ACME Case Study: Predicting Customer Response

Background

ACME, a company selling sports products, wants to promote its new product: **the XL Original Orange Baseball Cap**.
To test customer interest, ACME sent a **test mailing to 10,000 randomly selected customers** and recorded their responses. Two datasets are provided:

- `train_acme_customers.sav` – training dataset containing customers who received the test mailing and their responses.
- `test_acme_customers.sav` – testing dataset containing customers who did not receive the test mailing (response field undefined).

Both datasets include the following fields:

- **customer_id** – customer's identification number
- **gender** – customer's gender
- **email_address** – customer's e-mail address
- **postal_code** – customer's postal code
- **recency_01_01_2011** – last order date before Jan 1, 2011
- **frequency_01_01_2011** – number of orders before Jan 1, 2011
- **monetary_value_01_01_2011** – total purchase amount before Jan 1, 2011
- **has_received_test_mailing** – flag whether the customer received the Feb 1, 2011 test mailing
- **response** – whether the customer ordered the XL Original Orange Baseball Cap (only valid for training dataset customers)
- **orderdate** – date the cap was ordered (only for respondents)
- **days_between_test_and_order** – days between test mailing and orderdate (only for respondents)
- **ordered_within_month** – whether the order happened within one month after mailing (only for respondents)

Tasks

1. Data Overview

- Import the **training dataset** into IBM SPSS Modeler.
- Run a **Table node** to summarize the data.
- How many records are in the training dataset?
- How many fields are in the training dataset?

2. Test Mailing Customers

- Select only customers who were in the **test mailing**.
- How many customers were included in the test mailing?

3. Predictive Modeling

- Build a **CHAID decision tree model** to predict `response`, using:
- `gender`
- `recency_01_01_2011`
- `frequency_01_01_2011`
- `monetary_value_01_01_2011`
- Answer the following:
- Which field is used as the **first split**?
- Which group shows the **highest response rate**? What is the probability of responding for this group?

4. Model Output

- Run a **Table node downstream of the model nugget**.
- Identify the two new fields added by the model.
- What do these fields represent?

5. Applying the Model

- Apply the model to the **testing dataset** (customers who did not receive the test mailing).
- How many customers are predicted to respond positively (`predicted = T`)?

6. Exporting Results

- Export the selected customers to a text file named `customers_to_contact.txt`.
- Include only the following fields:
- `customer_id`
- predicted category (rename to `predicted_category`)
- confidence score (rename to `confidence_score`)