**Data to Decision**

Spring 2019

Daniel Egger

Problem Due Monday, February 11

Nathan Scheperle

Nes31

Note: students may work in groups on the problem but are responsible for submitting their own answers. Type answers directly in this word document, rename the file with your name, and upload to Sakai Dropbox before noon on Monday.

**Applying the Threshold Model and Forecasting Future Error Rates**

Computer chips that exceed certain size tolerances on the assembly line tend to be defective. You are given size measures for each chip, wirth binary outcomes (defective/not defective). You are given 400 typical cases (drawn at random from a stable production process).

The cost per FP classification is estimated at $50-100 and the cost per FN classification is $150-300.

Partition your 400 cases into a *training set* and *test set*, based on ensuring that with 99% confidence the error rate on future data will not exceed the error rate on test data by .10. Partition the data by assigning individual cases at random to the two sets in your desired proportions.

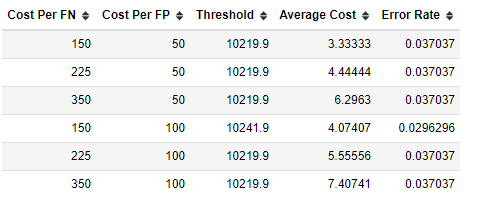
1. How large is your test set? Why?

Using Hoeffding's Inequality, the required number of observations in the test dataset to ensure an error rate below .10.10 at the 99%99% confidence level is:

𝑛≥𝑙𝑜𝑔(2/𝛼)2𝑡2=𝑙𝑜𝑔(2/.01)2⋅.12=𝑙𝑜𝑔(200).02≈265

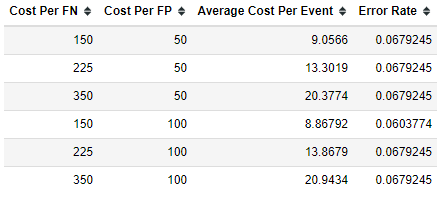
2. Evaluate the threshold model on your *training data only*, and identify six parameters (thresholds) that minimize the average per-chip cost of misclassification at each of six combinations of costs.

(A) Record these parameters. Also give (B) the average cost per event and (C) the error rate at each threshold on training set.

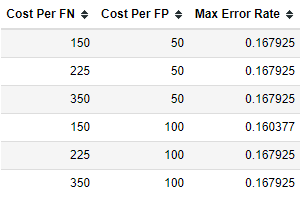


3. Evaluate the six thresholds derived from training data on the *test data* and record (B) the average cost per event and (C) the error rate at each threshold.

|  |
| --- |
|  |

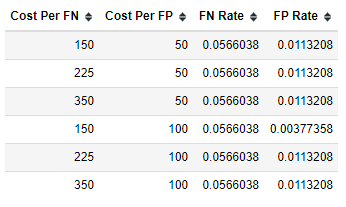


4. With 99% reliability, what is the expected maximum error rate on new data at each of the six thresholds?

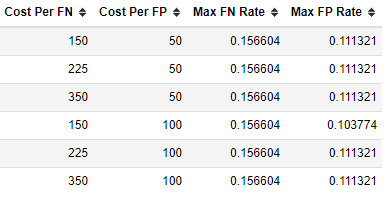


Note that the Hoeffding inequality can be used not only to generalize the overall error rate from test data results, but also the separate overall portion of False Positives (*not* the FP rate, g/b, but the portion g/(a+b)) and the overall portion of False Negatives (not the FN rate, f/a, but the portion of False Negatives f/(a+b)).

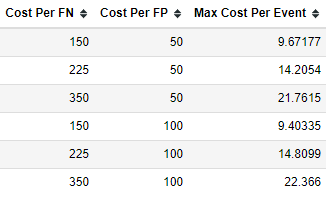
5. Calculate these portions for each of the six parameters on the test data.



6. With 99% confidence, what is the maximum portion of False Positives, and the maximum portion of False Negatives, that will be observed at each of the six thresholds?



7. With 99% confidence, what is the maximum cost per event that will be observed at each of the six thresholds?



**END**