**Predicting Worldwide Refugees**

Tejomay Gadgil (tg1906), Yash Gupta (yg1057)

Gina Holden (gh1407), Michael Stanley (mhs592)

Business Understanding

The United Nations High Commission for Refugees (UNHCR) is the worldwide authority for refugee aid. One of the main challenges they face is *resource allocation*. Resources available for processing and resettling refugees is extremely limited, and therefore every dollar counts. The UNHCR is responsible not only for supplying housing, infrastructure, transportation, food, and resources to refugee camps, but also for giving refugees support and resources for integration into their new culture. Unfortunately, there is never enough aid to support everyone. Therefore, any *improved ability to forecast* *the number of refugees* from any given country in the upcoming year would massively benefit the UNHCR’s ability to plan and allocate their limited resources to support refugees in the most efficient manner.

The UNHCR oversees processing of refugees into a third country. For example, a Burmese refugee who has fled to Thailand may be brought forward as a candidate for resettlement in the United States. These third countries — typically the United States, Canada, Australia, and Sweden — are limited in their resources to interview and admit refugees. The UNHCR coordinates the presentation of applicants and schedules interviews with the admitting third countries. Therefore, any increased ability to forecast the number of applicants from each country will help the UNHCR to more efficiently plan the trips that the United States, and other countries admitting refugees, take to Africa, Middle East, etc. The trips that United States officials go on are for interviewing and making a decision on applicants. It is clear that increased efficiency in interview and trip planning will allow the admitting countries to admit as many applicants as is politically possible.

Currently, UNHCR workers forecast applicants using a number of crude heuristics and human judgement calls. From domain experience (Gina works with United States Citizen and Immigration Services), we understand their main predictor to be the *applicant count* *from the previous year*. While this is not a bad start, we believe this model is too simplistic. Specifically, we believe we can improve on this by incorporating *geopolitical measures* such as economic conditions, and measures of civil and political unrest. We believe by combining these measures with the applicant count from the previous year, we will be able to paint a clearer *world picture*. This will allow us to train machine learning models which better understand the country conditions which drive the number of refugees from a given country.

Data Understanding

An instance in our model is one origin country for a given year. Our target feature, the number of applicants from each country in each year, is provided by the UNHCR (http://popstats.unhcr.org/en/asylum\_seekers).

Our feature data, the unique aspect of our project, combines a few sources of geopolitical indicators with the number of applicants from the previous years. We found economic and demographic data from the World Bank, who provide GDP per capita, life expectancy, and unemployment per year for each country.

The second group of geopolitical indicators from the Global Database of Events, Language, and Tone (GDELT) event database (<https://www.gdeltproject.org/data.html>). They maintain an extensive record of events in countries around the world. Through some cursory analysis, we found correlations for a few columns in the GDELT events to our feature variable: Goldstein Score, (<http://web.pdx.edu/~kinsella/jgscale.html>), AvgTone, and QuadClass. Since refugee claims must display a reasonable fear of returning to their home country, and this fear has to fall under a protected class (religion, tribe, race, etc.) we believe any data related to political, social, or economic strife would improve our model.

Our feature columns are from the year previous to the year of the target data. For a given country and year, a row in the feature data will have the following columns (plus a few more): GDP, Population, Country, Count Applied Previous Year. For example, for predicting the number of refugees that applied to the UNHCR from Pakistan in 2017, the features would include observations from Pakistan in 2016.

Data Preparation

Our target data from UNHCR specifies the number of applicants per year from 224 countries from 2000 to 2018. Applicant counts are not available for every country-year: 275 (6.8% of the 4,032 possible samples) country-years were missing data.

Using two country codes standards, FIPS and ISO, we were able to merge this applicant data with the GDELT database, which contains country-specific measures of geo-political instability. We selected GDELT features via correlation calculations, ending up with Goldstein Scale, Avg Tone, Num mentions, and Quad Class. These features capture the tone of events in each country per year, ranging from conflict-inducing to cooperative. Since records in GDELT are events that happened in every country (GDELT contains over 11 million records for 2000-2018), we had to aggregate that data in useful ways to generate usable features for our country-year samples. We created a variety of thresholded counts and averages, described more thoroughly in Appendix B. There is an opportunity for further feature engineering work here: weighting events by severity, considering when in the year events occurred, or filtering specific types of events. The GDELT data was missing 40 country-years at this step, causing our final dataset to lose some more data.

Next, we incorporated economic data from the World Bank. This took a fair amount of translating and cleaning. First, the .csv files that were available for download were separate for each feature and required merging. So we had to expand that data out to fit the year and country structure of the target and GDELT data. Moreover, many values were missing and had to be imputed. For GDP, we filled in the value using the closest years GDP data for that country. For the missing values for unemployment and life expectancy, there were no values at all for the countries with missing values and thus we were not able to use the same cleaning strategy as the GDP data. Instead we filled in missing values by taking the average across all countries for that year. This cleaned economic data was merged in by country code and year to create additional feature columns to our training data. As some of the countries in the economic data did not map to our country codes (Tibet, for example), we lost 68 country-year samples on this merge.

The target variable is the number of refugee applicants per origin country for a given year. For example, we are in December of 2019 the UNHCR is (last minute) planning for 2019. The UNHCR would like to predict the number of applicants they will have from each country in 2019.

After merging all of the data together, we saw that our linear models had non-constant variance. The variance had a cone shape, increasing as the target variable increased. Our model was not as good at predicting applicant count when they are very large, and cross-validation performance sporadically changed depending on how the outliers were allocated. We transformed our target variable to ln(y) which improved the constant variance as well as R^2. This transformation also corrected the skew, as is apparent in Figure 2 of Appendix A.

Since we are doing regression, we want to be careful not to include any high-leverage and high variance observations in our training data as these points could significantly skew our model. We found a couple of high leverage points but we think they are fine to keep because they are also low variance. We have one or two extremely high variance points that are actually fairly high leverage as well so we removed those. Next, we looked at removing features that were collinear, in the hopes of improving generalizability. We used the variance\_inflation\_factor method from statsmodels package to remove redundant features. This trimmed our feature set down by about half.

We additionally performed the following feature engineering:

* **Logarithm transform to remove feature skew**: Population, unemployment, and GDP per capita are highly skewed features. Transforming them via the natural logarithm generates a feature that more closely resembles a normal distribution. See Figure 2 in Appendix A.
* **Binning of current population**: National populations vary by three orders of magnitude, so we created three bins: under 4 million, between 4 million and 40 million, and above 40 million. The thresholds of these bins were chosen by observing inflection points in the population distribution.
* **Goldstein Net**: Goldstein\_Pos\_X (and Goldstein\_Neg\_X ) is a count of the events in the GDELT database with a Goldstein score above X (below -X). Goldstein\_Net\_X is the difference Goldstein\_Pos\_X - Goldstein\_Neg\_X. Intuitively, we might expect positive and negative events to “cancel each other out” in their effect on applicants: a conflict starting might be negated by a conflict ending, for example.

Modeling and Evaluation

Predicting applicant counts is a regression problem. The problem could be shaped into a classification problem by, for example, bucketing countries by the number of applicants, but the business problem requires a precise, granular prediction of the count of applicants, so classification would be less valuable.

For this business case, performance is more important than explainability, so we are free to consider any regression model that generalizes well. Given that our dataset contains fewer than 5,000 records, we did not attempt to train any neural network models, or other models dependent on larger training datasets. We trained and optimized a variety of models, primarily focused on linear and ensemble models:

* Linear Regressor
* Elastic Net Regressor
* Decision Tree Regressor
* Random Forest Regressor
* GradientBoostingRegressor
* XGBoost Regressor

These models were evaluated and compared as described in the next section.

Metrics and Model Evaluation

We have used R2 as the primary performance metric because this is a classic regression problem. Underestimating and overestimating applicant counts per country have roughly the same “cost” in terms of business value, so there is no preference for the direction of prediction error that would need to be accounted for in the metric. R2 measures the proportion of variance in the target variable that the model captures, which aligns with the task of predicting the number of applicants per country per year.

We held out 2017 as our test set, but used a random split of the remaining years for training and cross validation. Making predictions for the most recent year of data most closely reflects the business problem our model would face in deployment. While the data is inherently time-based, we have defined our features as point-in-time: the value of that feature *as of the relevant year*, to avoid any leakage across years. Because we considered so many different models and combinations of hyperparameters, we chose our specific model with a cross-validation set before applying it to the 2017 test data.

Baseline

Our focus in this project was to bring 1) disparate, external sources and 2) advanced machine learning models to bear on the problem of predicting asylum seeker applicants. To understand the impact of both concepts, we built a baseline model using the same algorithms tested for our model but utilizing only previous year applicants as a feature. We do not have evidence that UNHCR or others have ever built any model more complex than linear regression, but we considered other models in our baseline to separate the benefits of model complexity from the rich feature set.

Like the final model, the baseline model was trained on 2000-2016 data and tested on 2017 data, with the following results:

|  |  |
| --- | --- |
| **Baseline Models** | **Test R2** |
| Linear Regressor | 0.419 |
| Elastic Net Regressor | 0.367 |
| Support Vector Regressor (rbf kernel) | 0.582 |
| Decision Tree Regressor | 0.545 |
| Random Forest Regressor | 0.676 |
| GradientBoostingRegressor | 0.791 |
| XGBoost Regressor | 0.792 |

We consider the Linear Regression (**R2** = 0.419) to be the true baseline of predictive work in the field, but the benefits of model complexity (even without adding features) are readily apparent in the enormous performance improvement gained from moving to tree-based ensemble models.

The training-cross-validation-test framework described above is robust to overfitting and does not give preference to any particular type of model, so our evaluation framework is simply to optimize for R2 through feature engineering, model selection, and hyperparameter tuning. We would consider any model that exceeds the baseline R2 to be successful.

Results and Final Model

We optimized the hyperparameters of each of the aforementioned regression models, resulting in the following performance on the training, cross-validation, and 2017 test set:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Selection** | **Train R2** | **CV R2** | **2017 Test R2** |
| Linear Regressor | 0.759 | 0.665 | 0.702 |
| Elastic Net Regressor | 0.690 | 0.648 | 0.634 |
| Support Vector Regressor (rbf kernel) | 0.718 | 0.684 | 0.665 |
| Decision Tree Regressor | 0.999 | 0.879 | 0.912 |
| Random Forest Regressor | 0.992 | 0.936 | 0.932 |
| Gradient Boosting Regressor | 0.975 | 0.946 | 0.935 |
| XGBoost Regressor | 0.964 | 0.943 | 0.930 |

The Gradient Boosting Regressor was chosen as the best performing model during cross-validation (we show the performance of each optimized model type on the 2017 test data for comparison). The Gradient Boosting Regressor clearly outperforms the baseline model, with an R2 of 0.935 on 2017 data compared to the baseline R2 range of 0.419-0.792, indicating that the additional features and model complexity generate substantial performance benefits. This magnitude of improvement in performance means that granular, asylum seeker prediction should improve and the UNHCR should be able to better optimize its global resource allocation.

The results also indicate that tree-based ensemble models outperform linear models for this problem. Intuitively, it is reasonable that a number of the features in the dataset have non-linear relationships with applicant count. For example, GDP per capita is the second most important feature for the GBR model, but one would imagine that the difference in GDP per capita between the United States and Canada does not impact the applicant count from those countries much, if at all. Instead, we would imagine that there are minimal applicants from any country above some GDP per capita threshold. Similar logic can be applied to population, unemployment, and other features in the data, and these non-linear relationships are much more easily captured in ensemble models than in linear models. Appendix B provides both feature descriptions and feature importance from the final Gradient Boost Model

Deployment

The proposed deployment of the model by the UNHCR will be at the end of 2019 to predict the number of refugee applicants for the year 2020. The information from the model will help UNHCR better understand where to put their resources. The model will be run once a year and will be retrained every year by adding the previous year’s data. The current models r squared will be used as the baseline model to compare results. Each year, the models R-squared would be compared to previous years’ and the baseline model’s R-squared to see if anything has drastically changed or affected the model. As the model is only used once a year due to data update schedules during the year, the model would only be evaluated after a year once the actual applicant number data is available for that year. Since there is no data available per month or per quarter, the ability to monitor the model during the year is unavailable.

One of the ethical considerations is for the UNHCR to generalize geographical areas as high applicant areas or low applicant areas based on the model’s predictions. For example, if the model is predicting Afghanistan as high in applicants for the next year, the UNHCR should not generalize the middle east as a high applicant geographical area. Each country should only be representative of itself. Another example is if due to some improvements or changes in a country, the number of applicants drastically change. Each year might be different and thus the model should not be used to create a generalized idea about a country’s applicants. It might be very low one year and the very next year it might be much higher.

One of the major risks with this model due to the scarcity of day to day data is the ability to continuously evaluate and monitor the model rather than evaluate it at the end of the year when the years data is available. The only way to mitigate this issue is to be able to collect data on a more regular basis such as monthly or quarterly.

Sources

“The GDELT Project.” *GDELT*, Dec. 2019, [www.gdeltproject.org/data.html](http://www.gdeltproject.org/data.html).

“GDP (Current US$).” *Data*, Dec. 2019, data.worldbank.org/indicator/NY.GDP.MKTP.CD.

“Life Expectancy at Birth, Total (Years).” *Data*, Dec. 2019, data.worldbank.org/indicator/SP.DYN.LE00.IN.

“Unemployment, Total (% of Total Labor Force) (Modeled ILO Estimate).” *Data*, Dec. 2019, data.worldbank.org/indicator/SL.UEM.TOTL.ZS.

United Nations. “Data.” *UNHCR*, Dec. 2019, www.unhcr.org/en-us/data.html.

Appendix A - Figures

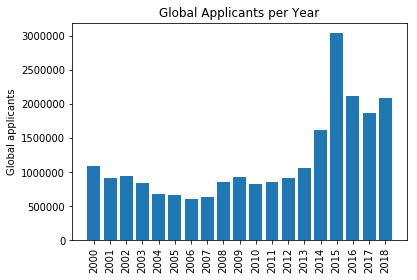


Figure 1: Global Applicants per year. Source: UNHCR

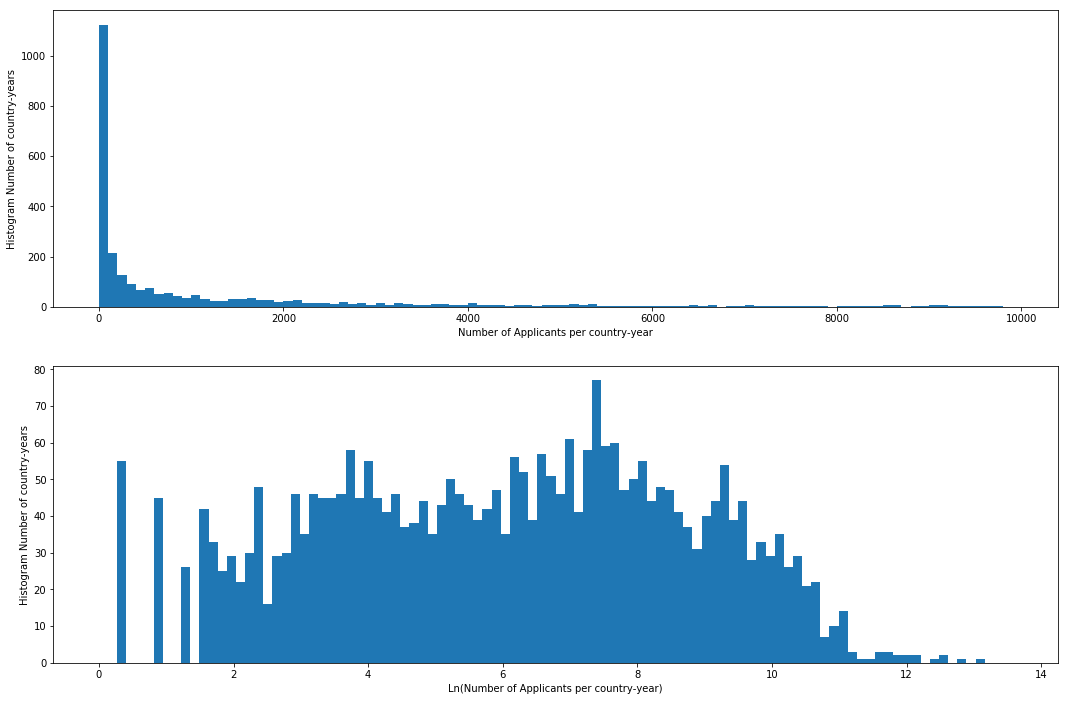


Figure 2: Histograms of target variable and log of target variable: Applicants per country per year. Source: UNHCR

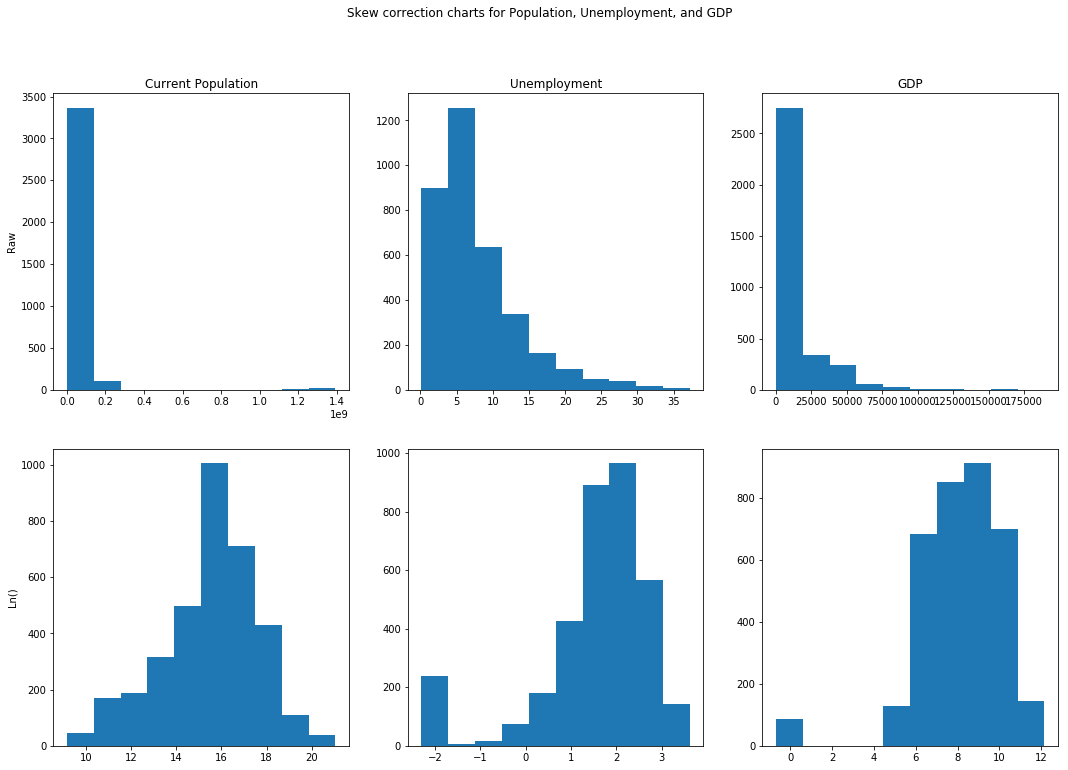


Figure 3: Skew correction: Histogram of features and log of features for Current Population, Unemployment, and GDP Per Capita

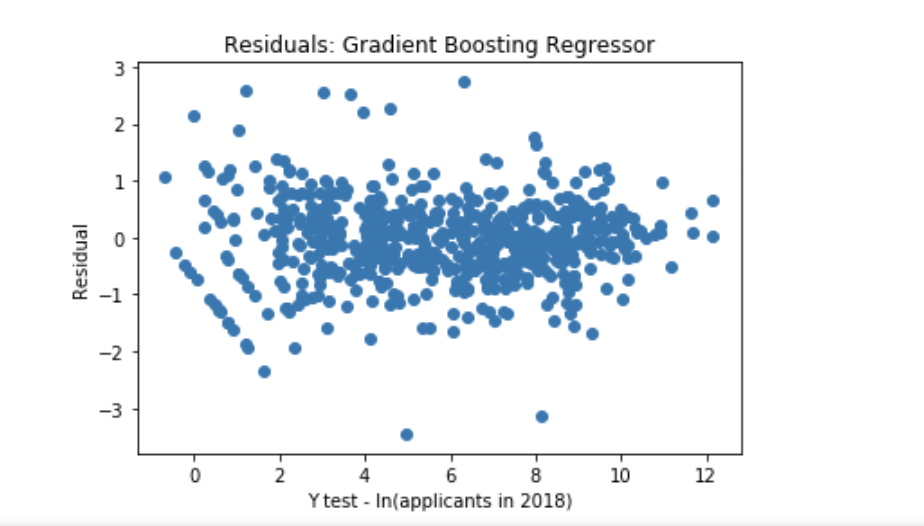


Figure 4: Residuals vs. target variable for Gradient Boosting Regressor (final model)

Appendix B - Feature Description and Importance

*Note: All features are for the specific country in the year immediately preceding the prediction year. Additional features were evaluated, but removed for lack of explanatory power or severe collinearity.*

UNHCR - Asylum Seeker Applicants

**'Applied during year'** - The number of asylum seeker applicants from the country in the previous year.

WorldBank - Economic Factors

**'Med\_pop'** - Binary variable that is 1 if the country’s population was between 4 million and 20 million, 0 otherwise.

**'Large\_pop'** - Binary variable that is 1 if the country’s population was greater than 20 million, 0 otherwise.

**'GDP Per Capita'** - Gross domestic product per capita.

**'Log\_gdp\_per\_capita'** - Natural log of ‘GDP Per Capita’.

**'Log\_unemployment'** - Natural log of unemployment rate.

**'Life expectancy'** - Life expectancy.

GDELT - Event Database

*Additional GDELT documentation available here:* [*http://data.gdeltproject.org/documentation/GDELT-Data\_Format\_Codebook.pdf*](http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf)

**'NumMentions'** - The average number of mentions across all source documents for all events in the country in the previous year.

**'AvgTone'** - GDELT’s “AvgTone” metric scores the tone (positive or negative) of mentions of an event in all sources. AvgTone is the average of that metric across all events.

**'Extreme\_Pos\_Tone\_Events'** - The count of events in GDELT database with AvgTone greater than 10 (considered extremely positive).

**'Extreme\_Neg\_Tone\_Events'** - The count of events in GDELT database with AvgTone < -10 (considered extremely negative).

*Note: Goldstein Score is a measure of the “theoretical potential impact” of each event, scored from -10 (extremely negative impact) to 10 (extremely positive impact).*

**'Goldstein\_Neg\_5'** - The count of events in GDELT database with Goldstein score < -5.

**'Goldstein\_Pos\_5'** - The count of events in GDELT database with Goldstein score > 5.

**'Goldstein\_Pos\_7'** - The count of events in GDELT database with Goldstein score > 7.

**'Goldstein\_Pos\_8'** - The count of events in GDELT database with Goldstein score > 8.

**'Goldstein\_Pos\_9'** - The count of events in GDELT database with Goldstein score > 9.

**'Goldstein\_Net\_5'** - The difference: 'Goldstein\_Pos\_5' - 'Goldstein\_Neg\_5'

**'Event\_Code\_19\_20'** - The count of events in GDELT under Cameo code 19 (“Fight”) or 20 (“Use Unconventional Mass Violence”).

Feature Importance

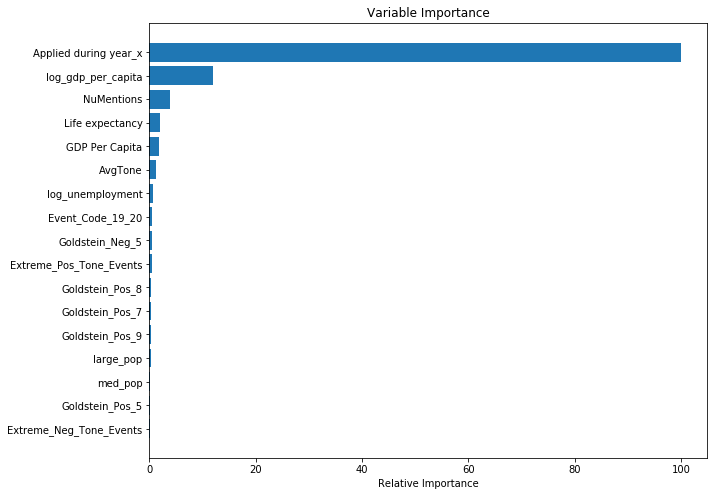


Figure 5: Feature Importance from Gradient Boosting Regressor (final model)

Appendix C - Team Contributions

Tejomay Gadgil (tg1906) - Data cleaning, merging data, editing final paper

Yash Gupta (yg1057) - Economic data, data cleaning, final paper

Gina Holden (gh1407) - Problem statement, gdelt data, feature transforming and selection, final paper

Michael Stanley (mhs592) - UNHCR data, feature engineering, baseline model definition, hyperparameter tuning, final paper