ITBA
Instituto Tecnológico de Buenos Aires

KAGGLE CHALLENGE
-
SUBMISSION OF A PREDICTIVE MODEL

13/11/2023

82.05 - Análisis Predictivo - Segundo parcial

Elsa DOYEN (65990)

ITBA

AGENDA

01
Data presentation and project objective

02
Data cleaning

03
Testing several models

04
Model selected

05
Conclusion

DATA PRESENTATION AND PROJECT OBJECTIVE

Data :

- 977 541 entries
- 1 response variable "averageRating
- 27 variables (numeric and categorical)

Objective :

obtain the best prediction of the "averageRating" variable in the test dataset

averageRating	numVotes	titleType	isAdult	startYear	endYear	runtimeMinutes	genres_x	directors	...	genr
4.4	15	movie	0.0	1951	0	91	Comedy,Musical	nm0883334	...	
7.0	990	tvSeries	0.0	2007	2021	30	Action,Adventure,Animation	nm2291816,nm3088555,nm4930005,nm1746040	...	
8.1	41	tvEpisode	0.0	2011	0	44	Documentary,History,War	nm0414025	...	
4.6	48	movie	0.0	1969	0	84	Drama	nm2977268	...	
5.6	28	movie	0.0	2010	0	130	Comedy,Drama	nm2366663	...	

DATA CLEANING

averageRating	0
numVotes	0
titleType	0
isAdult	0
startYear	0
endYear	0
runtimeMinutes	0
genres_x	2
directors	0
writers	0
seasonNumber	539298
episodeNumber	539298
ordering	606918
language	606918
attributes	606918
isOriginalTitle	606918
adult	930171
budget	930171
genres_y	930171
original_language	930183
popularity	930172
production_companies	930172
production_countries	930172
revenue	930172
runtime	930383
status	930242
tagline	953696
video	930172

Number of missing values in the dataset

Nan processing :

- Remove variables with more than 75% Nan
- Complete possible variables
- Processing difficult variables ("directors" and "writers")

12 remaining explanatory variables

ITBA TESTING SEVERAL MODELS

Model	R2
Linear Regression	0.19669
Random Forest	0.47565
Gradient Boosting	0.25007
KNN	0.14138

→ Selection of Random Forest model

MODEL SELECTED

Random Forest model

Hyperparameters to test :
- n_estimators
- max_depth
- min_sample_split

Choices :
- n_estimators = 150
- max_depth = 30
- min_sample_split = 2

MODEL SELECTED

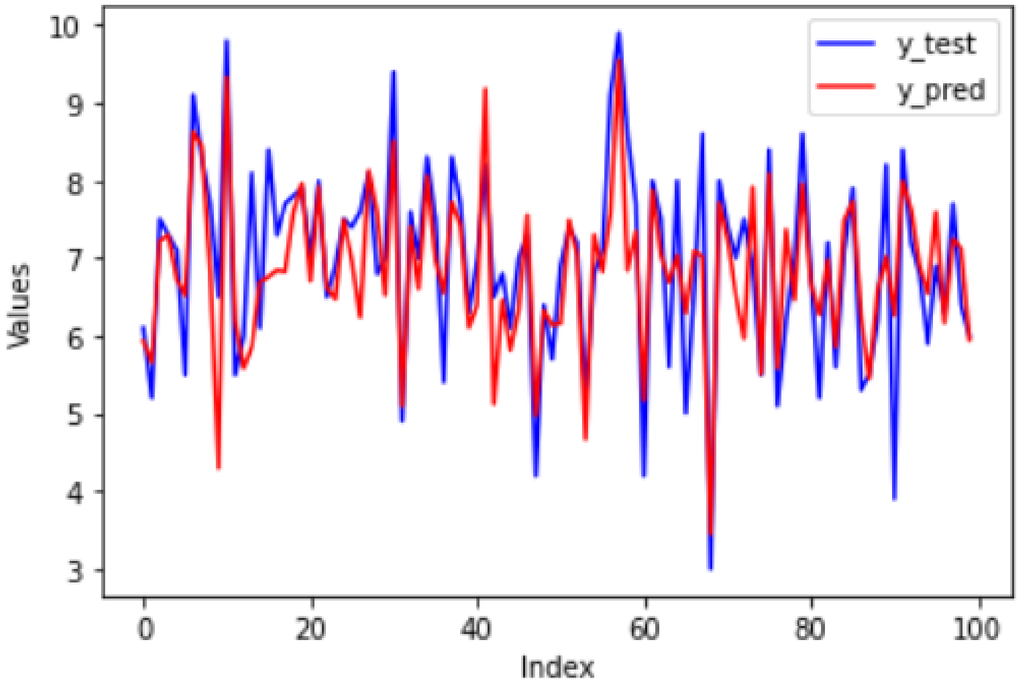
Random Forest model

```python
# Initialization
model = RandomForestRegressor(n_estimators=150, max_depth = 30, random_state=42)

# Training
model.fit(X_train, y_train)

# Prediction
y_pred = model.predict(X_test)

# Evaluation of model performance
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse} and R-squared {r2}')
```

```
Mean Squared Error: 1.016742553319181 and R-squared 0.48356767729733185
```

Score: 0.48490

Public score: 0.47891



Visualization of y_test and y_pred

Conclusion

Limits and improvements