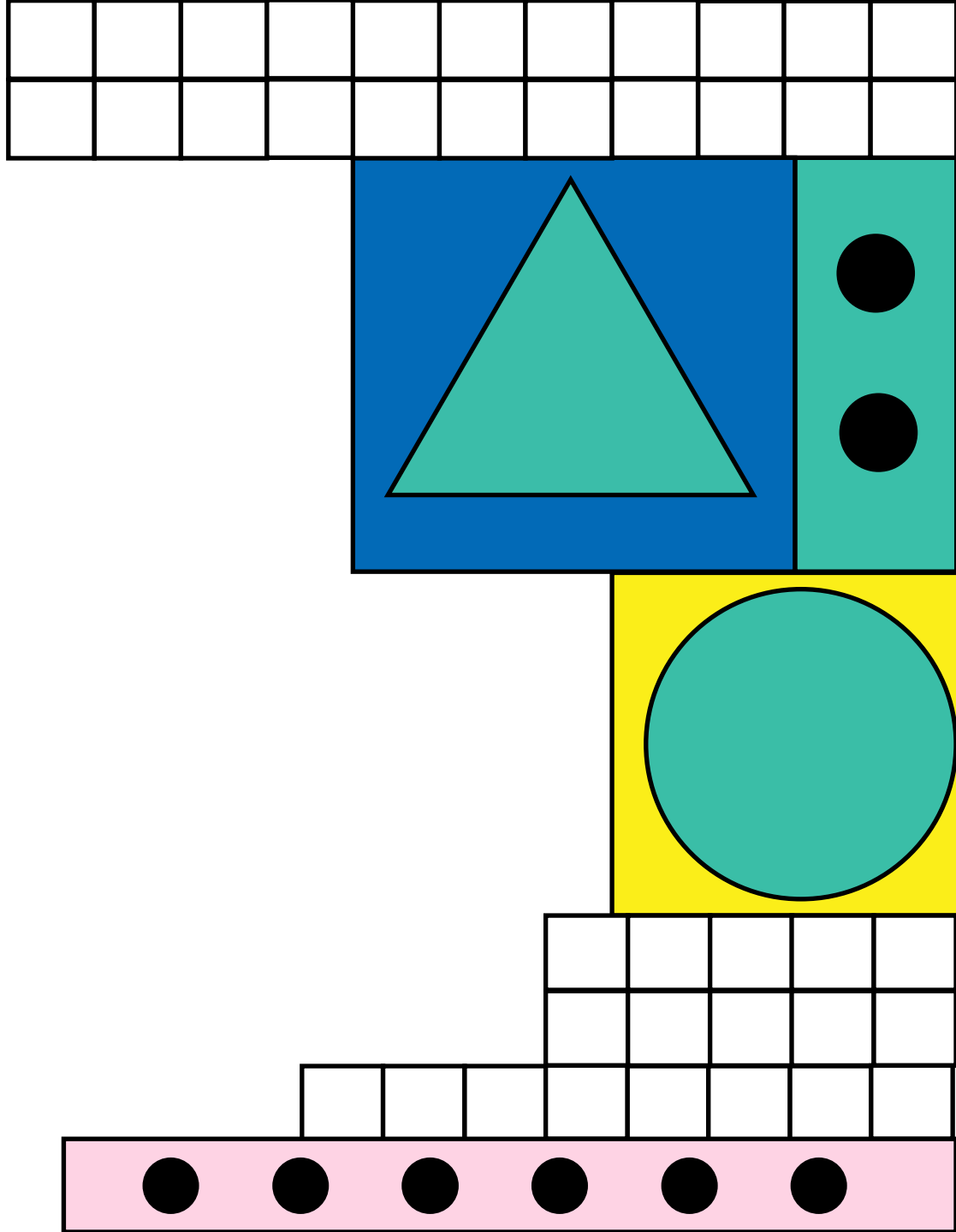
Portfolio

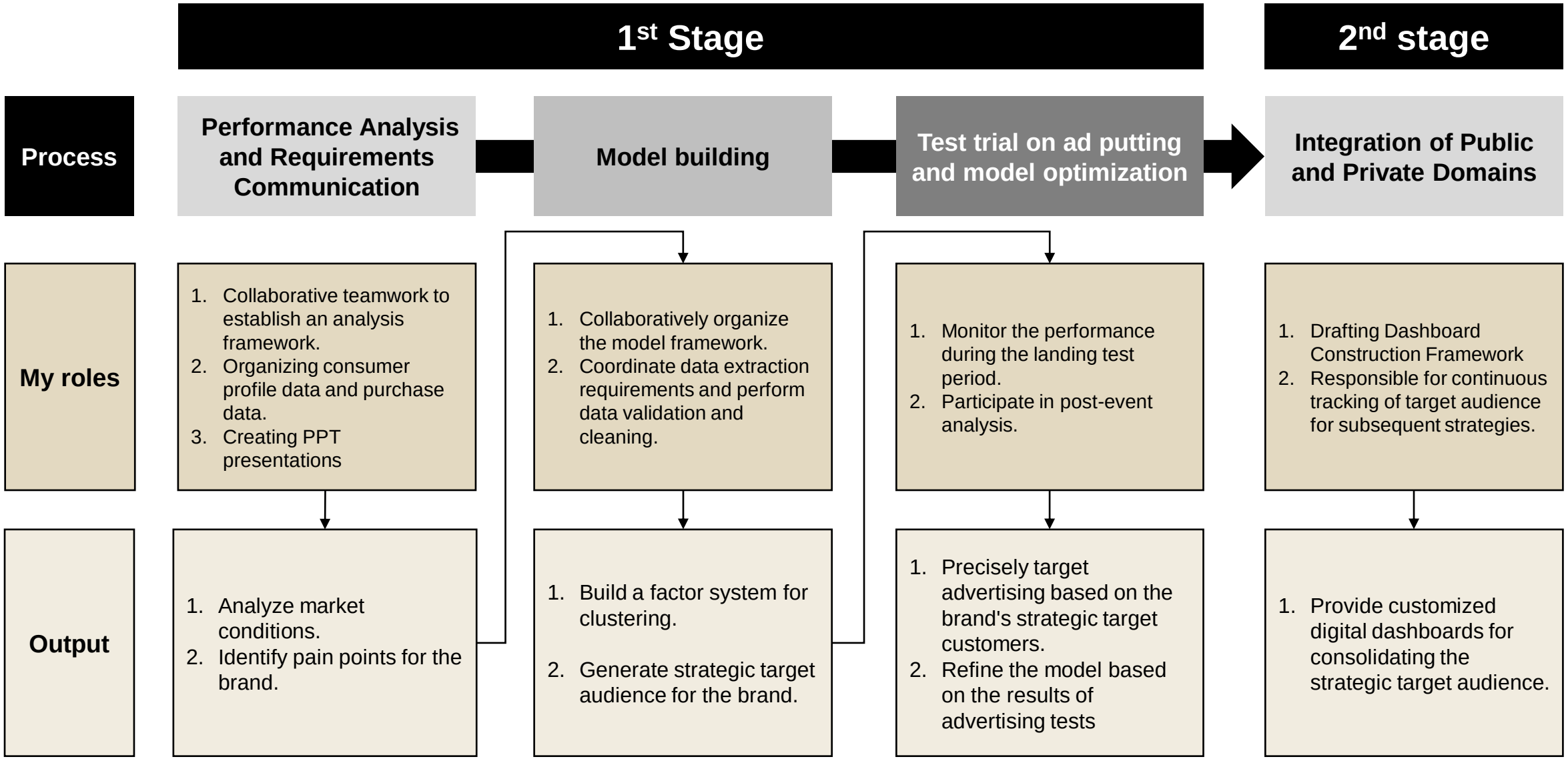
-- CHEN Yunshan --

Digital Hub Project

- full-time job project -



Project Description: collaborates with Tmall Ecological Experiment Lab and brand partners to meet the digital operational needs of various brands.



Purpose

- **High-end customers recruitment;**
- **Strategic planning based on growth curve;**

Painpoints

- **Multi-brands;**
- **Definition of industry premium customers not suitable for Brand A;**

How to segment the customers? How to find out the target group?

1 Category Mindset

WHAT CATEGORY

Factors:

- Total purchase amount
- Purchase frequency
- Avg. item price
- ...

2 Brand Mindset

WHAT BRAND

Factors:

- Prefer global brands
- Prefer domestic brands
- Prefer emerging brands
- …

3 Purchase Window

WHERE

Factors:

- Promotion window
- Gift seasons window
- Non-promotion window
- ...

4 Purchase Channel

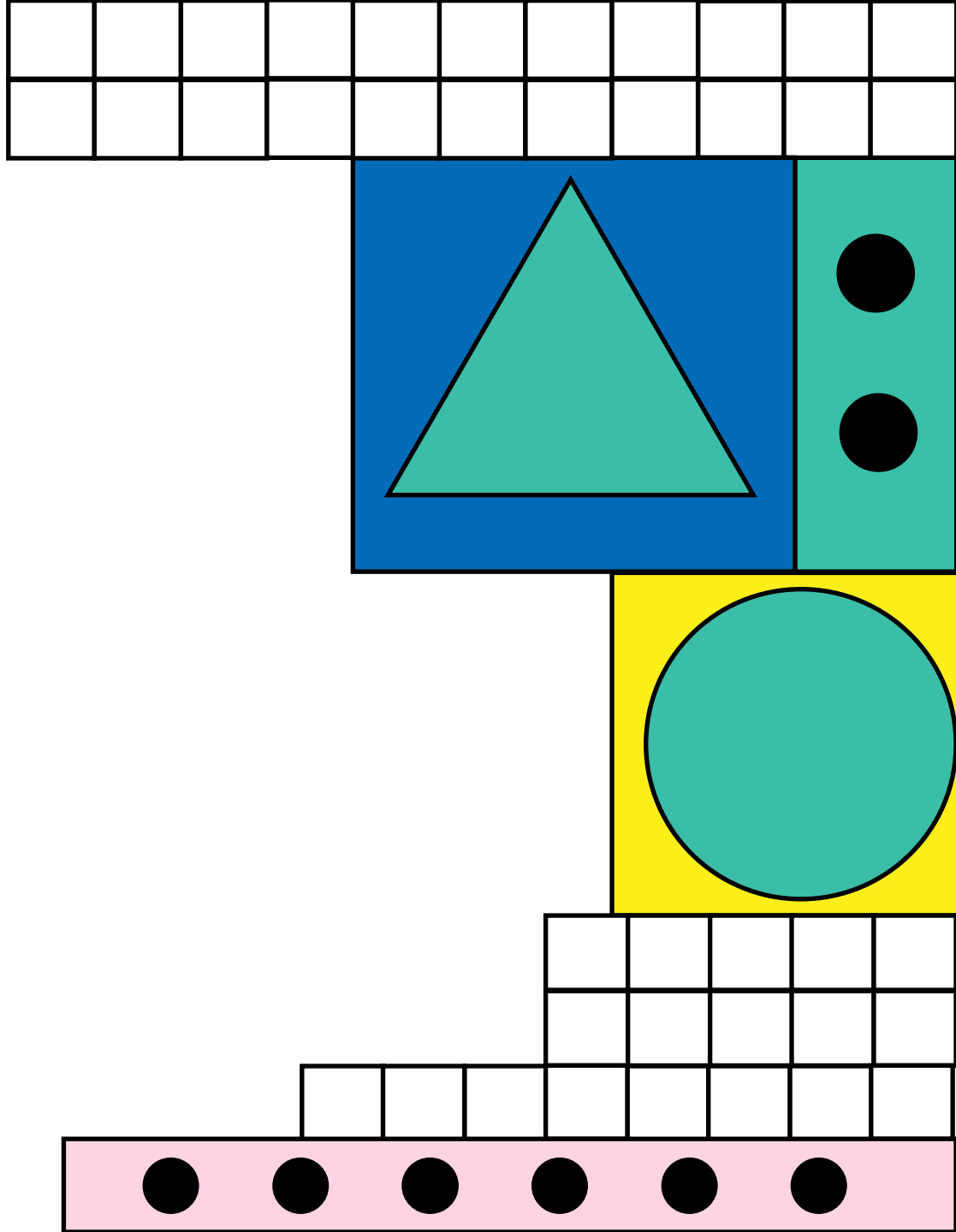
HOW

Factors:

- T-mall platform
- Taobao platform
- T-mall global platform
- ...

Click or Not

-school project using R language-



Business Problem Definition

QUESTION

- What kind of customers(with what attributes) would tend to click on the advertisement?

HOW

- Using machine learning methodology to predict customers'behaviors
  - **Logitstic regression**
  - **Bayesian classification**

This is a group assignment in BA program. I was in charge of the Bayesian classification coding.
Detail can be found in https://github.com/Chenyunshan33/Click-or-Not

Dataset Preparation

Dataset Understanding

- **Data size**: 110,000
- **Data source**: Taobao
- **Data structure:**
  - **ad_feature.csv**: a description of the advertiser who placed the ad
  - **raw_sample.csv**: statistics on the basic information of the ad placement
  - **user_profile.csv**: a description of the user profile

Data Preprocessing

- **Data linkage:** By linking user_id and adgroup_id, we merge raw_sample, ad_feature and user_profile into one dataset
- **Missing value:** Removing missing values and undefined attributes
- **Attributes Stratification**
  - Age level
  - Price
  - Time

Training & Test Data

- Randomly selected **20,000** of 110,000 pieces of data
- Split data set into **training and test data** (7:3)
- Build a **logistic regression** and a **Bayesian classification** model classification model respectively

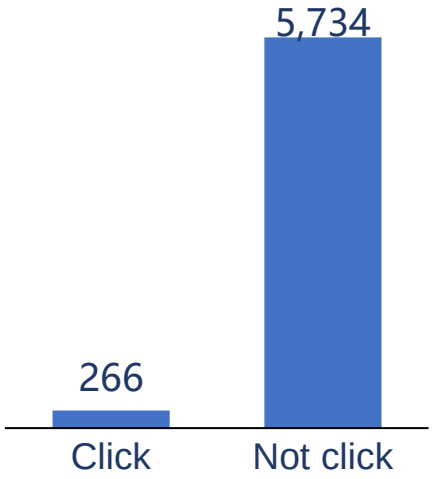
Problem: can only predict “0”

Test_dataset Confusion Matrix

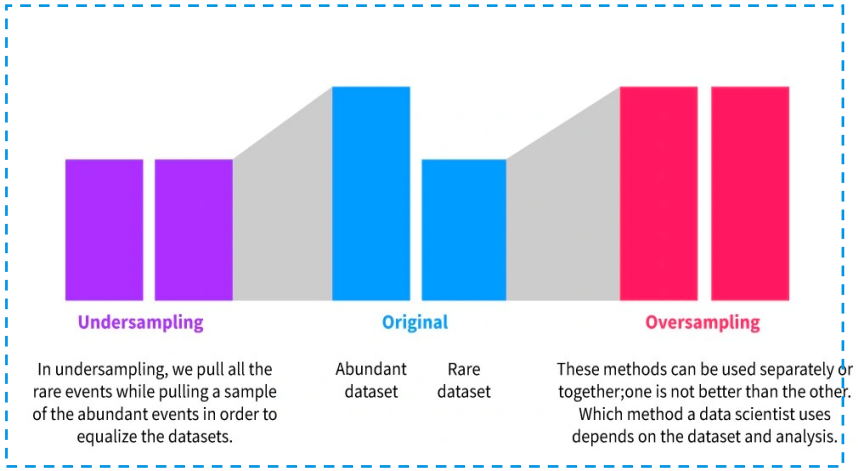
Predicted \ Actual	1	0
1	0	0
0	266	5734

* 0 – not click 1 - click

Response value distribution in dataset



Method to handle imbalanced data



- Model can **only predict “0”**, which means it can not distinguish click customers

Significant volume difference

- **Method:** Oversampling
- **Basic Idea:** randomly duplicates minority data points in order to increase its count

Model Comparison: using Naïve Bayes to model

Logistic regression

Predicted \ Actual	1	0
1	36	987
0	230	4747

* 0 – not click 1 - click

Accuracy	0.797
Specificity	0. 828
Sensitivity	0.135
Balanced Accuracy	0.482

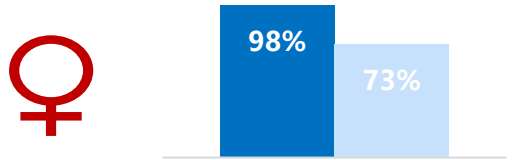
Bayes classifier

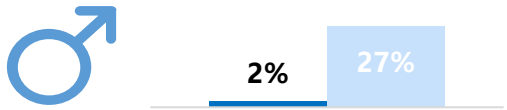
Predicted \ Actual	1	0
1	187	612
0	108	5093

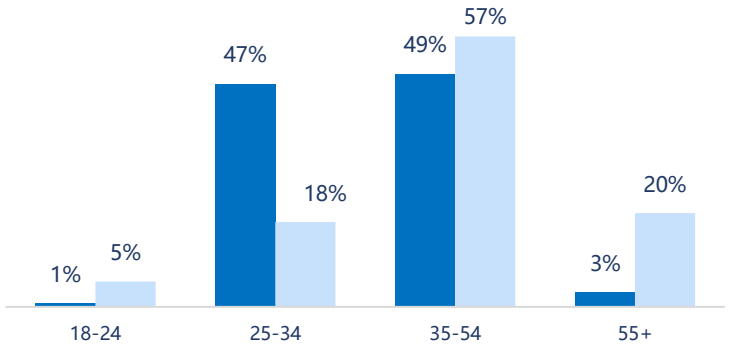
* 0 – not click 1 - click

Accuracy	0.880
Specificity	0.893
Sensitivity	0.634
Balanced Accuracy	0.763

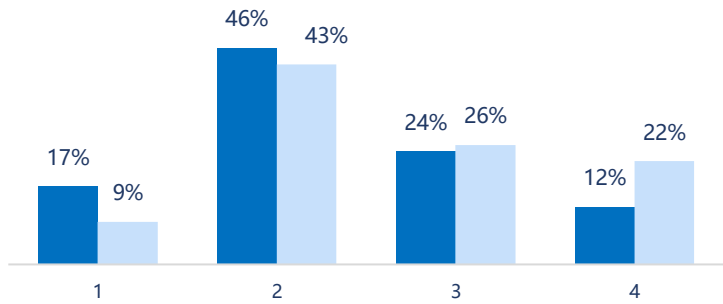
Predictor Profile: women and young generation who are in the middle purchase power are more likely to click



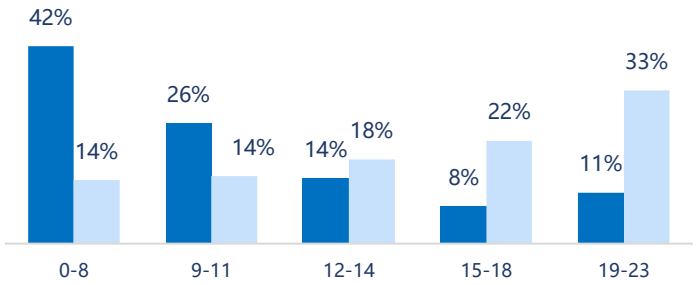




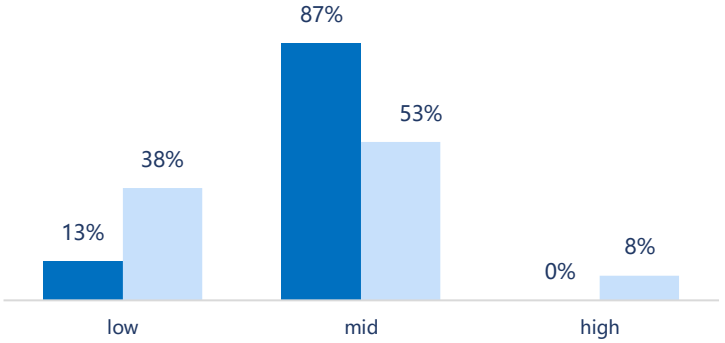
Age



User Class



Time Period



Purchase Value Level

Predict to click

Predict not to click