Video Games Sales Analysis

A Machine Learning Project

I'll begin by laying out the problem that I decided to solve through machine learning. There is a huge wealth of data out there about video games – video game sales, video game popularity, the best genres, etc. However, I wanted to get to the bottom of the matter about what is truly the best-selling type of video game, and to do that I believe machine learning is the perfect method, as it allows a massive amount of data to be read, understood and fed back to you, without any sort of bias. This would also be incredibly useful for development companies to use in terms of deciding what they want to make next and get an estimate of how popular their next game could be, given some statistics.

I found data on Kaggle about video game sales, including the data for Rank, Year, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Name, and Platform. I thought this data would be good as it includes everything important to a video games popularity and statistics globally on how popular it was through its number of sales, in detail.

I loaded it into my python notebook and read through it's basic stats using .head and .shape.

I then began cleaning the data by first checking for any NA/NULL values in the data, and the only ones were in columns that I already planned to remove.

I continued cleaning the data, removing the columns that I believed would have little to no bearing on the final result of the popularity of the video game. These were: Publisher, Year, and Name. I also removed Rank as it was unnecessary. I removed the unimportant columns as I don't believe the average consumer considers the name unless it's a sequel, rarely the publisher, and the year of it's release is unimportant past the actual year of release.

I further cleaned the data by plotting a scatter graph to find any outliers in terms of global sales. I found one, which was a Genre=Sports and Platform = Wii game called, you guessed it, Wii Sports, with 80 million sales. I removed this as it's a clear outlier, and ended up with much better, more normalized data.

I then turned the genre and platform columns into something readable to the machine by encoding them using LabelEncoder, turning them into numerical values instead of text.

I then checked the skew of the data, to see how far up or down the data skewed, and then used the square root to normalize the data for the sales values, as they skewed positive, (Biostats, 2017) to transform the skewed data and make it more easily usable in the machine learning algorithm.

When splitting the data into training and test data, I decided upon an 80-20 split, as from my research, it seems to be the most commonly occurring split, due to the Pareto Principle (Wikipedia, 2022) which states that 80% of effects come from 20% of causes. It's just an idea, but it did seem to apply during my testing as I tried a few different splits, and none of them were as accurate as 80-20. You can't have a test set too big or a training set too small.

After this, I tested Linear Regression, Lasso Regression, and Random Forest Regressor to find which would be the best one to use. (Brownlee, 2023) (Michael Parzinger, 2022) I decided in the end that Linear Regression would be the best to use for the final product, as it had the highest degree of accuracy compared to the other two.