

Starbucks capstone project

Thomas Heinemann

December 26, 2023

1 Project overview

This is a recommendation engine project proposed by Starbucks to find out which demographic groups of customers respond best to certain types of offers. The underlying datasets include transactional and customer data, as well as information about the types of offers sent. The full set of files, including the main Jupyter file with the detailed task, can be found at [1]. In addition, an approach for a simpler version of such a problem can be found here [2].

References

- [1] All project files: https://github.com/thomasheinemann/starbucks_promotion_type_exercise/
- [2] Simpler version of this project: https://github.com/thomasheinemann/starbucks_promotion_exercise/

2 Introduction

To begin with we first present the original introduction and task:

This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). Some users might not receive any offer during certain weeks.

Not all users receive the same offer, and that is the challenge to solve with this data set.

Your task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.

Every offer has a validity period before the offer expires. As an example, a BOGO offer might be valid for only 5 days. You'll see in the data set that informational offers have a validity period even though these ads are merely providing information about a product; for example, if an informational offer has 7 days of validity, you can assume the customer is feeling the influence of the offer for 7 days after receiving the advertisement.

You'll be given transactional data showing user purchases made on the app including the timestamp of purchase and the amount of money spent on a purchase. This transactional data also has a record for each offer that a user receives as well as a record for when a user actually views the offer. There are also records for when a user completes an offer.

Keep in mind as well that someone using the app might make a purchase through the app without having received an offer or seen an offer.

So how do we proceed? A simple statistical evaluation based on the frequency of offer completions or views per customer may not work, as we do not have huge data sets for each person with and without receiving, viewing and completing a considered offer type. Also, a customer being researched may not have exactly the same profile as those provided. We therefore improve the statistics by reducing the value set of the customer profile data. However, this is not enough, so we train models to predict as accurately as

possible whether the customer is worth promoting or not. The advantage of this is that these models also make good use of information from other more or less closed customer profiles.

In the following sections, we provide definitions, present the modelling approach, summarise the results for each type of offer and conclude with our main findings in the conclusion section.

3 Definitions

Main offer types to be considered:

- BOGO (buy one get one free) offer: If the customer buys the praised product of this offer then this person gets the same product on top for free.
- discount offer: If the customer spends a certain amount then this person receives a discount.
- informational offer: These offers just show the customer an advertisement for a product, e.g., a drink. In the given data, however, there is no offer completion record.

Offer instance: An offer sent for one specific customer, at one specific time (offer_received_time). So each offer instance is addressed with the key [offer_id, person, offer_received_time] which is found being unique.

Offer events:

- offer received
- offer viewed (can also happen after offer is completed or no longer valid)
- transaction = the customer makes a transaction of a certain amount
- offer completed = the difficulty challenged by the offer has been overcome with the current transaction

Offer event times: Times of the offer events above are further denoted as "offer_received_time", "offer_viewed_time", "transaction_time", "offer_completed_time".

Offer event record (OER): An OER consists of the offer receiving event, a possible offer viewing event as well as an offer completed event, and all transaction events belonging to one specific offer instance. Accordingly like each offer instance, each OER is addressed with the unique key [offer_id, person, offer_received_time]. For each person we also define one OER related to no-offer times. These are addressed as [offer_id="no offer", person, offer_received_time=0]. Each OER has an offer validation interval starting at its offer_received_time and finishes with the

$$\text{offer_time_out} = \text{offer_received_time} + 24 \times \text{duration}.$$

The duration marks the maximum offer-specific time interval in days, in which the offer can be valid, if not completed earlier.

States of an offer instance:

Each offer instance is characterized by the binary state variables "promoted", "viewed" and "completed". Their calculation requires the corresponding OER and is shown in the following table.

state variables of an offer instance	calculation in terms of logical expressions
promoted	$\text{offer_id} \neq \text{"no offer"}$
viewed	$\text{offer_viewed_time} \leq \text{offer_end_time}$ with: $\text{offer_end_time} = \text{Min}(\text{offer_time_out}, \text{offer_completed_time})$
completed	$\text{offer_completed_time} \leq \text{offer_time_out}$

Consumption during offer: It is a quantity referring to one OER. It is defined as the sum of all transactions done in an OER where each transaction lies in the time interval [offer_received_time, offer_time_out].

4 Modeling approach

4.1 Data flow

The data flow along our prediction pipeline starts with the original customer profile attributes and finishes with a decision on whether to promote that person or not. The first step consists in cleaning NaN values and introducing categorical variables for the gender attribute. In the second step, model 1 predicts an early viewing preference used to sort out "lazy viewers". For predicting the final promotion decision, we used in step 3 the model 2 which is specifically fitted for the each offer type separately.

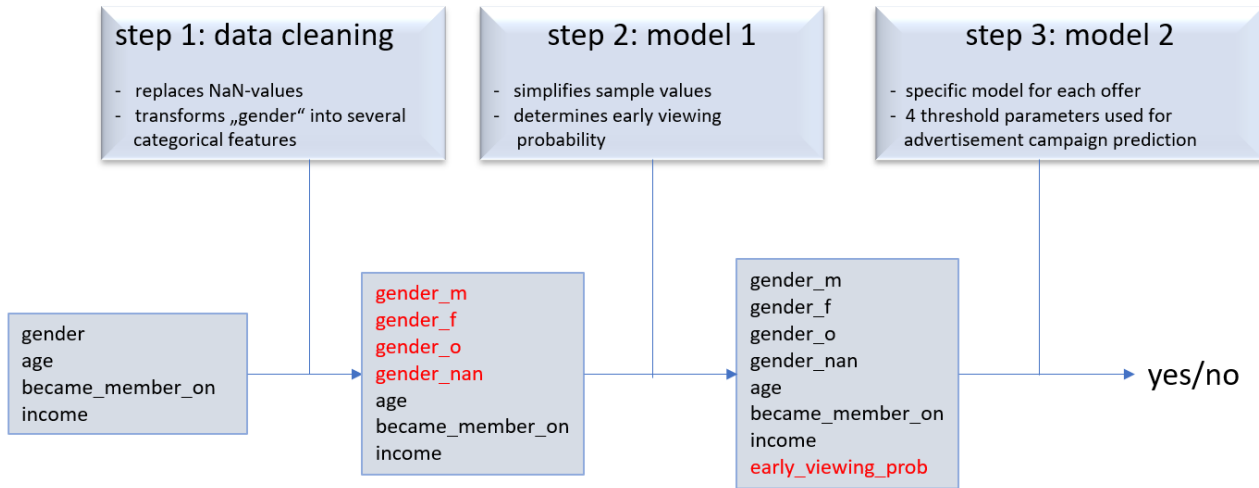


Figure 1: Data flow showing a prediction pipeline for customer profile attributes "gender", "age", "became_member_on", "income" towards a "yes/no"-prediction for an advertisement campaign. For each offer type a differently fitted version of model 2 is used. The models 1 and 2 have been fitted using customer profile attributes and the customers' offer event records (OERs).

4.2 Step 1: data preparation

The customer profile data is first undergoing a NaN-value treatment for the income. In particular, customers who do not provide an income, are remarked as having an income of -10 k dollar. Another issue being treated within this step is the gender attribute which is categorical without order and thus splitted into one-hot-encoded attributes for each class label yielding "gender_m", "gender_f", "gender_o", and "gender_nan" (denotes if no gender is provided).

4.3 Step 2: model 1

Model 1 has the task of simplifying the value sets of the customers' integer features "age", "become_member_on" and "income" using a transformer to achieve 10-year-steps for the "age", 1-year steps for "become_member_on" and 10 k dollar steps for "income". Then, in a second transforming process, a test-set customer yields, within the framework of model 1, a viewing preference encoded in the attribute, "early_viewing_prob" (probability for the offer to be viewed within 48 h). The underlying model of the second transformer is fitted with all the cleaned customer profile data of the profile training set along with informations of their offer instances' states. The offer instances to be considered comprise only the customers' non-overlapping ones. This constraint was chosen because multi-offer effects may be quite complex and would constitute a further investigation step probably requiring more data.

4.4 Step 3: model 2 (offer specific model)

In this step, an offer specific model comes into play. Accordingly, for each offer type, a separate instance of this model is used. The model itself makes use of cleaned customer profile data and all the customers' OERs of non-overlapping offer instances of the considered offer type and their no-offer OERs. The "offer completion" states for the no-offer instances are calculated in this model as if the customer was promoted

with the offer type considered in this model. To be precisely, a difficulty had to be overcome that scales with the ratio between the no-offer and offer duration.

With respect to states of offer instances, we generally distinct six since an offer can be sent or not, can be viewed (if promotion exists) or not, and completed or not. A grouping of these six classifications into four different labels forms the base of the current model and is shown in the following table.

label	states of offer instances		
	promoted	viewed	completed
I	yes	yes	yes
II	no	no	yes
	yes	no	yes
III	yes	yes	no
IV	no	no	no
	yes	no	no

The goal of model 2 is to make use of the upper 4-label classification of offer instance states to predict customers who complete the offer only if viewed. These customers ideally have only offers with states labeled I and IV, respectively. In Fig. 2 a Venn diagram is shown that depicts ideal customer sets along with the one we are interested in. We hereby define an ideal customer as one for who the offer completion status solely depends on the viewing status or is even independent of the latter. This definition leads to the four different types of ideal customers shown in the figure.

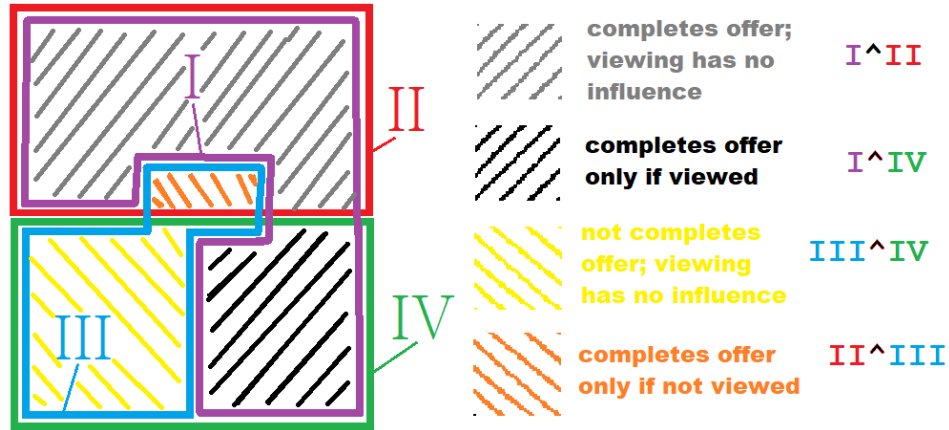


Figure 2: A Venn diagram representing a grouping of ideal customers into four disjunct sets filled with differently colored diagonal stripes is shown. In each set only two distinct offer instance state labels occur (represented by Roman numbers).

Unfortunately, our customers are not ideal, so each customer behaves differently with respect to each singular offer instance. We can therefore only predict what behavior the customer is most likely to reveal by assuming a correlation between the customer's profile and their offer instances' states.

In order to determine whether a customer is "completing the offer only if viewed", we create one predictor to distinguish between customer profiles having states with label II vs. those having states with label IV and in analogy a second one to distinguish between I vs. III. Using their prediction probabilities for customers with label I as well as label IV states, one can achieve a good estimate for the black-striped area. To do so, the threshold parameters named as p_1 and p_2 for both predictors have to be overcome.

However, some customers even get repelled from buying when viewing an advertisement (see orange-striped area). These customers are not in the black-striped area but since the border areas are fuzzy we further included a third predictor (with threshold parameter p_3) distinguishing customer sets with label $I \cup IV$ vs. $II \cup III$. This allows a separate tuning of the border between the orange and black-striped area as well as the border between the orange and gray-striped area. In addition, we implemented a fourth threshold

parameter p_4 for the early viewing probability as determined in model 1, for sorting out lazy viewers in the results set. This may prevent having too much customers being unknowingly rewarded.

The threshold parameters are optimized for each model using a 3-fold cross validation method. The scoring metric for the validation sets as well as the test set, used for each model 2, is defined in the following.

4.5 Scoring metric

The scoring metric was chosen to take two important aspects into account. On the one hand we want to maximize the number of customers who complete an offer when receiving it. The underlying metrics to maximize is the "incremental response rate" ("irr"). On the other hand, we aim at having a gain in consumption by maximizing a shifted version of the "net incremental revenue" ("nir"). This prevents our recommendation algorithm to promote too many customers whose reward makes them spend less as they are too satisfied. As a scoring metrics we accordingly used a compromise being a product of both former metrics, i.e.

$$\text{scoring metric} = irr \times nir \quad (1)$$

Both metrics are defined in the following.

Incremental response rate "irr" (for offer type X):

$$irr = \frac{n_{prom, compl}}{n_{prom}} - \frac{n_{no\ prom, compl}}{n_{no\ prom}} \quad (2)$$

$n_{prom, compl}$: number of instances of offer type X which are completed

n_{prom} : number of instances of offer type X

$n_{no\ prom, compl}$: number of no-offer instances whose consumption rate would have completed an offer of offer type X

$n_{no\ prom}$: number of no-offer instances

Shifted net incremental revenue "nir" (for offer type X): We take the sum of the consumptions of completed offers (of offer type X) minus the sum of consumptions of no-offer instances whose consumption rate would fulfill an offer completion event of offer type X. Each consumption is normalized by its corresponding offer/no-offer duration.

$$nir = \sum_{\substack{i \in \text{completed offer} \\ \text{instances in} \\ \text{test set}}} \frac{\text{consumption}(i)}{\text{duration}(i)} - \sum_{\substack{k \in \text{"completed"} \\ \text{no-offer instances} \\ \text{in test set}}} \frac{\text{consumption}(k)}{\text{duration}(k)} - \text{shift} \cdot n_{\text{test}} \quad (3)$$

$\text{consumption}(x)$: consumption during offer/no-offer instance x

$\text{duration}(x)$: duration of offer/no-offer instance x

n_{test} : number of all offer/no-offer instances (in the test set)

n_{train} : number of all offer/no-offer instances (in the training set)

$$\text{shift} = \frac{1}{n_{\text{training}}} \left(\sum_{\substack{i \in \text{"completed"} \text{ offer} \\ \text{instances in} \\ \text{training set}}} \frac{\text{consumption}(i)}{\text{duration}(i)} - \sum_{\substack{k \in \text{"completed"} \\ \text{no-offer instances} \\ \text{in training set}}} \frac{\text{consumption}(k)}{\text{duration}(k)} \right)$$

5 Results

5.1 Diagram for the test set

We present in this subsection a cluster analysis among all customers using five histograms covering gender, income, age, membership begin, and viewing preference and thereby stack the bin counts for each cluster

on top of each other using different colors. The clusterization has been performed in a five dimensional space spanned by all the former-mentioned demographic attributes represented by one histogram each. The optimal number of clusters have been detected using the kmeans++ algorithm with the help of the silhouette score and is limited hereby to five for simplicity.

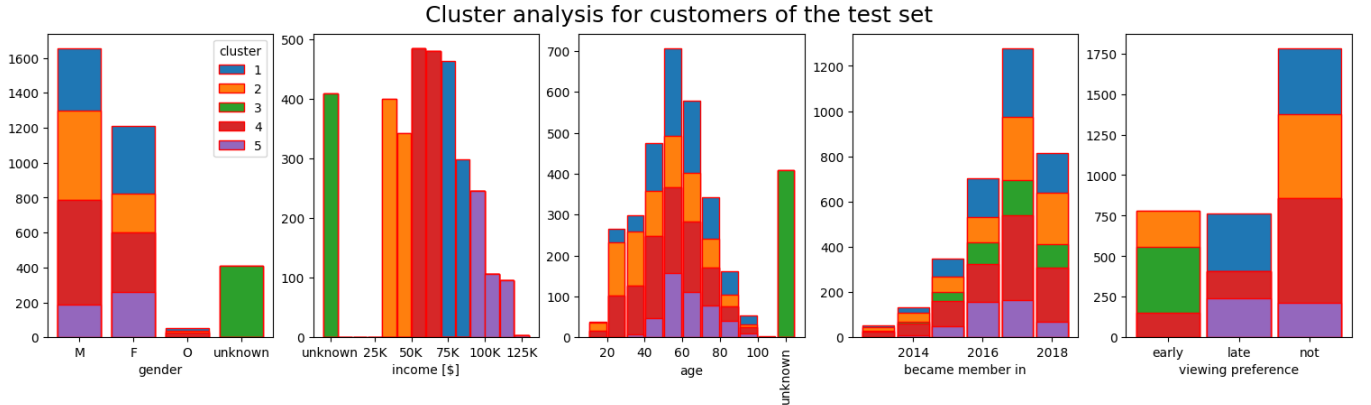


Figure 3: Cluster analysis for all customers in the test set using 5 clusters. With respect to the clusterization, we can recognize that people, who do not provide information about gender, income and age are entirely early viewers. Customers earning more than 70 k are slightly dominated by women. Those earning more than 90 k dollar are not under 40 years old and do not tend to view offers within two days.

5.2 Diagrams with observations for each offer type

In this subsection, we briefly analyze for each offer type the customers we recommend for a promotion and also those we do not want to recommend according to the here presented strategy. Each figure represents the results of one offer type analysis and consists of three rows of diagrams which cover the demographic information gender, income, age, membership begin, and viewing preference. The upper row of diagrams displays data of all customers and highlights the customers who should be promoted according to the found strategy. The mid row comprises only data of customers to be promoted, whereas the bottom row covers only customers not to be promoted. The customers' demographic data in the corresponding histograms have been clustered in the five dimensional space spanned by all the former-mentioned demographic attributes. The clusterization has been performed separately for the target and not-to-target customers.

5.2.1 Diagrams for "bogo" offers

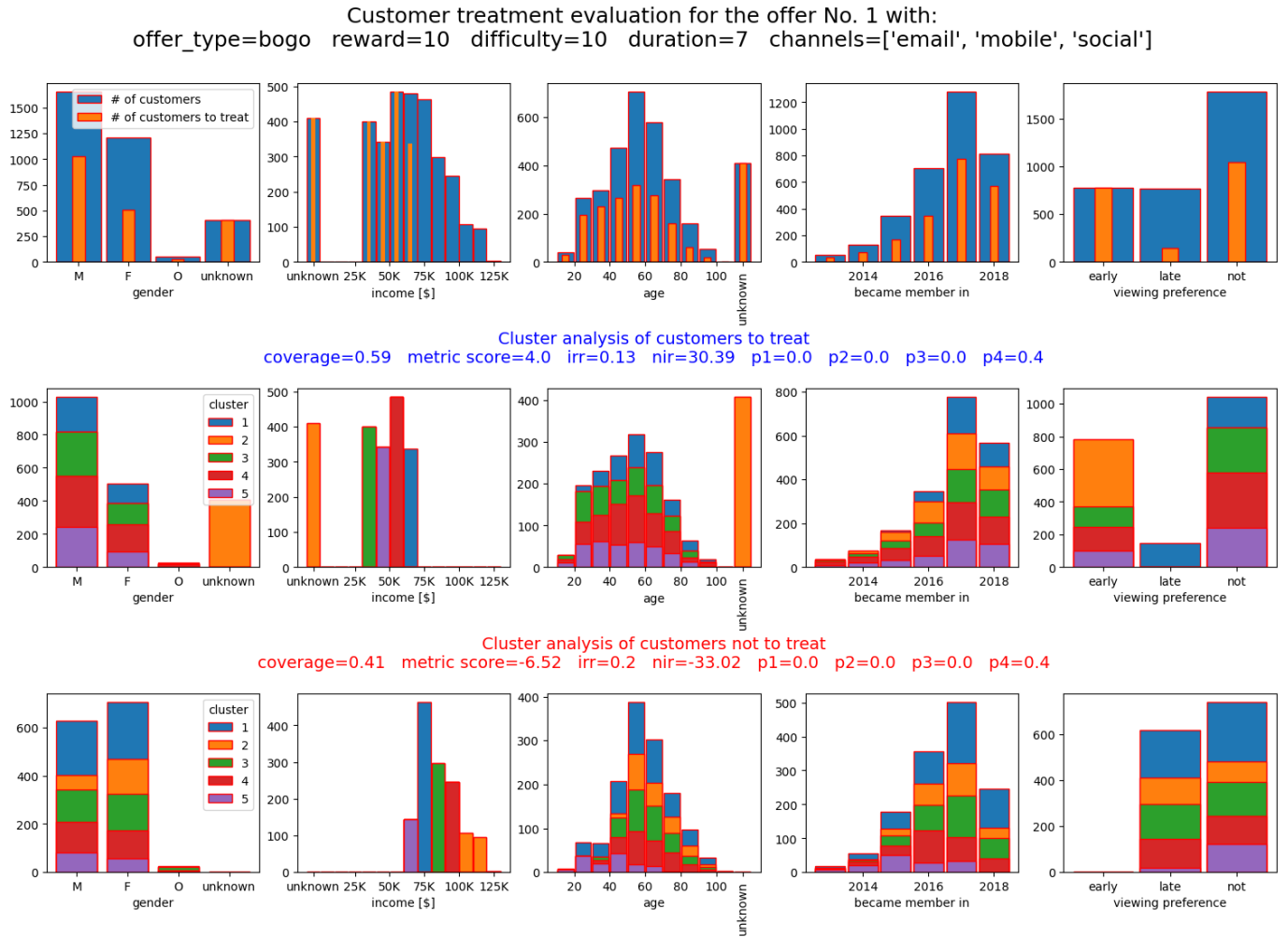


Figure 4: Customers with incomes up to 70 k dollars are worth being addressed by this offer. Those who do not provide demographic information should be included as well. The considered customers tend to view early or rather not.

Customer treatment evaluation for the offer No. 2 with:
offer_type=bogo reward=10 difficulty=10 duration=5 channels=['web', 'email', 'mobile', 'social']

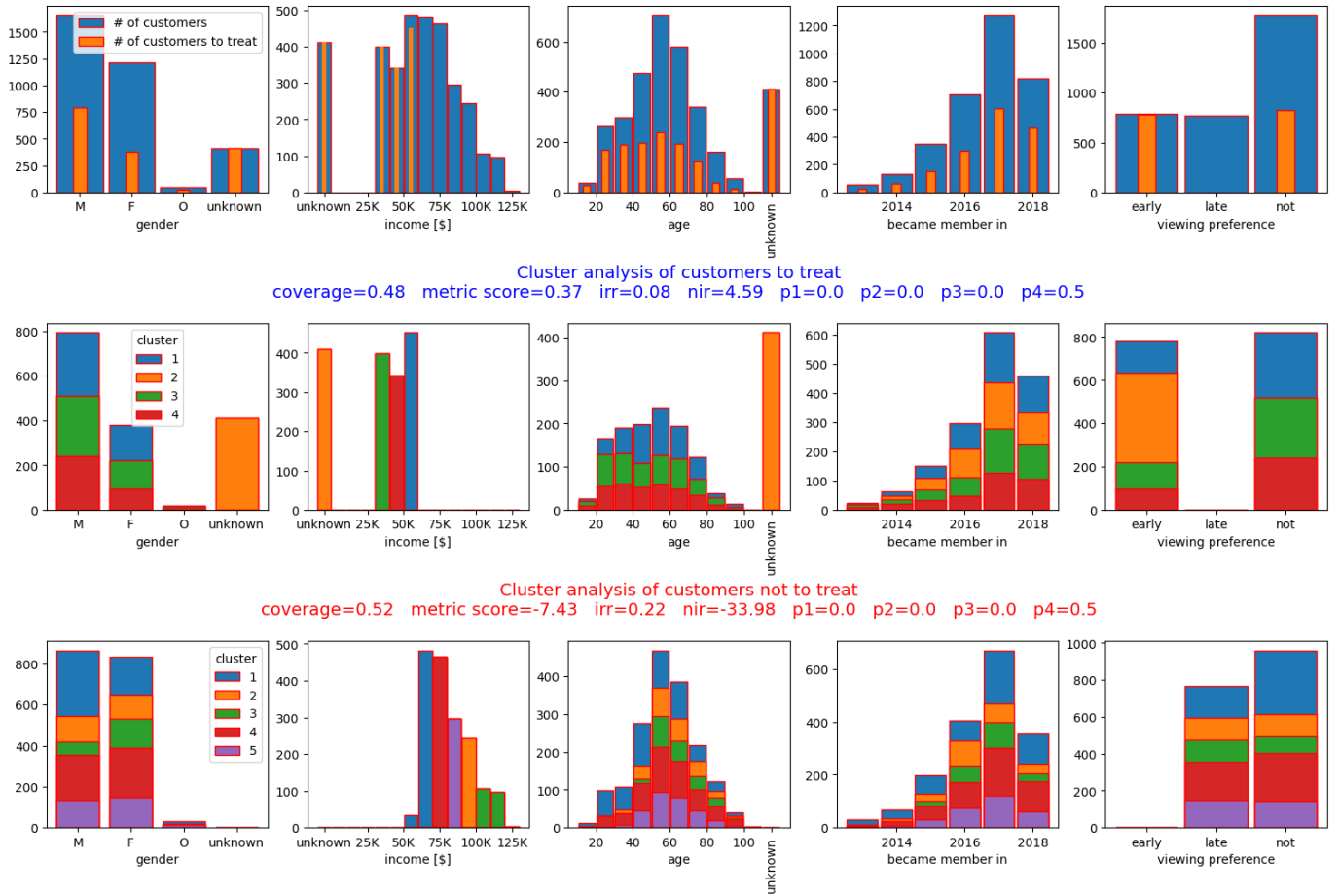


Figure 5: Customers with incomes up to 60 k dollars are worth being addressed by this offer. Those who do not provide demographic information should be included as well.

Customer treatment evaluation for the offer No. 3 with:
offer_type=bogo reward=5 difficulty=5 duration=7 channels=['web', 'email', 'mobile']

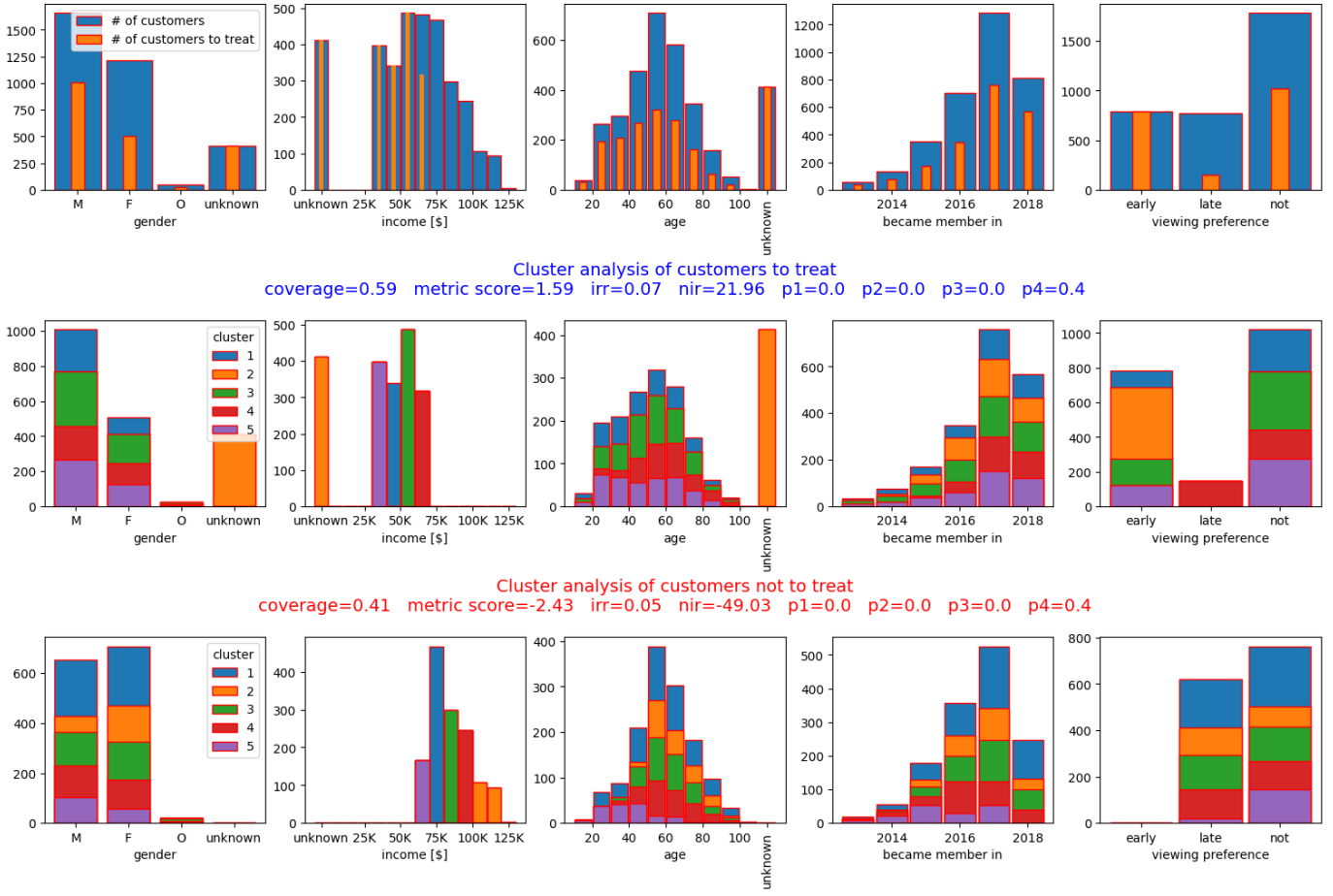


Figure 6: Customers with incomes up to 70 k dollars are worth being addressed by this offer. Those who do not provide demographic information should be included as well.

Customer treatment evaluation for the offer No. 4 with:
offer_type=bogo reward=5 difficulty=5 duration=5 channels=['web', 'email', 'mobile', 'social']

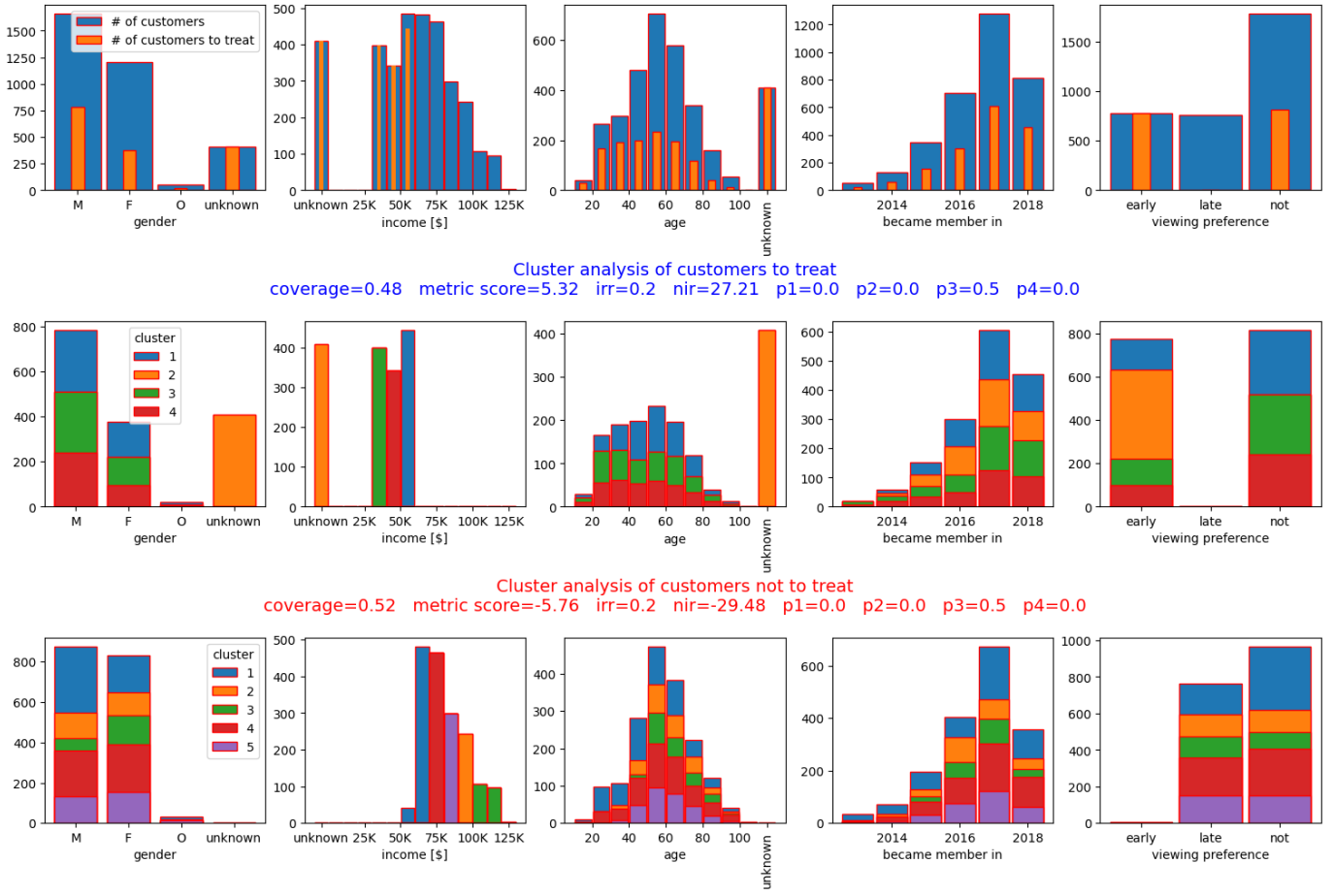


Figure 7: Customers with incomes up to 60 k dollars are worth being addressed by this offer. Those who do not provide demographic information should be included as well.

5.2.2 Diagrams for "discount" offers

Customer treatment evaluation for the offer No. 5 with:
offer_type=discount reward=5 difficulty=20 duration=10 channels=['web', 'email']

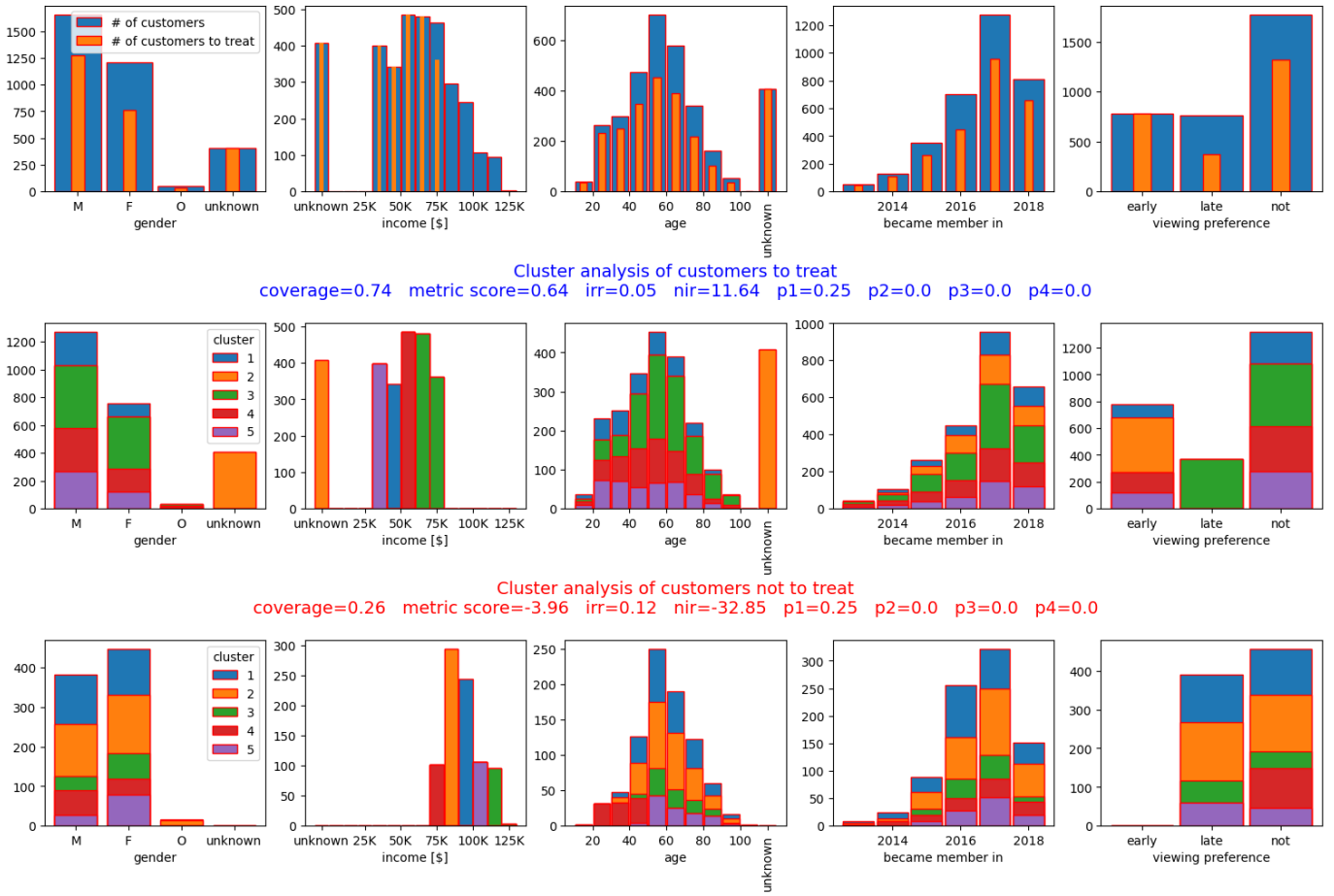


Figure 8: Customers with incomes up to 80 k dollars are worth being addressed by this offer. Those who do not provide demographic information should be included as well.

Customer treatment evaluation for the offer No. 6 with:
offer_type=discount reward=3 difficulty=7 duration=7 channels=['web', 'email', 'mobile', 'social']

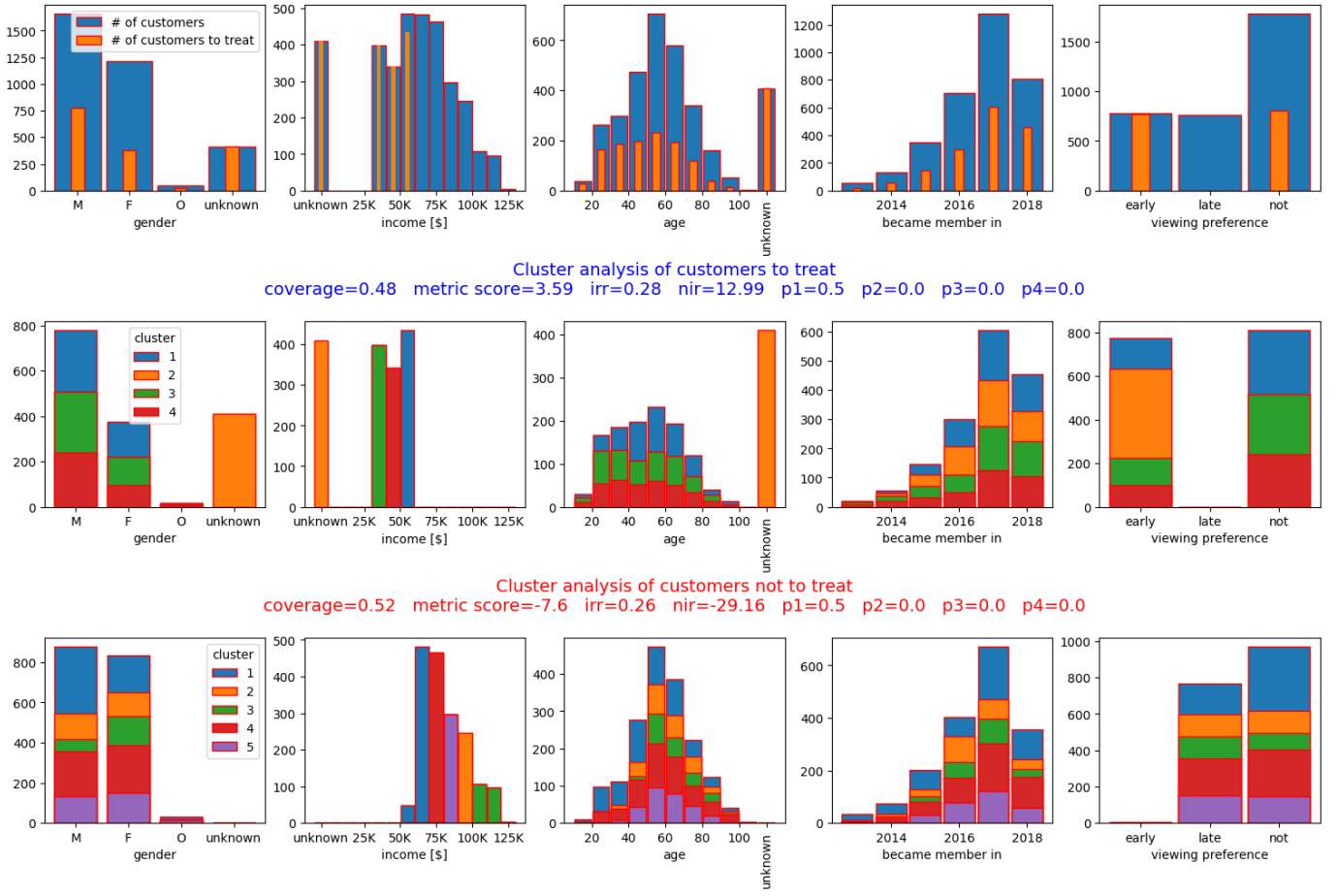


Figure 9: Customers with incomes up to 60 k dollars are worth being addressed by this offer. Those who do not provide demographic information should be included as well.

Customer treatment evaluation for the offer No. 7 with:
offer_type=discount reward=2 difficulty=10 duration=10 channels=['web', 'email', 'mobile', 'social']

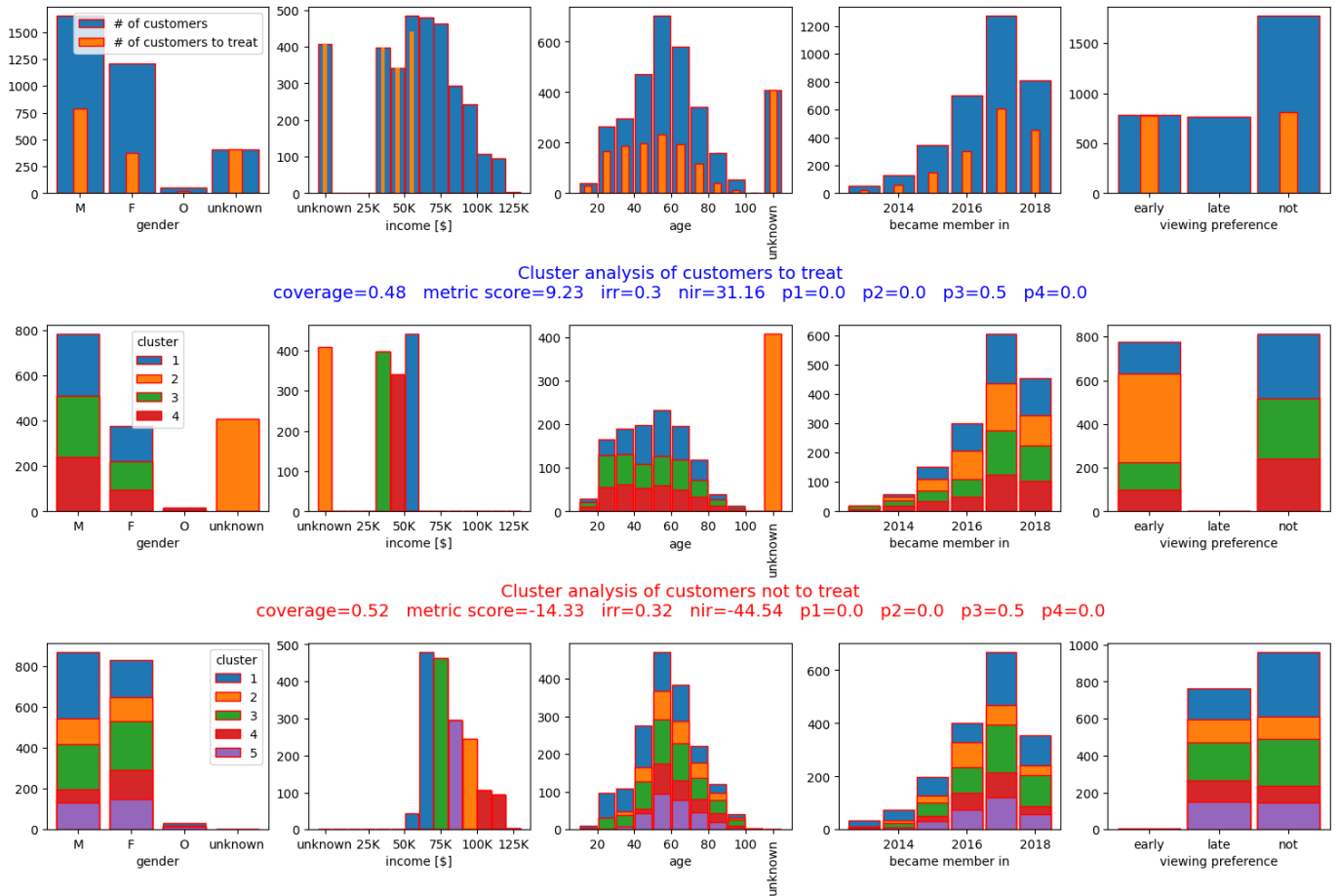


Figure 10: Customers to be addressed are basically the same as in offer type 6. However, this offer type has a stronger value in the metric score.

Customer treatment evaluation for the offer No. 8 with:
offer_type=discount reward=2 difficulty=10 duration=7 channels=['web', 'email', 'mobile']

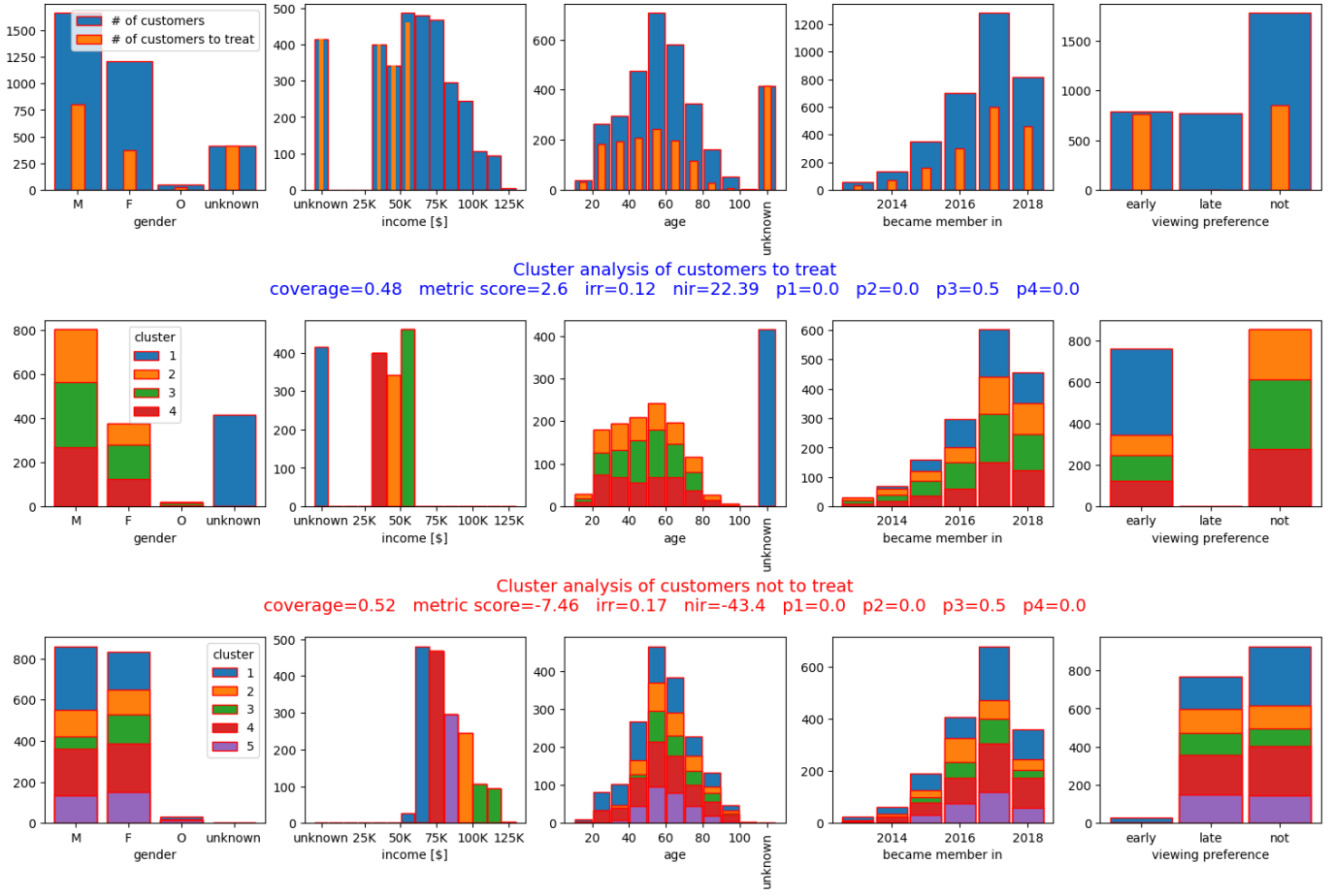


Figure 11: Customers to be addressed are basically the same as in offer type 6. However, this offer type has a weak value in the scoring metric.

5.2.3 Diagrams for "informational" offers

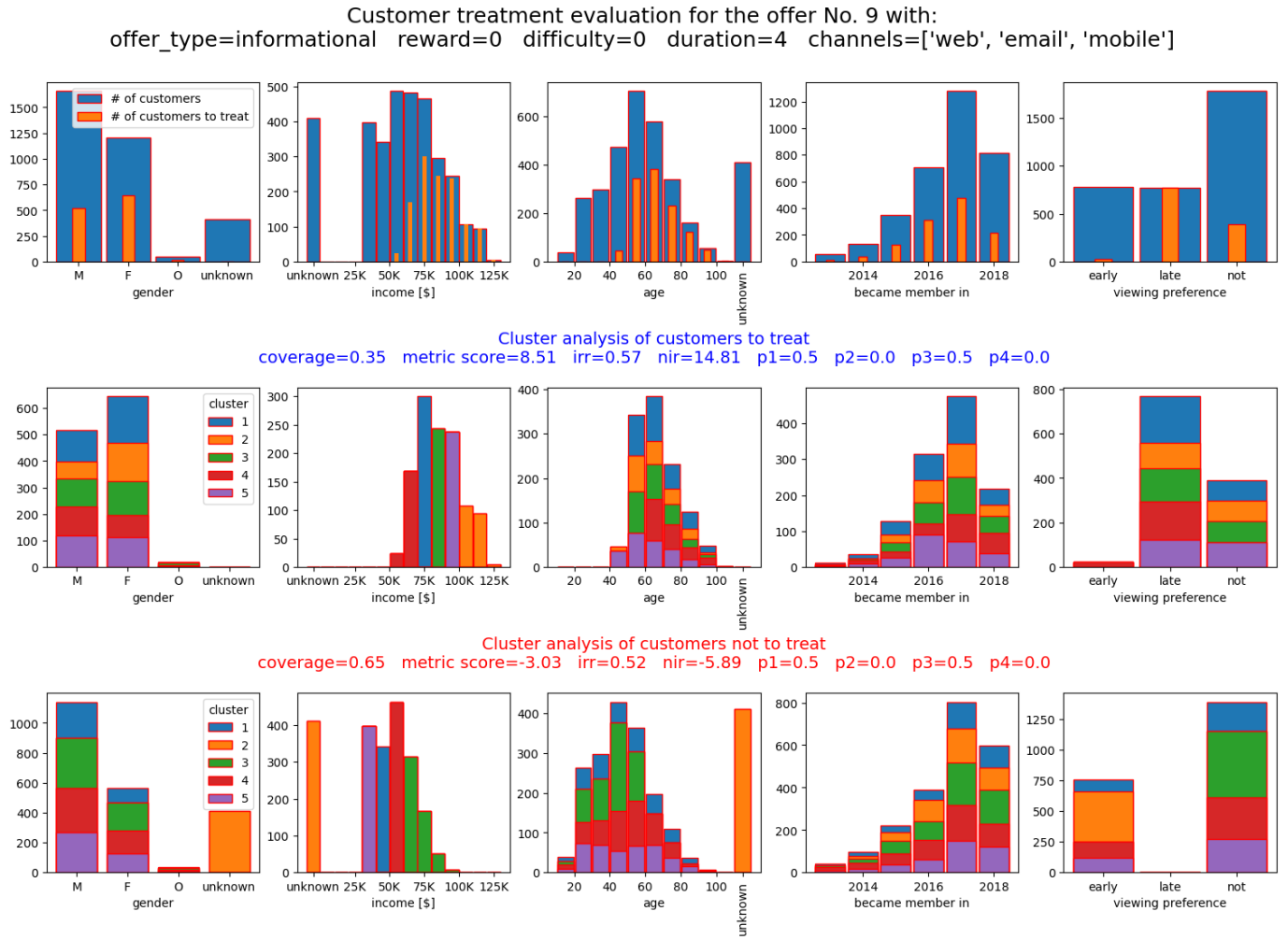


Figure 12: All customers earning at least 60 k dollars should be addressed by this offer. These customers tend to view lately or rather not. Customers not providing demographic information should be excluded.

Customer treatment evaluation for the offer No. 10 with:
offer_type=informational reward=0 difficulty=0 duration=3 channels=['email', 'mobile', 'social']

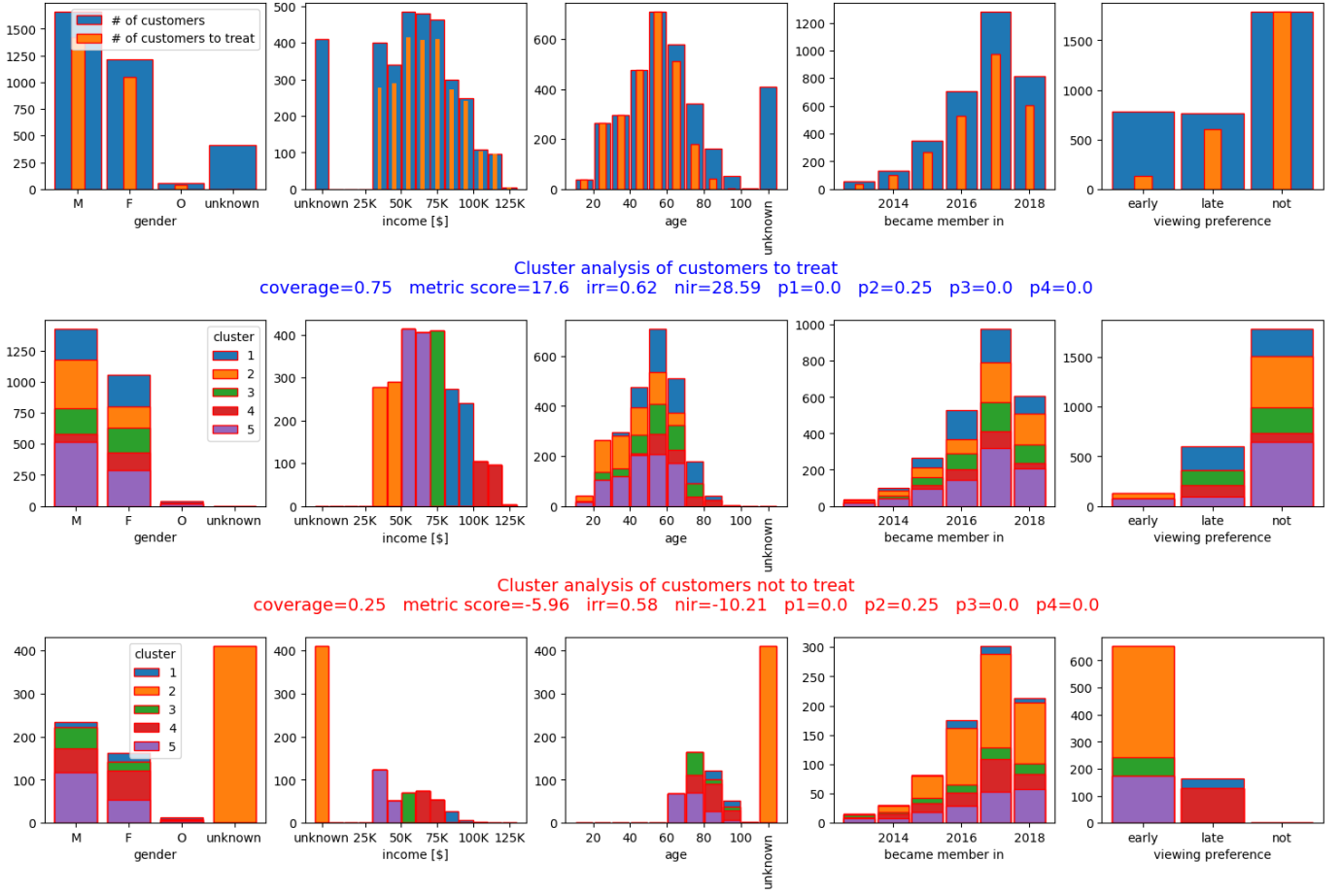


Figure 13: Primarily young and mid-aged customers are worth being addressed by this offer. These customers tend to view lately or rather not. Customers not providing demographic information should be excluded. From those earning 70 k dollars or less, only those who tend to view lately are within the target group.

6 Conclusion

BOGO and discount offers are primarily suitable for low and mid-income customers. The best BOGO offer is offer type No. 4 and designed for customers earning up to 60 k dollar. It has a 5-day vailidity period and a 5 dollar difficulty to overcome. The best discount offer is offer type No. 7 possessing a 7-day vailidity period, a 10 dollar difficulty and a 2 dollar reward. Also this offer type is designed for customers earning up to 60 k dollar but has a metric score being twice compared to the best BOGO one. Nonetheless, both money-saving offers, BOGO and discount, seem not designed for customers in the higher income region. With respect to the informational offers, however, we can recognize that these are primarily suitable for customers who earn a lot and do not tend to view offers immediately. Customers who do not provide demographic information are, contrary to the other offer types, not included in the target group. It might be that people who do not provide information about themselves wish not to be informed by the app without having a direct benefit. There might be also a tendency that addressing offers through social networks may increase the metric score.