Final Report

Adam Dameron

January 16, 2023

Abstract

1 Outline Methods

- Dataset (History, purpose, etc)
- Types of models created
- Why models require complete data and fewer variables (over fitting, computing power, etc)

2 Introduce Dataset

- Summarize dataset and define variables (https://cran.r-project.org/web/packages/naniar/vignettes/getting-started-w-naniar.html)
- Explain what NA means in relation to the data set, and reasons NA may be present - Briefly describe/define missing (or complete) entries and explain why this matters
- Remind that model can only be created with the complete entries

3 Manipulating CTDC Dataset

The CTDC data set contains 63 variables and 0 of its entries are complete. This means that without any data manipulation, creating a model is impossible. Creating a model is possible only if we have a significantly larger number of complete entries. One way to manipulate the data set into having a larger number of complete entries is by removing entire columns from the data set altogether.

While this process is efficient at creating complete entries, it is important to note that we would be losing some information in the process. For example, if we remove the variable that indicates if a victim was trafficking in the mining industry, then our final model cannot incorporate that information. This is why it is important a balance is struck between having complete entries, while still maintaining as much information as possible.

Carefully choosing which variables to remove will help ensure we are aware of what information we are losing, and that will be accounted for that in the analysis of the final model. There are two ways in which variables (columns) will be omitted from the data set. The first way is by analyzing what the variable represents. By understanding what the variable means, we can logically conclude if it will be helpful, or if it should be removed. A second approach is to quantify the amount of incomplete entries that are caused by the variable, or a group of variables. By looking at missing values, we can see what variables are missing in tandem with each other. This would help us to better understand the structure of the data set, and provide a better understanding of what causes the missing values to appear.

3.1 Logically Removing Variables

There are two variables that can be identified in the data set that meet this criteria. These variables are ”Data Source” and ”Year Of Registration.” Both of these variables are representative of the manner and time in which a case was added to the data set. Data source is whether the case was reported over a hot line managed by IOM, or through a case manager on the victim's behalf. The year of registration is the year in which a case was added to the data set. Since these two variables only describe the reporting process, they will not be helpful in the process of identifying victims within a country, and can be removed without having a negative impact on the effectiveness of any models.

A second type of variable that can be removed are those which serve to summarize other data contained within the data set. There are a few examples of variables which are concatenated versions of other variables and provide a written text summary. These variables are:

- Means of Control Concatenated
- Type of Exploit Concatenated
- Type of Labour Concatenated
- Recruiter Relationship

In a similar category as the previous variables, the ”Majority Status” variable serves to identify whether or not a victim was an adult at the time they were exploited. However, the ”Age Broad” variable already covers age information, and including ”Majority Status” would essentially serve as a summary variable of the age information. While it is true that the age of majority is different in various countries, the Age Broad variable is a more specific representation of the characteristics of the victim.

As a result of these findings, the following variables will be removed:

- Data Source
- Year Of Registration
- Means of Control Concatenated
- Type of Exploit Concatenated

- Type of Labour Concatenated
- Recruiter Relationship
- Majority Status
- Majority Status at Exploit
- Majority Entry

After removing these columns, our data set now has 22 complete cases. This is an improvement, but it is certainly not enough to make any meaningful model. However, quantitative methods will yield better results.

3.2 Quantitatively Removing Variables

The second way that individual variables can be removed is by using quantitative methods. There are many processes to complete this task, and a handful of them will be applied to this data set.

One way is to simply look at all the different values that the variable takes. If we see that all the entries for a variable in the data set are either NA or 1, then it is clear that any complete row will have a value of 1 for that variable. This means that the model will only take in the value of 1 for that feature in each row. Thus the variable will have a null effect on the model. This process led to the removal of:

- Is Forced Military
- Is Organ Removal
- Type of Labour Mining/Drilling

After removing these variables, there are still only 22 complete entries. However, the removal of these variables can do nothing but help us, as there is no way they can have an effect on our model. Unfortunately, these variables are the only type that we can remove and have no negative consequences. Any other variables we omit will have downsides, and it would be beneficial to try to find which variables are having a significant effect on the number of missing values, and to remove those. Since each absent entry is given a value of NA, we can count the number of NA's in each column to get a sense of which ones are contributing the most to the lack of complete entries.

Figure 1: Created with data from