SIOP Machine Learning Competition 2020

About the Data

This data set was provided by the Global Selection and Assessment Team at Walmart.

These data were collected as part of a validation study of a pre-employment test for entry-level positions in retail stores.

The original dataset was modified for this Machine Learning competition in the following way:

- All personal information and demographics were removed
- A new, contrived variable was created called “Protected_Group.” This variable is to be used a surrogate for ethnicity, gender, or other protected class as part of the “fairness” analysis. This variable was modeled to have substantial (but realistic) group differences for both predictor and criterion variables
- All of the item descriptions and response text was removed. Only an *example* of each type of item is provided, but none of the actual assessment content is revealed
- Items were reordered in the data file so there is no relationship between the ordering of items in the actual assessment form and their order in this data file
- A very small percentage of statistical outliers was removed from the dataset.

Otherwise, the data exists and behaves as it did in the original validation study. The predictor variables are data from actual assessment items, and the criterion variables are the actual performance ratings. We have included a variable called “Retained,” which is a binary indicator of whether the employee was still employed a certain number of days after hire. All of the relationships among the variables and the magnitude of the effects remain intact from the original study.

This dataset contains *N* hires made in a group of retail stores over a period of several months. Special *validation-only* performance ratings were collected for a sample *(N=)* of these hires.

To be included in the performance rating sample, employees had to have sufficient tenure so that raters observed employee performance long enough to make an accurate assessment. For this reason, no performance data exists for employees who exited the organization within their first few weeks.

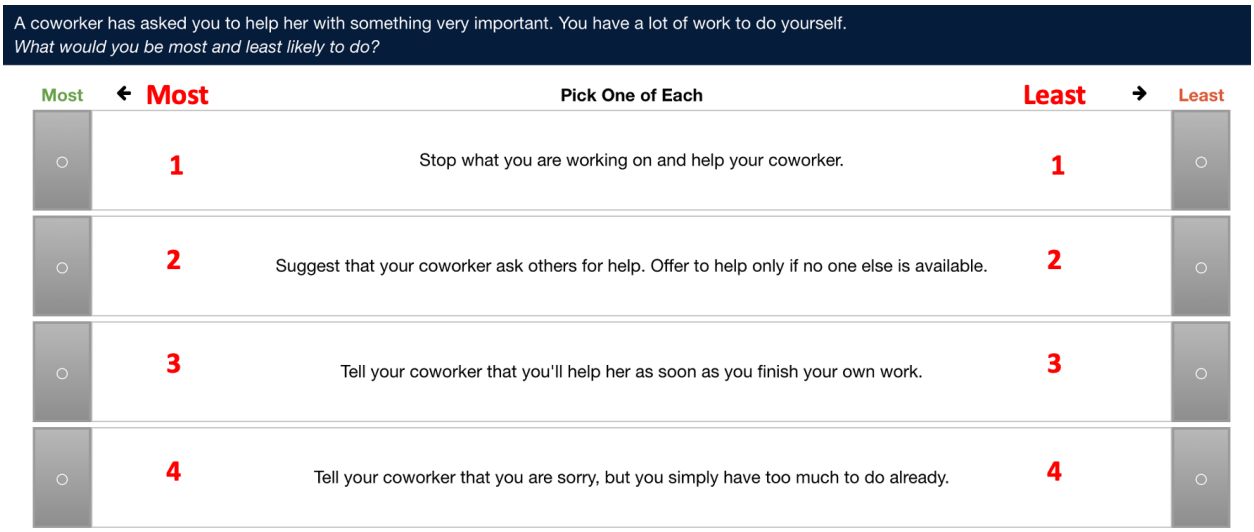
We included data for the entire cohort because part of the task is to create a predictor of *retention*; to do this effectively requires data for all exits, not just for employees who were retained long enough to have their performance evaluated.

Variables in Data File:

Variable	Description	Notes
UNIQUE_ID	Unique Individual Identifier	
Criterion Variables		
Overall_Performance	Overall Performance Rating	1-5 Scale (higher=more favorable)
Technical_Skills	Technical Skills Rating	1-5 Scale (higher=more favorable)
Teamwork	Teamwork Rating	1-5 Scale (higher=more favorable)
Customer_Service	Customer Service Rating	1-5 Scale (higher=more favorable)
Hire_Again	Would you hire this employee again?	1=Definitely would not 2=Probably would not 3=Not sure 4=Probably would 5=Definitely would
High_Performer	Was employee a “high performer?”	1 = High 0 = Not High
Retained	Was employee retained for a period of n days or did he/she exit prior to that time?	1 = Retained 0 = Termed
“Fairness” Variable		
Protected_Group	This is an artificially contrived variable intended to be used surrogate for protected class variable, such as gender, ethnicity, or age. This variable is to be used in any fairness/adverse impact analyses.	1 = Protected Group 0 = Not

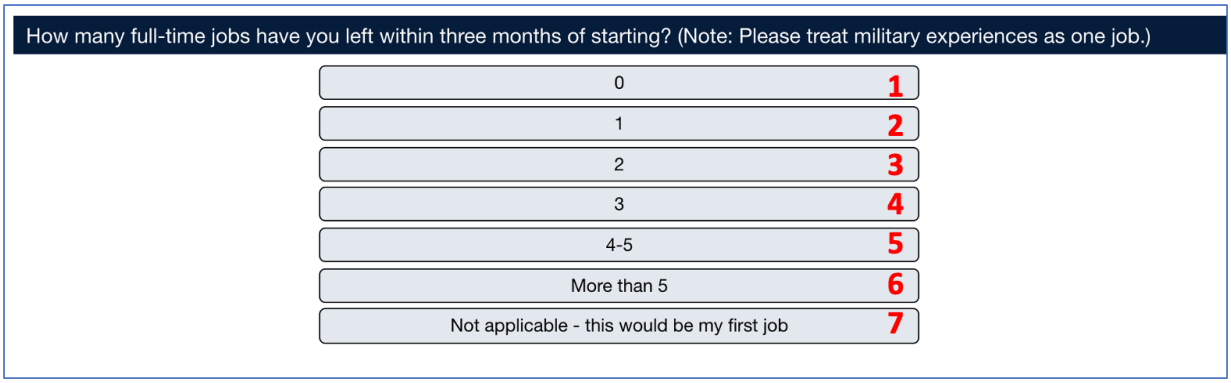
Predictor Variables		
Situational Judgment Items (SJ_Most_1 through SJ_Time_9)	There are nine of these items. Each Item produces three variables. The first variable indicates which of the actions the candidate was most likely to take. The second variable indicates the action the candidate was least likely to take, and the third item indicates the time (in seconds) the candidate took to answer the question.	Items are coded 1 to 4 for both "Most" and "Least" likely. The "Time" variable is in seconds.

Situational Judgment Item Example:



Scenario Interpretation (Scenario1_1 through Scenario2_Time)	Candidates view a scenario, then have to respond with the number of times that each of the response options occurred during the scenario. There are 8 options per scenario (e.g., Scenario_1 through Scenario_8) followed by a variable (e.g., Scenario1_Time) indicating the time (in seconds) the candidate took to answer the question. There are two of these scenarios.	The values in Scenario1_1 thru Scenario1_8 and Scenario1_1 thru Scenario2_8 can range between 0 and 9. Values for Scenario1_Time and Scenario2_Time represent time taken in seconds.
Biodata/Work History Items (Biodata_01 thru Biodata_20)	Candidates read an item and select a single response.	The number of response options for these items varies between 5 and 8.

Biodata/Work History Item Example:



Personality/Work Style Items (PScale01_Q1 thru PScale13_Q5)	These items have a bi-polar response format as shown below. Anchors appear above and below the item and the respondent chooses a single response. The items are grouped and labeled by subscale. For example, items PScale01_Q1 thru PScale01_Q4 are four items intended to measure the same personality-related subscale. There are 13 of these subscales with varying numbers of items per subscale	These items are coded 1 to 4 as shown in the example below.

Personality/Work Style Item Example: