# 1. Understanding the Data

**Data Observations:**

* **Features**: Var1 to Var230 — 230 variables (very wide format).
* **Target**: Label — Appears to be either numerical (possibly regression) or categorical (if encoded).
* **Mixed data types**: Some columns are numeric, others appear to be categorical
* **Presence of missing values**: Some fields are empty

# 2. Solution Outline

**Step 1: Data Preprocessing & Cleaning**

**Tasks:**

* **Missing Values:** Identify missing/null values and decide:
  + Drop columns with high % missingness (e.g., > 80%)
  + **Impute others:**
    - **Numerical:** Mean
    - **Categorical:** Mode or ‘Others category
* **Data Types**: Convert to proper types
  + **Alphanumeric variables →** categorical
  + **Numeric variables →** int
* **Constant columns:** Remove columns with a single unique value**.**

**Step 2: Exploratory Data Analysis (EDA)**

**Goals:**

* **Univariate Analysis:**
  + **Histograms of numeric columns**
  + **Value counts for categorical columns**
* **Correlation Matrix (numeric features):**
  + **Identify highly correlated variables (drop one of each pair if correlation > 0.9)**
* **Target Distribution:**
  + **Label is:** Binary

**Step 3: Feature Engineering**

**Approaches:**

* **Categorical Encoding:**
  + **Low-cardinality:** One-Hot
  + **High-cardinality:** Label Encoding
* **Outlier Handling**
  + Apply Winsorization
* **Feature Selection:**
  + Fit a RF model to get the feature importance scores and select the features that covers 80% of the importance

**Step 4: Modeling**

Choose model based on target type:

**Classification:**

* **Models**: Random Forest, XGBoost, LightGBM
* Split into train and test set using 70%:30% splitting
* **Evaluation Metrics**:
  + F1-score (imbalanced classification)
  + AUC-ROC
  + Confusion Matrix

**Step 5: Hyperparameter Tuning**

Can Use:

* **RandomizedSearchCV**
* **Optuna**
* **GridSearchCV** (if feasible)

**Step-6**: **Apply Oversampling Technique**

* Since the class distribution is highly imbalanced and required F1-score is not achieved, can apply SMOTE
* **SMOTE** generates synthetic samples for the minority class to balance the class distribution without simply duplicating existing data
* Applied on Training data