Transnational Education Review
May 2023
Volume: 1, No: 1, pp. 3 - 16
ISSN: 2753-8656 (Print) | ISSN: 2753-8664 (Online)
journals.tplondon.com/ter

TRANSNATIONAL PRESS® LONDON

Received: 15 November 2022 Accepted: 14 April 2023
DOI: https://doi.org/10.33182/ter.v1i1.2883

Proposing Location-based Predictive Features for Modeling Refugee Counts

Esther Ledelle Mead[1], Maryam Maleki[2], Mohammad Arani[3], and Nitin Agarwal[4]

Abstract

Machine learning models to predict refugee crisis situations are still lacking. The model proposed in this work uses a set of predictive features that are indicative of the sociocultural, socioeconomic, and economic characteristics that exist within each country and region. Twenty-eight features were collected for specific countries and years. The feature set was tested in experiments using ordinary least squares regression based on regional subsets. Potential location-based features stood out in our results, such as the global peace index, access to electricity, access to basic water, media censorship, and healthcare. The model performed best for the region of Europe, wherein the features with the most predictive power included access to justice and homicide rate. Corruption features stood out in both Africa and Asia, while population features were dominant in the Americas. Model performance metrics are provided for each experiment. Limitations of this dataset are discussed, as are steps for future work.

Keywords: *Data Science; Machine Learning; Predictive Modeling; Refugee Crisis*

Introduction

Refugee crisis situations have become increasingly common, and an ability to predict these situations can give governments and humanitarian organizations the power to better prepare to provide humanitarian aid and, ultimately, prevent further crises. A viable machine learning model to predict the number of migrants anticipated to be on the move via land or sea can enable humanitarian groups to better prepare for administering aid and security. According to the United Nations High Commissioner for Refugees (UNHCR), "79.5 million people had been driven from their homes across the world at the end of 2019" [1]. Twenty-six million of those displaced people were designated as "refugees" by the UNHCR, meaning that they were actually forced to leave the boundaries of their home country due to "violence or persecution". In this study we aim to determine whether a specific set of sociocultural, socioeconomic, and economic features can be used to predict refugee counts by country of origin. This research builds upon the work of Mead et al. [2] who set the stage for developing a proactive approach to refugee crisis management by identifying some of the numerous motivational factors that lead to refugee migration and attempting to use those variables to build a viable machining learning model to predict refugee crisis situations. This work is

[1] Esther Ledelle Mead, Southern Arkansas University, United States. E-mail: esthermead@saumag.edu

[2] Maryam Maleki, California State University, United States. E-mail: maleki@mail.fresnostate.edu

[3] Mohammad Arani, University of Arkansas at Little Rock, United States. E-mail: marani@ualr.edu

[4] Nitin Agarwal, University of Arkansas at Little Rock, United States. E-mail: nxagarwal@ualr.edu

motivated by our understanding that each region has varying availability regarding resources to meet human needs, and additional varying characteristics regarding sociocultural and socioeconomic factors. Due to these variations, we propose an analysis of predictive features based on region to determine whether there are any significant differences between them. The key contributions of this work include the additional advancement of scientific studies to predict refugee counts, and a continued analysis of a broader range of predictive features such as those relating to violence, human rights, economic well-being, and personal security. The remainder of this work is organized as follows: Section 2 reviews the extant literature related to attempts to model refugee counts based on various predictive features. Section 3 discusses the data collection and preprocessing methodology and describes the dataset and features used in the experiments conducted in this work. Section 4 provides a discussion of the results and analysis of the experiments. Section 5 provides the overall conclusions and limitations of this research, as well as a discussion of our ideas for future work.

Literature Review

Previously published studies relevant to the problem of modeling refugee counts primarily discuss early attempts made by researchers, a realization of the need for a broader range of proposed predictive features, and a need for a better understanding of the changing societal landscape and the features that reflect that change. Richmond [3] proposed a need for the creation of a migration multivariate model based on political and other factors that he referred to as "predisposing factors". Richmond's contribution, however, stops at a simple proposal of what could comprise such a migration model. Schmeidl [4] attempted to outline some root causes for forced migration and create a model analyzing their predictive power. Schmeidl used the term "facilitator" variables that she stated serve to "increase the likelihood of refugee flight", and the term "obstacles" that she stated serve to decrease the likelihood of refugee flight. The author concluded, however, that her method was only able to model small-scale refugee counts based only on a small set of political violence features and was unable to account for extreme refugee counts. She noted that future work needs to include a much broader set of proposed predictive features. Arango [5] discussed human migration from a high-level comparative theory point of view and pointed out that a robust and efficient refugee prediction model was still lacking. The author stated that this lack is most likely due to the continual change in the nature of migration and its complexity, and that researchers must try to account for this change and complexity.

Davenport et al. [6] claimed that refugees flee their home countries primarily due to "threats to personal integrity". Their work suffered from a small sample and from high levels of standard error for all proposed predictive variables except for "net migration" and "polity change". Although no model was proposed, Larkin [7] briefly discussed how North Korean refugees fleeing to China predominantly stated in survey responses that their reason for fleeing the home country was due to the high inflation rates. Bayesian methods for forecasting were used by Bijak [8] using a case study of migration between Germany and Poland based on the few economic predictive features of population, gross domestic product (GDP), and unemployment rate. Bijak's work suffered, however, from the limitations of the narrow nature of both the proposed predictive features and the location characteristics. Using Haiti as a case study, Shellman and Stewart [9] analyzed forced migration based on "civil violence, poor economic conditions and foreign interventions". The work suffered, however, from a lack of ground-truth data as well as the narrow nature of the proposed predictive features. Martineau

[10] attempted to develop a model based on the economic features of GDP, purchasing power parity, population, population density, and various sociocultural measures of a country's independence. Martineau stated that his model accurately predicted refugee counts based most strongly on the independence feature. The independence feature, however, was feature-engineered based on an arbitrary process, rather than on some ground-truth publicly available metric, which introduces limitations relative to reproducibility and scalability. Black et al. [11] proposed a basic framework for considering how to incorporate features related to environmental change into the discussion of what drives human migration. Although no refugee count prediction model was presented, the work contributed a valuable discussion of what the authors argue are the "five drivers of migration": economic, political, demographic, social, and environmental.

Using Venezuela as a case study, social media data was analyzed by Mead et al. [12] to determine whether social media platforms such as blogs were being used by citizens to discuss motivational factors for migration. The authors showed that keyword trend lines revealed that the online discussion did contain numerous "quality of life" features as being important to an individual's reason for fleeing Venezuela, including "need food", "need water", "need medicine", "high prices", and "inflation". Other researchers have added to the literature pool with their discussions of more varied techniques to try to predict refugee counts, such as in-person monitoring surveys [13], agent-based models [14], [15], and epidemiological models [16]. Unfortunately, these works either suffer from limitations regarding data sparsity, outliers, and a lack of ground truth data, or require extensive data collection based on both human surveys and geospatial data. These authors emphasize that future approaches to refugee count prediction must include a variety of features and techniques, and that it is imperative that researchers work closely with humanitarian organizations [17]. Previous work by Mead et al. [2] set the stage for developing a proactive approach to refugee crisis management by identifying some of the numerous motivational factors that lead to refugee migration and attempting to use those variables to build a viable machining learning model to predict refugee crisis situations. Although their work showed promise for the potential predictive power of numerous sociocultural, socioeconomic, and economic features, the authors note the main limitations of their model, which were related to the small size of the datasets used in their experiments and asserted the need for continued data collection and model finetuning and evaluation.

Methodology

This section provides the methodological details of our data collection and data preprocessing steps and the nature of the experiments. We also describe the contents of our dataset, giving visual representations of the location components and the numeric predictive feature set.

We created a broad dataset that includes 28 proposed predictive features regarding the prediction of refugee counts. The ground truth for this data comes from the UNHCR [18] and The World Bank [19] for the target feature of refugee count by country of origin for a given year. The ground truth for the 28 predictive features were collected from various sources. Each of the 28 predictive features are recorded metrics for a specific country for a specific year. A complete list of data sources is available upon request. Three of the features in the dataset created for this work are location-specific: country, region, and subregion. The region and subregion categories are according to the methodology used by the Statistics

Division of the United Nations [20]. A total of 138 countries in 5 regions and 16 subregions were included in this analysis. Additional discussion of our analysis of the location components is provided in Section 4. A complete region and subregion list is available upon request. Table 1 provides a list of the numeric predictive features used in our experiments. (*Note: Corruption perception index (cpi) and Corruption (c) are two independent measures.)

TABLE I
NUMERIC PREDICTIVE FEATURES SET OF THE DATASET USED IN THE EXPERIMENTS

Code	Feature	Type
gpi	global peace index	Sociocultural
cpi	corruption perception index	Sociocultural
hr	homicide rate	Sociocultural
sdfc	satisfied demand for female contraception	Sociocultural
em	early marriage	Sociocultural
c	corruption*	Sociocultural
dvam	domestic violence against minorities	Sociocultural
mc	media censorship	Sociocultural
fe	freedom of expression	Sociocultural
fr	freedom of religion	Sociocultural
aj	access to justice	Sociocultural
wpr	women's property rights	Sociocultural
ppge	political power gender equality	Sociocultural
bw	access to basic water	Socioeconomic
bs	access to basic sanitation	Socioeconomic
iu	internet users	Socioeconomic
e	access to electricity	Socioeconomic
mp	mobile phone users	Socioeconomic
hc	access to healthcare	Socioeconomic
p	population	Economic
pgr	population growth rate	Economic
pd	population density	Economic
u	unemployment rate	Economic
gdp	gross domestic product	Economic
gdpg	gross domestic product growth rate	Economic
i	inflation rate	Economic
annipc	annual net national income per capita	Economic
le	life expectancy	Economic

Some of the features collected are sociocultural in nature in that they directly reflect the society or cultural aspects within a particular country. The sociocultural features used in these experiments include global peace index, corruption perception index, homicide rate, satisfied demand for female contraception, early marriage rate, corruption, discrimination and violence against minorities, media censorship, freedom of expression, freedom of religion, access to justice, women's property rights, and political power gender equality. Others of the features collected are socioeconomic in nature in that they reflect the aspects of a particular country in terms of monetary requirements or costs to live in that country. The socioeconomic features used in these experiments include access to basic water, access to basic sanitation, internet users, access to electricity, mobile phone users, and access to healthcare. And some

other features collected are economic in nature in that they directly reflect specific aspects of a country that are normally used within the field of economic theory. The economic features used in these experiments include population, population growth rate, population density, unemployment rate, gross domestic product (GDP), GDP growth rate, inflation rate, annual net national income per capita, and life expectancy.

Once the data for each feature was collated into one dataset, Python and Pandas were used to conduct data preprocessing steps such as cleaning and standardization. We removed all rows that contained nulls [21]. These rows will be reinstated into the dataset in future work as more data becomes available to the public. Scikit-learn was used to apply a standard scaler to each numeric feature in the dataset [22]. The post-processed dataset consisted of 539 rows for 138 countries for years 2014-2017. The processed dataset was used as the base for two experiments using ordinary least squares (OLS) regression via statsmodel [23]: one on the entire dataset, and the other based on regional subsets of the dataset.

Results and Analysis

Applying This section provides a discussion of our results and analysis. We first provide and discuss the results of the OLS experiment based on the entire dataset. We then provide and discuss the results of the OLS regional experiment.

OLS Experiment Using Entire Dataset

When entering the ground truth metrics for the 28 predictive features of a hold-out row for Afghanistan in 2014, the refugee count prediction using the OLS model was 790,157. The true refugee count for Afghanistan in 2014, however, was 2,596,270 (the ground truth refugee count from our dataset). This is a difference of 1,806,113 refugees. Nonetheless, each of the 28 proposed predictive features included in the model show potential for predictive power. The top five in terms of strength being global peace index (gpi), access to electricity (e), access to basic water (bs), media censorship (mc), and access to healthcare (hc), respectively (Table 2). The complete OLS output for this experiment is contained in the Appendix.

TABLE II
OLS REGRESSION RESULTS: TOP FIVE PREDICTIVE FEATURES IN TERMS OF COEFFICIENT STRENGTH[a]

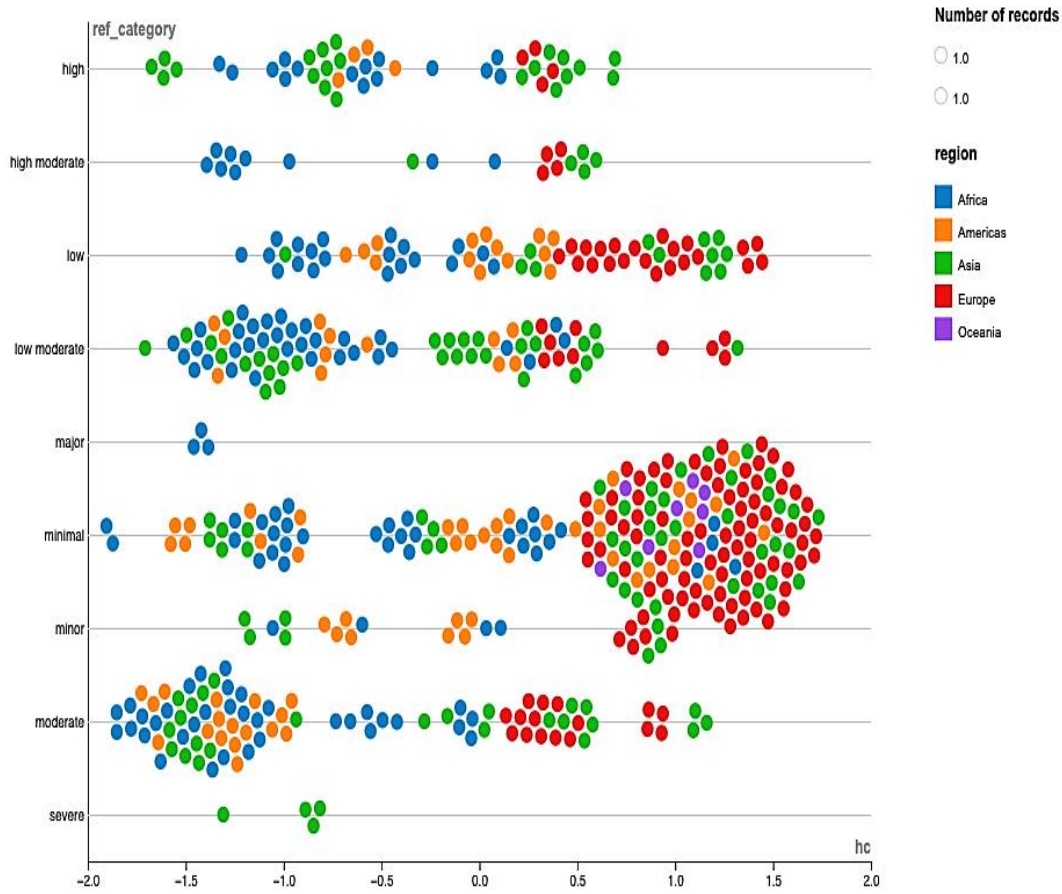
R-squared: 0.362	F-statistic: 10.33	Df Model: 28
Adj. R-squared: 0.327	No. Observation: 538	Covariance Type: nonrobust
feature:	*coefficient:*	*standard error:*
global peace index	0.5433	0.060
access to electricity	0.5369	0.099
access to basic water	-0.4388	0.09
media censorship	0.3630	0.100
access to healthcare	0.3407	0.070

[a]These are excerpts from the OLS regression output when using the entire dataset as the input (numeric features). Metrics are expressed in standard scaler form. The full output is available upon request.

Table 2 also shows the overall predictive performance of the model was quite low. The R-squared value reveals that only 36.2% of the variation in y (refugee count) is explained by our features (X1 through X28). Since the R-squared metric has the limitation of becoming unreliable as more predictor features are added, we use the Adjusted R-squared. The Adj. R-squared value in this experiment reveals that only 32.7% of the variation in y (refugee count) is explained by our features (X1 through X28). The Prob (F-statistic) gives the overall significance of the regression. This is the probability that the null hypothesis is true, which is that all coefficients are equal to zero and therefore have no effect on refugee count. The Prob (F-statistic) value in our experiment reveals that there is almost zero probability that the null hypothesis is true. This means that even though our Adj. R-squared value is low, our regression based on our features is meaningful. Some of the other OLS output metrics, however, reveal that there are some issues with our dataset, such as evidence of skewness and autocorrelation. These components will be discussed further in Section 5.

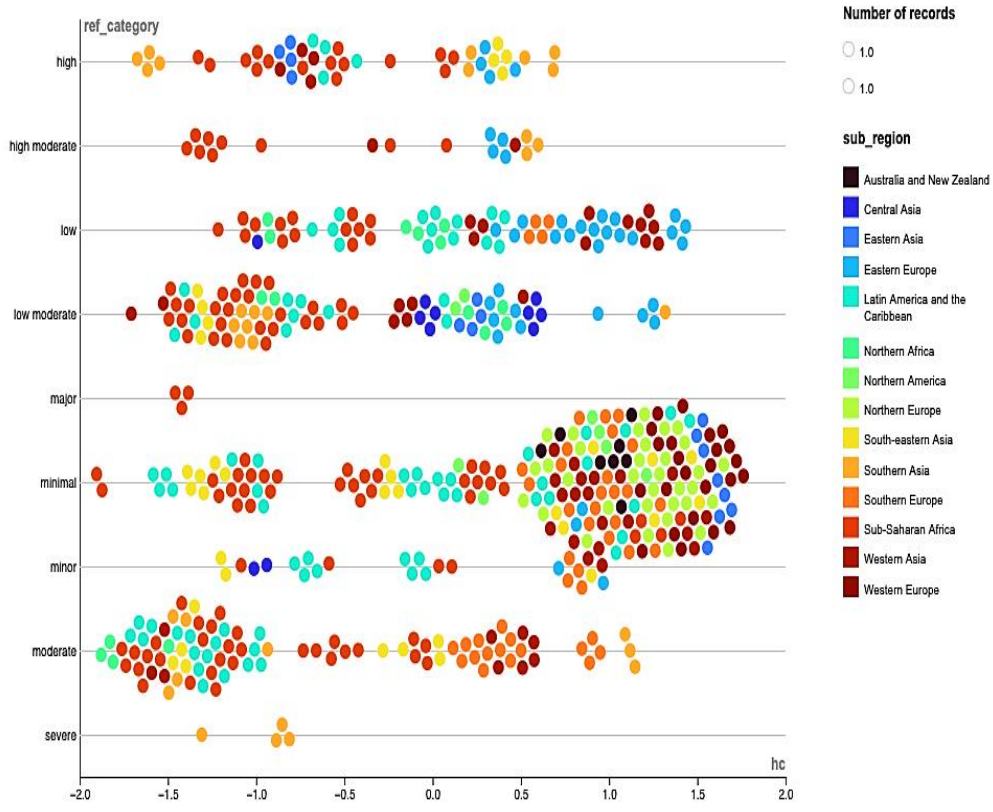
Figures 1 and 2 provide a visual representation of an example location analysis component of the dataset used in our experiments. Additional figures are available upon request. Figure 1 shows the predictive feature of access to healthcare (hc) and refugee category by region.

Figure 1. Access to healthcare and refugee category by region



Refugee category (ref_category) is a feature that we feature-engineered based on the ground-truth data for the target feature of refugee count. Refugee categories fit all the data for refugee count into nine mutually exclusive categories (minimal to severe) based on quantity buckets. Figure 2 shows the predictive feature of access to healthcare (hc) and refugee category by subregion. The next section provides a discussion of the OLS experiment based on region.

Figure 2. Access to healthcare and refugee category by subregion



OLS Experiment Using Regional Subsets of the Dataset

Tables 3 through 6 each provide a summary of the results of our OLS analysis based on region. The regions included in the analysis are Africa, the Americas, Europe, and Asia. The fifth region in our dataset, Oceania, had too few observations to warrant OLS regression. The OLS analysis for the region of Africa (Table 3) revealed that 74.6% of the variation in refugee count was explained by our feature set, and the top three most powerful predictive features in terms of coefficient were population, global peace index, and corruption perception index, respectively.

TABLE III
OLS REGRESSION RESULTS FOR AFRICA: TOP THREE PREDICTIVE FEATURES IN TERMS OF COEFFICIENT STRENGTH

R-squared: 0.793	F-statistic: 16.73	Df Model: 28
Adj. R-squared: 0.746	No. Observation: 151	Covariance Type: nonrobust
feature:	*coefficient:*	*standard error:*
P	0.6625	0.329
gpi	0.5671	0.059
cpi	0.2939	0.196

The OLS analysis for the region of the Americas (Table 4) revealed that 81.2% of the variation in refugee count was explained by our feature set, and the top three most powerful predictive features in terms of coefficient were population, population density, and global peace index, respectively.

TABLE IV
OLS REGRESSION RESULTS FOR THE AMERICAS: TOP THREE PREDICTIVE FEATURES IN TERMS OF COEFFICIENT STRENGTH

R-squared: 0.872	F-statistic: 14.42	Df Model: 28
Adj. R-squared: 0.812	No. Observation: 88	Covariance Type: nonrobust
feature:	*coefficient:*	*standard error:*
p	-0.7016	0.117
pd	-0.4570	0.251
gpi	0.5470	0.071

The OLS analysis for the region of Europe (Table 5) revealed that 88.4% of the variation in refugee count was explained by our feature set, and the top three most powerful predictive features in terms of coefficient were access to justice, access to basic sanitation, and homicide rate, respectively.

TABLE V
OLS REGRESSION RESULTS FOR EUROPE: TOP THREE PREDICTIVE FEATURES IN TERMS OF COEFFICIENT STRENGTH

R-squared: 0.907	F-statistic: 39.18	Df Model: 28
Adj. R-squared: 0.884	No. Observation: 141	Covariance Type: nonrobust
feature:	*coefficient:*	*standard error:*
aj	-0.2710	0.052
bs	0.2290	0.052
hr	-0.1972	0.062

The OLS analysis for the region of Asia (Table 6) revealed that just 59.2% of the variation in refugee count was explained by our feature set, and the top three most powerful predictive features in terms of coefficient were corruption, access to basic sanitation, and access to basic water, respectively.

TABLE VI
OLS REGRESSION RESULTS FOR ASIA: TOP THREE PREDICTIVE FEATURES IN TERMS OF COEFFICIENT STRENGTH

R-squared: 0.670	F-statistic: 8.552	Df Model: 28
Adj. R-squared: 0.592	No. Observation: 147	Covariance Type: nonrobust
feature:	*coefficient:*	*standard error:*
e	1.9275	0.551
bs	-1.7748	0.505
bw	-1.5162	0.387

Conclusion and Future Work

If able to predict, mitigate, and prepare, some refugee situations could be managed effectively so that they do not become a humanitarian crisis. This work provides evidence for numerous location-based features as having the potential for the prediction of refugee counts. The evidence for these predictive features provides a valuable contribution to the overall study of migration motivational factors and to the continued attempt to conduct predictive modeling of refugee counts. Ordinary least squares regression via statsmodel was used in two experiments in this work: one on the entire dataset, and the other on regional subsets of the dataset. Location-based features with potential refugee count predictive power stood out in our results, such as the global peace index, access to electricity, access to basic water, media censorship, and access to healthcare. The proposed model performed better when applied to regional subsets of the dataset. Out of the five regions that we analyzed using OLS regression based on our proposed 28 predictive sociocultural, socioeconomic, and economic features, the model produced the best performance when the region was Europe and access to justice and homicide rate were among the features with the most predictive power. Corruption features stood out as having the most potential for predictive power, however, in both Africa and Asia, while population features were the most dominant in the Americas. Each of these relationships will be further explored in our ongoing and future work. The contributions of this work include an analysis based on region for a broad set of proposed predictive features from a broad set of categories. This work also provides the additional contribution of the proposed predictive features being more representative of the changes that the societal landscape has undergone over the years. The limitations of this work, however, are inherent in the datasets used in the experiments in that they are small in nature. We suspect that this is the reason behind the poor quality of some of the OLS performance and evaluation metrics such as AIC, BIC, Durbin-Watson, and Omnibus. A complete list of these metrics is available upon request. These metrics revealed that our dataset suffers from skewness and that there is also evidence for possible autocorrelation. These issues will be addressed in future work via not only the addition of new data as it becomes available to the public, but also experimentation with data transformations and additional feature-engineering.

Appendix

This appendix includes the full output for the Ordinary Least Squares (OLS) experiment on the raw, standard scaler, and min/max scaler versions of the full dataset (Fig. 3, 4, and 5). Based on the ground truth from the dataset, the metric outputted for "Predicted Number of Refugees" for the Raw, Standard Scaler, and Min/Max Scaler versions of the dataset should have been 2596270, 10.81150162, and 0.973733313, respectively.

Figure 3. Full output for the Ordinary Least Squares (OLS) experiment on the raw version of the full dataset.

Raw:

Predicted Number of Refugees:
[790157.58737093]

OLS Regression Results

Dep. Variable:	ref	R-squared:	0.362
Model:	OLS	Adj. R-squared:	0.327
Method:	Least Squares	F-statistic:	10.33
Date:	Sun, 13 Jun 2021	Prob (F-statistic):	1.64e-34
Time:	04:02:42	Log-Likelihood:	-7231.8
No. Observations:	538	AIC:	1.452e+04
Df Residuals:	509	BIC:	1.465e+04
Df Model:	28		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.549e+05	1.3e+05	2.723	0.007	9.89e+04	6.11e+05
p	-3.587e-05	6.17e-05	-0.581	0.561	-0.000	8.54e-05
pgr	-1.857e+04	1.01e+04	-1.834	0.067	-3.85e+04	1327.055
pd	15.5971	11.954	1.305	0.193	-7.888	39.082
u	-2901.1928	1654.912	-1.753	0.080	-6152.491	350.106
gdp	-3.333e-09	5.17e-09	-0.644	0.520	-1.35e-08	6.83e-09
gdpg	-3950.8102	2492.115	-1.585	0.114	-8846.908	945.288
gpi	2.83e+05	3.12e+04	9.063	0.000	2.22e+05	3.44e+05
i	-2525.7382	1472.698	-1.715	0.087	-5419.053	367.576
annipc	0.5433	1.271	0.427	0.669	-1.953	3.040
cpi	548.3265	3606.583	0.152	0.879	-6537.294	7633.948
bw	-6295.0932	1366.715	-4.606	0.000	-8980.190	-3609.996
bs	-1890.2134	945.864	-1.998	0.046	-3748.492	-31.935
iu	-943.0207	738.458	-1.277	0.202	-2393.821	507.779
e	4447.1043	823.966	5.397	0.000	2828.311	6065.898
hr	-914.8948	907.672	-1.008	0.314	-2698.139	868.349
mp	-737.4794	344.331	-2.142	0.033	-1413.964	-60.995
sdfc	698.1518	592.811	1.178	0.239	-466.505	1862.808
em	-1860.1357	1033.370	-1.800	0.072	-3890.330	170.059
c	-246.3660	3513.622	-0.070	0.944	-7149.353	6656.621
dvam	-2.298e+04	6347.769	-3.621	0.000	-3.55e+04	-1.05e+04
mc	7.739e+04	2.14e+04	3.624	0.000	3.54e+04	1.19e+05
le	-8136.9276	5147.181	-1.581	0.115	-1.82e+04	1975.407
hc	7.52e+04	1.55e+04	4.860	0.000	4.48e+04	1.06e+05
fe	-9.708e+04	1.11e+05	-0.873	0.383	-3.16e+05	1.21e+05
fr	-3.107e+04	1.48e+04	-2.100	0.036	-6.01e+04	-2004.685
aj	-1.193e+05	6.82e+04	-1.748	0.081	-2.53e+05	1.48e+04
wpr	-5.015e+04	1.43e+04	-3.497	0.001	-7.83e+04	-2.2e+04
ppge	271.5926	2.12e+04	0.013	0.990	-4.14e+04	4.19e+04

Omnibus:	742.513	Durbin-Watson:	0.499
Prob(Omnibus):	0.000	Jarque-Bera (JB):	134623.144
Skew:	7.121	Prob(JB):	0.00
Kurtosis:	79.175	Cond. No.	3.59e+13

Figure 4. Full output for the Ordinary Least Squares (OLS) experiment on the standard scaler version of the full dataset

Standard Scaler:
Predicted Number of Refugees:
[3.14233176]

OLS Regression Results

Dep. Variable:	ref	R-squared:	0.362
Model:	OLS	Adj. R-squared:	0.327
Method:	Least Squares	F-statistic:	10.33
Date:	Sun, 13 Jun 2021	Prob (F-statistic):	1.64e-34
Time:	04:02:52	Log-Likelihood:	-576.97
No. Observations:	538	AIC:	1212.
Df Residuals:	509	BIC:	1336.
Df Model:	28		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0142	0.031	-0.454	0.650	-0.076	0.047
p	-0.0254	0.044	-0.581	0.561	-0.111	0.061
pgr	-0.0981	0.054	-1.834	0.067	-0.203	0.007
pd	0.0460	0.035	1.305	0.193	-0.023	0.115
u	-0.0700	0.040	-1.753	0.080	-0.148	0.008
gdp	-0.0275	0.043	-0.644	0.520	-0.111	0.056
gdpg	-0.0565	0.036	-1.585	0.114	-0.127	0.014
gpi	**0.5433**	0.060	9.063	0.000	0.426	0.661
i	-0.0621	0.036	-1.715	0.087	-0.133	0.009
annipc	0.0343	0.080	0.427	0.669	-0.123	0.192
cpi	0.0440	0.290	0.152	0.879	-0.525	0.613
bw	**-0.4388**	0.095	-4.606	0.000	-0.626	-0.252
bs	-0.2323	0.116	-1.998	0.046	-0.461	-0.004
iu	-0.1156	0.090	-1.277	0.202	-0.293	0.062
e	**0.5369**	0.099	5.397	0.000	0.341	0.732
hr	-0.0419	0.042	-1.008	0.314	-0.123	0.040
mp	-0.1089	0.051	-2.142	0.033	-0.209	-0.009
sdfc	0.0572	0.049	1.178	0.239	-0.038	0.153
em	-0.0949	0.053	-1.800	0.072	-0.198	0.009
c	-0.0200	0.285	-0.070	0.944	-0.579	0.539
dvam	-0.1952	0.054	-3.621	0.000	-0.301	-0.089
mc	**0.3630**	0.100	3.624	0.000	0.166	0.560
le	-0.1060	0.067	-1.581	0.115	-0.238	0.026
hc	**0.3407**	0.070	4.860	0.000	0.203	0.478
fe	-0.1026	0.117	-0.873	0.383	-0.333	0.128
fr	-0.1103	0.053	-2.100	0.036	-0.213	-0.007
aj	-0.1226	0.070	-1.748	0.081	-0.260	0.015
wpr	-0.1825	0.052	-3.497	0.001	-0.285	-0.080
ppge	0.0007	0.058	0.013	0.990	-0.113	0.115

Omnibus:	742.513	Durbin-Watson:	0.499
Prob(Omnibus):	0.000	Jarque-Bera (JB):	134623.138
Skew:	7.121	Prob(JB):	0.00
Kurtosis:	79.175	Cond. No.	42.2

Figure 5. Full output for the Ordinary Least Squares (OLS) experiment on the min/max scaler version of the full dataset

Min/Max Scaler:
Predicted Number of Refugees:
[0.29631352]

OLS Regression Results

Dep. Variable:	ref	R-squared:	0.362
Model:	OLS	Adj. R-squared:	0.327
Method:	Least Squares	F-statistic:	10.33
Date:	Sun, 13 Jun 2021	Prob (F-statistic):	1.64e-34
Time:	04:03:01	Log-Likelihood:	728.58
No. Observations:	538	AIC:	-1399.
Df Residuals:	509	BIC:	-1275.
Df Model:	28		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1641	0.048	3.403	0.001	0.069	0.259
p	-0.0186	0.032	-0.581	0.561	-0.082	0.044
pgr	-0.0589	0.032	-1.834	0.067	-0.122	0.004
pd	0.0463	0.035	1.305	0.193	-0.023	0.116
u	-0.0303	0.017	-1.753	0.080	-0.064	0.004
gdp	-0.0244	0.038	-0.644	0.520	-0.099	0.050
gdpg	-0.0757	0.048	-1.585	0.114	-0.170	0.018
gpi	0.2633	0.029	9.063	0.000	0.206	0.320
i	-0.0624	0.036	-1.715	0.087	-0.134	0.009
annipc	0.0159	0.037	0.427	0.669	-0.057	0.089
cpi	0.0167	0.110	0.152	0.879	-0.199	0.232
bw	-0.1521	0.033	-4.606	0.000	-0.217	-0.087
bs	-0.0665	0.033	-1.998	0.046	-0.132	-0.001
iu	-0.0339	0.027	-1.277	0.202	-0.086	0.018
e	0.1559	0.029	5.397	0.000	0.099	0.213
hr	-0.0291	0.029	-1.008	0.314	-0.086	0.028
mp	-0.0494	0.023	-2.142	0.033	-0.095	-0.004
sdfc	0.0221	0.019	1.178	0.239	-0.015	0.059
em	-0.0426	0.024	-1.800	0.072	-0.089	0.004
c	-0.0078	0.111	-0.070	0.944	-0.225	0.210
dvam	-0.0776	0.021	-3.621	0.000	-0.120	-0.035
mc	0.1125	0.031	3.624	0.000	0.052	0.173
le	-0.0422	0.027	-1.581	0.115	-0.095	0.010
hc	0.1029	0.021	4.860	0.000	0.061	0.145
fe	-0.0343	0.039	-0.873	0.383	-0.111	0.043
fr	-0.0434	0.021	-2.100	0.036	-0.084	-0.003
aj	-0.0404	0.023	-1.748	0.081	-0.086	0.005
wpr	-0.0705	0.020	-3.497	0.001	-0.110	-0.031
ppge	0.0003	0.027	0.013	0.990	-0.052	0.053

Omnibus:	742.513	Durbin-Watson:	0.499
Prob(Omnibus):	0.000	Jarque-Bera (JB):	134623.143
Skew:	7.121	Prob(JB):	0.00
Kurtosis:	79.175	Cond. No.	166.

Acknowledgment

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

UNHCR, "Global Trends: Forced displacement in 2019," 2019. https://www.unhcr.org/5ee200e37.pdf (accessed Jun. 15, 2021).

E. Mead, M. Maleki, R. Erol, and N. Agarwal, "Proposing a Broader Scope of Predictive Features for Modeling Refugee Counts," in Companion Proceedings of the Web Conference 2021, Apr. 2021, pp. 154–160, doi: 10.1145/3442442.3453457.

A. H. RICHMOND, "Reactive Migration: Sociological Perspectives On Refugee Movements," J. Refug. Stud., vol. 6, no. 1, pp. 7–24, 1993, doi: 10.1093/jrs/6.1.7.

S. Schmeidl, "Exploring the Causes of Forced Migration: A Pooled Time-Series Analysis, 1971-1990," Soc. Sci. Q., vol. 78, no. 2, pp. 284–308, Jun. 1997, [Online]. Available: http://www.jstor.org/stable/42864338.

J. Arango, "Explaining Migration: A Critical View," Int. Soc. Sci. J., vol. 52, no. 165, pp. 283–296, Sep. 2000, doi: 10.1111/1468-2451.00259.

C. Davenport, W. Moore, and S. Poe, "Sometimes You Just Have to Leave: Domestic Threats and Forced Migration, 1964-1989," Int. Interact., vol. 29, no. 1, pp. 27–55, Jan. 2003, doi: 10.1080/03050620304597.

J. Larkin, "Why refugees flee," Far East. Econ. Rev., vol. 166, no. 9, p. 14, 2003.

J. Bijak, "Bayesian methods in international migration forecasting," 2005.

S. M. Shellman and B. M. Stewart, "Predicting Risk Factors Associated with Forced Migration: An Early Warning Model of Haitian Flight," Civ. Wars, vol. 9, no. 2, pp. 174–199, Jun. 2007, doi: 10.1080/13698240701207344.

J. S. Martineau, "Red Flags: A Model for the Early Warning of Refugee Outflows," J. Immigr. Refug. Stud., vol. 8, no. 2, pp. 135–157, May 2010, doi: 10.1080/15562941003792093.

R. Black, W. N. Adger, N. W. Arnell, S. Dercon, A. Geddes, and D. Thomas, "The effect of environmental change on human migration," Glob. Environ. Chang., vol. 21, pp. S3–S11, Dec. 2011, doi: 10.1016/j.gloenvcha.2011.10.001.

E. L. Mead, M. N. Hussain, M. Nooman, S. Al-khateeb, and N. Agarwal, "Assessing situation awareness through blogosphere: a case study on Venezuelan socio-political crisis and the migrant influx," in The Seventh International Conference on Social Media Technologies, Communication, and Informatics (SOTICS 2017), 2017, pp. 22–29.

R. Nair et al., "A machine learning approach to scenario analysis and forecasting of mixed migration," IBM J. Res. Dev., vol. 64, no. 1/2, pp. 7:1-7:7, Jan. 2020, doi: 10.1147/JRD.2019.2948824.

G. A. Hébert, L. Perez, and S. Harati, "An agent-based model to identify migration pathways of refugees: the case of Syria," in Agent-Based Models and Complexity Science in the Age of Geospatial Big Data, Springer, 2018, pp. 45–58.

C. Vanhille Campos, D. Suleimenova, and D. Groen, "A Coupled Food Security and Refugee Movement Model for the South Sudan Conflict," 2019, pp. 725–732.

V. Ninković and Z. Keković, "DYNAMIC MIGRATION FLOW MODELLING.," Secur. Dialogues, vol. 8, no. 2, 2017.

M. Carammia and J.-C. Dumont, "Can we anticipate future migration flows," Migr. Policy Debates, OECD/EASO, no. 16, 2018.

United Nations High Commissioner for Refugees (UNHCR), "UNHCR - Refugee Data Finder." https://www.unhcr.org/refugee-statistics/ (accessed Jun. 16, 2021).

The World Bank, "Refugee population by country or territory of origin." https://data.worldbank.org/indicator/SM.POP.REFG.OR (accessed Jun. 16, 2021).

"UNSD — Methodology." https://unstats.un.org/unsd/methodology/m49/ (accessed Jun. 16, 2021).

Analytics India Mag, "5 Ways To Handle Missing Values In Machine Learning Datasets." https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/(accessed Jun. 16, 2021).

Scikit-learn.org, "sklearn.preprocessing.StandardScaler." https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (accessed Jun. 16, 2021).

Statsmodel.org, "statsmodel.regression.linear_model.OLS." https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html (accessed Jun. 16, 2021