Prediction of video game Sales based on linear regression and random forest machine learning algorithm.

Name: Zul Absar Ali

Student ID: 221056932

ECS784U/P – Data Analytics

1. Introduction

The very aim of this project is to make sales predication for the gaming industry. Before presenting the report, here is a brief introduction about the video games. From the past few decades video games saw a huge spike and became very popular. At first it was popular among the teenager but slowly people from every age group seems enjoy playing video game. There was arcade games from late 70s but the origin of 3d games starts from mid 90s the most popular consoles was Nintendo 64 and PlayStation 1. The first game which was 3D was battleship (Jordan, 2022). Since then video game industry gained popularity. Due to covid 19 the grow of video game grew even more which made this more lucrative industry. The overall income from video games (excluding esports) was US$214.2 billion in 2021, and it is expected to increase by 8.4% CAGR to US$321.1 billion in 2026. (PricewaterhouseCoopers, 2022).

As this industry have a huge potential companies started to invest money however the at the beginning this companies had limited insights and feedback from the costumer there prediction was hugely dependent on sales. To make AA(inde game category)medium size budget video game a company puts almost 250,000$ to develop the game (Startup Info, 2023) so the firms want a rock solid insight before investing the money. this problem can be solve by using machine learning predication so the aim of this project is to report a comprehensive report to prediction the sales of a video game. Such analysis approach is necessary for a company to be in the market and have a extra edge as compared to their competitor. A excellent report includes the popular genre, hype of the game, market surreys and platform to release the game etc.

2. Literature review:.

The paper investigates the relationship between independent variables and global sales of video games, focusing on the number of critics rating the game, their average score, and non-intuitive results related to user ratings. (Quest journal 2019).

To increase sales in video games, a recommendation system can be used to suggest items that match players' interests. This can be achieved through machine learning algorithms that predict the rating of a product by a user. The paper evaluates and compares two such algorithms, extremely randomized trees and deep neural networks, which show potential as effective video game recommendation engines (Paul Bertens et.al, 2018)

This study analyse video game achievement data to identify patterns in game completion rates and their correlation with external factors beyond game length. The findings can help project managers and product owners make informed decisions regarding project scope, potentially reducing game budgets and enhancing production efficiency.
(Eric Bailey et.al, 2019)

According to the study, using convolutional neural networks is the most effective method for predicting the value of players and identifying high-spending individuals, known as "whales." By detecting these valuable players early, developers can work to keep them engaged and maximize their spending, ultimately resulting in increased revenue.
(Tirath prasad sahu et.al, 2016)

paper proposes a novel method for predicting sales of video games using connectionist and subspace decomposition methods. The approach employs neural networks trained with back-propagation algorithm to forecast weekly sales while considering various influencing factors. To evaluate the significance of these factors, Principal Component Analysis is employed. The system's performance is evaluated and compared against baseline sales, and the results are presented and analyzed in terms of prediction accuracy.
(Sid-Ahmed Selouani et.al, 2009)

3. Data processing

3.1Data source:

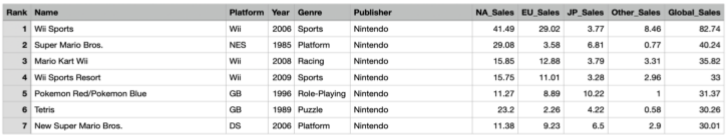
Data source : Kaggle

Link: https://www.kaggle.com/datasets/gregorut/videogamesales

The data is about video game sales which consists of 11 columns such as ranking, platform, year, genre, publisher, Na_sales (north America sales), Eu_sales (Europe sales), Jp_sales (Japan sales), other sales and the global sales. Almost 16,500 video game titles from various platforms and geographical areas are represented in the dataset's sales statistics.

This dataset is useful for analyse the popular genre or for video games. It can also be helpful to target a specific platform for the game like(PlayStation, Xbox) or to predict sales for particular areas mentioned in the file or else to prediction the global sales.

3.2 Pre-process data :



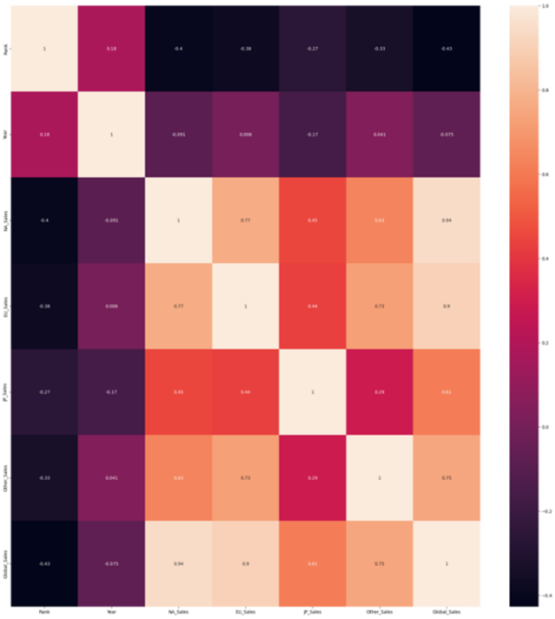
Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37
6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01

Rank	Ranking
Name	Video game name
Platform	Console name
year	Game released
genre	Game category
Publisher	Game developer
Na sales	north America sales
Eu sales	Europe sales

Jp sales	Japan sales
Other sales	Sales of other countries
Global sales	World-wide sales

3.3 Bivariate analysis**:**

The analysis of two variables is referred to as bivariate analysis. As "bi" signifies "two," it is easy to remember. To comprehend the link between two variables, bivariate analysis is used. (Zack 2021).



Fig(1)

In this part we used heatmap to check the correlation between the variables and it shows that there are 4 strong correlation in the dataset It means that it will be challenging to establish their relationship since modifications to one will affect the other.

3.4 dropping the columns :

The exploration highlights that few columns in the data set was not that useful for the machine learning model as a result dropping those columns with null vales will help to get more precise data and learning accuracy. The columns was ‘years’ and ‘ Publisher’

3.5 Outliners:

Outliers are datapoints in a dataset that contain atypical findings with the usual observations. These values can provide unusual accuracy ratings, distort measurements, and hide the true results from the user. Outliers can also be the result of anomalies and errors in the dataset.
(Nichani, 2020)

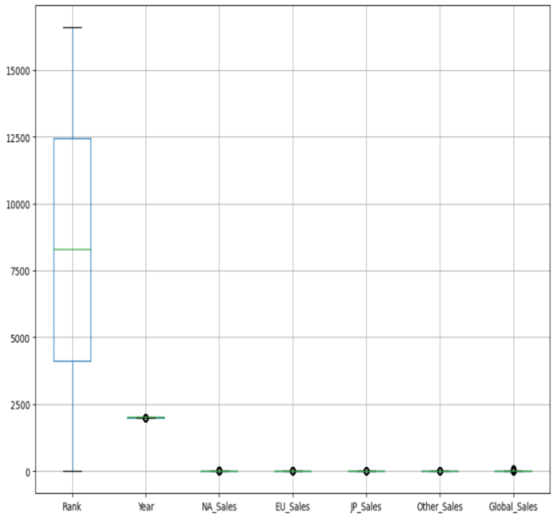


Fig (2)

At first we separated the numerical from the dataset and we got (rank, year, Na_sales, Eu_sales, Jp_sales, other_sales, global sale)
Then performed the IQR function for each numerical column. (Jiawei Yang et.al)

3.6 Converting categorical data into numerical data :

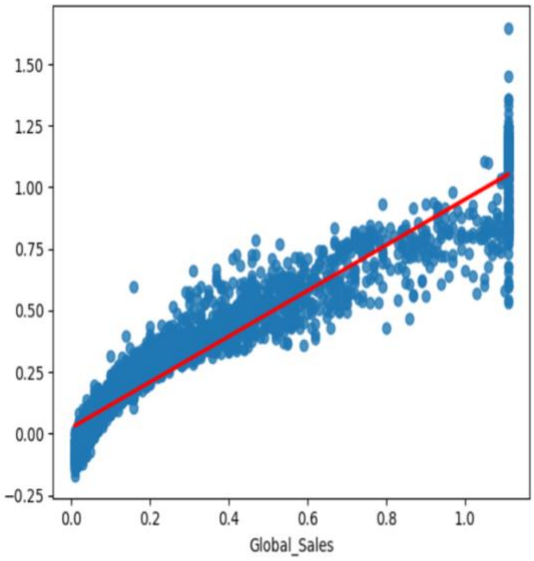
Algorithms for machine learning are created to operate on numerical data, not categorical data. As a result, before using machine learning techniques, categorical data must be transformed into numerical data.

4. Machine learning :

By analysing the available data and optimising a performance criterion that is based on the nature of the problem, machine learning techniques may be used to automatically create a computer model of these complicated interactions. Training is the automated process of constructing a model, while training data is the information utilised for training. The trained model can offer fresh perspectives on how input variables are translated into outputs and be used to create predictions for unique input values that weren't included in the training set of data. (Yalin Baştanlar)

4.1. Linear regression:

Linear regression is a machine learning model that predicts a target variable through one or more independent values, this method is used to determine the relationship between variables and their outcomes. (Xiaogang Su et,al.2012)



Fig(3)

the above fig is a output of linear regression. A measurement of the average absolute difference between actual and expected values is the mean absolute error (MAE). In the above output the (MAE) is 0.071 which seems to be a good number, MAE explores the difference between predictions made and the actual values in the dataset, therefore the result implies that there is little difference in the predictions being made. Talking about the mean square error (MSE) is 0.0098 which is relativity low and suggest that the model is a good fit with the data.

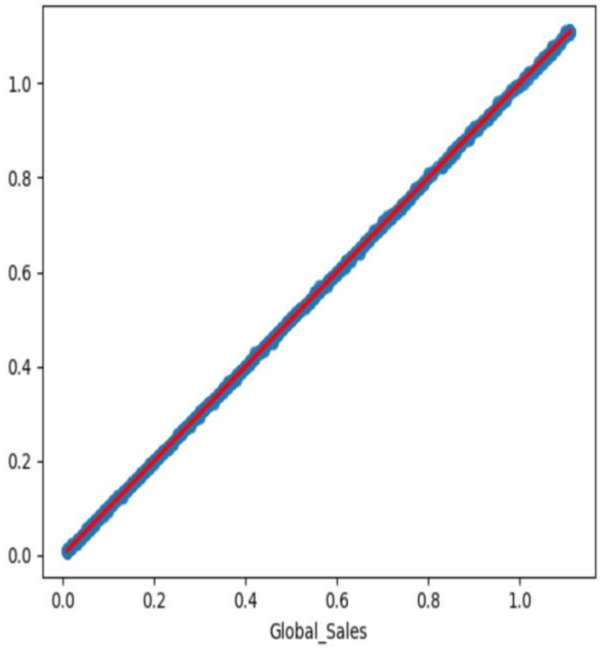
Now coming to R square the value of R square is 0.925 this is quite high and shows that a sizable proportion of the variance in the target variable is explained by the mode

In general, based on the outputs of (MSE),(MAE) and (R square) metrics, the model appears to have done well in predicting the target variable.

4.2 K Neighbour Classifier:

K Neighbour Classifier uses proximity of datapoints to make predictions and classifications about the groupings of the datapoint, it works on the assumption of similarity, where datapoints located next or nearer to each other are similar. (Kieran Greer et.al,2003)

Fig



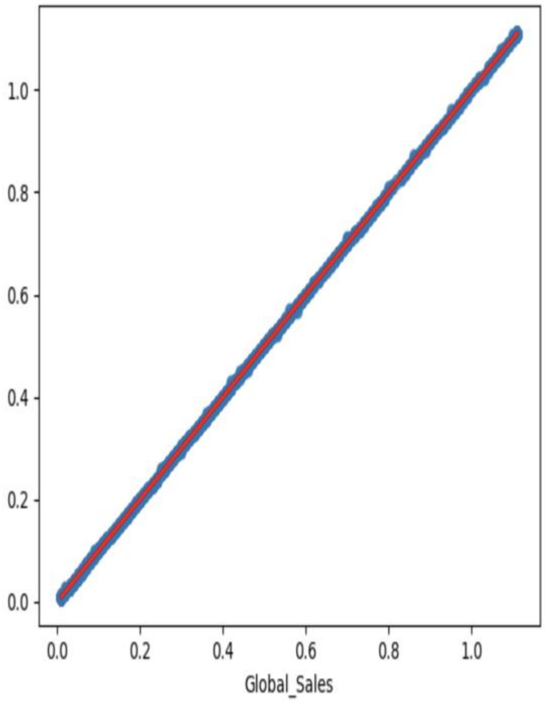
fig(4)

The KNN model scored a mean absolute error of 0.0000242 which is lower than the linear regression model, which implies that the predictions made by this model are closer to the actual values in the dataset. Also the KNN model scored a mean squared error of 1.11 , the MSE measures the average squared distance between the predicted value and actual value, the KNN scores higher in this parameter, which implies that Linear regression is more precise. The KNN scores a R square value of 0.99, which is good because it demonstrates that there is good variance between the variables in the model.

4.3 Random Forest Regression

Random Forest Regression is a machine learning model that is a combination of decision tree algorithms to make predictions, the Random Forest Regression takes the average of the decision trees . The one of the key advantages of Random Forest Regression is that its well known for its effective handling of large datasets as well as its accuracy.

In contrast to all the machine learning models the random forest regression model has scored the highest mean absolute error score of 8.78, which means that there is a greater difference between the predicted value and the actual value of the dataset, also in comparison to the other models the Random Forest scored the highest Mean Squared Error, which implies that the predicted value and actual value have the greatest average squared distance compared to Linear Regression and Decision Trees. However the model scored a r square value of 0.9 indicating good fit and good amount of variance.



Fig(5)

5. Concluding Remarks and limitation:

After selection the dataset. There has been some extensive data exploration done. we can see the use Bivariate analysis to find the correlation between the variable. Outliner approached was also used to clear the datapoint which was off the limit. Conversion of categorical data to numerical data was important for the machine learning process. After completing the data exploration and data cleaning stage we did the machine learning part for this we used 3 machine learning techniques 1. Linear regression 2. KKN 3. Random forest

The random forest model and KNN model appear to have pretty low MAE and MSE and high R Square values based on the results given, demonstrating that they may be suitable options for generating reliable predictions on this dataset. Even though it was still performing good, the linear regression model had greater MAE and MSE and a lower R Square value than the other two models.

Now moving towards limitation, there has been some limitation When It comes to data set because the initial rows and columns was 11658 rows and 11 columns and the dataset had few data which where irreverent and cause cause complication. Another drawback of the dataset comes forms the platform column the reason behind it is few platform are become outdated

And releasing any games on such platform will bad approach. From past few years we have saw a great spike in the digital platform where user can purchase the game or any special collectable using their id unfortunately there is no data for online purchased made .

References :

Bertens, P. *et al.* (1970) *[PDF] a machine-learning item recommendation system for video games: Semantic scholar*, *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. Available at: https://www.semanticscholar.org/paper/A-Machine-Learning-Item-Recommendation-System-for-Bertens-Guitart/80a3e1d693adec7c7d094abed0e7f818f8cb5dd9 (Accessed: March 15, 2023).

Marcoux, J. *et al.* (2009) *A hybrid subspace-connectionist data mining approach for sales forecasting in the video game industry: Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering - volume 05*, *Guide Proceedings*. Available at: https://dl.acm.org/doi/10.1109/CSIE.2009.1001 (Accessed: March 15, 2023).

PRASAD SAHU, T.I.R.A.T.H. (2016) *https://sci-hub.ru/10.1109/BigData.2018.8622151*. Available at: https://sci-hub.ru/10.1109/BigData.2018.8622151 (Accessed: September 28, 2016).

Predicting global video-game sales - quest journals (no date). Available at: https://www.questjournals.org/jrbm/papers/vol7-issue3/I07036064.pdf (Accessed: March 15, 2023).

Sci-hub | | 10.1016/j.entcom.2021.100456 (no date). Available at: https://sci-hub.se/10.1016/j.entcom.2021.100456 (Accessed: March 15, 2023).

Jordan, J. (2022) *What's the first 3D game in the world?, NarraSoft*. Available at: https://narrasoft.com/what-is-the-first-3d-game-in-the-world/ (Accessed: March 11, 2023).

Linear regression - su - 2012 - wiley online library (no date). Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1198 (Accessed: March 15, 2023).

Müller, A.C. and Guido, S. (2018) “Introduction,” in *Introduction to machine learning with python: A guide for data scientists*. Sebastopol: O'Reilly Media.

Nichani, P. (2020) *Outliers in machine learning, Medium*. Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/outliers-in-machine-learning-e830b2bd8660 (Accessed: March 13, 2023).

PricewaterhouseCoopers (no date) *Global Entertainment & Media Outlook 2022–2026 perspectives report, PwC*. Available at: https://www.pwc.com/gx/en/industries/tmt/media/outlook/outlook-perspectives.html (Accessed: March 11, 2023).

Sci-Hub | Deep Learning Decoding of mental state in non-invasive brain ... (no date). Available at: https://sci-hub.ru/10.1145/3371425.3371441 (Accessed: March 15, 2023).

Sci-Hub | Introduction to machine learning. methods in molecular ... (no date). Available at: https://sci-hub.ru/10.1007/978-1-62703-748-8_7 (Accessed: March 15, 2023).

Sci-Hub | KNN model-based approach in classification. lecture notes in ... (no date). Available at: https://sci-hub.ru/10.1007/978-3-540-39964-3_62 (Accessed: March 15, 2023).

Startup Info (2023) *How much does it cost to develop a game?, Startup Info*. Available at: https://startup.info/how-much-does-it-cost-to-develop-a-game/ (Accessed: March 11, 2023).

Zach (2021) *A quick introduction to bivariate analysis, Statology*. Available at: https://www.statology.org/bivariate-analysis/ (Accessed: March 13, 2023).