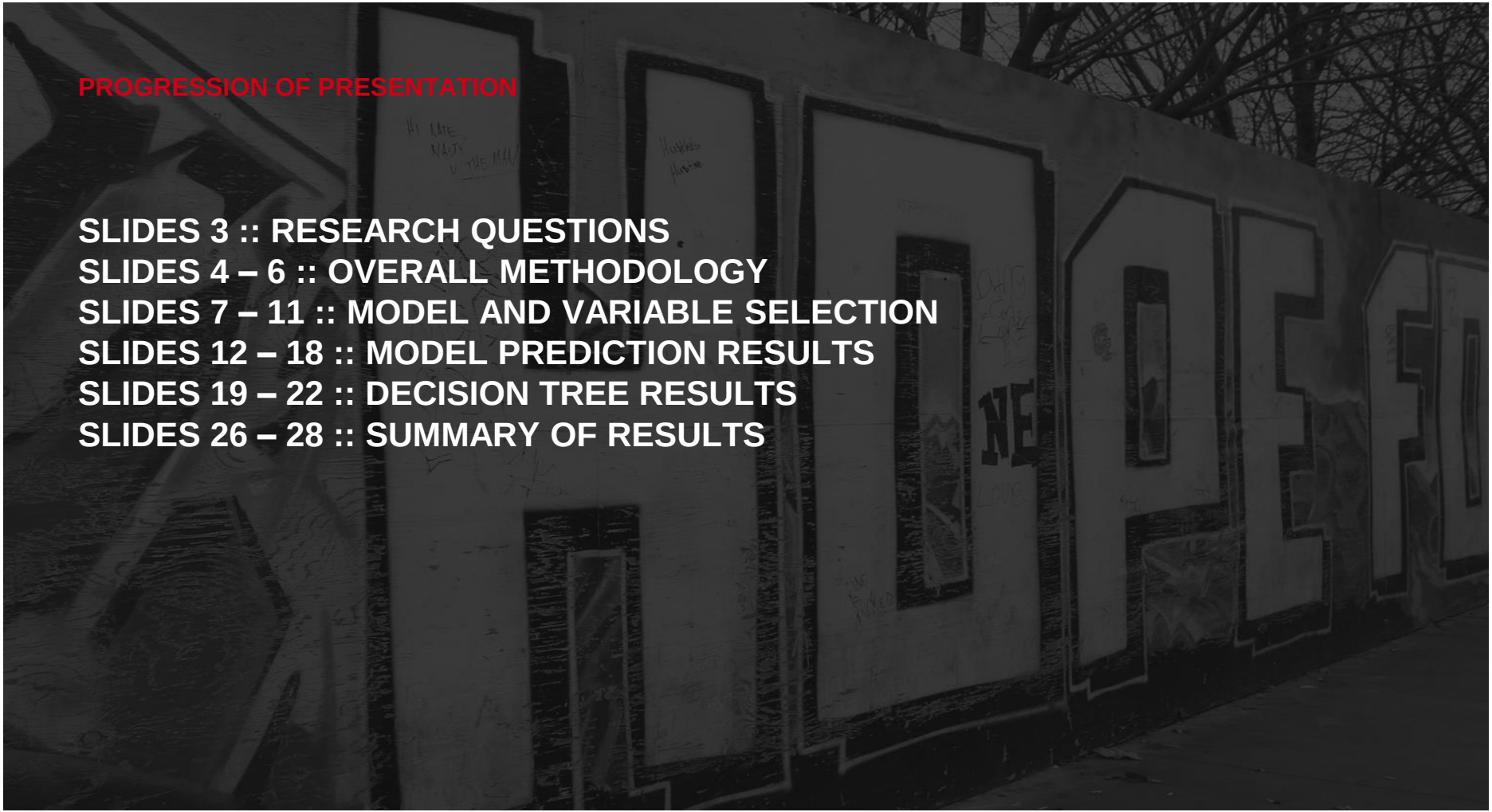
ARPIL 2018

ROAD HOME: PREDICTING FAMILY RETURNS TO SHELTER

PROGRESSION OF PRESENTATION



SLIDES 3 :: RESEARCH QUESTIONS
SLIDES 4 – 6 :: OVERALL METHODOLOGY
SLIDES 7 – 11 :: MODEL AND VARIABLE SELECTION
SLIDES 12 – 18 :: MODEL PREDICTION RESULTS
SLIDES 19 – 22 :: DECISION TREE RESULTS
SLIDES 26 – 28 :: SUMMARY OF RESULTS

RESEARCH QUESTIONS

1. What variables were identified as being more influential on predicting returns to shelter vs. not returning to shelter?

2. How accurate were these variables and the selected models in predicting returns to shelter?

3. Bundled together, what family characteristics were most likely to return to shelter vs. not returning to shelter?

4. On an individual level, how likely would a family return to shelter based on their first enrollment characteristics?

OVERALL METHODOLOGY

- This family shelter analysis explored what collection of family variables were most likely to predict a family's return to shelter. The analysis only included families that were in shelter from 7/1/2011-12/31/2014. Families were excluded from the analysis if anyone in that family had received shelter services before 7/1/2011 or after 12/31/2014. The data predominately included information from the head of household and their family during their first stay.

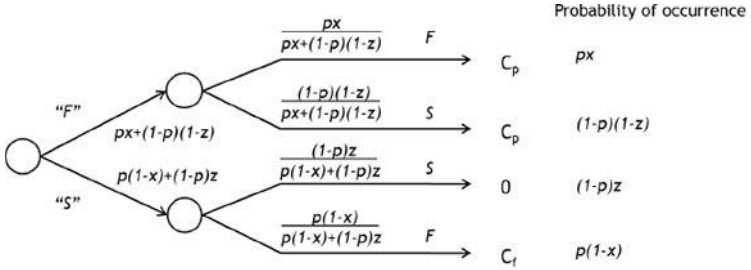
OVERALL METHODOLOGY CONTINUED

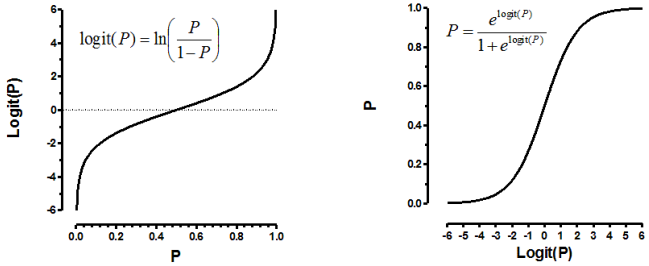
- Families had been identified as receiving shelter services more than once if they had been away from the family shelter for more than 30 consecutive days and then reentered a family shelter. 109 variables were initially created for the beginning of the analysis. These variables were then trimmed down to 39 due to insufficient data on certain variables and multicollinearity when running the logistic regression. 1,083 families were used for the analysis. In the dataset, 69% of the families did not return to shelter and 31% of the families returned to shelter.

OVERALL METHODOLOGY CONTINUED

- There were several statistical and machine learning techniques used for the analysis including: Akaike Information Criterion, Decision Tree learning, and Logit Regression. These techniques were used to identify influential variables related to families that returned to shelter, what groups were most likely to return to shelter, and given the most influential variables, how predictive were they to return to shelter individually as a family.

$$AIC = -2logL + 2q$$





QUESTION #1

What variables were identified as being more influential on predicting returns to shelter vs. not returning to shelter?

METHODOLOGY

- Three models were tested against one another to see which one was better at predicting returns to shelter: All Included, Akaike, and the Lean model. Each of these models had a different amount of variables attached to them. Variable selection is a key component when improving the logistic regression model and its predictions. If there are too many or too few variables it can often lead to overfitting, unreliable coefficients and unconfident p-values.

LOGIT REGRESSION RESULTS: ALL INCLUDED

- Using all of the variables, the logit regression was confident in several variables including:

1. Head of Household Disabling Condition (Binary)
2. Head of Household Age
3. Household Size (including head of household)
4. Certain Head of Household Races (White and Native Hawaiian)
5. First Shelter Stay Night Total

Logit Regression All Included

	Dependent variable:
	StayedInShelterMoreThanOnce
DisablingCondition	1.1***
	(0.2)
VeteranStatus	0.1
	(0.6)
AGE	-0.02**
	(0.01)
GenderMale	-0.003
	(0.3)
TotalFirstHouseholdIncome	-0.000
	(0.000)
TotalFirstHouseholdSize	0.2***
	(0.1)
SPFirstEnrollment	0.2
	(0.3)
DomViolenceExpValueNo	0.01
	(0.2)
DomViolenceExpValueRefused	1.1
	(1.5)
DomViolenceExpValueUnknown	1.7
	(1.5)
DomViolenceExpValueYes	-0.1
	(0.2)
AlcoholAbuse	10.0
	(535.4)
ChronicHealthCondition	11.1
	(535.4)
DrugAbuse	11.2
	(535.4)
MentalHealth	10.8
	(535.4)
BarriersSum	-11.4
	(535.4)
ChildSupportTotal	0.002
	(0.002)
EarnedIncomeTotal	0.001
	(0.002)
OtherIncomeTotal	0.002
	(0.002)
SocialSecurityDisabilityInsuranceTotal	-0.000
	(0.003)
FoodStampsTotal	0.002
	(0.002)
SSITotal	0.001
	(0.002)
TANFTotal	0.001
	(0.002)
UnemploymentInsuranceTotal	0.001
	(0.002)
TotalIncome	-0.002
	(0.002)
RaceAmerican Indian / Alaskan Native and Black	0.4
	(1.0)
RaceAmerican Indian / Alaskan Native and White	0.6
	(0.8)
RaceAsian	0.4
	(0.9)
RaceAsian and White	0.8
	(1.1)
RaceBlack	-0.1
	(0.4)
RaceBlack and White	0.4
	(0.8)
RaceHispanic	0.1
	(0.3)
RaceNative Hawiian / Other Pacific Islander	-1.2**
	(0.6)
RaceOther Multi Racial	0.4
	(0.6)
RaceWhite	-0.6**
	(0.3)
ChildSupport	0.3
	(0.5)
EarnedIncome	0.4
	(0.3)
OtherIncome	-0.7
	(0.5)
SocialSecurityDisabilityInsurance	1.5
	(1.7)
FoodStamps	0.2
	(0.3)
SSI	-0.3
	(0.7)
TANF	0.2
	(0.8)
UnemploymentInsurance	0.6
	(1.1)
AnyHeadofHouseholdIncome	-0.1
	(0.4)
AnyHouseholdIncome	0.2
	(0.3)
FirstShelterStayNightTotal	0.01***
	(0.002)
Constant	-1.1**
	(0.5)
Observations	1,083
Log Likelihood	-597.8
Akaike Inf. Crit.	1,289.6
Note:	*p<0.1; **p<0.05; ***p<0.01

LOGIT REGRESSION RESULTS: AKAIKE

- Using the Akaike Information Criterion, the Logit regression pulled several influential variables including:

1. Head of Household Disabling Condition (Binary)
2. Head of Household Total Barriers
3. First Shelter Stay Night Total
4. Household Size (including head of household)
5. Head of Household Age
6. Head of Household Race
7. Head of Household SSI Monthly Total
8. Head of Household Monthly Earned Income Total
9. Head of Household Monthly Earned Income (Binary)

Logit Regression Akaike

	Dependent variable:
	StayedInShelterMoreThanOnce
DisablingCondition	1.1***
	(0.2)
BarriersSum	-0.4***
	(0.1)
FirstShelterStayNightTotal	0.01***
	(0.002)
TotalFirstHouseholdSize	0.2***
	(0.1)
AGE	-0.02**
	(0.01)
RaceAmerican Indian / Alaskan Native and Black	0.4
	(1.0)
RaceAmerican Indian / Alaskan Native and White	0.5
	(0.7)
RaceAsian	0.3
	(0.9)
RaceAsian and White	0.9
	(1.1)
RaceBlack	-0.2
	(0.3)
RaceBlack and White	0.5
	(0.7)
RaceHispanic	0.03
	(0.3)
RaceNative Hawiian / Other Pacific Islander	-1.1**
	(0.6)
RaceOther Multi Racial	0.5
	(0.6)
RaceWhite	-0.6**
	(0.3)
SSITotal	-0.001*
	(0.000)
EarnedIncomeTotal	-0.001**
	(0.000)
EarnedIncome1	0.5
	(0.3)
Constant	-1.0**
	(0.4)
Observations	1,083
Log Likelihood	-604.2
Akaike Inf. Crit.	1,246.5
Note:	*p<0.1; **p<0.05; ***p<0.01

LEAN REGRESSION RESULTS: LEAN

- This model was leaned up with the following variables below. All of the variables were confident in their coefficient values.

1. Head of Household Disabling Condition (Binary)
2. Head of Household Total Barriers
3. First Shelter Stay Night Total
4. Household Size (including head of household)
5. Head of Household Age
6. Head of Household Race
7. Head of Household SSI (Binary)
8. Head of Household Monthly Earned Income Total

Logit Regression Lean

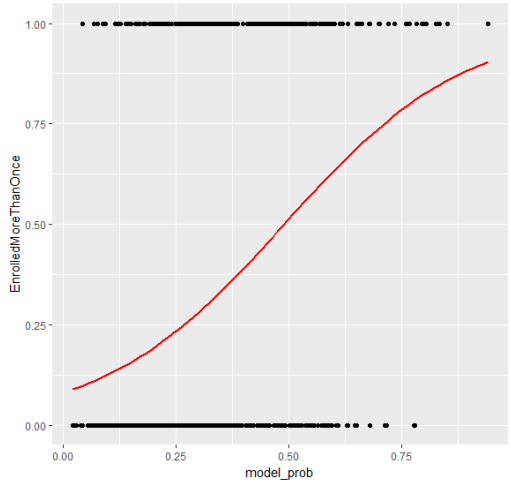
	Dependent variable:
	StayedInShelterMoreThanOnce
DisablingCondition	1.1***
	(0.2)
BarriersSum	-0.5***
	(0.1)
FirstShelterStayNightTotal	0.01***
	(0.002)
TotalFirstHouseholdSize	0.2***
	(0.1)
AGE	-0.02**
	(0.01)
SSI	-0.5*
	(0.3)
EarnedIncomeTotal	-0.000*
	(0.000)
Constant	-1.1***
	(0.3)
Observations	1,083
Log Likelihood	-617.9
Akaike Inf. Crit.	1,251.7
Note:	*p<0.1; **p<0.05; ***p<0.01

QUESTION #2

How accurate were these variables and the selected models in predicting returns to shelter?

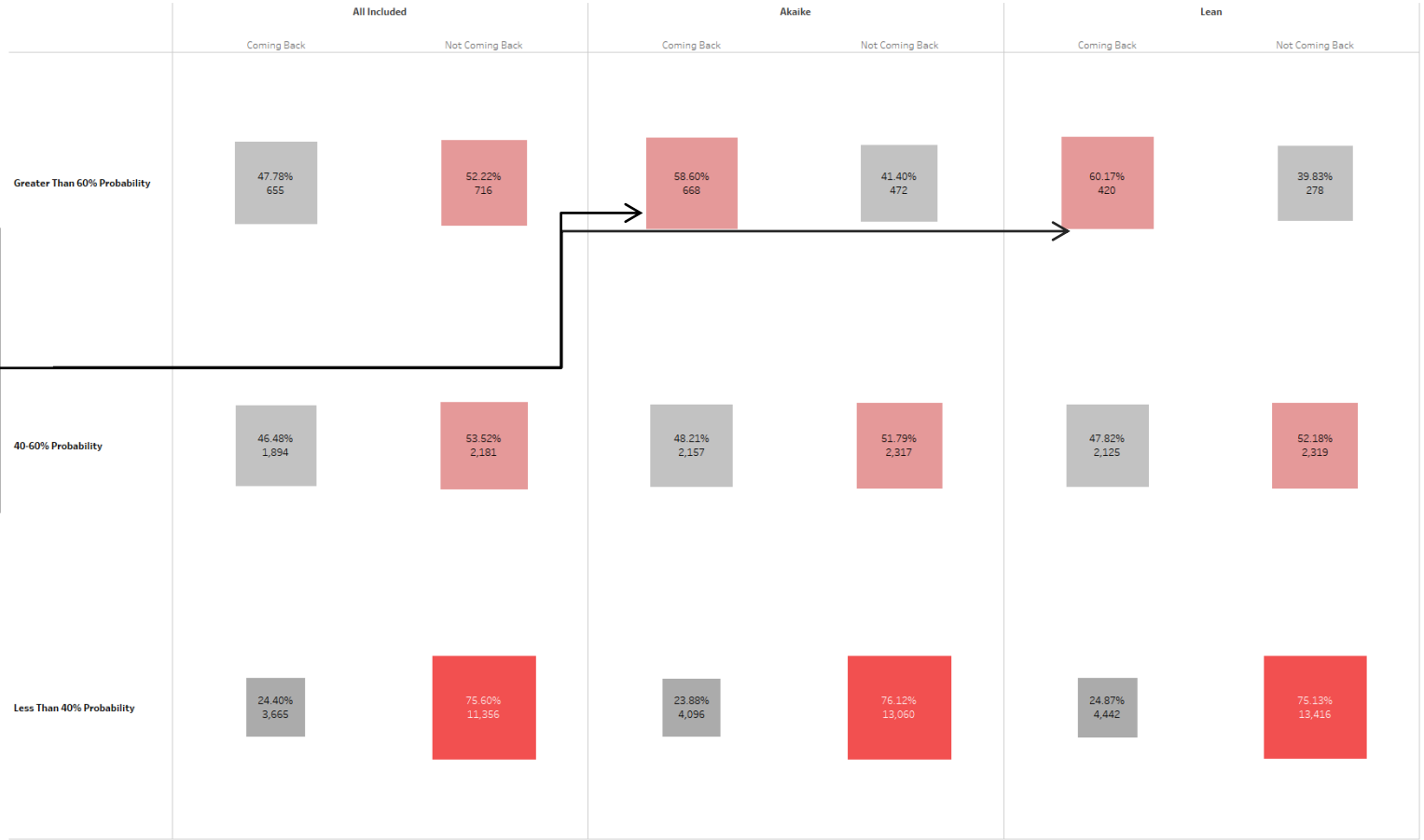
METHODOLOGY

- These three models were tested against one another after using a designated trained dataset with the model and then using a sample of the dataset (that had not been trained) to see how predictive the model was. This simulation was done 100 times to ensure validity and confidence in the final model selection.



MODEL PREDICTIONS: RETURN TO SHELTER

The Lean and the Akaike model both performed significantly better than the All Included model. Considering that roughly 30% of those returned to shelter, the Lean models predictability rate over 60% is nearly double the rate than if it were left to predicting the sample.



MODEL PREDICTIONS: RETURNS TO SHELTER

The Lean Model tended to perform especially well when the predictions were greater than 75% and less than 25%.

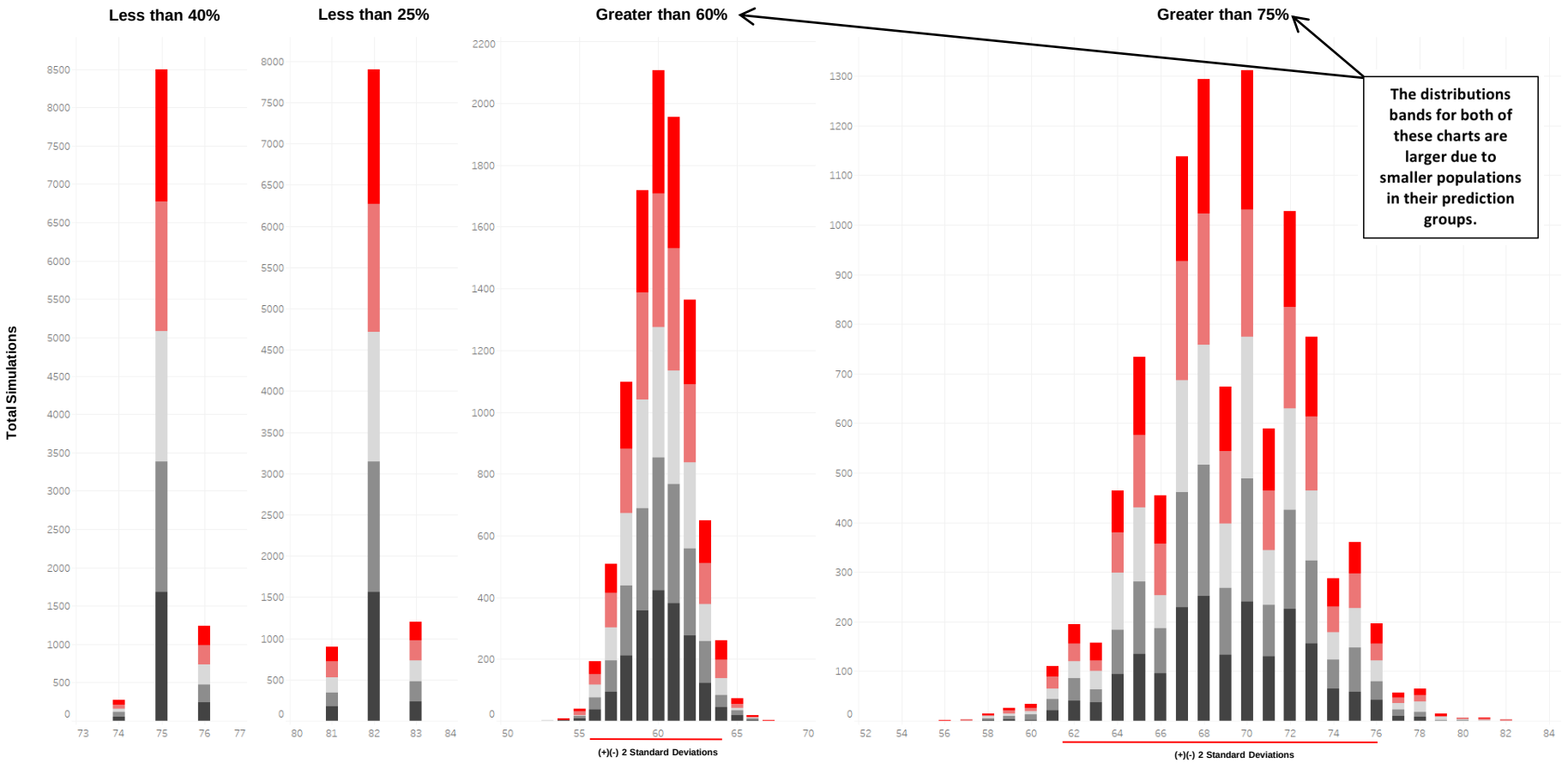
QUESTION #2 CONTINUED

How accurate were these variables and the selected models in predicting returns to shelter?

METHODOLOGY

- The lean model was selected because it seemed to be the most accurate and simplest model. To further ensure the validity of the numbers that one could expect for each percentage group from the Lean model (under 25%, under 40%, greater than 60% and greater than 75%), each subgroup, assuming a binomial distribution, was simulated 10,000 times. These totals give a distribution bar of the confidence of how often each of these numbers would pull up within each group.

LEAN MODEL: SIMULATION RESULTS



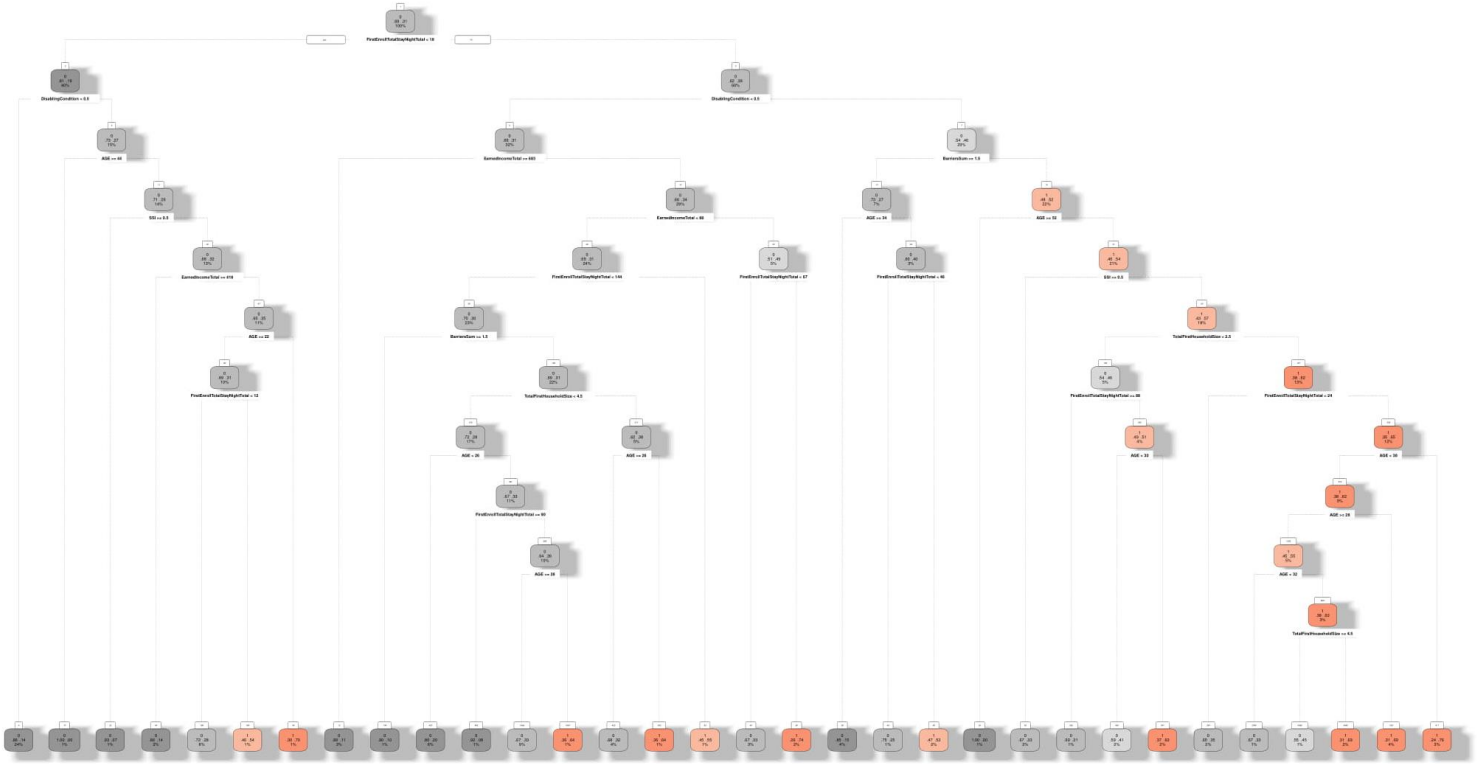
QUESTION #3

Bundled together, what family characteristics were most likely to return to shelter vs. not returning to shelter?

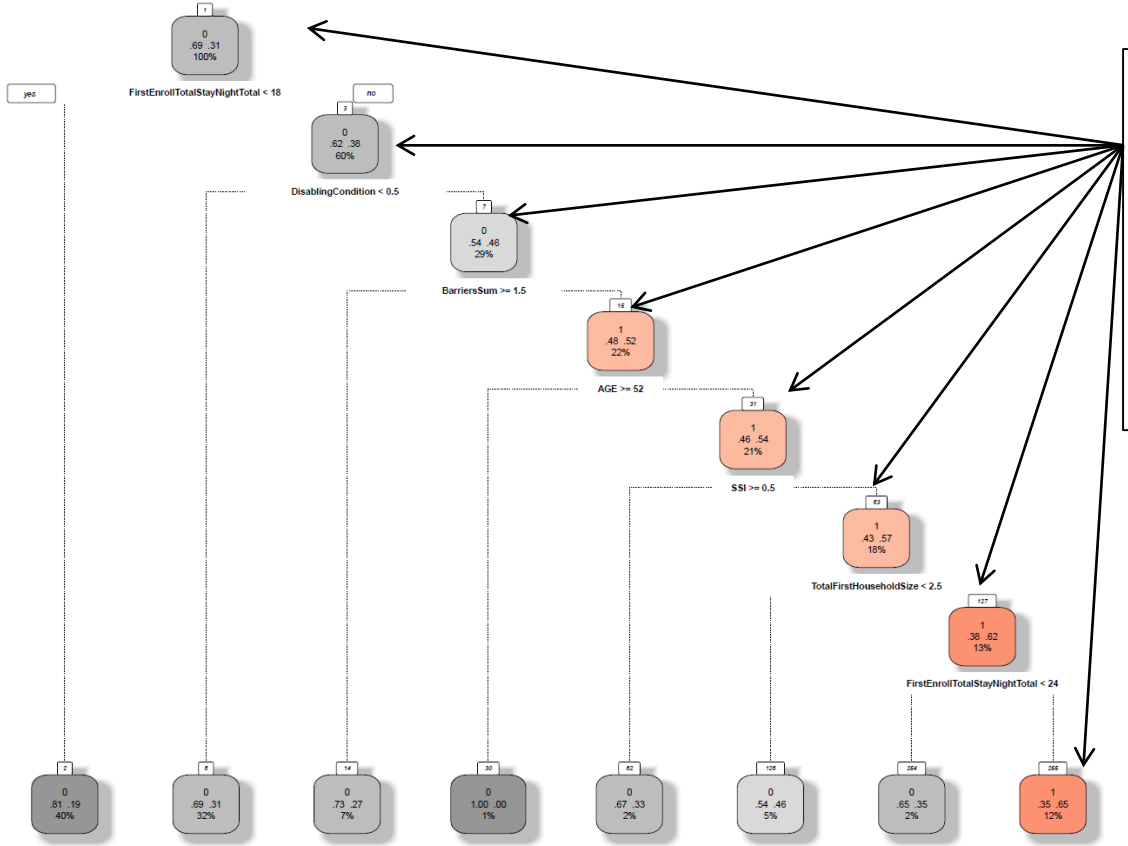
METHODOLOGY

- Using the lean model, a machine learning decision tree was used to identify subgroups that were most likely to return to shelter. These subgroups can give the family shelter a better idea of what target groups are more likely to return to shelter vs. not returning to shelter.

DECISION TREE MODEL: IDENTIFYING AT-RISK POPULATIONS THAT WERE LIKELIER TO RETURN TO SHELTER



DECISION TREE MODEL: A CLOSER LOOK



Families with head of households that stayed in shelter longer than 24 days, had a disabling condition, had less than two barriers, was less than 52 years old, did not have SSI, had a household size greater than two were twice as likely to return to shelter than the population. This group also represented 12% of the total 1,083 population used in the sample.

QUESTION #4

On an individual level, how likely would a family return to shelter based on their characteristics during first enrollment?

METHODOLOGY

- For individual families, a dashboard using the Shiny application was created to identify the probability of that family returning to shelter. This dashboard used exponential log weights to calculate the prediction for each individual family and their return to shelter. This is the same prediction that was used for the results found on slides 12-15.

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

DASHBOARD: PREDICTING RETURNS TO SHELTER

Predicting Returns to Shelter

Head of Household Age

28

Length of Stay in Family Shelter (Total Days)

150

Total Household Size (including head of household)

4

Does the head of household have SSI?

No

Does the head of household have a disabling condition?

Yes

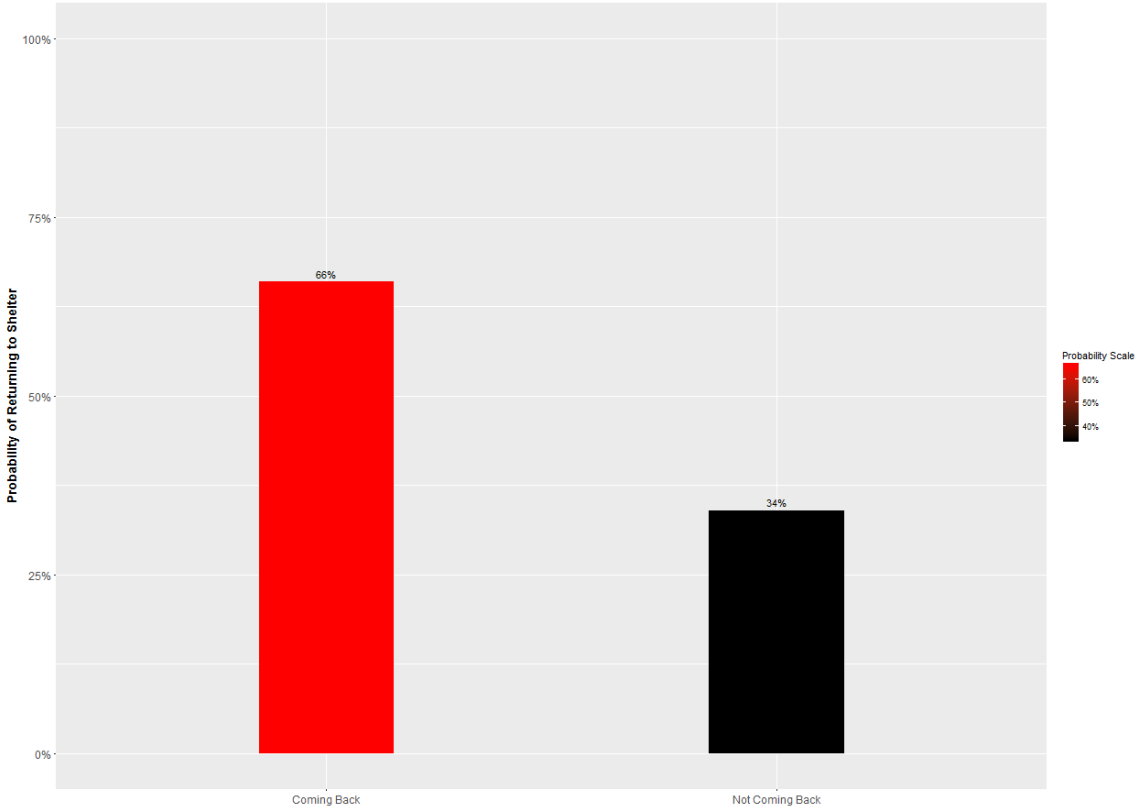
Total Head of Household Earned Monthly Income (Note: put 0 if the head of household does not have any income)

500

Total Barriers (Note: put 0 if the head of household does not have any barriers)

0

Note: Barries include alcohol abuse, physical disability, HIV/AIDS, developmental disability drug abuse, mental health, chronic health condition depression, injected drug use, schizophrenia, and spice.



Click anywhere on the photo to be redirected to the dashboard.

Summary of Results

SUMMARY OF RESULTS

- Three predictive models were tested against one another to assess their predictive accuracy of identifying when a family returned and did not return to shelter. The “Lean” logit regression model was chosen due to what seemed to be the most predictive model. Using decision tree modeling, the Lean model found key family subgroups that were much more likely to return to the shelter. For example, families that have stayed in the shelter longer than 24 days, had a disabling condition, had less than one barrier, was under 52 years old, did not have SSI, had a household size greater than two were twice as likely to return to shelter than the population. Considering that this subgroup consisted of 12% of the total datasets population, this is an especially influential demographic group that has returned to the shelter.

SUMMARY OF RESULTS CONTINUED

- The Lean model was also especially more predictive when the logistic predictions were greater than 60%, where it predicted nearly double the rate of returns to shelter than the original dataframe. The model also improved its predictive accuracy when the predictions were less than 25% and greater than 75%. If employees at the shelter choose to use the decision tree or the “Predicting Returns to Shelter” dashboard to gain information about the family that has received shelter services for the first time, they can use the information to allocate the necessary labor, resources and potentially housing to the identified families that are at a higher risk of returning to shelter.