PREDICTIVE ANALYTICS FOR MARKETERS: BEST PRACTICES AND CHALLENGES

MAY 26, 2016

2 PM EDT

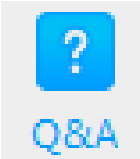
ENVIRONICS ANALYTICS

AGENDA

- Predictive Analytics Today
- Best Practices
  - The Top Ten
  - The Challenges

HOUSEKEEPING

- You will be in “listen mode” only
- Questions held until end
- Presentation will be available after the webinar environicsanalytics.ca/webcasts
- Use the Q&A feature in your WebEx interface



TODAY'S PRESENTER

Richard Boire

Senior Vice President,
Boire Filler

Environics Analytics

- 30 years of experience in relationship and database marketing, and predictive analytics
- Founder of analytics and database consultancy Boire Filler Group, which was recently acquired by Environics Analytics
- Active in the marketing community, he has taught applied statistics, data mining and database marketing at several Canadian colleges and universities

PREDICTIVE ANALYTICS TODAY

- Relatively new business discipline
- Grounded in data mining
- Requires structure
  - Develop consistent standards and practices
  - Provide comparisons and best practices
  - Establish a more rigorous approach to yield better learning and knowledge for practitioners

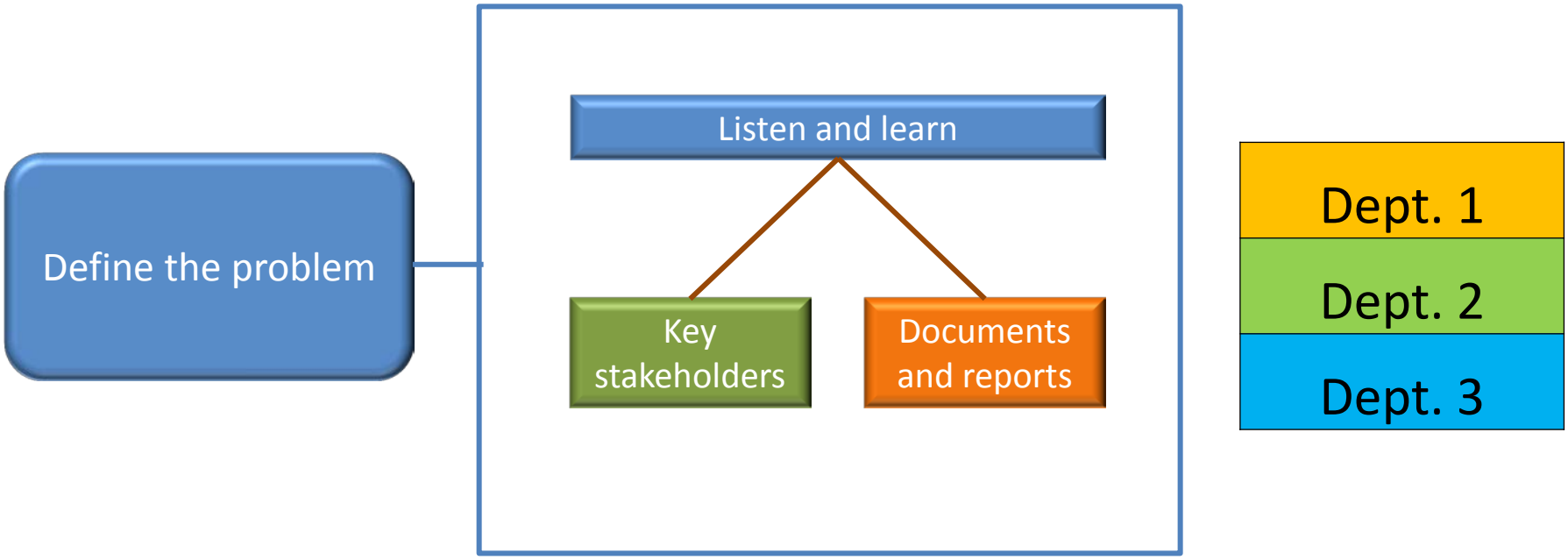
PREDICTIVE ANALYTICS:
BEST PRACTICES AND CHALLENGES

PREDICTIVE ANALYTICS BEST PRACTICES

1. Identify the Real Business Problem
2. Look for Quick Wins
3. Become Familiar with the Data
4. Use Statistics Judiciously
5. Establish Performance Benchmarks from the Start
6. Interpret Results Carefully
7. Use Art & Science to Build Solutions
8. Implement Solutions Carefully
9. Integrate Big Data Knowledge into your Business
10. Measure and Track Results

- . . . and their associated “DO’s” and “DON’Ts”

1. IDENTIFY THE REAL BUSINESS PROBLEM



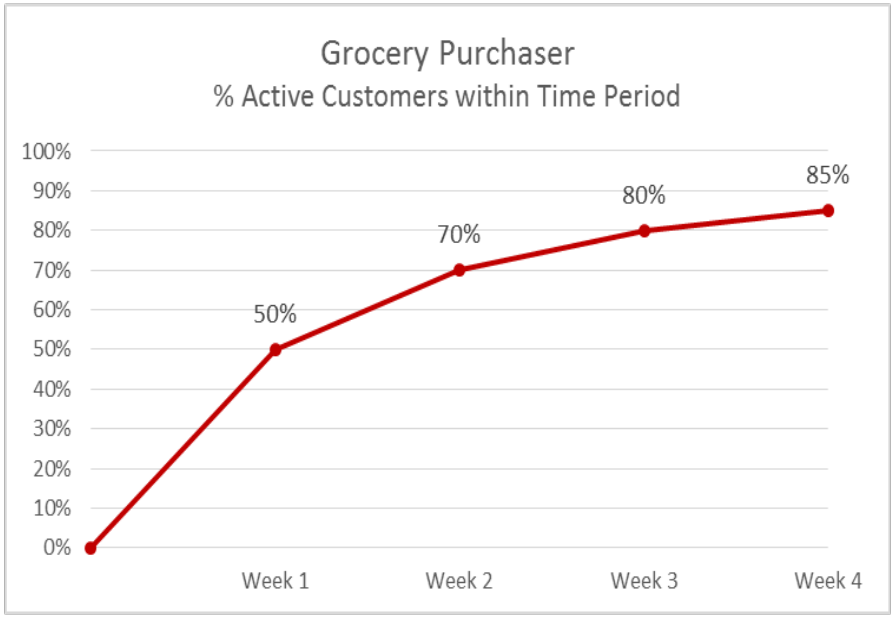
- **DO** be careful of silos between departments
- **DON'T** jump to conclusions

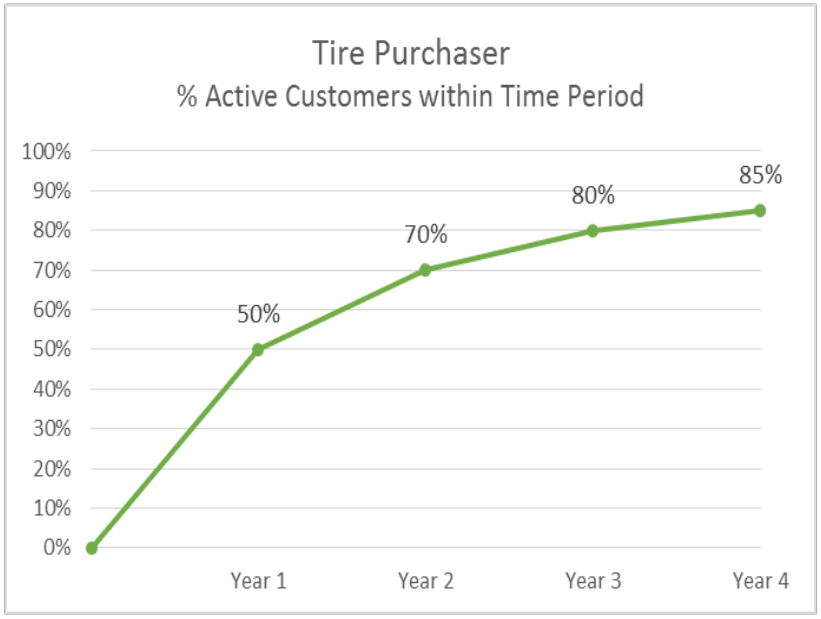
BE SURE TO ASK THE RIGHT QUESTION

- Telco built response model to acquire more customers
- New customers acquired more effectively, but overall customer base did not grow
- Analysis of new customer performance in first three months revealed increasingly higher rates of defection
- Model should have optimized both acquisition *and* retention
- **The right business question was not addressed at the outset**

RETENTION MODEL: FIRST QUESTION

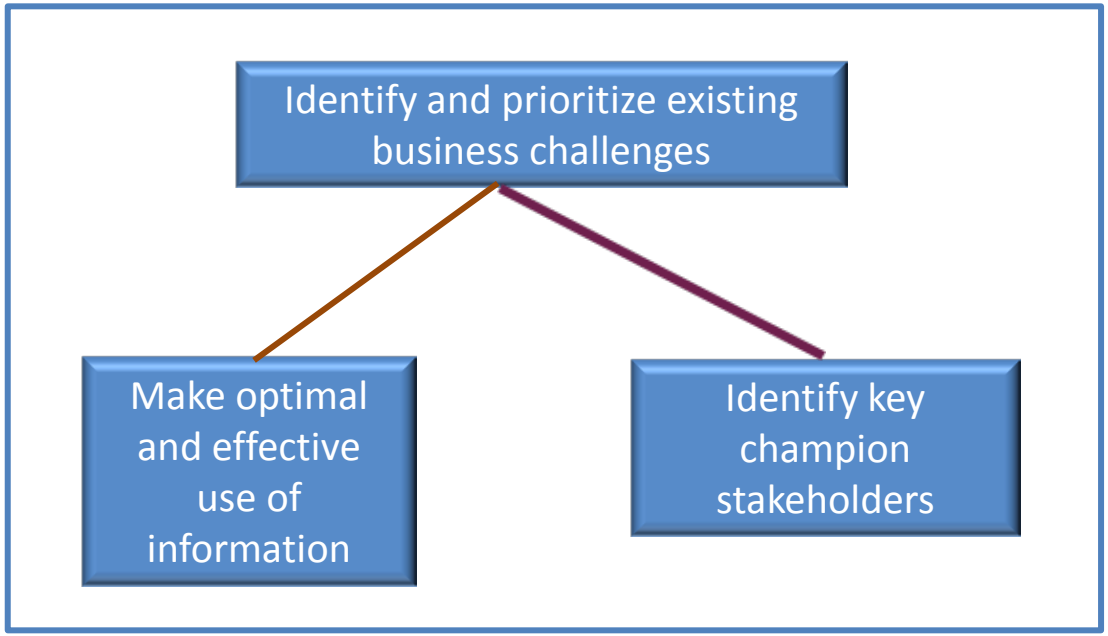
- How do we define retention across different organizations?





2. LOOK FOR QUICK WINS

Creating a quick win



- **DO** avoid exercises that cannot clearly demonstrate cost benefits

CREATING A QUICK WIN

- Retention issues are typically a challenge for most marketers
- High-value retention model offers significant savings

Potential Benefits - Targeting Top 25% of Model

	Quantity Promoted	Defection Rate of Group	# of Potential Defectors.	Save Rate	# of Cust. Saved	Avg. Value	Total Saved Quarterly
Savings without model	100,000	1.19%	1186	20%	237	$300	$71,100
Savings with model	100,000	3.12%	3117	20%	623	$300	$186,900
Diff: model vs. no model		1.93%	1931		386		$115,800

CHALLENGE: CREATING A QUICK WIN USING SOCIAL MEDIA

- Profile high value social media engagers
- How much $$ could I realize?
  - Clearly not a quick win yet
  - Business still needs to demonstrate revenue
  - Determine the impact of reaching out to these digital engagers

Variable	ENGAGER
Number of Friends adjusted for tenure	0.169272
# of Friends	0.117881
# of months since last twitter behaviour	-0.11391
Source was other:	0.105733
Source was Iphone	-0.06422
# of followers adjusted for tenure	0.05387
Source was Android	-0.04031
Tenure in years	-0.04026
# of replies per tweet	-0.01704
Sentiment was positive	0.014762
# of followers	0.014744
average Sentiment Score	0.014702
# of friends per tweet foor the user	-0.01448
source was blackberry	-0.00827
source was web	-0.00549
source was facebook	-0.00372
# of followers per tweet	-0.00321
# of video photos per tweet	0.000167
# of retweets per tweet	5.08E-05

3. BECOME FAMILIAR WITH THE DATA

- Gain better understanding of data's integrity and completeness
- Identify potential gaps and how to handle
  - Define and create new variables that will be valuable for the analytical phase of the project
  - Complete some preliminary analyses
- Consider:
  - Meta data
  - Frequency distributions
  - Random data dump

➢ **DON'T** make assumptions about the data

KNOW YOUR DATA: DATA AUDIT REPORTS

- Data audit is done on 53,235 records
- By looking at number of unique values and number of missing values, you can begin to understand data
- Data indicates that relatively few customers have email (70%) while tenure (creation date) appears to be a good variable with few missing values (3%)

Data Audit Report - Column MetaData

Structure			Value Distribution				
Ordinal	Column Name	Column Type	Unique Values	FREQs	Missing	Mis %	Non Missing
1	NAMEUPPER	varchar	48,962		134	0%	53,101
2	PROFILENO	int	53,177		59	0%	53,176
3	PROFILETYPE_LINKCODE	varchar	64	Y	0	0%	53,235
4	NAME	varchar	48,910		134	0%	53,101
5	FIRSTNAME	varchar	12,025		3,690	7%	49,545
6	MIDDLEINIT	varchar	823		44,632	84%	8,603
7	LASTNAME	varchar	21,678		3,077	6%	50,158
8	COURTESYTITLE	varchar	93	Y	15,884	30%	37,351
9	ADDRESS1	varchar	36,252		10,852	20%	42,383
10	ADDRESS2	varchar	2,795		49,254	93%	3,981
11	CITY	varchar	3,469		10,878	20%	42,357
12	STATE	varchar	758		10,084	19%	43,151
13	ZIP	varchar	27,667		11,807	22%	41,428
14	COUNTRY	varchar	1,217		19,509	37%	33,726
15	STMTREMARKS	varchar	1,867		50,326	95%	2,909
16	UNAPPLIEDBALANCE	varchar	57	Y	1,725	3%	51,510
17	PHONE	varchar	32,037		17,088	32%	36,147
18	FAX	varchar	3,053		49,673	93%	3,562
19	EMAIL	varchar	13,731		37,219	70%	16,016
20	TITLE	varchar	647		52,519	99%	716
21	BUSINESSTYPE	varchar	121		53,105	100%	130
22	NOTES	varchar	45	Y	53,191	100%	44
23	ADDITIONALNOTES	varchar	20	Y	53,216	100%	19
24	OTHER	varchar	9	Y	53,227	100%	8
25	TRAVELPREF	varchar	6	Y	53,229	100%	6
26	INTERFACEID	varchar	980		51,568	97%	1,667
27	CREATIONDATE	datetime	3,041	Y	1,822	3%	51,413
28	CREDITLIMIT	varchar			53,235	100%	0

➢ **DON'T** forget to look at key arithmetic diagnostics

KNOW YOUR DATA: META DATA REPORT OF TWITTER DATA

- From this data we know:
  - Location (place) information is meaningless
  - Only one language is being used
  - 329,121 unique persons

Column Name	Missing	Mis %	Non Missing	Unique Values
created_at	0	0%	822,177	209,390
entities_urls	0	0%	822,177	114,003
in_reply_to_screen_name	0	0%	822,177	13,327
in_reply_to_status_id	0	0%	822,177	26,744
in_reply_to_status_id_str	0	0%	822,177	26,744
in_reply_to_user_id	0	0%	822,177	28,700
in_reply_to_user_id_str	0	0%	822,177	28,700
lang	0	0%	822,177	1
place	816,208	99%	5,969	1,415
source	0	0%	822,177	3,919
text	0	0%	822,177	822,177
user_created_at	0	0%	822,177	129,976
user_followers_count	0	0%	822,177	68,365
user_friends_count	0	0%	822,177	41,694
user_id_str	0	0%	822,177	329,121

Table illustrates a small sample of the fields or variables that we might analyze

4. USE STATISTICS JUDICIOUSLY

- The appropriate technique will depend on the business problem

Statistical Tool	Business Application
Correlation Analysis	• Exploratory – identify key variables
CHAID	• Exploratory – identify key variables • Can also be used to build final model
Factor Analysis	• Exploratory – reduce data and help to identify key variables
Cluster Analysis	• Define distinct homogenous groups of customers
Multiple Regression	• Build Final Model
Logistic Regression	• Build Final Model
Neural Nets	• Build Final Model

➢ **DON'T** assume PhDs in math and computer engineering know everything

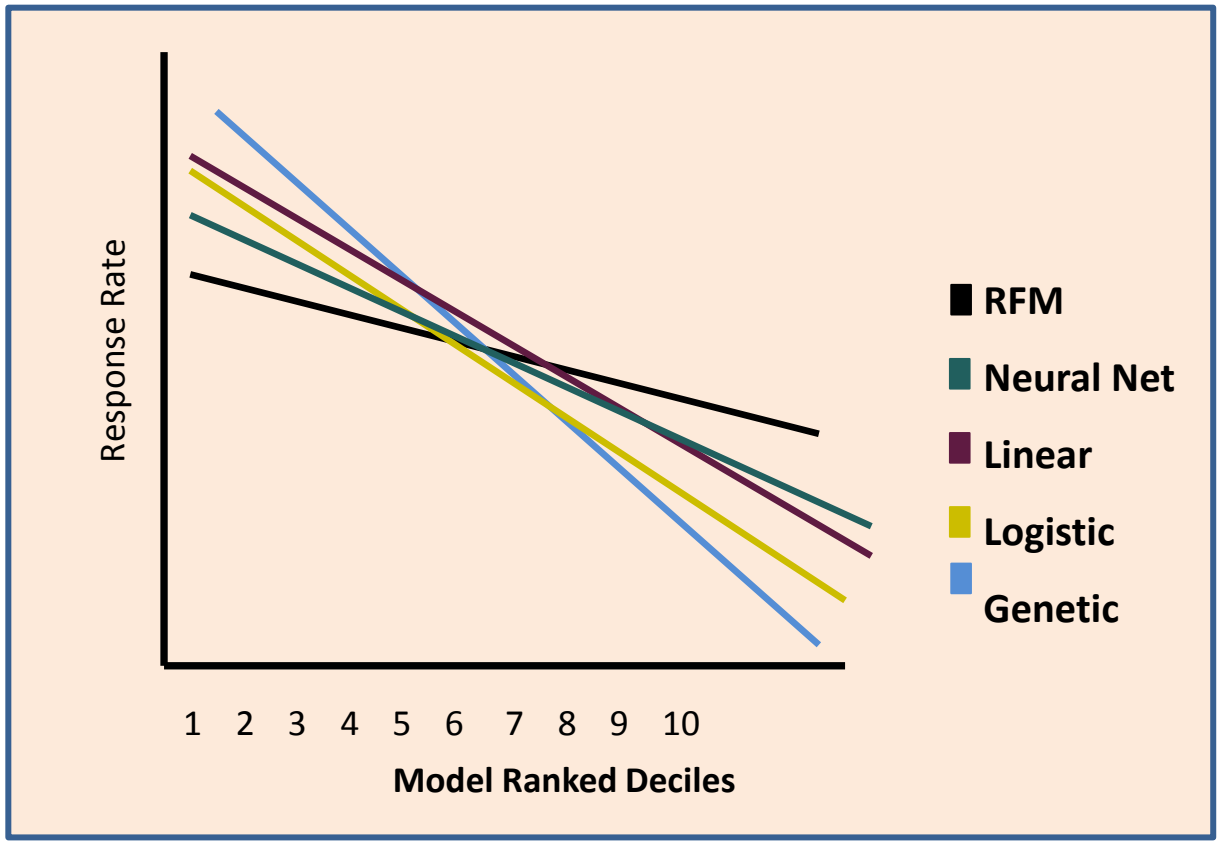
5. ESTABLISH PERFORMANCE BENCHMARKS FROM THE START

- Example: the gains chart looks at lift results

% of Validation Sample	Validation Names	Response Rate	% of Total Responders	Response Rate Lift	Interval ROI	Modelling Benefits
0-10%	20,000	3.50%	23%	233	145%	$26,667
10-20%	40,000	3.00%	40%	200	75%	$40,000
20-30%	60,000	2.75%	55%	183	58%	$50,000
30-40%	80,000	2.50%	67%	167	22%	$53,333
40-50%	100,000	2.25%	75%	150	-13%	$50,000
.	.					
.	.					
.	.					
90-100%	200,000	1.50%	100%	100	-58%	$0

- **DON'T** be consumed by looking at residuals (difference between predicted estimates and observed estimates)

GAINS CHART: COMPARES DIFFERENT SOLUTIONS WITH DIFFERENT TECHNIQUES



- Judge models by their rank-ordering capability by looking at the slope of the line

6. INTERPRET RESULTS CAREFULLY

- Challenge: Multicollinearity
  - Despite an opposite mathematical relationship in the response equation, higher income and higher education lead to higher response

Correlation Years of Education and Income on Response Rate		
	Years of Education	Income
Correlation Coefficient	0.11	0.12
Confidence Interval	99%	99.50%

Response = .50 +.00001*income -.03*yrs. of education

➢ **DON'T** trust mathematical output without some level of comprehension

6. INTERPRET RESULTS CAREFULLY

- Challenge: Outliers
  - Transformation of total customer spend variable is required to demonstrate the relationship between customer response and total customer spend

	Spend
Correlation Coefficient	0.009
Confidence Interval	10%

Spend	% of File	Response Rate
$ 0-25	25%	1%
$ 25-50	25%	2%
$ 50-75	25%	3%
$75 - 1,000	25%	4%

6. INTERPRET RESULTS CAREFULLY

- Challenge: Overstatement of results

Mortgage Insurance Model	
% of Names scored by model and ranked into half deciles	% of Mortgage Insurance Buyers
0-5%	80%
5-10%	5%
10-15%	5%
15%-100%	10%

Is there a problem here?

➢ **DON'T** accept results that are too good to be true

6. INTERPRET RESULTS CAREFULLY

- Challenge: Overstatement of results

Correlation Report on customer likelihood to purchase mortgage insurance	
Variable	Correlation Coefficient
ever bought insurance	0.75
1 or more lending products	0.2
have a line of credit product	0.18
have a car loan	0.17
have an RRSP	0.16
have an RESP	0.15
have an investiment product	0.16
live in Toronto	0.15
live in Ontario	0.14
have a chequing account	0.0002

- Are we sure the results are still valid?
- May need to investigate this variable more closely
- Consider upfront segmentation

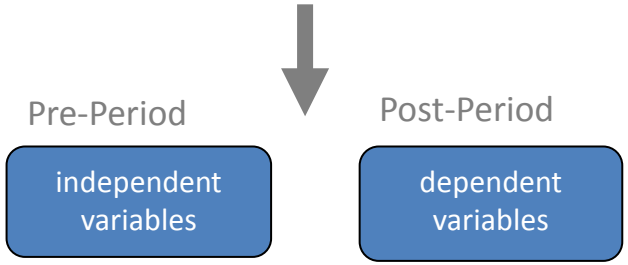
Variable Contribution Report	
Model Variable	% Contribution to Model
ever bought insurance	85%
1 or more lending product	8%
have an investment product	5%
have a credit card	1%
live in Ontario	1%

6. INTERPRET RESULTS CAREFULLY

- Challenge: Overstatement of results

  - Analytical file improperly created when information used to create the dependent variable was also used to create the "ever bought" insurance variable

  - Need to create proper analytical file



response	ever bought insurance	1 or more lending products	have an investment	have a credit card	live in Ontario
yes	yes	Yes	Yes	Yes	Yes
yes	yes	Yes	No	Yes	No
yes	yes	Yes	Yes	No	Yes
yes	yes	Yes	Yes	No	No
yes	yes	Yes	No	Yes	Yes
yes	yes	No	Yes	Yes	No
yes	yes	Yes	Yes	No	Yes
yes	yes	Yes	No	No	No
yes	yes	Yes	Yes	Yes	Yes
yes	yes	Yes	Yes	Yes	No
yes	yes	No	Yes	No	Yes
No	No	No	No	No	No
No	No	Yes	No	No	No
No	No	No	Yes	No	No
No	No	Yes	No	No	Yes
No	yes	No	No	No	Yes
No	yes	No	No	No	No
No	No	No	No	Yes	No
No	No	No	No	Yes	No
No	Yes	No	Yes	Yes	No

7. USE ART AND SCIENCE TO BUILD SOLUTIONS

- Challenge
  - Retailer collects no information on its customers
  - Market research indicates the key drivers of purchase behaviour are high income, females and immigrants
- Solution
  - Using an indexing approach, create postal code index variable based on three Statistics Canada variables

	Income	% Female	% 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 x 1.25)+(.33 x 1.15)+(.33 x 2) = 1.45

7. USE ART AND SCIENCE TO BUILD SOLUTIONS

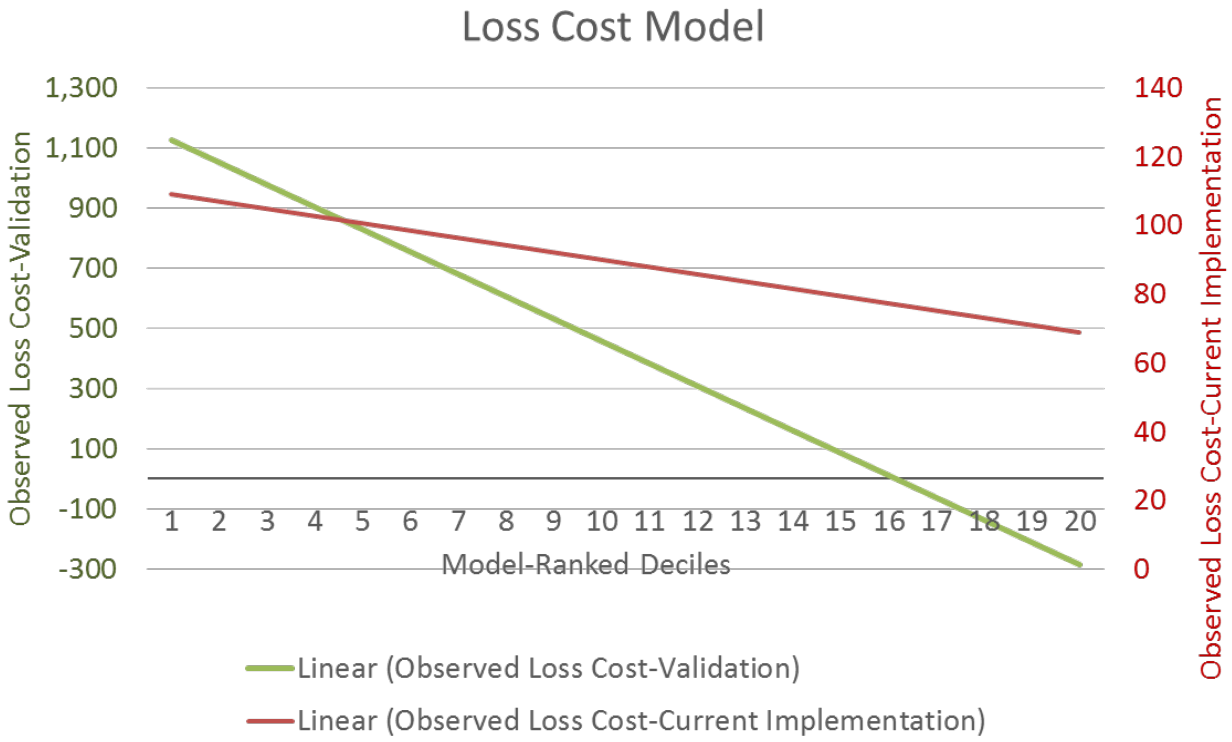
- Index scheme can then be used to score each postal code
- 800,000 postal codes in Canada are then ranked into 20 half deciles based on descending index score

% of File	# of Postal Codes	Min Index in Interval	# of Prospects
0-5%	40,000	5.50	80,000
5-10%	40,000	5.00	60,000
10-15%	40,000	4.80	90,000
…			
95-100%	40,000	0.05	30,000
Total	800,000		3,000,000

➢ **DON'T** use lack of data or inability to use advanced techniques as barriers to at least test different initiatives

8. IMPLEMENT SOLUTIONS CAREFULLY

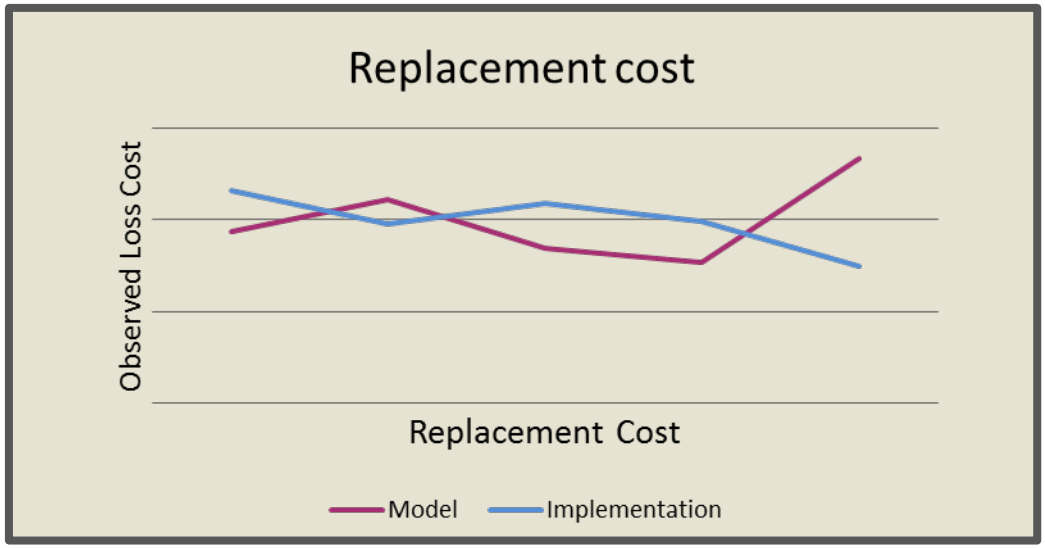
- Challenge: Loss Cost model
  - Loss cost is the loss amount of the claim/premium amount for that policy exposure period

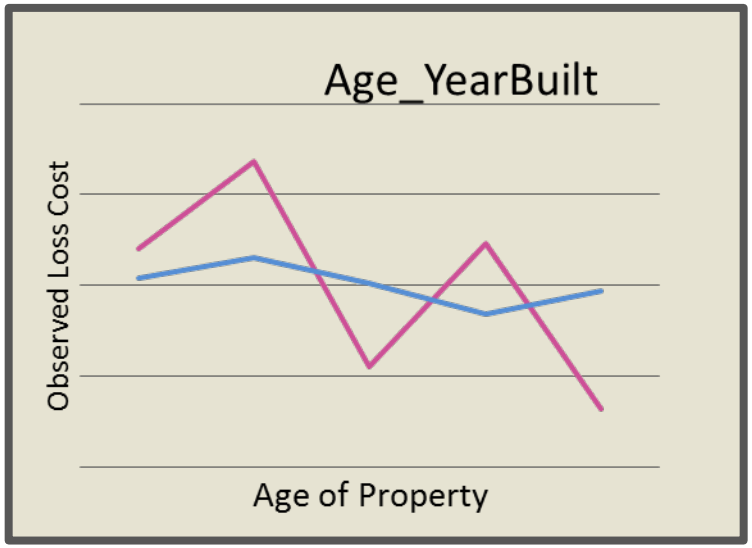


➢ **DON'T** make assumptions about the data

8. IMPLEMENT SOLUTIONS CAREFULLY

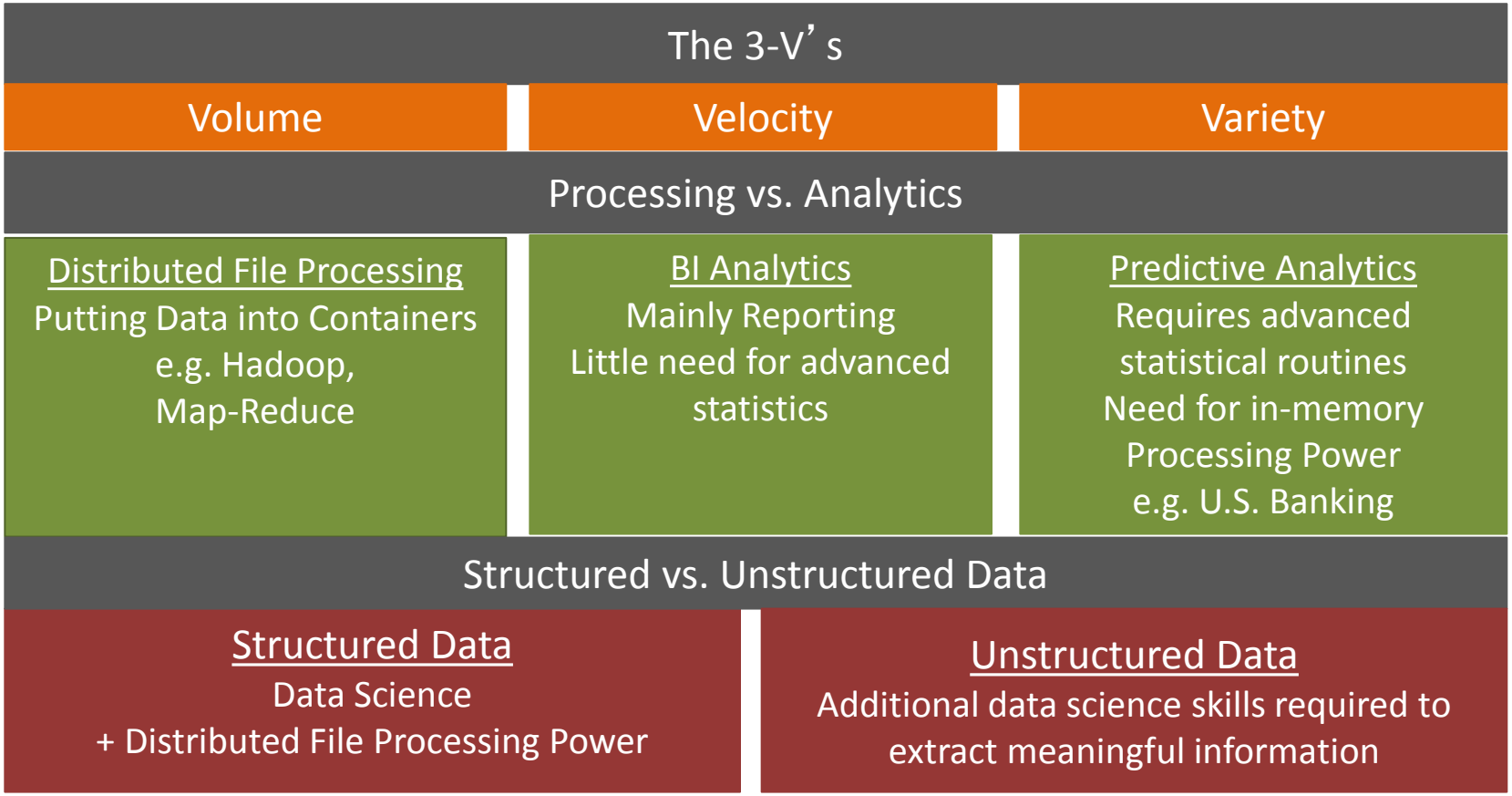
- Challenge: Loss Cost model
  - Key variables were analyzed
  - Why are key model variables not performing?
  - Audit of current data indicated strong presence of apartments
  - No apartment data were in model development file





9. INTEGRATE BIG DATA KNOWLEDGE INTO PREDICTIVE ANALYTICS

- Business Problem: Irrelevant vs. Relevant Data



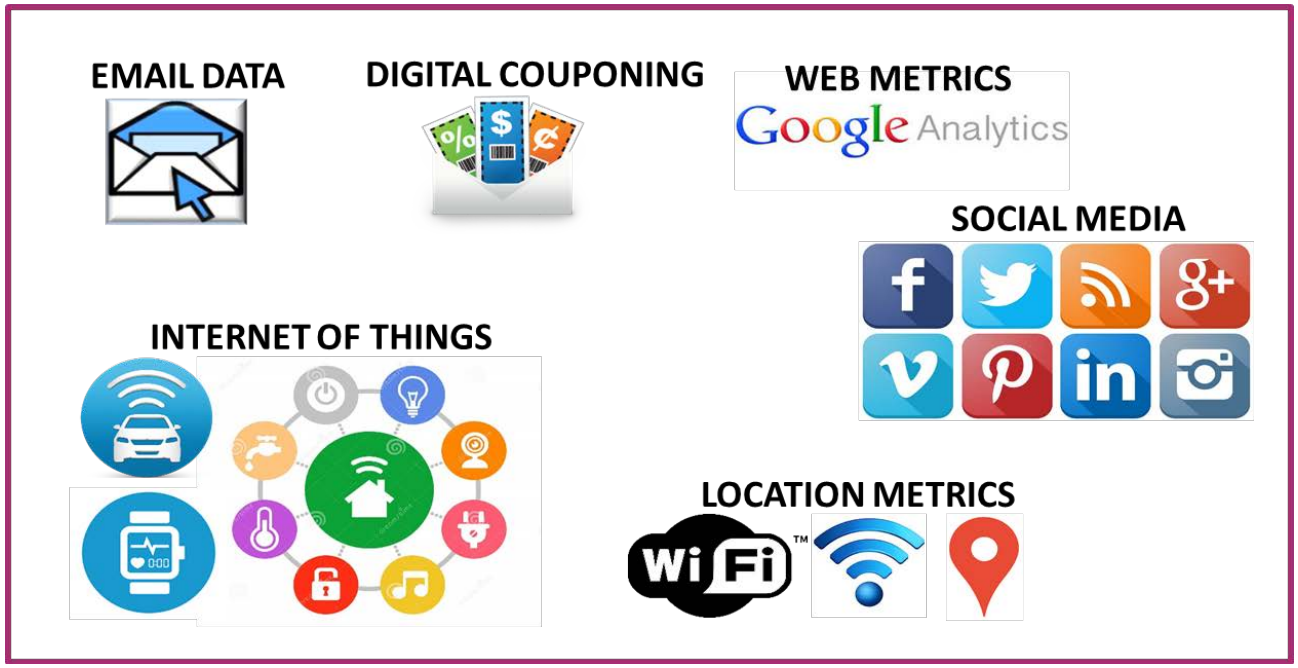
➢ **DON'T** become paralyzed by data overload

9. INTEGRATE BIG DATA KNOWLEDGE INTO PREDICTIVE ANALYTICS

- Challenge: Semi-Structured/Unstructured

What is the one view record of interest?

- The Person or Customer
- Institution
- Devices
- Areas / Geographies



Digital Data Silos must also be Integrated with each other and with Traditional Data

10. MEASURE AND TRACK RESULTS

- Customer Migration:
  A Different View
- Series of reports designed to:
  - Determine actual customer migration patterns of Carded Patrons between two set periods of time
  - Compare this to the predicted migration pattern
  - If variance is significantly different, look at which original profile variables are still impacting migration versus those that are not

Actual			New Segment (Current)			
		Est # Customers	Gold	Silver	Reward	Lapsed
Old Segment Pre Period	Gold	50,000	50%	30%	15%	5%
	Silver	150,000	20%	30%	30%	20%
	Reward	300,000	5%	10%	50%	35%
	Total	500,000				

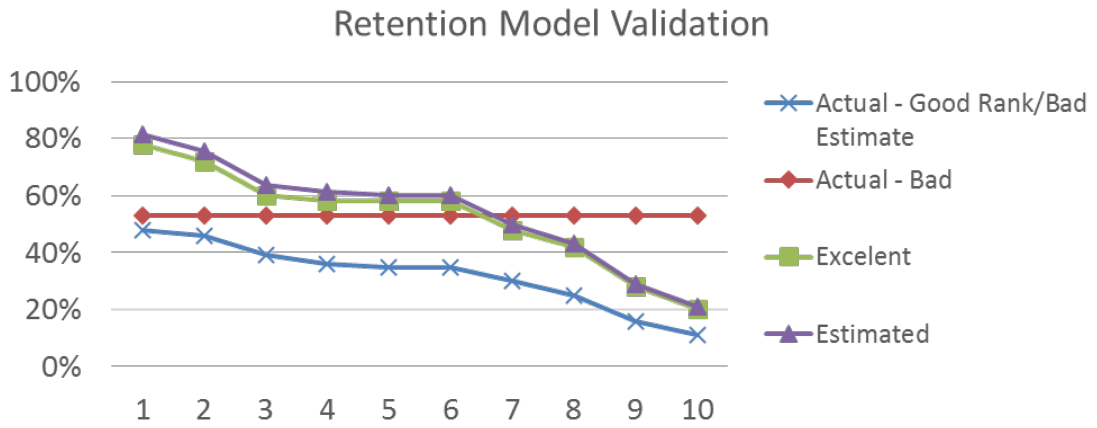
Predicted			New Segment (Current)			
		Est # Customers	Gold	Silver	Reward	Lapsed
Old Segment Pre Period	Gold	50,000	60%	20%	15%	5%
	Silver	150,000	15%	25%	35%	25%
	Reward	300,000	10%	15%	50%	25%
	Total	500,000				

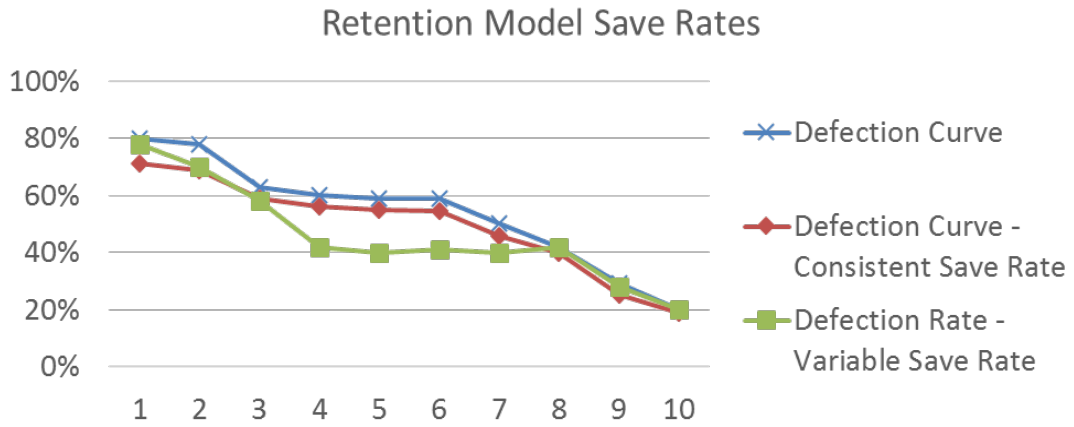
Variance			New Segment (Current)			
		Est # Customers	Gold	Silver	Reward	Lapsed
Old Segment Pre Period	Gold	50,000	-10%	10%	0%	0%
	Silver	150,000	5%	5%	-5%	-5%
	Reward	300,000	-5%	-5%	0%	10%
	Total	500,000				

➢ **DON'T** develop solutions that cannot be measured and tracked

10. MEASURE AND TRACK RESULTS

- Challenge: Retention





This report looks at the effectiveness of different models as a ranking tool and an estimating tool

This report looks at how effective specific marketing activities are at preventing defection

- Red line indicates save rates are consistent regardless of rank
- Green line indicates marketing activities have the biggest impact in the middle ranges of the model

FINAL THOUGHTS

- Big Data world has definitely impacted Predicted Analytics
- But at the end of the day, our challenge remains the same:
  - Identifying the right business problem
  - Using the right information and deriving the right insights

IN SUMMARY: DO'S AND DON'TS OF PREDICTIVE ANALYTICS

DO	DON'T
1. Identify the real business problem	• Jump to conclusions
2. Look for quick wins	• Create silos between departments
3. Become familiar with the data	• Avoid exercises that cannot clearly demonstrate cost benefit
4. Use statistics judiciously	• Make assumptions about data
5. Establish performance benchmarks from the start	• Forget to look at key arithmetic diagnostics
6. Interpret results carefully	• Assume that PhDs in mathematics and computer engineering know everything
7. Use art and science to build solutions	• Be distracted by predicted estimates
8. Implement solutions carefully	• Trust in math and business domain knowledge
9. Integrate Big Data knowledge into your business	• Accept results that are too good to be true
10. Measure and track results	• Use science and math as barriers
	• Become paralyzed by data overload
	• Develop solutions that cannot be measured and tracked

TO FIND OUT MORE

Richard Boire

Senior Vice President, Boire Filler

Environics Analytics

richard.boire@environicsanalytics.ca

647.800.1435

environicsanalytics.ca