

JOEL SHAPIRO

Predicting Employee Attrition

You have at your disposal a data set of 1470 employees from an anonymous company. For each employee, there is an outcome measure called “Attrition” and a handful of other individual level descriptors. Your job is to create a statistical model that best predicts employee attrition. You must use logistic regression only!

These employees have been split into a training set of 1170 individuals and a validation set of 300 individuals, 52 of whom left the company (indicated by Attrition = yes). Your model should, as accurately as possible, predict attrition for the 300 members of the validation set.

To do, please follow the following steps:

Step 1. Create a predictive model with “attrition” as your outcome, and include whatever explanatory variables (and combinations thereof) you want. Do this ONLY for the 1170 employees who are designated as “training=1.”

Step 2. Use this model to predict the likelihood of attrition for the 300 person validation set, for whom “training=0.”

Step 3. Separate these 300 into deciles based on your predicted probabilities of them leaving, and find the total number in each decile who actually left the company. You may find the following helpful:

Step 4. When you have the total number of actual attritions for each decile, report those decile totals to me. In addition, one group member must email me any and all code so that I can replicate your predictive model. (Make sure to include any code you may have used to generate new variables.)

Results: The team with greatest cumulative total of actual attrition from decile 10 (highest predicted attrition) to 1 (lowest predicted attrition) will be declared the winner.

****Important note 1: Make sure you are clear which decile is the “top” and which is the “bottom.” I will ask you to report your results to me as the sum for the decile with the highest predicted attrition, then the second highest decile, and so on.*

****Important note 2: You may create and test more than one model if you’d like, but – for this exercise – you may not change the composition of the training or validation sets. Also, you may only give me one set of decile totals. Once you have reported it to me, your submission is final.*
