Model Card for Event Recommendation Engine

Model Description

Input:

The input to this model consists of user and event-related data. The main data files used include:

1. **users.csv:** Contains demographic information of users (e.g., user ID, birth year, gender, location, timezone).
2. **events.csv:** Contains event-specific data (e.g., event ID, event start time, event location).
3. **user_friends.csv:** Contains user connections to other users, allowing a social network to be considered.
4. **event_attendees.csv:** Contains data on event participation, showing how many users were interested or not interested in attending the event.
5. **train.csv:** Used to train the model, containing user-event interactions (whether a user expressed interest in an event or not).
6. **test.csv:** Contains user-event pairs for which the model will predict the user's interest.

Output:

The model outputs a **probability score** that indicates the likelihood of a user showing interest in a given event. Events are then ranked by this score, and a **top-200 recommendation list** is generated for each user based on these scores.

Model Architecture:

The model utilizes **XGBoost (Extreme Gradient Boosting)**, a powerful and widely used machine learning algorithm for classification tasks. XGBoost is known for its efficiency in handling large datasets and its ability to manage feature interactions automatically.

Key features of the architecture include:

- **Hyperparameter tuning** using Optuna, an optimization library, to achieve the best model performance.
- Features such as user demographics, event characteristics, and past event attendance behavior are combined for model training.

The data is preprocessed to handle missing values and categorical variables (e.g., gender, location). Features are encoded to make them suitable for XGBoost's input requirements, and missing values in important columns like birth year are imputed.

Performance

The model was evaluated using **Mean Average Precision at 200 (MAP@200)**, which measures how well the model ranks the top 200 events for a user.

Performance Summary:

- **MAP@200 score:** 0.6014

This score indicates that the model is able to rank events quite effectively, ensuring that relevant events are recommended to users.

Key Metrics:

- **Training Dataset:** The model was trained on train.csv containing thousands of user-event interactions.
- **Test Dataset:** The model was evaluated on test.csv, which contains unseen user-event pairs.

Limitations

The model has several limitations:

1. **Cold Start Problem:** The model struggles to make accurate predictions for new users or events with little historical data, as it relies heavily on past behavior.
2. **Limited Contextual Understanding:** The model does not account for contextual information such as changes in user preferences or event timing (e.g., a user may be interested in different events depending on their current circumstances).
3. **Feature Engineering Constraints:** Some categorical features, like location, are treated with one-hot encoding, which may not capture the full relationship between different locations.

Trade-offs

1. Performance vs. Interpretability:

The XGBoost algorithm, while powerful, is less interpretable compared to simpler models like linear regression. Although XGBoost provides high performance, understanding why a particular prediction was made is more difficult.

2. Computation Time vs. Accuracy:

Hyperparameter tuning with Optuna is time-consuming but provides better performance. This trade-off was necessary to optimize the model, but the computational cost could be reduced by limiting the number of trials or using less exhaustive search methods.

3. Handling Missing Data:

Imputation of missing values (e.g., birth year) helps maintain the model's performance but may introduce bias, especially if the missing data is not random.

4. Generalization vs. Specificity:

The model generalizes well across users with extensive event history but might underperform for users with fewer interactions, where more personalized approaches could be more effective.