Data Cleaning and Augmentation Documentation: ConnectX Communications

Background

ConnectX Communications is a national telecom provider offering internet, mobile, and TV services. They have **multiple call centers across the U.S.**, and thousands of customers call daily for support. Lately, leadership is concerned about **inconsistent customer satisfaction (CSAT)** scores and growing **negative sentiment** in customer feedback.

Problem Statement

ConnectX Communications is experiencing inconsistent CSAT scores and negative sentiment. The business needs to identify root causes and improve customer experience.

Data Cleaning Process (Excel)

The original dataset contained one-month worth of dataset (**32,941 records)**. The cleaning process was performed manually in **Microsoft Excel** and involved the following steps:

1. CSAT Score Filtering

- **Problem:** 20,670 records (over 62%) had missing csat_score values.
- **Action:** Filtered out all rows where csat_score was blank.
- **Result:** Final cleaned dataset used for analysis had **12,271 complete records** with valid CSAT scores.

2. Removing Duplicates

- **Checked** for duplicate rows using Excel's **Remove Duplicates** function.
- No exact duplicates were found, so no rows were removed.

3. Date Format Consistency

- Ensured all date fields (call_date, response_time, call_duration) were consistently formatted as **Date** types.

4. Sentiment Label Cleanup

- Ensured values in sentiment column were limited to valid categories:
  - Very Positive, Positive, Neutral, Negative, Very Negative.

Column Augmentations

- **Call Date Breakdown:**
  - Created a new column to extract **day of the week** using =TEXT(date, "dddd").
  - Created another column for **dates only** using =DAY(date).
- **Call Duration Binning:**
  - Grouped raw call_duration values into:
    - Short = 5 -10 mins
    - Mid = 11 - 20 mins
    - Long = 12 - 30 mins
    - Very Long = 31 - above
  - Used nested IF formula in Excel to categorize each call.
    - =IF(M2<=10, "Short", IF(M2<=20, "Mid", IF(M2<=30, "Long", "Very Long")))

Final Output

A clean, enriched dataset of **12,271 records**, ready for in-depth analysis. Dataset includes:

- Valid CSAT scores
- Engineered features for time analysis (day of week, day)
- Grouped call durations
- Standardized categories for sentiment, reason, channel, and state

---

Data cleaning by: Vanesa Gate - Data Analyst

Email: nesagate@gmail.com