Final Report

Adam Dameron

January 12, 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
- Briefly describe number of missing (or complete) entries
- Remind that model can only be created with the complete entries

Manipulating CTDC Dataset

2.1 Removing Variables

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 more complete entries into forming is by removing entire columns from the data set altogether.

While this process is effecient 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 we strike a balance between 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 account for that in our final model analysis. There are two ways in which variables will be omitted from the data set. The first way is qualitatively, by understanding what the variable means and represents, 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.

2.1.1 Qualitatively

There are two variables that can be identified in the dataset that meet this criteria. These variables are ”Datasource” and ”Year Of Registration.” Both of these variables are representative of the manner and time in which a case was added to the dataset. Datasource is whether the case was reported over a hotline 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 dataset. 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.

A second type of variable that can be removed are variables which serve to summarize other data contained within the dataset. 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 lead to multicollinearity within the data. Not to mention the fact that many reported cases are missing a value for ”Majority Status.”