DATA SCIENCE TASK

Task:

- **Visualize the relation between features (you can design your own new features based on the given data)**
- **Develop an ML model which, given the name of a director, predicts the release year of his next movie along with its probable genres**

Our Data(Columns):

```
[5 rows x 28 columns]
Index(['color', 'director_name', 'num_critic_for_reviews', 'duration',
       'director_facebook_likes', 'actor_3_facebook_likes', 'actor_2_name',
       'actor_1_facebook_likes', 'gross', 'genres', 'actor_1_name',
       'movie_title', 'num_voted_users', 'cast_total_facebook_likes',
       'actor_3_name', 'facenumber_in_poster', 'plot_keywords',
       'movie_imdb_link', 'num_user_for_reviews', 'language', 'country',
       'content_rating', 'budget', 'title_year', 'actor_2_facebook_likes',
       'imdb_score', 'aspect_ratio', 'movie_facebook_likes'],
      dtype='object')
```

Data I considered(columns)

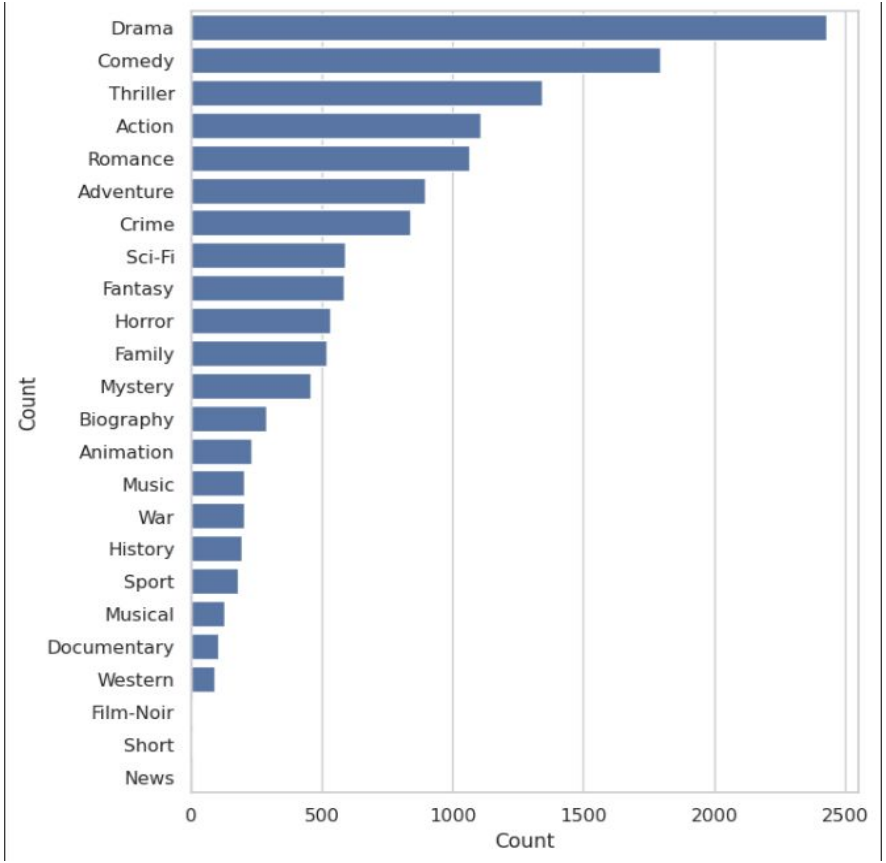
- director_name
- duration
- gross
- genres
- movie_title
- Num_voted_users (total voted users i.e critics + user)
- plot_keywords
- title_year
- imdb_score
- Total_facebook_likes (All possible likes)
- user_reviews

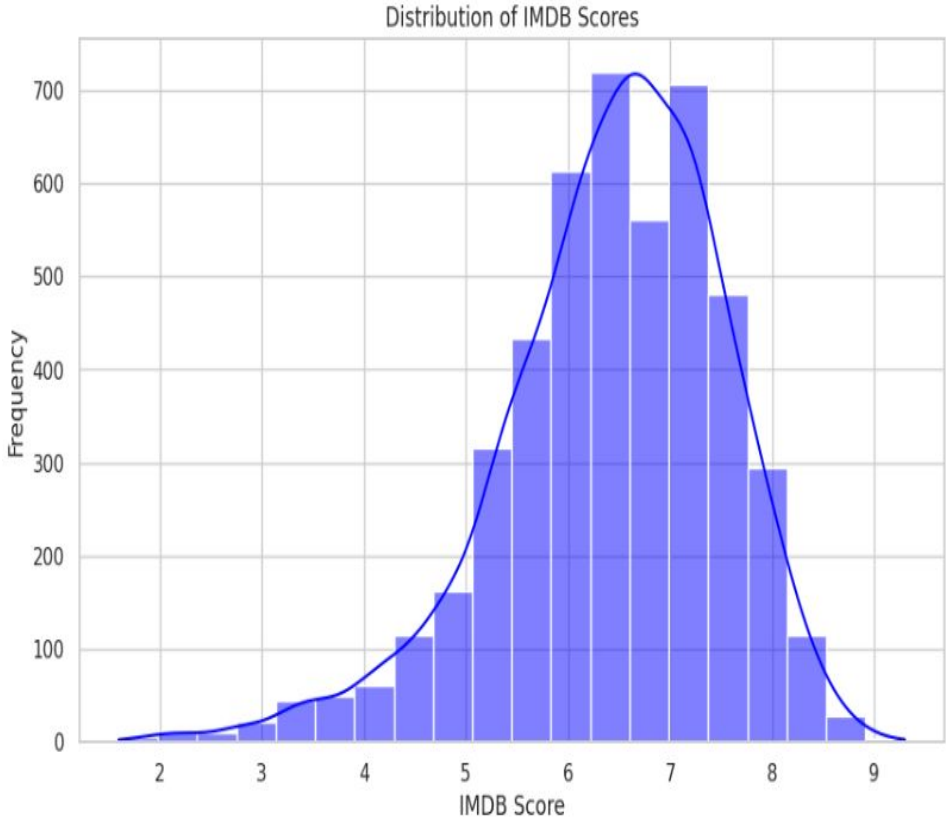
Features engineered:

- Cumulative movies (total movies made by a director over the years by grouping by the title year)
- genre_counts _for_each_director
- time_between_films
- avg _gap_between_movie

And few others..

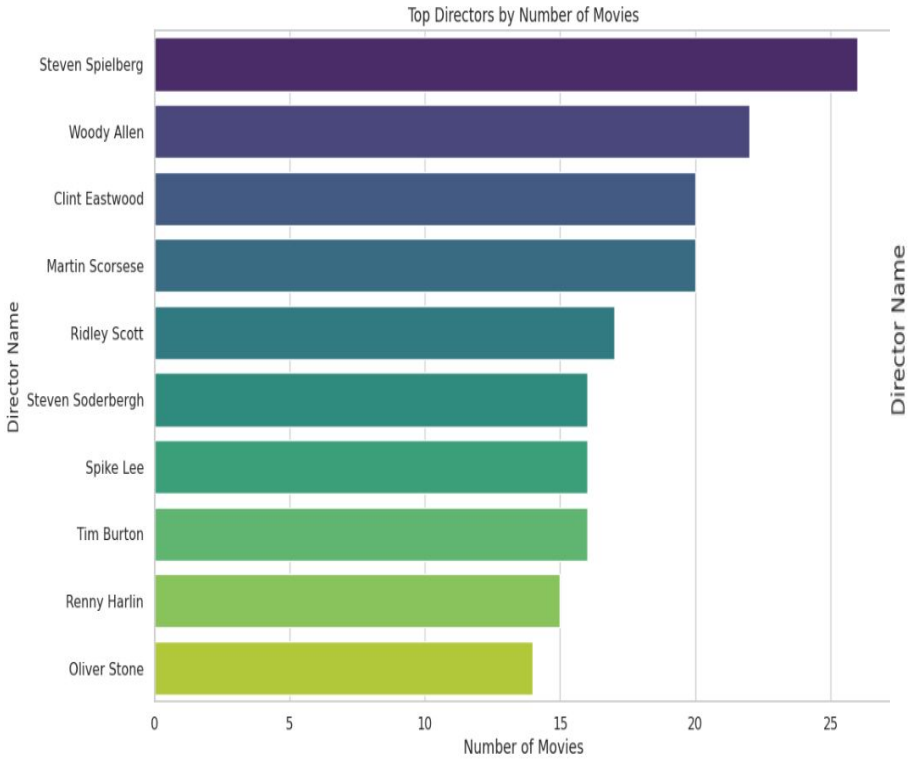
PLOTS

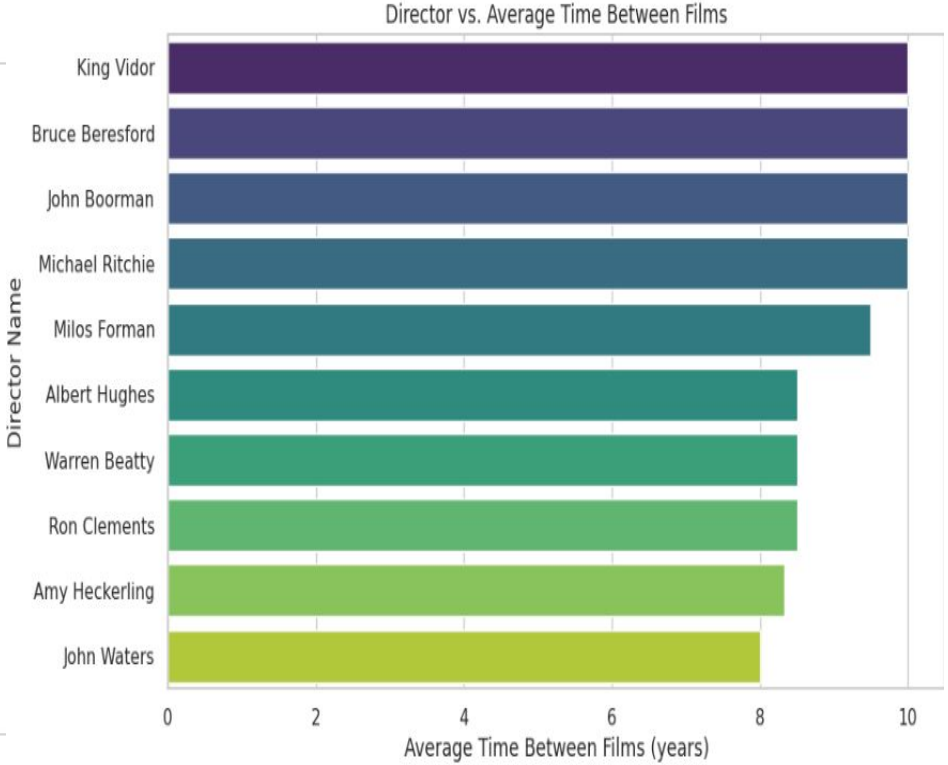




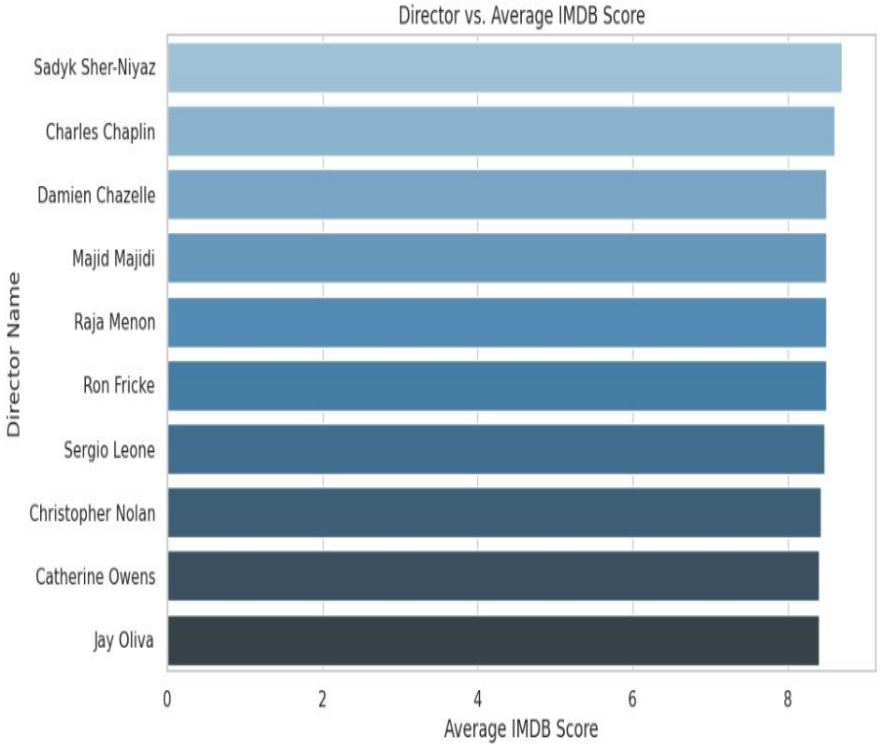
occurrences of each genre

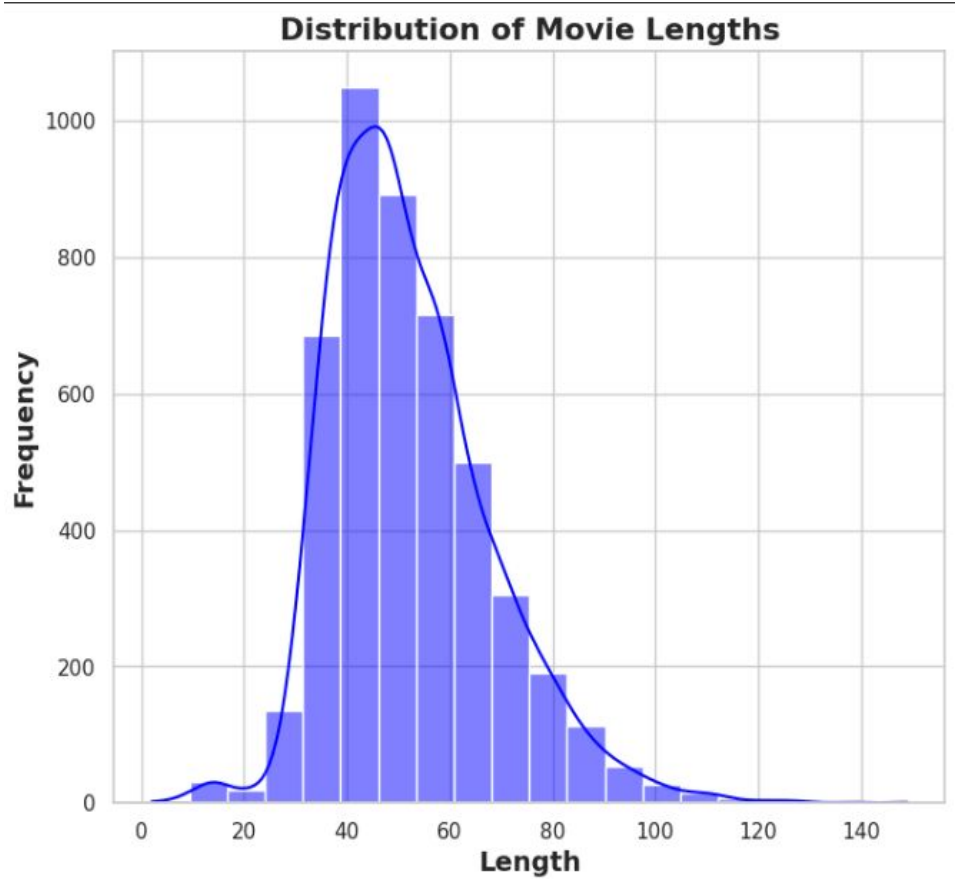
PLOTS

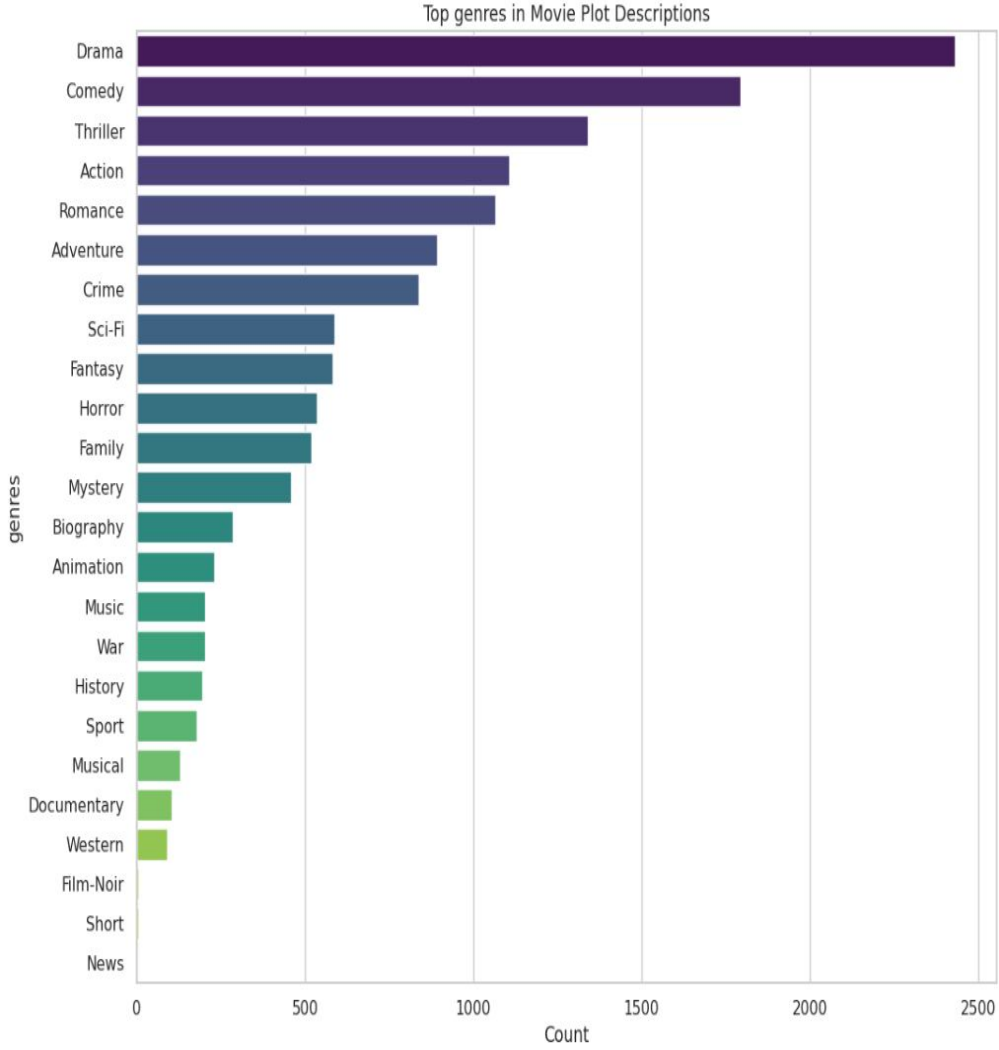

Top Directors by Number of Movies
Steven Spielberg
Woody Allen
Clint Eastwood
Martin Scorsese
Ridley Scott
Steven Soderbergh
Spike Lee
Tim Burton
Renny Harlin
Oliver Stone
Director Name
0
5
10
15
20
25
Number of Movies

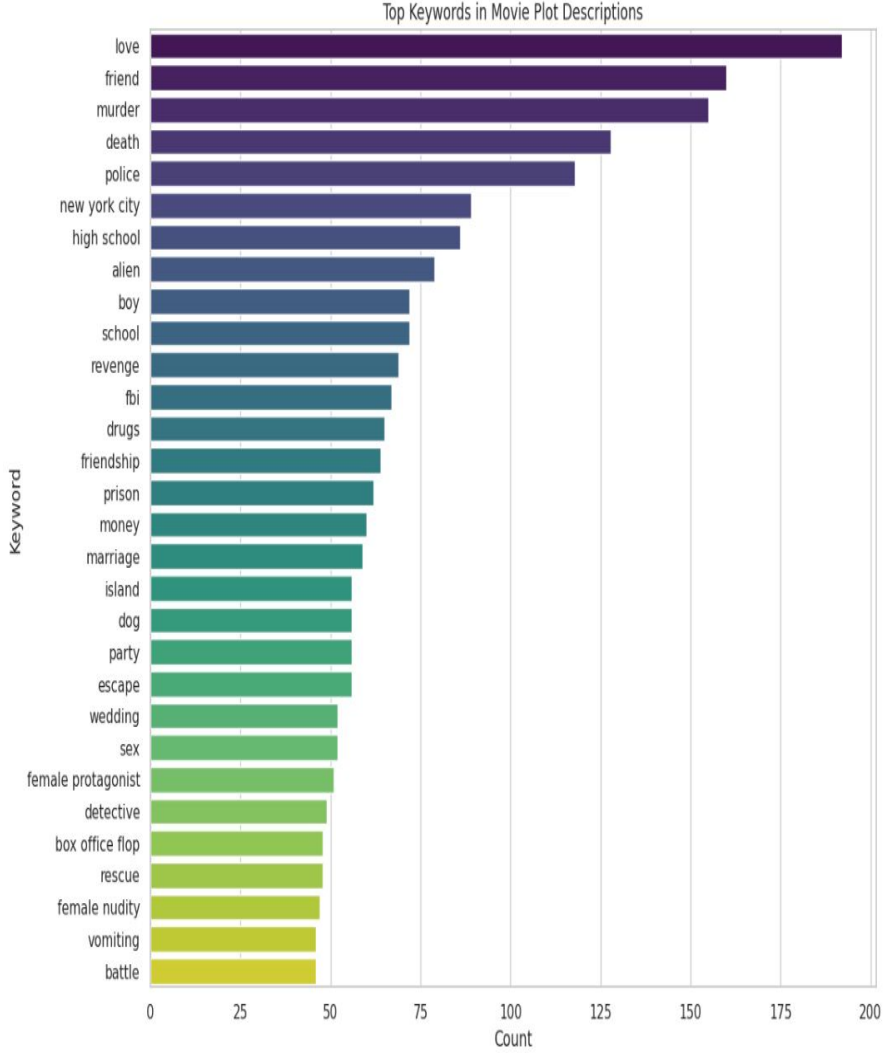

Director vs. Average Time Between Films
King Vidor
Bruce Beresford
John Boorman
Michael Ritchie
Milos Forman
Albert Hughes
Warren Beatty
Ron Clements
Amy Heckerling
John Waters
Director Name
0
2
4
6
8
10
Average Time Between Films (years)

PLOTS


Director vs. Average IMDB Score
Sadyk Sher-Niyaz
Charles Chaplin
Damien Chazelle
Majid Majidi
Raja Menon
Ron Fricke
Sergio Leone
Christopher Nolan
Catherine Owens
Jay Oliva
Director Name
0
2
4
6
8
Average IMDB Score


Distribution of Movie Lengths
1000
800
600
400
200
0
Frequency
0
20
40
60
80
100
120
140
Length


Top genres in Movie Plot Descriptions
Drama
Comedy
Thriller
Action
Romance
Adventure
Crime
Sci-Fi
Fantasy
Horror
Family
Mystery
Biography
Animation
Music
War
History
Sport
Musical
Documentary
Western
Film-Noir
Short
News
genres
0
500
1000
1500
2000
2500
Count


Top Keywords in Movie Plot Descriptions
love
friend
murder
death
police
new york city
high school
alien
boy
school
revenge
fbi
drugs
friendship
prison
money
marriage
island
dog
party
escape
wedding
sex
female protagonist
detective
box office flop
rescue
female nudity
vomiting
battle
Keyword
0
25
50
75
100
125
150
175
200
Count

Model Building for genre classification

1.Preprocessing & Feature Engineering:

- Used TfidfVectorizer to transform plot keywords into TF-IDF features.
- Encoded director names with LabelEncoder.
- Calculated interaction features like IMDb score multiplied by duration and log of number of voted users.

2.Feature & Target Preparation:

- Combined TF-IDF features with encoded director names to form the feature matrix X_genres.
- Used MultiLabelBinarizer to transform genres into a binary matrix y_genres.

3.Data Splitting:

Split the data into training and testing sets using train_test_split.

4.Model Training:
Trained a MultiOutputClassifier with a RandomForestClassifier on the training data.

Other models used:
SVM Classifier
Naive Bayes
Multiple Logistic Regressor

Evaluation:
F1 score :0.64

Model Building for Release Year Prediction

1.Preprocessing & Feature Engineering:

- Calculated non-negative time gaps and average gaps between films.
- Calculated interaction features and transform director names as done in the classifier model setup.
- Calculated interaction features like IMDb score multiplied by duration and log of number of voted users.

2.Feature & Target Preparation:

- Prepared feature matrix X_release_year including IMDb score, duration, interaction features, and average gap.
- Set target y_release_year as current title year plus the average gap.

3.Data Splitting:

Split the data into training and testing sets using train_test_split.

```
Example1: {'Name of Director': 'Christopher Nolan', 'Next movie release': 2014, 'Genres': ('Action', 'Thriller')}
Example2: {'Name of Director': 'James Cameron', 'Next movie release': 2019, 'Genres': ('Action', 'Drama')}
Example1: {'Name of Director': 'Baz Luhrmann', 'Next movie release': 2013, 'Genres': ('Romance', 'Adventure')}
Example2: {'Name of Director': 'Robert Zemeckis', 'Next movie release': 2010, 'Genres': ('Action', 'Adventure')}
Example1: {'Name of Director': 'Peter Sohn', 'Next movie release': 2013, 'Genres': ('Action', 'Adventure')}
Example2: {'Name of Director': 'John Lasseter', 'Next movie release': 2009, 'Genres': ('Action', 'Drama')}
```

4.Model Training:
Trained a RandomForestRegressor on the training data.

Other models used:
SVC Model
Multiple Linear Regressor

Evaluation:
rmse:4.17