BC Hydro Customer Classification Model Using SAS 9.2

VanSug Business Analytics Meeting

Scott Albrechtsen
BC Hydro, Load Analysis

November 2, 2011

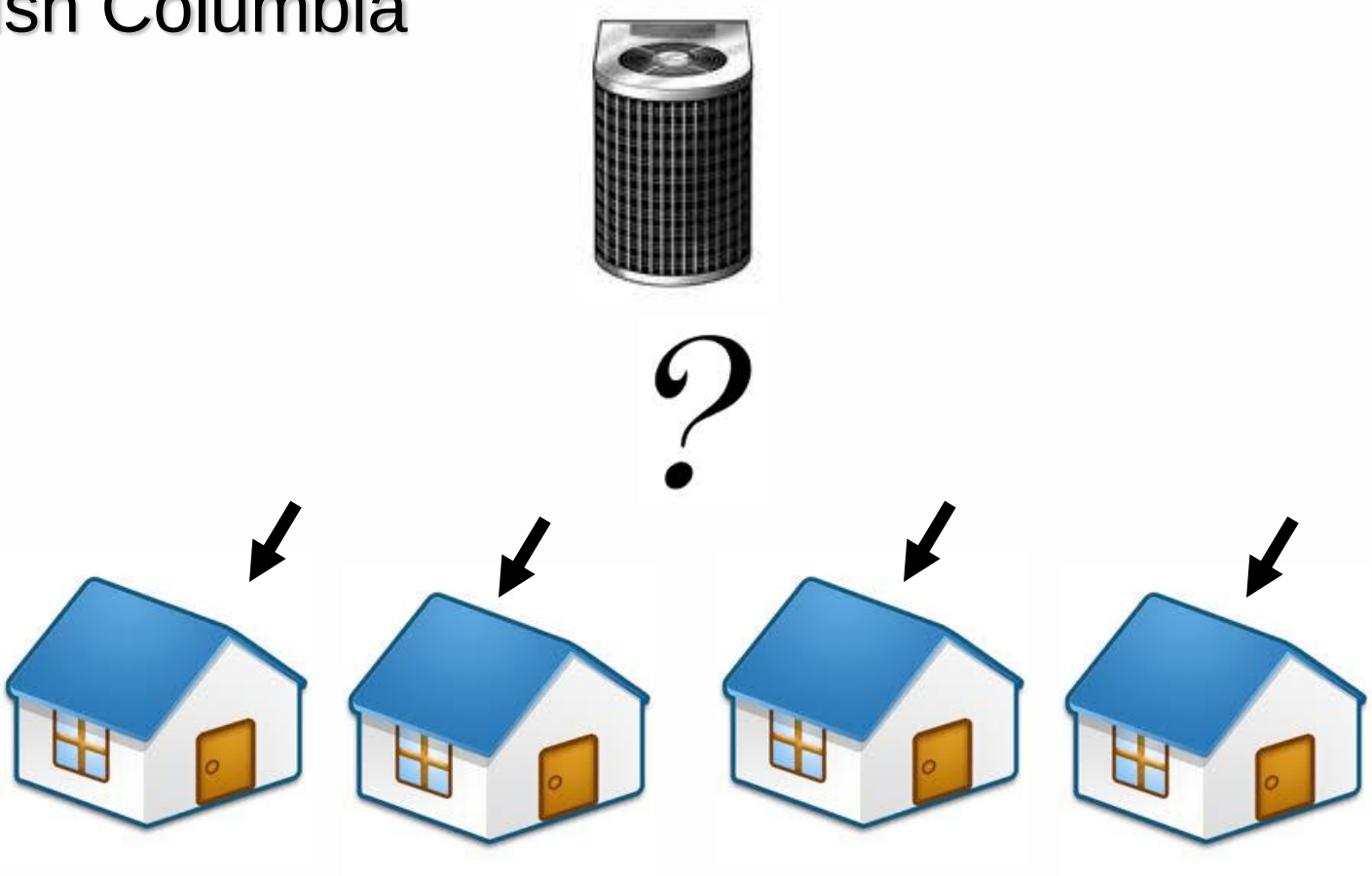
BC hydro

Agenda

- Business Problem & Why We Care
- The Data Sources Used
- The Modeling Concept
- The Logistic Regression
- The Scoring Process
- The Results

Business Problem

BC Hydro does not know which residential customers have Central Air Conditioning (A/C) in British Columbia

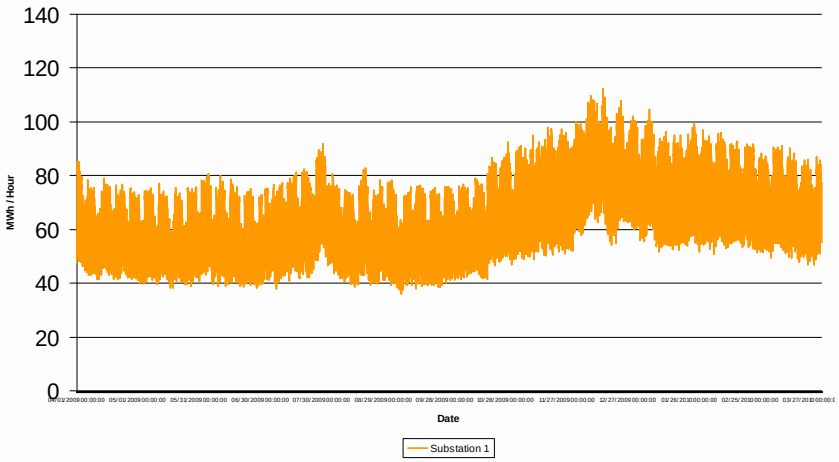


Why Do We Care?

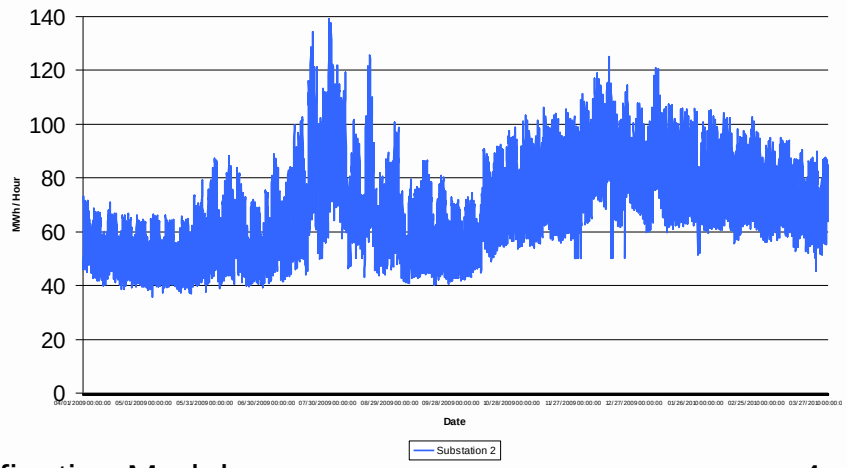
- Modeling which customers have Central A/C is important for
  - Infrastructure Planning
  - Load Forecasting
  - Benchmarking
  - Customer Segmentation
  - Weather Impact Assessments

The high A/C customer penetration on Substation 2 has a 47 MW Higher Summer Peak Power demand!

Substation with low A/C Penetration



Substation with High A/C Penetration



Why Do We Care?

BC hydro

- We can't answer the following questions:
  - "What are customer Central A/C penetration levels by substation?"
  - "How do we forecast for a very hot summer? A cooler summer?"
  - "I have A/C but I feel my home is very efficient. How do I compare to other A/C homes in my area."
  - "Why do homes in the South Interior consume differently than in Vancouver?"

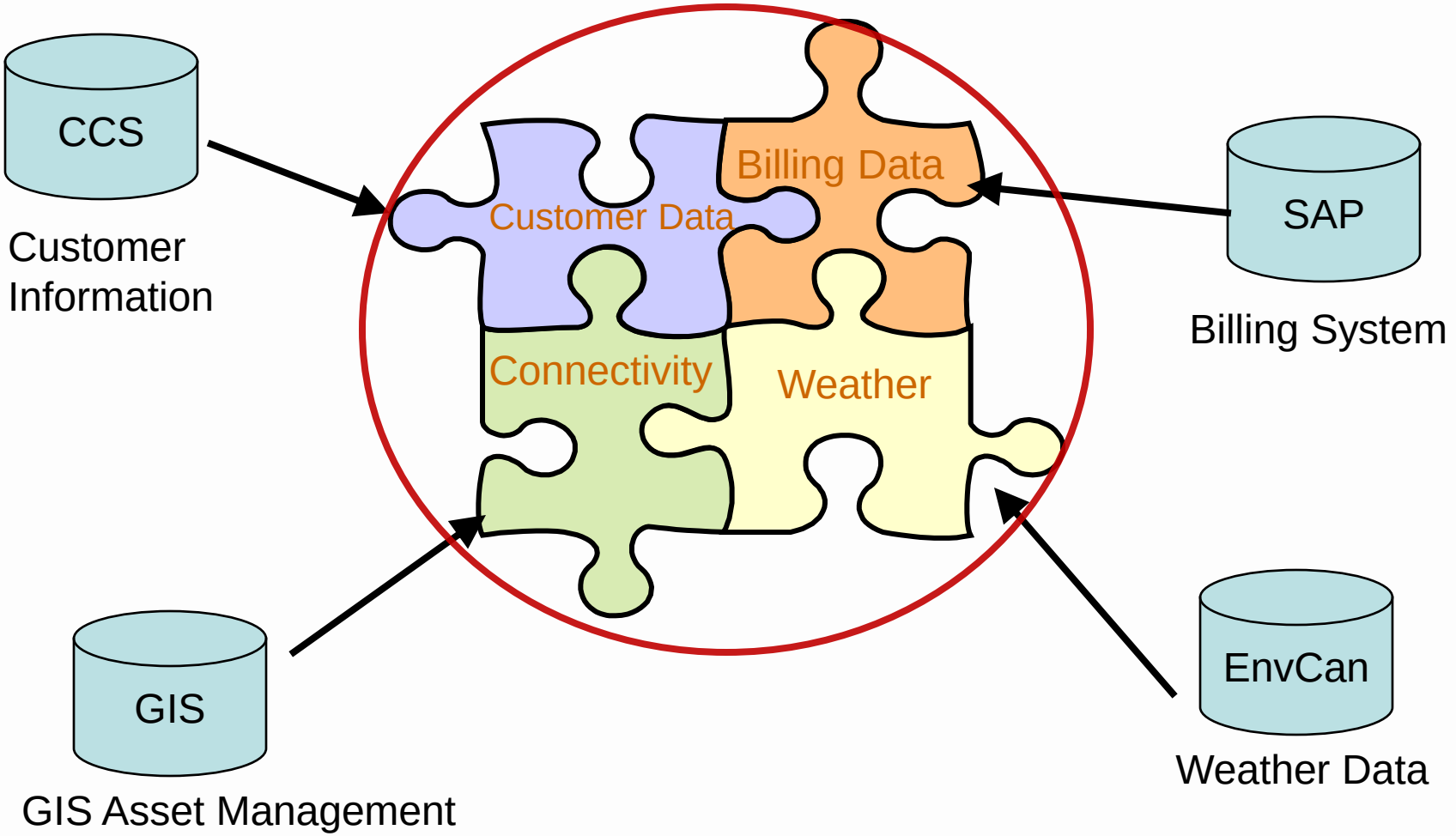
Solution?

- Create a predictive model to score all customers with a probability of having Central A/C

92% 28% 12% 82%

- Determine a cut-off matrix to classify each customer as having A/C or not

Population Data Sources Used



The Modeling Concept

BC hydro

≈10,000 Accounts

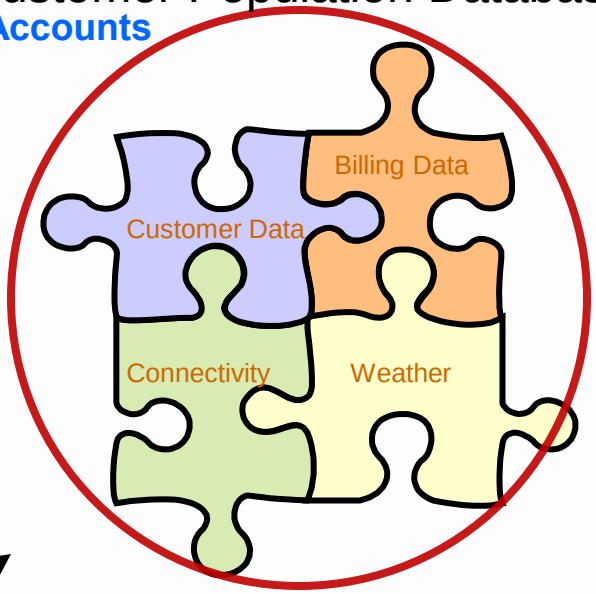
Customer Survey Data

A statistical sample of accounts where the appliance stock is known

Join

Customer Population Database

≈1.5 Million Accounts

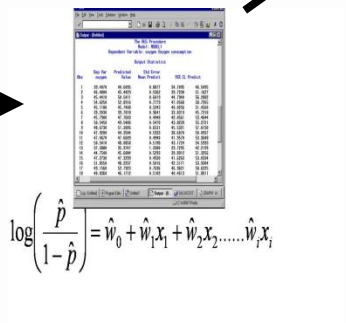


Produce

Score

Model

Sample Data for modeling



Classify

All Customers Classified

Where did we get the sample data?

- Stratified customer energy survey of ≈10,000 residential households
- Weighted to represent BC population & reduce survey bias
- Large appliance inventory was taken

BC hydro

FOR GENERATIONS

Survey ID: 000000
Pass Key: 000000

ATTN: <NAME>
<MAILING ADDRESS>
<MAILING TOWN, PROVINCE, POSTAL>

SERVICE ADDRESS
<SERVICE ADDRESS>
<SERVICE TOWN>

June 2010

Dear Participant:

As per an introductory letter previously sent to your household, you have been selected at random from our list of BC Hydro customers to be part of a very important energy survey.

36. Please indicate below the number of each home cooling appliance in your home. For each item that your home has, indicate the number of hours per day it is used strictly in the months that it typically is used in.

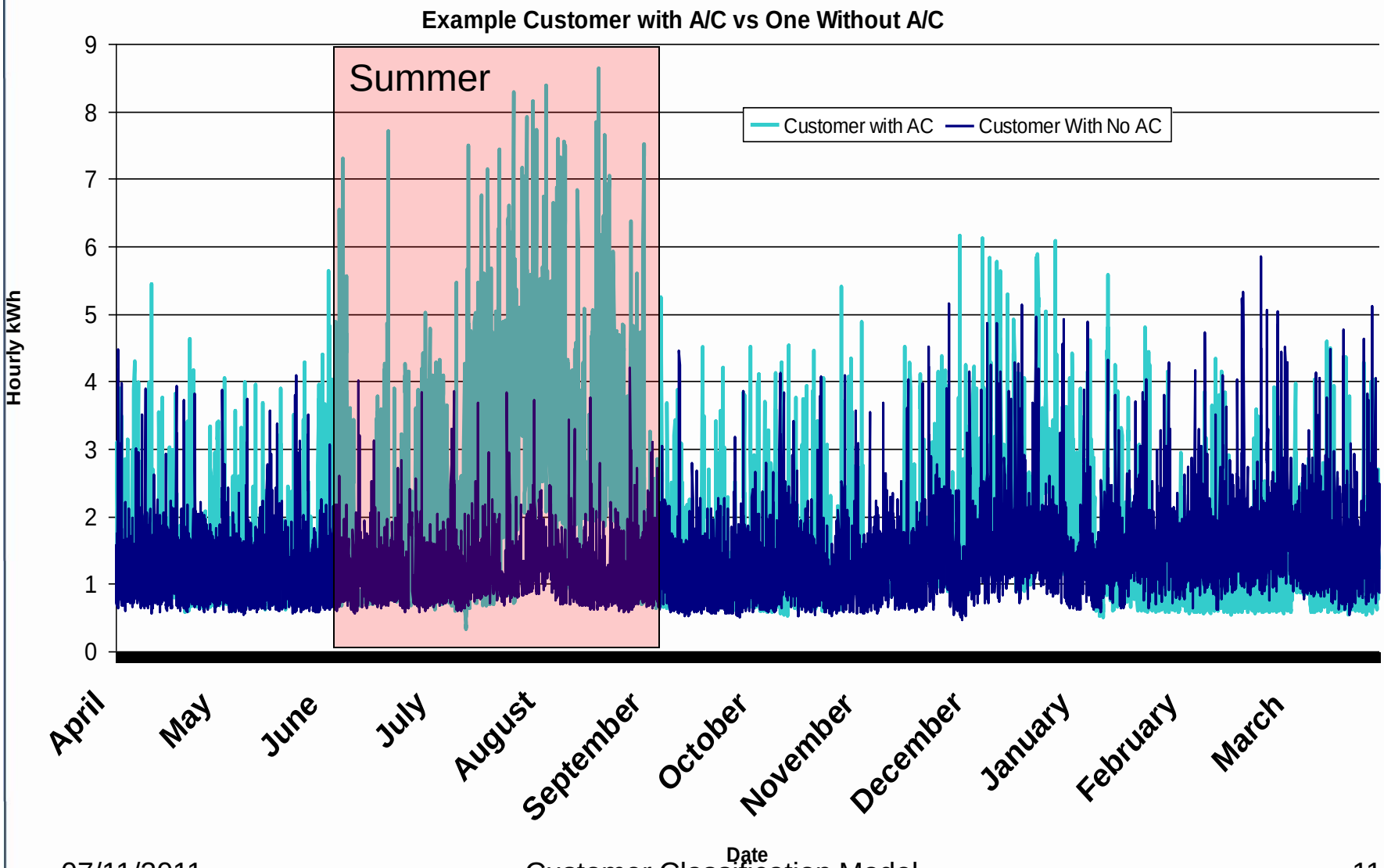
Be sure to indicate "0" if your home does not have the item.

	Number in use	Average number of hours on per day when used
a. Central air conditioners	____	____
b. Portable air conditioners	____	____
c. Room air conditioners	____	____
d. Humidifiers (in regular use)	____	____
e. Dehumidifiers (in regular use)	____	____
f. Portable fans		____

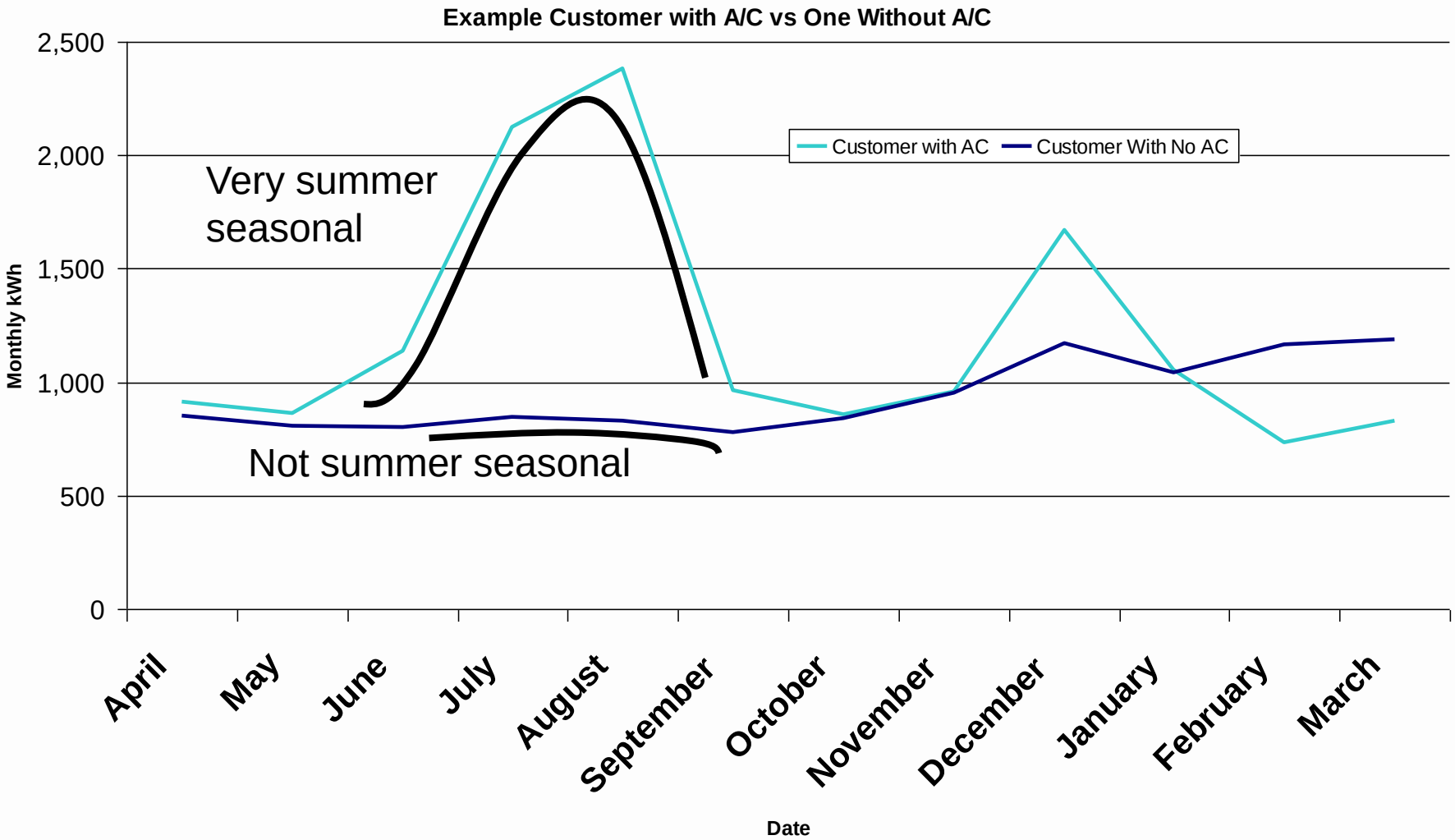
A Behavioural Model

- The A/C Classification Model is a Behavioural Model
- Variables are created to quantify:
  - Summer seasonal consumption patterns
  - Reaction to warm weather
  - Dummies for part of the province
    - Lower Mainland, South Interior, Van Island, North
  - Flags for type of household
    - Detached home, Row House, Apartment etc.
- Assumption: Homes who behave like other air conditioned homes have a high likelihood of having A/C

Difference between customers (AC & Non-AC)



Difference between customers (AC & Non-AC)



Model Inputs

Consumption Seasonality Variables

- Ratio of July & August Consumption to April & May
- Ratio of December & January Consumption to July & August
- If Summer ratio >1.5 Dummy
- Max(July , August) – Max(April, May)
- If Summer to shoulder kWh > 150 Dummy
- Average Summer kWh
- CV of Summer kWh
- Std.Deviation of kWh
- Average Summer kWh / Annual kWh

Weather Related Variables

- Correlation of Cooling Degree Days to monthly kWh
- If Correlation CDD>0.5 Dummy

Dummies for Location

- Regional Dummies
  - Lower Mainland, South Interior, Van Island, North

Dummies for Building Type

- Building Type Dummies
  - Single Family Dwelling, Row House, Mobile Home

Dependant Variable:

- If a house has A/C then Y=1, otherwise Y=0

Data Structure

Have A/C?
1=Yes
0=No

Accounts (Cases)	Y (target)	X_1	X_2	X_3	X_4	...	X*k*
1							
2							
3							
4							
5							
⋮							
n							

Data Manipulation

BC hydro

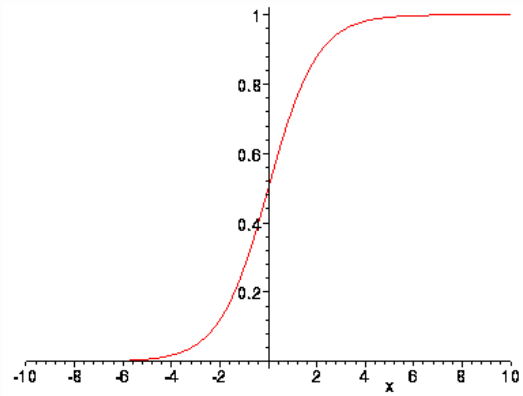
- Imputing for missing values
  - Can not have missing values in model variables
    - Hole-filling, variable exclusion, missing data indicators
    - data steps, proc stdsize, etc.
- Clustering Categorical Inputs
  - District or geographical areas
  - Rate types
    - proc cluster
- Variable Clustering
  - Principal components
    - proc varclus

Logistic Regressions

- Generalized linear model
- Estimated via Maximum likelihood
- Regression coefficients are estimated within the logistic function

$$\hat{p} = \frac{1}{1+e^{-z}}$$

Logistic Function

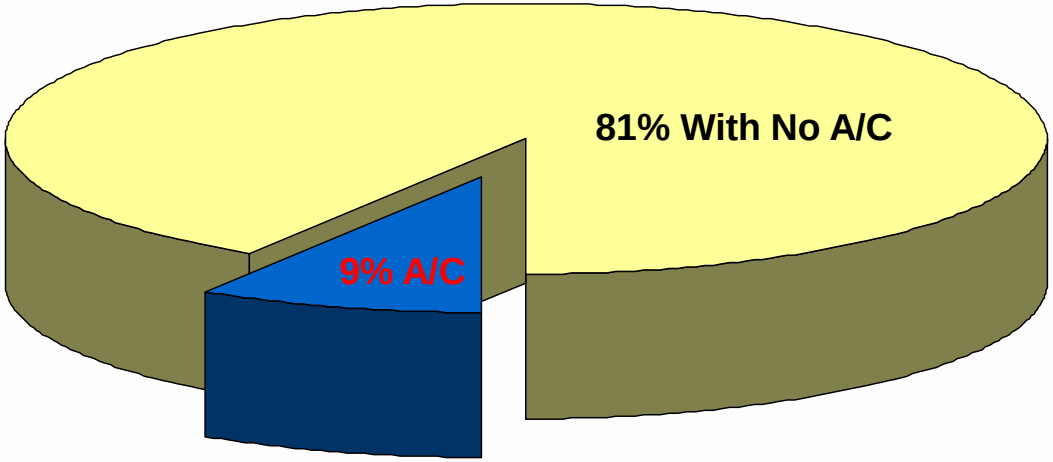


$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = z = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 \ldots\ldots \hat{w}_i x_i$$

Prediction is constrained between 0 and 1

The Sample Data Set

- In BC, only 9% of homes have Central Air Conditioning
  - Central A/C is a “rare event”
- We want our model data to have higher instances of A/C
  - We’ll keep all the A/C cases & sample out some of the non-cases
  - Creates a higher A/C % sample data set
  - We’ll correct for this during the scoring



The Sample Data Set

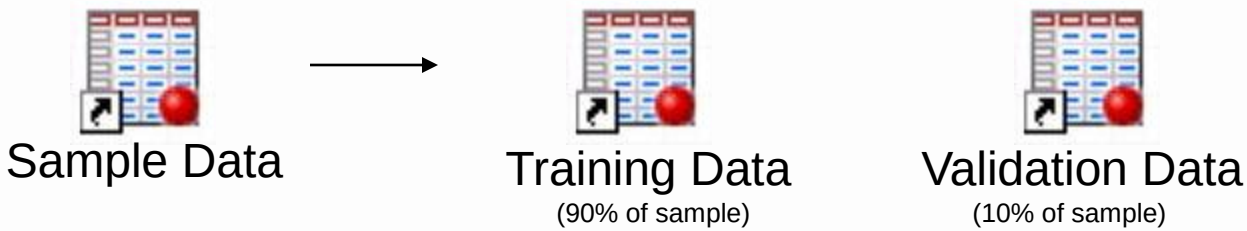
```
proc surveyselect data=whole_sample noprint
   samprate =.25 out =seg_no_AC seed = 121254 outall;
   where AC=0;
   strata region; run;

*Create the sample data set to be used;

data sample_set;
   set whole_sample(where =(AC=1)) seg_no_AC(where =(Selected=1)) ;
    drop Selected SelectionProb SamplingWeight; run;
```

- Now we have a sample data set with a higher proportion of A/C for modeling

The “Training & Validation” Data Sets




```
proc surveyselect data=sample_set noprint
   samprate =.90 out =sample_out seed = 5123344 outall;
   strata AC; run;

*Separate the data sets;

data train valid;
   set sample_out;
         if selected then output train;
         else output valid; run;
```

Logistic Regressions in SAS

BC hydro

- In SAS there are multiple ways to model logistic regressions but the most common is **proc logistic**

```
proc logistic data = train descending outest = model_out;
   title 'Fitting A/C model using Logistic Regression' ;
   model AC = &all_inputs.
                        /          ctable PPROB = (0 to 1 by .1)
                                   lackfit scale = n
                                   risklimits aggregate rsquare
                                   outroc = ROC;
                                   output out = model_inc1;
   run; quit;
```

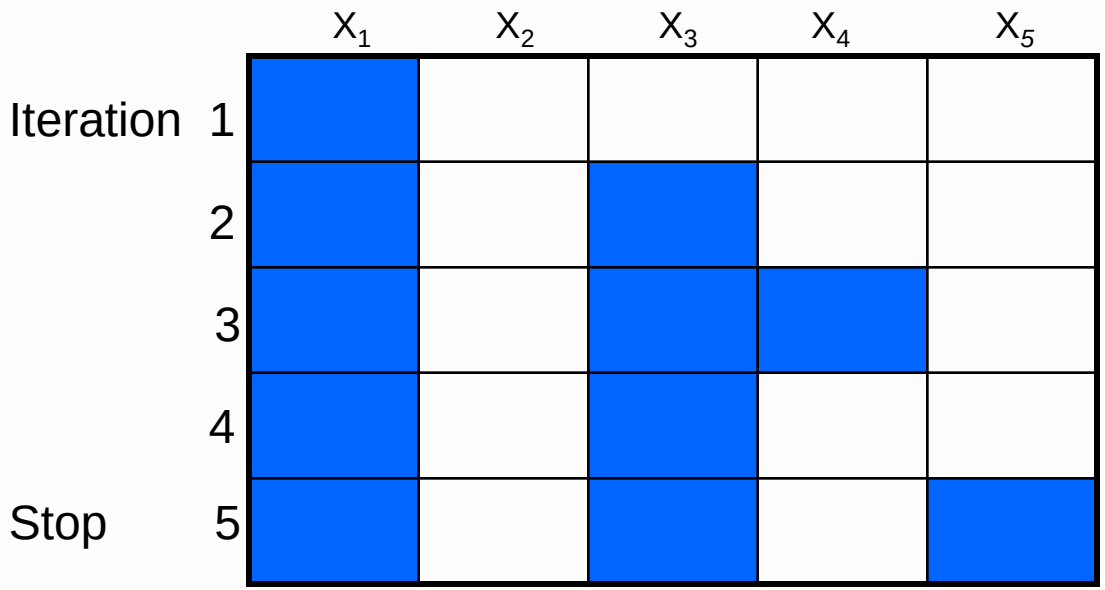
Variable Selection

- Potentially 100s of explanatory variables available for a model!
  - Don't want to over-fit
  - Don't want to over-complicate
  - Want to eliminate some co-linearity

- Solution: Variable Selection criteria

Variable Selection

- Variable selection in logistic regression
  - Find subsets of all the explanatory variable inputs statistically significant in predicting the target
  - Different methods: backwards, forewords, stepwise selection

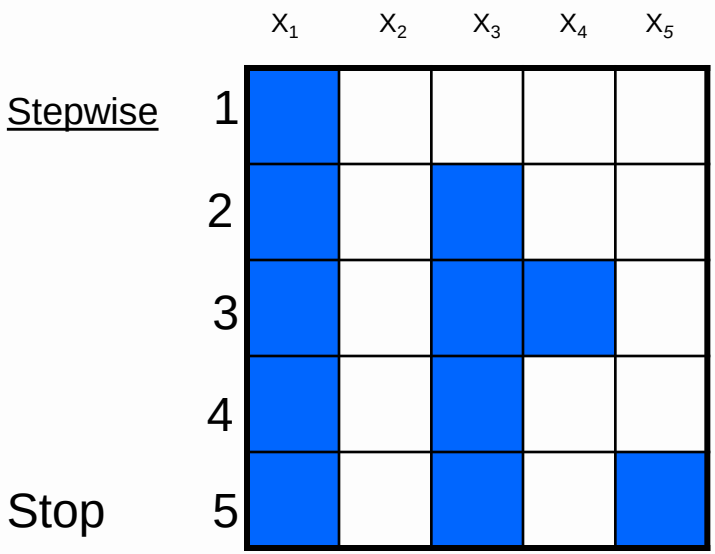


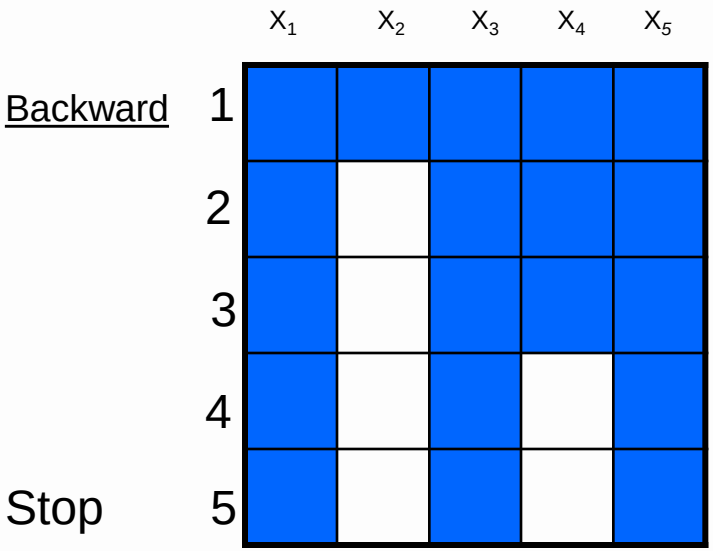
Variable Selection

```
proc logistic data = train descending outest = model_out;
   title 'Fitting A/C model using Logistic Regression' ;
   model AC = &all_inputs.
```

/ **selection** = stepwise

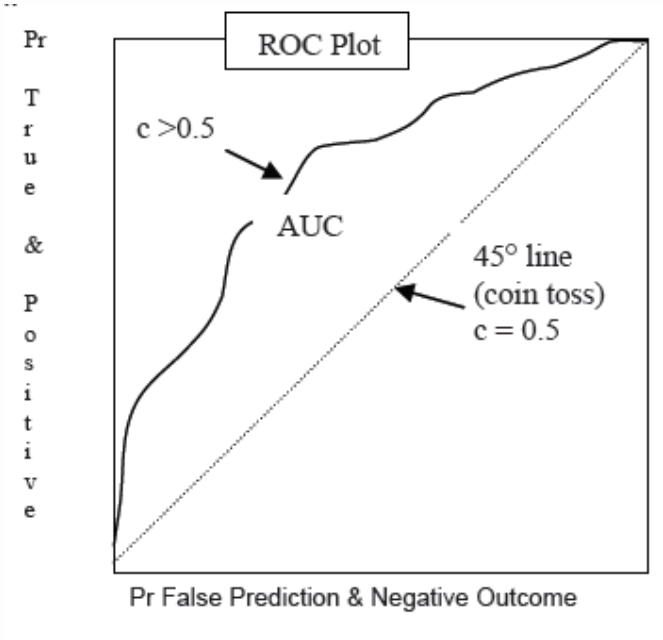
```
ctable PPROB = (0 to 1 by .1)
lackfit scale = n
risklimits aggregate rsquare
outroc = ROC;
output out = model_inc1; run; quit;
```





The ROC Curve & C Statistic

- The ROC Curve:
  - Trade off between sensitivity (captured response fraction) & specificity (false positive fraction)
- Each point along the ROC curve correspond to a fraction of cases.
  - The steeper the ROC the better the model
- The C statistic is the area under the ROC and depicts the predictive power of the model



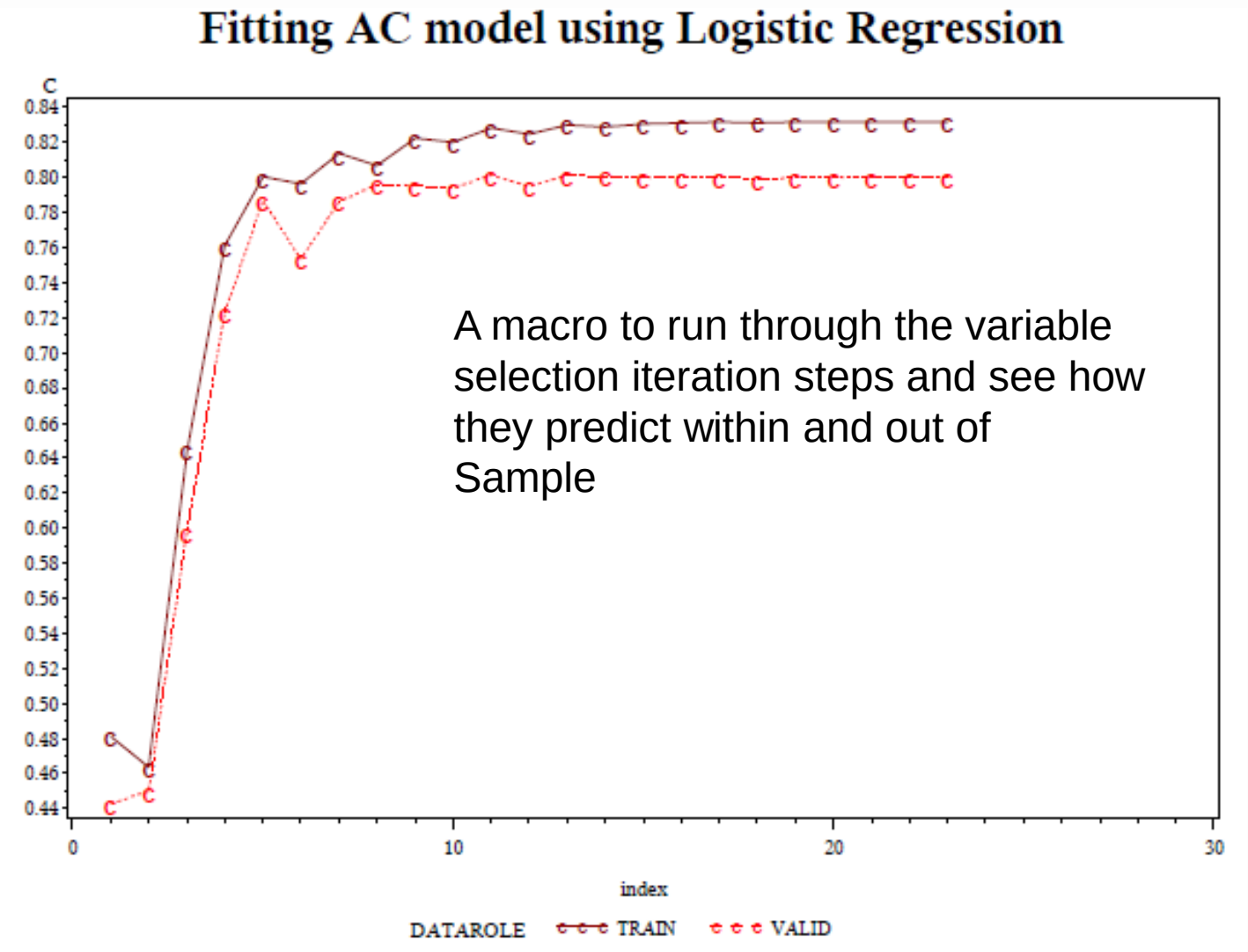
Scoring the Validation Data Set

BC hydro

- Validation data set will assess how well the model predicts out of sample
  - Cross-validation
  - Model assessment

```
proc logistic data = train descending outest = model_out;
   title 'Fitting A/C model using Logistic Regression' ;
   model &y. = &all_inputs.
                        /          selection = stepwise
                                   ctable PPROB = (0 to 1 by .1)
                                   lackfit scale = n
                                   risklimits aggregate rsquare
                                   outroc = ROC;
                                   output out = model_inc1;
                        score data =valid  out = Scored_valid;
run; quit;
```

Variable Selection



A macro to run through the variable selection iteration steps and see how they predict within and out of Sample

Champion Model

BC hydro

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	83.5	Somers' D	0.672
Percent Discordant	16.3	Gamma	0.673
Percent Tied	0.2	Tau-a	0.330
Pairs	5783250	c	0.836

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	4791.1211	4844	0.9891	0.7028
Pearson	5441.9136	4844	1.1234	<.0001

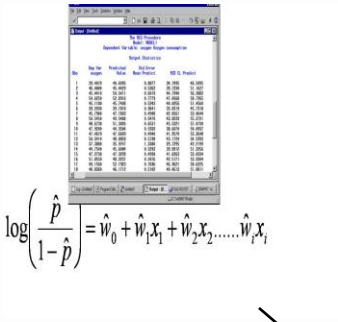
R Square 0.5234

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Average Summer kWh	1.001	1.001	1.001
Std. Deviation of kWh	0.999	0.998	0.999
Coefficient of Variation Summer kWh	13.6	5.996	30.848
Max(July, August) - Max(April, May) kWh	1.001	1.001	1.001
If Correlation with CDD>0.5 Dummy	1.5	1.268	1.774
If Summer to Shoulder Month kWh > 150	0.436	0.356	0.534
If in the Southern Interior Dummy	3.798	3.257	4.429
If a ROW House Dummy	0.183	0.111	0.303

Scoring the Population

Population Data

Champion Model



The whole point of
this modeling process!

Customer A/C Population
Data Set

Correct for Oversampling

BC hydro

- We know from other studies that only 9% of all BC has Central A/C

	Total						Lower Mainland						Vancouver Island						Southern Interior						North					
	'01	'03	'04	'06	'08	'10	'01	'03	'04	'06	'08	'10	'01	'03	'04	'06	'08	'10	'01	'03	'04	'06	'08	'10	'01	'03	'04	'06	'08	'10
Central air conditioner	5	6	4	8	9	9	3	3	2	5	5	5	1	2	5	5	7	9	28	31	39	36	35	36	1	1	1	2	4	4
Average number of units *	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.0	2.1	1.0	1.0	1.0

- We must adjust for oversampling when scoring the population database

Scoring the Population (Method #1)

BC hydro

```
proc logistic data = train descending;
    model &y. = &champ_vars.
                                /
                                        ctable PPROB = (0 to 1 by .1)
                                        lackfit scale = n
                                        risklimits aggregate rsquare
                                        outroc = ROC;
    score data =pop_set  out = scored_pop_set priorevent= .09;
run; quit;
```

Correct for the oversampling and adjust the scored population to 9% of A/C within the population

Scoring the Population (Method #2)

```
proc logistic data = train descending outest = model_out;
   model &y. = &champ_vars.; run;

proc score data = pop_set out= scored_pop_set score = model_out
    type = parms;
   var &y. = &champ_vars. ; run;

%let rho=0.39; *This this the proportion of A/C homes in the training sample;
%let pi=0.09; *This is the proportion of A/C to be scored in the population;

data pop_set;
    set pop_set;
   offset=log(( (1-&pi. )* &rho. ) / ( &pi. *(1- &rho. )));
   p=1 / (1+exp( - (AC – offset)));
   if p>= .50 then AC_pred=1; else AC_pred=0;
run;
```

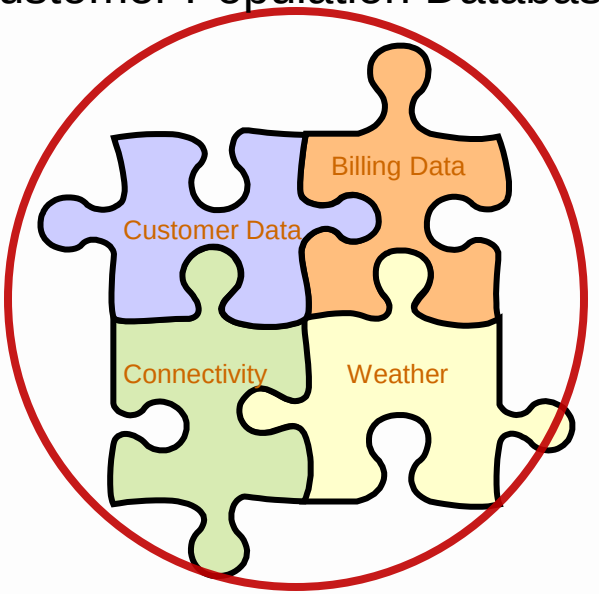
Resulting Data Set

- A complete residential population data set classified as having Central A/C or not
- This can be joined with other databases for infinite analysis

Account ID	Predicted A/C Central	Predicted Probability
1	0	0.04271
2	0	0.08184
3	1	0.84603
4	0	0.38308
5	0	0.18924
6	1	0.73029
7	0	0.29275
8	1	0.61423
9	0	0.40234
10	0	0.29793
11	0	0.33425
12	1	0.60665
13	0	0.1589

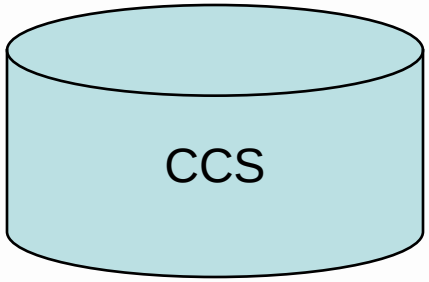
Scored Data Migration

Customer Population Database



Classify

All Customers Classified



Customer Information

≈1.5 Million Accounts

Applications

– "What are customer A/C penetration levels by substation?"

Substation Name	Sub Code	Number of Customers	Proportion of Central A/C
Mainwaring	MAN	39,699	0.28%
Whalley	WHY	30,375	1.25%
McLellan	MLN	25,906	1.95%
Clayburn	CBN	22,462	2.16%
George Tripp	GTP	21,741	0.67%
Newell	NEL	19,515	0.87%
Horsey	HSY	19,261	0.30%
Vernon Terminal	VNT	18,154	**53.45%**
Douglas St	DUG	14,647	**49.82%**
White Rock	WRK	13,611	1.44%
Westbank	WBK	13,541	**62.22%**

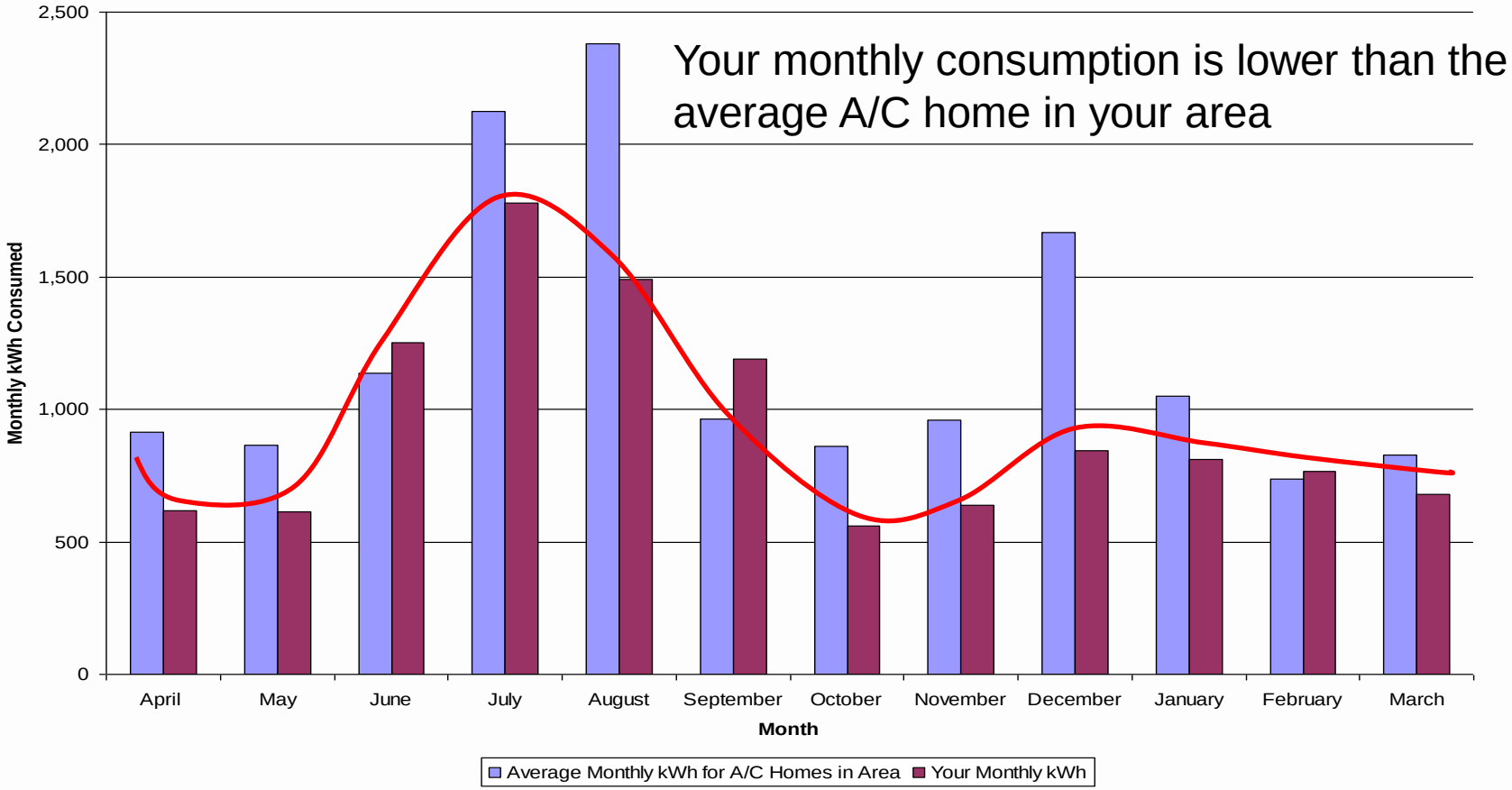
Applications

– "Why do homes in the South Interior consume differently than in Vancouver?"

City	Customers	Proportion of Central A/C
SURREY	100,253	1.56%
VANCOUVER	94,087	1.01%
VICTORIA	75,624	0.63%
RICHMOND	41,546	0.44%
KAMLOOPS	26,914	**56.85%**
MAPLE RIDGE	20,898	3.20%
VERNON	16,796	**50.94%**
WEST VANCOUVER	12,153	5.39%
CRANBROOK	7,464	20.19%

Applications

– "I have A/C but I feel my home is very efficient. How do I compare to other A/C homes in my area." *(Benchmarking)*



The End! Questions?