DATA CONSIDERATIONS – EXTRA CONTENT

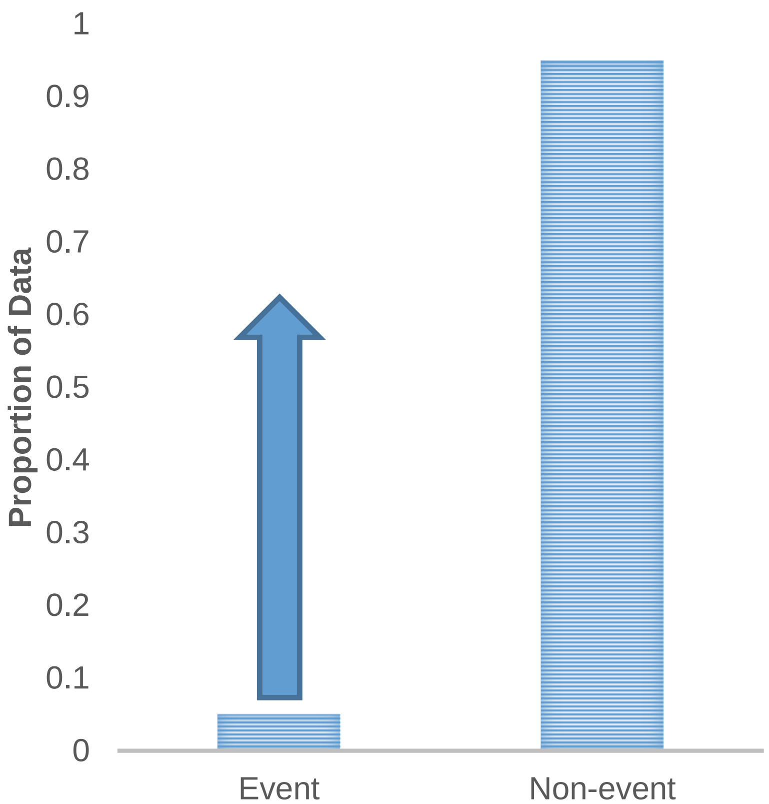
Dr. Aric LaBarr

Institute for Advanced Analytics

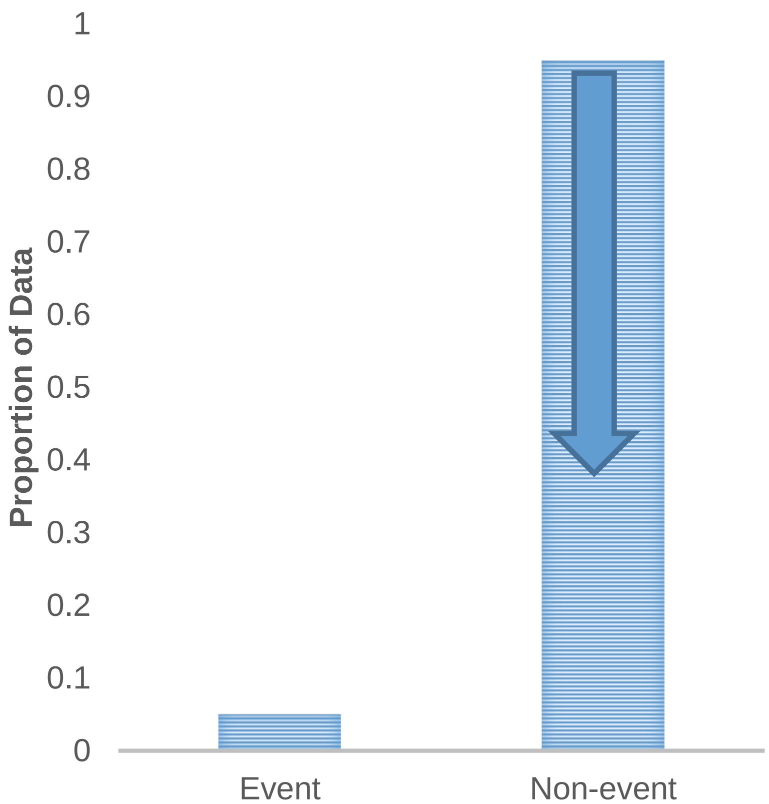
RARE EVENT MODELING IN R

Rare Event Sampling Correction

Oversampling



Undersampling



Rare Event Sampling Correction

Oversampling

- Duplicate current event cases in training set to balance better with non-event cases.
- Keep test set as original population proportion.

Undersampling

- Randomly sample current non-event cases to keep in the training set to balance with event cases.
- Keep test set as original population proportion.

Rare Event Sampling – R

```
train_id <- sample(seq_len(nrow(churn)),
                   size = floor(0.7*nrow(churn)))

train <- churn[train_id,]
valid <- churn[-train_id,]

table(train$churn)

##
## FALSE  TRUE
##  1995   107

table(valid$churn)

##
## FALSE  TRUE
##   855    47
```

Rare Event Sampling – R

```
prop.table(table(train$churn))

##
##     FALSE      TRUE
## 0.9490961 0.0509039

inputs <- train[,1:18]
target <- train[,19]
over_sam <- ubOver(X = inputs, Y = target)
train_o <- cbind(over_sam$X, over_sam$Y)
train_o$churn <- train_o$`over_sam$Y`
train_o$`over_sam$Y` <- NULL

table(train_o$churn)

##
## FALSE  TRUE
##  1995  1995
```

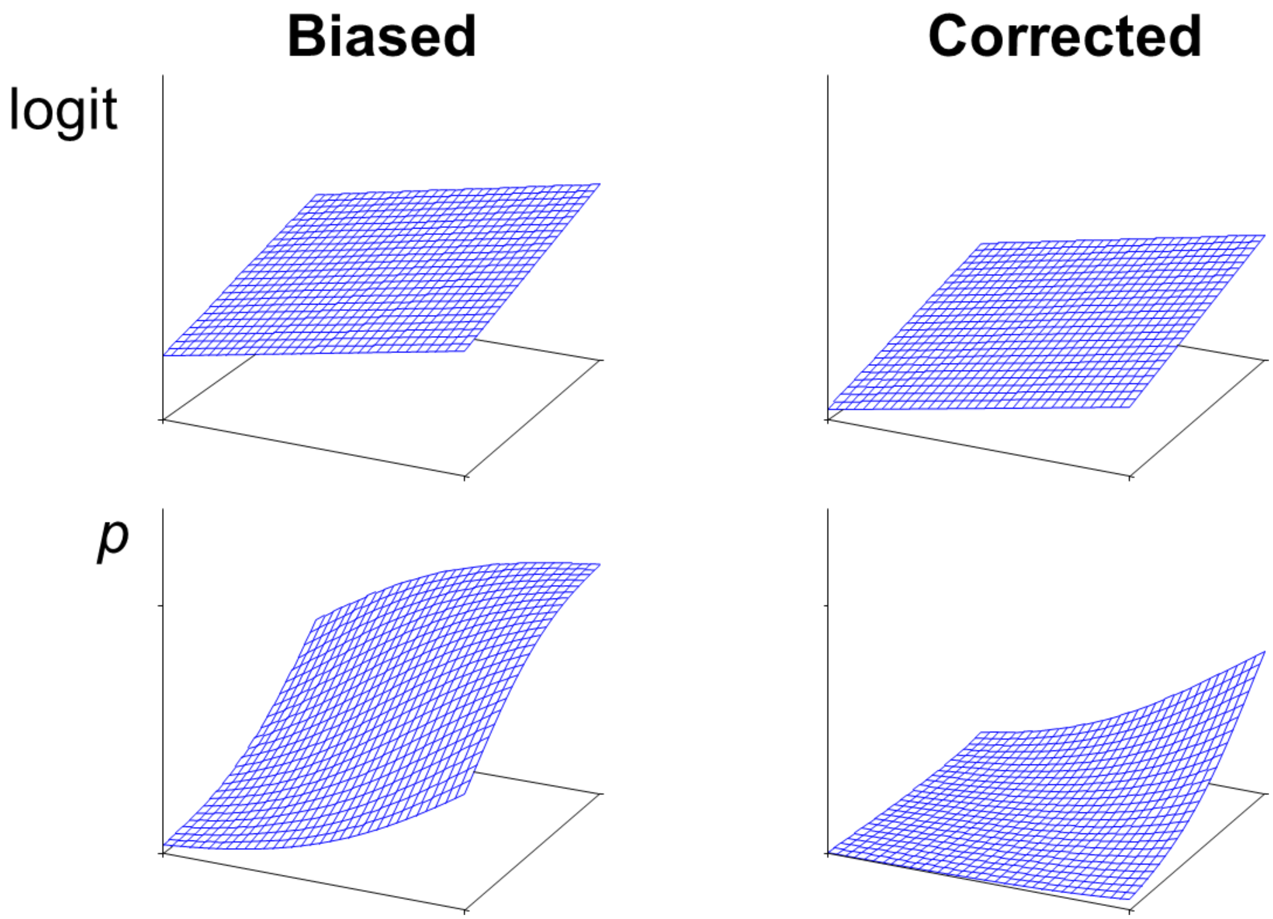
Rare Event Sampling – R

```
inputs <- train[,1:18]
target <- train[,19]
under_sam <- ubUnder(X = inputs, Y = target)
train_u <- cbind(under_sam$X, under_sam$Y)
train_u$churn <- train_u$`under_sam$Y`
train_u$`under_sam$Y` <- NULL

table(train_u$churn)

##
## FALSE  TRUE
##   107   107
```

Effect of Oversampling



Adjustments to Oversampling

- When the sample proportion is out of line with the population proportion, adjustments need to be made to correct the bias.
- 2 Methods:
  1. Adjusting the intercept
  2. Weighting observations

Adjust Intercept – R

```r
logit.model <- glm(churn ~ factor(international.plan) +
                           factor(voice.mail.plan) +
                           total.day.charge +
                           customer.service.calls,
                   data = train_u,
                   family = binomial(link = "logit"))
summary(logit.model)
```

Adjust Intercept – R

```
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -4.71202    0.77589  -6.073 1.25e-09 ***
## factor(international.plan)yes  2.91300    0.58964   4.940 7.80e-07 ***
## factor(voice.mail.plan)yes    -0.30174    0.43242  -0.698    0.485
## total.day.charge               0.09769    0.01829   5.341 9.26e-08 ***
## customer.service.calls         0.63437    0.12268   5.171 2.33e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 296.67  on 213  degrees of freedom
## Residual deviance: 217.88  on 209  degrees of freedom
## AIC: 227.88
```

Adjust Intercept – R

```
valid_p_bias <- predict(logit.model, newdata = valid,
                        type = "response")
valid_p <- (valid_p_bias*0.5*(154/3004))/
           ((1-valid_p_bias)*0.5*(2850/3004) +
            valid_p_bias*0.5*(154/3004))
```

Weighting Adjustment – R

```
train_u$weight <- ifelse(train_u$churn == 'TRUE', 1, 18.49)

logit.model.w <- glm(churn ~ factor(international.plan) +
                             factor(voice.mail.plan) +
                             total.day.charge +
                             customer.service.calls,
                     data = train_u,
                     family = binomial(link = "logit"),
                     weights = weight)
summary(logit.model.w)
```

Weighting Adjustment – R

```
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -7.29928    0.52017 -14.033  < 2e-16 ***
## factor(international.plan)yes   2.57544    0.28830   8.933  < 2e-16 ***
## factor(voice.mail.plan)yes     -1.08471    0.33129  -3.274  0.00106 **
## total.day.charge                0.10566    0.01363   7.754 8.90e-15 ***
## customer.service.calls          0.42969    0.07811   5.501 3.78e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 843.97  on 213  degrees of freedom
## Residual deviance: 681.96  on 209  degrees of freedom
## AIC: 686.82
```

CONTRASTS

Testing Individual Contrasts

- Instead of testing all possible combinations of odds ratios, you may only be interested in certain comparison, or a linear combination of comparisons.
- These are called **contrasts**.

- For example:
  - Group A vs. Group B
  - Group A vs. the average of Group B and Group C
  - Etc.

Testing Individual Contrasts – SAS

```
proc logistic data=churn_t;
      class international_plan(ref='no')
            voice_mail_plan(ref='no')
            customer_service_calls(ref='0') / param=ref;
      model churn(event='TRUE') = international_plan
                                  voice_mail_plan
                                  total_day_charge
                                  customer_service_calls;
      weight weights;
      oddsratio customer_service_calls_c / diff=all;
run;
quit;
```

Testing Individual Contrasts – SAS

Odds Ratio Estimates and Wald Confidence Intervals			
Odds Ratio	**Estimate**	**95% Confidence Limits**	
customer_service_calls 1 vs 0	1.082	0.125	9.397
customer_service_calls 2 vs 0	0.950	0.096	9.437
customer_service_calls 3 vs 0	1.246	0.088	17.694
customer_service_calls 4 vs 0	26.009	1.517	445.849
customer_service_calls 5 vs 0	13.653	0.634	293.830
customer_service_calls 6 vs 0	19.742	0.218	>999.999
customer_service_calls 7 vs 0	>999.999	<0.001	>999.999
customer_service_calls 1 vs 2	1.140	0.136	9.579
customer_service_calls 1 vs 3	0.869	0.071	10.572
customer_service_calls 1 vs 4	0.042	0.003	0.610
customer_service_calls 1 vs 5	0.079	0.004	1.482
customer_service_calls 1 vs 6	0.055	<0.001	4.643
customer_service_calls 1 vs 7	<0.001	<0.001	>999.999
customer_service_calls 2 vs 3	0.762	0.058	9.965
customer_service_calls 2 vs 4	0.037	0.002	0.556
customer_service_calls 2 vs 5	0.070	0.004	1.339
customer_service_calls 2 vs 6	0.048	<0.001	4.602
customer_service_calls 2 vs 7	<0.001	<0.001	>999.999
customer_service_calls 3 vs 4	0.048	0.002	1.029
customer_service_calls 3 vs 5	0.091	0.003	2.451
customer_service_calls 3 vs 6	0.063	<0.001	6.951
customer_service_calls 3 vs 7	<0.001	<0.001	>999.999
customer_service_calls 4 vs 5	1.905	0.070	51.556
customer_service_calls 4 vs 6	1.317	0.010	170.540
customer_service_calls 4 vs 7	<0.001	<0.001	>999.999
customer_service_calls 5 vs 6	0.692	0.005	106.159
customer_service_calls 5 vs 7	<0.001	<0.001	>999.999
customer_service_calls 6 vs 7	<0.001	<0.001	>999.999

Testing Individual Contrasts – SAS

```
proc logistic data=churn_t;
   class international_plan(ref='no')
         voice_mail_plan(ref='no')
         customer_service_calls(ref='0') / param=ref;
   model churn(event='TRUE') = international_plan
                               voice_mail_plan
                               total_day_charge
                               customer_service_calls / clodds=pl;
      weight weights;
      test customer_service_cal1 = customer_service_cal2;
      test customer_service_cal1 = 0.25*customer_service_cal4 +
                                   0.25*customer_service_cal5 +
                                   0.25*customer_service_cal6 +
                                   0.25*customer_service_cal7;
run;
quit;
```

Testing Individual Contrasts – SAS

Linear Hypotheses Testing Results			
Label	**Wald Chi-Square**	**DF**	**Pr > ChiSq**
Test 1	0.0145	1	0.9043
Test 2	0.0000	1	0.9951

Testing Individual Contrasts – R

```
train_u$fcsc <- factor(train_u$customer.service.calls)

logit.model.w.2 <- glm(churn ~ factor(international.plan) +
                               factor(voice.mail.plan) +
                               total.day.charge +
                               fcsc,
                       data = train_u,
                       family = binomial(link = "logit"),
                       weights = weight)

summary(glht(logit.model.w.2, linfct = mcp(fcsc = "Tukey")))
```

Testing Individual Contrasts – R

```
## Linear Hypotheses:
##              Estimate Std. Error z value Pr(>|z|)
## 1 - 0 == 0   -0.4461     1.0427  -0.428    0.999
## 2 - 0 == 0   -0.2486     1.0901  -0.228    1.000
## 3 - 0 == 0    0.1342     1.2275   0.109    1.000
## 4 - 0 == 0    0.8636     1.1639   0.742    0.987
## 5 - 0 == 0    2.3143     1.5825   1.462    0.727
## 6 - 0 == 0   20.5210  1865.2667   0.011    1.000
## 2 - 1 == 0    0.1975     1.0625   0.186    1.000
## 3 - 1 == 0    0.5803     1.1901   0.488    0.999
## 4 - 1 == 0    1.3097     1.1705   1.119    0.904
## 5 - 1 == 0    2.7604     1.5486   1.782    0.508
## 6 - 1 == 0   20.9671  1865.2667   0.011    1.000
## 3 - 2 == 0    0.3828     1.2285   0.312    1.000
## 4 - 2 == 0    1.1122     1.2322   0.903    0.965
## 5 - 2 == 0    2.5629     1.5755   1.627    0.617
## 6 - 2 == 0   20.7696  1865.2668   0.011    1.000
## 4 - 3 == 0    0.7294     1.3468   0.542    0.998
## 5 - 3 == 0    2.1801     1.6573   1.315    0.814
## 6 - 3 == 0   20.3868  1865.2668   0.011    1.000
## 5 - 4 == 0    1.4507     1.6966   0.855    0.973
## 6 - 4 == 0   19.6573  1865.2668   0.011    1.000
## 6 - 5 == 0   18.2067  1865.2670   0.010    1.000
## (Adjusted p values reported -- single-step method)
```

Rare Event Modeling **Contrasts**

?