



VIDEO GAMES SALES PREDICTION USING ML

Dr.K.Rama, IQAC Coordinator, Loyola Academy, ramakovur@gmail.com

Abstract: Playing video games for many years has led to a large volume of gaming data that consist of gamer likes and their playing behavior. 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 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 datasets to enhance accuracy and efficiency in predictions. In this paper, we briefly analyzed 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.

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 [1] followers by using historical sales data. This study involves extracting the video game sales data and analyzing which game has more sales globally when compared to other countries. 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. 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 per-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 datasets 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. 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 objectives of this research work is to find the trending sales by using machine learning algorithms. Sales prediction [2-6] is an essential part of business organizations. It provides



relevant information that can be used to make strategic business decisions. 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. 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.

METHODOLOGY

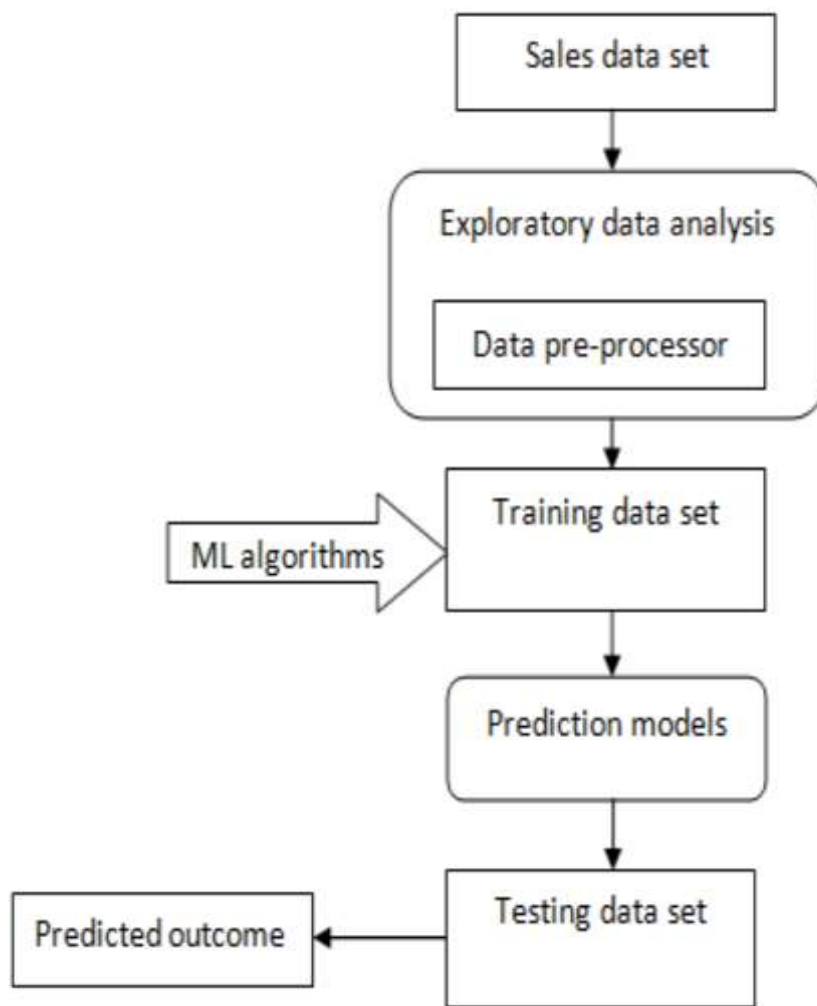


Fig.1. System architecture

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.

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 modeling 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 affecting 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 datasets attributes.

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.

PREDICTIONS MODELS:

Apply various possible prediction modeling 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. □

Linear regression:

Linear regression is commonly used for predictive modelling techniques. 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.

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. 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.

Random Forest:

Random Forest is a supervised machine learning algorithm [7] creates randomly a forest with several trees. 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. 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.

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. 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.



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.

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 datasets. The formulae for the calculation of the performance metrics are given below in table.

FINAL RESULT:

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 y_j - y^{\wedge}_j $
Root Mean Square Error	$\sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - y^{\wedge}_j)^2}$
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

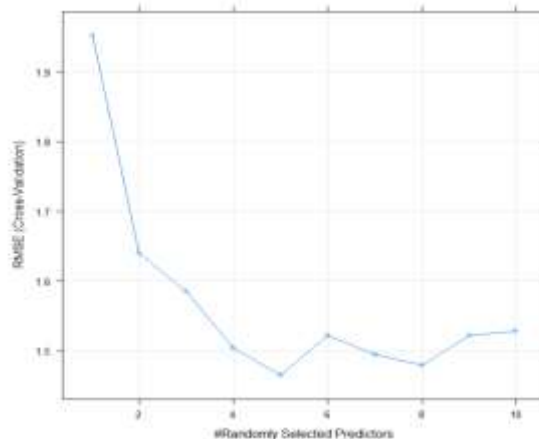


Fig.2. Random forest output graph

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 predictionresult 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.

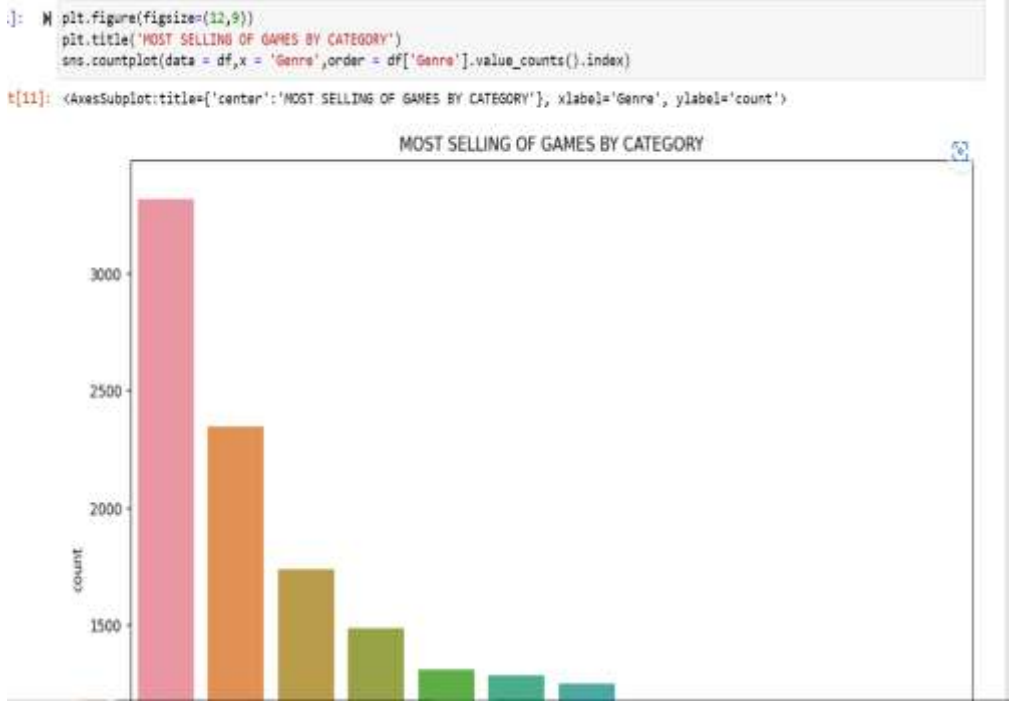


Fig.3. Sales Prediction of Games

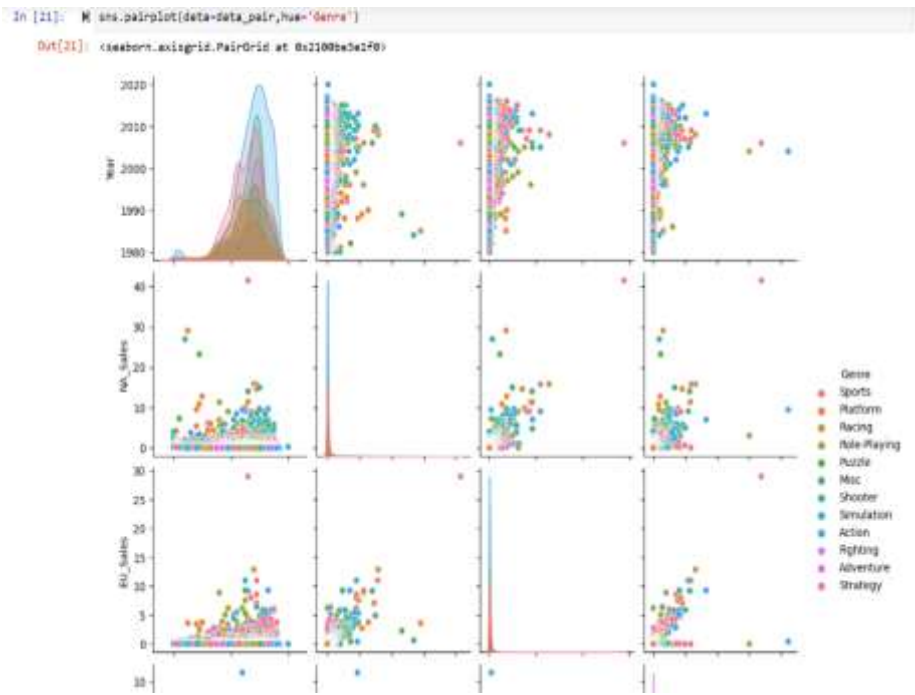


Fig.4. Graphical visual



Fig.5. Standard deviation

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.

It helps in analyzing the real-time stock prices in the market with best accurate results. 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. To keep with the results random forest technique offers the best result among alternative algorithms. This project can be implemented as a real time project and will get the datasets through online in this future work.

REFERENCES

1. M. Á. López-Cabarcos, D. Ribeiro-Soriano and J. Piñeiro-Chousa, "All that glitters is not gold. the rise of gaming in the covid-19 pandemic", *Journal of Innovation & Knowledge*, vol. 5, no. 4, pp. 289-296, 2020.
2. P. Meulstee and M. Pechenizkiy, "Food sales prediction:“ if only it knew what we know", *2008 IEEE International Conference on Data Mining Workshops*, pp. 134-143, 2008.
3. M. Li, S. Ji and G. Liu, "Forecasting of chinese e-commerce sales: an empirical comparison of arima nonlinear autoregressive neural network and a combined arima-narnn model",



Mathematical Problems in Engineering, vol. 2018, 2018.

4. M. Gao, W. Xu, H. Fu, M. Wang and X. Liang, "A novel forecasting method for large-scale sales prediction using extreme learning machine", *2014 Seventh International Joint Conference on Computational Sciences and Optimization*, pp. 602-606, 2014.
5. J. Marcoux and S.-A. Selouani, "A hybrid subspace-connectionist data mining approach for sales forecasting in the video game industry", *2009 WRI World Congress on Computer Science and Information Engineering*, vol. 5, pp. 666-670, 2009.
6. P. Ramos, N. Santos and R. Rebelo, "Performance of state space and arima models for consumer retail sales forecasting", *Robotics and computer-integrated manufacturing*, vol. 34, pp. 151-163, 2015.
7. B. Pavlyshenko, "Using stacking approaches for machine learning models", *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, pp. 255-258, 2018.