Anatomy of an Oscar

Genre

Director

Actors

Country

Language

Box Office Sales

Ranking

Rotten Tomatoes Score

Metacritic

IMDB

Step 1

Data Mining

- Downloaded a CSV list of movies from IMDB (https://www.imdb.com/list/ls057823854/)
  - Due to limitations of the API we separated the CSV into multiple CSVs
  - Created a list of titles based on the movies in the CSVs.
  - Created a loop that ran the titles through our API (https://www.OMdbapi.com) to pull data about each movie.
- Exported a list of Oscar award winning movies to CSV from wikipedia (https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films)

Step 1

Data Mining

- Exported a list of Box Office numbers to CSV from the numbers (https://www.the-numbers.com/box-office-records/worldwide/all-movies/cumulative/all-time)
- Merged the data together
  - Since the Open Movie Database was missing Box Office and Oscar Awards we needed to merged the movies, oscars and box office tables together.

Step 2

Converting Data Types

- Data set initially included strings, objects, and multiple strings in one cell
- First, dropped unnecessary columns
- Next, dropped dollar signs from sales columns, slashes from rating columns, commas, etc.
- Converted objects to numeric to integer, depending on column
- Converted multiple strings in cell to integer via hashing

Step 3

Transforming Categorical Features

- The machine learning algorithms that we wanted to use required numerical inputs and therefore categorical features needed to be transformed into numerical features.
- Needed to split columns that had multiple categorical features.
- From sklearn import preprocessing.
- Encoded the categorical features to a numerical values.
- Final cleaning of data, confirming all objects were float or integers and converted all NaN to 0.

Best Instructor:

Aash Anand

YES.