USDA

United States Department of Agriculture

One Team, One Purpose

Food Safety and Inspection Service

Protecting Public Health and Preventing Foodborne Illness

Data Mining for Developing Efficient Food Hazard Sampling Plans

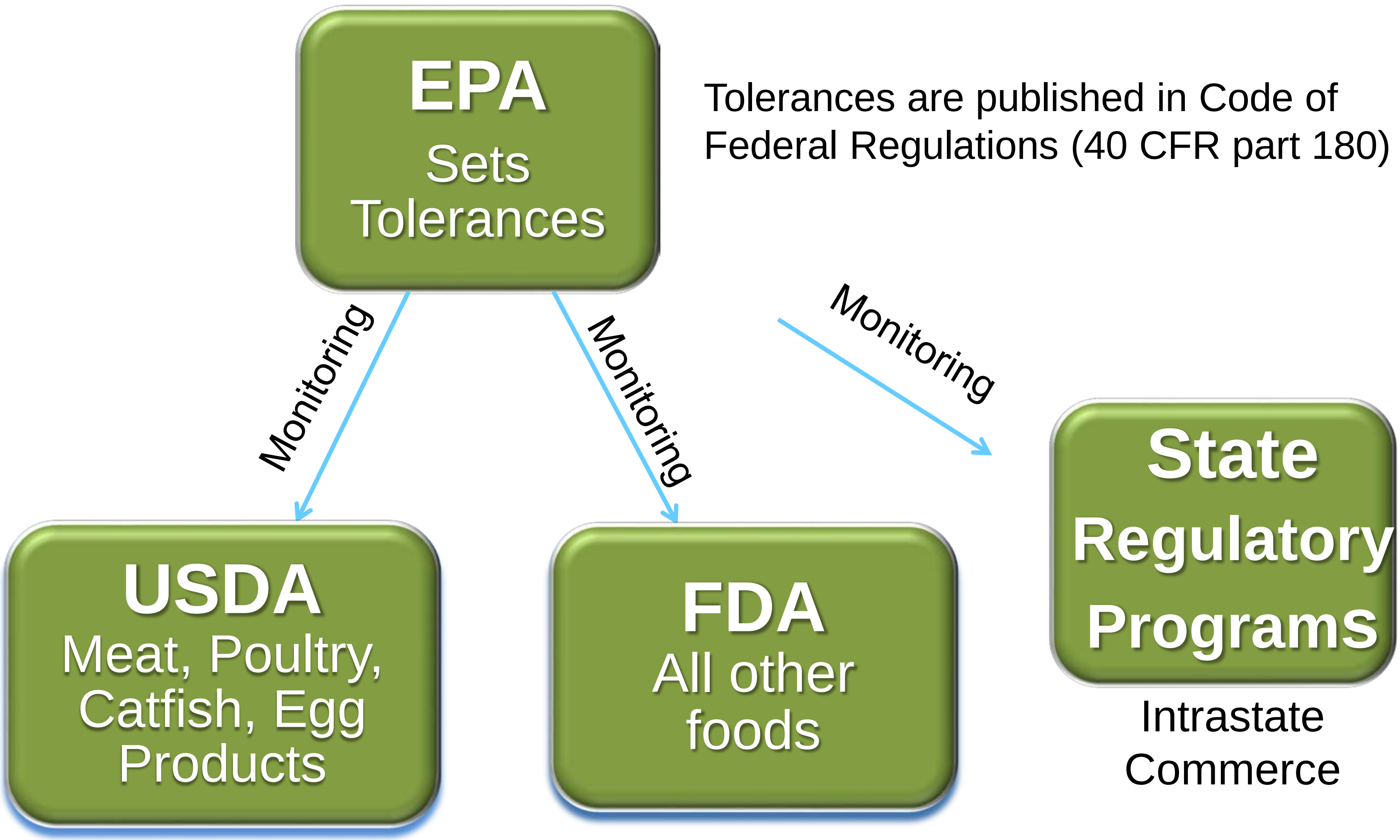
John J Johnston PhD MBA
Scientific Liaison/Interdisciplinary Scientist

International Association for Food Protection Conference
Salt Lake City, UT – July 10, 2018

Presentation Overview

- Review of U.S. pesticide monitoring programs
- Data mining project
  - Objective
  - Methods
  - Results
  - Conclusions
  - Future work

Monitoring Pesticides in Domestically Produced Foods



Interstate Commerce

Monitoring Pesticides - Enforcement

- USDA and FDA sample products and hold pending results:
  - Pesticide concentration is $\leq$ US tolerance = non-violative
  - Pesticide concentration > US tolerance = violation
  - Pesticide detected with no tolerance = violation

USDA Pesticide Data Program (Non-Regulatory)

- USDA AMS leads the Pesticide Data Program (PDP)
  - Provides pesticide exposure data for use by EPA in risk assessments and pesticide re-registration
  - Testing performed by State Departments of Agriculture and USDA

Data Mining

- The process of extracting patterns from large data sets by combining statistics and artificial intelligence with database management to permit improved decision making.

Objective

- Proof of Concept: illustrate how data mining can be applied to develop sampling plan resulting in increased probability of identifying foods with pesticide violations.

Methods – Project database

- Database 2015 USDA AMS Pesticide Data Program Analytical Results
  - 10,187 Sample cases
    - Sample case = Produce Sampling event
  - 2,333,852 results cases
    - 107 – 425 results cases associated with each sample case (mean = 229)
      - Results file case = analytical results for 1 pesticide analyte

Methods – Project database

- Analytical results flagged as either:
  1. Non-detect
  2. Detect pesticide $\leq$ tolerance
  3. Detect pesticide > tolerance
  4. Detect pesticide with no tolerance < LOQ
  5. Detect pesticide with no tolerance $\geq$ LOQ

Methods – Project database

- Analytical results flagged as either:

1. Non-detect
2. Detect pesticide $\leq$ tolerance
3. Detect pesticide > tolerance
4. Detect pesticide with no tolerance $\leq$ LOQ
5. Detect pesticide with no tolerance > LOQ

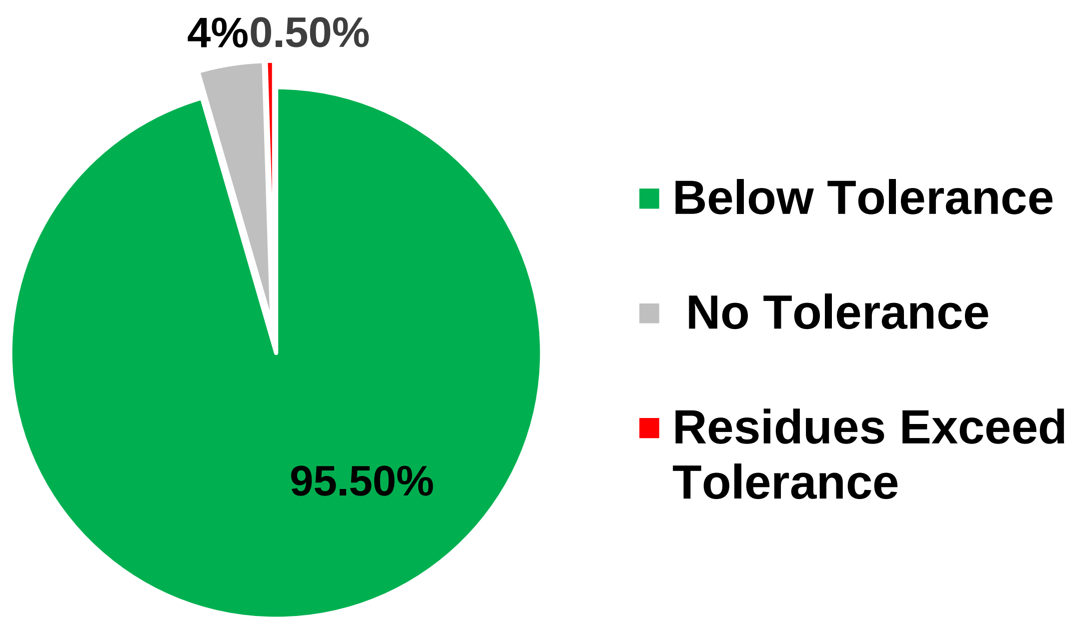
Presumed non-violative Presumed violative

Methods

- PDP sample and results files imported into Excel to facilitate data preparation
  - Data partitioning
  - Data reduction
  - Replacement
  - Spurious values
  - Data Transformations
  - Impute Data

Methods

- Excel sample file was converted into SAS file and imported into SAS Enterprise Miner
  - Target variable: violation



Methods

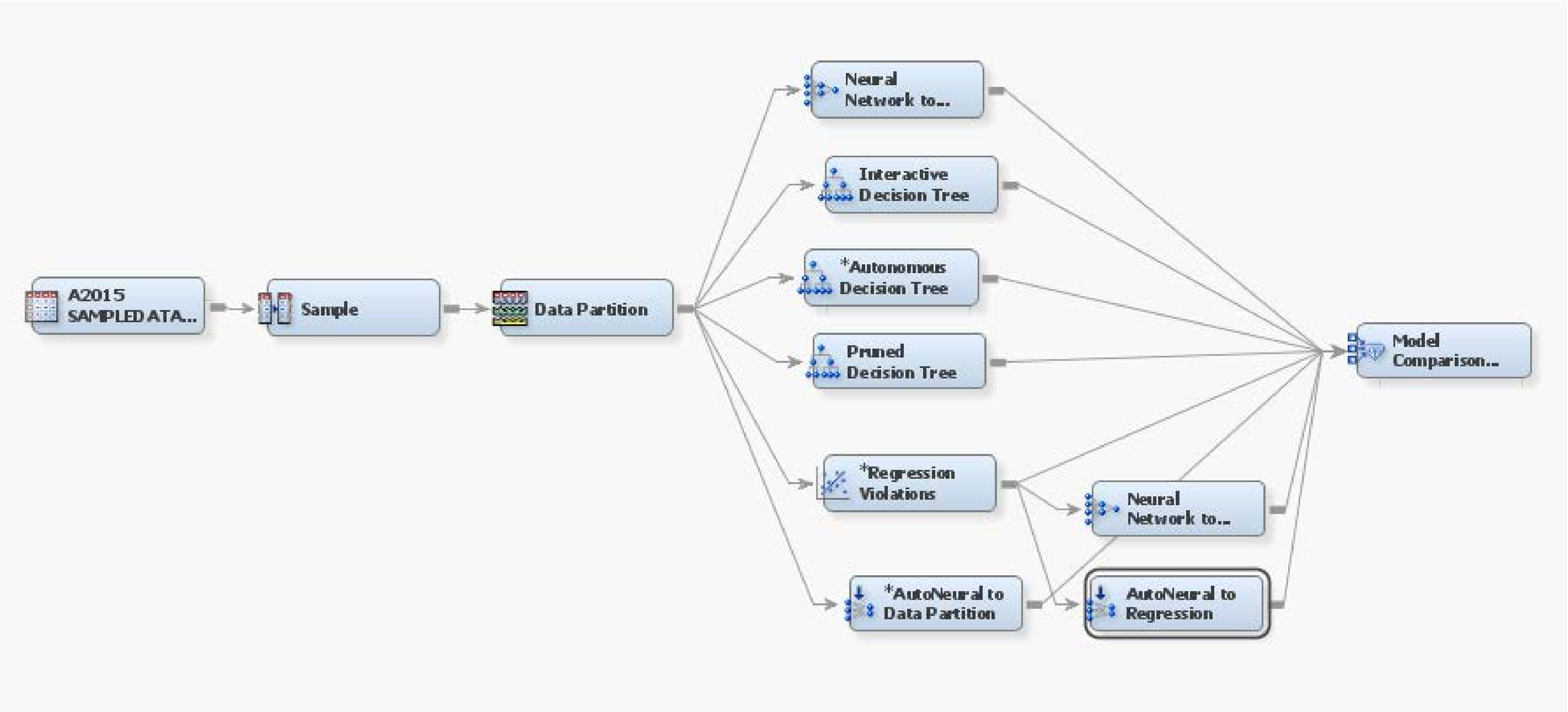
- Input variables:
  - Country of origin
  - Distribution state
  - Average latitude
  - Average temperature
  - Commodity
  - Commodity type
  - Claim
  - Distribution facility type

Methods

- Model Comparison
  - Models evaluated:
    - Decision trees
    - Regression models
    - Neural network models
  - Target = violation
  - Evaluation criteria: Misclassification rate

Methods

- Model Comparison



Results

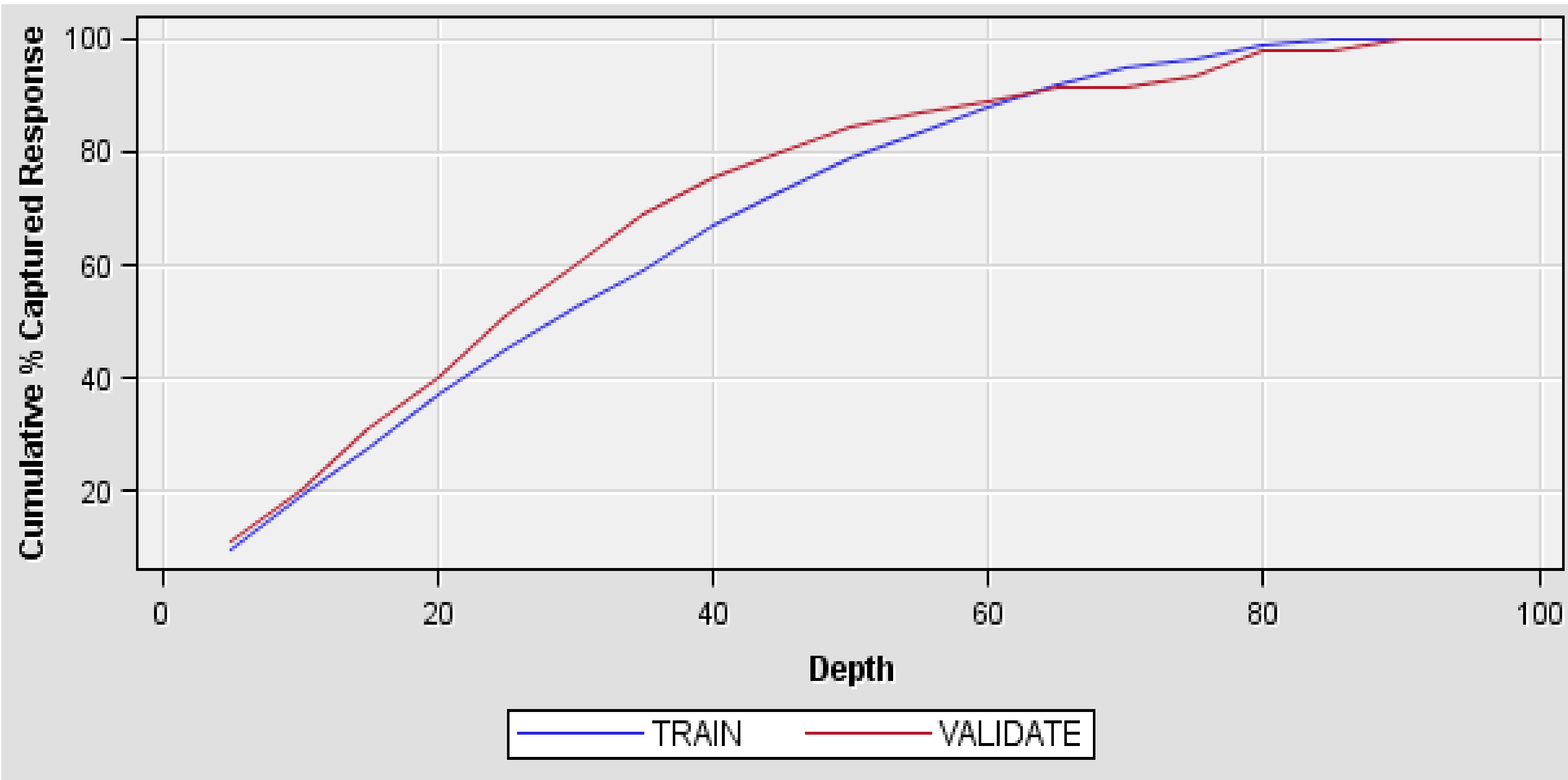
- Model Comparison

Winning Model →

Model Node	Pevious Node	Model Description	Target Variable	Validation Misclassification Rate
Regression	Data Partition	Partition to Regression	TotalViolation	0.1556
AutoNeural	Data Partition	Partition to AutoNeural Network	TotalViolation	0.1667
Regression	Auto Neural Network	AutoNeural to Regression	TotalViolation	0.1667
Neural Network	Data Partition	Data Partition to Neural Network	TotalViolation	0.1889
Neural Network	Regression	Regression to Neural Network	TotalViolation	0.2111
Autonomous Decision Tree	Data Partition	Data Partition to Autonomous Decision Tree	TotalViolation	0.2667
Interactive Decision Tree	Data Partition	Data Partition to Interactive Decision Tree	TotalViolation	0.2667
Pruned Decision Tree	Data Partition	Data Partition to Pruned Decision Tree	TotalViolation	0.2667

Results

- Cumulative Captured Response



Highest Predicted Probability of Violation Lowest

Results

- Regression results

Effect	DF	Chi-Square	Pr > ChiSq
AverLatitude	1	2.6619	0.1028
AverTemp	1	0.6715	0.4125
Claim	2	10.9993	0.0041
Commodity	19	146.8542	<.0001
Country	14	155.5018	<.0001
Month	11	11.3926	0.411

Results

- Odds ratio

Commodity	Odds Ratio
cherries	0
cherries frozen	0
corn fresh	0.001
apples	0.002
grape fruit	0.002
peanut butter	0.002
corn frozen	0.004
oranges	8.222
pears	11.748
grapes	12.562
potatoes	12.902
cucumbers	14.244
peaches	16.919
lettuce	55.881
green beans	71.149
tomatoes	79.942
nectarines	99.259
strawberries	118.527
spinach	999

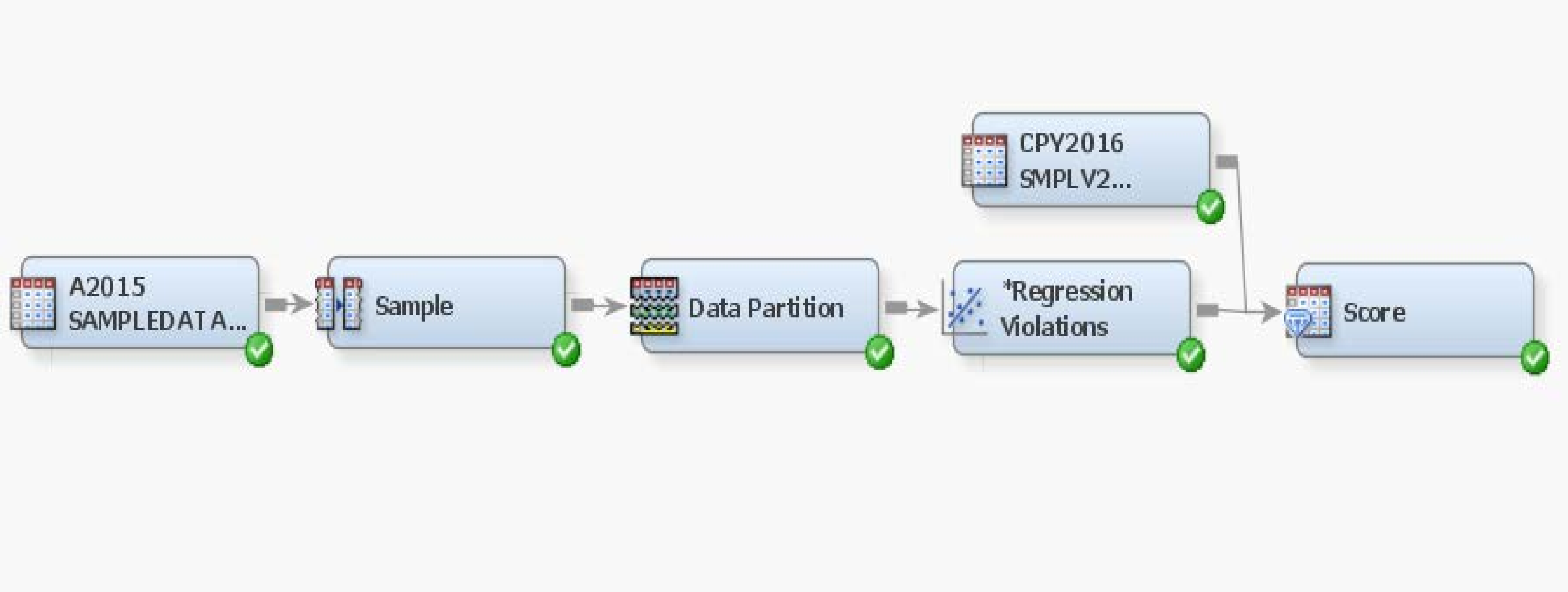
Results

- Odds ratio

Country	Odds Ratio
Guatemala	0
Netherlands	0.001
Honduras	0.002
Peru	0.004
Dominican Republic	0.006
Nicaragua	0.006
South Africa	0.006
Argentina	0.009
Australia	0.012
Italy	9.758
New Zealand	20.871
USA	27.101
Mexico	39.056
Chile	213.811
Greece	219.764
Canada	659.711
Turkey	999

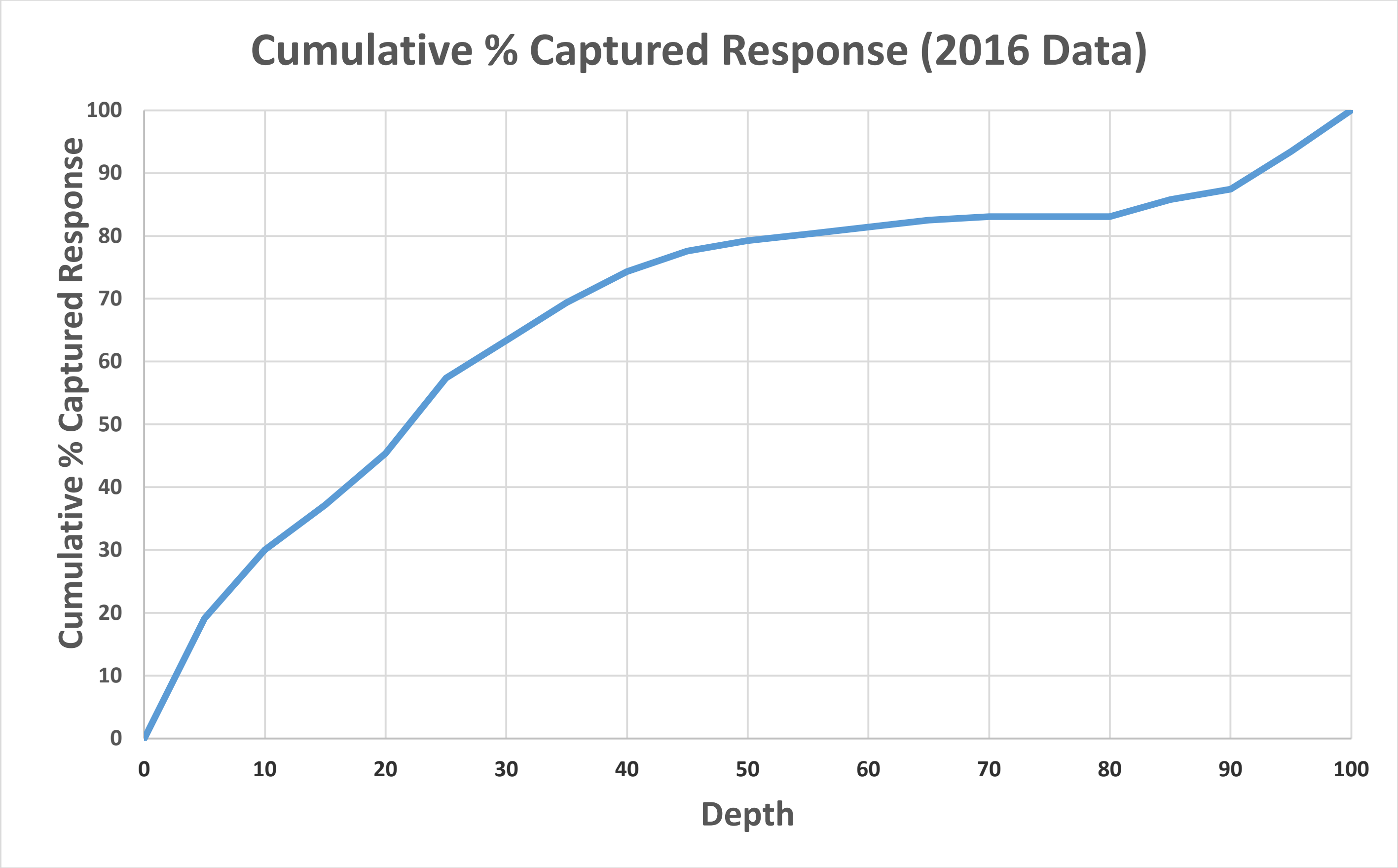
Results

- Scoring



Results

- Cumulative Captured Response



Highest Predicted Probability of Violation Lowest

Conclusions/Significance

- Data mining was used to successfully identify samples with higher probability of pesticide violations
- Model output could be used to develop Agency sampling plans that would increase the efficiency of pesticide monitoring
  - 80% of current violations could be detected by analyzing only 50% of current sample volume.
  - Remaining 50% of resources could be used to expand the variety of monitored commodities.

Future

- This approach is probably applicable to other prevalence (binary) food safety monitoring programs
- Other commodities
  - Meat
  - Poultry
  - Eggs
  - Fish
  - Dairy
- Other analytes of potential concern
  - Veterinary drugs
  - Environmental contaminants
  - Microbial hazards

Acknowledgements

- USDA AMS Pesticide Data Program
  - Diana Haynes
  - Roger Fry
  - Dawn Fay
- Leo Vijayasarathy, Computer Information Systems, College of Business, Colorado State University