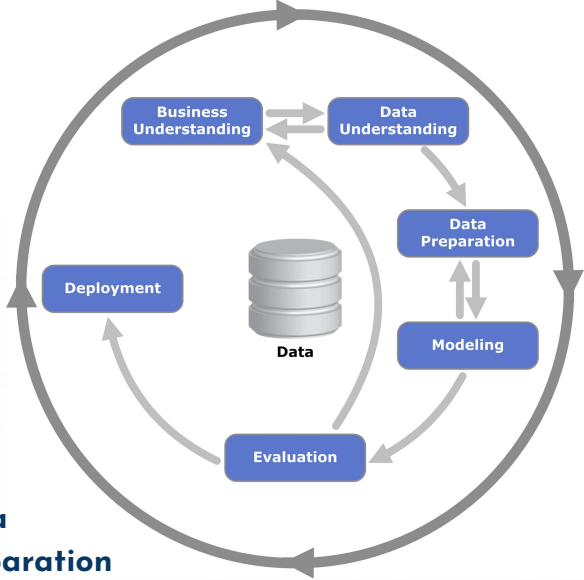
# Customer Characteristic Analysis

***Team Data Pundit***

**David Ishimwe Ruberamitwe, Richard Jensen, Yifei Wang, Laurie Ye, Yuki Ying**

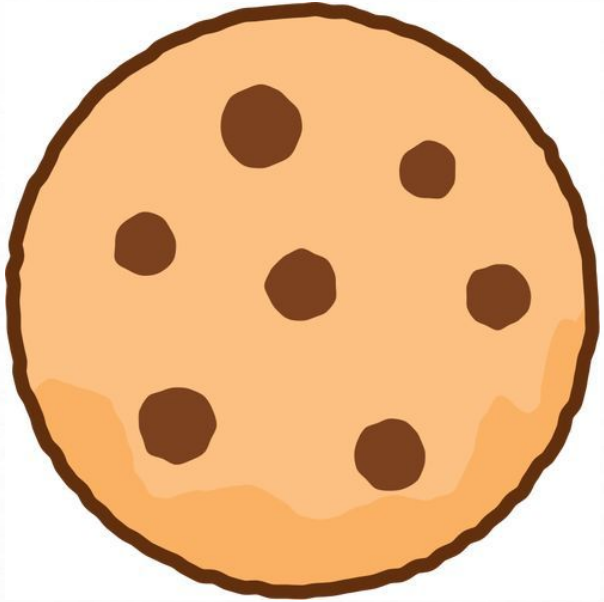
# Table of Contents



# 01

# Business Understanding

# Our Company...



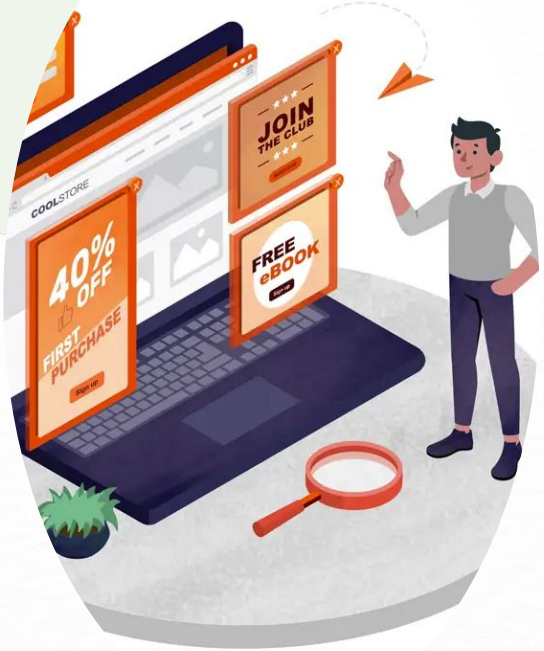
# Targeted marketing is important...

**80%** of internet users are more likely to click on a targeted ad

**76%** of customers feel that targeted advertising helps them discover new products.


*Source: https://blog.gitnux.com/targeted-advertising-statistics/*

# Who should we target in our marketing campaign to ensure the maximum ROI?




# Objective & Goal

**Objective**: Understand the **characteristics of the customers** that are more likely to respond to our marketing campaigns and target them in our new campaign

**Goal**: Maximize ROI in our new campaign

# 02

# Data Understanding

# Data Understanding: Variables

**Variables: 29 variables**

| ID | Year_Birth | Education | Marital_ Status | Income | Kidhome | Teenhome | Dt_ Customer | Recency | Complain |
|---|---|---|---|---|---|---|---|---|---|
| integer | integer | categorical | categorical | integer | integer | integer | date | integer | binomial |
| ID of the Customer | Customer's birth year | Customer's education level | Customer's marital status | Customer's yearly household income | Number of children in customer's household | Number of teenagers in customer's household | Date of customer's enrollment with the company | Number of days since customer's last purchase | Whether the customer complains |

| MntWines | MntFruits | MntMeat Products | MntFishP roducts | MntSweetP roducts | MntGold Prods | NumDealsPu rchases | Accepted Cmp1 | AcceptedC mp2 | AcceptedC mp3 |
|---|---|---|---|---|---|---|---|---|---|
| integer | integer | integer | integer | integer | integer | integer | binomial | binomial | binomial |
| Amount spent on wine in last 2 years | Amount spent on fruits in last 2 years | Amount spent on meat in last 2 years | Amount spent on fish in last 2 years | Amount spent on sweets in last 2 years | Amount spent on gold in last 2 years | Number of purchases made with a discount | whether customer accepted the offer in the 1st campaign | whether customer accepted the offer in the 2nd campaign | whether customer accepted the offer in the 3rd campaign |

| AcceptedCmp4z | AcceptedCmp5 | Response | NumWeb Purchases | NumCatalog Purchases | NumStore Purchases | NumWebVisits Month |
|---|---|---|---|---|---|---|
| binomial | binomial | binomial | integer | integer | integer | integer |
| whether customer accepted the offer in the 4th campaign | whether customer accepted the offer in the 5th campaign | Whether customer accepted the offer in the last campaign | Number of purchases made through the company's website | Number of purchases made using a catalogue | Number of purchases made directly in stores | Number of visits to company's website in the last month |

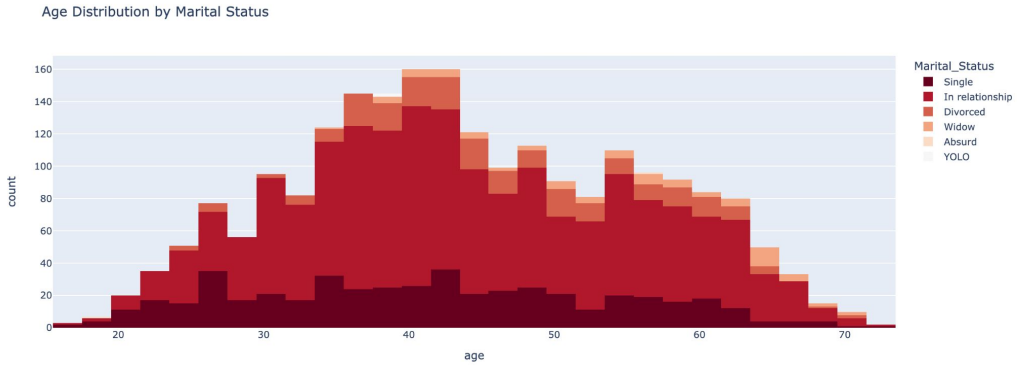
**Target Variable:** Response
**Definition:** Binomial Variable

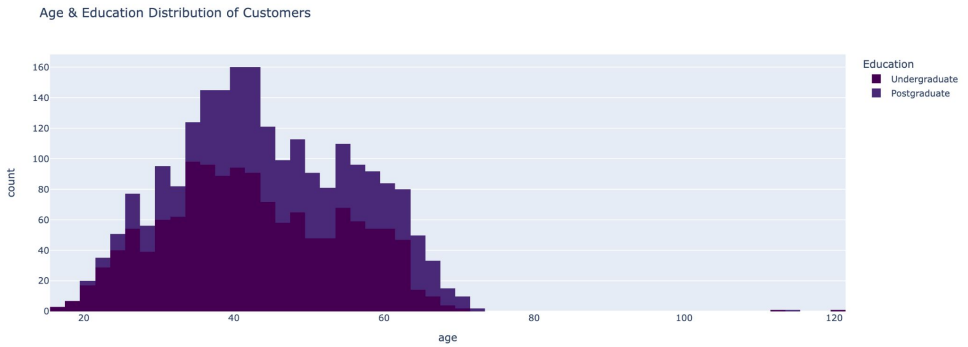
1==customer accepted the offer in the last campaign
0==customer did not accept the offer in the last campaign

**Distribution:** The dataset shows imbalanced distribution of the target variable. 1906 of 2240 records are recorded as 0 while the remaining 334 recorded as 1.

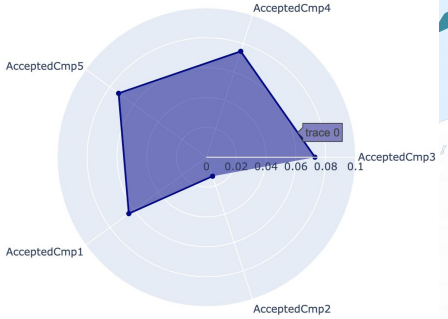
# Data Understanding: Data Exploration

## Distribution of Demographics

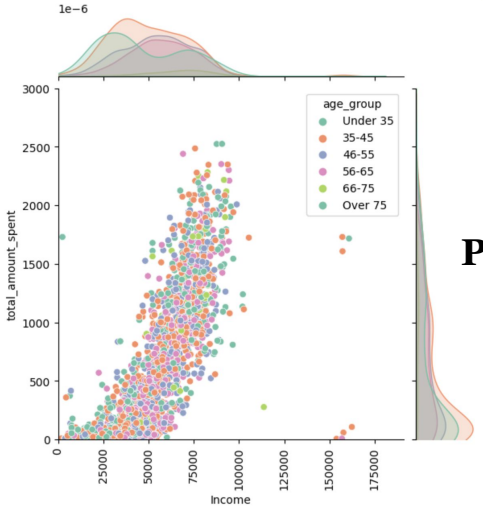



## Promotion Response




## Purchase Behavior



# 03

# Data Preparation

# Data Preparation: Data Preprocessing

- Combine duplicate categories of some variables  into single category
- Extract year part from the variable “Dt_Customer” and assign it to current year
- Create “Age” column
  6 bins: 'Under 35', '35-45', '46-55', '56-65', '66-75', 'Over 75'
- Drop “ID” column
- Replace missing values in “Income” with the mean value
- Apply label encoder to “Education”, “Marital_Status”, “Dt_Customer”

# 04

# Modeling

# Modeling (Pros & Cons)

## Decision Tree

**A flow-like tree structure, supervised learning method for classification and prediction.**

**Pros**

- Easy to understand for executives
- Easy to implement by dividing and conquering
- Convenient to use
- Computationally cheap

**Cons**

- Unstable tree structure: the decision tree is highly dependent upon the customer sample
- Overfitting: tree is like to overfit by utilizing many features that are less relevant

## k-NN

**A simple algorithm that stores all the available cases and classifies the new data based on a similarity measure.**

**Pros**

- Can be used both for Classification and Regression
- k-NN is pretty intuitive and simple
- Non-parametric

**Cons**

- Curse of Dimensionality
- Costly Computation
- Very Sensitive

## Logistic Regression

**A statistical approach that uses linear discriminant function to predict the likelihood of a scenario for solving categorical binary classification problems.**

**Pros**

- Yields better generalization accuracy for smaller training-set sizes
- Able to conduct feature selection
- Easily update with new data

**Cons**

- Prone to overfitting due to its sensitivity to outliers
- Unable to capture difficult relationships
- Can’t handle missing data

# Class Imbalance

We first built models with default parameters to test for the best technique to deal with class imbalance.

The best class imbalance technique is the **SMOTE** (Synthetic Minority Oversampling Technique) **for all three classification models.**

In SMOTE, we randomly increase minority class examples in aim to balance class distributions.

# Modeling (Parameter Optimization)

## Decision Tree

## k-NN

## Logistic Regression

**Parameter Optimization**

criterion: entropy
max_depth: 19
min_samples_leaf: 2
min_samples_split: 2
Nested cross validation:
shuffled sample and 10 folds

metric: manhattan
n_neighbors: 4
weights: distance
Nested cross validation:
shuffled sample and 10 folds

C: 0.1
penalty: l1
solver: saga
Nested cross validation:
shuffled sample and 10 folds

# Comparing the Model Results

## Decision Tree

Accuracy (Out-of-Sample): 0.818
**Precision** (Out-of-Sample): **0.385**
Recall (Out-of-Sample): 0.37

## k-NN

Accuracy (Out-of-Sample): 0.868
**Precision** (Out-of-Sample): **0.622**
Recall (Out-of-Sample): 0.28

## Logistic Regression

Accuracy (Out-of-Sample): 0.888
**Precision** (Out-of-Sample): **0.805**
Recall (Out-of-Sample): 0.33

**Our best model based on our chosen evaluation metric (precision) is the logistic regression.**

# 05

## Evaluation

# Why Do We Use Precision?

## Decision Tree

**Accuracy**
(Out-of-Sample): 0.818
**Precision**
(Out-of-Sample): 0.385
**Recall**
(Out-of-Sample): 0.37

## k-NN

**Accuracy**
(Out-of-Sample): 0.868
**Precision**
(Out-of-Sample): 0.622
**Recall**
(Out-of-Sample): 0.28

## Logistic Regression

**Accuracy**
(Out-of-Sample): 0.888
**Precision**
(Out-of-Sample): 0.805
**Recall**
(Out-of-Sample): 0.33

**Logistic Regression has the best Precision, but NOT the best Recall.**

# Precision vs. Recall

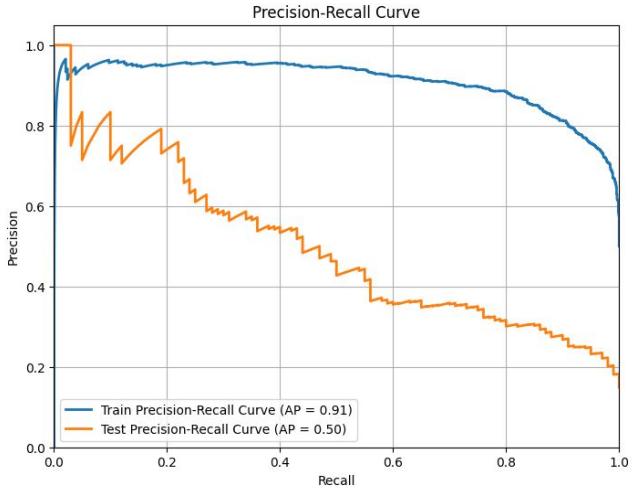
The Business Scenario informs our priority of **Precision**.

The cost of a false positive is higher than the cost of a false negative.

- Missing targeting a particular responsive customer is okay
- Falsely targeting many unresponsive customers with marketing (who will never actually buy) is very costly for business

This is represented by our Logistic Regression model results.

- Low recall = false negatives
- High precision = very few costly false positives



# Why Does Decision Tree Perform Poorly?

## Decision Tree

**Accuracy**
(Out-of-Sample): 0.818
**Precision**
(Out-of-Sample): 0.385
**Recall**
(Out-of-Sample): 0.37

- Our Dataset is relatively small (2240)
- Decision Trees are weaker with small datasets
- Logistic Regression specifically designed for binary classification
- Many categorical features using dummy variables

# 06

# Deployment

# Deployment

## Result Deployed in Real Life

- Internal: Integration with Marketing Platform / Dashboard Creation
- External: Promotion / Referral Program

## Potential Issues

- Data Drift
- Over-segmentation

## Ethical Considerations

- Transparency
- Data Privacy
- Bias

## Risks

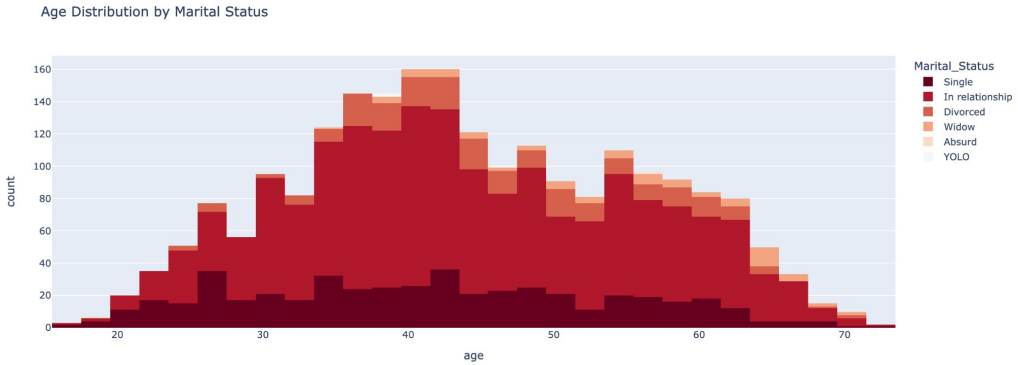
- Poor Model Performance (Implement A/B testing)
-

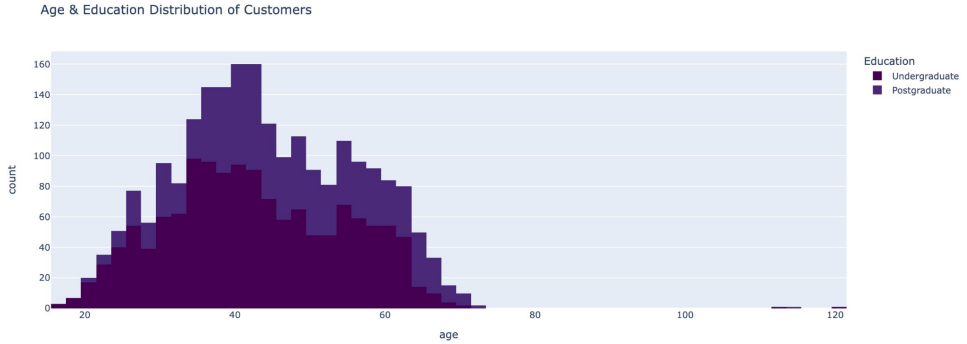
# Questions

# Appendix

# Appendix 1: Data Exploration

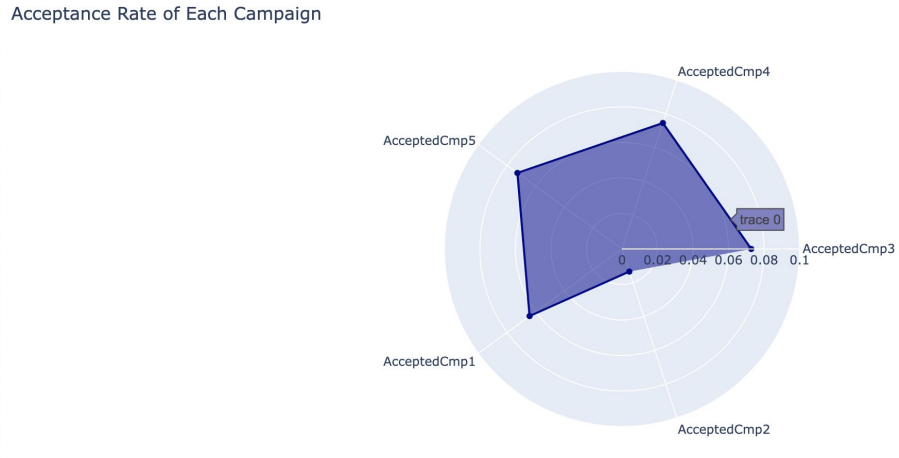
## Distribution of Demographics



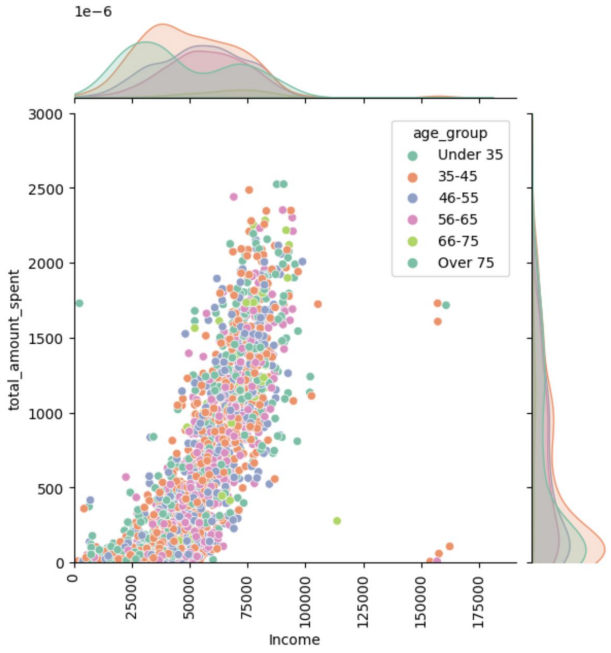



# Appendix 2: Data Exploration (Cont.)

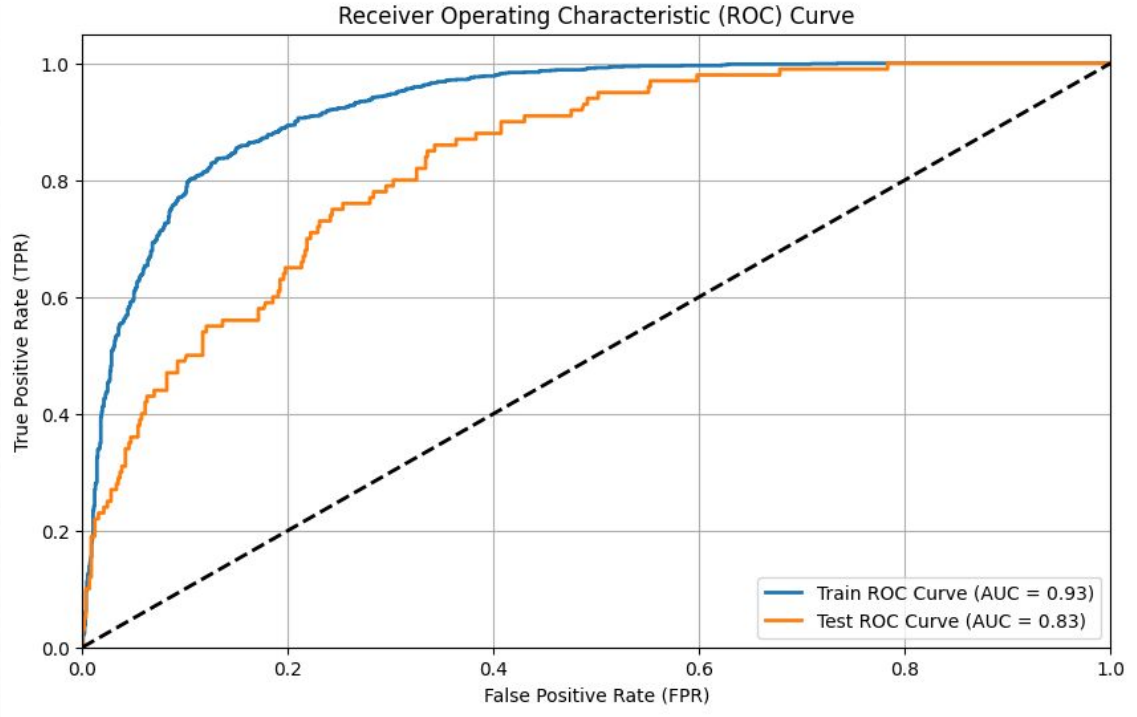
## Promotion Response



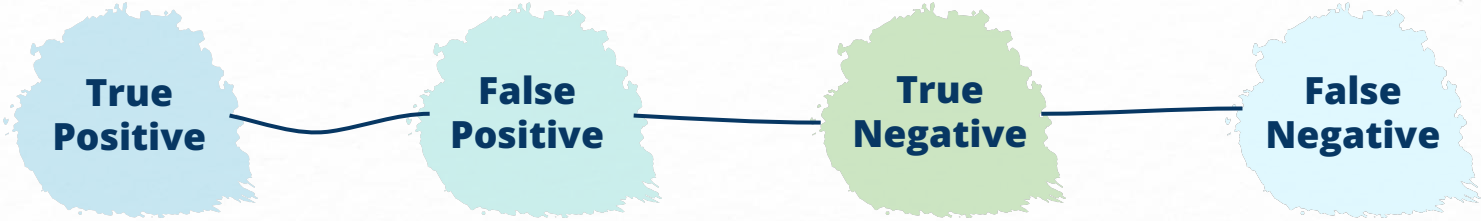

## Purchase Behavior



# Appendix 3: ROC Curve



# Appendix 4: Confusion Matrix




| True Positive | False Positive | True Negative | False Negative |
| --- | --- | --- | --- |
| We targeted the customer and they responded to our campaign | We targeted the customer but they did not respond to our campaign | We did not target the customer and they did not respond to our campaign | We did not target the customer and they responded to our campaign |

# Appendix: Source

Dataset: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Targeted Marketing Statistics: https://blog.gitnux.com/targeted-advertising-statistics/