Costa Rica Institute of Technology

Bigdata

Probability of winning an Oscar in a movie based on the IMDb and Rotten Tomatoes ratings

Students:

```bash
Jose Martinez
```

Input Data

Features Oscar Data Set

- `year_film`: year of the film (integer)
- `year_ceremony`: year of the ceremony (integer)
- `ceremony_name`: name of the ceremony (string)
- `category`: category of the ceremony (string)
- `name`: name of the film (string)
- `film`: name of the film (string)
- `winner`: winner of the ceremony (string)

Features IMDB Data Set

- `imdb_title_id`: IMDB title id (string)
- `title`: title of the movie (string)
- `original_title`: original title of the movie (string)
- `year`: year of the movie (integer)
- `date_published`: date of publication of the movie (string)

- `genre`: genre of the movie (string)
- `duration`: duration of the movie (integer)
- `country`: country of origin of the movie (string)
- `language`: language of the movie (string)
- `director`: director of the movie (string)
- `writer`: writer of the movie (string)
- `production_company`: production company of the movie (string)
- `actors`: actors of the movie (string)
- `avg_vote`: average vote of the movie (float)
- `votes`: number of votes of the movie (integer)
- `budget`: budget of the movie (float)
- `usa_gross_income`: gross income of the movie in USA (float)
- `worlwide_gross_income`: gross income of the movie in the world (float)
- `metascore`: metascore of the movie (integer)
- `reviews_from_users`: number of reviews from users of the movie (float)
- `reviews_from_critics`: number of reviews from critics of the movie (float)

Features Rotten Tomatoes Data Set

- `rotten_tomatoes_link`: link to the Rotten Tomatoes page of the movie (string)
- `movie_title`: title of the movie (string)
- `movie_info`: information about the movie (string)
- `critics_consensus`: critics consensus about the movie (string)
- `content_rating`: content rating of the movie (string)
- `genres`: list of genres of the movie (string)
- `directors`: list of directors of the movie (string)
- `authors`: list of authors of the movie (string)
- `actors_rt`: list of actors of the movie (string)
- `original_release_date`: original release date of the movie (string)

- `streaming_release_date`: streaming release date of the movie (string)

- `runtime`: runtime of the movie (integer)

- `production_company_rt`: production company of the movie (string)

- `tomatometer_status`: status of the movie on the tomatometer (string)

- `tomatometer_rating`: rating of the movie on the tomatometer (integer)

- `tomatometer_count`: number of votes of the movie on the tomatometer (integer)

- `audience_status`: status of the movie on the audience (string)

- `audience_rating`: rating of the movie on the audience (integer)

- `audience_count`: number of votes of the movie on the audience (integer)

- `tomatometer_top_critics_count`: number of top critics of the movie on the tomatometer (integer)

- `tomatometer_fresh_critics_count`: number of fresh critics of the movie on the tomatometer (integer)

- `tomatometer_rotten_critics_count`: number of rotten critics of the movie on the tomatometer (integer

Target Variable

- `winner`: winner of the ceremony (string)

Create database

1. Create database

  - cd db/
  - ./run_image.sh

Execute preprocessing

1. Create and docker image

- ./execute_image.sh

2. Enter part1_preprocessing folder

- cd part1_preprocessing/

3. Run preprocessing.py

- ./execute.sh

Execute test

1. Create and docker image

- ./execute_image.sh

2. Enter part1_test folder

- cd part1_test/

3. Run whole tests

- pytest

4. Run whole suite of tests

- pytest test_oscar.py
- pytest test_imdb.py
- pytest test_rotten.py

5. Run specific test

- pytest -k TEST_NAME

Write to DB

1. Create and docker image

- ./execute_image.sh

2. Enter part2_write_db folder

- cd part2_write_db/

3. Run write_to_db.py

- ./execute.sh

Execute model

1. Create and docker image

- ./execute_image.sh

2. Load jupyter notebook

- ./load_jupyter_notebook.sh
- Open url notebook in a browser

3. Enter part3_and_4_sparkml folder

- cd part3_and_4_sparkml/

4. Run the whole jupyter notebook:

- model_project.ipynb