NETFLIX MOVIE RECOMMENDATION SYSTEM

Netflix is an application that keeps growing bigger and faster with its popularity, shows and content. This is a story telling through its data along with a content-based recommendation system and a wide range of different graphs and visuals

Problem Statement :

Recommendation systems have been around with us for a while now, and they are so powerful. They do have a strong influence on our decisions these days. From movie streaming services to online shopping stores, they are almost everywhere we look. If you are wondering how they know what you might buy after adding an “x” item to your cart, the answer is simple: Power of Data.

Recommendation systems are a very interesting field of machine learning. Recommendation system recommends the movie based on the users movie choices.

Data Source :

MovieLens Dataset : https://grouplens.org/datasets/movielens/

This dataset contains 100,000 ratings and 3,600 tag applications applied to 9,000 movies by 600 users.

FEATURE DESCRIPTION :

- MovieId (Quantitative) : Uniquely Identifies the movie
- Title (Qualitative) : Title of the Movie
- Genres (Categorical) : Defines a movie based on narrative elements
- UserId (Quantitative): Uniquely Identifies the user
- Ratings (Quantitative) : Rating of the movie
- Tag (Qualitative): Defines the type of movie
- TimeStamp (Quantitative): Timestamp of the rating

SAMPLE DATASET :

Movies.csv

<table>
<tr><th>movieId</th><th>title</th><th>genres</th></tr>
<tr><td>1</td><td>Toy Story (</td><td>Adventure|Animation|Children|Comedy|Fantasy</td></tr>
<tr><td>2</td><td>Jumanji (1</td><td>Adventure|Children|Fantasy</td></tr>
<tr><td>3</td><td>Grumpier (</td><td>Comedy|Romance</td></tr>
<tr><td>4</td><td>Waiting to</td><td>Comedy|Drama|Romance</td></tr>
<tr><td>5</td><td>Father of t</td><td>Comedy</td></tr>
<tr><td>6</td><td>Heat (199</td><td>Action|Crime|Thriller</td></tr>
<tr><td>7</td><td>Sabrina (1</td><td>Comedy|Romance</td></tr>
<tr><td>8</td><td>Tom and H</td><td>Adventure|Children</td></tr>
<tr><td>9</td><td>Sudden De</td><td>Action</td></tr>
<tr><td>10</td><td>GoldenEye</td><td>Action|Adventure|Thriller</td></tr>
<tr><td>11</td><td>American I</td><td>Comedy|Drama|Romance</td></tr>
<tr><td>12</td><td>Dracula: D</td><td>Comedy|Horror</td></tr>
<tr><td>13</td><td>Balto (199</td><td>Adventure|Animation|Children</td></tr>
<tr><td>14</td><td>Nixon (19</td><td>Drama</td></tr>
<tr><td>15</td><td>Cutthroat</td><td>Action|Adventure|Romance</td></tr>
<tr><td>16</td><td>Casino (19</td><td>Crime|Drama</td></tr>
<tr><td>17</td><td>Sense and</td><td>Drama|Romance</td></tr>
<tr><td>18</td><td>Four Roon</td><td>Comedy</td></tr>
<tr><td>19</td><td>Ace Ventu</td><td>Comedy</td></tr>
<tr><td>20</td><td>Money Tra</td><td>Action|Comedy|Crime|Drama|Thriller</td></tr>
<tr><td>21</td><td>Get Shorty</td><td>Comedy|Crime|Thriller</td></tr>
<tr><td>22</td><td>Copycat (1</td><td>Crime|Drama|Horror|Mystery|Thriller</td></tr>
<tr><td>23</td><td>Assassins (</td><td>Action|Crime|Thriller</td></tr>
<tr><td>24</td><td>Powder (1</td><td>Drama|Sci-Fi</td></tr>
<tr><td>25</td><td>Leaving La</td><td>Drama|Romance</td></tr>
<tr><td>26</td><td>Othello (1</td><td>Drama</td></tr>
<tr><td>27</td><td>Now and T</td><td>Children|Drama</td></tr>
<tr><td>28</td><td>Persuasior</td><td>Drama|Romance</td></tr>
</table>

Ratings.csv

userId	movieId	rating	timestamp
1	1	4	9.65E+08
1	3	4	9.65E+08
1	6	4	9.65E+08
1	47	5	9.65E+08
1	50	5	9.65E+08
1	70	3	9.65E+08
1	101	5	9.65E+08
1	110	4	9.65E+08
1	151	5	9.65E+08
1	157	5	9.65E+08
1	163	5	9.65E+08
1	216	5	9.65E+08
1	223	3	9.65E+08
1	231	5	9.65E+08
1	235	4	9.65E+08
1	260	5	9.65E+08
1	296	3	9.65E+08
1	316	3	9.65E+08
1	333	5	9.65E+08
1	349	4	9.65E+08
1	356	4	9.65E+08
1	362	5	9.65E+08
1	367	4	9.65E+08
1	423	3	9.65E+08

Tags.csv

	A	B	C	D	E
1	userId	movieId	tag	timestamp	
2	2	60756	funny	1.45E+09	
3	2	60756	Highly quo	1.45E+09	
4	2	60756	will ferrell	1.45E+09	
5	2	89774	Boxing sto	1.45E+09	
6	2	89774	MMA	1.45E+09	
7	2	89774	Tom Hardy	1.45E+09	
8	2	106782	drugs	1.45E+09	
9	2	106782	Leonardo	1.45E+09	
10	2	106782	Martin Scc	1.45E+09	
11	7	48516	way too lo	1.17E+09	
12	18	431	Al Pacino	1.46E+09	
13	18	431	gangster	1.46E+09	
14	18	431	mafia	1.46E+09	
15	18	1221	Al Pacino	1.46E+09	
16	18	1221	Mafia	1.46E+09	
17	18	5995	holocaust	1.46E+09	
18	18	5995	true story	1.46E+09	
19	18	44665	twist endir	1.46E+09	
20	18	52604	Anthony H	1.46E+09	
21	18	52604	courtroom	1.46E+09	
22	18	52604	twist endir	1.46E+09	
23	18	88094	britpop	1.46E+09	
24	18	88094	indie recor	1.46E+09	
25	18	88094	music	1.46E+09	
26	18	144210	dumpster	1.46E+09	
27	18	144210	Sustainabil	1.46E+09	
28	21	1569	romantic c	1.42E+09	
29	21	1569	wedding	1.42E+09	

All these datasets are merged through and to be used for our recommendation system.

TOOLS USED :

Python : Python is an interpreted, high-level, general-purpose programming language used for performing the statistical analysis. When applying the technique of Web Scraping, Python coding will scrap the internet for selected data.

Open CV : OpenCV is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then It sees. The library is cross-platform and free for use under the open-source BSD license.

Pandas : Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Numpy : NumPy is a python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Seaborn : Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics

Methodology:

There are 5 major steps involved in the building a ML model for Movie Recommendation System.This encapsulates the following steps:

- Data loading
- Data cleaning
- Data Analysis
- Recommendation model

Pearson's correlation is used to recommend the movies based on the users choice of movies .

Pearson's correlation : Peason's Correlation, sometimes just called correlation, is the most used metric for this purpose, it searches the data for a linear relationship between two variables.

Analyzing the correlations is one of the first steps to take in any statistics, data analysis, or machine learning process, it allows data scientists to early detect patterns and possible outcomes of the machine learning algorithms, so it guides us to choose better models.

EVALUATION METRIC :

Personalization :

Personalization is a great way to assess if a model recommends many of the same items to different users. It is the dissimilarity (1- cosine similarity) between user's lists of recommendations.

Intra-list Similarity:

Intra-list similarity is the average cosine similarity of all items in a list of recommendations. This calculation uses features of the recommended items (such as movie genre) to calculate the similarity.

Coverage:

Coverage is the percent of items in the training data the model is able to recommend on a test set. In this example, the popularity recommender has only 0.05% coverage, since it only ever recommends 10 items. The random recommender has nearly 100% coverage as expected. Surprisingly, the collaborative filter is only able to recommend 8.42% of the items it was trained on.