Predicting Video Games Global Sales

By Nouf Alotaibi

Email: noufmitla@gmail.com

Abstract

This is the second project for the T5 Data Science Bootcamp, which is about building a linear regression models that address a useful prediction using *Scikit-learn*, and using a data scraped from a website with *Requests*, *BeautifulSoup*, and *Selenium*. This project discusses the investigation I did to see the relationship between the global sales of games and other features of games to predict the games' global sales and success.

1. Design

Video games have become an integral part of the online culture. They became famous when they moved from large arcade machines to personal computers and consoles [1]. In this project, I investigated the relationship between the global sales of games and user/critic scores to predict the games' global sales and success. Also, I tested the effect of global sales on the game's play score, publisher, developer, genre, and other features. Moreover, I got familiar with Machine Learning, web scraping techniques, and modeling methods such as Linear Regression and different Regression algorithms.

2. Dataset

For this project, I gathered the data from two sources:

- **Metacritic:** *Metacritic* is a website that aggregates reviews of films, TV shows, music albums, and video games. For each product, the scores from each review are averaged. It is also known as the leading online review aggregation site for the video game industry. I used *Selenium* and *BeautifulSoup* libraries to scrape the data of the best video games of all times from [2], I gathered about 2,000 video games records with 11 features. The table below illustrates the Metacritic video games dataset's features and their types.

Column Name	Column type
game_name	Object
platform	Object
publisher	Object
developer	Object
release_date	Object
critics_rating	Object
num_critics_rating	Object
users_rating	Object
num_users_rating	Object
game_rate	Object
genre	Object

- **Whatoplay:** *Whatoplay* is like a portal to help discover video games. I used *BeautifulSoup* library to scrape the data of the best video games [3], I gathered about 12,825 video games records with 3 features. The table below illustrates the Whatoplay video games dataset's features and their types.

Column Name	Column type
game_name	Object
platform	Object
play_score	float64

- **Video Games Sales from Kaggle:** *"Video Game Sales"* dataset that was uploaded to Kaggle.com, see [4]. The data set contains 11 features and 16,598 records, each of which is a game released between 1980 and 2020.

Column Name	Column type
Rank	int64
Name	Object
Platform	Object

Year	float64
Genre	Object
Publisher	Object
NA_Sales	float64
EU_Sales	float64
JP_Sales	float64
Other_Sales	float64
Global_Sales	float64

3. Feature Engineering

- Creating dummy variables for the categorical features such as *game genre*, *game rate*, and *game platform*.
- Handling outliers in the target variable *global_sales* per platform.
- Scaling the continuous values in the target variable *global_sales*.
- Creating a new feature *platform_count*, that represents the number of games per platform.

4. Tools

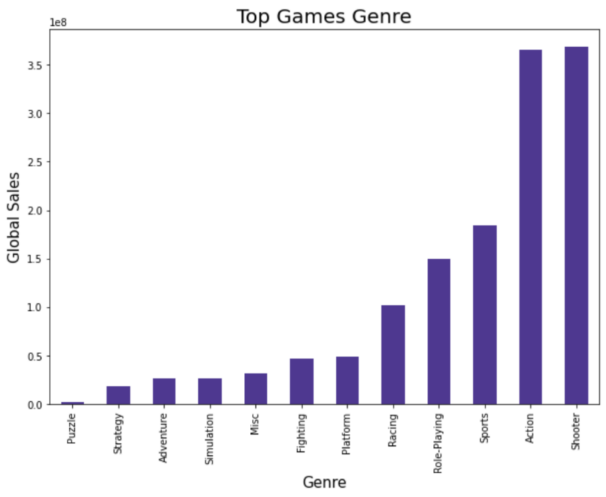
These are the technologies and libraries that I will be using for this project:

- **Technologies:** Python, Jupyter Notebook.
- **Libraries:** NumPy, Pandas, Matplotlib, Seaborn, Requests, BeautifulSoup, Selenium, Statsmodels, Scikit-learn.

5. Communications

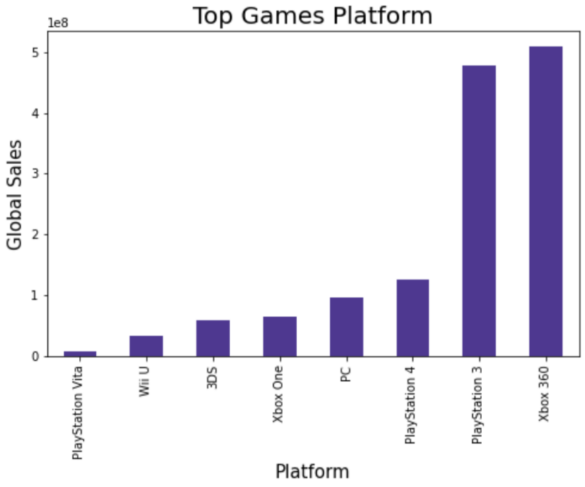
Top Video Games Genre

Shooter genre is leading, followed by Action, Sports, and Role-Playing.



Top Video Games Platform

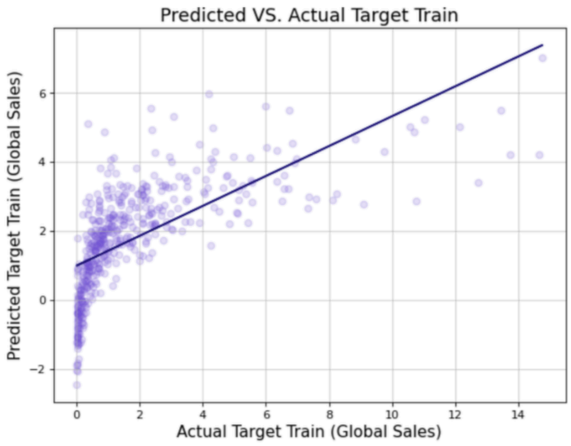
Xbox 360 is leading, followed by PS3, PS4, and PC.



Linear Regression Baseline Model

```
Training Score = 0.433532
Validation Score = 0.368182
```

global_sales has a high variance.



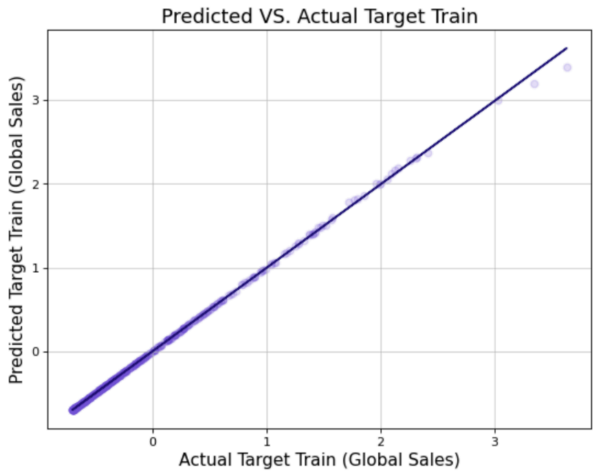
Exp	Regression Algorithem	Training Score	Validation Score
1	Linear Regression	0.512163	0.454959
2	Ridge Regression	0.511885	0.462712
3	Lasso Regression	0.510166	0.468811
4	Random Forest Regression	1.0	0.996715
5	Random Forest Regression	0.510166	0.987361

With cross validation

Data split into %90 training and %10 testing.

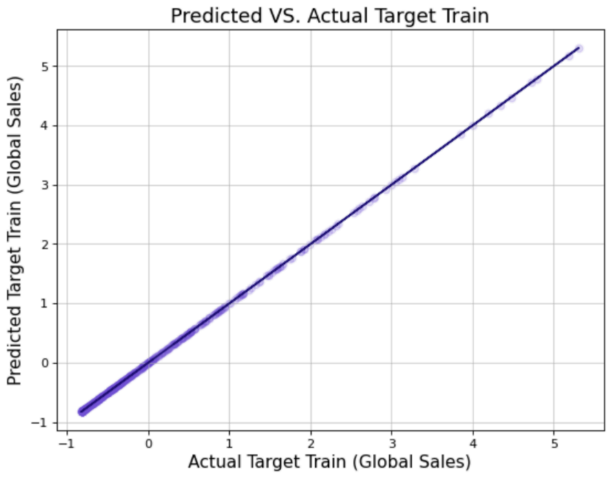
Random Forest Regression

```
Training Score = 0.999579
Validation Score = 0.987361
```



Decision Tree Regression

```
Training Score = 1.0
Validation Score = 0.996715
```



6. Resources

[1] VideoGameCons.com. (2018). 2018 Video Game Convention Calendar| VideoGameCons.com. [online] Available at: https://videogamecons.com/calendar/calendar.php?year=2018.

[2] "Best Video Games of All Time", Metacritic, 2021. [Online]. Available at: https://www.metacritic.com/browse/games/score/metascore/all/all/filtered?page=0.

[3] "Best video games of all time," *Whatoplay.com*. [Online]. Available: *https://whatoplay.com/best,*

[4] "Video Game Sales", Kaggle.com, 2021. [Online]. Available at: https://www.kaggle.com/gregorut/videogamesales.