Database Query Assistant • Chat Transcript

Exported: 11/25/2025, 5:21:10 PM

👤 **You**

what are your capabilities?

🤖 **Assistant**

As a helpful database assistant, I can:

1. **Answer questions about the database schema**: Provide details on tables, columns, data types, primary keys, foreign keys, and indices.
2. **Explain relationships between tables**: Describe how tables like `ads` and `feeds` are related (e.g., via `user_id`).
3. **Generate SQL queries**: Create MySQL queries to retrieve, filter, aggregate, or join data based on your specific questions. This includes complex queries for data analysis, recommendations, and statistical summaries.
4. **Provide information from documentation**: Access and utilize provided context to answer questions about data ranges, value meanings, and other relevant details.
5. **Explain synthetic data generation processes**: Describe how synthetic data is generated and the characteristics of the synthetic tables (`synthetic_train_smote`, `synthetic_train_adasyn`).
6. **Generate synthetic data**: If explicitly asked to "generate" or "create" synthetic data, I can simulate this process by specifying the table and number of rows.

👤 **You**

can you tell me how the synthetic smote and adasyn datasets were created

🤖 **Assistant**

The `synthetic_train_smote` and `synthetic_train_adasyn` datasets were created using advanced oversampling techniques to address class imbalance in the original training data, specifically for predicting ad click behavior (binary classification).

Here's a breakdown of the generation process and techniques:

1. **Data Preprocessing**:

- The original `ads` and `feeds` training data were preprocessed.
- Categorical variables were label-encoded (converted to numeric codes).

- Categorical variables were label-encoded (converted to numeric codes).
- List/array columns (like JSON fields) were processed to become numeric features.
- The resulting datasets for both SMOTE and ADASYN are entirely numeric with no missing values.

2. **Oversampling Techniques Applied**:
   - **SMOTE (Synthetic Minority Oversampling Technique)**:
     - This technique generates synthetic samples for the minority class (positive ad clicks, `label=1`).
     - It works by interpolating between existing minority class examples.
     - It uses k-nearest neighbors (with `k=5`) to identify similar examples and create new data points that are combinations of these neighbors.
     - The goal of SMOTE was to produce a perfectly balanced dataset, resulting in `synthetic_train_smote` having a 50/50 distribution for `label` 0 and 1.
   - **ADASYN (Adaptive Synthetic Sampling)**:
     - ADASYN is an adaptive version of SMOTE.
     - It focuses on generating more synthetic examples for "harder-to-learn" minority class samples, meaning it gives more weight to minority examples that are surrounded by majority examples.
     - It automatically determines the number of synthetic samples to generate for each minority example based on its learning difficulty, also using `n_neighbors=5`.
     - ADASYN aims to produce a near-balanced dataset, and `synthetic_train_adasyn` has an approximately 50/50 class distribution (e.g., 49.98% class 0, 50.02% class 1).

Both datasets are designed to maintain the statistical properties and relationships of the original data while providing a balanced class distribution suitable for training machine learning models.

👤 **You**

can you create a side by side bar chart of the proportion of no clicks and clicks between the ads dataset and the smote dataset

🤖 **Assistant**

NO_SQL_NEEDED:A side-by-side bar chart comparing the proportion of 'no clicks' (label 0) and 'clicks' (label 1) between the `ads` dataset and the `synthetic_train_smote` dataset would show the following:

- **For the `ads` dataset**:
- The bar for 'no clicks' (label 0) would be very tall, representing approximately **98.45%** of the data.
- The bar for 'clicks' (label 1) would be very short, representing approximately **1.55%** of the

data. This indicates a significant class imbalance in the original `ads` dataset.

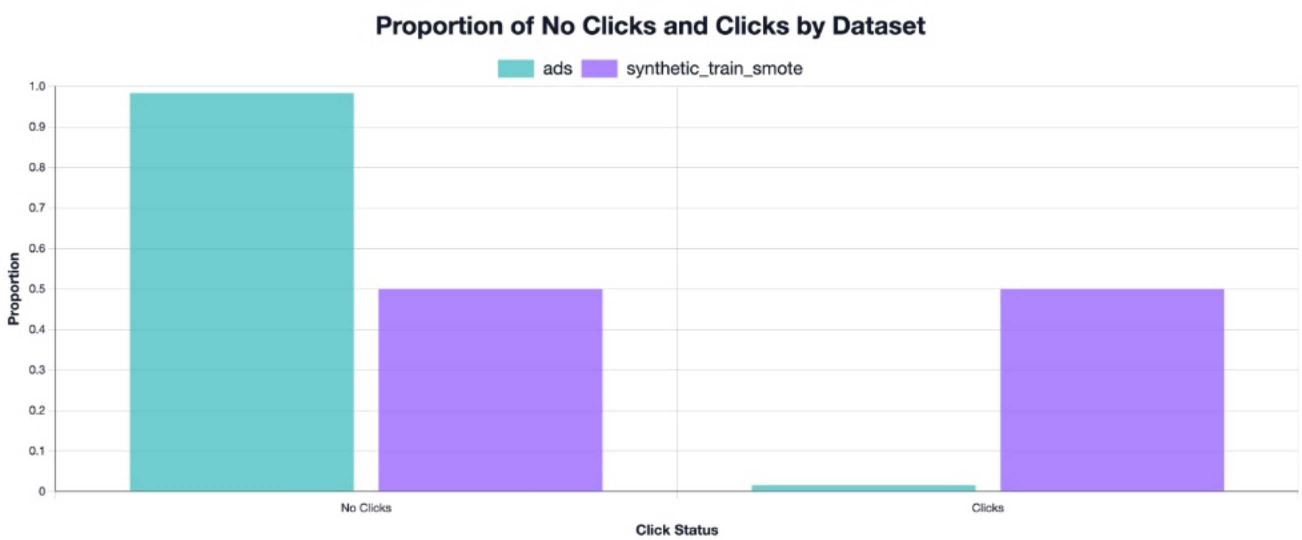
- **For the `synthetic_train_smote` dataset**:
- The bar for 'no clicks' (label 0) would be of equal height to the bar for 'clicks' (label 1), both representing exactly **50.00%** of the data. This demonstrates that the SMOTE technique successfully balanced the dataset, creating an equal proportion of both classes.

In summary, the chart would visually highlight how the `synthetic_train_smote` dataset addresses the severe class imbalance present in the original `ads` dataset by artificially increasing the proportion of the minority 'click' class.

SQL Queries Executed:

```
SELECT 'ads' AS dataset, label, COUNT(*) * 1.0 / (SELECT COUNT(*) FROM ads)
AS proportion FROM ads GROUP BY label UNION ALL SELECT
'synthetic_train_smote' AS dataset, label, COUNT(*) * 1.0 / (SELECT COUNT(*)
FROM synthetic_train_smote) AS proportion FROM synthetic_train_smote GROUP
BY label
```

4 row(s) returned



---

dataset	label	proportion
ads	0	0.98448
ads	1	0.01552
synthetic_train_smote	0	0.50000
synthetic_train_smote	1	0.50000