Ghanim Alghanim
Ali Altamimi

Predicting Game's Rating

Abstract

The purpose of this research was to employ regression models to predict average game sales in order to better understand what the community loves and promote the game-making industry's sales. We worked with the information supplied by VGChartz on their website. To get the best model feasible for this data, numerous methods have to be used. We used seaborn to visualize the outcome after acquiring the model.

Data

There are 61,000 games in the dataset, each with 18 characteristics, six of which are categorical. The game's rating, genre, console, publisher, and release date are just a few of the highlights. Due to a substantial number of missing data, the dataset was reduced to 400 games after data cleaning.

Algorithms

1. Splitting the Data Frame into 8 with different categorical combinations
2. Converting categorical features to binary dummy variables
3. Combining particular dummies and ranges of numeric features to highlight strong signals and illogical values for game rating identified during EDA

Models

Model Evaluation and Selection

The 400-piece dataset was divided into 267/133 for training and testing. We utilized Linear, Lasso, and Ridge Regression models to train them using the splitted data. All of the scores shown below were obtained using the test split, and the actual result was compared to the anticipated result. The R2, also known as the coefficient of determination, was used to calculate the regression rate.

	Linear Score	**Lasso Score**	**Ridge Score**
No Categorical	0.172062	0.172069	0.172088
Genre	0.160681	0.160806	0.163956
Console	0.242307	0.242381	0.244139
Rating	0.240552	0.240547	0.240355
Console & Rating	0.302434	0.302504	0.304803
Genre & Console	0.221529	0.222290	0.233919
Genre & Rating	0.193233	0.193353	0.198735
All Categorical	0.236200	0.236593	0.253576

We can conclude the best Data Frame is with only splitting Console and Rating with Ridge model since it has the highest R^2Score, the reason the Score is low is because the Data does not have high correlation with the other features.

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- BeautifulSoup and Requests for web scraping