

Description of project goals:

The dataset was scraped and collected from Kaggle, We do analysis of resulting data using classification. The models of classification were based on the data collected and the feature engineering necessary to analyze the data in classification. This project goal is to classify each video game in the dataset by ESRB rating. To do this we used Logistic Regression, Random Forest , k-neighbor , Naive bayes and Decision Tree then we applied voting . The classification metrics that we used on our models to evaluate the prediction scores was Accuracy and F1 .

Data Used:

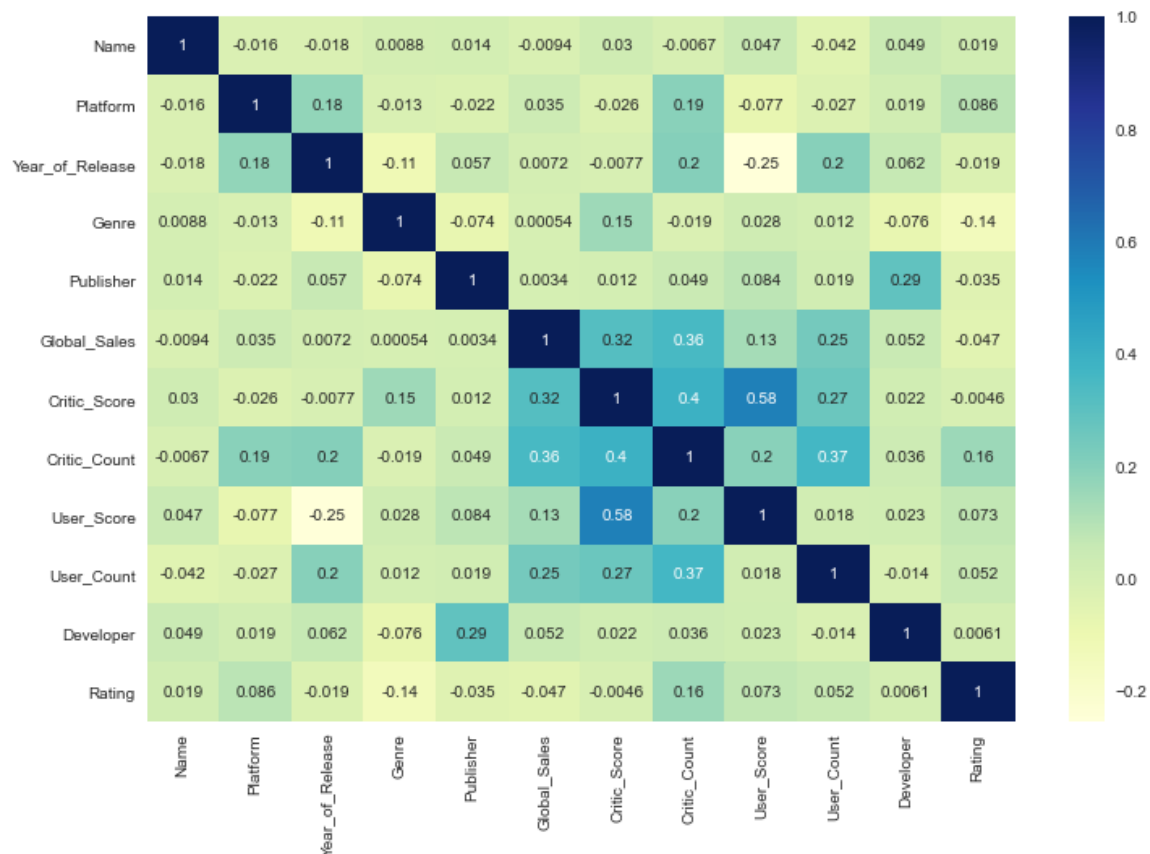
The data was collected from the Kaggle webpage,It has 15 **features**:

Name, Platform, Year of Release, Genre, Publisher, NASales, EUSales, JPSales, OtherSales, Global_Sales, Critic_Score, Critic_count, User_Score, Developer,

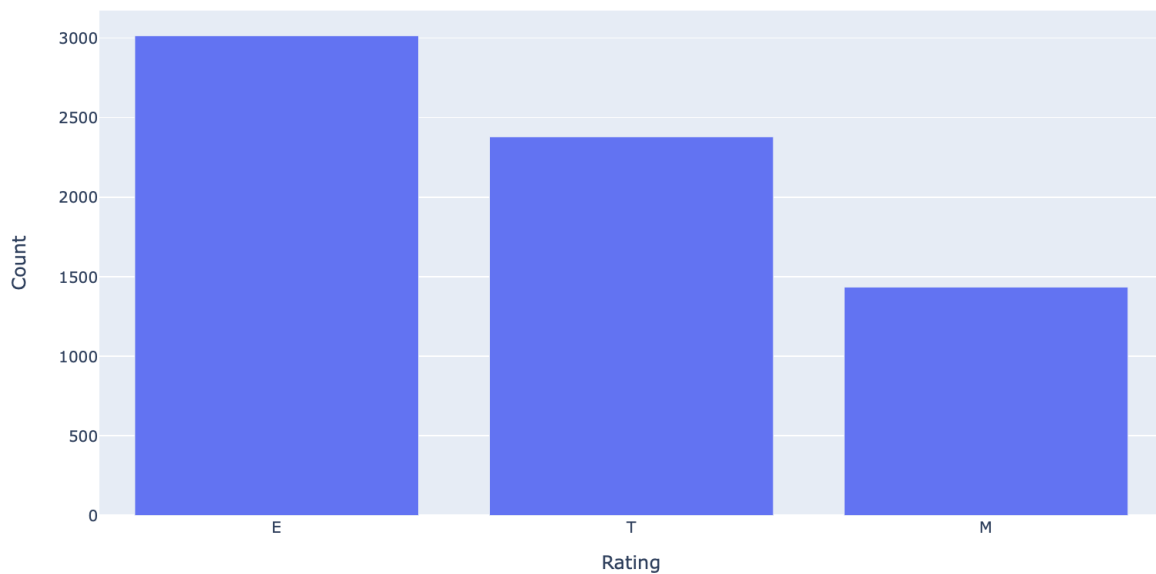
Target :Rating.(M,T,E)

Data peproseedsing :

Renaming columns for ease of use ,Dropping games without a year of release or genre,and Dropping all seles features in new data frame which used in our modeling,Convert(User_Score, Yearof_Release, User_Count, Critic_Count) to int,andEliminating all rows with NaN value



This heat map represent correlation between future we not there are low corr between futures and y tager (rating)



The bar chart represents the number of games for each class and as we note the data does not have an unbalanced problem.

Tools Used:

- Numpy
- Pandas
- Seaborn
- SKlearn:modeling ,splitting data, matrices
- Matplotlib

Methodology:

Five different models were applied:The Linear Regression Model,TheDecision Tree Model,Random Forests Model,K-Nearest Neighbors Model,Naive Bayes Model,and Voting Classifier

Result:

The results show that Random Forest & knn provide the best results with F1 score of 0.43,0.52and the Voting Classifier result was 0.46 inF1 score this scour regarding to add svm model in Voting

| Logistic Regression | |
|---------------------|------|
| Accuracy | F1 |
| 0.53 | 0.36 |

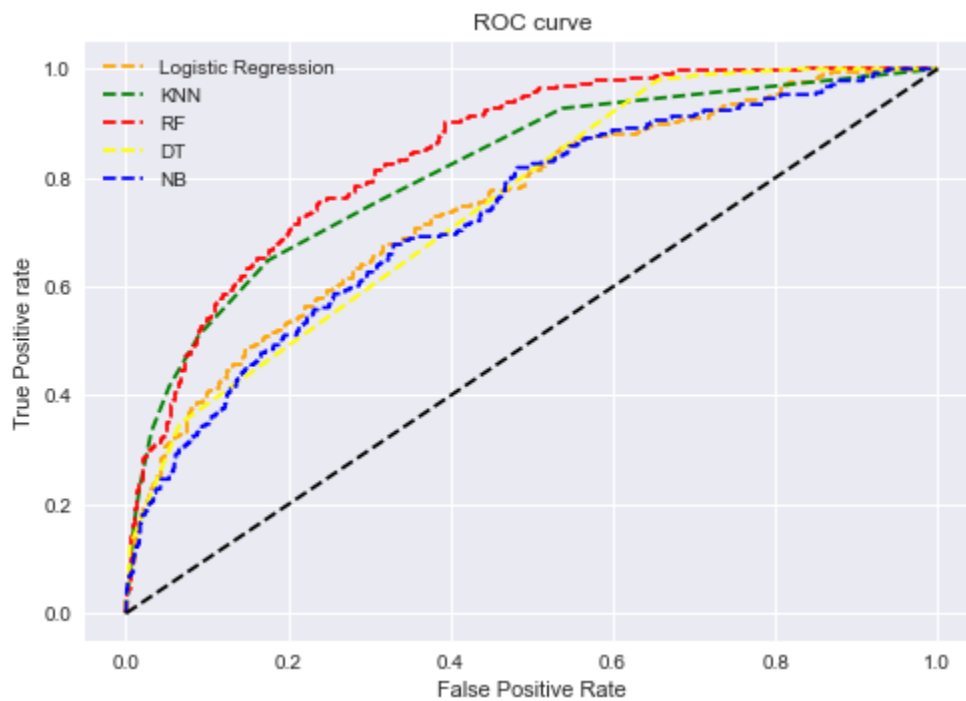
| Random Forest | |
|---------------|------|
| Accuracy | F1 |
| 0.57 | 0.43 |

| Naive Bayes | |
|-------------|------|
| Accuracy | F1 |
| 0.51 | 0.33 |

| Decision Tree | |
|---------------|------|
| Accuracy | F1 |
| 0.55 | 0.43 |

| K-Neighbor | |
|------------|------|
| Accuracy | F1 |
| 0.60 | 0.52 |

| Voting Classifier | |
|-------------------|------|
| Accuracy | F1 |
| 0.672 | 0.46 |



From The ROC above the Random Forest & knn were a better classifier than others.

Future work:

a bigger dataset could be beneficial for unsupervised learning methods, and New features like review text, number of uninstalls, and number of updates could provide new insight that can improve the modeling results.

By :

Rana Alqahtani Ohoud Albabtain Khulud Amer.