Predict Global Selse of Video Games

Introduction

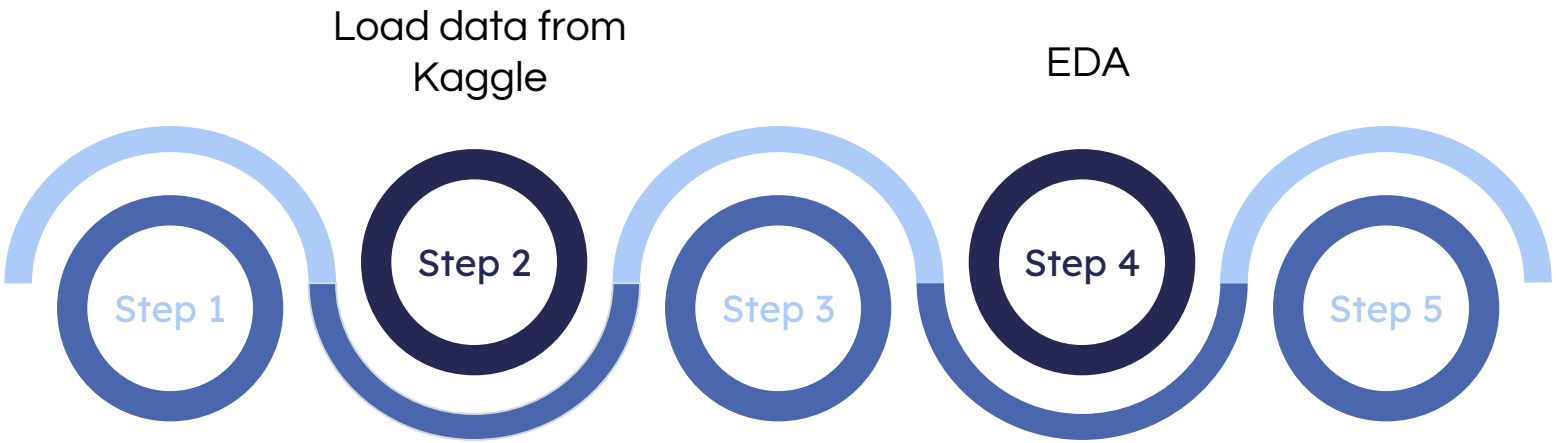
IMDb is an online database of information related to video gmaes, movies ,television series, Etc...

IMDb

Objective

The goal of this project is to predict global sales of the video games using linear regression model based on features of the data from IMDb website and Kaggle.

WorkFlow


Load data from Kaggle
EDA
Step 1
Step 2
Step 3
Step 4
Step 5
Web Scraping multiple pages
Merge datasets
Linear regression

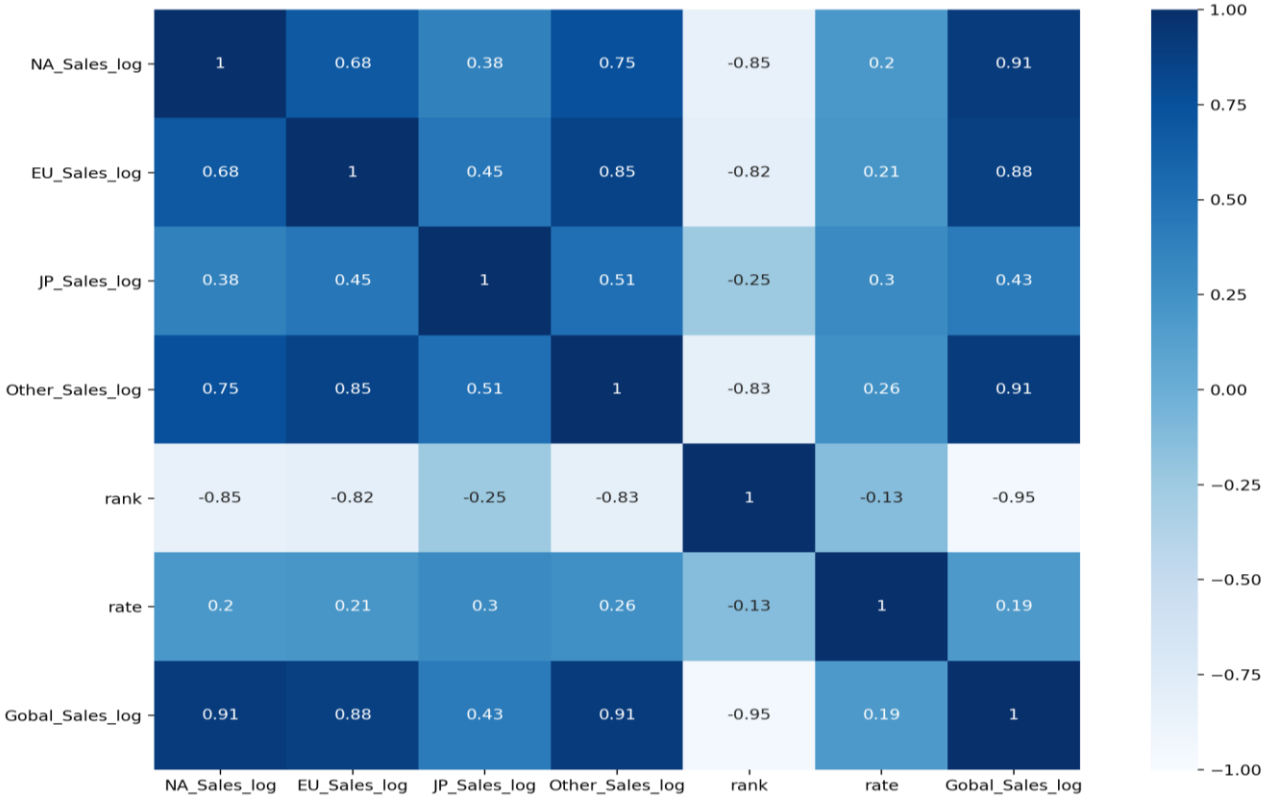
Data Discreption

We scrap 10 pages from IMDb then marge it with another dataset from Kaggle after that we got 658 rows and 7 columns .

Data :

- **Features:**
  - rate - Number of rate.
  - Rank - Ranking of overall sales.
  - NA_Sales - Sales in North America (in millions).
  - EU_Sales - Sales in Europe (in millions).
  - JP_Sales - Sales in Japan (in millions).
  - Other_Sales – Sales in the rest of the world (in millions).
- **Target:**
  - Global_Sales - Total worldwide sales.

EDA

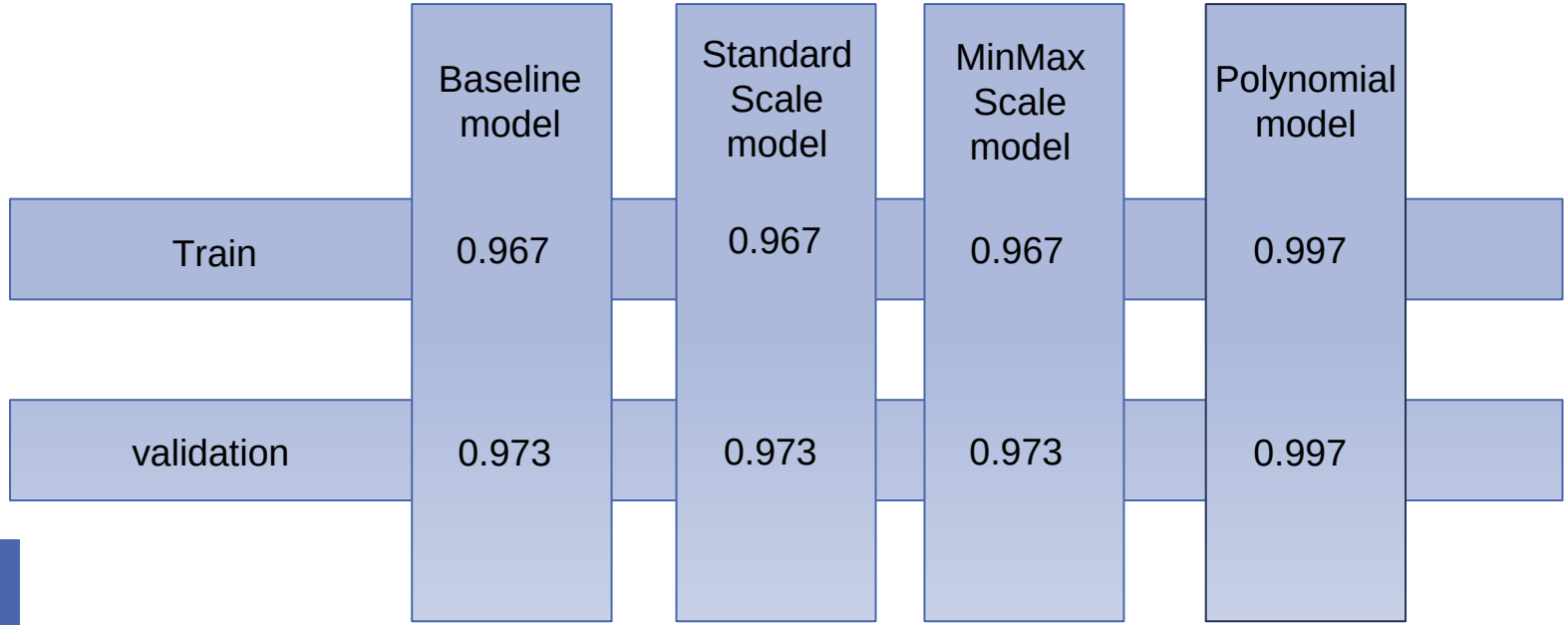

NA_Sales_log
EU_Sales_log
JP_Sales_log
Other_Sales_log
rank
rate
Gobal_Sales_log
1 0.68 0.38 0.75 -0.85 0.2 0.91
0.68 1 0.45 0.85 -0.82 0.21 0.88
0.38 0.45 1 0.51 -0.25 0.3 0.43
0.75 0.85 0.51 1 -0.83 0.26 0.91
-0.85 -0.82 -0.25 -0.83 1 -0.13 -0.95
0.2 0.21 0.3 0.26 -0.13 1 0.19
0.91 0.88 0.43 0.91 -0.95 0.19 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Linear Regression

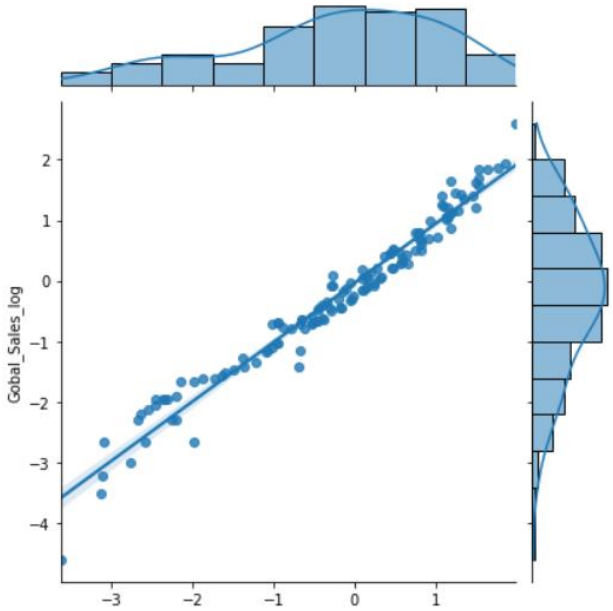
Features Engineering

The features engineering that we used is data transform using Logarithmic transformation we use it to make the data more normalized.

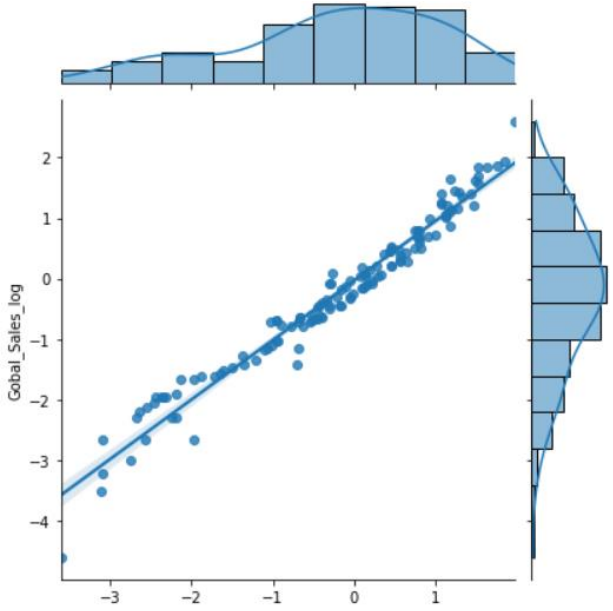
experiments



Models

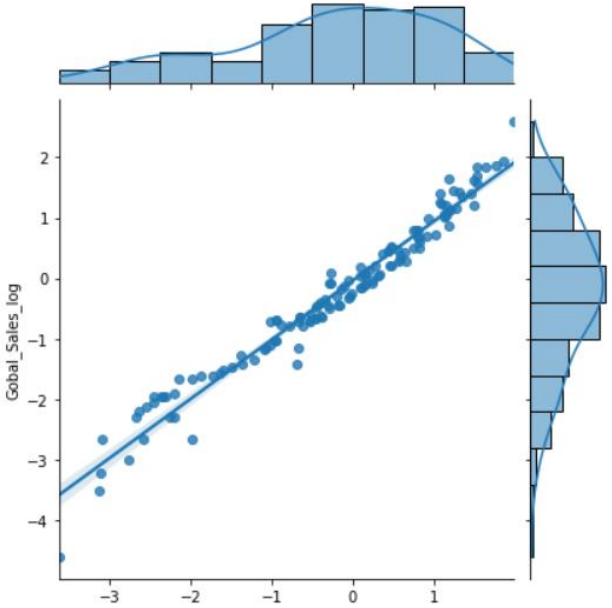


Baseline model

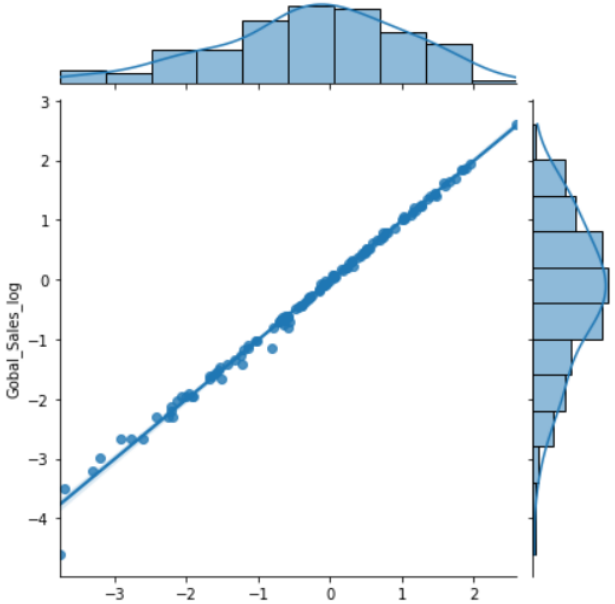


Standard Scale model

Models



MinMax Scale model

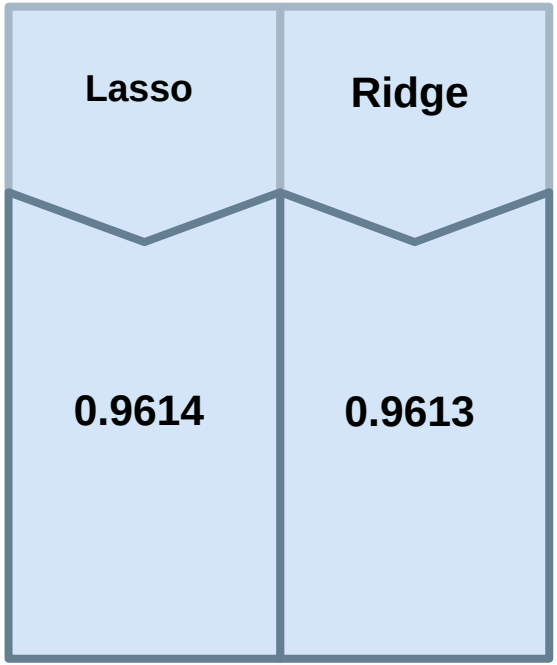


Polynomial model

Test Model (Unseen data)

After we do experiment, we noticed the best experiment from r^2 results is polynomial model we got r^2 = 0.99 which mean the model well trained on the data .

Lasso & Ridge



Conclusion

At the end from R-Squared Test results, we can assume that the best model is polynomial model because they have the highest r^2 in both test and train

Thank you