



INSIGHT  
PHILANTHROPY  
RESULTS

# EXPLORE

PD25

## New Depths

August 19-22, 2025

Hilton Baltimore Inner Harbor Hotel, Baltimore, Maryland



John Sammis, CCS Fundraising

# AN EXPLORATORY APPROACH TO PREDICTIVE MODELING

**EXPLORE**  
PD25  
*New Depths*

While you're waiting, complete your  
session evaluations in the mobile  
app!



# PRESENTER

John Sammis

- CCS Fundraising
- Senior Vice President, Data Analytics Practice
- 30 + years of statistical Analysis experience
- 20+ years experience building predictive models for fundraising organizations
- BS/Chem Engineering/Clarkson University
- MBA/Cornell University



# WHAT IS PREDICTIVE MODELING?

**EXPLORE**  
PD25  
*New Depths*

# PREDICTIVE MODELING

## What

- Statistically characterize “good donors” to find constituents who look like good donors but aren’t “good donors” yet
  - Replace “good donors” with other goals
    - Planned givers
    - Volunteers

## Why

- Find new prospects
- Evaluate assigned portfolio
- Annual Fund segmentation
- Other



# PREDICTIVE MODELING

## How

- Map organizational objective to dependent variable
  - Lifetime giving, last 10 years giving, frequency of giving, etc.
- Define cohort to be scored
  - Non-deceased individuals, young alums, cancer service line donors, etc.
- Prepare independent variables for the model
  - Categorical variables
  - Continuous variables



# TOOLS AND DATA

Much effort focused on finding the optimal tools

- Linear regression, logistics regression, classification trees, Random Forest, XGBoost etc.
- Important to match the tool to the data and objective

But there should be just as much focus on exploring, diagnosing and transforming the data, both before and during the modeling process





# EXPLORATORY DATA ANALYSIS (EDA)

**EXPLORE**  
PD25  
*New Depths*



# JOHN TUKEY

Famous Princeton statistician and father of Exploratory Data Analysis (EDA)

- Early statistics (through 1970's)
- Data explosion (1980's+)
  - Growth in “serendipitous” data
  - EDA: a set of tools and principles for finding patterns, outliers and subgroups in data
    - Data-driven decisions
    - Emphasized “looking” at your data
    - Revise hypothesis and collect new data



# EDA AND MODELING

Before building the model:

- Categorical variables
  - Goal – generate indicator variables (1/0)
  - Frequency tables
    - How many records are populated with data
    - How many categories
    - Are there are any oddities

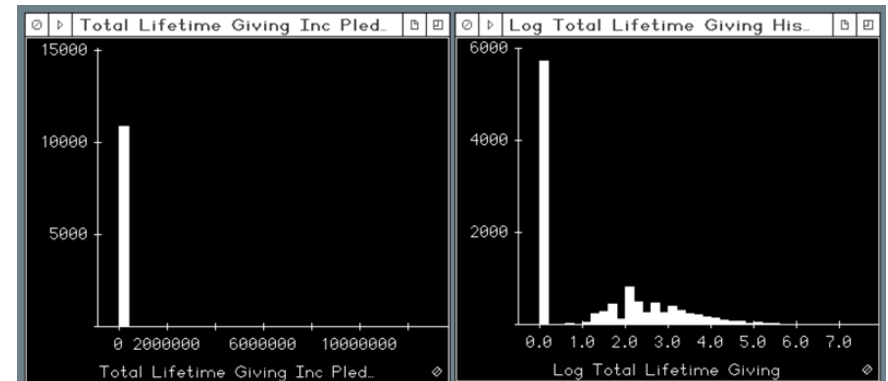
Frequency breakdown of			Marital Status
No Selector			
10921 total cases of which 5509 are missing			
Total Cases		5412	
Number of Categories		6	
Group	Count	%	Cumulative %
Divorced	415	7.668	7.668
Married	4163	76.922	84.590
Separated	60	1.109	85.698
Single	582	10.754	96.452
Widow	182	3.363	99.815
Widow(er)	10	0.185	100.000



# EDA AND MODELING

Before building the model:

- Continuous variables
  - Goal – generate:
    - Variables that are unimodal and symmetric (reasonably bell-shaped)
    - Indicator variables (1/0)
  - Summary statistics
  - Histograms or normal probability plots

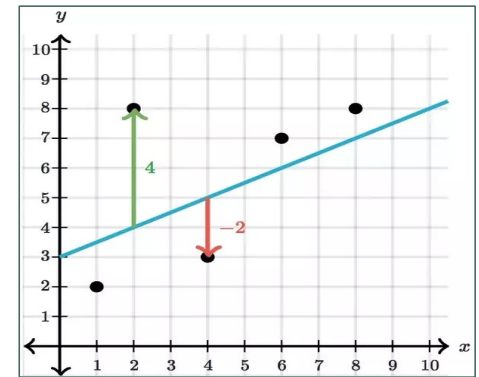
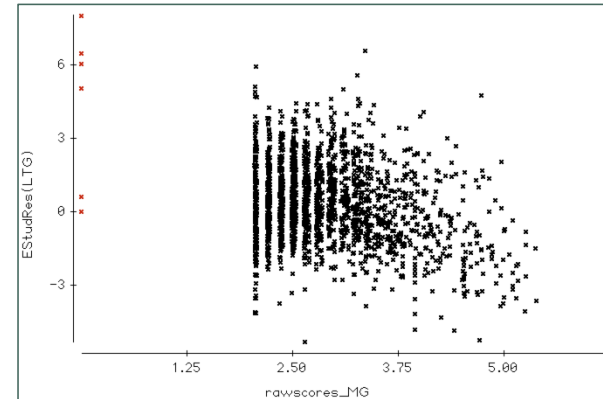




# EDA AND MODELING

While building the model:

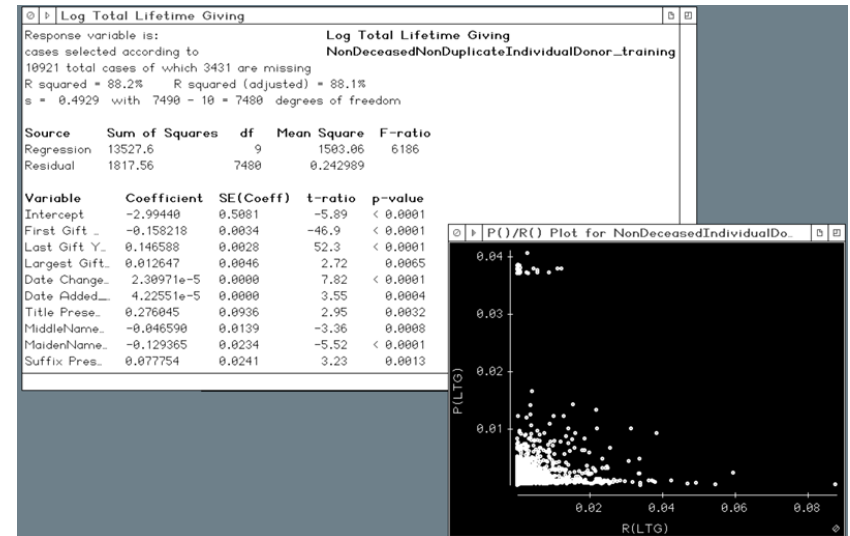
- Diagnostics
  - Residual:
    - Vertical difference between the Actual Value and the Predicted Value
    - Outlier residuals could indicate typos or subgroups that need to be treated specially
    - Typically display in a scatterplot versus the actual values.



# EDA AND MODELING

While building the model:

- Diagnostics
  - Leverage:
    - The influence of an individual point on a statistical model
    - Think teeter-totter:
    - Extreme points in the x-space will meaningfully affect the model
    - Leverages, and their corresponding plots help identify high-leverage points and sub-groups, including those points that are extreme in the multidimensional space



# EDA AND MODELING

While building the model:

- Diagnostics
  - Others:
    - Partial Regression Plots
    - DFFITS
    - Cook's Distance
    - Hadi's Influence







# INTERACTIVE MODELING

**EXPLORE**  
PD25  
*New Depths*

# INTERACTIVE MODELING & DIAGNOSTICS

- Create Diagnostic plots
- Add potential predictor variables a few at a time
- Evaluate the diagnostic plots
  - Are there outliers, odd patterns or subgroups?



# FIXING THE PROBLEMS

- What do we do when we find patterns, outliers or subgroups?
  - Look for typo or miscoding and fix
  - Impute with median when true missing
  - Try a transformation, or apply binning
  - Remove the offending variable from the model
  - Move the case from the training set to the validation set

*Live example*







# WRAP UP

**EXPLORE**  
PD25  
*New Depths*

# KEY TAKEAWAYS

- Properly exploring and adjusting data is important for building good models
- Errors and anomalies should be fixed
  - Typos
  - Transformations
  - Binning
  - Imputation
  - Variable exclusion
  - Case removal
- Ongoing process before and during the modeling process





A dark blue background with a complex, light blue topographic map pattern of contour lines.

# THANK YOU!

Please complete your session  
evaluations in the mobile app.

A green-tinted aerial photograph of a city street grid, showing buildings, roads, and parks.

apra