DATA 630-9040

Assignment 3

7/3/2021

Prepared by Kierra Young

Prepared for Professor Bati

**Introduction**

**Objective**:

The objective of this analysis is to “predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics”( Contraceptive Method Choice Data Set). The type of analysis that is being conducted in an explanatory analysis and the specific modeling type that will be used is a Decision tree. With the explanatory analysis and Decision tree model the type of questions that could be answered are the following “Does age play a role in a woman using contraceptive methods?”, “If a woman has or does not have an education is she more likely to use a contraceptive method or not?”, “If a women is working is she more or less likely to using contraception methods”, “Do living conditions affect chose to use or not use a contraceptive method?”, and “Are religious women more or less likely to use contraceptive methods?”

**Problem Domain**:

The background information for problem domain of this dataset is that many women the women that were coming into the facility for interviews with either not pregnant or did not know they were pregnant according to the dataset. In terms of context, the idea is seeing if there are any demographic factors that lead to a woman using or not using certain contraception methods at a given time. Some interesting statistics is that

“U.S. women aged 15–49 in 2018 were sexually active and not seeking to become pregnant. Sexually active couples who do not use any method of contraception have approximately an 85% chance of experiencing a pregnancy over the course of a year. The average number of children U.S. adults think is ideal is 2.7.3 To achieve this family size, a sexually active woman must use contraceptives for roughly three decades”( Contraceptive Use in the United States).

**Method Rationale**:

The method of using a decision tree classifier is so that all the items in the dataset can be out into target categories. It can help to accurately predict the target class for each of the variables in the contraception method dataset. It is applicable for the problem at hand because it can easily identify the variables that will affect the no use, short term use, or long-term use of the contraception methods.

**Analysis**:

**Data**

“This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. The problem is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics”(Contraceptive Methods Dataset). The explanation of the variables are the following and self-explanatory: Wife's age, Wife's education, Husband's education, Number of children ever born, Wife's religion, Wife's now working, Husband's occupation, Standard-of-living index(low, standard, medium), Media exposure, and Contraceptive method use (no use, short-term use, long- term use). For deeper explanation, all these variables are independent, and they will be used to predict the target variable Contraceptive Methods.

**Exploratory Analysis**

Str and Summary

The str shows the data frame of the contraceptive methods dataset. It shows that there were 1473 observations of 10 variables. All the variables are integers and labeled categorial. Also, none of the variables are missing in the dataset. As far as the summary since all of the variables are integers all of them were able to show the min, max, 1st Qu , Median , Mean , and the 3rd Qu.

Target Variable Chosen

The target variable from this dataset that was chosen was “Contraceptive Method”. This variable was chosen because it is the only variable that is dependent amongst the entire datasets. This variable has the data of no use, short term use and long-term use in it. This target variable will be the main indicator of predicting the main demographics when people use contractive methods or not.

Bar Chart

The bar chart located in the appendix as figure 1 shows the wife’s age, the contraceptive method of use, and the percentage of women that did use, do not use, or used for short term. The black coloring on the bar columns means that a woman was pleased  means that they did not use a contraceptive method. The bar columns that are pink means that the woman used the contraceptive methods for a long time, and the bar columns that are green are for a woman that uses the contraceptive methods for a short term. From this chart, it can easily be said that women over the age of 40 are very much less likely to use contraceptive methods versus females that are younger than 40. I also wanted to point out that females between the age of 16 and 19 were highly likely to use contraceptive methods for the short term and any other type of use. Also women that were 25 were the highest group to use short term contraceptive methods and they are less likely to use long-term contraceptive methods.

Histogram

The histogram shown in the appendix as figure 2, shows the data sets contraceptive methods on a scale from 1 to 3. On the left-hand side you can see the total method used percentage wise, which is basically how many people in the dataset used a certain method type. From this histogram it  can be seen that method type number one which is “no use” has the highest percentage of contraceptive method choice amongst the women in the data set at over 40%. Method number two which is “long term use'' has the lowest usage of method types in the data set which means a lot of women decided not to use long-term prevention methods for pregnancy with a percentage between 20 to 25% . Coming in as number two method type 3 which is “short term use” has a percentage of bearing over 35%.

**Preprocessing**:

For pre-processing the contraceptive methods data set, all the data was there, so  there were no missing values, and all the values were already integers, so it was easy to make the histogram and  bar chart. The most that needed to be done was setting the seed information so that way I could split the data in the train and test data. For splitting the information, the most that I had to do was write  the train and test data code. The training and test data are split Because I planned on making a model with both train and test. The training data was split into 70% while the test data  was split into 30%. Since this was not a linear model, I did not look for any outliers. In addition, the model was also split into train and test data so that way when I build my confusion matrix and my decision tree model, I will get the most accurate results possible.

**Algorithm Intuition**:

For intuition of the algorithm,  classification was the algorithm method chosen. The reason for using this method is so that I can classify the demographics of women that either decided to use contraceptive methods for a long time or a short period of time or did not decide to use it at all. I want to investigate that target variable and see all the demographic variables that play a part in a woman choosing one of those contraceptive methods. As far as intuition, I will be able to classify all those variables into different categories and see which ones are the strongest predictors for what a woman might choose in the future.

**Model Fitting**

I fit the model by using the formula for a ctree to be built. As far as fitting the model I used the training data set to build the decision tree model. There is a formula that was used one that can make the “the formula is an expression that contains the dependent variable followed by ~ (tilde) and independent variables. The dependent variables are delimited by + (plus sign)”(Decision Tree with R Docs). To explain, the formula was placed as “myFormula<-ContraceptiveMethod~” which sets the input to the contraceptive method variable class. The next part of the formal was “ cmc\_ctree <- ctree(myFormula, data = train.data)”. This sets the dataset to the main data to a ctree and my formula will equal the train data to deliver the classification tree.

**Result**

**Output**:

The result of the ctree model is that the conditional inference tree had 8 terminal nodes for the contraceptive training data set. The target response was the variables ContraceptiveMethod, and the inputs shown were WifeAge, WifeEducation, HusbandEducation, NumChildren, WifeReligion, WifeWorking, HusbandOccupation, LivingStandardIndex, and MediaExposure. The number of observations found were 1044. WifeAge is the first splitting attribute. If the value is <= 37 or is missing, then the variable NumChildren is the second splitting attribute. Next of Wife age is >37 then Wife education is the second splitting attribute when it comes to women choosing a contraceptive method.

Likewise, the nodes are also important as they do make the branches on the decision tree. Figure 3 in the Appendix showcases the actual ctree model. For the actual decision tree diagram the first node begins with the wife's age. There are two branches of wife’s age. The first branch goes to greater than 37 or equal to 37 and then goes to number two which is the number of children. Number 2 leads to node 3 which gives a hierarchy of the type of uses from three to one. Node number two also goes to node number four which is also the number of children which leads to no number eight which gives the hierarchy of methods used. Node number five, which is the wife's age, just leads to Node 6 and Node 7. Then on the other side of the wife's age, which is node number one it leads to less than 37. The method used depends on the wife's education which is node number nine and then it goes to number 10 which is standard of living index. Node 10 leads to node 11 and 12 for the method used. Then wife’s education also leads to node number 13 which leads to Node 14 and 15 for the hierarchy of terms of use.

The objective has been met. The goal was to find the variables that target a women deciding if she is more likely to no use, short term use, long term use of the contraceptive method. The ctree shows all the main indicators in a categorical fashion.

**Model Properties**

To explain the model properties, the first part is the formula, and it outputs the quantity of leaf nodes, dependent and independent variables, quantity of observations in the training set, and the variables that were split in the dataset (Decision Tree Classification).The terminal nodes are marked with an \* sign beside each node number. This training set had 8 terminal nodes. Likewise, the actual tree diagram consists of 15 nodes and 8 terminal nodes. Each of the nodes connect to another node given the direction that the branch is taken. For example, if wife’s education is equal to or greater than 2 and the living standard is greater than or equal to 3 the node for 11 would be either 2 or 3, 2 is for long term use and 3 is for short term contraceptive use. That is how one would understand the model properties.

**Evaluation**

The output for the confusion matrix was rather difficult to understand. The command for the matrix did work, but the output was not the most legible. The data kept giving 9 roles and only 4 columns. I tried using three more different codes and some with the caret package, however it said that the values had to be factors instead of int so that did not work either. An example is below:

#9. building a confusion matrix

> table(predict(cmc\_ctree), train.data$ContraceptiveMethod)

1 2 3

1 67 0 0

1.08333333333333 23 0 1

1.18518518518519 70 7 4

1.64935064935065 44 16 17

1.7037037037037 15 5 7

1.91907514450867 57 73 43

2.03688524590164 95 45 104

2.29344729344729 81 86 184

> prop.table(table(predict(cmc\_ctree), train.data$ContraceptiveMethod))

1 2 3

1 0.0641762452 0.0000000000 0.0000000000

1.08333333333333 0.0220306513 0.0000000000 0.0009578544

1.18518518518519 0.0670498084 0.0067049808 0.0038314176

1.64935064935065 0.0421455939 0.0153256705 0.0162835249

1.7037037037037 0.0143678161 0.0047892720 0.0067049808

1.91907514450867 0.0545977011 0.0699233716 0.0411877395

2.03688524590164 0.0909961686 0.0431034483 0.0996168582

2.29344729344729 0.0775862069 0.0823754789 0.1762452107

> #10. making a table for test data

> testPred <- predict(cmc\_ctree, newdata = test.data)

> table (testPred, test.data$ContraceptiveMethod)

testPred 1 2 3

1 18 0 2

1.08333333333333 6 0 1

1.18518518518519 22 2 1

1.64935064935065 13 7 13

1.7037037037037 15 1 3

1.91907514450867 19 28 15

2.03688524590164 43 28 46

2.29344729344729 41 35 70

**#10.New code method for confusion matrix**

> install.packages('caret')

Error in install.packages : Updating loaded packages

> install.packages("caret")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Warning in install.packages :

package ‘caret’ is in use and will not be installed

> #**11. another method for confusion matrix**

> confusionMatrix(cmc, test.data, positive = NULL, dnn = c('Prediction', 'Reference'))

Error: `data` and `reference` should be factors with the same levels.

**> #12. Last new method for the confusion matrix**

> confusionMatrix(cmc$ContraceptiveMethod, testPred)

Error: `data` and `reference` should be factors with the same levels.

>

**Conclusion**

**Summary**

Some of the key findings here was that Wife’s Age is the main indicator if a woman that is married will have no use, short term use, or long-term use for a contraceptive method. Also coming in second place a women’s number of children also greatly affect the outcome of a women using contraceptives for short term, long term and possibly no use. Likewise, a women’s education plays a major role on if she will use contraceptives as well being followed by standards of living and the number of children she already has.

**Limitations**:

Confusion matrix and converting int to factors was the biggest limitation for this analysis. I researched multiple ways to turn an int into a factor but was unsuccessful and many of the codes for confusion matrix only worked on factor variables. So, the confusion matrix was the hardest part to interrupt or understand. Likewise with the confusion matrix the first command printed out a matrix that was 4 by 9 with decimals in it and I was very unfamiliar with what was happening or why it was occurring.

**Improvement Areas**:

More research on converting int to factors and looking up more ways to code a confusion matrix in R and see what other packages available would-be great improvements for me in the future. Likewise, to look for more possible codes that can use int variables for a confusion matrix. Most of the codes I saw only worked for factor variables.

**Appendix**

Figure 1. Bar chart

Chart, bar chart

Description automatically generated

Figure 2. Histogram

Chart, bar chart, histogram

Description automatically generated

Figure 3. Ctree

Diagram

Description automatically generated

References

*Contraceptive Use in the United States by Demographics*. Guttmacher Institute. (2021, May 28). <https://www.guttmacher.org/fact-sheet/contraceptive-use-united-states>.

UCI Machine Learning Repository: Contraceptive Method Choice Data Set. (n.d.). https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice.

UMGC. (2021, July). *Decision Tree Classification* . *Data 630*.