AHMET FURKAN KARAOGLU - CSE603
JEPBAROVA DURSUN - CSE603

INTRODUCTION

Video game prototypes started showing themselves in the 1960s. But Atari's Pong game (1972) gave a big boost to the video games industry to start. Between 1970-1983 years, the videogame industry started growing with the home consoles and Atari machines. One of the well-known games from these years are Pac-Man, Donkey Kong, and Pong. Cause of the new technology between the years 1985-2000 the video game industry changed quickly. The companies who made large amounts of investments on home consoles started losing money. Because the personnel computers started getting more popular in the video game industry (Commando, Wolfenstein, Doom). But Nintendo spent a lot of money and effort to build Nintendo Entertainment System (NES) console. While personal computers were getting more common on the video game market, Nintendo released NES and stayed on the market as a home console. And they started releasing high quality games for their console. Such as Super Mario Bros, Duck Hunt, Excitebike and more. From 2001- present the video game industry spread into a lot of platforms. The computers and consoles became more powerful and started handling more complex tasks. Besides these platforms, mobile platforms and handheld consoles started getting more popular too.

As we can see, the video game industry has changed a lot over the years. Sometimes the Atari machines were popular, sometimes computers, sometimes consoles, and others. Even within the different years, different genres of games were getting popular. Our goal is to find out if there are any relations/ effects between, the genre of the games, the release date, or the platform they are coming out on the video games sales.

BACKGROUND RESEARCH

In order to understand the sales of video games, it is important to consider various attributes such as the game's platform, release year, genre, and maybe even publisher. For example, certain genres of games such as first-person shooters and sports games tend to be more popular on certain platforms, such as the Xbox and PlayStation. For example, Call of Duty, FIFA, PES and Fortnite are the mostly played games on consoles. This is an effect of combination of genre and the released platform.

There is research that is done on the same subject but more features. In this research there was platform, year, genre, critic scores, critic numbers, user scores, user count and ratings. The critic scores and numbers, and the user scores and count had important effects on sales between the regions. But only for the North America and Japan the effect oy the release year of the game had a big effect. But in the years when we get closer to the present, the significance of the release data lost its impact on sales. But there was an interesting detail of the user score effect on the sales. Because the higher user score gets it had negative effect on sales for the global sales. (*Factors That Impact Video Game Sales*, n.d.)

For the platforms feature 3DS platform had positive effect more than other platforms for all the regions. But every important console for Japan had negative selling on video games sales even the Wii platform which came out from Japan and had positive sales on other countries. Besides that, the PlayStation was popular mostly in Europe and had positive sales. (*Factors That Impact Video Game Sales*, n.d.)

In conclusion these data can provide valuable information for both game developers and publishers in terms of which types of games are popular and which platforms they should be developed for. Also, the release date of the games could be important on video games sales (not only release year but the season it is released too).

METHODS

The data for this project was collected from Kaggle website and contains information on various video games, including their name, platform, publisher release year, genre, publisher, global and other region sales. The dataset has a total of 16,598 records and contains information on games that were released between 1980 and 2020.

First from all the data we detected the outliers to remove them and continue our model with more consistent data. To detect outliers, we used Quartile 1 (25%) and Quartile 3 (75%). After detecting them we set them as “None” to remove. To compare what changed on the columns from data frame, the before and after plot histograms printed (Figure 1).

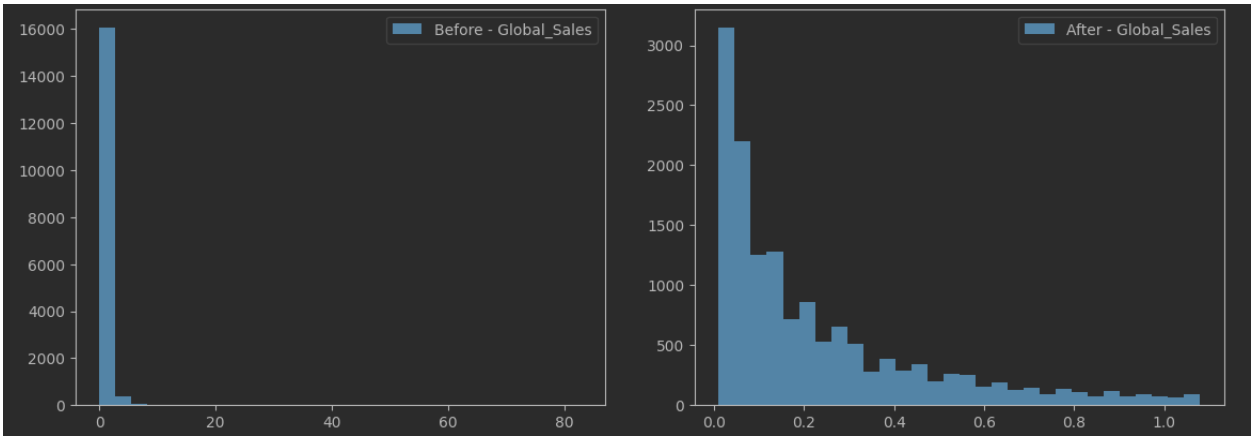

Figure 1

After removing the outliers, the unwanted stable columns got dropped. These are "Rank", "Name", "Publisher", "NA_Sales", "EU_Sales", "JP_Sales", "Other_Sales". Other sales columns are dropped because the model will check only for sales in Global. The model does not necessarily need the Name and Publisher columns. This model will only check if there is any effect of the Year, Platform and Genre on video games sales on Global Sales. After dropping all these unwanted columns, two dictionaries are created for Genre and Platform features. These dictionaries convert the string data from these two features to float data, so all the data types in cleaned data will be same and will make further analysis much easier (Figure 2).

	Platform	Year	Genre	Global_Sales
1937	7.0	1998.0	2.0	1.06
1938	2.0	2006.0	10.0	1.06
1939	5.0	2010.0	5.0	1.06
1940	2.0	2002.0	7.0	1.06
1941	1.0	2008.0	3.0	1.06
...	...	...	...	...
16593	10.0	2002.0	8.0	0.01
16594	11.0	2003.0	5.0	0.01
16595	2.0	2008.0	7.0	0.01
16596	1.0	2010.0	12.0	0.01
16597	10.0	2003.0	8.0	0.01

Figure 2

To split the data into training and test sets, the data was randomly shuffled and then split into two sets. From this data 80% of the data is being used for training and 20% of the data is being used for testing. This split ensures that the model is trained on a diverse set of data and is then evaluated on unseen data, which helps to prevent overfitting.

The decision tree regression model is implemented using the scikit-learn library in Python. This library provides a decision tree algorithm that can be used to train and make predictions with the train and predict sets we created (*1.10. Decision Trees*, n.d.). We used maximum depths “3, 4, 5, 6, 7, 10, 20, 25, 50, 125” and minimum sample leaves “5, 10, 20, 30, 40, 50” to have different models. For each model the Mean Squared Error (MSE) scores found and visualized to select best model from the models. To choose the best model from the all the decision tree models the numpy agrsort method used to sort every score we get from the scores array and we chose the one has best score as best model.

To compare how good our prediction model is, we created a plot scatter with the real data and predicted data for all the features.

Later the Mean Square Error (MSE) and Mean Absolute Error (MAE) are visualized for the best model we have, to understand how good our predicted data is. The MSE is a measure of the average squared difference between the predicted and actual values, while the MAE is a measure of the average absolute difference between the predicted and actual values (Karaderili, 2022).

At the end the Genre, Platform and Year features are individually compared to see what their effect on Global Sales is. To see that another train and test sets are created, and their Mean Absolute Error calculated individually for each feature.

RESULTS

We also displayed the model's predictions on the test set in order to assess the model's performance further. The predicted global sales and the actual global sales were shown in a scatter plot, and it was found that the model's predictions were mostly accurate. Most of the predictions

are seen to stay on a diagonal line and in the real data this diagonal line is mostly in same level and direction (Figure 3).

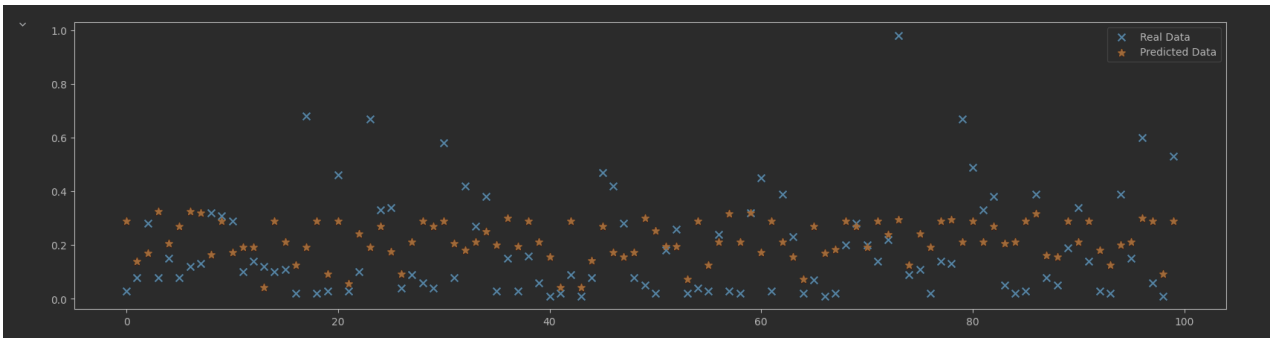

Figure 3

The MSE for the model was 0.0524, meaning that on average, the estimates for global sales were wrong by roughly 0.0524 units (Figure 4). The MAE for the best model was 0.1712, which means that on average, the predictions made by the best model regarding global sales were off by roughly 0.1712 units (Figure 5).

```
Mean Squared Error = 0.05244569664438616.
```

Figure 4

```
Mean Absolute Error = 0.17123234932145673
```

Figure 5

Later, the Mean Absolute Error calculated for each of the Genre, Platform and Year features separately to see their effect on global sales and how good our model can predict the Global Sales (Figure 6). Genre had 0.1802, Platform had 0.1770 and the Year had 0.1831 score from the predictions. We can see our model individually predicted better with “Platform” feature between all these three features.

```
{'Genre': 0.18025442403485903,
 'Platform': 0.1770962756675629,
 'Year': 0.1831890182780367}
```

Figure 6

DISCUSSION AND CONCLUSIONS

The decision tree regression model was able to estimate the global sales of video games with reasonable accuracy. The Mean Squared Error and Mean Absolute Error of the model were both small and demonstrated that the model's predictions were mostly accurate.

According to the feature individual study for each feature, the game's Genre, Platform, and Release Year of distribution had good impact on the game's global sales. On individual feature study the impact of each feature on global sales weren’t that significant. But at the end when all three were used together in our model we had the best MSE score. That shows us all three features together have the best prediction with our model. This shows that these elements are important in influencing video game sales and should be considered when creating and promoting new titles.

To be clear, this model is based on a particular dataset and might not apply well to other datasets or circumstances. Additionally, because our model is based on historical data, it might not be able to consider changes in the market or other outside factors that could affect the sales of video games. Even using different models, features and datasets may yield new information. So, when analyzing the findings of this investigation, it's critical to consider these limitations.

Overall, this study offers helpful information about the factors that affect video game sales and shows how decision tree regression may be used to forecast global sales. The results of this study can help future game developers and marketers create more successful and profitable games, which will ultimately benefit the entire industry.

References:

1. *Video Game Sales*. (2016, October 26). Kaggle. https://www.kaggle.com/datasets/gregorut/videogamesales
2. *Factors that impact video game sales*. (n.d.). https://www.causeweb.org/usproc/sites/default/files/usclap/2017-2/Factors_that_Impact_Video_Game_Sales.pdf
3. *1.10. Decision Trees*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/tree.html
4. Karaderili, S. (2022, March 11). *My Notes on MAE vs MSE Error Metrics*. HackerNoon. https://hackernoon.com/my-notes-on-mae-vs-mse-error-metrics