Final Project

IMDB Dataset Analysis

Omar Gaballah	201801697
Ahmed AbdelSalam	201801597
Mazen Hassan	201801897
Ibrahim Hamada	201800739

Agenda

- Introduction
- Dataset
- Data Analysis
- Model Analysis
- AWS application
- Conclusion

Introduction

Introduction

In this project, we are analyzing the IMDB reviews dataset. Moreover, we will be applying some Machine Learning models and sentiment analysis on reviews to predict some of the labels of the datasets such as: rating, and spoiler alert. We want to do extensive data analysis on the movies dataset, which should give us some crucial insights that would help us learn plenty of information about the data and apply machine learning models to predict how much help a review is and sentiment analysis on the reviews.

Dataset

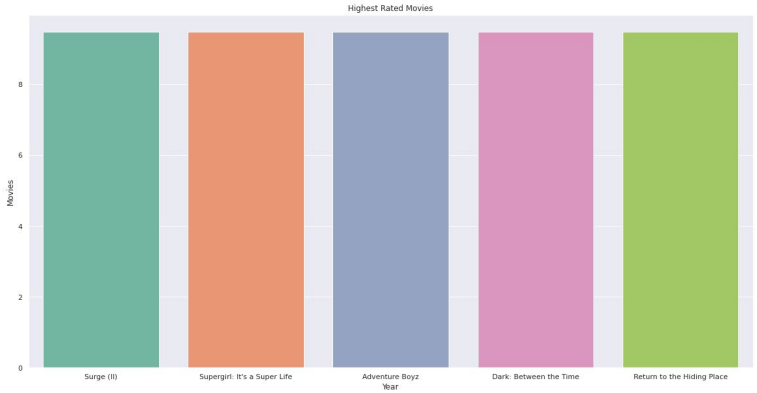
Dataset

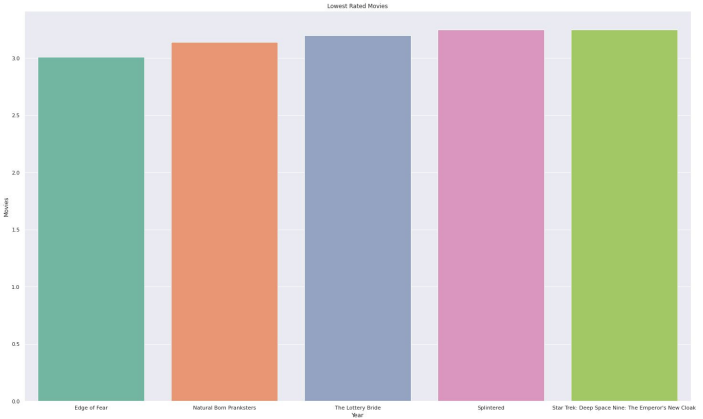
Total reviews: 5, 571, 499
Total shows: 453, 528
Total Users: 1, 699, 310

Content	Details
review_id	It is generated by IMBb and unique to each review
reviewer	Public identity or username of the reviewer
movie	It represents the name of the show (can be - movie, tv-series, etc.)
rating	Rating of movie out of 10, can be None for older reviews
review_summary	Plain summary of the review
review_date	Date of the posted review
spoiler_tag	If 1 = spoiler & 0 = not spoiler
review_detail	Details of the review
helpful	List of people find the review helpful.

Data Analysis

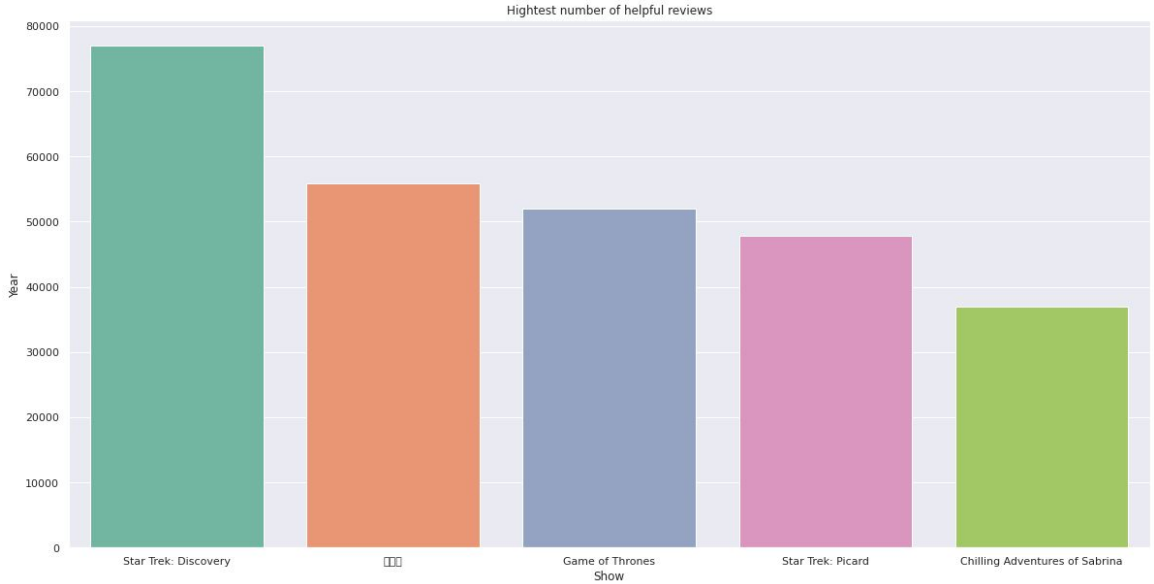
1.The Lowest & Highest Rated Movies





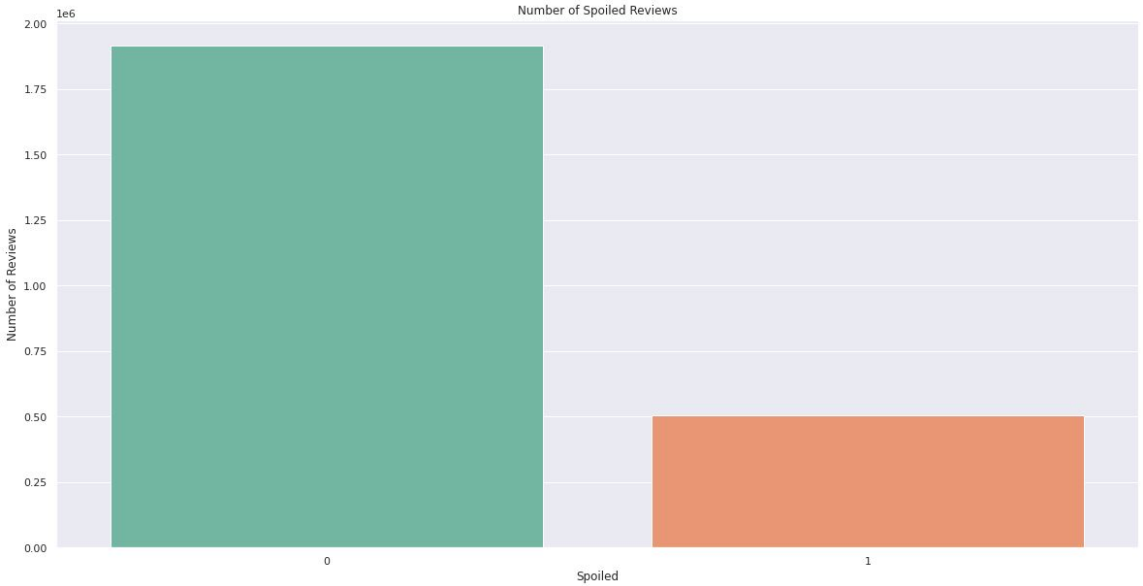
We found that the least 5 rated movies:Edge of Fear,The lottery bride, Star Trek Deep Space, Splintered, Catherine the Great. , we searched for a reference to compare our results with and we searched on IMDB website to compare and found similar results.

2. The highest number of helpful Reviews



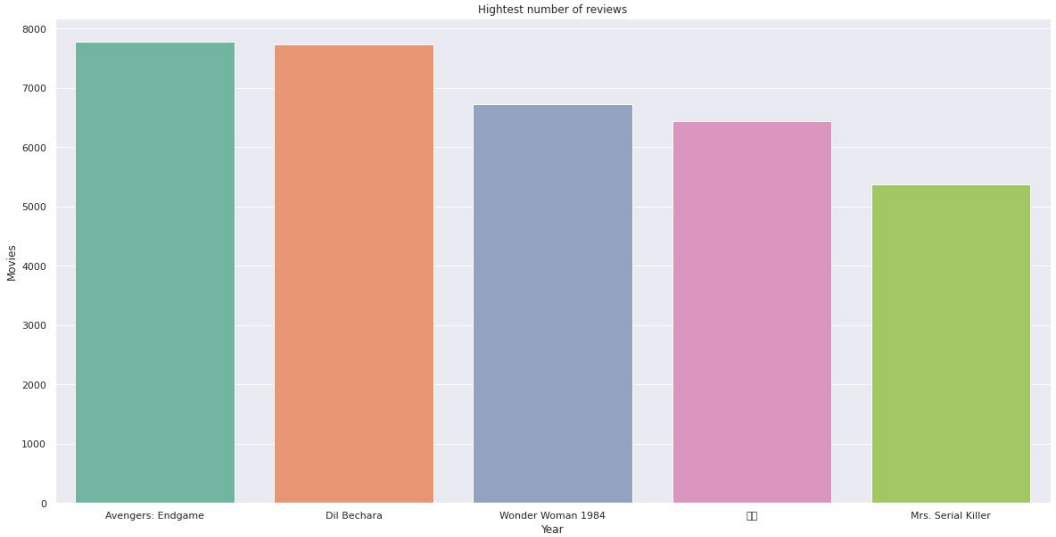
We found that the most 5 Helpful rates movies are: Star Trek which indicates its popularity and with a high helpful community, as well as Game of Thrones Show.

3. Number of Spoiled Reviews



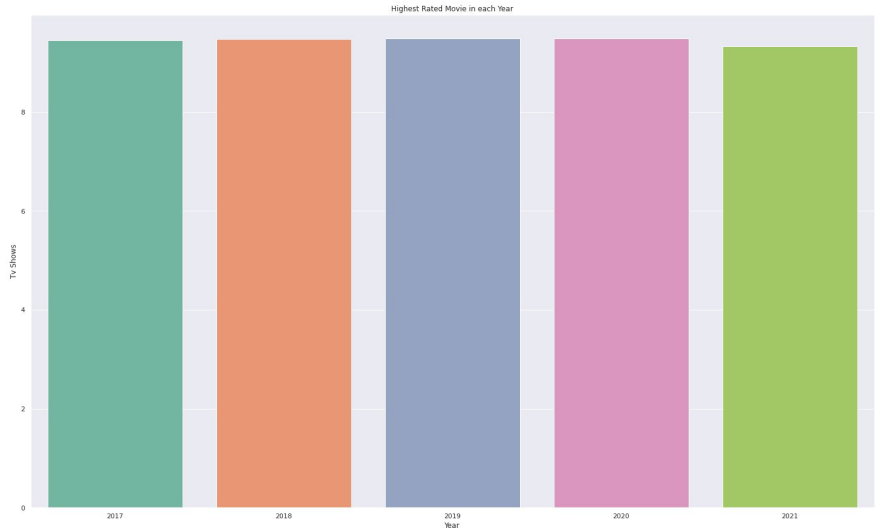
We found that the average Spoilers is around 25% of the comments, therefore it might affect the other people experience before watching the Movie.

4.The Lowest & Most Rated Movies & Series



We found that there is No Correlation between number of reviews and the average rating of a movie, as we can’t find any of these top reviewed movies, they’re not in top rated , movies, as Avengers: Endgame has ~8/10 rating.

5.The Highest Rate Show in Each Year



We found out the Top 5 Rate Show in Each Year: Cobra Kai:in 2021, Surge (II): in 2020, Adventure Boyz in 2019, Tijuana Jackson: Purpose Over Prison in 2018, and Rockaway in 2017, Moreover We can see the Highest rate was lower in 2021 than previous years, and this can be because of the high hopes after the global lockdown, or the slightly worse version of the media industry as it got effect by the lockdown at that time.

Machine Learning Models

SparkML

- In this part, we used SparkML to apply some ML model on the dataset in order to predict some important information.

- We used the Logistic Regression Model, the Naive Bayes Model, and the Support Vector Classifier (SVC) to predict the spoiler_tag given a particular input review.

- We used Linear Regression Model and Decision Tree Regressor Model to predict the Rating of a given review.

STEPS

1) Data Extraction

2) Data Preprocessing

3) Model Training

Model Performances when predicting Spoiler Tag given a review

Model	Logistic Regression	Naïve Bayes	Support Vector Machine
Accuracy	0.7919	0.7840	0.8011
F1-score	0.7181	0.7821	0.7851

Model Performances when predicting Rating given a review

Model	Decision Tree Regressor	GBT Regressor
RMSE	0.21	0.17

AWS Application

AWS’ Environment

- Easy reproducible
- Unified
- Secure

AWS Over Traditional Clusters

Instantaneous Scaling for either

1. the number of nodes
2. Storage Units

Conclusion

Conclusion

1. Automated movie recommendation: to identify similar movies based on movie attributes such as director, cast and/or genre. This type of automated suggestion system would enable IMDB users to discover new films they may enjoy.

2. Movie ratings prediction: Regression algorithms like linear regression and logistic regression can be used to predict ratings for upcoming movies, based on users' history of giving ratings for other IMDb movies. Such predictions will help film releasing companies to set their expectations and plan the publicity accordingly.

3. AI-assisted movie reviewing: By analyzing comments from other reviewers and past content from movie critics, an AI-powered model can suggest high quality reviews which contain useful insights about a given film that complement specific insights from humans writing the copy.

Therefore, with more data, analysis and model optimization, we can perform more powerful and further applications.