Model Validation

We will begin by populating the Logistic Regression Template worksheet with the data from the validation dataset. To start, be sure that you have already created the same recoded variables on the Validation Data sheet as you did on the Calibration Data sheet. Next, in L2, reference the recoded homeownership status variable from Validation Data sheet by entering the formula:

='Validation Data'!R2

Copy this formula down the worksheet. Next, in cell M2, we will reference the recoded loan status variable from the Validation data sheet:

='Validation Data'!U2

Copy this formula down the worksheet. In cell N2, we will calculate the probability that the loan is considered risky using the results of the logistic regression. Enter the following formula:

=EXP($I$1+$I$2*L2)/(1+EXP($I$1+$I$2*L2))

Copy this formula down the worksheet. To calculate the error, we will first calculate a measure of error based on the naïve assumption that all loans have the same likelihood of being risky. In this scenario, we will use the proportion of loans that are risky in our calibration dataset as our naïve estimate. We will calculate the error as |Observed-Predicted|. To calculate the absolute error associated with using this estimate, in cell O2, enter the formula:

=ABS(M2-AVERAGE($C$2:$C$50001))

In cell P2, we will calculate the absolute error using the results of our logistic regression:

=ABS(M2-N2)

Copy the formulas in O2 and P2 down the worksheet.

To calculate the mean absolute error under these models, we will average over the errors associated with individual loans. In cell S1, the error can be calculated as:

=AVERAGE(O2:O10001)

Similarly, in S2, the error from our model can be calculated as:

=AVERAGE(P2:P10001)

In cell S3, the percentage improvement from using our model can be calculated as:

=1-(S2/S1)