Exploring Global Video Games Sales & Ratings

Contents

Introduction

- The video game industry has been experiencing significant growth over the past few years, with global sales reaching record levels.

- The growth of the video game industry can be attributed to the increasing popularity of mobile gaming, the rise of esports, and the release of new gaming consoles such as the PlayStation 5 and Xbox One.

- With the rise of new technologies such as virtual reality and cloud gaming, the video game industry is expected to continue its growth trajectory in the years to come.

About the Dataset

- This dataset provides detailed information about the games and game sales that are categorized by genre, publisher, platform and ratings.

- This dataset gives researchers the power to profoundly analyze trends within the gaming market with informative and expansive details regarding popularity.

- The dataset is used to create regression models to predict sales, visualize data and gain insights into the game details and sales. It can be used to explore the most popular genres, the most popular platforms, and more.

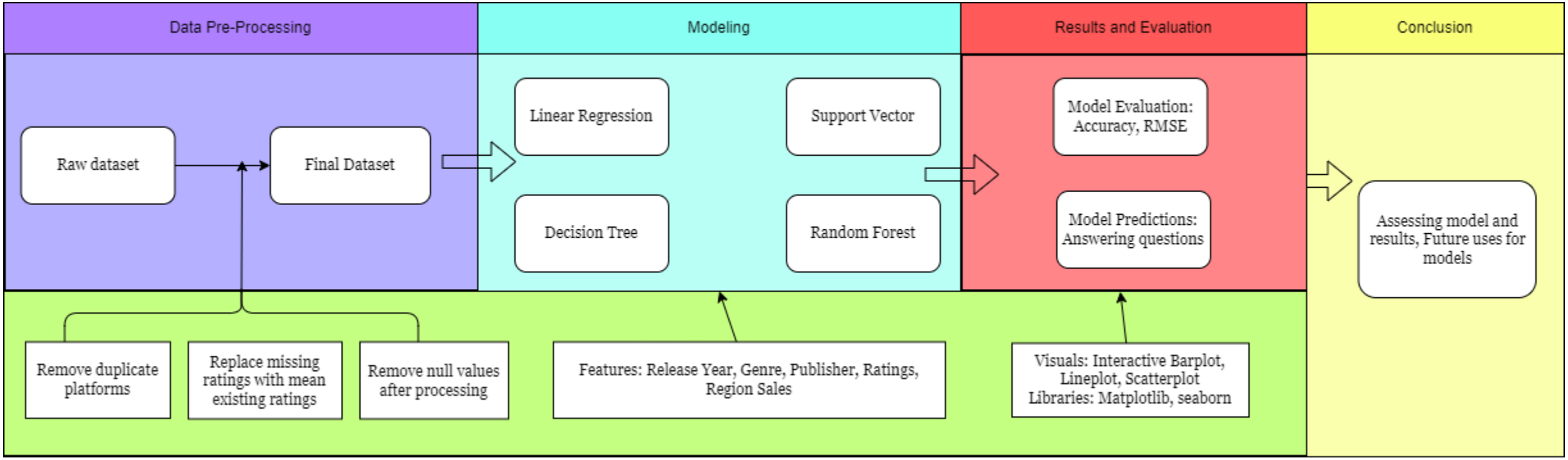
Research Questions:

1. How does a ML model's predictions based on ratings compare to actual global sales of games?

2. What is the projected trend for video game sales by top publisher in each genre for the year 2017 based on historical data from 1980 to 2016?

3. What is the impact of changing release year of a game on global sales, as predicted by a ML model?

Approach


Data Pre-Processing
Modeling
Results and Evaluation
Conclusion
Raw dataset
Final Dataset
Linear Regression
Support Vector
Decision Tree
Random Forest
Model Evaluation: Accuracy, RMSE
Model Predictions: Answering questions
Assessing model and results, Future uses for models
Remove duplicate platforms
Replace missing ratings with mean existing ratings
Remove null values after processing
Features: Release Year, Genre, Publisher, Ratings, Region Sales
Visuals: Interactive Barplot, Lineplot, Scatterplot
Libraries: Matplotlib, seaborn

Data Pre-processing:

- In the dataset, Game Names was the first column and we checked for any null values and found 2 out of 16,000 so we decided to drop it.
- We then checked for any duplicate fields with same name with platforms, and found few games duplicated with same platforms. We added all the game sales and dropped the other duplicate field.
- Next, we found out the null values in Critic and User score to be more than 8000. To decrease it as much as we can, we made a function to group same game names irrespective of their platform and replaced null scores with mean of existing scores. Then, we dropped the remaining missing fields from the dataset.

Null Values in dataset

Name	2
Platform	0
Year_of_Release	269
Genre	2
Publisher	54
NA_Sales	0
EU_Sales	0
JP_Sales	0
Other_Sales	0
Global_Sales	0
Critic_Score	8582
Critic_Count	8582
User_Score	6704
User_Count	9129
Developer	6623
Rating	6769

Data Pre-processing:

- After analyzing the data, we discovered that there were only a few missing values in the "Release Year" and "Publisher" columns. To fill in the missing years, we used web to search for the release dates of those games and compiled a list.

- Then, we wrote a function to replace the missing years and missing Publishers in the dataset with the values from the list we had prepared using web.

- As a final step, we decided to remove any remaining missing values in the dataset, as we had exhausted all options to retain them.

Null Values after Preprocessing

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
User_Count	0
Developer	0
Rating	0

Features and Feature Selection:

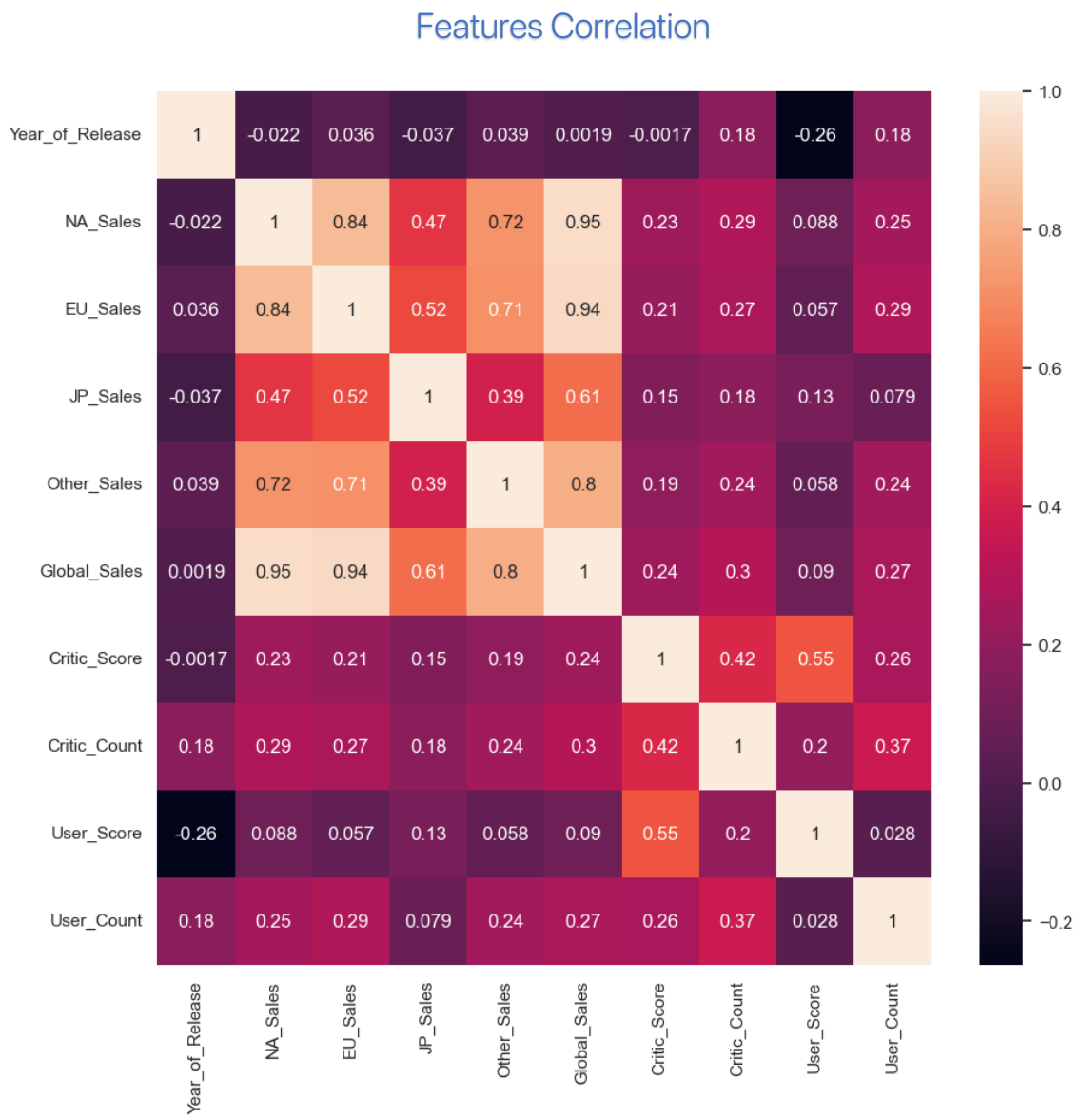
- We are using Release Year, Genre, Publisher, User Score, Critic Score, User and Critic Count and one region sales as the features.
- We will utilize some of the various features discussed above to address each of our research questions.
- Next, we will use the features we talked about to make predictions about how many video games are sold around the world. By looking at things like where people are buying games and what types of games are popular, we can make smart predictions about what might happen in the future.

Features Description

	Critic_Score	Critic_Count	User_Score	User_Count	Global_Sales
count	6823.000000	6823.000000	6823.000000	6823.000000	6823.000000
mean	70.269383	28.928917	71.857541	174.711564	0.765791
std	13.869575	19.224943	14.399744	587.511255	1.695998
min	13.000000	3.000000	5.000000	4.000000	0.010000
25%	62.000000	14.000000	65.000000	11.000000	0.110000
50%	72.000000	25.000000	75.000000	27.000000	0.290000
75%	80.000000	39.000000	82.000000	88.000000	0.750000
max	98.000000	113.000000	96.000000	10665.000000	35.520000

Features Correlation

- A moderately positive correlation between Critic_Score and User_Score is observed. This could suggest that high user scores are influenced by high critic scores, or vice versa.
- The high correlation between NA Sales and Global Sales is observed. This could be due to the huge gaming community and widespread interest in a diverse range of gaming genres.
- Game Developers and publishers should consider catering to the preferences and interests of North American consumers in order to maximize their sales potential.

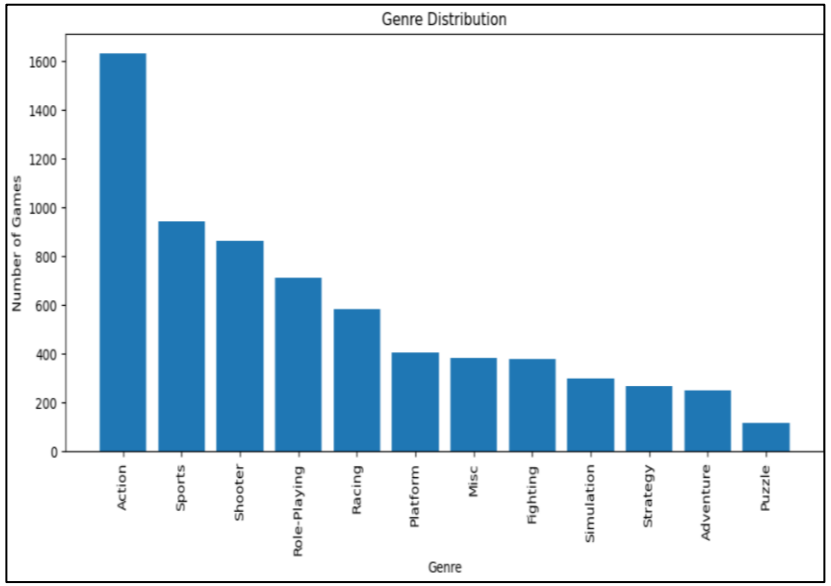


Exploratory Data Analysis:

Game Year Analysis

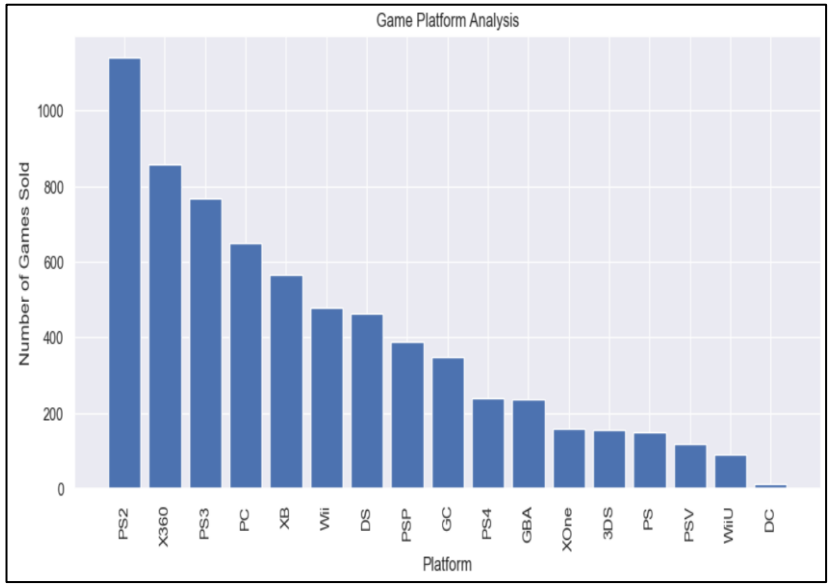
- By analyzing sales trends over time, we can identify market trends, such as the rise of mobile gaming or the increasing popularity of online multiplayer games.
- From the visualization, we can see the trend of game sales globally over time and we can see that an all-time high sales was recorded in year 2008 (672 million).

Genre Analysis



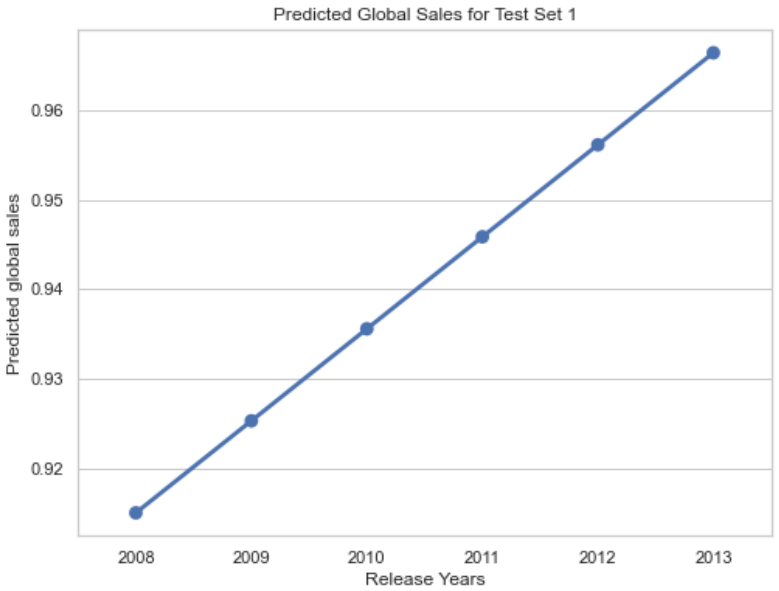
- Different genres of games have varying levels of popularity and this visual can help in marketing and game development strategies.
- We can observe from the plot that the most popular genre in our dataset is Action, with very few games falling under the Puzzle genre.

Game Platform Analysis

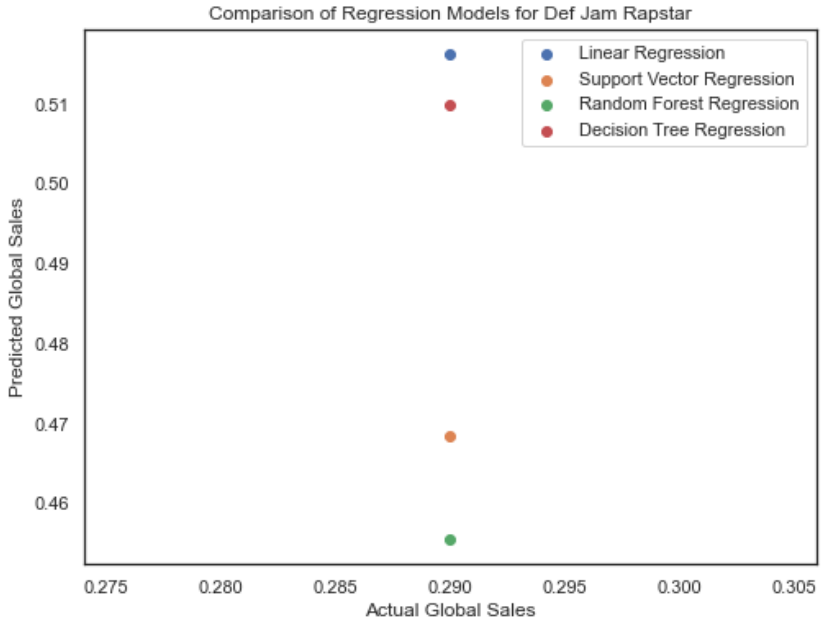


- With the rise of mobile gaming and the popularity of consoles like the PlayStation and Xbox, understanding platform preferences can provide valuable insights into consumer behavior.
- We can observe from the plot that the most popular platform in our dataset is PlayStation2, with very few games on the DC Platform.

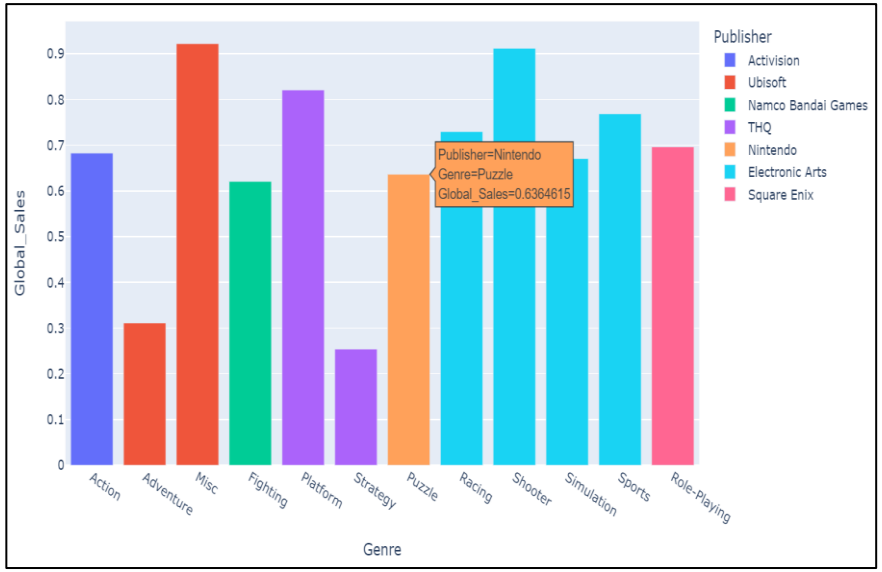
Exploratory Data Analysis:



- In this plot we can see the global sales of a game that released in 2008.
- We used our model to forecast the game's sales within the next 5 years, based on its release year.



- In this plot we can see the actual sales vs predicted sales of a particular game.
- We have used Linear, Support Vector, Random Forest, Decision Tree Regression models for the plot.



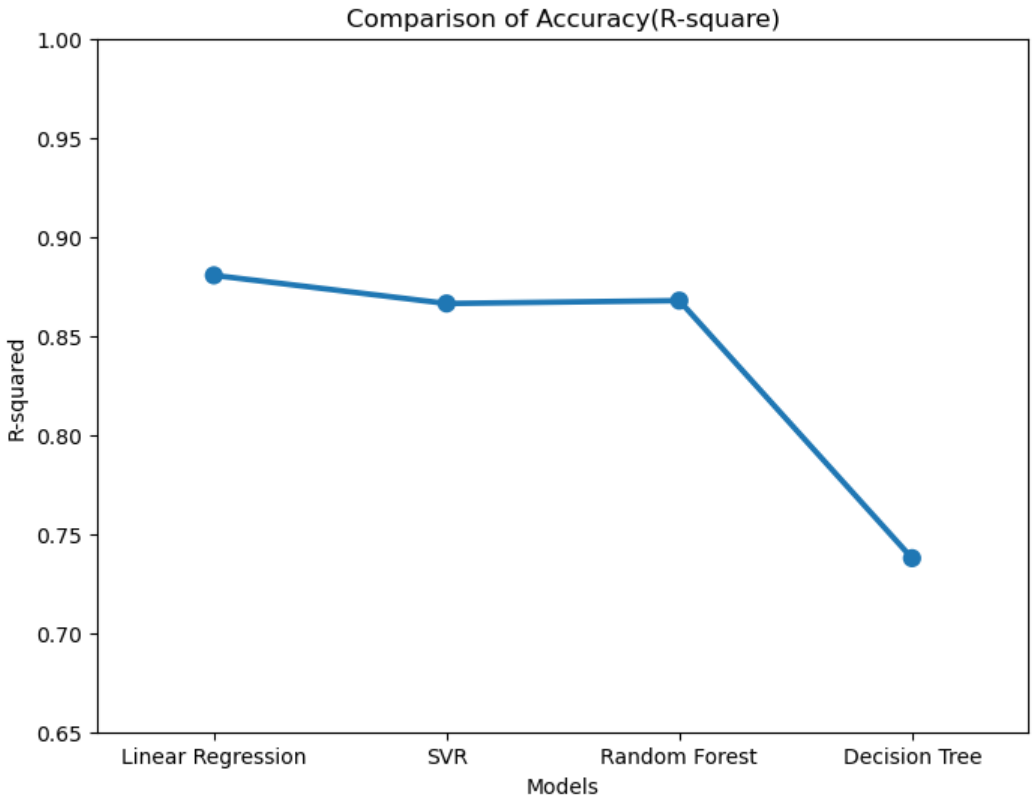
- This interactive bar plot displays the highest average global sales publishers by genre.
- We will further predict the sales of top publishers by genre using a regression model and compare it with the actual sales.

ML models Evaluation

- We utilized regression models to analyze our dataset and predict global sales of video games based on various features.

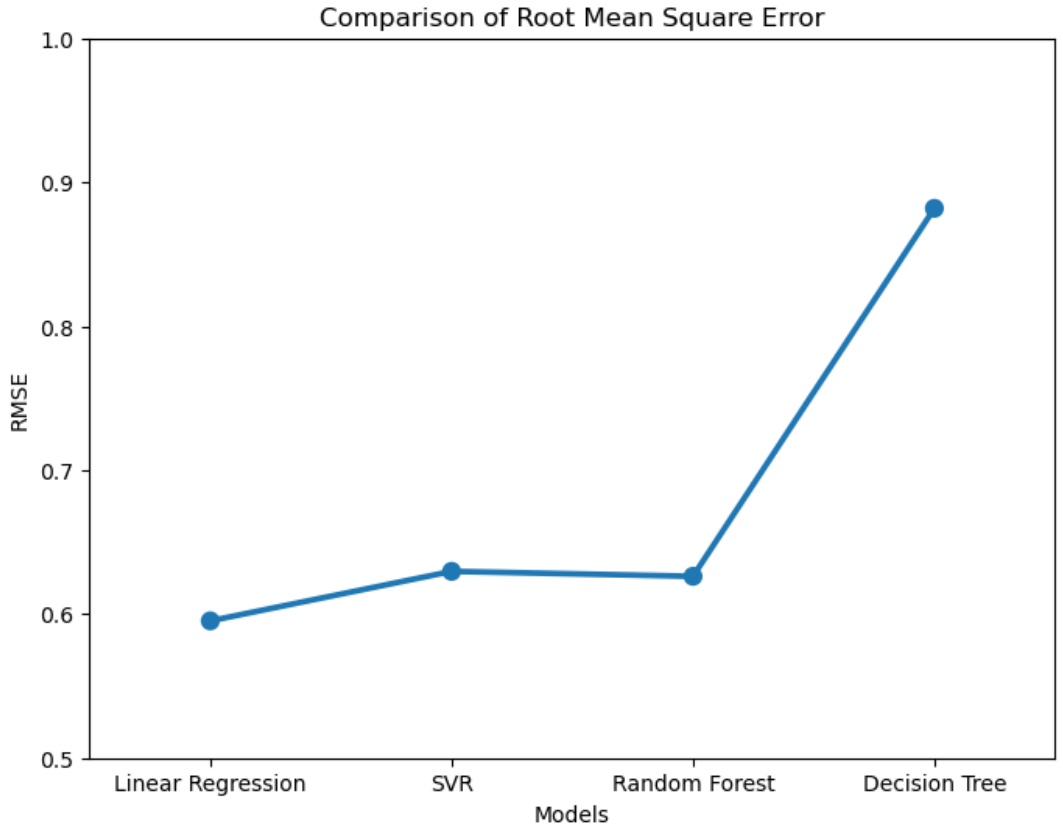
- The models we used were Linear Regression, SupportVectorRegression(Linear), DecisionTreeRegressor, RandomForestRegressor.

- Linear Regression Model's accuracy was the highest(88%), followed by RandomForestRegressor(86.7%), SVR (86.6%) and least of them was DecisionTreeRegressor (73%).



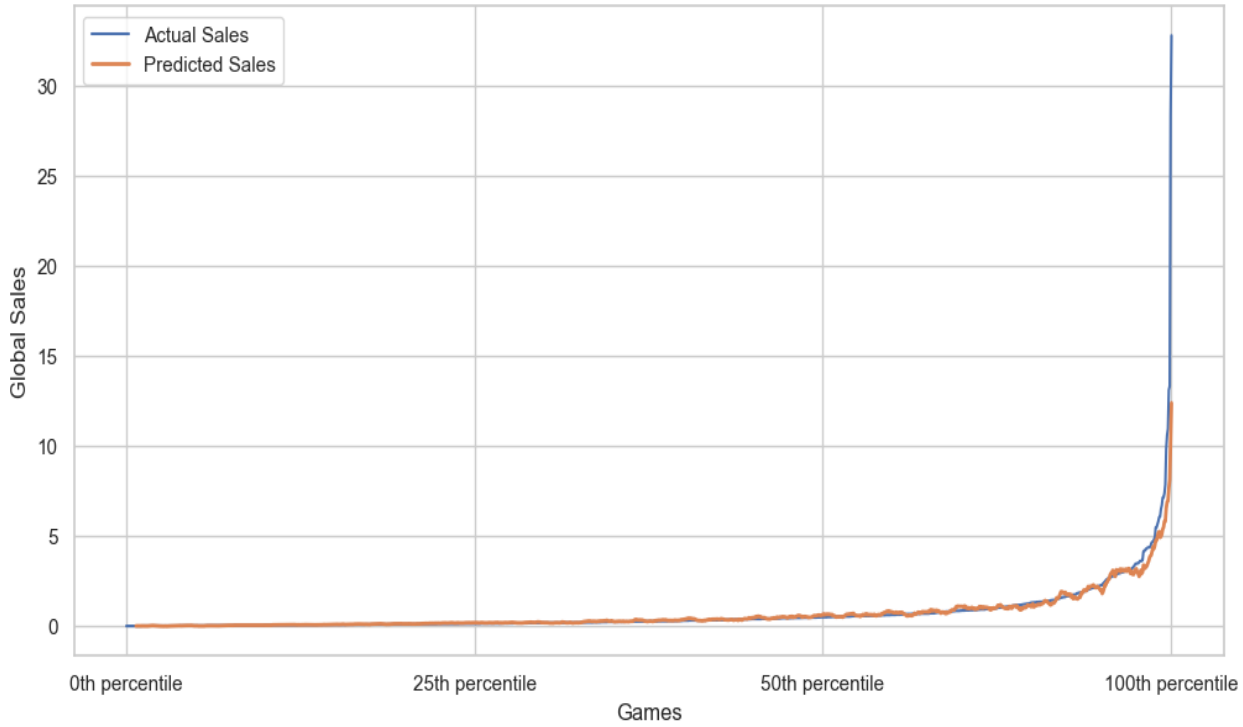
ML models Evaluation

- We can see the highest Root Mean Square Error is for DecisionTreeRegressor which had the least accuracy (R-2 Score) among the four models.
- It means that the model's predictions have a large average deviation from the actual values.
- By examining the accuracy and RMSE plots for each model, we can gain insight into the performance of each model and identify which model provides the best fit for our dataset.
- We will be using Linear Regression for answering our research questions.



Results

- Our first research question talks about how a ML model's predictions based on ratings compare to the actual global sales.
- We decided to use Critic and User Scores and North America Sales as features for our linear regression model to predict Global sales.
- Once the model was fit, we generated a new dataframe that included the game name, actual global sales, and predicted sales based on the model.
- Then we created a line graph which compared the actual global sales vs predicted global sales.

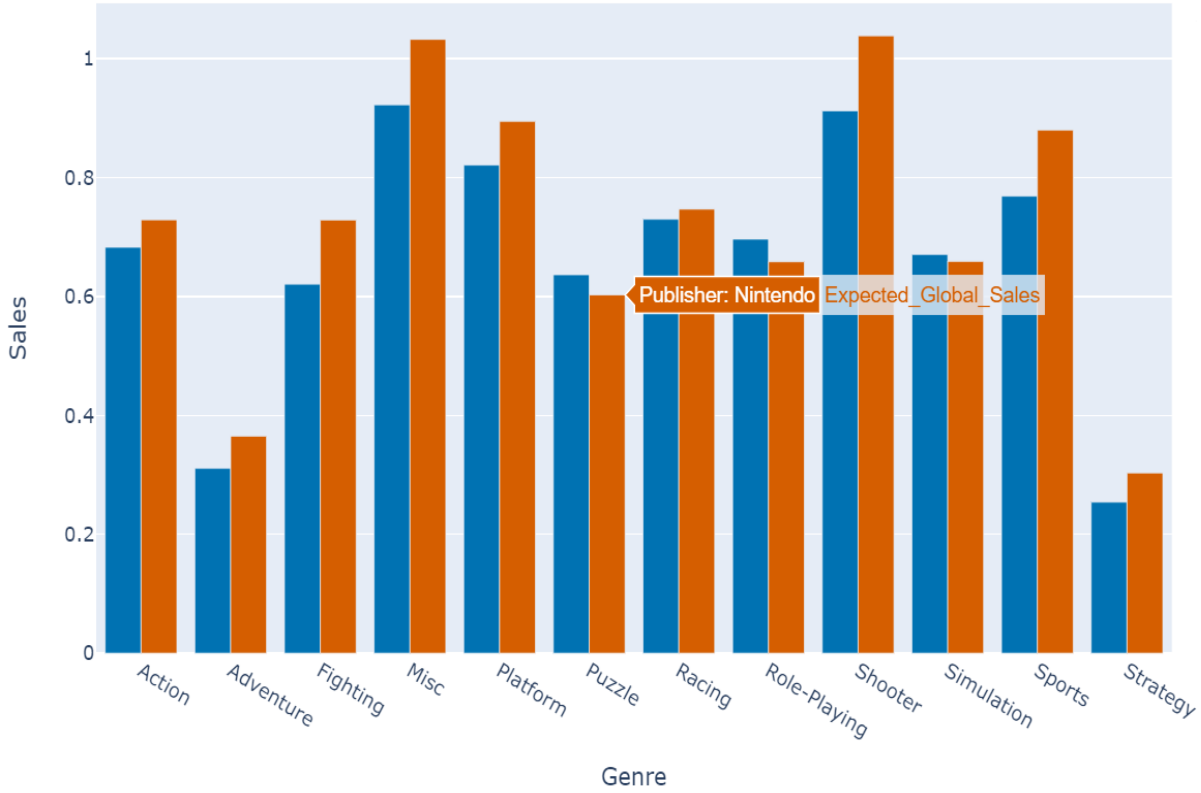


Results

- Our second research question focuses on analyzing the sales trend of the top publishers for each genre in the year 2017, based on the data spanning from 1980 to 2016.

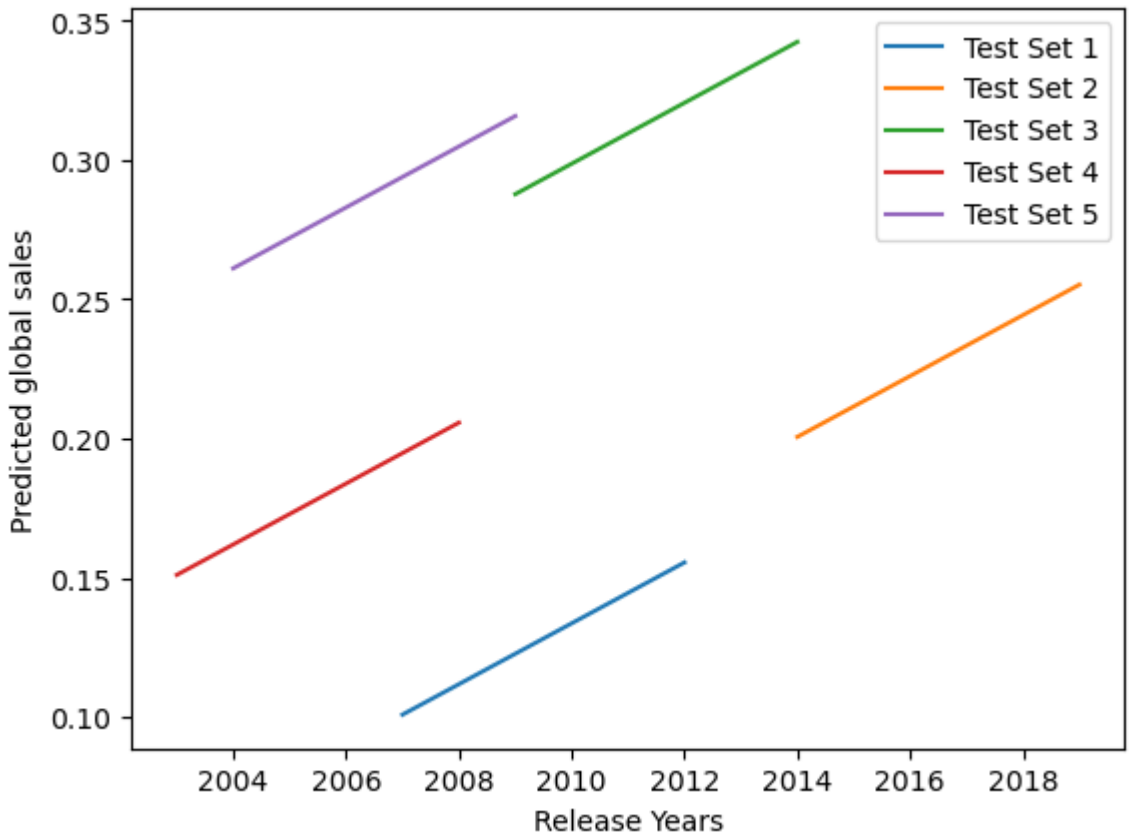
- We decided to use Release Year, Genre, Publisher and North America Sales as features for our linear regression model to predict Global sales. We used LabelEncoder for text data.

- Our visual shows that the sales of almost all publishers in genres increased in 2017. However, the sales of publishers for Puzzle, Role-Playing, and Simulation genres decreased slightly in the same year.



Results

- Our third research question talks about how changing the year of release affects the sales of a game. We took 5 random games as a sample.
- We decided to use Release Year, Critic and User Scores and North America Sales as features for our linear regression model to predict Global sales.
- Once the model was fit, we generated 5 test sets each with release year accompanied with 5 consecutive years, then predicted sales based on the model.
- Our visualization indicates that the sales of all games are projected to increase in the next five years if they are released during that period. This trend was observed in all the test cases.



Conclusion

- Understanding the video game industry through exploration and modeling of global game sales data can help guide business decisions in a variety of industries.

- The dataset we analyzed allows us to identify trends in game sales and predict how games might be affected if released in the upcoming years. By analyzing historical data, we can make informed projections about the market and make strategic decisions about game development and release timing.

- As the video game industry continues to evolve and grow, the importance of data-driven decision making will only increase, making global game sales data an important tool for success.

Conclusion

- Despite the growth of the industry, there are still challenges to consider, such as piracy and market saturation, and it is important for game developers and publishers to continue innovating and adapting to changing player preferences.

- The emergence of new VR and AR games has the potential to expand the player base and increase engagement. Future work could involve analyzing these new platforms and games to better understand their impact on the gaming industry and how they can be leveraged for growth.

- Overall, the global video game industry shows no signs of slowing down, and there are many opportunities for growth and innovation in the years to come.

References

[1] Global Games Sales & Ratings | Kaggle: https://www.kaggle.com/datasets/thedevastator/global-video-game-sales-ratings

[2] scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation. (n.d.). Scikit-learn: Machine Learning in Python &Mdash; Scikit-learn 1.2.2 Documentation. https://scikit-learn.org/stable/

[3] Matplotlib — Visualization with Python. (2023, February 14). Matplotlib — Visualization With Python. https://matplotlib.org/

[4] seaborn: statistical data visualization — seaborn 0.12.2 documentation. (n.d.). Seaborn: Statistical Data Visualization — Seaborn 0.12.2 Documentation. https://seaborn.pydata.org/

[5] Plotly. (n.d.). Plotly Python Graphing Library. https://plotly.com/python/

Thank you