T5 DATA SCIENCE BOOTCAMP

Predicting Video Games Global Sales

By Nouf Alotaibi

How video games' sales have been evolving through the years

In this Presentation

01 Problem Statement

02 Dataset

03 Methodology

04 Exploratory Data Analysis

05 Baseline Model

06 Feature Engineering

07 Regression Algorithms

08 Results & Insights

09 Conclusion

Problem Statement

Video games have become an integral part of the online culture.

Nintendo has been at the forefront of this online movement with video game consoles like the *Gameboy*, with *Microsoft* and *Sony* following closely behind with the *Xbox* and *PlayStation*.

Dataset

Metacritic - Web Scraping

2,000 video games records with **11** features.

Whatoplay - Web Scraping

12,825 video games records with **3** features.

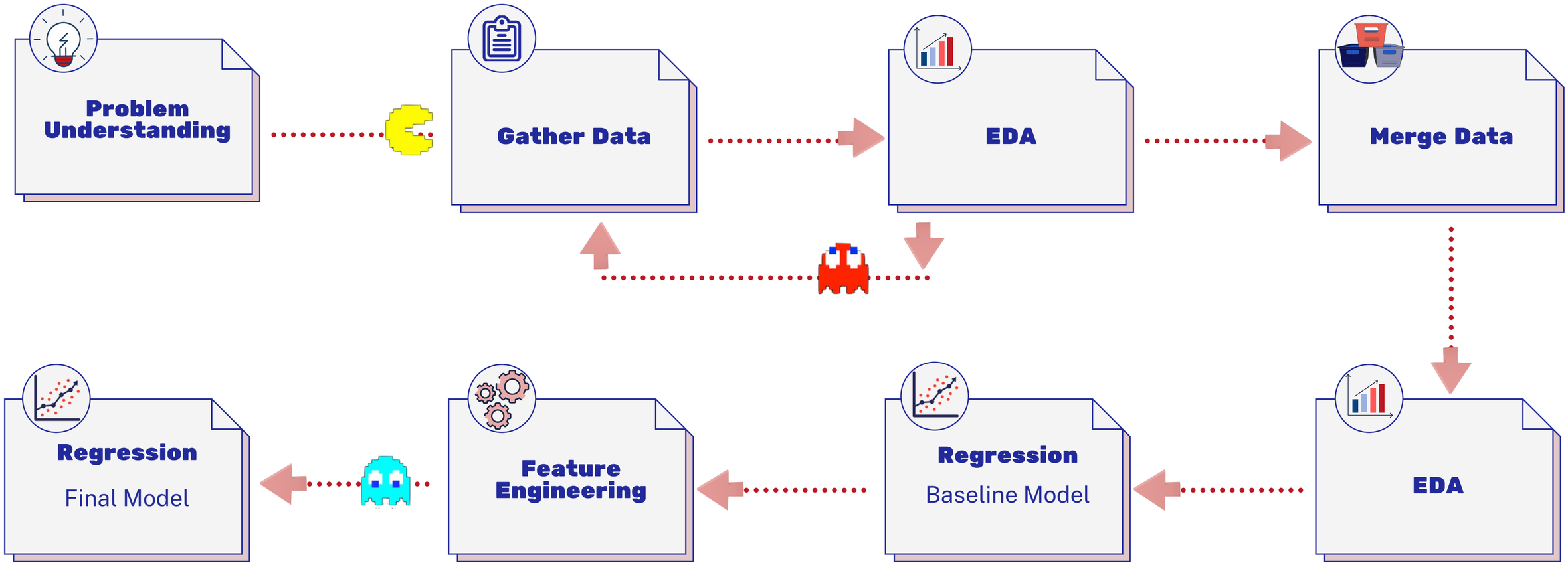
Kaggle

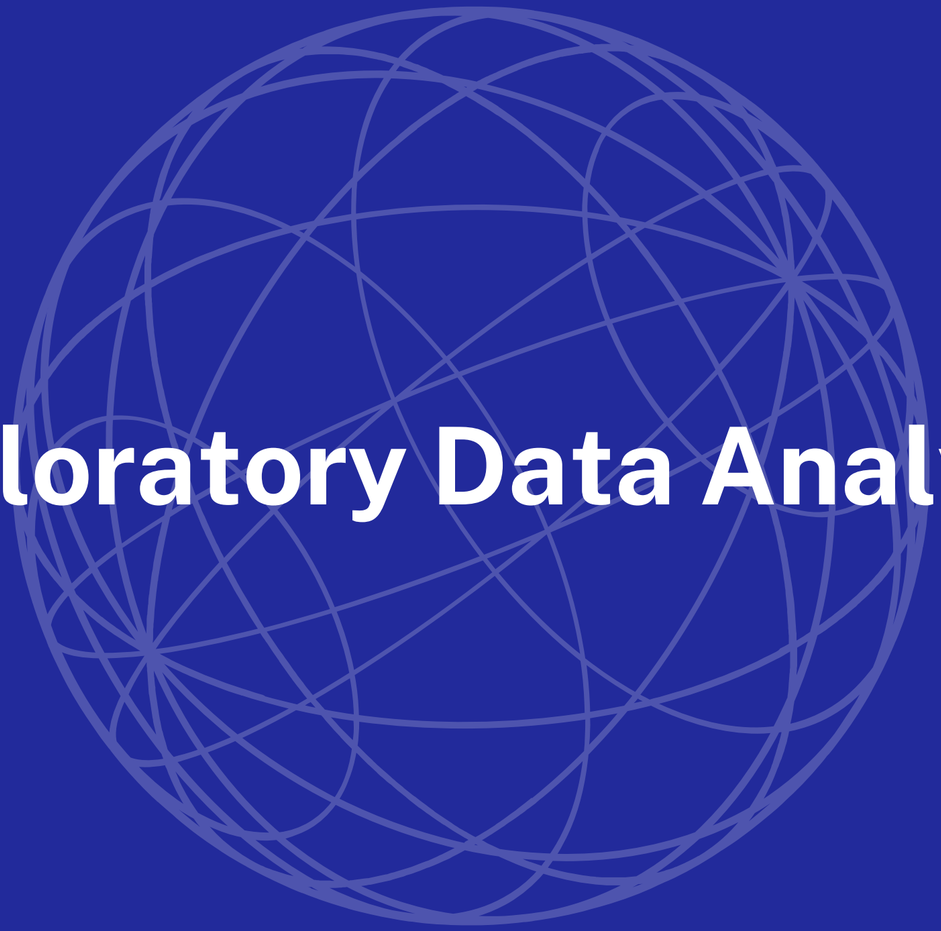
16,598 video games records with **11** features, including target variable **global_sales**.

Video Games Dataset

800 video games records with **22** features.

Methodology

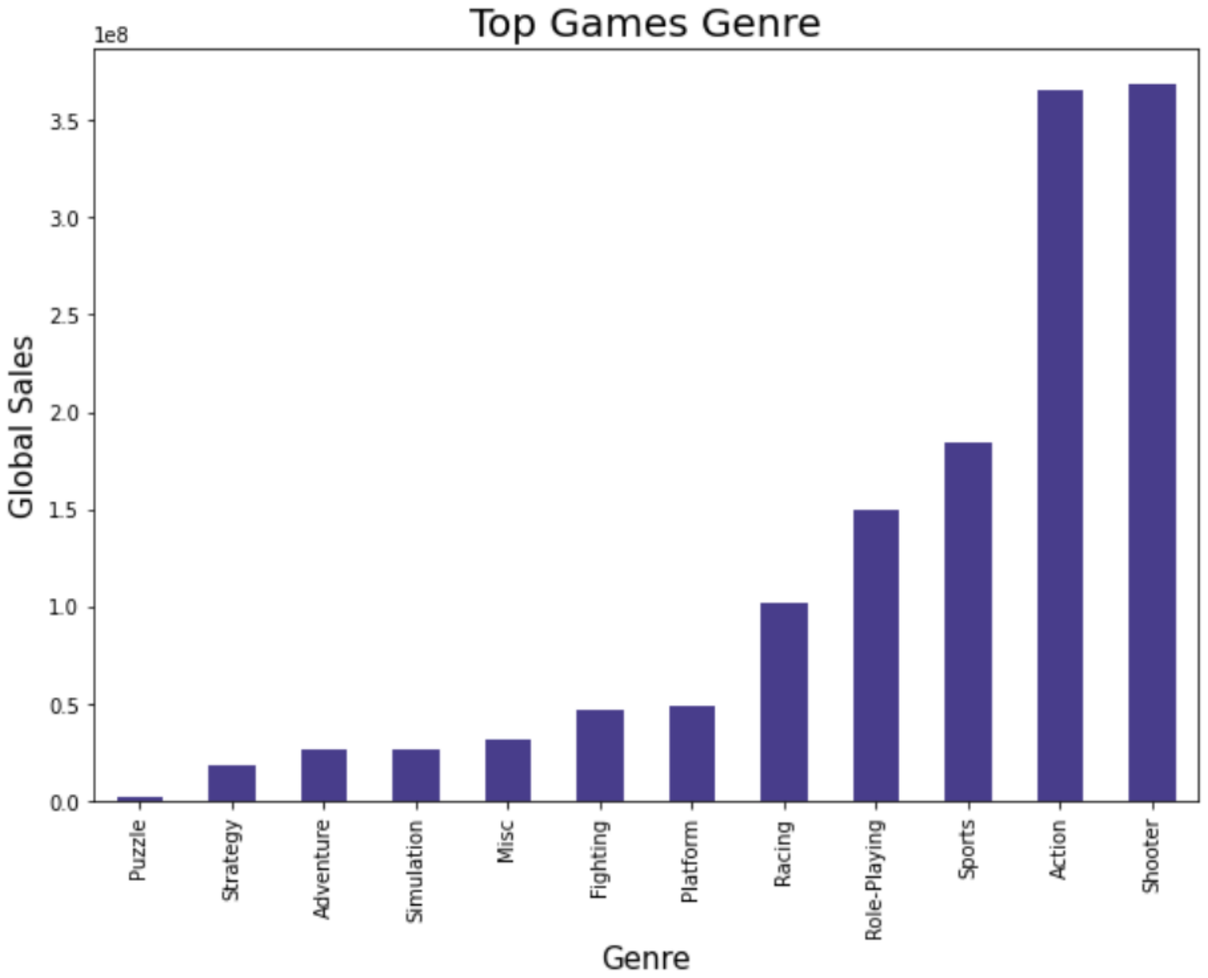




Exploratory Data Analysis

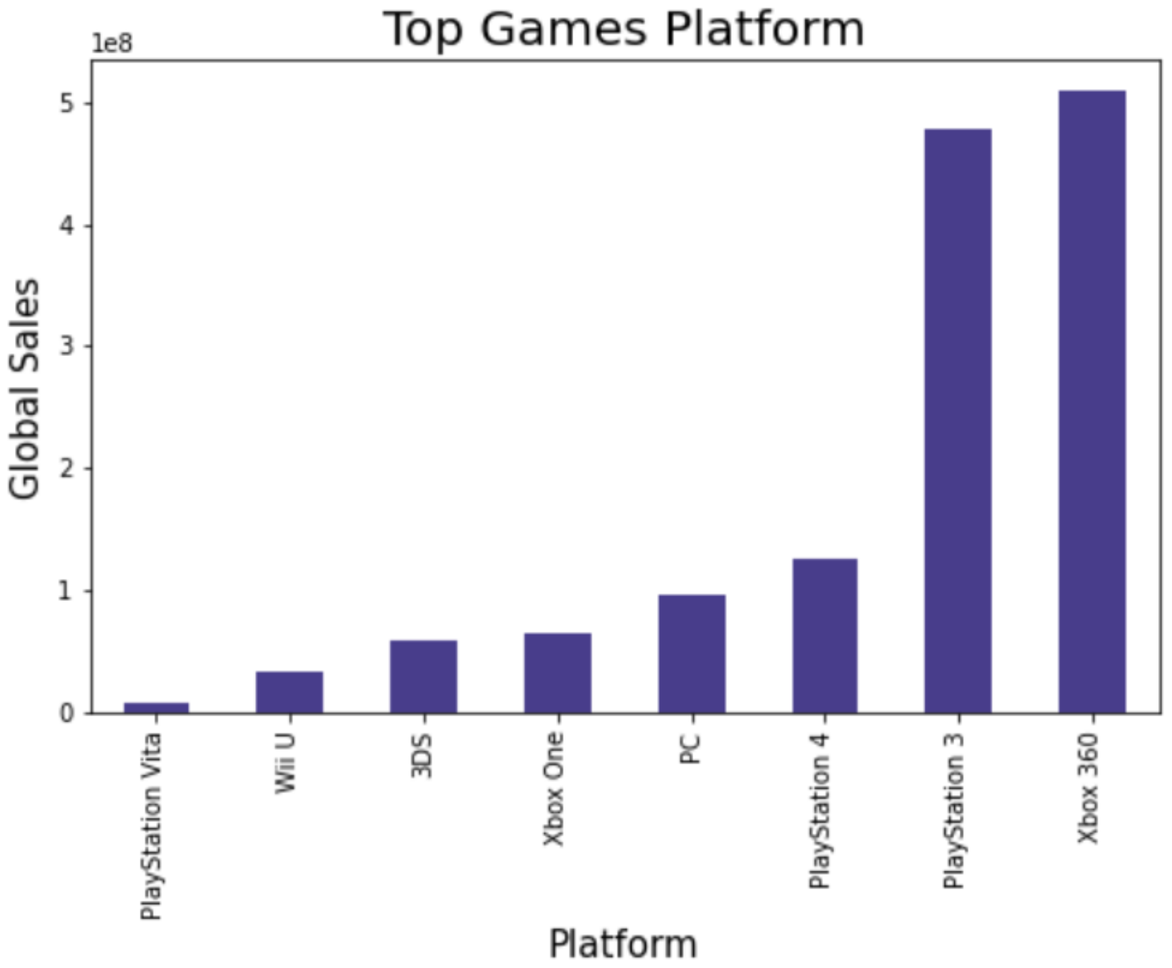
Top Video Games Genre

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.

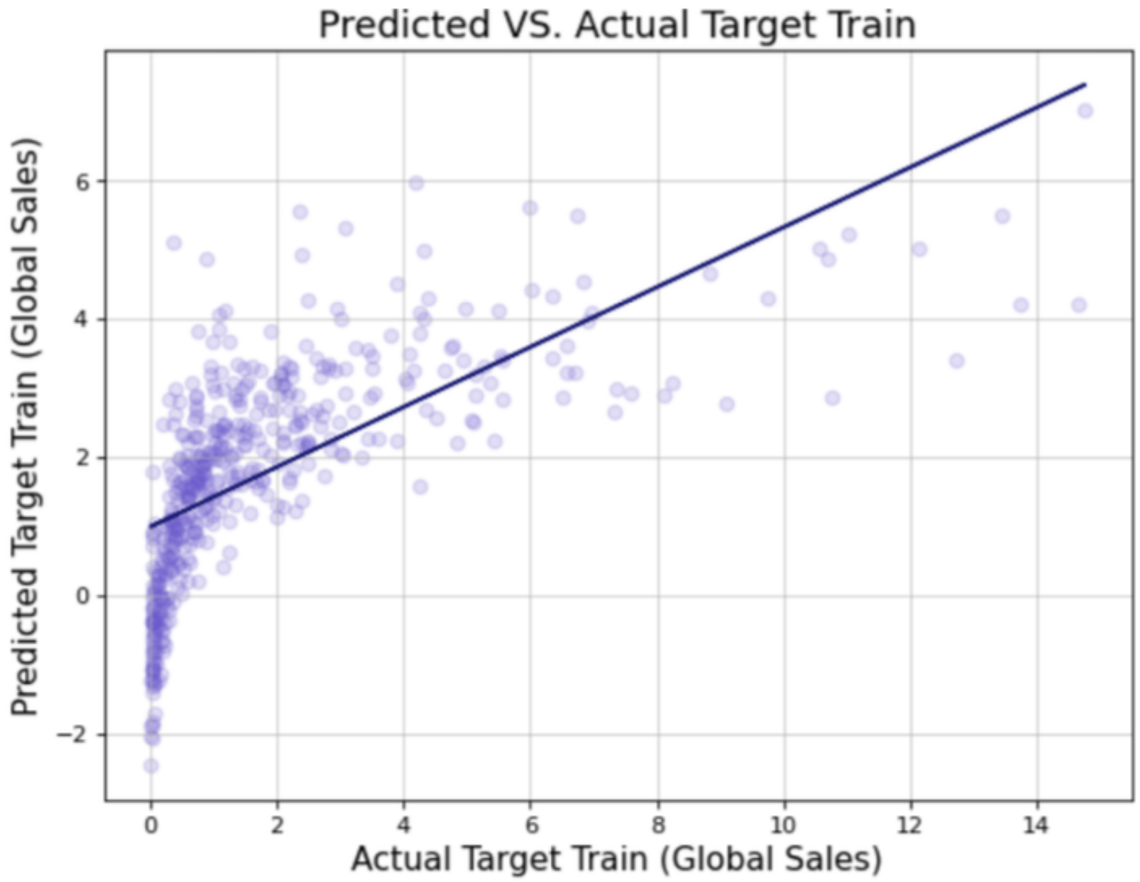


Baseline Model

Linear Regression Baseline Model

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

***global_sales* has a high variance.**



Feature Engineering

Create dummy variables for the categorical features

such as game genre, game rate, and game platform.

Handle outliers in the target variable global_sales

outliers were removed per platform.

Scaling

Scale the continous values in the target variable global_sales

Create a new feature platform_count

represents the number of games per platform.

Regression Algorithems

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

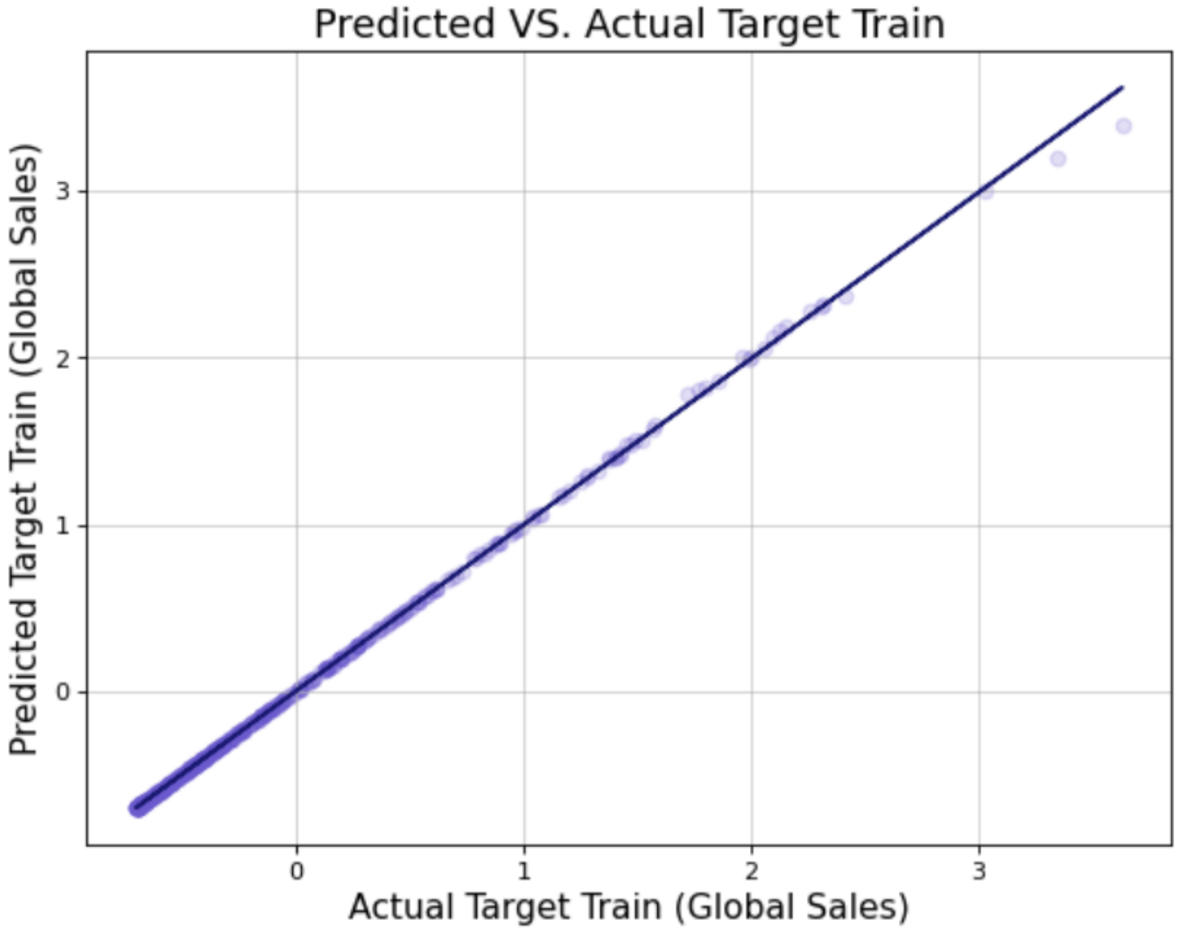
With cross validation

Data split into %90 training and %10 testing.

Random Forest Regression

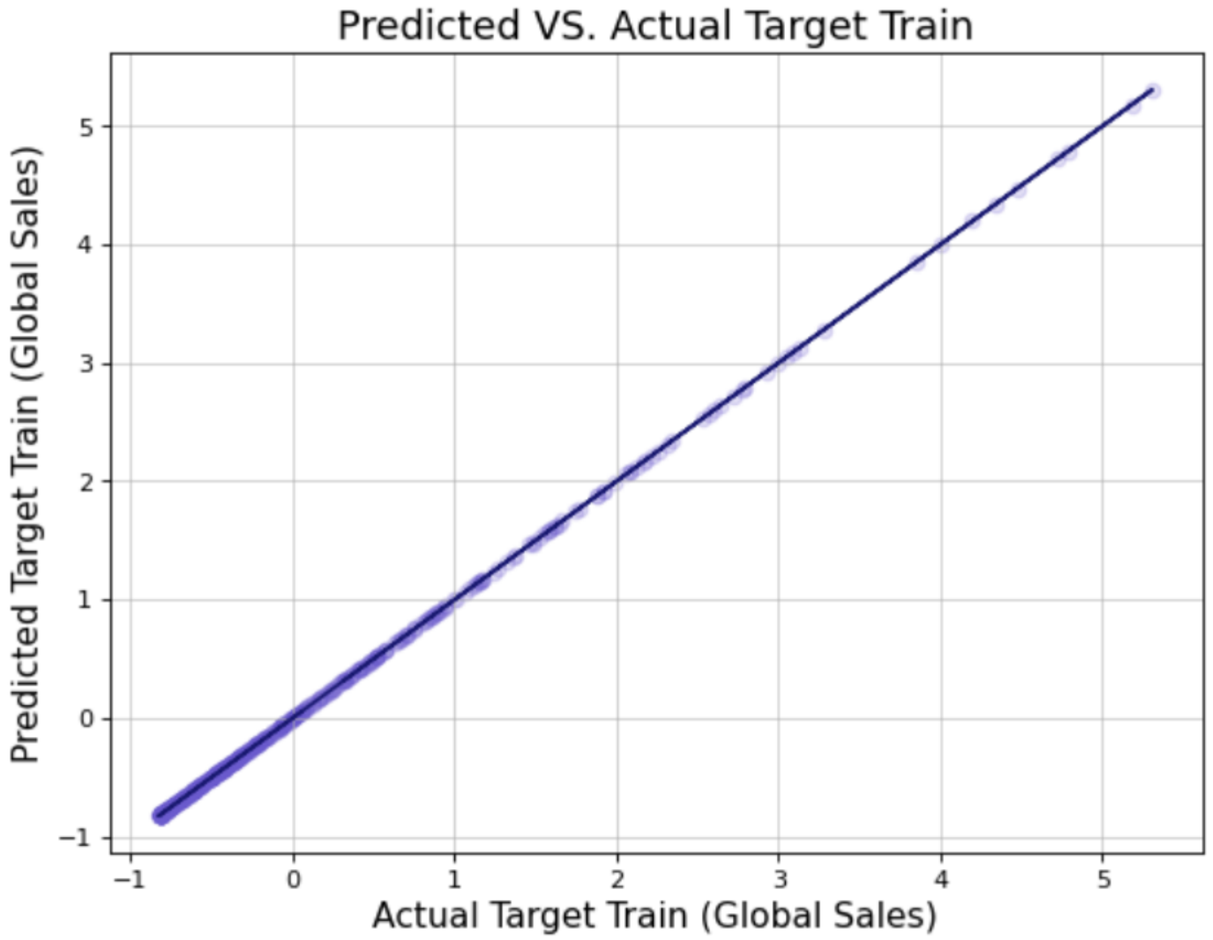
Training Score = 0.999579

Validation Score = 0.987361



Decision Tree Regression

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



Conclusion

Random Forest Regression algorithm has the best results!

- The target variable 'global_sales' with a large spread of values may result in making the learning process unstable.
- Consider improving the model stability and performance by scaling.
- Consider other algorithms for future work.

Thank you

Do you have any questions?