Evaluation on historical datasets

Evaluating recommender systems on historical datasets is a common approach to assess their performance using past user interactions and preferences. Here's a step-by-step guide on how you might conduct such an evaluation:

1. **Data Preparation:**
   - **Dataset Selection:** Choose a historical dataset that includes user-item interactions, such as ratings, clicks, purchases, or other relevant interactions.
   - **Data Cleaning:** Preprocess the data to handle missing values, remove duplicates, and address any other issues.
2. **Train-Test Split:**
   - **Temporal Split:** If the dataset has a temporal component, split it chronologically into a training set (earlier interactions) and a test set (later interactions).
   - **Random Split:** Alternatively, you can randomly partition the data into training and test sets.
3. **Model Training:**
   - **Algorithm Selection:** Choose one or more recommender algorithms suitable for the task, such as collaborative filtering, content-based filtering, matrix factorization, or hybrid methods.
   - **Parameter Tuning:** Fine-tune the parameters of the selected algorithms based on the training data.
4. **Prediction Generation:**
   - **Generate Predictions:** Use the trained model(s) to generate predictions for the test set, predicting user preferences for items not seen during training.
5. **Evaluation Metrics:**
   - **Select Metrics:** Choose appropriate evaluation metrics based on the nature of the recommendation task. Common metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Precision, Recall, and ranking-based metrics like Normalized Discounted Cumulative Gain (NDCG).
   - **Compute Metrics:** Calculate the selected metrics using the predicted ratings or rankings and the actual user feedback from the test set.
6. **Analysis of Results:**
   - **Quantitative Analysis:** Examine the values of the evaluation metrics to assess the accuracy, ranking quality, and other performance aspects of the recommender system.
   - **Comparison:** If using multiple algorithms, compare their performance to identify the most effective one for your specific goals.
7. **Coverage and Diversity:**

- **Item Coverage:** Evaluate the coverage of recommended items to ensure that the recommender system suggests a diverse set of items.
- **Diversity Metrics:** Assess the diversity of recommendations using metrics like intra-list diversity or inter-list diversity.

8. **Novelty and Serendipity:**
    - **Novelty Analysis:** Assess the novelty of recommended items, measuring how often the system suggests items that users haven't interacted with before.
    - **Serendipity Assessment:** Consider the system's ability to surprise users with unexpected but relevant recommendations.
9. **User Feedback:**
    - **Surveys or Interviews:** Collect user feedback through surveys or interviews to gain qualitative insights into user satisfaction, preferences, and the perceived quality of recommendations.
10. **Ethical Considerations:**
    - **Privacy and Fairness:** Consider ethical aspects such as user privacy and fairness in recommendations. Ensure that the evaluation process aligns with ethical guidelines.
11. **Iterative Improvement:**
    - **Feedback Loop:** Use the evaluation results to iteratively improve the recommender system. This may involve adjusting algorithm parameters, incorporating user feedback, or updating the system based on changing user preferences.

Remember that the evaluation of recommender systems is an ongoing process, and the choice of metrics and methods should align with the specific goals and characteristics of the recommendation task. Additionally, consider the limitations of historical datasets, such as the potential for data sparsity, the cold start problem, and changes in user preferences over time.