Predicting Global Video-Game Sales

Abstract

As it is known, video games have spread today in homes worldwide and are suitable for all ages, additional they are spending for made of these things are a huge annually. Although video games have been criticized for making people more depressed, antisocial, or overweight. But now researchers are finding that games can actually change us for the better and improve both our body and mind.

So, I had selected this data to analyze some questions including some ways use it to learn the different machine predict to global sales for video games based on other sales.

QUESTIONS

Q1: Are there a different local Sales related to each other in countries?
Q2: What is the highest selling type of video game in global sales and North America?
Q3: Who are the most popular publishers?
Q4: Which is the most popular Platform in Global Sales?
Q5: who are the highest sales?

Design

In the beginning, this subject is my passion of video games and I had been excited to know the most popular types that sell. So, I collected this data from Kaggle.com.

At this point, the data is presented, cleaned, and preprocessing, and the EDA is used to understand and analyze the relationships between features and their importance in global sales and to figure out the answers for questions they have been asked. Through these results, any programmer in the future can develop or programed video games according to the increasing demand in sales.

Data

a) Description of Dataset:

Dataset I obtained was from 'kaggle.com' that contains 16,598 datapoints and 11 features for each, including features:
Ranking of overall sales
The games name
Genre
Publisher
Sales in North America

Sales in Europe
Sales in Japan
Sales in rest of the world
Sales in global.

b) Dataset source:

Sales Of Video Games | Kaggle

Tools

There are tools I will used in the datasets, such as:

- Data Processing: Panda, NumPy.
- Modeling: Scikit-Learn.
- Visualizations: matplotlib, seaborn.

Algorithm

In the case of global sales predicting, I have used of two models to get the best results of (LinearReression and RandomForestRegression) and determine the results value of R- squared for LinearReression = 0.99 and RandomForestRegression = 0.88. Even so the LinearReression has surpass the RandomForestRegression. anyhow the random forest is likely to improve when changing the N-estimators value but will take a long time for improvement.

MVP goal

- Data has been cleaned up by removing duplicates, missing data, and some unimportant features.
- EDA: to display the relationship between features and compare between Actual & predict values.