Bad, mediocre, good or outstanding? 🎲
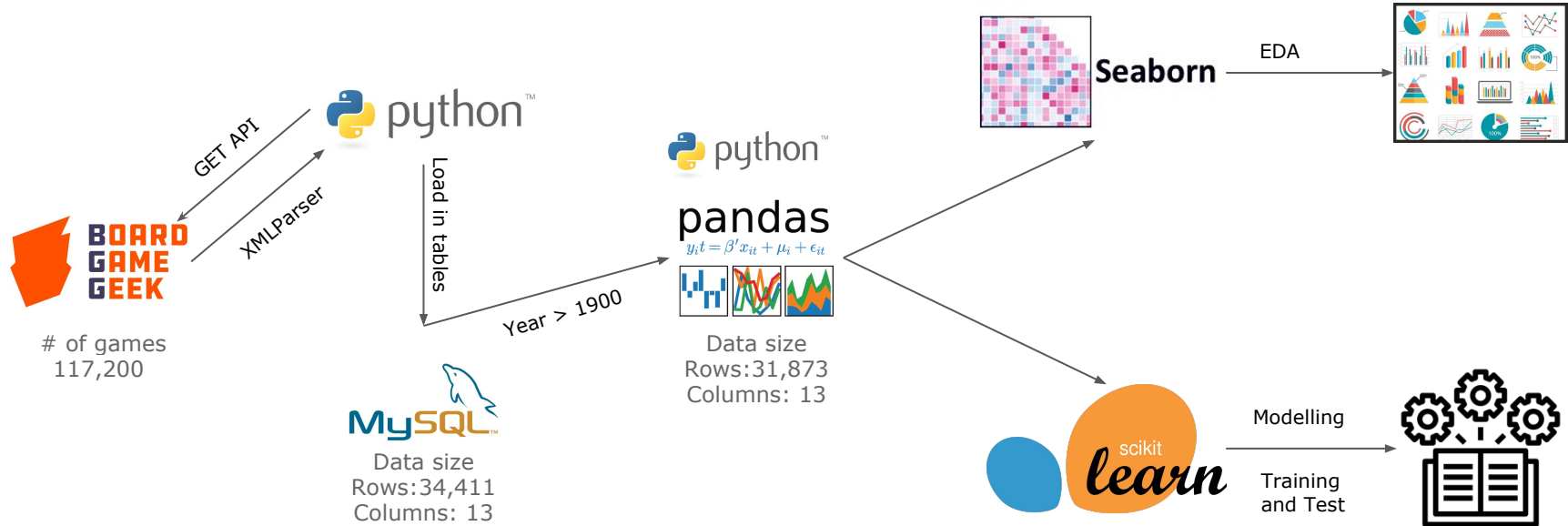
Identifying features and predicting board game ratings.

By: Janki Chauhan
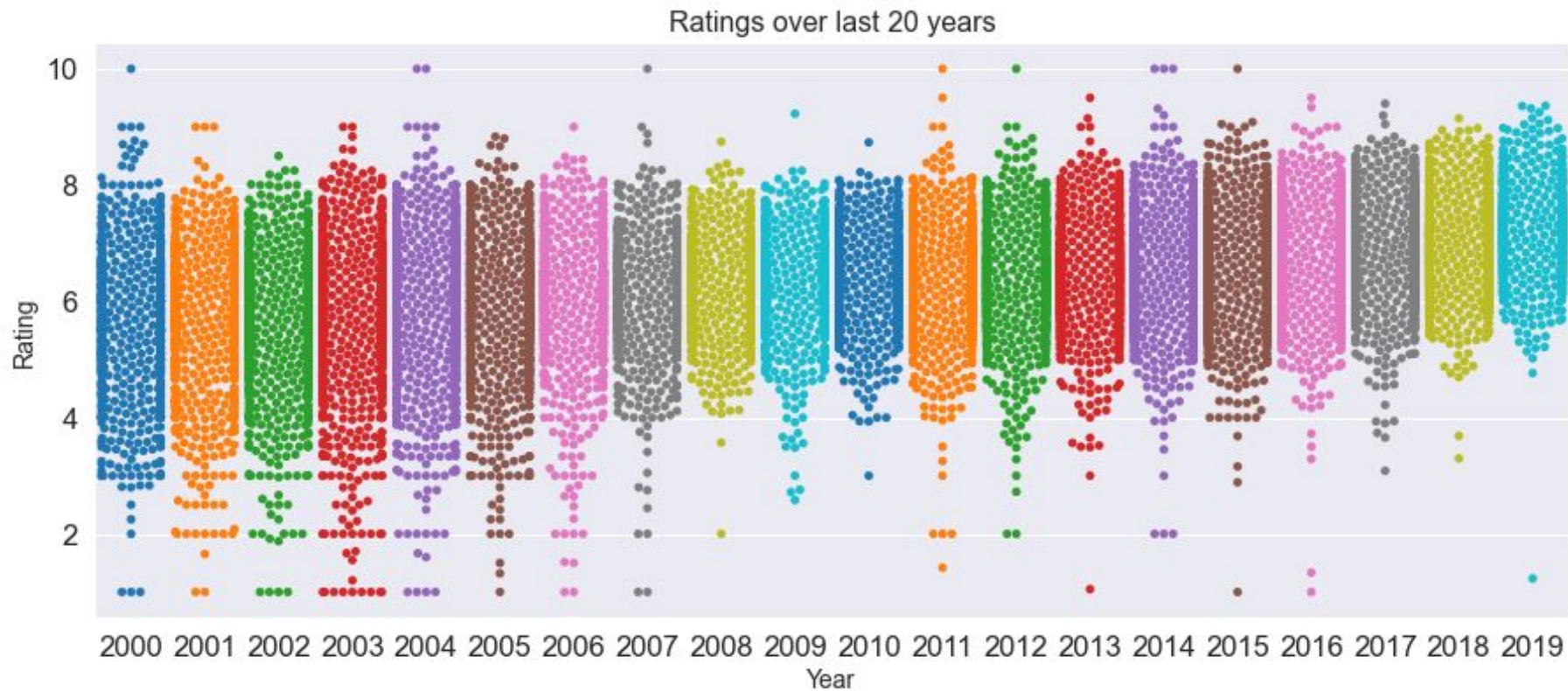
[Project github]

Data Pipeline
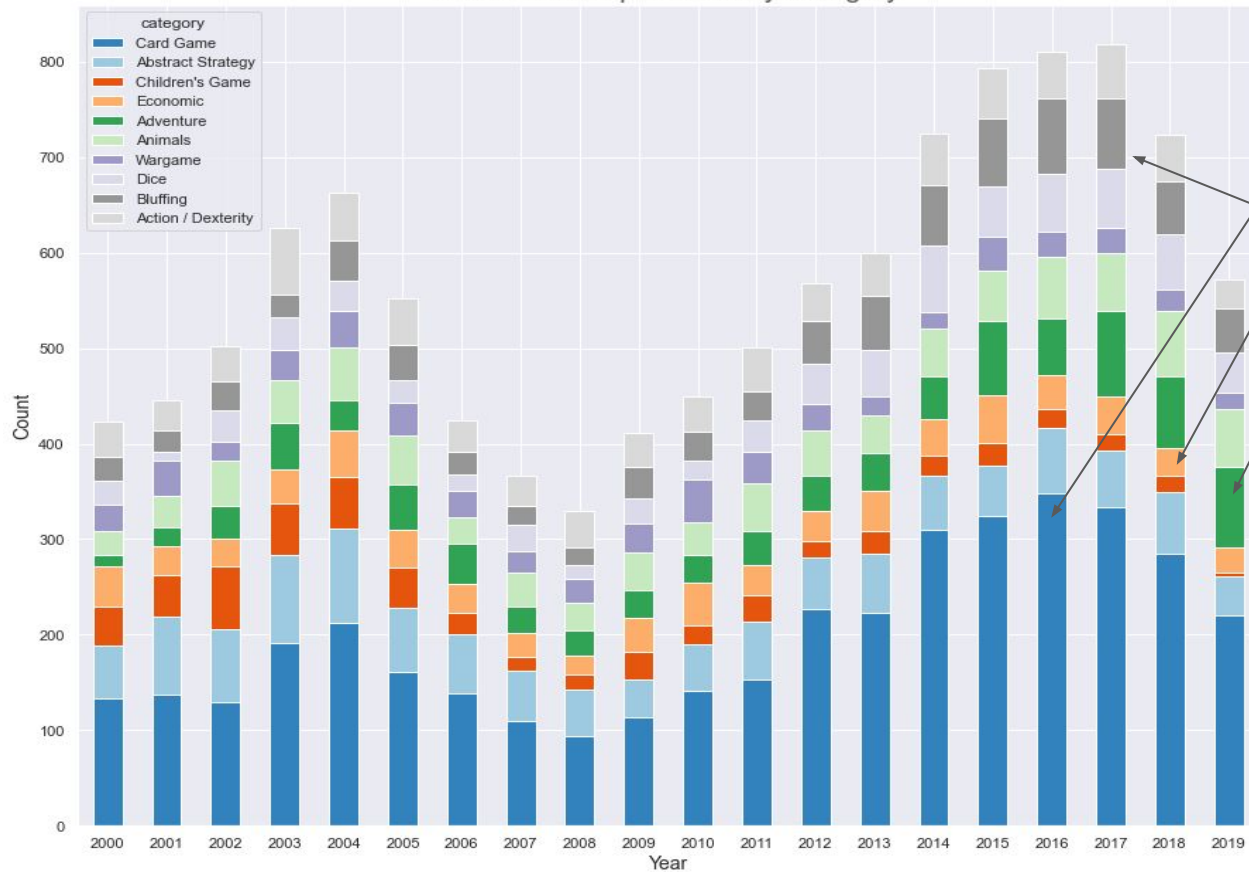

GET API

XMLParser

BOARD GAME GEEK

# of games
117,200

Load in tables

Year > 1900

MySQL
Data size
Rows:34,411
Columns: 13

pandas

$$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$$

Data size
Rows:31,873
Columns: 13

Seaborn

EDA

scikit learn

Modelling

Training
and Test

Exploratory Data Analysis
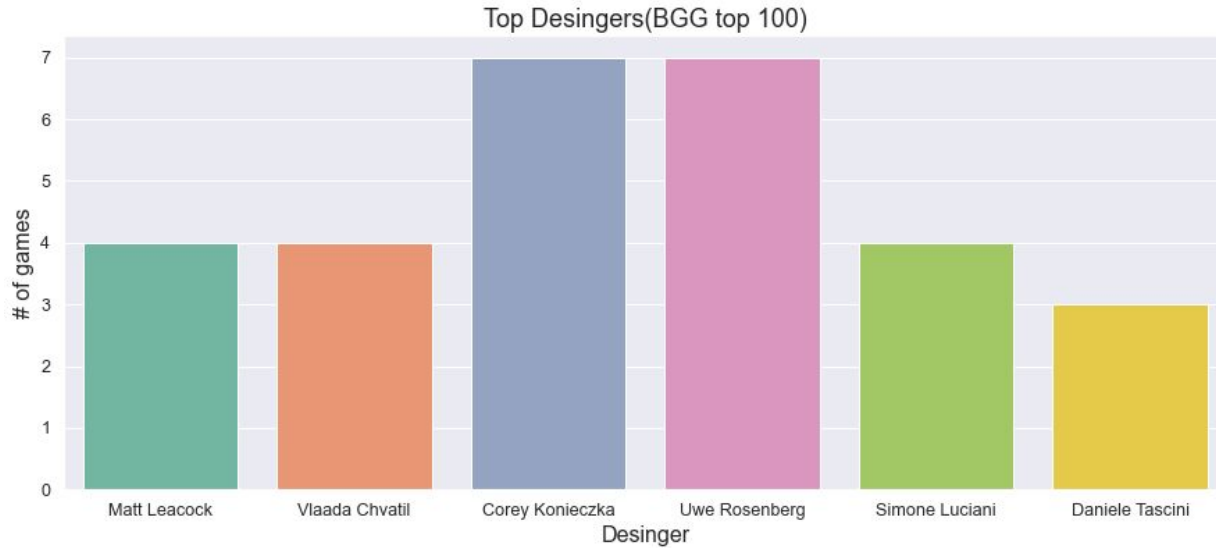

Ratings over last 20 years

Exploratory Data Analysis

Board Games published by Category

- Card games are most popular category
- There has been decline in children's game in recent years
- While increase in bluffing and adventure games.

Exploratory Data Analysis

Top Desingers(BGG top 100)

Game designers with more than 2 game in BGG top 100

Modelling

Predicting rating of a board game

Linear Regression:

- Default parameters
- **MSE: 4.24**

Random Forest:

- No. of estimators: 1000
- Max features: 10
- Bootstrap: True
- **MSE: 3.03**

Gradient Boost:

- Learning rate: 0.2
- No. of estimators: 1000
- Max features=10
- **MSE: 3.01**

Model Tuning

Clustering:

- Use for categorical features
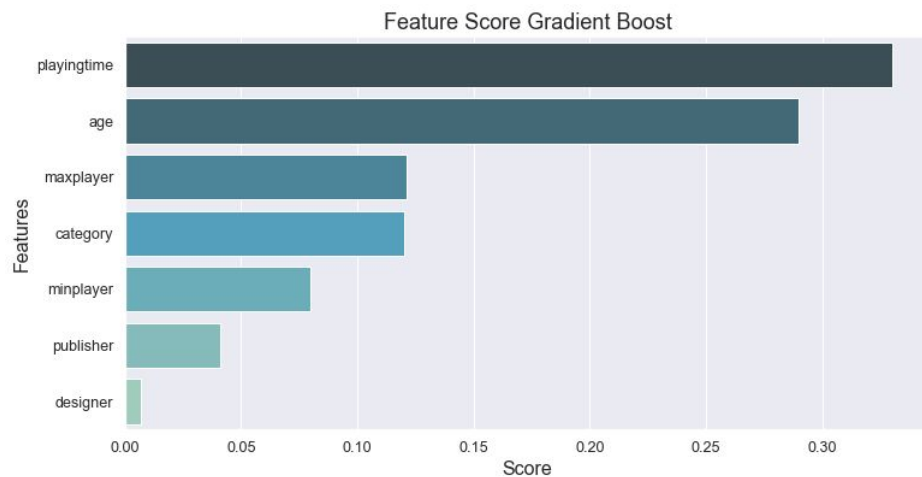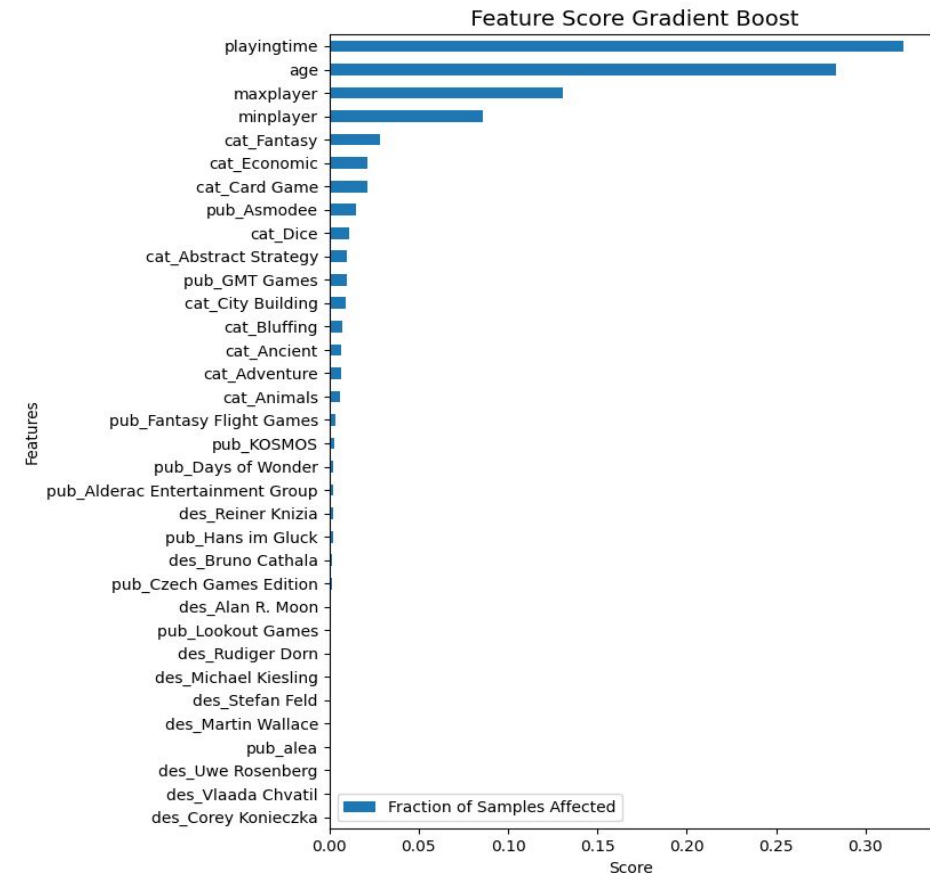- Kmode: 15 clusters

Random Forest:

- Bootstrap: True
- Max features: 20
- No. of estimators:2000
- No cluster, pick top 15 for categorical features
- **MSE: 2.92**
- **R2: 0.22**

Gradient Boost:

- Learning rate: 0.05
- Max features: 15
- No. of estimators: 2000
- No cluster, pick top 15 for categorical features
- **MSE: 2.83**
- **R2: 0.21**

Results

Feature Score Gradient Boost

Feature Score Gradient Boost

Grouped publishers, categories and designers to get better visual understanding

Future Analysis

- There is an increase in ratings over the years. Is it because the production quality has gotten better? Are publishers and game designers targeting niche groups?
- How does kickstarter factors in game ratings, it definitely creates awareness before the game goes to production.

Questions?

Thank you!

[Project github]