CSE_A18_2

by Cse_a18_2 Cse_a18_2

Submission date: 08-Jul-2021 04:30PM (UTC+0530)
Submission ID: 1617106382
File name: CSE_A18_Final_Documentation.docx (2.15M)
Word count: 5281
Character count: 32590

A Mini Project with Seminar On

VIDEO GAME SALES PREDICTION

Submitted in partial fulfillment of the requirements for the award of the

Bachelor of Technology

In

Department of Computer Science and Engineering

By

M SAI KIRAN	**18241A0527**
M S KUSHAL RAJ	**18241A0526**
P V S KARTHIKEYA	**17241A05S4**
GOLKONDA MAHESH	**18241A0511**

Under the Esteemed guidance of

Dr S Kranthi Kumar,

Professor

Department of Computer Science and Engineering

GOKARAJU RANGARAJU INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Autonomous)

Bachupally, Kukatpally, Hyderabad-500090

GOKARAJU RANGARAJU INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Autonomous)

Bachupally, Kukatpally, Hyderabad-500090

CERTIFICATE

This is to certify that the major project entitled "**Video Game Sales Prediction**" is submitted by **M Sai Kiran(18241A0527), M S Kushal Raj(18241A0526), P V S Karthikeya(17241A05S4), Golkonda Mahesh(18241A0511)** in partial fulfillment of the award of degree in BACHELOR OF TECHNOLOGY in Computer Science and Engineering during academic year 2020-2021.

INTERNAL GUIDE	HEAD OF THE DEPARTMENT
Dr S Kranthi Kumar	**Dr. K. MADHAVI**
Professor	Professor

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

There are many people who helped us directly and indirectly to complete our project successfully. We would like to take this opportunity to thank one and all. First, we would like to express our deep gratitude towards our internal guide **Dr S Kranthi Kumar, Prof.** Department of CSE for his support in the completion of our dissertation. We wish to express our sincere thanks to **Dr. K. Madhavi, HOD, Department of CSE** and to our principal **Dr. J. Praveen** for providing the facilities to complete the dissertation. We would like to thank all our faculty and friends for their help and constructive criticism during the project period. Finally, we are very much indebted to our parents for their moral support and encouragement to achieve goals.

M Sai Kiran(18241A0527)

M S Kushal Raj(18241A0526)

P V S Karthikeya(17241A05S4)

Golkonda Mahesh(18241A0511)

DECLARATION

We hereby declare that the industrial major project entitled "**Video Game Sales Prediction**" is the work done during the period from **8th March 2021 to 8th July 2021** and is submitted in the partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering from Gokaraju Rangaraju Institute of Engineering and Technology (Autonomous under Jawaharlal Nehru Technology University, Hyderabad).The results embodied in this project have not been submitted to any other university or Institution for the award of any degree or diploma.

M Sai Kiran(18241A0527)

M S Kushal Raj(18241A0526)

P V S Karthikeya(17241A05S4)

Golkonda Mahesh(18241A0511)

ABSTRACT

Previous video games success prediction is done assumed the game was already released and made predictions based on that knowledge. In addition to that, they focus mostly on pre-2010 console games when there was no incentive to study the PC games market.

Around 10,000 to 12,000 games are being developed and are being uploaded in steam platform. Huge number of games are being developed every year, for relaxation or excitement to people around the world. Out of the complete populace on the planet today, around 40% of them playing computer games. As per IDC information worldwide computer game income is relied upon to expand 20% to $179.7 billion by 2021 making the computer game industry a greater moneymaker than the worldwide film and North American games businesses joined. We will be fostering a model that can be utilized by game makers to give information for creating games, which will drift and producing more income.

In this project we use Random Forest Algorithm and Linear Regression Algorithm for the better accuracy as compared to existing video game sales predictions. Random forest is the most used prediction technique in supervised ML(it predicts categorical and continuous outcome). Linear Regression is one of the supervised ML algorithm where the predicted output is nonstop and has a steady slant. It's utilized to anticipate values within a continuous range. Logistic Regression produces discrete/categorical output.

We will be using the Video Game Sales Dataset from Kaggle which is a trusted source and which is free to use. In this project we can implement a model by following features 1. collecting data and importing the suitable libraries to the platform called Jupyter Notebook(anaconda3) using Python language, 2.Analysing data by graphical representation such as Heatmap and Histograms, 3.Data wrangling is used to clean the data by removing unwanted columns from the dataset, 4. We construct model on the train information and foresee the yield on the test information, 5.Accuracy check : we can find the actual and predicted values by using existing methods called RMSE and MAE. Results of Random Forest and Linear regression algorithms, we collect the outputs of and find the averages for better accuracy of video game sales prediction. Results of Logistic Regression, we calculate the hit rate of the test dataset.

Video Game Sales Prediction is helpful in predicting the revenue of games. This method is helpful for several companies which are planning in developing Video Games which will be top-grossing.

TABLE OF CONTENTS

Chapter-1
INTRODUCTION

The PC games industry has a seen an increase in sales and the number of releases after Valve launched its Steam Store. This Steam store saw a major growth after 2012. Greenlight is a program that allows the developers to relatively easily release their games on Steam without a publisher which had been very difficult until then.

The gaming market is certainly one of the upcoming industries of the modern age and one of those that are most influenced by the advancement in technology. With the different technologies like AR/VR in consumer products like gaming consoles and even smartphones, the gaming sector shows great value. In this project, we use machine learning models to predict the sales of video games depending on given factors.

Industries based on video games needs the forecast of revenue in an outstanding business sector development. Over the most recent 10 years in the USA the pay coming from PC and computer games expanded a ton. Thus, we need to foresee the nature of video games that can generate more revenue by utilizing previous revenue data. This examination includes collection of revenue data of video games and performing analysis on the data to find out the video game which has more deals universally when contrasted with different nations. In this we utilized diverse ML methods to foresee the video game revenue. The developed method is helpful for different companies which are keen on foreseeing the game deals.

The goal of this is to find out the factors which are affecting the success of video games, collect the data of PC games as no suitable one is available and, finally, and predict a game's success based on different descriptive information such as genre, price, developer, or game features. Video games are no more a specialized section of population product having little portion of customers. Exploration has shown the that computer game customers are an altogether different gathering and in varying backgrounds. The gaming business has affected for all intents and purposes an immense customer section, enormous measure of the current populace has grown up with computer games and played them for amusement and schooling from various perspectives.

Goal

A significant part of the current work is engaged in two significant ways:

- Foreseeing the global sales of the game.
- Predicting if the game is hit.

Objective

The plan of our work is to:

- Analysis of video game sales
- Predicting the global sales of the game
- Predicting if the game is hit.

Methodology

Machine learning

Machine Learning is a sort of AI which permits programming applications to deliver more precision at expectation of results without performing express programming to do as such. AI calculations utilize past chronicled information as contribution to it and it gives expectation of new yield esteems.

Machine Learning is prevailing on the grounds that it's anything but a perspective on patterns in client conduct and business functional examples, and furthermore it helps being developed of new items. Large numbers of the present various organizations, like LinkedIn, Google and Uber, make ML a principle part of their tasks. It's anything but a huge serious differentiator for some organizations on the lookout. The term Machine Learning alludes to the computerized identification of various significant examples in information. Machine Learning is a technique for examination of information that mechanizes scientific model structure. In the new years and years, Machine Learning has become a typical instrument in practically any undertaking that requires data assortment from enormous informational collections. We are circled by an ML based innovation: web crawlers apparatuses figure out how to present to us the exact outcomes, against spam programming figures out how to separate our email messages, and online installment exchanges are ensured by a product that figures out how to recognize fakes. Advanced cameras figure out how to perceive faces and perform insightful individual help applications on PDAs figure out how to perceive voice orders. Engine vehicles are furnished with mishap anticipation frameworks that are fabricated utilizing diverse ML calculations.

Classical ML is utilized to arranged by how a Machine Learning calculation figures out how to turn out to be more exact in its expectations. The sort of Machine Learning calculation utilized by information researchers relies upon what kind of information they need to anticipate. One normal property of these calculations is that, in various to more customary employments of PCs, in these cases, because of the diverse intricacy of the examples that should be determined, a developer can't give an unequivocal, fine point by point expressing of how such assignments ought to be executed. Machine Learning calculations are worried about giving projects the capacity to learn and adjust.

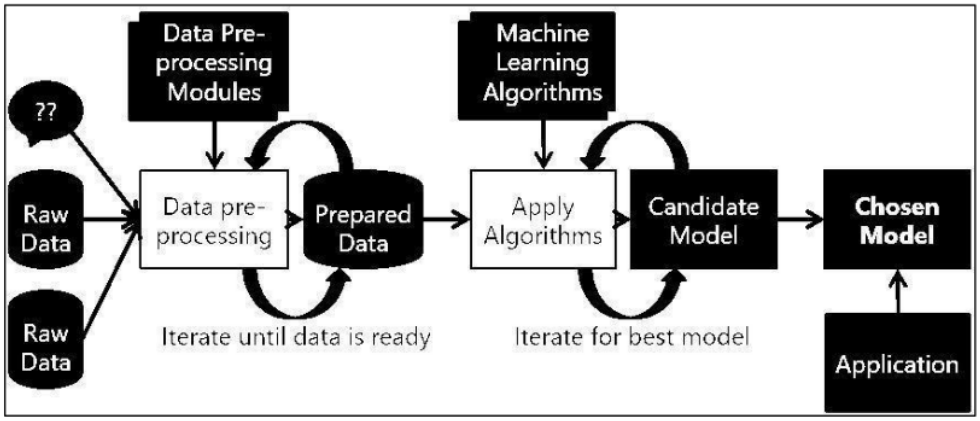


Fig 1.1-ML Diagram

The inputs to our algorithms are

- Critic_Count
- User_Score
- User_Count
- Rating

The output is prediction of Global Sales that is probably going to have happened. We evaluate various calculations, such Linear Regression, Logistic Regression and Random Forests.

Our Dataset

Kaggle website contains the dataset which is utilized by us which is from steam platform. However, the dataset has numerous invalid qualities and to perform Machine Learning this information can't be utilized all things considered. Subsequently the information should be prepared

Attributes of our data

- Name
- Platform
- Year of Release
- Genre
- Publisher
- NA_Sales
- EU_Sales
- JP_Sales
- Other_Sales
- Global_Sales
- Critic_Score
- Critic_Count
- User_Score
- User_Count
- Rating

Preprocessing

Before we create the modules, we need to remove unwanted date from the data which we collected and then we build the model.

1.Importing required libraries:

We have imported NumPy, pandas, seaborn, sklearn, matplotlib libraries.

2.Importing the dataset:

Downloaded Kaggle dataset has been imported.

3.Data cleaning:

Filling the null values and removing unwanted data.

Table 1.2: Dataset after Preprocessing

VIDEO GAME SALES PREDICTION

1.IMPORTING THE LIBRARIES

```
In [2]: # Importing essential libraries
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
```

2.IMPORTING THE DATASET

```
In [3]: #Importing the Kaggle Dataset
        df = pd.read_csv("Video_Games_Sales_as_at_22_Dec_2016.csv")
```

3.DATA CLEANING

```
In [8]: df.head()
```

Out[8]:

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_S
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0	
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	

```
In [6]: df.isnull().sum()
```

```
Out[6]: Name              2
        Platform          0
        Year_of_Release   269
        Genre             2
        Publisher         54
        NA_Sales          0
```

```
In [10]: df = df.fillna({
             'Year_of_Release' : 2007,
         })

In [11]: df['Developer'].fillna(method = 'ffill' ,inplace = True)
         df['Critic_Score'].fillna(method = 'ffill' ,inplace = True)
         df['Critic_Count'].fillna(method = 'ffill' ,inplace = True)
         df['User_Score'].fillna(method = 'ffill' ,inplace = True)
         df['User_Count'].fillna(method = 'ffill' ,inplace = True)
         df['Rating'].fillna(method = 'ffill' ,inplace = True)

In [12]: df = df.dropna()

In [13]: df.isnull().sum()

Out[13]: Name              0
         Platform          0
         Year_of_Release   0
         Genre             0
         Publisher         0
         NA_Sales          0
         EU_Sales          0
         JP_Sales          0
         Other_Sales       0
         Global_Sales      0
         Critic_Score      0
         Critic_Count      0
         User_Score        0
```

Methodology

Methodology may be known as a range from a primary quantitative approach towards a primary qualitative approach.

While a Machine Learning Model is building, it isn't constantly a case that we perform all and designed information. And keeping in view that performing any activity with information, it is required to preprocess it and put in an arranged format. So, for this, we perform data preprocessing task

- Linear Regression Model
- Random Forest Model
- Logistic Regression Model

Role and Responsibilities

Role	Name	Responsibilities
Data Entry & Tester	G Mahesh	· Data Entry · Data Preprocessing Documentation · Testing
Data Entry & Tester	M S Kushal Raj	· Data Entry · Data Preprocessing Documentation · Testing
Data Scientist	**PVS** Karthikeya	· Data Entry · Data Preprocessing · Documentation · Machine Learning · Data Analysis · Data Mining · Data Visualization

Data Scientist	M Sai Kiran	· Data Entry · Data Preprocessing · Documentation · Machine Learning · Data Analysis · Data Mining · Data Visualization

Project Contribution:

Potential in market

Due to boost in information analysis, data science, and diverse regions, statistical surveying bunches that play out the forecasts on video games came to fruition. One such organization is Newzoo which works with various enormous clients like Facebook, Microsoft, Google, Pokémon, and various Sports Interactive to help them in fostering a superior investigation. Organizations are getting a handle on the information and aptitude of statistical surveying organizations to build their development generally and into new sections.

Under ideal conditions, this undertaking can be utilized by video game delivering organizations to improve comprehension of the video game deals in the market to perceive what sorts of games are performing better in market and which conditions creates more deals. These distributers and engineers are in the enthusiasm of creating income and deals, so they perform to see which elements are generally significant for computer game deals. Preferably, the forecasts of our investigation will assist engineers with choosing in which stage, classification, or nation to deliver their future computer games which impacts the video game deals.

By using our project, the video game companies can learn the changes of the gaming market and can develop the games which satisfies the user preferences which in turn produces a great revenue for the company.

Innovativeness

The primary thought behind this task is that video game deals are unsurprising; Our objective in this undertaking is to play out the anticipating of future computer game deals dependent on past deals, in light of information from Kaggle dataset around 16,719 distinct games from a wide range of stages. These information come from Kaggle, a site that contains the diverse informational collection. Utilize a huge scope of information of the games from the entirety of the distinctive computer game classes to really improve perspective available. We are playing out the forecast of the computer games by consolidating the consequence of two ML models and we are foreseeing the hit of the computer game which will be useful for various organizations for fostering the video games.

Usefulness

By using our project, the gaming companies can learn the changes of the gaming market and can develop the game which satisfies the user preferences which will generate more revenue. Under ideal conditions, this project can be utilized by computer game designers to improve comprehension of the computer game market to perceive what kinds of games are liked and which components sway deals. These analysts and engineers are occupied with producing income and deals, so they need to see which elements are generally significant for computer games deals. Ideally, the forecasts of our undertaking will help computer game designers to choose in which platform, genre, or nation to deliver their future computer games for creating more deals for the game.

2 Report Organization

The leftover segment of the report is organized as follows:

- **Chapter 2** gives itemized business and specialized necessities
- **Chapter 3** gives investigation and plan of this task
- **Chapter 4** gives Construction, execution subtleties of this venture
- **Chapter 5** gives Conclusion and future extension just as future use of this task

Chapter-2

Engineering Requirement

Requirement of functional

It is an illustration of the service that the software should offer to the user. It defines a software system or its component. A functional requirement is nothing but they are the inputs to the software system, its behavior, and outputs. The functional requirements describe about how an application performs functionally.

A useful prerequisite is a necessity in regards to a yield of conduct that will be given by an element of the framework. Practical Requirements will be prerequisites that indicates a capacity that a framework or framework segment should have the option to perform.

Requirement of non-functional

Non-Functional necessities are the contributions to the application which are not straightforwardly with explicit usefulness conveyed by the framework. The non-utilitarian necessities are predominantly about execution, data, economy, the executives and security intensity and administrations. They might be likewise identified with various properties like unwavering quality, ease of use and so on

- Generate maximum accuracy:

  Accuracy is one the method for evaluation of different models. Accuracy is also known as the fraction of predictions to know whether our model got right or not.

- Provide visualized analysis:

  Visual investigation is basically the mix of information examination and perceptions. This technique to tackling issues is respected with coordinating intelligent visual portrayals with basic examination cycles to viably work with undeniable level, complex exercises, for example, thinking and information driven dynamic.

- Ease of use:

  The degree to which an application can be utilized by indicated clients to accomplish determined objectives with adequacy, effectiveness, and fulfillment in a predefined setting of utilization.

- Availability:

  Availability is technique to ascertain how much the clients can rely upon the framework during working occasions.

- Reliability:

  The degree that a item, framework, or administration will play out its planned capacity satisfactorily for a specific timeframe, or will work in a characterized climate without disappointment.

- Maintainability:

  Degree of a product application or part can be adjusted to address issues, further develop execution or different qualities, or adjust to a changed conditions.

Software Environment:

Operating System - Microsoft Windows 8.1

Scripting language - Python

Packages Required:

The following modules are used in our project:

Pandas:

It is an accurate, fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

It is one of the Python libraries.

It is used to for the analysis of data.

NumPy:

It is a general array processing package. It is a library used for working with arrays. It also has different functions for working in linear algebra domain, and matrices.

It defines a high- performance multidimensional array object, and tools for working with the arrays. It is the primary package for scientific computing in Python.

we have different lists that serve the purpose of arrays, but they are slow when compared to NumPy to process.

It provides an object of array type that is up to 50x faster the than traditional Python lists.

Matplotlib:

It is a fundamental visualization library in Python for two dimensional plots of arrays. It is one of the multi-platform data visualization libraries built on NumPy arrays and built to work with the broader SciPy stack.

It is a collection of different functions that make matplotlib work like MATLAB.

Each pyplot function makes some change to a figure: such as, creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

Sklearn:

It is the most useful and secured library for machine learning in Python.

It describes a choosing of appropriate tools for machine learning and statistical modeling including different classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Seaborn:

Seaborn is one of the Python data visualization libraries based on matplotlib.

It provides a high-level interface for generating attractive and informative statistical graphics

Seaborn is used in exploring and understanding the data.

Its API helps us to focus on what the different elements of plots mean, rather than on the details of how to generate them

Chapter-3

ANALYSIS AND DESIGN

Use case diagram

It is the general portrayal of the framework. It is an arrangement of steps depicting a collaboration between a client and a framework.

Consequently, it is a bunch of circumstances integrated by some objective. The utilization case graphs are drawn for showing the functionalities of the framework.

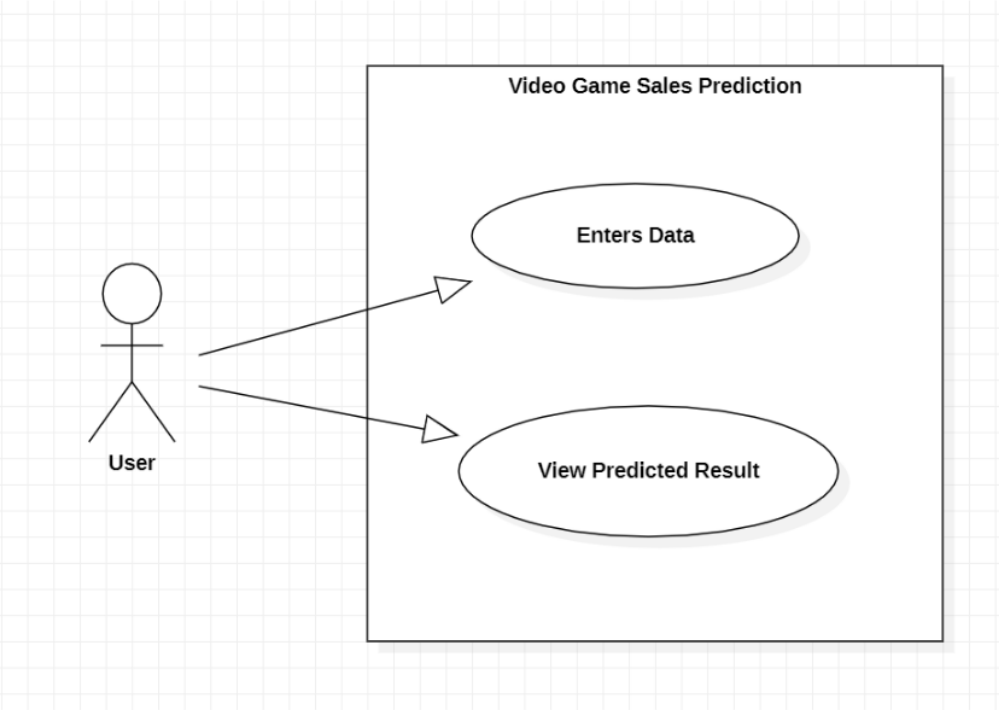


Fig 3.1-Representation of Use case

Sequence diagram:

Sequence diagrams is interrelated with utilization case representation acknowledge in the sensible perspective on the framework under programming advancement. It is otherwise called occasion charts or occasion situations.

The arrangement chart shows in which way the items cooperates with one another. The means of occasions that are addressed.

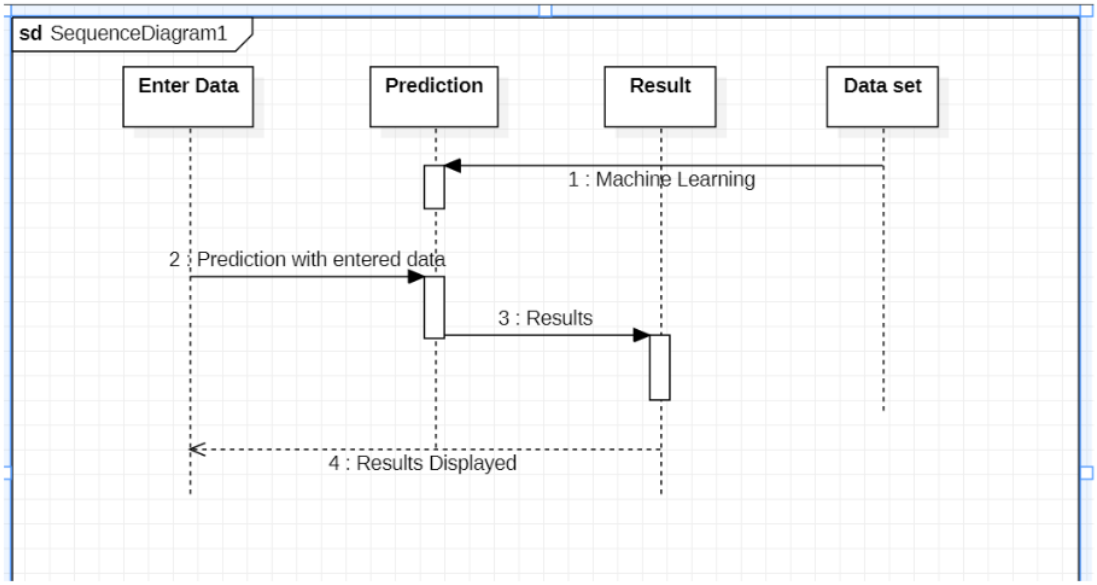


Fig. 3. 4. Sequence diagram

System architecture:

System architecture will help the designers and engineers in representation a system or application's high-level, overarching layout to ensure the framework addresses the needs of their customers.

The framework compositional plan is the cycle for perceiving the subsystems developing the framework and system for subsystem control and correspondence. The point of the framework engineering configuration is to produce the general design of the product framework.

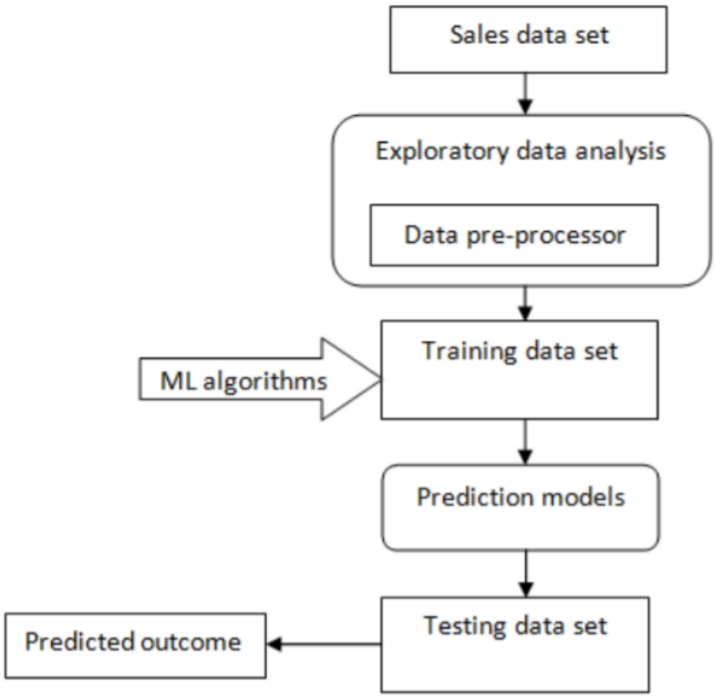


Fig. 3.5. Representation of system structure

Chapter-4

CONSTRUCTION

Implementation

Our project is implemented using python language in a platform called Anaconda. It is used for the building of different Machine learning models.

Anaconda is one of the distributions of Python. Anaconda is one of the new distributions of Python. Continuum Analytics is also a name of Anaconda. Anaconda comprises of in excess of 100 new bundles. Anaconda is utilized for various purposes like logical processing, information science, measurable investigation, and ML.

Technology in python is easy to implement in Anaconda, because it removes the accompanying tasks:

- Introducing Python on different stages.
- Distinguishing out various conditions.
- Performing such that it does not have right advantages.

The information is available in Kaggle website which has dataset which are reliable. Implementation of the idea started from the Problem statement which is given in some random Hackathon for the prediction of Global Sales. The data was sorted and the null values are filled and remaining useless data is dropped for better designing of the machine learning project.

The sections were done just to make the Machine Learning model to realize what all it has to do with the information and what really the yield will be delivered. When the Machine Learning is performed with various calculations and diverse interaction, execution of various calculations is estimated and the calculation which is ideal is utilized for the expectation i.e., Random Forest, Linear Regression, Logistic Regression.

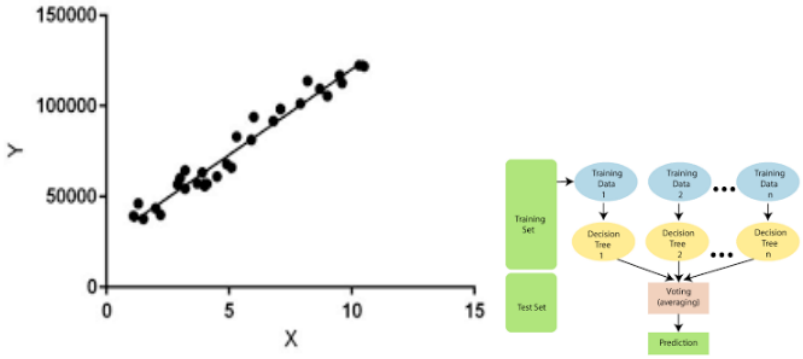
EXISTING APPROACH:

Algorithm linear regression:

The algorithm linear regression is generally utilized and one of the simplest and most well-known Machine Learning calculations.
The algorithm linear regression is performed for prescient displaying methods.

The fundamental motivation behind this calculation is to acquire a numerical condition for persistent factors Y when we have at least one X factors.

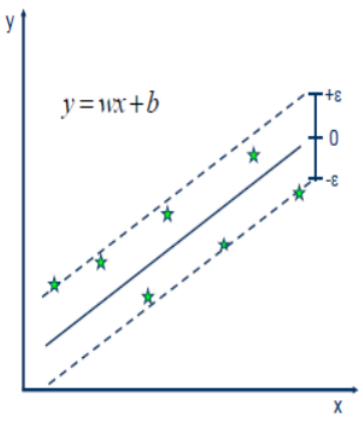
This calculation creates a connection between two factors one variable is anticipated variable and another is result variable whose worth is gotten from the prescient variable.



Algorithm Support Vector Regression:

It utilizes algorithm of classification to anticipate a constant variable. In any case, other models are utilized to diminish the inaccuracy between anticipated and real value.

It ensures to keep perfect line between the already generated inaccuracy value.



• Solution:

$$\min \frac{1}{2}\|w\|^2$$

• Constraints:

$$y_i - wx_i - b \le \varepsilon$$

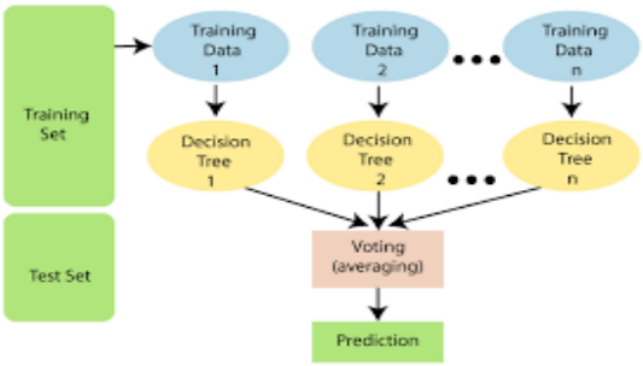
$$wx_i + b - y_i \le \varepsilon$$

Algorithm random forest:

14

It is one of the machine learning algorithms of supervised type that generates haphazardly a forest containing many trees.

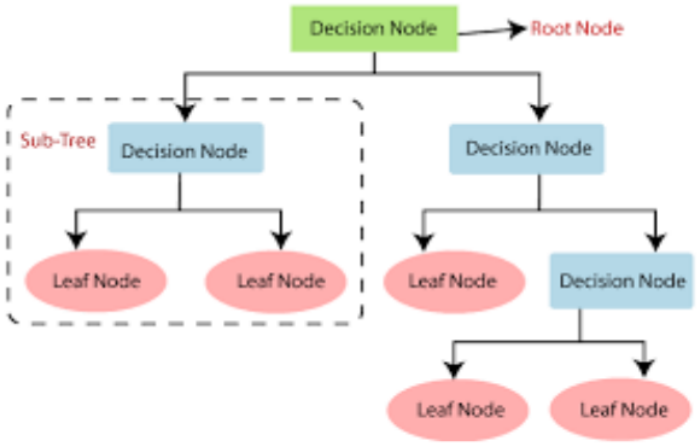
It is used in calculation as opposed to decision tree, these trees are not difficult to utilize and work productively with the information, however it gives less exactness due to over fitting.



Decision Tree:

It is one of the managed ML trees which portrays about what the information is and what is the important yield as indicated by the information.

The fundamental objective of this calculation is to conjecture the worth of an objective variable. It contains rules are as contingent articulations, for example, in the event that else.



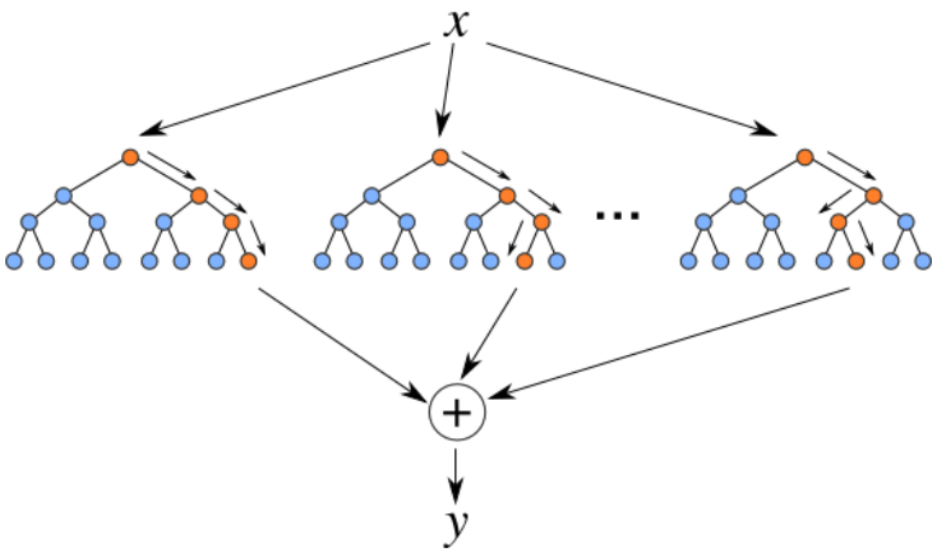
PROPOSED APPROACH:

Implementation Details

With the end goal of better execution and working we utilize fitting Algorithm and methods are utilized. Following is the calculation utilized:

Algorithm random forest regression:

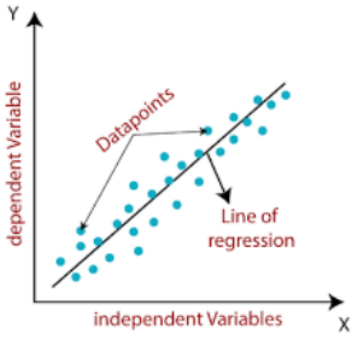
It is one of the machine learning algorithms of supervised type that generates haphazardly a forest containing many trees. It is used in calculation as opposed to decision tree, these trees are not difficult to utilize and work productively with the information, however it gives less exactness due to over fitting.



Representation of random forest algorithm

Algorithm linear regression:

The algorithm linear regression is generally utilized and one of the simplest and most well-known Machine Learning calculations.
The algorithm linear regression is performed for prescient displaying methods.



Representation of linear regression

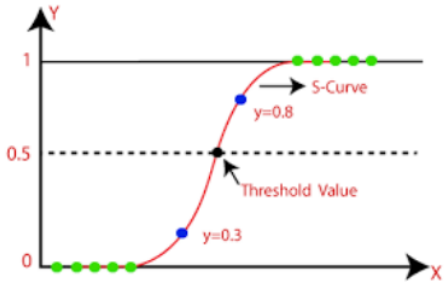
Algorithm logistic regression:

Logistic regression is a supervised learning technique. Logistic regression is one of the most popular Machine Learning algorithms.

It is used for forecasting the categorical dependent variable using a given set of independent factors.

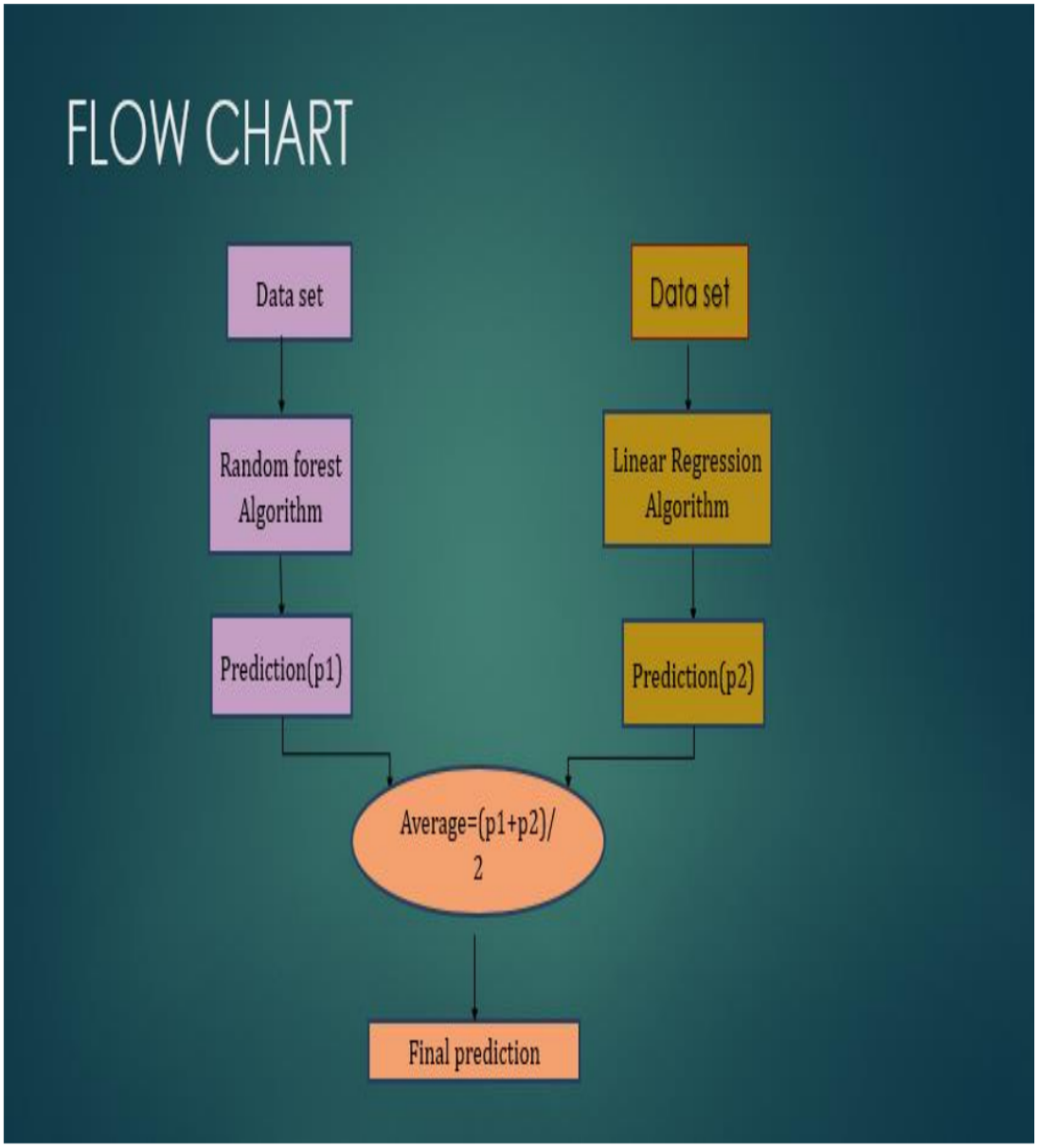
Logistic regression forecasts the output of a categorical dependent variable. Therefore, the predicted value must be a categorical or discrete value.

It can be either YES or NO, 0 or 1, TRUE or False, etc. but rather than giving the exact value as 0 and 1.

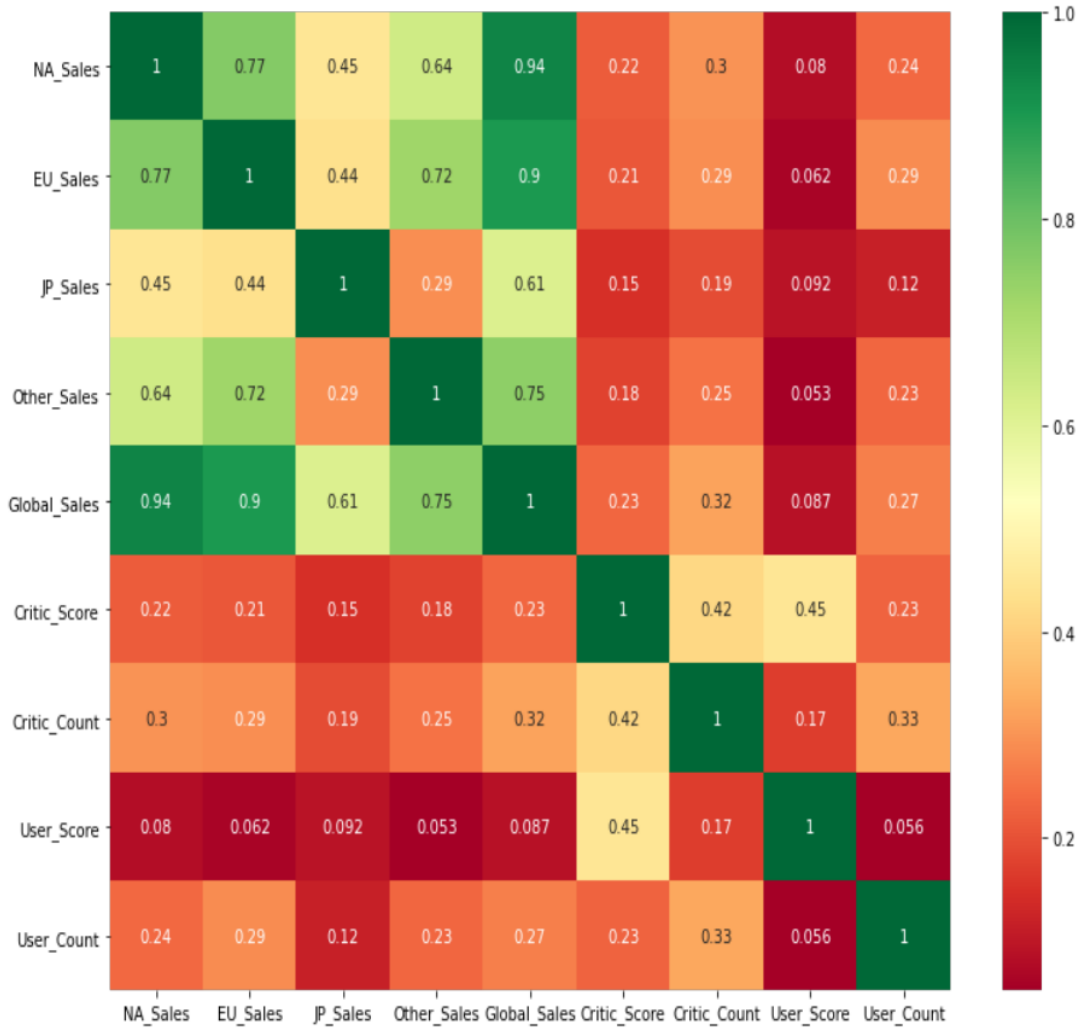


Logistic Regression

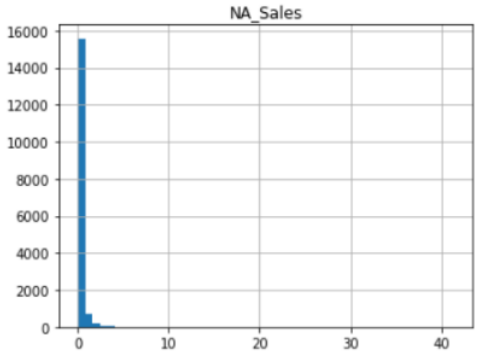
FLOWCHART OF IMPLEMENTATION:



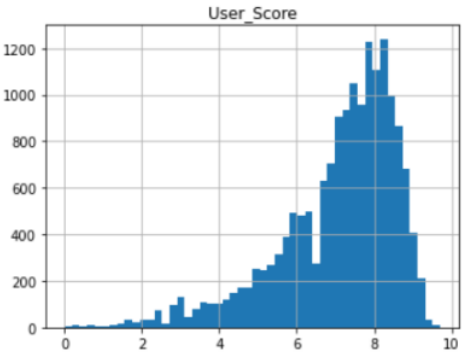
Data Visualization:



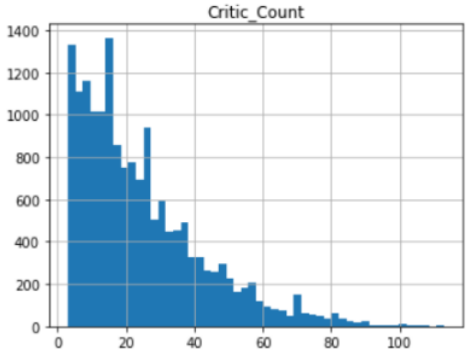
Heatmap



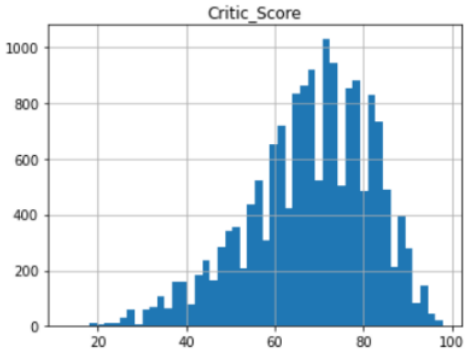
Histogram of NA_Sales



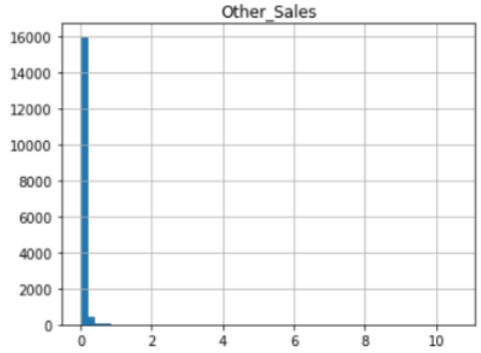
Histogram of User_Score



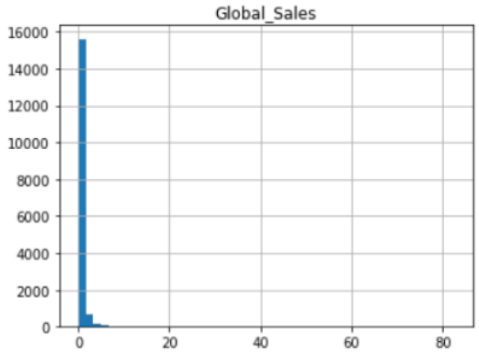
Histogram of Critic_Count



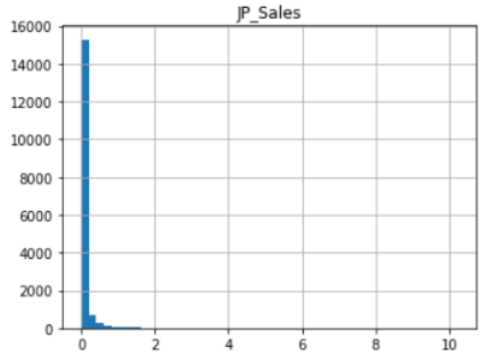
Histogram of Critic_Score



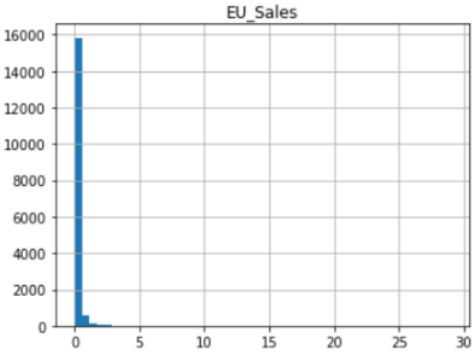
Histogram of Other_Sales



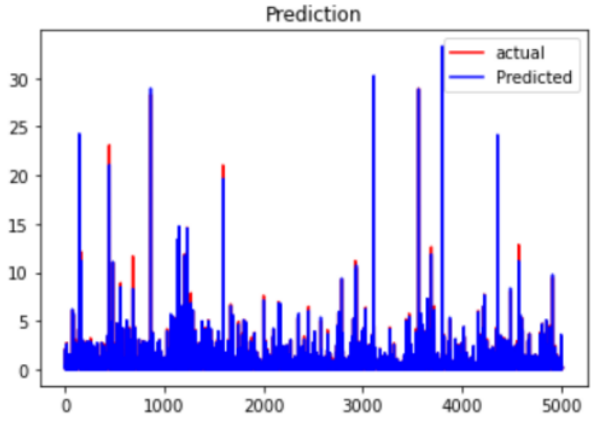
Histogram of Global_Sales



Histogram of JP_Sales



Histogram of EU_Sales



Prediction

Software Details

- Anaconda Distribution (v5.1)
- Python (3.6.5)
- Jupyter Notebook

Hardware Details

- Operating system: Windows 7 or newer, 64-bit macOS 10.9+, or Linux.
- System architecture: 64-bit x86, 32-bit x86 with Windows or Linux.
- CPU: Intel Core 2 Quad CPU Q6600 @ 2.40GHz or greater.
- RAM: 4 GB or greater.

Testing:

Programming testing is the way toward setting up a framework fully intent on perceiving any mistakes, holes or missing necessity versus the genuine prerequisite. Programming testing is mostly arranged into two kinds

- testing of functional.

- testing of non-functional.

Performing test exercises: Testing ought to be proceeded as right on time as conceivable to limit the expense and time to revamp and create programming that is without mistake so it tends to be reached to the customer.

In Software Development Life Cycle (SDLC), testing can be performed from the information assortment stage and endures till the product is out there in creations.

For instance, in the Waterfall model, testing is performed from the testing stage which is only beneath in the tree; yet in the V-model, testing is done in corresponding to the improvement stage.

Programming testing is an assessment prompted create accomplices information about the idea of the item thing or organization under test. Programming testing can create an objective, programmed point of view on the item to permit the business to recognize and appreciate the threats of programming running.

Test strategies consolidate the way toward running a program or application with the arrangement of discovering programming blunders (botches or various disfigurements), and valuating that the item thing is good for use.

It contains the running of an item part or structure portion to evaluate something like one properties of interest.

At the point when everything is said in finished, these highlights show how much the portion or system under test:

- Encounters the conditions that teaches its structure, advancement,
- performs viably to an enormous scope of wellsprings of data,
- plays in its anything but a sufficient time,
- it is sufficiently viable,
- implemented, executed in expected conditions,
- fulfil the overall expectation its accomplices need.

As the quantity of likely tests for even broad programming parts is essentially boundless, all item testing utilizes a few methodologies to pick tests that are attainable for the open time and resources. Accordingly, programming testing by and large endeavor to execute a program or application with the assumption for finding programming errors. The movement of testing is a recursive technique as when one error is fixed, it can illuminate other, more significant bugs, or can even make new ones.

White box testing

It is programming trying system in which inner construction, plan and coding of programming are tried to check stream of information yield and to further develop plan, convenience and security. In this testing an inner point of view of the framework, just as programming abilities, are utilized to configuration experiments. The analyzer picks contributions to practice ways through the code and decide the normal yields.

Working process in this testing is:

- Input
- Processing
- Proper test planning
- Output

3

Black box testing

It is a product testing technique in which the functionalities of programming applications are tried without knowing about interior code structure, execution subtleties and inner ways. Its chiefly centers around information and yield of programming applications and it is altogether founded on programming necessities and details.

Proposed Modules and their Functionality:

- 1.**IMPORTING REQUIRED LIBRARIES:**

  we have imported NumPy, pandas, sklearn, matplotlib libraries.

- 2.**IMPORTING THE DATASET:**

  downloaded Kaggle dataset has been imported.

- 3.**DATA CLEANING:**

  filling the null values and removing unwanted data.

- 4. **BUILDING LINEAR REGRESSION MODEL:**

  training the linear regression model using train data set

- **5. BUILDING RANDOM FOREST MODEL:**

  training the random forest model using train data set

- **6. BUILDING LOGISTIC REGRESSION MODEL:**

  training the logistic regression model using train data set

- **7. ACCURACY CALCULATION:**

we can find the actual and predicted values by using existing methods called RMSE, r^2 and MAE.

- **8. PREDICT ON A NEW DATA FEATURES:**

  Results of Random forest and Linear regression algorithms, we collect the outputs of and find the averages for better accuracy and classification of video game sales prediction. Results of Logistic Regression, we calculate the hit rate of the test dataset.

Implementation:

16

Module 1:

VIDEO GAME SALES PREDICTION

1.IMPORTING THE LIBRARIES

In [2]:
```
# Importing essential libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Module 2:

2.IMPORTING THE DATASET

In [3]:
```
#Importing the Kaggle Dataset
df = pd.read_csv("Video_Games_Sales_as_at_22_Dec_2016.csv")
```

Module 3:

3.DATA CLEANING

In [8]:
```
df.head()
```

Out[8]:

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Sc
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0	
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
Name                2
Platform            0
Year_of_Release     269
Genre               2
Publisher           54
NA_Sales            0
```

In [10]:
```
df = df.fillna({
    'Year_of_Release' : 2007,
})
```

In [11]:
```
df['Developer'].fillna(method = 'ffill' ,inplace = True)
df['Critic_Score'].fillna(method = 'ffill' ,inplace = True)
df['Critic_Count'].fillna(method = 'ffill' ,inplace = True)
df['User_Score'].fillna(method = 'ffill' ,inplace = True)
df['User_Count'].fillna(method = 'ffill' ,inplace = True)
df['Rating'].fillna(method = 'ffill' ,inplace = True)
```

In [12]:
```
df = df.dropna()
```

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
Name              0
Platform          0
Year_of_Release   0
Genre             0
Publisher         0
NA_Sales          0
EU_Sales          0
JP_Sales          0
Other_Sales       0
Global_Sales      0
Critic_Score      0
Critic_Count      0
```

Module 4:

4.BUILDING LINEAR REGRESSION MODEL

In [15]:
```
from sklearn.linear_model import LinearRegression
regressor_MultiLinear = LinearRegression()
regressor_MultiLinear.fit(X_train,y_train)
```

Out[15]:
```
LinearRegression()
```

In [16]:
```
s_pred = regressor_MultiLinear.predict(X_test)
games_in_test_set = games_in_test_set.reshape(-1, 1)
s_pred = s_pred.reshape(-1, 1)
prediction = np.concatenate([games_in_test_set, s_pred, y_test], axis = 1)
prediction = pd.DataFrame(prediction, columns = ['Name', 'Predicted_Global_Sales', 'Actual_Global_Sales'])
```

Module 5:

5.BUILDING RANDOM FOREST MODEL

In [19]:
```
from sklearn.ensemble import RandomForestRegressor
regressor=RandomForestRegressor(n_estimators=10,random_state=0)
regressor.fit(X_train,np.ravel(y_train,order='C'))
y_pred = regressor.predict(X_test)
```

In [20]:
```
y_pred = regressor.predict(X_test)
games_in_test_set = games_in_test_set.reshape(-1, 1)
y_pred = y_pred.reshape(-1, 1)
predictions = np.concatenate([games_in_test_set, y_pred, y_test], axis = 1)
predictions = pd.DataFrame(predictions, columns = ['Name', 'Predicted_Global_Sales', 'Actual_Global_Sales'])
```

Module 6:

6.BULDING LOGISTIC REGRESSION MODEL

In [28]:
```
df['hit'] = np.where(df['Global_Sales'] > .47 ,1 ,0)
```

In [29]:
```
df = df.drop(['NA_Sales','EU_Sales','JP_Sales','Other_Sales','Critic_Count','User_Count','Rating'], axis = 1)
```

In [31]:
```
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
df['Genre']= label_encoder.fit_transform(df['Genre'])
```

In [32]:
```
a= df.iloc[:, [0,1,3,4]].values
b= df.iloc[:, 5].values
```

In [33]:
```
from sklearn.model_selection import train_test_split
a_train, a_test, b_train, b_test= train_test_split(a, b, test_size= 0.3)
```

In [34]:
```
from sklearn.linear_model import LogisticRegression
logisticRegr = LogisticRegression()
logisticRegr.fit(a_train[:,1:],b_train)
```

Out[34]:
```
LogisticRegression()
```

Module 7:

7.ACCURACY CALCULATION

RANDOM FOREST MODEL

In [38]:
```
from sklearn.metrics import  mean_squared_error
print("Score of the model:",end=" ")
r_sq = regressor.score(X_train,np.ravel(y_train,order='C'))
print(r_sq)
rmse = math.sqrt(mean_squared_error(y_test, y_pred))
print(f"Root Mean Squared Error of the model : {rmse:.3f}")
```

```
Score of the model: 0.9673619390973538
Root Mean Squared Error of the model : 0.237
```

LINEAR REGRESSION MODEL

In [39]:
```
from sklearn.metrics import  mean_squared_error
print("Score of the model",end=" ")
print(1 - (1-regressor_MultiLinear.score(X_train, y_train))*(len(y_train)-1)/(len(y_train)-X_train.shape[1]-1))
rmse = math.sqrt(mean_squared_error(y_test, s_pred))
print(f"Root Mean Squared Error of the model : {rmse:.3f}")
```

```
Score of the model 0.9999893733060171
Root Mean Squared Error of the model : 0.005
```

Test Table

Test Cases

In [142]:
```
predictions.iloc[1000:1010]
```

Out[142]:

	Name	Predicted_Global_Sales	Actual_Global_Sales
1000	Samurai Warriors: State of War	0.193	0.18
1001	Dynasty Tactics 2	0.099	0.11
1002	Sloane to MacHale no Nazo no Monogatari 2	0.091	0.09
1003	Major League Baseball 2K10	0.518	0.51
1004	GT Advance 3: Pro Concept Racing	0.069	0.08
1005	Imagine: Wedding Designer	1.301	1.28
1006	Natural Doctrine	0.081	0.1
1007	Elebits	0.344	0.32
1008	OutRun 2006: Coast 2 Coast	0.041	0.04
1009	Mashiro Iro Symphony: *mutsu-no-hana	0.03	0.03

In [143]:
```
prediction.iloc[170:180]
```

Out[143]:

	Name	Predicted_Global_Sales	Actual_Global_Sales
170	Yamasa Digi Portable: Matsuri no Tatsujin - Wi...	0.020578	0.02
171	The Lost Chronicles of Zerzura	0.060317	0.05
172	Ou to Maou to 7-nin no Himegimitachi: Shin Ous...	0.040341	0.04
173	Road Rash	1.270237	1.27
174	Bejeweled Twist	0.090992	0.1
175	Scaler	0.040323	0.04
176	M&M's Kart Racing	0.1203	0.12
177	Routes PE	0.010177	0.01
178	World Trigger: Borderless Mission	0.040535	0.04
179	Puzzler Collection	0.980954	0.99

In [140]:
```
answer.iloc[120:130]
```

Out[140]:

	Name	Predicted_Global_Sales	Actual_Global_Sales
120	PlayStation All-Stars Battle Royale	0.439795	0.45
121	Venetica	0.110507	0.12
122	Driver: Parallel Lines	0.06202	0.07
123	Jikkyou Powerful Pro Yakyuu 2000 Kaimakuban	0.287527	0.29
124	Yogi Bear: The Video Game	0.137723	0.14
125	Fatal Fury Special	0.55804	0.56
126	Master Jin Jin's IQ Challenge	0.141688	0.14
127	Mario Strikers Charged	2.578048	2.58
128	Backyard Wrestling 2: There Goes the Neighborhood	0.12444	0.12
129	Pajama Sam: Don't Fear The Dark	0.100144	0.1

In [144]:
```
pd.DataFrame(predict, columns = ['Name', 'Predicted hit', 'Actual hit']).iloc[220:230]
```

Out[144]:

	Name	Predicted hit	Actual hit
220	Style Lab: Makeover	0	0
221	Alice: Madness Returns	0	0
222	Power Rangers Samurai	0	0
223	Hisshou Pachinko*Pachi-slot Kouryaku Series DS...	0	0
224	Dante's Inferno	0	0
225	Rock Band Unplugged	1	0
226	Gunslingers	0	0
227	Winning Post 7: Maximum 2007	0	0
228	TOCA Touring Car Championship	1	0
229	Tony Hawk's Underground 2	1	0

Test Cases

Chapter-5

CONCLUSION AND FUTURE SCOPE

Conclusion

Video Game Sales Prediction is helpful in predicting the revenue of games. This methodology is valuable to a few enterprises which are keen on creating Video Games which will be top-netting. Revenue foreseeing is a significant piece of the essential arranging measure. It permits an organization to anticipate how the organization will act later on. Foreseeing revenue of an organization isn't just for arranging new plans, yet additionally permit knowing the negative patterns that show up in the expectation. At last, we infer that expectation of deals on video games has done and we saw which game has more deals in the market internationally.

Keywords: Accuracy, Classification, Random Forest, Linear Regression, Logistic Regression, Categorical, and Regression.

Regression Future Scope

The future scope is that we can perform the same with different machine learning algorithms and we can also use the different dataset to train the model so that it can generate more accurate results. The GUI can be also developed so that it looks more interactive. Further we can also find for a method to calculate the score for our model. An intriguing finding is that the count of users and the count of critics, rating a game on Metacritic is more significant in clarifying the outcomes than the actual scores.

This means that the quickly developing computer game industry: as the incentive for the year builds, unit revenue increment for the most part because of the actual market developing.

Expected Outcome

Random Forest and Linear Regression Algorithm for the better accuracy as compared to existing video game sales predictions. Random forest is a prediction technique in supervised ML(it predicts categorical and continuous outcome). Linear Regression is one of the supervised ML algorithm where the predicted output is nonstop and has a steady slant. It's utilized to anticipate values within a continuous range. Logistic Regression produces discrete/categorical output. Computer game Sales forecast is a significant piece of the essential arranging measure. It permits the gaming organizations to anticipate the revenue of the game. Anticipating revenue of the game isn't just for arranging new freedoms, yet additionally helps in knowing the negative patterns that show up in the forecast.

References

- Dataset Source: https://www.kaggle.com/sidtwr/videogames-sales-dataset?select=Video_Games_Sales_as_at_22_Dec_2016.csv
- https://vgsales.fandom.com/wiki/Video_game_industry
- https://www.questjournals.org/jrbm/papers/vol7-issue3/I07036064.pdf
- https://analyticsindiamag.com/video-game-sales-prediction-hackathon/
- https://www.sas.com/en_in/insights/analytics/machine-learning.html#:~:text=Machine%20learning%20is%20a%20method,decisions%20with%20minimal%20human%20intervention.

CSE_A18_2

ORIGINALITY REPORT

19%	9%	2%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	3%
2	innovate.mygov.in Internet Source	2%
3	Submitted to Midlands State University Student Paper	2%
4	www.slideshare.net Internet Source	2%
5	Submitted to Softwarica College Of IT & E-Commerce Student Paper	1%
6	Submitted to University of Johannsburg Student Paper	1%
7	Submitted to Roehampton University Student Paper	1%
8	Submitted to Gitam University Student Paper	1%
9	Submitted to Federation University Student Paper	1%

10	Submitted to Chiang Mai University Student Paper	1 %
11	Submitted to Anglia Ruskin University Student Paper	1 %
12	grietinfo.in Internet Source	1 %
13	Submitted to Taylor's Education Group Student Paper	<1 %
14	Pramod Gupta, Naresh K. Sehgal. "Introduction to Machine Learning in the Cloud with Python", Springer Science and Business Media LLC, 2021 Publication	<1 %
15	Submitted to University of Wollongong Student Paper	<1 %
16	www.dimensionaleclipsing.com Internet Source	<1 %
17	www.globenewswire.com Internet Source	<1 %
18	Submitted to SVKM International School Student Paper	<1 %
19	Submitted to Trident University International Student Paper	<1 %
20	Submitted to University of Salford Student Paper	<1 %

21	Iqra Basharat, Ali Raza, Mamuna Fatima, Usman Qamar, Shoab Ahmed. "A Framework for Classifying Unstructured Data of Cardiac Patients: A Supervised Learning Approach", International Journal of Advanced Computer Science and Applications, 2016 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		