Modeling Rare Events
Over/Undersampling, Priors, Decision Weights
Undersampling/Oversampling and Prior Probabilities

Can be accounted for automatically in SAS EM
Undersampling and Prior Probabilities

- Say you have a rare event as target (<10% of data)
  - Fraud
  - Catastrophic failure
  - 10%+ single day change in value of stock market index

- May have trouble modelling because a model is accurate for classifying everything as nonevent!

- Potential Solution: Create a biased sample
Undersampling and Prior Probabilities

- Potential Solution: Create a biased sample
  - **Undersample:** under-represent common events in training data.
    - Keep all rare events and only a fraction of common events
    - Ratio of Common:Rare events is up for debate.
      - 70:30 ought to be fine.
      - 50:50 is sometimes encouraged.
  - **Oversample:**
    - replicate the rare events in training.
    - do this *after* the training/validation split so don’t have the same observation in both training and validation set!
    - OR, use a hybrid technique like **SMOTE** (Chawla, 2002) that creates new data points *like* the rare events (not exact replicates)
Undersampling and Prior Probabilities

- Models provide **posterior probabilities** for events.

- The accuracy of the posterior probabilities rely on a representative sample.

- If we bias our sample, must adjust the posterior probabilities to account for this.
Why Adjustment is Necessary

Predict voting machine failure. Only 100 voting machines failed out of 10,000.

Undersample. Dataset has 100 failures and 100 non-failures.

- Failures: 100
  - Last inspection date ≤3 years: Failures: 10, Non-failures: 90 (p=0.1)
  - Last inspection date >3 years: Failures: 90, Non-failures: 10 (p=0.9)
Why Adjustment is Necessary

Does a new machine with last inspection date >3 years really have a 90% probability of failing?
Why Adjustment is Necessary

- We’d have to go back to the data to answer this question.
- Assuming the 100 non-failures chosen were random, representative sample, we expect inspection date to be ≤ 3 years 90% of the time.
- That is 8,910 non-failing machines with inspection date ≤ 3 years.
- Similarly, 10% of non-failures have expect inspection date >3 years ago. This is 990 machines.

<table>
<thead>
<tr>
<th></th>
<th>≤ 3 years</th>
<th>&gt;3 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failures</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>Nonfailures</td>
<td>8910</td>
<td>990</td>
</tr>
</tbody>
</table>

\[
P(\text{Failure} \mid \text{last inspection date} > 3 \text{ years}) = \frac{90}{90+990} = 8\%
\]

(Still failing at 8 times the rate of recently inspected machines)
Summary: Adjusting for Undersampling

- Let \( l = l_1, l_2, \ldots, l_L \) be the levels of the target variable
- Let \( i = 1, 2, \ldots, n \) index the observations in the data
- Let \( \text{OldPost}(i, l) \) be the posterior probability from the model on oversampled data
- Let \( \text{OldPrior}(l) \) be the proportion of target level in the oversampled data
- Let \( \text{Prior}(l) \) be the correct proportion of target level in true population

\[
\text{NewPost}(i, l) = \frac{\text{OldPost}(i, l) \frac{\text{Prior}(l)}{\text{OldPrior}(l)}}{\sum_{j=1}^{L} \text{OldPost}(i, l_j) \frac{\text{Prior}(l_j)}{\text{OldPrior}(l_j)}}
\]
Entering Priors and Decision Weights into SAS EM
Entering Priors into SAS EM

- Priors are also adjusted in the “decisions” on a dataset panel.
- Click “Build” when first opening the prompt, then click priors tab.
Only Some Output Uses the Prior Information

- In SAS EM, accounting for priors **has no effect on**:
  - Growing decision trees
  - Misclassification Rate (The cutoff probability is still 0.5 by default)

- **Priors do affect**:
  - *Pruning* decision trees
  - Once we account for a prior, a given split may not have a reasonable gain

- **Net Effects**:
  - Increasing a prior probability increases the posterior probability
  - Decreasing a prior decreases the posterior probability
  - Changing prior will have more noticeable effect if the original posterior is near 0.5 than if it is near 0 or 1.
Oversampled Data with No Priors

Predicted Probabilities come from the decision tree as expected
Predicted Probabilities are adjusted according to the priors. The tree is pruned according to those adjustments too. Default cutoff probability is still 0.5!
### Oversampled Data with Priors

<table>
<thead>
<tr>
<th>Predicted: Creditability=good</th>
<th>Predicted: Creditability=bad</th>
<th>Unadjusted P: Creditability=good</th>
<th>Unadjusted P: Creditability=bad</th>
<th>Residual: Creditability=good</th>
<th>Residual: Creditability=bad</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4892703626609</td>
<td>0.5107296137339</td>
<td>0.10526315789473</td>
<td>0.89473684210526</td>
<td>-0.10526315789473</td>
<td>0.105263157894739997</td>
<td>BAD</td>
</tr>
<tr>
<td>0.4892703626609</td>
<td>0.5107296137339</td>
<td>0.10526315789473</td>
<td>0.89473684210526</td>
<td>-0.10526315789473</td>
<td>0.105263157894739997</td>
<td>BAD</td>
</tr>
<tr>
<td>0.4892703626609</td>
<td>0.5107296137339</td>
<td>0.10526315789473</td>
<td>0.89473684210526</td>
<td>-0.10526315789473</td>
<td>0.105263157894739997</td>
<td>BAD</td>
</tr>
<tr>
<td>0.4892703626609</td>
<td>0.5107296137339</td>
<td>0.10526315789473</td>
<td>0.89473684210526</td>
<td>-0.10526315789473</td>
<td>0.105263157894739997</td>
<td>BAD</td>
</tr>
<tr>
<td>0.4892703626609</td>
<td>0.5107296137339</td>
<td>0.10526315789473</td>
<td>0.89473684210526</td>
<td>-0.10526315789473</td>
<td>0.105263157894739997</td>
<td>BAD</td>
</tr>
<tr>
<td>0.4892703626609</td>
<td>0.5107296137339</td>
<td>0.10526315789473</td>
<td>0.89473684210526</td>
<td>-0.10526315789473</td>
<td>0.105263157894739997</td>
<td>BAD</td>
</tr>
<tr>
<td>0.4892703626609</td>
<td>0.5107296137339</td>
<td>0.10526315789473</td>
<td>0.89473684210526</td>
<td>-0.10526315789473</td>
<td>0.105263157894739997</td>
<td>BAD</td>
</tr>
<tr>
<td>0.4892703626609</td>
<td>0.5107296137339</td>
<td>0.10526315789473</td>
<td>0.89473684210526</td>
<td>-0.10526315789473</td>
<td>0.105263157894739997</td>
<td>BAD</td>
</tr>
</tbody>
</table>

**Default cutoff probability is still 0.5!**
Oversampled Data with No Priors
Oversampled Data with Priors

Balance of current acc...

NO RUNNING ACCOU...

>=200 DM, <=200 DM

Duration in months

< 22.5 Or Missing

>= 22.5

Payment of previous cr...

HESISTANT PAYMEM... NO PROBLEMS WITH...

Telephone

YES

NO Or Missing
To emphasize rare events in a modelling context, we may want to increase the “profit” of making a correct prediction of the rare event.

The easiest way to do this is to weight the decisions with a profit (or cost – make errors negative) matrix:

- Prior: RareEvent = 0.02, CommonEvent = 0.98

<table>
<thead>
<tr>
<th>Decision Weights to emphasize correct classification of rare events</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>RareEvent</td>
</tr>
<tr>
<td></td>
<td>1/0.02 = 50</td>
</tr>
<tr>
<td></td>
<td>CommonEvent</td>
</tr>
</tbody>
</table>
Creating Inv. Prior Decision Weights

If first time, click “Build”
Creating Inv. Prior Decision Weights

![Decision Processing - CREDIBILITY_priors_invDecisionWeights](image)

- **Decisions Tab**: The 'Decisions' tab is selected.
- **Default with Inverse Prior Weights**: The option 'Default with Inverse Prior Weights' is highlighted.
- **Table Content**:
  - **Decision Name**: DECISION1, DECISION2
  - **Label**: GOOD, BAD
  - **Cost Variable**: < None >
  - **Constant**: 0.0

Options available at the bottom:
- **Add**
- **Delete**
- **Delete All**
- **Reset**
- **Default**
- **OK**
- **Cancel**
Creating Inv. Prior Decision Weights

Select a decision function:
- Maximize
- Minimize

Enter weight values for the decisions:

<table>
<thead>
<tr>
<th>Level</th>
<th>DECISION1</th>
<th>DECISION2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOOD</td>
<td>1.05263157</td>
<td>0.0</td>
</tr>
<tr>
<td>BAD</td>
<td>0.0</td>
<td>20.0</td>
</tr>
</tbody>
</table>
Oversampled Data with Priors and Decision Weights

Decision Tree:
- Balance of current account:
  - 0: BAD (1.00%)
  - ≥200 DM: BAD (0.70%)
  - < 200 DM: GOOD (89.10%)
- Further running credits:
  - ≤200 DM: GOOD (99.57%)
  - ≥200 DM: BAD (0.43%)
- Purpose of credit:
  - Used car, new car:
    - ≥200 DM: BAD (1.18%)
    - ≤200 DM: GOOD (99.82%)
  - Retraining, repair, other:
    - ≥200 DM: BAD (0.43%)
    - ≤200 DM: GOOD (99.57%)
- Most valuable asset available:
  - Savings contract:
    - ≤200 DM: GOOD (99.13%)
    - ≥200 DM: BAD (0.87%)
  - No assets or missing:
    - ≤200 DM: GOOD (99.57%)
    - ≥200 DM: BAD (0.43%)
- Duration in months:
  - < 22.5 or missing:
    - BAD (5.33%)
  - ≥22.5:
    - BAD (7.61%)
- Value of savings or stock:
  - < 22.5 or missing:
    - BAD (4.16%)
  - ≥22.5:
    - BAD (2.34%)
- Payment of previous credit:
  - Hesitant payer, no problems:
    - BAD (9.29%)
  - No savings, less than 100:
    - BAD (11.87%)
  - Greater than 100:
    - BAD (9.29%)
  - Missing:
    - BAD (6.49%)

Counts and percentages are for each decision node.
Oversampled Data with Priors

Cutoff probability is now the population probability of the rare event, here $p=0.05$!

<table>
<thead>
<tr>
<th>Predicted: Creditability=bad</th>
<th>Unadjusted P: Creditability=good</th>
<th>Unadjusted P: Creditability=bad</th>
<th>Residual: Creditability=good</th>
<th>Residual: Creditability=bad</th>
<th>Decision: Creditability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.26923076923076</td>
<td>0.25</td>
<td>0.75</td>
<td>0.75</td>
<td>-0.75</td>
<td>BAD</td>
</tr>
<tr>
<td>0.26923076923076</td>
<td>0.25</td>
<td>0.75</td>
<td>-0.25</td>
<td>0.25</td>
<td>BAD</td>
</tr>
<tr>
<td>0.26923076923076</td>
<td>0.25</td>
<td>0.75</td>
<td>-0.25</td>
<td>0.25</td>
<td>BAD</td>
</tr>
<tr>
<td>0.26923076923076</td>
<td>0.25</td>
<td>0.75</td>
<td>-0.25</td>
<td>0.25</td>
<td>BAD</td>
</tr>
<tr>
<td>0.26923076923076</td>
<td>0.25</td>
<td>0.75</td>
<td>-0.25</td>
<td>0.25</td>
<td>BAD</td>
</tr>
<tr>
<td>0.26923076923076</td>
<td>0.25</td>
<td>0.75</td>
<td>-0.25</td>
<td>0.25</td>
<td>BAD</td>
</tr>
</tbody>
</table>