Agenda

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

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

• Determine a cut-off matrix to classify each customer as having A/C or not
Population Data Sources Used

- CCS: Customer Information
- GIS: GIS Asset Management
- SAP: Billing System
- EnvCan: Weather Data
- Customer Data
- Billing Data
- Connectivity
- Weather

Population Data Sources Used
The Modeling Concept

A statistical sample of accounts where the appliance stock is known

Sample Data for modeling

Model

Score

Join

≈10,000 Accounts

Customer Survey Data

≈1.5 Million Accounts

Customer Population Database

Billing Data

Customer Data

Connectivity

Weather

Modeling Concept

All Customers Classified

07/11/2011

Customer Classification Model

Using SAS 9.2
Where did we get the sample data?

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

---

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.

<table>
<thead>
<tr>
<th>Appliance</th>
<th>Number in use</th>
<th>Average number of hours on per day when used</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Central air conditioners</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Portable air conditioners</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. Room air conditioners</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. Humidifiers (in regular use)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Dehumidifiers (in regular use)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f. Portable fans</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• 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)

Example Customer with A/C vs One Without A/C

Summer

Hourly kWh

Customer with AC
Customer With No AC

07/11/2011

Customer Classification Model
Using SAS 9.2
Difference between customers (AC & Non-AC)

Example Customer with A/C vs One Without A/C

Very summer seasonal

Not summer seasonal

Date

Monthly kWh

Customer with AC

Customer With No AC

07/11/2011

Customer Classification Model

Using SAS 9.2
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

1  
2  
3  
4  
5  
...  
n

Y (target)  
X₁  
X₂  
X₃  
X₄  
...  
Xₖ

---

07/11/2011  
Customer Classification Model  
Using SAS 9.2
• **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

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

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

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

```sas
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

Sample Data

Training Data
(90% of sample)

Validation Data
(10% of sample)

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

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

![Diagram showing variable selection process]

- Iteration 1
- Stop
- Variables: X₁, X₂, X₃, X₄, X₅
**Variable Selection**

*proc logistic*

```sas
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;
```

07/11/2011
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

- 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

#### Odds Ratio Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Summer kWh</td>
<td>1.001</td>
<td>1.001</td>
</tr>
<tr>
<td>Std. Deviation of kWh</td>
<td>0.999</td>
<td>0.998</td>
</tr>
<tr>
<td>Coefficient of Variation Summer kWh</td>
<td>13.6</td>
<td>5.996</td>
</tr>
<tr>
<td>Max(July, August) - Max(April, May) kWh</td>
<td>1.001</td>
<td>1.001</td>
</tr>
<tr>
<td>If Correlation with CDD&gt;0.5 Dummy</td>
<td>1.5</td>
<td>1.268</td>
</tr>
<tr>
<td>If Summer to Shoulder Month kWh &gt; 150</td>
<td>0.436</td>
<td>0.356</td>
</tr>
<tr>
<td>If in the Southern Interior Dummy</td>
<td>3.798</td>
<td>3.257</td>
</tr>
<tr>
<td>If a ROW House Dummy</td>
<td>0.183</td>
<td>0.111</td>
</tr>
</tbody>
</table>
Scoring the Population

Population Data

Champion Model

The whole point of this modeling process!

Customer A/C Population Data Set

\[ \log \left( \frac{\hat{p}}{1-\hat{p}} \right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_n x_n \]
Correct for Oversampling

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

<table>
<thead>
<tr>
<th></th>
<th>Total '01 '02 '03 '04 '05 '06 '07 '08 '09 '10</th>
<th>Lower Mainland '01 '02 '03 '04 '05 '06 '07 '08 '09 '10</th>
<th>Vancouver Island '01 '02 '03 '04 '05 '06 '07 '08 '09 '10</th>
<th>Southern Interior '01 '02 '03 '04 '05 '06 '07 '08 '09 '10</th>
<th>North '01 '02 '03 '04 '05 '06 '07 '08 '09 '10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central air conditioner</td>
<td>5.0 6.0 4.0 8.0 9.0 9.0 5.0 5.0 5.0 1.0</td>
<td>1.0 2.0 5.0 5.0 7.0 9.0 28.0 31.0 39.0 36.0 35.0 35.0 1.0 1.0 1.0 2.0 4.0 4.0</td>
<td>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 1.0 1.0 1.0 1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average number of units*</td>
<td>1.0 1.0 1.0 1.0 1.0 1.0 1.1 1.0 1.0 1.0</td>
<td>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 1.0 1.0 1.0 1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- We must adjust for oversampling when scoring the population database
proc logistic data = train descending;
  model &y. = &champ_vars.
     / ctable PPORB = (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

<table>
<thead>
<tr>
<th>Account ID</th>
<th>Predicted A/C Central</th>
<th>Predicted Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.04271</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.08184</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.84603</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.38308</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.18924</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.73029</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.29275</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.61423</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0.40234</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0.29793</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0.33425</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.60665</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0.1589</td>
</tr>
</tbody>
</table>
Scored Data Migration

Customer Population Database

- Customer Data
- Billing Data
- Connectivity
- Weather

Classify

All Customers Classified

≈1.5 Million Accounts

CCS

Customer Information

07/11/2011

Customer Classification Model
Using SAS 9.2
“What are customer A/C penetration levels by substation?”

<table>
<thead>
<tr>
<th>Substation Name</th>
<th>Sub Code</th>
<th>Number of Customers</th>
<th>Proportion of Central A/C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainwaring</td>
<td>MAN</td>
<td>39,699</td>
<td>0.28%</td>
</tr>
<tr>
<td>Whalley</td>
<td>WHY</td>
<td>30,375</td>
<td>1.25%</td>
</tr>
<tr>
<td>McLellan</td>
<td>MLN</td>
<td>25,906</td>
<td>1.95%</td>
</tr>
<tr>
<td>Clayburn</td>
<td>CBN</td>
<td>22,462</td>
<td>2.16%</td>
</tr>
<tr>
<td>George Tripp</td>
<td>GTP</td>
<td>21,741</td>
<td>0.67%</td>
</tr>
<tr>
<td>Newell</td>
<td>NEL</td>
<td>19,515</td>
<td>0.87%</td>
</tr>
<tr>
<td>Horsey</td>
<td>HSY</td>
<td>19,261</td>
<td>0.30%</td>
</tr>
<tr>
<td>Vernon Terminal</td>
<td>VNT</td>
<td>18,154</td>
<td>53.45%</td>
</tr>
<tr>
<td>Douglas St</td>
<td>DUG</td>
<td>14,647</td>
<td>49.82%</td>
</tr>
<tr>
<td>White Rock</td>
<td>WRK</td>
<td>13,611</td>
<td>1.44%</td>
</tr>
<tr>
<td>Westbank</td>
<td>WBK</td>
<td>13,541</td>
<td>62.22%</td>
</tr>
</tbody>
</table>
“Why do homes in the South Interior consume differently than in Vancouver?”

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

Your monthly consumption is lower than the average A/C home in your area.
The End! Questions?