SALES PREDICTION ON VIDEO GAMES USING MACHINE LEARNING

[1]. BODDURU KEERTHANA, [2]. Dr. K.VENKATA RAO,

[1]. M.TECH SCHOLAR, [2]. PROFESSOR,

DEPARTMENT OF COMPUTER SCIENCE & SYSTEMS ENGINEERING,

ANDHRA UNIVERSITY, VISAKHAPATNAM.

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.

Keywords: sales prediction, extract knowledge, machine learning, performance measures, linear regression, support vector regression, random forest, decision tree.

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. Sales and market predictions are two different aspects which determine the client and

market demand respectively [3]. Sales prediction provides relevant information that can be used to make strategic business decisions. In the next sections we formulate the review of the related work, methodologies with detailed descriptions, comparison work, results and discussions. The paper ends with conclusion and future enhancement.

2. 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 instances.

Akshay Krishna and Akhilesh V, the authors proposed, "Sales – forecasting for retail stores using machine learning techniques",[6] this paper tries to predict the sales of a retail store using different machine learning techniques and tries to determine the best algorithm suited to a problem statement. They implemented normal regression techniques and as well as boosting techniques. Finally they conclude that boosting algorithm have better results than the regular regression algorithms.

Paul Bertens, Anna Guitart, the authors proposed, "Games and Big Data: A Scalable Multi-Dimensional Churn Prediction Model", [7] this article presents an approach to predict game churn based on survival ensembles. This method provides accurate predictions on both the level at which each player will leave the game and their accumulated playtime until the moment. This model is well suited to perform real time analyses of churners, even for games with millions of daily active users.

Gopalakrishnan T, Ritesh Choudhary and Sarada Prasad, the authors proposed, "Prediction of Sales Value in Online shopping using Linear Regression", [8] the aim of this paper is to analyze the sales of a big superstore, and predict future sales for helping them to increase their profits and make their brand even better and competitive as per the market trends by generating customer satisfaction as well. The technique used for prediction of sales is the Linear Regression Algorithm, which is a famous algorithm in the field of Machine Learning.

3. METHODOLOGY :

There are few steps can be performed for gathering data, analysis and modelling to get best predictions.

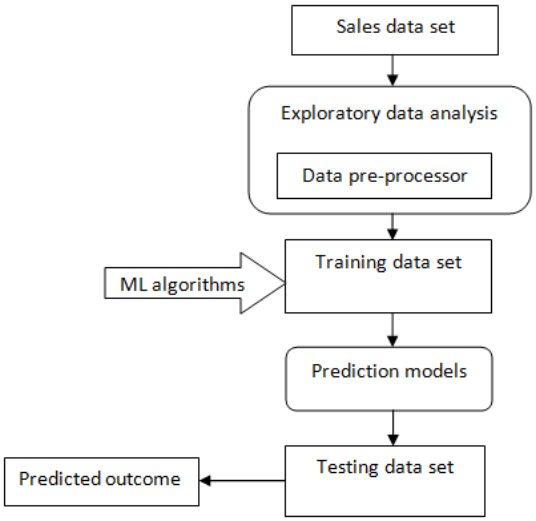

Fig.1. System architecture

3.1 DATASET DESCRIPTION:

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.

3.2 DATA PREPARATION AND ANALYSIS PROCEDURE:

- Statistically exploring the data- a check list:
  - Check data dimensions
  - Rows, columns and column names
  - Data types and unique values per column.
- Cleaning the data: -
  - Look for any missing data.
  - Identify and convert categorical values to numerical representation or convert numerical to categorical representation using dummy variables if suitable for modelling and check for distinct values in categorical columns.
- Statistically overview of data: -
  - Check head, tails of data to see complete required data loaded.
  - Identify numerical columns and look for insights like median, mean, mode etc.
  - Understand the relationship of columns and how they are effecting each other.
  - Check correlation and chi-square.

    Correlation - shows relation of numerical columns

    Chi-square - shows relation of categorical columns
- Graphical representation of data: -
  - Perform visualization on dataset attributes.

3.3 SPLITTING DATASET: -

After completion of exploratory data analysis we have to split the dataset into training and testing with split ratio 80 and 20. Training set contains 80% of data and test set contains 20% of data. In this project, we have used four machine learning algorithms such as linear regression, support vector regression, random forest and decision tree. These algorithms are comes under supervised learning. Now apply these algorithms on training set to train the model and perform predictions on testing data.

3.4 PREDICTIONS MODELS:

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 [10].

- **Linear regression**: Linear regression is commonly used for predictive modelling techniques [8]. The main theme of this algorithm is to find a mathematical equation for continuous variables Y when we have one or more X variables. This algorithm establishes a relation between two variables one variable is predicted variable and another one is result variable whose value is derived from the predictive variable.

$$Y = aX + b \quad \text{... Eq.4}$$

Function: model = lm (formula, data)

Where Y is result variable
X is predicted variable
a and b are coefficients

- **Support vector regression**: Support vector regression uses svm classification algorithm to forecast a continuous variable. But other regression models are used to minimize the error between predicted value and actual value [11]. SVR tries to fit best line among the predefined error value. Svr have few important key words such as **kernel, hyper plane, boundary line, support vector.**

Support vector regressions have two types

- Linear SVR - $\sum_{i=1}^{N}(\alpha_i - \alpha_i^*).(x_i, x) + b$ **... Eq.5**
- Non linear SVR - $\sum_{i=1}^{N}(\alpha_i - \alpha_i^*).K(x_i, x) + b$ **... Eq.6**
- Kernel function - polynomial:- $K(x_i, x_j) = (x_i.x_j)^d$ **... Eq.7**

- Gaussian radial basis function:-

$$K(x_{i,}x_j) = \exp\left[-\frac{\|x_i - x_j\|^2}{2\,\sigma^2}\right] \quad \text{...Eq.8}$$

- **Random Forest: -** Random Forest is a supervised machine learning algorithm creates randomly a forest with several trees [3]. Why we use random forest instead of decision tree, decision trees are easy to implement and work efficiently with training data, but it gives less accuracy this happens due to over fitting [12]. Over fitting occurs when a model trains the data to such an extent that is negatively impacts the performance of the model on new data. For this reason random forest comes into way.

  Function: train (formula, dataset, method="rf",trControl=trcontrol()) [where "rf" is random forest method].

- **Decision Tree: -** A decision tree is a type of supervised machine learning tree which explains about "what the input is and what is the relevant output according to the our data [13]". The main objective of this algorithm is to predict the value of a target variable. Mostly the decision tree rules are in the form of conditional statements i.e. "if-then-else". Decision trees are used for both classification and regression problems [4].

  Function: train(formula, dataset, method="rpart", trcontrol = trcontrol())

3.5 PREDICTION OUTCOME:

Prediction outcome will gives predicted values of a dataset after applying machine learning algorithms. Among all the algorithms the best algorithm with accuracy can be determined. In order to keep with the results random forest technique offers the best result among alternative algorithms.

4. RESULT AND DISCUSSIONS: -

For performing the quantitative analysis we have taken few methods, the performance metric value needed to be computed and they are to be compared with the other. Hence, for performing the calculations of the performance metric there are a few formulas which can be utilized for achieving the performance value from the dataset. The formulae for the calculation of the performance metrics are given below in table.

Table.1. quantitative analysis

METRIC	FORMULA
Accuracy	(TP+TN) / (TP+TN+FP=FN)
Error Rate	100-Accuracy
Mean Absolute Error	$\frac{1}{n}\sum_{j=1}^{n}\lvert y_j - y\hat{}_j\rvert$
Root Mean Square Error	$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - y\hat{}_j)^2}$
Precision	TP/(TP+FP)
F-value	2/(1/P+1/R)

Table.2. Accuracy and RMSE values

S.NO	ALGORITHM	RMSE VALUE	ACCURACY
1	Random Forest	1.4648	0.9605
2	Support vector regression	1.9773	0.8154
3	Decision Tree	2.8762	0.8036
4	Linear regression	2.4830	0.5734

FINAL RESULT:

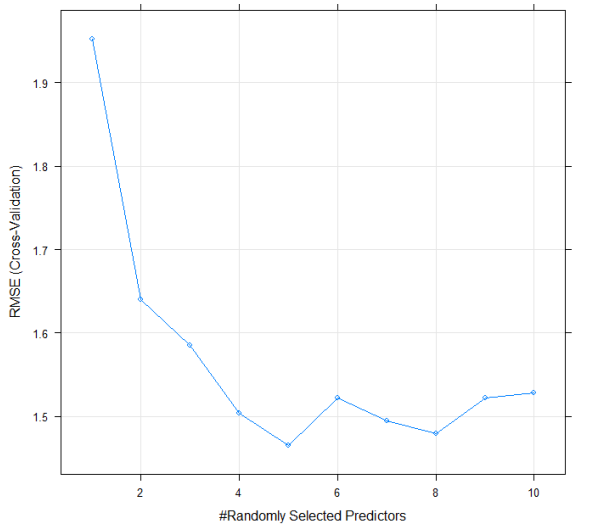

Fig.2. Random forest output graph

5. MODEL COMPARISON:

In this section, we contrast the predictive effects of linear regression, support vector regression, random forest and decision tree models. We choose linear regression as the baseline model and introduce its prediction

result into the result comparison. The baseline model (linear regression)'s accuracy result, which is 75%, is not ideal. Each of the models we select performs better than the baseline model. Random forest, Decision tree and support vector regression are 96%, 80%, 85% respectively. It is observed that random forest model performs the best among them, with the result of 96%. Compared with the other prediction models random forest model plays a prominent role in improving video game sales prediction accuracy in the field of large-scale product sales data.

6. 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.

7. 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.

[8] Gopalakrishnan T, Ritesh Choudhary and Sarada Prasad, "Prediction of Sales Value in Online shopping using Linear Regression", 2018, 10.1109/CCAA.2018.8777620, IEEE.

[9] N. Ansari, M. Talreja and V. Desai, ―Data Mining in Online Social Games,‖ In Proceedings of International Conference on Advances in Computing, pp. 801-805. Springer India, 2012.

[10] A. Alfons. cvTools: Cross-validation tools for regression models, 2012. R package, version 0.3.2.

[11] Mahdevari S, Shahriar K, Yagiz S, et al. A support vector regression model for predicting tunnel boring machine penetration rates[J]. International Journal of Rock Mechanics and Mining Sciences, 2014, 72: 214-229.

[12] P. Boinee, A. D. Angelis, and G. Foresti, "Meta random forests," International Journal of Computational Intelligence, vol. 2, no. 3, pp. 138-147, 2005.

[13] T. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, 1998.