# 2015 Analytic Challenge

## *HEAD-TO-HEAD COMPETITION*


John Young

SVP, Analytic Consulting

EPSILON


&THEN A DMA EVENT

# SPONSOR - EPSILON AT A GLIMPSE

A Global Marketing Services Provider…

- **7,000** Associates globally
- **70+** Offices
- **150+** Marketing databases
- **1.5B** Individual records
- **278M+** Device IDs
- **47B+** Email messages per year
- **50B+** Bid requests per day
- **250M+** Memberships managed

… Helping Brands Bond with Consumers

# THE DMA ANALYTICS COMMUNITY

***Mission …***

Provide services, opportunities and resources that advance members' educational, social and professional development -- for marketing strategists, analytic practitioners as well as managers and executives responsible for leveraging analytics to drive return on marketing investment

---

***What the Community Provides …***

- **Monthly Webinars and Town Halls**: Events focused on hot topics, best practices and emerging trends in the analytics field, presented by analytic experts
- **Analytics Journal:** Annual publication featuring thought leadership in data analytics for marketing -- new approaches to analytics, advances in statistical methodologies and optimization, attribution, big data, behavioral, mobile, and predictive analytics
- **Analytics Advantage Blog Series:** Resource where analytic professionals and marketers find case studies and success stories to aid in powering data-driven marketing. Marketing strategists find ideas to better partner with and utilize the talents of their analytic teams

# THE ROLE OF THE ANALYTIC CHALLENGE

- Launched by the DMA Analytics Council in 2006
- Raise the visibility of analytics as a critical enabler of better business outcomes
- Allow practitioners to go “head-to-head” in building a model to support a real-world marketing challenge
- Share best practices and facilitate the exchange of ideas – allowing practitioners to raise their game and increase the value of their work

# THIS YEAR'S CHALLENGE

- A direct marketer of consumer goods seeks to increase repeat purchases of its flagship product -- looking for a targeting tool to increase precision of marketing to 1X buyers

- Challenge participants were asked to build a model that identifies 1X buyers with the *greatest likelihood* of making a repeat purchase

- Participants were supplied with both first-party and third-party data:
  - First party → customer characteristics & prior purchase behavior
  - Third-party → EPSILON TotalSourcePlus … hundreds of attributes capturing consumers' demographic, financial, behavioral, and lifestyle characteristics

- Solutions evaluated based on lift achieved at the 6th decile of a modeling hold-out sample

# THE COMPETITORS

## Suppliers

Acme Explosives
Alight Analytics
Allant Group
Alliance Data Systems
Analysis
Bisnode Belgium
Catalyst Direct
CDG Consulting Group
Cogensia
Cognilytics Software & Consulting
Contemporary Analysis
Customer Analytics India
DataLab USA
DM Group srl
DX Marketing
eleventy marketing group
Eric Novak & Associates
FCB
Focus Optimal
Focus USA
iknowtion
KBMG
Marketing Metrix
Merkle
MRM End to End
MSC - A Valid Company
Ogilvy CommonHealth
Ogilvy One Paris
Outsell
Rapid Insight Inc.
Saatchi & Saatchi Wellness
Semcasting
SIGMA Marketing Insights
Sparkroom
Strategic America
The frank Agency
The Lukens Company
Web Decisions
Whereoware
Wunderman

## Corporations

Best Buy
Capstone Associated Services
FedEx
Foremost Insurance
H&R Block
IBM
Protective Life
Springleaf Financial
Transamerica

## Academia & Organizations

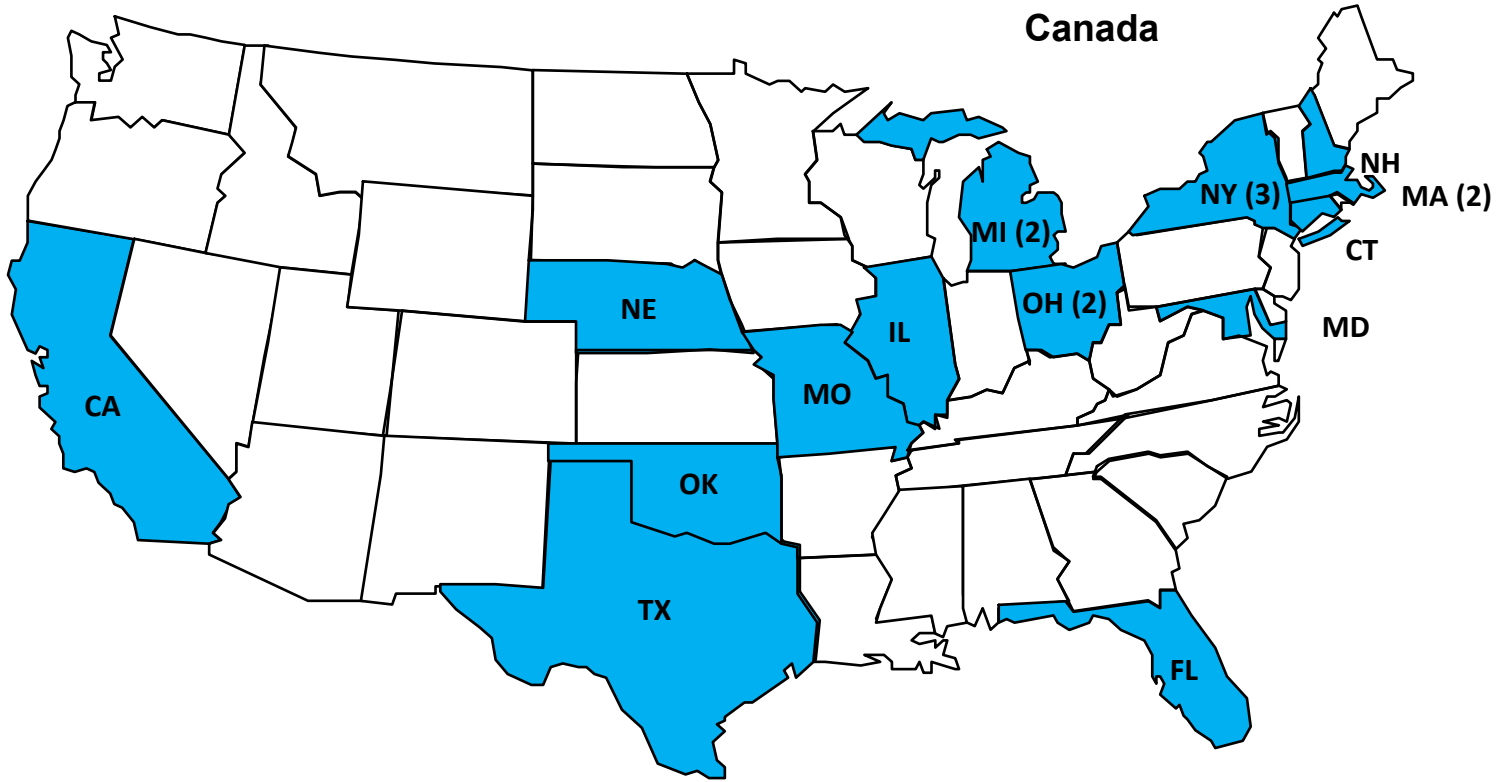
Golden Gate University
Indiana University South Bend
Montclair State University
State University of New York at New Paltz
The University of Alabama
University of Connecticut
University of Southern California
USGA

# THE 23 FINAL COMPETITORS

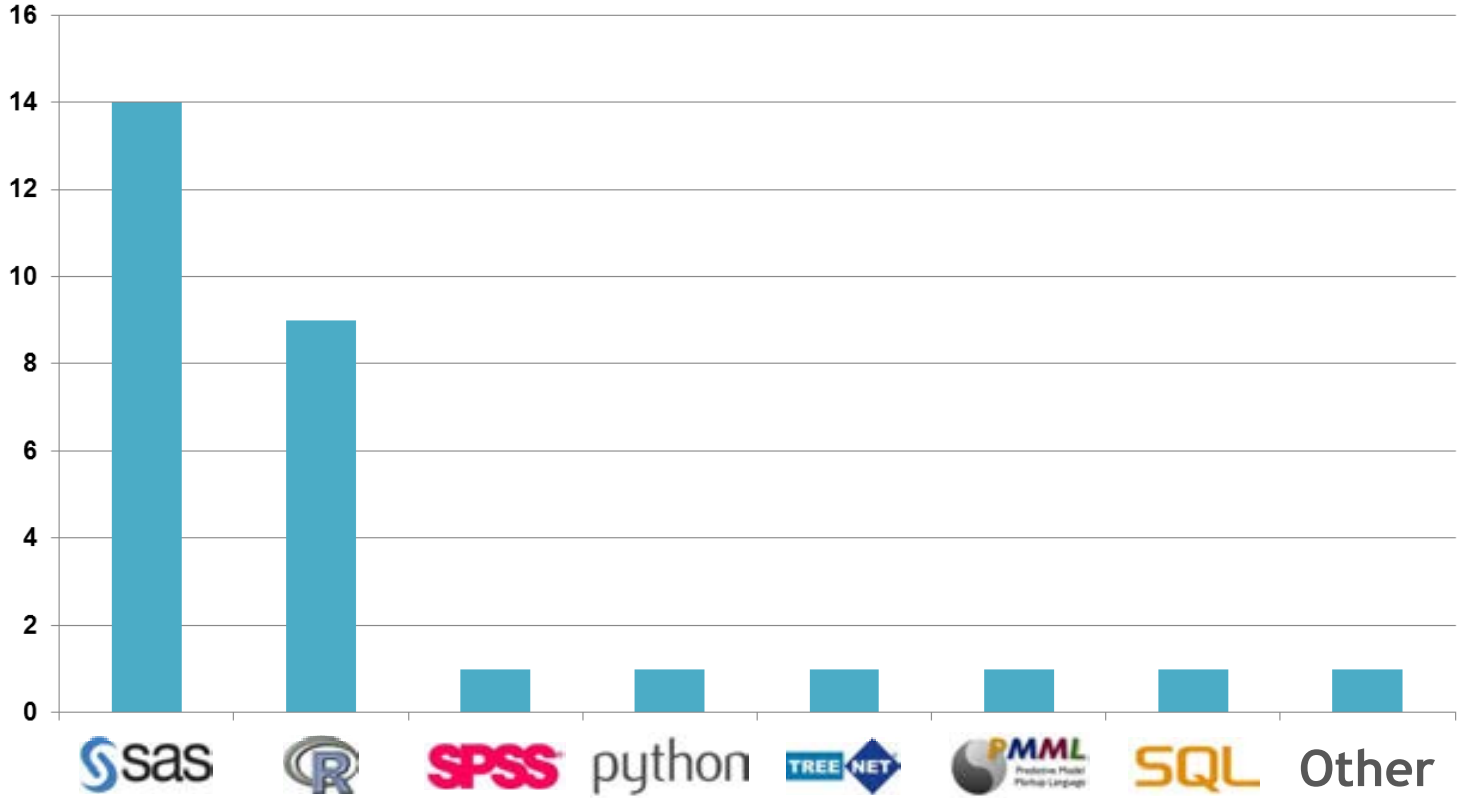
-Acme Explosives
-Allant Group
-Alliance Data Systems
-Best Buy
-Bisnode Belgium
-Catalyst Direct
-Cognilytics Software & Consulting
-Contemporary Analysis
-DataLab USA
-DX Marketing
-eleventy marketing group
-Focus Optimal
-Foremost Insurance
-H&R Block
-KBMG
-Marketing Metrix
-Merkle
-MSC – A Valid Company
-Rapid Insight Inc.
-Semcasting
-State University of New York at New Paltz
-Transamerica
-University of Connecticut


&THEN A DMA EVENT

# GEOGRAPHIC BREAKDOWN OF COMPETITORS



# SOFTWARE USED BY COMPETITORS




* These are not mutually exclusive.

# MODELING TECHNIQUES USED BY COMPETITORS


* These are neither exhaustive nor mutually exclusive.

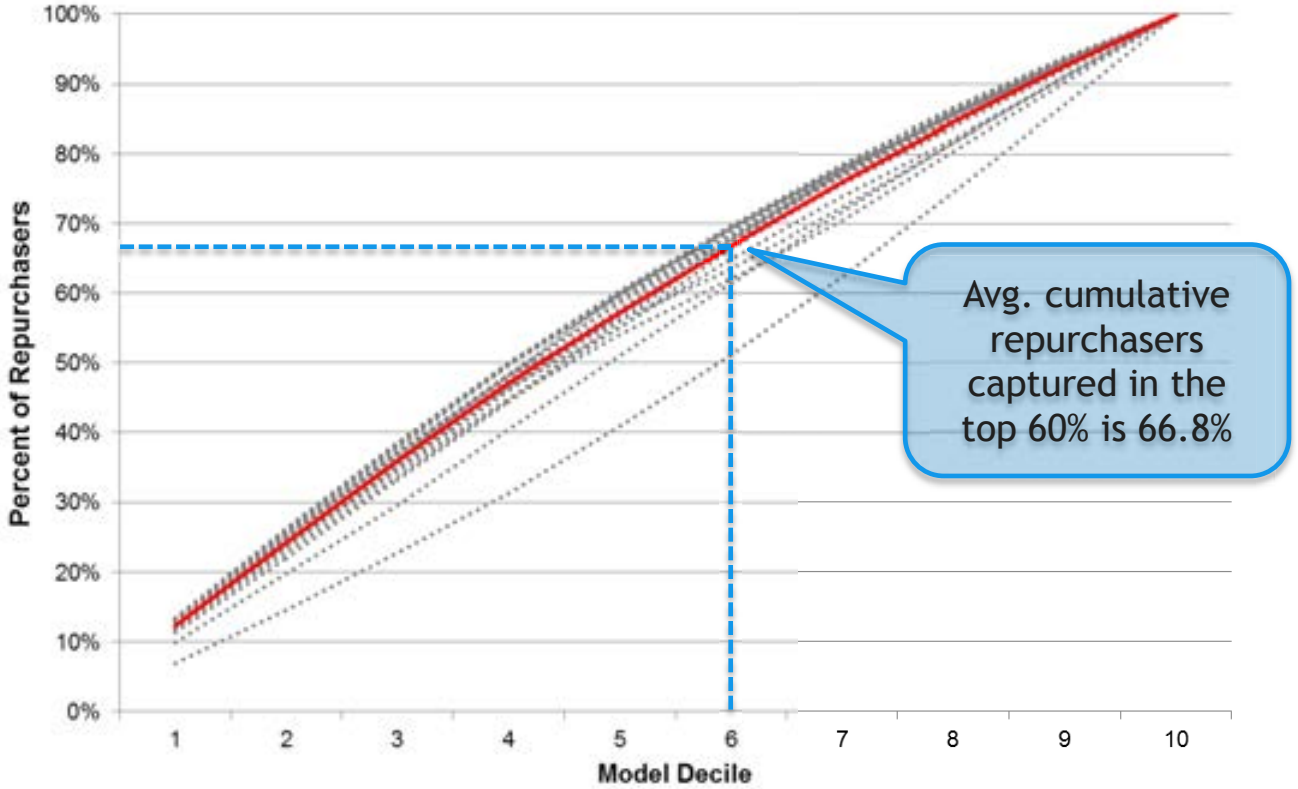
# THE EPSILON EVALUATION COMMITTEE

**Qizhi Wei**
Vice President

**Danny Jin**
Director

**Wenli Zhou**
Statistician

# COMPARISON OF MODEL PERFORMANCE




* Performance shown on the validation sample

# THE WINNERS


UCONN
UNIVERSITY OF CONNECTICUT
A VALID Company
MSC
DATABASE SOLUTIONS
Marketing. Strategic. Customized.
DATALAB USA
1
2
3

# 2015 Analytic Challenge

KARAN SARAO

UCONN

&THEN A DMA EVENT

# TEAM

- Karan Sarao

# ANALYTIC SOFTWARE USED

- Data Preparation – SAS
- Model Building – R
- Hardware
  - Acer Aspire 5750
  - 6 GB RAM

# SOLUTION OVERVIEW

## Data Preparation

**Missing Value Treatment**

- Nominal – New Category
- Numeric/Ordinal – Replace with 0 (Value)

**New Variable Creation**

- Multiple derived Variables

## Model Tuning and Stacking

Training / Blending /Testing Split

Caret Function to tune Multiple Model parameters

Stacking and Testing to optimize sequence

## Final Modeling

2 Stage Modeling process adopted

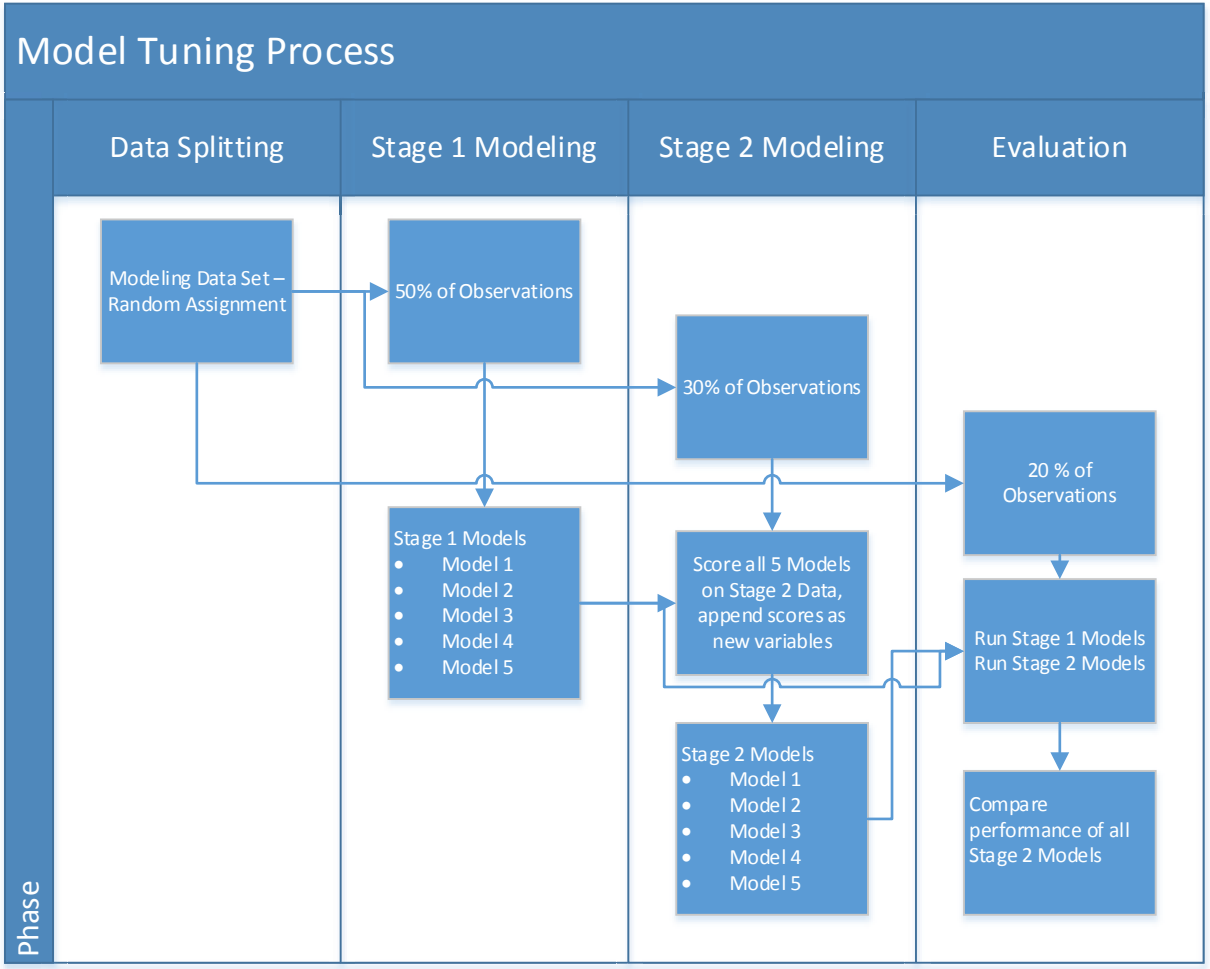
Initial set of optimized models created in Stage 1

Scores incorporated into final blended Model in Stage 2

## Scoring

2 Stage scoring process followed

# SOLUTION OVERVIEW - Continued (Model Tuning)



# DATA TRANSFORMATIONS

- Mix of Linear and Non Linear (Tree Based) Models
  - Cover each others weakness
  - Tree based models are invariant to order preserving transformations (no need for Log/Exponent etc.)
- More focus on feature engineering, new variables created as below →
  - SHIP_RATIO → (ORDER_SH_AMT+ORDER_ADDL_SH_AMT)/ORDER_GROSS_AMT (Does shipping cost as a ratio of the initial order have any influence)
  - PAYMT_RATIO=(ORDER_SH_AMT+ORDER_ADDL_SH_AMT+ORDER_GROSS_AMT)/PAYMENT_QTY (What is amount of each payment)
  - REV_RATIO=TOTAL_REV_PRIOR_TO_A/TENURE (Revenue ratio per unit tenure)
  - REV_PER_ORDER=TOTAL_REV_PRIOR_TO_A/TOTAL_ORDERS_PRIOR_TO_A (Revenue per order)
  - FIRST_ORDER_RATIO=ORDER_GROSS_AMT/ITEM_QTY
  - FIRST_PAYMENT_RATIO=ORDER_GROSS_AMT/PAYMENT_QTY
  - ORDER_FREQ=TENURE/TOTAL_ORDERS_PRIOR_TO_A
  - ORDER_DUE_RATIO=RECENCY/ORDER_FREQ
  - ORDER_DUE_RATIO_2=(RECENCY-ORDER_FREQ)/ORDER_FREQ
  - ORDER_DUE_RATIO_3=(RECENCY-ORDER_FREQ)/RECENCY
  - All divide by zero exceptions set to 0

# Stage 1 Models

## Multiple Models trained on 50% of the data

- Random Forests (randomForest)
- AdaBoost (ada)
- Gradient Boosting Machines (gbm)
- eXtreme Gradient Boost (xgboost)
- Logistic Regression (variables selected by studying glmnet output)
- Regularized Logistic Regression (glmnet)

## Several of the above models have tunable parameters

- Caret package in R used to cycle through various combinations of input parameters using multiple folds
- Problem statement specifies rank order primacy, hence ROC metric maximized

# Stage 2 Models

- All 5 Models built in stage 1 used to score both Stage 2 and evaluation data
- 5 score columns added  back to the data set (stage 2 and evaluation)
- 4 Models created again on Stage 2 dataset
- Stage 1 and Stage 2 models are scored on evaluation dataset
- ROC (AUC) calculated for the models on evaluation dataset
- Best Model identified – xgboost (Stage 2)

| Model | Stage 1 (AUC) On EvaluationSet | Stage 2 (AUC) On EvaluationSet |
|---|---|---|
| **xgboost** | **0.646** | **0.647** |
| logit | 0.641 | 0.646 |
| gbm | 0.636 | 0.644 |
| glmnet | 0.641 | 0.642 |
| ada | 0.637 | 0.642 |
| random forest | 0.617 | NA |

# Final Model Building

- Data split as 50-50 between Stage 1 modeling and Stage 2 blending
- Xgboost used to blend in Stage 2
- Initial 5 models score  the submission dataset and scores merged back to create dataset for sixth model
- Blend Model used to generate the final submission score

# TOP VARIABLES

- Mix of ready and derived variables
- Ranking of top variables can be difficult to quantify across multiple modeling techniques/blends
- Plain logistic regression with these variables can create a Model with comparable performance (~.64 AUC)

| Important Variables |
| --- |
| TXN_CHANNEL_CD |
| PAYMENT_QTY |
| RUSH_ORD_FLAG |
| SHIP_RATIO |
| FIRST_ORDER_RATIO |
| DEMOGRAPHIC_SEGMENT |
| ORDER_GROSS_AMT |
| RETAIL/CATALOG_SPENDING_QUINTILE |
| REV_PER_ORDER |
| HH_INCOME |
| PAYMT_RATIO |
| ETHNICITY |
| LANGUAGE |

# KEYS TO SUCCESS

- Derived Variables
  - Create as many behavioral/pattern variables as possible
  - Ratios such as revenue/order, order frequency, shipping cost to total cost etc.
- Cross Validation for controlling overfit
  - K fold (maximum possible) validation runs
  - Tune parameters (control depth and boosting rounds to maximize test ROC)
  - Use grid search for optimum parameter search or employ Caret package

# 2015 Analytic Challenge


Aaron Davis

EVP, Analytics DATALAB USA


&THEN A DMA EVENT

# TEAM

- Aaron Davis – Presenter
- Adam Bryan – Participant/Coordinator
- Jeremy Walthers – Participant
- Julia Wen - Participant

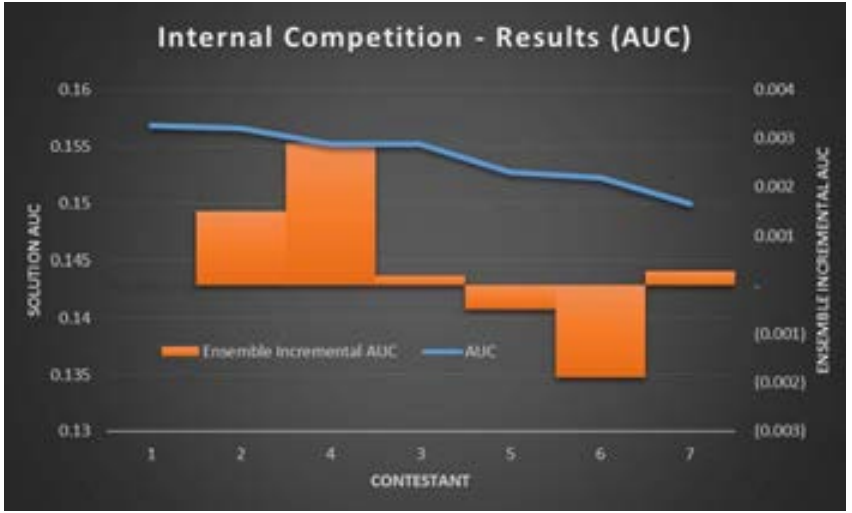
&THEN a DMA event

DATA LAB USA

# SOLUTION OVERVIEW - Internal Competition

- **Stage #1 – Internal Competition**
  Members of DataLab's Analytics Team developed their best individual solutions.
  - Seven contestants used a mix of DataLab's established best practices and new/non-standard approaches.
  - One week time frame. All work done outside of working hours.
  - Results evaluated on a 20% internal validation.
    - 1st Place – Judged based on best raw discrimination – AUC
    - 2nd/3rd Place – Judged based on incremental performance gain using crude ensemble approach with 1st place solution.

- Results:
  - Very minor differences in performance between candidate solutions.
  - Best performing solution utilized DataLab's proprietary methodology for hyper parameter tuning and variable selection. Model was developed in less than two hours.
  - Solutions leveraging different approaches in general showed larger incremental gains when ensemble with the 1st place solution.



# SOLUTION OVERVIEW - Final Entry

- **Stage #2 – Final Entry**
  - **One Day Timeframe**

  - **Two competing paths:**
    - **Ensemble 1st & 2nd Place Models – Gain performance by leveraging multiple models**
      - **More work to code solution for entry**
      - **Less representative of a real world solution**
      - **Typically requires holding out portion of data in sub-models**
      - **Requires more coordination between team members**
    - **Refit 1st Place Model – Gain performance by leveraging 100% of Available Experience**
      - **Increase in performance likely less than ensembling**
      - **Easy to code solution**
      - **More representative of a real world solution**
      - **Less time intensive**

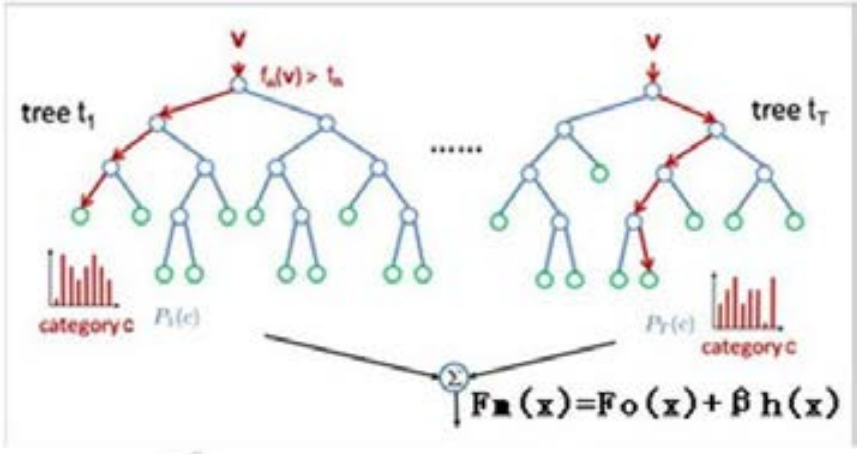
  - **Chose to submit single model fit on 100% of available experience**
    - **Expected .002 increase in AUC**

# MODELING TECHNIQUE(S)

- **Data Prep:**
  - High Order Categorical Data - one hot encoding of high level categorical features
  - Missing Data - Surrogation & creation of dummy flags to identify missing features.

- **Algorithm:** Gradient Boosted Decision Trees

- **DataLab Predictive Modeling Toolkit:** Massively parallelized heuristic methods for parameter tuning & feature selection
  - Optimal parameters/features result of 1000's of predictive model experiments

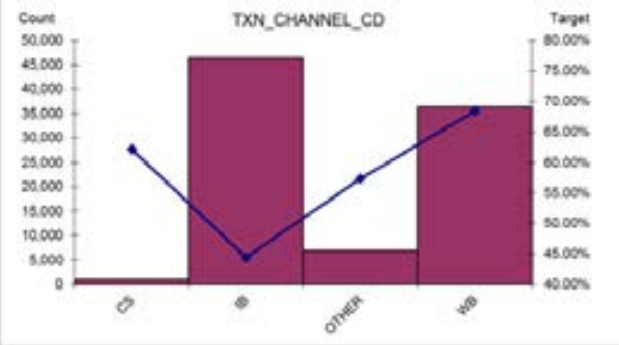


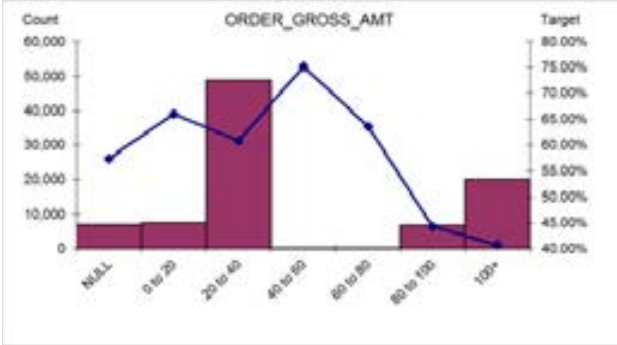

Microsoft SQL Server

SPM
SALFORD PREDICTIVE MODELER
CART MARS TREENET Random Forests

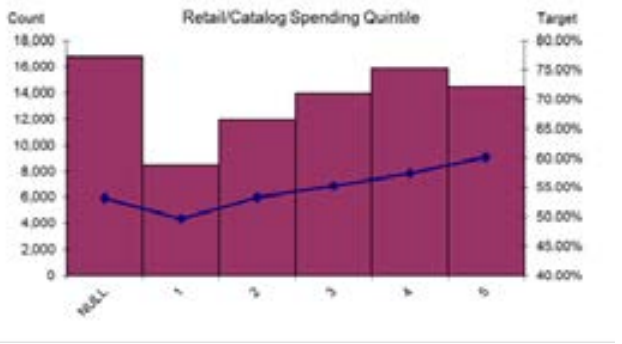
R

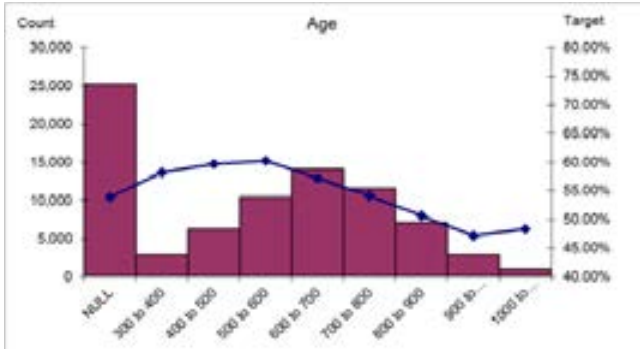
# TOP VARIABLES







# KEYS TO SUCCESS

- Team based solution
- Domain experience
- Data Prep
- Parameter Tuning

# 2015 Analytic Challenge


**Scott Ross**

Senior Database Architect


A VALID Company

MSC

DATABASE SOLUTIONS

Marketing. Strategic. Customized.

&THEN A DMA EVENT

# TEAM

- Scott Ross
- Gary Abrams
- Nataly Slobodsky

***We deliver custom marketing database solutions for our B2C & B2B clients that enable them to better engage with their customers***

- Public / Global
- US Offices in Chicago & Los Angeles
- 35+ years in business
- Average Client Tenure > 12 years

# ANALYTIC SOFTWARE USED

- ModelMax
- R

# SOLUTION OVERVIEW

- **DISCOVERY**
  - Look for fields that have issues (inconsistent data, possibly unreproducible values)
- **TRANSFORMATION**
  - Convert variables that need help to be more predictive
- **BUILD**
- **TEST**
  - Look at the resulting lift, and the curvature of that lift for gains and consistency
  - Ideally there is a forward test, using a following mailing
- **REPEAT**
  - Remove variables that prove statistically insignificant
  - Create more granular transformations on variables that are marginally significant

# DATA TRANSFORMATIONS

- **Horizontal binning**
  - Allow significance of categorical values to come through.
- **Ages (adult and children)**
  - Continuous numerical day values for higher granularity.
- **Continuous values that are non-linear**
  - Transform via formula, or with binning.

# VARIABLE SELECTION/REDUCTION

- **1-way ANOVA**
  - Identify significant variables.
- **Means plots**
  - Discover the nature (linear and non-linear) of relationships on continuous variables
- **Correlation matrices**
  - Identify co-linear variables
- **Stepwise process on the variables**
  - Identify the most important predictors.

- **Ended up throwing all this out, and used the “kitchen sink”.**

# MODELING TECHNIQUE & TOP VARIABLES

**Technique:** Binary Logistic Regression Model

**TOP VARIABLES**

1. PAY_TYPE_CD [Method of Payment]
2. ITEM_QTY [Purchase Quantity of Initial Order]
3. Ethnicity [Ethnic background]
4. PAYMENT_QTY [bin of 1 payment was pertinent]

# KEYS TO SUCCESS

What we would have done with more time

- Data reduction
- Forward Testing
- Ensemble Modeling
  - Would have loved to include a performance model
- Looked for relationships between variables to create additional calculated fields


QUESTIONS?
&THEN A DMA EVENT
EPSILON