TECHNIQUES FOR IMPROVING YOUR PREDICTIVE MODELS

2013 RAPID INSIGHT USER CONFERENCE

CAITLIN GARRETT, STATISTICAL ANALYST

MODEL BUILDING IS AN ITERATIVE PROCESS

STAGE ONE

IMPROVING THE APPROACH

START WITH THE QUESTION

- Make sure it’s specific
- Consider end user
- Create a plan of attack

GET COLLABORATIVE

- Talk to decision makers, end users, data pullers, IT, etc.
- When you’re knee-deep, it helps to bounce ideas

ON THE PROCESS

- Model building is an iterative process
- Not linear – get used to going back and forth between Veera & Analytics

GETTING YOUR DATA

- Work with the data in its most raw form
- At some point, you'll probably have to pull the data again/differently.

STAGE TWO

IMPROVING THE DATA

GET TO KNOW YOUR DATA

- Being proactive now will pay off later
- Understand variables & codes
- Review it
- Get help if you need it

REALLY GET TO KNOW YOUR DATA

- Review for accuracy
- Review for completeness
- Any gaps?
- Check for missings or outlying values

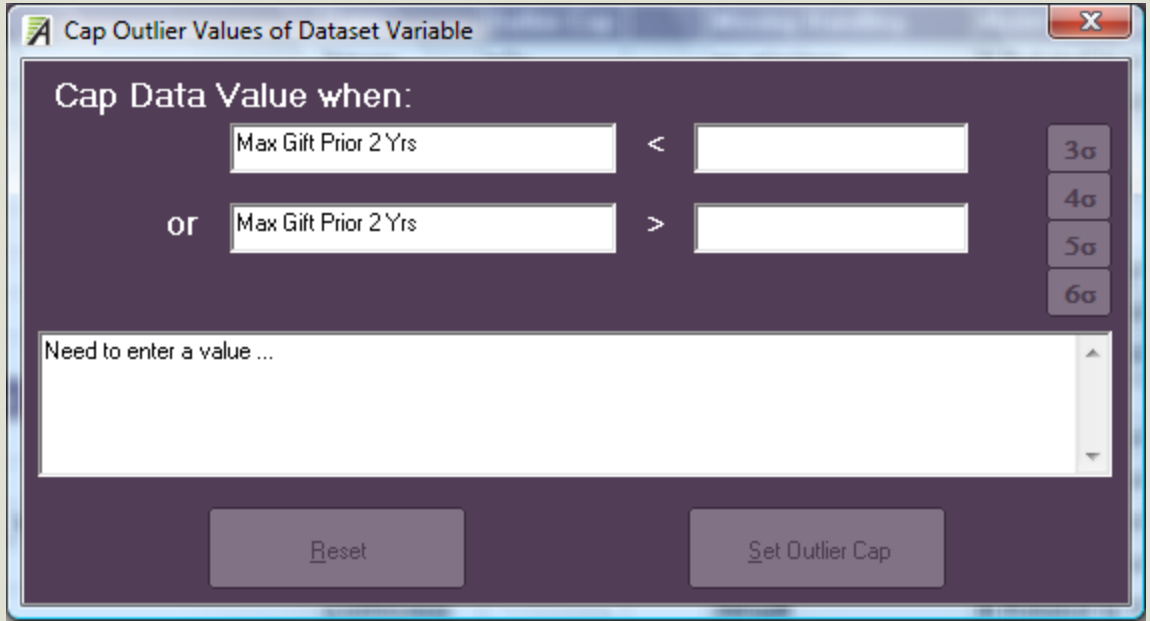
MISSING VALUES

- Sometimes the absence of data is as important as the data itself
- Can indicate engagement
- Create flags

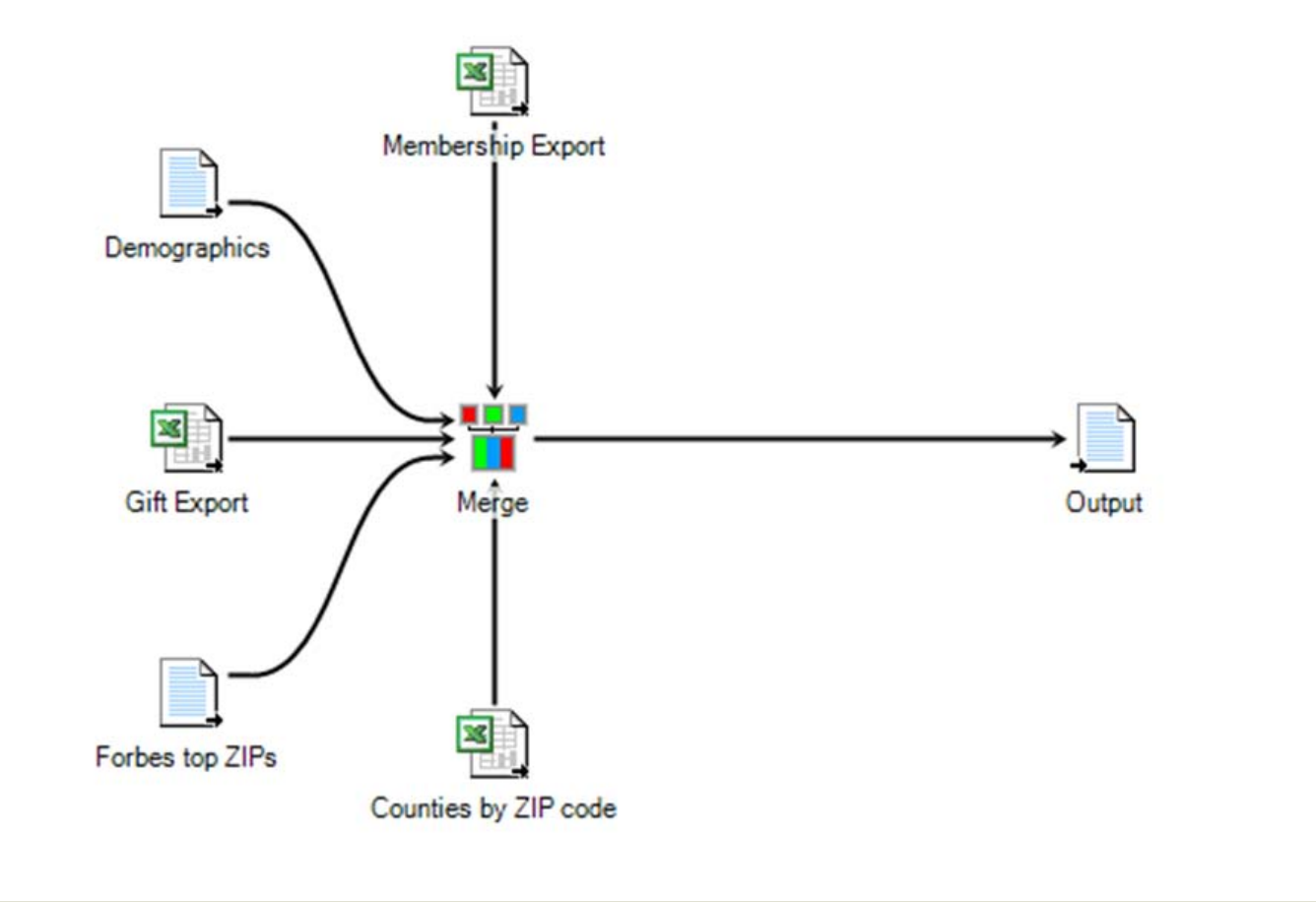
variable	mean	std dev	min	max	coeff var	missing	# obs	range	var_type	# distinct	text
Campus	n/a	n/a	n/a	n/a	n/a	8895	1568	n/a	categorical	5	yes
CampusActivity	n/a	n/a	n/a	n/a	n/a	10046	417	n/a	categorical	6	yes
CollegeName	n/a	n/a	n/a	n/a	n/a	6613	3850	n/a	categorical	3	yes
Degree	n/a	n/a	n/a	n/a	n/a	10262	201	n/a	categorical	7	yes
Max Gift Prior 2 Yrs	220.07	326.21	0	1,000.00	1.482	0	10463	1,000.00	continuous	---	no
Max Gift Prior 5	366.47	351.91	100.00	1,000.00	0.9603	0	10463	900.00	continuous	---	no
Max Gift Prior Yr	132.13	266.08	0	1,000.00	2.014	0	10463	1,000.00	continuous	---	no

OUTLIERS

- “Numerically distant from the rest of the data”
- Can throw off an analysis
- Consider capping gift amount



GET MORE DATA



CREATING VARIABLES

- Make sure you have the basics, then add from there
- Get creative!
- In a hurry? Create in Analytics, Auto-mine on the spot.
  - Don’t forget to repeat in Veera

CREATING VARIABLES (CONT.)

- Create true binary variables (0/1) from Y/N variables
- Dates can’t come into a model
- Dates: use ‘days between’ or ‘years between’ functions in a Transform
  - Graduation year -> “Years Since Graduation”

RECENT INTERESTING VARIABLES

Years since graduation

Earned a degree?

Number of previous phone refusals

Number of events attended

Lived on-campus as a student?

Name prefix is doctor

Number of student activities

Number of cross-references in DB

ON POPULATION SIZE

- Hard to generalize a “right size”
- Depends on what you’re modeling,
- Litmus test: make sure your dataset is “filled out”
  - Plenty of data points to fit different variable categories
  - Example: 100 records, gender vs. state

LOOK HISTORICALLY

- Build your dataset such that you're seeing what a donor looked like historically.
- Trying to find out what they looked like *before* the gift rather than after

RAPID INSIGHT BEST PRACTICES

- Do all of your data prep in Veera, build your model, then score in Veera
- Logistic model (binary outcome) – use the smaller response category as your y-variable
  - Ex: retention vs. attrition
- Keep your original question in mind

STAGE THREE

IMPROVE YOUR ANALYSIS

AVOID THE EASY ONES

- ❑ Don’t include IDs as predictors
- ❑ Beware of duplicates
- ❑ Check your population size
- ❑ Not accounting for missing values
- ❑Take out anachronistic variables – those that you would only know about after the response

DESCRIPTIVE STATISTICS

- Lots of insight into your data
- Correlation, Profiling, Frequency, Means
- Get a good feel for the relationships that exist within your dataset

CORRELATION

- Look at the correlations with giving
- Positive or Negative? Large or small?
- Rule of thumb: +/- . 1 is statistically significant

	Days Since Last Gift	Gave 10k	Max Gift Prior 2 Yrs	Mem_Total_Years	Years Since Graduation
Days Since Last Gift		-0.0642	-0.1911	-0.1085	-0.1709
Gave 10k	-0.0642	1	0.2516	0.1659	0.0715
Max Gift Prior 2 Yrs	-0.191	0.2516	1	0.3213	0.0998
Mem_Total_Years	-0.1085	0.1659	0.3213	1	0.0833
Years Since Graduation	-0.1709	0.0715	0.0998	0.0833	1

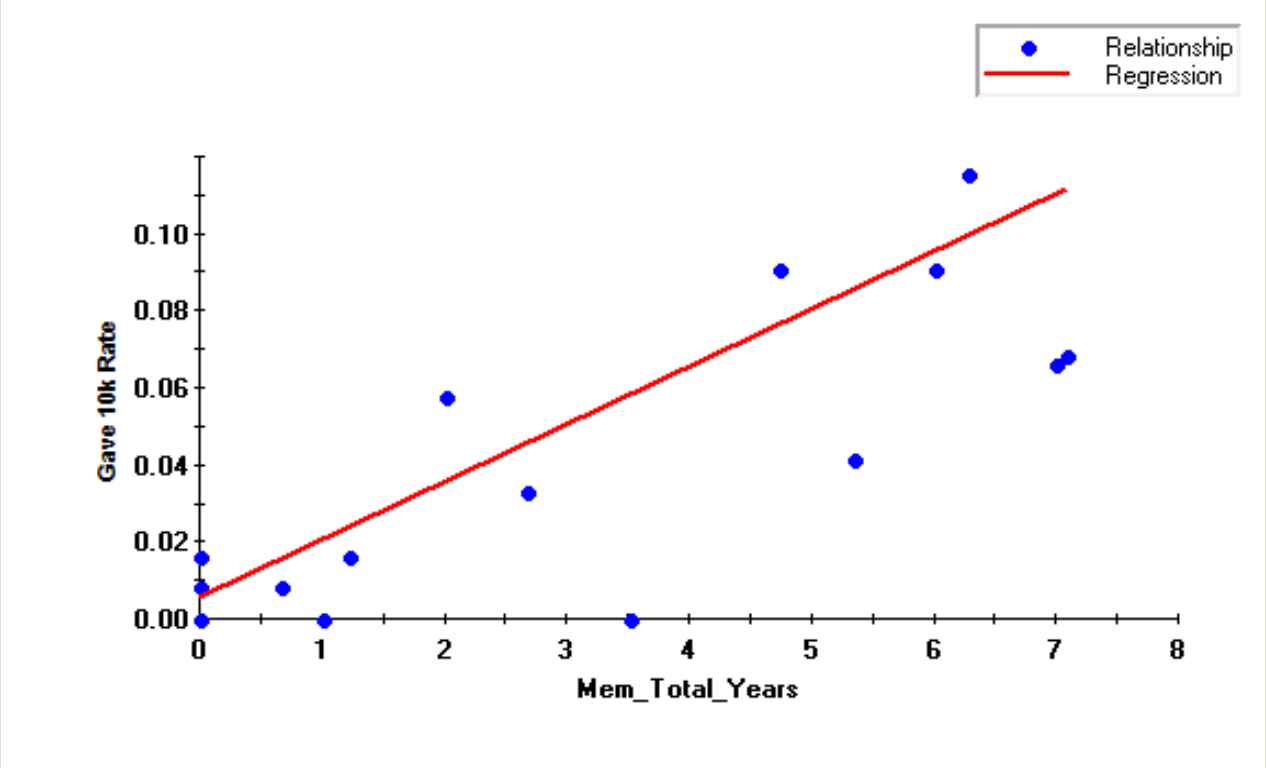
PROFILING

- Look for differences between givers and non-givers

Continuous Variables		Gave 10k=0	Gave 10k=1
Days Since Last Gift		2,650.197	2,086.887
Given Age		59.796	73.462
Max Gift Prior 2 Yrs		208.758	815.701
Max Gift Prior 5		356.422	895.641
Max Gift Prior Yr		127.240	389.785
Mem_Consecutive_Years		2.173	4.641
Mem_Total_Years		2.258	4.718
Total Gifts Prior 2 Yrs		0.939	2.559
Total Gifts Prior 5		2.004	4.672
Total Gifts Prior Yr		0.517	1.087
Total Giving Prior 2 Yrs		262.260	849.364
Total Giving Prior 5		1,389.998	7,036.305
Total Giving Prior Yr		148.687	433.062
Years Since Graduation		29.454	40.071

“VIEW RELATIONSHIPS”

- After Automated Mining
- Visuals help understanding relationships



AFTER MODELING

- Look at the variables in your model. Do they make sense?
- Can you explain it to others?

Variable	Coef	S.E.	Wald chi-sqr	p-value
Intercept	-6.113	0.1881	1,055.72	0.0000
Max Gift Prior 2 Yrs	0.00215	0.000269	63.92	0.0000
Total Giving Prior 5	0.000222	0.000026	75.07	0.0000
Binary(County,Jefferson)	2.375	0.8148	8.49	0.0036
Square(Total Gifts Prior 2 Yrs)	0.03126	0.00807	15.01	0.0001
Binary(ZIP,97838)	1.515	0.5266	8.28	0.0040

PARSIMONY

- “Adopting the simplest assumption in the interpretation of data”
- If you have a choice between two models, go for the one with less variables (assuming they both make sense)

SMALL POPULATIONS

- Slice data historically
- Get creative with your data sources
- Adjust the p-values in Analytics

STAGE FOUR

IMPROVE YOUR APPLICATION

VALIDATE YOUR MODEL

- Some variables will retain their predictive power and some will decay
- Use a decile analysis
  - By decile, look at how many gifts have come in – should correspond

POPULATION TO SCORE

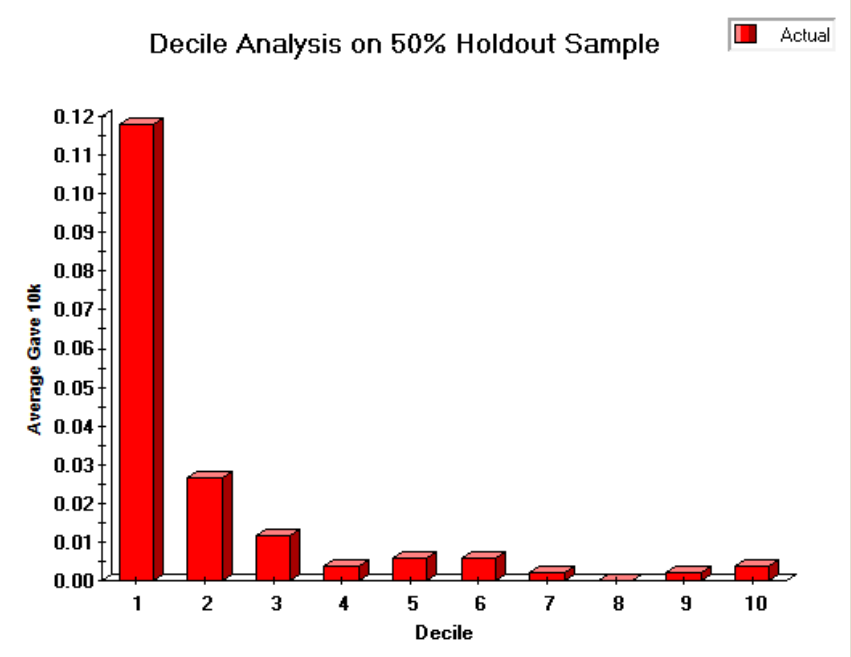
- Make sure this population is similar to the population you modeled
  - The same assumptions have to hold
- Different models for donors and non-donors due to giving history
- Consider the population you'd eventually want to score, and build a model tailored to that population

ON SCORING

- Don’t dive too deep
  - The most quality records are in the first few deciles
  - Selecting records from latter deciles can decrease productivity
- No set cut-off of a good vs. bad score
  - Let your decile analysis tell you where the cutoff is

ON SCORING

- No set cut-off of a good vs. bad score
  - Let your decile analysis tell you where the cutoff is



RESOURCES

- Online courses (Coursera, etc.) to help build dataset
- Conferences: APRA, NEDRA, etc.
- The blog
- Rapid Insight team
- Prospect DMM!