NETFLIX DATA

PROJECT AT UNIFIED MENTOR

Flow Chart

MongoDB

DATA INGESTION → E.D.A → PREPROCESSING → MODEL TRAINING → MODEL EVALUATION

Best Model

MODEL EVALUATION → DOCKER IMAGE

Github Action

DOCKER IMAGE → AWS EC2 → Web App

ABOUT DATASET

01 Netflix is a popular streaming service that offers a vast catalog of movies, TV shows, and original contents. Thisdataset is a cleaned version of the original version which can be found here.

02 The data consist of contents added toNetflix from 2008 to 2021. The oldest content is as old as 1925 and the newest as 2021. This dataset will becleaned with PostgreSQL and visualized with Tableau.

03 The purpose of this dataset is to test my data cleaning and visualization skills.

This project involves loading, cleaning, analyzing, and visualizing data from a Netflix dataset. We'll use Python libraries like Pandas, Matplotlib, and Seaborn to workthrough the project. The goal is to explore the dataset, derive insights, and prepare for potential machine learning tasks

STEPS FOLLOW

01 **Treat the Nulls**

02 **Treat the duplicates**

03 **Populate missing rows**

04 **Drop unneeded columns**

05 **Extra Features / Split Columns**

ABOUT DATASET

Libraries

```
#importing basic libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
%matplotlib inline
warnings.filterwarnings("ignore")
import sweetviz  as sv
import ydata_profiling as yd
from wordcloud import   WordCloud
from sklearn.model_selection import GridSearchCV , RandomizedSearchCV
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import learning_curve
from sklearn.tree import plot_tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier , GradientBoostingClassifier , AdaBoostClassifier
from  xgboost import XGBClassifier
from sklearn.model_selection import train_test_split , KFold , cross_val_score
import os ,sys
```

Data Types

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8790 non-null   object
 1   type          8790 non-null   object
 2   title         8790 non-null   object
 3   director      8790 non-null   object
 4   country       8790 non-null   object
 5   date_added    8790 non-null   object
 6   release_year  8790 non-null   int64 
 7   rating        8790 non-null   object
 8   duration      8790 non-null   object
 9   listed_in     8790 non-null   object
dtypes: int64(1), object(9)
```

8790 Rows

10 Columns

Nulls and Duplicated Values

```
#there is no duplicate values inside a dataset
df.duplicated().sum()
```

```
np.int64(0)
```

```
#checking there is any null values inside a data
df.isnull().sum().sum() #no null value
```

```
np.int64(0)
```

Features Available

```
df.head()
```

Python

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 min	Documentaries
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 min	Children & Family Movies, Comedies
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies

PROJECT OVERVIEW

This project aimed to analyze and model data for predictive analysis on content types, specifically classifying them as "Movie" or "TV Show." The core goals were as follows:

1. **Conduct Exploratory Data Analysis (EDA) to understand the dataset, identify important features, and assess any data irregularities.**
2. **Implement Data Preprocessing techniques, including handling missing values, transforming categorical data, and performing feature engineering to make the data compatible for model training.**
3. **Use Machine Learning Model Training to classify content type, tuning the model for optimal performance, and evaluating it based on selected performance metrics**

EXPLORATORY DATA ANALYSIS

- Data Structure and Feature Analysis: We examined the initial data structure, types, and counts, highlighting essential features such as director, country, rating, duration, and listed_in.
- Summary Statistics: Basic statistics provided insights into numerical data distributions and highlighted the presence of categorical features needing transformation
- Visual Exploration: Data distributions for numerical features and value counts for categorical features were visualized, providing a quick understanding of feature variability and significance.
- Correlation Analysis: Correlations among features were evaluated to uncover relationships, giving insights for feature engineering and model selection.

Findings from EDA

- Feature Importance: Key features like country, rating, and listed_in are potentially influential in predicting content type.
- Data Quality: Certain columns contained missing values, notably in director and country, which were addressed during preprocessing.

DATA PREPROCESSING

- Handling Missing Values: Missing values were managed using SimpleImputer to fill gaps, ensuring no loss of data in critical columns.
- Encoding Categorical Features: Label encoding and one-hot encoding were used to convert categorical features into a machine-readable format. Features like country and rating were one-hot encoded, while content_type was label-encoded for binary classification.
- Duration Transformation: The duration feature was converted to a numerical format, allowing it to be used in the model without compatibility issues.
- Data Splitting: The dataset was split into training and test sets to ensure that model evaluation was conducted on data the model hadn’t seen before.

K**ey reprocessing Outcomes**

- Prepared Data: The dataset was now entirely numerical and compatible with machine learning models, improving the accuracy of predictive analysis.
- Balanced Classes: Balancing techniques ensured an even representation of classes, optimizing model performanc

MODEL TRAINING AND HYPERPARAMETER TUNING

The chosen model, XGBoost, was trained on the processed data to classify content type:

- Model Selection: XGBoost, known for high accuracy and efficiency, was selected due to its adaptability in handling categorical data through encodings.
- Hyperparameter Optimization: Hyperparameter tuning was performed using Randomized Search, which was faster than exhaustive Grid Search. This approach allowed optimal parameter selection while minimizing training time.
- Training Process: Model training involved fitting the data to XGBoost, capturing underlying patterns in the features relevant to content type classification.

MODEL EVALUATION

- Metrics: We evaluated the model using accuracy, precision, recall, and F-beta scores, assessing the effectiveness of predictions. A confusion matrix was also generated, providing a clear view of correct and incorrect predictions.
- Model Performance: The model achieved satisfactory performance, with balanced precision and recall scores, indicating reliable predictions for both content classes.

PROJECT REPORT

SUMMARY

Type: Movies and TV Shows
Countries: US, India, UK, and more

1. **Growth Trend**: Netflix releases increased steadily from 2017 to 2020 but have slowed down since then. As of July, there were still more TV shows and movies being released.

2. **Popular Categories** Documentaries and International Movies are very popular, showing Netflix's wide variety of content. Comedy and Drama also get a good amount of attention, appealing to different tastes.

3. **Genre Variety**: The range of genres suggests Netflix is trying to reach more people with different interests.

4. **Top TV Shows**: Most popular TV shows are International, with the US leading, followed by India and the UK.

5. **Country Influence**: The US has the most content on Netflix, contributing the largest share. This strong presence shows Netflix's big influence worldwide, especially in the Drama and Kids categories.

6. **Movies vs. TV Shows**: The dataset reveals that Netflix offers more movies than TV shows overall.

In this project, we successfully built a classification model to predict content type (Movie/TV Show) based on various features. Key takeaways include:

1. Effective Feature Engineering: Handling categorical data effectively through encoding and transforming duration values contributed significantly to model success.
2. Optimized Model: The use of XGBoost with Randomized Search ensured a robust model configuration, achieving high performance in classification accuracy and reliability.
3. Future Improvements: Further refinement could include exploring ensemble models or advanced feature selection techniques to improve accuracy even more.

THANK YOU