Universidad Francisco Marroquin

Machine Learning Models

Christian Medina Armas

Videogame Sales Predictions

Victor Andre Farfán Miranda

20170473

Contents

Introduction

This report contains the analysis and predictions made using a Videogame Sales with Ratings dataset. My initial attempt is to predict the Global sales of a videogame based on the User Score and the Critic Score given by the Metacritic website. As we go on I notice that my initial attempt was not going anywhere as it is not linear by any meas, so a new approach is taken where I use a single continent sales count in order to try to predict the Global Sales.

The dataset contains the amount of sales of a videogame in each continent or area and the global sales amount which is the sum of the sales of all continent and areas. It also contains the Critic Score and the number of critics that gave that score, and the same for User Scores.

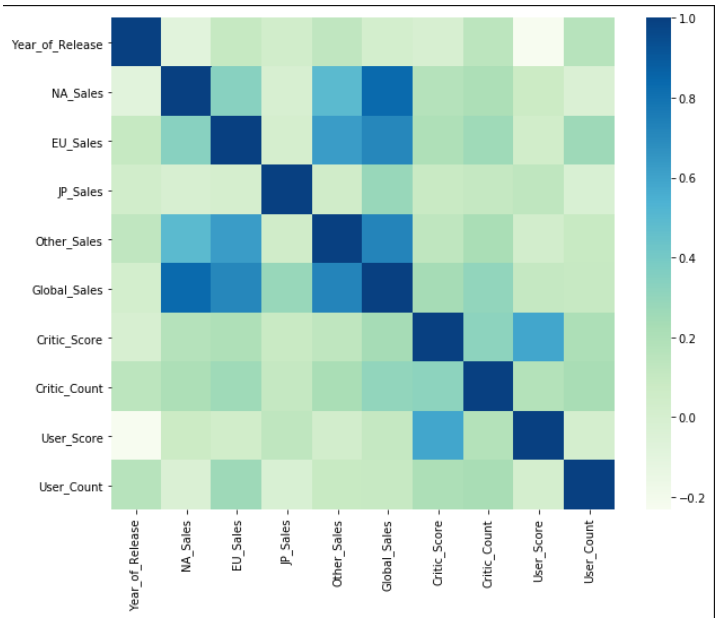
The following algorithms are used for the predictions:

- Linear Regression
- Lasso Regression
- Ridge Regression
- Polynomial Regression

Data

This videogame sales dataset contains information about all games that sold over 100,000 copies that were released between 1980 and 2019. There is a total of 16,719 observations and 16 variables. The information about each videogame includes the name of the videogame, year of release, platform it was released, the publisher, sales by continent, global sales, the score given by critics, score given by users, and the amount of critics (Metacritic staff) and the amount of users (Metacritic subscribers) which gave their score to the videogame. All this information was retrieved by scraping vgchartz.com for the videogame sales information and Metacritic.com for retrieving the videogame scores

Correlation



As we can see in this correlation matrix the global sales are not correlated to the User or Critic Score, so my first approach of using this variable to predict the global sales is not possible to do. Instead I will use the North America Sales in order to predict the Global Sales of each videogame by platform. I will focus on using North American Sales because it is the variable with the higher correlation.

Figure 1 Correlation matrix of all features in the dataset.

Methods

Data Analysis

The dataset was retrieved from Kaggle. All analysis was made using python as the coding language. From python the following libraries were used:

- For loading the data and data wrangling:
  - Pandas
  - Numpy
- For visualizing the data:
  - Plotly
  - Matplotlib
  - Seaborn
- For further analysis and predictions:
  - Scikit-learn
  - Scipy

- Statsmodels

Data cleaning

Unfortunately, there are missing observations as Metacritic only covers a subset of the platforms. This NA values were distributed in the following columns:

- Name
- Year of Release
- Publisher
- Critic and User Score
- Critic and User count

In this case I decided to remove all missing values. I made this decision because there was a 17% of missing values, with each value corresponding to 1 observation and 1 variable, but when checking the amount of observations with at least 1 missing value in the critic or user score it was a 40% of the observations. Imputing this amount of data with a mean or mode would have added a lot of bias to the dataset and after deleting them I still had 6,825 which was a good amount for the predictions.

User scores were also in a range to 1-10, and Critic Scores were in a range of 1-100. I changed the User Scores, so it fits the format of 1-100 that the Critic Scores have.

Predictions

In order to make predictions I will use the following algorithms:

- Linear Regression
- Lasso Regression
- Ridge Regression
- Polynomial Regression

Results

Linear Regression

Using the simple linear regression from statsmodels the Linear regression returned an R^2 error of 0.92 and a RMSE of 0.60.

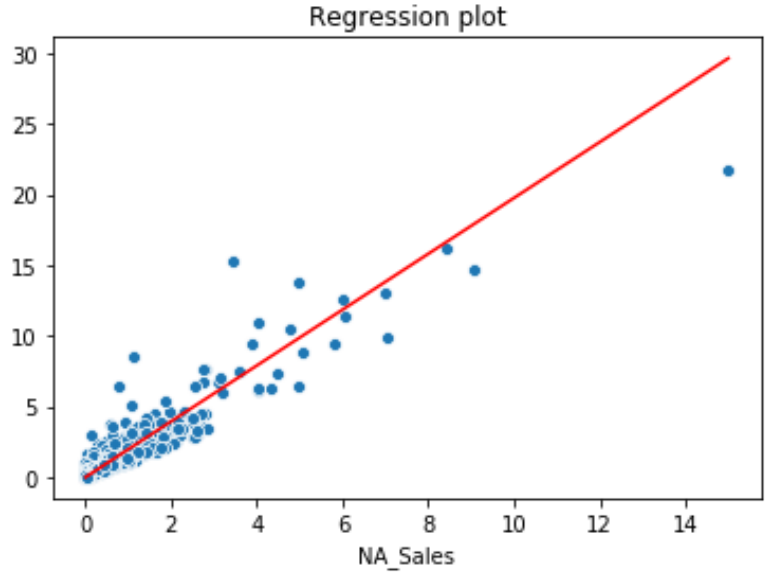


Figure 2 Simple Linear Regression using North America sales to predict the Global Sales.

Lasso Regression

For the Lasso Regression I manually tried to find the alpha values that could include all or at least most of the points of the graphic. I found that setting alpha=-1 the prediction line would be above almost all the points and setting alpha=1 the prediction line would be below of almost all points. With this interval I iterated in this range using a step of 0.1 in order to find the alpha value that gave me the best R^2 score. After this process I found that using an alpha value of 0.20 R^2 error is of 0.87 and the RMSE error is 0.57.

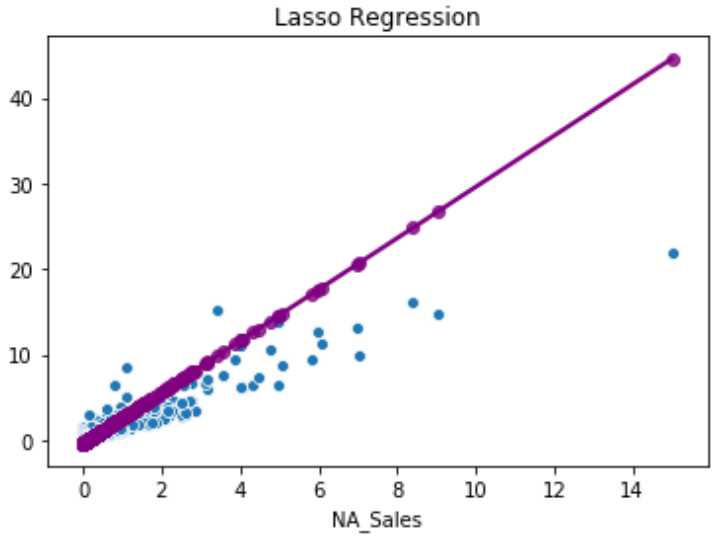


Figure 3 Lasso Regression using North America sales to predict the Global Sales.

Ridge Regression

I made the same iteration over an interval process to find the best value for the Ridge Regression but in this case the alpha values interval was between -10,000 and 10,000 and using a step of 1. In this process I found that the best alpha value was 668, which returned an R^2 error of 0.87 and a RMSE of 0.57.

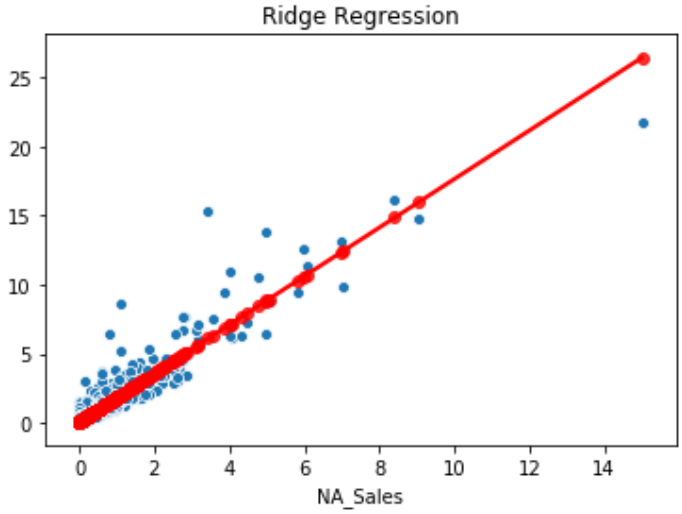


Figure 4 Ridge Regression using North America sales to predict the Global Sales.

Linear Regressions Comparison

As described before, Lasso and Ridge Regressions returned the same R^2 error and RMSE error on its best-case scenario even though the predicting line does not have the same slope. There is also a huge concentration (almost 75% of the data) focused on the lower left corner of the graph which could also contribute to this different slope on each regression line.

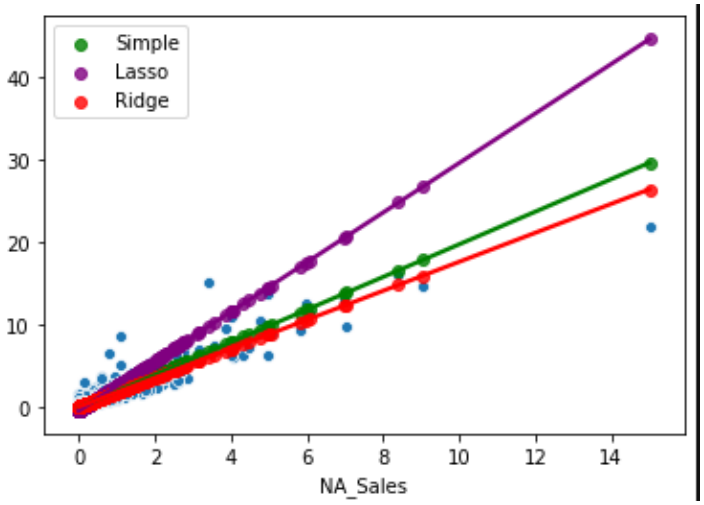


Figure 5 Comparison of a Simple Linear Regression, Lasso Regression and Ridge Regression.

Polynomial Regression

In order to try to make a more effective algorithm I also tried a Polynomial regression still using the North America sales to make the prediction of the Global Sales. Testing using increments of 1 in the regression degree I found that after degree 3 there was no improvement on the R^2 error. I still kept using a degree of 4 as it was the last one I tried. This polynomial regression returned an R^2 error of 0.92 and a RMSE error of 0.63. R^2 error is the same as the simple Linear Regression and the RMSE error is worse, so this model is worse for this data.

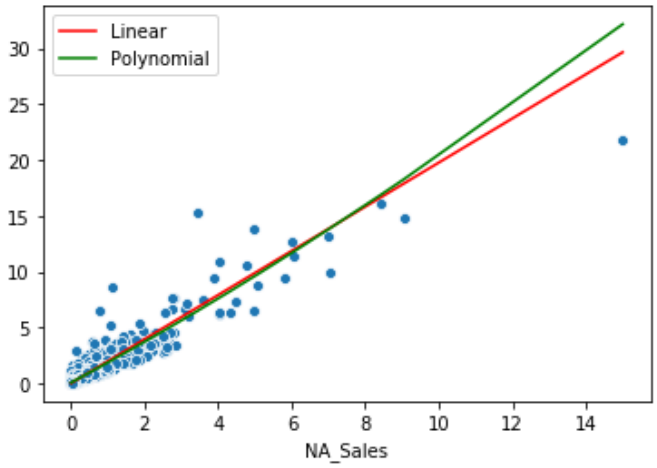


Figure 6 Polynomial regression of degree 4 using North America sales to predict Global Sales

Regressions comparison selecting only sales under 5 million

In another attempt to improve the accuracy of the model I selected only games with under 5 million global sales. I used the same method to find the best possible Lasso and Ridge Regression. The results were the following:

	R^2 error	RMSE error	Alpha
Simple Linear Regression	0.82	0.39	
Lasso Regression	0.77	0.39	-1.77
Ridge Regression	0.77	0.38	-29

Table 1 Regression results for Simple Linear Regression, Lasso Regression and Ridge Regression.

Lasso and Ridge Regression continue being almost equally good, this time the difference in the RMSE error resides in some different decimals but are still almost the same. Also as we can see in the graphic the prediction line is almost the same for the 3 algorithms.

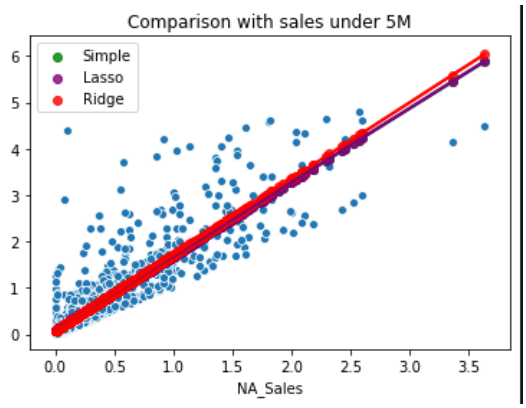


Figure 7 Comparison of Simple Linear Regression, Lasso Regression and Ridge Regression but only with games that sold under 5M worldwide.

Use Japan Sales instead of North America sales

Using Japan sales to predict is worse as the data is not as linear as it was with the North America sales but it still has a high concentration on the lower values of the graphic and there are few games that sell more than around 5 million copies. Results were R^2 error of 0.40 and RMSE error of 1.45.

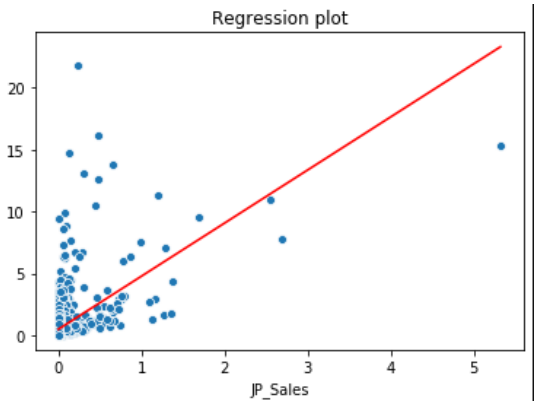


Figure 8 Simple Linear Regression using Japan sales to predict Global Sales

Conclusion

To conclude my report, I can say the following statements:

- Removing the outliers of the dataset improved the better as it returned a RMSE error of almost half than when using the full dataset.
- Polynomial regression made a worse model, probably because the data is completely linear.
- Lasso and Ridge regression where the best predicting models for this dataset.
- We can have a good guess of how much a videogame will sell globally based on how well it sells in North America.
- Lasso and Ridge regression will have the same R^2 error and RMSE error when we select the best alpha for those algorithms.

Appendix

Dataset recovered from: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

Variable dictionary:

- Name: Name of the game.
- Platform: Console on which the game is running.
- Year_of_Release: Year of the game released.
- Genre: Game's category.
- Publisher: Company which published the game.
- NA_Sales: Game sales in North America (in millions of units).
- EU_Sales: Game sales in the European Union (in millions of units)
- JP_Sales: Game sales in Japan (in millions of units)
- Other_Sales: Game sales in the rest of the world, i.e. Africa, Asia excluding Japan, Australia, Europe excluding the E.U. and South America (in millions of units).
- Global_Sales: Total sales in the world (in millions of units).
- Critic_Score: Aggregate score compiled by Metacritic staff.
- Critic_Count: The number of critics used in coming up with the Critic_score.
- User_Score: Score by Metacritic's subscribers.
- User_Count: Number of users who gave the user_score.
- Developer: Party responsible for creating the game.
- Rating: The ESRB ratings (E.g. Everyone, Teen, Adults Only..etc).