Case Study: Netflix Recommendation System

Business Understanding:

Goal:

The main goal is to improve user engagement by recommending movies or TV shows that are personalized for each user. By understanding users' past viewing behavior and ratings, we can suggest content they are more likely to enjoy.

Impact:

- **User Satisfaction:** Users get content they like, leading to a better experience.
- **Retention:** Happy users are more likely to stay subscribed and continue using the platform.
- **Revenue:** Improved engagement can indirectly lead to higher revenue through subscriptions and loyalty.

Data Understanding:

Data Understanding is the **second phase of the Data Science process** (after Business Understanding). It involves **examining and exploring the data** collected to understand its structure, quality, patterns, and relevance to the problem you want to solve.

The goal is to **gain insights into what the data contains, identify potential issues, and prepare for further analysis or modeling**.

Key points:

1. **Purpose:**
    - To know what kind of data is available.
    - To check if the data is sufficient to solve the business problem.
    - To detect any errors, missing values, or inconsistencies.
2. **Typical Questions in Data Understanding:**
    - How many users, movies, or items are in the dataset?
    - What are the characteristics of the data (numerical, categorical, text, timestamps)?
    - Are there missing or duplicate values?
    - What is the distribution of key variables (e.g., ratings, genres)?
    - Which features are most important or frequent?
3. **Activities involved:**
    - **Data Collection Review:** Ensuring all required data is available.
    - **Data Profiling:** Summarizing the data (counts, unique values, min/max, mean, standard deviation).
    - **Exploratory Analysis:** Visualizing distributions, trends, and relationships.
    - **Data Quality Assessment:** Checking for missing, inconsistent, or duplicate entries.
4. **Outcome:**
    - A clear understanding of the data structure and quality.
    - Insights into patterns or anomalies that may influence modeling.
    - Foundation for **data preparation, feature engineering, and model building**.

```python
import pandas as pd
import zipfile

# Load zip file
zip_path = r'D:\ml-latest-small.zip'
with zipfile.ZipFile(zip_path) as z:
    movies = pd.read_csv(z.open('ml-latest-small/movies.csv'))
    ratings = pd.read_csv(z.open('ml-latest-small/ratings.csv'))

# View first few rows
print(movies.head())
print(ratings.head())

# Explore dataset
print("Number of users:", ratings['userId'].nunique())
print("Number of movies:", movies['movieId'].nunique())
print("Number of ratings:", ratings.shape[0])

# Rating distribution
print(ratings['rating'].value_counts())
```

Output :

```
   movieId                               title  \
0        1                    Toy Story (1995)
1        2                      Jumanji (1995)
2        3             Grumpier Old Men (1995)
3        4            Waiting to Exhale (1995)
4        5  Father of the Bride Part II (1995)

                                        genres
0  Adventure|Animation|Children|Comedy|Fantasy
1                   Adventure|Children|Fantasy
2                              Comedy|Romance
3                        Comedy|Drama|Romance
4                                      Comedy
   userId  movieId  rating  timestamp
0       1        1     4.0  964982703
1       1        3     4.0  964981247
2       1        6     4.0  964982224
3       1       47     5.0  964983815
4       1       50     5.0  964982931
Number of users: 610
Number of movies: 9742
Number of ratings: 100836
rating
4.0    26818
3.0    20047
5.0    13211
3.5    13136
4.5     8551
2.0     7551
2.5     5550
1.0     2811
1.5     1791
0.5     1370
Name: count, dtype: int64
```

Data Preparation:

Data Preparation is the **third phase of the Data Science process** (after Business Understanding and Data Understanding). It involves **cleaning, transforming, and structuring raw data** so that it is ready for analysis, visualization, or modeling. This step is critical because **poorly prepared data can lead to incorrect insights or inaccurate models**.

Purpose of Data Preparation:

- Ensure the data is **accurate, complete, and consistent**.

- Convert data into a format suitable for **analysis and machine learning algorithms**.
- Enhance the quality and usability of the dataset for better decision-making.

Key Activities in Data Preparation

1. **Handling Missing Data**
    - Missing values can affect analysis and predictions.
    - Techniques:
        - **Removal:** Delete rows or columns with missing values if they are few.
        - **Imputation:** Fill missing values with meaningful substitutes (mean, median, mode, or a placeholder).
2. **Transforming Data**
    - Convert raw data into meaningful formats.
    - Example: Convert timestamps into human-readable dates to analyze trends over time.
3. **Encoding Categorical Variables**
    - Many datasets have categorical data (e.g., genres of movies).
    - Techniques like **one-hot encoding** transform categories into numerical columns so algorithms can process them.
4. **Data Normalization and Scaling** (optional in some cases)
    - Adjust numerical data to a common scale without distorting differences.

- Helps improve the performance of certain machine learning algorithms.

5. **Feature Engineering**
    - Creating new variables or features that improve predictive power.
    - Example: Extracting year or month from a timestamp for trend analysis.

Outcome of Data Preparation

After this step, the dataset is:

- **Clean:** Errors and missing values handled.
- **Structured:** Properly formatted for analysis or modeling.
- **Ready for Modeling:** Suitable for machine learning algorithms, visualization, and insights extraction.

```
# Check missing values
print(movies.isnull().sum())
print(ratings.isnull().sum())

# Convert timestamp to datetime
ratings['timestamp'] = pd.to_datetime(ratings['timestamp'], unit='s')

# One-hot encode genres
movies['genres'] = movies['genres'].str.split('|')
genres_df = movies['genres'].explode('genres')
print(genres_df.head())
```

Output:

```
movieId    0
title      0
genres     0
dtype: int64
userId       0
movieId      0
rating       0
timestamp    0
dtype: int64
0    Adventure
1    Animation
2     Children
3       Comedy
4      Fantasy
Name: genres, dtype: object
```

Modeling:

Modeling is the **fourth phase of the Data Science process** (after Data Preparation). It involves **selecting and applying mathematical or computational algorithms** to the prepared data to **identify patterns, make predictions, or provide recommendations**.

The goal of modeling is to **build models that can generalize well on unseen data** and help solve the business problem effectively.

Key Activities in Modeling

1. **Choosing the Right Algorithm**
    - Depends on the problem type:
        - **Classification:** Predict categories (e.g., like/dislike a movie).
        - **Regression:** Predict continuous values (e.g., movie ratings).
        - **Clustering:** Group similar items or users (e.g., segment users by taste).

        - **Recommendation Algorithms:** Collaborative filtering, content-based filtering, hybrid methods.

2. **Splitting Data**
    - Divide data into **training** and **testing sets** (sometimes also a validation set).
    - Ensures the model can **generalize** to new data and prevents overfitting.
3. **Training the Model**
    - Feed the prepared data into the chosen algorithm.
    - The model **learns patterns, relationships, or preferences** from the training data.
4. **Tuning Hyperparameters**
    - Adjust algorithm settings to improve model performance (e.g., number of neighbors in KNN, learning rate in gradient boosting).
5. **Making Predictions**
    - Use the trained model to predict outcomes for the test data or new unseen data.
6. **Evaluating Model Performance**
    - Measure accuracy and effectiveness using metrics such as:
        - **RMSE, MAE** for regression (rating prediction).
        - **Precision, Recall, F1-score** for classification.
        - **Hit rate, precision@k** for recommendations.

Outcome of Modeling

- A trained and validated model that can **predict, classify, or recommend effectively**.
- Provides insights into patterns, relationships, or trends in the data.
- Forms the basis for **decision-making or system deployment** (e.g., a recommendation engine on Netflix).

```python
# Create user-item matrix
user_movie_matrix = ratings.pivot(index='userId', columns='movieId', values='rating')

# Fill missing values with 0 (simple approach)
user_movie_matrix = user_movie_matrix.fillna(0)

# Find similarity between users using correlation
user_similarity = user_movie_matrix.T.corr()

# Recommend movies for a user (user 1)
user_id = 1
similar_users = user_similarity[user_id].sort_values(ascending=False)[1:6]  # Top 5 similar users
print("Most similar users to user 1:\n", similar_users)
```

Output:

```
Most similar users to user 1:
 userId
266    0.344983
313    0.333875
368    0.324041
57     0.323948
39     0.320120
Name: 1, dtype: float64
```

Content based filtering

```python
# Example: Recommend based on genres
# Pick user 1's liked movies (rating >= 4)
liked_movies = ratings[(ratings['userId']==1) & (ratings['rating']>=4)]
liked_genres = movies[movies['movieId'].isin(liked_movies['movieId'])]['genres'].explode()
print("User 1 likes genres:\n", liked_genres.value_counts())
```

Output:

```
User 1 likes genres:
 genres
Action       76
Adventure    74
Comedy       70
Drama        64
Thriller     43
Fantasy      41
Crime        39
Children     37
Sci-Fi       32
Animation    27
Romance      24
Musical      20
War          20
Mystery      13
Horror        9
Western       6
Film-Noir     1
Name: count, dtype: int64
```

Evaluation:

Evaluation is the **fifth phase of the Data Science process** (after Modeling). It involves **assessing the performance and effectiveness of the trained model** to determine if it meets the business objectives.

The goal of evaluation is to **measure accuracy, reliability, and usefulness** of the model before deploying it in real-world applications.

Key Activities in Evaluation

1. **Selecting Evaluation Metrics**
    - Metrics depend on the problem type:
        - **Regression:** Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)
        - **Classification:** Accuracy, Precision, Recall, F1-Score
        - **Recommendation Systems:** Precision@K, Recall@K, Mean Average Precision (MAP), Hit Rate

2. **Testing the Model**
    - Evaluate the model on **test data** that was not used during training.
    - Ensures the model **generalizes well** to new, unseen data.
3. **Analyzing Results**
    - Compare predicted vs. actual outcomes.
    - Identify strengths, weaknesses, and possible biases in the model.
4. **Model Comparison**
    - Often multiple models are trained. Evaluation helps **select the best-performing model**.
5. **Iterative Improvement**
    - Based on evaluation, you may **tune hyperparameters, add features, or try different algorithms** to improve performance.

Outcome of Evaluation

- Determines if the model is **accurate, reliable, and ready for deployment**.
- Provides **insights into model strengths and weaknesses**.
- Ensures alignment with **business goals and objectives** (e.g., recommending movies users are likely to enjoy).

```python
# Split data (80% train, 20% test)
from sklearn.model_selection import train_test_split
train, test = train_test_split(ratings, test_size=0.2, random_state=42)

# You can compute RMSE for simple predictions
import numpy as np
# Predict mean rating for simplicity
mean_rating = train['rating'].mean()
rmse = np.sqrt(((test['rating'] - mean_rating) ** 2).mean())
print("RMSE (baseline):", rmse)
```

Output:

```
RMSE (baseline): 1.0488405992661316
```

Deployement :

deployment is the **final phase of the Data Science process** (after Evaluation). It involves **integrating the trained and validated model into a real-world environment** so that it can generate predictions or recommendations for actual users.

The goal is to make the model's output **accessible and usable** by business systems, applications, or end-users to support decision-making.

Key Activities in Deployment

1. **Model Integration**
    - Deploy the model into a production system (e.g., a web app, mobile app, or backend service).
    - Example: Netflix integrates its recommendation model into its platform so users instantly see personalized movie suggestions.
2. **Scalability and Performance Optimization**
    - Ensure the model can handle **large-scale data and high traffic**.
    - Use cloud platforms or distributed systems for real-time recommendations.
3. **Monitoring and Maintenance**
    - Continuously monitor performance (e.g., accuracy, response time).

    - Detect issues like **model drift** (when data patterns change over time).

4. **Feedback Loop**
    - Collect feedback from users (e.g., ratings, likes/dislikes).
    - Use this feedback to **retrain and improve the model**.
5. **Automation and Scheduling**
    - Automate regular model retraining with new data.
    - Schedule periodic updates for continuous improvement.

Outcome of Deployment

- The model is **live and functional** in the real-world environment.
- End-users or business systems can **interact with predictions or recommendations in real-time**.
- Establishes a cycle of **monitoring, feedback, and retraining** for long-term effectiveness.