



# 10 Tips for Building Successful Predictive Analytics Solutions

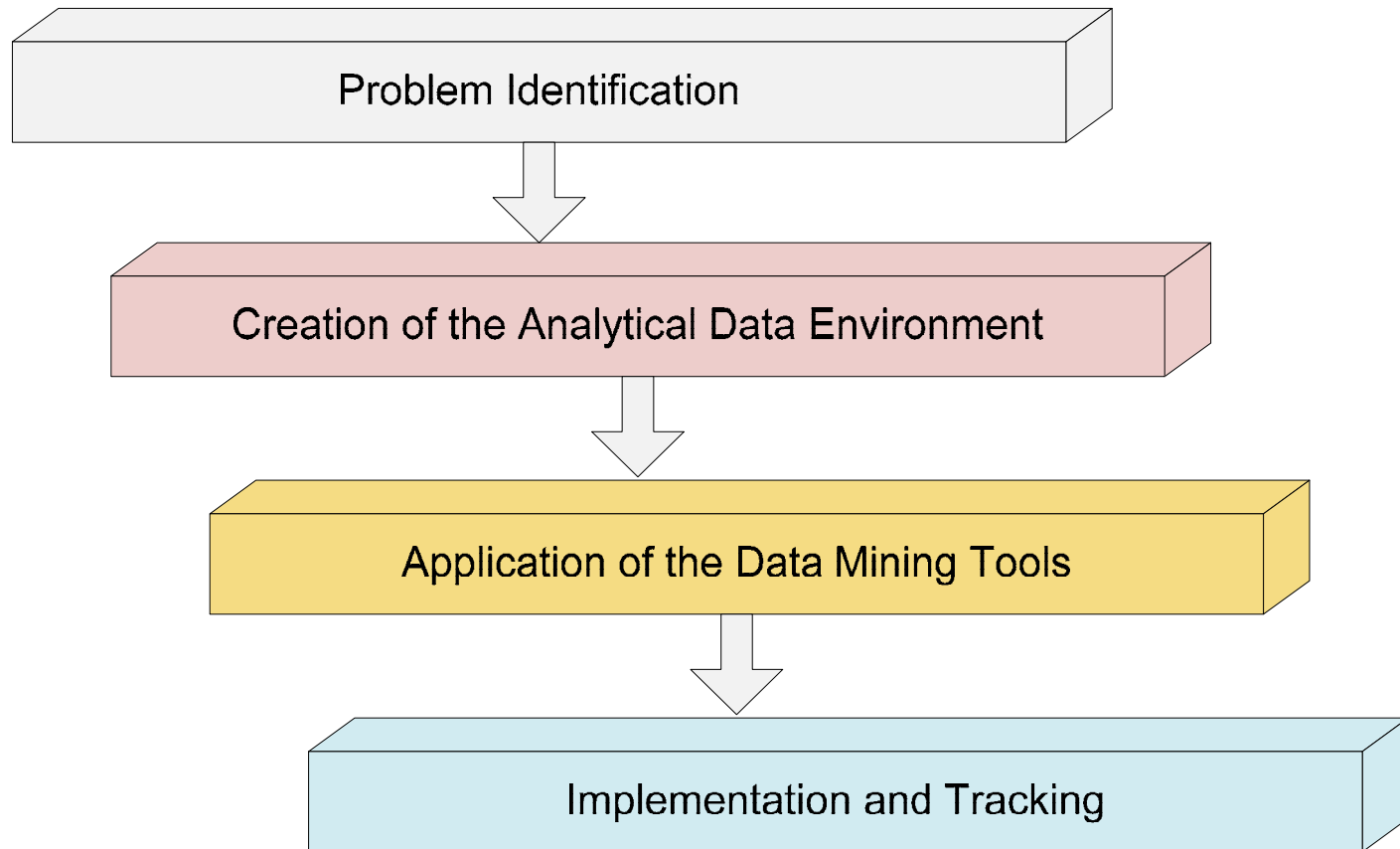
Boire Filler Group

# Background

- Predictive Analytics - Context
  - To Make better predictions about future events or activities
    - Relies on advanced statistics that help you make better business decisions
    - Identifies characteristics and or key areas to assist you in targeting customers or prospects
      - Customer Acquisition, Cross-selling, Up-selling, Retention, loyalty, etc.
- A Growing Area
  - More data, better tools available
  - Provides greater accountability
    - Program Measurement
    - Optimization of marketing \$\$\$ spending

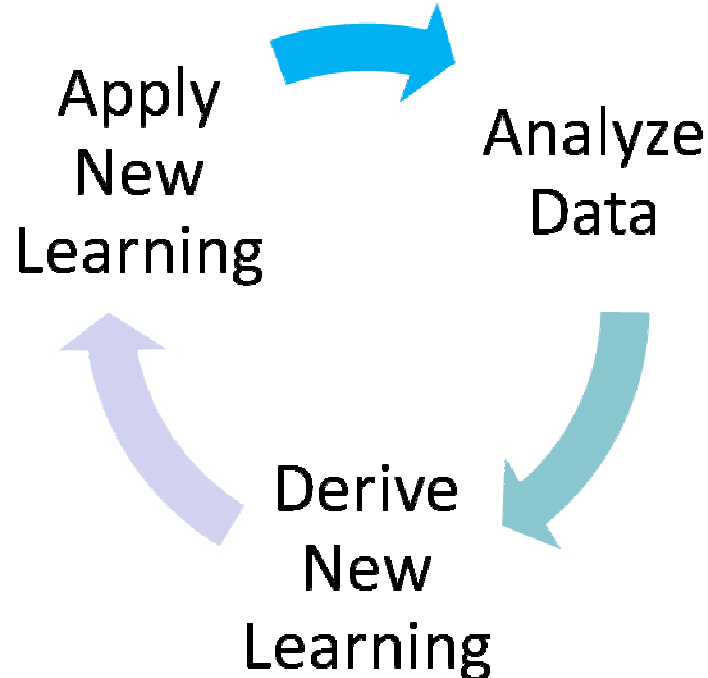
# Predictive Analytics

- A discipline that requires **STRUCTURE** and **PROCESS**
  - At BFG we utilize the following four-step process to manage projects:



# The Process is Dynamic and Flexible

- Our “Tips” have come from many years of “derived” learning
  - Data discovery/audit process
    - Intensity will vary based on data
  - New learning or unexpected findings generated from every exercise
  - Alter plans/approach based on learning



Issues and concerns are always addressed as part of this ongoing cycle of learning.

## Ten Tips

---

1. Identify the Real Problem
2. Get Stakeholders Onside
3. Understand the Data
4. Identify Quick Wins
5. Use Statistics Judiciously
6. Combine Art and Science
7. Establish Performance Benchmarks
8. Interpret Results Correctly
9. Relative Results vs Absolute Results
10. Action Tools and Measure Performance

# Tip 1: Identify the Real Problem

- Gather facts:
  - Listen, Listen, Listen!
  - Then dig a little deeper
  - This might require
    - Stakeholder interviews
    - Gathering all information and reports that are pertinent to the issue
    - Conducting new analysis with available data
- Once the problem is identified.....can analytics solve the problem?
  - e.g. Company with a 50% Defection Rate

## Tip #1: Identify the Real Problem - Example

- A Telco has experienced 50% increase in their overall acquisition costs in the last 2 years
- Their acquisition campaigns have been using direct mail and telemarketing as their key distribution vehicles for about 10 years
- These channels now comprise 75% of the marketing budget
- Defection among New customers within the 1st three months of service has increased by 50% in the past two years

# Tip 1: Identifying the Real Problem (continued)

- Key Information:
  - Increasing acquisition costs
  - Direct marketing/telemarketing are key marketing vehicles
  - Attrition is increasing
- What is the problem?
  - Identify the “right” customers to acquire
  - We need to identify what the 'right' customer means
  - “Right” in this case means that we identify prospects that are:
    - Most likely to respond to an acquisition campaign
    - Are most likely to remain a customer after 3 months
- The Business Problem/Challenge is:
  - Develop an acquisition model that optimizes the likelihood of a prospect becoming a new customer and remaining an active subscriber after 3 months

## Tip 2: Getting Stakeholders Onside

- Predictive Analytics projects always involve stakeholders from different disciplines:
  - Analytics
  - I/T
  - Marketing
  - Finance
  - Operations
  - Senior Management
- Degree of Involvement depends on complexity of project
  - Tactical vs. Strategic
- Definition of Success
  - Senior Management direction/buy-in
- Requires Ongoing Communications

## Tip 3: Understand the Data

- Practitioners constantly work with new data sets where integrity and quality of data needs to be checked and reviewed
  - Understanding the quality and availability of data is critical
- At BFG, we have automated this “Data Discovery” process
  - Loading of Data into BFG systems
  - Evaluation of data based on Initial Data diagnostics
  - Frequency Distribution Reports for each field

# Tip 4: Understand the Data (continued)

<b>Data Set Name</b>	LIB1.RPL_CSTC_PHONE_HIST	<b>Observations</b>	438411
<b>Member Type</b>	DATA	<b>Variables</b>	7
<b>Engine</b>	V9	<b>Indexes</b>	0
<b>Created</b>	Monday, March 20, 2006 01:39:05 PM	<b>Observation Length</b>	256
<b>Last Modified</b>	Monday, March 20, 2006 01:39:05 PM	<b>Deleted Observations</b>	0
<b>Protection</b>		<b>Compressed</b>	NO
<b>Data Set Type</b>		<b>Sorted</b>	NO
<b>Label</b>			
<b>Data Representation</b>	WINDOWS_32		
<b>Encoding</b>	wlatin1 Western (Windows)		

Direct Output from Proc Contents procedure In SAS

Engine/Host Dependent Information	
<b>Data Set Page Size</b>	16384
<b>Number of Data Set Pages</b>	6959
<b>First Data Page</b>	1
<b>Max Obs per Page</b>	63
<b>Obs in First Data Page</b>	57
<b>Number of Data Set Repairs</b>	0
<b>File Name</b>	C:\SASTemp\CST\rpl_cstc_phone_hist.sas7bdat
<b>Release Created</b>	9.0101M3
<b>Host Created</b>	XP_PRO

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
1	AGRMT_ID	Num	8		BEST32.
6	NFH_SEQ_ID	Num	8		BEST32.
2	SYSTEM_DT	Num	8	MMDDYY10	MMDDYY10
4	TRXN_DETAILS	Char	200		\$200.00
3	TRXN_DT	Num	8	MMDDYY10	MMDDYY10
5	USER_ID	Char	20		\$20.00

# Tip 4: Understand the Data (continued)

Data Diagnostics Report – determines how well different fields are populated

Variable Name	No. of Observations	No. of Unique Values	No. of Missing Value	No. of Non-Missing Value
AGRMT_ID	438411	180119	0	438411
NFH_SEQ_ID	438411	438411	0	438411
SYSTEM_DT	438411	1060	0	438411
TRXN_DETAILS	438411	398791	0	438411
TRXN_DT	438411	1048	0	438411
USER_ID	438411	53	0	438411

Frequency Distribution Report – determines the different possible outcomes/values in each field (i.e. for each variable)

txn_details	Frequency	Percent	Cumulative Frequency	Cumulative Percent	Percent of Unique Sub_ID
.	231744	52.86%	231744	52.86%	60.33%
<b>CHANGE PLAN</b>	66463	15.16%	298207	68.02%	17.31%
<b>OTHER</b>	61377	14.09%		82.11%	16.08%
<b>FSR</b>	45594	10.41%	405178	92.52%	11.88%
<b>ADD UNITS</b>	13985	3.19%	419163	95.71%	3.64%
<b>UNKNOWN</b>	11223	2.56%	430386	98.27%	2.92%
<b>STATEMENT</b>		1.71%	437882	99.97%	1.95%
<b>CANTONESE</b>		0.03%	438411	100.00%	0.03%

## Tip 4: Understand the Data (continued)

- Examine the value of key metrics over time to determine or identify significant business/market changes

	Period 1	Period 2	Period 3
<b>Record Count</b>			
<b>Average Purchase Amount</b>			
<b>Average Age</b>			
<b>Average Tenure</b>			
<b>Etc.</b>			

# Tip 4: Identify Quick Wins

- Best Example: Identify and Target Best Customers
- In most cases, this can be done by simply looking at purchase behaviour in last 12 months

% of Customers (in descending order)	# of Customers	Average Value(\$)	Total Spend	% of Total from Segment	Value Group	% of Value from Group
0-10%	100,000	\$3,400	\$340,000,000	34.0%	High (top 20%)	56.00%
11-20%	100,000	\$2,200	\$220,000,000	22.0%		
21-30%	100,000	\$1,400	\$140,000,000	14.0%	Medium (next 20%)	24.00%
31-40%	100,000	\$1,000	\$100,000,000	10.0%		
41-50%	100,000	\$750	\$75,000,000	7.5%	Low (bottom 60%)	20.00%
51-60%	100,000	\$500	\$50,000,000	5.0%		
61-70%	100,000	\$400	\$40,000,000	4.0%		
71-80%	100,000	\$200	\$20,000,000	2.0%		
81-90%	100,000	\$100	\$10,000,000	1.0%		
91-100%	100,000	\$50	\$5,000,000	0.5%		
	1,000,000	\$1,000	\$1,000,000,000	100.0%		

## Tip 5: Use Statistics Judiciously

- Understand why you are using certain statistical applications and what you hope to glean from them
- Traditional applications in Predictive Modeling:
  - Multivariate analysis
  - Linear, Logistic, CHAID, Neural Net
- Statistics more basic analytics exercises (understanding relationships between variables, segmentation, etc):
  - Correlation Analysis, Basket Analysis, Factor Analysis, CHAID

# Tip 5: Use Statistics Judiciously – Correlation Analysis

Correlation Analysis is a uni-variate technique that measures the strength of the relationship between Variables and the behaviour we are trying to predict.

The example below shows how variables in the file are correlated with auto insurance claims.

Rank	SAS Variable Name	Variable Description	Correlation (+ or -)
1	FSA_HIGH	High Claims FSA	0.1194
2	per_totper	FAMILY: % of Total number of persons in private households - 20% Sample Data of TOTAL Population	-0.0700
12	per_grtot20	% of EDUCATION: Total population 20 years and over by highest level of schooling of Total Population	-0.0626
15	Claim_Numbers_prev	Historical Number of Claims (BEFORE PRE Period)	0.0576
20	per_abnapop2	AB_NON_POPULATION: % of Total non-Aboriginal population of TOTAL Population	-0.0537
21	per_pft	FEMALE POPULATION: % of Female Total Population of TOTAL Population POPULATION_BY_IMM_STATUS: % of Total Immigrants 20 years and over, age at immigration of TOTAL Population	-0.0532
22	per_ti20_	Population	-0.0529
23	TOTRIDLIMITS	TOTAL_RIDER_LIMITS	0.0518
24	per_pftot	FEMALE POPULATION: % of Female Total Population of TOTAL Population	-0.0516
26	white_collar	% of White Collars of TOTAL Lab.Force 15yrs.+ by Industry	-0.0490
34	IBCCoverDesc_BASIC	IBCCoverDesc=BASIC	-0.0449
35	per_pt0_4	POPULATION: % of 0-4 Years of Age (Total Population) of TOTAL Population	-0.0443
36	per_pt35_39	POPULATION: % of 35-39 Years of Age (Total Population) of TOTAL Population	-0.0440
37	per_gs2nd	GENERATIONS: % of Total Population 15 years and over, 2nd generation of TOTAL Population	-0.0436
38	IBCTerr_HIGH	High Claims IBCTerr	0.0435
39	per_fiwinc1	% of POP.BY TOT.INC With income of Total Population	-0.0434
43	Centre_HIGH	High Claims Centre	0.0427
44	IBC_160850	EVER_HAD IBCCCLASS=160850	0.0422
46	per_grutot	% of EDUCATION: University of Total Population	-0.0421
48	FSA_LOW	Low Claims FSA	-0.0416
49	per_pf18_24	FEMALE POPULATION: % of 18-24 Years of Age (Female Population) of TOTAL Population LABOUR FORCE BY UNPAID CHILDCARE: % of Pop 15 yrs & over by hours of unpaid childcare of TOTAL Pop	-0.0413
50	per_cctot	15yrs+ by hours of unpaid childcare	-0.0413
51	per_vmtot	VISIBLE MINORITY: % of Total visible minority population of TOTAL Population	-0.0407
54	IBCTerr_111	IBCTerr=111	-0.0399
55	PCRV_CNT	PC_RIDERCNT=RVC	0.0396
56	per_ch0_6	FAMILY: % of Under 6 years of age of TOTAL Population	-0.0394
57	Canceled	Ever Canceled (1/0)	0.0394



# Tip 5: Use Statistics Judiciously - Factor Analysis

Factor Analysis is a valuable data reduction technique that can be applied when there is an excessive number of variables in a file. The technique allows one to determine new variables based on broad grouping of different variables. In the simplistic example below, 9 variables are grouped into 3 factors based on a statistical threshold called the “Eigen value”.

In this example, we use the technique against 9 variables and produce an “affluence” variable by combining the variables income, education and wealth, and broad product categories ABC and DEF.

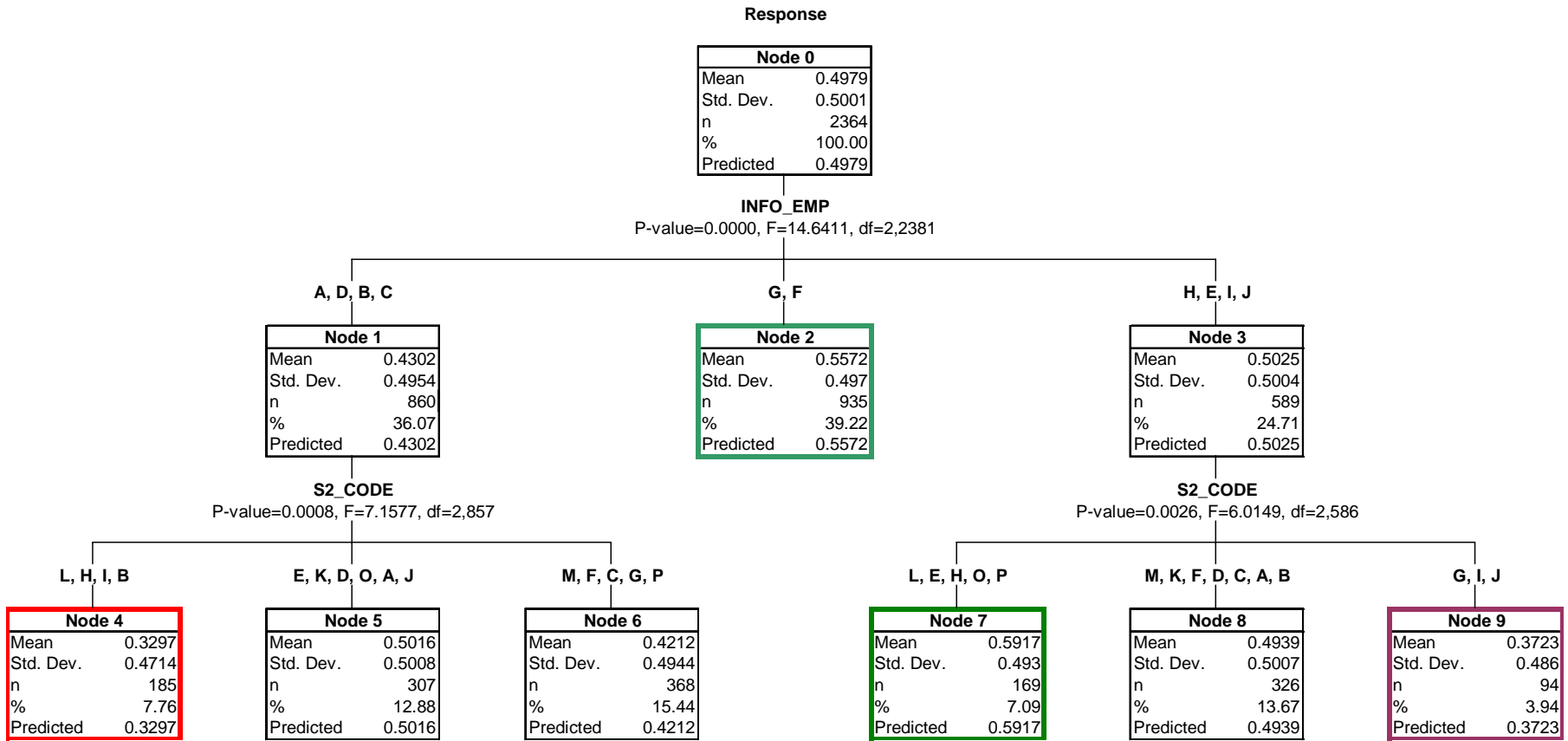
The groupings are determined based on the highest (absolute) variable scores that are identified in each factor.

	Factor1	Factor 2	Factor 3
1) Income	0.905	0.255	0.255
2) Education	0.855	0.373	0.212
3) Wealth	0.956	0.303	0.185
4) Product A	0.303	0.855	0.205
5) Product B	0.295	0.805	0.245
6) Product C	0.323	0.755	0.285
7) Product D	-0.105	-0.355	0.755
8) Product E	-0.155	-0.405	0.705
9) Product F	-0.085	-0.304	0.725

■ What are these results telling us?

# Tip 5: Use Statistics Judiciously – CHAID Analysis

CHAID analysis is a type of multivariate decision tree technique. It's advantages are that its output is highly visual and easy to interpret. CHAID is often used in the context of direct marketing for the selection of groups of prospects/customers to optimize a given behaviour (response, retention, conversion, etc). CHAID is generally applied when large amounts of data are available.



## Tip 6: Combine Art and Science

- Be Creative
  - Look for all available data sources
  - Leverage existing customer data where possible
    - Fish where the Fish are, Fish where the new Fish are
    - Sometimes the most obvious things are right in front of you!!
  - Augment internal data with external data – like research
- Example:
  - Retailer collects no customer data, but market research has identified three **distinct** variables which represent their shoppers:
    - Above average income
    - Products appeal to New Immigrants
    - Families - Larger HH size

# Tip 6: Use Art and Science to Build Solutions

## Potential Solution:

- Using a “RecencyFrequencyMoney” index approach, create postal code index based on three Statistics Canada Variables:
  - Median taxfiler income of postal code
  - % 3 Plus HH’s
  - % of population landed immigrants within postal code

	Income	% 3+ Household	% Landed Immig.
Average Postal Code	\$40,000	52%	5%
M5A 1J2	\$50,000	60%	10%
Index	1.25	1.15	2

The index for M5A 1J2 is  $(.33 \times 1.25) + (.33 \times 1.15) + (.33 \times 2) = 1.45$

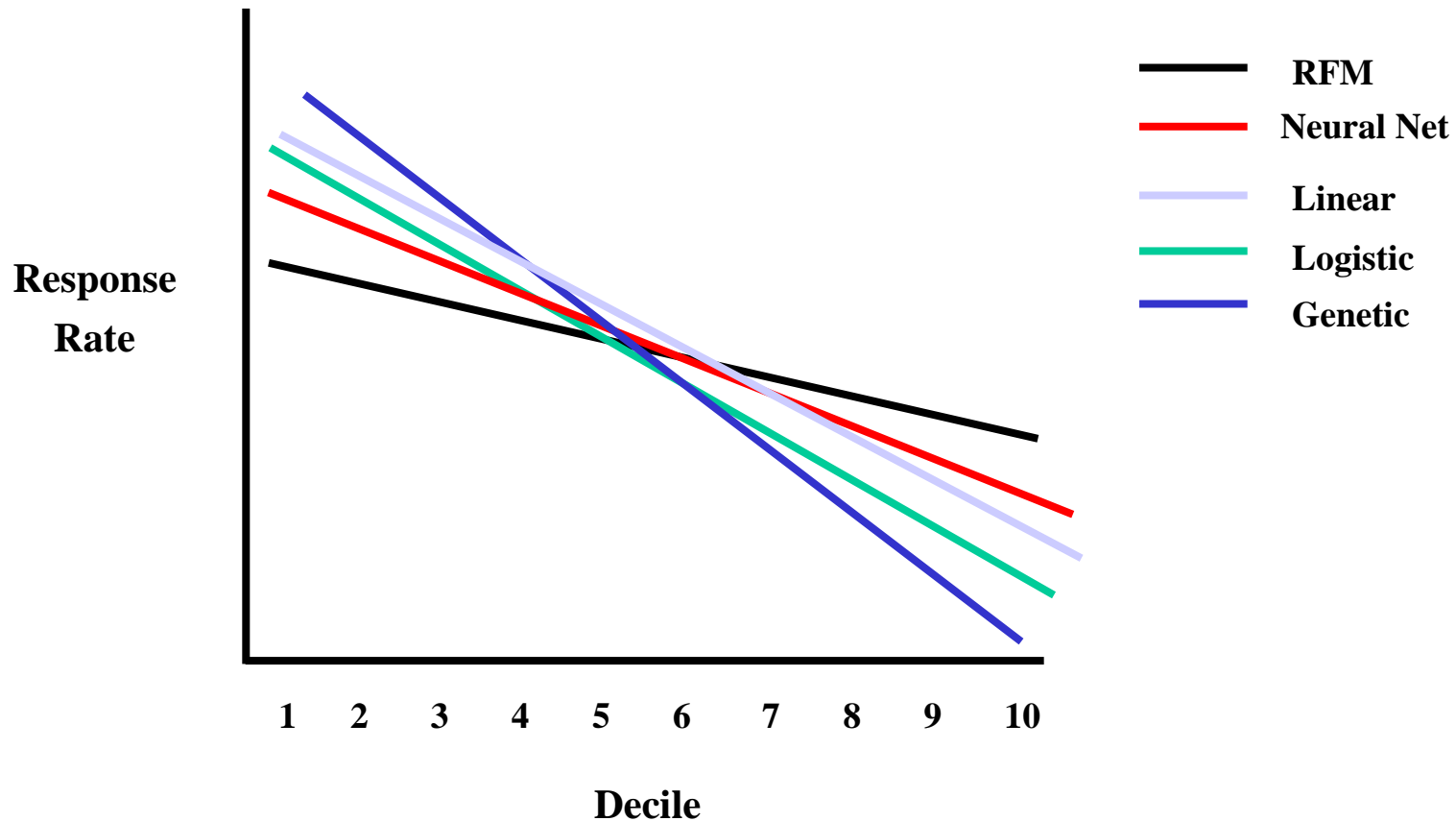
## Tip 7: Establish Performance Benchmarks

- Identify success before you start
  - What are you hoping to achieve?
  - Be specific!! Have an understanding of the performance “lift” you want to achieve
- The Gains Chart:

% of Validation Sample	Validation Names	Response Rate	% of Total Responders	Response Rate Lift	Interval ROI	Modelling Benefits
0-10%	20000	3.50%	23%	233	145%	\$26,667
10-20%	40000	3.00%	40%	200	75%	\$40,000
20-30%	60000	2.75%	55%	183	58%	\$50,000
30-40%	80000	2.50%	67%	167	22%	\$53,333
40-50%	100000	2.25%	75%	150	-13%	\$50,000
.						
.						
.						
90-100%	200000	1.50%	100%	100	-58%	\$0

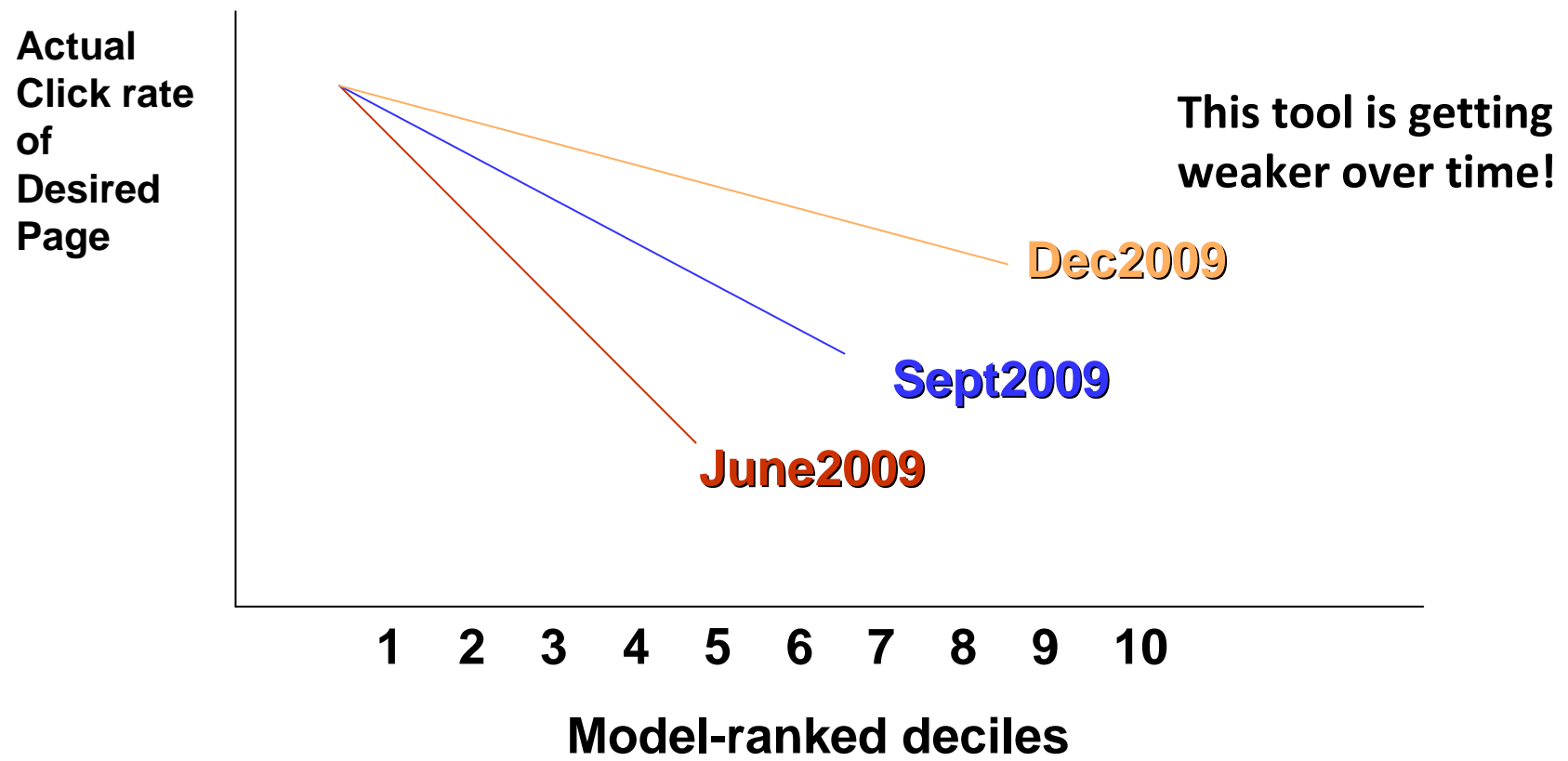
# Tip 7: Performance Benchmarks – Different Techniques

- Compare results of different analytical techniques
  - Techniques will vary in effectiveness and should be selected based on the nature of the solution required, data availability and resources



# Tip 7: Performance Benchmarks – Stay Current

- Evaluate effectiveness of solution at different points in time or over different Campaigns:



# Tip 8: Interpret Results Correctly – Overstated Results

- Be very wary of results that are “overstated”
  - If results look too good to be true...they might be!
- Example

% of Names Promoted	% of Mortgage Insurance Buyers
0-5%	80%
5-10%	5%
10-15%	5%
.....	10%
95-100%	1%

- Further investigation revealed that records in the analytical file created to develop the model were assigned incorrectly
  - The same Customers that purchased the product in the **Pre-period** (period during which we observe Customer characteristics) were also included in the **Post-period** (period in which we observe the purchase behaviour)

## Tip 8: Interpret Results Correctly – Multicollinearity

- Be wary of Multicollinearity
  - Different variables that are correlated with each other and the behaviour we are predicting erode the power of a predictive tool
  - E.g. Education and Income

Response Variable	Education Level	Income
Correlation Result	0.35	0.33
Confidence Level	99%	99%

- These variables are both so strongly correlated to each other that results are not as good as they could be if collinearity issues were minimized

# Tip 8: Interpreting Results Correctly – Address Outliers

- Look out for Outliers
  - Extreme values of a given variable result in skewed distribution around the mean
  - Mitigates actual impact of variable against the behaviour we are predicting:

Response Variable	Spending (last Yr)
Correlation Result	0.009
Confidence Level	10%

Correlation with uncapped data would tell you there is no relationship between spending and response

Spending Level	Response Rate
0 - 250	1%
251-500	2%
501-750	3%
750 - 1000 (Capped)	4%

When we cap the outliers and relook at the data we see a very strong relationship

# Tip 9: Relative Differences vs Absolute Results

- Tools are generally designed to predict relative differences in behaviour not absolute results

% of Validation Sample	Validation Names	Response Rate	% of Total Responders	Response Rate Lift	Interval ROI	Modelling Benefits
0-10%	20000	3.50%	23%	233	145%	\$26,667
10-20%	40000	3.00%	40%	200	75%	\$40,000
20-30%	60000	2.75%	55%	183	58%	\$50,000
30-40%	80000	2.50%	67%	167	22%	\$53,333
40-50%	100000	2.25%	75%	150	-13%	\$50,000
.						
.						
.						
90-100%	200000	1.50%	100%	100	-58%	\$0

- Predictive tool estimates that the top 10% of Customers will perform 2.33 times better than the lowest 10%, but not necessarily at a 3.5% response rate
  - Many factors might impact response upon implementation
    - Creative, offer, seasonality, competitive and external environment

# Tip 10: Action Solutions and Measure Performance

- Make sure there is a plan of action to implement solutions after they are developed
- Exercise great care in implementing solutions
  - QC to check individual Records and distribution of values

	Income	Age	Tenure	Calculated Score	Correct Score
Record 1	30,000	50	3	0.45	0.7206
Record 2	60,000	35	6	0.54	1.0812
Record 3	80,000	30	8	0.57	1.2916
Record 4	29,000	65	2	0.36	0.5404

% of List	Minimum Score (validation sample)	Minimum Score (current score)
0-10%	0.08	0.04
10-20%	0.07	0.03
20-30%	0.06	0.02
30-40%	0.05	0.01
40-50%	0.04	0.004
etc...		

# Tip 10: Action Solutions and Measure Performance (cont'd)

- Set Up Measurement and Tracking Matrix

	Control	Test Piece 1	Test Piece 2	Do Not Promote
Modeled List	330,000 <b>Cell 1</b>	10,000 <b>Cell 2</b>	10,000 <b>Cell 3</b>	
Non-model List	40,000 <b>Cell 4</b>			10,000 <b>Cell 5</b>

From above results, redefine and revise objectives for next campaign

## Ten Tips Summary

---

1. Identify the Real Problem
2. Get Stakeholders Onside
3. Understand the Data
4. Identify Quick Wins
5. Use Statistics Judiciously
6. Combine Art and Science
7. Establish Performance Benchmarks
8. Interpret Results Correctly
9. Relative Differences vs Absolute Results
10. Action Tools and Measure Performance

# Questions?

Larry Filler, Partner

[larryf@boirefillergroup.com](mailto:larryf@boirefillergroup.com)

PHONE: (905) 837-0005 x223

Richard Boire, Partner

[richb@boirefillergroup.com](mailto:richb@boirefillergroup.com)

PHONE: (905) 837-0005 x224

BOIRE FILLER GROUP  
1101 KINGSTON ROAD, SUITE 310  
PICKERING, ONTARIO  
L1V 1B5

Visit us at [www.boirefillergroup.com](http://www.boirefillergroup.com)