Reality-Competition Shows: Are You In Or Are You Out?

Rebecca Kurtz

Ball State University

rpkurtz@bsu.edu

April 17, 2018

Overview

Title

- Is it possible to predict the winner of a reality-competition program?
- At what point in the competition can a winner be determined?
- What are the primary factors used to determine the winner?

Reality-Competition Programs

Definition by the Academy of Television Arts & Sciences:

- A reality series with a minimum of 6 episodes.
- A reality series with a competition component

Additional requirements to be considered in this study:

- For each elimination, the judges sort the contestants into five mutually exclusive categories: WIN, HIGH, IN, LOW, OUT.
- The series must average one elimination per episode/week.
- The season must be produced in the United States.
- All contestants must be 18 years old or older.

Timepoints Considered

- t_0: Before First Episode
- t_{Ep1}: After First Episode
- t_{25}: 25th Quantile
- t_{50}: 50th Quantile
- t_{75}: 75th Quantile
- $t_{Ep2ndLast}$: After Second to Last Episode

Competition Results

Skin Wars Season 2

	EPISODES									
CONTESTANTS	1	2	3	4	5	6	7	8	9	10
Lana	WIN	IN	HIGH	LOW	LOW	HIGH	WIN	HIGH	LOW	WIN
Avi	IN	WIN	HIGH	HIGH	HIGH	WIN	LOW	WIN	WIN	OUT
Aryn	IN	HIGH	IN	WIN	HIGH	HIGH	LOW	LOW	HIGH	OUT
Cheryl Ann	HIGH	HIGH	LOW	LOW	LOW	IN	LOW	LOW	OUT	
Rio	IN	IN	WIN	LOW	HIGH	LOW	HIGH	OUT		
Kyle	HIGH	IN	HIGH	IN	WIN	LOW	OUT			
Dawn Marie	IN	LOW	LOW	HIGH	LOW	OUT				
Sammie	IN	IN	LOW	IN	OUT					
Fernello	LOW	IN	IN	OUT						
Rachel	LOW	LOW	OUT							
Rudy	IN	OUT								
Macio	OUT									

Variables Considered

- TOP_Ep1
- BOTTOM_Ep1

- LeastBOTTOM
- AnyBOTTOM
- MostBOTTOM

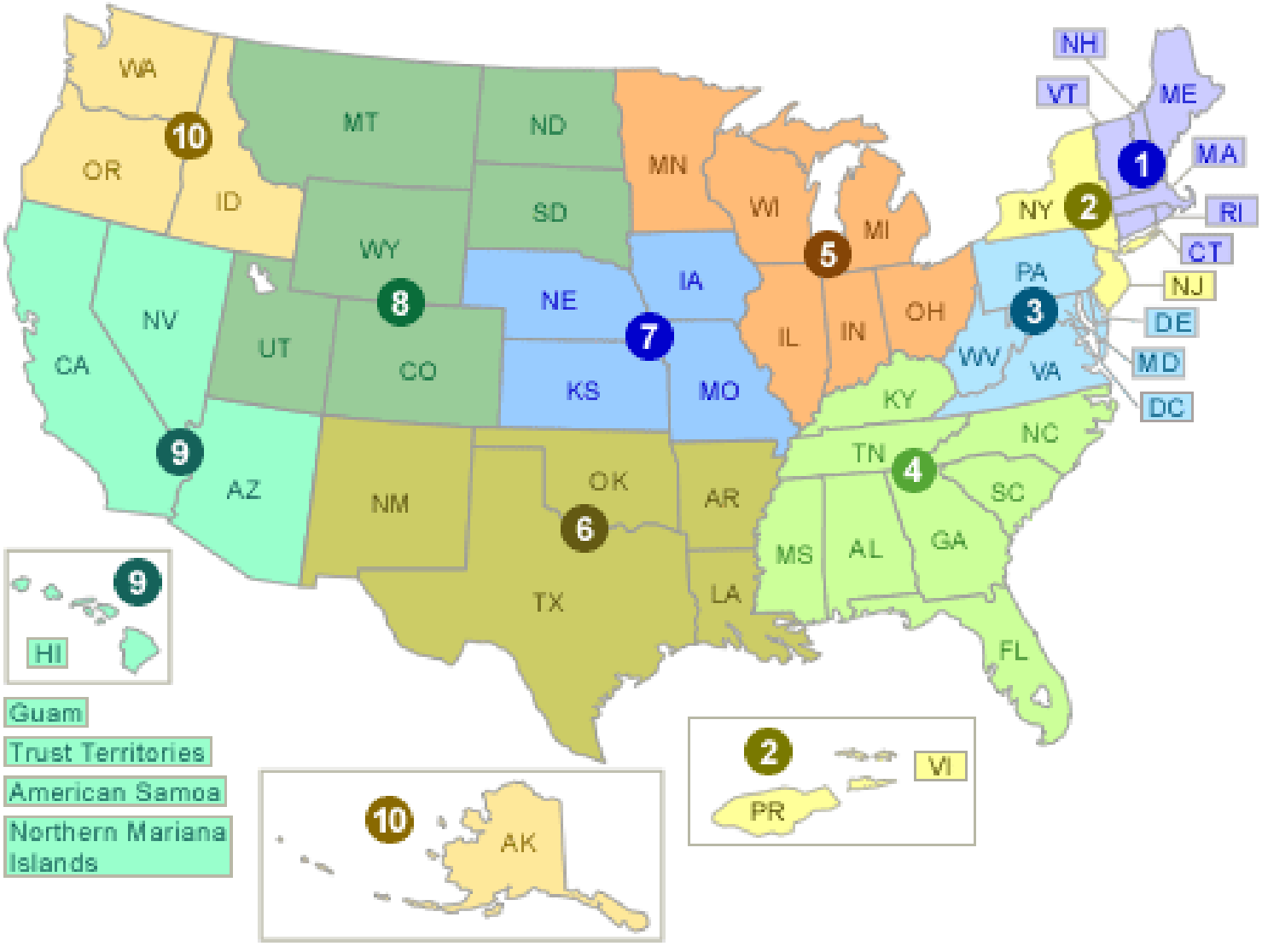
- LeastHIGH
- AnyHIGH
- MostHIGH

- LeastWIN
- AnyWIN
- MostWIN

Variables Considered

- Age
- Gender
- Residence
  - Top Metropolitan Area
  - Federally Defined Regions

Federal Regions



http://travelquaz.com/map-of-i-s.html/where-you-live-acid-rain-clean-air-markets-us-epa-2

Overview

Methods:

- Multinomial Logistic Regression
- Modified Random Forest

Logistic Regression

$$\widehat{\pi}(\boldsymbol{x_i}) = exp(\alpha + \boldsymbol{\beta^T x_i})$$

- $\widehat{\pi}(\boldsymbol{x_i})$ probability that contestant i wins a competition
- $\boldsymbol{x_i} = (x_{i,1}, x_{i,2}, ..., x_{i,P})$ contestant i's values for each of the variables
- P total number of variables being considered
- $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_P)$ vector of coefficients

MNL Regression Overview

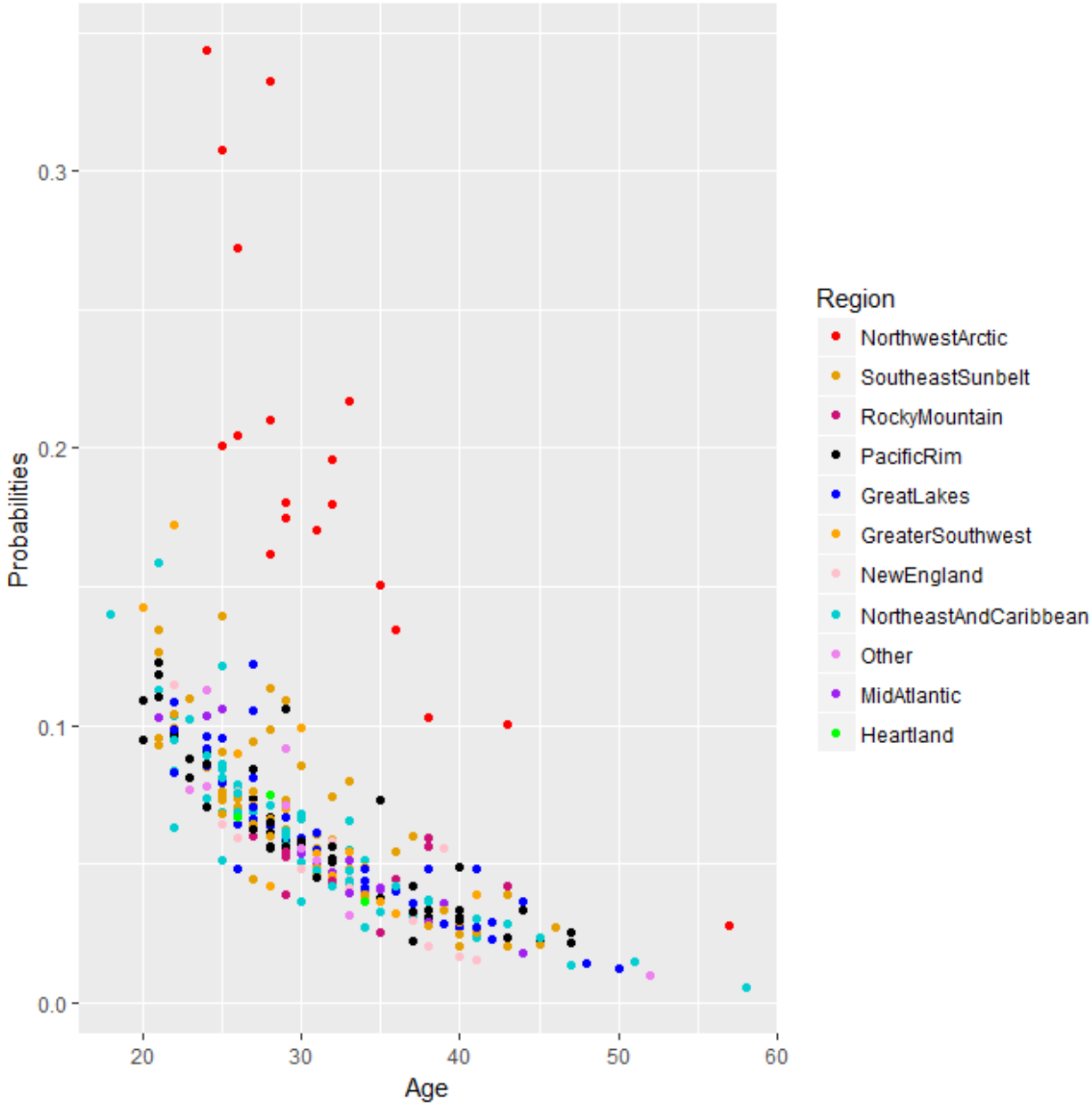
$$\widehat{\pi}_j(\boldsymbol{x_{i,j}}) = \frac{exp(\alpha + \boldsymbol{\beta^T x_{i,j}})}{\sum_{i \in N_j} exp(\alpha + \boldsymbol{\beta^T x_{i,j}})}$$

- $\widehat{\pi}_j(\boldsymbol{x_{i,j}})$ probability of contestant i winning competition j
- $\boldsymbol{x_{i,j}} = (x_{i,j,1}, x_{i,j,2}, ..., x_{i,j,P})$ values for contestant i in competition j
- N_j set of contestants in competition j

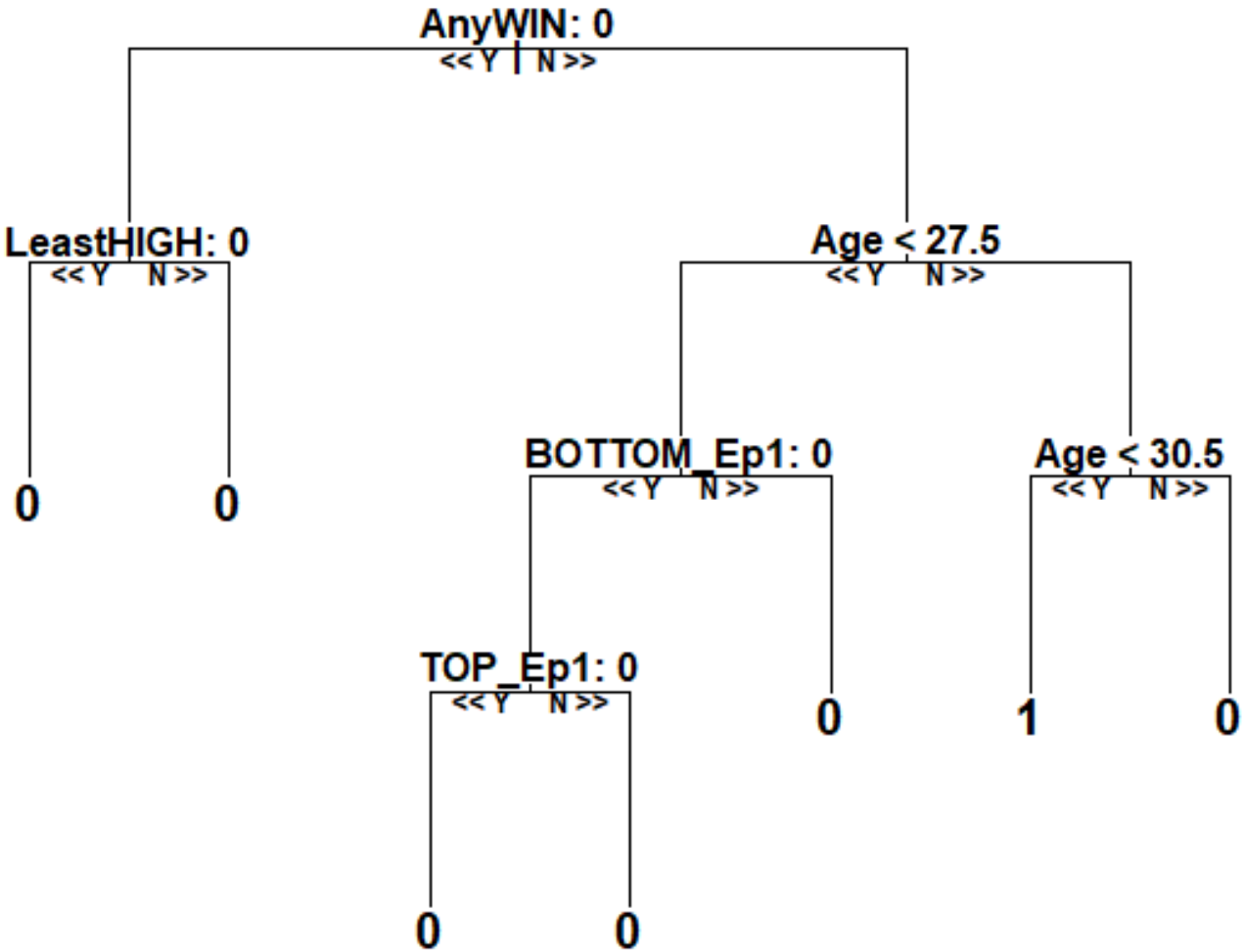
MNL Regression Predictions

- Forward Stepwise Selection using AIC
- Cross Validation
- The contestant with the highest probability is the winner for competition j
  - $\widehat{\pi}_j(\boldsymbol{x_{i,j}}) > \widehat{\pi}_j(\boldsymbol{x_{k,j}})$ for all $k \neq i$
  - $\widehat{Y}_{i,j} = 1$

MNL Probability vs Age



Simple Decision Tree Example



Random Forest Overview

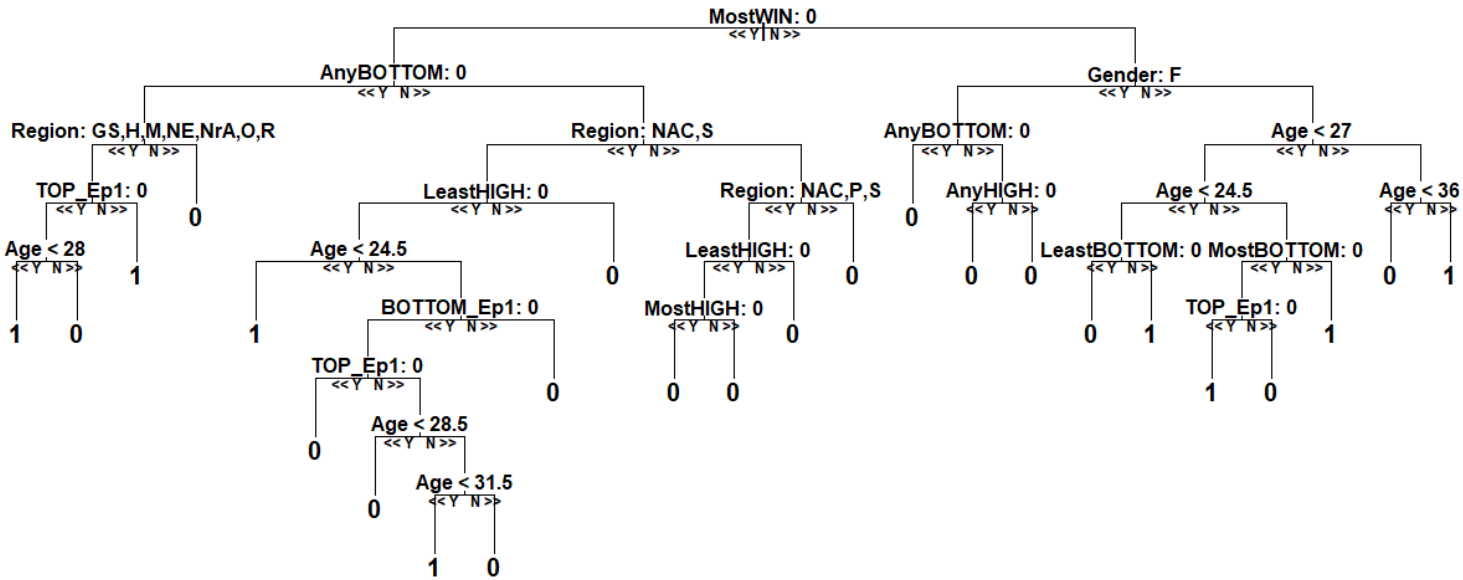
- Gini Index:
  - $G_p = m_{p,0} - m_{p,0}^2 + m_{p,1} - m_{p,1}^2$
  - $m_{p,k}$ proportion of training observations in the response category k when splitting variable p
- Modified Bootstrap Sample of Observations
- Random Subset of $\sqrt{P}$ Variables

Random Forest Predicted Values

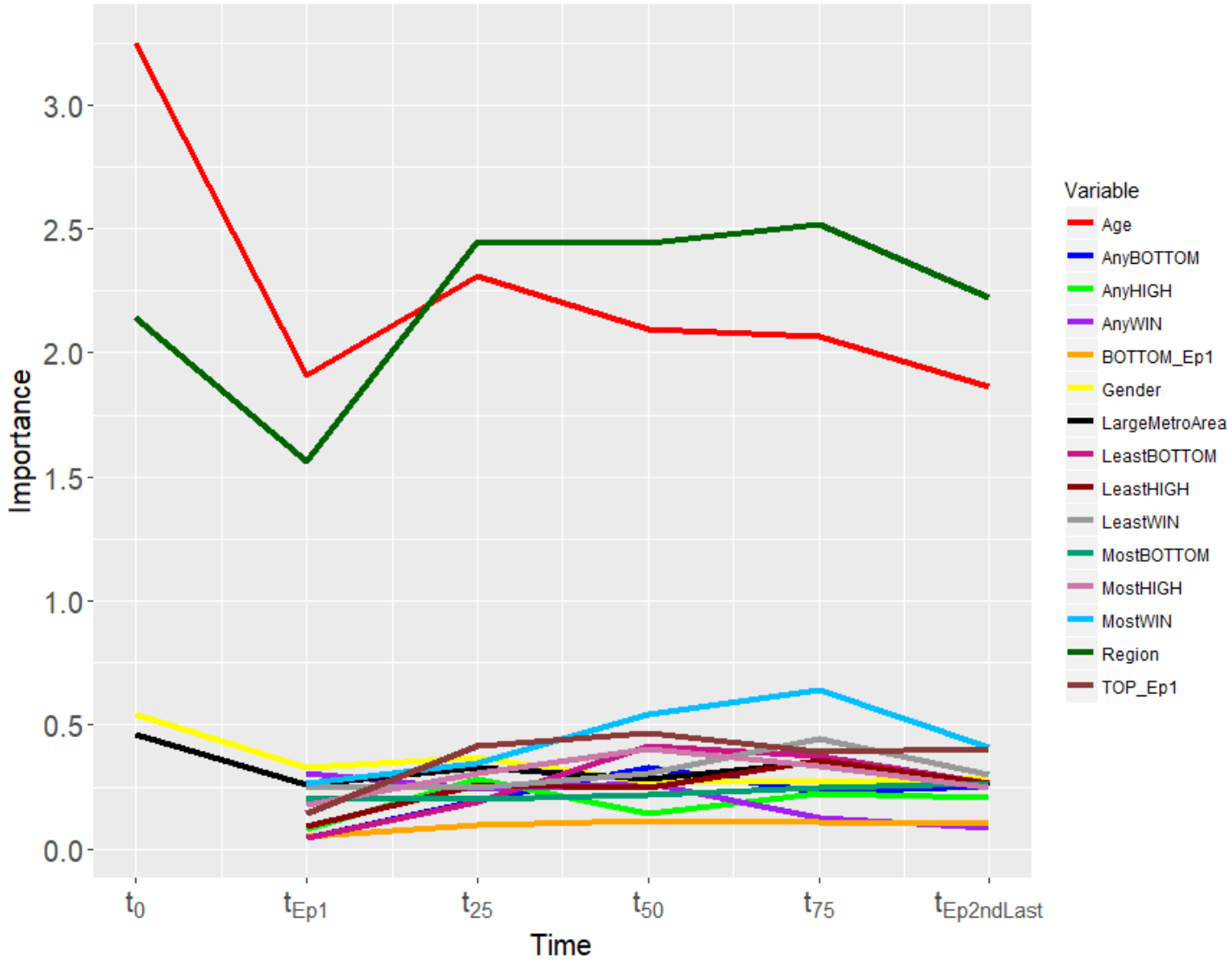
$$\widehat{\pi}_j(\mathbf{x_{i,j}}) = \frac{1}{|B_j|} \sum_{b \in B_j} \widehat{Y}_{i,j,b}$$

- $\widehat{\pi}_j(\mathbf{x_{i,j}})$ probability of contestant i winning competition j
- $\widehat{Y}_{i,j,b}$ predicted value of contestant i in competition j for tree b
- B_j set of all trees that do not include competition j in the training set
- The contestant with the highest probability is categorized as the winner for competition j
  - $\widehat{\pi}_j(\boldsymbol{x_{i,j}}) > \widehat{\pi}_j(\boldsymbol{x_{k,j}})$ for all $k \neq i$
  - $\widehat{Y}_{i,j} = 1$

Random Forest Tree Example



Random Forest Variable Importance

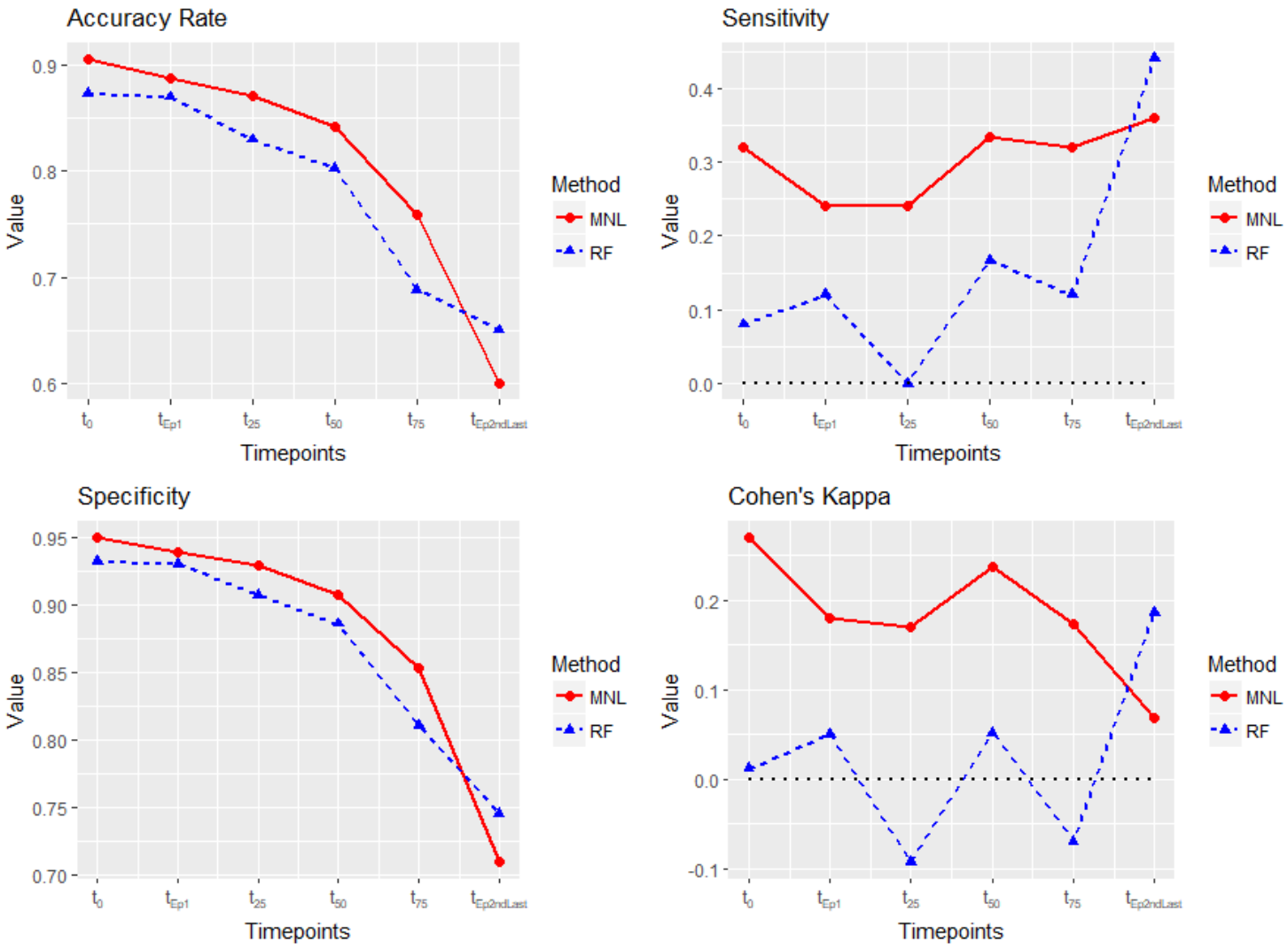


Classification Evaluation

		Observed Values	
		$Win\ (Y = 1)$	$Lose\ (Y = 0)$
Predicted Values	$Win\ (\hat{Y} = 1)$	a	b
	$Lose\ (\hat{Y} = 1)$	c	d

- **Accuracy Rate**: Percent of contestants correctly identified
- **Sensitivity**: Percent of contestants who won correctly identified
- **Specificity**: Percent of contestants who lost correctly identified
- **Cohen's Kappa**: Measure of model's performance compared to random guessing

Results


Accuracy Rate
Value
0.9
0.8
0.7
0.6
t_0
t_{Ep1}
t_{25}
t_{50}
t_{75}
$t_{Ep2ndLast}$
Timepoints
Method
MNL
RF
Sensitivity
Value
0.4
0.3
0.2
0.1
0.0
Timepoints
Method
MNL
RF
Specificity
Value
0.95
0.90
0.85
0.80
0.75
0.70
Timepoints
Method
MNL
RF
Cohen's Kappa
Value
0.2
0.1
0.0
-0.1
Timepoints
Method
MNL
RF

Summary

- Is it possible to predict the winner of a reality-competition program?
- How early in the competition can a winner be determined?
- What are the primary factors used to determine the winner?