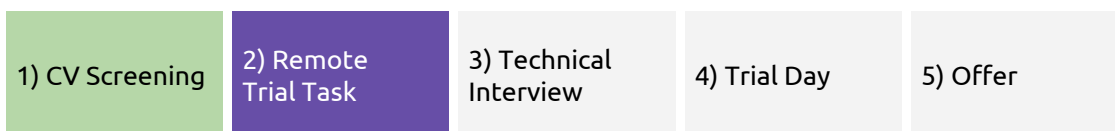


SO1 Remote Trial Task

Recruiting process

Thank you again for applying to SO1. As a reminder, the recruiting process for a ML position at SO1 consists of five consecutive steps:



We really liked your CV so in the next three steps (starting with this “remote task”) we want to assess whether you have the necessary skills to be successful at SO1:

1. theoretical knowledge on statistics, probability theory, and machine learning
2. experience in applying ML algorithms to real-world problems
3. good coding skills

The **remote trial task** below consists of two subtasks. In the first task you need to develop a formal probabilistic description of consumer behavior. In the second task we want you to implement a model to predict future consumer purchases. We are looking forward to your solution!

If you pass the remote task, the next step is a **technical interview**. During the interview we will discuss questions on machine learning (theory, methods, and applications), probability theory, and linear algebra. We will also ask a couple of follow-up questions on the remote task.

If you pass the technical interview stage as well, the final step in our recruiting process will be an onsite or remote **trial day** (based on where you are located). You will work closely with our ML team at the SO1 office and either spend more time on part 2 of the trial day tasks or implement a simple agent for individualised coupons within our “supermarket gym”—SO1’s homemade simulated environment for testing and developing reinforcement learning agents.

Task 1: Formalisation of Consumer Shopping Behaviour

A consumer i shops in week t at a retailer r . The retailer carries several categories c . In each category c , the retailer offers multiple products j . The consumers decision process, that is the probability that consumer i purchases q_{ijt} units of product j in week t , can be modeled by

$$P(Q_{ijt} = q_{ijt}) = P(I_{ict} = 1) P(C_{it} = j | I_{ict} = 1) P(Q_{ijt} | I_{ict} = 1 \wedge C_{it} = j) \quad (1)$$

where $P(I_{ict} = 1)$ is the probability of i purchasing in c , $P(C_{it} = j | I_{ict} = 1)$ is the probability that i purchases j conditioned on purchasing in category c , and $P(Q_{ijt} | I_{ict} = 1 \wedge C_{it} = j)$ is the probability that i purchases q_{ijt} units of j in t conditioned on purchasing product j (in category c). Whether the consumer buys in a category or not is independent from her purchase decisions in other categories. If she purchases in a given category, she picks exactly one product j . The probability of choosing product j in category c depends on (1) her (time-invariant) product preferences, (2) the products' prices in week t , and (3) whether product j was advertised or not (a boolean indicator). Of the chosen product j , the consumer buys n units (n is a positive integer).

Your task:

Which probability distributions would you use to model each of the three components: probability of category purchase $P(I_{ict} = 1)$, the probability of product choice within a category $P(C_{it} = j | I_{ict} = 1)$, and the probability for the purchased amount $P(Q_{ijt} | I_{ict} = 1 \wedge C_{it} = j)$? In your opinion, what drives the category purchase incidence? What factors besides the three mentioned above could influence product choice?

Task 2: Predict Consumer Purchases

The data set attached to this task contains the purchase histories (i.e., shopping baskets) of 2,000 consumers over 49 weeks across 5 categories (*train.csv*). In simulating the basket data we assumed that consumers only buy one unit of a product in a given week. The data set also contains the price consumers paid for one unit of product j in week t and a boolean variable that indicates whether the purchased product was advertised (1) or not (0). We also provide the week 50 promotion schedule (discounts and advertising) for all products (*promotion_schedule.csv*).

Your task:

Use the data to build a ML model for consumer purchases. With the trained model, predict week 50 purchases for all 80,000 possible consumer-product combinations (40 products x 2,000 consumers) in the data. Feel free to use any non-parametric or "black box" model you consider appropriate (i.e., you don't have to follow the formalisation from part 1). Please provide your predictions as a .csv file (cf. *prediction_example.csv*) that contains the columns i (consumer), j (product), and prediction. We will benchmark your predictions against observed purchases using the AUC metric. *Please indicate your expected AUC score.*