Stop Human Trafficking: An Analytical Approach
MGT 6203
Team 064
Lucas Chen, Amit Dommeti, Roy Garcia, Garrison Winter
TA: David Firrincieli

Background:
Human trafficking is typically considered a form of modern-day slavery. Human Trafficking, as defined by Mariam Webster, is "organized criminal activity in which human beings are treated as possessions to be controlled and exploited (as by being forced into prostitution or involuntary labor)." It has an everlasting impact on affected individuals, leading to lifelong damage to the survivor's mental and physical health. As depicted by Love Justice, there are "an estimated 50 million people trapped in slavery around the world" because of human trafficking. There is an increasing amount of public attention that is being driven toward the problem of human trafficking. There is a shortage of analytics solutions that can assess decisions related to preventing human trafficking.

Problem Statement:
Human trafficking has been a worldwide problem and using the 130k+ records of human trafficking data, our team aims to determine the demographics of the most susceptible populations to human trafficking which can help organizations better allocate their funding to mitigate this issue.

Research Question:

- What is the most used tactic for abducting individuals?
- How has human trafficking trended over the last 2 decades?
- Has human trafficking leaned or shifted towards a specific group of people? (specific gender, age group, and others…)

Initial Approach:
Using the dataset, our team initially performed a time series analysis to understand trafficking trends as well as segmented out certain groups that may be at highest risk of being targeted. The team conducted analysis through Python. We also performed model analysis while considering other approaches.

Unexpected Challenges
We had to change the original dataset proposed to a new one due to this one having more datapoints in certain years. This new dataset also has a holistic and better representation of global human trafficking while showing a better protection to the privacy of the victims among other benefits such as:

- To resolve the challenges in previous datasets, the team finalized on a new dataset that has 130K+ records globally.
- The distribution of human trafficking by year is a lot more evenly distributed than the previous dataset.

- Dataset now also contains more detail on the relationship between victim and perpetrator, region of citizenship, etc
- The data contains k-anonymity (“safety in numbers”) for group privacy and differential privacy (“safety in noise”) for individual privacy – working with sensitive data, we must consider the ethics behind it.
- The dataset was created in collaboration with Microsoft Research and the CTDC to create an accurate Global Synthetic dataset that contains data from 156,000 victims and trafficking survivors across 189 countries and territories.
    - Allows for complete data that protects the privacy of individuals so their information/status is not publicly exposed
    - Preserves all statistical properties and relationships in original data

Data model:

Our data model comes from the CTDC site; the file comes in .tsv format – it is a complex dataset which contains 39 columns with the Year of Registration, Gender, Age Range, Majority Status, Traffick Months, Citizenship, among many other columns that provide great insights such as Means of Control (e.g. Physical Abuse, Psychological Abuse, False Promises,…), type of labor (if is Forced Labor or Sexual), type of labor (Agriculture, Construction,…), type of sex trafficking (pornography, prostitution,..), and the recruiter relationship with the victim (partner, friend, family,…). Each scenario is in a separate column, which helps with analysis. We had a lot of data to analyze and find the best analysis scope and model.

Data cleaning process:

Most of the data is already cleaned but will refer to the CTDC Codebook for all data cleansing methods. On top of those we performed some extra steps such as the creation of an Index Column, replacing all values that are empty or –99 to null values, replacing the citizenship with country of exploitation, filtering accordingly to the necessary data for each model tested/proposed.

Initial Findings:

Working with global data with numerous locations creates multiple problems so we decided to focus first on a global scale and evaluate later if it makes sense to go down to a country level.

We did an exploratory analysis of the top countries of trafficking volume compared to the bottom countries of trafficking volume to see if there were any key trends that applied to heavily trafficked countries vs lightly trafficked countries. One thing to note is that the bottom countries had a lot fewer data points, so we pulled a lot more countries from the bottom portion and used the top 5

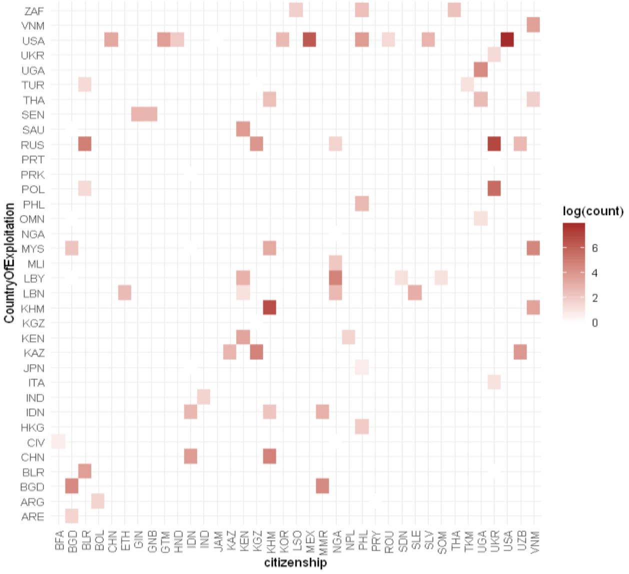


Figure 1 – Heatmap of the Log(count) of victims’ citizenship vs country of exploitation

countries in the top portion. First, we broke down the countries based on country of citizenship. In terms of age, both the top countries and bottom countries exhibited similar trends with the 9-17 and 30-38 age groups having the highest volume. However, the top countries had the highest volume in the 30-38 age group while the bottom countries had the highest volume in the 9-17 age group. This could be due to a larger youth population in the highly trafficked countries.

With the gender breakdown, both sets of countries had about twice as many females getting trafficked vs male. This figure makes sense since sex trafficking is one of the most popular forms of human trafficking and females are more desired than males in this realm.

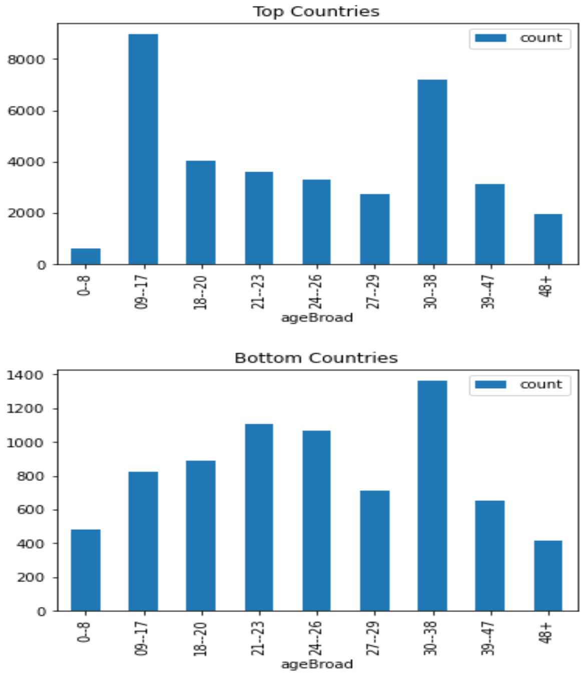


Figure 2 - Age breakdown of Top vs Bot Countries

The top countries exhibit a shorter amount of overall time trafficked with both countries having the highest volume in 0-1 year but the second highest volume in top countries was for 1-2 years while for the bottom countries it was 2-5 years. One explanation for this trend could be that the high-volume countries have a lot more humans being trafficked so you can purchase new humans more often.

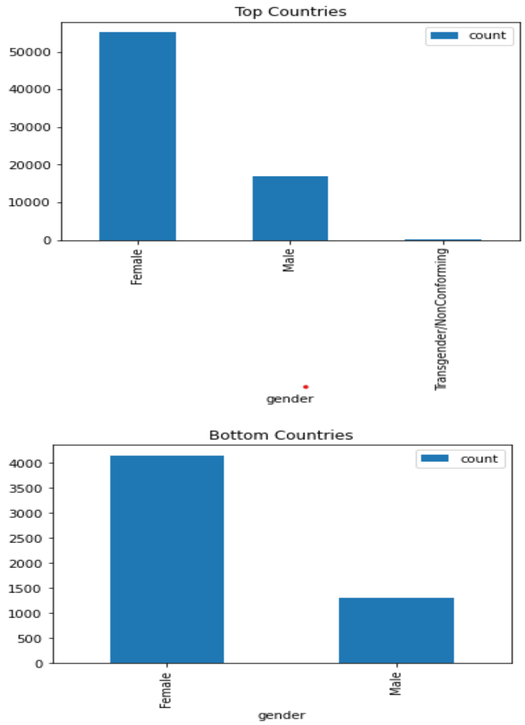


Figure 3 – EDA with gender groupings comparing Top and Bottom countries distributions.

Next, we broke down the countries by the country of exploitation. The top five countries shifted but USA remained in the top 5 for both categories. The age breakdown reversed from the citizenship analysis with the 9-17 age group being the highest volume in the top countries while the 30-38 age group being the bottom countries. The gender ratio for both sets of countries remained about the same with females being trafficked at twice the rate of males. Lastly, both top and bottom countries exhibit the same trend as citizenship with top countries trending towards shorter trafficking times than bottom countries.

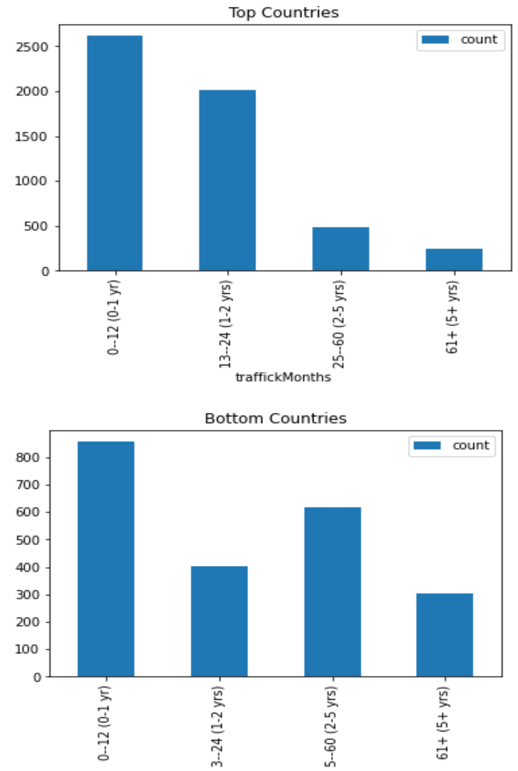


Figure 4 – Comparing time trafficked for top and bottom countries.

Research

After the initial findings we decided to conduct even more research to start understanding what the experts are saying about Human Trafficking:

- Research has shown that human trafficking is one of the fastest growing forms of transnational crime. The clandestine nature of trafficking makes it much harder to get evidence of actual trends happening worldwide. This project will attempt to create trends based on the dataset we discovered.
- Traffickers take advantage of the unequal status of women and children in disadvantaged countries and capitalize on the cheap labor and the promotion of sex labor in some countries.
- About half of all humans trafficked are under the age of 18.
- Other big populations that are at risk of being trafficked are runaways and homeless youth as these groups don't have a strong sense of family in their normal lives.
- Within the next 10 years human trafficking is expected to surpass drug and arms trafficking in its incidence.
- The annual revenue exchanged from human trafficking is projected to total to 32 billion US dollars.

This research proves the importance of this problem in the current time and why there should be a greater focus on it.

Models Proposed

The development of the following models, we believe, serves a greater cause than just the requirements listed within this course. The main reason we decided to take on a more sensitive topic was due to the overall threat it poses to many daily. However, this threat remains to be actively addressed on a national, and especially a global, scale. Our research can hopefully shed an ounce of light on how impactful human trafficking is on everyone and yet still purvey the message that this is only partial for what has been unfolded for the entire story.

With our models we can potentially represent the target associated with the most attacks and shed awareness for those who may be in sight of an attacker. Through our clustering model we can easily visualize the groupings of attack by specific dimension settings. With the logistic regression model, we can estimate the likelihood of human trafficking in specific populations. Finally, we can set a time series forecast going into 3 years past our latest data, all combine to help law enforcement and anti-trafficking organizations allocate resources effectively to prevent and combat human trafficking. By identifying potential hotspots and trends, authorities can focus their efforts on areas where trafficking is likely to occur or increase. Having all these tools at our disposal there can be a stronger fight against human trafficking.

Prophet Model

Based on our initial research, our primary focus was to create a time series analysis via Prophet. The output is not favorable for extensive research to be conducted due to the date attribute only being present at the year dimension. We were able to notice this with how inputting of the dat only yielded trend lines relatively and not true seasonality changes over each year. This would have been more interesting if the data points spanned across more than just ~20 years/data points. We were able to dive deeper into the model with specific trend lines (ie gender trend lines for both Male and Female categories), yet again the output was rudimentary. Going forward we do believe the data is still useful, but due to sample size we do not recommend forecasting segments. We can segment our forecast with the results from the Logit models.

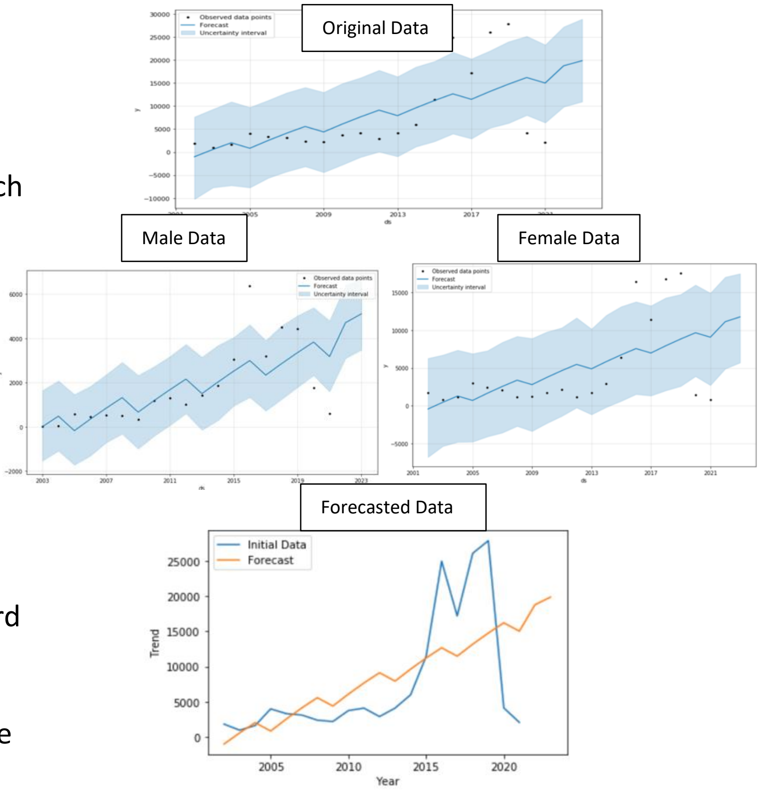
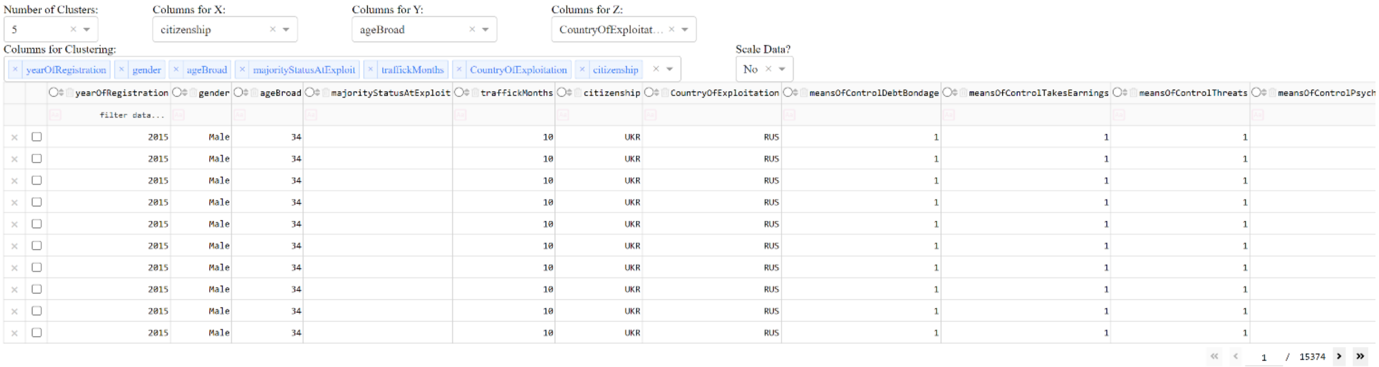


Figure 5 – Prophet model results for all data, only Male, only female, and final forecast

Clustering Model


Number of Clusters: 5
Columns for X: citizenship
Columns for Y: ageBroad
Columns for Z: CountryOfExploitat...
Columns for Clustering: yearOfRegistration, gender, ageBroad, majorityStatusAtExploit, traffickMonths, CountryOfExploitation, citizenship
Scale Data? No

		yearOfRegistration	gender	ageBroad	majorityStatusAtExploit	traffickMonths	citizenship	CountryOfExploitation	meansOfControlDebtBondage	meansOfControlTakesEarnings	meansOfControlThreats	meansOfControlPsych
		filter data...										
×	□	2015	Male	34		10	UKR	RUS	1	1	1	
×	□	2015	Male	34		10	UKR	RUS	1	1	1	
×	□	2015	Male	34		10	UKR	RUS	1	1	1	
×	□	2015	Male	34		10	UKR	RUS	1	1	1	
×	□	2015	Male	34		10	UKR	RUS	1	1	1	
×	□	2015	Male	34		10	UKR	RUS	1	1	1	
×	□	2015	Male	34		10	UKR	RUS	1	1	1	
×	□	2015	Male	34		10	UKR	RUS	1	1	1	
×	□	2015	Male	34		10	UKR	RUS	1	1	1	
×	□	2015	Male	34		10	UKR	RUS	1	1	1	

Figure 6 - Cluster model parameters selection, and data table. You can also filter with the column headers.

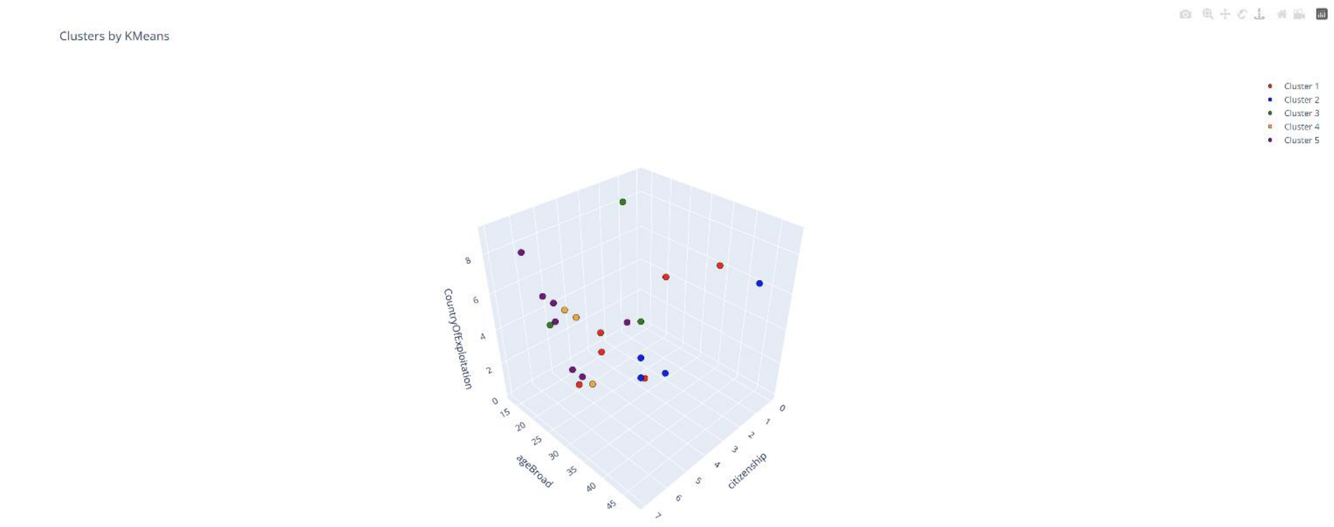


Figure 7 - Clustering model running. This is the graph output after setting the dimensions.

Data Cleansing:

With this model, our focus centered around the capability to generate a clustering model based on an ever-changing user input. Before getting into the actual clustering model, we had to conduct data cleansing for string valued columns that could be easily interpreted into float values. This was done specifically for the columns 'traffickMonths' and 'ageBroad'. For 'traffickMonths' we removed all string values via Regex based on the following pattern: 'r\d+'. This pattern simply states to grab the start of the digit all the way to the end of the values indices. For 'ageBroad' we made an even simpler cleaning process which split the string value based on '--' and summed the values on both sides of the delimiter then divided by the count of all values. This returned to us the average of the column in now float format. For all remaining columns except, 'citizenship', 'CountryOfExploitation', 'gender', and 'majorityStatusAtExploit', we then casted the columns as float type.

Front End Build:

The next phases of the model dealt with actual dashboard front development. First, we decided to utilize the *Plotly* library to do all the visualization lift and application interface. This dealt with us needing to define all slicers which would have an impact of a typical clustering model. The slicers included 'Number of Clusters', 'X', 'Y', 'Z', 'Scale Data?', and 'Columns for Clustering'. The only nuance that must be mentioned is that 'X', 'Y', and 'Z' must be present with 'Columns for Clustering', but there can be more columns included within this selection to possibly provided variety to a new way of clustering. Also, we included a table which can be interacted upon as more of a data specific filtration method. An example would be based on trying to only surface "USA" through filtering the 'citizenship' column with that specific value.

Back End Build:

For the backend portion of the dashboard, we primarily just relied upon the selections chosen through the above filters. Upon selection, the slicer values were then read into our application and pushed into the **update_clusters** function. This function is used to update the visualization based on the inputs placed into the application. When read into the function the data is then manipulated first through the dropping of all nulls values for the specific columns received from 'Columns for Clustering'. With all the columns now reprised of all null values, the columns are then determined if needing to be label encoded primarily if it is not a float column data type. Dependent upon the 'Scale Data?' filter, the data is then pushed through the StandardScaler() function or simply passes without standardizing. Finally, the last portion is to create a 2-D map of the selected 'X', 'Y', and 'Z' columns to then pass into KMeans() function which is determined to be by the 'Number of Clusters' filter. Thus, the result is displayed through go.Figure(), specifically go.Scatter3d(). To utilize this application upon run, go to http://127.0.0.1:8050/.

Logit Model for global estimates of diverse types of slavery

Despite diligent efforts in collecting data, which includes interviewing individuals about their family networks, there is a lack of dependable information on the entire population affected by modern slavery.

Survey data falls short in capturing enough cases when estimating certain populations, specifically children in modern slavery and individuals subjected to forced commercial sexual exploitation. Interviewing a representative sample of these groups poses challenges due to ethical concerns surrounding interviewing children and the sensitive nature and stigma associated with forced commercial sexual exploitation and the modern slavery of children.

Consequently, while the survey data is considered sufficiently reliable for measuring adult forced labor exploitation and forced marriage, it is not deemed sufficiently reliable for measuring forced commercial sexual exploitation and the exploitation of children through forced labor. To complement the survey data, we want to use **odds ratio** to extrapolate the prevalence of these hidden populations from the CTDC dataset.

To calculate the likelihood of experiencing forced commercial sexual exploitation compared to forced labor exploitation, we utilize a logit model that examines the correlation between the type of exploitation (forced commercial sexual exploitation versus forced labor exploitation) as the dependent variable (Y) and factors such as gender, majority status, and the interaction between gender and majority status as the independent variables (X).

Data preparation

For this analysis we had to prepare the data to match the data available in the surveys. Surveys only contain data for male and female victims and for adults and minor.

Our dataset contains several types of genders and majority status. We limited the scope to male and female and to minor and adult to match the data available in the surveys.

For the dependent variable we transformed the null values (NaN) in the isSexualExploit column to 0 and 1 to be able to fit the Logistic Regression correctly.

Logit Model

The model that we are proposing is this one:

$$\ln\left(\frac{p}{1-p}\right) = b0 + b1 * gender + b2 * \ majorityStatusAtExploit +$$

$$b3 * gender * \ majorityStatusAtExploit$$

After fitting the model we obtained the following output:

For gender we have two levels: female and male. For this model female is the base case. For Majority Status we also have two levels: Adult and minor/ For this model Adult will be the base case. For the interaction term, a registry matching female and adult will be the base case. Following this, we can create a summary table like the one below:

						(b0+b1+b2+ b3)	p/1-p
Genre	**Majority Status at Exploit**	**b0**	**b1**	**b2**	**b3**	**ln(p/1-p)**	**Odds ratio**
Female	Minor	0.19016		1.89411		2.08427	8.038721
Female	Adult	0.19016				0.19016	1.209443
Male	Minor	0.19016	-4.68387	1.89411	3.19678	0.59718	1.816988
Male	Adult	0.19016	-4.68387			-4.49371	0.011179

Chart 1 – Logit model results for model expressed above.

In the last column we are reporting the obtained Odds Ratio, if the ratio is greater than 1 means that a victim matching that genre and majority status is more likely to fall to forced commercial sexual exploitation than forced labour exploitation. These are Minor females. Adult females, and Minor males. The only ones that are more likely to fall in forced labour exploitation are Adult males.

This means that for Men (Adult, male) the odds of being a victim of forced commercial sexual exploitation are 0.011 times lower than being a victim of forced labour exploitation, which is the same as saying that for every 100 men who were victims of forced labour, it is likely that 1 of them were forced to commercial sexual exploitation.

These ratios will help us to estimate the forced commercial sexual exploitation of e.g adult males and females by using their corresponding odds ratio.

From a survey shared by the International Labour Organization (ILO) and the International Organization for Migration (IOM) in 2022 they state that Men (Adult, male) forces to labour in Thousands are equal to 10,656 and Women (Adult, female) 5,361. By applying their odds ratio to fall in forced commercial sexual exploitation (0.011179) we get that approximately 120 men will be forced to do so.

Other great analysis that can be obtained from this logit model is how male stack vs female to be sexually exploit. By applying the next calculation **(exp(model1$coefficients[-1])-1) * 100** to each coefficient obtained in the model we can obtained the odds of being sexually commercialized.

This means that the male's odds of being sexually commercialized are 99.07 smaller than adult female's odds of being sexually commercialized. If you are a minor, you have 564.66 greater odds of being forced to commercial sexual exploitation and so on.

After running multiple Logit models (can be found in the Appendix) we obtained the next probabilities:

Victim type	**Sexual Exploitation**	**Forced Labor**	**Recruiter relation: Intimate Partner**	**Recruiter relation: Friend**	**Recruiter relation: Family**
Girl (minor, female)	88.94%	7.38%	21.72%	9.78%	72.85%
Woman (adult, female)	54.74%	28.59%	60.94%	35.88%	2.72%
Boy (minor, male)	64.50%	26.19%	0.06%	0.26%	72.85%
Man (adult, male)	1.11%	79.99%	0.35%	99.54%	2.72%

Chart 2 – Probabilities obtained by running multiple logit models to determine the probabilities of being sexually exploited, forced to labor, probabilities of being recruited by an intimate partner, a friend, or a family member.

The odds and probabilities obtained from these Logit models can be extrapolated with the forecast from the prophet model or with surveys from official entities to segment the data.

Concluding thoughts

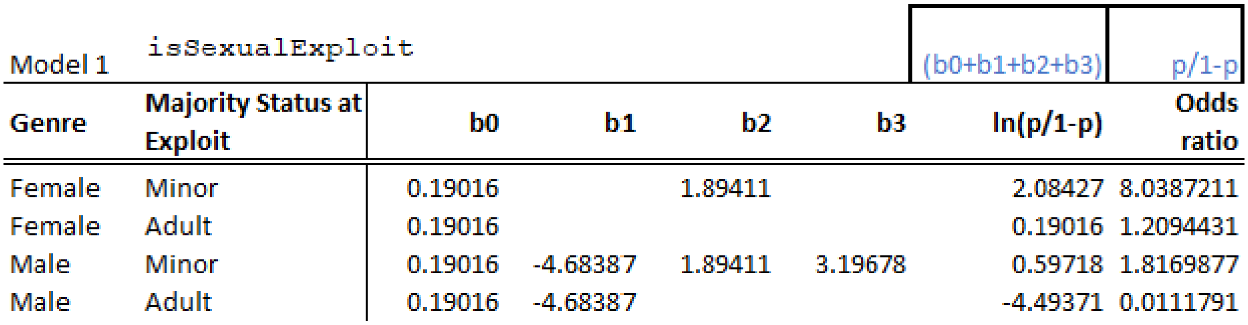
From our insights gathered through the multiple Logit models, we were able to get a clear breakdown of affected populations, broken down by gender, hard labor, and relation to the victim – as a result, we have been effectively able to answer our research questions Our models supported our exploratory data analysis, with females being much more likely to be trafficked in all categories except forced labor compared to males. Also, minors were much more likely to be trafficked with supported our initial findings with the 9-17 age group having the highest volume. Looking into the impact of the analysis, various global groups, like the ILO and IOM (as referred above) can allocate funds accordingly to better protect groups at highest risk.

Works Cited

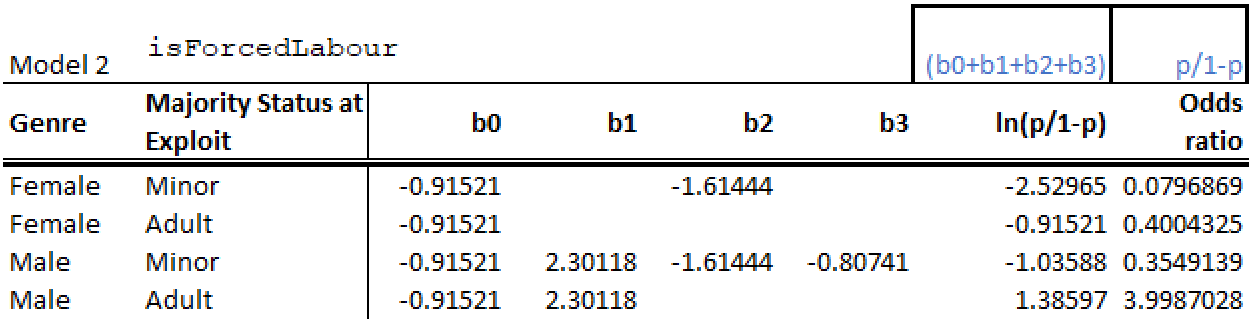
- Weitzer, R. (2014). New Directions in Research on Human Trafficking. *The ANNALS of the American Academy of Political and Social Science*, *653*(1), 6–24. https://doi.org/10.1177/0002716214521562
- Clawson, Heather J., et al. "Human trafficking into and within the United States: A review of the literature." *Washington, DC: Office of the Assistant Secretary for Planning and Evaluation, US Department of Human and Heath Services. Retrieved December* 25 (2009): 2009.
- Wheaton, Elizabeth M., et al. “Economics of Human Trafficking.” *International Migration*, vol. 48, no. 4, 2010, pp. 114–141, https://doi.org/10.1111/j.1468-2435.2009.00592.x.

Appendix

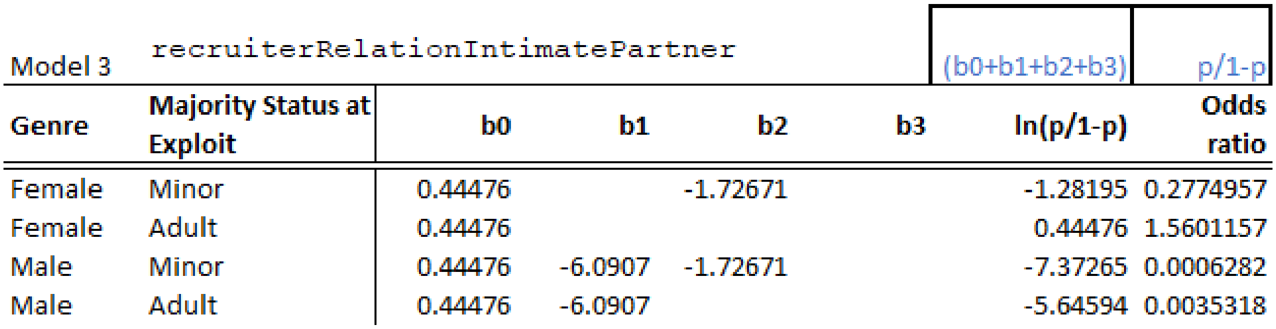
Tables with the multiple logit models results.



Model 1	isSexualExploit					(b0+b1+b2+b3)	p/1-p
Genre	**Majority Status at Exploit**	**b0**	**b1**	**b2**	**b3**	**ln(p/1-p)**	**Odds ratio**
Female	Minor	0.19016		1.89411		2.08427	8.0387211
Female	Adult	0.19016				0.19016	1.2094431
Male	Minor	0.19016	-4.68387	1.89411	3.19678	0.59718	1.8169877
Male	Adult	0.19016	-4.68387			-4.49371	0.0111791



Model 2	isForcedLabour					(b0+b1+b2+b3)	p/1-p
Genre	**Majority Status at Exploit**	**b0**	**b1**	**b2**	**b3**	**ln(p/1-p)**	**Odds ratio**
Female	Minor	-0.91521		-1.61444		-2.52965	0.0796869
Female	Adult	-0.91521				-0.91521	0.4004325
Male	Minor	-0.91521	2.30118	-1.61444	-0.80741	-1.03588	0.3549139
Male	Adult	-0.91521	2.30118			1.38597	3.9987028



Model 3	recruiterRelationIntimatePartner					(b0+b1+b2+b3)	p/1-p
Genre	**Majority Status at Exploit**	**b0**	**b1**	**b2**	**b3**	**ln(p/1-p)**	**Odds ratio**
Female	Minor	0.44476		-1.72671		-1.28195	0.2774957
Female	Adult	0.44476				0.44476	1.5601157
Male	Minor	0.44476	-6.0907	-1.72671		-7.37265	0.0006282
Male	Adult	0.44476	-6.0907			-5.64594	0.0035318

Model 4	`recruiterRelationFriend`					(b0+b1+b2+b3)	p/1-p
Genre	**Majority Status at Exploit**	**b0**	**b1**	**b2**	**b3**	**ln(p/1-p)**	**Odds ratio**
Female	Minor	-0.58064		-1.64087		-2.22151	0.1084452
Female	Adult	-0.58064				-0.58064	0.5595401
Male	Minor	-0.58064	5.95128	-1.64087	-9.67257	-5.9428	0.0026247
Male	Adult	-0.58064	5.95128			5.37064	215.00042

Model 5	`recruiterRelationFamily`					(b0+b1+b2+b3)	p/1-p
Genre	**Majority Status at Exploit**	**b0**	**b1**	**b2**	**b3**	**ln(p/1-p)**	**Odds ratio**
Any	Minor	-3.5771		4.5639		0.9868	2.6826363
Any	Adult	-3.5771				-3.5771	0.0279567