Classification Tree Analysis:

A Useful Statistical Tool for Program Evaluators

Meredith L. Philyaw, MS
Jennifer Lyons, MSW(c)

Why This Session?

Stand up if you...

Consider yourself to be a data analyst, frequently work with quantitative data in your job or are really just interested in statistics.

Work with quantitative data some...not as much as a data analyst per say….and you would like to learn a new method.

Hate statistics with a passion but you're in this session because working with quantitative data is a necessary evil in program evaluation. (It's okay...we've all felt this way at some point)

Other reasons?

Session Outline

Overview of Classification Tree Analysis (CTA)

Walk-through of performing a CTA

Group Activity: Presenting the results of a CTA to your client

Wrap-up/resources for continued learning

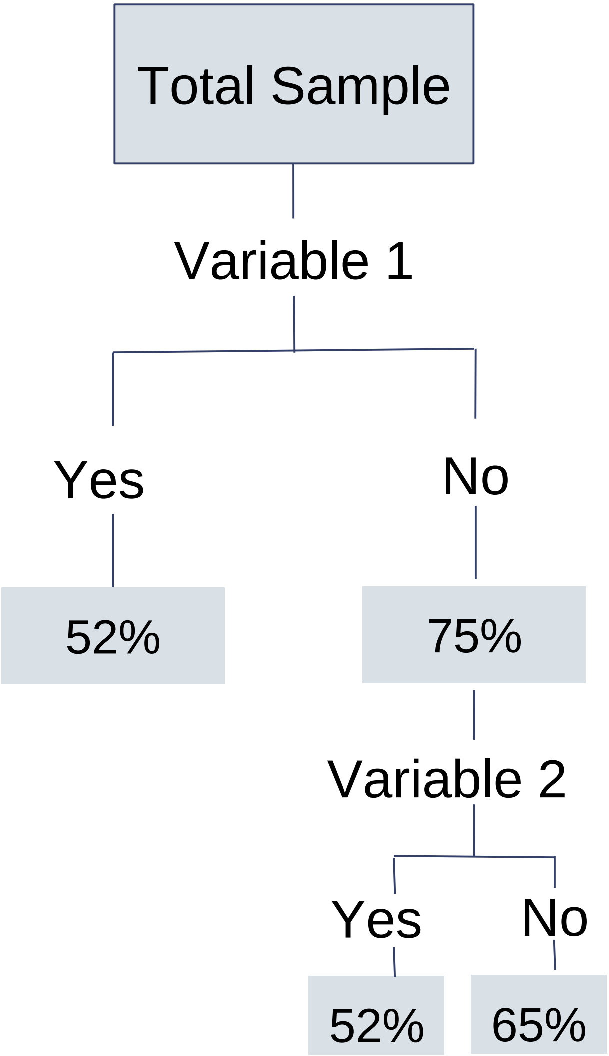
What is Classification Tree Analysis?

Identifies a set of characteristics that best differentiates individuals based on a categorical outcome variable

Generates a multi-level tree diagram

The order in which variables appear in the tree matters!

Creates **exhaustive** and **mutually exclusive** subgroups of individuals



Data Considerations

Do you have an outcome variable that can be measured categorically?

Is there variation in the outcome variable among your sample?

Do you have variables that are theoretically related to your outcome variable?

What is your sample size?

Is it possible to measure your variables so the right-hand side variables precede the outcome variable?

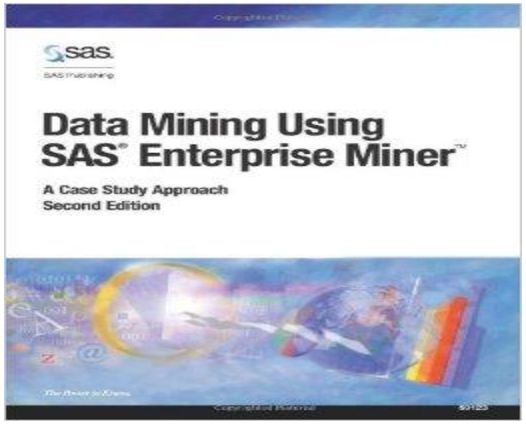
What Types of Evaluation Questions Can CTA Answer?

What factors best differentiate treatment attenders from non-attenders?

What characteristics predict health improvement from baseline to follow-up?

Others?

What software can I use?

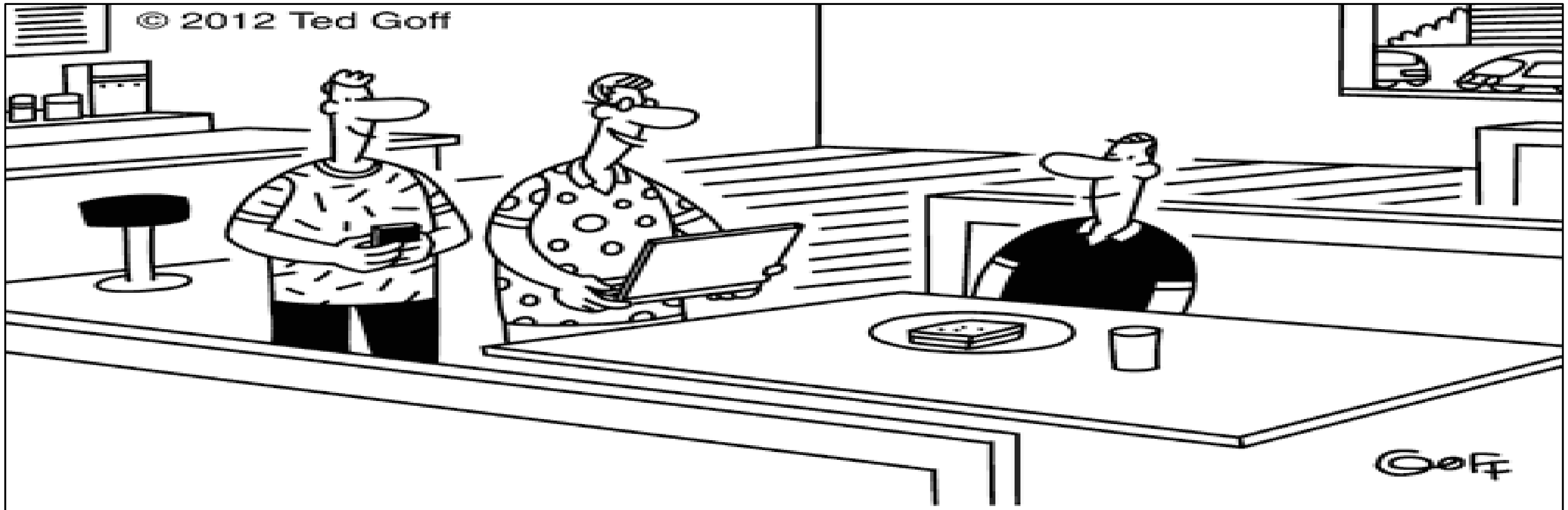

sas. SAS Publishing

Data Mining Using SAS® Enterprise Miner™

A Case Study Approach
Second Edition

The Power to Know.

Validation and CTA



© 2012 Ted Goff

“Twitter and Facebook can’t predict the election, but they did predict what you’re going to have for lunch: a tuna salad sandwich. You’re having the wrong sandwich.”

Validation Approaches

1. Hold-out sample

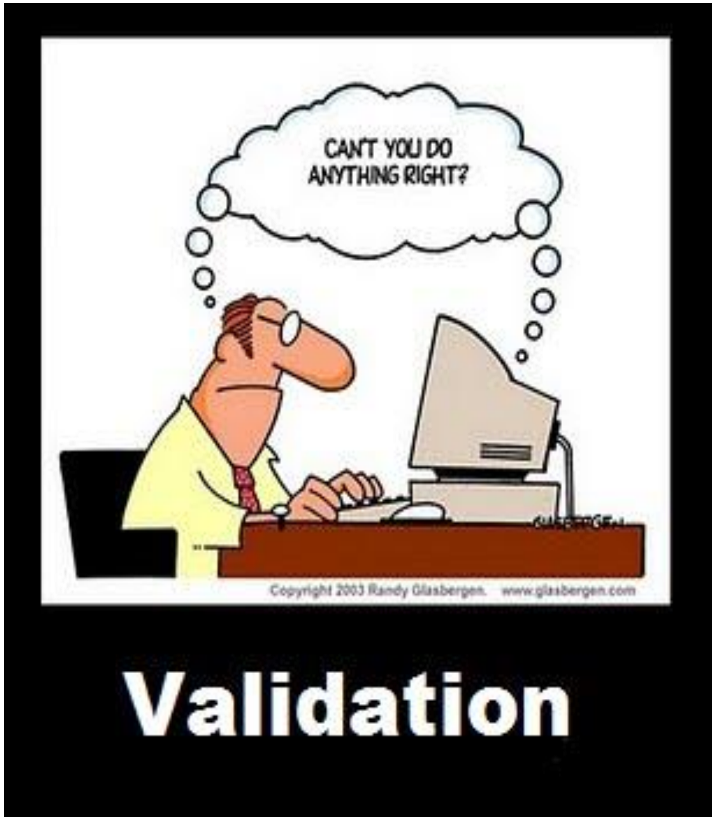
    80% training sample

    20% testing sample

2. You can also add in a validation sample

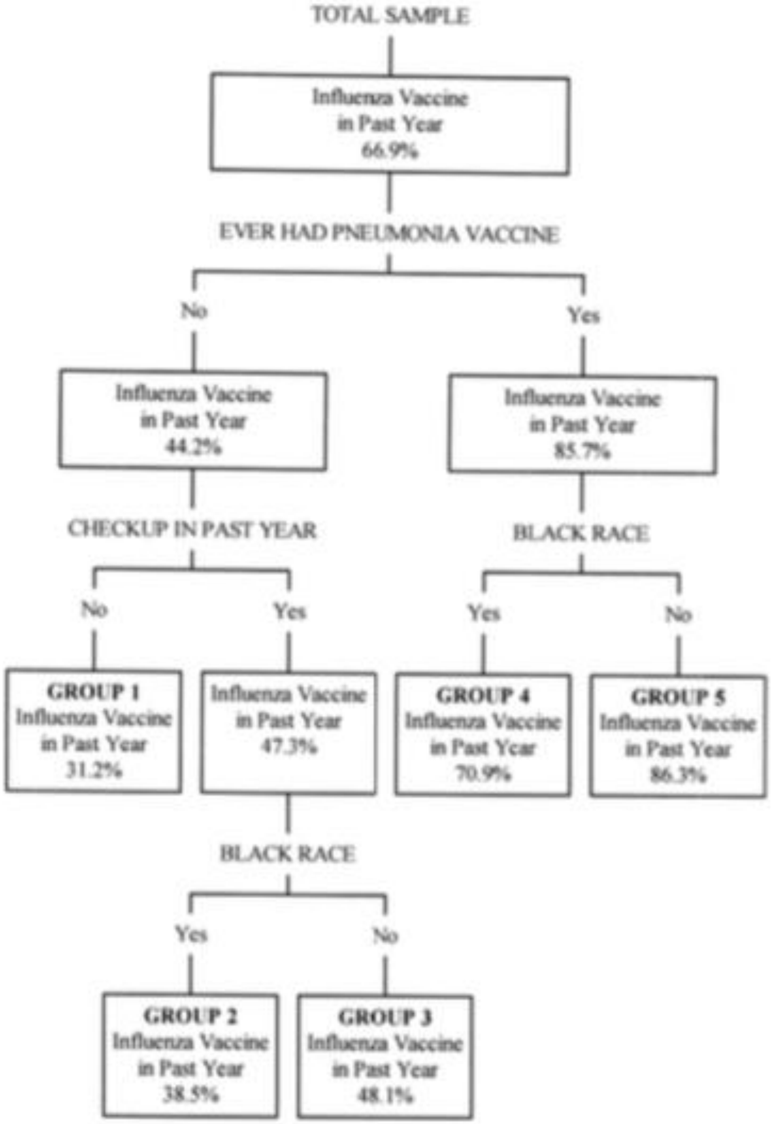
3. K-fold cross validation

K=5 or k=10 is typically used



Copyright 2003 Randy Glasbergen. www.glasbergen.com

Interpreting the Output of CTA



TOTAL SAMPLE

Influenza Vaccine in Past Year 66.9%

EVER HAD PNEUMONIA VACCINE

No

Influenza Vaccine in Past Year 44.2%

CHECKUP IN PAST YEAR

No

GROUP 1 Influenza Vaccine in Past Year 31.2%

Yes

Influenza Vaccine in Past Year 47.3%

BLACK RACE

Yes

GROUP 2 Influenza Vaccine in Past Year 38.5%

No

GROUP 3 Influenza Vaccine in Past Year 48.1%

Yes

Influenza Vaccine in Past Year 85.7%

BLACK RACE

Yes

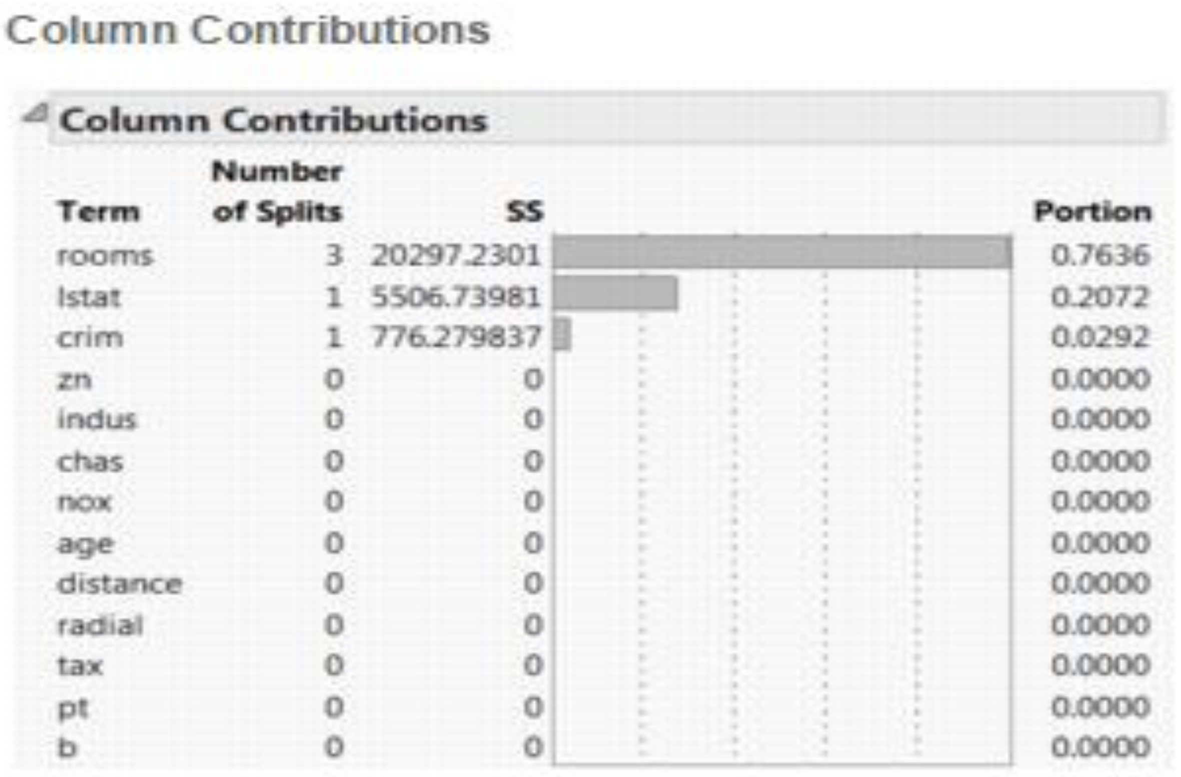
GROUP 4 Influenza Vaccine in Past Year 70.9%

No

GROUP 5 Influenza Vaccine in Past Year 86.3%

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine*, *26*(3), 172-181.

Column Contributions



Column Contributions

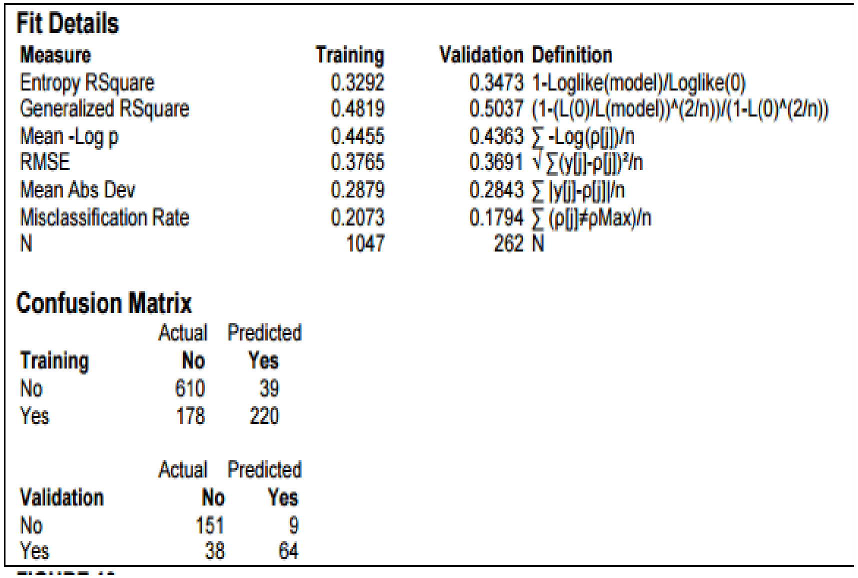
Column Contributions

Term	Number of Splits	SS	Portion
rooms	3	20297.2301	0.7636
lstat	1	5506.73981	0.2072
crim	1	776.279837	0.0292
zn	0	0	0.0000
indus	0	0	0.0000
chas	0	0	0.0000
nox	0	0	0.0000
age	0	0	0.0000
distance	0	0	0.0000
radial	0	0	0.0000
tax	0	0	0.0000
pt	0	0	0.0000
b	0	0	0.0000

Decision Tree

http://www.jmp.com/support/help/Examples_of_Partitioning_Methods.shtml

Evaluating Tree Performance



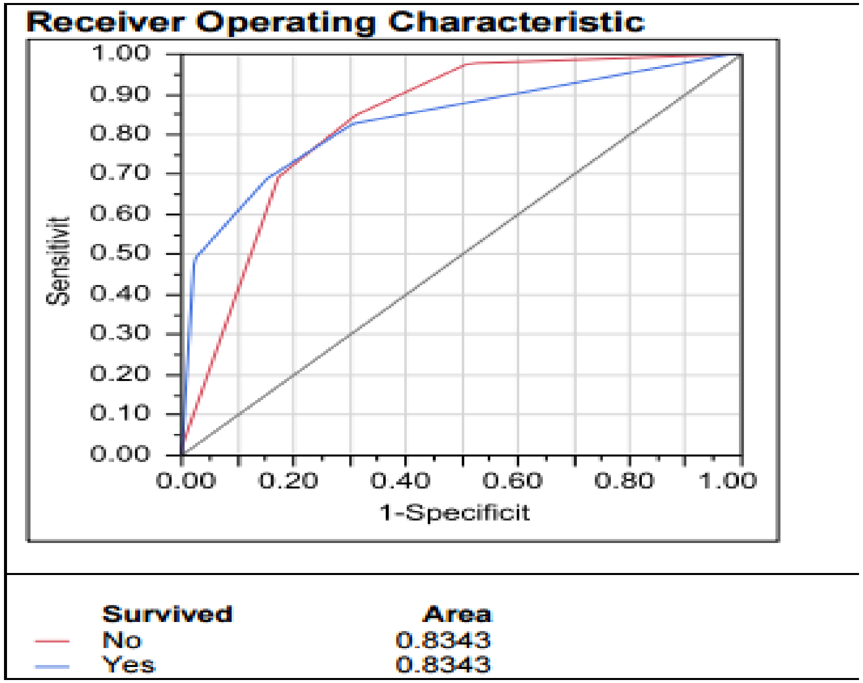
Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.3292	0.3473	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4819	0.5037	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4455	0.4363	∑ -Log(ρ[j])/n
RMSE	0.3765	0.3691	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2879	0.2843	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2073	0.1794	∑ (ρ[j]≠ρMax)/n
N	1047	262	N

Confusion Matrix

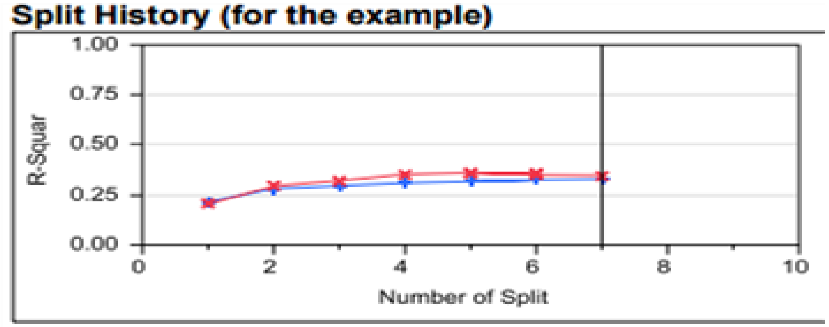
	Actual	Predicted
Training	**No**	**Yes**
No	610	39
Yes	178	220

	Actual	Predicted
Validation	**No**	**Yes**
No	151	9
Yes	38	64



Receiver Operating Characteristic

	Survived	Area
🟥	No	0.8343
🟦	Yes	0.8343



Split History (for the example)

Validation Data in Red

CTA Using JMP

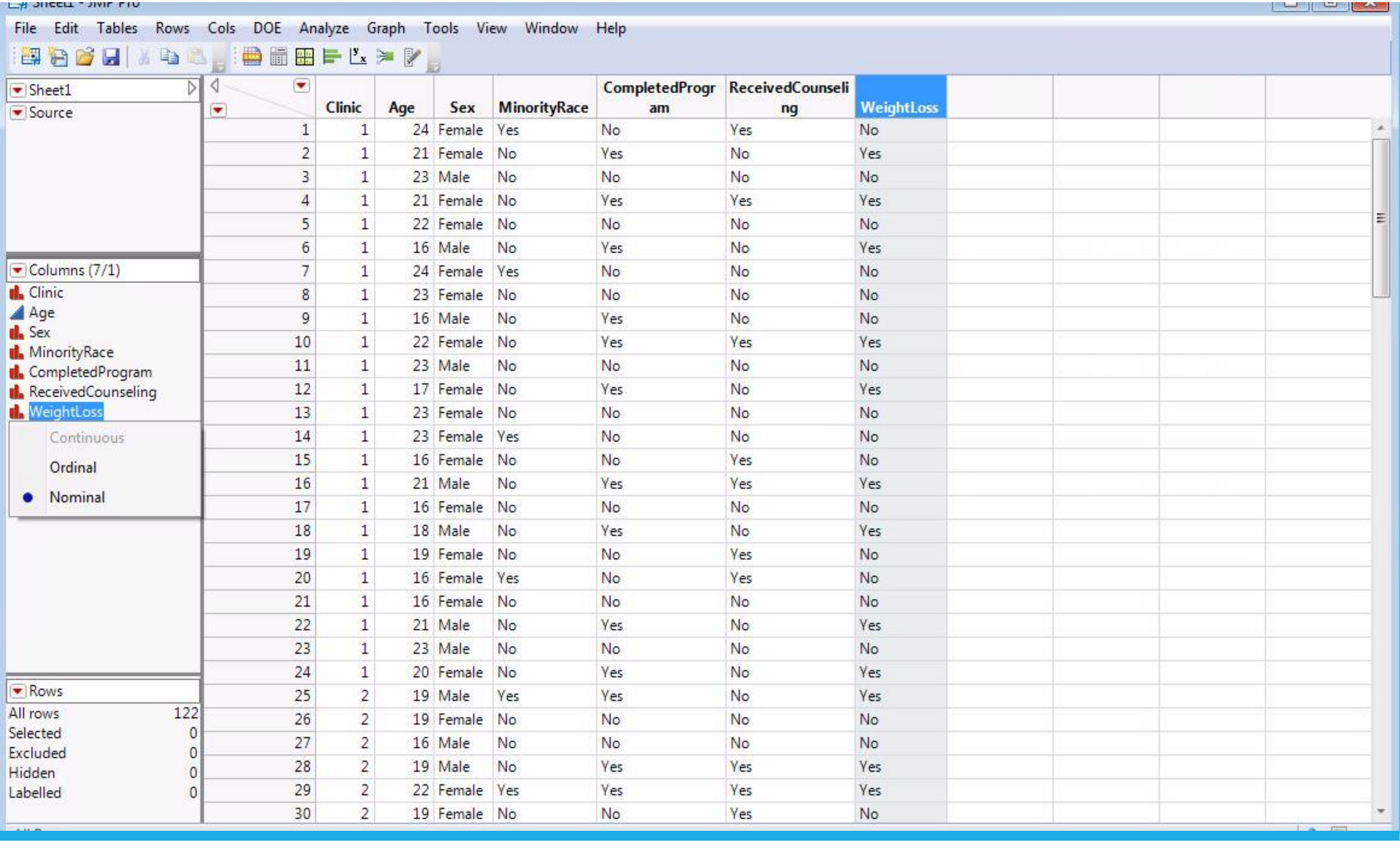
Case Scenario

You are the evaluator for a multi-site clinical intervention designed to promote weight loss among patients with diabetes

The intervention's funder wants to know:

What factors predict weight loss at 3-month follow-up?

Variables of Interest



Columns (7/1)

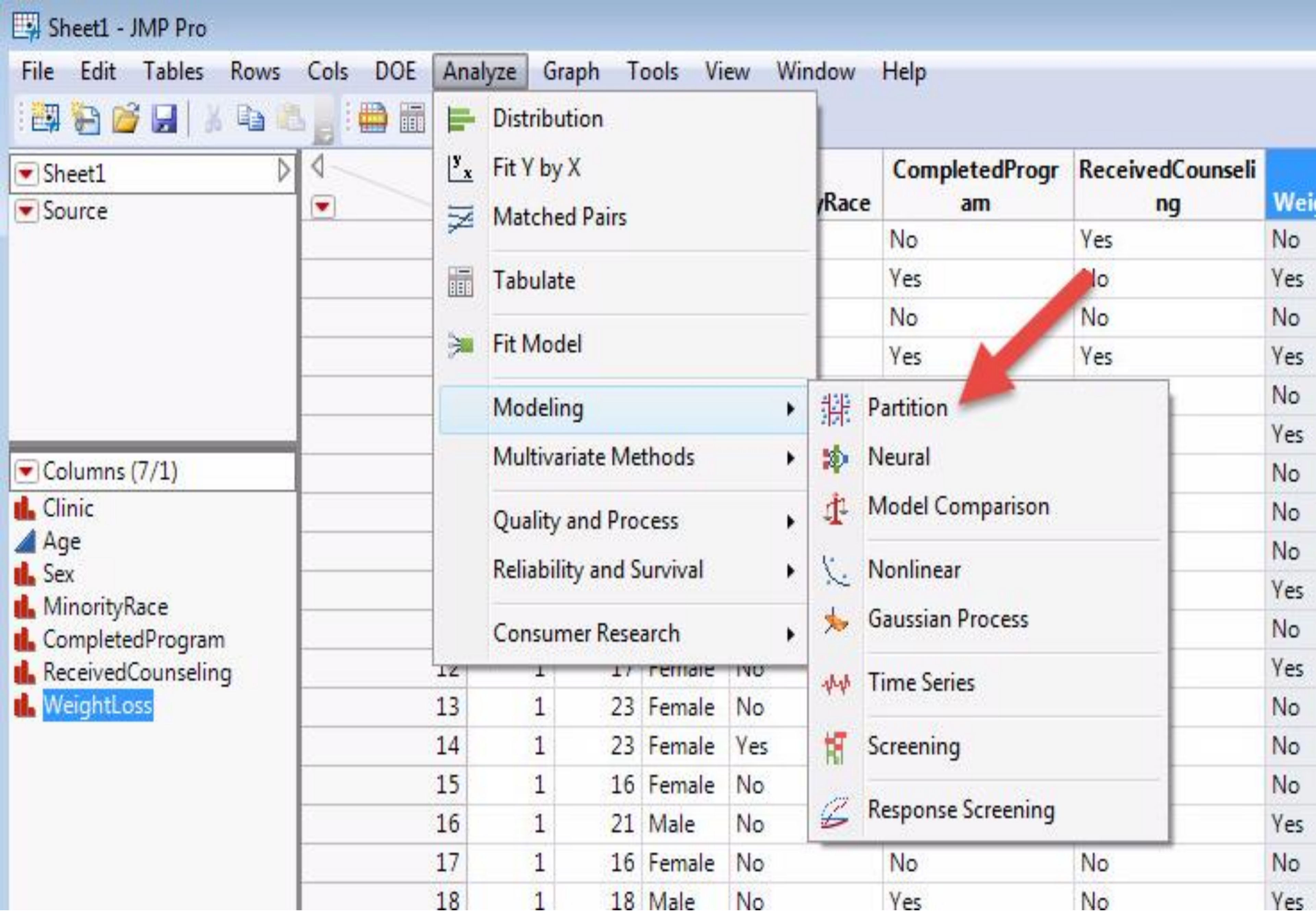
- Clinic
- Age
- Sex
- MinorityRace
- CompletedProgram
- ReceivedCounseling
- WeightLoss

Continuous
Ordinal
● Nominal

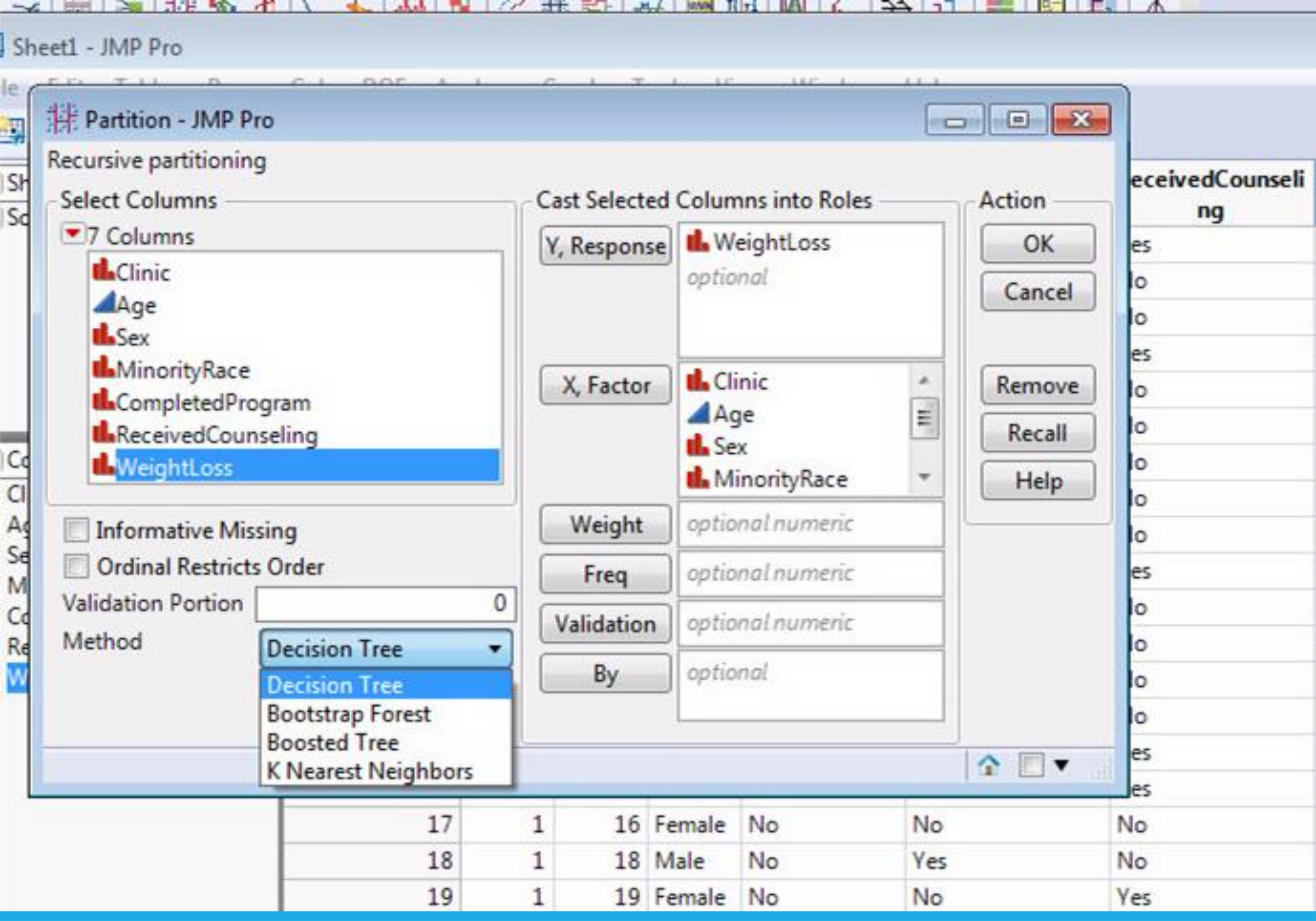
Rows

All rows	122
Selected	0
Excluded	0
Hidden	0
Labelled	0

	Clinic	Age	Sex	MinorityRace	CompletedProgram	ReceivedCounseling	WeightLoss
1	1	24	Female	Yes	No	Yes	No
2	1	21	Female	No	Yes	No	Yes
3	1	23	Male	No	No	No	No
4	1	21	Female	No	Yes	Yes	Yes
5	1	22	Female	No	No	No	No
6	1	16	Male	No	Yes	No	Yes
7	1	24	Female	Yes	No	No	No
8	1	23	Female	No	No	No	No
9	1	16	Male	No	Yes	No	No
10	1	22	Female	No	Yes	Yes	Yes
11	1	23	Male	No	No	No	No
12	1	17	Female	No	Yes	No	Yes
13	1	23	Female	No	No	No	No
14	1	23	Female	Yes	No	No	No
15	1	16	Female	No	No	Yes	No
16	1	21	Male	No	Yes	Yes	Yes
17	1	16	Female	No	No	No	No
18	1	18	Male	No	Yes	No	Yes
19	1	19	Female	No	No	Yes	No
20	1	16	Female	Yes	No	Yes	No
21	1	16	Female	No	No	No	No
22	1	21	Male	No	Yes	No	Yes
23	1	23	Male	No	No	No	No
24	1	20	Female	No	Yes	No	Yes
25	2	19	Male	Yes	Yes	No	Yes
26	2	19	Female	No	No	No	No
27	2	16	Male	No	No	No	No
28	2	19	Male	No	Yes	Yes	Yes
29	2	22	Female	Yes	Yes	Yes	Yes
30	2	19	Female	No	No	Yes	No



				Race	CompletedProgram	ReceivedCounseling	Wei
					No	Yes	No
					Yes	No	Yes
					No	No	No
					Yes	Yes	Yes
							No
							Yes
							No
							No
							No
							Yes
							No
12	1	17	Female	No			Yes
13	1	23	Female	No			No
14	1	23	Female	Yes			No
15	1	16	Female	No			No
16	1	21	Male	No			Yes
17	1	16	Female	No	No	No	No
18	1	18	Male	No	Yes	No	Yes



Sheet1 - JMP Pro

Partition - JMP Pro

Recursive partitioning

Select Columns

7 Columns

- Clinic
- Age
- Sex
- MinorityRace
- CompletedProgram
- ReceivedCounseling
- WeightLoss

Cast Selected Columns into Roles

Y, Response: WeightLoss

X, Factor: Clinic, Age, Sex, MinorityRace

Weight: optional numeric

Freq: optional numeric

Validation: optional numeric

By: optional

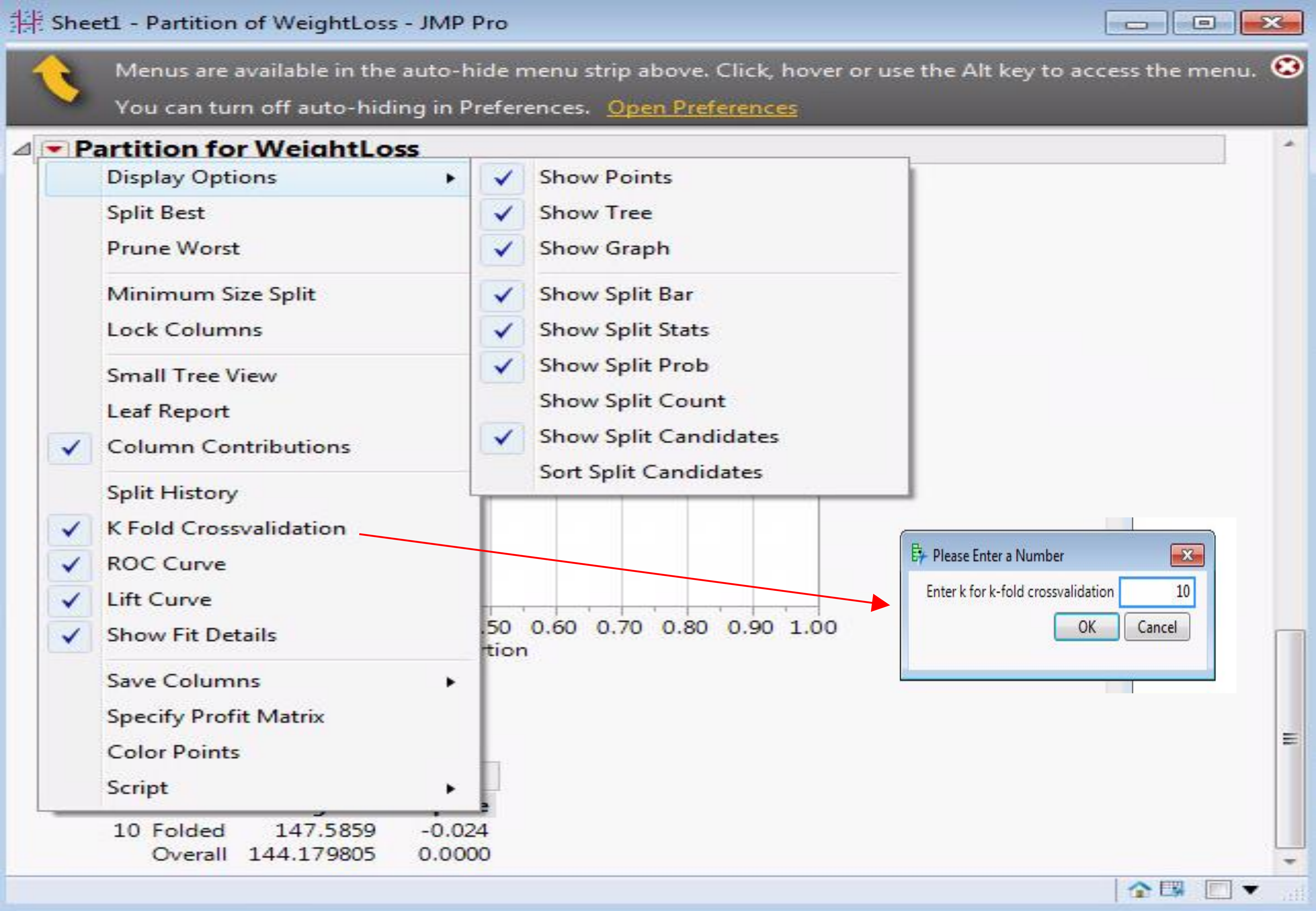
□ Informative Missing

□ Ordinal Restricts Order

Validation Portion: 0

Method: Decision Tree

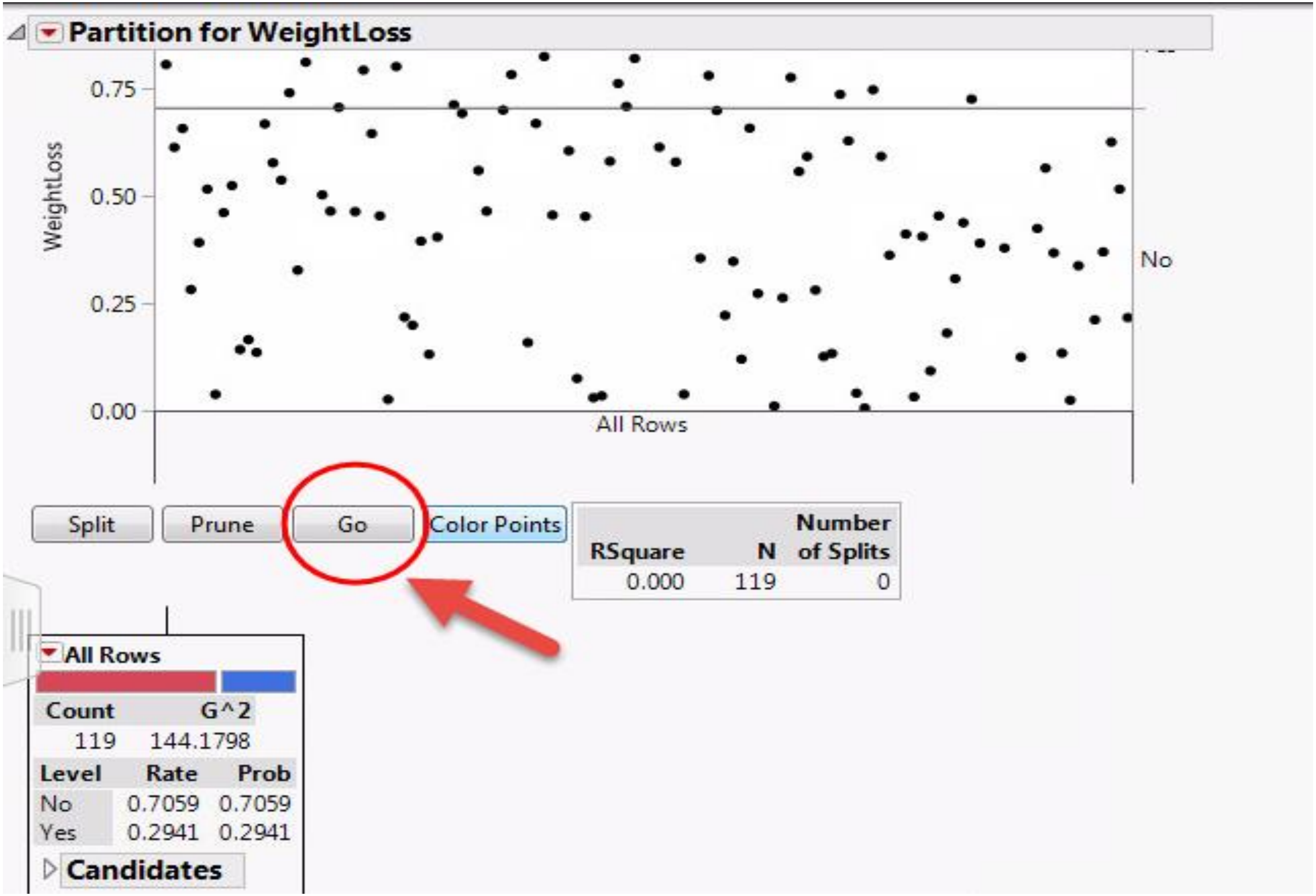
17	1	16	Female	No	No	No
18	1	18	Male	No	Yes	No
19	1	19	Female	No	No	Yes



Menus are available in the auto-hide menu strip above. Click, hover or use the Alt key to access the menu.
You can turn off auto-hiding in Preferences. Open Preferences

Enter k for k-fold crossvalidation: 10

```
10  Folded    147.5859    -0.024
    Overall   144.179805   0.0000
```



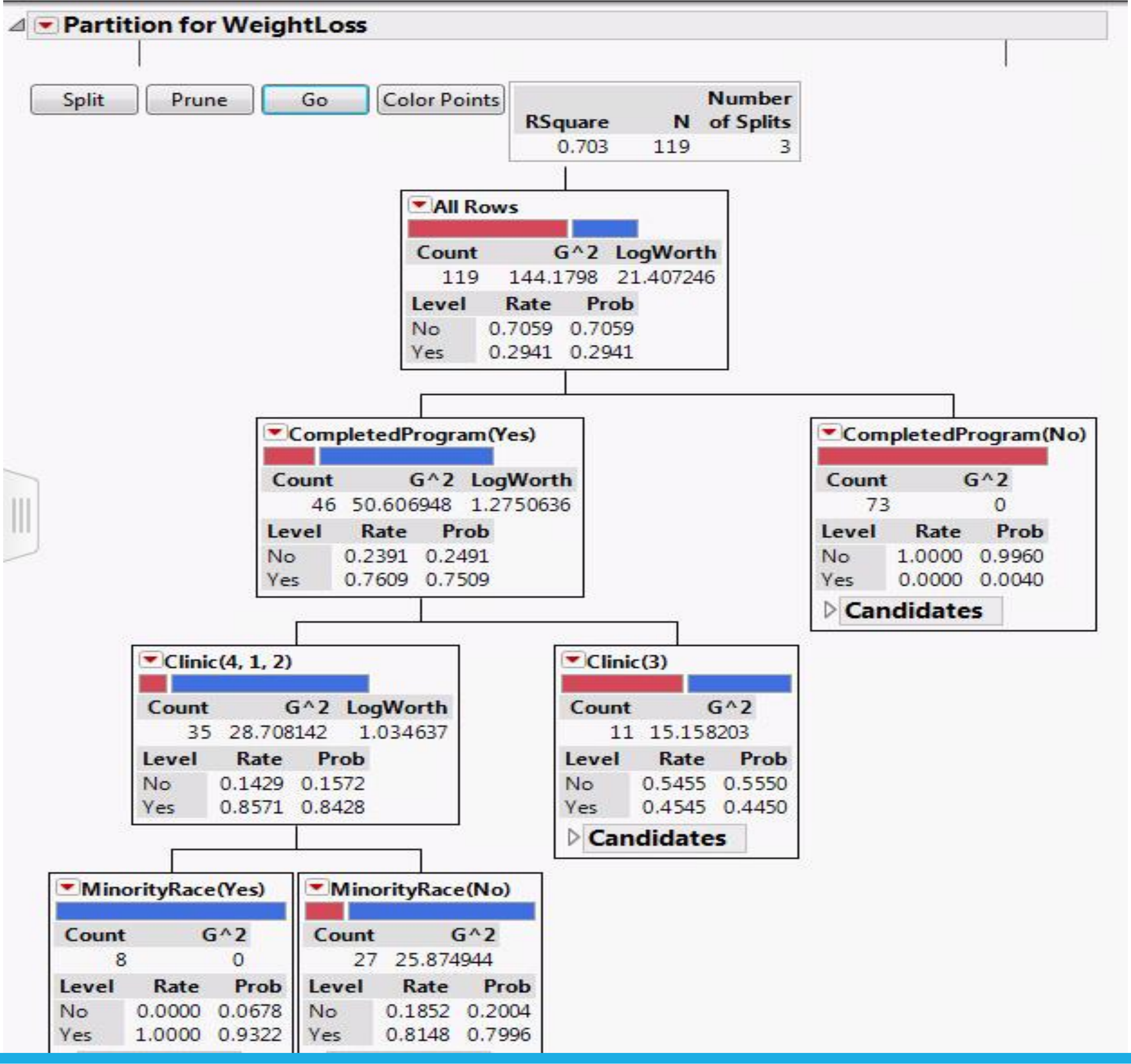
RSquare	N	Number of Splits
0.000	119	0

All Rows

Count	G^2
119	144.1798

Level	Rate	Prob
No	0.7059	0.7059
Yes	0.2941	0.2941

Candidates



Partition for WeightLoss

RSquare	N	Number of Splits
0.703	119	3

All Rows

Count	G^2	LogWorth
119	144.1798	21.407246

Level	Rate	Prob
No	0.7059	0.7059
Yes	0.2941	0.2941

CompletedProgram(Yes)

Count	G^2	LogWorth
46	50.606948	1.2750636

Level	Rate	Prob
No	0.2391	0.2491
Yes	0.7609	0.7509

CompletedProgram(No)

Count	G^2
73	0

Level	Rate	Prob
No	1.0000	0.9960
Yes	0.0000	0.0040

Candidates

Clinic(4, 1, 2)

Count	G^2	LogWorth
35	28.708142	1.034637

Level	Rate	Prob
No	0.1429	0.1572
Yes	0.8571	0.8428

Clinic(3)

Count	G^2
11	15.158203

Level	Rate	Prob
No	0.5455	0.5550
Yes	0.4545	0.4450

Candidates

MinorityRace(Yes)

Count	G^2
8	0

Level	Rate	Prob
No	0.0000	0.0678
Yes	1.0000	0.9322

MinorityRace(No)

Count	G^2
27	25.874944

Level	Rate	Prob
No	0.1852	0.2004
Yes	0.8148	0.7996

Column Contributions

Term	Number of Splits	G^2	Portion
CompletedProgram	1	93.5728567	0.9072
Clinic	1	6.74060264	0.0653
MinorityRace	1	2.83319859	0.0275
Sex	0	0	0.0000
Age	0	0	0.0000
ReceivedCounseling	0	0	0.0000

Crossvalidation

k-fold		-2LogLike	RSquare
10	Folded	44.2701686	0.6930
	Overall	41.0331469	0.7033

Fit Details

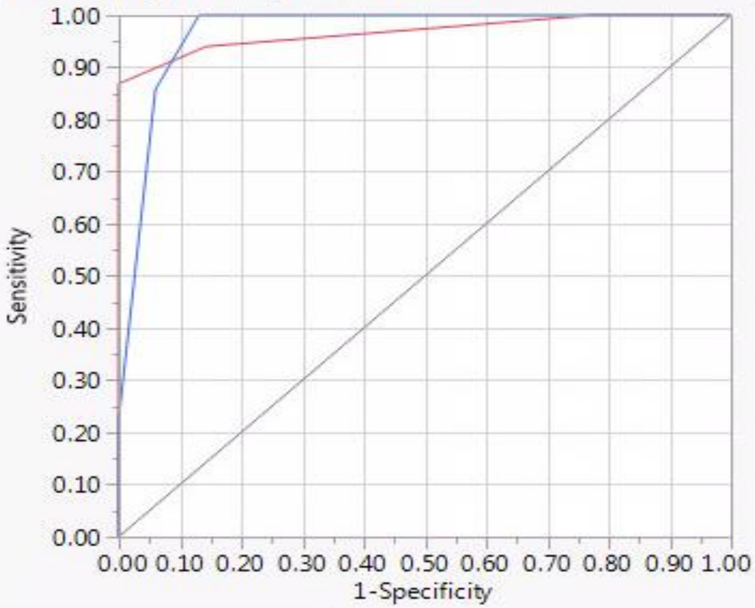
Measure	Training	Definition
Entropy RSquare	0.7033	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.8166	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1798	∑ -Log(ρ[j])/n
RMSE	0.2399	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.1234	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0840	∑ (ρ[j]≠ρMax)/n
N	119	n

Confusion Matrix

Training

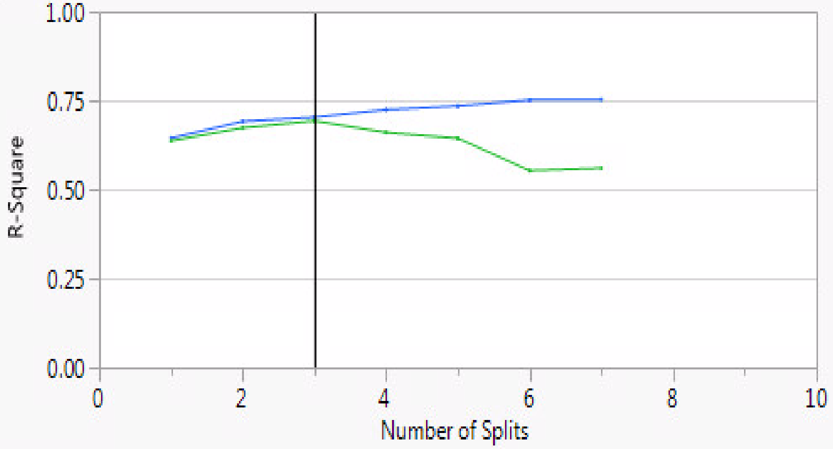
Actual	Predicted	
WeightLoss	No	Yes
No	79	5
Yes	5	30

Receiver Operating Characteristic



WeightLoss	Area
No	0.9677
Yes	0.9677

Split History



K-Fold in Green

Next Steps

Experiment with different approaches for modeling the data.

Select the model that works best.

Decide on how to present the results, depending on your venue and audience.

Limitations to Mention

If you can’t draw causal relationships from the data, be sure to mention this!

Other variables not included in the model may also impact your outcome variable

Group Exercise

In groups of 3-4, come up with a plan for explaining the results of the CTA on your handout to a client with limited statistical knowledge. Be sure to think about:

How you would explain the method
How you would present the results
What conclusions you would draw
What limitations you would mention

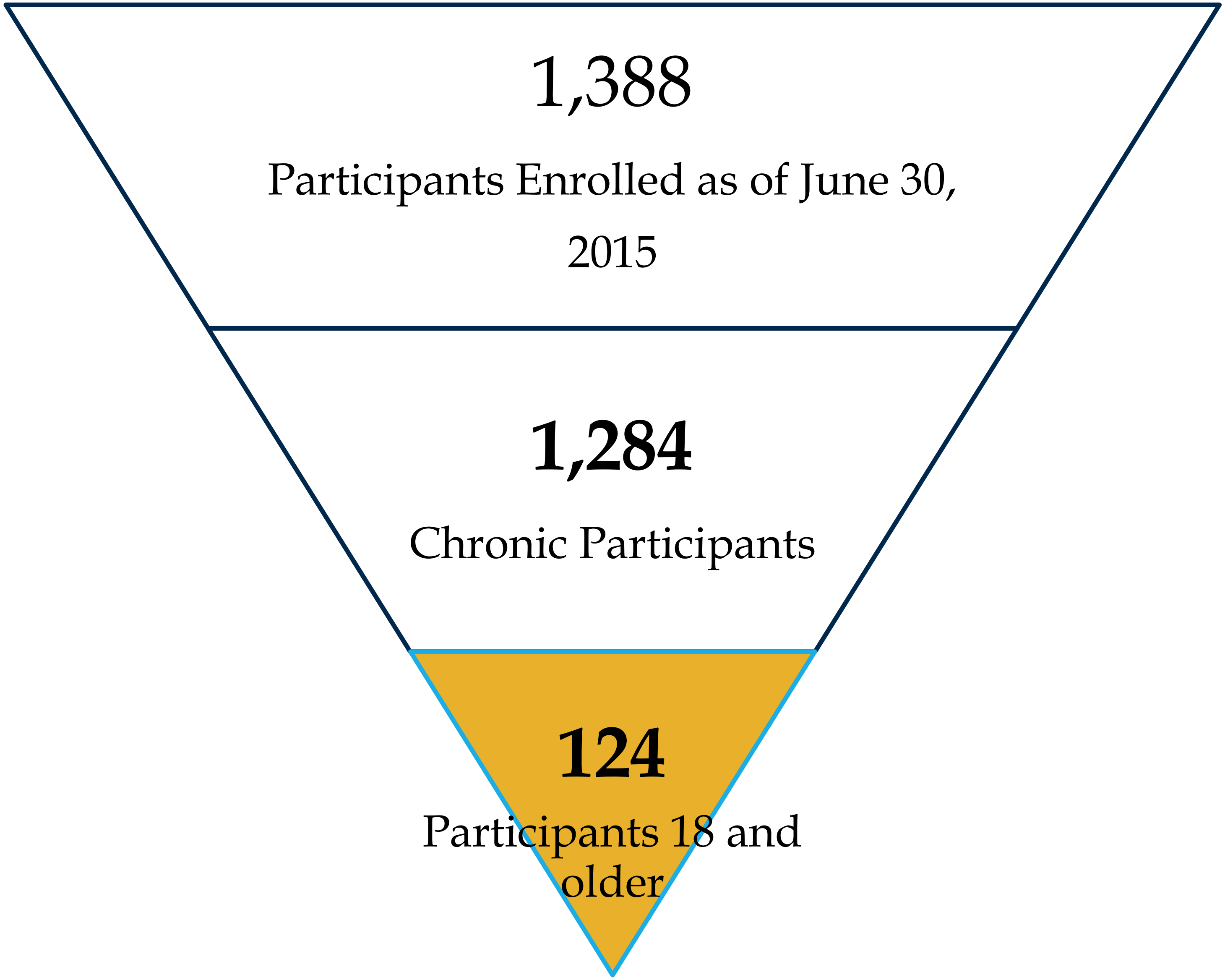
Report Back

Study Aim

For clients in a permanent supportive housing program, what characteristics at intake assessment predict housing retention after 1 year?

Methods

Sample Inclusion Criteria



1,388

Participants Enrolled as of June 30, 2015

1,284

Chronic Participants

124

Participants 18 and older

Measures

Measure	Description of Measure	Variable Values
Outcome Variable		
Housing Retention	This measure captures whether or not an individual retained housing after one year of being housed in permanent supportive housing.	Yes, No
Predictors		
Gender	Binary measures were created for each indicated gender (Woman, Man, Transgender)	Yes, No
Race	Binary measures were created for each indicated race (White, Black, Asian, AKNA/AI, NHPI, Other, Multiracial).	Yes, No
Age	Participants were grouped into age categories	Yes, No
Mental Health Diagnosis	This measure captures whether or not a person has a diagnosed mental health disorder.	Yes, No
Substance Abuse Disorder	This measure captures whether or not a person has a diagnosed with a substance abuse disorder.	Yes, No
Veteran Status	This measure captures whether or not a person is a veteran, determined by a presence of DD-214 documentation.	Yes, No

Analytic Strategy

- Examined frequencies of key variables.

- Conducted a classification tree analysis using JMP.

  - A classification tree analysis is a data mining technique that identifies what combination of factors (e.g. demographics, behavioral health comorbidity) best differentiates between individuals based on a categorical variable of interest, such as treatment attendance.

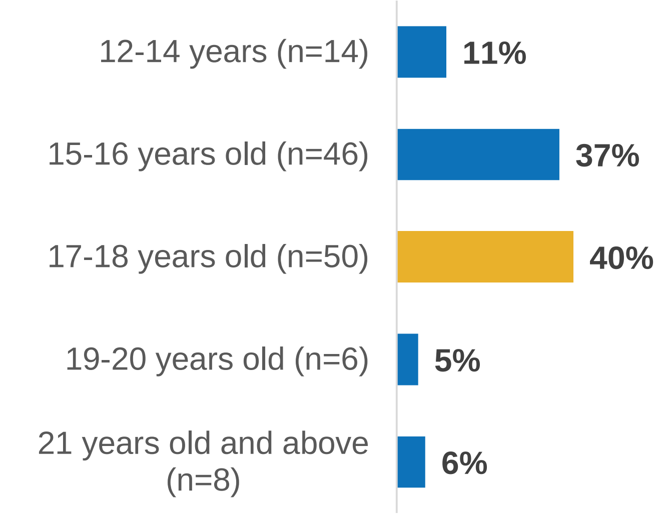
  - 10-fold cross-validation was used to improve the predictive power of the tree.

- Statistics (e.g. R^2, misclassification rate) were examined to evaluate the performance of the final classification tree.

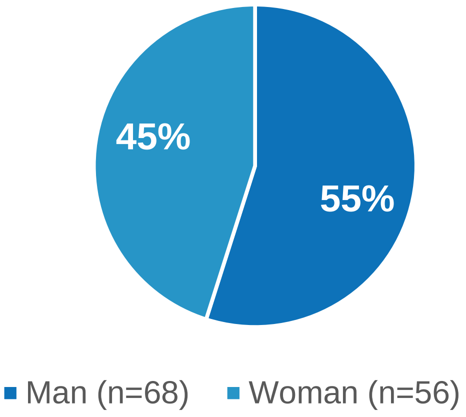
Results

Sample Characteristics

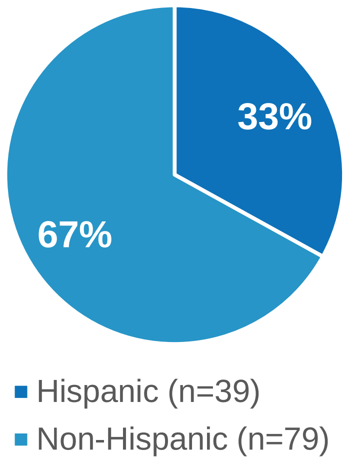
Age



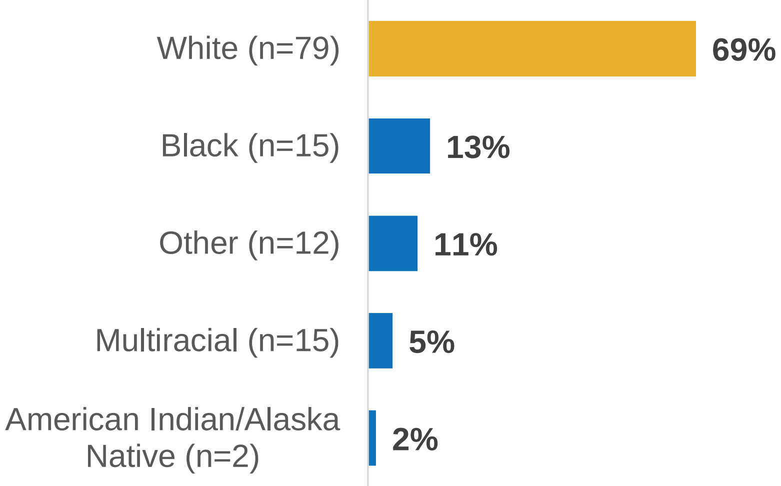
Gender



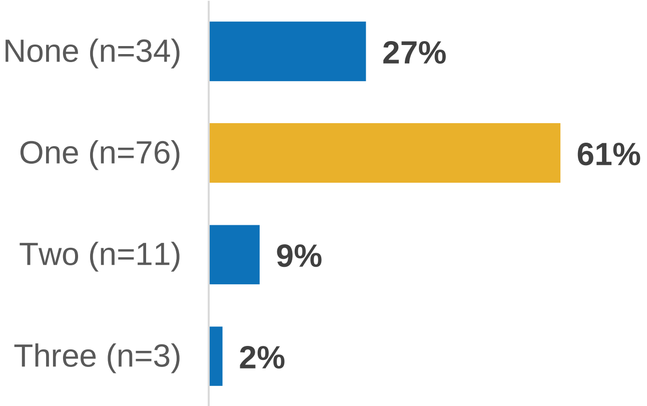
Ethnicity



Race (n=114)

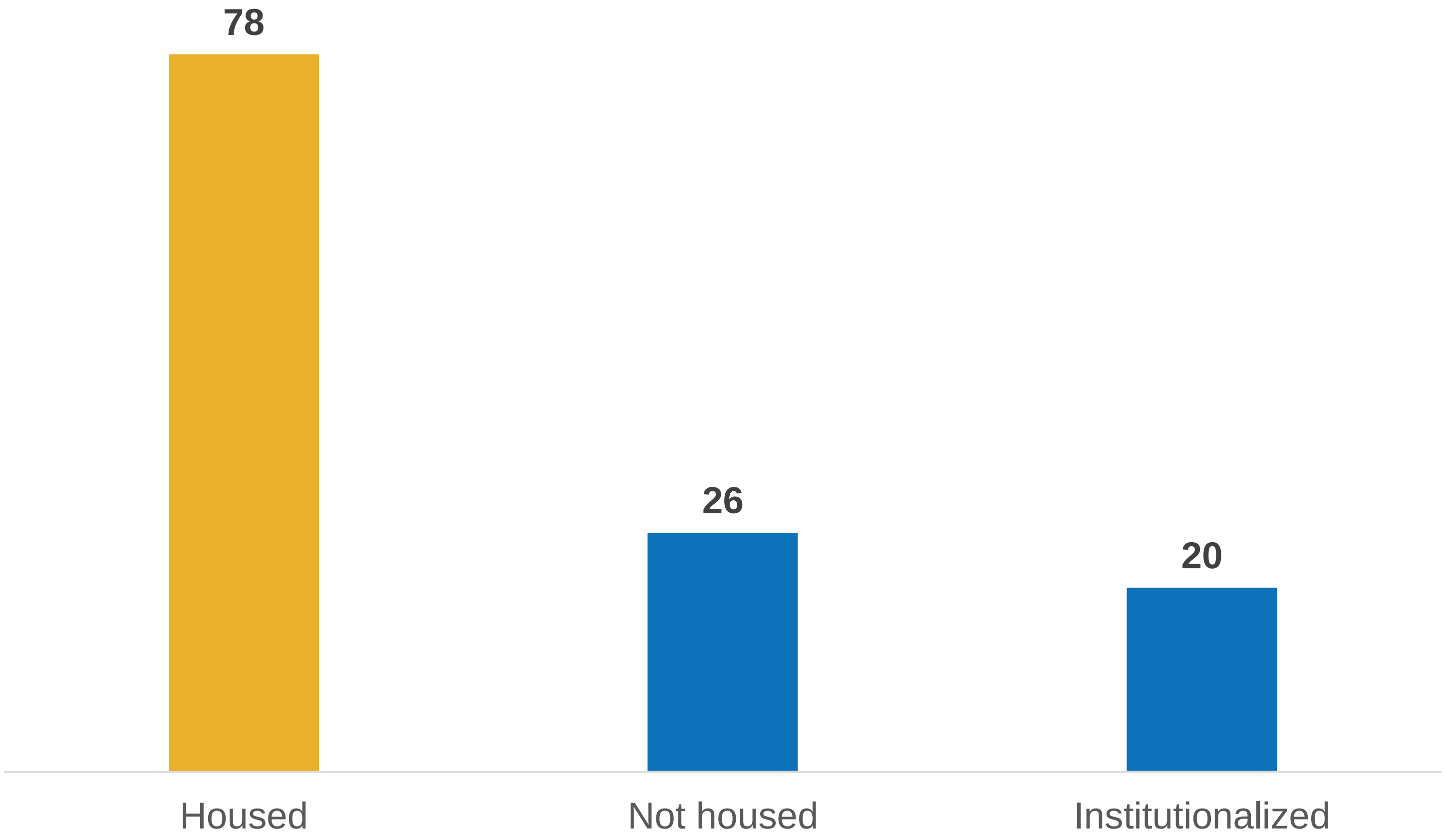


Number of Mental Health Diagnoses



Treatment Attendance

63% of people experiencing chronic homelessness retained housing at 1 year follow-up.



Classification Tree Results

5 factors significantly impacted treatment attendance among referred participants:

Mental Health Substance Abuse

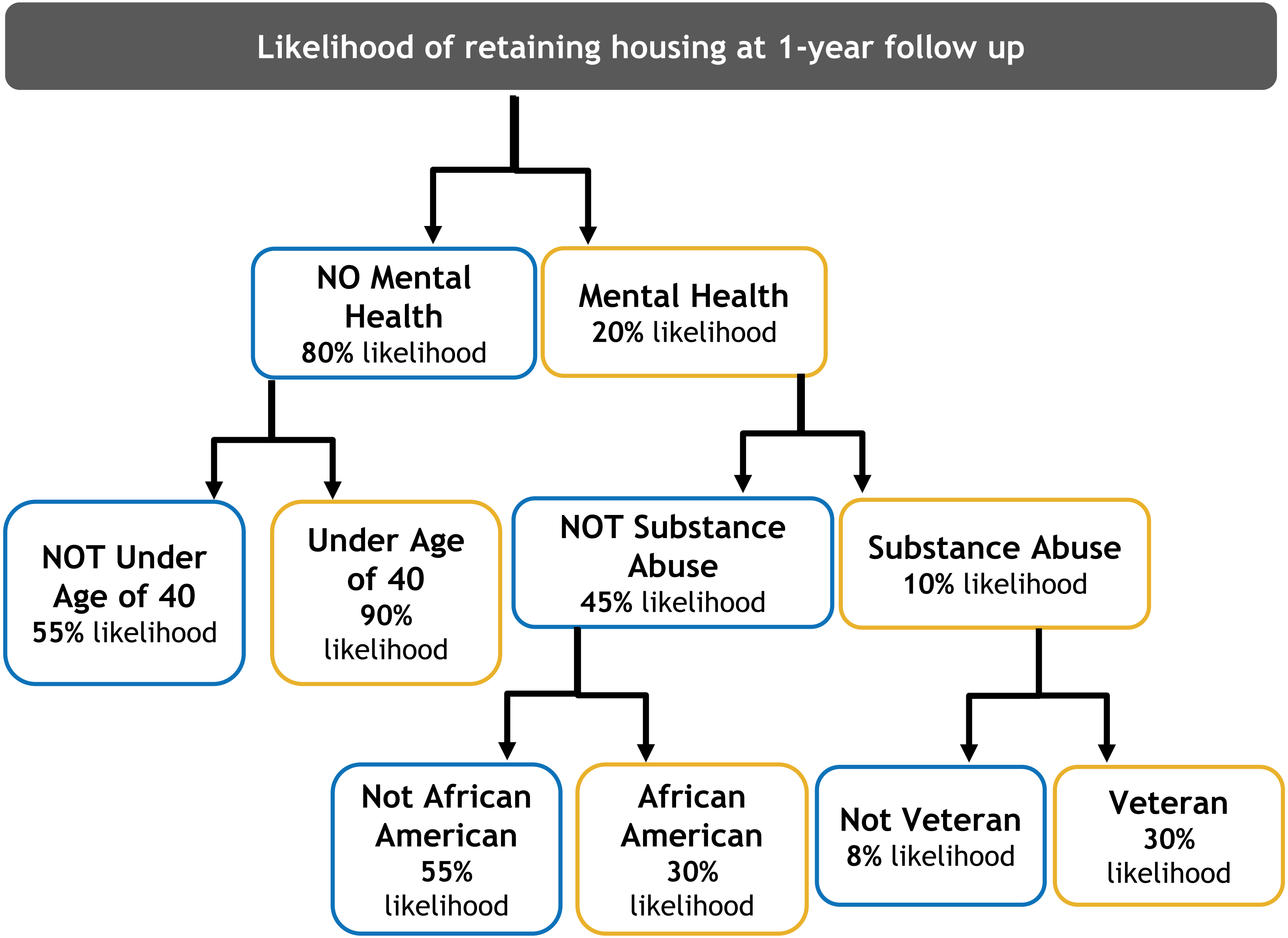
Veteran Status Age

Race

K-fold	R Square
10-Folded	0.23
Overall	0.37

The misclassification rate is **0.18**

Classification Tree Results


Likelihood of retaining housing at 1-year follow up

NO Mental Health
80% likelihood

Mental Health
20% likelihood

NOT Under Age of 40
55% likelihood

Under Age of 40
90% likelihood

NOT Substance Abuse
45% likelihood

Substance Abuse
10% likelihood

Not African American
55% likelihood

African American
30% likelihood

Not Veteran
8% likelihood

Veteran
30% likelihood

Key Conclusions

- Chronically homeless participants who **have a mental health diagnosis**, **have a substance abuse disorder**, and are **not a veteran** are the **least likely (8% likelihood)** to retain housing after one year.

- Chronically homeless participants who **do not have a mental health diagnosis** and **who are under the age of 40** are the **most likely (8% likelihood)** to retain housing after one year.

- Others?

Limitations

- Organization's data quality
- Other factors not included in the analysis could also impact the likelihood of housing retention at follow-up
- Given the small sample size used in this analysis, caution should be applied when generalizing the results of this analysis to larger samples.

Resources for Continued Learning

JMP Website:

http://www.jmp.com/support/help/Partition_Models.shtml#1296905

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine*, *26*(3), 172-181.

Youtube videos

https://www.youtube.com/watch?v=xj-Orr3KTSM

Thank you!

Feel free to reach out to us:

Meredith Philyaw
mphilyaw@med.umich.edu

Jennifer R. Lyons
jrnulty@umich.edu

Additional Slides

Comparing CTA and Regression

Classification Tree Analysis

More holistic view of what factors influence whether or not an individual attains a desired outcome

Easy to account for nested data

Results are presented in an user-friendly format

Results can vary each time you run the model

All right-hand side variables are treated as independent variables

Logistic Regression

Shows the impact of each right-hand side variable on the outcome variable after adjusting for other variables in the model

Multilevel modeling is required if you have nested data

Interaction terms can be difficult to interpret

Results are consistent each time you run the model

You can theoretically differentiate between your IV, confounders and covariates