DATA CONSIDERATIONS

Dr. Aric LaBarr

Institute for Advanced Analytics

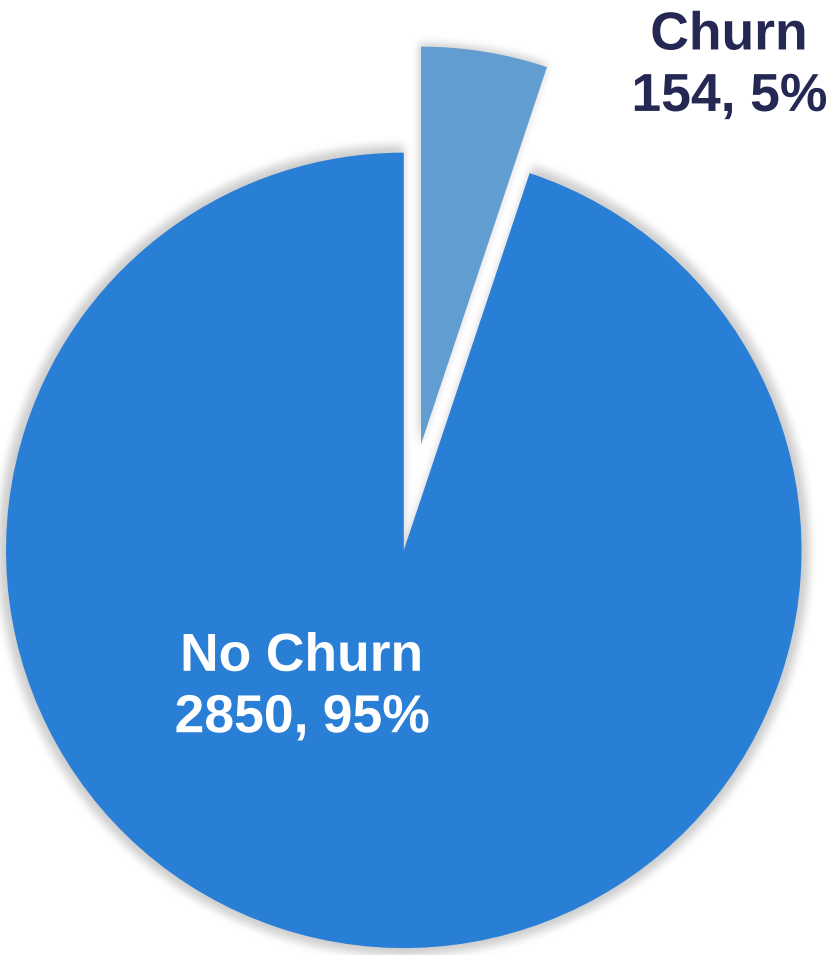
RARE EVENT MODELING

Rare Event Modeling

- 5% or smaller in a category can lead to classification problems.
- Common Situations:
  - Fraud
  - Default
  - Marketing Response
  - Weather Event

Telecomm Churn Data Set

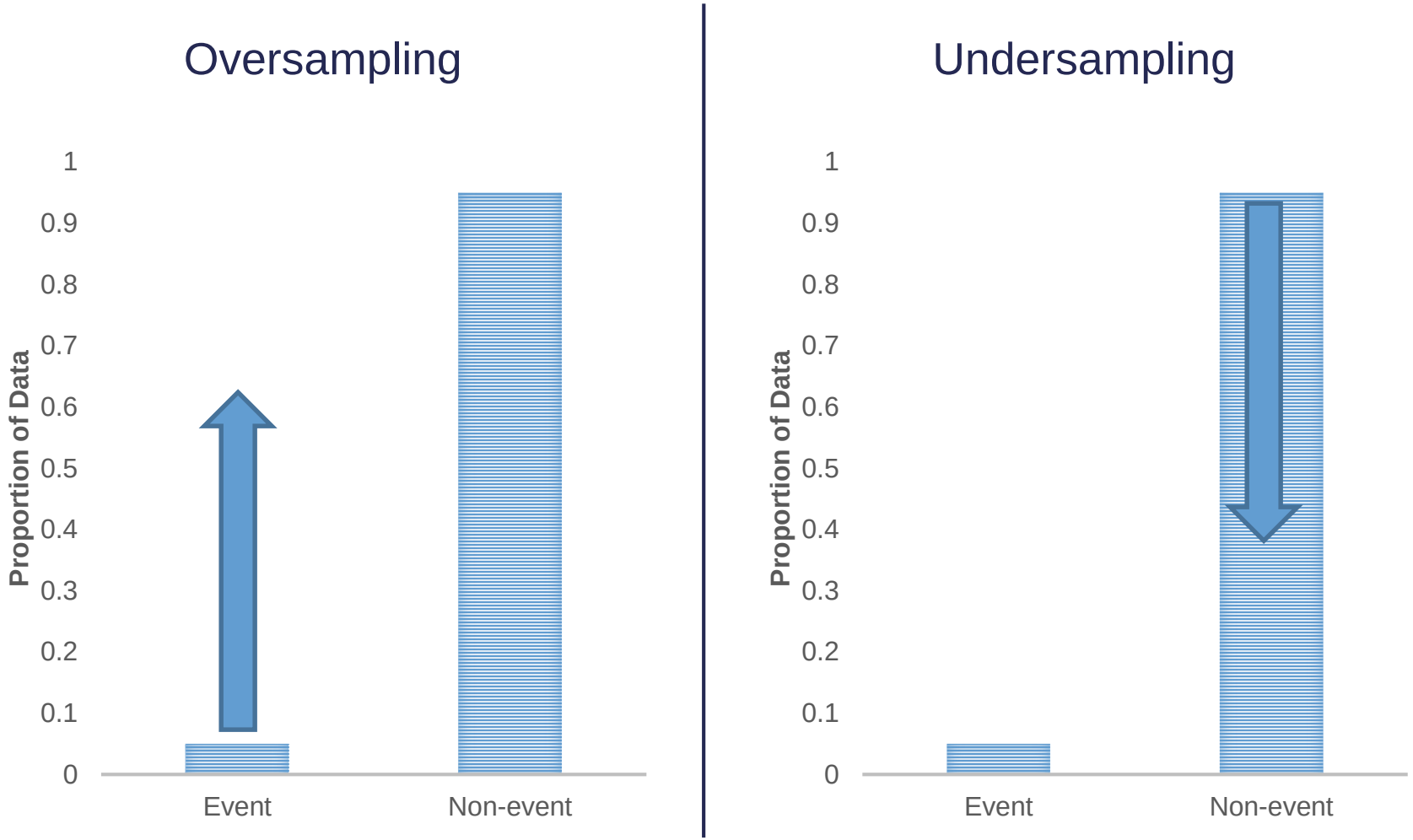
- Model the association between various factors and a customer churning (leaving the company)
- 3004 observations in the data set



Telecomm Churn Data Set

- Model the association between various factors and a customer churning (leaving the company)
- Predictors:
  - **account_length:** length of time with company
  - **international_plan:** yes, no
  - **voice_mail_plan:** yes, no
  - **customer_service_calls:** number of service calls
  - **total_day_minutes:** minutes used during daytime
  - **total_day_calls:** calls used during daytime
  - **total_day_charge:** cost of usage during daytime
  - Same as previous three for evening, night, international

Rare Event Sampling Correction



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 – SAS

```
proc surveyselect data = logistic.tele_churn noprint
                  out=churn_over seed=12345
                  sampsize=(100 100) outall
                  stratumseed=restore;
   strata churn;
run;
```

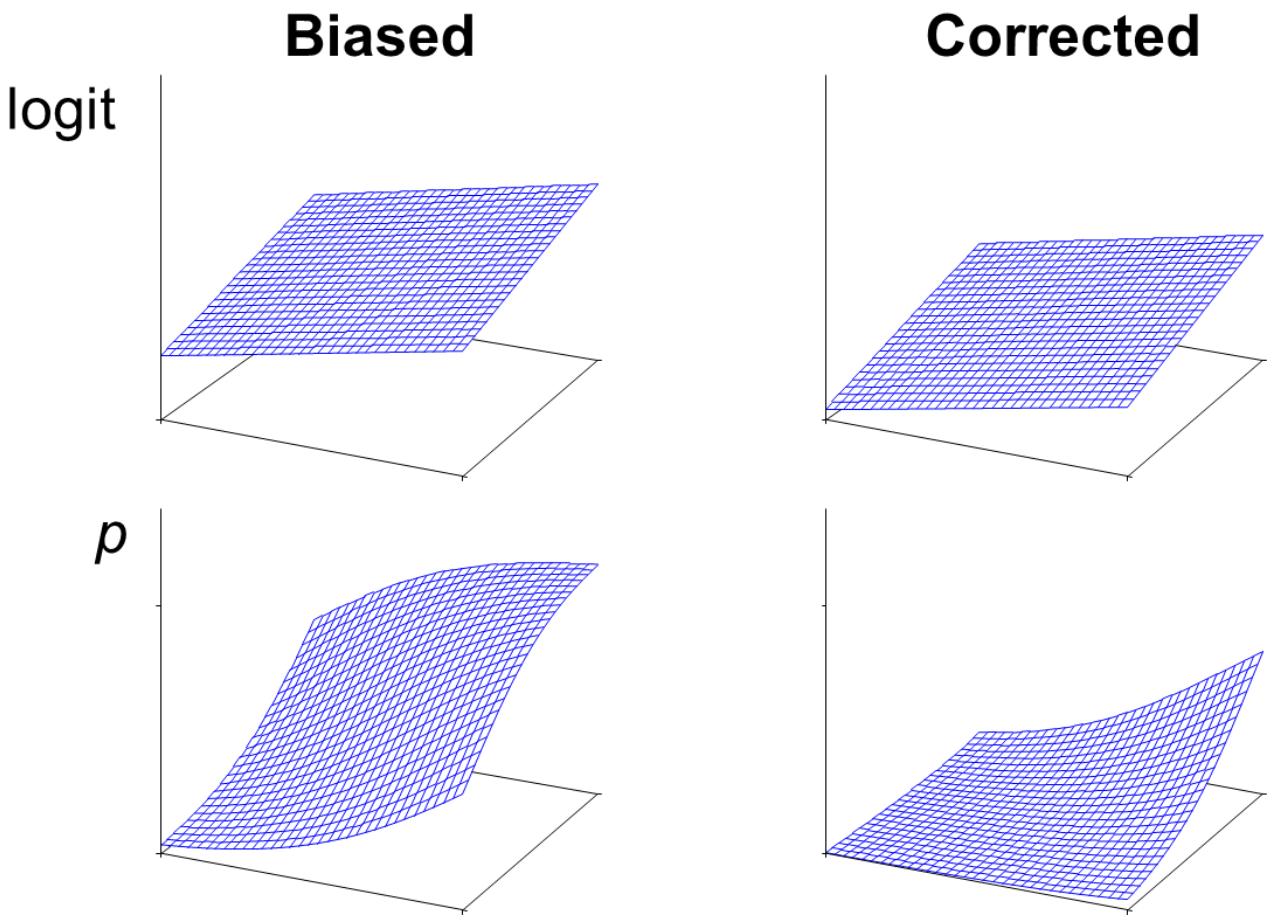
Rare Event Sampling – SAS

The FREQ Procedure

Frequency
Col Pct

Table of churn by Selected			
churn	**Selected(Selection Indicator)**		
	0	**1**	**Total**
FALSE	2750 98.07	100 50.00	2850
TRUE	54 1.93	100 50.00	154
Total	2804	200	3004

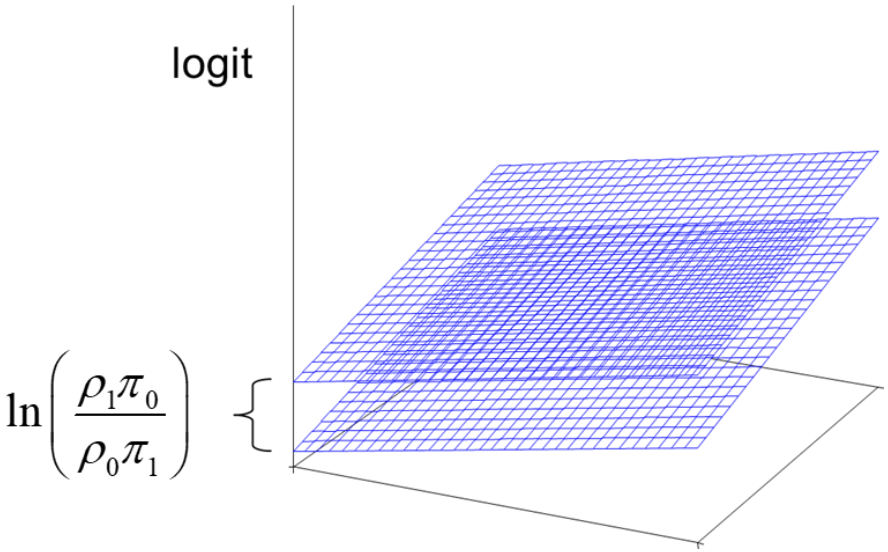
Effect of Oversampling



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

Adjusting the Intercept



- Need to correct for the bias created by oversampling.
- Adjustment is only applied to intercept.
- This adjusts the predicted values:

- Population proportion: π_1, π_0
- Sample proportion: ρ_1, ρ_0
- Unadjusted predictions: $\hat{p}_i^*$

$$\hat{p}_i = \frac{\hat{p}_i^* \rho_0 \pi_1}{(1-\hat{p}_i^*)\rho_1\pi_0 + \hat{p}_i^* \rho_0 \pi_1}$$

Weighting Observations

- Instead of adjusting the model after it is built, weighting observations adjusts while the model is being built.
- Uses **weighted MLE** instead – each observation has potentially different weight to the MLE calculation.
- Need to create a weight variable in the oversampled data set:

$$weight = \begin{cases} \pi_1/\rho_1, & y = 1 \\ \pi_0/\rho_0, & y = 0 \end{cases}$$

Weighting Observations

- Instead of adjusting the model after it is built, weighting observations adjusts while the model is being built.
- Uses **weighted MLE** instead – each observation has potentially different weight to the MLE calculation.
- Need to create a weight variable in the oversampled data set:

$$weight = \begin{cases} \pi_1/\rho_1, & y = 1 \\ \pi_0/\rho_0, & y = 0 \end{cases}$$

OR

$$weight = \begin{cases} 1, & y = 1 \\ \rho_1\pi_0/\rho_0\pi_1, & y = 0 \end{cases}$$

When to Use Which Technique?

	Model Correct	Model Misspecified
Small Sample ($n \leq 1000$)	Adjust Intercept	Weighted Observations
Large Sample ($n > 1000$)	Either	Weighted Observations

Adjust Intercept – SAS

```
proc freq data=logistic.tele_churn noprint;
    table churn / out=priors(drop=percent
                             rename=(count=_prior_));
run;

proc logistic data=churn_t;
    class international_plan(ref='no')
          voice_mail_plan(ref='no') / param=ref;
    model churn(event='TRUE') = international_plan
                                voice_mail_plan
                                total_day_charge
                                customer_service_calls
                                / clodds=pl;
    score data=churn_v prior=priors out=churn_scored1;
run;
quit;
```

Adjust Intercept – SAS

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
international_plan	1	24.5737	<.0001
voice_mail_plan	1	5.6354	0.0176
total_day_charge	1	17.1895	<.0001
customer_service_calls	1	27.7822	<.0001

Adjust Intercept – SAS

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-3.9665	0.7663	26.7894	<.0001
international_plan	**yes**	1	2.7737	0.5595	24.5737	<.0001
voice_mail_plan	**yes**	1	-1.0891	0.4588	5.6354	0.0176
total_day_charge		1	0.0754	0.0182	17.1895	<.0001
customer_service_cal		1	0.6943	0.1317	27.7822	<.0001

Adjust Intercept – SAS

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
international_plan yes vs no	1.0000	16.018	5.771	52.942
voice_mail_plan yes vs no	1.0000	0.337	0.131	0.804
total_day_charge	1.0000	1.078	1.042	1.119
customer_service_cal	1.0000	2.002	1.568	2.634

Weighting Adjustment – SAS

```
data churn_t;
      set churn_t;
      weights = 0.1026;
      if churn = 'FALSE' then weights = 1.8974;
run;

proc logistic data=churn_t;
      class international_plan(ref='no')
            voice_mail_plan(ref='no') / param=ref;
      model churn(event='TRUE') = international_plan
                                  voice_mail_plan
                                  total_day_charge
                                  customer_service_calls
                                  / clodds=pl;
      weight weights;
      score data=churn_v out=churn_scored2;
run;
quit;
```

Weighting Adjustment – SAS

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
international_plan	1	8.8884	0.0029
voice_mail_plan	1	1.4210	0.2332
total_day_charge	1	3.5723	0.0588
customer_service_calls	1	6.3759	0.0116

Weighting Adjustment – SAS

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-6.6834	1.6097	17.2391	<.0001
international_plan	**yes**	1	2.4699	0.8285	8.8884	0.0029
voice_mail_plan	**yes**	1	-1.2089	1.0142	1.4210	0.2332
total_day_charge		1	0.0760	0.0402	3.5723	0.0588
customer_service_cal		1	0.6111	0.2420	6.3759	0.0116

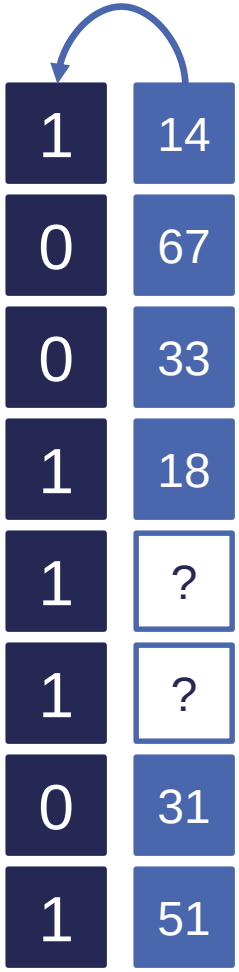
Weighting Adjustment – SAS

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
international_plan yes vs no	1.0000	11.822	2.172	62.024
voice_mail_plan yes vs no	1.0000	0.299	0.024	1.657
total_day_charge	1.0000	1.079	1.000	1.173
customer_service_cal	1.0000	1.842	1.138	2.997

MISSING VALUES

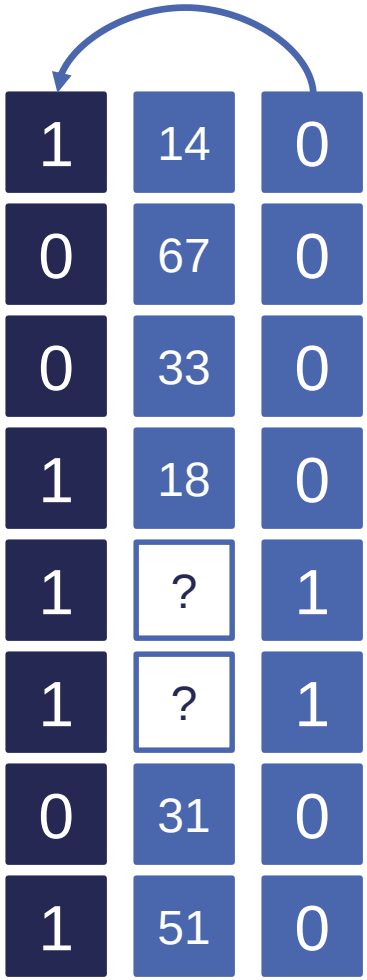
Target Dependent?

- Missing values in predictor variables are not necessarily bad.
- Might be randomly missing which doesn't necessarily pose a model problem.
- Are they dependent on the **target** variable?



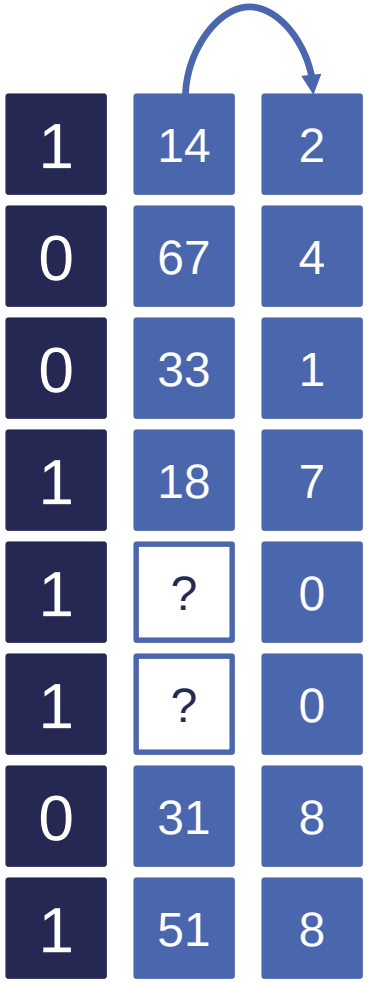
Target Dependent?

- Missing values in predictor variables are not necessarily bad.
- Might be randomly missing which doesn't necessarily pose a model problem.
- Are they dependent on the **target** variable?
- Create **missing value variable**, is it significant?



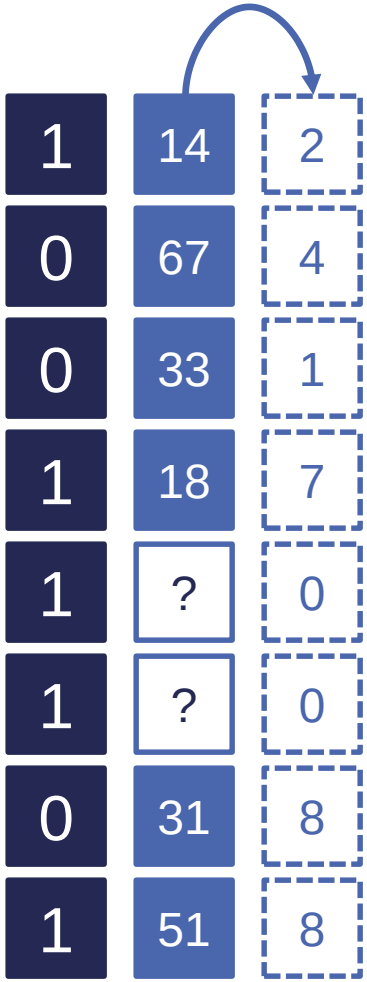
Predictor Dependent?

- Missing values in predictor variables are not necessarily bad.
- Might be randomly missing which doesn't necessarily pose a model problem.
- Are they **dependent** on an **independent** variable?

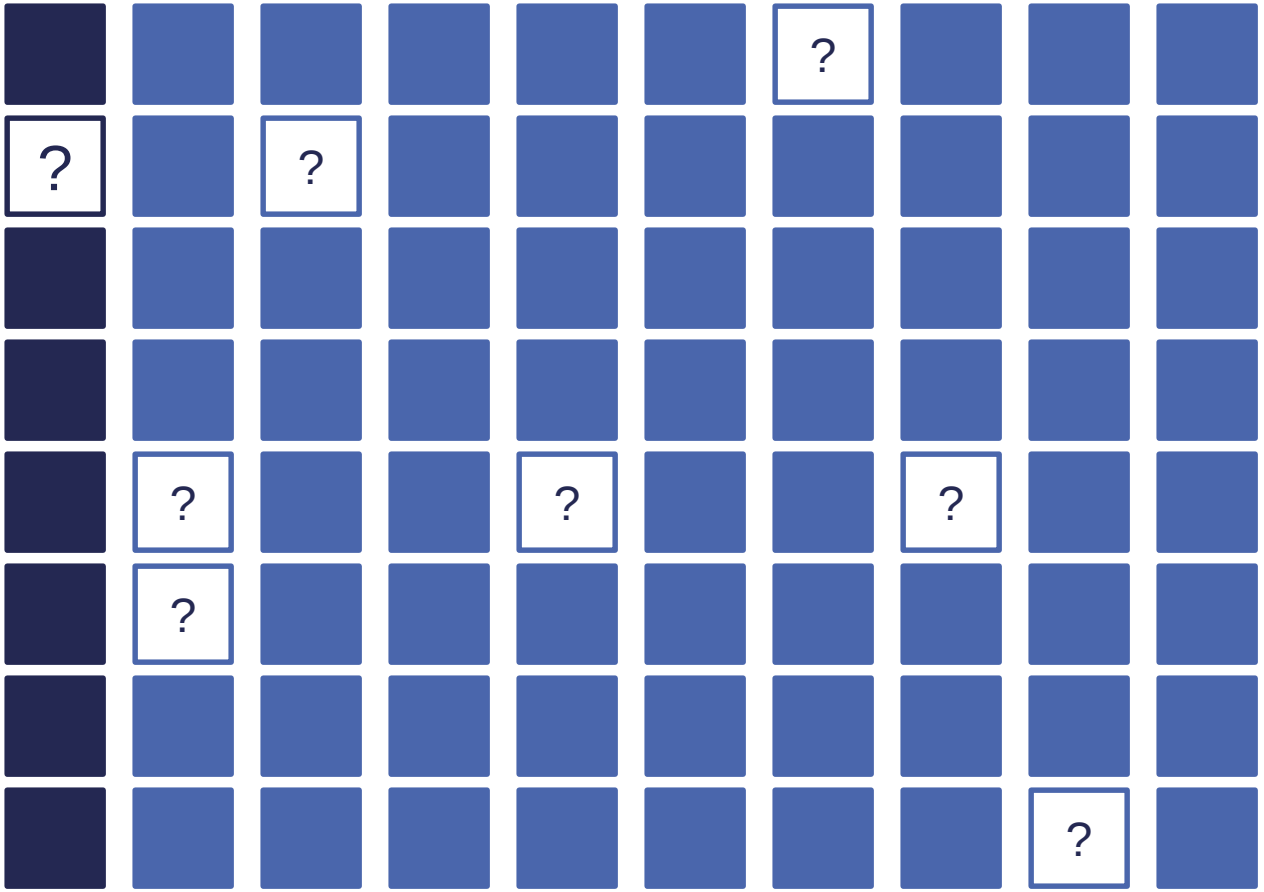


Predictor Dependent?

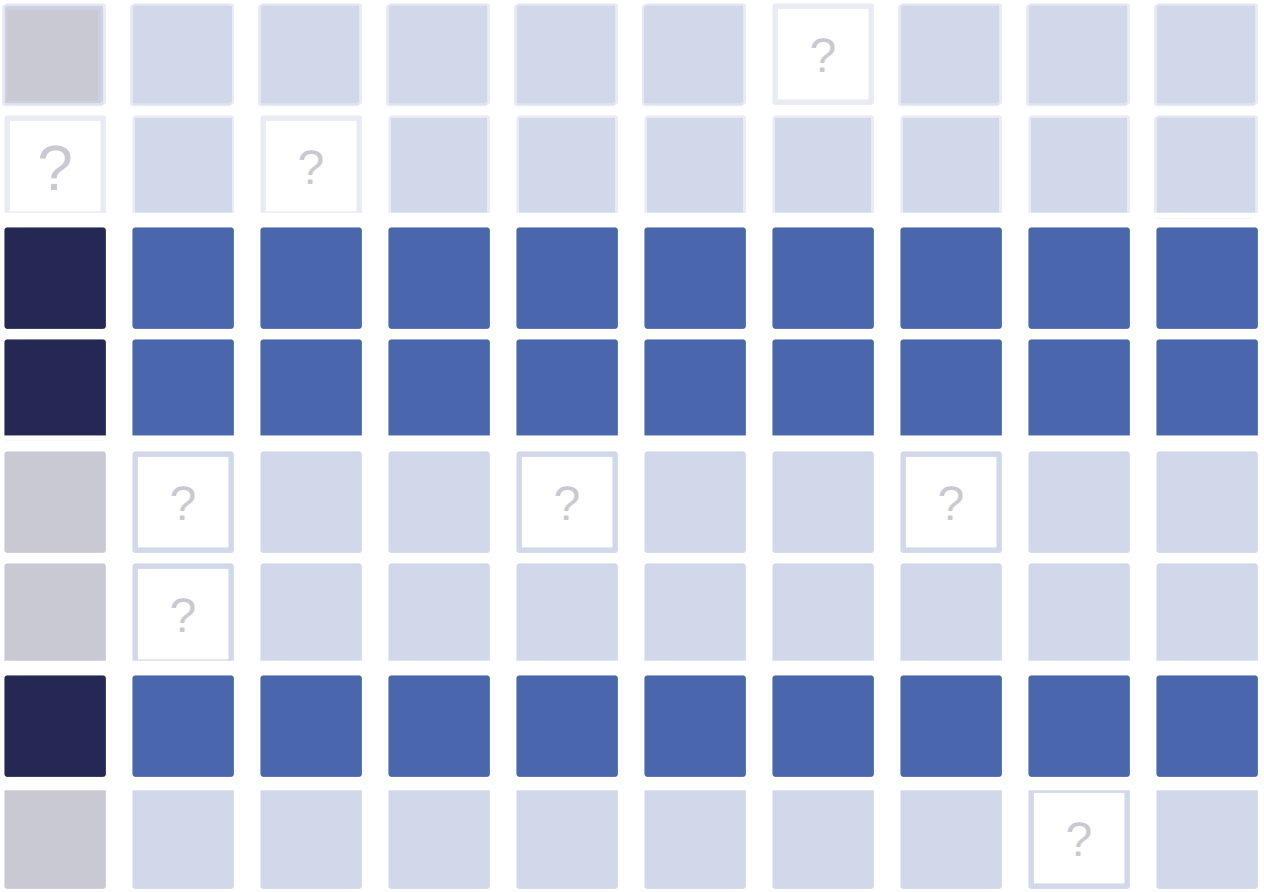
- Missing values in predictor variables are not necessarily bad.
- Might be randomly missing which doesn't necessarily pose a model problem.
- Are they **dependent** on an **independent** variable?
- Is that **independent** variable recorded?



Complete Case Analysis



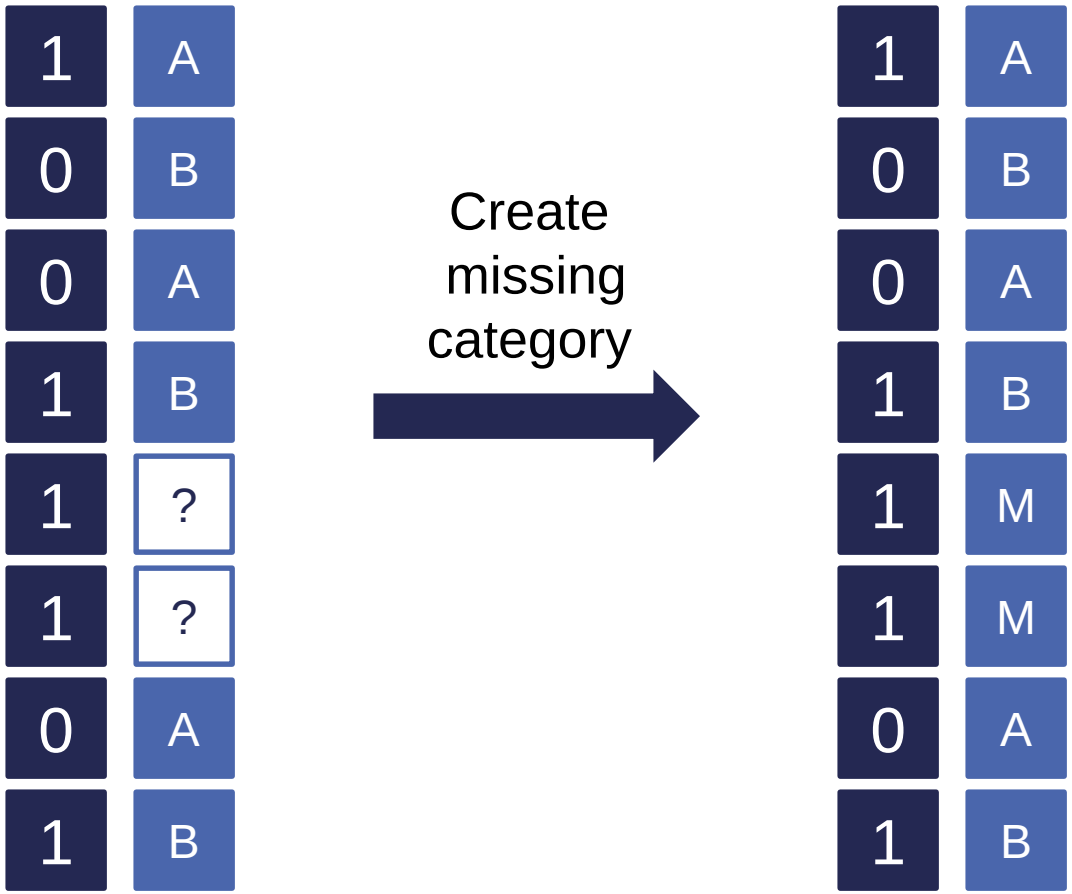
Complete Case Analysis



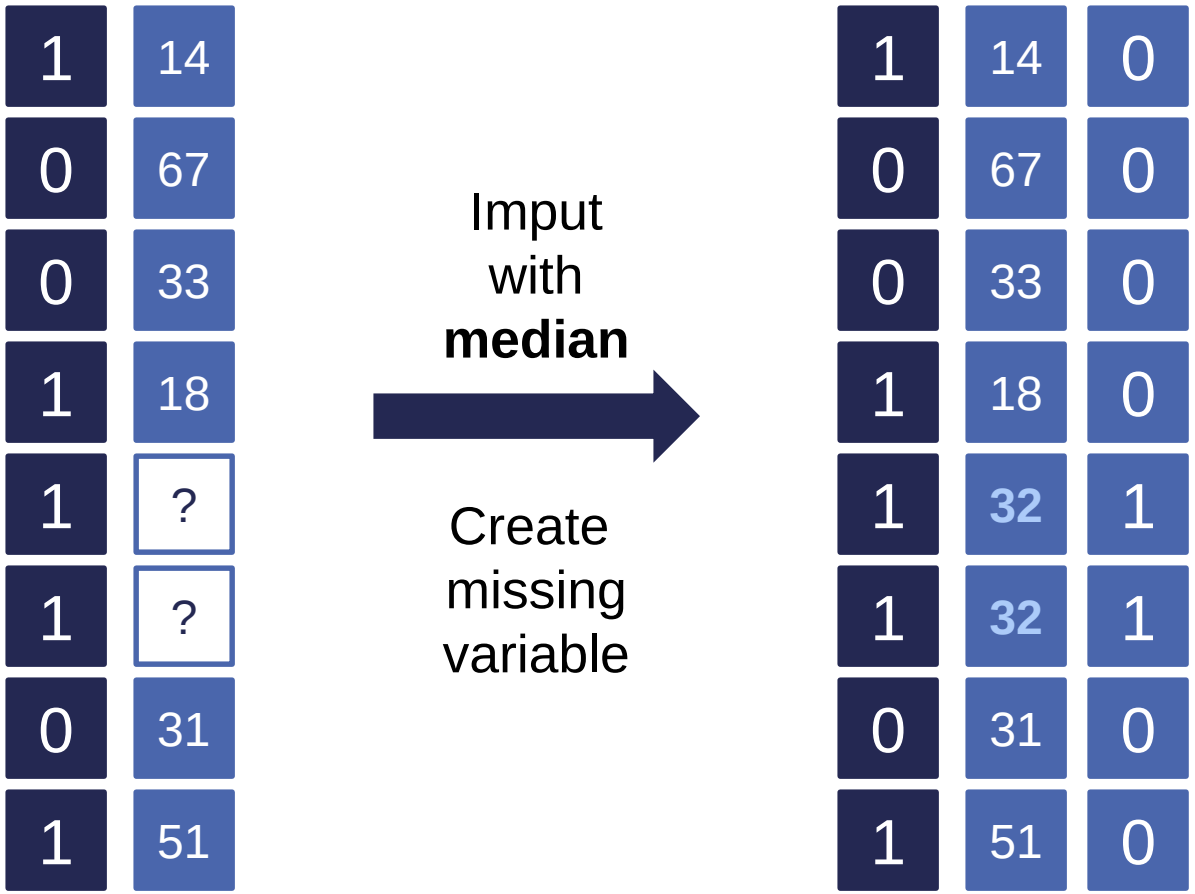
Scoring New Values?

- Complete cases analysis isn't necessarily bad.
- How to handle scoring new observations with missing values?
- Imputation might be necessary.

Categorical Imputation?



Continuous Imputation



General (not Strict) Imputation Rules

- If variable has more than 50% missing, consider deleting from analysis.

- **Categorical**:
  - Create missing value category for categorical variables.

- **Continuous**:
  - Impute missing values for continuous variables (median is a popular choice)
  - Create a missing value binary variable for each of the continuous variables you impute.

CONVERGENCE PROBLEMS

Categorical Variables – 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;
run;
quit;
```

Categorical Variables – SAS

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
international_plan	1	7.9167	0.0049
voice_mail_plan	1	1.2282	0.2678
total_day_charge	1	1.9227	0.1656
customer_service_calls	7	10.5543	0.1593

Categorical Variables – SAS

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
international_plan yes vs no	1.0000	11.072	1.921	60.907
voice_mail_plan yes vs no	1.0000	0.278	0.017	1.962
total_day_charge	1.0000	1.059	0.979	1.156
customer_service_cal 1 vs 0	1.0000	1.082	0.117	11.674
customer_service_cal 2 vs 0	1.0000	0.950	.	11.401
customer_service_cal 3 vs 0	1.0000	1.246	.	.
customer_service_cal 4 vs 0	1.0000	26.009	1.582	575.742
customer_service_cal 5 vs 0	1.0000	13.653	0.486	334.790
customer_service_cal 6 vs 0	1.0000	19.742	.	>999.999
customer_service_cal 7 vs 0	1.0000	>999.999	.	.

Categorical Variables – SAS

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
international_plan yes vs no	1.0000	11.072	1.921	60.907
voice_mail_plan yes vs no	1.0000	0.278	0.017	1.962
total_day_charge	1.0000	1.059	0.979	1.156
customer_service_cal 1 vs 0	1.0000	1.082	0.117	11.674
customer_service_cal 2 vs 0	1.0000	0.950	.	11.401
customer_service_cal 3 vs 0	1.0000	1.246	.	.
customer_service_cal 4 vs 0	1.0000	26.009	1.582	575.742
customer_service_cal 5 vs 0	1.0000	13.653	0.486	334.790
customer_service_cal 6 vs 0	1.0000	19.742	.	>999.999
customer_service_cal 7 vs 0	1.0000	>999.999	.	.

Categorical Variables – SAS

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-5.3575	1.6969	9.9676	0.0016
international_plan	**yes**	1	2.4044	0.8545	7.9167	0.0049
voice_mail_plan	**yes**	1	-1.2811	1.1560	1.2282	0.2678
total_day_charge		1	0.0577	0.0416	1.9227	0.1656
customer_service_cal	**1**	1	0.0789	1.1028	0.0051	0.9429
customer_service_cal	**2**	1	-0.0517	1.1716	0.0019	0.9648
customer_service_cal	**3**	1	0.2199	1.3538	0.0264	0.8710
customer_service_cal	**4**	1	3.2584	1.4498	5.0514	0.0246
customer_service_cal	**5**	1	2.6140	1.5659	2.7868	0.0950
customer_service_cal	**6**	1	2.9827	2.2991	1.6831	0.1945
customer_service_cal	**7**	1	18.4993	4416.3	0.0000	0.9967

Categorical Variables – R

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

summary(logit.model.w)
```

Categorical Variables – R

```
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -6.96906    1.89291  -3.682 0.000232
***
## factor(international.plan)yes     2.58133    0.96015   2.688 0.007178 **
## factor(voice.mail.plan)yes       -1.29504    1.11910  -1.157 0.247184
## total.day.charge                  0.11776    0.04691   2.510 0.012063 *
## factor(customer.service.calls)1  -0.44612    1.04269  -0.428 0.668757
## factor(customer.service.calls)2  -0.24863    1.09007  -0.228 0.819577
## factor(customer.service.calls)3   0.13420    1.22748   0.109 0.912943
## factor(customer.service.calls)4   0.86362    1.16391   0.742 0.458089
## factor(customer.service.calls)5   2.31430    1.58254   1.462 0.143632
## factor(customer.service.calls)6  20.52095 1865.26674   0.011 0.991222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
##     Null deviance: 86.595  on 213  degrees of freedom
## Residual deviance: 65.565  on 204  degrees of freedom
## AIC: 39.538
```

Linear Separation

- **Complete linear separation** occurs when some combination of the predictors perfectly predict **every** outcome:

	Yes	No
Group A	100	0
Group B	0	50

- **Quasi-complete separation** occurs when the outcome can be perfectly predicted for only a subset of the data:

	Yes	No
Group A	77	23
Group B	0	50

Linear Separation

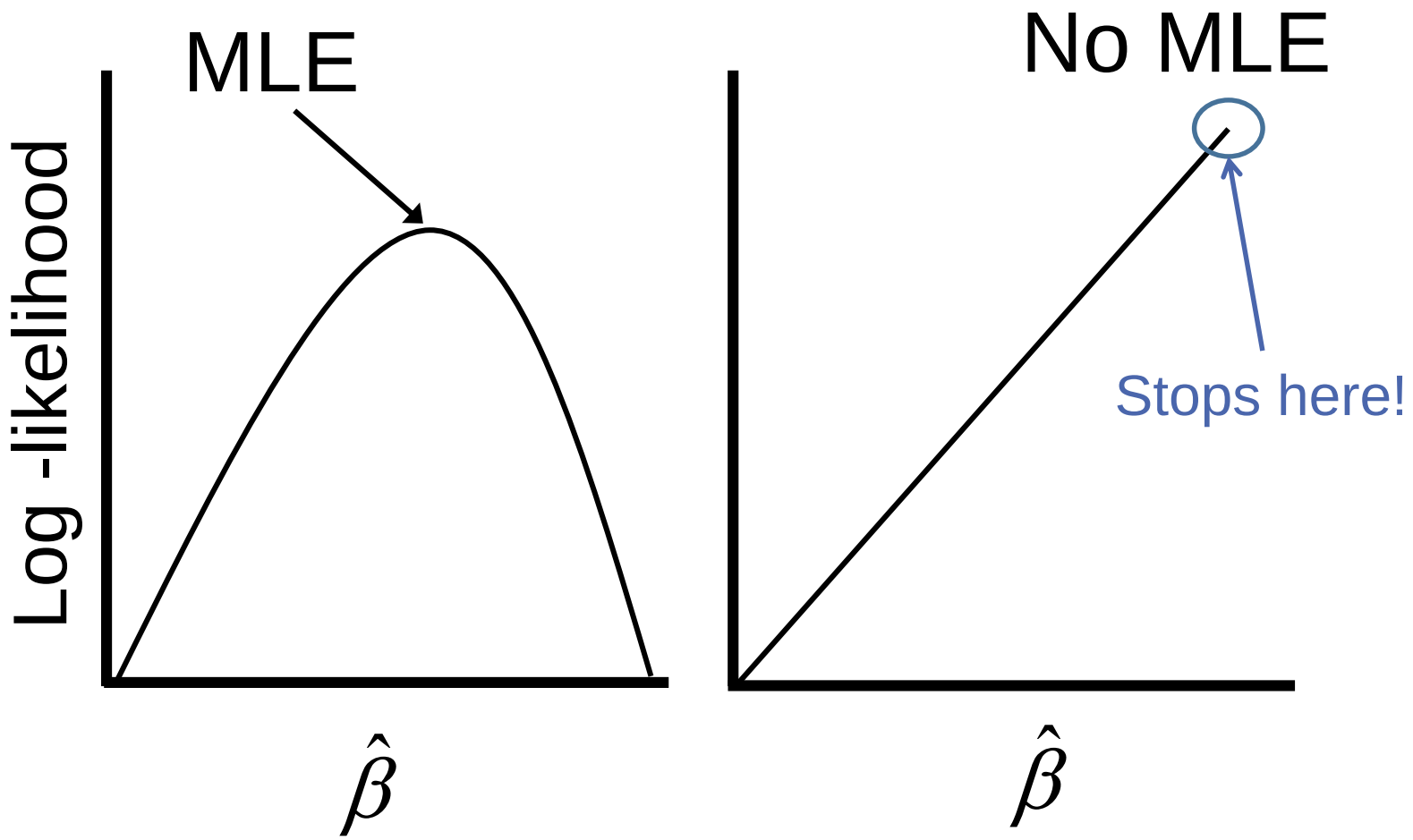
- **Complete linear separation** occurs when some combination of the predictors perfectly predict **every** outcome:

	Yes	No	Logit
Group A	100	0	∞
Group B	0	50	−∞

- **Quasi-complete separation** occurs when the outcome can be perfectly predicted for only a subset of the data:

	Yes	No	Logit
Group A	77	23	1.39
Group B	0	50	−∞

Problems with Convergence



Linear Separation – SAS

- SAS Warning Message:
  - WARNING: There is a complete separation of data points. The maximum likelihood estimate does not exist.
  - WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Linear Separation – SAS

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
international_plan yes vs no	1.0000	11.072	1.921	60.907
voice_mail_plan yes vs no	1.0000	0.278	0.017	1.962
total_day_charge	1.0000	1.059	0.979	1.156
customer_service_cal 1 vs 0	1.0000	1.082	0.117	11.674
customer_service_cal 2 vs 0	1.0000	0.950	.	11.401
customer_service_cal 3 vs 0	1.0000	1.246	.	.
customer_service_cal 4 vs 0	1.0000	26.009	1.582	575.742
customer_service_cal 5 vs 0	1.0000	13.653	0.486	334.790
customer_service_cal 6 vs 0	1.0000	19.742	.	>999.999
customer_service_cal 7 vs 0	1.0000	>999.999	.	.

Linear Separation – SAS

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-5.3575	1.6969	9.9676	0.0016
international_plan	**yes**	1	2.4044	0.8545	7.9167	0.0049
voice_mail_plan	**yes**	1	-1.2811	1.1560	1.2282	0.2678
total_day_charge		1	0.0577	0.0416	1.9227	0.1656
customer_service_cal	**1**	1	0.0789	1.1028	0.0051	0.9429
customer_service_cal	**2**	1	-0.0517	1.1716	0.0019	0.9648
customer_service_cal	**3**	1	0.2199	1.3538	0.0264	0.8710
customer_service_cal	**4**	1	3.2584	1.4498	5.0514	0.0246
customer_service_cal	**5**	1	2.6140	1.5659	2.7868	0.0950
customer_service_cal	**6**	1	2.9827	2.2991	1.6831	0.1945
customer_service_cal	**7**	1	18.4993	4416.3	0.0000	0.9967

Linear Separation – R

- R Warning Message:

Linear Separation – R

```
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -6.96906    1.89291  -3.682 0.000232
***
## factor(international.plan)yes     2.58133    0.96015   2.688 0.007178 **
## factor(voice.mail.plan)yes       -1.29504    1.11910  -1.157 0.247184
## total.day.charge                  0.11776    0.04691   2.510 0.012063 *
## factor(customer.service.calls)1  -0.44612    1.04269  -0.428 0.668757
## factor(customer.service.calls)2  -0.24863    1.09007  -0.228 0.819577
## factor(customer.service.calls)3   0.13420    1.22748   0.109 0.912943
## factor(customer.service.calls)4   0.86362    1.16391   0.742 0.458089
## factor(customer.service.calls)5   2.31430    1.58254   1.462 0.143632
## factor(customer.service.calls)6  20.52095 1865.26674   0.011 0.991222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
##     Null deviance: 86.595  on 213  degrees of freedom
## Residual deviance: 65.565  on 204  degrees of freedom
## AIC: 39.538
```

Solutions

- Possible Solutions:
  - Penalized maximum likelihood.
  - Collapse the categories of the predictor variable to eliminate the 0 cell count.
  - Eliminate the category altogether – probably not reasonable since the category seems important!
  - Add a very small constant to the cell counts.

Solutions

- Possible Solutions:
  - Penalized maximum likelihood.
  - Collapse the categories of the predictor variable to eliminate the 0 cell count.
  - Eliminate the category altogether – probably not reasonable since the category seems important!
  - Add a very small constant to the cell counts.

Penalized Likelihood – 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
                                    / firth;
        weight weights;
run;
quit;
```

Penalized Likelihood – SAS

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
international_plan yes vs no	8.878	1.996	39.491
voice_mail_plan yes vs no	0.426	0.072	2.525
total_day_charge	1.048	0.977	1.125
customer_service_cal 1 vs 0	1.016	0.159	6.482
customer_service_cal 2 vs 0	0.934	0.131	6.654
customer_service_cal 3 vs 0	1.371	0.151	12.435
customer_service_cal 4 vs 0	18.640	1.344	258.420
customer_service_cal 5 vs 0	11.596	0.666	201.894
customer_service_cal 6 vs 0	19.526	0.530	719.576

Penalized Likelihood – R

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

summary(logit.model.w)
```

Penalized Likelihood – R

```
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -6.10051    1.63131  -3.740 0.000184 ***
## factor(international.plan)yes     2.30352    0.87299   2.639 0.008323 **
## factor(voice.mail.plan)yes       -0.93552    0.91055  -1.027 0.304219
## total.day.charge                  0.10157    0.04079   2.490 0.012773 *
## factor(customer.service.calls)1 -0.43647    0.91200  -0.479 0.632235
## factor(customer.service.calls)2 -0.21697    0.95968  -0.226 0.821132
## factor(customer.service.calls)3  0.17770    1.05547   0.168 0.866302
## factor(customer.service.calls)4  0.75983    1.06167   0.716 0.474181
## factor(customer.service.calls)5  2.21092    1.48426   1.490 0.136336
## factor(customer.service.calls)6  5.12629    3.53274   1.451 0.146757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.462  on 213  degrees of freedom
## Residual deviance: 66.954  on 204  degrees of freedom
## Penalized deviance: 59.44305
## AIC:  45.797
```

Solutions

- Possible Solutions:
  - Penalized maximum likelihood.
  - Collapse the categories of the predictor variable to eliminate the 0 cell count.
  - Eliminate the category altogether – probably not reasonable since the category seems important!
  - Add a very small constant to the cell counts.

Thresholding – Ordinal Option

Level	Sample Size	0	1
A	1562	1230	332
B	970	917	53
C	223	206	17
D	111	101	10
E	85	81	4
F	50	40	10
G	23	22	1
H	17	17	0
I	12	11	1
J	5	5	0

Thresholding – Ordinal Option

Level	Sample Size	0	1
A	1562	1230	332
B	970	917	53
C	223	206	17
D	111	101	10
E	85	81	4
F	50	40	10
G	23	22	1
H	17	17	0
I	12	11	1
J	5	5	0

Recombine to single new level, OTHER.

Clustering Levels – Nominal Option

	0	1
A	28	7
B	16	0
C	94	11
D	23	21

Clustering Levels – Nominal Option

	0	1
A	28	7
B	16	0
C	94	11
D	23	21

Most common categories

	0	1
A	28	7
B/C	110	11
D	23	21

Clustering Levels – Greenacre Method

	0	1
A	28	7
B	16	0
C	94	11
D	23	21

$\chi^2 = 31.7$

	0	1
A	28	7
B/C	110	11
D	23	21

$\chi^2 = 30.7$

	0	1
A/B	44	7
C	94	11
D	23	21

$\chi^2 = 28.9$

	0	1
A	28	7
C	110	11
B/D	39	21

$\chi^2 = 18.3$

Clustering Levels – Greenacre Method

	0	1
A	28	7
B	16	0
C	94	11
D	23	21

$\chi^2 = 31.7$

Least amount information lost

	0	1
A	28	7
B/C	110	11
D	23	21

$\chi^2 = 30.7$

Combining Categories – SAS

```
data churn_t;
   set churn_t;
   customer_service_calls_c = put(customer_service_calls, 2.);
   if customer_service_calls > 3
   then customer_service_calls_c = '4+';
run;

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

Combining Categories – SAS

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
international_plan	1	7.9899	0.0047
voice_mail_plan	1	1.4798	0.2238
total_day_charge	1	1.9598	0.1615
customer_service_cal	4	11.0065	0.0265

Combining Categories – SAS

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-5.2962	1.6611	10.1657	0.0014
international_plan	**yes**	1	2.4044	0.8506	7.9899	0.0047
voice_mail_plan	**yes**	1	-1.2762	1.0491	1.4798	0.2238
total_day_charge		1	0.0560	0.0400	1.9598	0.1615
customer_service_cal	**1**	1	0.0744	1.1007	0.0046	0.9461
customer_service_cal	**2**	1	-0.0559	1.1677	0.0023	0.9618
customer_service_cal	**3**	1	0.2087	1.3518	0.0238	0.8773
customer_service_cal	**4+**	1	3.0129	1.1695	6.6366	0.0100

Combining Categories – R

```
train_u$customer.service.calls.c <-
        as.character(train_u$customer.service.calls)
train_u$customer.service.calls.c[
        which(train_u$customer.service.calls > 3)] <- "4+"

table(train_u$customer.service.calls.c, train_u$churn)
```

```
##
##       FALSE TRUE
##    0     17   23
##    1     46   21
##    2     24   18
##    3     14   12
##    4+     6   33
```

Combining Categories – R

```
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -6.17785    1.76972  -3.491 0.000481 ***
## factor(international.plan)yes      2.29115    0.92121   2.487 0.012879 *
## factor(voice.mail.plan)yes        -1.22099    1.05706  -1.155 0.248056
## total.day.charge                   0.09754    0.04454   2.190 0.028534 *
## factor(customer.service.calls.c)1 -0.50788    1.02479  -0.496 0.620178
## factor(customer.service.calls.c)2 -0.29731    1.06926  -0.278 0.780969
## factor(customer.service.calls.c)3  0.04458    1.20849   0.037 0.970572
## factor(customer.service.calls.c)4+ 1.38862    1.00660   1.380 0.167737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 86.595  on 213  degrees of freedom
## Residual deviance: 69.247  on 206  degrees of freedom
## AIC: 37.024
```

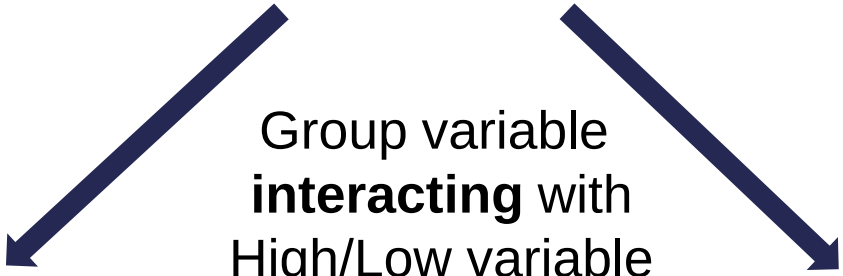
Watch out for Interactions!

	Yes	No
Group A	77	23
Group B	16	50

Group variable seems good

Watch out for Interactions!

	Yes	No
Group A	77	23
Group B	16	50



Group variable **interacting** with High/Low variable

	Yes	No
High	43	11
Low	0	41

	Yes	No
High	34	12
Low	16	9