Topic 10. Classification, Oversampling, and Lift

Case 3: Donor Recapture

using Transaction, Overlay, and Census Data

Reading Assignment

Berry and Linoff (2000)

- Pages 196–201. Oversampling.

We Were Warned

Remember that the project description, cty_doc.txt, warned us that attempts to use classification to predict donor giving would fail because donors who have a high probability of responding give small gifts.

Plan

First we review how lift charts are constructed, then we will look at a variety of estimation strategies involving classification and over sampling.

The presentation is by means of a series of lift charts.

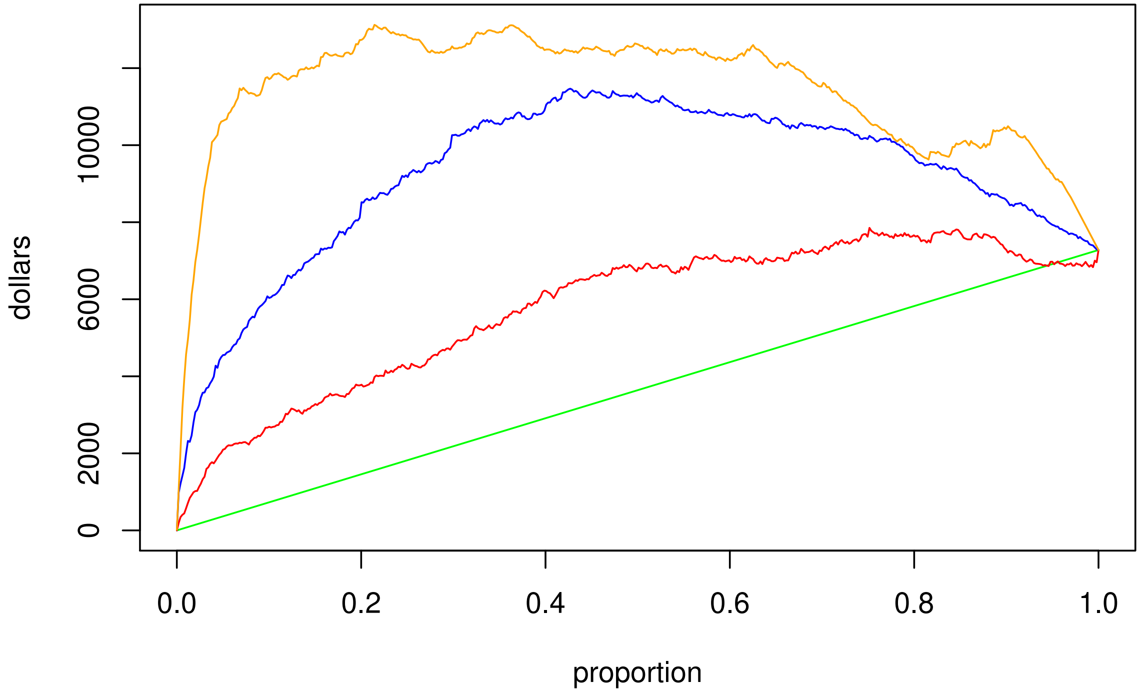
The titles and the footnotes of the charts tell the whole story.

Lift Charts

A lift chart for n cases is constructed as follows:

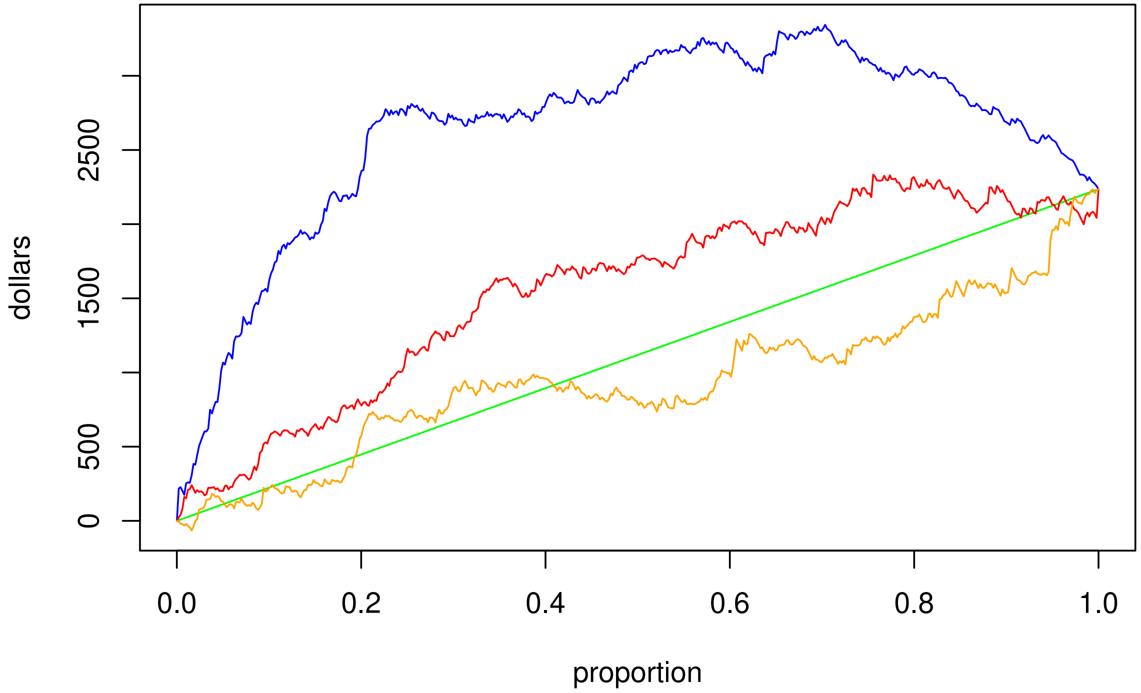
- In a column labeled y put the values of the target.
  - For a predictive model, y will be numeric.
  - For a binary classification model, y will be 0's and 1's.
  - A different target such as revenue can be substituted for either.
- In a column labeled $\hat{y}$ put the predicted target.
  - For a predictive model, $\hat{y}$ is numeric.
  - For a classification model, $\hat{y}$ is the conditional probabilities (confidence scores).
- Sort both columns, y and $\hat{y}$, on $\hat{y}$ in descending order.
- Plot the cumulative sum of y against $1/n, 2/n, ..., n/n = 1$ or against these values times 100.

Fig 105. Learning Set Conditional Probability Lift Chart, No Oversampling



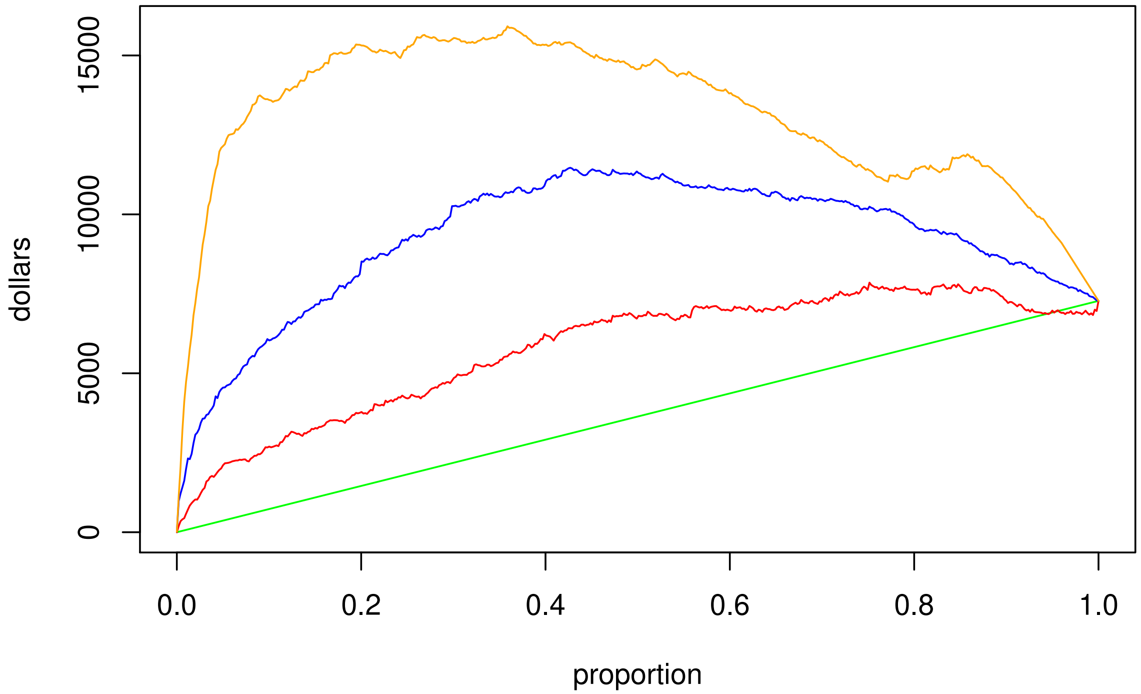
The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a classification regression and the yellow the same for a classification tree.

Fig 106. Validation Set Conditional Probability Lift Chart, No Oversampling



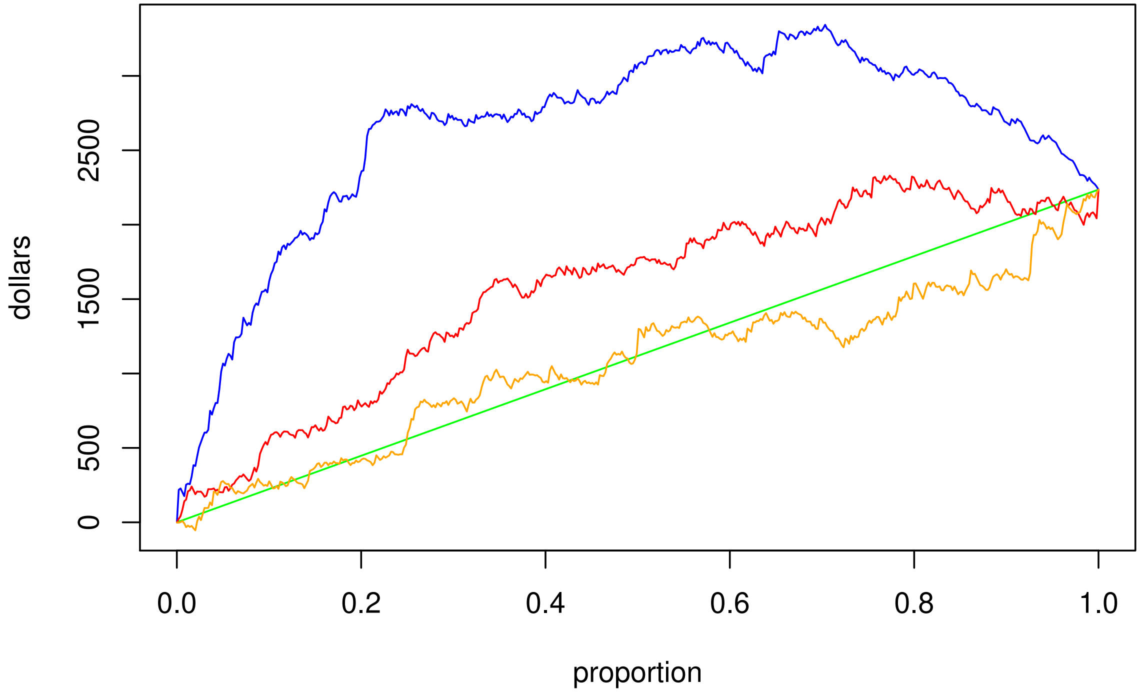
The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a classification regression and the yellow the same for a classification tree.

Fig 107. Learning Set Conditional Probability Lift Chart, 10% Oversampling



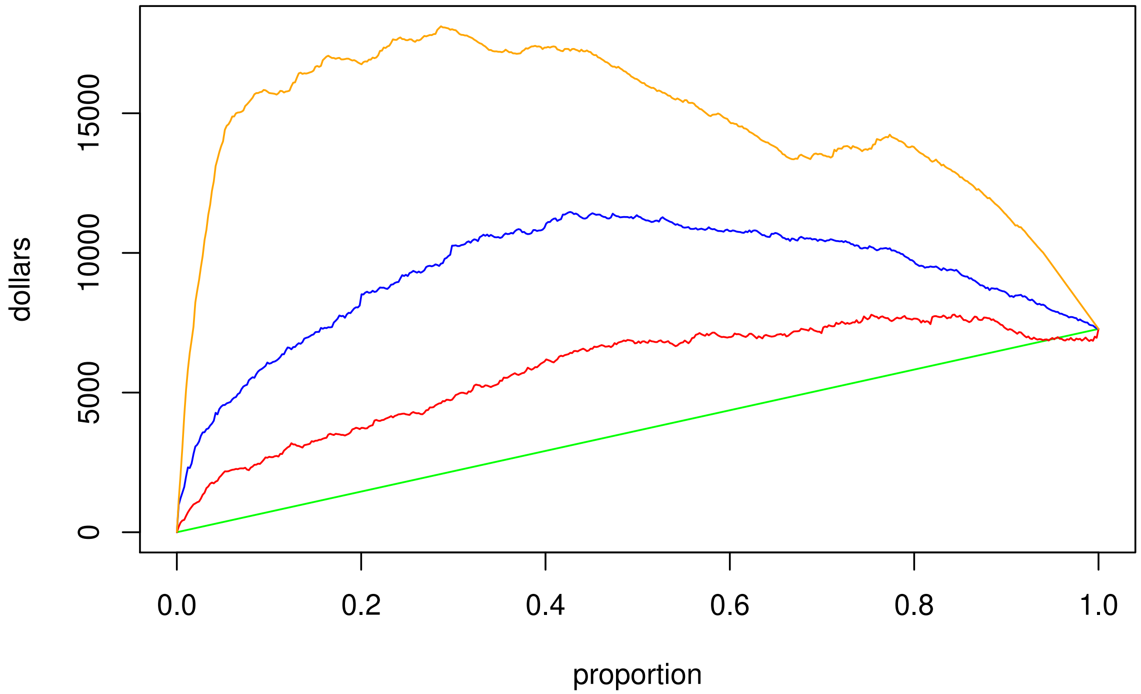
The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a oversampled classification regression and the yellow the same for a oversampled classification tree.

Fig 108. Validation Set Conditional Probability Lift Chart, 10% Oversampling



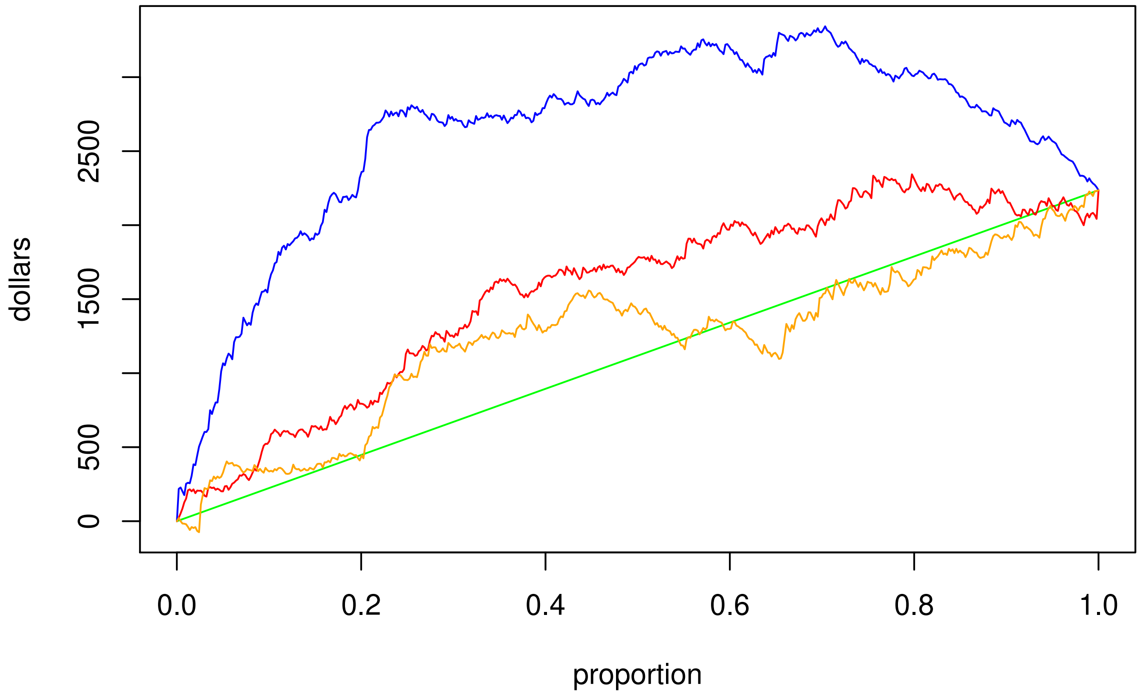
The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a oversampled classification regression and the yellow the same for a oversampled classification tree.

Fig 109. Learning Set Conditional Probability Lift Chart, 30% Oversampling



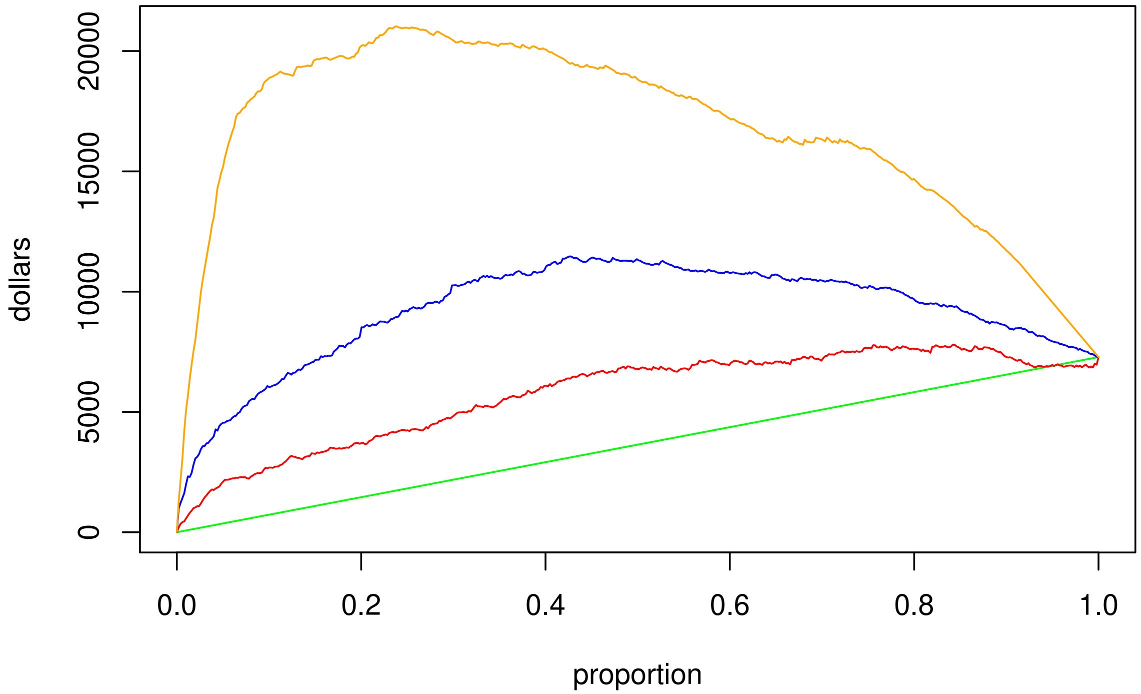
The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a oversampled classification regression and the yellow the same for a oversampled classification tree.

Fig 110. Validation Set Conditional Probability Lift Chart, 30% Oversampling



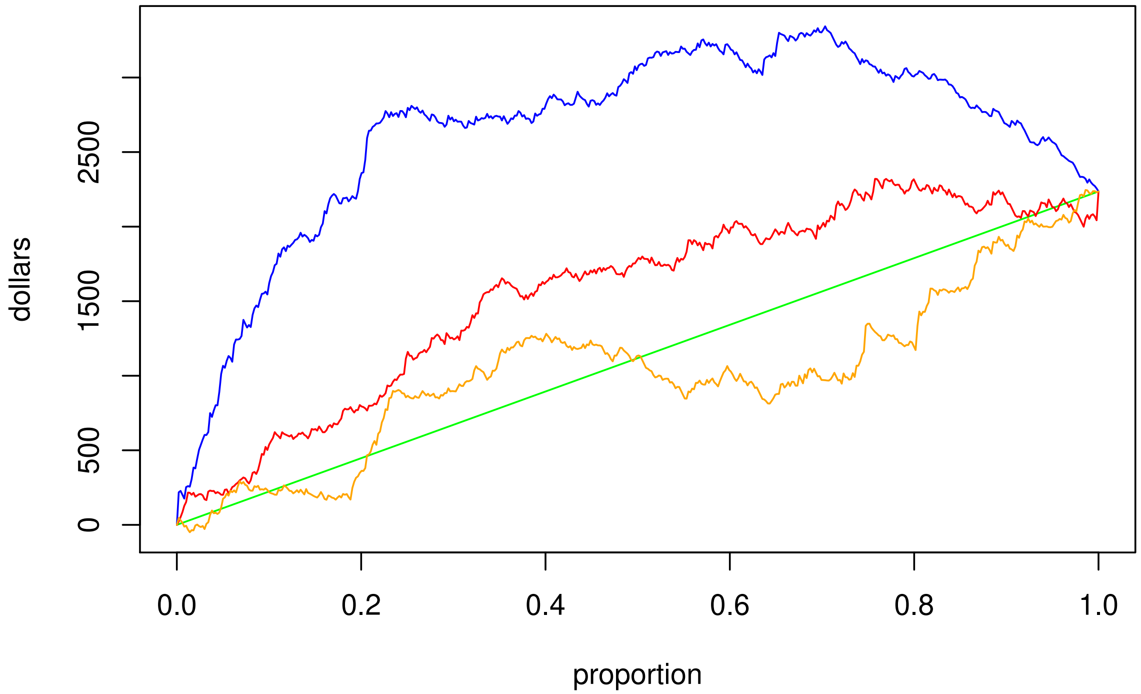
The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a oversampled classification regression and the yellow the same for a oversampled classification tree.

Fig 111. Learning Set Conditional Probability Lift Chart, 50% Oversampling



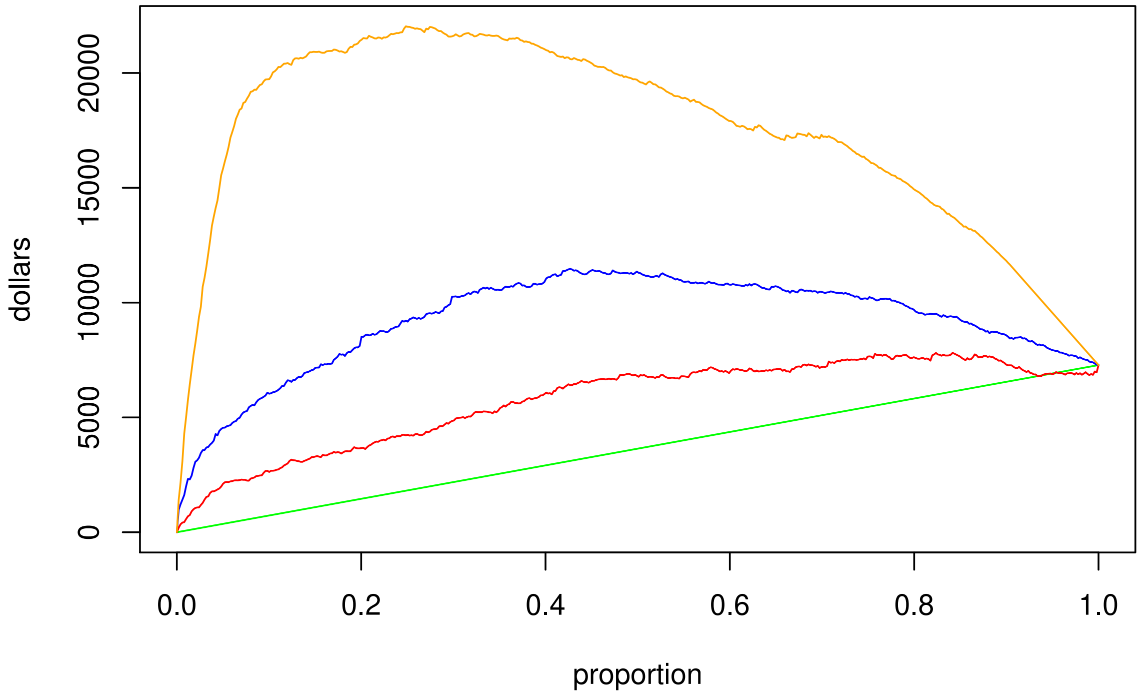
The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a oversampled classification regression and the yellow the same for a oversampled classification tree.

Fig 112. Validation Set Conditional Probability Lift Chart, 50% Oversampling



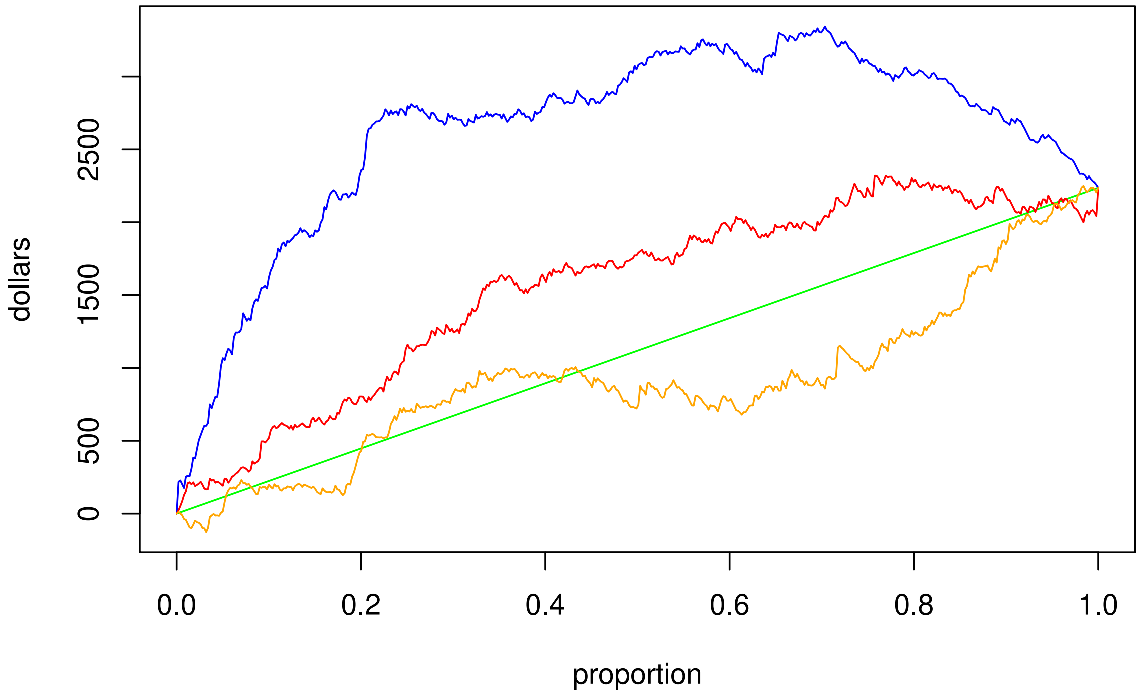
The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a oversampled classification regression and the yellow the same for a oversampled classification tree.

Fig 113. Learning Set Conditional Probability Lift Chart, 80% Oversampling



The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a oversampled classification regression and the yellow the same for a oversampled classification tree.

Fig 114. Validation Set Conditional Probability Lift Chart, 80% Oversampling



The green curve shows expected net revenue if persons were mailed solicitations in random order. The blue curve shows net expected revenue if persons are sorted according to a predictive regression. The red curve is the same for a oversampled classification regression and the yellow the same for a oversampled classification tree.

Blank page