APC 2022/23

TV shows on Netflix, Prime Video, Hulu and Disney+

Cristian Pérez Díaz

Github

Table of contents

- Dataset
- Model Selection
- Conclusion

Dataset

Atributes

- 12 atributes in total:
    - 5 numerical
    - 5 binary
    - 1 categorical
        - Age_all
        - Age_7+
        - Age_13
        - Age_16
        - Age_18

Atributes

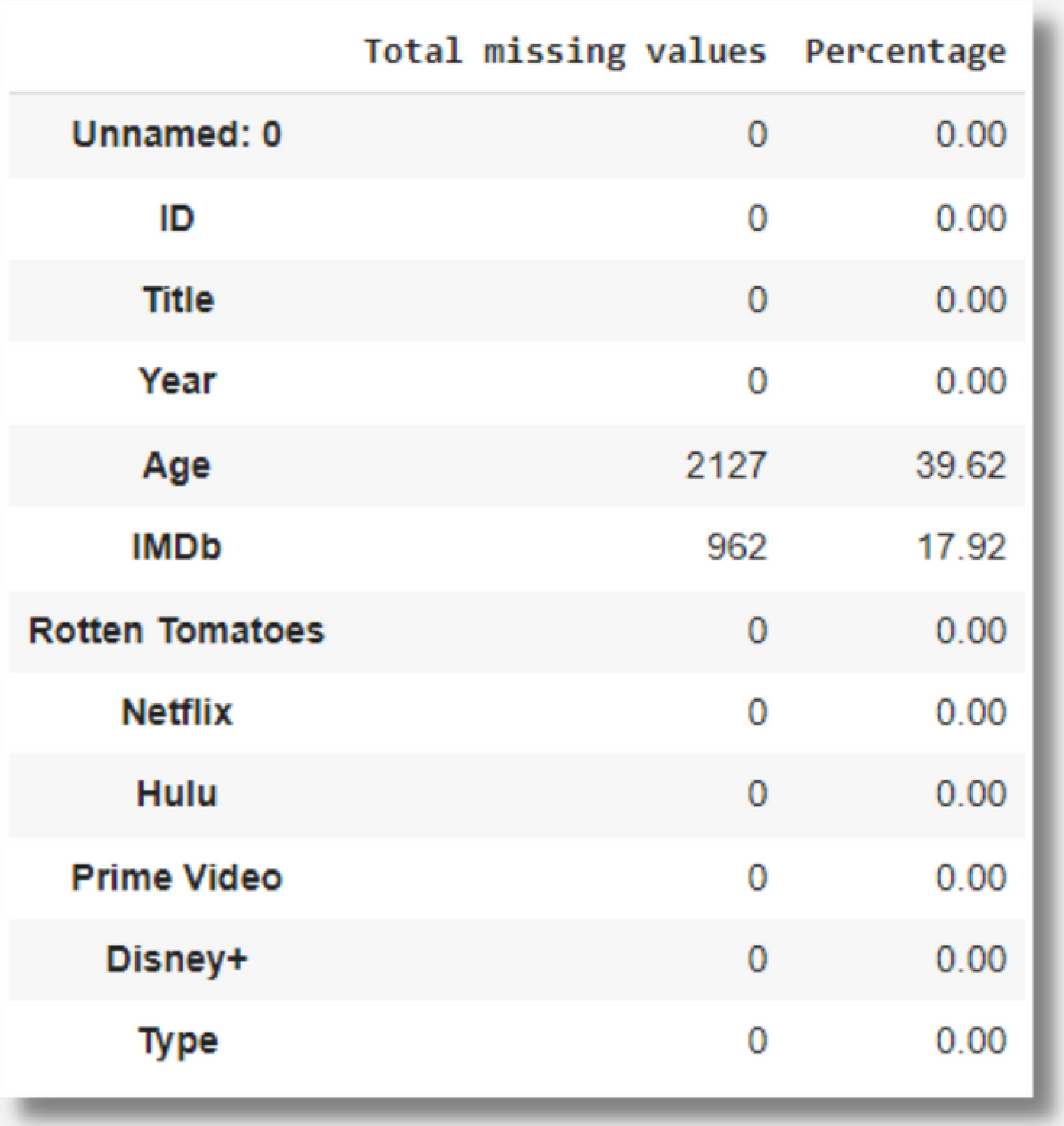
- Unnamed: 0 : Row ID
- ID : Unique TV show ID
- Title : Title of Movie/Show
- Year : The year in which the tv show was produced
- Age : Target age group
- IMDb : IMDb rating
- Rotten Tomatoes : Rotten Tomatoes rating
- Netflix : Whether the tv show is found on Netflix
- Hulu : Whether the tv show is found on Hulu
- Prime Video : Whether the tv show is found on Prime Video
- Disney+ : Whether the tv show is found on Disney+
- Type : Movie or TV Show

Atributes

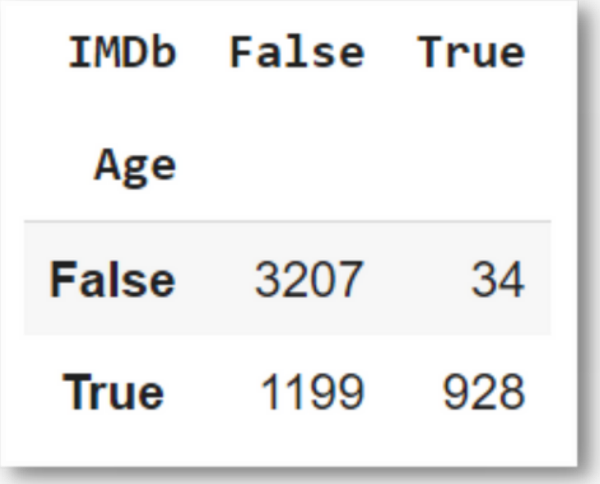
- Unnamed: 0 : Row ID
- ID : Unique TV show ID
- Title : Title of Movie/Show
- Year : The year in which the tv show was produced
- Age : Target age group
- IMDb : IMDb rating
- Rotten Tomatoes : Rotten Tomatoes rating
- Netflix : Whether the tv show is found on Netflix
- Hulu : Whether the tv show is found on Hulu
- Prime Video : Whether the tv show is found on Prime Video
- Disney+ : Whether the tv show is found on Disney+
- Type : Movie or TV Show

Atributes

- Columns Age & IMDb with nulls



	Total missing values	Percentage
Unnamed: 0	0	0.00
ID	0	0.00
Title	0	0.00
Year	0	0.00
Age	2127	39.62
IMDb	962	17.92
Rotten Tomatoes	0	0.00
Netflix	0	0.00
Hulu	0	0.00
Prime Video	0	0.00
Disney+	0	0.00
Type	0	0.00



IMDb	False	True
Age		
False	3207	34
True	1199	928

Atributes

- Columns IMDb & Rotten Tomatoes

IMDb	'6/10'
Rotten Tomatoes	'60/100'

Atributes

- Columns IMDb & Rotten Tomatoes

IMDb
Rotten Tomatoes

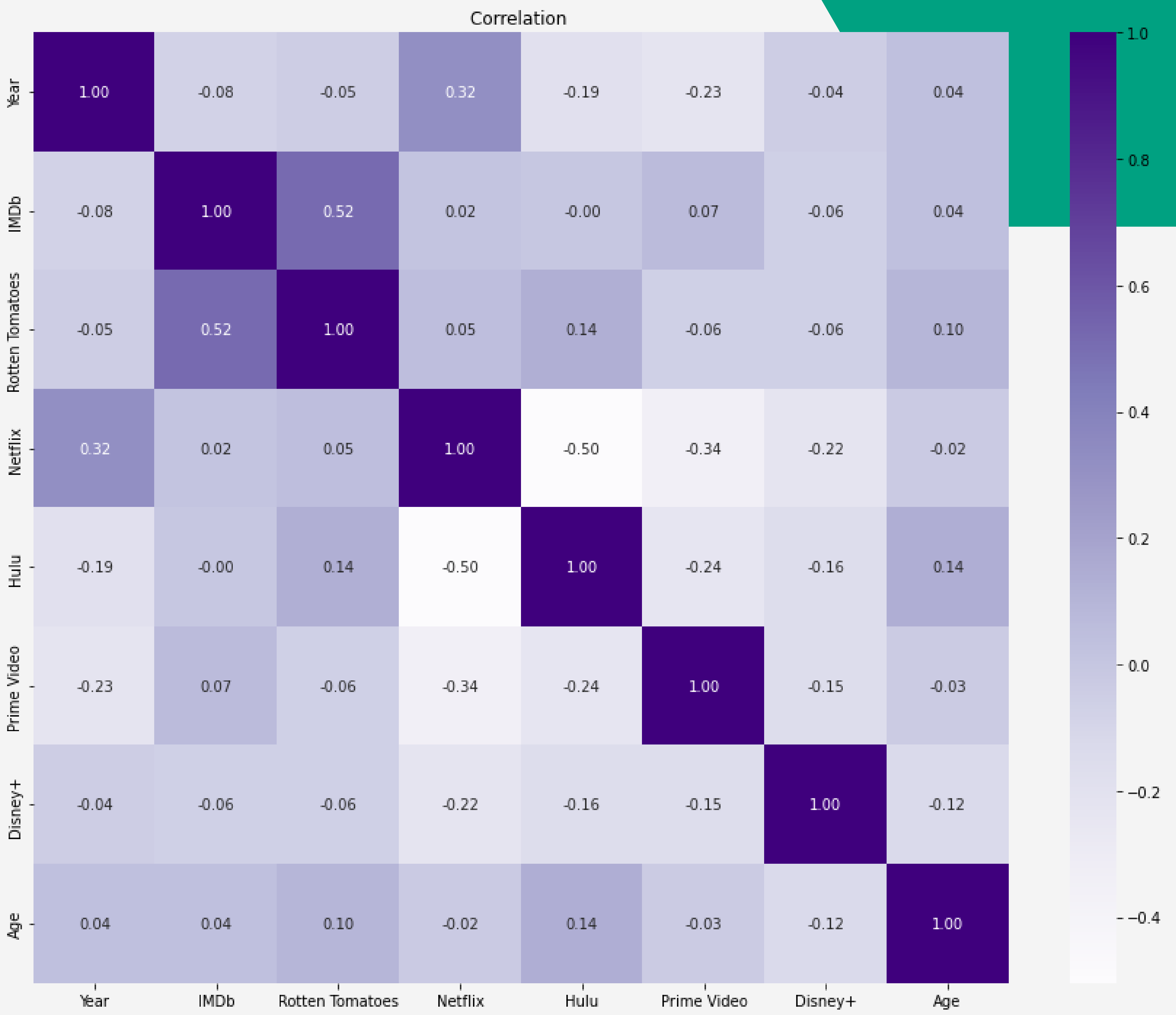
6

Atributes

- Columns Age converted to numeric:
  - Age_all ->1
  - Age_7 ->10
  - Age_13 ->10000
  - Age_16 -> 1000
  - Age_18 -> 100

Correlation Matrix

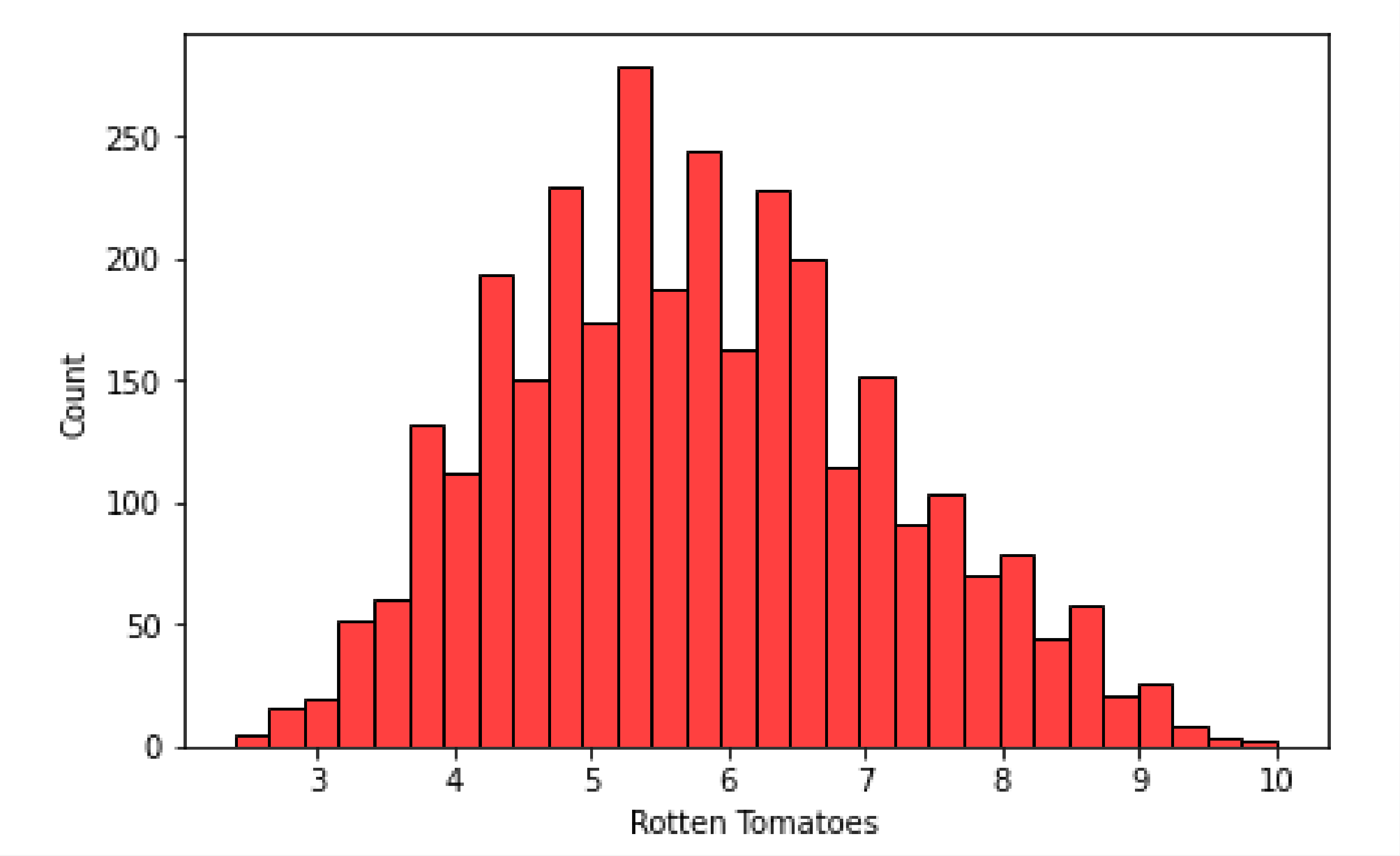
- High interesting correlations:
  - Rotten Tomatoes & IMDb
  - Year & Netflix



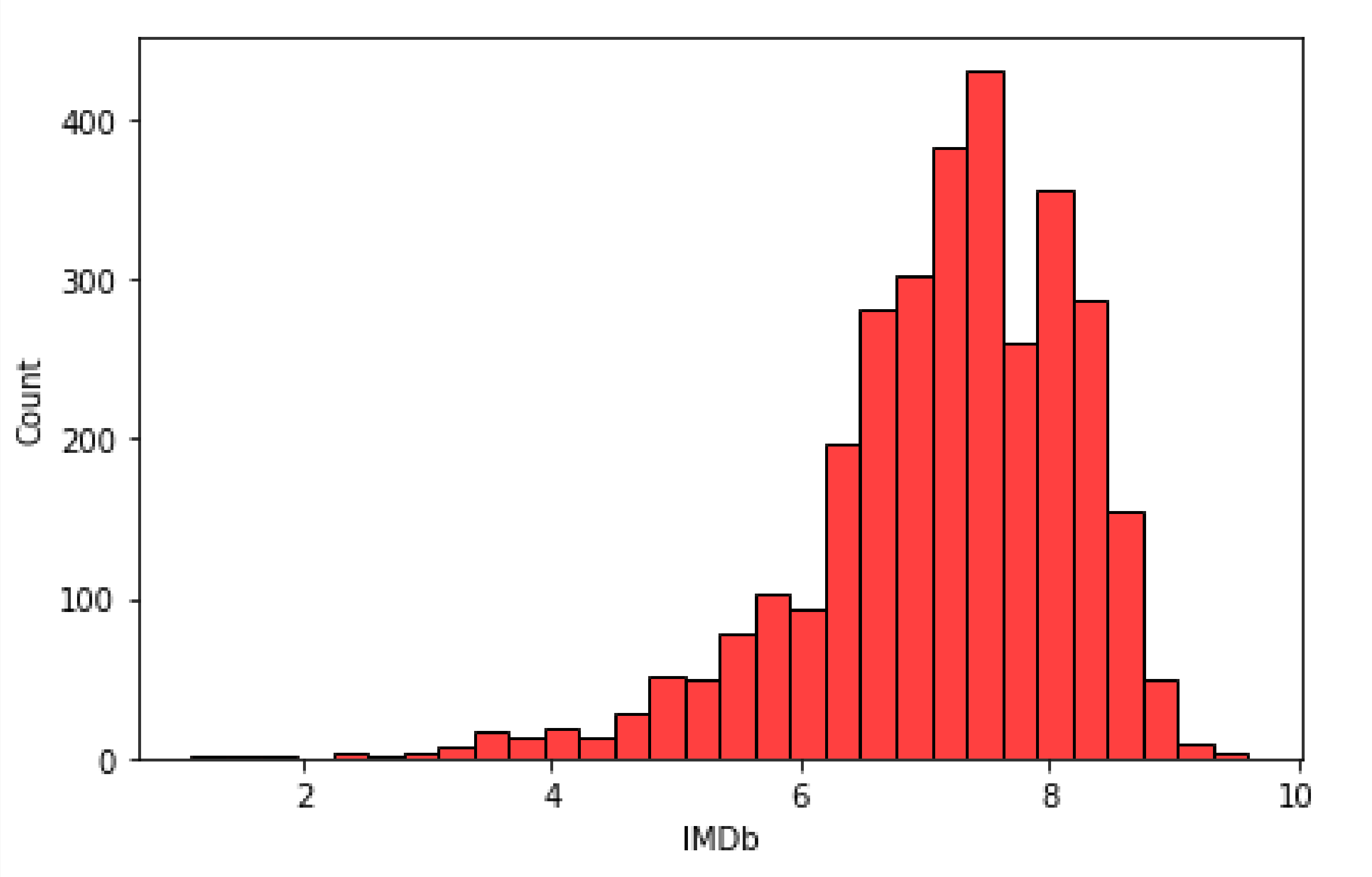
	Year	IMDb	Rotten Tomatoes	Netflix	Hulu	Prime Video	Disney+	Age
Year	1.00	-0.08	-0.05	0.32	-0.19	-0.23	-0.04	0.04
IMDb	-0.08	1.00	0.52	0.02	-0.00	0.07	-0.06	0.04
Rotten Tomatoes	-0.05	0.52	1.00	0.05	0.14	-0.06	-0.06	0.10
Netflix	0.32	0.02	0.05	1.00	-0.50	-0.34	-0.22	-0.02
Hulu	-0.19	-0.00	0.14	-0.50	1.00	-0.24	-0.16	0.14
Prime Video	-0.23	0.07	-0.06	-0.34	-0.24	1.00	-0.15	-0.03
Disney+	-0.04	-0.06	-0.06	-0.22	-0.16	-0.15	1.00	-0.12
Age	0.04	0.04	0.10	-0.02	0.14	-0.03	-0.12	1.00

Histograms

- Rotten Tomatoes



- IMDb



Model selection

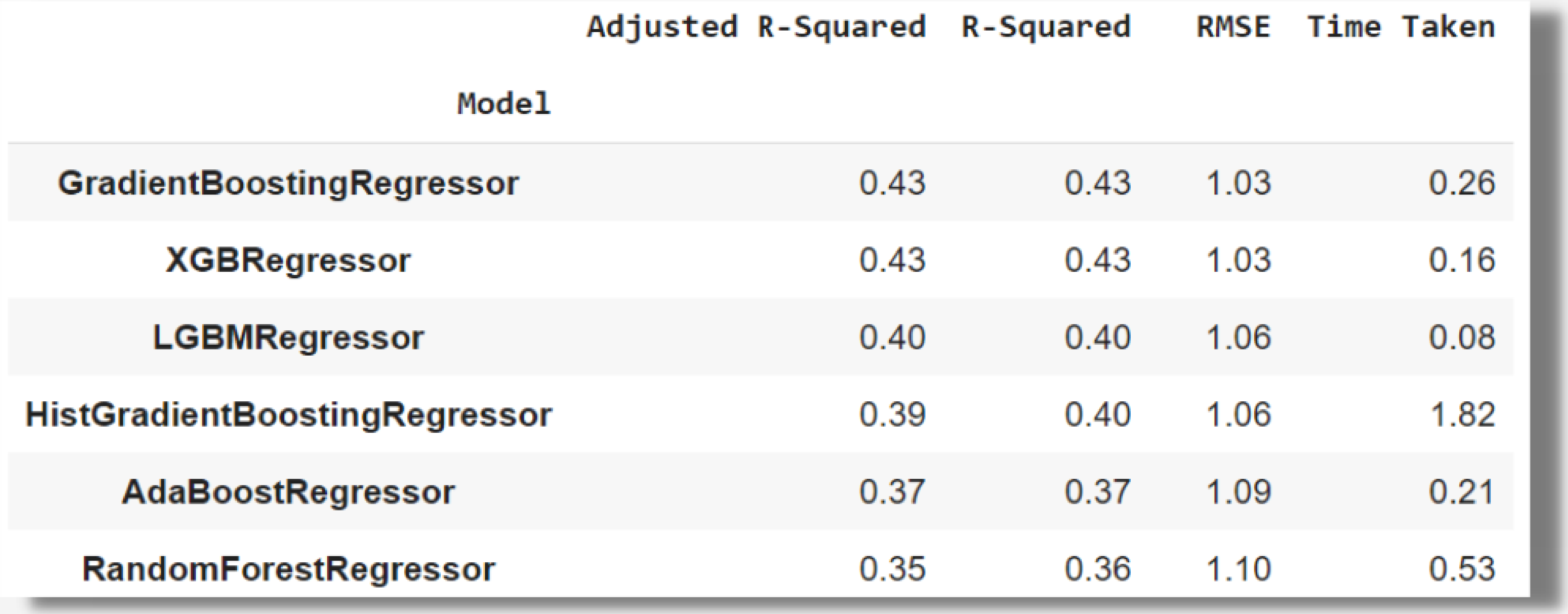
LazyPredict

- Library that allows you to test many algorithms to see which ones might be the best.

- Targets tested:
    - Age
    - Netflix
    - Rotten Tomatoes
    - Year

LazyPredict

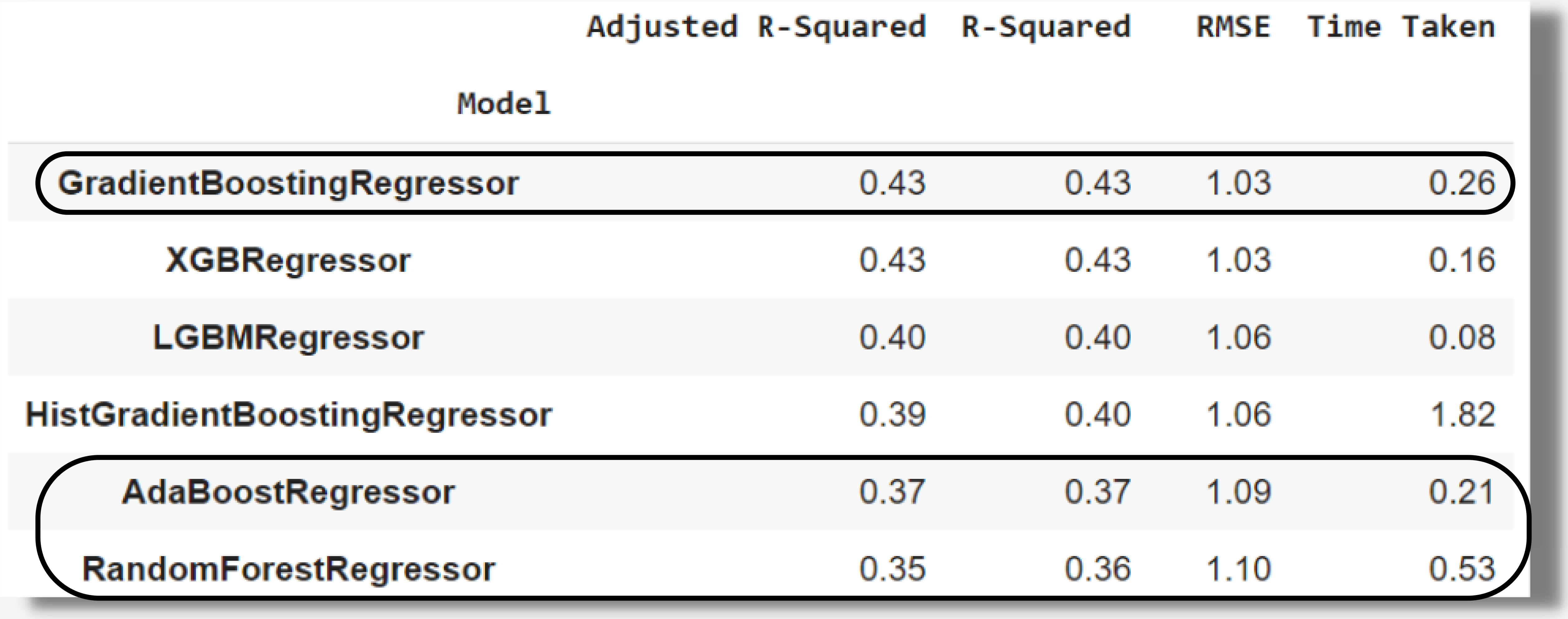
- Target atribute -> Rotten Tomatoes



	Adjusted R-Squared	R-Squared	RMSE	Time Taken
Model				
GradientBoostingRegressor	0.43	0.43	1.03	0.26
XGBRegressor	0.43	0.43	1.03	0.16
LGBMRegressor	0.40	0.40	1.06	0.08
HistGradientBoostingRegressor	0.39	0.40	1.06	1.82
AdaBoostRegressor	0.37	0.37	1.09	0.21
RandomForestRegressor	0.35	0.36	1.10	0.53

LazyPredict

- Target atribute -> Rotten Tomatoes



	Adjusted R-Squared	R-Squared	RMSE	Time Taken
Model				
GradientBoostingRegressor	0.43	0.43	1.03	0.26
XGBRegressor	0.43	0.43	1.03	0.16
LGBMRegressor	0.40	0.40	1.06	0.08
HistGradientBoostingRegressor	0.39	0.40	1.06	1.82
AdaBoostRegressor	0.37	0.37	1.09	0.21
RandomForestRegressor	0.35	0.36	1.10	0.53

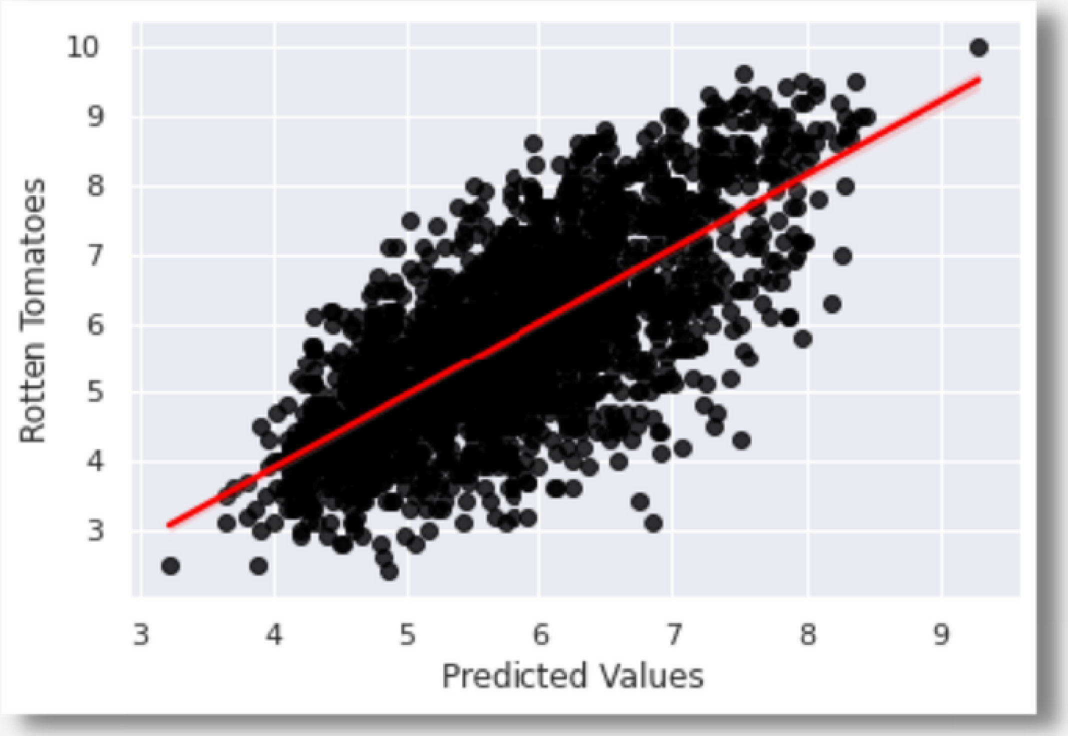
Validation with Train set

- Default parameters for every model
- R2 & MSE for measuring performance

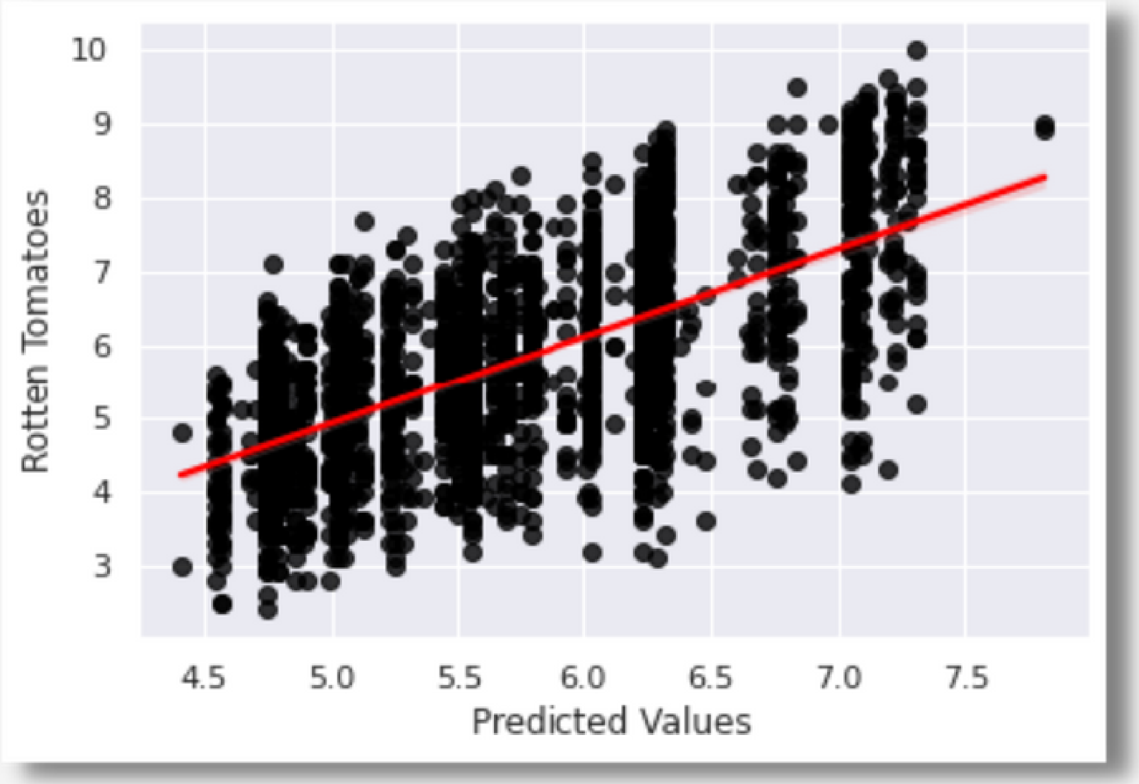
	R2	**MSE**
GradientBoostingRegressor	0.514	0.932
AdaBoostRegressor	0.412	1.126
RandomForestRegressor	0.837	0.311

Validation with Train set

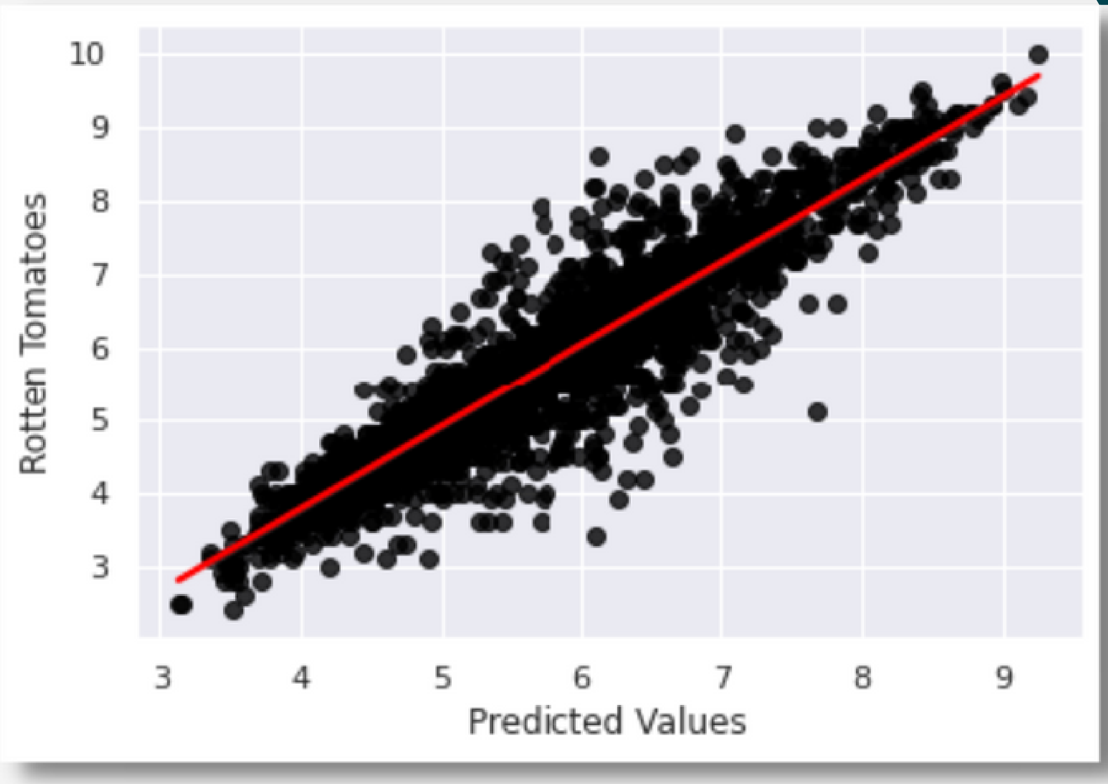
- GradientBoostingRegressor



- AdaBoostRegressor



- RandomForestRegressor



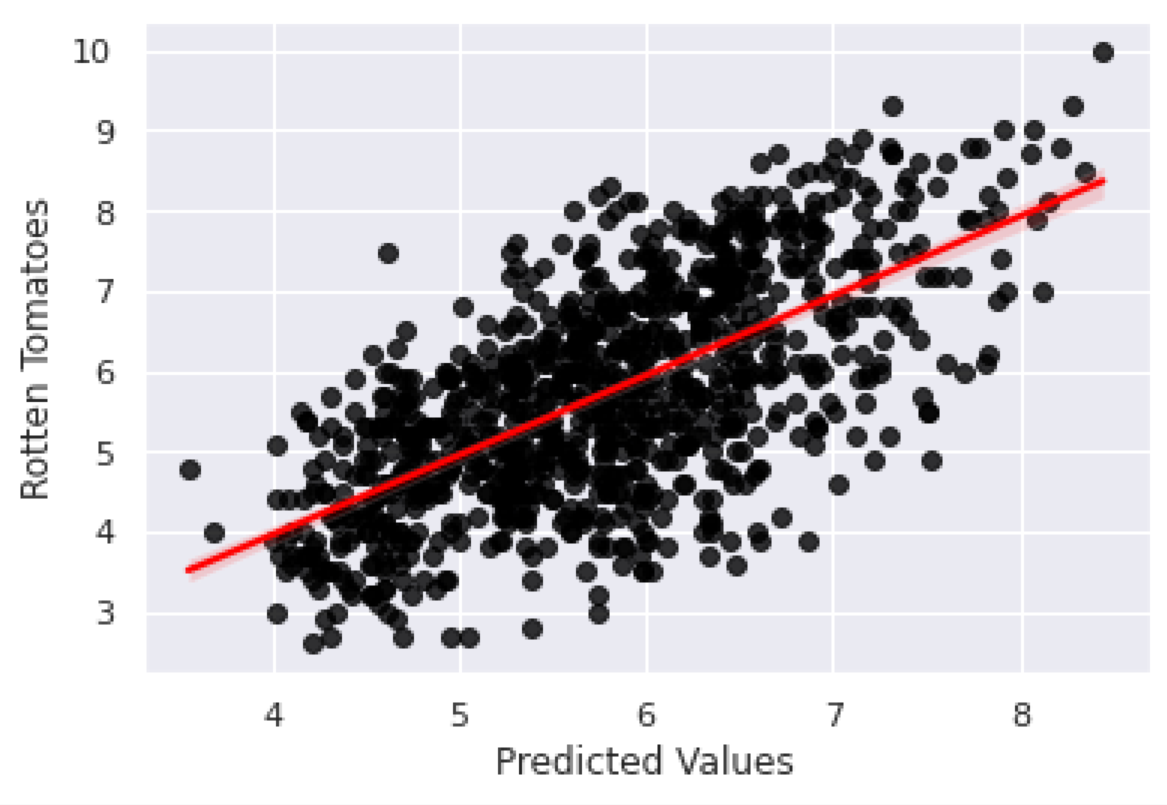
Validation with Test set

- Default parameters for every model
- R2 & MSE for measuring performance

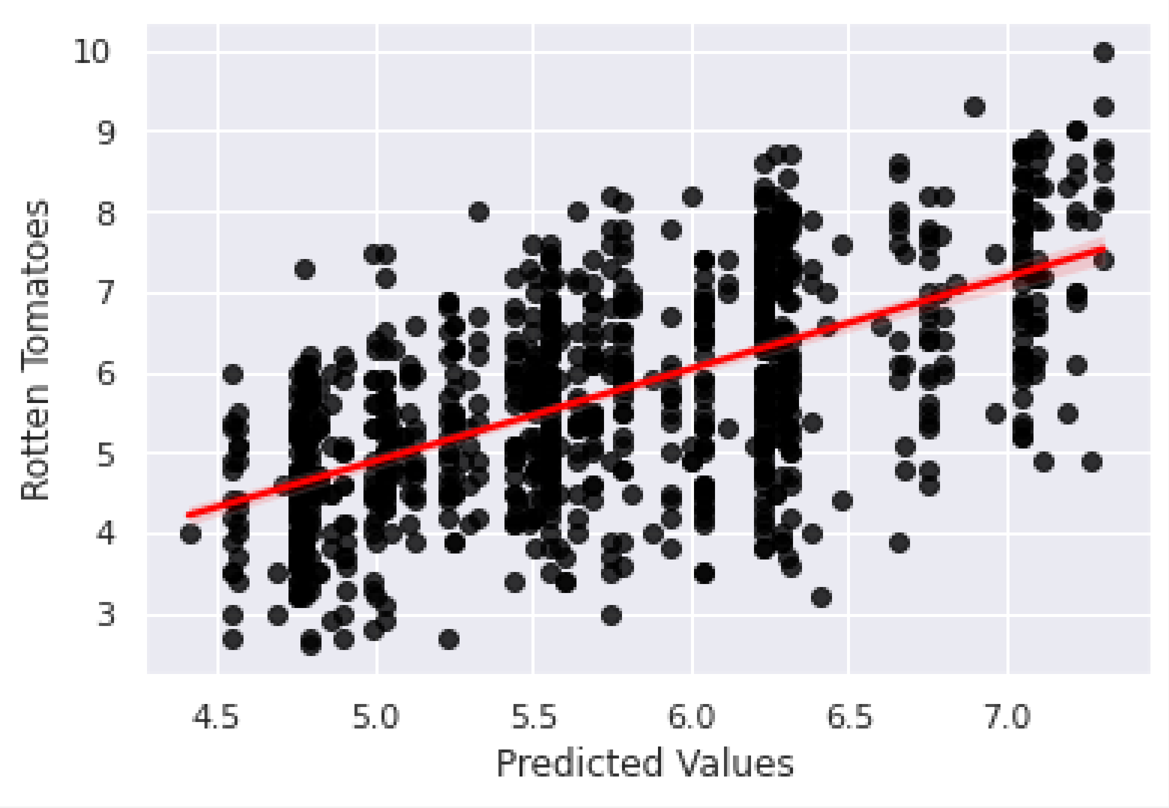
	R2	MSE
GradientBoostingRegressor	0.432	1.063
AdaBoostRegressor	0.376	1.169
RandomForestRegressor	0.351	1.215

Validation with Test set

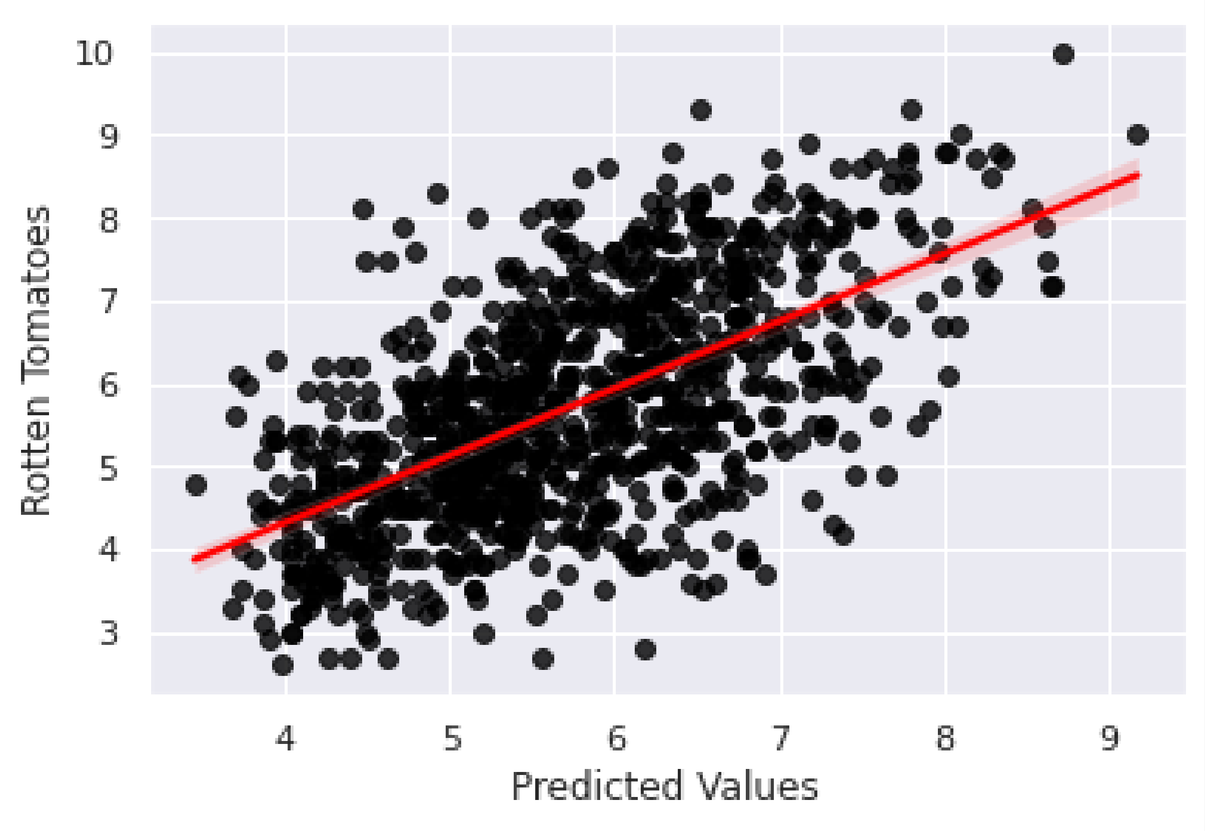
- GradientBoostingRegressor



- AdaBoostRegressor



- RandomForestRegressor



Comparison

- Train set validation

	R2	MSE
GradientBoostingRegressor	0.514	0.932
AdaBoostRegressor	0.412	1.126
RandomForestRegressor	0.837	0.311

- Test set validation

	R2	MSE
GradientBoostingRegressor	0.432	1.063
AdaBoostRegressor	0.376	1.169
RandomForestRegressor	0.351	1.215

Crossvalidation

- With Crossvalidation worst results obtained in each of the 3 algorithms we tested.
  - Worse R2
  - Worse MSE

Hyperparameters search

- GridSearch for every algorithm
- Best parameters:
  - GradientBoostingRegressor:
    - {'learning_rate': 0.01, 'max_depth': 4, 'n_estimators': 500, 'subsample': 0.5}
  - AdaBoostRegressor:
    - {'learning_rate': 0.04, 'n_estimators': 100}
  - RandomForestRegressor:
    - {'max_depth': 6, 'min_samples_split': 2, 'n_estimators': 100}

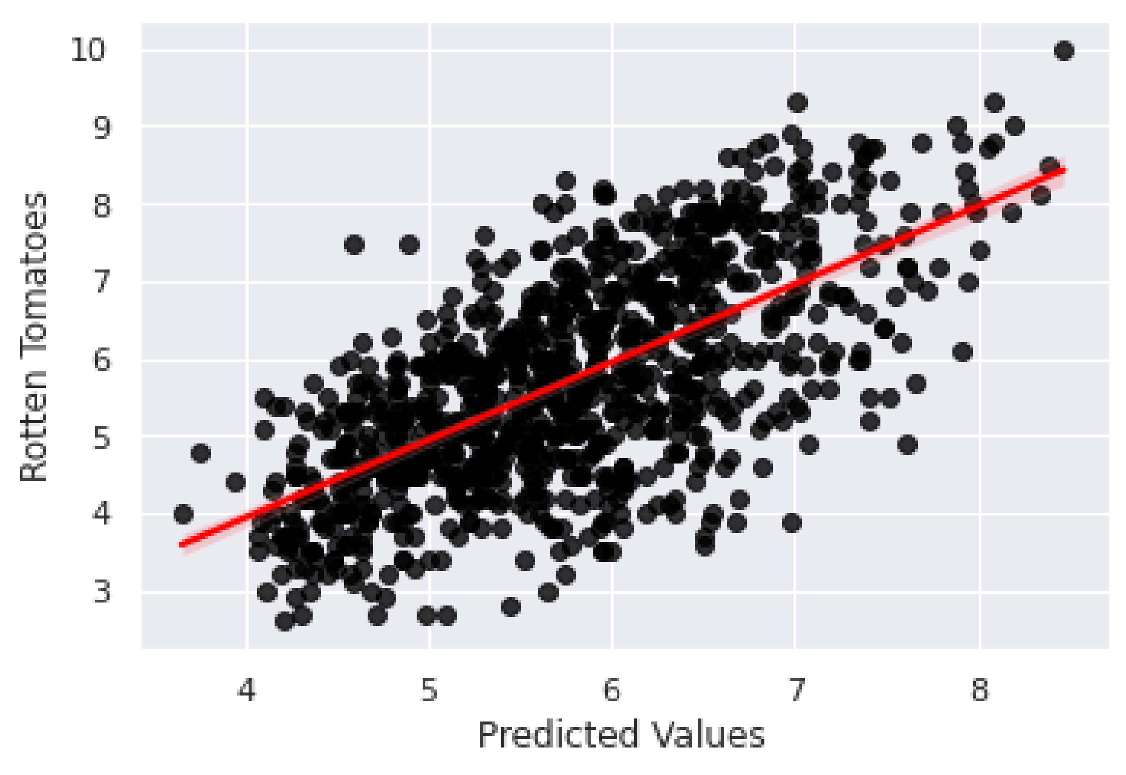
Final models

- GridSearch for every algorithm

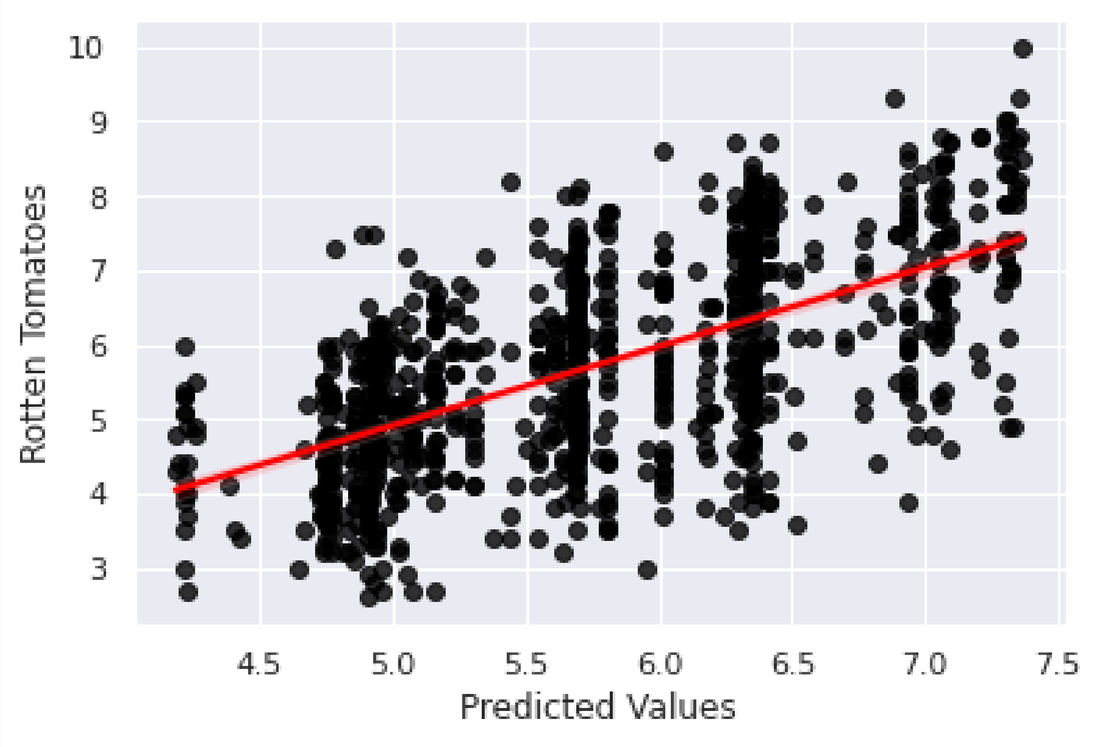
	R2	MSE
GradientBoostingRegressor	0.434	1.060
AdaBoostRegressor	0.374	1.173
RandomForestRegressor	0.424	1.079

Final models

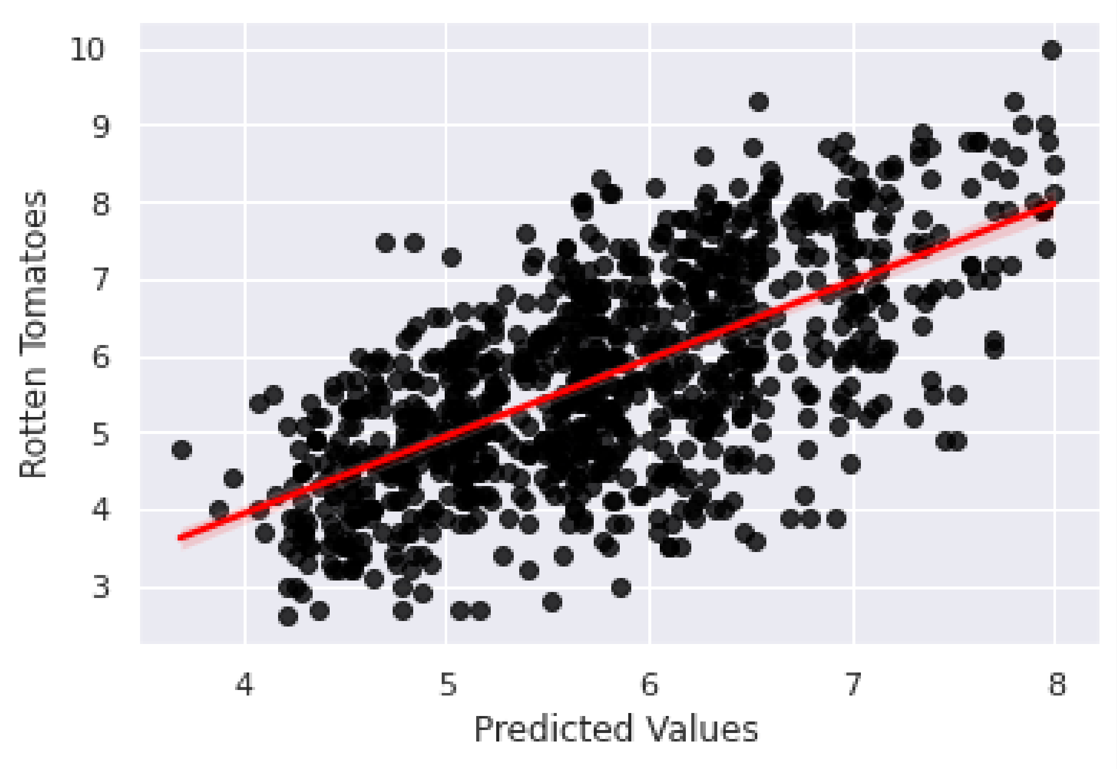
- GradientBoostingRegressor



- AdaBoostRegressor



- RandomForestRegressor



Comparison

- Tuned hyperparameters

	R2	MSE
GradientBoostingRegressor	0.434	1.060
AdaBoostRegressor	0.374	1.173
RandomForestRegressor	0.424	1.079

- Default Hyperparameters

	R2	MSE
GradientBoostingRegressor	0.432	1.063
AdaBoostRegressor	0.376	1.169
RandomForestRegressor	0.351	1.215

Conclusions

- Best Model
  - GradientBoostingRegressor
- CrossValidation has not helped
- Possibilty to expand Dataset with IMDb & Rotting Tomatoes APIs