with commodities single sale attributes such as sale volume, sale money, season factor, but all were not considered as the important factor called profit. Profit is the key component for all retail enterprises to succeed. So this paper used SPV model and ID3 decision tree algorithms. And on this basis they predicted sales state of the commodity. Finally they conclude that SPV model is the best model.

Vishal shrivastava, the author proposed, "A study of various clustering algorithms on retail sales data" [5], this paper discusses the four major clustering algorithms k-means, density based, filtered, farthest first clustering algorithm and comparing the performances of these principle clustering algorithm on the aspect of correctly class wise cluster building ability of algorithm. The results are listed on datasets of retail sales using weka interface and compute the correctly cluster building instances in proportion with incorrectly formed cluster. A comparison of these four algorithms is given on the basis percentage of incorrectly classified

Journal of Engineering Sciences

instances.

3. EXISTING SYSTEM

In this project we choose video game sale data, our dataset consists of 11 variables and 500 samples with a combination of categorical and numeric variables. They are Rank, Name of video game, Platform, Year, Genre, Publisher, North American Sales, Europe Sales, Japan Sales, Other Sales and Global Sales. In this dataset name, platform, year, genre, publisher are unused variables, so drop those variables. Now find correlation between variables to know the input variable and target variable for applying ml algorithms.

4. PROPOSED SYSTEM

Apply various possible prediction modelling algorithms to see which provides best results. Linear Regression, Decision Tree, Random Forest and Support vector regression algorithms were used on video game sales data

After loading the dataset in our Python IDE, the very first step would be data cleaning. Since the data is very raw in nature, getting proper visualization before cleaning will be hard. Hence, this is the first step in our architecture. After the data cleaning is performed, we perform data exploration (EDA) on the cleaned dataset. We have performed exploration using seaborn library as well as the Plotly Library.

After completing the data exploration, lot of valuable inference about the data can be gained. From the visualization we find that nearly half of the entries do not have scored attributes. Hence, we decide to make two different models in which the first model will remove all the entries in which score is not present while the other model will use weighted factor to control the scoring of the other entries. But before proceeding towards training the model, respective encoding schemes needs to performed on the categorical entities.

After splitting the data respectively, we plan to apply 8 different machine learning algorithms on the new dataset to find which model gives us the minimum error and the best fit for the data. The respective errors for the models are calculated and the model with the least error is taken. We further try to minimize the error output by finding out which are the right hyperparameters for the data. We used randomized searching after this to which the number of iterations required to get the best output after placing the hyperparameter grid respectively. The final output after fitting the respective hyperparameters are hence calculated and compared with the initial error value. This procedure followed for both the model and the final inferences are presente

Journal of Engineering Sciences

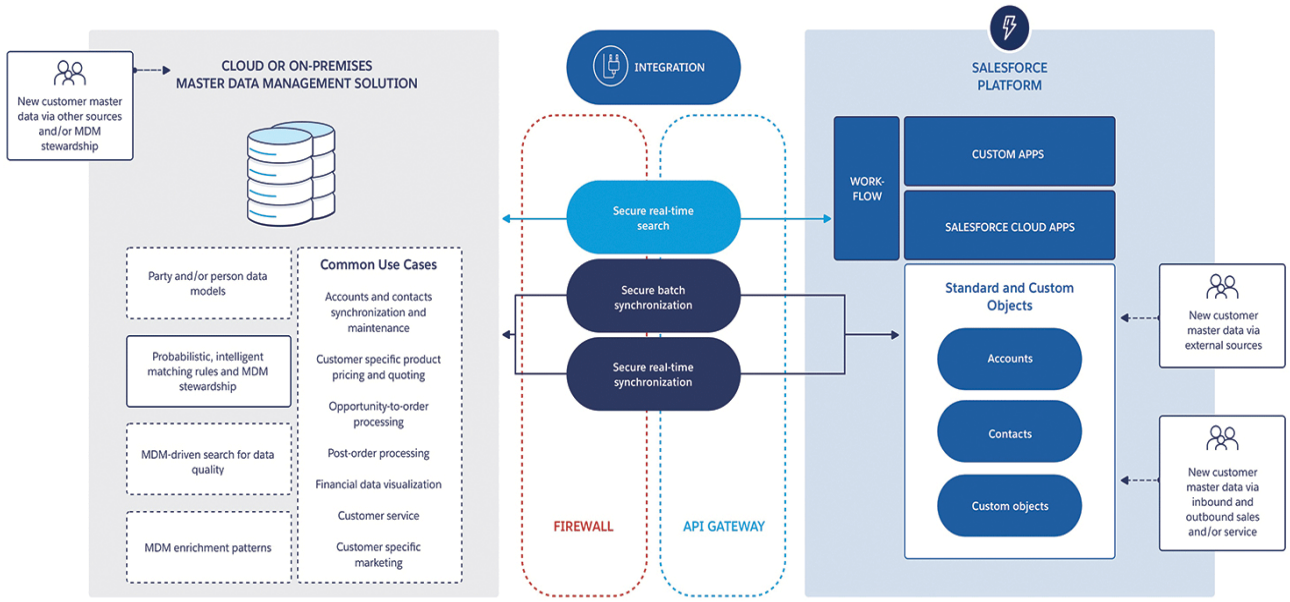


FIG 1 – SYSTEM ARCHITECTURE

5. METHODOLOGIES

MODULE

A .Machine Learning Algorithm

The ML algorithm is a logic that grasp one step ahead when exposed to more information/data. When ML is exposed to training data it produces model. To build a Model, the Machine Learning Algorithm used here Linear Regression (Supervised Learning). It predicts the output values based on the input data fed.

This algorithm builds a model based on the training data producedand predicts the new data.

B .Dataset

The RStudio is used to import the dataset. Dataset can be in excel or in CSV format. The dataset is reviewed and normalized. Normalization is changing the value of numeric columns of the dataset to common values and fit into a specific range.

The attributesof the dataset are shown in the figure.

Journal of Engineering Sciences

Rank - Ranking of overall sales
Name - The games name
Platform - Platform of the games release (i.e. PC,PS4, etc.)
Year - Year of the game's release
Genre - Genre of the game
Publisher - Publisher of the game
NA_Sales - Sales in North America (in millions)
EU_Sales - Sales in Europe (in millions)
JP_Sales - Sales in Japan (in millions)
Other_Sales - Sales in the rest of the world (in millions)
Global_Sales - Total worldwide sales.

Data cleaning

At this stage, by using RStudio we import dataset and remove redundant, missing, duplicate, and unnecessary data for further processing.

This stage is the most time-consuming stage in Data Science because to prevent wrongful prediction and get rid of the inconsistencies of data.

Data Exploration and Analysis

In this stage, we detect patterns, trends, and behavior in the data or dataset. This process makes further analysis easier because it excludes irrelevant data point and searches for no results data.

It uses visualization which makes it easy to analyze. From our analysis,we concluded Platform attribute has mainly affected the Video Game Sales in North America.

Platform vs NA_Sales (North America)

Machine Learning Algorithm

The ML algorithm is a logic that grasp one step ahead when exposed to more information/data.

When ML is exposed to training data it produces model.

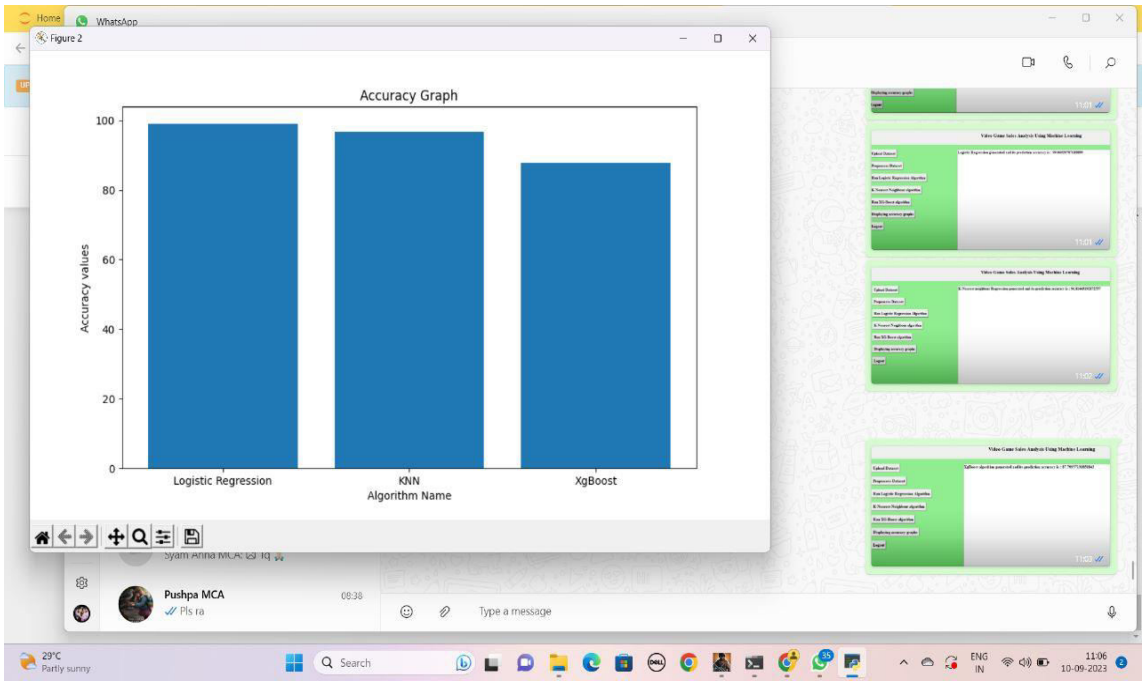
To build a Model, the Machine Learning Algorithm used here Linear Regression (Supervised Learning).

It predicts the output values based on the input data fed. This algorithm builds a model based on the training data producedand predicts the new data

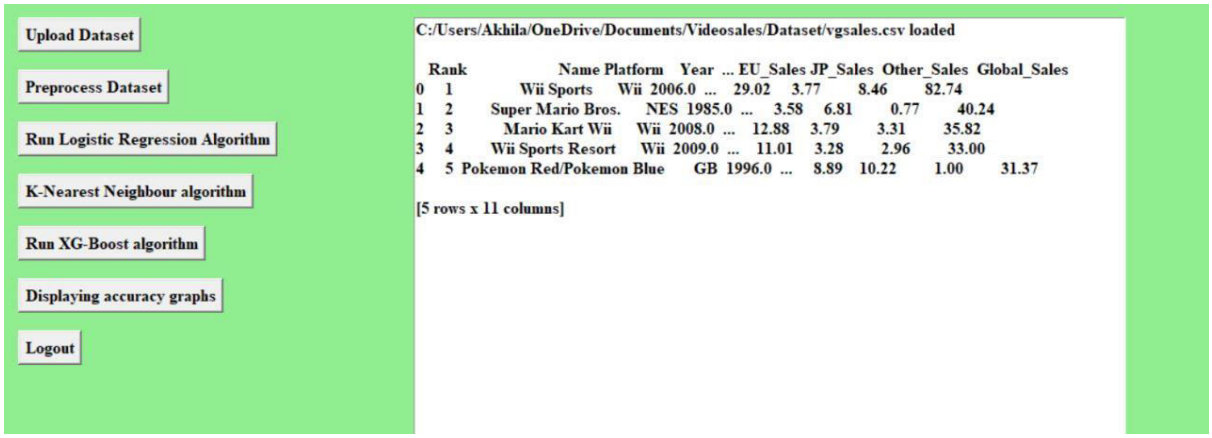
6. RESULTS AND DISCUSSION SCREEN SHOTS

HOME SCREEN

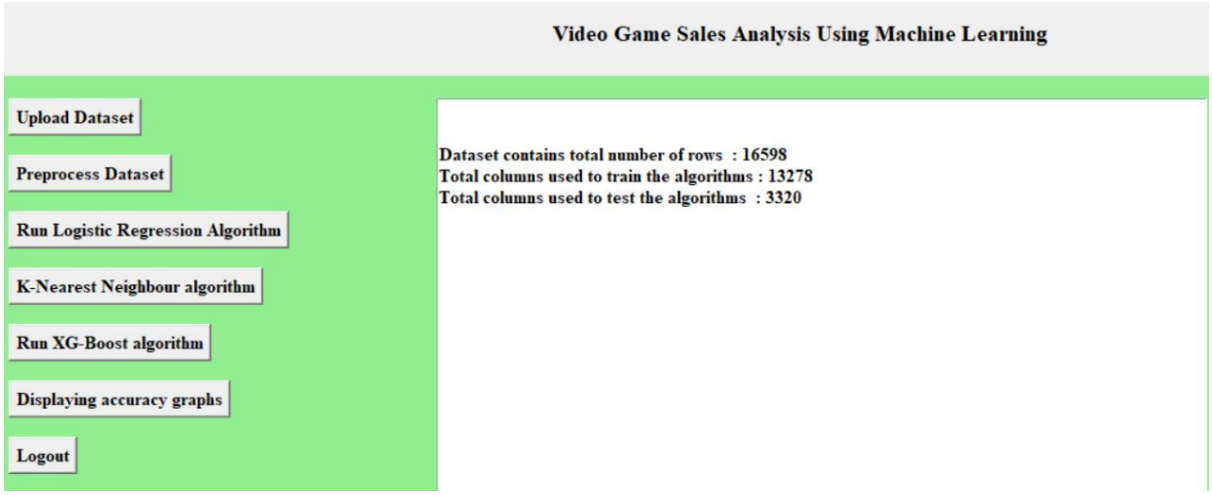
Journal of Engineering Sciences



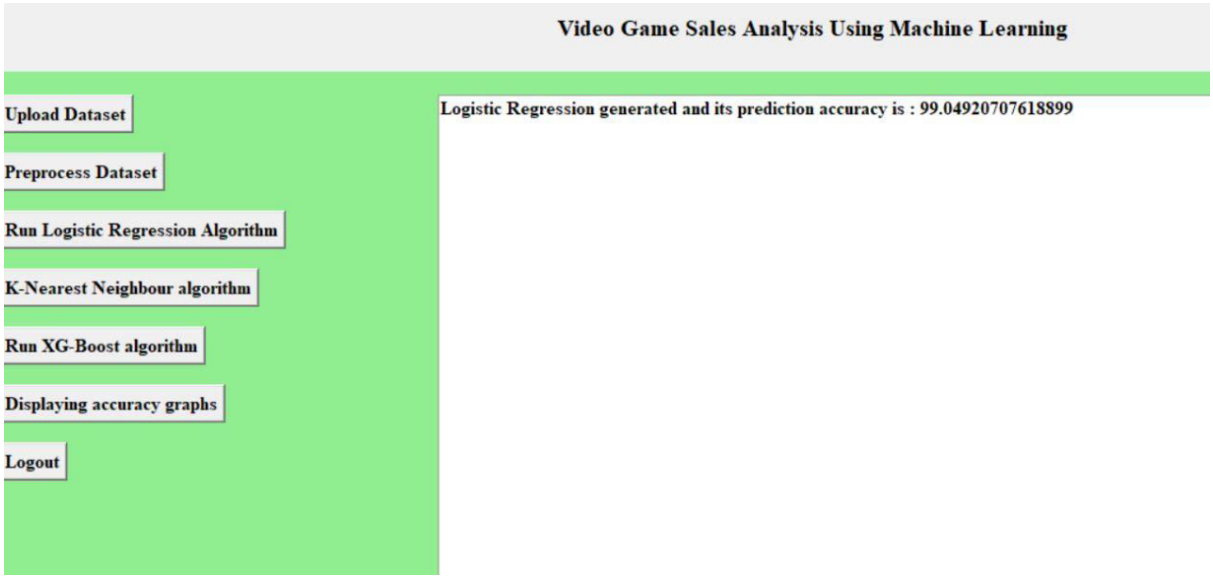
OUTPUT SCREEN 1



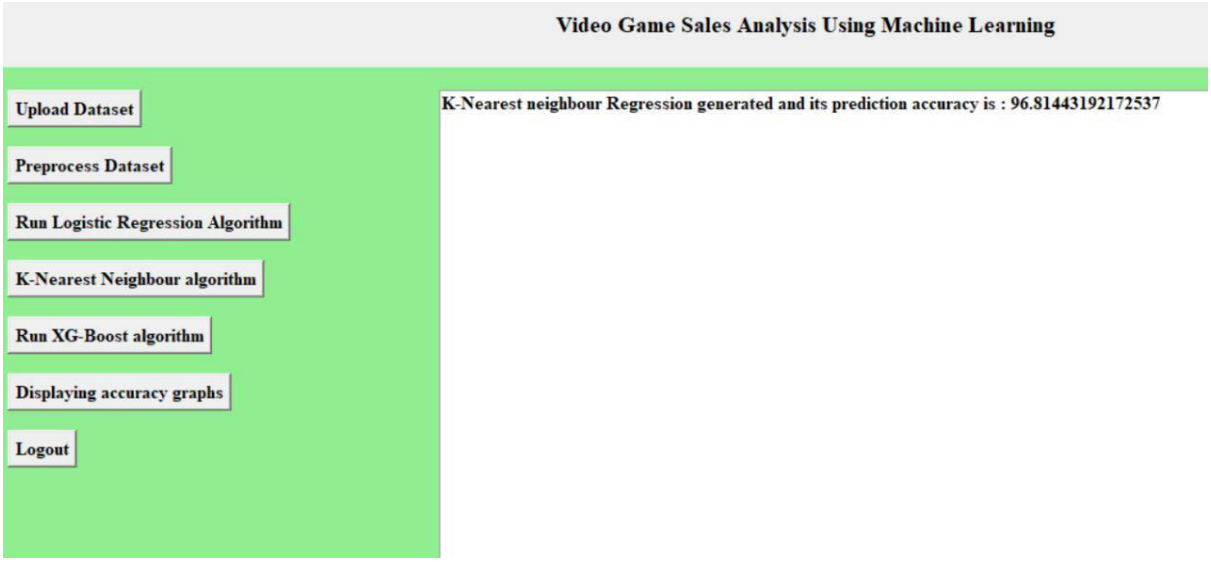
Journal of Engineering Sciences

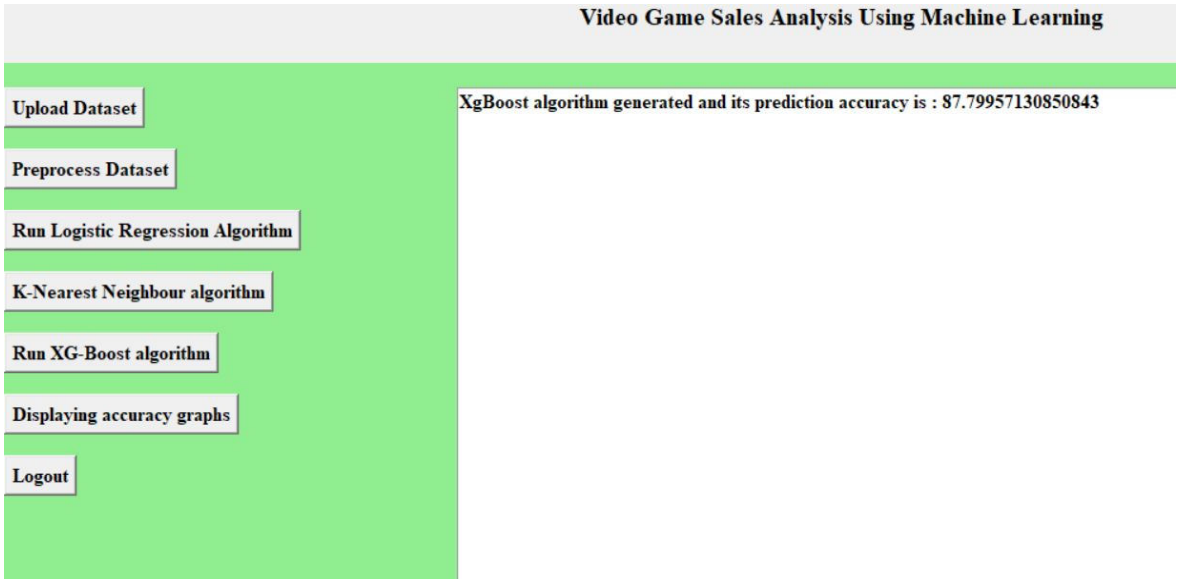


DESCRIPTION:



OUTPUT SCREEN 2





7. CONCLUSION AND FUTURE SCOPE

CONCLUSION

Sales prediction is a crucial part of the strategic planning process. It allows a company to forecast how the company will perform in the future. Predicting sales of a company is not only for planning new opportunities, but also allow knowing the negative trends that appear in the prediction. Finally we conclude that prediction of sales on video games has done and we observed which game has more sales in the market globally. For predicting sales of video games we applied several machine learning algorithms

(Linear regression, Random Forest, Decision tree Support vector regression). Among all these algorithms random forest gave us the best accurate result with minimum error rate.

FUTURE SCOPE

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

8. REFERENCES

[1] Julie Marcous and Sid-Ahmed Selouani, "A hybrid subspace-connectionist data mining approach for sales forecasting in video game sales industry", 2008, 978-0-7695-3507-4/08, IEEE.

[2] Hycinta Andrat and Nazneen Ansari, "Integrating data mining with computer games", 2016, ISBN:978- 1-5090-1666-2/16, IEEE.

[3] David Buckley, Ke Chen and Joshua Knowles, "Predicting skill from game play input to a first person shooter", 2013, 978-1-4673-5311-3/13, IEEE.

[4] Jing Zhang and Juan Li, "Retail Commodity Sale Forecast Model Based on Data Mining", 2016, 10.1109/INCoS.2016.42, IEEE.

[5] Vishal shrivastava, "A study of various clustering algorithms on retail sales data", 2012, Vol 1, ISSN 2319-2720.

[6] Akshay Krishna and Akhilesh V, "Sales – forecasting for retail stores using machine learning techniques", 2018, 10.1109/CSITSS.2018.8768765, IEEE.

[7] Paul Bertens, Anna Guitart, "Games and Big Data: A Scalable Multi-Dimensional Churn Prediction Model", 2017, 978-1-5386-3233-8/17, IEEE.