Brandon Mather

Predicting Game Sales Based on Genre Popularity

Summary:

This paper presents an approach to predicting video game sales based on genre popularity. Given the video game industry's nature and its multi-billion-dollar market value, accurately forecasting the success of new game releases is important for developers and publishers. By analyzing historical sales data, this study identifies trends in genre popularity and employs a predictive model to forecast future sales.

Business Problem:

The video game industry is a multi-billion-dollar market where predicting the success of a new game can help make business decisions. Understanding the trends in genre popularity over the years can provide insight into future game success. The approach used in this paper will help game developers and publishers make informed decisions about which genres to focus on for future releases.

History:

The video game industry has grown dramatically since it first started. The market has seen numerous shifts in consumer preferences. Major players in the industry, such as Nintendo, Sony, and Microsoft, have continued to adapt to these changes, often predicting and shaping trends. However, predicting the success of a new game remains a challenge due to the nature of consumer interests and technological advancements. Using historical sales data, we can try to

identify patterns and predict future success based on genre, offering a data-driven approach to the competitive landscape of the video game industry.

Data Explanation:

The dataset being used is a list of video game sales data, including features such as the game's name, platform, release year, genre, publisher, and regional sales figures (North America, Europe, Japan, and other regions). The primary target variable is Global sales, which is the sum of all the regional sales.

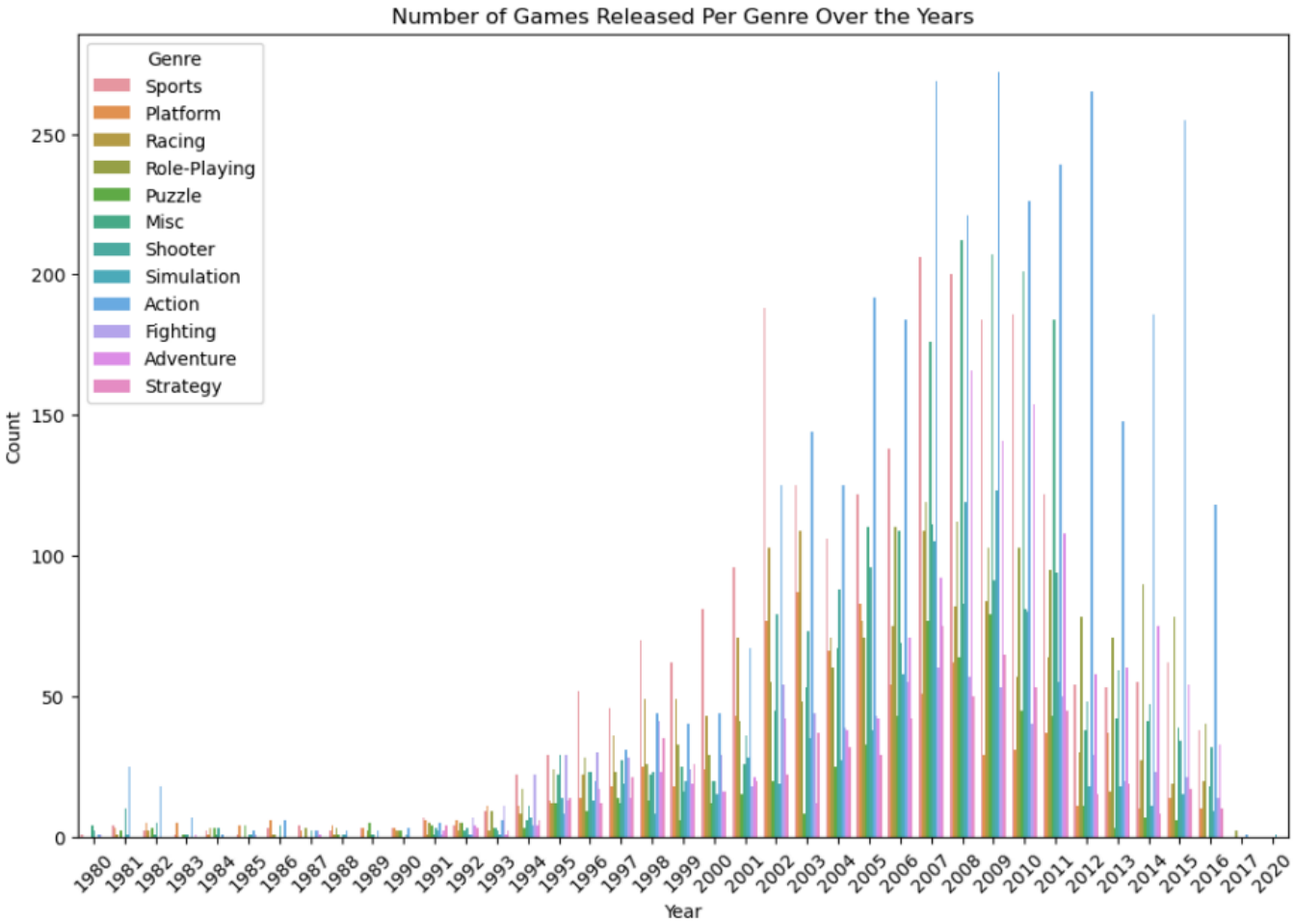
Methods:

The methods that were used were preprocessing, Exploratory Data Analysis (EDA), and predictive modeling. Preprocessing involved several steps including handling missing values and converting the year column to datetime. The EDA involved visualizing the data to identify trends, patterns, and relationships. To predict the success of new games, a Random Forest Regressor model was used. This method combines multiple decision trees to improve predictive accuracy. The process involved selecting relevant features, training the model on the historical data, and assessing the model performance using Mean Squared Error (MSE) and R-squared score, then using the trained model to predict future sales based on genre and other relevant features.

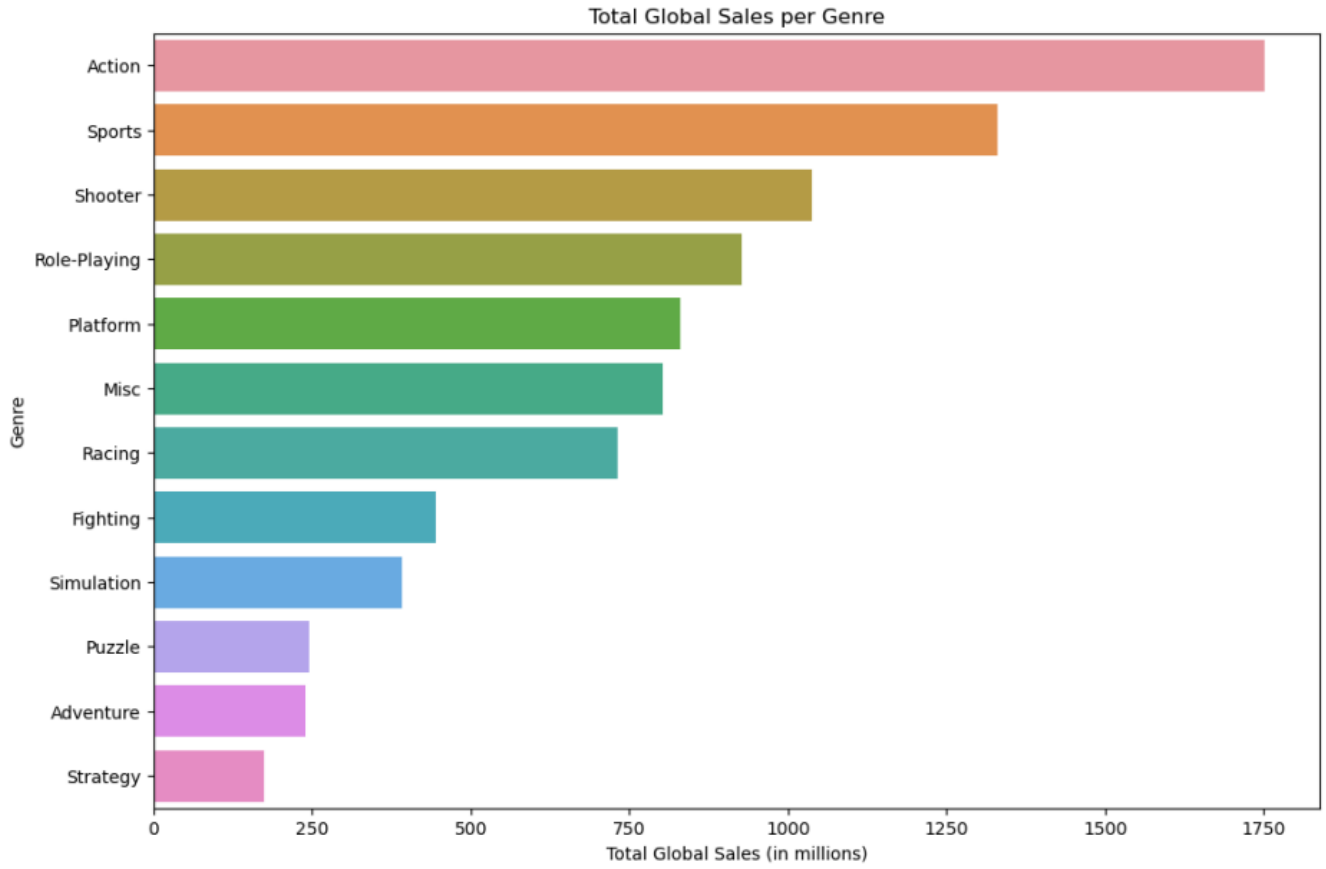
Analysis:

The first thing that needed to be done before the analysis was the preprocessing. What we had to solve were missing values and converting the year to datetime for the model itself. The only missing values were in the Year and Publisher category, I didn't feel there were any good replacements for those, so I got rid of those rows, seeing as a small number of games.

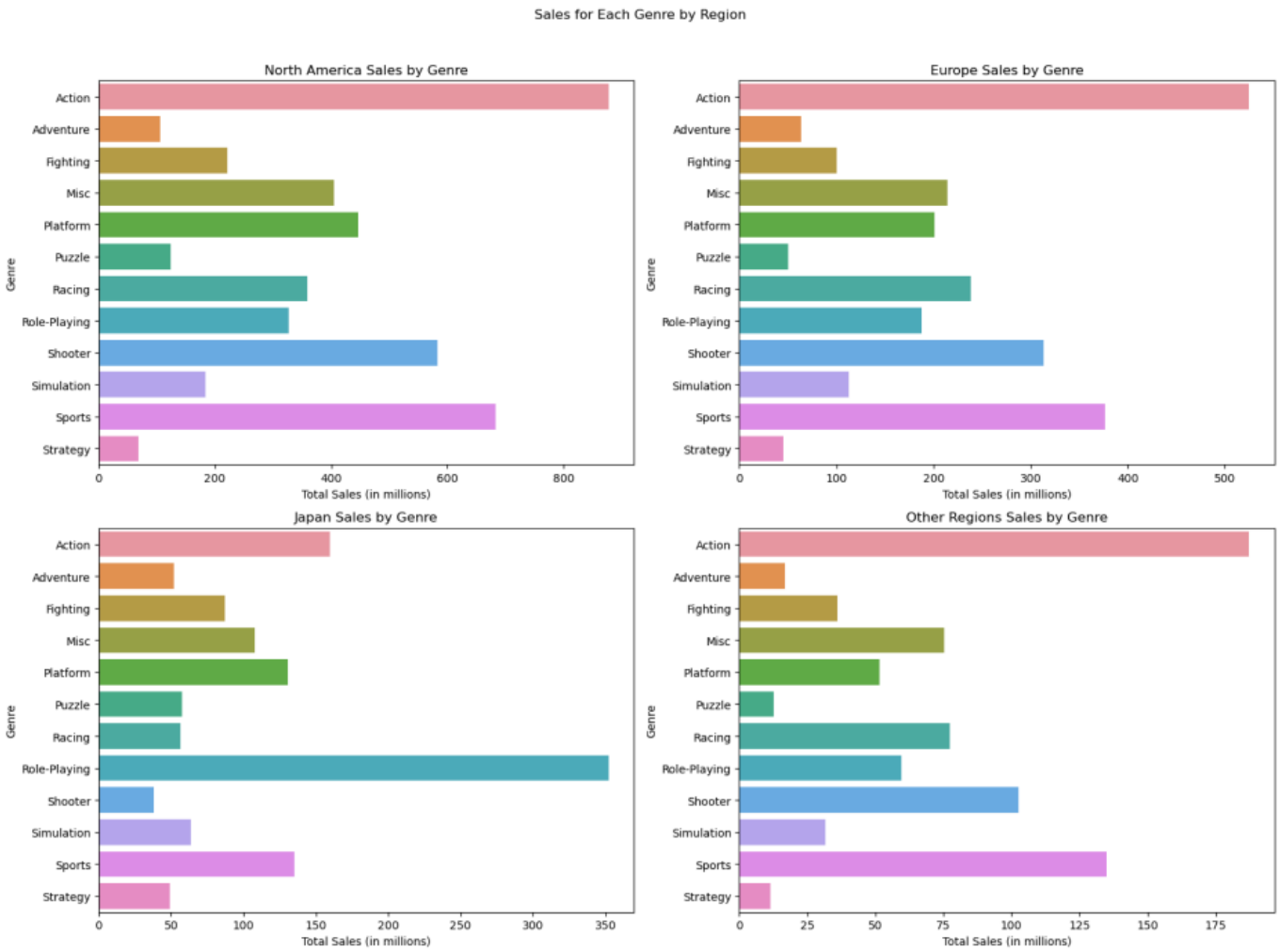
After cleaning the data, I went into my Exploratory Data Analysis (EDA). Since the focus of the model is based on genre popularity, I first decided to see how the number of games per genre have gone up or down over the years:



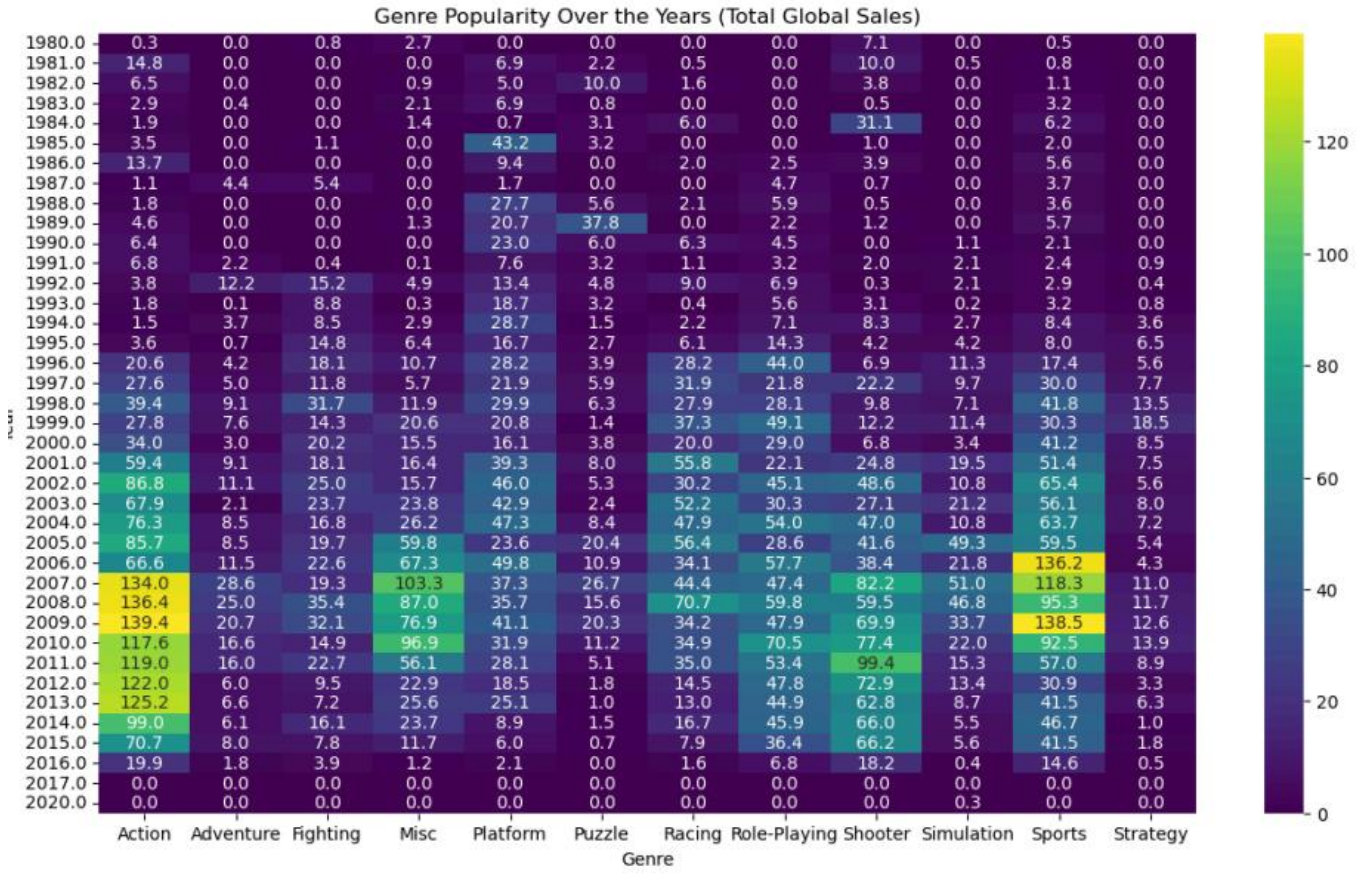
Here we can see that games in general started to skyrocket in the mid 2000's and can see the Action genre having some of the most releases each year in that time. Now that we can see how many games are releasing, lets see how they are selling:



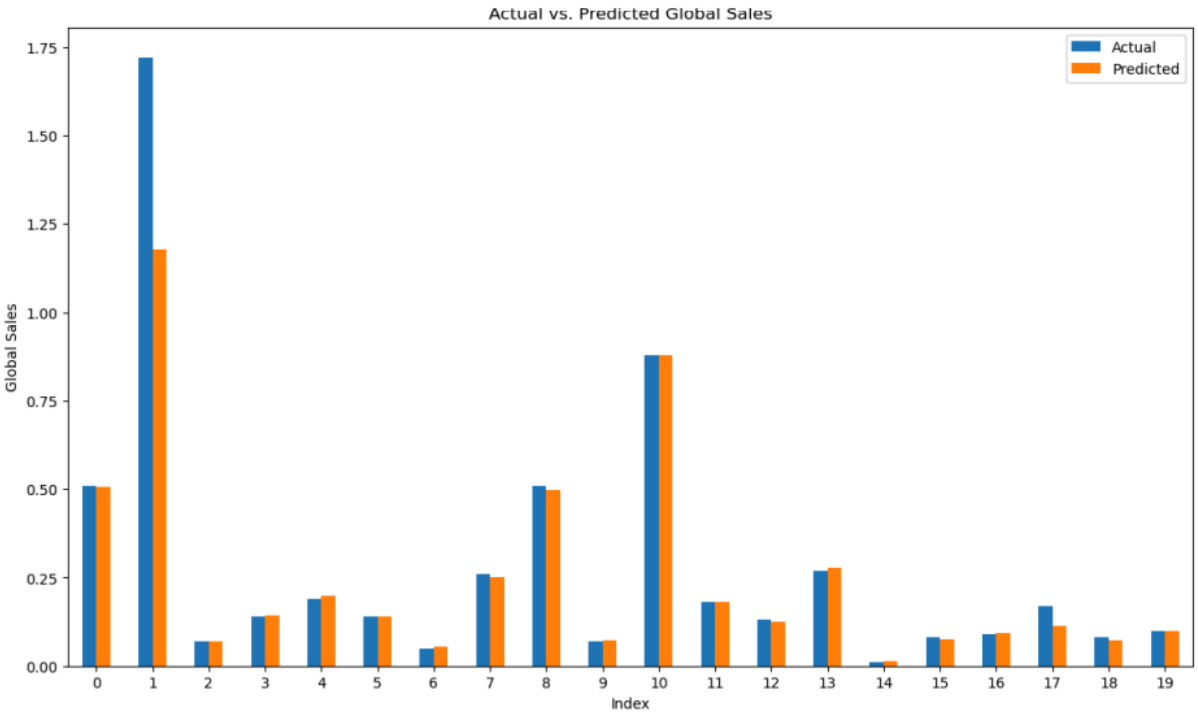
Here we can see globally that Action, Sports, and Shooter are the most popular genres in sales. This could be a correlation with why we started to see a rise in Action game releases in the 2000's. While it's important to see global sales, it also could be helpful to see how each region looks in sales:

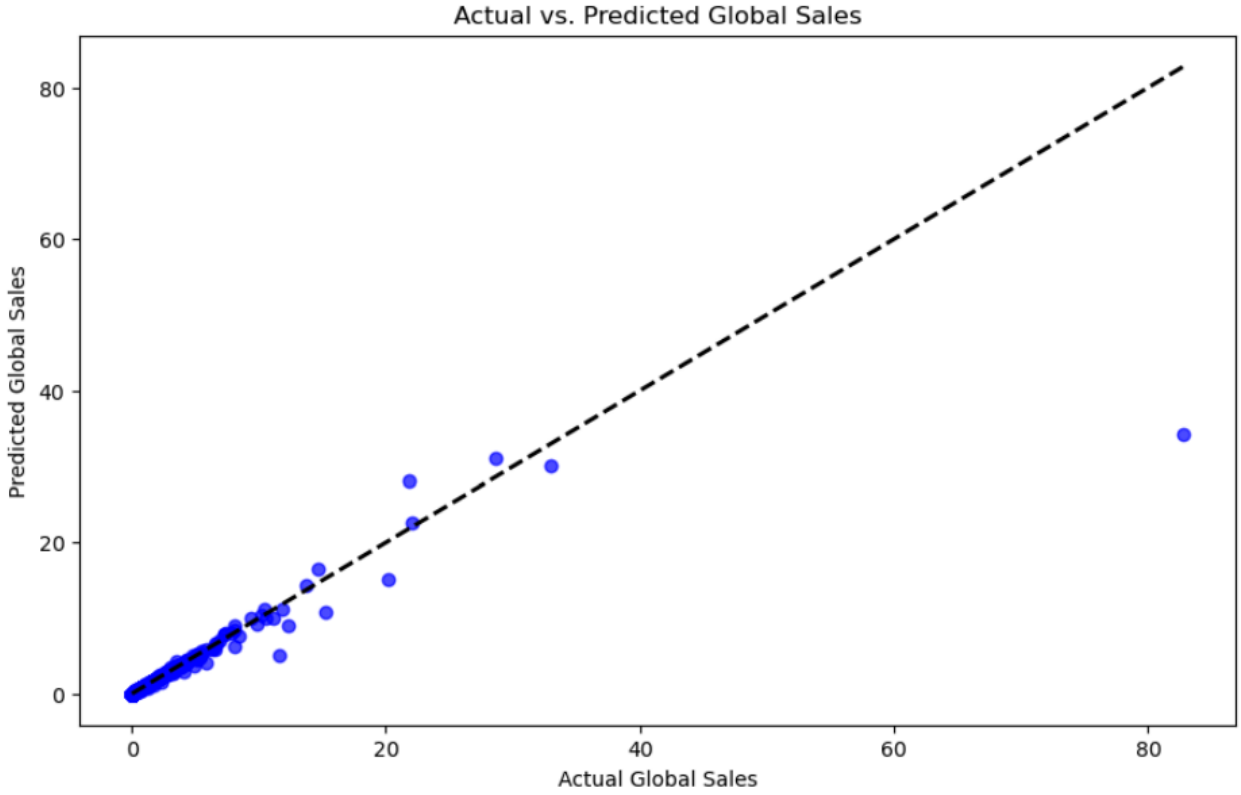


Here for example, you can see that while the Action genre is overall the highest selling, that's not true for every region. You can see in Japan that Role-Playing is the highest selling. The last thing I wanted to see in my EDA was the genre popularity over the years. This heatmap will show what were able to see before in the number of releases:



After my EDA I went into my modeling, I used a Random Forrest Regressor model to predict the sales based on genre popularity. The Random Forest model, while robust, exhibited an R2 Score suggesting room for improvement in accuracy. This implies that while genre and year are significant predictors, other factors also significantly impact a game's success. I also decided to visualize the accuracy of my model:


Actual vs. Predicted Global Sales
Actual
Predicted
Global Sales
Index
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19


Actual vs. Predicted Global Sales
Predicted Global Sales
Actual Global Sales
0
20
40
60
80

Conclusion:

This analysis confirms that genre popularity significantly influences game sales. Historical data can provide valuable insights into future trends. This data-driven approach can help game developers and publishers make informed decisions about game development and marketing strategies.

Assumptions:

Several assumptions were made in this analysis. It is assumed that historical sales data is a reliable indicator of future trends, that genre popularity is a primary driver of games sales, and that the dataset is representative of the market.

Limitations:

There were some limitations in this analysis. Data coverage might not include all games released, leading to potential biases. External factors such as marketing, game quality, and economic conditions were not accounted for either.

Challenges:

Challenges encountered included ensuring the accuracy of the dataset, identifying relevant features to improve model performance, and choosing the appropriate model for optimal performance.

Future Uses:

Future application of this model includes real-time sales prediction by integrating the model with real-time data to provide ongoing sales forecasts, genre-specific marketing strategies

based on predicted genre trends, and platform-specific analysis by extending the model to predict sales based on platform trends.

Recommendations:

It's recommended to incorporate additional data sources, such as marketing spend and user reviews, to improve model accuracy. Exploring advanced modeling techniques like neural networks could capture more complex patterns. Continuous monitoring and regularly updating the model with new data will help maintain its accuracy.

Implementation Plan:

The implementation plan involves four phases. Phase one focuses on data collection and cleaning. Phase two involves model development, training the predictive model and evaluating it. Phase three covers integrations, where the model is integrated with existing systems, and a user interface for real-time predictions is developed. Phase four focuses on monitoring and maintenance with new data and monitoring its performance.

Ethical Assessment:

Ethical considerations include ensuring data privacy by complying with privacy regulations and guidelines, maintaining transparency in how predictions are made and used, and addressing potential biases in the data to ensure fair and accurate predictions.

References:

GregorySmith. (2016, October 26). *Video game sales*. Kaggle. https://www.kaggle.com/datasets/gregorut/videogamesales

10 Questions:

1. What motivated this study on predicting video game sales based on genre popularity?

   ANSWER: The motivation behind this study stems from the significant financial stakes in the video game industry, where accurately predicting the success of games can greatly influence decision making.

2. How was the data for this study collected and prepared?

   ANSWER: The data was collected from public video games sales data.

3. Why was a Random Forest Regressor chosen for the predictive model?

   ANSWER: This type of model was chosen because it is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and robustness. It can handle many input features and is less prone to overfitting compared to other models.

4. How accurate is the predictive model, and what metrics were used to evaluate it?

   ANSWER: The model's accuracy was evaluated using Mean Squared Error (MSE) and R-squared (R2) score. The model showed good accuracy, indicated in its R2 score, there is room for improvement.

5. What strategies are in place to address biases?

   ANSWER: Strategies include regularly monitoring the predictions for signs of bias and take corrective actions when necessary, maintaining transparency about the data sources

and algorithms used, and incorporating fairness metrics to evaluate the model's performance.

6. How does the model handle privacy and data security, and what measures are taken to ensure compliance?

   ANSWER: The model complies with data protection regulations ensuring that data handling practices meet legal requirements.

7. How can game developers and publishers use the insights from this study?

   ANSWER: They can use the insights to prioritize genres with high sales potential, tailor marketing strategist, and allocate resources more effectively. The model can also aid in planning for future game releases by identifying emerging trends and consumer preferences.

8. What ethical considerations were considered in this study?

   ANSWER: Considerations included ensuring data privacy, maintaining transparency in the prediction process, and addressing any potential biases in the data.

9. What are best practices for implementing and maintaining the recommendation model in a real-world setting?

   ANSWER: There are several best practices that can be followed including continuously collecting and integrating new data into the model, monitoring the model's performance using MSE and R2 to identify any decline in accuracy, ensure the model is scalable to

handle increasing data volumes, and develop a user interface that can be interacted with to obtain predictions easily.

10. Do you have any unanswered questions that warrant further investigation?

ANSWER: Unanswered questions would include: How do external factors affect game sales?, What role do emerging gaming platforms play?, how can user engagement metrics improve predictions?, and How do regional preferences evolve over time?.