Sample task for AI/ML Engineer

Personalized Feed Generation System

Data Collection and Preprocessing

For this project, I generated synthetic user interaction data using the Faker library to simulate a realistic dataset. The dataset comprises the following fields:

- **User_ID**: Unique identifier for each user.
- **Post_ID**: Unique identifier for each post.
- **Interaction_Type**: Type of interaction (like, comment, share, seen).
- **Interaction_Timestamp**: Timestamp of the interaction.
- **Post_Content**: Text content of the post.
- **Post_Timestamp**: Timestamp when the post was created.

Preprocessing Steps:

1. **Data Generation**: Created 500 users and 2000 posts per user, resulting in 1,000,000 records.
2. **Timestamp Conversion**: Converted timestamp strings to datetime objects for proper time-based calculations.
3. **Data Sorting**: Sorted the dataset by Interaction_Timestamp to organize interactions chronologically.

Feature Engineering

To enhance the predictive capability of the model, I engineered several features:

1. **Interaction Features**: Counted each type of interaction (like, comment, share, seen) for each (User_ID, Post_ID) pair.
2. **Content Features**: Applied TF-IDF vectorization on the Post_Content to extract meaningful text features.
3. **Temporal Features**: Calculated the age of each post in hours (Post_Age) by finding the difference between Interaction_Timestamp and Post_Timestamp.

Model Selection and Training

I selected a Logistic Regression model for its simplicity, interpretability, and efficiency in binary classification tasks. The process involved the following steps:

1. **Feature Preparation**: Selected relevant features including interaction counts, Post_Age, and TF-IDF features.
2. **Data Splitting**: Split the data into training and test sets (80/20 split).
3. **Feature Scaling**: Standardized the features using StandardScaler.
4. **Model Training**: Trained the Logistic Regression model on the scaled training data.

5. **Evaluation**: Assessed the model using accuracy, precision, recall, and cross-validation scores.

Model Performance:

- Accuracy: 87.82%
- Precision: 95.12%
- Recall: 88.30%
- Cross-validation scores: [0.87878125 0.87765 0.87829375 0.87786875 0.87679375]
- Mean CV score: 87.79%
- These metrics demonstrate that the model performs well in predicting user interactions.

Sample Output

I developed a function to generate personalized feeds for users by calculating a weighted sum of relevance and recency scores for each post. The relevance score is predicted by the Logistic Regression model, while the recency score is based on the age of the post.

Example Personalized Feed for a Sample User:

1. **Post ID**: 6b5d5bd8-d3a3-4a7d-9cc7-a85981404001
    - **Content**: These describe admit economic enter agency. Measure worker senior woman mouth ago page.Tough Americ...
    - **Final Score**: 0.7006
2. **Post ID**: a0dbdf9d-8b80-4f46-ad54-ef74c89c1f0e
    - **Content**: Official provide action generation five. News challenge draw evening. Rest into old international th...
    - **Final Score**: 0.7005
3. **Post ID**: 6f7d5c14-d2d5-4d22-bc00-f8be3e5b284c
    - **Content**: Speak but century seat statement best. Provide owner stand right. Will out appear should.Event me b...
    - **Final Score**: 0.6999

These posts were ranked highly due to their high relevance and recent timestamps, ensuring that the user sees content that is both pertinent and timely.