Refugee Resettlement in the United States

By: Saad Monem and Patrick Thornton

GitHub Repository: https://github.com/smonem/Capstone-Project-Haven.git

I. Introduction

The United States has taken in over 3 million refugees since the signing of the Refugee Act in 1980 [1][2]. Despite this long history, much of the contemporary discussion around refugees has become increasingly politicized. As the wealthiest nation in the world, we firmly believe that the US still has a responsibility to take in refugees, especially as the majority today end up in less affluent nations such as Turkey and Jordan [3]. Yet, there exists a gap in formal research on the lives of refugees after they resettle in the US. The subject of our report is studying refugee resettlement and its outcomes within the United States, with a focus on understanding what factors play a key role in the "success" of a refugee's resettlement.

Refugee resettlement in the US is a complex process; if there is no existing family then refugees are sponsored by non-profit resettlement agencies that attempt to relocate them to suitable locations [4]. The resettlement agency provides services for several months while the Department of State grants a one-time cash assistance [4]. As there is a lack of comprehensive research on the effectiveness of the process, our project seeks insight through an analysis of data gathered from the Annual Survey of Refugees (ASR). This survey is conducted by the Office of Refugee Resettlement, to understand how well a resettlement is going through a slate of questions about the refugee's current life [5]. Other research such as Naseh et al.'s study on poverty among US refugees has also been conducted using this survey, although our analysis aims to be more holistic than prior studies [6].

Our topic directly deals with refugees, a very vulnerable population, and thus we have approached our project with a heightened awareness of ethical issues. As our work could have a very direct impact on their lives, such as by influencing how refugees are resettled or perceived by others, we acknowledge that we have a heightened responsibility to avoid producing biased models or inaccurate conclusions that could lead to negative impacts.

II. Data Source

Data gathering on refugees in the US is not widely conducted nor comprehensive. The Annual Survey of Refugees represents the largest official effort on refugee data-gathering, interviewing a random sample of approximately 1,500 resettled households each year. The questionnaire consists of almost 40 primary questions, which typically branch into a series of sub-questions as well. We combined multiple years of data, ranging from 2016 to 2019, resulting in a total of over 6,000 households.

Due to the nature of the data, which involved a wide set of features and a questionnaire format, in which questions may not be asked depending on previous answers, we were required to perform a considerable amount of data preprocessing. Using the surveys from four separate years meant we needed to append the different year's date. Some years had questions to remove. For a consistent feature space, we took the union of the column names in each year, and dropped any columns that didn't exist across all years. From here the data was filtered to only include the primary respondents (head of household) and unnecessary weighting pre-included with the ASR was removed. The columns of sub-questions that were not useful to our project were then dropped. Afterwards a significant amount of manual data cleaning and imputation was done to prepare the data and fill in empties with logically appropriate answers wherever possible. In the cases where data could not be imputed, these respondents were dropped. Once data preprocessing was completed, we performed several additional steps such as creating the target variable (as described below in the methodology section), splitting features based on whether they were known pre or post resettlement, and performing a standard train-test-validation split.

III. Methodology

As there is no specific, existing criteria on how "successful" a resettlement is, we sought to produce our own target variable constructed from factors in the ASR data. The main factors that we consider include: is the person employed (if they are in the labor force)? Do they make an income above the US poverty line for their given household size? And are they a permanent resident of the US or applying to be one? The criteria that we chose was based on simplicity and ease of understanding; we mainly are looking at whether the refugee is able to support themselves and their family (if eligible to work), and if they are planning for legal and permanent integration into the US. Our criteria also does not penalize anyone who is ineligible to work, such as if they are students, disabled, elderly, or have familial responsibilities precluding them from employment. We examined cases where people exited the labor force due to reasons such as limited understanding of English or other difficulties finding employment, yet the number of these cases was small and we accounted for them.

Our target attempts to be an objective as possible measure of how successful a resettlement was, yet we acknowledge that this variable leaves out various subjective factors that are outside the scope of this project. We would like to clarify that this metric is not intended to judge any individual on a personal level, but is only constructed to help determine what factors influence successful resettlement and integration.

After assigning the target labels, we applied two standard machine learning approaches using unsupervised and supervised ML In the unsupervised portion we experimented with several dimensionality reduction methods and settled on using UMAP to produce a 2D representation of the data, then applied this along with clustering to identify key groups of our population. Our goal was to find and define groups with common traits, so as to understand what indicators were most important, and if a tailored resettlement experience could be effective.

For our supervised learning model, we tested several binary classification models implemented using Scikit-Learn. Given the goals of our project, interpretability was a major concern and thus we focused on models that offer high interpretability, i.e. non-deep learning models such as random forest and logistic regression. Six different types of models were tested and compared using cross-fold validation, with grid search being used to tune these models' hyperparameters. The selection of the final model type was determined based on the overall performance and explainability of the model, given the project's goal of finding trends within the data that can be used to improve the resettlement experience.

IV. Results

Unsupervised ML:

The focus of our unsupervised learning efforts were on differentiating between various types of refugees in the data, with the goal being to understand if we could potentially identify that a person coming into the US belongs to a more at-risk group which needs additional support or resources. To begin with, the data was filtered to only features known pre-resettlement and processing steps identical to those described in the supervised learning section were performed. UMAP was then applied to reduce the data to a 2D space, with the hyperparameters being visually tuned via examining a grid of different representations. Hyperparameters of *n_neighbors*: 100 and *min_dist*: 0 was found to produce visually well-separated clusters, HDBSCAN, a robust clustering method that works well with complex data, was then applied to produce the final representation.

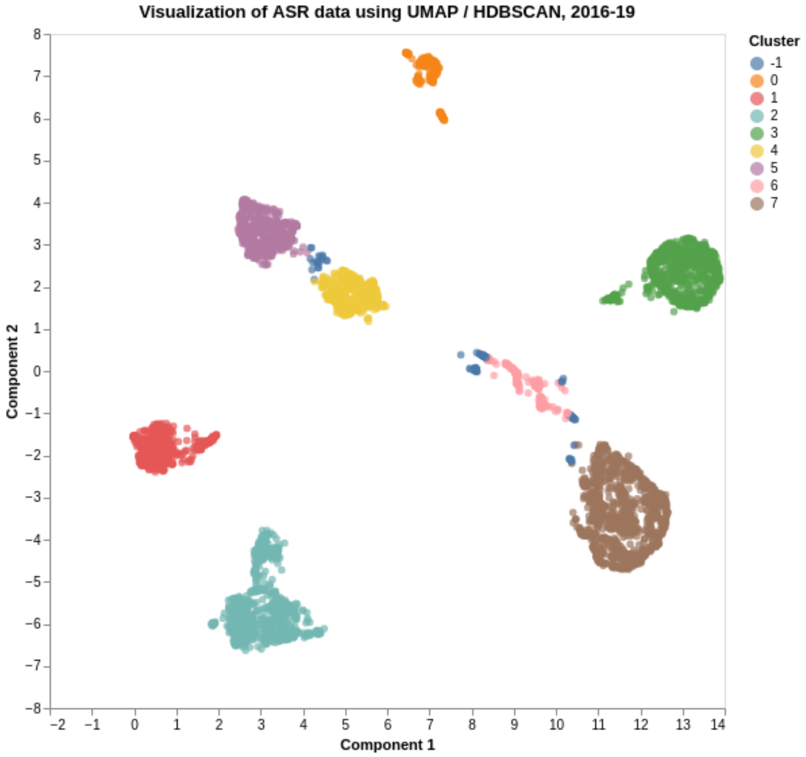


Figure 1: Visualization of clusters found in ASR data using pre-resettlement features

4.1.1 Cluster Interpretation

As shown above, we identified 8 distinct clusters among respondents. Mean resettlement success rate was found to be in a large range, from ~53% in the least successful cluster to ~90% in the most. Averages of the numerical features for each cluster are shown in the table below:

Cluster	*Size*	*Household Size*	*Age*	*Gender (M:0/F:1)*	*Years of Education*	*English Fluency*	*English Instruction (N:0/Y:1)*	*Year Entered US*	***Resettlement Success Rate***
0	269	4.35	41.95	0.06	7.92	0.22	0.09	2016.05	0.54
5	464	3.44	43.21	1	4.21	0.28	0	2014.94	0.61
4	556	3.4	40.6	1	12.97	0.85	0	2015.06	0.63
1	850	3.57	42.4	0.01	5.9	0.42	0	2014.48	0.67
2	565	3.32	35.55	1	11.93	1.03	1	2015.05	0.67
7	1158	3.36	40.28	0	13.42	1.2	0	2014.27	0.73
3	972	2.97	37.33	0	11.87	1.2	1	2014.65	0.76
6	227	2.21	36.81	0.14	7.49	0.47	0.05	2013.74	0.9

Table 1: Cluster averages for pre-resettlement features

At a glance, there appears to be a strong correlation between household size, education, English fluency, and resettlement success. Barring the last cluster, which will be explained more below, the most successful group on average were educated male refugees with small households and decent prior fluency in English. The least successful clusters, in comparison, had the largest households, fewest years of education, and the least proficiency in English. Prior English language training seemed to consistently boost average success by a few percentage points when comparing clusters, even when the clusters without English instruction had more years of education compared to their peers. We can also see when comparing groups such as 2 (all English-trained women) and 3 (all English-trained men) that women tend to do noticeably worse when compared to their male peers.

Cluster	*County of Citizenship*	*Ethnicity*	*Highest Level of Education*	*Prior Occupation*	***Resettlement Success Rate***
0	syria (98.88%)	arab (87.73%)	primary (71.38%)	self-employed (62.08%)	0.54
5	somalia (20.04%)	other (55.17%)	none (64.01%)	not employed (64.87%)	0.61
4	iraq (36.69%)	other (37.77%)	secondary (49.82%)	not employed (35.97%)	0.63
1	iraq (36.82%)	other (38.35%)	primary (53.65%)	self-employed (38.82%)	0.67
2	other (16.99%)	other (54.34%)	Secondary (40.53%)	not employed (27.79%)	0.67
7	iraq (58.64%)	arab (45.68%)	secondary (47.67%)	self-employed (29.71%)	0.73
3	iraq (23.56%)	other (51.85%)	secondary ((40.84%)	self-employed (34.98%)	0.76
6	burma (79.30%)	chin (79.30%)	none (35.24%)	employed (43.17%)	0.9

Table 2: Cluster modes for pre-resettlement categorical features, with percentages

When examining the categorical pre-resettlement features in terms of modes, we see similar trends in terms of education (with the exception of group 6). The most successful groups tended to have at least a secondary education prior to migrating, while the least successful were most likely to have only a primary or no education. This information also helps us understand more about the last cluster which somewhat does not follow the trends of the rest: they appear to be a small group from before the current wave of migration and consist mostly of people from Burma.

Their high success rate could be explained as due to having more time to integrate into the country compared to other groups or the presence of an existing community which aided integration. As we can see, the rest of the clusters found generally do not trend strongly along country or ethnic lines with the exception of cluster 0 (almost exclusively Syrians).

Altogether, these trends provide strong evidence on the importance of education as a key factor in resettlement success as well as English fluency. Groups which did have prior English instruction consistently outperformed similar groups which did not by ~4 percentage points. This suggests that current training programs may not be adequately filling the language gap and more resources could be devoted to those who did not have formal English education prior to resettlement. The negative association between household size and resettlement success also suggests that more may need to be done for larger refugee families resettling in the US.

4.1.2 Post-Resettlement English fluency for clusters

Cluster	*English Fluency (Pre)*	*English Fluency (Post)*	*Attended Training in Past Year*	*Attending Training Now*	***Resettlement Success Rate***
0	0.22	1.1	0.45	0.23	0.54
5	0.28	0.88	0.34	0.21	0.61
4	0.85	1.58	0.37	0.2	0.63
1	0.42	1.09	0.24	0.1	0.67
2	1.03	1.73	0.39	0.21	0.67
7	1.2	1.97	0.25	0.1	0.73
3	1.2	1.91	0.33	0.15	0.76
6	0.47	1.15	0.14	0.07	0.9

Table 3: Cluster averages for features related to Eng. fluency and training

Although we see a marked improvement in English in all groups post-resettlement, a few still remain around an average level of 1 (not well). We also see that the majority of refugees post-resettlement have not attended an English training program either now or in the past 12 months, even for groups with lower English fluency such as 0, 4, and 3. It is possible that these programs may not be accessible to many refugees or that they may not have the time or availability to attend. Further investigation into the reasons behind why many refugees do not attend English training even when they could benefit is likely needed, and presents an avenue for future research.

Supervised ML:

Our supervised learning involves three main components: predicting outcomes based on features prior to resettlement, incorporating actionable features post-resettlement, and conducting a brief follow-up analysis.

In the initial phase, a logistic regression model assesses our ability to proactively identify individuals requiring additional support for a successful move. Expanding on this, we integrate features representing actions within the US immigration system or by the respondent. Using a random forest model, we gain insights into recommended actions for families adapting to their new life. Finally, we extract key insights from both models and explore the relationship between features and our target variable.

4.2.1 Proactive resettlement analysis

During the survey respondents answer various questions about their life, including some background questions like country of origin and family size. As the respondents are refugees from the past five years, factors such as family size or marital status are unlikely to have changed significantly. We focused on these specific answers, resulting in 14 features. To reduce further, we computed correlation matrices for both numerical and categorical features.

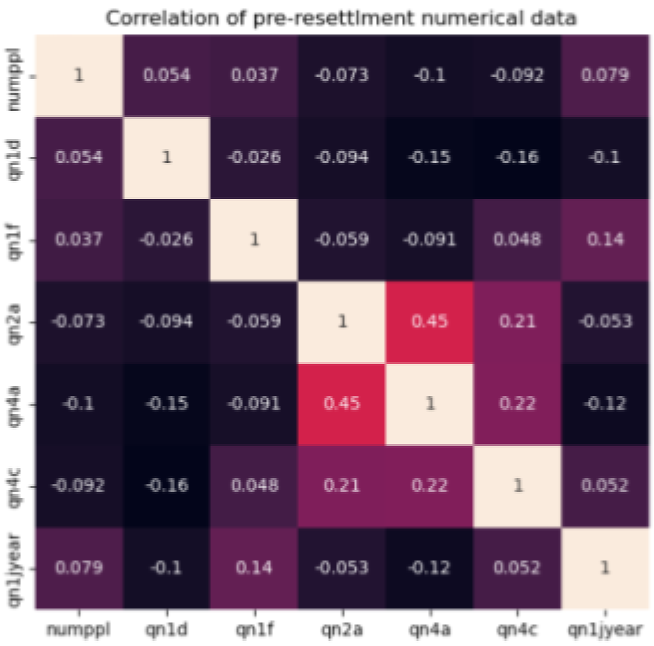


Figure 2: Correlation matrix of proactive analysis' numerical features

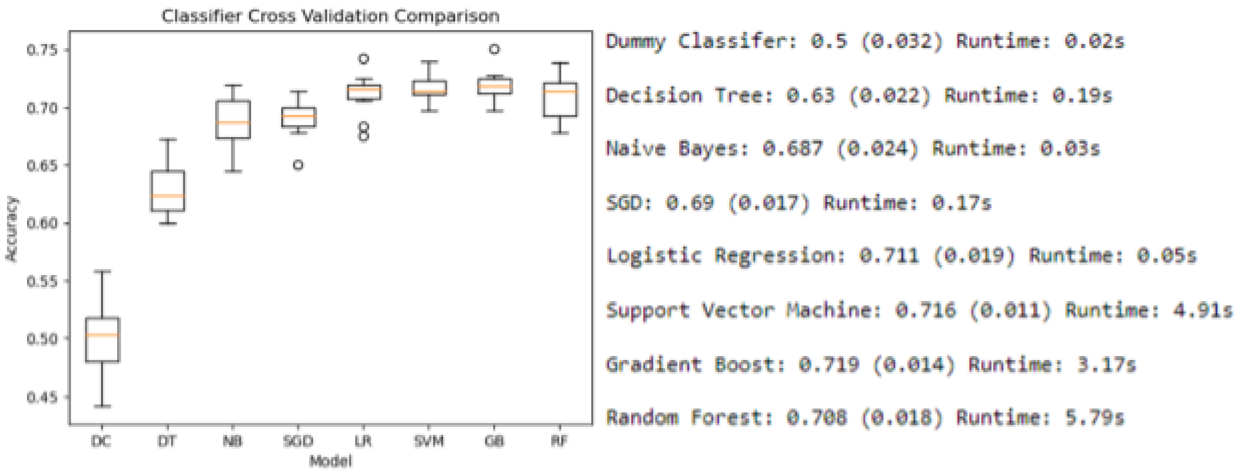
Utilizing Pandas' correlation function and Seaborn's heat matrix, we visualize the relationships between numerical features. The resulting image shows minimal correlation among all features, indicating their suitability for inclusion in our model.

To understand information sharing among nominal columns, we employ Cramer's V, leveraging Pearson's chi-squared statistic. Cramer's V measures the association between features based on the occurrences of observation combinations. To apply this to our variables, we utilized a function that calculates this value, and adds it to a matrix. Country of origin (qn1h) and country of birth (qn1g) exhibit a strong correlation (Cramer's V = 0.96). This comes as no surprise considering the imputation relation, along with a clear intuitive connection.

The level of shared information led us to drop the country of birth value. Additionally, a notable correlation (Cramer's V = 0.56) between country of origin (qn1h) and ethnicity (qn1i) is observed. This value isn't so extreme that it needs to be dropped, and could provide some good insights if important in the model.

Now with 13 variables, we must convert our data into a numerical form. Four features already meet this criteria, while the rest require encoding. For our binary and ordinal columns, we used a label encoder, where each value was mapped to an integer. Features with higher cardinaily couldn't be encoded in this way as the models would interpret their values as being relatively larger or smaller than one another. One-hot encoding (OHE) proved problematic, leading to overly wide datasets and model overfitting. Instead we applied target encoding, replacing each value with the mean of the target variable for that value. This encoding was done after splitting our data into train, test, and validation sets using a 70:15:15 ratio. This approach maintains the relation with the target while keeping the dataset compact. After converting features to numerical types, we normalized them using sklearn's StandardScaler.

To decide on a model, we conducted a test by comparing each model's average accuracy score over five randomly initiated models. Tested models include: Dummy Classifier (DC), Decision Tree (DT), Naive Bayes (NB), Stochastic Gradient Descent (SGD), Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boost (GB), and Random Forest (RF).



```
Dummy Classifer: 0.5 (0.032) Runtime: 0.02s

Decision Tree: 0.63 (0.022) Runtime: 0.19s

Naive Bayes: 0.687 (0.024) Runtime: 0.03s

SGD: 0.69 (0.017) Runtime: 0.17s

Logistic Regression: 0.711 (0.019) Runtime: 0.05s

Support Vector Machine: 0.716 (0.011) Runtime: 4.91s

Gradient Boost: 0.719 (0.014) Runtime: 3.17s

Random Forest: 0.708 (0.018) Runtime: 5.79s
```

Figure 3: Accuracy comparison between model types

Considering the accuracy and runtime of the logistic regression model, it was a clear choice. To tune the model, we conducted a grid search, identifying the hyper-parameter values that most accurately predicted our train data outcome. The final hyper parameter values we used were: 'C': 1, 'penalty': 'l1', 'solver': 'liblinear'. This selection got us to an accuracy score of 0.68 when predicting our test score.

One concern in this project was the size of the dataset, with around 4,000 records, it is small by data science standards. To understand its impact on accuracy, we conducted a sample size test.

Results indicated that accuracy level off around 1,000 samples with some volatility. This volatility decreases with more samples, indicating a larger sample could be beneficial. However these results show sample size isn't a crippling fault.

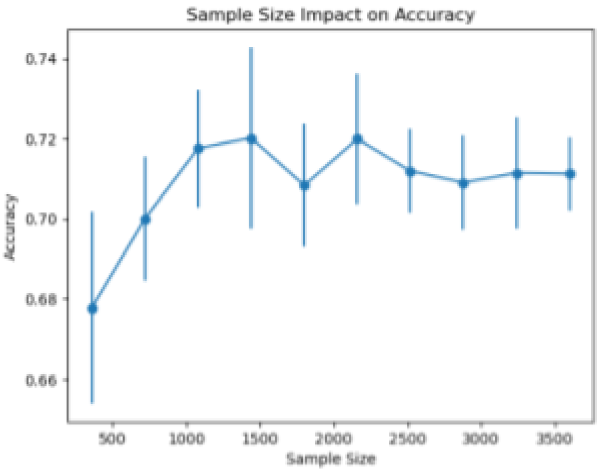


Figure 4: Model Accuracy vs Sample Size

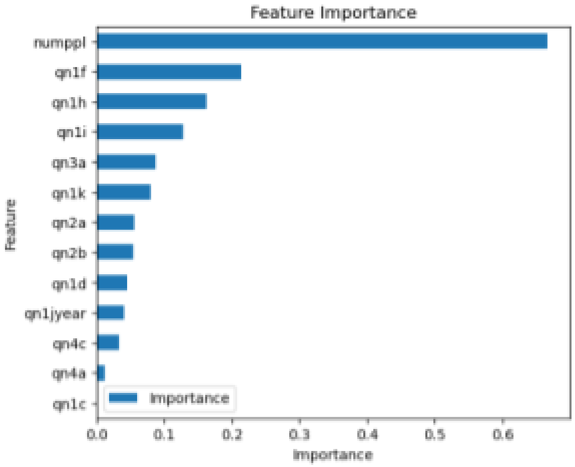


Figure 5: Logistic Regression Feature Importance

To understand feature importance in a logistic regression model, examining correlation coefficients reveals relations between input and predictions.

The biggest factor is family size, followed by gender (qn1f) and country of origin (qn1h). Our model shows that larger family sizes have a negative impact on the success of a resettlement, as does the respondent being a female (qn1f=1). Year of arrival, marital status (qn1c), and English instruction prior to resettling (qn4c) minimally impacts the prediction.

Our proactive resettlement analysis achieved a 20% higher accuracy than what could be done by random guess (represented by the dummy classifier).

4.2.2 Impact of actions taken after resettlement

To include insights from post-resettlement data, we incorporate housing details, in-country movement, English education enrollment, and government assistance into the feature set used in the proactive analysis.

Similar to the previous model, we calculated correlation matrices for the numerical and categorical variables, dropping highly correlated features, which was only the country of birth.

The dataset was split into train, test, and validation sets, and encoded via label and target encoding where applicable.

Instead of using a logistic regression model, we chose a random forest model, given increased categorical variables for better explainability. After applying a grid search, our model achieved an accuracy of 0.75 on our test data while using the hyper-parameter settings: 'criterion': log_loss, 'max_depth': 10, 'max_features': 'log2', 'n_estimators': 50, a 0.02 improvement from the base version, surpassing other models, including logistic regression.

Achieving a score of 0.72 indicates that the new variables improve the accuracy. To understand if the model uses the new features as the cornerstones of its predictions, or as compliments to the pre-resettlement features, we again look at the feature importance of our model.

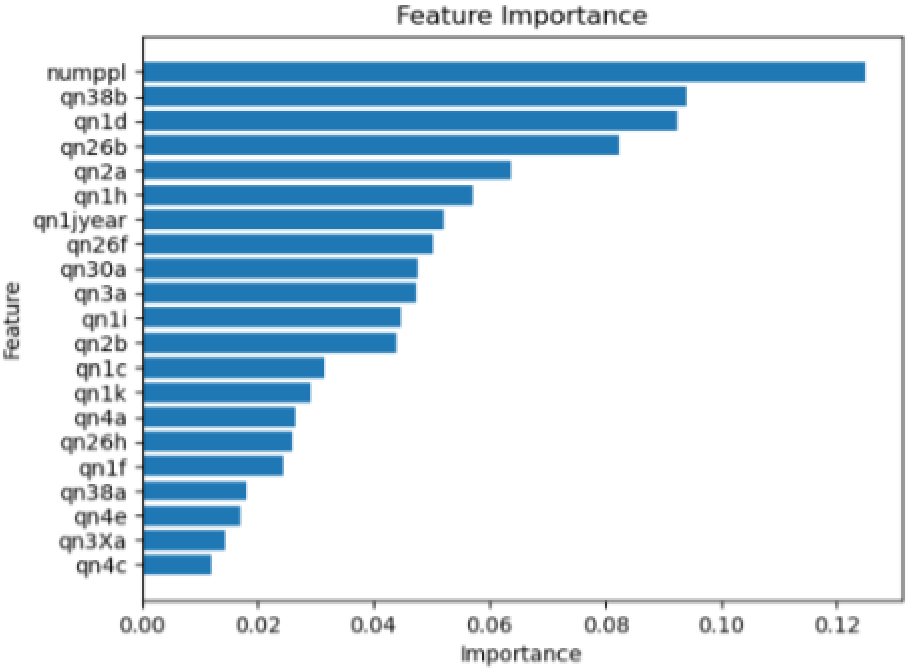


Figure 6: Random Forest Feature Importance

Family size remains important in predicting the success of a resettlement, while other features such as country of origin, ethnicity, and gender play lesser roles. In their place, the model relies on the cost of housing (qn38b), the respondent's age (qn1d), and the length of time spent at their current residence (qn26b). Notably, these new categories have higher cardinality than several lower-ranked features.

To better understand how our model used these features, we can look at one of the trees in the forest. For the visual to be readable, we limit the depth to 3.

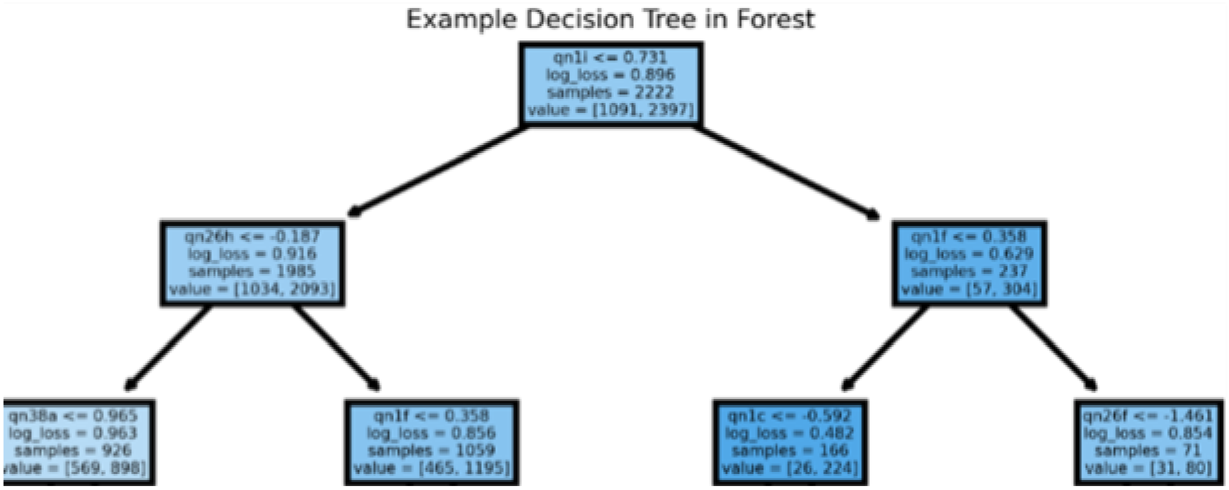


Figure 7: Tree plot of the first tree in random forest model

The visual shows an example of how 1 of the 50 trees in this model lands on a prediction. In the full model, the prediction would be synthesized from the output of the whole forest.

While the most important feature of the model is family size, the tree initially splits on ethnicity (qn1i). During the data preparation phase, we encoded the country names by replacing them with the mean of the target variable for that ethnicity, and normalized it. In this initial split it uses a cut-off value of 0.731, with 1,091 records belonging below that value, and 2,397 above it. Due to our model using bootstrapping, the values won't add up to the sample in each node. Encoded into the visual is also the extremity the node favors the majority class, represented by the color where the redder boxes show a heavier prediction toward t_resettlement=0, and blue equalling 1. Lastly we can see the information gain at each node via the log_loss.

From adding in post resettlement features, we can better identify refugees who need additional assistance. We also learned that money spent on housing and longevity at a residence plays a large role in this prediction.

4.2.3 Follow-Up Analysis

Interpreting our results, we were left with questions about specific features effect on resettlement success that we address through targeted analysis.

Our first focus will be on a person's country of origin, and ethnicity, both playing a large role in our models. We begin by looking at the average rate of successful resettlements based on these. Some subgroups were removed due to insufficient sample sizes.

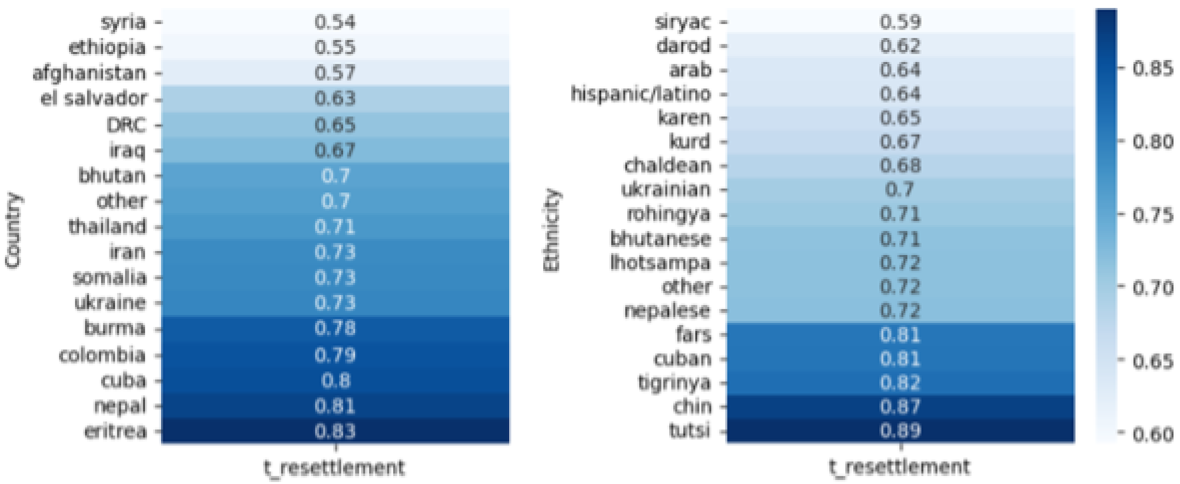


Figure 8: Average resettlement success rate by Country and Ethnicity

One apparent pattern is individuals from the middle east, have a lower rate of success, except for Iran. Surprisingly, Iraq ranks lower despite being the most represented country, challenging expectations of having a larger community. Bhutan and Burma being so different is also surprising given their proximity, and that their refugee crises are similarly recent.

Breaking down the data by country or ethnicity and resettlement regions is stretching an already small dataset very thin. More analysis on these features were done, but at this level of granularity, it is hard to endorse any trend. However, there is potential to investigate this further with more data, along with comparing these results with current data about these populations within the regions.

V. Discussion

5.1 Final thoughts

This experiment proves it’s possible to develop a model using information known at the time of resettlement to predict resettlement success. The model can be enhanced by adding in features post-resettlement such as English class participation or time spent at current residence. Focusing on specific features like country and ethnicity can help better understand what trends we should expect given these priors, but slices at these sizes lead to unreliable conclusions.

5.2 Future work

A common conception is that settled individuals acclimate better to their new location if there is an existing community from a similar background. We were not able to directly test this hypothesis in our analysis due to the anonymized nature of the ASR data, which lacks detailed information on the location of resettlement. One of the goals of future research will be to attempt to assess the validity of this belief, either using data from different sources or through combining complimentary data from censuses conducted at a similar time.

Citations:

National Archives Foundation. (n.d.). Refugee Act of 1980. https://www.archivesfoundation.org/documents/refugee-act-1980/

U.S. Department of State. (n.d.). Refugee Admissions https://www.state.gov/refugee-admissions/

International Rescue Committee. (2023, September 26). Refugee facts, statistics and FAQs. https://www.rescue.org/article/facts-about-refugees-key-facts-faqs-and-statistics

U.S. Department of State. (n.d.). Reception and Placement https://www.state.gov/refugee-admissions/reception-and-placement

The Administration for Children & Families. (2022, November 4). Annual Survey of Refugees. https://www.acf.hhs.gov/orr/programs/refugees/annual-survey-refugees

Naseh, M., Lehman Held, M., Gilbertson, A., & Shrestha, L. (2022). Multidimensional Deprivation amongst Refugees in the USA. *The British Journal of Social Work*, *53*(4), 2120–2139. https://doi.org/10.1093/bjsw/bcac200

Statement of Work:

Patrick	**Saad**
Project Ideation, Data Preparation, Supervised Learning, Report Writing	Unsupervised Learning, Methodology Development, Report Writing

Appendix:

Team 24 Report Appendix