Netflix Movies and TV Shows

Genre Prediction

Presented By : Priyasri Sankaran

Introduction

Motivation of the project

- Real life application
- Saves time
- Machine learning Techniques
- Challenges

OVERVIEW

- Problem
- Exploratory Data Analysis
- Analysis
- Objective
- Preprocessing
- Thank You
- Dataset
- Implementation

PROBLEM

Why to predict Genre:-

Benefits the users

Recommendation system

Marketing development

Given:

Metadata about the Netflix movies and TV Show Genre Prediction

OBJECTIVES

- Prediction using the key features
- Machine Learning Techniques
- Exploratory Data Analysis
- Evaluating the performances/ Analysis

Dataset

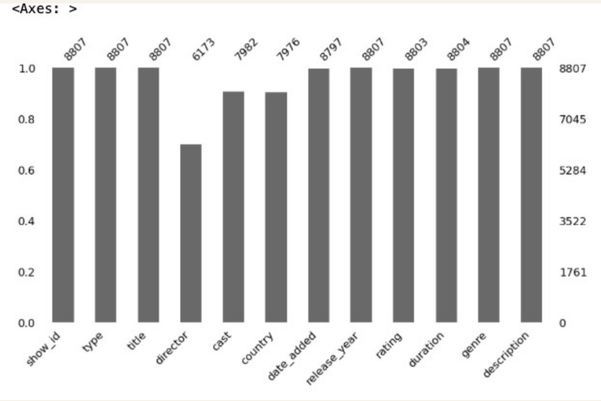
1. Data Source: Kaggle.com

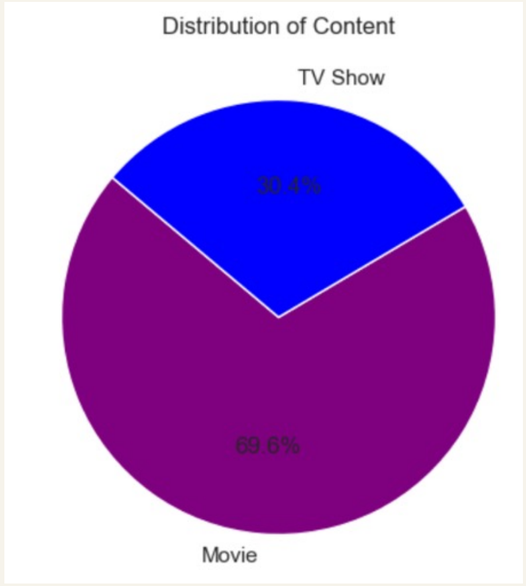
2. Data Size-8807

3. Features –show_id, type, title, director, description, cast, country, data_added, release_year, rating, genre and duration.

Exploratory Data Analysis

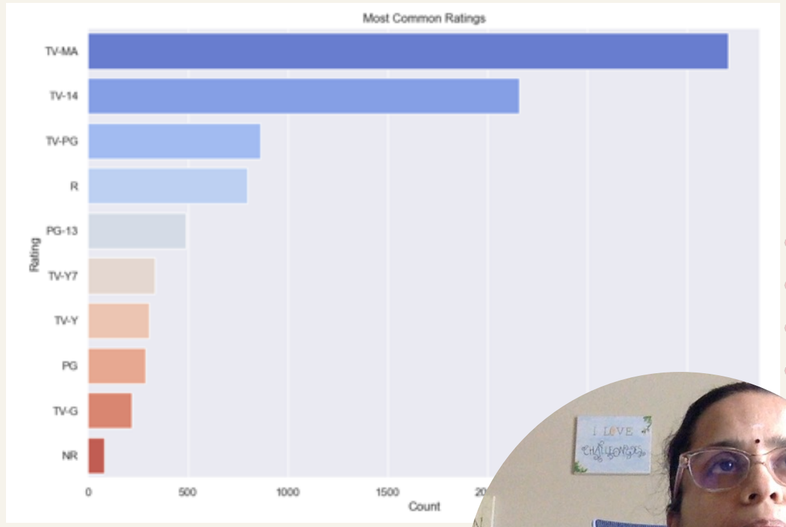
The above image shows the missing values





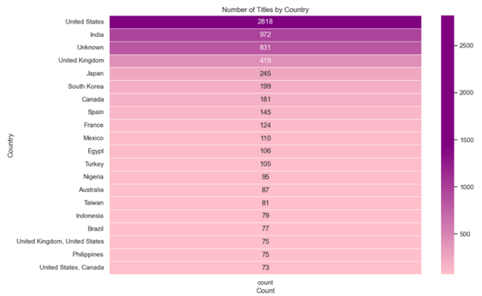
Distribution of Movies and TV Shows

Most Common Rating



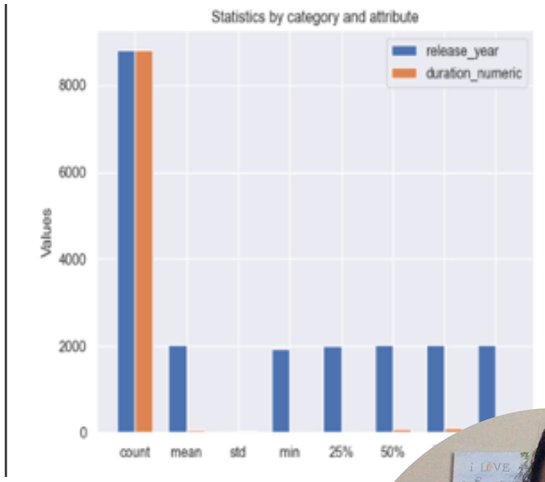
Exploratory Data Analysis

Plot- Titles by count by country

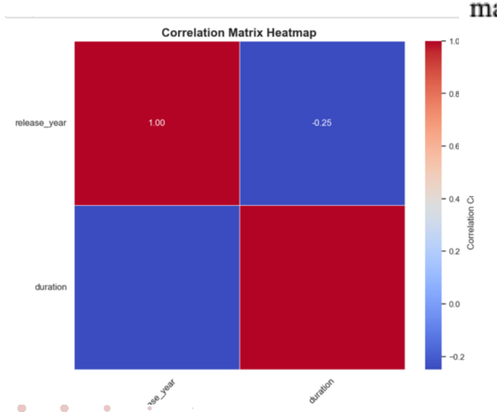


Statistics by duration and release year

	release_year	duration_numeric
count	8807.000000	8804.000000
mean	2014.180198	69.846888
std	8.819312	50.814828
min	1925.000000	1.000000
25%	2013.000000	2.000000
50%	2017.000000	88.000000
75%	2019.000000	106.000000
max	2021.000000	312.000000



Correlation Matrix Heatmap



Preprocessing

- Data involves text
- Preprocessing is an essential step for building a model.
- Well processed data is easy to work with.
- The performance will be better without compromising the information.

Preprocessing

- NLTK-Tokenizing
- Punctuations
- Stopwords
- Lowercase
- Grouping and categorizing

IMPLEMENTATION

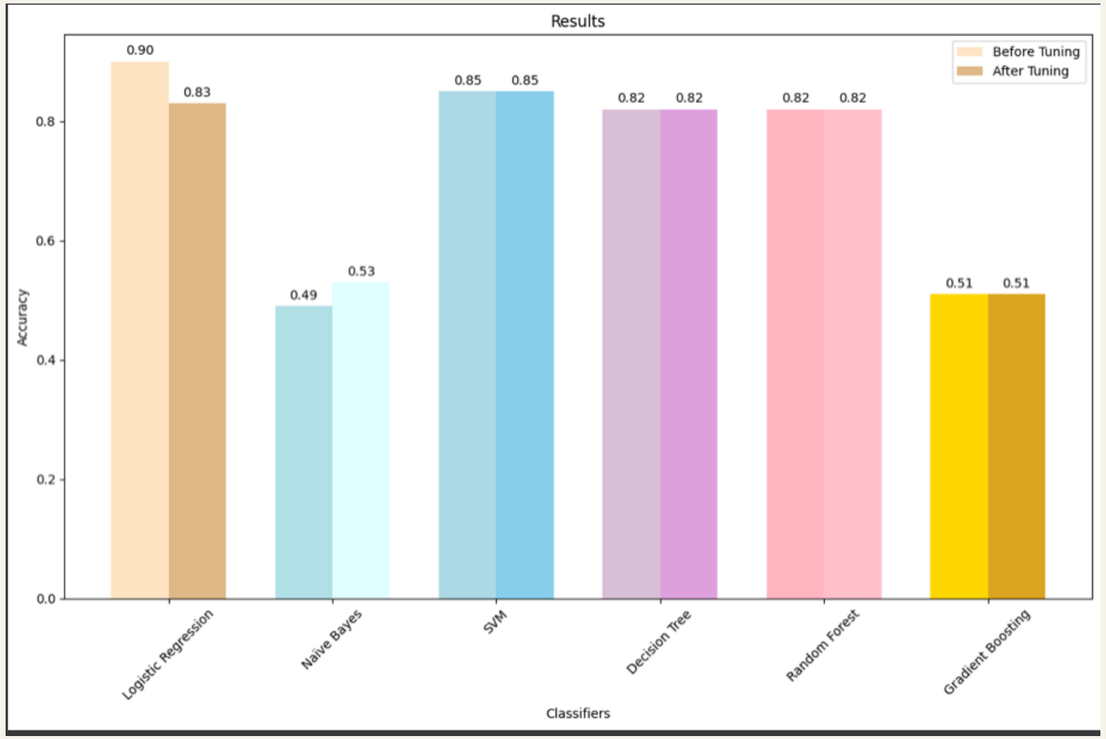
- Logistic Regression
- Random Forest
- Support Vector Machine
- Naïve Bayes
- Gradient Boosting
- Decision Tree

Standardized the data and hyperparameter tuned. It helped the performance when I stand
data. Hyperparameter tuning also helped improve in model development. Performance m
applied.

Analysis

The accuracy of the SVM is 0.85. We notice the Logistic regression accuracy before and after the hyperparameter tuning, we notice that the accuracy is reduced to 0.83. It might be because it's struggling to capture the data structure. Then decision tree and random forest with both with accuracy value of 0.82 . Naïve Bayes performance is moderate with accuracy of 0.53. Gradient boosting gives the accuracy of 0.51. Overall performance of SVM and Logistic Regression is good ~ 85% accuracy. Below is the visual of the combined results for comparison. Also performed precision, recall, F1 score and cross validation.

RESULTS


Results
Before Tuning
After Tuning
Accuracy
0.0
0.2
0.4
0.6
0.8
0.90
0.83
0.49
0.53
0.85
0.85
0.82
0.82
0.82
0.82
0.51
0.51
Logistic Regression
Naive Bayes
SVM
Decision Tree
Random Forest
Gradient Boosting
Classifiers

Lesson Learnt

- Preprocessing and how it impacts model development
- Categorical data handling
- Exploratory Data Analysis(EDA)
- Model development with various algorithms
- Helps me solve some challenges
- Libraries are wonderful resource

UTD | Fall 2024

THANK YOU

Presented By : Priyasri Sankaran

Slide design from Canva.com